Development of computational methodologies for antibody design

Jinwoo Leem

St Anne's College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Michaelmas 2016

Abstract

Antibodies are proteins of the adapative immune system. Structural diversity in an antibody's two variable domains, V_H and V_L, allow it to bind almost any molecule with high affinity and specificity. This thesis focuses on characterising such variations to develop computational tools for antibody design, with a particular interest toward engineering antibodies as therapeutics.

First, we describe a method to predict the binding affinities of antibody–antigen interactions. Using the contacts at the antibody–antigen interface, we show promising results, but the performance is too poor for extensive design applications. Since several factors can influence antibody binding, we investigate V_H–V_L pairing, one of the largest sources of antibody structural variation. Based on our data, we describe a structure–based mechanism to describe V_H–V_L pairing. In particular, the high conservation of contacts at the V_H–V_L interface in over 6000 antibody sequences provides support for random V_H–V_L pairing.

Following this analysis, we introduce our antibody modelling pipeline, ABodyBuilder. We demonstrate that ABodyBuilder can rapidly build accurate models, and is useful for mapping the antibody structural landscape from sequence. Furthermore, ABodyBuilder calculates the model's expected accuracy in order to help the decision–making process for users during antibody design. To complement ABodyBuilder's current setup, an antibody–specific rotamer library and side chain prediction algorithm are described. Although the maximum achievable accuracy is near 100%, the actual accuracy is closer to 80%, suggesting that the algorithm needs further refinement before full integration into the ABodyBuilder pipeline.

Finally, we discuss how the tools presented in the thesis can be improved, and applied to other problems in computational antibody design. We also present an overview on the potential avenues for expanding the work herein.

Development of computational methodologies for antibody design

Jinwoo Leem

St Anne's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2016

For my father, who has taught me the value of optimism.
For my mother, who has shown me the merits of dedication.
For my brother, who has supported me in all my highs and lows.

Acknowledgements

Dominus illuminato mea. God has been the light of my life at Oxford, and has blessed me with opportunities, mentors, and friends. Without Him, I would not be where I am, and who I am today.

This thesis would not have been possible without my supervisor, Professor Charlotte M. Deane. She has motivated me from day one, giving me great advice and direction; it has been an honour and privilege to be her student. I would also like to thank my industrial supervisors Guy Georges and Jiye Shi for their collaboration and guidance. Finally, I am grateful for the generous funding provided by the EPSRC, Roche, and UCB.

Many thanks go out to fellow ImmunOPIGlets: Konrad, James, Claire (fellow thesis survivor), Cristian, and Jaro – it has been a pleasure to work with you all these last four years. Eleanor, and Hannah, thank you for being crossword buddies, fellow cake eaters, and generally helping me smile more.

Through rowing for three years, I have met amazing individuals from all walks of life. In particular, I would like to thank Anna Kotova, Henrik Hannemann, Alistair Martin, Julian Nowag, Caitlin Armstrong, and members at St Anne's boat club. You have all been a pleasure to row with, and thanks for supporting and motivating me on and off the water. Maybe we'll make a super 2– or 4+ some day and sweep every regatta in the country.

Another shout–out is dedicated to the St Ebbe's community, who have been so warm and understanding in my academic journey. This thesis is dedicated to our 'Thesis' group, with special thanks to Richard Brash, Andi Wang, (Thomas) Austin and Martha Snowbarger for your prayers.

Most importantly, I would like to thank Jae Won Suh. She has been graciously understanding, recognising the demands of my hectic schedule, my unusual sense of humour, and embracing my days of frustration. Undeservedly, she has always responded with care, kindness, and love. Thank you, and I love you.

Abstract

Antibodies are proteins of the adapative immune system. Structural diversity in an antibody's two variable domains, V_H and V_L, allow it to bind almost any molecule with high affinity and specificity. This thesis focuses on characterising such variations to develop computational tools for antibody design, with a particular interest toward engineering antibodies as therapeutics.

First, we describe a method to predict the binding affinities of antibody–antigen interactions. Using the contacts at the antibody–antigen interface, we show promising results, but the performance is too poor for extensive design applications. Since several factors can influence antibody binding, we investigate V_H–V_L pairing, one of the largest sources of antibody structural variation. Based on our data, we describe a structure–based mechanism to describe V_H–V_L pairing. In particular, the high conservation of contacts at the V_H–V_L interface in over 6000 antibody sequences provides support for random V_H–V_L pairing.

Following this analysis, we introduce our antibody modelling pipeline, ABodyBuilder. We demonstrate that ABodyBuilder can rapidly build accurate models, and is useful for mapping the antibody structural landscape from sequence. Furthermore, ABodyBuilder calculates the model's expected accuracy in order to help the decision–making process for users during antibody design. To complement ABodyBuilder's current setup, an antibody–specific rotamer library and side chain prediction algorithm are described. Although the maximum achievable accuracy is near 100%, the actual accuracy is closer to 80%, suggesting that the algorithm needs further refinement before full integration into the ABodyBuilder pipeline.

Finally, we discuss how the tools presented in the thesis can be improved, and applied to other problems in computational antibody design. We also present an overview on the potential avenues for expanding the work herein.

Contents

List of Figures

List of Tables

List of Abbreviations

PDB . . . . . Protein Data Bank

DNA . . . . . Deoxy–ribonucleic acid

mRNA . . . . . Messenger ribonucelic acid

2–D, 3–D . . . . . Two–dimensional, Three–dimensional

NMR . . . . . Nuclear magnetic resonance

EM . . . . . Electron microscopy

kDa . . . . . Kilodalton

Ig . . . . . Immunoglobulin

$\mathbf{V_H}$,$\mathbf{V_L}$ . . . . . Variable heavy, variable light

$\mathbf{C_H}$,$\mathbf{C_L}$ . . . . . Constant heavy, constant light

Fv . . . . . Variable fragment

CDR . . . . . Complementarity–determining region

Fab . . . . . Antigen–binding fragment

Fc . . . . . Crystallisable Fragment

scFv . . . . . Single–chain variable fragment

FR . . . . . Framework region

SLC . . . . . Surrogate light–chain

BCR . . . . . B–cell receptor

RMSD . . . . . Root–mean square deviation

GDT . . . . . Global distance test

NK Cells . . . . . Natural killer cells

ADCC . . . . . Antibody–dependent cellular cytotoxicity

RAG recombinase Recombination–activating gene recombinase

RSS . . . . . Recombination signal sequence

SHM . . . . . Somatic hypermutation

AID . . . . . Activation–induced cytidine deaminase

GC . . . . . Germinal centre

SAbDab, SAbPred Structural antibody database; Structural antibody prediction

PEARS . . . . . Position–dependent antibody rotamer swapper

AMA . . . . . Antibody modelling assessment

PIGS . . . . . Prediction of Immunoglobulin structures

CCG . . . . . Chemical computing group

PSSM . . . . . Position–specific substitution matrix

ESSS . . . . . Environment–specific substitution score

NGS . . . . . Next–generation sequencing

ΔG, $\Delta\Delta$G . . . . . Change in free energy, or change in free energy changes

K_D . . . . . Dissociation constant

PPI . . . . . Protein–protein interaction

MD . . . . . Molecular dynamics

RAPDF . . . . . Residue–specific all–atom probability discriminatory function

CAPTAIN . . . . . Computational Affinity Prediction Tool for Antibody–Antigen Interactions

SASA . . . . . Solvent–accessible surface area

ANARCI . . . . . Antigen receptor numbering and receptor classification

Tm . . . . . Melting temperature

PC, PCA . . . . . Principal component, Principal components analysis

CPU . . . . . Central processing unit

DEE . . . . . Dead–end elimination

GMM . . . . . Gaussian mixture model

AIC . . . . . Akaike Information Criterion

Never say never, because limits, like fears, are often just an illusion.

— Michael Jordan

1 Introduction

Contents

In vertebrate organisms, the adaptive immune system can specifically recognise almost any foreign molecule or pathogen (collectively known as 'antigens'). In particular, a class of antigen–binding proteins, known as antibodies, detect antigens and recruit the immune system for defence against them. The ability of antibodies to bind many different molecules has attracted interest, *e.g.* as a specialised case–study of protein–protein interactions, and as a platform for designing bespoke biotherapeutics.

This thesis concerns the building and refining of computational methodologies for antibody design. The work presented harnesses the wealth of antibody structural data in the Protein Data Bank (PDB; Berman *et al.*, 2000).

This Chapter will provide the context for the thesis. It will start with an introduction to the structure and biological function of proteins. The Chapter

will then cover the structure and biological role of antibodies, followed by an overview of methodologies in experimental and computational antibody design. To finish, the Chapter will provide an outline of the rest of the thesis.

1.1 Proteins: the engines of biology

Proteins are the primary functional units of a cell. The sequence of DNA nucleotides is transcribed to messenger RNA (mRNA); the mRNA is then translated by ribosomes into a chain of amino acids. This chain usually folds into a three–dimensional (3D) protein structure, which determines its function. These functions range from, but not limited to, biochemical catalysis (enzymes), transport (membrane proteins), and molecular recognition (receptors).

1.1.1 Amino acids and protein sequences

Amino acids form the building blocks of a protein. Each block, or 'residue', is composed of four atoms in its 'backbone': nitrogen (N), the alpha carbon (Cα), the carbonyl carbon (C), and the carbonyl oxygen (O). Attached to the Cα is a functional group, or side chain (R) (Figure 1.1A).

Starting from the N–terminal end of the protein (*i.e.* the end that features a free NH_2 group and is first translated by the ribosome), each successive residue is covalently bonded by a peptide bond to the previous residue (Figure 1.1A). Every non–terminal amino acid in a protein has two torsion angles (Figure 1.1B). The ϕ torsion angle refers to the angle about the N–Cα bond, whereas the ψ angle refers to the torsion angle about the Cα–C bond. Each amino acid adopts a series of ϕ/ψ angles, and each amino acid is flexible within steric constraints (Figure 1.1C).

The chemical structure of the side chain affects the conformational freedom of the amino acid. For instance, Gly does not have a side chain, and is thus much more flexible in comparison to Phe, whose side chain has an aromatic phenyl ring. In addition to size, each amino acid side chain has unique chemical features (Appendix Figure A.1).

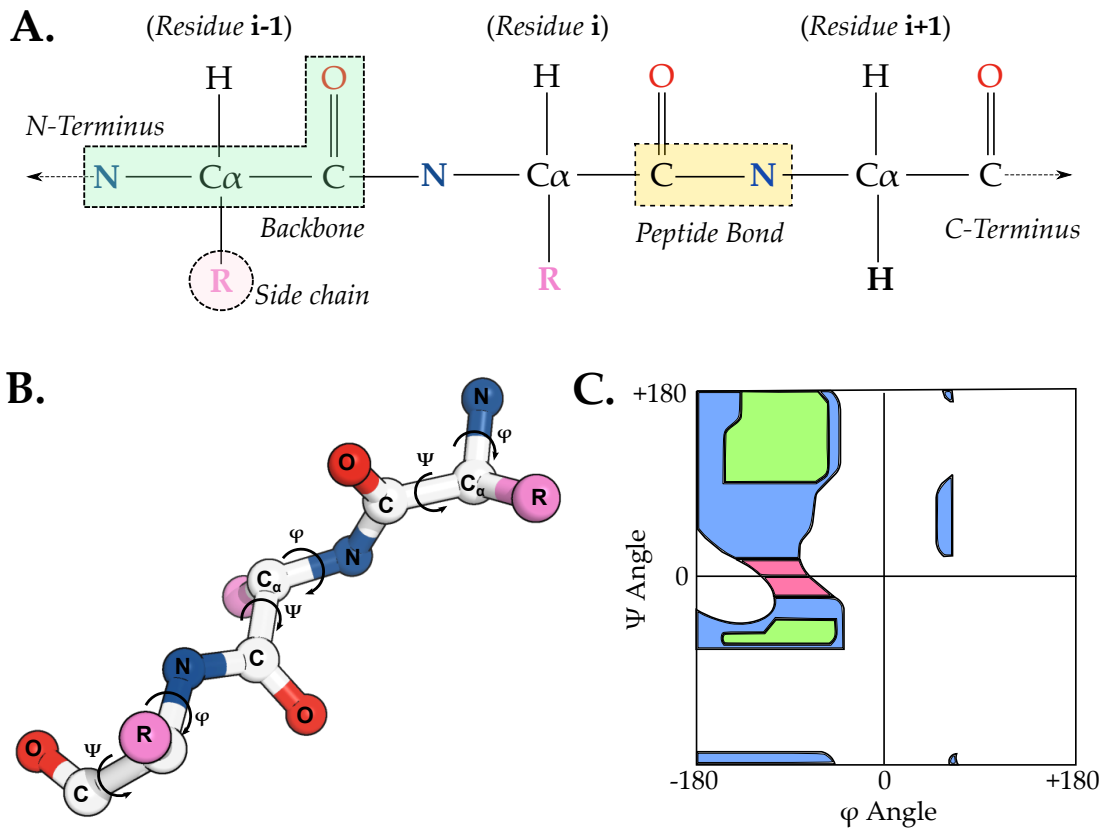


Figure 1.1: A. Two–dimensional structure of three amino acid residues. The 'backbone' of an amino acid, highlighted in green, consists of nitrogen (N; blue), the α–carbon (Cα), carbonyl carbon (C), and oxygen (O; red). Apart from glycine ($i+1$th residue), each amino acid has a distinct chemical group at the side chain (R; pink). Each residue is linked to another by a peptide bond, highlighted within the yellow quadrangle. **B.** 3D structure of three amino acid residues. Each non–terminal amino acid residue has two torsion angles along the backbone: the ϕ angle and the ψ angle. **C.** Ramachandran plot of Ala residues. In proteins, Ala residues can adopt a limited combination of ϕ and ψ angles due to steric constraints. Combinations are coloured in decreasing order of steric allowance: green, blue, and pink. Plot adapted from Nelson and Cox (2004).

These chemical properties may also play key roles in a protein's function. For example, Ser residues of proteases are responsible for cleaving proteins, and Tyr residues in antibodies are often a key part of the binding to the antigen. Ultimately, a protein is defined by its amino acid sequence, which is thought to determine its fold (Anfinsen, 1973), and function.

For simplicity, it is common to use the single–letter alphabet to represent a protein's sequence (Appendix Figure A.1). For example, a protein with

the sequence Gln–Val–Leu can be written as 'QVL'. Comparing two (or more) protein sequences offers an avenue to find the relationship between proteins. In principle, proteins with similar sequences are assumed to share structural and functional characteristics (Sander and Schneider, 1991; Orengo *et al.*, 1994).

1.1.2 Protein structures

The amino acid chain usually folds into a series of local, or 'secondary' structure elements. These secondary structures are often organised in 3D space to form a 'tertiary' structure. There are three major types of secondary structures: α–helix, β–strand, and loops (Appendix Figure A.2). The α–helix is characterised by hydrogen bonds between the backbone oxygen of the ith residue and the backbone nitrogen of the i+4th residue (Appendix Figure A.2). On the other hand, multiple β–strands connect together in parallel or anti–parallel directions to form β–sheets. Similar to the α–helix, both types of β–sheets are characterised by their hydrogen bond patterns (Appendix Figure A.2). Loops are an irregular secondary structure that serve multiple purposes, and connect the other secondary structures.

The collection of secondary structures that form a stable, independent unit is known as a 'domain' (Alberts *et al.*, 2008). A protein domain can be classified as predominantly α–helical, β–sheet, or a mix of both; they are also grouped into 'families' according to topological and sequence similarities (Murzin *et al.*, 1995; Orengo *et al.*, 1997).

Smaller proteins often have only one domain, and it is possible for a single chain to have multiple domains. A classic example is an antibody, where one antibody chain has multiple 'immunoglobulin' domains (Edelman, 1973). Once a single chain is folded, it can associate with other chains to form a 'quaternary' structure (Appendix Figure A.2). Multiple antibody chains form interactions to form a quaternary structure (Section 1.2).

1.1.2.1 Acquisition of protein structural data

Analysing protein structures has been facilitated by the availability of data in the PDB (Berman *et al.*, 2000). As of September 2016, over 110,000 protein structures have been released. Ninety percent of the protein structural data in the PDB has been acquired via X–ray crystallography. Other methods such as solution–state nuclear magnetic resonance (NMR) and cryo electron microscopy (cryo–EM) contribute a small, but significant, portion to the PDB.

In an X–ray crystallography experiment, proteins are first crystallised by concentrating the protein in crystallisation media. Within the crystal, proteins form an ordered lattice. An X–ray beam is then directed at the crystal, and atoms within the protein diffract the X–ray; the detected electron diffraction is transformed into an electron density map. Finally, a model is fitted onto the electron density to infer the protein's atomic coordinates (Alberts *et al.*, 2008). Although X–ray crystallography can provide a high–resolution image of the target protein, crystallisation may force the protein into a non–native environment that can affect its structure (Sander and Schneider, 1991). Furthermore, crystallisation is a non–trivial task (Wlodawer *et al.*, 2013).

In solution–state NMR, a protein is exposed to an external magnetic field and electromagnetic radiation. This excites certain atoms (*e.g.* hydrogen) within the protein, which then emit different spectra of electromagnetic radiation depending on its local chemical environment (*i.e.* amino acid). The hydrogen atoms' emission spectra can also be used to infer the relative distances between atoms. Multiple models are fitted to the spectrum, which represent the possible 3D states of the protein (Alberts *et al.*, 2008; Wüthrich, 1990). Although NMR can provide information on the conformational variance of a protein, the method is limited to small (∼30–50kDa) proteins.

Finally, cryo–EM is becoming an increasingly popular method for studying protein structures. Proteins are flash frozen, then examined in an electron microscope. The images are used to reconstruct the atomic coordinates of the protein. Currently, cryo–EM generates lower resolution images in comparison to

X–ray crystallography, though it is useful for obtaining the structures of larger macromolecular complexes (Amunts *et al.*, 2015).

1.2 Antibody structural biology

In most vertebrates, antibodies are characterised as a Y–shape protein, consisting of four protein chains: two 'heavy' chains and two 'light' chains. The heavy chain has a higher molecular weight, and can be one of several isotypes: immunoglobulin M (IgM), IgD, IgE, IgA, and IgG. The choice of the isotype is dependent on biological context. The light chain has two isotypes: κ and λ; as the name implies, a light chain has a lower molecular weight. Often, one light chain pairs up with one heavy chain, and two sets of heavy–light chain pairs form the antibody's quaternary structure (Figure 1.2A). An antibody with a IgG heavy chain is considered an IgG antibody, regardless of whether it is paired to a κ or λ light chain.

Throughout the thesis, antibodies will be assumed and illustrated in the IgG form, unless otherwise stated. As discussed later in Section 1.3, the structure of antibodies provides the foundations for their function in the immune response.

1.2.1 Quaternary structure of antibodies

In an IgG molecule, the two heavy chains are held together by inter–chain disulphide bonds and non–covalent bonds (Chothia *et al.*, 1998). *In vivo*, both heavy chains and both light chains of an IgG are identical. However, synthetic IgG antibodies can be engineered to have up to four different chains in one antibody (Lewis *et al.*, 2014). Antibodies can also be synthesised and/or engineered in various formats (Figure 1.2B).

1.2.2 Antibody chain architecture

Each antibody chain has a series of folded units (domains) known as immunoglobulin (Ig) domains. An Ig domain is known as a variable (V) or constant (C) domain. As their names suggest, the amino acid sequences of two

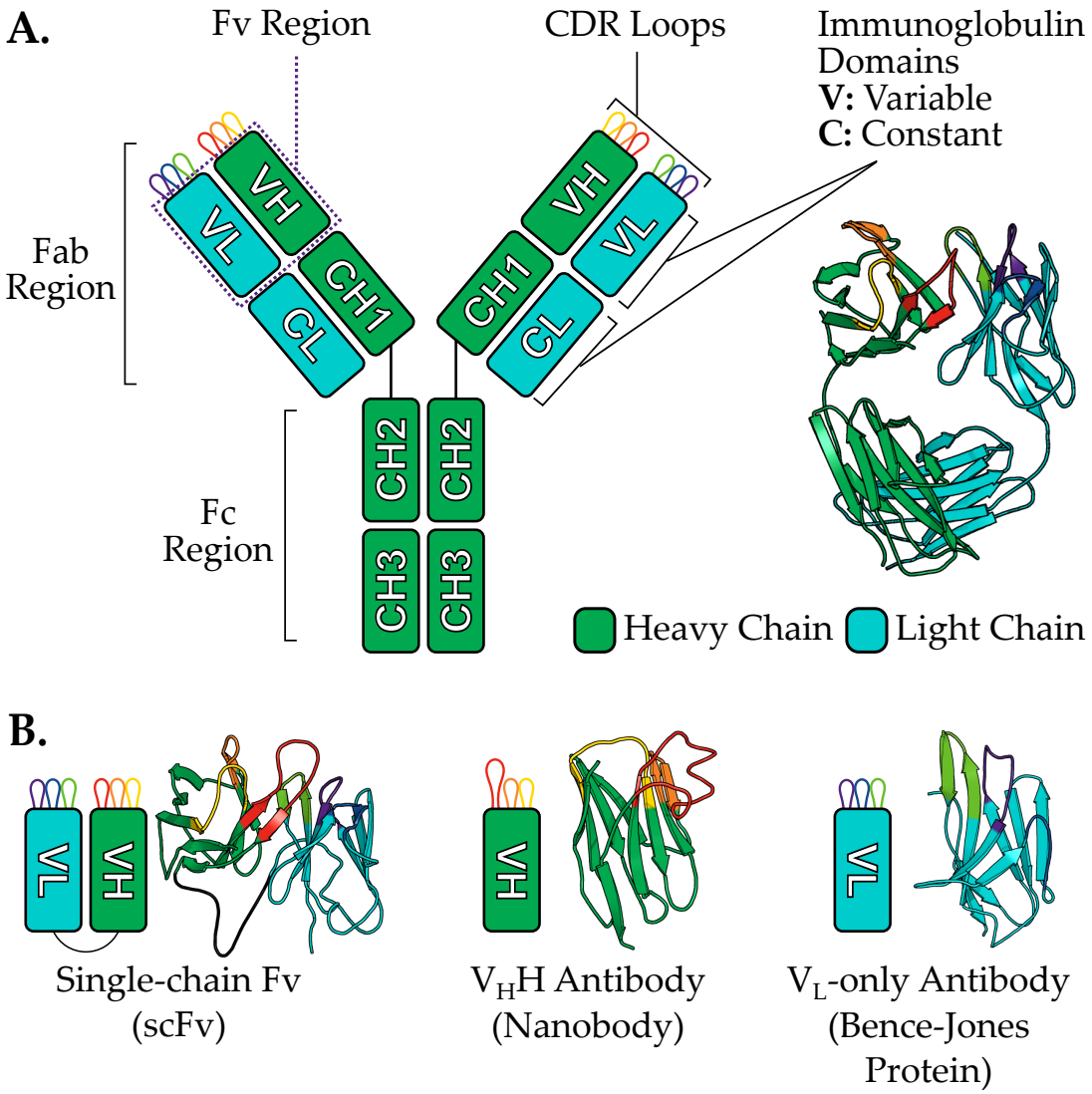


Figure 1.2: **A.** An immunoglobulin G (IgG) antibody molecule has four protein chains: two heavy (green) and two light (cyan) chains. Each chain has a series of immunoglobulin domains: variable (V_H, V_L), and constant (C_H1, C_H2, C_H3, C_L). The V_H domain has three 'complementarity–determining' region (CDR) loops: CDRH1, CDRH2, and CDRH3. The V_L domain also has three CDR loops: CDRL1, CDRL2, and CDRL3. The six loops combine to form the majority of the antigen–binding site. The V_H and V_L domains associate with each other via non–covalent interactions; together, this V_H–V_L pair is known as the variable fragment (Fv). Most antibody structural data in the PDB are solved as antigen–binding fragments (Fab), which consists of the Fv, C_H1 and C_L. The tail of the antibody is known as the crystallisable fragment (Fc), which is formed by C_H2 and C_H3. **B.** Antibodies come in many shapes and sizes. The single–chain variable fragment (scFv) is an engineered antibody format, where the V_H and $V_L$domains are held together by a protein loop (*i.e.* a linker). Nanobodies, such as camelid antibodies, only have the heavy chain, and thus only the V_H domain; they usually have a long CDRH3 loop (red; $\geq$15 residues). Some antibodies are only made of the V_L domain; these are known as Bence–Jones proteins.

different antibodies' variable domains can be different. However, their respective constant domains' sequences are almost identical (assuming the antibodies are of the same isotype) (Sela-Culang *et al.*, 2013).

In an IgG antibody, the heavy chain has four Ig domains (V_H, C_H1, C_H2, C_H3), and the light chain has two Ig domains (V_L, C_L) (Figure 1.2A). The Y–shape of the antibody can be divided to two fragments by treating the antibody with the papain enzyme: the antigen–binding fragment (Fab; V_H, V_L, C_H1, C_L), and the crystallisable fragment (Fc; C_H2, C_H3) (Schroeder and Cavacini, 2010). Within the Fab fragment, the V_H and V_L domains are collectively known as the variable fragment (Fv) region.

Within each variable domain, there are three linear stretches of amino acid residues whose sequences show the largest difference between different variable domains. These regions are known as the 'complementarity–determining' regions (CDRs) (Wu and Kabat, 1970; Chothia and Lesk, 1987). The V_H domain has CDRH1, CDRH2 and CDRH3, and the V_L domain has CDRL1, CDRL2, and CDRL3. Collectively, the first, second and third CDRs of both domains are known as CDR1, CDR2, and CDR3. The regions flanking the CDRs are known as the 'framework' region (FR). Each domain has four FRs, known as FR1, FR2, FR3, and FR4.

The sequence diversity of the CDRs is reflected in antibody structures. The CDRs' structures vary significantly between antibodies (Figure 1.3). The six CDR loops combine to form the majority of the antigen–binding site, known as the paratope (Kunik *et al.*, 2012; Krawczyk *et al.*, 2013, 2014).

1.2.3 Tertiary structure of immunoglobulin domains

The Ig domain is a β–sheet sandwich structure made of $\sim$100 amino acids. Each domain has two anti–parallel β–sheets, which are held together by a disulphide bond. The V_H and V_L domains pack together to form the V_H–V_L interface (Figure 1.4A). In the variable domains, the β–sheet facing the interface (*i.e.* the interfacial sheet) typically has five strands, whereas the peripheral sheet

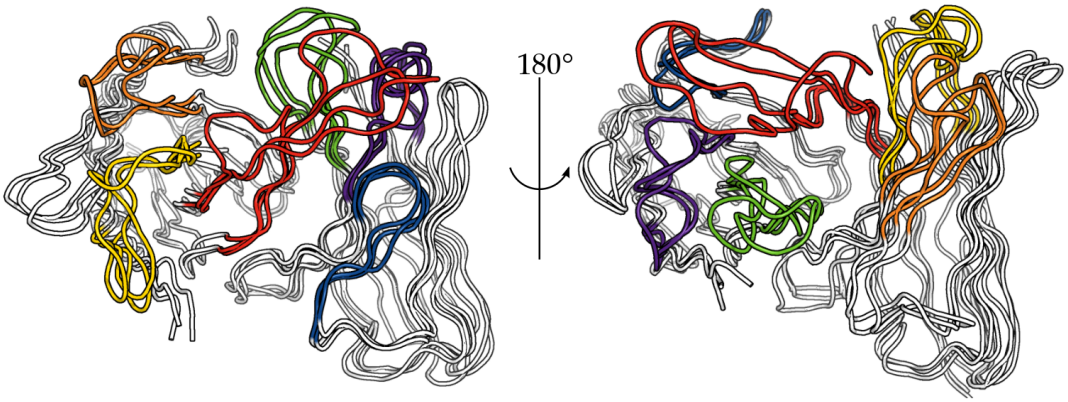


Figure 1.3: The CDR loops show the most structural variation between different antibodies. The following colour scheme is used for the CDRs: purple (CDRL1), blue (CDRL2), light–green (CDRL3), yellow (CDRH1), orange (CDRH2), and red (CDRH3). FRs are coloured white. Of these, CDRH3 shows the greatest variation.

(away from the interface) has four (Figure 1.4C). The constant domains have seven β–strands (4 + 3 strands; Chothia *et al.*, 1985)). The number of β–strands is not static between different variable domains. Some structures may have extra strands (*e.g.* near the N–terminus), or lack some strands (*e.g.* strands flanking CDRH2 and CDRL2).

1.2.4 Annotation of variable domains

The original concept of FRs and CDRs is owed to work by Wu and Kabat (1970). Here, the authors aligned 77 V_L domains' sequences. They observed that certain positions in their alignment showed greater variation in amino acid frequencies over other positions. These positions were denoted as the CDRs. From here, the positions of their alignment established the foundations of a 'numbering scheme' for antibodies, and the boundaries for demarcating CDRs.

1.2.4.1 Variable domain numbering

The goal of an antibody 'numbering scheme' is to annotate an antibody sequence in order to facilitate the comparison of two (or more) sequences. In principle, a given position should capture features such as amino acid frequencies, and

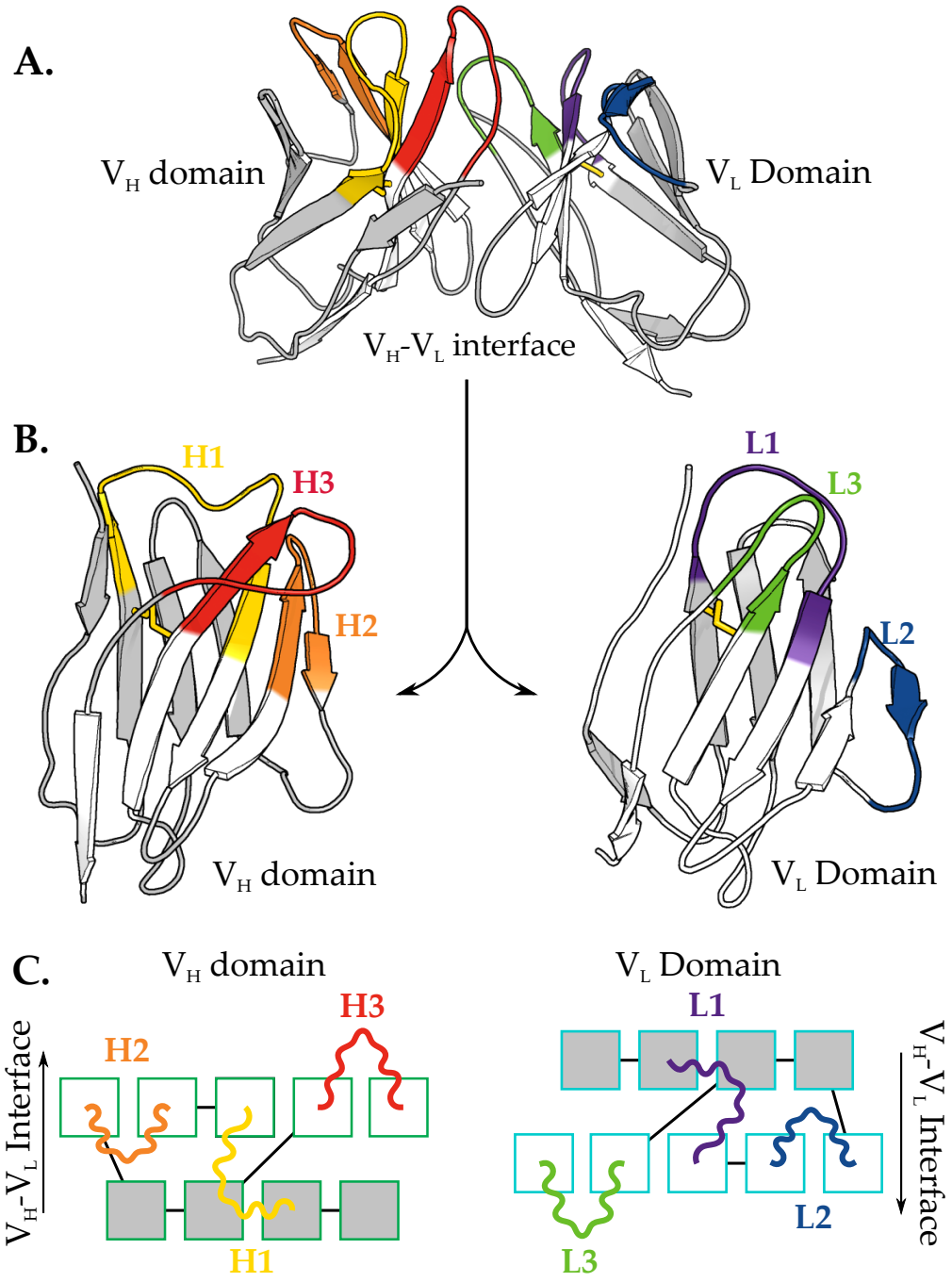


Figure 1.4: Immunoglobulin domain structures in antibodies. **A.** 3D structure of an antibody Fv. CDRH1 (yellow), CDRH2 (orange), CDRH3 (red), CDRL1 (purple), CDRL2 (blue), CDRL3 (green) are coloured under North *et al.* (2011)'s CDR definitions (Section 1.2.4.2). **B.** The V_H and V_L domains are β–sandwiches, where the interfacial sheet (white) has five strands, and the peripheral sheet (grey) has four strands. Each sheet is connected by a disulphide bond (yellow sticks). **C.** Each β–strand of the V_H and V_L domains is represented as a square; CDR loops are coloured as in **A**. The location of the V_H–V_L interface is represented by the arrow. Based on Chothia and Lesk (1987).

Table 1.1: Comparison of major antibody numbering schemes.

	Kabat	Chothia	AHo	IMGT
V_H positions	1–113	1–113	1–149	1–128
V_H insertions	35, 52, 82, 100	31, 52, 82, 100		111, 112
V_L positions	1–107	1–107	1–149	1–127
V_L insertions	27, 95, 106	30, 95, 106		111, 112

the position's role on an Ig domain's structure. Ideally, a position p should represent a particular location with respect to an Ig domain's 3D structure.

There are four major numbering schemes for annotating antibody variable domains: the Kabat (Kabat *et al.*, 1983), Chothia (Chothia and Lesk, 1987), AHo (Honegger and Plückthun, 2001), and IMGT (Lefranc *et al.*, 2003) numbering schemes. For each scheme, the general notation for a position is to use the chain, followed by the position number. If an insertion occurs in the domain, typically a letter is used. For example, heavy chain position 24 would be 'H24', whereas light chain position 95 with two consecutive insertions would be 'L95A' and 'L95B'.

The Kabat scheme is the oldest numbering scheme, and is based solely on sequence analyses. Under the scheme, each variable domain has insertions in certain positions (Table 1.1; Kabat *et al.*, 1983). The Chothia scheme is largely based on the Kabat scheme, though insertions for the CDRH1 and CDRL1 loops were corrected from the Kabat scheme following structural analyses (Table 1.1; Chothia and Lesk, 1987).

Despite the historical and widespread use of the Kabat and Chothia schemes, there are two disadvantages. The annotations for the V_H and V_L domains are different, making it difficult to establish structural equivalence between the two domains (Figure 1.5A). Arguably, the bigger limitation of both schemes is their unidirectional insertion notation. This is particularly problematic when comparing CDR loops of different lengths (Figure 1.5B).

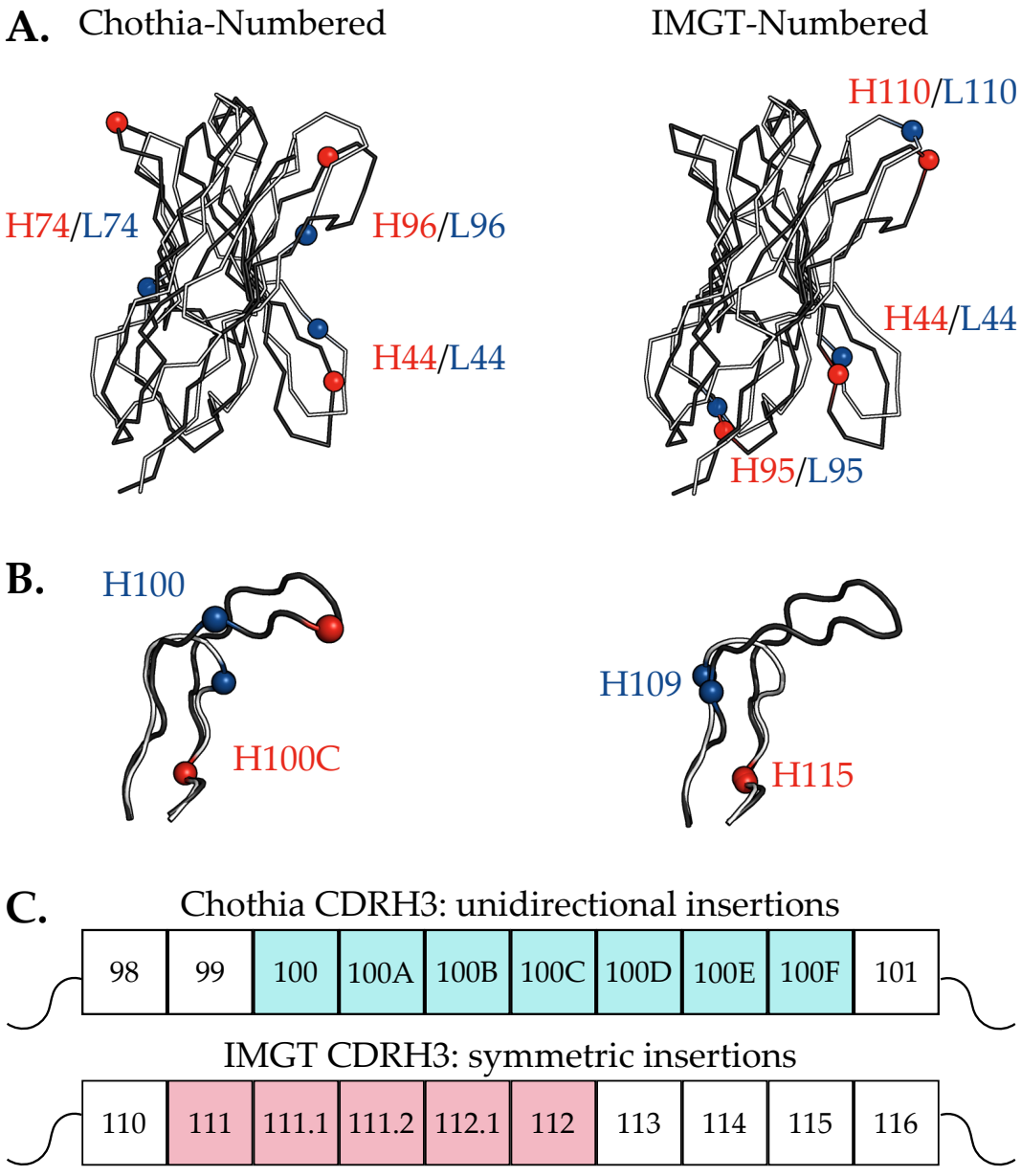


Figure 1.5: Comparison of the Chothia and IMGT numbering schemes. **A.** The light chain (white) of an antibody has been superimposed onto the heavy chain (black). The Cα atom of the superimposed structures for given positions are shown as red (heavy chain) or blue (light chain) spheres. In the IMGT scheme, a position on either domain refers to a similar location within the β–sandwich. However, this is not necessarily the case in Chothia numbering (*e.g.* H74/L74). **B.** Superimposing two CDRH3 loops shows that a position in the Chothia scheme (*e.g.* H100C) can refer to very different areas of a CDRH3. In the IMGT scheme, CDRH3 positions tend to have good structural equivalence between antibodies. **C.** Unidirectional insertion in the Chothia scheme (H100A–F) compared to symmetric insertion in the IMGT scheme, centred around residues H111 and H112.

Table 1.2: Numbering method for CDRH3 loops in the IMGT scheme.

Length										
7	...	109	110						114	...
13	...	109	110	111			113	113	114	...
15	...	109	110	111	111.1	112.1	112	113	114	...

The AHo and IMGT numbering schemes are more recent methodologies for antibody numbering (Honegger and Plückthun, 2001; Lefranc *et al.*, 2003). Both methods use the same numbering for the V_H and V_L domains to establish structural equivalence (Figure 1.5). Both methods also use a larger base range of residue numbers, leading to a different number for almost every residue in the sequence (Table 1.1). In the IMGT scheme, insertions and deletions are handled symmetrically around one central residue (Table 1.2). For example, when annotating CDRH3 loops in the IMGT scheme, insertions are added in increasing order until the centre residue (*i.e.*, approximately the apex of the CDRH3 loop). From here, insertions are added in decreasing order (Table 1.2). In both AHo and IMGT schemes, excess positions are deleted as necessary. For instance, in the IMGT scheme, a CDRH3 loop that is seven residues long will only have numbers H105–H110, then H114–117 (*i.e.*, H111–H113 are deleted; Table 1.2).

Throughout the thesis, all references to antibody positions will be in the IMGT numbering scheme unless otherwise stated. A slight variation will be made, where instead of decimals, insertions will be represented by letters. For instance, H112.2 will be represented as H112B.

1.2.4.2 CDR boundaries

Historically, the numbering schemes in Section 1.2.4.1 were developed to identify hypervariable positions, *i.e.* the CDR loops, of an antibody sequence (*e.g.*, Wu and Kabat, 1970; Chothia and Lesk, 1987). The CDR loops can be defined from three different perspectives: sequence variability, structural variability, or

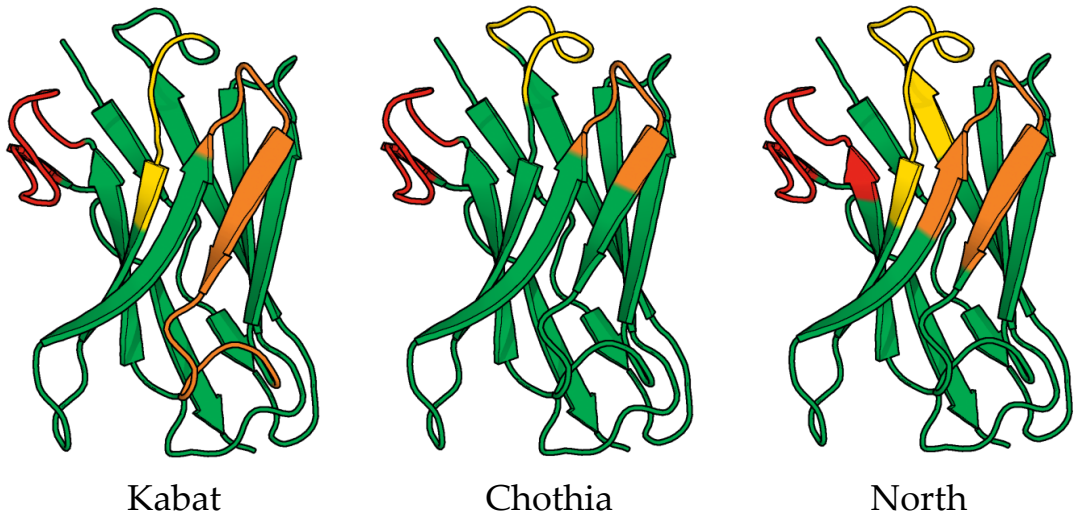


Figure 1.6: Comparison of CDR loop definitions on a V_H domain. CDRs are coloured in the same scheme as Figure 1.4.

relevance to antigen binding. Based on these definitions, a different set of IMGT positions from the V_H or V_L will be classified as a CDR (Figure 1.7).

The Kabat CDR boundaries are based on the variations of antibody amino acid sequences (Kabat *et al.*, 1983). As this definition of CDRs was developed without structural information, the correspondence between certain CDRs and loop structures is low. For example, in the case of the Kabat–defined CDRH1 and CDRH2, these CDRs cover a portion of the interfacial β–strands (Figure 1.6).

The Chothia definition is arguably the first structure–based definition of the CDR loops. Chothia and Lesk (1987) analysed six antibody structures, and identified conserved positions within the β–sheets. This established the anchor points of the CDR loops – the residues flanking the CDR loop. The residues between the anchors were labelled as the CDRs. Other definitions, such as those by North *et al.* (2011), defined new boundaries for the CDR loops based on Honegger and Plückthun (2001)'s structural alignment. The IMGT definition is a hybrid definition that uses sequence and structural variation to demarcate the CDRs (Lefranc *et al.*, 2003).

Finally, the CDRs have also been defined in the context of antigen binding. MacCallum *et al.* (1996) analysed positions that were involved in forming interatomic contacts with the antigen. The CDRs were then defined on the basis of contact profiles. Kunik *et al.* (2012) also use a contact–based definition to define the antigen–binding residues, which have a high overlap with the Chothia–defined CDRs.

A comparison of the Kabat, Chothia, IMGT and North CDR definitions are shown in Figure 1.7. There are overlaps between the various definitions, and each definition has its merits. For the remainder of the thesis, CDRs will be defined as those set by North *et al.* (2011) unless otherwise stated. The IMGT numbering and North *et al.* (2011)'s CDR definitions are represented as a Collier de Perles diagram (Figure 1.8) (Lefranc *et al.*, 2009).

1.2.5 Classification of CDR loops

Although the CDRs have an enriched level of sequence variability, the extent of structural diversity is relatively limited for CDRH1, CDRH2, CDRL1, CDRL2, and CDRL3. These five CDR loops have been clustered into groups of common, or 'canonical', structures (*e.g.* Chothia and Lesk, 1987). Collectively, these five CDRs are also known as the canonical CDRs.

Early investigations manually clustered the CDR loops into the canonical forms. These decisions were based on visual inspection of a limited number of structures. The underlying assumption in the canonical model is that certain residues, are determinants of the loops' shapes (Chothia *et al.*, 1989). Currently, it is well–established that this is a broad generalisation, with several outliers that do not fit in any cluster (Martin and Thornton, 1996; North *et al.*, 2011; Nowak *et al.*, 2016).

CDR categorisation is now often automated (*e.g.* Martin and Thornton, 1996; North *et al.*, 2011). The general paradigm for clustering is to separate each of the canonical CDRs into length bins, and using sequence information to define

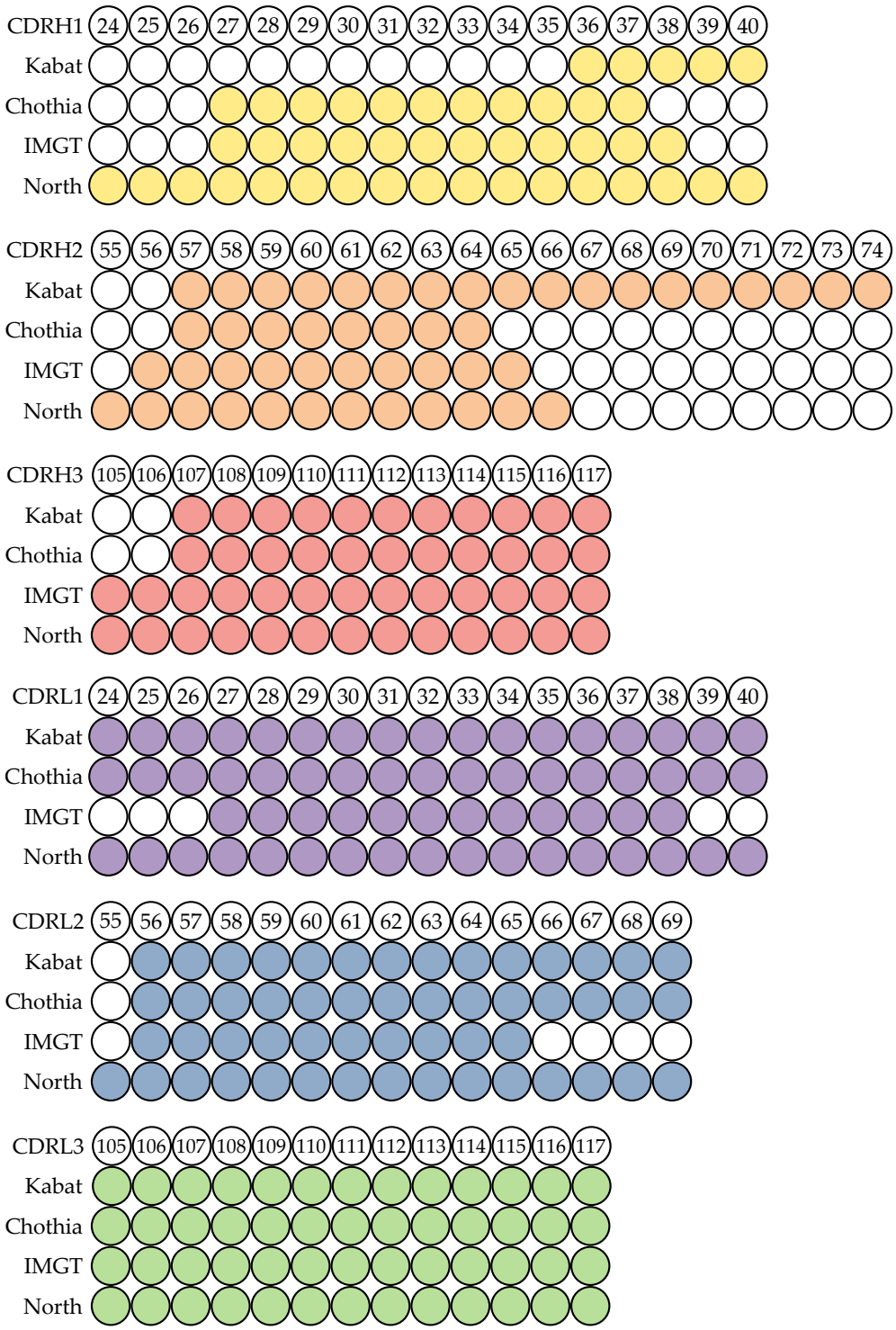


Figure 1.7: Comparison of the CDR boundaries. For each CDR loop, the IMGT numbers are written in white circles on the top row. Circles are filled in for each CDR definition under the IMGT scheme. For example, the Kabat–defined CDRH1 loop spans from H36–H40, whereas the Chothia–defined CDRH1 spans between H27–H37.

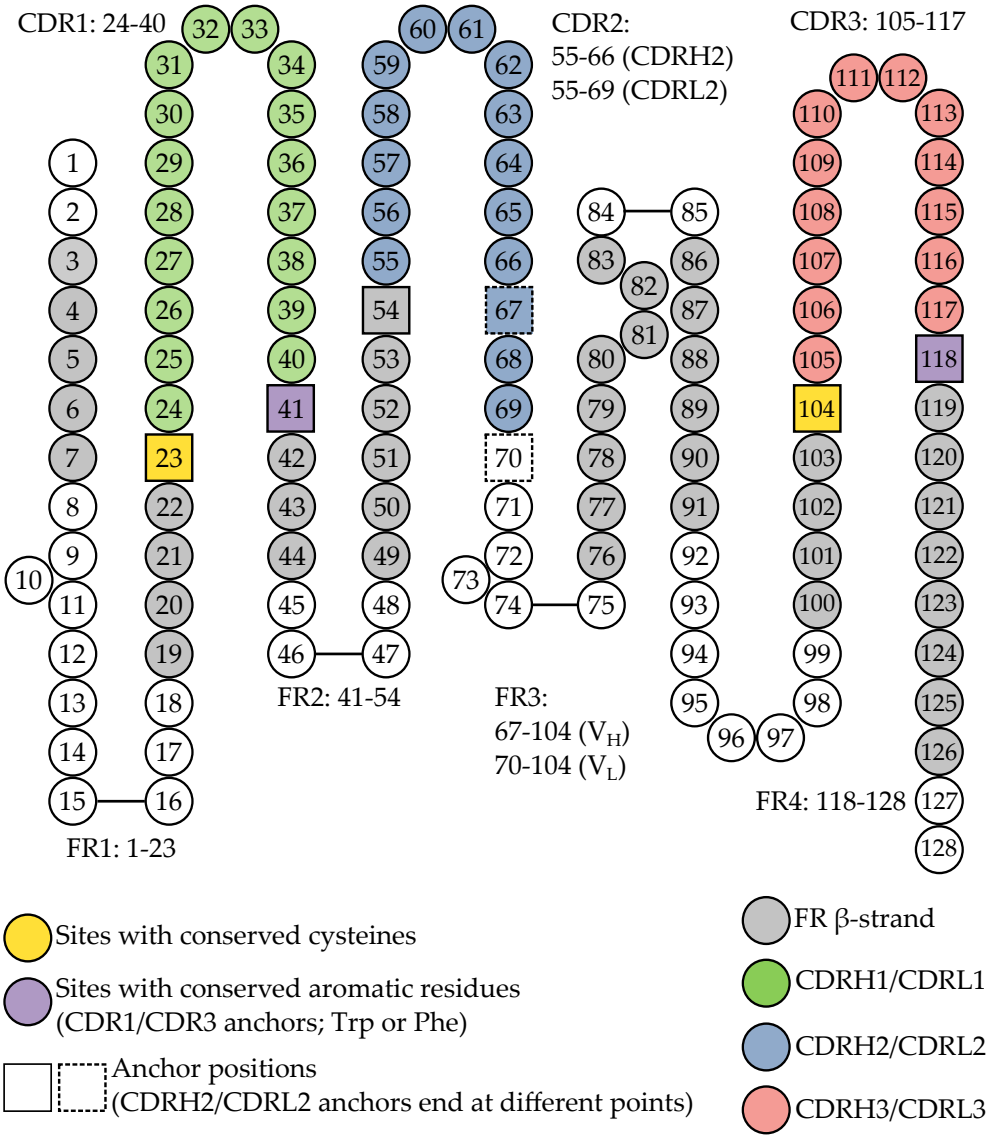


Figure 1.8: Collier de Perles diagram (Lefranc *et al.*, 2003), showing IMGT positions in the variable domains with North *et al.* (2011)'s CDR definitions. FR positions that are often deleted (*e.g.* position 10 in V_H) are offset from the chain. FR positions that form β–strands are coloured grey (Chothia and Lesk, 1987). Positions with conserved Cys are coloured yellow, and aromatic residues at the end of CDRH1, CDRH3, CDRL1, and CDRL3 loops are coloured purple. CDR loop anchors are labelled as solid squares, except for the C–terminal anchor of the CDRH2 and CDRL2 loops (dotted squares).

memberships for loops. An exception is the methodology by Nowak *et al.* (2016), where CDRs are categorised by a length–independent method.

The CDRL3 and CDRH3 loops also been investigated independently, given their greater diversity, role in antigen binding and influence on V_H–V_L pairing (Kuroda *et al.*, 2009; Townsend *et al.*, 2016). Currently, CDRL3 structures can be clustered into canonical forms (*e.g.* Kuroda *et al.*, 2009; Teplyakov and Gilliland, 2014), and CDRH3 loops can not. Several rules have been developed to classify CDRH3 loops according to structural motifs (Shirai *et al.*, 1999; Kuroda *et al.*, 2008). However, these rules have poor generality, and often make incorrect assignments.

1.2.6 V_H–V_L interface

The V_H–V_L interface has been studied extensively for its potential role in stability (Masuda *et al.*, 2006; Hsu *et al.*, 2014), antigen binding (Foote and Winter, 1992; Nakanishi *et al.*, 2008; Fera *et al.*, 2014), and V_H–V_L pairing (Igawa *et al.*, 2010; Lewis *et al.*, 2014). The interfacial β–sheets of the V_H and V_L fold into a concave, barrel–like structure at the interface (Chothia *et al.*, 1985; Novotný and Haber, 1985). Furthermore, the β–sheets pack against each other with extensive conformational flexibility (Abhinandan and Martin, 2010; Dunbar *et al.*, 2013).

1.2.6.1 Contacts at the V_H–V_L interface

Early analyses of the V_H–V_L interface focussed on characterising residues that influenced the association of the variable domains. Chothia *et al.* (1985) described positions H50, H113, H118, L50, L116, and L118 as the key residues in V_H–V_L packing; these positions were twisted toward the centre of the V_H–V_L interface in all three structures of their dataset. These six key positions have since been revised, based on an expanded dataset of 23 structures (Vargas-Madrazo and Paz-García, 2003). Despite the small dataset sizes in both studies, the common theme is that the interface is formed from the interaction of FR and CDR residues.

Several types of interactions stabilise the V_H–V_L interface (Novotný and Haber, 1985). Interdomain hydrogen bonds, such as the hydrogen bond between two Gln residues at H44 and L44, appear to play a key role in V_H–V_L packing (Novotný and Haber, 1985; Chothia *et al.*, 1985; Kuroda and Gray, 2016).

A recent analysis on 547 V_H–V_L structures confirmed the importance of hydrogen bonds in the association of V_H and V_L (Kuroda and Gray, 2016). Despite the hydrophobicity of the V_H–V_L interface, the H44–L44 Gln–Gln hydrogen bond and the 'asymmetric' hydrogen bond network is highly conserved. Residues in the V_H domain, particularly CDRH3 residues, tend to use backbone atoms to form interdomain hydrogen bonds, whereas V_L residues tend to use side chain atoms. This asymmetric usage of atoms in the V_H and V_L may be deliberate. Kuroda and Gray argue that diverse side chains are embedded in the CDRH3 for antigen binding, and thus it is likely that the remaining invariant set of backbone atoms would be used to form stabilising interdomain contacts.

1.2.6.2 Interfacial contacts and V_H–V_L pairing

The H44–L44 Gln–Gln hydrogen bond is well–conserved, and this interaction has been mutated to control V_H–V_L pairing. For instance, the H44 and L44 have been mutated to Lys and Asp (or Glu), respectively, to generate specific V_H–V_L pairings (Igawa *et al.*, 2010; Lewis *et al.*, 2014).

In vivo, heavy chains may be screened on the basis of their ability to form interdomain contacts. In particular, there seems to be a focus on probing the CDRH3 loop to form such contacts. During the B–cell maturation process (Section 1.3.3), the B–cell uses the 'surrogate' light chain (SLC) to positively select for heavy chains that will form pairs with a V_L domain. Absence of the SLC has been linked to several diseases, though its precise role remains elusive (Mårtensson *et al.*, 2010).

In humans, there is only one SLC, which is comprised of the VpreB and $\lambda 5$ proteins (Melchers, 2005). VpreB and $\lambda 5$ have strong structural resemblance to

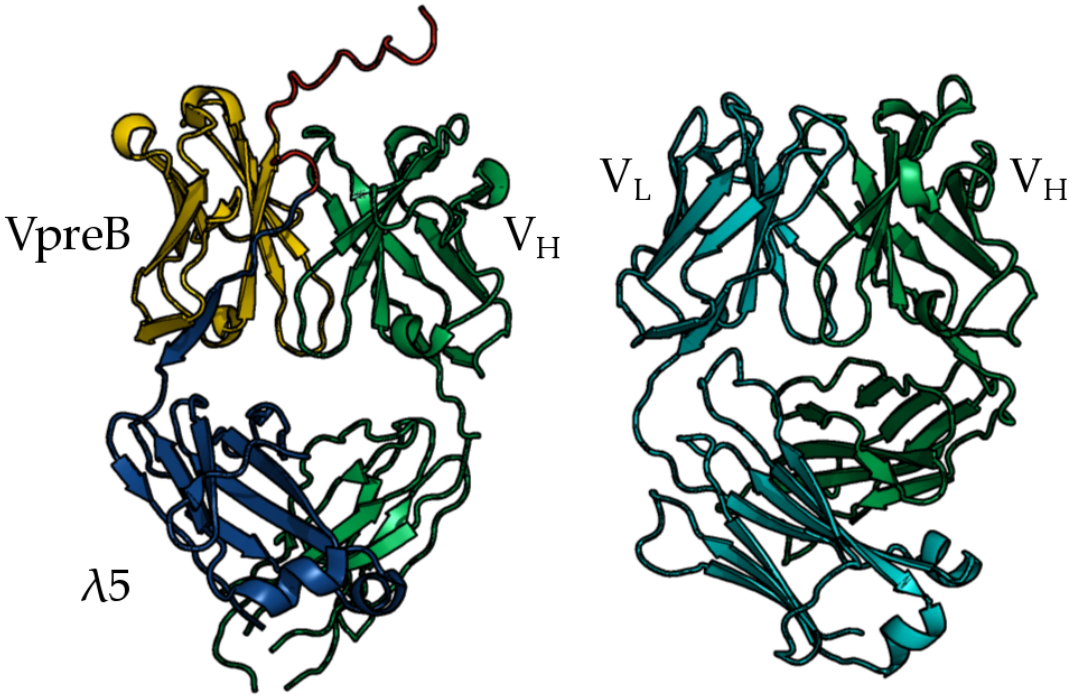


Figure 1.9: Comparison of the pre–BCR (left) and a Fab fragment (right). The structure of the SLC resembles a typical light chain. Most components of a V_L have an analogous counterpart in VpreB, with the exception of CDRL3 and FR4. The equivalent of CDRL3 is the 'unique region' of VpreB, and the equivalent of FR4 comes from the N–terminus of $\lambda 5$. The approximate unique regions of VpreB and $\lambda 5$ are shown in red. Several residues are missing from the unique regions of both VpreB and $\lambda 5$ in the structure by Bankovich *et al.* (2007).

a typical V_L and C_L, respectively (Figure 1.9). An antibody heavy chain pairs up with the SLC to form the pre–B–cell receptor (pre–BCR).

The SLC uses two 'unique regions', a flexible string of amino acids from the C–terminus of VpreB and the N–terminus of $\lambda 5$, for V_H screening. Although the SLC has fewer contacts across the entire interface with a V_H domain, most of its contacts are concentrated between the unique regions and the CDRH3 loop. The V_H–SLC interface also has a greater buried surface area than traditional V_H–V_L interfaces (Bankovich *et al.*, 2007). These two pieces of evidence suggest that the SLC may check for a specific set of contacts as the determinant of a heavy chain's ability for pairing. However, the lack of structural data of other pre–BCRs, and the fact that some heavy chains can pair with functional light

chains, but not the SLC, prompts further investigation (Smith and Roman, 2010). In Chapter 3, the thesis will cover the mechanism of V_H–V_L pairing in more detail, with an emphasis on the V_H–V_L interface structure.

1.2.6.3 Antibody orientation

The relative arrangement of the variable domains is described as the antibody's orientation (Abhinandan and Martin, 2010; Dunbar *et al.*, 2013; Marze *et al.*, 2016). The orientation is highly dependent on the composition of the V_H–V_L interface (Masuda *et al.*, 2006), and is known to affect an antibody's binding properties (Fera *et al.*, 2014) and 'humanness', *i.e.* its closeness to a human structure (Nakanishi *et al.*, 2008; Bujotzek *et al.*, 2016).

The analysis by Chothia *et al.* (1985) was one of the first to describe the concept of antibody orientation, quantifying it using a packing angle between the two variable domains. Since then, antibody orientation has been described in several ways, and quantified as a metric (Narayanan *et al.*, 2009; Chailyan *et al.*, 2011; Abhinandan and Martin, 2010; Dunbar *et al.*, 2013; Marze *et al.*, 2016).

Narayanan *et al.* (2009) and Chailyan *et al.* (2011) describe antibody orientation as a relative measure. A distance–based metric is used to describe how the arrangement of the variable domains varies between different antibodies. Narayanan *et al.* (2009) use an orientation root–mean square deviation (RMSD; Appendix C.3) value; only one of the domains is used for structural alignment, and the deviation of the opposing domain is computed as the o–RMSD. Chailyan *et al.* (2011) uses the global distance test (GDT) score, and cluster structures on the basis of their pairwise GDT scores, leading to three orientation groups.

An early attempt at defining an absolute sense of orientation was the work by Abhinandan and Martin (2010). Here, the orientation is only described by the V_H–V_L packing angle, which represents the torsion angle between two vectors fitted through the V_H and V_L domains. However, the vectors are fitted differently for each structure, making it difficult to compare the orientation

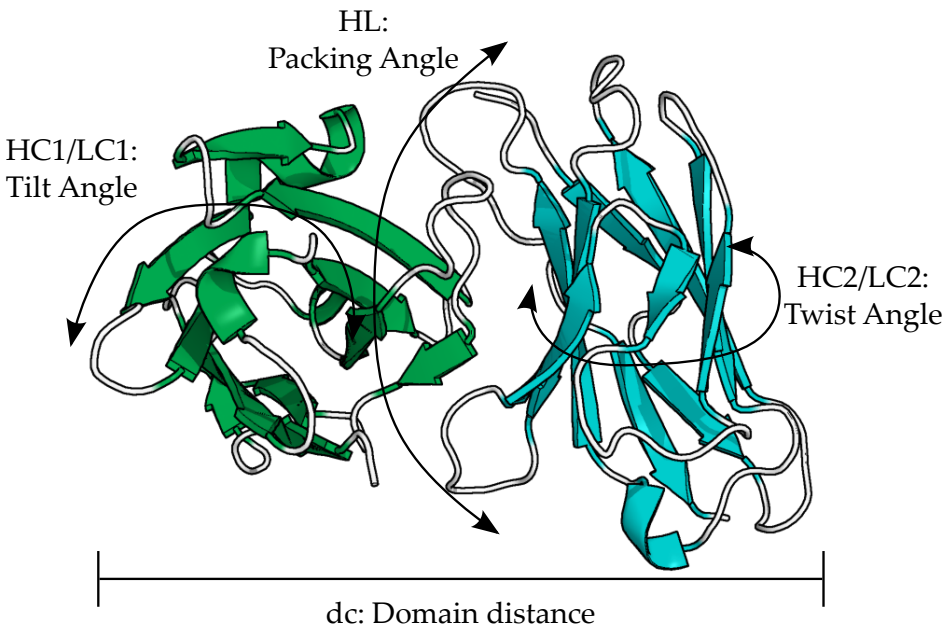


Figure 1.10: V_H–V_L orientation parameters as described by ABangle (Dunbar *et al.*, 2013). Each variable domain has a tilt (HC1/LC1) and twist angle (HC2/LC2). The HL angle refers to the packing angle between the two variable domains. The dc parameter refers to the distance between the variable domains.

between different antibodies. Furthermore, having only one metric does not fully capture the movements of the variable domains.

ABangle is a method that defines the V_H–V_L orientation in terms packing, tilt, and twist angles, along with the distance between the variable domains. In order to define these parameters, Dunbar *et al.* (2013) aligned a non–redundant set of antibody structures using the most structurally conserved sites of the V_H and V_L framework regions. The aligned coordinates were then used to define two reference planes, one for each variable domain. The planes can be fitted to any antibody structure in order to calculate the orientation between the two domains. Recently, Marze *et al.* (2016) have also described antibody orientation in a similar manner to ABangle. Their parameters are nearly identical to those of ABangle; the exception is that the twist angles are omitted in their description. In the thesis, an antibody's orientation will be described by its ABangle terms (Figure 1.10).

1.2.7 Antibody–antigen contacts

Each antibody has a unique paratope that allows it to bind a unique structural motif of the antigen, known as the 'epitope'. The CDR loops are often considered to define the antigen–binding site, despite the fact that segments of some CDR loops are not involved in contacting the antigen (MacCallum *et al.*, 1996). In fact, the precise definition of the paratope has been subject to debate (*e.g.* Padlan *et al.*, 1995; MacCallum *et al.*, 1996; Kunik *et al.*, 2012). With the increasing volume of structural data, paratopes have been defined according to the contacts between an antibody and its antigen. For instance, Kunik *et al.* (2012) used a multiple structural alignment of antibody–antigen complex structures to define a consensus set of binding residues, known as the 'antigen–binding regions'. However, these new definitions for the paratope often overlap with the CDR loops.

Restricting the binding site to a subset of residues, such as the antigen–binding regions, can limit the number of sites that need to be tested for mutation. Several methods are dedicated to predicting the paratope (*e.g.* Kunik *et al.*, 2012; Krawczyk *et al.*, 2013; Olimpieri *et al.*, 2013). For example, Paratome uses a combination of BLAST and structural alignment to predict the antigen–binding regions of an antibody. On the other hand, Antibody i–Patch annotates a per–residue score which reflects the likelihood that a residue on the antibody would bind a residue on the antigen. Similarly, proABC uses a random forest model to predict the probability that residues on the antibody would form contacts with the antigen. Ultimately, these methods aim to identify key regions of the antibody structure in order to help the decision–making process for antibody design.

Antibody–antigen interfaces are characterised by unique amino acid distributions that are different from general protein–protein interfaces (Raghunathan *et al.*, 2012; Krawczyk *et al.*, 2013). For instance, Tyr residues are enriched on the surfaces of antibodies (Clark *et al.*, 2006b). Other residues, such as Ser, are

often used by antibodies to form a combination of interactions that help it to bind specific types of antigens (Raghunathan *et al.*, 2012).

A wide range of interactions are formed between the antibody and antigen, including hydrogen bonds (Kuroda and Gray, 2016), hydrophobic, and electrostatic interactions (Clark *et al.*, 2006a). It is the combination of these interactions which affect the binding strength, or affinity, of an antibody–antigen interaction. The affinity can be quantified in many ways (Appendix Section C.2). One of the most common is K_D, the dissociation constant. The K_D value represents the ratio of the unbound antibody and antigen with respect to the bound antibody–antigen complex. A chief design aim is to improve an antibody's affinity by mutating the paratope; in Chapter 2, we will discuss our approach to predicting antibody–antigen binding affinities.

1.3 Antibodies and the adaptive immune response

B–cells are one of the main mediators of the adaptive immune response; they secrete antibodies following antigen stimulation. A single B–cell often produces one type of antibody, which matures to bind a specific epitope on the antigen. Following binding, antibodies recruit other components of the immune system to clear the antigen.

1.3.1 Biological function of antibodies

Antibodies bind a particular antigen with high specificity and high affinity. In principle, a given antigen has multiple epitopes that can be recognised by different antibodies (Schroeder and Cavacini, 2010; Krawczyk *et al.*, 2014). A 'specific' antibody is one whose paratope recognises one particular epitope. On the other hand, a 'polyspecific' antibody is one that can bind multiple antigens, typically with low affinity per antigen (Clark *et al.*, 2006b; Willis *et al.*, 2013).

Once bound to its antigen, antibodies directly neutralise the antigen, or recruit other members of the immune system for reinforcements (Figure 1.11).

The series of responses that follow recognition, collectively called effector functions, largely depend on the isotype of the antibody (Table 1.3).

Table 1.3: Types of effector functions triggered by human antibodies.

	IgG1	IgG2	IgG3	IgG4	IgA	IgM	IgD	IgE
Neutralisation	++	++	++	++	++	+		
Opsonisation	+++		++	+	+			
Complement activation	++	+	+++		+	+++		
ADCC (by NK cells)	++		++					
Activation of mast cells	+		+					+++

'+++': major effector function; '++': minor effector function; '+': very minor effector function. Adapted from (Janeway *et al.*, 2001). ADCC: antibody–dependent cellular cytotoxicity. NK cells: natural killer cells.

Neutralisation is the simplest, yet most direct method in which the antibody defends the host against antigens. For instance, antibodies can neutralise and prevent ligand molecules from interacting with cancer cells' receptor proteins (Weiner, 2015). After an antibody binds the ligand, the exposed Fc region of the antibody (Section 1.2) acts as a tag to recruit monocytes for phagocytosis (Figure 1.11A).

Similarly, antibodies can 'opsonise' the membranes of pathogens or aberrant cells. For example, cancer cells display an exclusive range of antigens on their membranes (Weiner, 2015). Antibodies recognise these antigens and form a coat (opsonise) along the cancer cell's membrane (Figure 1.11B). Opsonised cells are then recognised by monocytes or cytotoxic cells (*e.g.* natural killer cells) via the Fc regions.

Antibody Fc regions can also be recognised by proteins of the complement system (Janeway *et al.*, 2001). Complement proteins work in a cascade to form pores that destroy target cells via cell lysis.

1.3.2 Antibody diversification

In humans, B–cells can theoretically produce over 10^{13} possible antibodies. In practice, only 10^{11} are produced, owing to the number of B–cells in the body (Market and Papavasiliou, 2003; Georgiou *et al.*, 2014). Genes encoding antibody chains are partitioned into multiple gene segments, which later recombine to form a complete 'coding joint'. This is a parsimonious method of maintaining diversity; a relatively low number of gene segments is used to generate an immense number of different functional antibody chains. Antibody–coding genes are further modified by several enzymes for added layers of diversity (Tonegawa, 1983; Janeway *et al.*, 2001; Georgiou *et al.*, 2014).

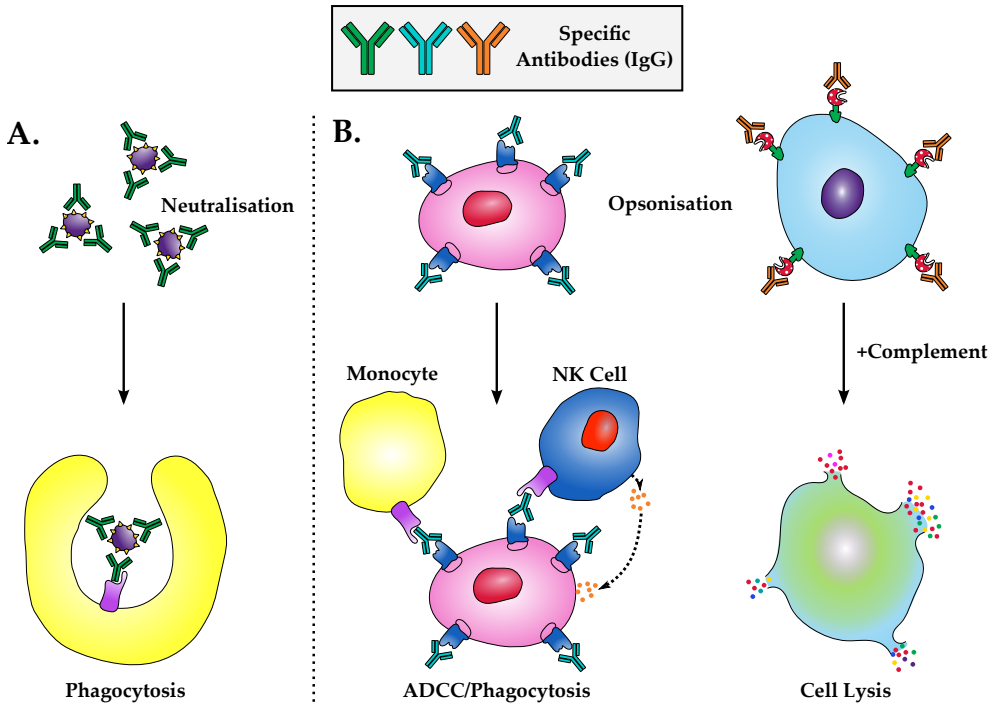


Figure 1.11: Effector functions triggered by IgG antibodies.**A.** Antibodies can bind soluble antigens that are found in the extra–cellular environment. Once the antigen is neutralised by antibodies, monocytes engulf the antibody–antigen complex (phagocytosis). **B.** Antibodies can also form a coat across the surface of a pathogen or aberrant cell (*e.g.* cancer cell). The 'opsonised' cell is then recognised by white blood cells for antibody–dependent cellular cytotoxicity (ADCC) or phagocytosis. **C.** Alternatively, antibodies can be recognised by components in the complement pathway, which trigger cell lysis.

1.3.2.1 V(D)J recombination

Three types of gene segments assemble together to form an antibody chain. They are the variable (V), diversity (D), and joining (J) gene segments. In humans, these gene segments are found in chromosome 14 (heavy chain), chromosome 2 (λ light chains) and chromosome 22 (κ light chains). The heavy chain gene contains 39 V gene segments, 27 D gene segments, and 6 J gene segments. The κ light chain has 40 V and 5 J gene segments, and the λ light chain has 32 V and 4 J gene segments (Schroeder and Cavacini, 2010).

V(D)J recombination generates combinatorial diversity; if every combination of V, D, and J genes can form a functional heavy chain (or V and J genes for light chains), this alone generates a large number of possible antibodies. Furthermore, random nucleotides are added during the joining of the V, D, and J genes, leading to junctional diversity (Market and Papavasiliou, 2003; Schroeder and Cavacini, 2010).

V(D)J recombination is mediated by the recombination–activating gene recombinase enzyme (RAG). These enzymes first bind to recombination signal sequences (RSS) downstream of the V, D and J gene segments. RAG then introduces double–stranded DNA breaks between the gene segment and the RSS, yielding a hairpin at the end of each gene segment. The hairpins are nicked, leaving an overhang of DNA. TdT then adds random, non–templated nucleotides to the single–stranded overhangs at the end of each gene segment. The overhangs are then paired to form the coding joint (Figure 1.12). The V–DJ junctions (or V and J for light chains) overlap with the CDR3 amino acid sequence. In other words, junctional diversity is one of the primary determinants of sequence and structural diversity in CDR3 (Janeway *et al.*, 2001; Schatz and Ji, 2011).

1.3.2.2 Notation of antibody genes

Each of the germline V, D, and J segments are often grouped into families by their nucleotide sequences' identity (Kirkham and Schroeder, 1994; Matsuda *et al.*, 1998; Vargas-Madrazo *et al.*, 1997). The IMGT database provides an

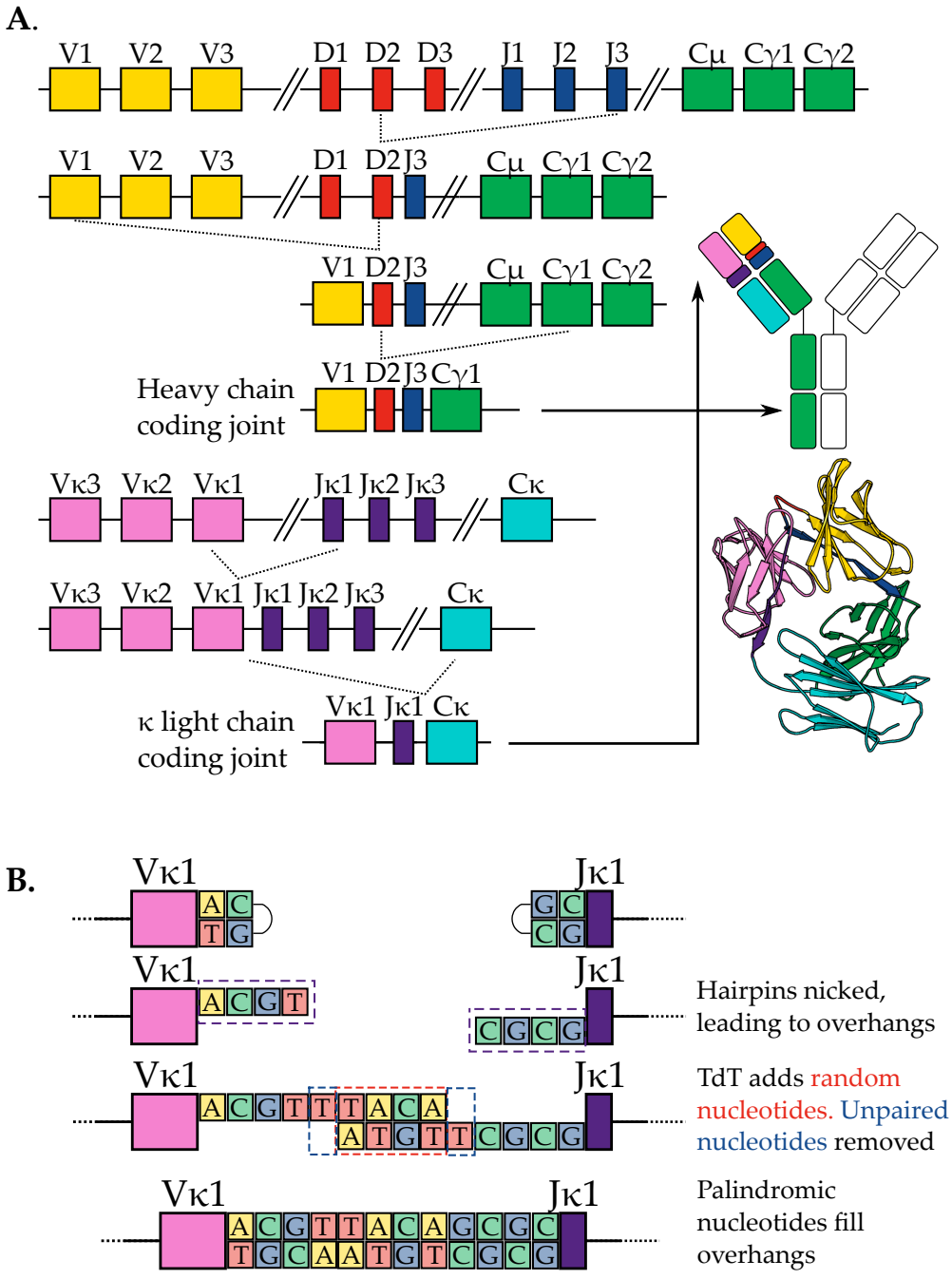


Figure 1.12: Mechanism of V(D)J recombination. **A.** V(D)J recombination starts with the heavy chain; the *D* and *J* gene segments are first joined together, followed by the *V* segment. Light chain recombination starts with the recombination of *V* and *J* gene segments in κ light chains. If the synthesised κ light chain is unsuitable (*e.g.* creates self–reactive antibodies), then λ light chains are used (not shown). The mechanism of VJ recombination is identical for λ light chains. The *V*, *D*, *J* gene segments ultimately encode for different components of the V_H and V_L domains. **B.** When two gene segments are joined by RAG, the hairpins are nicked. TdT randomly adds nucleotides to the ends of each overhang. Any unpaired nucleotides between the overhangs are removed by exonucleases. The segments then pair up, and the remainder of the overhang is filled by palindromic nucleotides.

ontology for these families, *i.e.* a notation for relating the germline origins of antibodies (Giudicelli and Lefranc, 1999; Lefranc *et al.*, 2009). In the thesis, the likely germline genes that encode for V_H and V_L domains will be given in this ontology (Figure 1.13).

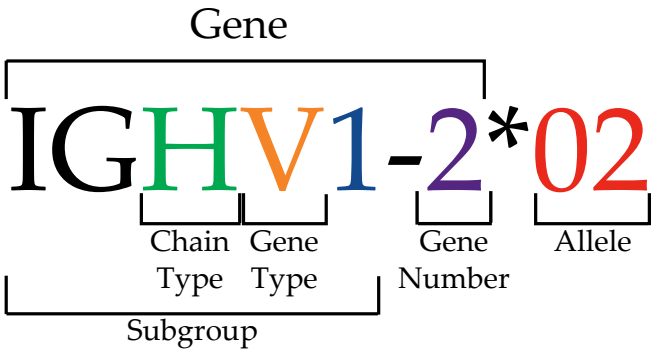


Figure 1.13: IMGT ontology for antibody germline genes under the classification concept (Giudicelli and Lefranc, 1999). For antibodies, there are three chain types: heavy (H), κ–light (K) and λ–light (L). The gene type is either *V*, *D*, *J*, or constant (*C*), where *C* is an antibody isotype. Each gene is identified by a number, and polymorphs of the gene have an allele number. A set of genes with $\geq$75% nucleotide sequence identity is collected under one subgroup. This example is a member of the IGHV1 subgroup, representing a *V* gene from the heavy chain with gene–2, allele–2.

1.3.2.3 Somatic hypermutation

Antigen–activated B–cells migrate to the germinal centres (GCs) for affinity maturation (Section 1.3.3). Somatic hypermutation (SHM) is the engine driving affinity maturation, where the antibody–coding gene is mutated at a rate of 10^{-3} mutations per base pair. In context, this represents a 10^6–fold higher mutation rate over basal mutation rates in somatic, non–lymphocyte cells (Peled *et al.*, 2008).

SHM is initiated by the activation–induced cytidine deaminase (AID). AID introduces a lesion in one of the DNA strands by deaminating a cytosine into a uracil. The U–G base pair can be handled in three different mechanisms: replication, base–excision repair and mismatch repair (Appendix Figure A.3). These

pathways are not mutually exclusive, and the combination of these pathways leads to the overall accumulation of mutations in SHM (Peled *et al.*, 2008).

As discussed in Section 1.3.3, B–cells producing higher–affinity antibodies survive. B–cells can also synthesise antibodies with lower affinity, or self–reactive antibodies as a result of SHM; in either case, the B–cell undergoes apoptosis in the GC. Analyses by Clark *et al.* (2006b) have shown that residues such as Tyr and Ser tend to be replaced by other residue types, such as His during this process. In terms of the antibody's structure, SHM has been observed to influence positions that contact the antigen (Cauerhff *et al.*, 2004; Barderas *et al.*, 2008; Burkovitz *et al.*, 2014). Mutations are not strictly limited to increasing affinity, and may have supporting roles that stabilise the antibody (Wang *et al.*, 2013). Furthermore, computational analyses have observed that SHM can reduce the flexibility of the antibody (Wong *et al.*, 2011; Willis *et al.*, 2013).

1.3.3 Clonal selection of antibodies

B–cells undergo rigorous control mechanisms in the bone marrow. The B–cell's IgM antibody must meet two criteria: the antibody's heavy chain must be able to form pairs with light chains (Section 1.2.6.2), and the antibody cannot be self–reactive (Janeway *et al.*, 2001; Nemazee, 2006). If a B–cell's antibody does not meet either criteria during its maturation process, it is eliminated by apoptosis.

After a B–cell passes these internal checks, it migrates to peripheral lymphoid organs (lymph nodes). Once it binds an antigen, it is activated, and further stimulated by helper T–cells in the lymph node. Antigen–stimulated B–cells either differentiate into plasmablasts for quick secretion of low–affinity antibodies, or participate in the germinal centre (GC) reaction for affinity maturation (Appendix Figure A.4, De Silva and Klein, 2015; Nutt *et al.*, 2015).

The GC is a transient microenvironment within lymph nodes, formed by an activated B–cell. Inside the GC, B–cells clone rapidly, and the clones participate in 'affinity maturation'. This is an iterative process where a B–cell clone's antibody–coding gene undergoes rapid mutation (somatic hypermutation; Section 1.3.2.3),

and each clone is selected by its antibody's affinity. Mutant clones with high–affinity antibodies survive by binding to follicular dendritic cells in the GC, whereas those with low–affinity or self–reactive antibodies are displaced and discarded via apoptosis. The surviving B–cell clone expressing the 'matured' antibody then switches its antibody isotype depending on stimuli (*e.g.* cytokines). Typically, antibodies swap to IgG isotypes. The surviving clone also differentiates into plasma cells or memory B–cells for immediate antibody synthesis or future immune response against the same antigen (De Silva and Klein, 2015).

1.4 Therapeutic antibody design

Since Ehrlich's 'magic bullet' proposal, the idea of developing highly–specific drugs has dominated drug design paradigms (Imai and Takaoka, 2006). Given the ability of antibodies to bind a wide range of targets with high specificity and high affinity, they have generated major interest as biotherapeutic candidates.

1.4.1 Antibodies in the clinic

Therapeutic antibodies are currently used to treat various diseases, including various forms of cancer (trastuzumab for breast cancer, cetuximab for colorectal cancer), Crohn's disease (vedolizumab), and asthma (mepolizumab) (Imai and Takaoka, 2006; Mullard, 2015, 2016). Antibodies are administered to patients whose immune systems are not responsive to antigens that cause these disease states (*e.g.*, receptor proteins on tumours). They form the largest class of biotherapeutics, and many are currently in Phase II or Phase III of clinical trials.

1.4.2 Antibody design objectives

A major objective of designing a new therapeutic antibody is to improve the antibody's affinity and specificity. There are two additional concerns that are pertinent to therapeutic antibody design: safety and stability (Hansel *et al.*, 2010; Jarasch *et al.*, 2015; Seeliger *et al.*, 2015). In theory, even a fully–human antibody

can trigger an immune response (Nelson *et al.*, 2010). Furthermore, the antibody must be stable, both *in vivo* and *ex vivo* (Jarasch *et al.*, 2015).

Antibodies can be modified from the standard IgG format to develop alternative, bespoke therapeutics. For example, antibody–drug conjugates use antibodies as the vehicle to localise toxic chemicals toward disease environments (Imai and Takaoka, 2006; Mullard, 2013). Other engineering campaigns discard the traditional IgG format. Popular formats include bispecific antibodies where each Fv of the antibody binds to a separate antigen. Nanobody–based therapeutics (*i.e.* camelid V_HH antibodies) are also gaining increased popularity due to its long CDRH3 and small size (Holliger and Hudson, 2005; Krah *et al.*, 2016).

1.4.3 Experimental methods for designing antibodies

Therapeutic antibodies are developed, matured, and isolated by a range of techniques. The main methods for antibody engineering include: hybridoma technology (Köhler and Milstein, 1975), antibody humanisation (Jones *et al.*, 1986), transgenic, 'humanised' mice (Jakobovits, 1995), and phage display (McCafferty *et al.*, 1990).

Hybridoma technology is one of the oldest methods of isolating a 'monoclonal' antibody, *i.e.* an antibody from a specific B–cell clone. Although it was once widely used to generate therapeutic antibodies, hybridomas synthesise murine (mouse) antibodies, leading to an immune response against the antibody drug (Köhler and Milstein, 1975).

Safer alternatives to murine antibodies include chimaeric and humanised antibodies. Chimaeric antibodies have mouse variable domains, but human constant domains, which limits the immune response (Imai and Takaoka, 2006). Humanisation takes this one step further, where the CDR loops from a murine antibody are excised and grafted into an otherwise human antibody construct (Jones *et al.*, 1986). Many therapeutic antibodies are now either either chimaeric or humanised (Nelson *et al.*, 2010; Hansel *et al.*, 2010; Mullard, 2016).

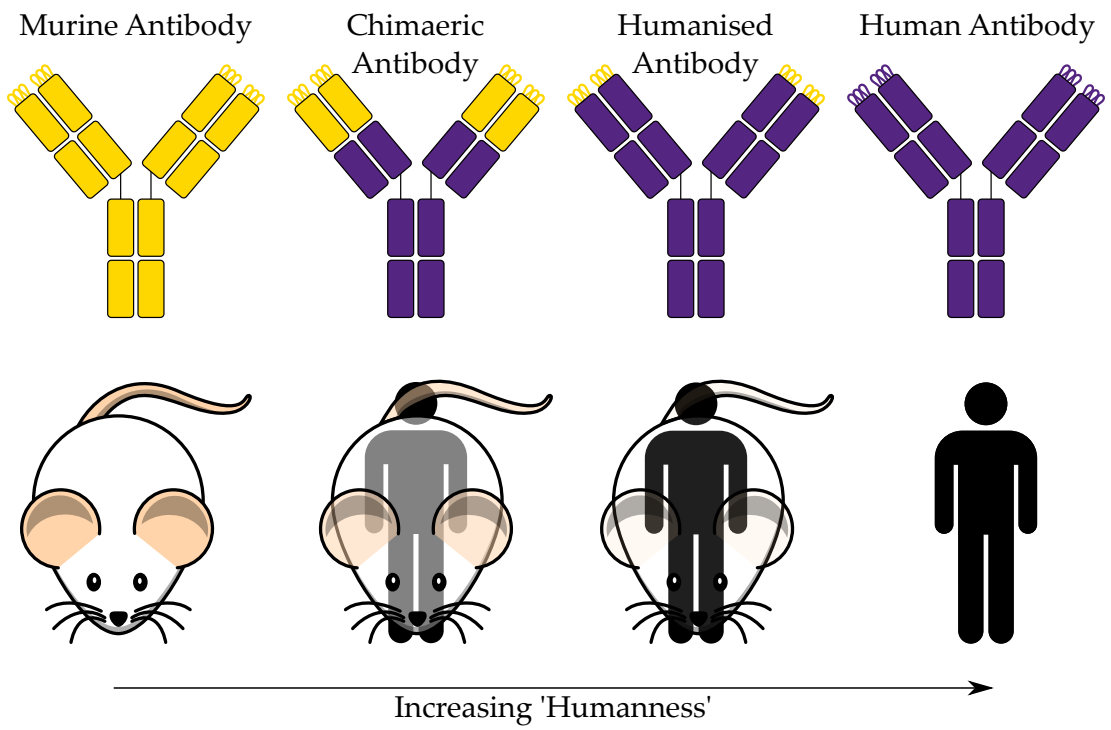


Figure 1.14: Therapeutic antibodies were first discovered as murine antibodies. Several efforts have been made to make them more 'human'. Chimaeric antibodies have human constant domains, but mouse variable domains. Humanised antibodies have human variable domains as well, but the CDRs are directly grafted from murine antibodies. Murine antibody components are coloured yellow, and human antibody components are coloured purple.

A comparison of murine, chimaeric, humanised and fully human antibodies is shown in Figure 1.14.

It has also become increasingly popular to use transgenic, or 'humanised' mice for engineering therapeutic antibodies. Antibody–coding genes in murine embryonic stem cells are knocked out and replaced by human antibody–coding genes (Jakobovits, 1995). These humanised mice are then immunised with the antigen of interest, and the isolated monoclonal antibody from the mouse is a fully human antibody (Marasco and Sui, 2007).

Phage display uses phage viruses to synthesise human antibodies on a large scale. Typically, $>10^7$ different antibody sequences are synthesised and screened (McCafferty *et al.*, 1990; Georgiou *et al.*, 2014). Phage viruses are used to express different antibody genes for a specific target, and viruses are iteratively

selected (Marasco and Sui, 2007). A particular advantage of phage display is the level of control that is available. For example, predictions from computational methodologies can be used to inform the design of antibody sequences, leading to bespoke phage libraries (*e.g.* Barderas *et al.*, 2008; Tiller *et al.*, 2013).

1.4.4 Next–generation sequencing based antibody discovery

A major advance in antibody discovery and informatics has been the increased availability and application of next–generation sequencing (NGS) methods (Mathonet and Ullman, 2013; Robinson, 2015). By using NGS equipment, DNA sequencing has become highly parallelised, thus generating thousands of sequences for analysis. NGS methods have been used to understand the adaptive immune response, particularly an individual's antibody repertoire (*e.g.* Wang *et al.*, 2015; Zhu *et al.*, 2013; DeKosky *et al.*, 2016). Until recently, the main disadvantage of NGS methods was that the information regarding V_H–V_L pairing was lost. However, by restricting sequencing to single B–cells and using novel PCR methods, it has become possible to retain some of this information (DeKosky *et al.*, 2013). In Chapter 4, we will discuss how our antibody modelling tool, ABodyBuilder, can potentially be coupled with NGS to provide an overview of the structural landscape of antibodies.

1.5 Bioinformatics–driven approaches to antibody design

Computational methods are increasingly being used to inform decision–making in antibody design (Kuroda *et al.*, 2012). Previously Lippow *et al.* (2007) have computationally matured the binding affinities of four antibodies, and Choi *et al.* (2015) have computationally humanised a mouse antibody. These studies, along with several others (*e.g.* Clark *et al.*, 2006a; Barderas *et al.*, 2008), demonstrate the growing use and importance of *in silico* antibody design.

1.5.1 Computational antibody design pipelines

Computational protein 'design' refers to engineering a protein *de novo*, whereas 're–design' refers to the modification of an existing protein (Jäckel *et al.*, 2008; Khoury *et al.*, 2014). In either scenario, the aim is to develop a new protein through mutations. Extending from these ideas, a computational antibody design pipeline aims to generate an antibody *de novo* (*e.g.* Pantazes and Maranas, 2010; Li *et al.*, 2014), or re–design an existing antibody (*e.g.* Lewis *et al.*, 2014). Most computational antibody 'design' pipelines have actually been used for re–design. However, there is growing interest in *de novo* computational design, especially in cases where the antibody's structure is not available (*e.g.* Miklos *et al.*, 2012).

Several computational antibody design pipelines have been released, though only a few are freely available (*e.g.* Kaufmann *et al.*, 2010; Pantazes and Maranas, 2010). Rosetta's protein design tool is capable of both *de novo* design and re–design. For example, Miklos *et al.* (2012) built a homology model of the anti–MS2 scFv, and introduced mutations to increase its affinity and stability. Lewis *et al.* (2014) used the structures of two anti–HIV antibodies (PDB: 3tv3, 3tcl) to introduce mutations at the C_H1–Cλ and V_H–V_L interfaces; these mutations were then used to produce bispecific antibodies. OptCDR is another freely available tool, which designs CDR loops *de novo* for a given antigen epitope. The protocol was used to design CDR loops for peptide (hepatitis C virus capsid peptide), hapten (fluorescein), and protein (vascular endothelial growth factor) antigens (Pantazes and Maranas, 2010). Since OptCDR only provides the CDR loops, OptMAVEn was recently developed as an extension of the OptCDR algorithm to model the remainder of the Fv (Li *et al.*, 2014).

For both Rosetta Antibody Design and OptMAVEn, there are several common components (Figure 1.15), which include, but not limited to: sequence analysis and design (*e.g.* Abhinandan and Martin, 2008), antibody structure prediction (*e.g.* Sivasubramanian *et al.*, 2009), and function prediction (*e.g.* epitope prediction, Krawczyk *et al.*, 2014). These components have often been developed

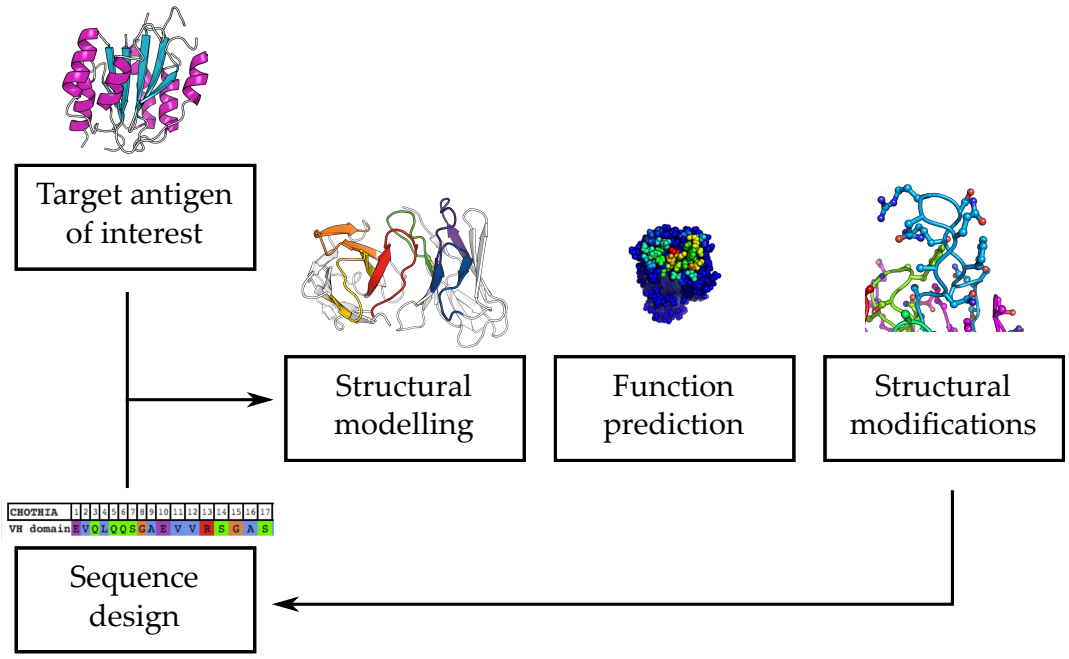


Figure 1.15: Computational antibody design pipeline for *de novo* antibody design. Several tools are available for designing antibodies *de novo*, such as Rosetta Antibody Design (Kaufmann *et al.*, 2010) and OptMAVEn (Li *et al.*, 2014). For *de novo* designs, the design campaign starts with a target antigen or an antibody sequence whose structure is not yet available. Subsequently, structural models are built for function prediction and subsequent mutations are introduced for improving the antibody. The designed candidate can then undergo further re–design by mutating the sequence.

independently, and have also been used as separate applications for various problems. This Thesis is focussed on developing these components to ultimately develop a fully automated, computational antibody design pipeline.

We have recently built SAbPred (Dunbar *et al.*, 2016), a computational antibody structural prediction suite. SAbPred is a collection of tools that are aimed at facilitating antibody design. For instance, SAbPred allows users to model antibody structures (ABodyBuilder, Chapter 4) and submit the model for subsequent predictions, *e.g.* paratope predictions by i–Patch (Krawczyk *et al.*, 2013). Unlike Rosetta and OptMAVEn, SAbPred currently does not offer a method for mutating antibody structures and sequences; in other words, it is not a design pipeline *per se*. However, a long–term aim for SAbPred is to introduce such design features, *e.g.* mutation of amino acids by PEARS (Chapter 5).

1.5.2 Antibody structure prediction

Predicting a protein's structure is key to understanding and designing proteins (Martí-Renom *et al.*, 2000). Likewise, having a structural model of an antibody is valuable for optimising many antibody design objectives (Section 1.4.2, Almagro *et al.*, 2014). Experimentally solving an antibody's structure is not always feasible; for example, obtaining a crystal for X–ray crystallography can be difficult. Thus, models provide a valuable source of data in these scenarios. Structural predictions can also help users investigate the impact of mutations or grafting loops. Often, modelling efforts have focussed on the Fv (*e.g.* Marcatili *et al.*, 2014; Sivasubramanian *et al.*, 2009), though there have been several investigations that focussed on predicting the CDR loops (*e.g.* Choi and Deane, 2011; Messih *et al.*, 2014), or the V_H–V_L orientation (Bujotzek *et al.*, 2015b).

Structure prediction can be broadly classified as 'template–free' or 'template–based' methods; both types of methods have been applied to antibody structure prediction. The latter is more widely used as antibody structures are highly similar and consistently share high sequence identity.

1.5.2.1 Template–based modelling of protein structures

Template–based, or homology modelling techniques, are based on finding an appropriate 'template' structure for a 'target' protein sequence (Martí-Renom *et al.*, 2000). The underlying assumption of homology modelling is that proteins with similar amino acid sequences tend to have similar structures. Template–based approaches can be used to model an entire protein (*e.g.* Šali and Blundell, 1993), or parts of a protein (*e.g.* loop regions, Deane and Blundell, 2001).

Template search and identification The initial step in a homology modelling pipeline is to search for a structural template. Templates can be selected from the entire PDB (Berman *et al.*, 2000) or a curated subset (*e.g.* Bujotzek *et al.*, 2015a). There are two main methods of searching a template. First, the target protein's sequence is compared to the sequences of proteins with known structures.

Typically, the structure with the highest sequence identity or sequence similarity is selected (Fiser, 2010). Alternatively, templates can be identified by structural alignment (*e.g.* Holm and Sander, 1995). For some tools (*e.g.* MODELLER, Šali and Blundell, 1993), multiple structural templates can be used to generate the model structure.

Model building Once a template is found, the model structure can be generated by several methods. Models can be formed by using spatial restraints (*e.g.* MODELLER Šali and Blundell, 1993). Here, the template(s) structure's profile (*e.g.* dihedral angles, hydrogen bonds) is used to place constraints on building a model for the target sequence. Effectively, the model is built with the aim of maximally satisfying all restraints (Fiser, 2010).

Another method for building models is 'rigid body assembly'. The aligned region of the template and target is copied from the template structure; this is used as a 'core' scaffold for further model building (*e.g.* Schwede *et al.*, 2003). The core scaffold can be a single template, a weighted average structure (Sutcliffe *et al.*, 1987), or a series of fragments that correspond to aligned regions between the target and template (*e.g.* Zhang and Skolnick, 2004a).

Subsequently, non–aligned segments, *e.g.* loops, are modelled; the predicted loop structure is 'grafted' between two 'anchor' points of the model structure (Figure 1.16). Loop prediction can be done *ab initio* (*e.g.* Šali and Blundell, 1993; Zhang and Skolnick, 2004a), or by a database search method (*e.g.* Fernandez-Fuentes *et al.*, 2006; Deane and Blundell, 2001). For example, FREAD is a database search method that selects fragments based on anchor Cα separations (Deane and Blundell, 2001). Predicted structures, or 'decoys', are filtered according to their backbone environment–specific substitution score (ESSS), and are ranked by their RMSD (Appendix C.3) to the anchors.

Once the core and loops are modelled, if they have not yet been added during these stages, the side chains are then added to complete the model (Martí-Renom *et al.*, 2000). When predicting the side chains, the aim is to place the side chains

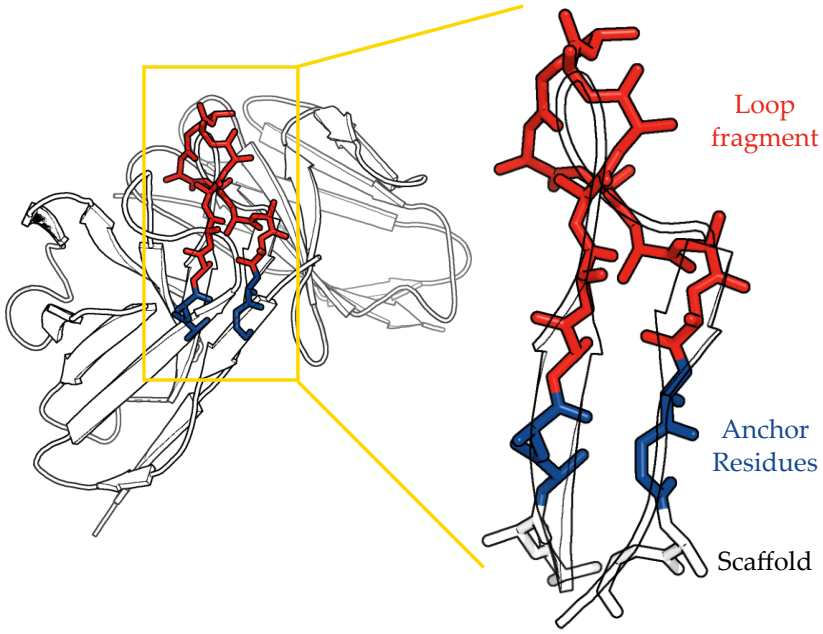


Figure 1.16: When modelling loops using template–based approaches, it is common to select a prediction fragment (red) to be grafted onto the scaffold (white) via the anchor residues (blue). For methods such as FREAD (Deane and Blundell, 2001), the anchors are defined as the two residues flanking the loop. Predictions, or decoys, are filtered by their anchor RMSD to the scaffold structure's anchors.

in the correct configuration, whilst avoiding van der Waal's clashes. Side chain prediction methods rely on a database of side chain conformations, known as a 'rotamer library' (*e.g.* Shapovalov and Dunbrack, 2011; Towse *et al.*, 2016). These libraries provide conditional probabilities that an amino acid's side chain would adopt a particular conformation, given certain structural features, *e.g.* the ϕ/ψ angles of the protein backbone (Shapovalov and Dunbrack, 2011).

Measurement of model accuracy In order to measure the accuracy of structural prediction, it is common to use the RMSD between two sets of coordinates, *i.e.* the native and model structures' coordinates (Appendix C.3). A caveat of the RMSD value is that it does not follow the rules of triangle inequality; furthermore, a single RMSD value can be generated from multiple possible predictions. Other methods are also used for evaluating model quality, such as the TM or GDT scores (Zhang and Skolnick, 2004b; Zemla, 2003); however,

RMSD is much more widely used for historical reasons and its simplicity.

Despite the success of template–based modelling methods, especially in antibody modelling (Almagro *et al.*, 2011, 2014), they have several limitations. Template choice, for example, has a major influence on the quality of the final model (Martí-Renom *et al.*, 2000). Furthermore, template–based modelling is limited by the database and can have poor coverage. For example, when predicting loop structures, a suitable template structure may not be available, in which case no prediction may be made (Choi and Deane, 2011).

1.5.2.2 Template–based modelling of antibody structures

The Antibody Modelling Assessment competitions in 2011 and 2014 were held to benchmark and compare antibody modelling methods (Almagro *et al.*, 2011, 2014). In the second competition (AMA–II), seven methods were benchmarked on their ability to model eleven Fvs as a blind test. The competition was divided into two stages: modelling the entire Fv (stage I), or only the CDRH3 loop, given the crystal structure of the remaining Fv (stage II). In stage I, most Fvs were modelled relatively well (average RMSD of the Fv backbone: 1.1Å), though the CDRL1, CDRL3, CDRH1, and the CDRH3 loops were often modelled with lower accuracy.

Three of the methodologies (RosettaAntibody, PIGS, and Kotai Antibody Builder) are freely available, whereas the others (Macromoltek, CCG, Accelrys, Schrodinger) are commercial. For each of these methods, they follow a four–stage workflow, with minor variations in the steps. Most pipelines are automated, though some allow manual intervention (Marcatili *et al.*, 2014; Fasnacht *et al.*, 2014). Initially, a template structure is chosen for the target antibody, either for the V_H and V_L domains separately, or for both domains combined. The V_H–V_L orientation is then modelled after choosing the framework template if necessary. In the third stage, the canonical CDR loops are modelled, followed by CDRH3. The models may also be refined. A simplified overview of our antibody modelling tool, ABodyBuilder (see Chapter 4), is shown in Figure 1.17.

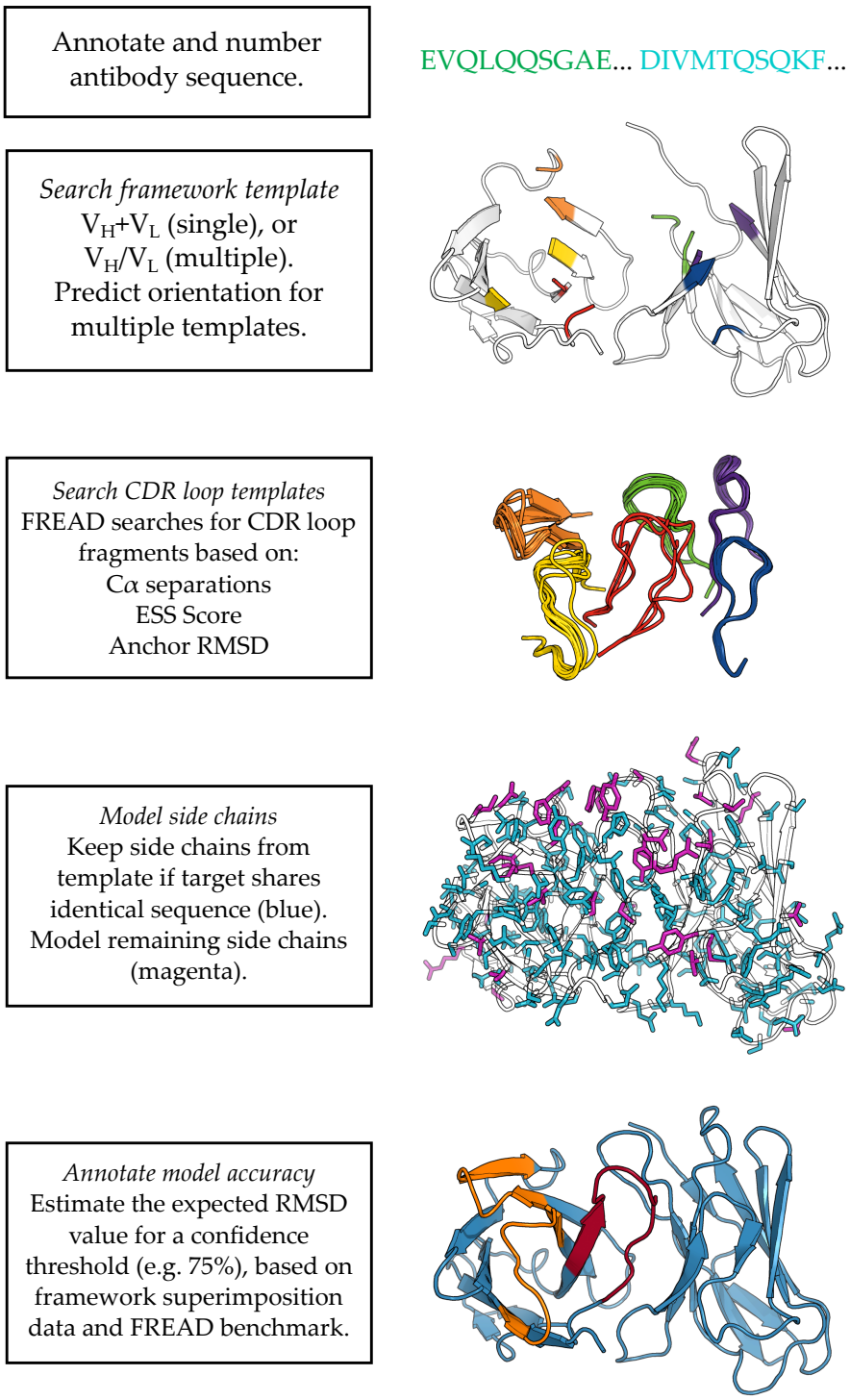


Figure 1.17: Antibody structure prediction typically involves four steps: framework template selection, orientation prediction, CDR loop prediction, and refinement (*e.g.* side chain prediction). This figure shows how ABodyBuilder predicts an antibody structure from sequence; other pipelines (*e.g.* PIGS, Marcatili *et al.*, 2014) follow a similar procedure. Unlike other antibody modelling tools, ABodyBuilder can estimate the model's accuracy as a conditional probability (Chapter 4).

Framework region templates are searched on the basis of sequence identity or sequence similarity. Pipelines often use a single framework template (*e.g.* Berrondo *et al.*, 2014; Shirai *et al.*, 2014) though some offer the option to use multiple templates (*e.g.* Marcatili *et al.*, 2014). MoFvAb is an exception, which uses a fragment–based method to assemble the V_H and V_L domains (Bujotzek *et al.*, 2015a).

Similar to framework region prediction, canonical CDRs are predicted by searching templates with sufficient sequence identity or similarity, often with a loop length restriction (*e.g.* Fasnacht *et al.*, 2014; Maier and Labute, 2014; Zhu *et al.*, 2014). Some pipelines also allow users to input the canonical class information for template selection (*e.g.* Marcatili *et al.*, 2014; Shirai *et al.*, 2014). Other methods use geometric criteria, *e.g.* loop anchor geometry, to predict the CDR loops' structures (*e.g.* Choi and Deane, 2011). The CDRH3 loop is often modelled last, and independently by database (*e.g.* Choi and Deane, 2011), machine learning (*e.g.* Messih *et al.*, 2014), *ab initio* (*e.g.* Zhu and Day, 2013; Finn *et al.*, 2016), or hybrid methods that combine database and *ab initio* methods (Marks *et al.*, 2016).

In model refinement, side chains are placed given the model's backbone geometry. This is typically done by side chain prediction methods (*e.g.* Krivov *et al.*, 2009), and/or energy minimisation (*e.g.* Zhu *et al.*, 2014; Sivasubramanian *et al.*, 2009). Through refinement, interatomic clashes are resolved to yield a physically sensible structure.

Prediction of Immunoglobulin Structures (PIGS) PIGS, by default, chooses a single antibody template with the highest sequence identity. PIGS also allows manual template selection along with various options. For instance, a user can choose to select a single framework template, as long as it has the same canonical CDR loops as the target sequence. PIGS also offers the ability to use multiple templates. Next, users can specify the method for modelling the six CDR loops. For example, the template antibody's CDR loops can be used if they

share canonical classes with the target. In the case of CDRH3, 'canonical' forms refer to whether the CDRH3 is predicted to have a bulge structure (Marcatili *et al.*, 2014). By default, the CDR loop with an identical canonical structure and maximal sequence similarity is used. Orientation is predicted thereafter based on a fingeprint of residues in the framework region. Finally, side chains are modelled by SCWRL (Krivov *et al.*, 2009). PIGS is available as both a standalone tool and web application (Marcatili *et al.*, 2014).

RosettaAntibody RosettaAntibody is a tool within the Rosetta modelling suite. It first selects two structural templates for the framework region based on the BLAST bit–score (Altschul *et al.*, 1990); the Fv structure with the highest sequence similarity is used as the orientation template. Templates for the canonical CDR loops are searched by a length–specific BLAST search, again using the template with the highest bit score. The CDRH3 loop is then modelled *de novo* by assembling three–residue long fragments. Simultaneously, the V_H–V_L orientation is adjusted by Rosetta's docking algorithm. Finally, the model structures are refined by gradient–based minimisation (Sivasubramanian *et al.*, 2009). RosettaAntibody is also available as a web application and a standalone tool.

Kotai Antibody Builder Kotai Antibody Builder selects two templates for the framework region based on weighted sequence identity, structural motifs (*e.g.* the framework region motif), and the MolProbity score (Yamashita *et al.*, 2014). The non–H3 CDR loops are selected based on their canonical class and a position–specific substitution matrix (PSSM) score. For the CDRH3 loop, structural rules are used along with the PSSM score for template search (Shirai *et al.*, 1999). Models are then built using MODELLER, and side chains are predicted using OSCAR–star (Liang *et al.*, 2011). Finally, models are minimised using myPresto (Shirai *et al.*, 2014). Currently, Kotai Antibody Builder is only available as a web application.

Macromoltek Macromoltek's SmrtMolAntibody tool is similar to RosettaAntibody. It selects two framework region templates and the canonical CDR loop templates based on BLAST; the CDRH3 is then modelled *ab initio* (Berrondo *et al.*, 2014). The V_H–V_L orientation is based on the orientation from the template light chain's native structure. Side chains are predicted using an energy function, and the structure is refined by energy minimisation.

Schrodinger Schrodinger's antibody modelling tool starts with a single template, based on a sequence similarity search. For all six CDR loops, the loops are clustered according to their anchor RMSD. The CDR loop with the highest sequence similarity is used as the representative. Subsequently, the side chains are predicted, and the structure undergoes energy minimisation by an implicit solvent energy model (Zhu *et al.*, 2014).

Chemical Computing Group (CCG) The CCG 'autoFv' pipeline selects one or two framework region templates based on sequence identity. All six CDR loops are selected by sequence similarity. The V_H–V_L orientation is set as the template's orientation if both V_H and V_L frameworks are from the same structure. Otherwise, the orientation of the highest–ranking single (V_H+V_L) template is used. Using the MOE homology modeller, the CDR loops are grafted onto the framework template, followed by side chain prediction and energy minimisation. Ten initial models are built, and clustered to build a final 'consensus' structure (Maier and Labute, 2014).

Accelrys Accelrys' antibody modelling pipeline offers multiple options for selecting the framework region template. Based on sequence similarity, either a single V_H+V_L template, a 'chimaeric' template from two different antibodies, or a consensus template from five different antibodies can be used. The canonical CDR loops were modelled using sequence–similar templates; if the template framework has an identical CDR loop, the template's CDR loop is used. The CDRH3 loop can either be modelled using the most sequence–similar template,

or *ab initio* using Looper. Finally, the model structure is refined via CHARMm energy minimisation (Fasnacht *et al.*, 2014).

MoFvAb MoFvAb is a fragment–based antibody modelling method (Bujotzek *et al.*, 2015a). It searches for a template for each of the seven regions of the antibody: the four framework regions (FR1–FR4), and the three CDR loops. Templates are ranked by four filters, such as sequence similarity, and the template region's RMSD to a consensus Fv structure. MoFvAb also offers manual intervention, and the option to model CDRH3 loops *de novo*. Side chains are then predicted by a neighbourhood–based method; a 4.0Å shell is placed around a query position, and chemically similar environments are searched to model the side chains. The orientation is then predicted by a random forest regression (Bujotzek *et al.*, 2015b), and the model undergoes energy minimisation.

1.5.3 Antibody sequence annotation

Numbering antibody sequences facilitates the comparison of two (or more) antibody sequences. For instance, sequence identity can be calculated by counting the number of matches between aligned IMGT positions. Numbering also helps to build multiple sequence alignments, and identify amino acid preferences at specific positions (*e.g.* Li *et al.*, 2014). These preferences can then be used to guide mutations for design.

Two common methods for numbering are Abnum and ANARCI (Abhinandan and Martin, 2008; Dunbar and Deane, 2016). Abnum numbers sequences by using the conserved anchors of the CDR loops as the starting points. In contrast, ANARCI uses a set of HMMs to number sequences (Eddy, 2004; Dunbar and Deane, 2016).

1.5.4 Antibody binding prediction methods

Computational methods have also been developed to predict how antibodies bind to their antigen. For instance, the paratope (*e.g.* Kunik *et al.*, 2012; Krawczyk *et al.*, 2013) and epitope (*e.g.* Krawczyk *et al.*, 2014) residues are predicted in order to enhance the accuracy of antibody–antigen docking protocols.

Another area of interest in computational antibody design is predicting antibody–antigen binding affinities via scoring functions (Kitchen *et al.*, 2004). Scoring functions are based on empirical or statistical parameterisations of structural data in the PDB. These methods collect structural features, *e.g.* frequencies of interatomic contacts, to evaluate a score, and thus predict binding affinity. In the context of computational affinity maturation, an accurate scoring function can be used to select for mutations which would ultimately enhance an antibody's binding affinity.

Traditionally, scoring functions have been designed for predicting the binding affinities of all types of protein–protein complexes (*e.g.* Moal and Fernandez-Recio, 2013; Vangone and Bonvin, 2015). Most methods have poor correlations to binding affinities (Kastritis and Bonvin, 2010), which may be due to the non–specific nature of these scoring functions (Ross *et al.*, 2013). In Chapter 2, we describe our knowledge–based, antibody–specific statistical potential for predicting antibody–antigen binding affinities.

1.6 Thesis overview

This Chapter is an introduction to the work presented in the Thesis, with a synopsis on antibody structures, biology, and therapeutic antibody design. The rest of the Thesis is composed of the following Chapters.

1.6.1 Chapter 2, Affinity prediction

Chapter 2 discusses the work on predicting the binding affinities of antibody–antigen complexes. Here we describe CAPTAIN, a statistical potential that uses

the weighted interatomic contacts between antibodies and antigens. Despite the correlation between binding affinities of antibody–protein complexes and their CAPTAIN scores, the method does not apply to antibody–peptide complexes. Further, the data sparsity in training CAPTAIN has prompted us to approach other aspects of antibody design, namely modelling.

1.6.2 Chapter 3, V_H–V_L pairing

Chapter 3 is an investigation on the V_H–V_L pairing of antibodies. We discuss the V_H–V_L pairing landscape based on antibody sequences from the public domain, and complement sequence analyses with structural data from SAbDab. We demonstrate that pairing frequencies can act as a proxy for predicting the thermal stability of antibodies. We also propose a structure–based mechanism to support random V_H–V_L pairing.

1.6.3 Chapter 4, ABodyBuilder

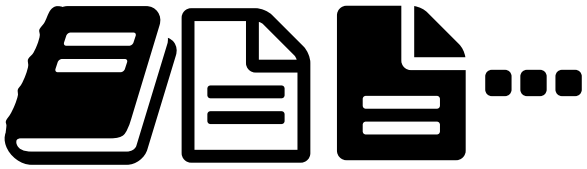
This Chapter describes our antibody modelling pipeline, ABodyBuilder (Leem *et al.*, 2016). The tool is able to model antibodies with comparable accuracy to other publicly–available methodologies, and provide estimates of model accuracy. Further, ABodyBuilder identifies sequence motifs that can potentially cause issues during experimental development. We also provide a commentary on the scalability of ABodyBuilder, especially in light of analysing NGS datasets.

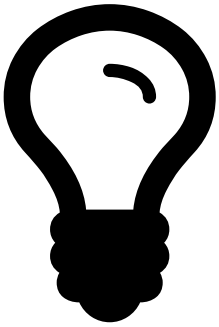
1.6.4 Chapter 5, Side chain prediction

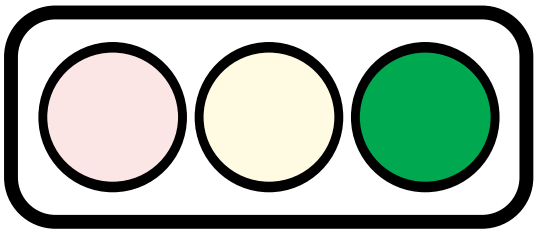
This Chapter describes our antibody–specific rotamer library, PEARL. We first describe the rotamer distributions of antibody side chains. Given the tendency for some amino acids' side chains to adopt a limited number of conformations, we use this information to model side chains on crystal and model antibody structures using our algorithm, PEARS. We comment on how side chain prediction can facilitate computational antibody design, and on some of the current algorithmic challenges for PEARS.

1.6.5 Chapter 6, Closing remarks

In Chapter 6, we conclude and summarise the Thesis, and place its findings in the context of therapeutic antibody design. We also comment on extensions for the work, with a particular interest in NGS and humanisation.







Failure taught me things about myself that I could have learned no other way.

— J.K. Rowling

2

A knowledge–based framework for antibody affinity prediction.

Contents

2.1 Introduction

In the previous Chapter, we reviewed the importance of antibodies, particularly as a platform for designing new biotherapeutics. A key design challenge is computational affinity maturation, where antibodies are mutated *in silico* to achieve a higher binding affinity. In this Chapter, we discuss a methodology for predicting the affinity of antibody–antigen interactions for use in computational affinity maturation.

2.1.1 Computational antibody affinity maturation

Natural affinity maturation can be experimentally emulated via iterative mutagenesis and selection, leading to an antibody with a desired set of binding properties. Identifying 'favourable' mutations for affinity maturation is a huge combinatorial problem (Barderas *et al.*, 2008). Typically, there are ∼60 residues across the six CDR loops. If each position can mutate to any of the 20 standard amino acids, this alone leads to 20^{60} antibody combinations.

The push to develop a high–throughput *in silico* framework is the potential to reduce this experimental search space. There exist a small number of previously published computational methods that can identify positions for mutation (*e.g.* Barderas *et al.*, 2008), and predict the effects of mutations (*e.g.* Lippow *et al.*, 2007; Kiyoshi *et al.*, 2014).

The simplest computational affinity maturation pipeline would first require the structure of the antibody–antigen complex (experimental or model). The pipeline would then assess the binding affinity of the complex, and predict how mutations would change the antibody's binding affinity. Once favourable mutants are identified, mutations can then be validated experimentally.

Following this paradigm, three studies have previously attempted computational affinity maturation on crystal structures. Clark *et al.* (2006a) matured the anti–VLA1 antibody AQC2 and achieved a ten–fold improvement. They also crystallised the mutant antibody to illustrate the structural impact of their four suggested mutations. Lippow *et al.* (2007) experimentally validated their predictions on two case studies, demonstrating a 140–fold affinity improvement for the anti–lysozyme antibody D44.1, and a 10–fold improvement for cetuximab. Their method was computationally expensive; testing 1080 single mutants of D44.1 required 24 hours on a 100–CPU cluster. Kiyoshi *et al.* (2014) used software packages from MOE and Discovery Studio to improve the 11K2 antibody's affinity by 4.6–fold. They proposed twelve mutations to increase affinity, though experimental validation showed that only five of these mutations were beneficial.

The central problem in these studies was the ability to predict the antibody's binding affinity. For example, one of the single mutants of D44.1 from Lippow *et al.* (2007) had a calculated affinity change ($\Delta\Delta G$) of –0.97kcal·mol^{-1} and an experimental $\Delta\Delta G$ of 0.45kcal·mol^{-1}. This difference of 1.42kcal·mol^{-1} between the calculated and experimental values represents an 11–fold error in estimating K_D values (see Appendix C.2 for conversion factor). Similarly, Kiyoshi *et al.* (2014) predicted that mutating an Asn residue on the CDRL1 loop would lead to a $\Delta\Delta G$ of –15.4kcal·mol^{-1}. However, the actual affinity change was only –1.0kcal·mol^{-1}, which represents a 10^9 fold error in K_D values.

2.1.2 Affinity prediction methods

Predicting the binding affinities of protein–protein interactions (PPIs), such as antibody–antigen interactions, is a formidable challenge (Kastritis and Bonvin, 2010; Pallara *et al.*, 2013). Previously, molecular dynamics (MD) simulations, machine learning methods, and scoring functions have been used to predict binding affinities.

2.1.2.1 Molecular dynamics for affinity prediction

MD simulations calculate the trajectories of atoms and molecules as forces; the force acting on each atom can be decomposed into bonded and non–bonded terms (Durrant and McCammon, 2011). The simulations can be 'all–atom', where each atom is considered for calculating the trajectories, or 'coarse–grained', where atoms are grouped into pseudo–molecular entities (May *et al.*, 2013). The ensemble average trajectories from MD simulations can then be used to calculate the binding energy.

MD has been successful at estimating the binding energy of protein–peptide interactions. The binding affinity of the major histocompatibility complex with twelve different peptides was successfully predicted by Wan *et al.* (2015), achieving a correlation up to 0.91. However, their analysis required over 4000 cores and nine hours of computation per run. On the other hand, May *et al.*

(2013) used a coarse–grained MD method to predict the binding affinities of two protein–protein complexes. Although their coarse–grained approximation only required 30 minutes, the method was only accurate for one of the two complexes. Currently, MD is not yet an established method for predicting the binding affinities of protein–protein complexes, especially at a large scale.

2.1.2.2 Machine learning methods in affinity prediction

A wide range of machine learning methods have been used in protein–protein binding affinity prediction (*e.g.* Moal *et al.*, 2011; Yugandhar and Gromiha, 2014). Such methods often use a large number of descriptors, and can be prone to overfitting errors. For example, the random forest model by Moal *et al.* (2011) relied on 200 descriptors for a training set of 57 protein–protein complexes. Although the random forest prediction showed strong correlation to binding affinities in the training set (Pearon's $r = 0.70$), the correlation dropped to $r = 0.48$ for the entire dataset consisting of both the training and test sets (Moal *et al.*, 2011). The SVM method by Yugandhar and Gromiha (2014) used a set of nine descriptors (from an initial set of 610) to separate high and low–affinity protein–protein complexes. However, the biological significance of certain predictors, *e.g.* α–helix content, was not described in detail, making it difficult to evaluate the practicality of their methodology.

2.1.2.3 Scoring functions in affinity prediction

Scoring functions are models that use a combination of terms to estimate the energy of a protein or protein–protein complex structure (*e.g.* Vangone and Bonvin, 2015). Scoring functions have been used for a range of applications, such as structure prediction (*e.g.* Samudrala and Moult, 1998a) and docking (*e.g.* Vreven *et al.*, 2011).

Scoring functions are classified as either force field–based, empirical, or statistical (knowledge–based) potentials (Kitchen *et al.*, 2004). Force–field based methods explicitly calculate physicochemical terms based on parameters from

molecular mechanics force fields. For example, the AutoDock scoring function calculates terms from the AMBER force field (Kitchen *et al.*, 2004). Empirical scoring functions calculate a sum of weighted terms (*e.g.* electrostatics, van der Waal's energy) which have been parameterised on a set of protein structures (*e.g.* Pierce and Weng, 2007; Cheng *et al.*, 2007; Vangone and Bonvin, 2015). Knowledge–based scoring functions, also known as statistical potentials, can be used as an independent function (*e.g.* Samudrala and Moult, 1998a), or can also be part of a term in an empirical function (Vreven *et al.*, 2011). These functions reflect the probability that an observed molecular feature (*e.g.* set of interatomic contacts) appear in a structure, given a reference state (Samudrala and Moult, 1998a; Cossio *et al.*, 2012).

In a systematic evaluation of scoring functions in a protein–protein binding benchmark, it was found that no scoring function was able to predict the affinities of protein–protein interactions (Kastritis and Bonvin, 2010). FireDock scores showed the strongest correlation to binding affinities (absolute Pearson's correlation, $|r| = 0.32$), followed by other empirical functions such as PyDock ($|r| = 0.22$) and ZRANK ($|r| = 0.18$).

Since this analysis by Kastritis and Bonvin (2010), new scoring functions have been developed to predict the binding affinities of protein–protein complexes, and have had mixed success (*e.g.* Vreven *et al.*, 2012; Vangone and Bonvin, 2015). Moal and Fernandez-Recio (2013) demonstrated that scores from their interatomic contact potential showed a $|r|$ value of 0.33 with the binding affinities of protein–protein complexes in the SKEMPI dataset. Vreven *et al.* (2012)'s ZAPP function predicted the binding affinities of enzyme–inhibitor complexes ($|r| = 0.66$), but showed poor correlations to the binding affinities of antibody–antigen complexes ($|r| = 0.24$). The PRODIGY function by Vangone and Bonvin (2015) showed a four–fold cross–validated correlation of $|r| = 0.73$ to the binding affinities of 81 complexes. Despite using an Akaike Information Criterion for model selection, the final model had the most number of parameters, which may make the function susceptible to overfitting errors.

2.1.3 Developing an antibody–specific scoring function

The analysis by Ross *et al.* (2013) suggested that scoring functions should be ligand–specific. Several scoring functions have now been developed for specific types of protein–ligand complexes. For instance, SPA–PN has been designed to predict the affinities of protein–nucleic acid interactions (Yan and Wang, 2013). Antibody–specific scoring functions have previously been developed (*e.g.* Brenke *et al.*, 2012; Krawczyk *et al.*, 2014), though these functions were used to assess docking decoys, rather than predict binding affinities.

In this Chapter, we follow Brenke *et al.* (2012) and Krawczyk *et al.* (2014) and develop an antibody–specific scoring function, but in our case for affinity prediction. We introduce CAPTAIN (Computational Affinity Prediction Tool for Antibody–Antigen Interactions), which is a weighted form of the RAPDF statistical potential (Samudrala and Moult, 1998a) that was trained using only antibody–antigen complexes. Previously, Wang *et al.* (2004) have weighted RAPDF to emphasise contact types that are more common in high–quality structural decoys. Extending from this formalism, we have weighted RAPDF by two weighting schemes, following on from the works of Moal and Fernandez-Recio (2013) and Robin *et al.* (2014).

Moal and Fernandez-Recio (2013) developed a scoring function weighted by experimental $\Delta\Delta G$, whereas Robin *et al.* (2014) suggested that $\Delta\Delta G$ predictions from FoldX can identify energetically important residues in an antibody–antigen complex. Thus, we weighted RAPDF by either theoretical $\Delta\Delta G$ values or by experimental binding affinities ($\ln K_D$) of antibody–antigen complexes. These weighting schemes should highlight the contact types that are predicted to make large contributions to the binding energy, or emphasise contacts that are more prevalent in high–affinity antibodies. To our knowledge, CAPTAIN is the first scoring function that is specifically tailored for predicting the binding affinities of antibody–antigen complexes. We demonstrate that using and weighting antibody–specific information is key to successful affinity prediction.

2.2 Methods

2.2.1 Datasets

Antibody–protein complexes with resolution $\leq$4.0Å were downloaded from SAbDab in October 2013 (Dunbar *et al.*, 2014). We defined an antigen as a protein antigen if it was longer than 50 residues. Complexes were clustered at 99% antibody sequence identity and 90% antigen sequence identity by CD–HIT (Li and Godzik, 2006). For each cluster, the antibody–antigen complex with affinity information (K_D) and the best resolution was chosen as the representative member. If none of the members in a cluster had K_D data, the antibody–antigen complex with the best resolution was used. Overall, the non–redundant set used to train our antibody–specific functions consists of 241 antibody–antigen complexes, 70 of which have K_D data (Figure 2.1).

For our test set, the 170 antibody–antigen complexes with K_D data in SAbDab were taken. Of the 170 complexes, 70 were excluded as they overlapped with antibodies in the training set. An additional ten complexes were excluded as their antibodies shared $\geq$99% sequence identity to a member of the training set. The final test set features 90 complexes, 55 of which are antibody–protein complexes, and 35 are antibody–peptide complexes (Figure 2.1).

Our protein–protein complex dataset consists of 1882 non–antibody protein–protein complexes selected from the PDB using the criteria from Hamer *et al.* (2010). The protein–protein set was used to train our general protein functions. The protein–peptide complex set consists of 221 complexes from the PDB, based on the same set of conditions from Hamer *et al.* (2010); peptides were defined as PDB chains with $\leq$50 residues. Both these sets were obtained in December 2014.

The binding affinity benchmark set (AffBM set) from Kastritis *et al.* (2011) consists of 124 non–antibody protein–protein complexes (*e.g.* enzyme–inhibitor complexes) with K_D data. Four complexes (PDB: 1de4, 1m10, 1nb5, 1uug) were removed, as in Moal *et al.* (2011). The remaining 120 complexes were used

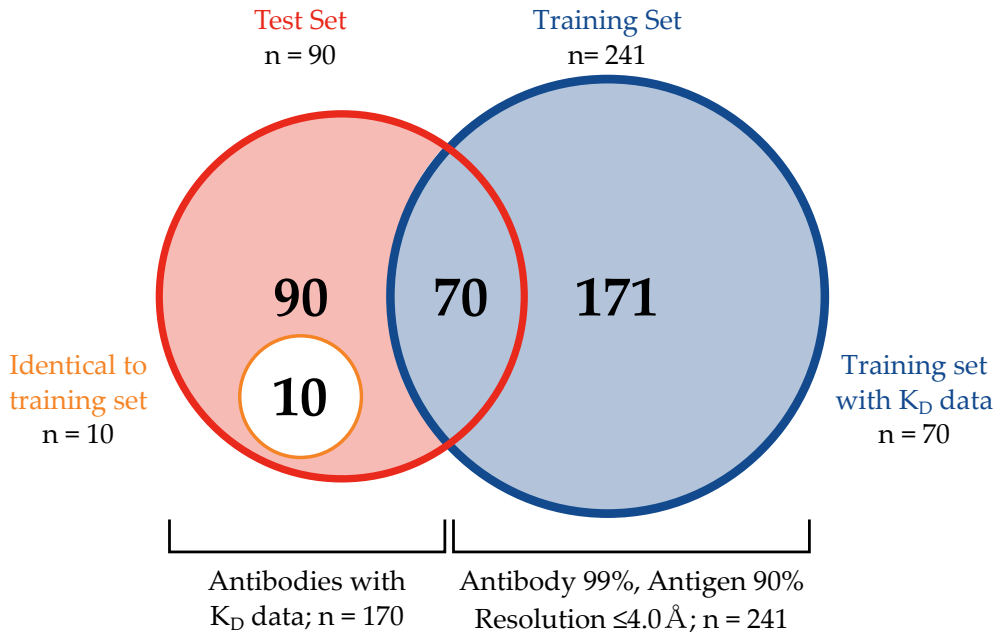


Figure 2.1: Datasets used for training and testing our antibody–specific scoring functions. The training set features 241 non–redundant antibody–antigen complexes from SAbDab, whereas the test set has 90 antibody–antigen complexes. Datasets are described in Section 2.2.1.

to test the ability of the methods on predicting the affinities of non–antibody protein–protein complexes.

2.2.2 Comparing binding interfaces

Two residues at the binding interface were considered a contact if any of their non–hydrogen atoms were within 4.5Å. Individual amino acid frequencies of antibody–protein, antibody–peptide, protein–protein and protein–peptide complexes were compared by a χ^2 test (Collis *et al.*, 2003).

2.2.3 Benchmarking methodologies for affinity prediction

We investigated 16 different scoring functions for predicting antibody–antigen binding affinities. We also considered the simple measure of the change in solvent–accessible surface area.

2.2.3.1 Change in solvent–accessible surface area

The change in solvent–accessible surface area upon complex formation, ΔSASA, was evaluated by NACCESS (Hubbard and Thornton, 1993). ΔSASA approximates the solvation effects of binding (Liang *et al.*, 2007), and has been used to approximate the binding affinities of general protein–protein interactions (*e.g.* Vangone and Bonvin, 2015). To evaluate the solvent–accessible surface area, NACCESS traces a 1.4Å–probe around the surface of an input structure. The solvent–accessible surface area was calculated for the antibody–antigen complex structure and the antibody alone. ΔSASA was then calculated as the difference between these two surface areas, *i.e.*,

$$\Delta\text{SASA} = \text{SASA}_{\text{complex}} - \text{SASA}_{\text{antibody}}.$$

2.2.3.2 Statistical potentials

We tested three different statistical potentials which rely on the DFIRE reference state (DCOMPLEX, dDFIRE, and GOAP; Liu *et al.*, 2004; Yang and Zhou, 2008; Zhou and Skolnick, 2011). DCOMPLEX uses the total sum of interatomic contact energies between two atoms within 4.5Å. dDFIRE uses the sum of contact energies over atoms within 15Å, and contains three angle terms. GOAP considers the distance and orientation of atoms, using five angle terms along with a distance term. GOAP failed to evaluate one complex (PDB: 2nz9). DockSorter is a statistical potential that uses the sum of docking precision scores (Krawczyk *et al.*, 2013). The docking precision score represents the likelihood that a docking algorithm (*e.g.* ZDOCK; Chen *et al.*, 2003) will place an antibody residue within 4.5Å of an antigen residue (Krawczyk *et al.*, 2013). DockSorter failed on three cases (PDB: 2fx8, 2fx9, 3eys).

2.2.3.3 Empirical scoring functions

We tested seven different empirical scoring functions (EMPIRE, FireDock, IRAD, ZRANK, PyDock, and Rosetta; Liang *et al.*, 2007; Andrusier *et al.*, 2007; Vreven *et al.*, 2011; Pierce and Weng, 2007; Cheng *et al.*, 2007; Chaudhury *et al.*, 2010)

EMPIRE, FireDock, IRAD, ZRANK, PyDock, and Rosetta are empirical scoring functions that are used for evaluating docking decoys, whereas PRODIGY has been specifically developed for predicting binding affinities. EMPIRE is a scoring function with parameterised terms for hydrogen bonds, electrostatics, and ΔSASA. FireDock is a decoy refinement and re–scoring algorithm; it features a set of parameterised terms for electrostatics, solvation and van der Waal's forces. For each complex, FireDock performs side chain and rigid–body optimisation prior to scoring. ZRANK is a scoring function consisting of weighted terms for electrostatics, van der Waal's forces, and desolvation. IRAD uses the terms from ZRANK, along with an additional desolvation energy term and a statistical potential term. PyDock is a scoring function that combines electrostatics, van der Waals, and desolvation terms. The Rosetta scoring function uses a combination of energy terms (*e.g.* π–π interactions, hydrogen bonds) with different weights for each term; four different weighting schemes ('score12', 'talaris2013', 'docking', and 'interface') were used. PRODIGY uses a combination of weighted terms based on the types of contacts at the binding interface, and residue characteristics at the non–interacting surface. EMPIRE failed on one case (PDB: 2nz9), FireDock failed on one (PDB: 2nz9), IRAD failed on two (PDB: 2hrp, 3ifl), and Rosetta failed on four cases (PDB: 2fx8, 2fx9, 3e8u, 3eys).

2.2.3.4 Other functions

FoldX is an empirical force field with terms for van der Waal's forces, electrostatics, solvation and hydrogen bonds (Schymkowitz *et al.*, 2005). The AnalyseComplex routine was used to estimate the binding affinity for each antibody–antigen complex. FoldX version 3.0b6 was used with default parameters.

We used Gromacs (v. 4.6.5) to perform MD energy minimisation for 500ps. Minimisation was performed in the GROMOS 53a6 force field with spc water molecules (Pronk *et al.*, 2013). Gromacs could not minimise the energies of 54 structures, due to missing atoms in the PDB structure (Appendix Table B.2). The minimised energy of the complex was used as the predictor of binding affinity.

2.2.4 Construction of CAPTAIN

CAPTAIN is a weighted statistical potential based on RAPDF (Samudrala and Moult, 1998a). RAPDF is a residue–specific, all–atom function that utilises the log ratio of interatomic contact probabilities. Atom types that were used for building our scoring function are listed in Appenidx Table B.1.

2.2.4.1 Unweighted residue–specific all–atom function (RAPDF)

Interatomic contacts were discretised into 1Å–wide bins, ranging from 3Å to 10Å. Contacts were symmetric; in other words, a contact between an antibody Tyr OH and an antigen His ND1 is the same as a contact between an antibody His ND1 and an antigen Tyr OH. Contacts within 3Å were placed into a single bin (0–3Å), leading to a total of eight distance bins (Figure 2.2). The score for an interatomic contact ij at distance bin d, $s(d_{ij})$, can be expressed as

$$[\text{RAPDF}]\ s(d_{ij}) = -\ln\left(\frac{\left(\frac{N(d_{ij})}{\sum_d N(d_{ij})}\right)}{\left(\frac{\sum_{ij} N(d_{ij})}{\sum_d \sum_{ij} N(d_{ij})}\right)}\right). \tag{2.1}$$

$N(d_{ij})$ is the number of contacts between atom types i and j in distance bin d. It then follows that $\sum_d N(d_{ij})$ represents the number of contacts for contact ij across all distance bins, whereas $\sum_{ij} N(d_{ij})$ represents the number of contacts across all contact types at distance bin d. $\sum_d \sum_{ij} N(d_{ij})$ represents the number of contacts for all contact types across all distance bins. To avoid zero counts, every contact type at every distance bin was given a default count of 1 (Samudrala and Moult, 1998a).

We built the RAPDF scores based on several different datasets. PPI–RAPDF was built from contacts between non–hydrogen atoms in non–antibody, protein–protein complexes from the protein–protein complex dataset (Section 2.2.1). Ab–RAPDF was built from contacts between the antibody and the antigen within the non–redundant set. In addition, both the protein–protein complex dataset and our non–redundant antibody set were combined to build PPI+Ab–RAPDF.

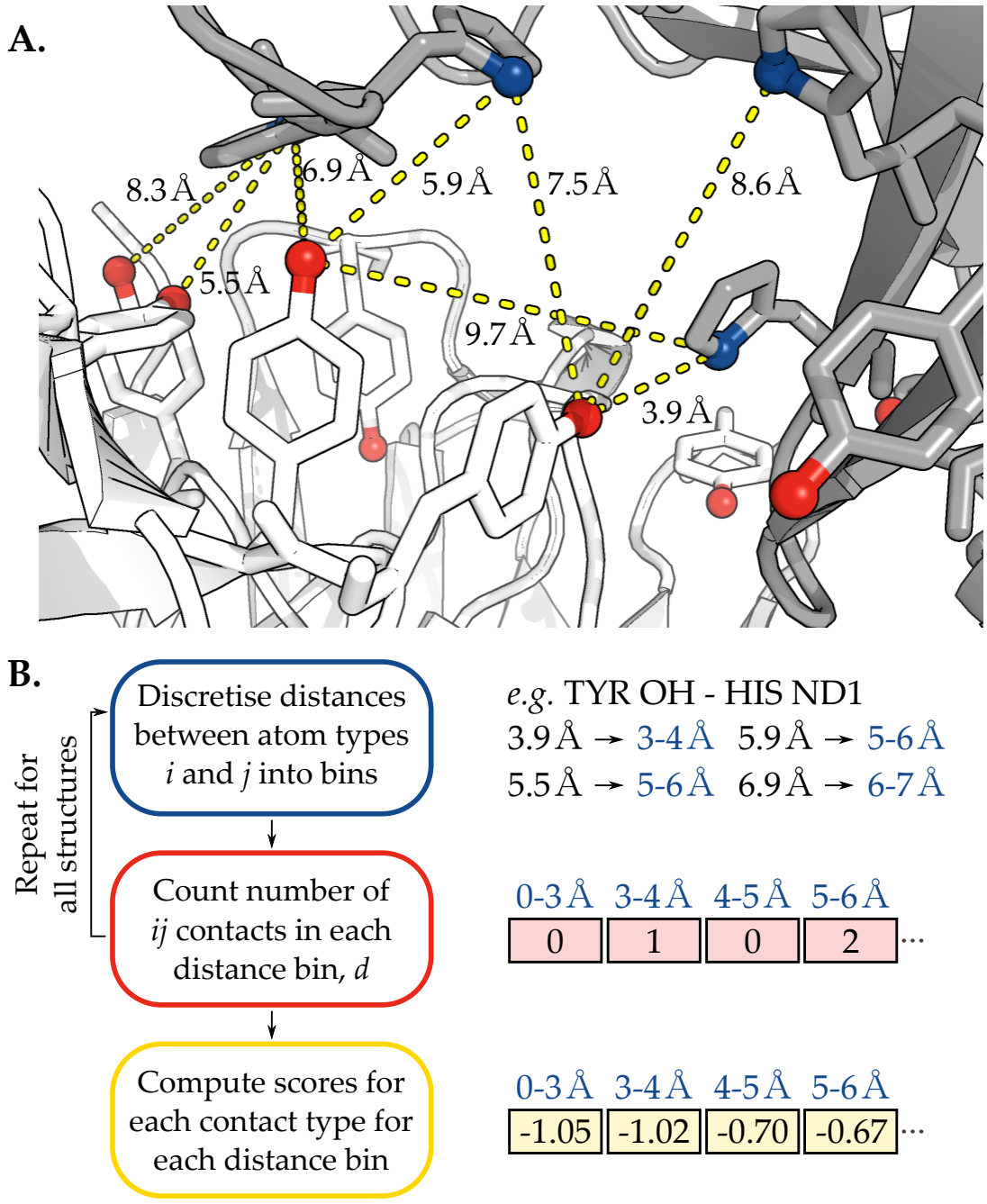


Figure 2.2: Construction of RAPDF using interatomic contacts. **A.** The antibody is shown in white, and the antigen in grey. The yellow dashes indicate contacts between Tyr OH atoms (red spheres) and His ND1 atoms (blue spheres). When building RAPDF, two atoms were considered to be in contact if they were within 10Å of each other. **B.** Every contact was discretised into 1Å wide bins, with the exception of contacts within ≤3Å, which were placed into one bin (0–3Å). In **A.**, there are eight Tyr OH–His ND1 contacts; one contact in bin 3–4Å (3.9Å), two in bin 5–6Å (5.5Å, 5.9Å), one in bin 6–7Å (6.9Å), one in 7–8Å (7.5Å), two in 8–9Å (8.3Å, 8.6Å) and one in 9–10Å (9.7Å). PDB: 3mxw.

2.2.4.2 Weighted residue–specific all–atom function

The Ab–RAPDF function was weighted by the affinities of antibody–antigen complexes in the training set. The weighted Ab–RAPDF is constructed in similar fashion to Equation 2.1. The weighted RAPDF follows Wang *et al.* (2004) by introducing a weight, w_x, to Equation 2.1.

In Wang *et al.* (2004), the weight of a decoy x was proportional to its normalised density score S_x. S_x is the average Cα RMSD between a decoy x and all other decoys in the dataset. S_x is then normalised to a value between –1 and 1 by the median S value (Wang *et al.*, 2004). The weight w_x was obtained by

$$w_x = e^{-cS_x},$$

where c is a constant. The weights were then used to weight the counts, leading to

$$s(d_{ij}) = -\ln\left(\frac{\left(\frac{\sum_x w_x N(d_{ij})_x}{\sum_d \sum_x w_x N(d_{ij})_x}\right)}{\left(\frac{\sum_{ij}\sum_x w_x N(d_{ij})_x}{\sum_d \sum_{ij} \sum_x w_x N(d_{ij})_x}\right)}\right). \tag{2.2}$$

2.2.4.3 Weighting by predicted $\Delta\Delta G$

Ab–RAPDF was first weighted using predicted affinity changes using FoldX; we refer to this version as Ab–fRAPDF (Schymkowitz *et al.*, 2005). For each antibody in the training set, an antibody residue r with at least one heavy atom within 10Å of the antigen (except alanine and glycine) was mutated to alanine, similar to Robin *et al.* (2014). The affinity change for each mutant, $\Delta\Delta G_r$, was calculated as

$$\Delta\Delta G_r = \Delta G_{\text{mutant}} - \Delta G_{\text{wild–type}}.$$

Two mutants were not used for weighting due to high values of $\Delta\Delta G$ ($|\Delta\Delta G| \geq$10kcal·mol^{-1}). To determine the FoldX–dependent weight for an antibody residue r, $w_{f,r}$, the $\Delta\Delta G$ values were first scaled to the range of 0–2 by

$$w_{f,r} = 2\frac{-\Delta\Delta G_r - \min(\Delta\Delta G)}{\max(\Delta\Delta G) - \min(\Delta\Delta G)}.$$

The $w_{f,r}$ values were then scaled as in Wang *et al.* (2004) by

$$w_{f,r} = e^{(c(w_{f,r}-1))},$$

where c is a constant; for our work, c was set to 4. Since we built RAPDF as an all–atom potential, we applied the same weight for every atom i in residue r. Thus, $w_{f,i} = w_{f,r}$. The weight was applied in a similar fashion to Equation 2.2, making the revised equation

$$\text{[Ab–fRAPDF] } s(d_{ij}) = -\ln\left(\frac{\frac{w_{f,i}N(d_{ij})}{\sum_d w_{f,i}N(d_{ij})}}{\frac{\sum_{ij} w_{f,i}N(d_{ij})}{\sum_d \sum_{ij} w_{f,i}N(d_{ij})}}\right). \tag{2.3}$$

2.2.4.4 Weighted by Experimental Affinity Data

Ab–RAPDF was also weighted using affinity–based weights, which are specific for each antibody–antigen complex in the training set, a. This form of Ab–RAPDF will be referred to as Ab–wRAPDF. In this scheme, every contact ij in complex a was given the same weight, $w_{k,a}$. The weights were set by

$$w_{k,a} = \frac{\ln K_{D,a} - \max(\ln K_D)}{0.5 \times \max(\ln K_D)}$$

where $K_{D,a}$ represents the affinity of an antibody–antigen complex a. Antibody–antigen complexes in the training set without affinity data (171 complexes) were given a $w_{k,a}$ value of 1. We revise Equation 2.2 to

$$\text{[Ab–wRAPDF] } s(d_{ij}) = -\ln\left(\frac{\frac{\sum_a w_{k,a}N(d_{ij})_a}{\sum_d \sum_a w_{k,a}N(d_{ij})_a}}{\frac{\sum_{ij} \sum_a w_{k,a}N(d_{ij})_a}{\sum_d \sum_{ij} \sum_a w_{k,a}N(d_{ij})_a}}\right). \tag{2.4}$$

Both the FoldX–dependent ($w_{f,r}$) and affinity–dependent ($w_{k,a}$) weights were used to create a combined weighted scoring function, leading to

$$\text{[Ab–wfRAPDF]}\ s(d_{ij}) = -\ln\left(\frac{\frac{\sum_a w_a w_{f,i} N(d_{ij})_a}{\sum_a \sum_d w_{f,i} N(d_{ij})_a}}{\frac{\sum_a \sum_{ij} w_{f,i} N(d_{ij})_a}{\sum_a \sum_d \sum_{ij} w_{f,i} N(d_{ij})_a}}\right). \tag{2.5}$$

To account for cases that have more contacts, a structure's total score was divided by its number of contacts.

2.2.4.5 Hybrid scoring method

The 'hybrid score' method allows an observed distribution of interatomic contacts, h_2, to be complemented with a saturated 'reference' distribution of interatomic contacts, h_1 (Sippl, 1990; Studer *et al.*, 2014). Mixing the two distributions depends on a convergence parameter, σ. This allows h_1 and h_2 to be mixed more finely, as opposed to a simple sum (as in the case for PPI+Ab–RAPDF). Thus, the target contact distribution g_{ij} between atom types i and j is approximated by

$$g_{ij} \approx \frac{1}{1+N\sigma} h_{1,ij} + \frac{N\sigma}{1+N\sigma} h_{2,ij}.$$

The negative log ratio of g_{ij} with respect to $h_{1,ij}$ represents the hybrid score, $hs(d_{ij})$. Thus,

$$\begin{aligned} hs(d_{ij}) &= -\ln\left(\frac{g_{ij}}{h_{1,ij}}\right) \\ &= \ln(1+N\sigma) - \ln(h_{1,ij} + N\sigma h_{2,ij}). \end{aligned} \tag{2.6}$$

We used the contact data from the protein–protein complex dataset (1882 structures) as $h_{1,ij}$. Our affinity–weighted set of contacts from the non–redundant training set (241 complexes) was used for $h_{2,ij}$. Higher values of σ increase the contribution of $h_{2,ij}$ to $hs(d_{ij})$; in contrast, lower values of σ favour the saturated dataset.

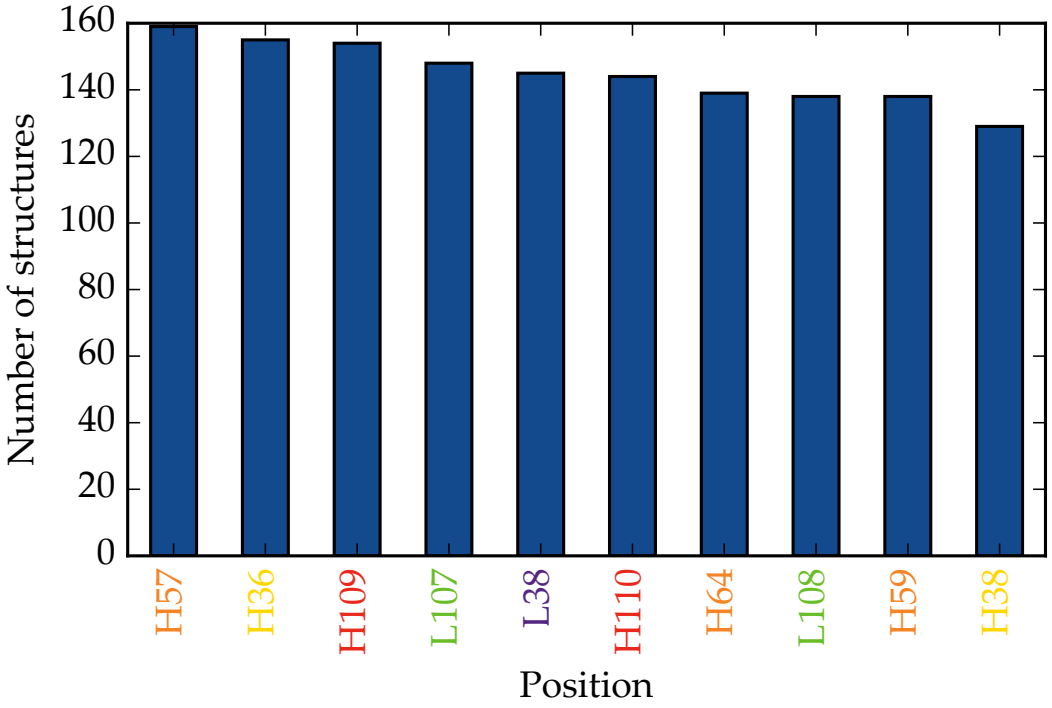


Figure 2.3: The ten most frequently–contacting positions in 214 paired antibodies of our non–redundant training set. A position was considered to be in contact if any of its heavy atoms were within 4.5Å of the antigen. Positions are coloured according to the scheme in Figure 1.3.

2.3 Results

2.3.1 Statistical analyses of binding interfaces

In our training set of 241 antibody–protein complexes, 214 are Fv structures. Of these, 180 use at least one residue from both the CDRH3 and CDRL3 loops to form a contact with the antigen (*i.e.*, within 4.5Å of the antigen). Seven of the most frequent contact positions were from V_H CDR loops, and three were from V_L CDR loops (Figure 2.3), suggesting that paired antibodies use a combination of V_H and V_L CDR loops to bind an antigen.

A χ^2 test showed that, overall, antibody–protein complexes had different amino acid frequencies compared to antibody–peptide, protein–protein and protein–peptide complexes (Table 2.1). In terms of individual amino acids,

Table 2.1: χ^2 test of amino acid usages at the binding interface.

	Antibody–peptide complexes	Protein–protein complexes	Protein–peptide complexes
Antibody–protein complexes	57.89 ($p = 8.30 \times 10^{-6}$)	4187.31 ($p < 10^{-6}$)	1084.50 ($p < 10^{-6}$)
Antibody–peptide complexes		1490.64 ($p < 10^{-6}$)	542.28 ($p < 10^{-6}$)
Protein–protein complexes			160.21 ($p < 10^{-6}$)

p–values were calculated based on a χ^2 distribution with 19 degrees of freedom.

antibody–protein and antibody–peptide complexes only showed a significant difference in their usage of His. When comparing the antibody–protein set and the protein–protein complex set, significant differences ($p \leq 2.63 \times 10^{-5}$) were observed for 17 amino acid types (Figure 2.4). The general difference is that antibodies use polar amino acids such as Tyr and Ser, whereas protein–protein complexes use hydrophobic and charged amino acids, such as Leu, Lys and Glu. These differences will be automatically captured in our antibody–specific potential.

2.3.2 Affinity prediction

The crystal structures of the 90 antibody–antigen complexes in the test set were scored with 16 different scoring functions (Section 2.2.3). We also calculated the change in solvent–accessible surface area (ΔSASA). We used the absolute Pearson's correlation coefficient ($|r|$) to measure the correlations between a score (or ΔSASA) and binding affinities ($\ln K_D$).

2.3.2.1 Benchmarking the performance of scoring functions

All methods had low correlations to $\ln K_D$ (Table 2.2). The best predictors were IRAD ($|r| = 0.264$), and ΔSASA ($|r| = 0.311$). These correlations compare to $|r| = 0.23$ for IRAD and $|r| = 0.54$ for ΔSASA when they were used to predict the affinities of general protein–protein complexes (Vreven *et al.*, 2012; Kastritis

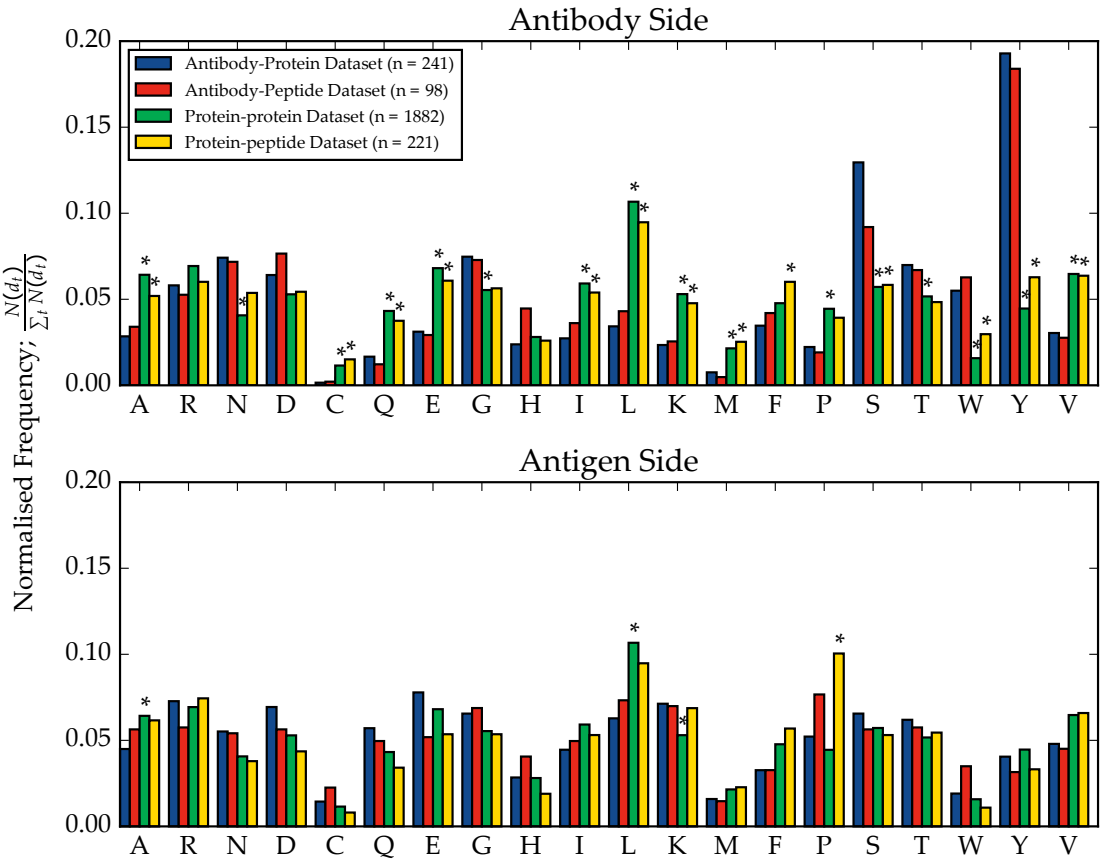


Figure 2.4: Normalised amino acid frequencies in the binding interfaces of each dataset. $N(d_t)$ represents the number of times that amino acid of type t is found within distance d (4.5Å), and $\sum_t N(d_t)$ represents the number of times that we see all amino acid types within d. Asterisks indicate significant differences ($p \leq 1.32 \times 10^{-6}$) with respect to the antibody–protein dataset.

et al., 2011). Subdividing the test set into protein–binders and peptide–binders also gave low $|r|$ values (Figure 2.6; Table 2.2).

2.3.2.2 Performance of standard RAPDF

RAPDF was first trained on the protein–protein dataset; this version, PPI–RAPDF, showed a correlation of $|r| = 0.292$ with antibody–antigen binding affinities (Figure 2.6, Table 2.3). ΔSASA and IRAD showed stronger correlations, suggesting that a standard RAPDF, without any antibody information, is poor at estimating antibody–antigen binding affinities.

Table 2.2: Absolute Pearson's correlation of predictions and binding affinity.

Method	All test set antibodies (n = 90)	Protein–binding antibodies (n = 55)	Peptide–binding antibodies (n = 35)
DCOMPLEX	0.224	0.001	**0.355**
dDFIRE	0.053	0.041	0.077
DockSorter	0.154	0.033	0.104
EMPIRE	0.073	0.065	0.227
FireDock	0.123	0.033	0.171
FoldX	0.083	0.039	0.141
GOAP	0.170	0.071	0.085
Gromacs	0.075	0.050	NA
IRAD	0.264	0.186	0.287
ΔSASA	**0.311**	0.057	0.336
PRODIGY	0.115	**0.204**	0.280
PyDock	0.026	0.012	0.106
Rosetta (Docking)	0.098	0.085	0.063
Rosetta (Interface)	0.117	0.076	0.209
Rosetta (Score12)	0.086	0.069	0.046
Rosetta (Talaris2013)	0.085	0.085	0.048
ZRANK	0.203	0.196	0.229

2.3.2.3 Antibody information improves performance

We next trained RAPDF on the non–redundant set of antibodies from SAbDab (Section 2.2.4); this version, Ab–RAPDF, showed a correlation of $|r| = 0.411$ to binding affinities. This improvement in performance supports the notion that training RAPDF on antibody–antigen complexes is critical. Furthermore, Ab–RAPDF scores showed a correlation of $|r| = 0.492$ to the subset of antibody–protein complexes (Table 2.3).

A major issue for Ab–RAPDF has been the lack of data in the training set. The most noticeable effect of a sparse set of counts is the tendency to have extreme values of $s(d_{ij})$, especially in the first two distance bins (0–3Å, 3–4Å; Figure 2.5).

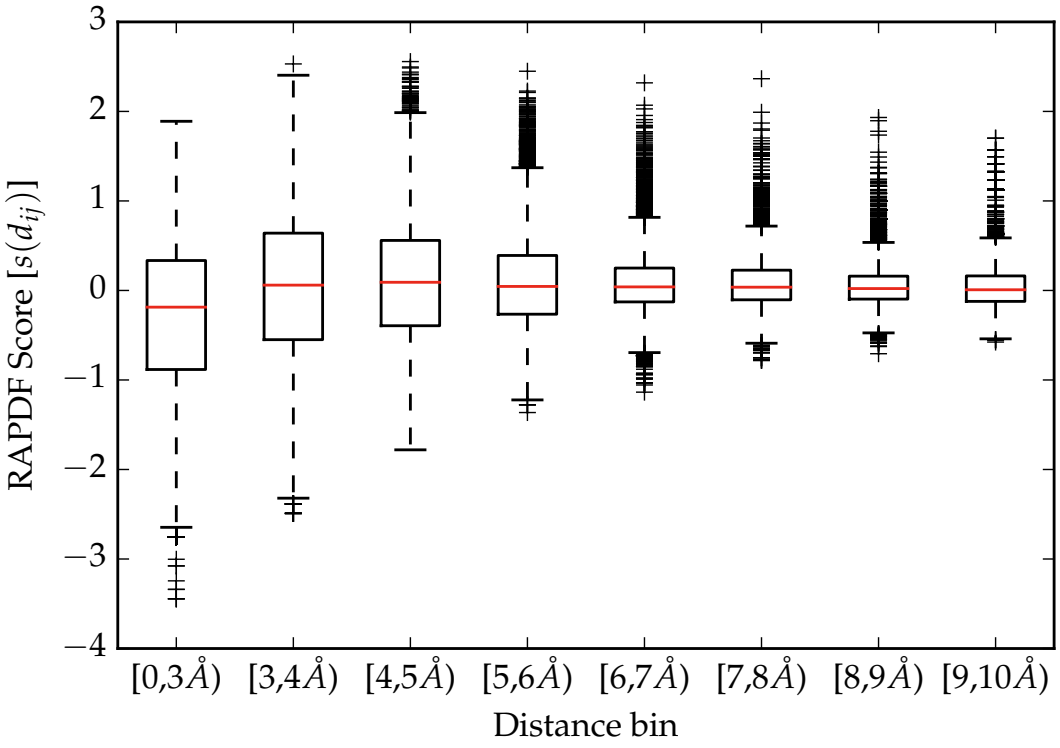


Figure 2.5: Ab–RAPDF score distributions depending on the distance bin. In the first two distance bins, the interquartile ranges for the RAPDF score distributions are 1.218 (0–3Å) and 1.190 (3–4Å). In comparison, the last distance bin (9–10Å) has an interquartile range of 0.284.

Thus, we trained RAPDF using a combination of complexes (PPI+Ab–RAPDF) to increase the size of our training set; however, PPI+Ab–RAPDF showed lower correlations with affinity ($|r| = 0.313$).

2.3.2.4 Weighting by affinity enhances prediction

So far, we have presented three different versions of RAPDF: PPI–RAPDF, Ab–RAPDF, PPI+Ab–RAPDF. However, Ab–RAPDF, which showed the highest correlation to $\ln K_D$, is still only weakly correlated. In order to overcome this, two weighting schemes were implemented into the RAPDF formula. Using a similar scheme to Wang *et al.* (2004) (Section 2.2.4, Equation 2.2), we applied weights that were dependent on predicted affinity changes (Schymkowitz *et al.*, 2005) or experimental binding affinities (Dunbar *et al.*, 2014).

For the first weighting scheme, FoldX (Schymkowitz *et al.*, 2005) was used to estimate binding affinity changes ($\Delta\Delta G$) of complexes in the training set; these changes were used to weight Ab–RAPDF. In principle, a mutation whose $\Delta\Delta G > 0$ indicates that the wild–type residue forms stabilising interactions with the antigen. Hence, antibody–antigen contacts involving antibody atoms with predicted $\Delta\Delta G > 0$ were up–weighted, and contacts involving atoms with predicted $\Delta\Delta G < 0$ were down–weighted. We refer to this weighted form of Ab–RAPDF with FoldX information as Ab–fRAPDF. Ab–fRAPDF showed weaker correlation to binding affinities of antibody–antigen complexes in the test set ($|r| = 0.399$). Ab–fRAPDF showed stronger correlation to the subset of antibody–protein complexes in the test set ($|r| = 0.536$; Figure 2.6), and outperformed the unweighted Ab–RAPDF potential. Thus, it appears that weighting contacts in the non–redundant set enhances our prediction of antibody–protein complexes.

The level of improvement was small, and there are two possible reasons. First, most mutations had been predicted to have $|\Delta\Delta G|$ <0.5kcal·mol^{-1}, meaning that most contacts were up– or down–weighted by only a marginal amount. Second, we chose to not mutate antibody alanines and glycines, similar to Robin *et al.* (2014). Glycine residues are prevalent in antibodies and are thought to play a key role in some antibody–antigen interactions (Birtalan *et al.*, 2008).

Our second weighting method used the experimental affinity data of the training set, rather than predicted affinity changes. Effectively, contacts that are prevalent in high–affinity antibody–antigen complexes are enriched, leading to lower RAPDF scores. We refer to the second weighted form of the Ab–RAPDF potential with experimental affinity data as Ab–wRAPDF. Ab–wRAPDF shows stronger correlation to binding affinities ($|r| = 0.431$; Figure 2.6). Similar to our other implementations of RAPDF, Ab–wRAPDF showed stronger performance on protein–binding antibodies ($|r| = 0.532$; Figure 2.6), but weakly correlated to the affinities of peptide–binding antibodies ($|r| = 0.088$).

The marginal improvement in prediction performance may be due to the lack of affinity data; less than a third of the non–redundant set (70 out of 241)

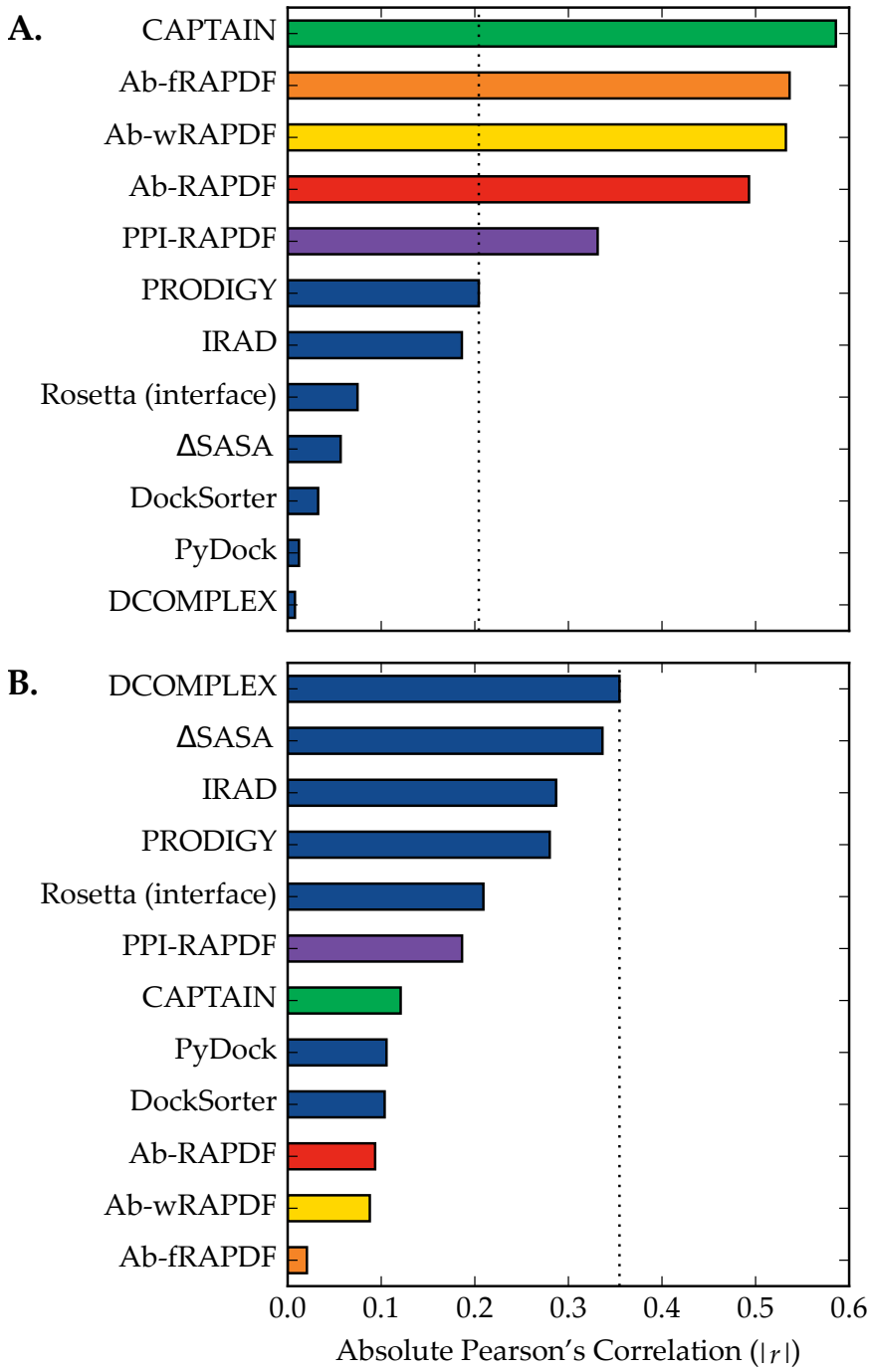


Figure 2.6: Absolute Pearson's correlation between scores against $\ln K_D$ for different affinity prediction methods. The dotted line indicates the best–performing method from the literature. Descriptions of the methods are given in the text. **A.** Antibody–protein complexes of our test set ($n = 55$). **B.** Antibody–peptide complexes of our test set ($n = 35$).

were weighted. Another possible explanation is that all contacts of an antibody–antigen complex were given the same w_a value. This may lead to false–positive weighting of contacts that are generally destabilising.

A combined weighting scheme was developed to create an affinity–dependent, contact–specific weighting system, *i.e.* using both $w_{k,a}$ and $w_{f,i}$ (Equation 2.5). The combined weighting method, which we refer to as Ab–wfRAPDF, showed a relatively weak correlation to the entire test set ($|r| = 0.416$). Again, Ab–wfRAPDF showed a stronger correlation to antibody–protein complexes ($|r| = 0.586$), which was stronger than either Ab–wRAPDF or Ab–fRAPDF.

To account for antibody–antigen complexes that have more contacts, we normalised the scores by the number of contacts between the antibody and antigen. We refer to the normalised Ab–wRAPDF scores as our potential, CAPTAIN, as it showed the best performance on subsets of the test set. CAPTAIN showed $|r| = 0.586$ to the binding affinities of antibody–protein complexes, and $|r| = 0.121$ to the binding affinities of antibody–peptide complexes (Figure 2.6, Table 2.3).

We also tested CAPTAIN on non–antibody–antigen complexes from the AffBM set (Section 2.2.1). As expected, CAPTAIN was unable to predict the affinities in this set ($|r| = 0.211$). In contrast, DCOMPLEX ($|r| = 0.363$) and ZRANK ($|r| = 0.258$) performed better. It seems that these scoring functions, which follow the 'one model fits all' approach, show moderate correlations (Ross *et al.*, 2013). However, they are unable to predict the affinities of antibody–antigen complexes, confirming the necessity of ligand–specific scoring functions.

2.3.2.5 Hybrid scoring does not improve affinity prediction

Following the approach by Studer *et al.* (2014), we introduced a convergence parameter σ to combine contact data from the protein–protein complex dataset and the affinity–weighted antibody contacts for CAPTAIN. We used a range of σ values to control the contribution of general protein–protein contacts in scoring. Larger σ values led to increased correlations to binding affinities

Table 2.3: Absolute Pearson's correlation between RAPDF and binding affinity.

RAPDF Version	All test set antibodies (n = 90)	Protein–binding antibodies (n = 55)	Peptide–binding antibodies (n = 35)
PPI–RAPDF	0.253	0.329	0.178
PPI+Ab–RAPDF	0.313	0.341	**0.193**
Ab–RAPDF	0.411	0.493	0.094
Ab–fRAPDF	0.398	0.541	0.010
Ab–wRAPDF	0.431	0.532	0.088
Ab–wfRADF	0.414	0.556	0.012
CAPTAIN (normalised Ab–wRAPDF)	0.401	**0.586**	0.121
Ab–fRAPDF (normalised)	0.410	0.575	0.098
Ab–wfRAPDF (normalised)	**0.435**	0.586	0.107
PPI+CAPTAIN, $\sigma = 0.001$	0.202	0.473	0.121
PPI+CAPTAIN, $\sigma = 0.01$	0.300	0.546	0.119
PPI+CAPTAIN, $\sigma = 0.1$	0.385	0.579	0.128
PPI+CAPTAIN, $\sigma = 1$	0.399	0.585	0.122
PPI+CAPTAIN, $\sigma = 10$	0.400	0.586	0.119

(Table 2.3). However, none of the hybrid scoring functions showed stronger performance than CAPTAIN.

2.4 Discussion

In this Chapter, we demonstrate that our weighted antibody–specific scoring function, CAPTAIN, shows the strongest correlation to the binding affinities of protein–binding antibodies compared to 16 other published scoring functions and ΔSASA.

Unlike previous methods which used a variety of protein–protein complex structures from the PDB (including antibody–antigen complexes), we used only antibody–antigen complexes for training CAPTAIN. χ^2 analyses demonstrated that antibodies have distinct preferences for specific amino acids (Figure 2.4), corroborating previous studies (*e.g.* Clark *et al.*, 2006b; Krawczyk *et al.*, 2013). Our potential captures these features at the binding interface, and our results

highlight the benefit of using antibody–specific data for successful affinity prediction, particularly antibody–protein complexes (Figure 2.6).

CAPTAIN showed promising results for antibody–protein complexes ($|r| = 0.586$), but predicts the affinities of peptide–binding antibodies poorly ($|r| = 0.121$). This could be attributed to the fact that only protein–binding antibodies were used for training the function. Using peptide–antibody information should improve the prediction of the affinities of peptide–binding antibodies. However, combining peptide–antibody complexes in training CAPTAIN could introduce noise that interferes with the prediction of protein–binding antibodies' affinities.

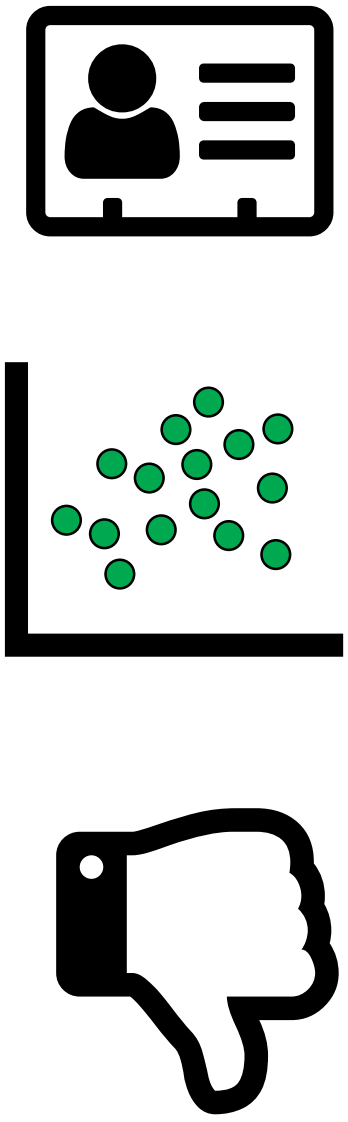
The main issue in building CAPTAIN was the sparsity of the contact data, leading to extreme values in the score distributions (Figure 2.5). Combining the contact data from the protein–protein complex dataset as a simple sum (as in the case of PPI+Ab–RAPDF), or as a hybrid score did not improve prediction (Sippl, 1990; Studer *et al.*, 2014). In fact, the hybrid scores only showed strong correlations at high σ values, reiterating the importance of antibody–specific information. One possible solution to reduce the sparsity of contact data is to produce antibody models and dock them to model antigen structures, leading to a simulated, saturated set of antibody–antigen contacts. Alternatively, reducing the number of contact types to a set of 'essential' antibody–antigen contacts may stabilise the score distributions, and avoid potential overfitting issues.

We also weighted the counts by using predicted $\Delta\Delta G$ values and experimental affinity data. Both methods improved the correlation of the scores with respect to $\ln K_D$, but when combined, their benefits were not additive. Affinity information, whether predicted or experimental, can enhance predictions, though we advocate the use of experimental data where possible. A more reliable prediction of $\Delta\Delta G$, or an increase in available affinity data are possible paths to improve CAPTAIN's predictions.

All methods, including CAPTAIN, showed poor correlations to the binding affinities of antibody–antigen complexes. CAPTAIN was the best performer

among these methods, though further work is necessary to make a reliable tool for applications such as affinity maturation. Throughout this Chapter, we have assumed that the binding affinity of an antibody is purely determined by contacts with its antigen. Considering other factors, such as conformational entropy (Haidar *et al.*, 2014), domain orientation (Dunbar *et al.*, 2013; Fera *et al.*, 2014), and germline V gene pairing could further improve affinity prediction, but the knowledge base is currently insufficient. Ultimately, we conclude that affinity prediction remains a formidable challenge, which requires more affinity data, a better understanding of how antibodies interact with their antigen, and parsimonious models that are robust to data sparsity issues.

In the next Chapter, we will cover the mechanism of V_H–V_L pairing in antibodies, as it provides one of the largest sources of structural variation. Analysing the sequence and structural variation of antibodies should help us understand how antibodies bind, and improve affinity for their targets. Subsequently, Chapters 4 and 5 focus on building high–resolution antibody models using the wealth of structural data in the PDB. We hope that in the long–term, these modelling approaches will provide complementary data that can enhance binding affinity predictions.



People may see things differently, but they don't really want to.

— Don Draper, *Mad Men*

3

All V_H–V_L pairs are equal; some are more equal than others.

Contents

3.1 Introduction

In the previous Chapter, we discussed the challenges of predicting antibody–antigen binding affinities. In order to understand the antibody structural landscape, we next looked at the largest source of structural variation: the pairing of the V_H and V_L. In this Chapter, we propose a potential structural mechanism for V_H–V_L pairing.

3.1.1 Pairing: a determinant of antibody function

The pairing of the V_H and V_L domains has an important role in the structure and function of antibodies. We showed in the previous Chapter that 'paired'

antibodies (*i.e.* Fvs, scFvs) often bind antigens using the CDR loops from both V_H and V_L domains, such as the CDRH3 and CDRL3 loops. Furthermore, the V_H and V_L domains pack together in a wide range of orientations, leading to a huge diversity in the shape of the antibody's binding site. This ultimately allows antibodies to bind many different types of antigens (Muyldermans *et al.*, 2001). Although the relationship between pairing and V_H–V_L orientation is not clear, the different pairings do have an influence (Jayaram *et al.*, 2012; Dunbar *et al.*, 2013). A recent analysis by Teplyakov *et al.* (2016) suggests that pairing can affect the structure of the CDRH3 loop, thus leading to changes in the antigen binding site.

In vivo, pairing acts as a screening mechanism during the B–cell maturation process (Sections 1.2.6.2, 1.3.3). For example, heavy chains are selected by their ability to pair with the SLC (Melchers, 2005). Furthermore, if a putative V_H–V_L pair is autoreactive, the pre–B–cell recombines light *V* genes until the new V_H–V_L pair is non–reactive (Nemazee, 2006).

From an antibody engineering perspective, V_H–V_L pairing attracts interest as it is known to modulate antibody stability and binding. Certain V_H–V_L pairs show greater thermostability, which is vital for antibody integrity and function (Ewert *et al.*, 2003; Tiller *et al.*, 2013). In addition, V_H–V_L pairing has been manipulated to engineer bispecific antibodies (*e.g.* Lewis *et al.*, 2014; Klein *et al.*, 2012; Fischer *et al.*, 2015).

3.1.2 Mechanism of V_H–V_L pairing

Despite the importance of V_H–V_L pairing, the precise mechanism of V_H–V_L pairing is unknown. The general belief is that V_H–V_L pairing is 'random' (de Wildt *et al.*, 1999; Glanville *et al.*, 2009; DeKosky *et al.*, 2015). Pairing has previously been shown to be flexible; a given heavy chain can pair with several different light chains, and vice–versa (Edwards *et al.*, 2003). On the other hand, several studies suggest a preference in pairing, *i.e.* certain *V* gene subgroups tend to form pairs (*e.g.* Jayaram *et al.*, 2012).

Randomness in pairing is often defined from a statistical perspective; the frequencies of paired IMGT subgroups (or alleles) is compared to their expected frequency by a χ^2 test of independence. If the frequencies of subgroup (or allele) pairs are not significantly different from the expected distribution, V_H–V_L pairing is then considered 'random' (*e.g.* Glanville *et al.*, 2009; DeKosky *et al.*, 2015). Statistical tests have also been used on subsets of paired subgroups. For example, Jayaram *et al.* (2012) constructed several 2×2 contingency tables to test for the enrichment or depletion of specific V_H–V_L pairs via Fisher's exact tests.

Recently, Townsend *et al.* (2016) have analysed the physicochemical properties (*e.g.* isoelectric points) of the CDRH3 and CDRL3 loops of antibody sequences from DeKosky *et al.* (2016). Kolmogorov–Smirnov tests indicated that, for instance, λ CDRL3 loops are more hydrophobic than κ CDRL3. However, these differences were not observed between the CDRH3 loops of κ and λ antibodies. Effectively, this test showed that CDRH3 loops can be paired with chemically different CDRL3 loops. However, this study did not extensively discuss how this association justifies random pairing. On the other hand, Kuroda and Gray (2016) have provided a more structure–driven view on the mechanism of pairing. Though the authors state that any chain can pair up with another, their conclusions are largely based on shape complementarity and the number of hydrogen bonds. There is little comment on antibody–specific characteristics that may influence pairing, such as the sequence of the CDR loops.

In this Chapter, we seek to address how antibodies pair, and explicitly define 'random' pairing as the ability for a given heavy chain to pair with any light chain. We first follow the works of Edwards *et al.* (2003), Jayaram *et al.* (2012), and others and describe V_H–V_L pairing on a sequence level, and determine preferences between V germline genes. Next, the structures of paired antibodies, along with the structure of the pre–BCR, are analysed in further detail to elucidate a structure–based mechanism for pairing. We also complement our observations with thermodynamic data from Teplyakov *et al.* (2016) to investigate the relationship between pairing frequency and antibody stability.

We demonstrate that the apparent preferences in V_H–V_L pairs seem to be due to the significant over– or under–representation of some pairs. In addition, it appears that 'common' pairs may be less stable. Therefore, while our data supports the conclusion that any V_H can associate with any V_L, we also suggest that there are costs in forming certain V_H–V_L pairs.

3.2 Methods

3.2.1 Annotation of antibody sequences

All input sequences were numbered by ANARCI (Dunbar and Deane, 2016). Briefly, ANARCI queries a sequence against a database of hidden Markov models (HMMs; Eddy, 2004). Each HMM represents a multiple sequence alignment of a particular allele, *e.g.* IGHV1–02*01. ANARCI then calculates the sequence identity between the query sequence and IMGT germline sequences to identify the most likely germline genes (Lefranc *et al.*, 2009). Germlines from a duplicate subgroup were grouped under one subgroup; for instance, IGHV1D is treated as IGHV1.

In some cases, the matching HMM and the corresponding germline gene do not necessarily agree. For example, a sequence can be matched to an HMM of a rat heavy *V* gene, but share the greatest sequence identity to a mouse heavy *V* germline gene. In this scenario, the sequence is deemed to be a mouse *V* gene sequence. An antibody is considered to be from a particular species only if both heavy *V* and light *V* genes are matched to the same species.

The numbering of the human SLC was based on the synthetic construct by Morstadt *et al.* (2008). Their 'single–chain' SLC was formed by truncating the unique regions of VpreB and $\lambda 5$, and joining them thereafter. Without the unique regions, the VpreB and $\lambda 5$ sequences are homologous to light *V* and *J* gene sequences, respectively. ANARCI was used to number the single–chain form of the SLC (PDB: 3bj9), and the numbering was mapped onto the pre–BCR structure (PDB: 2h32).

3.2.2 Datasets

3.2.2.1 Sequence datasets

We downloaded paired antibody sequences from DIG–IT (Chailyan *et al.*, 2012), abYsis (Swindells *et al.*, 2016), and SAbDab (Dunbar *et al.*, 2014) on 27 January, 2016. Sequences with missing framework positions (between H5–H122, L5–L122), unknown amino acids, or those that could not be numbered by ANARCI were discarded. This gave a redundant set of 13702 sequences. Sequences were then clustered by 99% Fv sequence identity using CD–HIT (Li and Godzik, 2006). We call this dataset the non–redundant (NR) set, in which there are 6295 paired antibody sequences.

Human antibody sequences from our redundant set and the NGS dataset from DeKosky *et al.* (2015) were combined to obtain a total of 205206 human antibody sequences. The NGS set from DeKosky *et al.* (2015) represent paired CDRH3–CDRL3 sequences from three human donors, obtained from single B–cell sequencing.

These sequences were clustered by 97% CDRH3 sequence identity, similar to DeKosky *et al.* (2015). We call this dataset the human CDRH3 (hCDRH3) dataset, in which there are 132988 non–redundant human CDRH3 sequences.

3.2.2.2 Structural datasets

One thousand thirty–five Fv structures, including scFv structures, with resolution $\leq$2.5Å were downloaded from SAbDab on 27 January, 2016 (Dunbar *et al.*, 2014). These structures were clustered by 90% Fv sequence identity to generate a non–redundant set of 527 structures; we call this dataset the V_H–V_L contact set. Together, there are 173 human and 298 mouse antibody structures; the remainder are either rabbit (9), rat (1), or hybrid structures (*e.g.* human/rhesus hybrid; PDB: 2a9m).

Sixteen antibodies (21 Fab structures) from Teplyakov *et al.* (2016) were used to investigate the effects of germline pairing on CDRH3 loop structures and antibody stability. Each antibody has a melting temperature (Tm) value, and

the same CDRH3 sequence, but different germline V gene pairings. Details of these structures are listed in Appendix Table B.3; four structures were removed due to missing coordinates in the structure, and three structures were removed as they were duplicate structures in the asymmetric unit of the crystal. We refer to the final set of 14 antibodies as the Teplyakov set.

3.2.3 Statistical analyses

Genes were collated at the IMGT subgroup level for statistical analysis. We constructed a $N \times M$ matrix to count the frequency of V_H–V_L subgroup pairs in our NR set. Here, N and M represent the number of different V_H–V_L subgroups. We then used the χ^2 test of independence to test for the dependence between heavy and light subgroups in terms of pairing. The χ^2 value represents a scaled squared difference between the observed and expected counts of a heavy–light subgroup pair, *e.g.* IGHV3:IGKV1. The χ^2 value is then calculated for every possible pair of heavy (h) and light (l) subgroups;

$$\chi^2 = \sum_{h \in H, l \in L} \frac{(O_{hl} - E_{hl})^2}{E_{hl}}. \tag{3.1}$$

O_{hl} represents the observed number of a pair hl, and E_{hl} represents the expected number of hl. H and L represent the set of all heavy and light subgroups, respectively.

Fisher's exact tests were also used on 2×2 contingency tables of individual pairs, similar to Jayaram *et al.* (2012):

$$\begin{matrix} hl & h'l \\ hl' & h'l' \end{matrix}$$

Here, h' and l' represent every heavy or light subgroup apart from h or l. For example, IGHV3′:IGKV1 represents heavy–light subgroup pairs formed from any heavy subgroup (apart from IGHV3) with the IGKV1 subgroup. The p–value of Fisher's exact test is calculated by

$$p = \frac{\binom{hl+h'l}{hl}\binom{hl'+h'l'}{hl'}}{\binom{T}{hl+hl'}} \tag{3.2}$$

where T is the total number of antibodies in the contingency table, *i.e.* $T = hl + h'l + hl' + h'l'$.

3.2.4 Entropy Scoring of V_H–V_L pairs

The entropy of a V_H subgroup is calculated from the conditional distribution of V_L subgroup counts. The normalised frequency of l for a given h is

$$p(l|h) = \frac{n_l}{\sum_{l \in L} n_l},$$

where n_l represents the number of light chains paired up with h. The entropy of the heavy chain, $E(h)$, is therefore

$$E(h) = -\sum_{l \in L} p(l|h) \log_2 p(l|h). \tag{3.3}$$

The same calculation is used for calculating the entropy of a V_L subgroup, where the conditional distribution of V_H subgroup counts (for a given V_L) is used. The entropy of a pair, $E(hl)$, is the sum of the individual chains' entropies, *i.e.*, $E(hl) = E(h) + E(l)$.

3.2.5 Contact distributions in V_H–V_L interfaces

Contacts at the V_H–V_L interface were analysed in the V_H–V_L contact set. Two residues were considered to be in contact if any one of their non–hydrogen atoms were within 5Å.

3.2.5.1 PCA of antibody regions

Similar to the approach by Scarabelli and Grant (2013), antibody structures were analysed by principal components analysis (PCA). First, structures were superimposed to a reference structure based on a core set of positions (*e.g.* the framework). Only Cα atoms were used for structural alignment.

Next, a $3N \times 3N$ covariance matrix C_{ij} was constructed given the coordinates (x, y, z) of the N equivalent Cα atoms between structures. Thus, if 70 Cα atoms

are aligned, C_{ij} is a 210×210 matrix. For each pair of coordinates i and j within the entire set of coordinates r,

$$C_{ij} = \langle \left((r_i - \langle r_i \rangle) \cdot (r_j - \langle r_j \rangle) \right) . \tag{3.4}$$

Here, r_i represents the ith coordinate of every aligned structure – for instance, the x coordinate of Cα at position H44. $\langle r_i \rangle$ denotes the average of the same coordinate over all structures.

Next, the eigenvectors and eigenvalues of C_{ij} were calculated; each eigenvalue w has an associated eigenvector v. The eigenvectors with the largest eigenvalues account for the majority of the variance between coordinates of the aligned Cα atoms. The percentage of accounted variance is the fraction of the kth eigenvalue with respect to the sum of all eigenvalues.

Structures were projected onto the principal component (PC) space by

$$q_k = (s - \langle s \rangle) \cdot v_k, \tag{3.5}$$

where q_k represents the projected position of structure s along the kth eigenvector, v_k. Here, $\langle s \rangle$ denotes the average coordinate (centroid) of the structure.

For the PCA, we used the highest resolution structure as the reference structure. All other structures were superimposed to the reference by using common framework region Cα atoms. Positions that were used for PCA are listed in Table 3.2.

3.3 Results

3.3.1 Germline pairings indicate pairing dependence

Similar to previous analyses (*e.g.* de Wildt *et al.*, 1999; Glanville *et al.*, 2009; Jayaram *et al.*, 2012), we analysed the pairing behaviour of antibodies based on germline V gene pairings. Since the number of sequences for each IMGT allele was very low, we examined pairings at the IMGT subgroup level (*e.g.* IGHV2 as opposed to IGHV2–01*03). Furthermore, IMGT subgroups are not equivalent

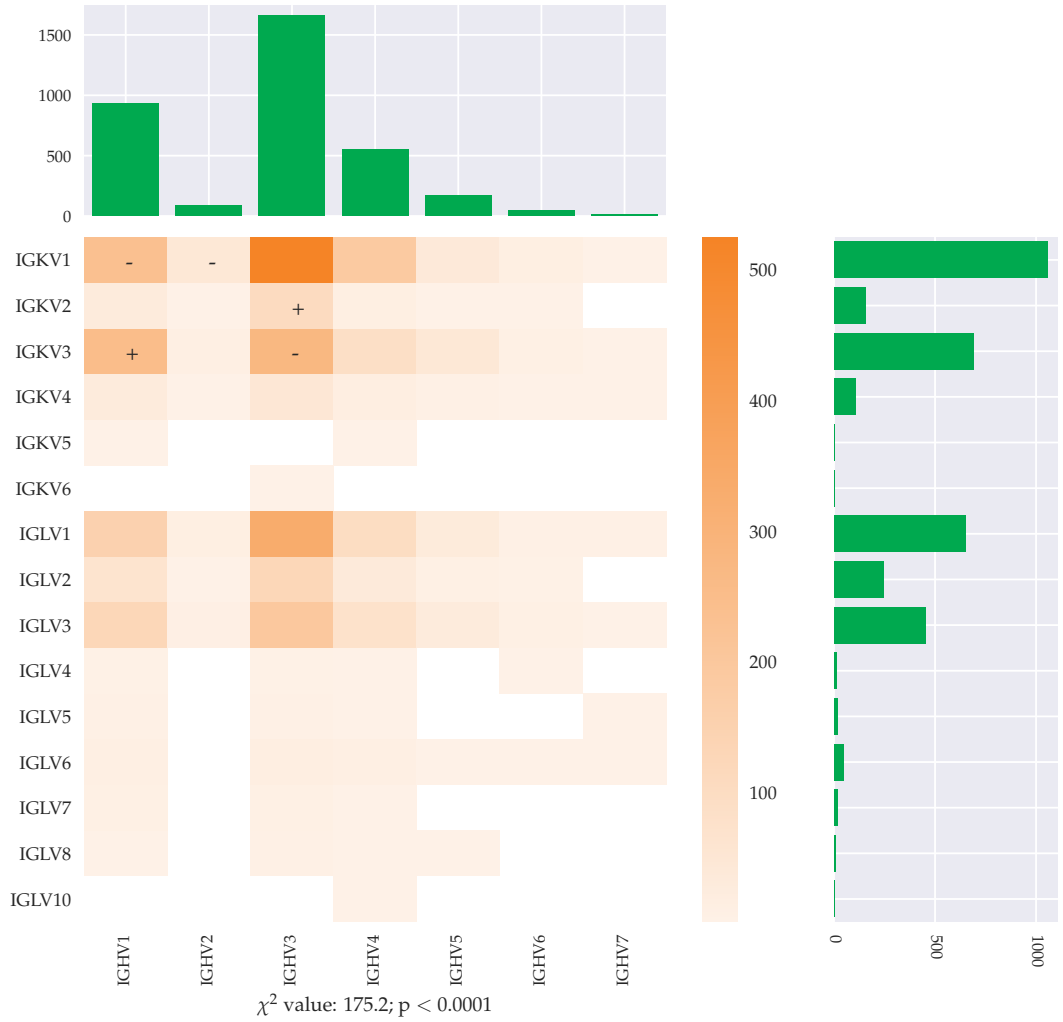


Figure 3.1: Heatmap of human antibodies' V_H–V_L subgroup pairs in the NR set, with histograms of the marginal distributions of V_H (top) and V_L (right) subgroups. Each cell in the heatmap represents the number of antibody sequences with a given V_H–V_L subgroup pair. For instance, the IGHV3:IGKV1 pair is the most common, hence it is coloured dark orange, whereas the IGHV2:IGKV7 pair is uncommon, and is thus coloured light orange. Pairs with no data (*e.g.* IGHV1:IGLV10) are coloured white. Pairs with significant differences (based on Fisher's exact test, Section 3.2.3) are indicated by + (enrichment) or – (depletion).

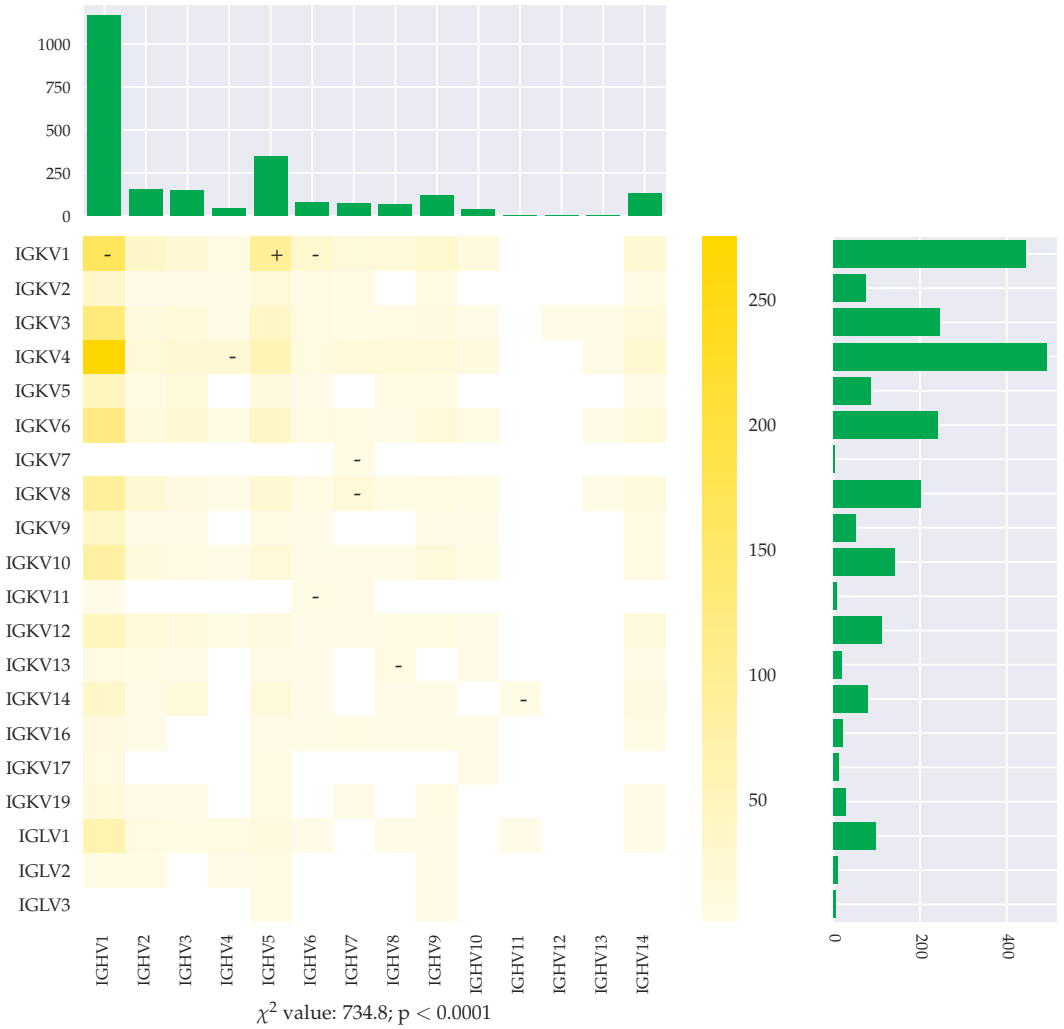


Figure 3.2: Heatmap of mouse antibodies' V_H–V_L pairs in the NR set, with histograms of the marginal distributions of V_H (top) and V_L (right). Please see Figure 3.1 for more details.

between species; *e.g.*, a mouse IGHV1 does not correspond to a human IGHV1. Hence, we divided the analyses into human and mouse antibodies.

In both human and mouse antibodies, the χ^2 test indicated a significant dependence between the heavy and light chains for pairing ($p < 0.0001$). However, the apparent V_H–V_L dependence may be due to the over–representation of a subset of heavy–light chain pairs. More importantly, the χ^2 test also requires an expected value of 5 per cell; 61 of the 105 possible human subgroup pairs (198 of 280 mouse subgroup pairs) did not meet this requirement, and thus

the results should be interpreted with caution. Similar to Jayaram *et al.* (2012), we considered merging several V_H and V_L subgroups. Although we merged duplicate subgroups such IGHV1 and IGHV1D, we felt that merging others (*e.g.* IGHV5, IGHV6, IGHV7) could affect the biological significance of the results. Removing 'sparse' rows did not change the results, and thus we used the original frequency table for analysis.

We also constructed multiple 2×2 contingency tables for Fisher's exact tests. Unlike Jayaram *et al.* (2012), we apply a Bonferroni correction factor. Despite the correction, five human subgroup pairs were considered significant: IGHV1:IGKV1, IGHV1:IGKV3, IGHV2:IGKV1, IGHV3:IGKV2, and IGHV3:IGKV3 (Figure 3.1). Among mouse antibody subgroup pairs, nine were significantly different (Figure 3.2). Collectively, these results suggest that the apparent subgroup dependence that is reported from the χ^2 test is likely due to the significant differences of a select set of V_H–V_L pairs.

3.3.2 Pairing: a proxy for thermal stability

V_H–V_L pairing is often manipulated in the interests of antibody stability (*e.g.* Tiller *et al.*, 2013). We investigated the relationship between the frequency of subgroup pairs and the thermal stability of antibodies. The frequency of pairing was represented by an entropy measure (see Section 3.2.4). For a given pair, its entropy will increase if its constituent heavy and light subgroups are more promiscuous. For example, the IGHV5:IGKV1 pair has low entropy as IGHV5 tends to only form a pair with IGKV1. Pairing entropy was calculated from the paired subgroup frequencies in the hCDRH3 set (DeKosky *et al.*, 2015).

For the 14 antibodies in the Teplyakov set, their pairing entropy had a negative correlation with melting temperature (Tm; Figure 3.3). This implies that pairs with low entropy are more thermostable than pairs with high entropy. Despite a significant correlation ($p = 0.02$), the small size of the dataset means that these results are preliminary.

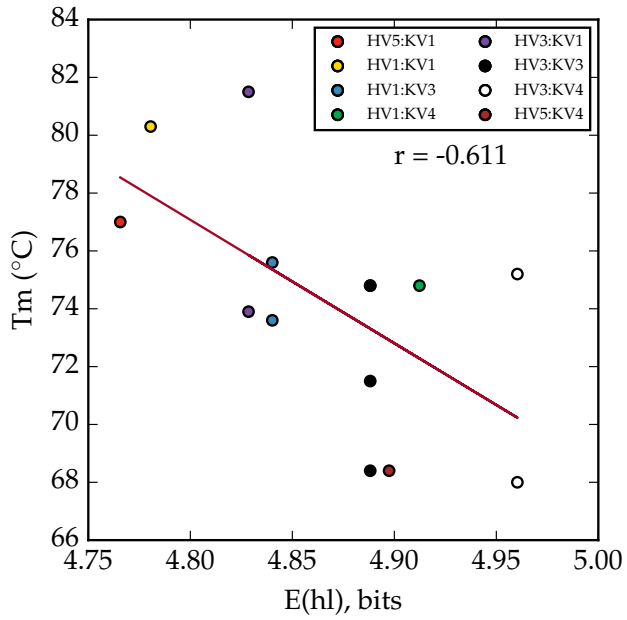


Figure 3.3: Entropy of pairing $E(hl)$ plotted against melting temperature (Tm) of antibodies in the Teplyakov set. A fitted linear model is shown with a red line. The Pearson's correlation is –0.611, with a p–value of 0.02.

3.3.3 Paired sequences can be very different

The pairwise sequence identities of the NR set were used to investigate the pairing landscape of antibodies at a finer resolution. Here, we considered two V_H or two V_L domains to be 'identical' if they shared $\geq$99% sequence identity. Thus, if there are two antibodies A and B sharing 99.5% sequence identity over the V_H but 82% sequence identity over the V_L, then we assume that A's V_H can pair with two V_Ls. Extending from this formalism, if there are four sequences with an identical V_H but four different light chains (each with $<$99% V_L sequence identity), then the V_H domain is considered to pair with four different V_Ls.

In most cases, a given V_H domain or V_L domain was observed to only pair up with one unique V_L or V_H domain. However, for certain V_H domains, we observed more than ten different light chains (Figure 3.4). The same phenomenon was observed for V_Ls, where certain V_Ls were paired with many different V_Hs (Figure 3.5). Our results corroborate with previous observations that certain

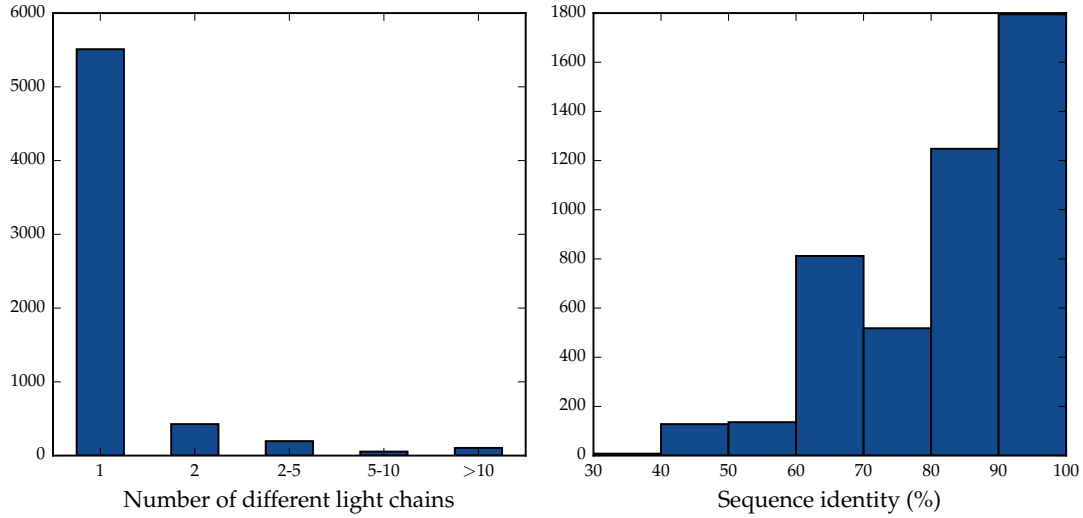


Figure 3.4: Number of different V_L domains that can pair with a given V_H domain (left), and the sequence identity between two V_L domains that both pair up with a given V_H (right). We considered two V_H domains to be 'identical' if they shared $\geq$99% sequence identity. Likewise, if two light chains shared $\geq$99% sequence identity, they are also considered 'identical'.

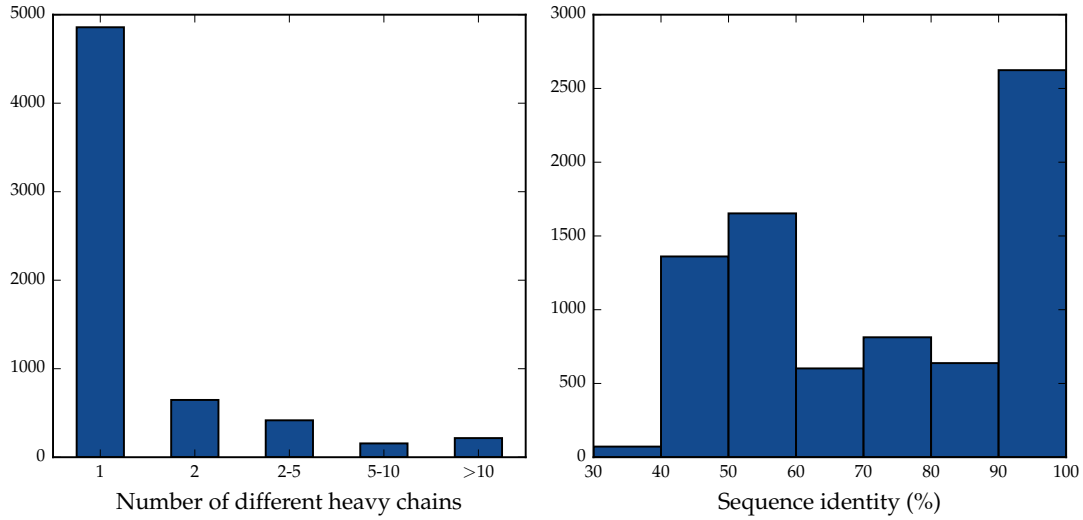


Figure 3.5: Number of different V_H domains that can pair with a given V_L domain (left), and the sequence identity between different V_H domains that both pair up with a given V_L (right). Please see Figure 3.4 for more details.

V_H and V_L domains are 'promiscuous' (Edwards *et al.*, 2003; DeKosky *et al.*, 2015).

As expected, many of the 'different' heavy or light chains were point mutants, with each domain often sharing >95% sequence identity (6769/10155 V_L sequence identity calculations; 6922/12621 V_H sequence identity calculations; Figures 3.4, 3.5). For a given V_H, the average sequence identity between its partner V_L domains was 92%, whereas the average sequence identity between V_H domains was 83%.

Very few V_L pairs shared <60% sequence identity, though we observed as low as 36% sequence identity between two V_Ls. We observed a spike between 60–70% V_L sequence identity (878 V_L pairs), which is likely due to mutations of more than two CDR loops. For example, antibodies AY942079 and AY942110 (from Chailyan *et al.*, 2012) are from the same IGLV1 subgroup, but have 17 mutations across all three CDRs.

In contrast, the distribution of V_H domains' sequence identities showed a distinct peak of cases where the V_H domain had 40–60% sequence identity, and two V_H domains can share as low as 32% sequence identity (Figure 3.5). We expected a wider distribution of sequence identities as the CDR loops of heavy chains tend to be more divergent.

The low V_H or V_L sequence identities (<40% identity) may arise from engineering; for example, 1a6v:JN (mouse IGHV:IGKV1) and 1nqb:AA (mouse IGHV:IGLV1) share 35% V_L sequence identity. Although 1a6v seems to represent an antibody from an immunised mouse (no publication found), 1nqb is an engineered scFv, meaning that the antibody would have paired via the linker (Pei *et al.*, 1997). However, in the case of 2ddq:HL (mouse IGHV1:IGKV1) and 2ipu:HL (mouse IGHV8:IGKV1), which share 32% V_H sequence identity, we discovered that this is a demonstration of natural flexibility in pairing. Both antibodies were raised from mice that were immunised with a hapten (2ddq:HL; Akagawa *et al.*, 2006) or peptide (2ipu:HL; Gardberg *et al.*, 2007).

3.3.4 Essential V_H–V_L contacts are not different

In >90% of structures in our V_H–V_L contact set, we observed 17 contacts at the V_H–V_L interface. These contacts involve a set of framework and CDR loop positions (Table 3.1). Using these contacts, we then extracted the amino acids that are present at the two contacting positions among antibodies in the NR set (Figures 3.6, 3.7).

Table 3.1: Essential contacts at the V_H–V_L interface.

V_H Position	V_L Positions
H42	L118
H44	L44, L103
H50	L118, L103, L50
H51	L118
H52	L118, L116, L115
H103	L50, L44
H115	L52
H116	L52
H118	L50, L42
H119	L49

Two positions were considered to be in contact if any of their atoms are within 5Å of each other. Contacts that are found in over 90% of structures in the V_H–V_L contact set are shown. Positions are coloured if they belong to CDRH3 (red) or CDRL3 (green).

Regardless of the species (human, mouse, rabbit) and light chain isotype, the contacts in Table 3.1 often involved the same, if not similar, amino acids. For instance, positions H118 and L50, which make a contact in every structure of our V_H–V_L contact set, are almost always a Trp–Pro contact in the NR set (Figure 3.6). The H50–L118 contact occurs in all but one structure of our dataset. Between these positions, there is almost always a Leu–Phe contact in the NR set. The contacts at the V_H–V_L interface are mostly hydrophobic, with some hydrogen bonds. For example, two conserved Gln residues at the

'base' of the Fv (H44/L44) form a hydrogen bond (Figure 3.7), and the Tyr residue at H103 also points toward the base (Figure 3.8A). We also detected an asymmetry in polar contacts, similar to Kuroda and Gray (2016). Heavy chains tend to use backbone atoms, whereas light chains use their side chain atoms for forming polar interactions. Considering the high conservation of contacts at these positions and the amino acids at these sites, we believe that these positions form a core set of 'essential' contacts that are key for V_H–V_L pairing.

Most variations in contact residues featured chemically similar amino acids. For instance, positions H118 and L50 feature a Trp and Pro, respectively. Five human antibodies, *e.g.* the anti–tumour antibody FW418207 (Chailyan *et al.*, 2012), featured Arg and Pro at these positions. Since no structure in SAbDab had an Arg and Pro at these sites, we constructed a model for these five antibodies using our modelling tool, ABodyBuilder (Chapter 4). In all five models, the Cδ of H118 was within 5Å of the Pro at L50, suggesting that the H118–L50 contact may still be established in spite of the change at H118.

The fact that the 'essential' contacts are so highly conserved across different V_H–V_L pairings suggests again that all pairings are possible. This is further reinforced by the residues seen in the pre–BCR structure.

3.3.4.1 The SLC has the same contacts as a typical light chain

The structure of the pre–BCR by Bankovich *et al.* (2007) provides additional clues toward the mechanism of pairing. In humans, there is only one SLC for the entire immune repertoire. This suggests that either the unique regions of the SLC are flexible enough to pair with several V_H structures (Melchers, 2005; Bankovich *et al.*, 2007), or that there is a generic, non–specific binding mechanism that allows any V_H domain to pair with any V_L domain.

In the pre–BCR structure, we observed the same set of essential contacts found in antibodies of our V_H–V_L contact set. For example, the H118–L50 contact was a Trp–Pro contact; in addition, the H44–L44 contact was a Gln–Gln hydrogen bond (Figure 3.8). The residues at these positions also had similar side

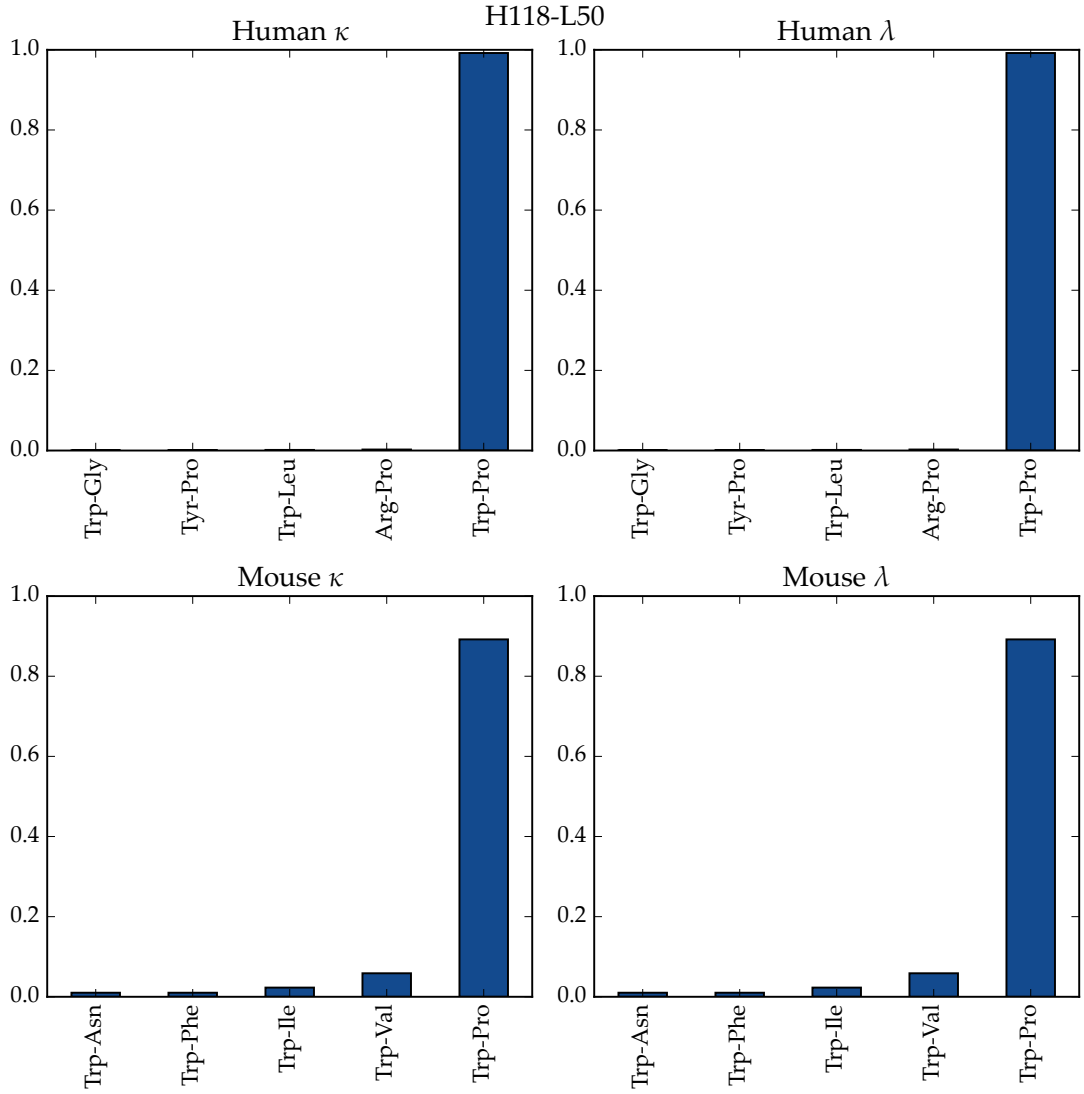


Figure 3.6: Amino acid pair distributions for the H118–L50 contact. We first observed that all structures in our V_H–V_L contact set have a contact between these two positions (Section 3.2.5). We then extracted the amino acids at H118 and L50 for different antibody types (*e.g.* human κ) from our NR set of 6295 sequences. Although a small proportion of mouse antibodies have a Trp–Val contact, these positions are highly conserved, with a Trp–Pro contact.

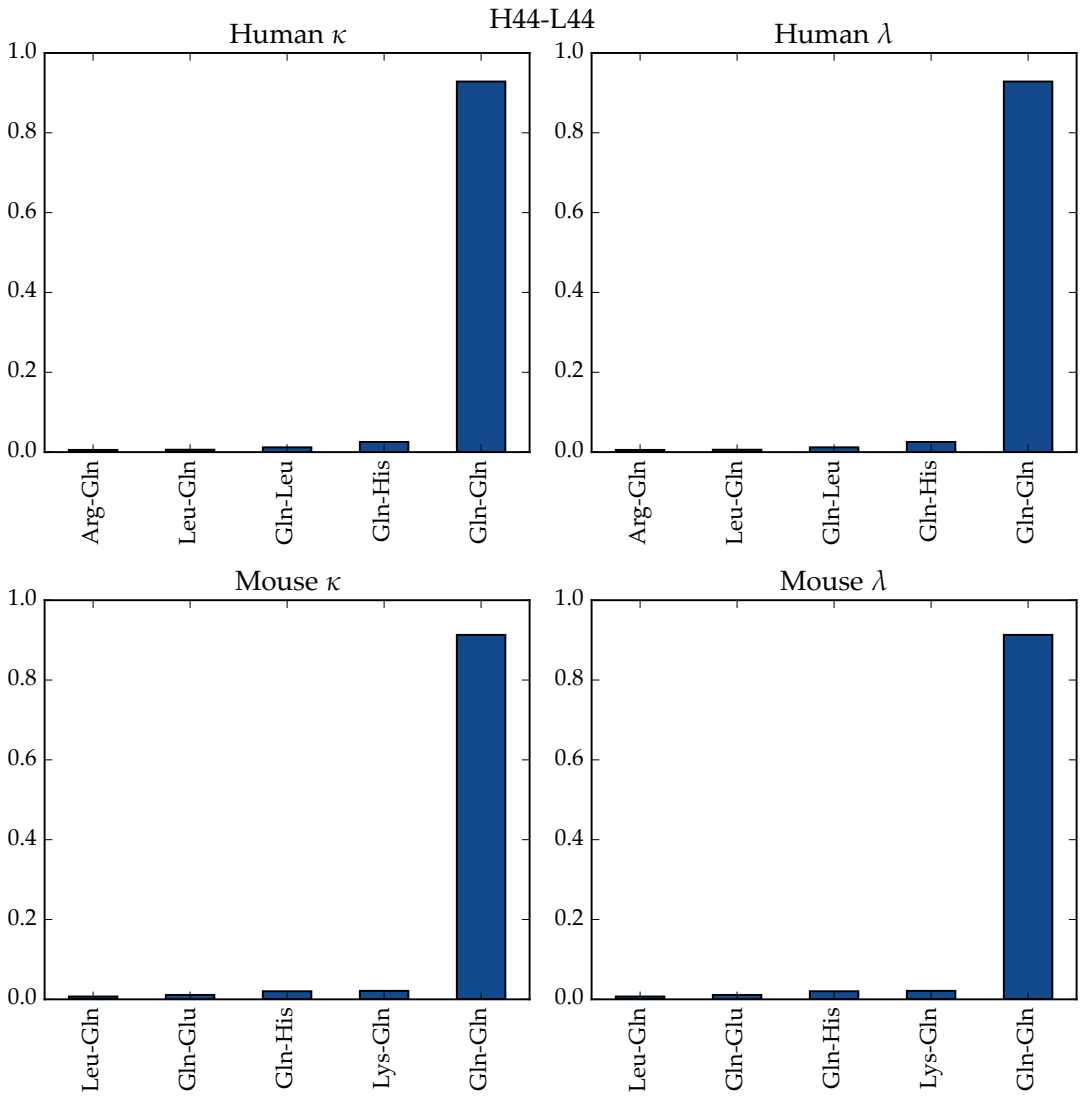


Figure 3.7: Amino acid pair distributions for the H44–L44 contact. Apart from three structures (PDB: 1ay1, 4m1d, 5cgy), the other 524 structures of our V_H–V_L contact set had a contact between these two positions. Similar to Figure 3.6, we observe a high conservation of a Gln–Gln contact for each antibody type.

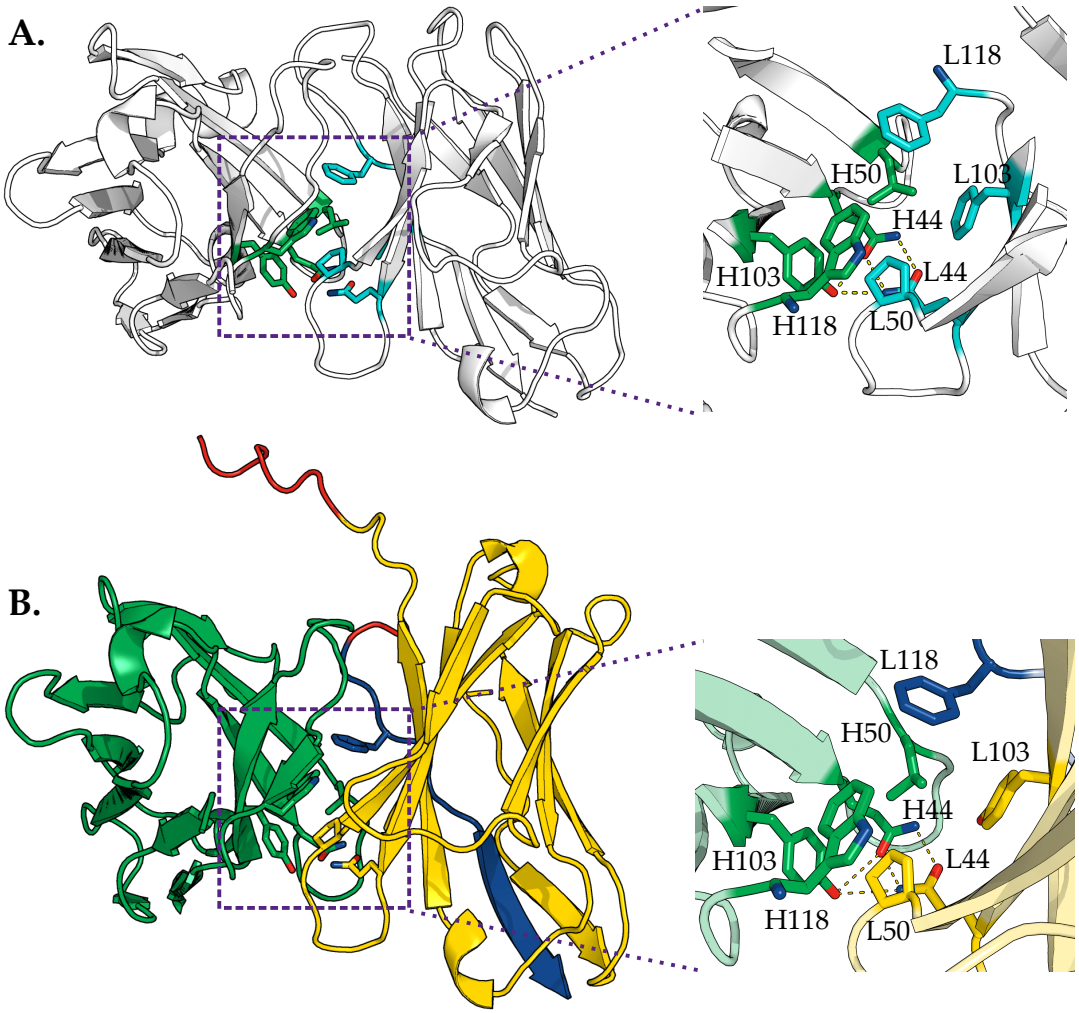


Figure 3.8: Comparing the contacting residues of a typical mouse antibody (**A.**) and a human pre–BCR structure (**B.**). **A.** Positions H44, H50, H103, H118, L44, L50, L103, and L118 have a cluster of highly conserved residues whose side chains point toward the interface, with a hydrogen bond between H44–L44. In the inset, we see that the side chains point toward the interface. **B.** The same positions have a similar structural and sequence profile in the human pre–BCR. The side chains of the SLC (yellow: VpreB, blue: $\lambda 5$) are also oriented in a similar manner to a typical light chain.

chain configurations, thus confirming the role of these residues in interacting with a V_H domain. Therefore, it seems that the satisfaction of these contacts provides one of the necessary conditions for an SLC to approve a heavy chain for pairing by light chains in the pre–B cell.

Table 3.2: Positions aligned for PCA of antibody structures (Section 3.2.5.1).

H6+§	H7+§	H8	H9	H11	H12	H13	H14	H15	H16
H17	H18	H19	H20	H21	H23	H41	H42	H43	H44
H45	H46*§	H47	H48	H49	H50	H51	H52	H53	H54
H67	H68	H69	H70	H71	H72	H74	H75	H76	H77
H78	H79	H80	H81	H82	H83	H84	H85	H86	H87
H88	H89+§	H90	H91	H92	H93	H94	H95	H96	H97
H98	H99	H100	H101	H102	H103	H104	H118	H119	H120
H121	H122	H123	H124	H125	H126	H127+	H4§	H5§	H128§
L3*§	L4*§	L5*§	L6*§	L7*§	L8*§	L9*§	L11	L12	L13
L14	L15	L16	L17	L18	L19	L20	L21	L22	L23
L41	L42	L43	L44	L45	L46	L47	L48	L49	L50
L51	L52	L53	L54	L70*§	L71*§	L72*§	L74	L75	L76
L77	L78	L79	L80	L83	L84	L85	L86	L87	L88
L89	L90	L91	L92	L93	L94	L95	L96	L97	L98
L99	L100	L101	L102	L103	L104	L118	L119	L120	L121
L122	L123	L124	L125	L1§	L2§	L10§	L126§	L127§	

+: position that was only used in human structures; *: position that was only used in mouse structures; §: position that was only used for PCA of the Teplyakov set (Table B.3).
For each position, the Cα coordinates were used for computing the principal components.

3.3.5 Pairs are structurally flexible

Structures in the V_H–V_L contact set were analysed by PCA to determine the flexibility of V_H–V_L pairs (Figures 3.9, 3.11). PCA has previously been used to monitor conformational dynamics of proteins such as kinesin (Scarabelli and Grant, 2013), and we used the method to determine the relationship between antibody pairing and framework region flexibility. We aligned human (or mouse) structures in the V_H–V_L contact set to the highest–resolution human (or mouse) structure, and used common framework region Cα atoms (Table 3.2) for computing the principal components (PCs).

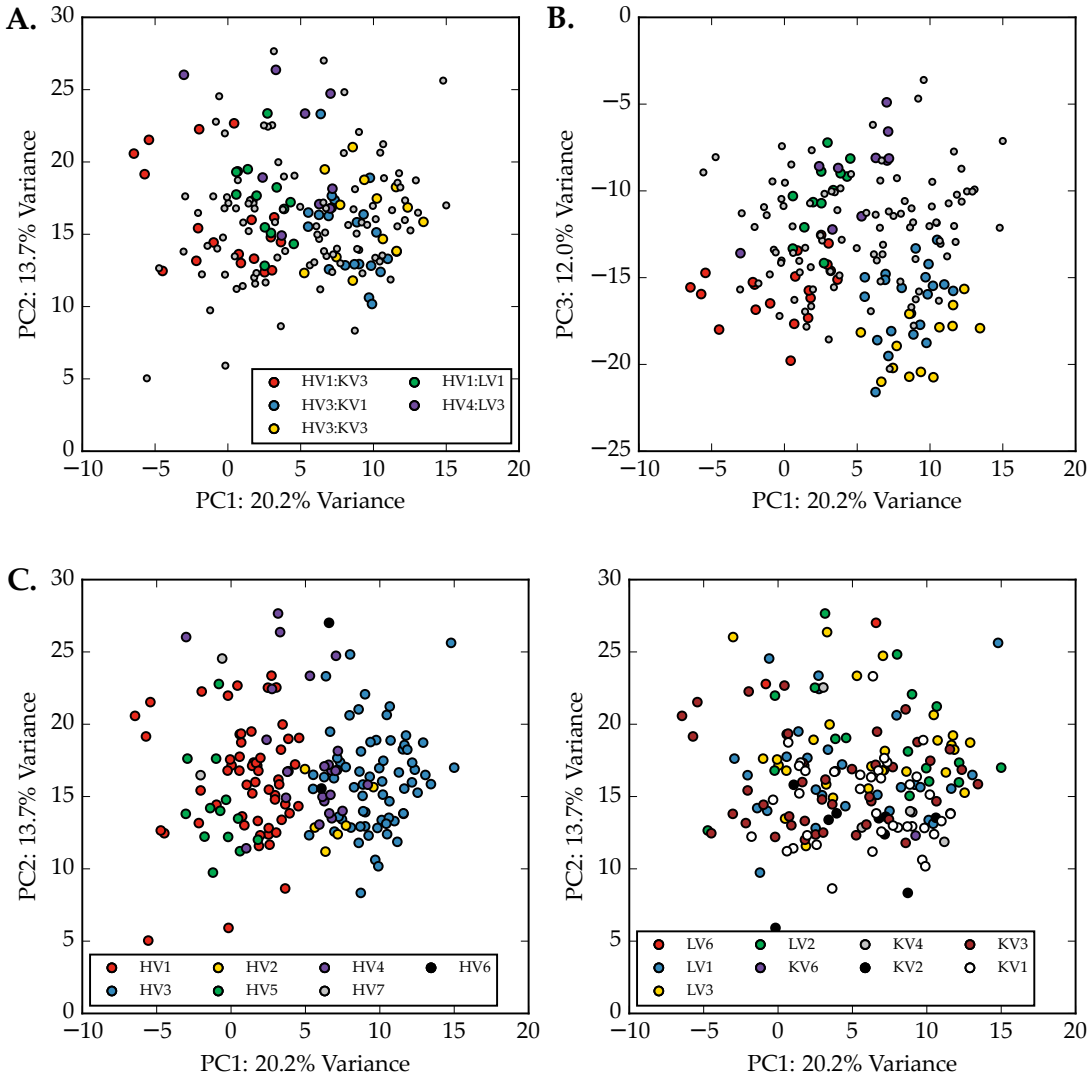


Figure 3.9: PCA plot of 173 human antibody structures in the V_H–V_L contact set, using framework positions for alignment (Table 3.2). The PCA procedure is described in Section 3.2.5.1. **A.** Antibodies projected onto the first and second principal components (PC1 and PC2), with five common pairs (>10 structures) labelled. Non–labelled pairs are plotted with grey circles. **B.** Antibodies projected onto the first and third PCs. **C.** Antibodies projected onto the first and second PCs, labelled either by their V_H subgroup (left) or V_L subgroup (right).

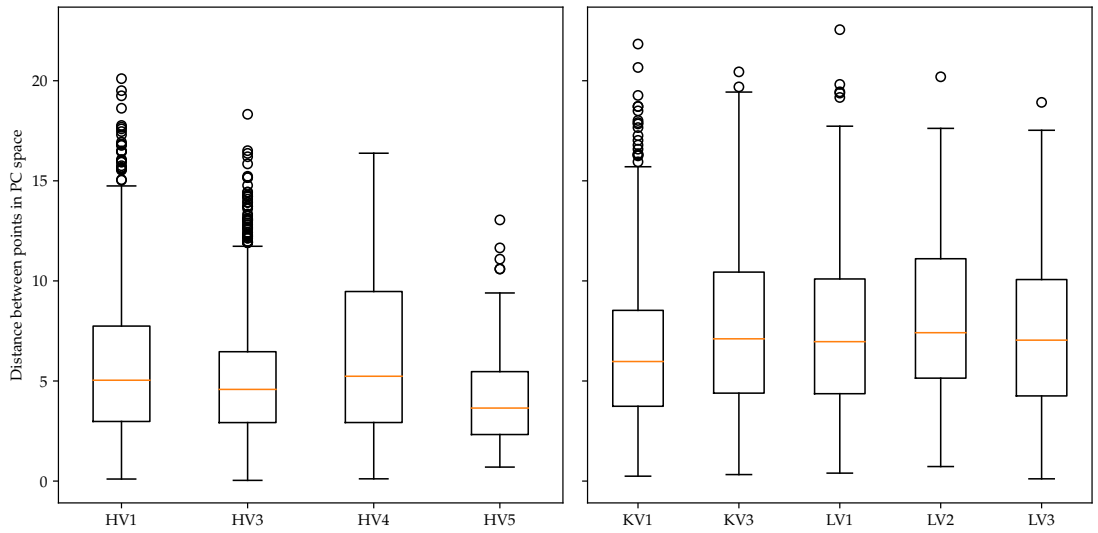


Figure 3.10: Distribution of pairwise distances between human antibody structures in the PC1–PC2 space (Figure 3.9). The pairwise Euclidean distances of antibodies within a particular V_H subgroup (left) or V_L subgroup (right) were calculated if there were $\geq$10 structures for each subgroup.

The distribution of human antibodies in the PC space suggest that antibody frameworks can adopt multiple conformations, regardless of its germline pairing. High–frequency pairs, such as IGHV3:IGKV1, localised in a specific area of the PC space, whereas low–frequency pairs such as IGHV4:IGLV3 were scattered (Figure 3.9A). In addition, several different pairs were found in one area of the PC space, *i.e.*, there is a high degree of overlap. Antibody pairing has previously been confirmed to have little relationship with orientation (Jayaram *et al.*, 2012; Dunbar *et al.*, 2013). Our PCA corroborates these observations, as pairs were distributed in various regions of the PC space. Mouse antibody structures also showed pairing–independent flexibility (Figure 3.11).

When we labelled the antibodies according to their subgroup types, we observed that antibodies from different V_H subgroups occupied distinct areas in the PC space (Figure 3.9C). For example, human IGHV1 and IGHV3 antibodies occupy two separate regions, with a clear boundary between them at $PC1 = 5$. However, separation was not perfect; there were some overlaps between different V_H subgroups, *e.g.* IGHV5 and IGHV1. On the other hand, when we labelled the antibodies by their V_L subgroups, there was no evidence for subgroup separation.

In other words, multiple light chains occupied various areas of the PC space with high overlap. In addition, antibodies of a particular V_H subgroup were clustered more tightly than those in a specific V_L subgroup, as shown by the distribution of distances between antibodies in the PC1–PC2 space (Figure 3.10). Thus, it seems that light chains are flexible and can fit onto several different heavy chains. This supports the concept of a 'promiscuous' light chain (DeKosky *et al.*, 2015). Although the separation effect was not as clear as we saw in human antibodies, mouse antibodies were also separated more distinctly in the PC space by their V_H subgroup, rather than their V_L subgroup (Figure 3.11). Similar to human antibody structures, mouse antibodies of a particular V_H subgroup were clustered more closely than those of an identical V_L subgroup (Figure 3.12).

These observations were further reinforced by PCA on the Teplyakov set (Figure 3.13, Appendix Table B.3). Similar to the PCA on human and mouse structures of the V_H–V_L contact set, the distribution of pairs in the PC space showed some overlap; for instance, the IGHV3:IGKV1 and IGHV3:IGKV3 pairs occupy similar regions of the PC space. However, annotating antibodies by their V_H subgroup showed clear separation (Figure 3.13C), where IGHV1, IGHV3, and IGHV5 antibodies were projected onto three different areas of the PC space. When annotating projected antibodies by their V_L subgroup, it was clear that there was more disorder, indicating inherent flexibility. Thus, we speculate that V_L domains are capable of 'wobbling' to fit and pair with different V_H domains.

3.3.5.1 Pairing effects on CDRH3 structures

Despite having an identical CDRH3 sequence, two of the antibodies in the Teplyakov set (PDB: 5i17:HL [IGHV1:IGKV3], 5i1i:HL [IGHV3:IGKV4]) have an 'extended' CDRH3 conformation, and all other structures have a 'kinked' CDRH3. We first grafted the CDRH3 loops from each structure onto a different framework in the Teplyakov set, leading to 182 different CDRH3–framework combinations (Figure 3.14).

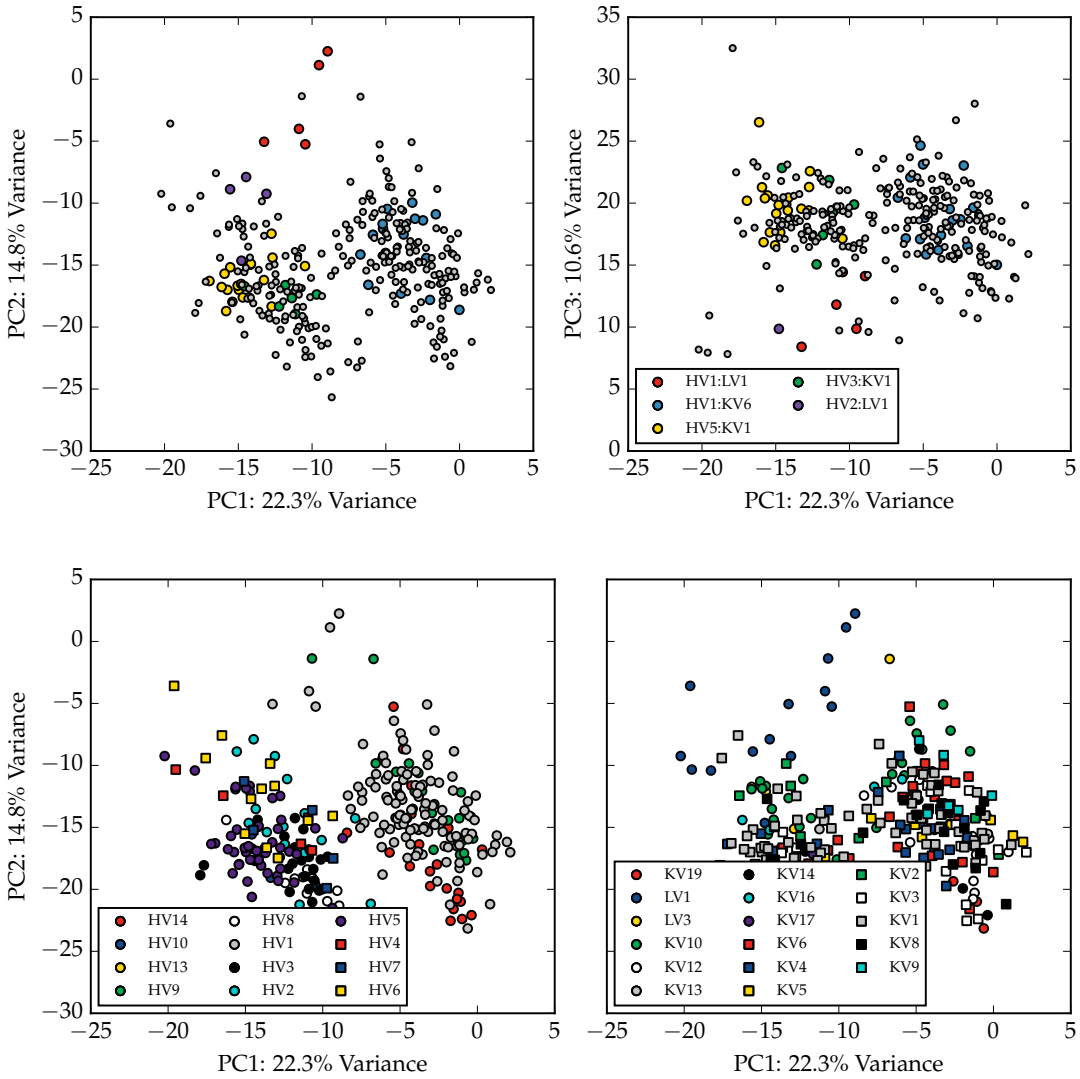


Figure 3.11: PCA plot of 298 mouse antibody framework structures in the V_H–V_L contact set. Please see Figure 3.9 for more details.

In 47% of cases, the grafted structures had less than five clashes. However, there was no apparent relationship between the number of clashes and the structures that were grafted. When the kinked CDRH3 loop from 5i19:HL was grafted onto the framework of 5i1i:HL, this led to 55 clashes, despite having the same V_H subgroup (IGHV3). In contrast, grafting the kinked CDRH3 loop from 5i18:HL onto this framework led to zero clashes. Although 5i18:HL and 5i1i:HL share the same V_L subgroup (IGKV4), grafting the CDRH3 loop from 5i1h:HL onto the framework of 5i17:HL (sharing IGKV3) led to 16 clashes. Thus,

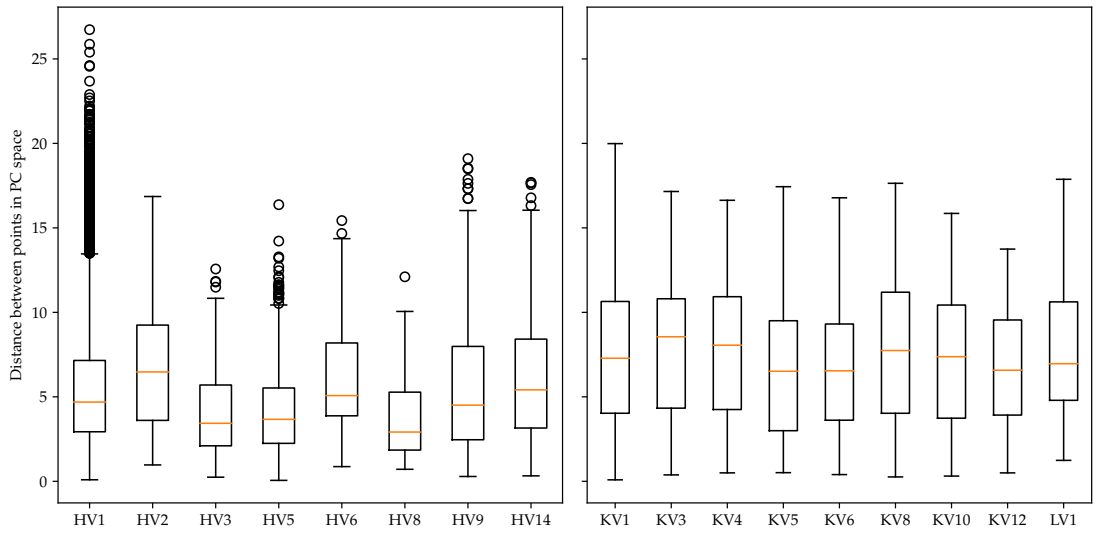


Figure 3.12: Distribution of pairwise distances between mouse antibody structures in the PC1–PC2 space (Figure 3.11). Please see Figure 3.10 for more details.

this suggests that two antibodies with the same V_H or V_L subgroup may not necessarily accommodate the same CDRH3 structure.

We also considered the relationship between pairwise differences of the ABangle parameters and the two CDRH3 conformations. Antibodies with extended CDRH3 loops were not different from those with kinked CDRH3 loops (Appendix Figure A.5). Furthermore, differences in orientation were not correlated to the number of clashes from grafting (Appendix Figure A.6). These two results indicate that one CDRH3 loop conformation may not be specific to a particular orientation.

Based on our observations, it is unclear how the CDRH3 loop forms different conformations for the various pairing environments. However, the lack of clashes in grafting indicate that some pairs can accommodate both kinked or extended structures.

3.4 Discussion

In this chapter, we have described V_H–V_L pairing in a structure–based context. Unlike previous analyses which have mainly focussed on the distribution of

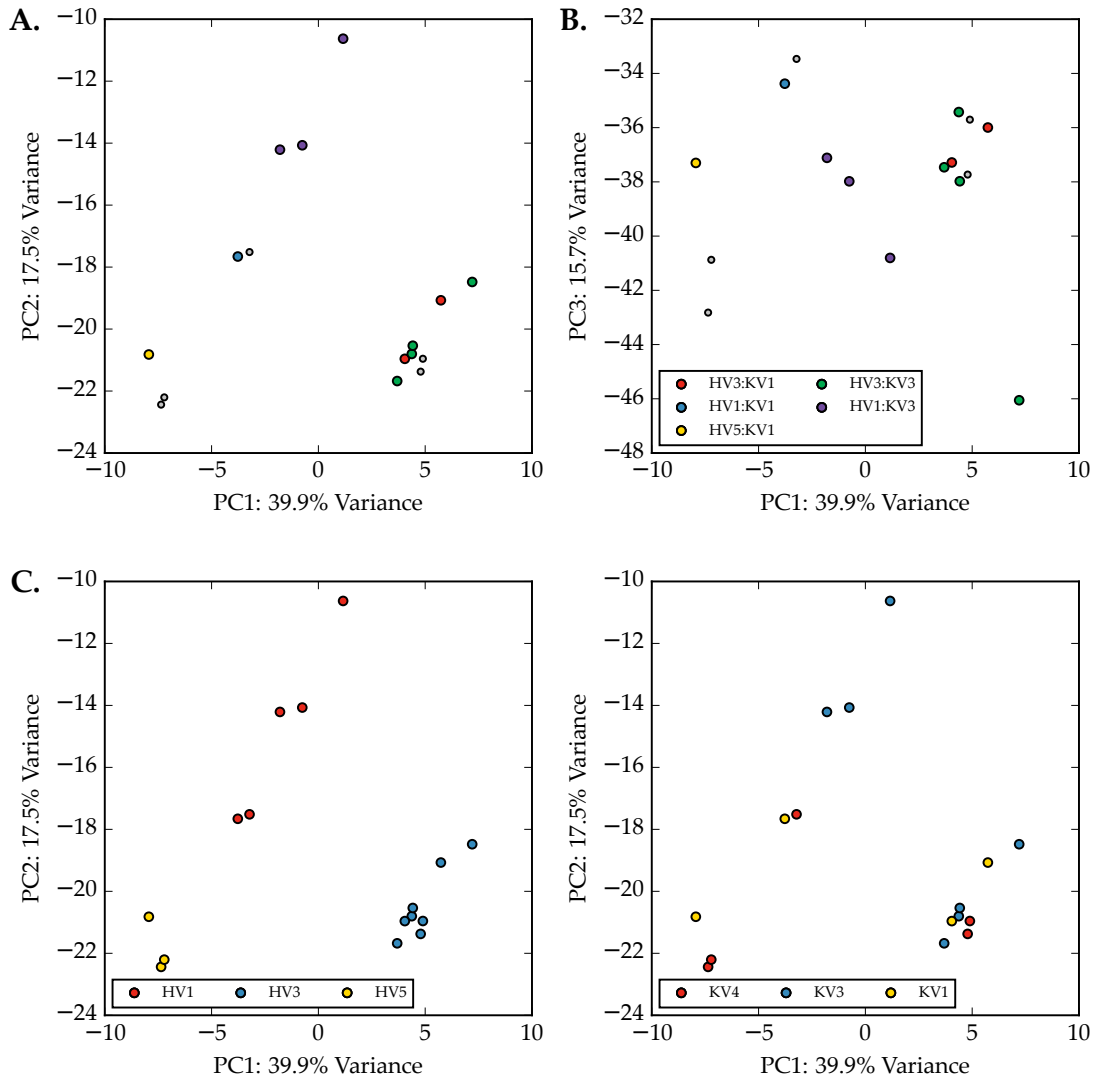


Figure 3.13: PCA plot of the antibody structures from the Teplyakov set. Please see Figure 3.9 for more details.

V gene germline pairings to describe antibody pairing, we use a combination of sequence and structural data to characterise antibody pairing. Given the conservation of contacts and the flexibility of antibody structures, we propose that V_H–V_L pairing is 'random' – that is, there is no clear evidence suggesting that any V_H domain could not pair with any V_L domain.

We first performed a set of statistical analyses, following previous studies (*e.g.* de Wildt *et al.*, 1999; Jayaram *et al.*, 2012; Glanville *et al.*, 2009; DeKosky *et al.*, 2015). Given the sparsity of our dataset (*e.g.* human antibodies with IGLV4

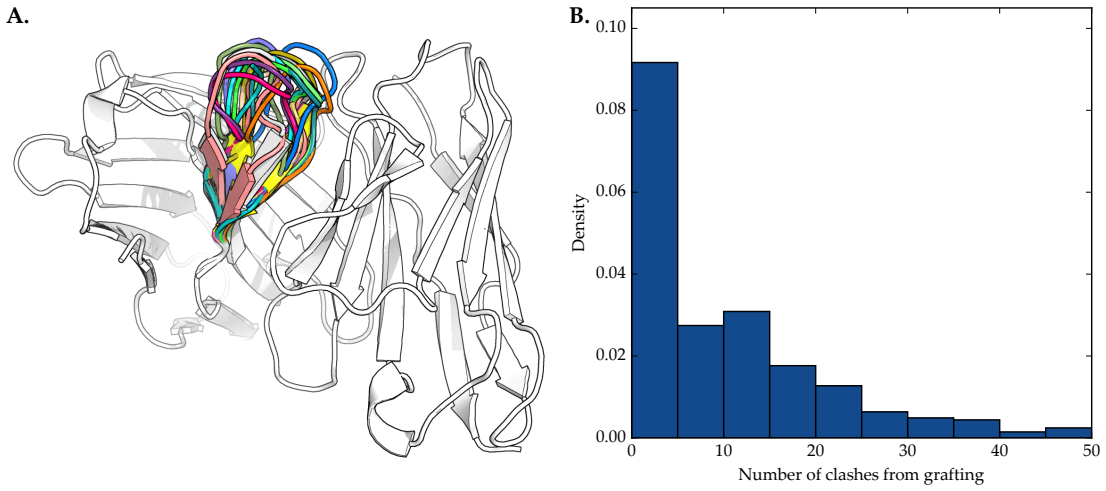


Figure 3.14: Grafting CDRH3 from each antibody in the Teplyakov set to a common framework. **A.** The framework of 4kmt:HL is shown (white) with 13 different CDRH3 loops, excluding the native CDRH3. **B.** Histogram of clashes from grafting different CDRH3 loops onto a single framework.

light chains), and the significant difference of a handful of pairs, we believe that germline *V* gene annotations alone provide an incomplete picture of V_H–V_L pairing. However, the pairing entropies showed a promising correlation to Tm values, and the relationship between pairing frequency and thermostability will need to be verified with a larger dataset.

The sequence identities of the various V_L domains for a given V_H (and vice–versa) were much more informative, providing an overview of the pairing landscape (Figures 3.4, 3.5). The number of different possible V_L or V_H partners for a given V_H or V_L supported the concept of a 'promiscuous' variable domain (Edwards *et al.*, 2003; Fischer *et al.*, 2015; DeKosky *et al.*, 2015). Despite the diversity of sequences in forming V_H–V_L pairs, we discovered a high level of conservation in V_H–V_L contacts. In fact, these contacts were also preserved in the pre–BCR. Therefore, we describe these contacts as 'essential' contacts (Table 3.1).

In comparison to the vast repertoire of heavy chains that must be screened for pairing, the low number of SLCs (one in humans, two in mice and rabbits), and the high sequence homology between SLCs (Ohnishi and Melchers, 2003; Morstadt *et al.*, 2008) reinforce the random pairing mechanism. For the filtering

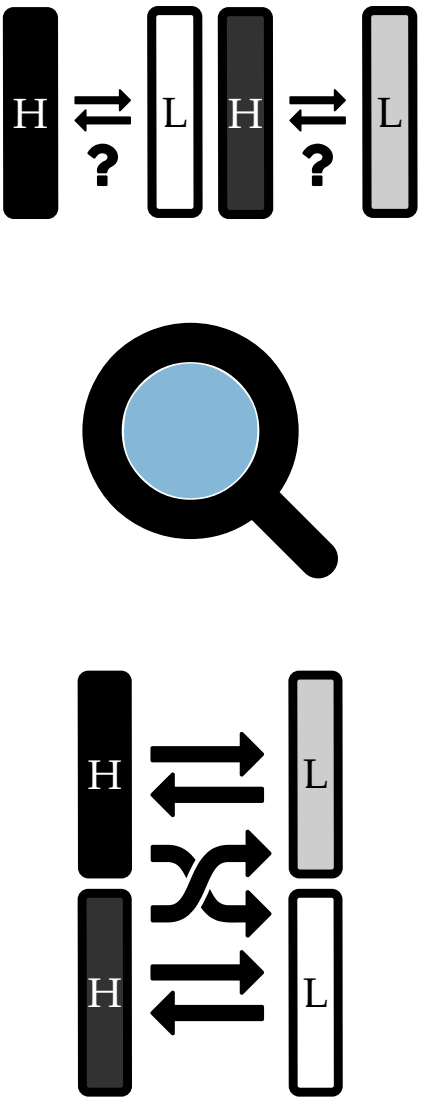
process, the interactions have to be non–specific, or involve a very limited set of residues, such that every V_H domain can be checked by the SLC. The 'essential' contacts are not only frequent (*e.g.* H118–L50), but often utilise the same residues. This generic method of screening V_H domains may explain why the SLC does not perfectly screen for all V_H domains (Smith and Roman, 2010).

For both V_H–V_L and V_H–SLC pairing, restricting the pairing mechanism to the essential contacts is a major simplification. Other contacts at the interface involving a different set of residues may enhance, or weaken, the interaction strength between the two domains. Furthermore, we have not discussed the role of loops in the context of pairing in detail. For instance, the CDRH3 loop and the SLC's unique region are flexible loops whose structures depend upon, and influence pairing (Melchers, 2005; Kuroda *et al.*, 2009; Teplyakov *et al.*, 2016). In the case of antibodies, a more dedicated analysis on CDRH3 variation and pairing is necessary. For pre–BCRs, more structures of pre–BCRs with different V_H domains will shed light on the importance of the essential contacts and the unique regions in the context of pairing.

The PCA on the V_H–V_L contact set and the Teplyakov set point toward the flexibility of light chains as a second driving force for random pairing. We propose that light chains 'wobble' to accommodate a variety of heavy chain structures for V_H–V_L pairing. Given the large sequence variation of V_H domains, a V_L domain would need to be malleable in order to complement the various sequences and shapes of the V_H domain.

Despite the apparent lack of a straightforward relationship between subgroups and orientation (Dunbar *et al.*, 2013), the PCA presented in this Chapter captures movements in the light chain due to a larger, high–resolution alignment. Our PCA aligns as many framework positions as possible; when we performed another PCA using only the ABangle–defined 'coreset' positions, there was no separation between V_H and V_L subgroups (not shown). In addition, the covariance matrix probes coupled changes of each atomic coordinate; thus, we obtain a more detailed view on the movement of each domain.

For a given heavy chain, it will pair with any light chain, and vice–versa. The association seems to depend on a series of conserved contacts, and the light chain 'wobbling' to fit onto the heavy chain. The grand challenge of determining how antibodies pair is the absence of a negative dataset. The best biological 'negative' example would be a V_HH antibody; however, there are no known studies that have attempted pairing a camelid V_H domain with a V_L. In order to generate a negative set, a one–class support vector machine may be a possible avenue (Schölkopf *et al.*, 2000). Alternatively, structural models of sequences in our NR set will provide insight on how antibodies pair, especially those with rare *V* gene germlines (Dunbar *et al.*, 2014). In the next Chapter, we will describe our antibody modelling pipeline, ABodyBuilder, and how it can be used to model thousands of sequences to create a structural map of the antibody sequence space.


H
L
?
H
L
?
H
L
H
L

Art begins in imitation, and ends in innovation.

— Mason Cooley

4

Automated antibody structure prediction with data–driven accuracy estimation.

Contents

The chapter is largely based on the work in the following publication:

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

4.1 Introduction

In the previous chapters, we covered the importance of an antibody's structure with respect to its biological function. In the context of antibody design, a structural model offers a platform to test further engineering decisions, such as mutations. Furthermore, with the explosion of available sequence data, *e.g.* from

NGS datasets, modelling can help map the sequence diversity of antibodies on a structural level (DeKosky *et al.*, 2016). Antibody modelling usually refers to predicting the structure of the entire Fv, which will be the focus of this Chapter.

As discussed in Chapter 1, a structural model of an antibody is valuable for several applications, such as computational affinity maturation and humanisation (Almagro *et al.*, 2014). In particular, antibody modelling is a key component of *de novo* computational antibody design pipelines, as it provides a system for testing the effects of sequence design decisions (*e.g.* Miklos *et al.*, 2012). There is a wide selection of antibody modelling tools, some of which are freely available, such as PIGS and RosettaAntibody (Marcatili *et al.*, 2014; Sivasubramanian *et al.*, 2009).

Most pipelines break down prediction to four major steps: predicting the framework region, V_H–V_L orientation, the CDR loops, and the side chains (Section 1.5.2.2, Figure 1.17). However, no current pipelines comment on the expected accuracy of the model, so a model's accuracy is only known *post hoc*, once a crystal structure has been determined. From a user's perspective, it is impossible to determine how useful a prediction will be. This is particularly problematic for the CDRH3 loop, which can be modelled extremely poorly (RMSD >5Å) in some targets (Almagro *et al.*, 2014). In addition, no current freely available pipeline comments on the *in vitro* 'developability' of the antibody (Jarasch *et al.*, 2015; Seeliger *et al.*, 2015). Antibodies are prone to several post–translational modifications that hinder the production and retention of the antibody. More importantly, these modifications can adversely affect the antibody's functional properties (Jarasch *et al.*, 2015; Seeliger *et al.*, 2015). Thus, knowledge of these sequence liabilities prior to experimental work could reduce the rate of failure of producing functional antibodies.

In this Chapter, we describe ABodyBuilder, an antibody modelling pipeline that uses our increasing knowledge–base of antibody structures (Dunbar *et al.*, 2014) to guide decision–making in modelling antibodies. The overall methodology behind ABodyBuilder is similar to other pipelines (see Chapter 1). We

select template structures based on sequence identity, and if necessary, predict the antibody's orientation based on the ABangle parameters (Dunbar *et al.*, 2013). All six CDR loops are then predicted by FREAD using a CDR–specific database (Deane and Blundell, 2001; Choi and Deane, 2010, 2011). The model's side chains are then completed by SCWRL4 (Krivov *et al.*, 2009). ABodyBuilder differs from other methods as it annotates the confidence of a model as the probability that a region (*e.g.* the framework) will be modelled within a specific RMSD threshold. ABodyBuilder also flags structural motifs within the model antibody which are known to hinder *in vitro* development (Jarasch *et al.*, 2015). Finally, ABodyBuilder is the only freely available software that is capable of modelling nanobodies (*e.g.* camelid V_HH antibodies). Unlike other pipelines that allow manual input (Fasnacht *et al.*, 2014; Marcatili *et al.*, 2014; Shirai *et al.*, 2014), ABodyBuilder is a rapid, fully automated method for antibody model generation, making it ideal for challenges such as modelling large, NGS datasets (DeKosky *et al.*, 2015, 2016; Robinson, 2015). We also show that ABodyBuilder produces models of similar quality to other leading methods in its fully automated mode, and describe how it provides meaningful information for antibody development.

4.2 Methods

4.2.1 Numbering Sequences and Structures

All input antibody sequences are numbered using the IMGT numbering scheme via ANARCI (Lefranc *et al.*, 2003; Dunbar and Deane, 2016), and the CDR loop positions are identified using North *et al.* (2011)'s definitions.

At the end of the modelling process, ABodyBuilder annotates model structures with all the major numbering schemes and CDR definitions (Section 1.2.4.1). For all prediction steps, sequence–identical antibodies were ignored.

Figure 4.1: Examples of unusual structures (PDB: 1oay, 1sjv). Some of the chains are missing atom coordinates (left, yellow), or in the case of 1sjv (right, white), the framework region following CDRH3 (red) trails off.

4.2.2 Datasets

Our initial dataset was a redundant set of antibodies with resolution $\leq$2.5Å, downloaded from SAbDab on 24 February, 2015 (Dunbar *et al.*, 2014). Structures which could not be numbered, or those with unusual structures (*e.g.* PDB: 1oay, PDB: 1sjv; Figure 4.1) were omitted, leaving 1170 structures (998 complete Fvs, 1104 V_H's and 1064 V_L's). For benchmarking FREAD and the rotamer modelling method, a non–redundant set of 541 antibodies was used (462 complete Fvs, 79 nanobodies; 522 V_H's and 481 V_L's overall), based on a 90% Fv sequence identity cutoff using CD–HIT (Li and Godzik, 2006). A 'blind test' set of 136 structures (with resolution $\leq$4.0Å; 108 complete Fvs and 28 nanobodies) that had been deposited in the PDB between 24 February 2015 and 20 December 2015 was used to blind test the ABodyBuilder pipeline. To test the efficiency of ABodyBuilder, ABodyBuilder was run on a non–redundant set of 6267 paired antibody sequences from DIG–IT (Chailyan *et al.*, 2012), abYsis (Swindells *et al.*, 2016), and SAbDab (Dunbar *et al.*, 2014) using a 99% Fv sequence identity cutoff.

4.2.3 Calculation of Model Accuracy

For all measurements of accuracy, we represent RMSD as the backbone RMSD, *i.e.* the RMSD between the backbone atoms of the model and native structures. Fv RMSD is calculated by superimposing all backbone atoms between the model and native structures. Similarly, the framework regions' RMSD is determined for each chain of the model by superimposing the framework backbone atoms to the corresponding chain in the native structure.

For each of the CDR loops, accuracy is calculated by first superimposing the respective chain's framework region backbone atoms, then calculating the RMSD between the loops, similar to the method used in the AMA–II competition (Almagro *et al.*, 2014). The only difference is in the definition of the CDR loops; we use North *et al.* (2011)'s definition as opposed to the Chothia–defined CDR loops (Chothia and Lesk, 1987; Almagro *et al.*, 2014). However, for calculating the CDR loops' RMSD for the AMA–II targets, the Chothia–defined CDR loops were used. To measure the accuracy of the CDRH3 loop, the heavy chains' framework regions are first superimposed, and the RMSD is then calculated between the model and native CDRH3 loops. Likewise, for the CDRL3 loop, the light chains' framework regions are superimposed before calculating the RMSD between the CDRL3 loops. In our initial analysis when we determined the order of CDR loop modelling, the RMSD between CDR loops was calculated after superimposing the backbone atoms of both chains' framework regions, a more stringent test.

4.2.4 Template Selection

As ABodyBuilder is a template–based modelling pipeline, the first step is template selection. In order to identify templates, ABodyBuilder searches SAbDab for structures with a resolution of 2.5Å or better that are close in sequence to the target. Template selection is based on sequence identity over the framework region – *i.e.*, residues that are outside of North *et al.* (2011)'s CDR definitions. ABodyBuilder can either select a single 'global' template

from one antibody (V_H and V_L framework structures and the orientation) or a 'hybrid' template, where two template structures – one for the V_H, and one for the V_L framework – are used.

4.2.5 V_H–V_L orientation prediction

If a 'global' template is selected, the V_H–V_L orientation is given by that template. For hybrid templates, the V_H and V_L domains are re–oriented using the orientation from the highest sequence identity global template. This re–orientation procedure is carried out as previously described (Bujotzek *et al.*, 2015b). Briefly, the Cα coordinates of the ABangle consensus structure are transformed according to the ABangle parameters from the global template. The heavy and light chains of the hybrid template are then superimposed to the rotated consensus structure.

4.2.6 CDR prediction

The CDR loops are modelled sequentially in an order determined by our ability to accurately predict them, and the contacts between them. The CDR loops are predicted by FREAD, which is a database search algorithm that selects for potential structures based on anchor Cα separations, its environment–specific substitution score (ESSS), and anchor RMSD (Deane and Blundell, 2001; Choi and Deane, 2010, 2011). For each CDR loop, FREAD first searches for fragments from a CDR–specific database. Each of the six CDR–specific databases contains a particular CDR loop's fragments; in other words, a CDRL1 database only contains CDRL1 loop fragments. The six CDR–specific databases were built using the redundant dataset to capture various conformations of sequence–identical loops. If a suitable decoy cannot be found for a given CDR loop, a second iteration of FREAD is performed on the missing CDR loop(s) using an Fv–specific database. The Fv–specific database contains all possible fragments of antibodies from the redundant dataset. If FREAD fails to find a decoy from

the CDR–specific or Fv–specific databases, the most sequence–similar, length–matched CDR loop with resolution $\leq$2.5Å is used as the template. Sequence similarity is determined by using the BLOSUM62 score (Henikoff and Henikoff, 1992) between the template and target CDR loops. If there are no length–matched templates, an *ab initio* prediction is performed using MODELLER (Šali and Blundell, 1993). Of the 3009 CDR loops in the non–redundant set, 164 loops were predicted by using the most sequence–similar loop, and seventeen required *ab initio* modelling. For our blind test set of 136 antibodies, 732 CDR loops were predicted, of which 74 were predicted using the most sequence–similar loop, and ten were modelled *ab initio*. In the AMA–II set, only one of the 66 was modelled by using the most sequence–similar loop. Finally, for the 37602 CDR loops from the set of 6267 paired antibody sequences (Section 4.2.2), 2166 were predicted by using a sequence–similar template, and 230 were modelled *ab initio*.

4.2.7 Side chain prediction

SCWRL4 is used to predict the side chain rotamers of residues that are not identical between the template and target (Krivov *et al.*, 2009). The model is then checked for backbone and side chain clashes; two atoms are considered to be clashing if the distance between them is less than 65% of the sum of their van der Waal's radii. For example, the van der Waal's radius of a carbon atom is 1.7Å; thus, two carbon atoms would be clashing if they are less than 2.21Å apart. If clashes are detected, models are first relaxed by MODELLER by only refining the clashing residues (Šali and Blundell, 1993). If clashes still exist, then all side chains are refined by MODELLER. Rotamer accuracies were calculated as the fraction of rotamers that were 'correct', that is, the fraction of rotamers within $\pm 40^{\circ}$ of the native rotamer.

4.2.8 Confidence measurements

The confidence we have in a segment of the model is calculated for a specific RMSD threshold as a conditional probability, given the sequence identity

of that segment. If the segment is a CDR loop, the loop length is used. Confidence measurements for the framework region are based on the pairwise superimpositions carried out on the redundant set. The probability that the framework region will be modelled with $\leq x$Å accuracy for a sequence identity bin s is calculated as

$$P(x|s) = \frac{P(x \cap s)}{P(s)} \tag{4.1}$$

and each bin is 1% wide. To obtain the probability of a sequence identity bin, P(s), we first calculate the sequence identity of antibodies a and b, $S(a,b)$, in the set of all antibodies, A. We then divided the pairs of antibodies with sequence identity $s \pm 2.5\%$, by the number of all possible antibody pairs, *i.e.*

$$P(s) = \frac{\sum_{a \in A} \sum_{b \neq a, b \in A} I(S(a,b) \in s \pm 2.5)}{|A|(|A|-1)} \tag{4.2}$$

where $|A|$ represents the cardinality of the set A, and I is the indicator function. The joint probability, $P(x \cap s)$, represents the probability that antibody pairs with sequence identity $s \pm 2.5\%$ will have RMSD $\leq x$Å. Thus,

$$P(x \cap s) = \frac{\sum_{a \in A} \sum_{b \neq a, b \in A} I(S(a,b) \in s \pm 2.5, RMSD(a,b) \leq x)}{|A|(|A|-1)} \tag{4.3}$$

For example, of the 608856 V_H–V_H superimpositions in the redundant set, 32904 are within sequence identity bin $s = 80\%$. Thus,

$$P(80) = \frac{32904}{608856} \approx 0.0540.$$

Of the 32904 superimpositions within this sequence identity bin, 24696 superimpositions had RMSD $\leq$1.0Å. Thus, the joint probability is calculated as

$$P(1.0 \cap 80) = \frac{24696}{608856} \approx 0.0406.$$

Thus, the conditional probability that a template framework region would have $\leq$1.0Å RMSD to a target antibody, given 80% sequence identity, is

$$P(1.0|80) = \frac{0.0406}{0.0540} \approx 0.752.$$

The framework regions' confidence measures are separate for V_H and V_L as we do not have a method that will reliably estimate the accuracy of orientation prediction (Dunbar *et al.*, 2013; Bujotzek *et al.*, 2015b). The CDR loops' confidence measures were based on the results from running FREAD on model framework structures of the non–redundant set. The conditional probability is calculated as a function of CDR loop length, l. As RMSD is dependent on length, it is possible to encounter situations where the CDR loops have higher confidence values at lower RMSD thresholds than the framework. For example, it is possible to have 75% confidence that CDRL1 is accurate to within 1Å, and 75% confidence that the framework is accurate to within 2Å. Furthermore, the lack of data per loop length bin for the CDR loops in comparison to a sequence identity bin for the framework may lead to poorer estimates of accuracy for the CDR loops.

This probability–based metric was preferred over estimating an RMSD value with error ϵ (*i.e.*, estimating that a region will be modelled at $x \pm \epsilon$Å) as the latter assumes that each sequence identity or length bin has the same distribution of RMSD values. Our data showed that this is not the case (Figure 4.2), and any measurement of ϵ (*e.g.* standard deviation) will vary for each bin. Thus, we opted for a distribution–free metric. Furthermore, the probability estimate allows a more dynamic expectation of a model's accuracy, as we can derive the accuracy for a wide range of RMSD thresholds (x) or probabilities ($P(x|s)$), depending on the user's application.

4.2.9 Sequence liabilities

The target antibody's sequence and structure is analysed for potential issues that can conflict with *in vitro* development. This is based on data collected from publications (Jarasch *et al.*, 2015; Seeliger *et al.*, 2015). Currently, ABodyBuilder flags eleven possible issues with antibody development; a full list is given in Table 4.1. For a predicted sequence liability, it is only visualised if the position's relative accessible surface area, calculated by DSSP, is greater than 10% (Kabsch and Sander, 1983).

Table 4.1: Sequence liabilities and their motifs that are highlighted by ABodyBuilder.

Liability	Motif*	Reference
Unpaired Cys	Free Cys	(Brych *et al.*, 2010)
N–linked glycosylation	Asn–X–Ser/Thr (X not Pro)	(Gavel and von Heijne, 1990)
Met oxidation	Free Met	(Jarasch *et al.*, 2015)
Trp oxidation	Free Trp	(Jarasch *et al.*, 2015)
Asn deamidation	Asn–Gly/Ser/Thr	(Sydow *et al.*, 2014)
Asp isomerisation	Asp–Gly/Ser/Thr/Asp/His	(Sydow *et al.*, 2014)
Lys glycation	Lys–Glu/Asp/Lys	(Jarasch *et al.*, 2015)
N–terminal Pyroglutamte	N–terminal Glu	(Liu *et al.*, 2011)
Integrin binding	Arg–Gly–Asp, Arg–Tyr–Asp	(Ruoslahti, 1996)
CD11c/CD18 binding	Gly–Pro–Arg	(Ruoslahti, 1996)
Fragmentation	Asp–Pro	(Vlasak and Ionescu, 2011)

*: The three–letter representation of amino acids is used to illustrate motifs (Appendix Figure A.1).

4.3 Results

4.3.1 Structure–based decisions in ABodyBuilder

The decisions behind the ABodyBuilder methodology are based on observations from antibody structural data in SAbDab (Dunbar *et al.*, 2014).

4.3.1.1 Framework Selection

The first stage in ABodyBuilder is the selection of a single template, or two templates (one for the V_H and one for the V_L), to model the framework region. In order to determine how sequence identity between template and target influences the accuracy of model building, the framework regions of all pairs of structures in our redundant set were superimposed. First, both chains were superimposed (Fv–Fv superimposition), and second, the heavy and light chains were superimposed separately (V_H–V_H or V_L–V_L). The RMSD between the pairs were compared to their sequence identities (Figure 4.2).

Given our observations, we use a single 'global' template (both V_H and V_L structures and the V_H–V_L orientation) if a single template structure for the target could be found with $\geq$80% sequence identity for both heavy and light

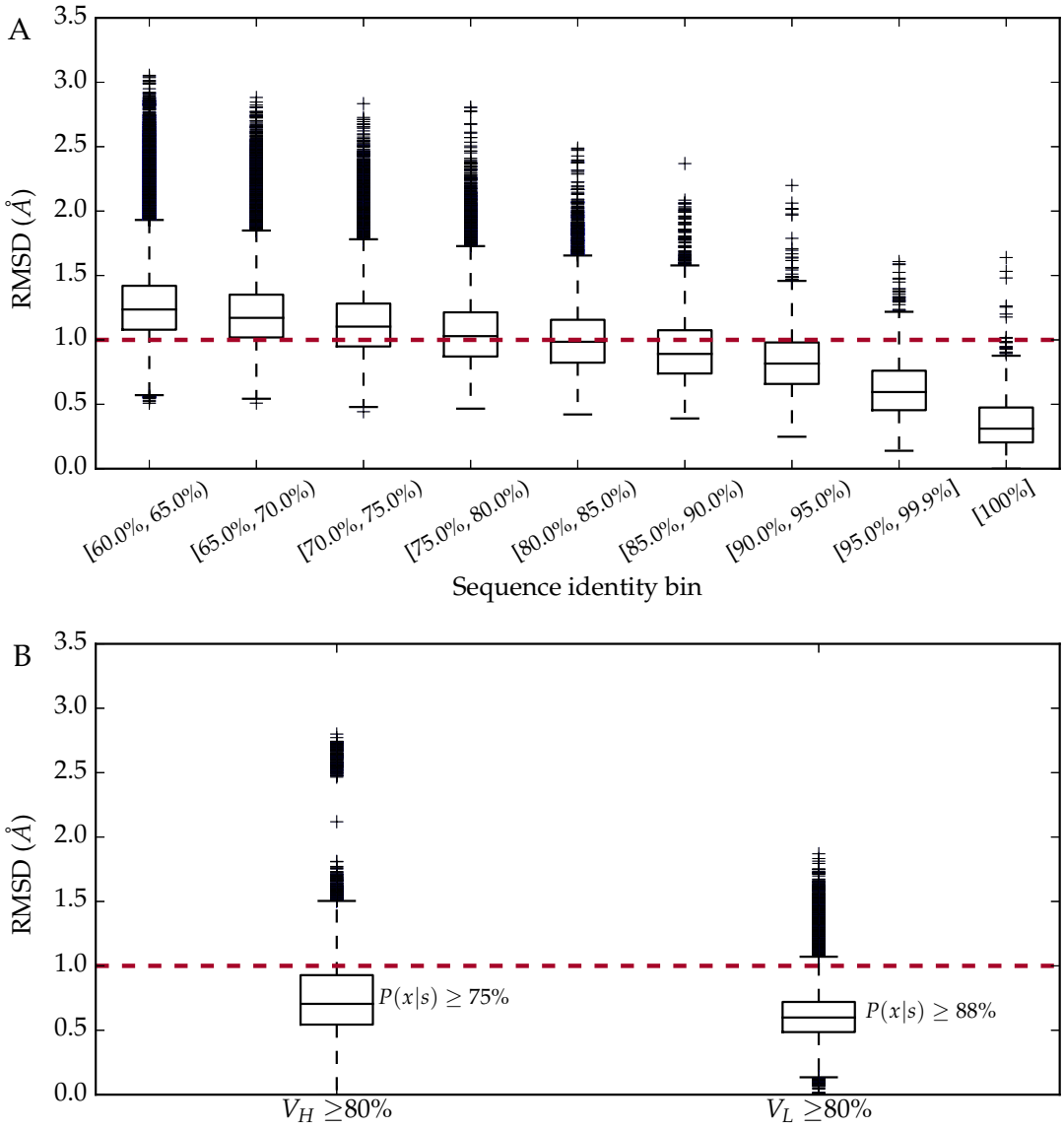


Figure 4.2: A. Boxplot of pairwise Fv–Fv framework region superimpositions in the redundant set; only pairs with sequence identity ≥60% are shown. **B.** Boxplot of pairwise V_H–V_H framework region superimpositions and V_L–V_L framework region superimpositions where sequence identity ≥80%.

Table 4.2: Two example cases of framework template selection.

Target Antibody	V_H Template (% SeqID)	V_L Template (% SeqID)	Orientation Template
1h0d:BA	1ejo:H (89%)	1ejo:L (86%)	1ejo:HL
12e8:HL	3nig:E (90%)	1i3g:L (95%)	1i3g:HL

SeqID: Sequence identity. In the case of 1h0d:BA, a global template is used as both chains of 1ejo:HL have $\geq$80% sequence identity to the target. In the case of 12e8:HL, a 'hybrid' template is used, as 1i3g:H has a sequence identity of 79.8%, and thus a second structure is used to predict the V_H domain. However, 1i3g:HL had the highest global sequence identity (87%) and is thus used for re–orientation of the V_H and V_L domains.

chains' framework regions. In this scenario, we expect to have a sub–Angstrom template for the V_H and V_L domains with a probability of 0.75. If either chain has <80% sequence identity to the target, two separate structures are used, and the orientation of the highest sequence identity global template is used (example template selections are given in Table 4.2).

4.3.1.2 Modelling the CDR loops

Once a template framework structure is selected, ABodyBuilder uses FREAD, a database method, to model the CDR loops (Deane and Blundell, 2001; Choi and Deane, 2010, 2011). A CDR–specific database was used for each CDR loop; if a suitable decoy was not found in the database, a Fv–specific database was used. If no decoy is still found, the most sequence–similar, length–matched CDR loop (based on its BLOSUM62 score) is used as the template. If no length–matched templates are found, the most sequence–similar loop is then used as the template for *ab initio* modelling by MODELLER (see Methods, Šali and Blundell, 1993).

Figure 4.3 shows the accuracy of individual CDR loop predictions from FREAD on template framework structures for our non–redundant set. In this initial assessment, the RMSD between the model and native CDR loops was calculated after superimposing both chains' framework regions' backbone atoms (*i.e.*, excluding the CDR loops). CDRL2 was modelled with the highest accuracy

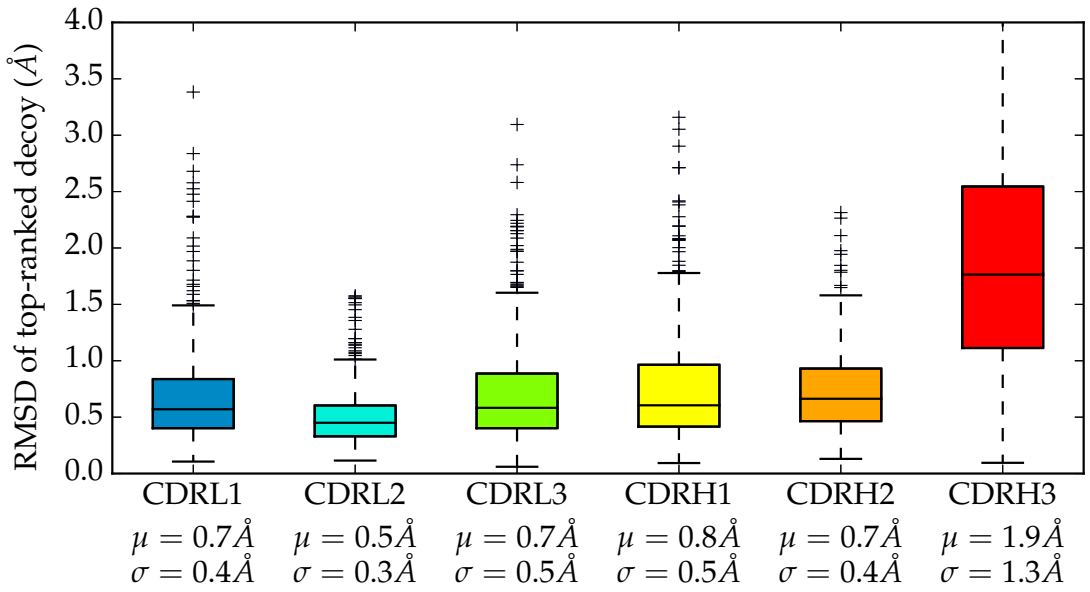


Figure 4.3: RMSD distributions of the top–ranked decoy from FREAD for each CDR loop. FREAD was used to model individual CDR loops on template framework structures of our non–redundant set. The RMSD was calculated by superimposing the backbone atoms of the framework regions of the template and target. Decoys with RMSD >4Å are not displayed. μ: mean; σ: standard deviation.

(average backbone RMSD 0.5Å), followed by CDRL1, CDRL3, CDRH2, CDRH1. CDRH3 was modelled with the lowest accuracy (average backbone RMSD 1.9Å).

The order of CDR loop modelling is important as each modelled CDR may influence the conformation of the next CDR loop. We used the accuracy of predicting individual CDR loops and the occurrence of Cβ–Cβ contacts between CDR loops (Figures 4.4, 4.5) to decide the ordering. The CDR loops are modelled in the following order: CDRL2, CDRH2, CDRL1, CDRH1, CDRL3, and CDRH3. The CDRL2 loop is modelled first as it is usually predicted with the highest accuracy. Next, CDRH2 is modelled as it is the best predicted CDR loop within the heavy chain, and is not in contact with CDRL2. CDRL1 and CDRH1 follow on as they are the next–best predicted CDR loops on the light and heavy chains, respectively. Finally, the CDRL3 is modelled before CDRH3. An alternative order of CDRL2, CDRL1, CDRL3, CDRH2, CDRH1 and CDRH3 was considered on the basis of FREAD accuracy per variable domain. However, results were

unaffected, so the proposed order was retained. When modelling a nanobody, the order is conserved, *i.e*. CDRH2, CDRH1, CDRH3.

4.3.1.3 Side chain modelling

At this stage in the ABodyBuilder methodology, we have a complete backbone, and potential side chain predictions where the template and target share identical residues. The side chains of the target antibody could be modelled by two different methods. 'Complete' prediction, where every side chain is predicted, or alternatively, 'partial' prediction where side chains of identical residues from the template are retained, and the remaining side chains are predicted. The side chains of the framework region and the CDR loops were either completely re–modelled or partially predicted using SCWRL4 (Krivov *et al.*, 2009). Side chain clashes occasionally arise from SCWRL4 predictions. We removed these by subjecting the clashing side chains to an initial round of relaxation by MODELLER (Šali and Blundell, 1993). If clashes are still found in the structure, all side chains are relaxed. The χ_1 angle and χ_{1+2} angle accuracies of the complete and partial predictions were compared. Our analyses showed that preserving side chains of common residues leads to better accuracy than re–modelling every side chain (Figure 4.6); however, MODELLER relaxation marginally decreases the χ_1 angle accuracy at the expense of resolving clashes in the model structure.

4.3.1.4 Confidence Estimates

ABodyBuilder estimates the confidence of the model antibody structure as the probability that a region (*e.g*. CDRL3) will be modelled within xÅ given the sequence identity or loop length (Figure 4.7). Thus, the confidence calculations can also be used to obtain the expected RMSD for a specified probability. The confidence measures are data–driven, and are based on the results from the pairwise framework region superimpositions or the FREAD predictions for the individual CDR loops. They are empirical approximations, and are not *de facto*

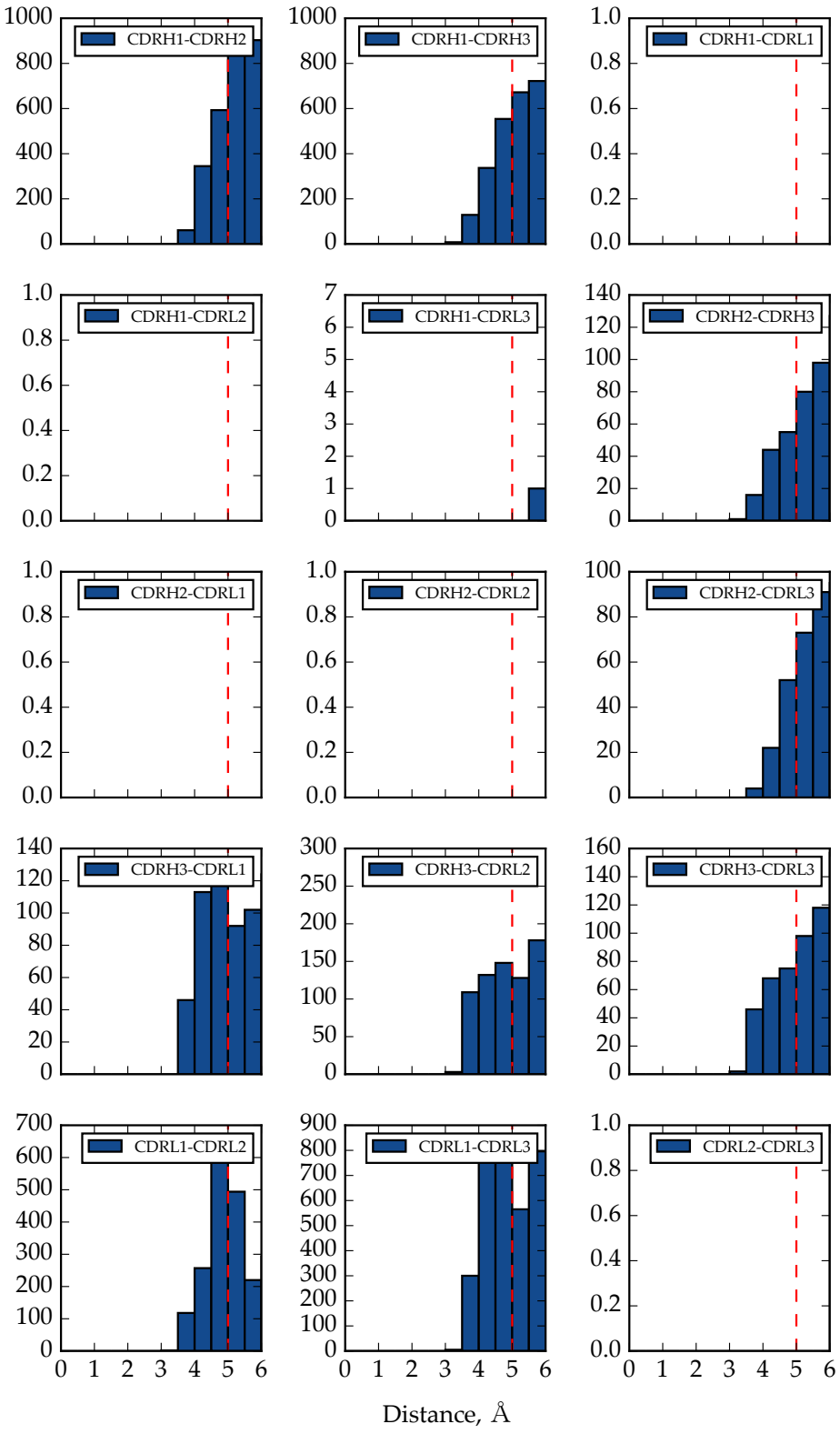


Figure 4.4: Histogram of Cβ–Cβ contacts between CDR loops of antibodies in the non–redundant set. The number of contacts within 5Å (red dotted line), and the accuracy of modelling CDR loops (Figure 4.3) were used to determine the order of CDR loop modelling.

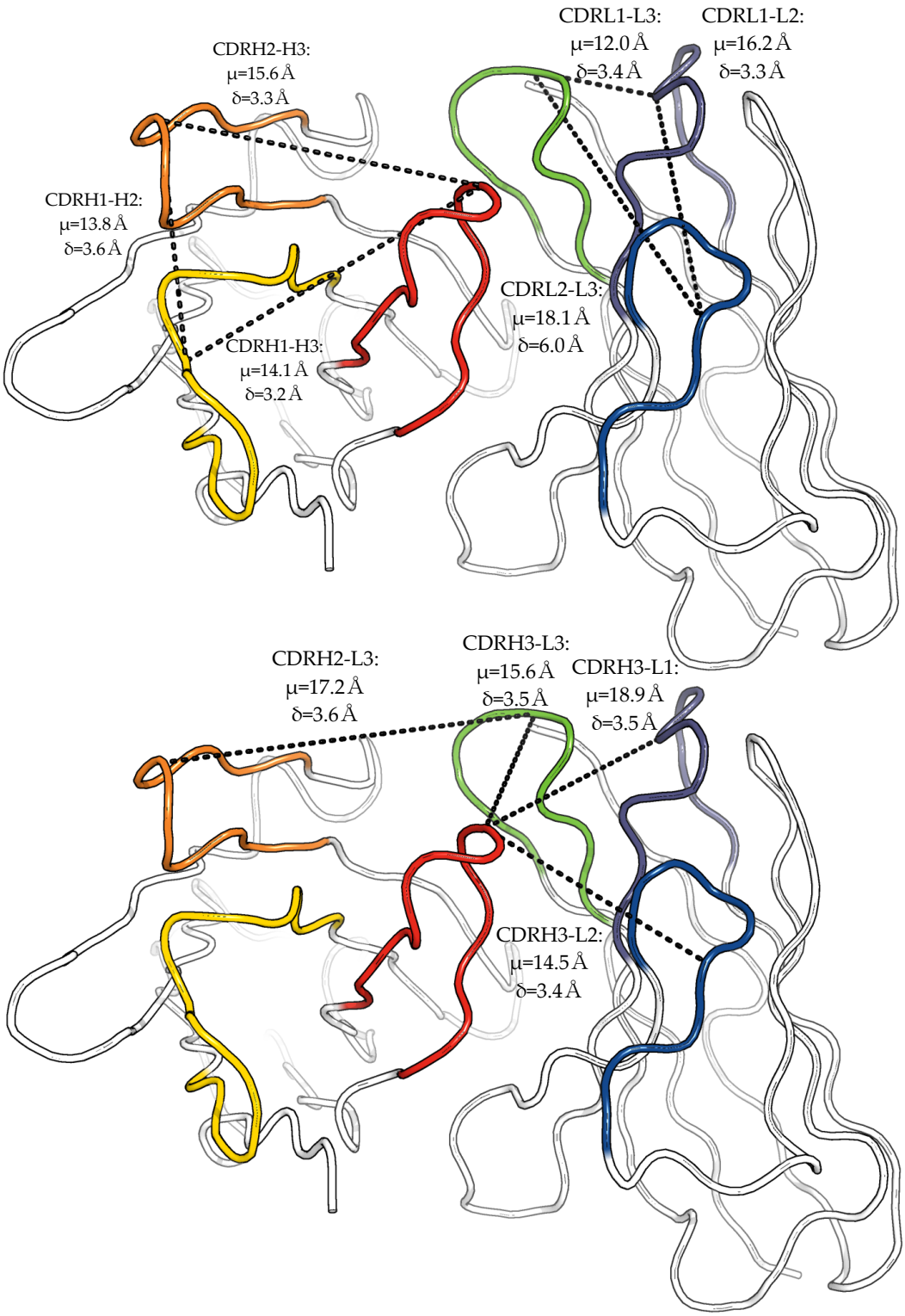


Figure 4.5: Mean and minimum Cβ–Cβ distances between CDR loops. If a pair of CDR loops' minimum Cβ–Cβ distance is >5Å (Figure 4.4), it is not shown. Top: Cβ–Cβ contacts between CDR loops within each variable domain. Bottom: Cβ–Cβ contacts between CDR loops between variable domains. μ: mean; δ: minimum distance.

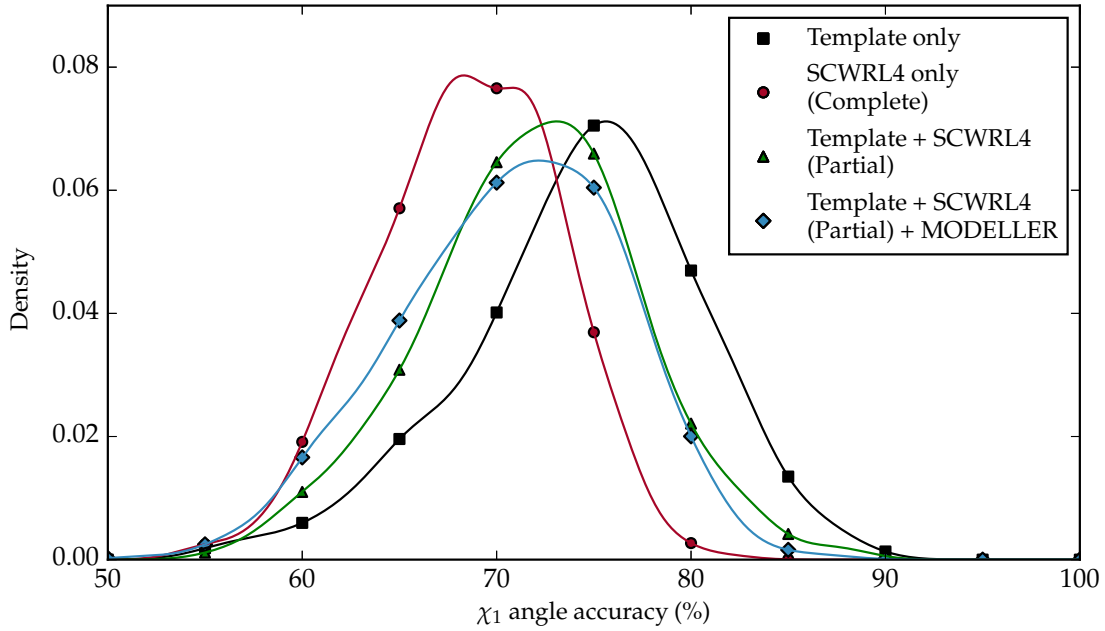


Figure 4.6: Density plot of χ_1 angle accuracy (%) for side chain prediction, using only the template's rotamers, completely re–modelling every side chain using SCWRL4, or using both the template rotamers where available, and SCWRL4's rotamers elsewhere. The χ_1 angle accuracy of models that were refined by MODELLER is also shown. Note that the χ_1 angle accuracy for 'Template Only' is calculated from fewer rotamer comparisons as it is based on comparing only the identical residues between the template and target.

RMSD values for the model. We use our ∼1.2 million pairwise superimpositions of framework regions (Figure 4.2) to estimate these confidence values.

We also estimate the accuracy of the CDR loops as a function of loop length. The backbone RMSDs of the top–ranked FREAD predictions on template framework regions for our non–redundant set were used to estimate the confidence of modelling a CDR loop.

4.3.2 Benchmarking ABodyBuilder

In order to benchmark ABodyBuilder, it was tested on the antibodies in the non–redundant set, and on our blind test set, a set of 136 structures that have been deposited in SAbDab since we built our methodology. When modelling these structures, sequence–identical antibodies were ignored. The accuracy of the models was calculated as described in Section 4.2.3. Briefly, the Fv and

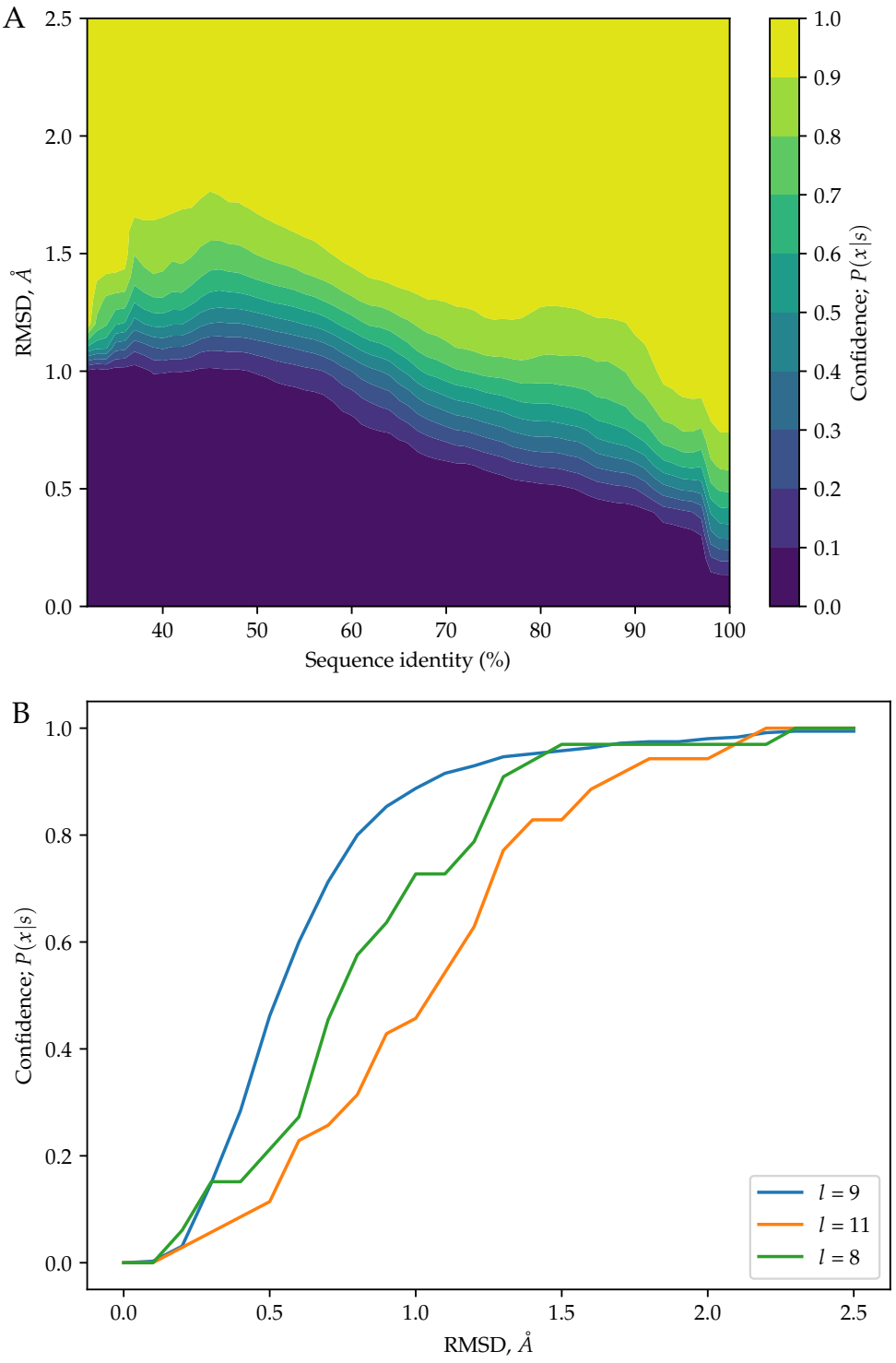


Figure 4.7: A. Conditional probability (Equation 4.1) contours for V_H framework region accuracy. Framework superimposition data on the redundant set (Figure 4.2) was used to calculate the probability that a framework region will be modelled within xÅ. The calculations depend on the sequence identity of the template. **B.** Conditional probability curve for CDRL3 loop accuracy. The calculations depend on the length of the CDRL3 loop, and the three most common CDRL3 loop lengths are shown.

framework regions' RMSD was calculated after superimposing the backbone atoms for both chains (Fv RMSD), or for each chain separately (framework region RMSD). For the CDR loops, the backbone atoms of each chain's framework region were superimposed, and the RMSD between the model and native CDR loops was calculated thereafter, similar to the method used in the AMA–II competition (Almagro *et al.*, 2014). Our RMSD calculations use North *et al.* (2011)'s CDR loop definitions, whereas the AMA–II competition calculated a model's RMSD using Chothia's definitions (Almagro *et al.*, 2014). As discussed later, when benchmarking ABodyBuilder on the AMA–II dataset, we used the Chothia definition to be consistent with the competition (Chothia and Lesk, 1987).

ABodyBuilder gives confidence estimates in its structural predictions. A complete antibody is modelled as eight regions – two framework regions, six CDR loops – and each of these is separately considered. Comparably, a nanobody's four regions (one framework, three CDR loops) are annotated individually. In all our models, a default confidence of 75% was used to calculate the expected RMSD. This value indicates, based on our framework region superimpositions and FREAD results on individual CDR loops, that there is a 75% chance that a component will be modelled within xÅ. These confidence measures were used to identify components that were difficult to model.

ABodyBuilder modelled the 462 Fvs in the non–redundant set with an average backbone RMSD of 1.3Å. On average, the canonical CDR loops were predicted with sub–Angstrom accuracy (average backbone RMSD 0.5Å, 0.4Å, 0.6Å for CDRL1, CDRL2, and CDRL3, respectively, and 0.6Å, 0.6Å, 1.9Å for CDRH1, CDRH2, and CDRH3, respectively). The RMSD values are lower than those in the original investigation using FREAD to predict CDR loops (Choi and Deane, 2011). It appears that the increase in available structural data has led to an improvement in CDR loop modelling. Sixty of the 79 nanobodies in the non–redundant set were V_HH antibodies, with an average domain RMSD of 2.6Å; the average backbone RMSD for the CDRH1, CDRH2, and CDRH3 loops were 1.4Å, 0.9Å, and 3.5Å, respectively. On the other hand, the domain RMSD

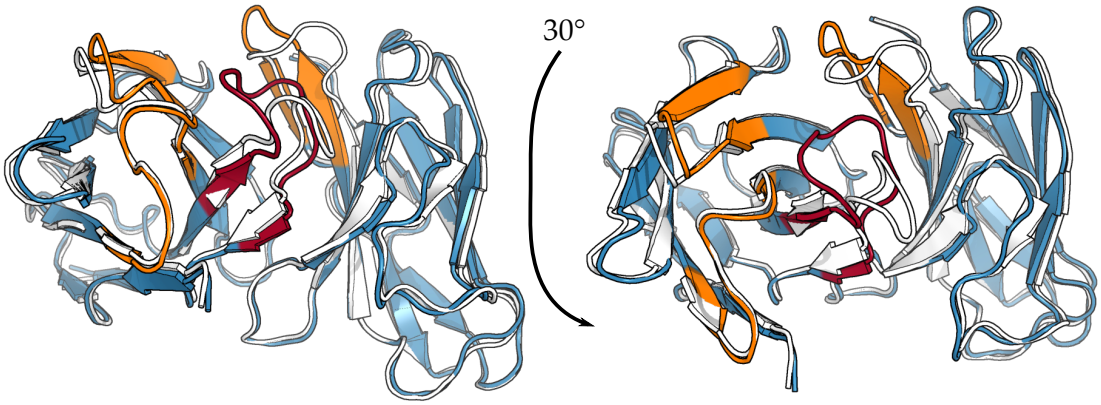


Figure 4.8: Example case where ABodyBuilder's confidence metric can indicate regions that may be poorly modelled. The native structure of 2vxv:HL is coloured white, and the model is coloured blue (expected region accuracy: $\leq$1.0Å with probability of 0.75), yellow (expected region accuracy: 1.0–2.5Å), or red (expected region accuracy: 2.5–4.0Å). The region with the worst expected accuracy (red; CDRH3) has the greatest RMSD with respect to the native structure.

of the 19 V_L–only nanobodies was 1.0Å; the average backbone RMSD of the CDRL1, CDRL2, and CDR3 loops were 0.6Å, 0.3Å, and 1.1Å respectively.

The confidence metric was particularly useful in identifying CDR loops that were modelled poorly. For instance, the CDRH3 loop of 2vxv:HL was estimated with 75% confidence to be modelled within 3.1Å, and its actual RMSD was 3.0Å (Figure 4.8). However, there are some cases where, at the default confidence level of 75%, ABodyBuilder can over– or under–estimate a region's accuracy. For example, ABodyBuilder was 75% confident that the CDRH1 loop of 3aaz:HL is modelled within 1.3Å, though its actual RMSD was 1.7Å. The framework regions' confidence measure is relatively robust, but in the case of the CDR loops, the lack of data can lead to less accurate confidence estimates.

ABodyBuilder was also used to build models of 136 structures (108 Fvs, 24 V_HH antibodies, 4 V_L–only antibodies) that were deposited in the PDB between 24 February 2015 and 20 December 2015 (our blind test set). Here, the average backbone Fv RMSD of Fvs was 1.5Å. Similar to the non–redundant set, the 'canonical' CDR loops of Fvs were predicted with sub–Angstrom accuracy (average backbone RMSD 0.6Å, 0.5Å, and 0.8Å for CDRL1, CDRL2 and CDRL3,

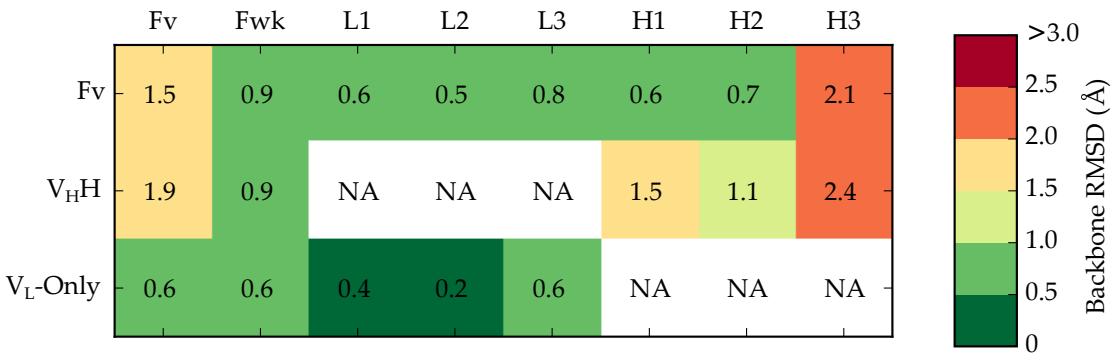


Figure 4.9: Heatmap of average backbone RMSD in the blind test set of 136 antibodies, divided into either Fvs, V_HH, or V_L–only antibodies. The RMSD of each component was calculated as described (Section 4.2.3). North *et al.* (2011)'s CDR definitions were used. Fwk: framework; H1, H2, H3, L1, L2, L3 refer to CDRH1, CDRH2, CDRH3, CDRL1, CDRL2, and CDRL3, respectively.

respectively; 0.6Å and 0.7Å for CDRH1 and CDRH2, respectively; Figure 4.9). For the 24 V_HH antibodies, the domain RMSD was 1.9Å, and the CDRH1 and CDRH2 loops were predicted with average backbone RMSDs of 1.5Å and 1.1Å (Figure 4.9). In contrast, the V_L–only antibodies were predicted with sub–Angstrom accuracy for the entire domain (0.6Å), and for the CDR loops (0.4Å, 0.2Å and 0.6Å for CDRL1, CDRL2, and CDRL3; Figure 4.9).

Despite the low averages, the blind test set posed several challenges. Some canonical CDR loops were poorly modelled (*e.g.* the CDRL1, CDRL2, and CDRL3 loops of one target, 4yfl:HL), and *ab initio* modelling was necessary for modelling the CDRL3 loop of 5c0n:CD. The average RMSD for the CDRH3 loops in the blind test set was 2.1Å for Fv antibodies, and 2.4Å for V_HH antibodies. Fifty–nine of the CDRH3 loops in this set were long ($\geq$15 residues), which may explain the high average backbone RMSD; furthermore, 51 of the CDRH3 loops were not modelled by FREAD (50 were modelled by a sequence similar template, one *ab initio*). For CDRH3 loops not modelled by FREAD, it was not possible to determine the confidence, but we suggest that they are likely to be modelled poorly. Overall, the results on these two datasets suggest that ABodyBuilder can generate high–quality models for most targets.

4.3.3 ABodyBuilder's performance on AMA–II targets

ABodyBuilder was also tested on the eleven antibodies from the AMA–II competition (Almagro *et al.*, 2014). To replicate the blind test conditions as far as possible, all structures that were deposited in the PDB after 31 March 2013 were omitted from the template search and FREAD databases. The accuracy of the models was calculated as described in Section 4.2.3, and here the Chothia–defined CDR loops were used as in the AMA–II competition (Almagro *et al.*, 2014).

The ABodyBuilder models were of similar quality to that of the methods used in AMA–II (Figure 4.10). The average RMSD for the whole Fv for our models was 1.2Å; this is comparable to other publically available pipelines: RosettaAntibody (1.1Å), Kotai Antibody Builder (1.1Å), and PIGS (1.5Å). Except for the CDRL2 loop where the average backbone RMSDs of ABodyBuilder models (0.3Å) were marginally lower (RosettaAntibody: 0.4Å, Kotai Antibody Builder: 0.3Å, PIGS: 0.5Å), the average backbone RMSD of all other CDR loops was far lower.

In particular, ABodyBuilder showed an improvement of >0.5Å RMSD for the CDRL3 and CDRH3 loops compared to the other pipelines. This is likely due to the choices made by FREAD. For Ab06 (PDB: 4m6o), ABodyBuilder selected 3hr5:HL as the template framework, as done by the Schrodinger group (Zhu *et al.*, 2014); however, ABodyBuilder used 1om3 over 2aab as its CDRH3 template. Despite the differences in environment–specific substitution scores (ESSS; 1om3: 26, 2aab: 47), FREAD's ranking based on anchor RMSD (1om3: 0.188Å vs. 2aab: 0.223Å) led to this selection, which was ultimately a better template for CDRH3. In many cases, ABodyBuilder generated the top, or joint–top prediction for a component of an antibody. However, there were some cases which were more challenging to ABodyBuilder in comparison to other pipelines, such as the CDRL1 loop of Ab05 (PDB: 4m6m, RMSD 2.8Å). Here, ABodyBuilder chose 1lgv as it had the lowest anchor RMSD (0.127Å), despite its low ESSS (31). Of the top 10 predictions, 3h42 had the highest ESSS (96) but was ranked fourth

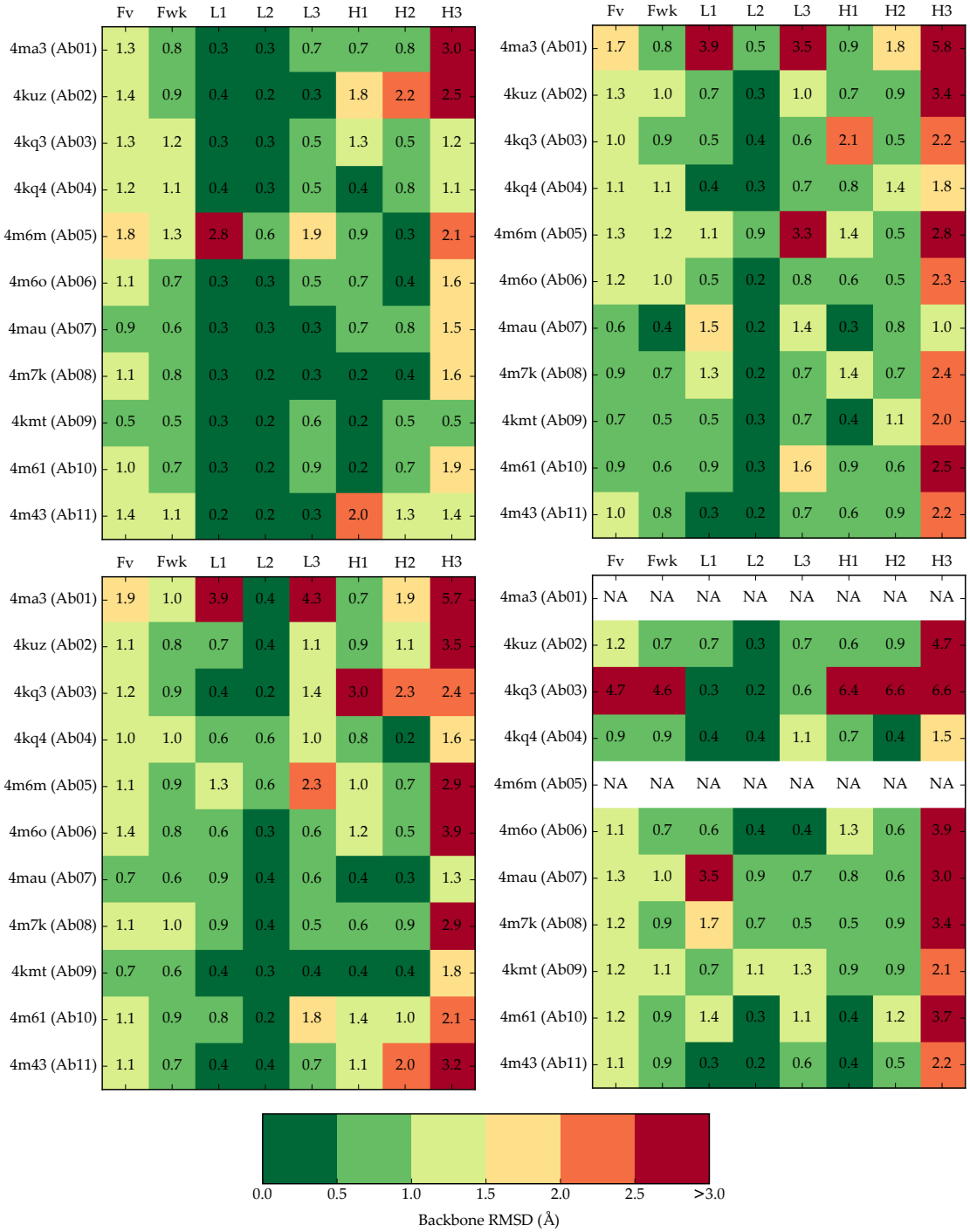


Figure 4.10: Backbone RMSD heatmap of different methods from the AMA–II competition, including ABodyBuilder (top–left), Kotai Antibody Builder (top–right), RosettaAntibody (bottom–left), and PIGS (bottom–right) (Shirai *et al.*, 2014; Sivasubramanian *et al.*, 2009; Marcatili *et al.*, 2014). The RMSD of each region was calculated as described in Section 4.2.3. ABodyBuilder was run using only structures that were deposited in the PDB by 31 March, 2013, and the templates for each component are described in Table 4.3.

Table 4.3: Template selection for the AMA–II set.

Antibody	Framework Template	SeqID(%)	Orientation Template	CDR Loop Templates					
				CDRH1	CDRH2	CDRH3	CDRL1	CDRL2	CDRL3
Ab01 (4ma3)	4jo2I–4jo2M	82.0	4jo2IM	3vfgH	2dquH	2pcpD	4jo2M	4jo2M	4jo2M
Ab02 (4kuz)	2w9dH–3mbxL	95.5	3o2dHL	2w9dH	2w9dH	1pg7Z	3mbxL	3mbxL	3hi5L
Ab03 (4kq3)	3macH–3eo9L	99.0	2cmrHL	3macH	3macH	1pg7I	3eo9L	3eo9L	3qpxL
Ab04 (4kq4)	3mxwH–3mxwL	85.0	3mxwHL	3mxwH	3mxwH	2g60H	3mxwL	3mxwL	3mxwL
Ab05 (4m6m)	2xwtA–2xwtB	90.5	2xwtAB	2xwtA	2xwtA	1wejH	1lgvA	2xwtB	2xwtB
Ab06 (4m6o)	3hr5H–3hr5L	93.0	3hr5HL	3hr5H	3hr5H	1om3K	3difC	3sgdI	3e8uL
Ab07 (4mau)	1f58H–1f58L	92.0	1f58HL	1f58H	1f58H	1fl5B	1f58L	1f58L	1f58L
Ab08 (4m7k)	1d5iH–2ap2A	95.0	1ap2BA	1d5iH	1d5iH	1xf3B	2ap2A	2ap2A	2ap2A
Ab09 (4kmt)	3nabH–3nabL	99.0	3nabHL	3nabH	1i3gH	3o2vH	3nabL	3nabL	2fr4A
Ab10 (4m61)	1kb5H–3ijhC	95.0	3ujtHL	1kb5H	1kb5H	1svzB	3ijhC	3ijhC	1orsA
Ab11 (4m43)	2w9dH–4gw5C	91.0	4dgiHL	2w9dH	4h0gA	4h0hB	4gw5C	4gw5C	4gw5C

SeqID: Sequence identity.

in terms of anchor RMSD (0.151Å); using this template would have led to a prediction with a backbone RMSD of 0.8Å.

4.3.4 Large–scale modelling of antibody sequences

Given the growing availability of large datasets of antibody sequences (Robinson, 2015), in particular from NGS (DeKosky *et al.*, 2015, 2016), a desirable aspect of an antibody modelling pipeline is the ability to rapidly generate models. To test the scalability of ABodyBuilder, the non–redundant set of 6267 (3490 human, 2373 mouse) paired antibody sequences from DIG–IT, AbYSis, and SAbDab were modelled. For any sequence that required *ab initio* intervention by MODELLER (Šali and Blundell, 1993), only one model was generated.

The average runtime for each sequence was 34 seconds, taking 222.9 seconds at most. In total, the entire set of 6267 paired sequences was modelled in 3552 CPU minutes (Figure 4.11), which is approximately 9.45 CPU hours per 1000 sequences. This compares to approximately 250,000 CPU hours per 1000 sequences that were modelled in a recent study modelling antibodies from an NGS dataset of human antibody sequences (DeKosky *et al.*, 2016). In this study, the framework region RMSD to a non–redundant set of crystal structures was 1.0±0.29Å. The CDR loops' RMSDs were only calculated for models of naïve antibodies, ranging from 0.8–2.4Å (DeKosky *et al.*, 2016). In contrast,

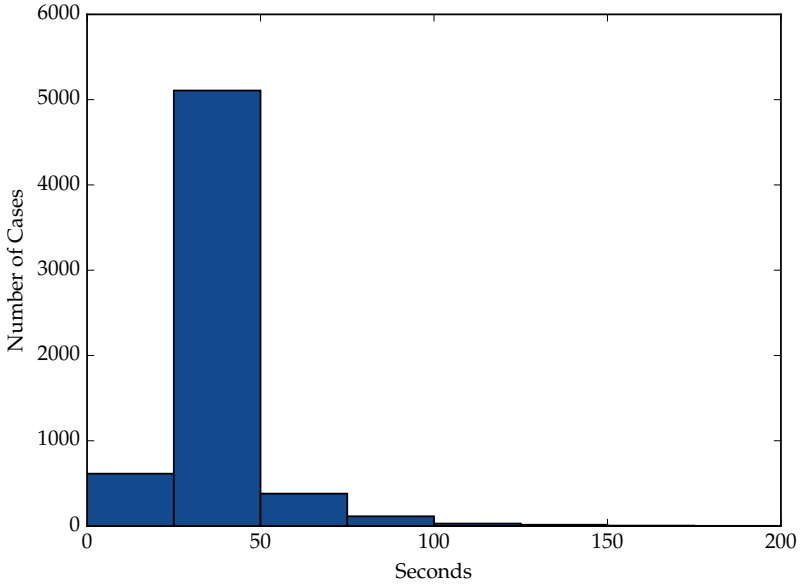


Figure 4.11: Time elapsed by ABodyBuilder in modelling the non–redundant set of 6267 paired antibody sequences. Time is measured from when the sequence is given as input to ABodyBuilder until it finishes re–numbering models into the IMGT numbering scheme. For each target requiring *ab initio* modelling by MODELLER (*e.g.* CDR loops), only one possible model was generated.

for every ABodyBuilder model, model confidence is annotated where possible. At the default 75% confidence level, the average expected RMSD values are 0.9Å and 0.7Å for the heavy and light chain framework regions, and 0.8Å, 0.7Å, 1.0Å, 1.1Å, 1.0Å, and 3.1Å for the CDRL1, CDRL2, CDRL3, CDRH1, CDRH2, and CDRH3 loops, respectively. The models are currently hosted on http://opig.stats.ox.ac.uk/webapps/abodybuilder/models, and the time elapsed for each model is shown.

4.3.5 Server output

As described above, ABodyBuilder can rapidly build accurate models of antibodies from sequence. Once a model has been generated, it can be downloaded by the user, or interactively analysed. ABodyBuilder is available as a web–app at http://opig.stats.ox.ac.uk/webapps/abodybuilder. For each model structure, an annotations page is created using PV (Figure 4.12; Biasini, 2015).

Here, the secondary structure elements, domains, solvent exposure, confidence measures, and the sequence liabilities can be visualised on the model structure. For example, if a particular target antibody has an N–linked glycosylation motif in its CDRL1 sequence (Asn–X–Ser/Thr, where X is not proline), this portion of the CDRL1 is then highlighted if its relative accessible surface area is >10% (Gavel and von Heijne, 1990; Kabsch and Sander, 1983). Each sequence liability is given a unique colour, allowing simultaneous visualisation of multiple liabilities (should they be present). A full list of annotated sequence liabilities is provided in Table 4.1.

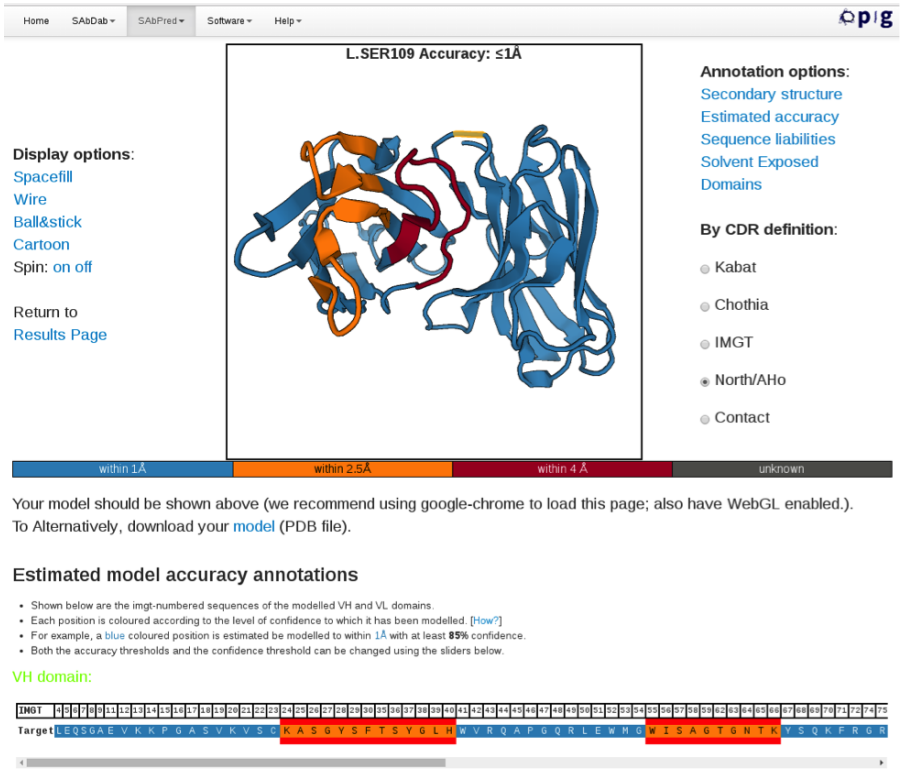


Figure 4.12: Screenshot of an example ABodyBuilder output (annotations page), generated using PV (Biasini, 2015). Users have the freedom to visualise specific features and download the model and accompanying data.

4.4 Discussion

In this Chapter, we have described an antibody modelling pipeline that builds antibody models that are comparable to other current freely available methodologies, but is the first to offer model accuracy estimates. ABodyBuilder can also model Fvs or nanobodies. ABodyBuilder follows the archetypal four–stage workflow that is characteristic of most antibody modelling pipelines. In order to design ABodyBuilder, we tested several strategies for every step and selected the best method given the currently available structural data.

The framework region's RMSD was rarely above 1.5Å in all the models generated in this chapter. The framework region had the highest average RMSD after CDRH3; this may be due to the length dependence of RMSD. However, the relatively high RMSD value may also stem from two features: V_H–V_L orientation, and the framework loops. Currently, ABodyBuilder re–orients the V_H and V_L chains using the ABangle parameters from the best available global template. However, predictions from machine learning methods (Bujotzek *et al.*, 2015b) or energy minimisation (Sivasubramanian *et al.*, 2009; Berrondo *et al.*, 2014; Marze *et al.*, 2016) could be used to enhance prediction in the future. The choice of the template could also be enhanced by the use of detailed structural features – for example, 'bulge' structures near the N–terminus of the V_H domain (Shirai *et al.*, 2014).

ABodyBuilder models the 'canonical' CDR loops with sub–Angstrom accuracy (on average) in the blind test and AMA–II datasets. In the AMA–II set, the CDRL1 and CDRL3 loops were modelled particularly well in comparison to the other methods (Figure 4.10), reinforcing the benefits of a knowledge–based approach (Almagro *et al.*, 2014). However, FREAD was incapable of generating a decoy in some cases (e.g. CDRL3 of 5c0n:CD). This is perhaps due to the lack of data in FREAD's databases; for example, 5c0n:CD represents a rabbit antibody. In 31 March 2013 (the deposition date cutoff for benchmarking with AMA–II targets), there were only six rabbit antibody structures, and as of 27 January

2016, there are only eleven in SAbDab (Dunbar *et al.*, 2014; Berman *et al.*, 2000). However, as FREAD's performance has already improved since the previous benchmark (Choi and Deane, 2011), and the ever increasing amount of antibody data in the PDB, we expect predictions to continue to improve over time.

The choice of a poor framework template could have affected the prediction of the CDRL1 loop of Ab05. The template's light chain framework region had a backbone RMSD of 1.5Å. Furthermore, the target structure has an unusual tilt angle, as previously commented (Bujotzek *et al.*, 2015a); our chosen template had a high deviation in tilt angle with respect to the target structure. Both features could have affected the choice of the CDRL1 loop by FREAD.

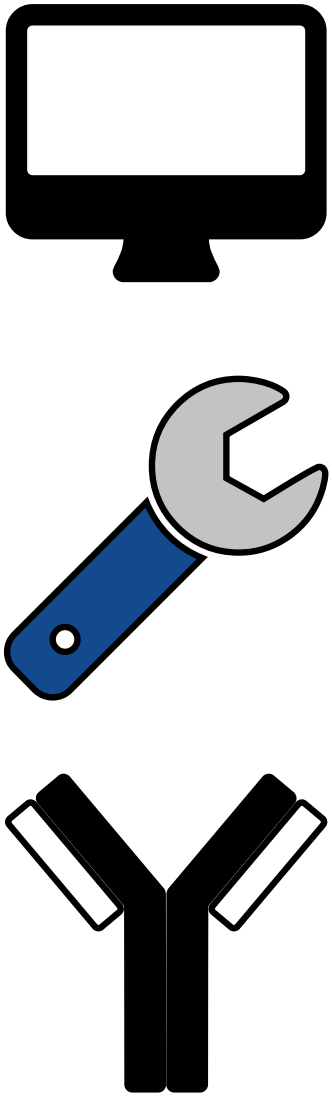
Once ABodyBuilder generates a model structure, it displays several features of the antibody model, including the confidence measurements and sequence liabilities. Both metrics are intended to assist users in pursuing *in vitro* development of their target antibody. The confidence measures successfully identified structures that were modelled poorly. This feature of ABodyBuilder will help users to identify cases where current methodologies may not be able to produce high–quality decoys. However, the confidence calculations assume that the distribution of antibody sequences in SAbDab is representative of all antibody sequences. Therefore, if a query sequence is very different from those in SAbDab, the assumptions underlying our confidence measures will not hold, and incorrect estimates may occur.

Currently, the side chains are predicted with modest accuracy (average χ_1 angle accuracy 71% using template's rotamers, SCWRL and MODELLER; Krivov *et al.*, 2009; Šali and Blundell, 1993). Unlike other components, the side chains do not have separate confidence measurements. Since the χ_1 angle is dependent on the backbone geometry (Krivov *et al.*, 2009), side chain confidence should be approximated by confidence estimates of the framework and CDRs. However, improving side chain prediction is of interest, due to its influence on antibody–antigen binding affinity (Chapter 2) and structure (*e.g.* orientation;

Dunbar *et al.*, 2013). In the next chapter, we will discuss the development of an antibody–specific side chain prediction methodology.

ABodyBuilder currently identifies 11 possible sequence liabilities that are known to affect antibody development (Table 4.1). In our non–redundant set, ABodyBuilder identified an N–linked glycosylation site in the V_H framework region of cetuximab (PDB: 1yy8, position H97) (Qian *et al.*, 2007), and Asp isomerisation sites in the CDRL1 of omalizumab (PDB: 2xa8, positions L34, L36) (Sydow *et al.*, 2014). However, these liabilities were identified solely on the basis of the sequence motif and solvent accessibility. Thus, ABodyBuilder will flag sequence motifs as liabilities despite the fact that they may only be problematic in certain conditions (*e.g.* acidic environments).

ABodyBuilder automatically builds high–quality models for most targets in $\sim$30 seconds, allowing users to quickly obtain a model structure for any antibody sequence. Our tool uniquely estimates the model's confidence and flags sequence liabilities; in particular, we demonstrate that the confidence estimate can help identify components that have been poorly modelled. ABodyBuilder serves to facilitate antibody design by translating candidate antibody sequences into structural prototypes for further study, such as antibody–antigen docking and antibody humanisation.



Simplicity is the ultimate sophistication.

— Leonardo Da Vinci

5 An antibody position–dependent library for side chain prediction.

Contents

5.1 Introduction

In the previous Chapter, we described our antibody modelling pipeline, ABodyBuilder. Side chains are predicted in the last stage, and this is currently the only component that does not rely on antibody–specific data. Here, we describe the development of an antibody–specific, position–dependent rotamer library.

5.1.1 The side chain prediction problem

An amino acid's side chain is described by a series of torsion angles, known as the χ angles. Apart from Ala and Gly, the remaining 18 amino acids have between one and four χ angles. For example, Met has three χ angles; they

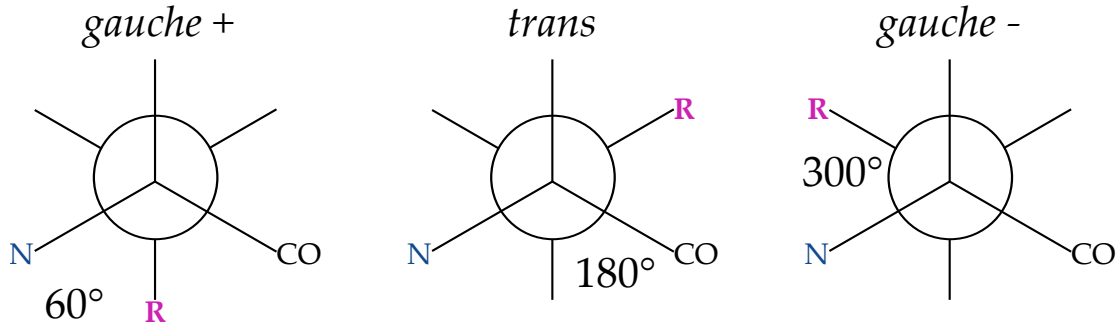


Figure 5.1: Newman projection of rotamers. The rotameric states ($g+, t, g-$) are defined according to their χ angle. The χ_1 angle, as shown here, represents the torsion angle about the Cα–Cβ bond. The $g+$ form peaks around 60°, the t form around 180°, and the $g-$ form around 300°.

represent the dihedral angle about the Cα–Cβ bond (χ_1), the Cβ–Cγ bond (χ_2), and the Cγ–Sδ bond (χ_3). Similar to the backbone's ϕ and ψ angles, the χ angles adopt a limited range of values, as some conformations can lead to steric hindrance. There are three favourable low–energy conformations, which peak at χ angle values of 60°, 180°, and 300°. These low–energy forms are known as the $g+$ (gauche +), t (trans), and $g-$ (gauche –) rotameric states, respectively (Figure 5.1; Shapovalov and Dunbrack, 2011).

The series of rotameric states for each χ angle is known as the amino acid's 'rotamer' (Shapovalov and Dunbrack, 2011). For instance, Ser only has one χ angle; thus there are only three rotamers for Ser residues. However, residues such as Met, with three χ angles, have 27 possible rotamers. A Met residue with $\chi_1 = g-, \chi_2 = g+, \chi_3 = t$ rotameric states is represented as the $\{g-, g+, t\}$ rotamer. On the other hand, groups of rotamers are represented by $*$. For instance, Met $\{g+, t, *\}$ represents all Met rotamers with $\chi_1 = g+, \chi_2 = t$.

Predicting a protein's side chains can have an impact in many applications, such as protein design, docking, and structure prediction (Krivov *et al.*, 2009; Harder *et al.*, 2010; Colbes *et al.*, 2016). In the context of antibody design, recovering near–native side chains could help structural prediction, *e.g.* in ranking CDR loop structure predictions. Furthermore, side chain prediction is

vital for predicting the structural impact of mutations during computational antibody design. For example, if a mutant side chain can be placed in the correct location, any resulting changes in antibody–antigen contacts can then be used to estimate differences in binding affinity. Ideally, the side chain prediction method should be able to predict side chains in model structures.

Predicting the rotamers of a side chain remains a huge combinatorics problem (Gaillard *et al.*, 2016). If we have a protein with 100 residues, where each residue has two χ angles (*i.e.* nine rotamers), then 9^{100} possible solutions exist. Thus, when developing side chain prediction methods, both accuracy and speed are highly desirable.

5.1.2 Rotamer libraries

Rotamer libraries are databases that describe the frequencies of specific rotamers, and additional information relevant to the rotamer. Most libraries encode rotamers by their average χ angles (*e.g.* Dunbrack and Cohen, 1997); thus, during prediction, side chains are reconstructed using an ideal set of bond lengths. In contrast, some libraries explicitly contain the coordinates for each observed rotameric structure (*e.g.* Xiang and Honig, 2001; Shetty *et al.*, 2003).

Often, rotamer libraries reflect the observed distribution of rotamers in a curated set of protein structures (*e.g.* Tuffery *et al.*, 1991; Lovell *et al.*, 2000; Dunbrack and Cohen, 1997). The Dynameomics rotamer library goes beyond the observed structural data from the PDB, and complements structural data with MD simulations to calculate rotamer probabilities (Towse *et al.*, 2016). Unlike other rotamer libraries, BASILISK is a generative model trained on structures from the PDB. It can then be used to predict χ angles and the log likelihood of a rotamer given the backbone (Harder *et al.*, 2010).

Each of the aforementioned rotamer libraries can be categorised as backbone–independent (*e.g.* Tuffery *et al.*, 1991), backbone–dependent (*e.g.* Shapovalov and Dunbrack, 2011), or 'alternative', *i.e.*, dependent on other structural features (*e.g.* secondary structures; Lovell *et al.*, 2000). Backbone–independent libraries simply

describe the distribution of a rotamer's χ angles (Tuffery *et al.*, 1991; Towse *et al.*, 2016). Backbone–independent libraries only offer a low–resolution overview of side chain conformations, and disregard the local protein environment (*e.g.* solvent exposure).

On the other hand, backbone–dependent libraries describe an amino acid's rotamer probabilities as a function of the backbone's ϕ/ψ angles, as the ϕ, ψ, χ angles are correlated (Dunbrack and Karplus, 1993). Typically, the ϕ/ψ angles are discretised into bins (*e.g.* $10^\circ \times 10^\circ$ bins), and conditional probabilities are calculated for each bin (*e.g.* Dunbrack and Cohen, 1997). Recent versions of backbone–dependent libraries have treated the backbone dihedrals as continuous variables (*e.g.* Harder *et al.*, 2010; Shapovalov and Dunbrack, 2011; Towse *et al.*, 2016), though some side chain prediction algorithms still use discretised libraries (Peterson *et al.*, 2014; Colbes *et al.*, 2016). Although backbone–dependent libraries are the most popular, they depend on sufficient coverage of backbone bins. Furthermore, they may not be suitable for side chain prediction on models, as the backbone can be inaccurate.

Alternative rotamer libraries use other structural features to describe rotameric probabilities. For example, rotamers have been classified depending on their secondary structures (*e.g.* Lovell *et al.*, 2000; Towse *et al.*, 2016); effectively, this groups a number of ϕ/ψ angle bins. Rotamer libraries can also be dependent on longer backbone conformations (Chinea *et al.*, 1995). In this method, protein fragments (five or seven residues long) were grouped based on a Cα RMSD threshold, and the amino acid type at the centre of the fragment. For both these types of libraries, a rotamer's probability is implicitly affected by the backbone.

For most rotamer libraries, probabilities are calculated using a non–redundant set of proteins from the PDB. This assumes that, for example, all proteins with identical backbone dihedral angles will have the same rotameric distribution. However, it may be that within different classes of proteins, different rotameric distributions may exist for identical backbone dihedral angles. Analyses on membrane proteins have shown that their side chains are more difficult to

predict (Peterson *et al.*, 2014). This suggests that either the energy functions underlying side chain prediction tools are not suited for membrane proteins, or that rotamer libraries do not fully recapture rotamer distributions in membrane proteins. Extending from these observations, and the fact that protein family–specific predictions lead to better performance (Ross *et al.*, 2013), it is also possible that current side chain prediction methods will perform sub–optimally for antibodies.

5.1.3 Side chain prediction methods

A wide range of tools have been developed to predict side chains (*e.g.* Krivov *et al.*, 2009; Miao *et al.*, 2011; Liang *et al.*, 2011). Most side chain prediction algorithms require three components: a rotamer library (*e.g.* Dunbrack and Cohen, 1997) for sampling, an energy function for filtering rotamers (*e.g.* Krivov *et al.*, 2009), and a search algorithm for determining the optimal solution (*e.g.* Desmet *et al.*, 1992).

SCWRL is one of the most popular side chain prediction methods, with over 1000 citations for SCWRL3 and SCWRL4 combined (Krivov *et al.*, 2009). SCWRL samples rotamers from a backbone–dependent library based on the rotamer's probability for the given backbone (Shapovalov and Dunbrack, 2011). A rotamer's energy is then calculated using a distance–dependent van der Waal's term and a hydrogen bond term. The possible rotamers for each residue are then represented as a graph, where residues are the vertices, and edges represent interactions between candidate rotamers in two residues. The graph then undergoes edge decomposition, dead–end elimination (DEE), and tree decomposition to determine a high–quality solution.

RASP (Miao *et al.*, 2011) is similar to SCWRL, with minor variations. It uses an earlier version of the backbone–dependent library from Dunbrack's group (Dunbrack and Cohen, 1997). Rotamer energies are determined using a variant of the OPUS–PSP energy function (Lu *et al.*, 2008). A graph is then constructed before DEE and Monte Carlo optimisation. OSCAR–star, which

showed the strongest performance in a benchmark by Peterson *et al.* (2014), uses the same rotamer library as RASP. OSCAR–star's energy function is comprised of distance and orientation–dependent terms, and the final solution is obtained via Monte Carlo simulated annealing (Liang *et al.*, 2011). BetaSCPWeb is a recently–developed side chain predictor which uses the backbone–dependent rotamer library from Shapovalov and Dunbrack (2011). It constructs a Voronoi diagram to predict a protein's side chains; however, it has lower accuracy than other methods, such as SCWRL (Ryu *et al.*, 2016).

Traditionally, a predicted side chain is considered 'correct' if its χ angle is within 40° of the native side chain (Peterson *et al.*, 2014; Colbes *et al.*, 2016). Therefore, the χ_1 accuracy represents the fraction of side chains with the correct χ_1 angle. Similarly, the χ_{1+2} accuracy represents the fraction of side chains where both the χ_1 and χ_2 angles are correct. Under these criteria, the benchmark by Peterson *et al.* (2014) reported a χ_1 accuracy between 80–89% and a χ_{1+2} accuracy between 57–72%,. However, a benchmark study by Colbes *et al.* (2016) reported χ_1 and χ_{1+2} accuracies of 86–89% and 75–80%, respectively. In both benchmarks, the accuracy values are based on predicting the side chains on crystal structures. Only a small number of studies have extensively tested such prediction on model structures (Chinea *et al.*, 1995; Lu *et al.*, 2008; Leem *et al.*, 2016). Both Lu *et al.* (2008) and Leem *et al.* (2016) show that predicting the side chains of model structures is more difficult.

In order to improve side chain prediction accuracy, Colbes *et al.* (2016) have argued that energy functions need to be more accurate. They calculated the 'maximum achievable accuracy' by sampling for the best possible rotamer from Dunbrack and Cohen (1997)'s library. Overall, the χ_1 and χ_{1+2} accuracies had theoretical maxima of $\geq$97% and $\geq$95%. Although these values effectively represent the coverage of each rotamer library, there are some caveats. By using two 'correct' predictions, the 'best' prediction can lead to clashes; in other words, not all predictions are feasible. Furthermore, this test will not reflect the maximum possible accuracy for model structures.

Given these values, Colbes *et al.* (2016) argue that existing libraries have a sufficient amount of rotamer data, and the challenge lies in scoring for the 'correct' rotamer. However, the study also showed that the correct rotamer is not necessarily the most probable. Thus, it is possible that a more accurate description of the rotamer landscape will provide better samples, which would ultimately improve scoring, and increase prediction accuracy.

In this Chapter, we present PEARL (Position–dEpendent Antibody Rotamer Library), our antibody–specific, position–dependent rotamer library. To our knowledge, no current rotamer libraries are specific for a particular protein family. Given the structural similarity of antibodies, we use the concept of position–dependence, where a position is defined in terms of an antibody's IMGT position. We first describe PEARL's coverage of the antibody rotamer landscape. Next, we apply our side chain prediction algorithm, PEARS (Position–dEpendent Antibody Rotamer Swapper), using the rotamer data from PEARL. Despite using antibody–specific information, PEARS shows lower accuracy than leading side chain prediction methods (*e.g.* OSCAR–star, Liang *et al.*, 2011) in predicting side chains on crystal structures. In contrast, PEARS outperforms other methods when predicting side chains on model structures. These results suggest that current side chain predictors are more applicable for crystal structures because the backbone is correct. However, PEARS may be a better technique when working with model structures as PEARL's method of data encoding is more robust to errors in the backbone from modelling.

5.2 Methods

5.2.1 Datasets

Antibody structures with resolution $\leq$2.5Å were downloaded from SAbDab in January 2016 (Dunbar *et al.*, 2014). Structures were clustered by CD–HIT (Li and Godzik, 2006), using a 90% sequence identity cutoff. This non–redundant set of 640 antibodies (617 V_H, 562 V_L chains) was used to build the PEARL library.

The same set of 640 antibodies was used to make two test sets. The 'crystal test set' is identical to the training set. We also made a 'model test set', where ABodyBuilder was used to generate a model structure for each structure in the crystal test set. Three structures in the crystal test set (1dzb:AA, 1yc7:B, 5d6c:HL) were removed as RASP was unable to predict the side chains on these structures (Section 5.2.5.2). However, all 640 model structures in the model test set were tested. For all side chain predictions, self–predictions were not allowed.

The dataset for the backbone–dependent rotamer library from Shapovalov and Dunbrack (2011) (henceforth referred to as Dunbrack RL) was downloaded in November 2015. The library contains 3983 protein chains; of these, nine antibody chains (750 rotamers) were removed.

5.2.2 Construction of PEARL

For every side chain in the non–redundant set of 640 antibodies, we calculated the χ angles using the atom types listed in Appendix Table B.4. Side chain χ angles were then discretised using rotamer definitions from the Dynameomics library (Towse *et al.*, 2016, Figure 5.2). Although some χ angles (*e.g.* χ_2 angle of Tyr) have previously been considered to be 'nonrotameric' (Shapovalov and Dunbrack, 2011), we defined rotamers for these angles, following Towse *et al.* (2016).

PEARL is organised hierarchically, starting from the IMGT position; at the next level, we categorise by amino acid type, then rotamer type. For each rotamer type, we have a list of observed rotameric structures with their coordinates and local tripeptide sequence (Figure 5.3). For example, position H21 has 62 Val $\{t\}$ rotameric structures. One of these is from 2cmr:HL; the Val in this structure is preceded by a Lys residue at H20, and a Ser residue at H22. Therefore, its local tripeptide sequence is KVS.

5.2.2.1 Density estimation of χ_1 angle modes

For each amino acid type α at IMGT position p, we estimated the χ_1 angle density by using a Gaussian mixture model (GMM). To estimate the density,

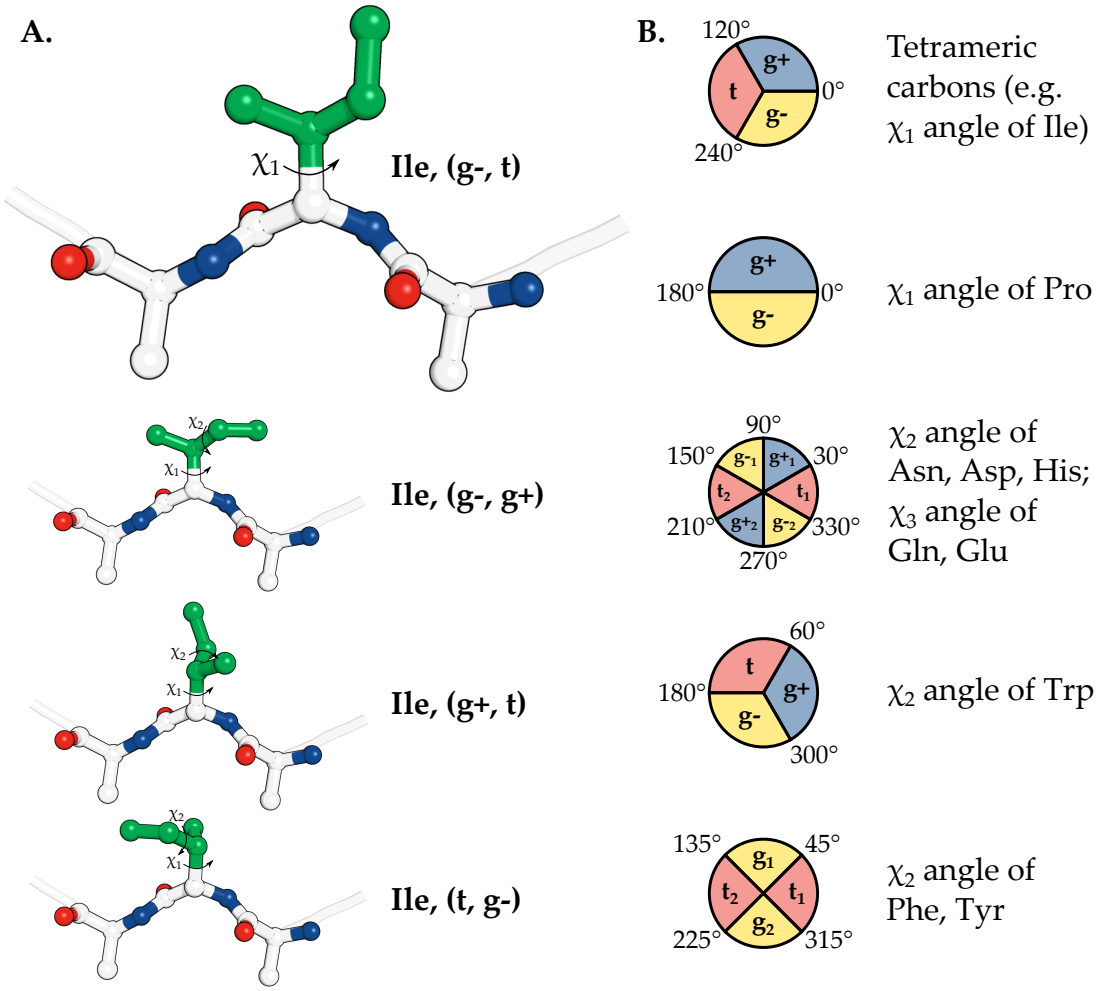


Figure 5.2: Definition of χ angles for building PEARL. **A.** The χ angles are torsion angles between atoms in all side chains, apart from Ala and Gly. The rotamer of a side chain refers to its set of discrete χ angles. We show four separate rotamers for an Ile residue. **B.** The χ angle definitions used in this Chapter follow those from Towse *et al.* (2016).

we imposed a minimum number of 20 rotameric structures per amino acid at an IMGT position. For example, at L109, 41 Asn residues are observed in our non–redundant set; in contrast, only eight Lys residues are observed at the same position. Thus, we estimated a χ_1 angle mode for L109 Asn, but not for L109 Lys. In total, there are 2384 'side chain types', *i.e.*, combinations of IMGT positions and amino acid types, *e.g.* H24 Lys. Of the 2384 side chain types, 1664 have less than 20 observations in our data.

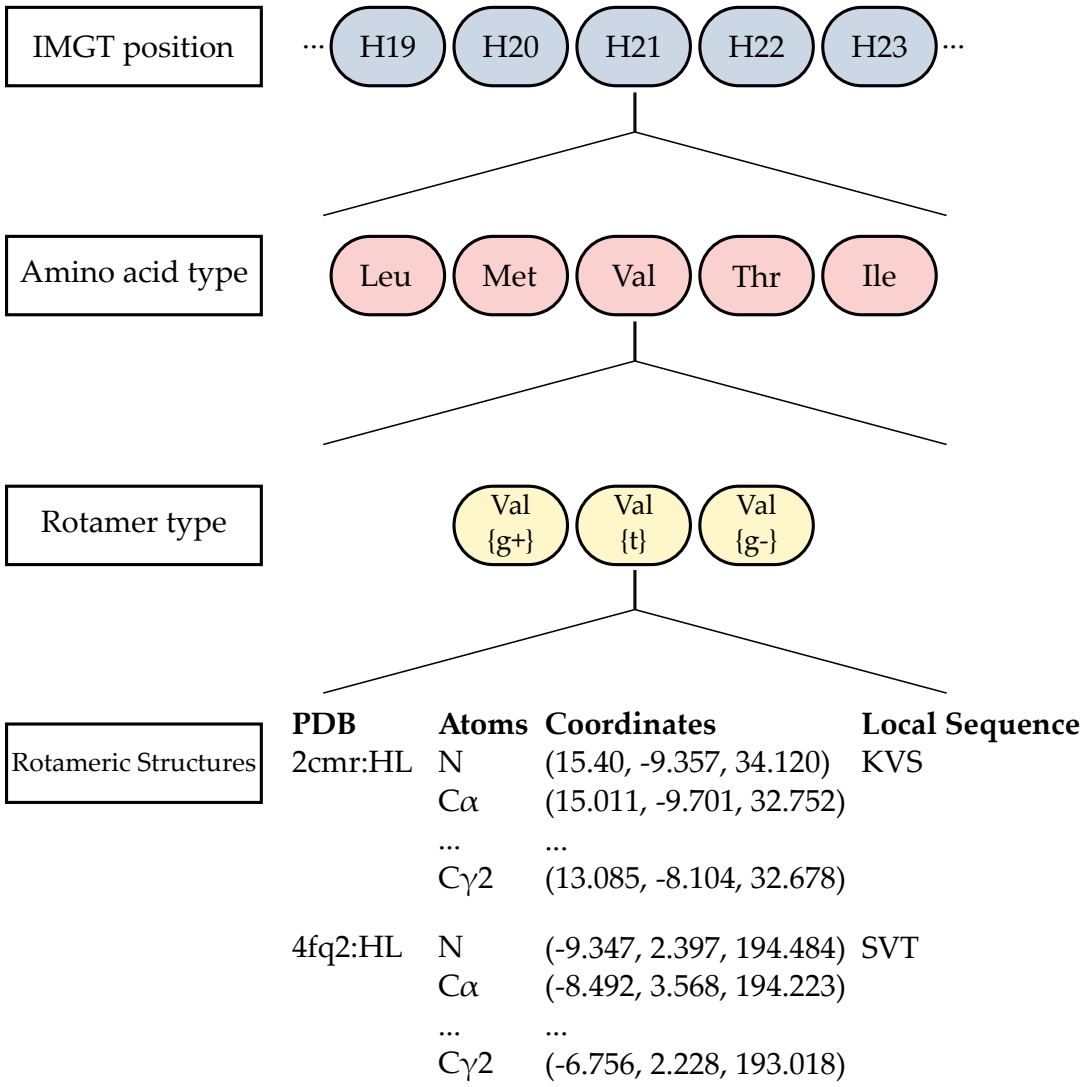


Figure 5.3: Organisation of rotamer data in PEARL. The library is categorised by IMGT position, amino acid type, then rotamer type. For each rotamer type, we have a list of coordinates for each observed rotameric structure from our non–redundant set of 640 antibody structures.

Briefly, a GMM represents a weighted sum of K Gaussian distributions,

$$p(\chi) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\chi|\mu_k, \sigma_k). \tag{5.1}$$

The kth Gaussian distribution has three parameters, $\{\pi_k, \mu_k, \sigma_k\}$, corresponding to the weight, mean, and covariance, respectively. We represent the K–long vector of weights as Π; depending on the values of π_k, we determine the mode, $\hat{\chi}$, for each α at p.

$$\hat{\chi}(\alpha, p) = \begin{cases} \text{Unimodal} & \text{if } \max(\Pi) \geq 0.8 \\ \text{Bimodal} & \text{if } \min(\Pi) \leq 0.1 \text{ and } \max(\Pi < 0.8) \\ \text{Trimodal} & \text{otherwise.} \end{cases} \tag{5.2}$$

The number of components, K, was decided by calculating the Akaike Information Criterion (AIC) for each GMM (Hastie *et al.*, 2009).

5.2.3 Maximum achievable accuracy estimation

We calculated the maximum achievable accuracy for PEARL and Dunbrack RL. For each side chain, we searched for a 'correct' rotameric structure from non–identical structures. Side chains were classified as χ_1 correct and χ_{1+2} correct if its χ angle(s) were within 40° of the native side chain. If the χ_1 angle is incorrect while the χ_2 angle is correct, we classified this as an incorrect prediction as the side chain would not be oriented in the correct location (Figure 5.4). Example selections are shown in Table 5.1.

The χ_1 accuracy represents the fraction of predictions with a correct χ_1 angle. Similarly, the χ_{1+2} accuracy represents the fraction of predictions where both the χ_1 and χ_2 angles are correct.

5.2.4 Side chain prediction algorithm

A flow chart of our side chain prediction methodology, PEARS, is given in Figure 5.5. When we select a rotameric structure from PEARL, we apply a filter based on the number of clashes, rotameric energy, local sequence identity, and if

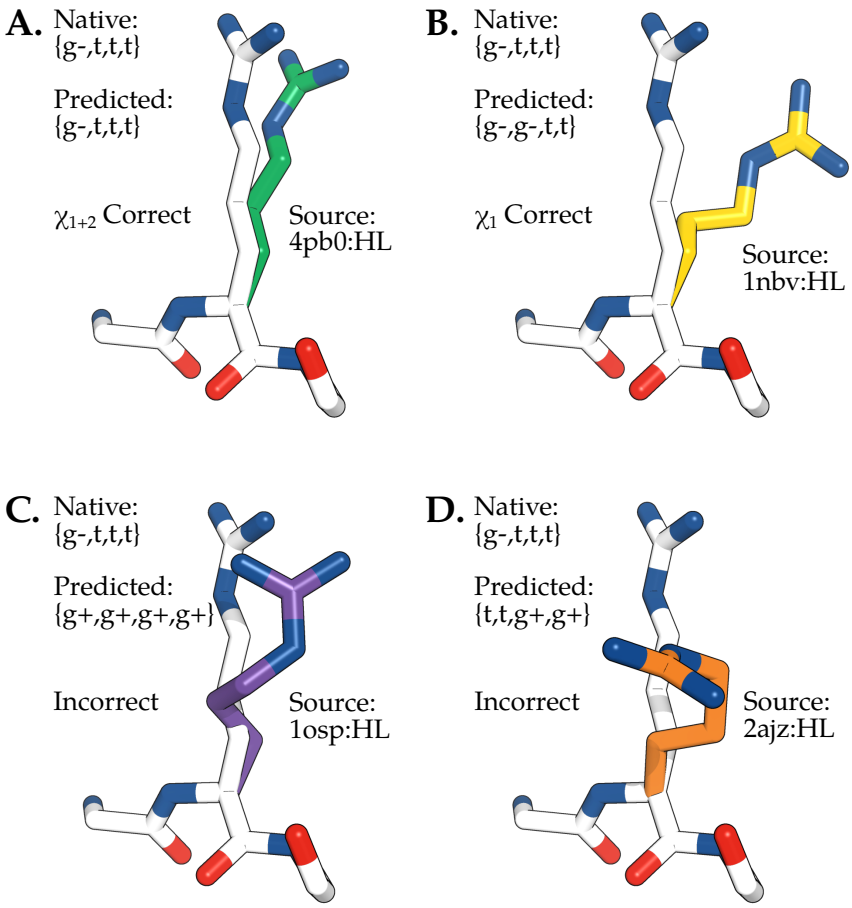


Figure 5.4: Classification of correct rotamers for an Arg residue at H57 (white; 2zkh:HL). To determine a 'correct' rotamer, the predicted χ angle must be within 40° of the native rotamer's χ angle. **A.** Rotamers are classified as χ_{1+2} correct if both χ_1 and χ_2 angles are correct. **B.** Rotamers are classified as χ_1 correct if only its χ_1 angle is correct. **C.** Example of an incorrect rotamer where both χ_1 and χ_2 angles have >40° deviation to the native side chain. **D.** If the predicted χ_1 angle is incorrect, yet the χ_2 angle is correct, this prediction is still considered to be incorrect. This is because the incorrect χ_1 angle points the side chain toward a different environment, making the 'correct' χ_2 angle meaningless.

Table 5.1: Determination of correct side chain predictions.

Native χ_1	Native χ_2	Template χ_1	Template χ_2	Verdict
100°	160°	120°	172°	χ_{1+2} correct
30°	150°	65°	27°	χ_1 correct
50°	60°	110°	50°	Incorrect

necessary, backbone RMSD. We first describe the order in which we make our predictions, followed by our filter that selects rotameric structures.

5.2.4.1 Order of side chain prediction

We first predict IMGT positions of a target with a unimodal χ_1 distribution. For example, at L116, Tyr residues are unimodal, with a peak at $\chi_1 = g-$ (Figure 5.8B). If the target structure has a Tyr at this position, we predict this residue, along with any other unimodal side chain types, first. When predicting unimodal side chain types, we only sample rotameric structures with the same χ_1 rotamer. Thus, for L116 Tyr, we only use rotameric structures whose χ_1 angle is in the $g-$ rotameric state. If none of the rotameric structures pass the backbone RMSD filter, Tyr structures with the $g-$ rotamer are selected from all positions in PEARL.

Once we predict the side chains of these unimodal side chain types, we then predict positions with the greatest number of predicted residues within a 4.5Å neighbourhood. In some cases, a position's environment may be populated entirely by unimodal side chain types. Hence, these positions are predicted next.

By this stage, PEARS is typically predicting the side chains of two or more positions that are dependent on each other. In these cases, we make three separate initial predictions for each position. Each of these three initial predictions represent a rotameric structure for a different χ_1 rotameric state. For residues such as Ser, we have three initial predictions, $\{g+\}$, $\{t\}$ and $\{g-\}$. In the case of Glu, which has three χ angles, we have the lowest–energy $\{g+, *, *\}$, $\{t, *, *\}$ and $\{g-, *, *\}$ rotamers as the initial predictions. In other words, we do not specify the rotameric states of the χ_2 and χ_3 angles. We then enumerate all possible configurations; thus, if we need to predict three positions (*e.g.* Ser, Arg, Tyr), with three initial predictions each, we evaluate all 27 possible solutions. Among these combinations, we choose the set of rotameric structures that minimises the number of clashes and the total energy in the target structure.

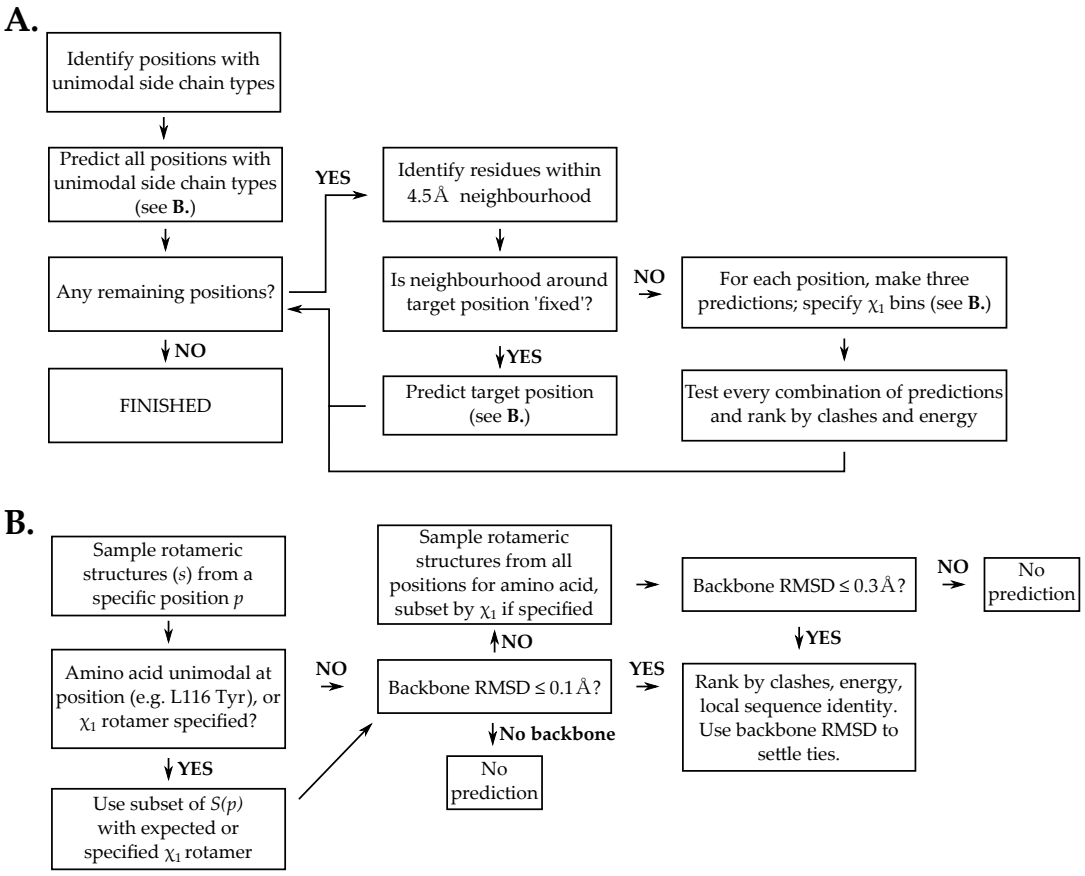


Figure 5.5: Schematic overview of PEARS in side chain prediction. **A.** We first predict the side chains of unimodal side chain types. Subsequently, for each remaining position, PEARS identifies the position's local neighbourhood defined by a 4.5Å shell. If its environment is completely fixed, *i.e.* all its surrounding residues have been predicted, the side chain at p is predicted as explained in **B.** In contrast, if ≥ 2 positions must be predicted simultaneously, all positions are given three independent predictions, one for each χ_1 rotamer. The combination of predictions are then ranked by the total number of clashes and total energy. This is repeated until all residues have been predicted on the target structure. **B.** A rotameric structure, s, is selected based on a set of filters. First, the χ_1 angle distribution is determined for the side chain type. Users can also specify the χ_1 bin for sampling. All s with backbone RMSD $\leq$0.1Å to the target structure backbone pass for further testing. If there are no rotameric structures after the first RMSD filter, rotameric structures are sampled from all positions, and the RMSD criterion is relaxed to 0.3Å. Rotameric structures are then ranked by the number of clashes, their rotameric energy, and their local tripeptide sequence identity. For target structures without a backbone, or cases where the backbone RMSD is always greater than 0.3Å, we do not make a prediction.

5.2.4.2 Selection of rotameric structures

We denote the set of rotameric structures at IMGT position p as $S(p)$. A candidate structure s from $S(p)$ is selected if its backbone RMSD to the native structure is $\leq$0.1Å. We then check for clashes between s and the rest of the target structure using a KD–tree algorithm (Section 5.2.4.3). If no rotameric structures are found from $S(p)$, the RMSD filter is relaxed to 0.3Å, and searched from all IMGT positions for the target amino acid type.

The energy of s (whose rotamer is r) at p, $E_p(s_r)$, is calculated using our implementation of the SCWRL3 energy function (Canutescu *et al.*, 2003). Finally, we calculate the sequence identity between the local tripeptide sequence of s and the local tripeptide sequence at p in the target structure.

Rotameric structures are ranked by the number of clashes, then their calculated energy, and finally by sequence identity. The rotameric structure with the lowest number of clashes and energy is grafted into the target structure. If there are any ties after applying all three filters, we use the rotameric structure with the lowest backbone RMSD.

5.2.4.3 Clash checking using KD–trees

A KD–tree (Bentley, 1975) is constructed for each input antibody structure (Figure 5.6). At each iteration, the antibody's atomic coordinates are split into two groups using the median value of a dimension. Thus, in the first iteration, we split the atomic coordinates according to the median value of the x–coordinate. In the next two iterations, we split the atomic coordinates by their y–, then z–coordinates. At the fourth iteration, we split the subsets by their x–coordinate again, and so on. This process can also be visualised as growing a binary tree, where each iteration creates two child nodes based on a separate dimension. The algorithm cycles through each dimension until a subset (*i.e.* a leaf in the tree) has $\leq L$ coordinates; we set L to 10.

For each new rotameric structure s at p, its potential nearest–neighbours are found from searching the KD–tree. The atoms of s and its nearest neighbours,

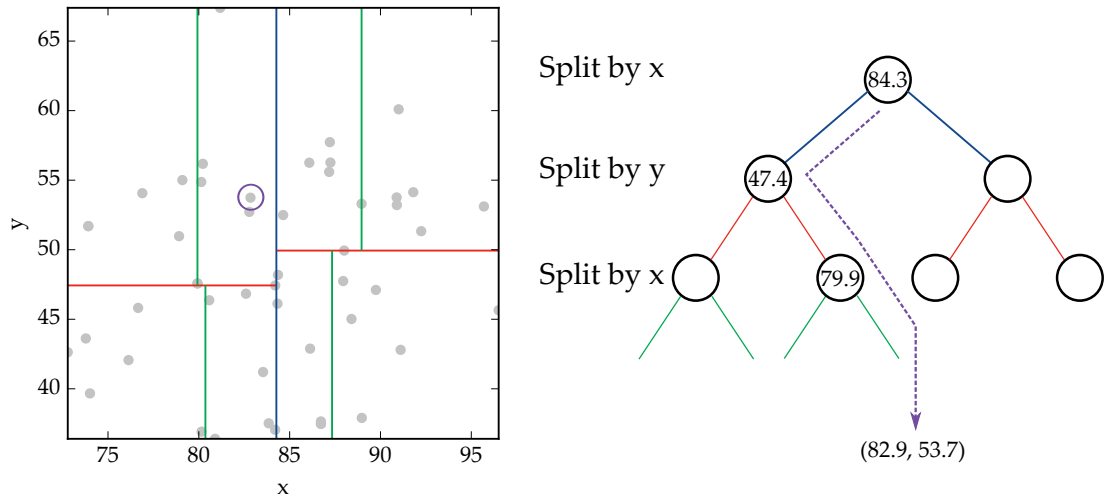


Figure 5.6: Two–dimensional example of a KD–tree. In the first iteration, we split the coordinates by the x dimension based on the median value of x (blue line). In the second iteration, we split by the median value of the y dimension (red lines) for both subsections. In the third iteration, we split again by the x–coordinate (green lines); in other words, each qudrant is split by the median value of x of each quadrant. These splits can also be visualised as a binary tree; the cycle continues until each leaf in the tree has $\leq L$ coordinates. When searching for the nearest neighbour(s) of a query (with coordinates [82.9, 53.7]), the query is searched through the binary tree.

n, are considered to clash if they are separated by less than 65% of the sum of their van der Waal's radii. The KD–tree is reconstructed after grafting a rotameric structure.

5.2.4.4 Energy function

The contact energy between atoms a and b, $E(a,b)$, is dependent on their distance, $d(a,b)$, and the sum of their van der Waal's radii, $d_{\max}(a,b)$. We used van der Waal's radii of 1.70Å, 1.55Å, 1.52Å, 1.80Å for C, N, O, and S atoms, respectively. We implemented our own version of the energy function from SCWRL3 (Canutescu *et al.*, 2003):

$$E(a,b) = \begin{cases} 0 & d(a,b) \geq d_{\max}(a,b) \\ 10 & d(a,b) \leq 0.8254\ d_{\max}(a,b) \\ 57.273\left(1 - \dfrac{d(a,b)}{d_{\max}(a,b)}\right) & \text{otherwise.} \end{cases} \tag{5.3}$$

$E_p(s_r)$ represents the energy of the rotameric structure s with rotamer r placed on IMGT position p. It is calculated as the sum of contacts, along with a probability term, $Pr(r|p)$. The probability term, $Pr(r|p)$, represents the probability of finding a rotamer r at p; this is scaled relative to the most frequent rotamer at p. If r is unobserved at p, the default value is set to 0.001. For calculating the probabilities, rotameric structures from the target structure were removed from PEARL. For example, there are 62 Val $\{t\}$ rotamers in PEARL, one of which is found in 2cmr:HL. When predicting H21 Val on 2cmr:HL, we removed the native rotamer from PEARL prior to calculating the probability of the Val $\{t\}$ rotamer at H21.

Let $BB(p)$ and $SC(p)$ represent the set of backbone and side chain atoms at p, respectively. Given two positions in the antibody p and p',

$$
\begin{aligned}
E_p(s_r) = -Z\frac{Pr(r|p)}{\max_{r\in R(p)} Pr(r|p)} \quad &+ \quad E_{p,p'}(SC,BB) + E_{p,p'}(SC,SC) \qquad (5.4)\\
E_{p,p'}(SC,BB) \quad &= \quad \sum_{|p-p'|>1}^{P} \sum_{a\in SC(p)} \sum_{b\in BB(p')} E(a,b)\\
E_{p,p'}(SC,SC) \quad &= \quad \sum_{p\neq p'}^{P} \sum_{a\in SC(p)} \sum_{b\in SC(p')} E(a,b)
\end{aligned}
$$

Here, Z is a constant, which is set to 3 (Canutescu *et al.*, 2003). The side chain–backbone contact energy, $E_{p,p'}(SC,BB)$, is only calculated between the side chain atoms of p and the backbone atoms of p' if p and p' are separated by at least one residue. The side chain–side chain contact energy, $E_{p,p'}(SC,SC)$, is calculated for every pair of positions.

5.2.4.5 Local tripeptide sequence identity

For each IMGT position p in the target structure, we obtain its local tripeptide sequence, which is the amino acid sequence at the $p-1$, p, and $p+1$ positions. This is then compared to the local tripeptide sequence of s (Section 5.2.2). This is only used if there are two or more rotameric structures that have the same number of clashes and the same energy.

5.2.5 Benchmarking accuracy

PEARS was benchmarked against three other side chain prediction methods, using two test sets (Section 5.2.1). χ_1 and χ_{1+2} accuracies were calculated using the criteria in Section 5.2.3. For PEARS, self–predictions were not allowed.

5.2.5.1 SCWRL

SCWRL4 (Krivov *et al.*, 2009) was used with default parameters. SCWRL samples from a backbone–dependent library (Shapovalov and Dunbrack, 2011), using an energy function that combines van der Waal's and hydrogen bond terms. The optimal solution is determined by DEE and graph decomposition (Krivov *et al.*, 2009).

5.2.5.2 RASP

RASP uses the backbone–dependent library from Dunbrack and Cohen (1997). The method uses a variant of the OPUS–PSP energy function, including a hydrogen bonding term. RASP uses a combination of DEE and a Monte Carlo search algorithm to determine the best solution (Miao *et al.*, 2011). RASP did not predict side chains for the crystal structures of 1dzb:AA, 1yc7:B, and 5d6c:HL as one residue in each of these structures did not have a complete backbone.

5.2.5.3 OSCAR–star

OSCAR–star uses the same rotamer library as RASP. The energy of a rotamer is determined by a series of distance and orientation–dependent terms. The best solution is determined using a Monte Carlo simulated annealing algorithm (Liang *et al.*, 2011).

5.3 Results

5.3.1 An antibody–specific rotamer library

PEARL was built using 640 non–redundant antibody structures downloaded from SAbDab (Dunbar *et al.*, 2014). This is a smaller dataset than other

rotamer libraries (*e.g.* Shapovalov and Dunbrack, 2011, 1179 chains *vs.* 3974 chains). Comparing PEARL to Dunbrack RL shows that antibodies have unique χ angle distributions compared to general non–antibody proteins. A two–sample Kolmogorov–Smirnov test between the backbone–independent χ_1 and χ_2 angle distributions of PEARL and Dunbrack RL showed significant differences ($p < 0.001$; Figure 5.7).

5.3.2 Position–dependent χ_1 distributions

PEARL groups rotamers by their structural position, rather than their backbone ϕ/ψ angle profile. For each side chain type (*i.e.* combination of IMGT position and amino acid type, *e.g.* H24 Lys), we calculate the probability of each rotamer type. PEARL treats variations in the χ angle as a backbone–independent property, making it particularly useful for scenarios where the backbone dihedrals may not be exact (*e.g.* predicting side chains on model structures).

5.3.2.1 Certain positions show limited range of χ_1 angles

For 255 side chain types, the χ_1 angle shows a unimodal distribution (Figure 5.8A). Generally, these unimodal side chains are found in the framework, and are associated to known conserved amino acids (Igawa *et al.*, 2010). For instance, the highly–conserved Gln residues at H44 and L44 (Chapter 3) show a unimodal χ_1 profile, adopting the $\chi_1 = t$ configuration. This phenomenon is also observed in the CDR loops. For example, L116 Tyr residues (CDRL3 loop) show a unimodal χ_1 angle distribution, with a peak correponding to the $g-$ rotamer (Figure 5.8B). PEARL captures the variation of χ_1 angles in a backbone–independent manner, which is partiularly useful for CDR loop positions as the backbone ϕ/ψ angles can vary extensively (Figure 5.8B).

A large number of side chain types do not have a defined χ_1 mode; in total, 4482 side chain types' χ_1 angle modes are undefined (Figure 5.8A). One reason for the lack of a mode annotation is the insufficient number of observations. We used a threshold of 20 observations to calculate the χ_1 angle mode, which

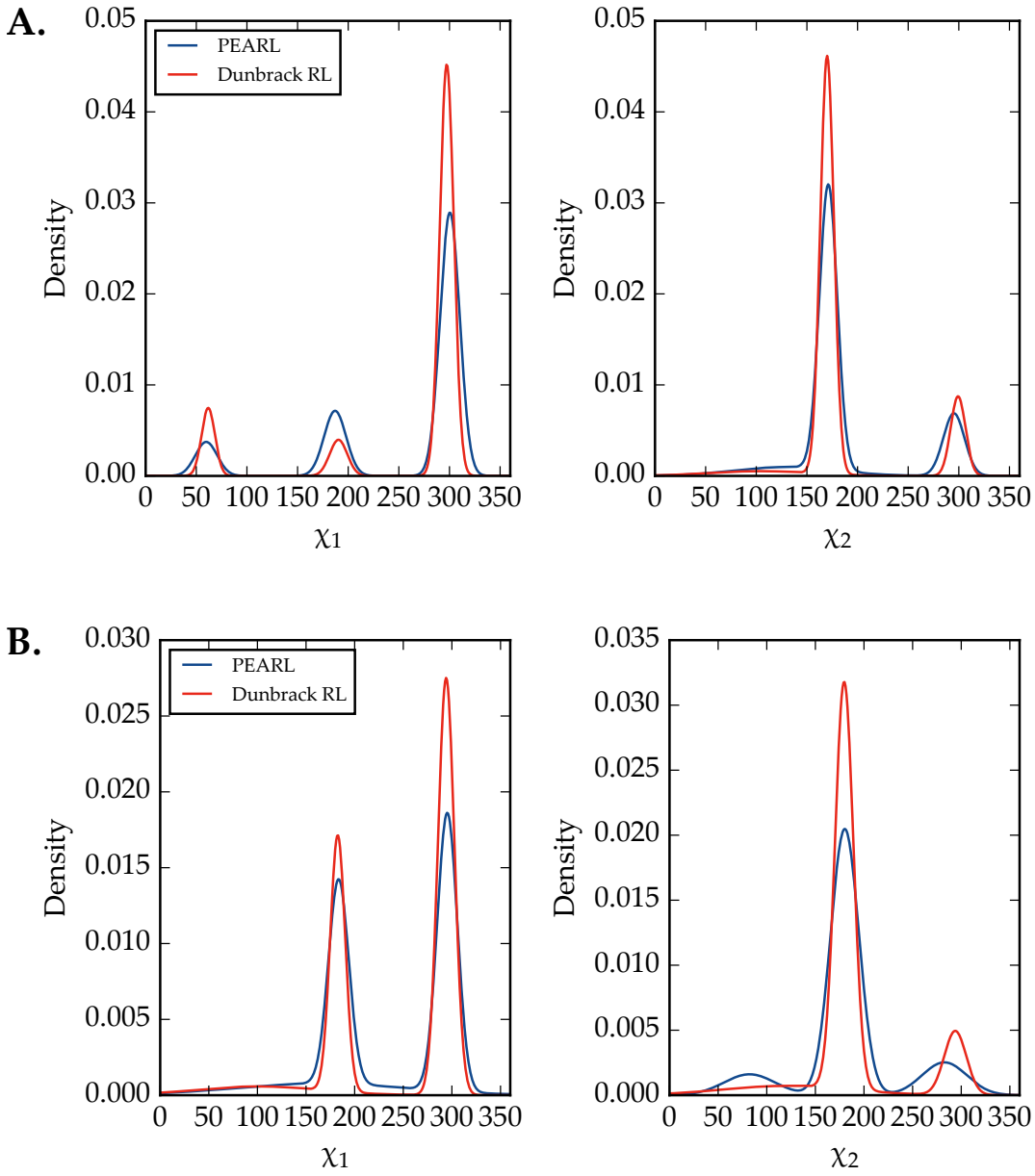


Figure 5.7: Backbone–independent distribution of χ_1 and χ_2 angles in Ile (**A.**) and Lys (**B.**). The Kolmogorov–Smirnov test for both χ angles for both amino acids indicated significant differences ($p < 10^{-6}$) between distributions.

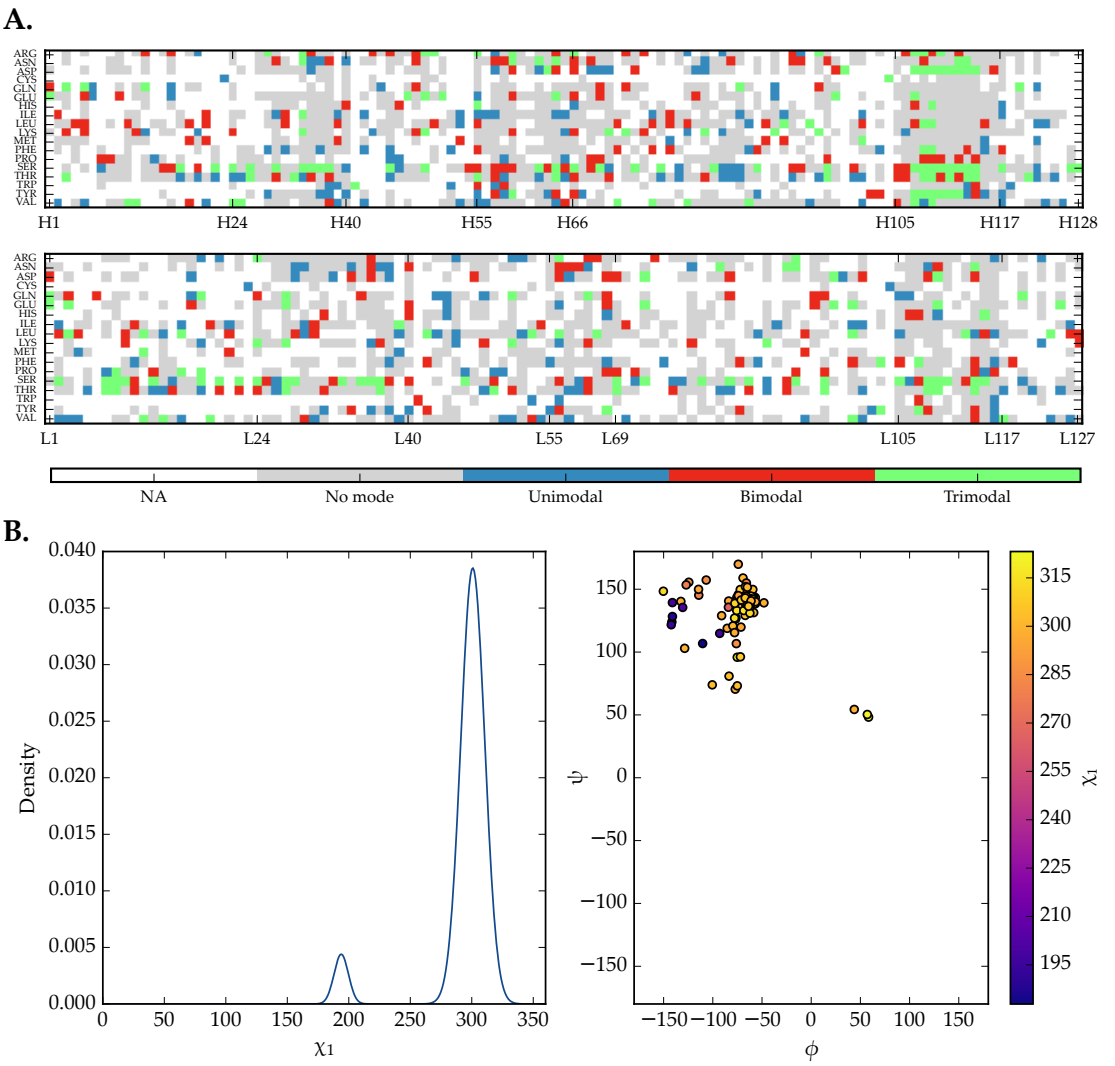


Figure 5.8: Rotamer preferences of amino acids are IMGT position–dependent. **A.** For each residue type at a given position, we determined the modality of the χ_1 angle distribution using a GMM. Amino acids at certain positions are coloured grey if they have <20 data points ('no mode'; *e.g.* L55 Arg), and white if they are never observed (*e.g.* L69 Val). For simplicity, positions with insertions (*e.g.* H111A) are not visualised; in terms of prediction, these positions are treated in the same manner as any other IMGT position. **B.** *Left:* The distribution of the χ_1 angle for L116 Tyr shows a peak at ∼300°, and a small peak around 180°. *Right:* The position–dependent library captures the χ_1 distribution of L116 Tyr, despite the position's wide range of ϕ/ψ angles. Points are coloured according to their χ_1 angle.

is a more relaxed, but similar, threshold set by Shapovalov and Dunbrack (2011). Based on this criterion, 1664 side chain types' χ_1 angle modes were undefined as they had 1–19 observations (coloured grey, Figure 5.8A). The remaining 2818 possible side chain types' χ_1 angles modes were not defined as they were never observed. This is likely due to the sequence conservation of antibodies. For instance, H23 and H104 are always Cys residues that form disulphide bridges. Therefore, we did not expect any other side chain type at these positions. Likewise, other positions such as H98 and L98 always feature an Asp. Furthermore, some positions are CDR positions that are only found in a small number of antibodies. For example, H32 is only found in two antibodies of our non–redundant set.

5.3.3 PEARL provides sufficient coverage

We tested the ability of PEARL to provide the correct rotamer for each side chain, similar to Colbes *et al.* (2016). Briefly, for each structure in the crystal test set, we find any rotameric structure in PEARL that would be a 'correct' prediction for the χ_1 and χ_2 angles. Again, we disallowed self–predictions. Effectively, this avoids prediction errors that can be attributed to the algorithm, *e.g.* the energy function. We then calculated the average χ_1 and χ_{1+2} accuracies to represent the maximum achievable accuracy.

Based on a 40° cutoff, PEARL achieved maximum achievable average χ_1 and χ_{1+2} accuracies of 99.97% and 99.92%, respectively. In other words, these accuracy values indicate that a perfect side chain prediction algorithm will achieve a χ_1 accuracy of 99.97%, on average. Furthermore, the χ_{1+2} accuracy will be 99.92% in this scenario. The corresponding maximum achievable χ_1 and χ_{1+2} accuracies using Dunbrack RL are 99.8 and 91.0%. Therefore, PEARL offers a slightly higher coverage, despite having a smaller dataset.

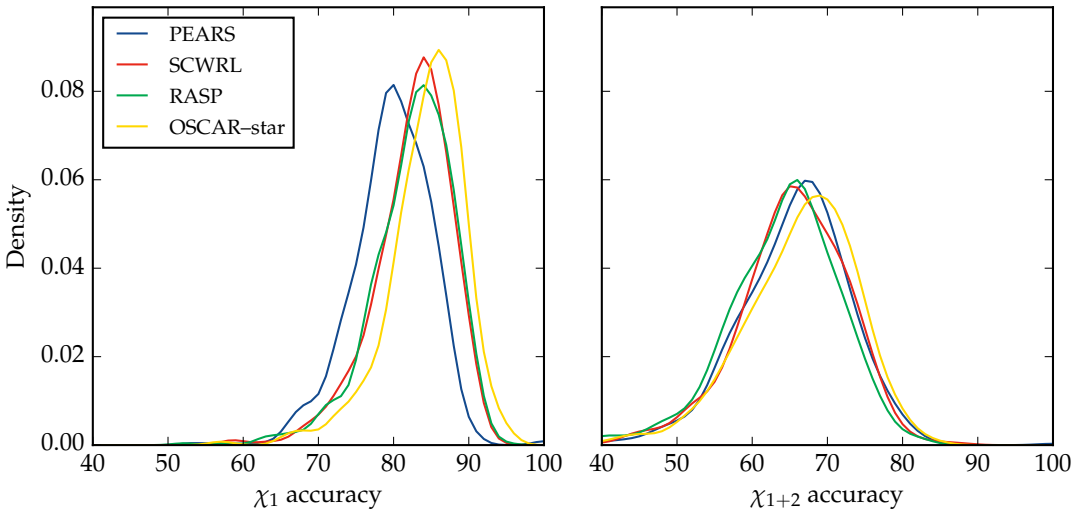


Figure 5.9: Density plots of χ_1 and χ_{1+2} accuracies from predicting side chains of the crystal test set. The χ_1 and χ_{1+2} accuracies were calculated using a 40° cutoff.

Table 5.2: Average χ_1 and χ_{1+2} accuracies for crystal and model test sets.

Method	Crystal test set		Model test set	
	χ_1	χ_{1+2}	χ_1	χ_{1+2}
SCWRL	82.5%	65.4%	73.2%	55.5%
RASP	82.5%	64.2%	72.9%	54.2%
OSCAR–star	**84.3%**	**66.4%**	73.8%	54.9%
PEARS	79.7%	65.6%	**77.1%**	**62.5%**

5.3.4 Crystal structures refined by PEARS

Currently, no study has benchmarked side chain prediction methods on a large set of antibody crystal structures. Thus, we initially predicted all side chains in the crystal test set. For each query structure, self–predictions were disallowed for PEARS. Among the four methods that we tested, OSCAR–star showed the strongest performance, with an average χ_1 accuracy of 84.3% and χ_{1+2} accuracy of 66.4%. Although PEARS had the poorest χ_1 accuracy (Figure 5.9), its χ_{1+2} accuracy was on par with other methods (Table 5.2).

We then examined the side chain types that PEARS predicted poorly (Fig-

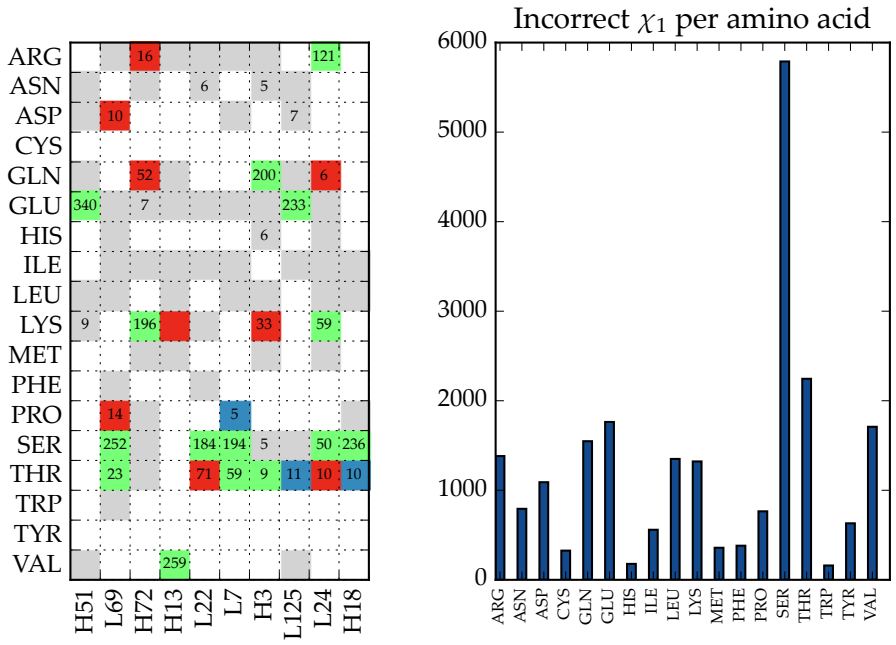


Figure 5.10: Sources of PEARS' errors in the crystal test set. **A.** For ten of the most poorly predicted positions by PEARS, we investigated their χ_1 mode. Each cell is coloured according to the same colour scheme in Figure 5.8A. The numbers at the centre of each cell indicate the number of incorrect predictions for the amino acid type at the position (only if there are $\geq$5 mistakes). **B.** Number of incorrect χ_1 rotamers per amino acid type; Ser and Thr residues are the most difficult to predict.

ure 5.10A). Among the ten IMGT positions with the largest number of incorrect χ_1 predictions, the majority were associated to side chain types with a trimodal χ_1 angle distribution. For instance, H51 Glu has a trimodal χ_1 distribution; out of 579 cases, 340 were predicted incorrectly. Likewise, Ser residues at L69, L22, L7, and H18 show trimodal χ_1 distributions, and in all cases, PEARS was unable to predict the χ_1 angle. In comparison, the number of incorrect predictions for unimodal side chain types, *e.g.* L125 Thr, was far lower.

Overall, Ser and Thr residues were the most challenging for PEARS (Figure 5.10B). Again, this seems to be associated to the multimodal χ_1 angle distributions of both amino acids. For example, there are 108 IMGT positions with $\geq$20 observations of Ser residues. Sixty–three of these are trimodal; however, only 18 positions with Ser have a unimodal χ_1 angle. Among 81 IMGT positions with $\geq$20 observations of Thr residues, 36 are unimodal, whereas

45 are multimodal.

OSCAR–star's worst predictions were observed for different side chain types (Figure 5.11A). In particular, OSCAR–star was unable to predict the χ_1 angle of 327 Pro residues at L50. This is a unimodal position which PEARS predicted with greater accuracy, making only 12 incorrect predictions. Likewise, while OSCAR–star could not predict the χ_1 angle in 55 cases with L51 Arg, PEARS only made 13 mistakes for the same side chain type. This shows the benefits of a position–specific approach, especially when the side chain has a unimodal χ_1 distribution.

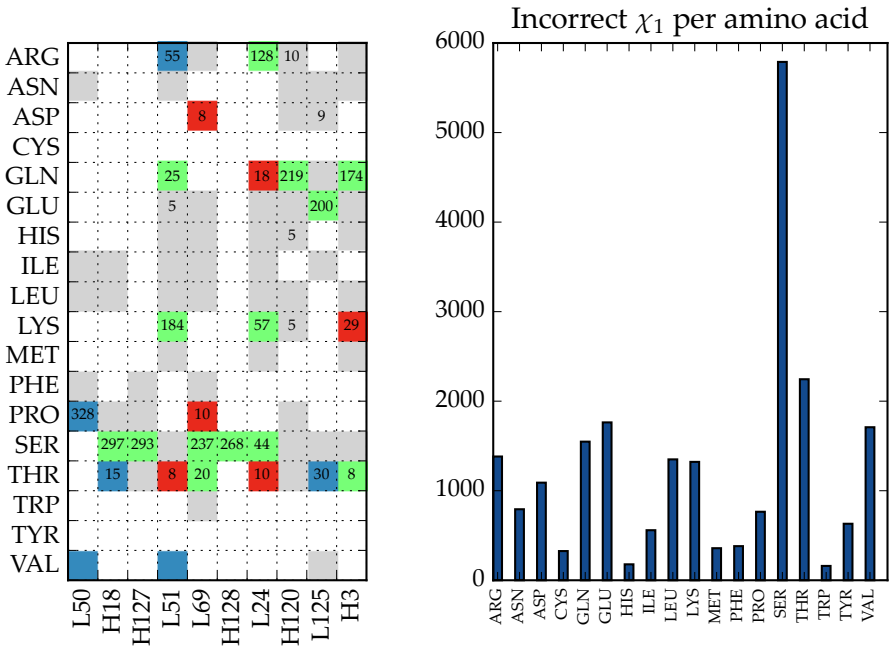


Figure 5.11: OSCAR–star's errors in the crystal test set. **A.** For ten of the most poorly predicted positions by OSCAR–star, we investigated the sources of error; cells are coloured and labelled in a similar scheme to Figure 5.10. **B.** Number of incorrect χ_1 rotamers per amino acid type; Ser and Thr residues are, again, the most difficult to predict.

There was some overlap between 'difficult' side chain types for both PEARS and OSCAR–star; for instance, H18 Ser and L24 Arg. Again, both side chain types have trimodal χ_1 distributions. For both scenarios, it is possible that two different rotamers can have similar $E_p(s_r)$ values, as the probability term will not strongly favour one rotamer at the target position. Furthermore, the energy

function does not seem to discriminate correct rotamers properly; improving its accuracy, *e.g.* by adding a hydrogen bond term, and/or an orientation–dependent term (Miao *et al.*, 2011), may enhance prediction quality at these difficult positions.

5.3.5 ABodyBuilder models' side chains predicted by PEARS

We next tested side chain prediction on the model test set. Predicting a model structure's side chains is a more stringent test, as the actual backbone geometry is unknown. However, we believe that this is a more relevant test, especially in the context of antibody design.

In contrast to our results on crystal structures, PEARS outperformed all other methods in predicting the side chains of model structures (Figure 5.12). In particular, the χ_{1+2} accuracy was significantly higher. Our results demonstrate the value of using antibody–specific information in predicting side chains on model structures. Earlier, we showed that the χ angle of specific side chain types is agnostic to the backbone dihedral angles (Figure 5.8B). Since PEARS does not explicitly rely on the backbone dihedral angles for sampling, it is more robust to modelling inaccuracies.

In order to test the impact of model prediction quality on PEARS' performance, we separated models based on their Fv backbone RMSD to the native structure. Regardless of prediction quality, both PEARS and OSCAR–star correctly predicted a large number of rotamers. However, PEARS slightly outperformed OSCAR–star, predicting more unique rotamers across model structures, especially for medium–quality (Fv backbone RMSD 1.0–2.5Å) models (Figure 5.13).

We also divided models according to their CDRH3 backbone RMSD, as the CDRH3 is often the most poorly modelled region (Almagro *et al.*, 2014), and is usually a target for mutations (*e.g.* Barderas *et al.*, 2008). Again, both PEARS and OSCAR–star correctly predicted a common set of rotamers across all model CDRH3 loops. For 'good' (CDRH3 backbone RMSD $\leq$1.0Å) models, both PEARS

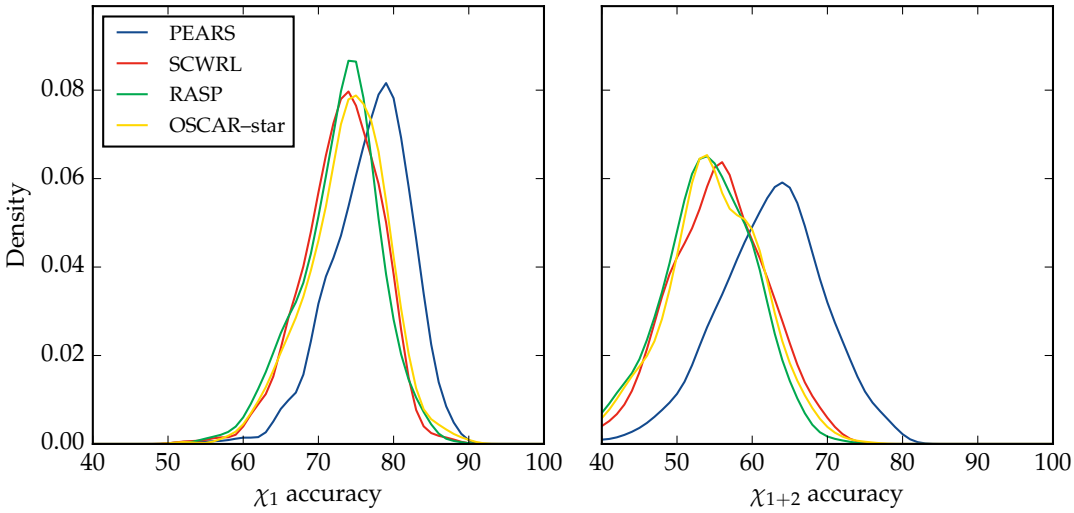


Figure 5.12: Density plots of χ_1 and χ_{1+2} accuracies in predicting the side chains of the model test set.

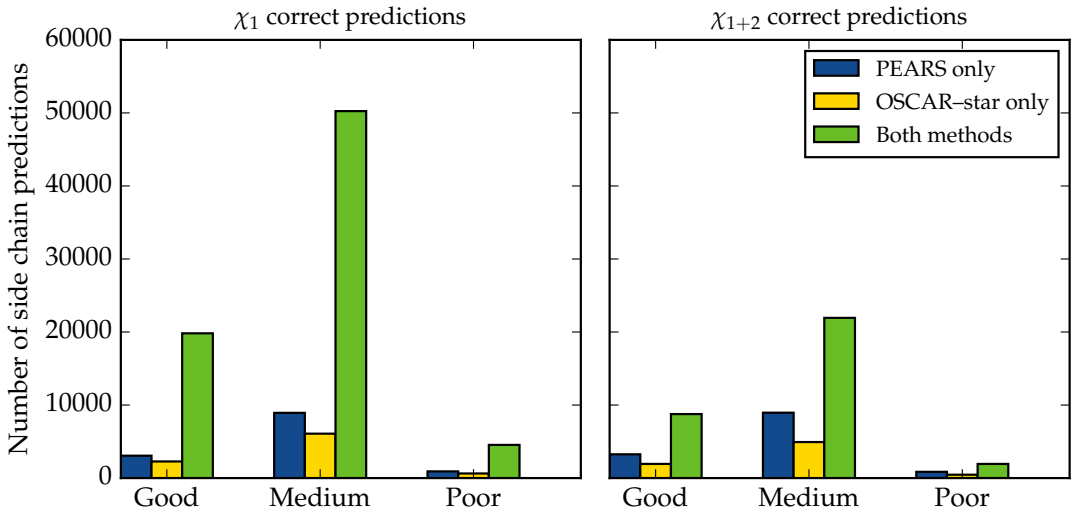


Figure 5.13: Fv model quality does not affect side chain prediction by PEARS. Each model was characterised by its Fv backbone RMSD: good (RMSD $\leq$1.0Å), medium (RMSD 1.0–2.5Å), and poor (RMSD >2.5Å). The bars show the number of rotamers that were only predicted by PEARS, OSCAR–star, or by both methods.

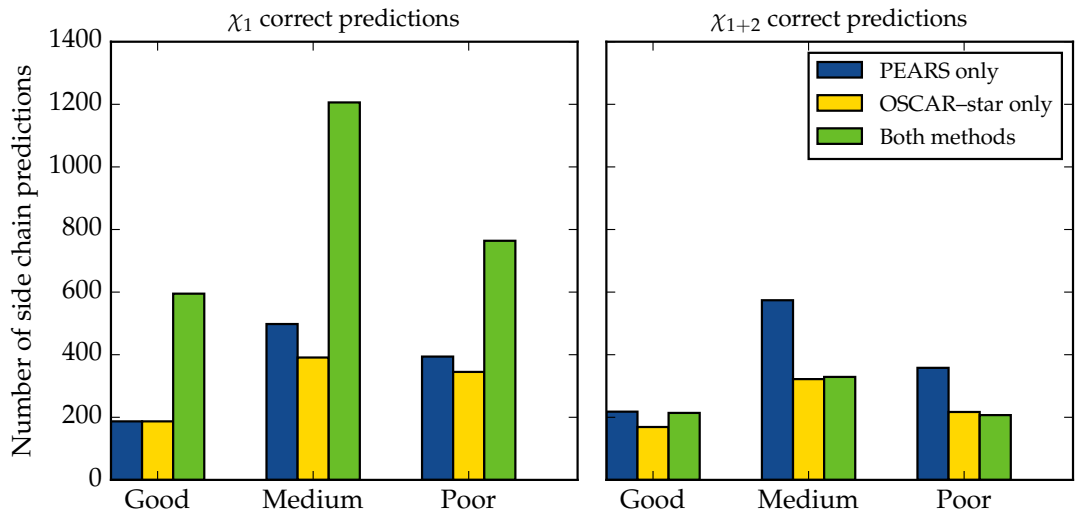


Figure 5.14: CDRH3 loop model quality does not affect side chain prediction by PEARS. Each model was characterised by its CDRH3 loop backbone RMSD, similar to the criteria in Figure 5.13.

and OSCAR–star predicted a similar number of correct χ_1 rotamers, suggesting that with an accurate set of backbone dihedral angles, the accuracies of the two methods converge. However, as the model quality deteriorates, especially for bad (CDRH3 backbone RMSD >2.5Å) models, PEARS had a much higher number of χ_1 and χ_{1+2} correct rotamers.

Both results highlight the advantage of our method. Since the predictions are based purely on the observed rotameric preferences at specific IMGT positions, rather than the accuracy of the backbone dihedral angles, it is robust to errors arising from modelling. Despite sampling rotamers only based on their χ_1 rotameric state, we nonetheless recover more χ_{1+2} correct rotamers, which will be useful for design applications.

5.4 Discussion

In this Chapter, we presented our position–dependent rotamer library, PEARL. We also developed PEARS, which uses the data in PEARL for a position–driven method of side chain prediction. To our knowledge, PEARS is the

first antibody–specific rotamer library; furthermore, it is the first that categorises rotamers according to their IMGT positions. Most rotamer libraries are based on the distribution of rotamers in general proteins (*e.g.* Lovell *et al.*, 2000; Shapovalov and Dunbrack, 2011). Although protein–specific libraries have been built before (*e.g.* Bhuyan and Gao, 2011), these libraries are query–specific, as opposed to family–specific, *e.g.* antibodies.

Antibodies have a unique distribution of χ angles in comparison to general proteins. Calculating the backbone–independent probabilities for our dataset showed clear differences against rotamers from general non–antibody proteins, suggesting that antibodies use a different set of rotamers.

For antibodies, a position–specific rotamer library is sensible as the IMGT positions are well–defined from large sequence alignments (Lefranc *et al.*, 2003). The principle behind the IMGT numbering is that a position should consistently refer to the same location in different antibodies. Extending from this assumption, we found that ∼10% of side chain types in PEARL (255/2384) have a 'unimodal' χ_1 angle distribution. Unimodal side chain types are found in both framework and CDR loop positions, though most side chains in the CDR loops tend to be multimodal.

PEARL's χ_1 modes are based on the weights of GMMs; although angle data is circular, we preferred using GMMs over other kernels (*e.g.* Harder *et al.*, 2010; Shapovalov and Dunbrack, 2011, both have used von Mises kernels) for two reasons. At high values of the concentration parameter ($\kappa > 10$), von Mises distributions converge to Gaussian distributions (Mardia *et al.*, 2007). In fact, Dunbrack RL used very high concentration parameters; the initial κ value for every amino acid was >35, suggesting that some bins' densities can be approximated by Gaussian kernels. Furthermore, since most χ angles form distinct peaks (*e.g.* 60°, 180°, and 300° for χ_1), a GMM would better capture variations centred around these χ angle peaks.

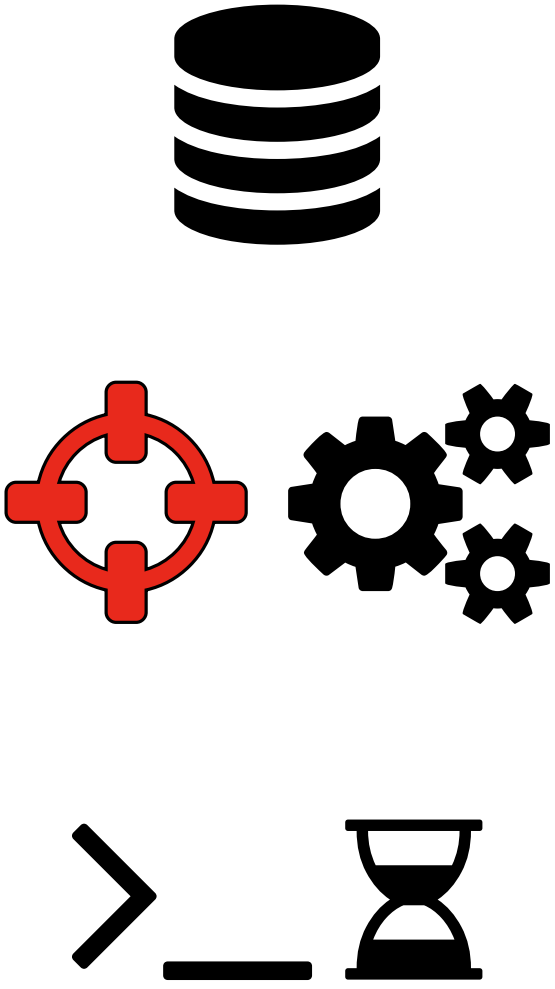
When predicting the rotamers on crystal structures, OSCAR–star had the highest χ_1 and χ_{1+2} accuracy. In fact, PEARS had the worst performance,

suggesting that PEARS' algorithm needs further development to select the correct rotameric structure. Several aspects can be addressed; for instance, the current energy function is not perfect; in some cases, the lowest energy rotameric structure was not the best prediction. Similarly, when predicting multiple positions at once, the configuration with the lowest energy was not always the best configuration. A potential issue is the fact that the current energy function does not incorporate a hydrogen bonding term (Canutescu *et al.*, 2003). Thus, we are aiming to develop our own antibody–specific energy function for selecting rotamers; for instance, RAPDF has previously been used to predict side chains (Samudrala and Moult, 1998b). In addition, considering contact orientation (*e.g.* Miao *et al.*, 2011) may further enhance scoring. A more important step in improving prediction quality would lie in developing structure–based rules that select the correct rotamer. For instance, searching for chemically–similar environments, as in the case of MoFvAb, may provide clues on how difficult side chains should be predicted, given its local protein environment (Bujotzek *et al.*, 2015a).

More importantly, a more rapid and accurate search method is critical. As a heuristic, when we need to predict $\geq$2 positions, we only make three initial predictions per position, corresponding to a different χ_1 rotameric state. However, this simplification can be highly problematic when we predict residues with multiple χ angles. For example, Arg can have 81 different rotamers, but we only choose the three lowest–energy rotamers based on the χ_1 rotameric state. In other words, we choose the lowest–energy {g+,*,*,*}, {t,*,*,*} and {g–,*,*,*} rotamers. Thus, the method risks precluding the correct rotamer from ever being tested. Furthermore, we currently test every possible configuration when we predict multiple positions. Thus, if five positions are found within a 4.5Å neighbourhood, we test all 3^5 combinations. An implementation of DEE (Desmet *et al.*, 1992), or a graph–based method (*e.g.* Krivov *et al.*, 2009; Miao *et al.*, 2011) would improve the speed of PEARS and allow it to efficiently test more candidate rotamers at each position.

PEARS had its biggest impact in predicting the side chains on the model test set. The advantage of a position–dependent approach is its robustness against uncertainties of the protein backbone. In particular, for medium and poor quality models, PEARS was able to provide more correct rotamers in the model structure. In fact, OSCAR–star and PEARS had similar accuracies for 'good' CDRH3 models, suggesting that OSCAR–star is more effective in cases where the backbone is accurate. In particular, given the current speed advantage of OSCAR–star over PEARS, it may be feasible to use both methods during structure prediction. In the previous Chapter, we showed that ABodyBuilder can estimate model accuracy. Thus, a possible implementation is to use backbone–dependent side chain prediction methods (*e.g.* OSCAR–star) for antibody regions that are expected to be modelled well, and PEARS elsewhere.

Overall, the results presented in this Chapter demonstrate the value of antibody–specific information in side chain prediction. By harnessing the unusual χ_1 angle distributions in antibody side chains, we can restrict the sampling space and provide high–quality predictions, especially for models. We also demonstrated that PEARL provides excellent coverage. The maximum achievable χ_1 and χ_{1+2} accuracies were near 100%; improving PEARS is therefore an algorithmic challenge. Ultimately, we envision that PEARS will be used as a component in ABodyBuilder, and provide a method for testing the effects of mutations from computational antibody design. In the next Chapter, we describe strategies and future directions for the work presented in the Thesis.



. . .the goal he set out initially to become the greatest of all time was a very fickle one. When [the young basketball player] realised that the most important thing in life is how your career moves and touches those around you, and how it carries forward to the next generation, then he realised that's what makes true greatness.

— Kobe Bryant

6 Conclusion

Contents

Antibodies are a key component of the adaptive immune response, and are a major class of biotherapeutics. Designing them in a computational framework is ideal, as it allows rapid and efficient testing of candidates. In this thesis, the structural variation of antibodies has been investigated to explain biological phenomena. These observations were then used to rationalise decisions for developing computational tools for antibody design. First, we proposed a knowledge–based method for predicting antibody–antigen binding affinities. We then focussed on elucidating a possible mechanism for V_H–V_L pairing as it is a major source of structural variation. Finally, we developed an antibody modelling pipeline, and a side chain prediction method to construct structural models for further computational or experimental tests.

6.1 Binding affinity prediction remains a challenge

In Chapter 2, we described an antibody–specific statistical potential, CAPTAIN. This is a weighted form of the RAPDF potential (Samudrala and Moult, 1998a). We demonstrate that weighting interatomic contacts by affinity data improves prediction performance. Although CAPTAIN showed the highest correlation to binding affinities in comparison to 16 other scoring functions, its performance is currently too weak for large–scale prediction. In particular, our function was unable to predict the binding affinities of peptide–binding antibodies, suggesting that the current training method is not suitable for peptide–binding antibodies.

Currently, CAPTAIN counts contacts across 207 different atom types over eight distance bins. In total, there are 15407 contact types; for most of these contacts, the data is sparse, leading to a wide distribution of scores. Although we attempted to saturate the dataset by using contacts from a set of general protein–protein complexes (Studer *et al.*, 2014), our scoring function showed weaker performance, confirming the importance of antibody–specific information. Thus, we propose that developing a set of antibody–antigen docked poses can generate a saturated dataset that can approximate and complement the observed distribution of contacts in SAbDab.

In its current stage, our method assumes that contacts are the sole determinant of an antibody's binding affinity, and disregards other structural features, such as loop flexibility (Haidar *et al.*, 2014). These variations must then be analysed in the context of an antibody's binding affinity. Further investigation on peptide–binding antibodies is also necessary to determine if they have a unique binding mechanism compared to protein–binding antibodies. Once these variations are identified, it may be possible to develop new scoring functions that use a linear combination of terms, *e.g.* contacts and orientation angles. Accurate affinity prediction still remains a topic of interest, not only for computational antibody design, but also for the general community in studying protein–protein interactions.

6.2 V_H–V_L pairing is random

The findings from Chapter 3 aim to clarify the ongoing debate surrounding the precise mechanism of V_H–V_L pairing. Understanding V_H–V_L pairing is not only important for rationalising structural variation in antibodies, but also for designing thermostable and/or bispecific antibodies (Jayaram *et al.*, 2012; Lewis *et al.*, 2014). Although there is an accepted paradigm that pairing is 'random' (*e.g.* Brezinschek *et al.*, 1998; DeKosky *et al.*, 2015), no study has extensively viewed V_H–V_L pairing from a structural context. In contrast to previous studies, we defined 'random' pairing as the ability for any V_H to pair with any V_L, and vice–versa. Both our sequence and structural data point toward the possibility that any V_H can pair with any V_L.

Similar to previous studies, we first examined V_H–V_L pairing from the level of IMGT *V* gene pairings. Despite the apparent dependence between the V_H and V_L subgroups, only a limited number of pairs showed significant enrichment. Furthermore, our pairings were sparse and highly skewed; thus, using subgroup information alone could not verify the 'randomness' in pairing. However, pairwise sequence identities showed that a single V_H domain can have >10 potential V_L partners, and vice–versa. Furthermore, these partners can range in sequence identities from ∼40 to ∼100%, supporting previous observations that V_H and V_L domains can pair with a diverse set of partners (Edwards *et al.*, 2003).

We then examined a non–redundant set of structures from SAbDab, featuring human, mouse, and rabbit structures. Here, we report a series of contacts that is ubiquitous in all structures, and each of these contacts almost always feature the same pair of amino acids from the V_H and V_L. For instance, the H118–L50 contact is almost always a Trp–Pro contact in all antibodies. The high degree of conservation suggests that these residues form a generic set of contacts at the base of the Fv, and other contacts, *e.g.* between CDR loops, may provide additional stability or destabilise the structure. This structural conservation was also observed in the pre–BCR. Given that a single SLC (or two, *e.g.* in mouse)

screens all V_H domains for their ability to form pairs, this provides further support for a generic mechanism that allows the V_H and V_L to pair together.

From the perspective of computational antibody design, pairing information could be used to quantify antibody stability. We have demonstrated some predictive ability on a small dataset of antibodies (Teplyakov *et al.*, 2016), though a larger–scale analysis is necessary. Given the random behaviour of pairing, it may be possible to design antibodies using many different combinations of V_H and V_L to target a specific antigen. There are several unanswered questions surrounding pairing, such as CDRH3 loop compatibility. It may be possible that different V_H–V_L pairs form unique microenvironments that allow CDR loops to form a range of conformations. Thus, a pairing–dependent clustering of CDR loops is a potential avenue for future research.

6.3 Antibody–specific data improves modelling

In Chapters 4 and 5, we described two methodologies for structure prediction: ABodyBuilder and PEARS. ABodyBuilder is a rapid, automated tool that models the structure of Fv's and nanobodies. PEARS is a side chain prediction algorithm that utilises the antibody–specific distribution of rotamers for its prediction.

ABodyBuilder follows a canonical four–stage workflow, similar to other antibody modelling pipelines (*e.g.* Sivasubramanian *et al.*, 2009; Marcatili *et al.*, 2014). Models are typically built in $\sim$30 seconds, making it scalable for larger datasets, *e.g.* our database of $>$6000 paired antibody sequences from Chapter 3. In addition, 'liability' sites, *e.g.* glycosylation sites, may lead to issues during development, and these are flagged on the model structure. ABodyBuilder's models are comparable to other pipelines in terms of accuracy, demonstrated by our benchmark on the AMA–II set of antibodies. However, the advantage of ABodyBuilder lies in its ability to estimate a model's accuracy. It warns users that specific regions of the model, *e.g.* the CDRL3 loop, are expected to be poorly modelled.

The accuracy metric represents the probability that a region is modelled within a specific RMSD threshold. These probabilities are based on the wealth of pairwise framework region superimposition data, or the performance of FREAD on individual CDR loops. Currently, an Fv structure has eight separate accuracy estimates: two for the framework regions, and six for the CDR loops. As the error in V_H–V_L orientation is difficult to quantify (Abhinandan and Martin, 2010; Bujotzek *et al.*, 2015b), we have not provided an explicit accuracy measure for describing the orientation. To further improve ABodyBuilder, a dedicated V_H–V_L orientation prediction method would be ideal (*e.g.* Bujotzek *et al.*, 2015b; Marze *et al.*, 2016); currently, we assume that the template's orientation sufficiently recaptures the orientation of the native structure. An additional route for improving ABodyBuilder is an antibody–specific side chain prediction method; existing methods (*e.g.* Krivov *et al.*, 2009) use rotamer libraries based on general proteins.

We explored the latter by developing our rotamer library, PEARL, and our side chain prediction algorithm, PEARS. In PEARL, we present an antibody–specific, position–specific library, which is the first of its kind. No rotamer library, to our knowledge, describes the distribution of rotamers for a specific family of proteins, such as antibodies. Although position–dependence has previously been explored (Chinea *et al.*, 1995), our method describes the rotamers for absolute positions in the antibody structure. In particular, we demonstrate that certain amino acid types have a unimodal χ_1 angle distribution at various IMGT positions. For these side chain types, prediction can be simplified by using the most common rotamer.

PEARS was used to predict the side chains of crystal and model structures. For crystal structures, backbone–dependent side chain prediction methods (*e.g.* Krivov *et al.*, 2009) were superior, whereas PEARS was more useful for model structures. Given that PEARL is inherently backbone–independent, our side chain predictions are more robust to inaccuracies in the backbone. At present, PEARS is at an early stage, and needs more algorithmic improvements. The

maximum accuracy test suggests that there is almost always an accurate (within 40°) rotamer in PEARL; thus, rotamer scoring must be improved to select the best possible rotamer. Furthermore, PEARS is much slower than most other side chain prediction protocols, especially when predicting >3 positions simultaneously. This suggests that a search method (*e.g.* DEE Desmet *et al.*, 1992) is necessary to eliminate 'poor' solutions from the search.

6.4 Future research avenues

6.4.1 NGS: the cornerstone of future analysis

NGS has fostered a new movement in antibody informatics, bringing an unprecedented volume of data for analysis and prediction. Throughout this thesis, we have alluded the significance of our results and the scalability of our prediction under the context of having an NGS dataset. As millions of sequences will be profiled by NGS, a major challenge will be translating these sequences into structural models. This will ultimately uncover the structural landscape of an organism's 'antibodyome', and also help discover antibodies, *e.g.* antibodies unique to HIV–resistant humans. There is a huge amount of potential in these datasets, which structural modelling (*e.g.* by ABodyBuilder) will complement and provide new ideas for future research. Other aspects of structural variation, such as V_H–V_L pairing, will be further clarified by the availability of such data.

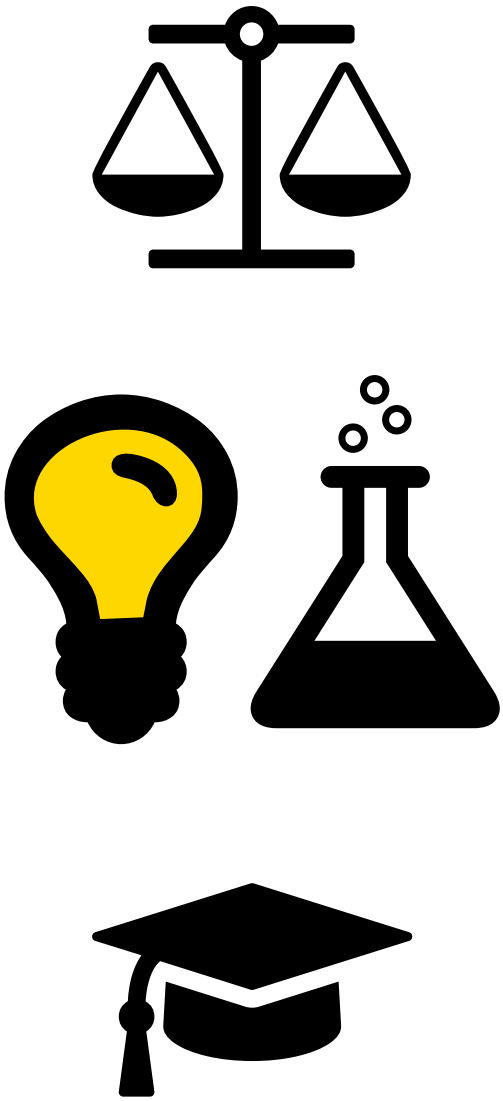
6.4.2 Computational antibody humanisation

For any antibody design campaign, especially in therapeutics, the aim is to develop a safe antibody; this is often achieved by obtaining human antibodies or 'humanising' antibodies from other organisms. This is not limited to Fv's, but also nanobodies. Although humanisation has not been a key focus of this thesis, this is a key property that must be satisfied, especially when designing antibodies for therapeutic use. Key structural features that define the 'humanness' of an antibody, such as V_H–V_L orientation (Bujotzek *et al.*, 2016), should be analysed

in greater detail before implementing strategies to humanise structures. Once these features are identified, a scoring system (*e.g.* Choi *et al.*, 2015) can be developed as an objective function for humanising antibodies. To profile structural 'humanness', we would first need to collect structural data from a humanisation campaign; the data can then be complemented by building models for sequences from various stages of humanisation. Structural changes can then be corrleated to degrees of humanness. Here, we envision that ABodyBuilder will play a central role by providing models for analysis and design.

6.5 Closing remarks

This thesis studies the structural variation of antibodies, and uses the unique profile of antibodies to understand mechanisms, and develop tools for prediction. CAPTAIN was first developed to predict binding affinities. Considering the complexity of the problem, the thesis revisits one of the core components of structural diversity: V_H–V_L pairing. The 'random' nature of pairing justifies the decision to pair almost any two antibody chains for further modelling and development. The increasing amount of antibody structural data in SAbDab has been harnessed to develop our modelling tools, ABodyBuilder and PEARS. Both methods provide high–quality models and annotations to facilitate computational antibody design. However, gaps in our current knowledge base must be addressed in order to realise a fully computational antibody design pipeline.



Bibliography

Abhinandan, K. R. and Martin, A. C. R. (2008). Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Mol Immunol*, **45**(14), 3832–3839. (Cited on pages 35 and 45.)

Abhinandan, K. R. and Martin, A. C. R. (2010). Analysis and prediction of VH/VL packing in antibodies. *Protein Eng Des Sel*, **23**(9), 689–697. (Cited on pages 18, 21, and 173.)

Akagawa, M., Ito, S., Toyoda, K., Ishii, Y., Tatsuda, E., Shibata, T., Yamaguchi, S., Kawai, Y., Ishino, K., Kishi, Y., Adachi, T., Tsubata, T., Takasaki, Y., Hattori, N., Matsuda, T., and Uchida, K. (2006). Bispecific Abs against modified protein and DNA with oxidized lipids. *Proc Natl Acad Sci USA*, **103**(16), 6160–6165. (Cited on page 90.)

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2008). *Molecular Biology of the Cell*. Garland Science, 5th edition. (Cited on pages 4 and 5.)

Almagro, J. C., Beavers, M. P., Hernandez-Guzman, F., Maier, J., Shaulsky, J., Butenhof, K., Labute, P., Thorsteinson, N., Kelly, K., Teplyakov, A., Luo, J., Sweet, R., and Gilliland, G. L. (2011). Antibody modeling assessment. *Proteins*, **79**(11), 3050–3066. (Cited on page 40.)

Almagro, J. C., Teplyakov, A., Luo, J., Sweet, R. W., Kodangattil, S., Hernandez-Guzman, F., and Gilliland, G. L. (2014). Second Antibody Modeling Assessment (AMA–II). *Proteins*, **82**(8), 1553–1562. (Cited on pages 37, 40, 108, 111, 125, 128, 133, and 162.)

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, **215**(3), 403–410. (Cited on page 43.)

Amunts, A., Brown, A., Toots, J., Scheres, S. H. W., and Ramakrishnan, V. (2015). The structure of the human mitochondrial ribosome. *Science*, **348**(6230), 95–98. (Cited on page 6.)

Andrusier, N., Nussinov, R., and Wolfson, H. J. (2007). FireDock: Fast interaction refinement in molecular docking. *Proteins*, **69**(1), 139–159. (Cited on page 57.)

Anfinsen, C. B. (1973). Principles that Govern the Folding of Protein Chains. *Science*, **181**(4096), 223–230. (Cited on page 3.)

Bankovich, A. J., Raunser, S., Juo, Z. S., Walz, T., Davis, M. M., and Garcia, K. C. (2007). Structural Insight into Pre–B Cell Receptor Function. *Science*, **316**(5822), 291–294. (Cited on pages 20 and 92.)

Barderas, R., Desmet, J., Timmerman, P., Meloen, R., and Casal, J. I. (2008). Affinity maturation of antibodies assisted by in silico modeling. *Proc Natl Acad Sci USA*, **105**(26), 9029–9034. (Cited on pages 30, 34, 50, and 162.)

Bentley, J. L. (1975). Multidimensional Binary Search Trees Used for Associative Searching. *Commun. ACM*, **18**(9), 509–517. (Cited on page 151.)

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), 235–242. (Cited on pages 1, 5, 37, and 134.)

Berrondo, M., Kaufmann, S., and Berrondo, M. (2014). Automated aufbau of antibody structures from given sequences using macromoltek's smrtmolantibody. *Proteins*, **82**(8), 1636–1645. (Cited on pages 42, 44, and 133.)

Bhuyan, M. S. I. and Gao, X. (2011). A protein–dependent side–chain rotamer library. *BMC Bioinformatics*, **12**(Suppl 14), S10–S10. (Cited on page 165.)

Biasini, M. (2015). pv: v1.8.1. (Cited on pages 131 and 132.)

Birtalan, S., Zhang, Y., Fellouse, F. A., Shao, L., Schaefer, G., and Sidhu, S. S. (2008). The Intrinsic Contributions of Tyrosine, Serine, Glycine and Arginine to the Affinity and Specificity of Antibodies. *J Mol Biol*, **377**(5), 1518–1528. (Cited on page 69.)

Brenke, R., Hall, D. R., Chuang, G.-Y., Comeau, S. R., Bohnuud, T., Beglov, D., Schueler-Furman, O., Vajda, S., and Kozakov, D. (2012). Application of asymmetric statistical potentials to antibody–protein docking. *Bioinformatics*, **28**(20), 2608–2614. (Cited on page 54.)

Brezinschek, H.-P., Foster, S. J., Dörner, T., Brezinschek, R. I., and Lipsky, P. E. (1998). Pairing of Variable Heavy and Variable κ Chains in Individual Naive and Memory B Cells. *J Immunol*, **160**(10), 4762–4767. (Cited on page 171.)

Brych, S. R., Gokarn, Y. R., Hultgen, H., Stevenson, R. J., Rajan, R., and Matsumura, M. (2010). Characterization of antibody aggregation: Role of buried, unpaired cysteines in particle formation. *J Pharm Sci*, **99**(2), 764–781. (Cited on page 116.)

Bujotzek, A., Fuchs, A., Qu, C., Benz, J., Klostermann, S., Antes, I., and Georges, G. (2015a). MoFvAb: modeling the Fv region of antibodies. *mAbs*, **7**(5), 838–852. (Cited on pages 37, 42, 45, 134, and 166.)

Bujotzek, A., Dunbar, J., Lipsmeier, F., Schäfer, W., Antes, I., Deane, C. M., and Georges, G. (2015b). Prediction of VH–VL domain orientation for antibody variable domain modeling. *Proteins*, **83**(4), 681–695. (Cited on pages 37, 45, 112, 115, 133, and 173.)

Bujotzek, A., Lipsmeier, F., Harris, S. F., Benz, J., Kuglstatter, A., and Georges, G. (2016). VH–VL orientation prediction for antibody humanization candidate selection: A case study. *mAbs*, **8**(2), 288–305. (Cited on pages 21 and 174.)

Burkovitz, A., Sela-Culang, I., and Ofran, Y. (2014). Large–scale analysis of somatic hypermutations in antibodies reveals which structural regions, positions and amino acids are modified to improve affinity. *FEBS J*, **281**(1), 306–319. (Cited on page 30.)

Canutescu, A. A., Shelenkov, A. A., and Dunbrack, R. L. (2003). A graph–theory algorithm for rapid protein side–chain prediction. *Prot Sci*, **12**(9), 2001–2014. (Cited on pages 151, 152, 153, and 166.)

Cauerhff, A., Goldbaum, F. a., and Braden, B. C. (2004). Structural mechanism for affinity maturation of an anti–lysozyme antibody. *Proc Natl Acad Sci USA*, **101**(10), 3539–3544. (Cited on page 30.)

Chailyan, A., Marcatili, P., and Tramontano, A. (2011). The association of heavy and light chain variable domains in antibodies: Implications for antigen specificity. *FEBS J*, **278**(16), 2858–2866. (Cited on page 21.)

Chailyan, A., Tramontano, A., and Marcatili, P. (2012). A database of immunoglobulins with integrated tools: DIGIT. *Nucleic Acids Res*, **40**(D1), D1230–D1234. (Cited on pages 81, 90, 92, and 110.)

Chaudhury, S., Lyskov, S., and Gray, J. J. (2010). PyRosetta: a script–based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, **26**(5), 689–691. (Cited on page 57.)

Chen, R., Li, L., and Weng, Z. (2003). ZDOCK: An initial–stage protein–docking algorithm. *Proteins*, **52**(1), 80–87. (Cited on page 57.)

Cheng, T. M.-K., Blundell, T. L., and Fernandez-Recio, J. (2007). pyDock: Electrostatics and desolvation for effective scoring of rigid–body protein–protein docking. *Proteins*, **68**(2), 503–515. (Cited on pages 53 and 57.)

Chinea, G., Padron, G., Hooft, R. W. W., Sander, C., and Vriend, G. (1995). The use of position–specific rotamers in model building by homology. *Proteins*, **23**(3), 415–421. (Cited on pages 140, 142, and 173.)

Choi, Y. and Deane, C. M. (2010). FREAD revisited: Accurate loop structure prediction using a database search algorithm. *Proteins*, **78**(6), 1431–1440. (Cited on pages 109, 112, and 118.)

Choi, Y. and Deane, C. M. (2011). Predicting antibody complementarity determining region structures without classification. *Mol Biosyst*, **7**(12), 3327–3334. (Cited on pages 37, 40, 42, 109, 112, 118, 125, and 134.)

Choi, Y., Hua, C., Sentman, C. L., Ackerman, M. E., and Bailey-Kellogg, C. (2015). Antibody humanization by structure–based computational protein design. *mAbs*, **7**(6), 1045–1057. (Cited on pages 34 and 175.)

Chothia, C. and Lesk, A. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol*, **196**(4), 901–917. (Cited on pages 8, 10, 11, 13, 14, 15, 17, 111, and 125.)

Chothia, C., Novotný, J., Bruccoleri, R., and Karplus, M. (1985). Domain association in immunoglobulin molecules. *J Mol Biol*, **186**(3), 651–663. (Cited on pages 9, 18, 19, and 21.)

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., Colman, P. M., Spinelli, S., Alzari, P. M., and Poljak, R. J. (1989). Conformations of immunoglobulin hypervariable regions. *Nature*, **342**, 877–883. (Cited on page 15.)

Chothia, C., Gelfand, I., and Kister, A. (1998). Structural determinants in the sequences of immunoglobulin variable domain. *J Mol Biol*, **278**(2), 457–479. (Cited on page 6.)

Clark, L. A., Boriack-Sjodin, P. A., Eldredge, J., Fitch, C., Friedman, B., Hanf, K. J., Jarpe, M., Liparoto, S. F., Li, Y., Lugovskoy, A., Miller, S., Rushe, M., Sherman, W., Simon, K., and Van Vlijmen, H. (2006a). Affinity enhancement of an in vivo matured therapeutic antibody using structure–based computational design. *Prot Sci*, **15**(5), 949–960. (Cited on pages 24, 34, and 50.)

Clark, L. A., Ganesan, S., Papp, S., and van Vlijmen, H. W. T. (2006b). Trends in antibody sequence changes during the somatic hypermutation process. *J Immunol*, **177**(1), 333–340. (Cited on pages 23, 24, 30, and 72.)

Colbes, J., Corona, R. I., Lezcano, C., Rodríguez, D., and Brizuela, C. A. (2016). Protein side–chain packing problem: is there still room for improvement? *Brief Bioinform*. (Cited on pages 138, 140, 142, 143, and 158.)

Collis, A. V., Brouwer, A. P., and Martin, A. C. (2003). Analysis of the antigen combining site: Correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *J Mol Biol*, **325**(2), 337–354. (Cited on page 56.)

Cossio, P., Granata, D., Laio, A., Seno, F., and Trovato, A. (2012). A simple and efficient statistical potential for scoring ensembles of protein structures. *Sci Rep*, **2**, 351. (Cited on page 53.)

De Silva, N. S. and Klein, U. (2015). Dynamics of B cells in germinal centres. *Nat Rev Immunol*, **15**(3), 137–148. (Cited on pages 30, 31, and 195.)

de Wildt, R. M. T., Hoet, R. M. A., van Venrooij, W. J., Tomlinson, I. M., and Winter, G. (1999). Analysis of Heavy and Light Chain Pairings Indicates that Receptor Editing Shapes the Human Antibody Repertoire. *J Mol Biol*, **285**(3), 895–901. (Cited on pages 78, 84, and 102.)

Deane, C. M. and Blundell, T. L. (2001). CODA: A combined algorithm for predicting the structurally variable regions of protein models. *Prot Sci*, **10**(3), 599–612. (Cited on pages 37, 38, 39, 109, 112, and 118.)

DeKosky, B. J., Ippolito, G. C., Deschner, R. P., Lavinder, J. J., Wine, Y., Rawlings, B. M., Varadarajan, N., Giesecke, C., Dorner, T., Andrews, S. F., Wilson, P. C., Hunicke-Smith, S. P., Willson, C. G., Ellington, A. D., and Georgiou, G. (2013). High–throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol*, **31**(2), 166–169. (Cited on page 34.)

DeKosky, B. J., Kojima, T., Rodin, A., Charab, W., Ippolito, G. C., Ellington, A. D., and Georgiou, G. (2015). In–depth determination and analysis of the human paired heavy– and light–chain antibody repertoire. *Nat Med*, **21**(1), 86–91. (Cited on pages 78, 79, 81, 87, 90, 99, 102, 103, 109, 130, and 171.)

DeKosky, B. J., Lungu, O. I., Park, D., Johnson, E. L., Charab, W., Chrysostomou, C., Kuroda, D., Ellington, A. D., Ippolito, G. C., Gray, J. J., and Georgiou, G. (2016). Large–scale sequence and structural comparisons of human naive and antigen–experienced antibody repertoires. *Proc Natl Acad Sci USA*, **113**(19), E2636–E2645. (Cited on pages 34, 79, 108, 109, and 130.)

Desmet, J., Maeyer, M. D., Hazes, B., and Lasters, I. (1992). The dead–end elimination theorem and its use in protein side–chain positioning. *Nature*, **356**(6369), 539–542. (Cited on pages 141, 166, and 174.)

Dunbar, J. and Deane, C. M. (2016). ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, **32**(2), 298–300. (Cited on pages 45, 80, and 109.)

Dunbar, J., Fuchs, A., Shi, J., and Deane, C. M. (2013). ABangle: Characterising the VH–VL orientation in antibodies. *Protein Eng Des Sel*, **26**(10), 611–620. (Cited on pages 18, 21, 22, 74, 78, 98, 104, 109, 115, and 135.)

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 Res*, **42**(D1), D1140–1146. (Cited on pages 55, 68, 81, 105, 108, 110, 116, 134, 143, and 154.)

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 Res*, **44**(W1), W474. (Cited on page 36.)

Dunbrack, R. L. and Cohen, F. E. (1997). Bayesian statistical analysis of protein side–chain rotamer preferences. *Prot Sci*, **6**(8), 1661–1681. (Cited on pages 139, 140, 141, 142, and 154.)

Dunbrack, R. L. and Karplus, M. (1993). Backbone–dependent Rotamer Library for Proteins Application to Side–chain Prediction. *J Mol Biol*, **230**(2), 543–574. (Cited on page 140.)

Durrant, J. D. and McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC Biol*, **9**(1), 71. (Cited on page 51.)

Eddy, S. R. (2004). What is a hidden Markov model? *Nat Biotechnol*, **22**(10), 1315–1316. (Cited on pages 45 and 80.)

Edelman, G. M. (1973). Antibody Structure and Molecular Immunology. *Science*, **180**(4088), 830–840. (Cited on page 4.)

Edwards, B. M., Barash, S. C., Main, S. H., Choi, G. H., Minter, R., Ullrich, S., Williams, E., Du Fou, L., Wilton, J., Albert, V. R., Ruben, S. M., and Vaughan, T. J. (2003). The Remarkable Flexibility of the Human Antibody Repertoire; Isolation of Over One Thousand Different Antibodies to a Single Protein, BLyS. *J Mol Biol*, **334**(1), 103–118. (Cited on pages 78, 79, 90, 103, and 171.)

Ewert, S., Huber, T., Honegger, A., and Plückthun, A. (2003). Biophysical Properties of Human Antibody Variable Domains. *J Mol Biol*, **325**(3), 531–553. (Cited on page 78.)

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*, **82**(8), 1583–1598. (Cited on pages 40, 42, 45, and 109.)

Fera, D., Schmidt, A. G., Haynes, B. F., Gao, F., Liao, H.-X., Kepler, T. B., and Harrison, S. C. (2014). Affinity maturation in an HIV broadly neutralizing B–cell lineage through reorientation of variable domains. *Proc Natl Acad Sci USA*, **111**(28), 10275–10280. (Cited on pages 18, 21, and 74.)

Fernandez-Fuentes, N., Oliva, B., and Fiser, A. (2006). A supersecondary structure library and search algorithm for modeling loops in protein structures. *Nucleic Acids Res*, **34**(7), 2085–2097. (Cited on page 38.)

Finn, J. A., Koehler Leman, J., Willis, J. R., Cisneros, A., Crowe, J. E., and Meiler, J. (2016). Improving Loop Modeling of the Antibody Complementarity–Determining Region 3 Using Knowledge–Based Restraints. *PLOS ONE*, **11**(5), e0154811. (Cited on page 42.)

Fischer, N., Elson, G., Magistrelli, G., Dheilly, E., Fouque, N., Laurendon, A., Gueneau, F., Ravn, U., Depoisier, J.-F., Moine, V., Raimondi, S., Malinge, P., Di Grazia, L., Rousseau, F., Poitevin, Y., Calloud, S., Cayatte, P.-A., Alcoz, M., Pontini, G., Fagète, S., Broyer, L., Corbier, M., Schrag, D., Didelot, G., Bosson, N., Costes, N., Cons, L., Buatois, V., Johnson, Z., Ferlin, W., Masternak, K., and Kosco-Vilbois, M. (2015). Exploiting light chains for the scalable generation and platform purification of native human bispecific IgG. *Nat Commun*, **6**, 6113. (Cited on pages 78 and 103.)

Fiser, A. (2010). *Template–Based Protein Structure Modeling*, pages 73–94. Humana Press, Totowa, NJ. (Cited on page 38.)

Foote, J. and Winter, G. (1992). Antibody framework residues affecting the conformation of the hypervariable loops. *J Mol Biol*, **224**(2), 487–499. (Cited on page 18.)

Gaillard, T., Panel, N., and Simonson, T. (2016). Protein side chain conformation predictions with an MMGBSA energy function. *Proteins*, **84**(6), 803–819. (Cited on page 139.)

Gardberg, A. S., Dice, L. T., Ou, S., Rich, R. L., Helmbrecht, E., Ko, J., Wetzel, R., Myszka, D. G., Patterson, P. H., and Dealwis, C. (2007). Molecular basis for passive immunotherapy of Alzheimer's disease. *Proc Natl Acad Sci USA*, **104**(40), 15659–15664. (Cited on page 90.)

Gavel, Y. and von Heijne, G. (1990). Sequence differences between glycosylated and non–glycosylated Asn–X–Thr/Ser acceptor sites: implications for protein engineering. *Protein Eng*, **3**(5), 433–442. (Cited on pages 116 and 132.)

Georgiou, G., Ippolito, G. C., Beausang, J., Busse, C. E., Wardemann, H., and Quake, S. R. (2014). The promise and challenge of high–throughput sequencing of the antibody repertoire. *Nat Biotechnol*, **32**(2), 158–168. (Cited on pages 26 and 33.)

Giudicelli, V. and Lefranc, M.-P. (1999). Ontology for immunogenetics: the IMGT–ONTOLOGY. *Bioinformatics*, **15**(12), 1047–1054. (Cited on page 29.)

Glanville, J., Zhai, W., Berka, J., Telman, D., Huerta, G., Mehta, G. R., Ni, I., Mei, L., Sundar, P. D., Day, G. M. R., Cox, D., Rajpal, A., and Pons, J. (2009). Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc Natl Acad Sci USA*, **106**(48), 20216–20221. (Cited on pages 78, 79, 84, and 102.)

Haidar, J. N., Zhu, W., Lypowy, J., Pierce, B. G., Bari, A., Persaud, K., Luna, X., Snavely, M., Ludwig, D., and Weng, Z. (2014). Backbone Flexibility of CDR3 and Immune Recognition of Antigens. *J Mol Biol*, **426**(7), 1583–1599. (Cited on pages 74 and 170.)

Hamer, R., Luo, Q., Armitage, J. P., Reinert, G., and Deane, C. M. (2010). i–Patch: Interprotein contact prediction using local network information. *Proteins*, **78**(13), 2781–2797. (Cited on page 55.)

Hansel, T. T., Kropshofer, H., Singer, T., Mitchell, J. A., and George, A. J. T. (2010). The safety and side effects of monoclonal antibodies. *Nat Rev Drug Discov*, **9**(4), 325–338. (Cited on pages 31 and 32.)

Harder, T., Boomsma, W., Paluszewski, M., Frellsen, J., Johansson, K. E., and Hamelryck, T. (2010). Beyond rotamers: a generative, probabilistic model of side chains in proteins. *BMC Bioinformatics*, **11**(1), 306. (Cited on pages 138, 139, 140, and 165.)

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer New York Inc, New York, NY, USA, 2nd edition. (Cited on page 147.)

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA*, **89**(22), 10915–10919. (Cited on page 113.)

Holliger, P. and Hudson, P. J. (2005). Engineered antibody fragments and the rise of single domains. *Nat Biotechnol*, **23**(9), 1126–1136. (Cited on page 32.)

Holm, L. and Sander, C. (1995). Dali: a network tool for protein structure comparison. *Trends Biochem Sci*, **20**(11), 478–480. (Cited on page 38.)

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), 657–670. (Cited on pages 11, 13, and 14.)

Hsu, H.-J., Lee, K., Jian, J.-W., Chang, H.-J., Yu, C.-M., Lee, Y.-C., Chen, I.-C., Peng, H.-P., Wu, C., Huang, Y.-F., Shao, C.-Y., Chiu, K., and Yang, A.-S. (2014). Antibody Variable Domain Interface and Framework Sequence Requirements for Stability and Function by High–Throughput Experiments. *Structure*, **22**(1), 22–34. (Cited on page 18.)

Hubbard, S. J. and Thornton, J. M. (1993). NACCESS: Computer program. (Cited on page 57.)

Igawa, T., Tsunoda, H., Kikuchi, Y., Yoshida, M., Tanaka, M., Koga, A., Sekimori, Y., Orita, T., Aso, Y., Hattori, K., and Tsuchiya, M. (2010). VH/VL interface engineering to promote selective expression and inhibit conformational isomerization of thrombopoietin receptor agonist single–chain diabody. *Protein Eng Des Sel*, **23**(8), 667–677. (Cited on pages 18, 19, and 155.)

Imai, K. and Takaoka, A. (2006). Comparing antibody and small–molecule therapies for cancer. *Nat Rev Cancer*, **6**(9), 714–727. (Cited on pages 31 and 32.)

Jäckel, C., Kast, P., and Hilvert, D. (2008). Protein Design by Directed Evolution. *Annu Rev Biophys*, **37**(1), 153–173. (Cited on page 35.)

Jakobovits, A. (1995). Production of fully human antibodies by transgenic mice. *Curr Opin Biotech*, **6**(5), 561–566. (Cited on pages 32 and 33.)

Janeway, C. A., Travers, P., Walport, M., and Shlomchik, M. J. (2001). *Immunobiology: the immune system in health and disease*. Garland Science, 5th edition. (Cited on pages 25, 26, 27, 30, and 195.)

Jarasch, A., Koll, H., Regula, J. T., Bader, M., Papadimitriou, A., and Kettenberger, H. (2015). Developability Assessment During the Selection of Novel Therapeutic Antibodies. *J Pharm Sci*, **104**(6), 1885–1898. (Cited on pages 31, 32, 108, 109, 115, and 116.)

Jayaram, N., Bhowmick, P., and Martin, A. C. R. (2012). Germline VH/VL pairing in antibodies. *Protein Eng Des Sel*, **25**(10), 523–530. (Cited on pages 78, 79, 82, 84, 87, 98, 102, and 171.)

Jones, P. T., Dear, P. H., Foote, J., Neuberger, M. S., and Winter, G. (1986). Replacing the complementarity–determining regions in a human antibody with those from a mouse. *Nature*, **321**(6069), 522–525. (Cited on page 32.)

Kabat, E. A., Wu, T. T., Bilofsky, H., Reid-Miller, M., and Perry, H. M. (1983). *Sequences of proteins of immunological interest*. National Institutes of Health, 3rd edition. (Cited on pages 11 and 14.)

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, **34**(5), 827–828. (Cited on page 206.)

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen–bonded and geometrical features. *Biopolymers*, **22**(12), 2577–2637. (Cited on pages 115 and 132.)

Kastritis, P. L. and Bonvin, A. M. (2010). Are Scoring Functions in Protein–Protein Docking Ready To Predict Interactomes? Clues from a Novel Binding Affinity Benchmark. *J Proteome Res*, **9**(5), 2216–2225. (Cited on pages 46, 51, and 53.)

Kastritis, P. L., Moal, I. H., Hwang, H., Weng, Z., Bates, P. A., Bonvin, A. M. J. J., and Janin, J. (2011). A structure–based benchmark for protein–protein binding affinity. *Prot Sci*, **20**(3), 482–491. (Cited on pages 55 and 65.)

Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H., and Meiler, J. (2010). Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You. *Biochemistry*, **49**(14), 2987–2998. (Cited on pages 35 and 36.)

Khoury, G. A., Smadbeck, J., Kieslich, C. A., and Floudas, C. A. (2014). Protein folding and de novo protein design for biotechnological applications. *Trends Biotechnol*, **32**(2), 99–109. (Cited on page 35.)

Kirkham, P. M. and Schroeder, H. W. (1994). Antibody structure and the evolution of immunoglobulin V gene segments. *Sem Immunol*, **6**(6), 347–360. (Cited on page 27.)

Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov*, **3**(11), 935–949. (Cited on pages 46, 52, and 53.)

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. (Cited on pages 50 and 51.)

Klein, C., Sustmann, C., Thomas, M., Stubenrauch, K., Croasdale, R., Schanzer, J., Brinkmann, U., Kettenberger, H., Regula, J. T., and Schaefer, W. (2012). Progress in overcoming the chain association issue in bispecific heterodimeric IgG antibodies. *mAbs*, **4**(6), 653–663. (Cited on page 78.)

Köhler, G. and Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature*, **256**(5517), 495–497. (Cited on page 32.)

Krah, S., Schröter, C., Zielonka, S., Empting, M., Valldorf, B., and Kolmar, H. (2016). Single–domain antibodies for biomedical applications. *Immunopharm and Immunot*, **38**(1), 21–28. (Cited on page 32.)

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 Eng Des Sel*, **26**(10), 621–629. (Cited on pages 8, 23, 36, 46, 57, and 72.)

Krawczyk, K., Liu, X., Baker, T., Shi, J., and Deane, C. M. (2014). Improving B–cell epitope prediction and its application to global antibody–antigen docking. *Bioinformatics*, **30**(16), 2288–2294. (Cited on pages 8, 24, 35, 46, and 54.)

Krivov, G. G., Shapovalov, M. V., and Dunbrack, R. L. (2009). Improved prediction of protein side–chain conformations with SCWRL4. *Proteins*, **77**(4), 778–795. (Cited on pages 42, 43, 109, 113, 120, 134, 138, 141, 154, 166, and 173.)

Kunik, V., Peters, B., and Ofran, Y. (2012). Structural consensus among antibodies defines the antigen binding site. *PLOS Comput Biol*, **8**(2), e1002388. (Cited on pages 8, 15, 23, and 46.)

Kuroda, D. and Gray, J. J. (2016). Shape complementarity and hydrogen bond preferences in protein–protein interfaces: implications for antibody modeling and protein–protein docking. *Bioinformatics*, **32**(16), 2451–2456. (Cited on pages 19, 24, 79, and 92.)

Kuroda, D., Shirai, H., Kobori, M., and Nakamura, H. (2008). Structural classification of CDR–H3 revisited: a lesson in antibody modeling. *Proteins*, **73**(3), 608–620. (Cited on page 18.)

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*, **75**(1), 139–146. (Cited on pages 18 and 104.)

Kuroda, D., Shirai, H., Jacobson, M. P., and Nakamura, H. (2012). Computer–aided antibody design. *Protein Eng Des Sel*, **25**(10), 507–521. (Cited on page 34.)

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), 1259–1268. (Cited on pages 47 and 142.)

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. *Dev Comp Immunol*, **27**(1), 55–77. (Cited on pages 11, 13, 14, 17, 109, and 165.)

Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G., and Duroux, P. (2009). IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res*, **37**(D1), D1006–D1012. (Cited on pages 15, 29, and 80.)

Lewis, S. M., Wu, X., Pustilnik, A., Sereno, A., Huang, F., Rick, H. L., Guntas, G., Leaver-Fay, A., Smith, E. M., Ho, C., Hansen-Estruch, C., Chamberlain, A. K., Truhlar, S. M., Conner, E. M., Atwell, S., Kuhlman, B., and Demarest, S. J. (2014). Generation of bispecific IgG antibodies by structure–based design of an orthogonal Fab interface. *Nat Biotechnol*, **32**(2), 191–198. (Cited on pages 6, 18, 19, 35, 78, and 171.)

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), 1–17. (Cited on pages 35, 36, and 45.)

Li, W. and Godzik, A. (2006). Cd–hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**(13), 1658–1659. (Cited on pages 55, 81, 110, and 143.)

Liang, S., Liu, S., Zhang, C., and Zhou, Y. (2007). A simple reference state makes a significant improvement in near–native selections from structurally refined docking decoys. *Proteins*, **69**(2), 244–253. (Cited on page 57.)

Liang, S., Zhou, Y., Grishin, N., and Standley, D. M. (2011). Protein side chain modeling with orientation–dependent atomic force fields derived by series expansions. *J Comput Chem*, **32**(8), 1680–1686. (Cited on pages 43, 141, 142, 143, and 154.)

Lippow, S. M., Wittrup, K. D., and Tidor, B. (2007). Computational design of antibody–affinity improvement beyond in vivo maturation. *Nat Biotechnol*, **25**(10), 1171–1176. (Cited on pages 34, 50, and 51.)

Liu, S., Zhang, C., Zhou, H., and Zhou, Y. (2004). A physical reference state unifies the structure–derived potential of mean force for protein folding and binding. *Proteins*, **56**(1), 93–101. (Cited on page 57.)

Liu, Y. D., Goetze, A. M., Bass, R. B., and Flynn, G. C. (2011). N–terminal Glutamate to Pyroglutamate Conversion in Vivo for Human IgG2 Antibodies. *J Biol Chem*, **286**(13), 11211–11217. (Cited on page 116.)

Lovell, S. C., Word, J. M., Richardson, J. S., and Richardson, D. C. (2000). The penultimate rotamer library. *Proteins*, **40**(3), 389–408. (Cited on pages 139, 140, and 165.)

Lu, M., Dousis, A. D., and Ma, J. (2008). OPUS–Rota: A fast and accurate method for side–chain modeling. *Prot Sci*, **17**(9), 1576–1585. (Cited on pages 141 and 142.)

MacCallum, R. M., Martin, A. C., and Thornton, J. M. (1996). Antibody–antigen Interactions: Contact Analysis and Binding Site Topography. *J Mol Biol*, **262**(5), 732–745. (Cited on pages 15 and 23.)

Maier, J. K. X. and Labute, P. (2014). Assessment of fully automated antibody homology modeling protocols in molecular operating environment. *Proteins*, **82**(8), 1599–1610. (Cited on pages 42 and 44.)

Marasco, W. A. and Sui, J. (2007). The growth and potential of human antiviral monoclonal antibody therapeutics. *Nat Biotechnol*, **25**(12), 1421–1434. (Cited on pages 33 and 34.)

Marcatili, P., Olimpieri, P. P., Chailyan, A., and Tramontano, A. (2014). Antibody structural modeling with prediction of immunoglobulin structure (PIGS). *Nat Protoc*, **9**(12), 2771–2783. (Cited on pages 37, 40, 41, 42, 43, 108, 109, 129, and 172.)

Mardia, K. V., Taylor, C. C., and Subramaniam, G. K. (2007). Protein Bioinformatics and Mixtures of Bivariate von Mises Distributions for Angular Data. *Biometrics*, **63**(2), 505–512. (Cited on page 165.)

Market, E. and Papavasiliou, F. N. (2003). V(D)J recombination and the evolution of the adaptive immune system. *PLOS Biol*, **1**(1), E16. (Cited on pages 26 and 27.)

Marks, C., Nowak, J., Klostermann, S., Georges, G., Dunbar, J., Shi, J., Kelm, S., and Deane, C. (2016). Sphinx: Merging knowledge–based and ab initio approaches to improve protein loop prediction. *Bioinformatics*. (Cited on page 42.)

Mårtensson, I.-L., Almqvist, N., Grimsholm, O., and Bernardi, A. I. (2010). The pre–B cell receptor checkpoint. *FEBS Lett*, **584**(12), 2572–2579. (Cited on page 19.)

Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F., and Šali, A. (2000). Comparative Protein Structure Modeling of Genes and Genomes. *Annu Rev Biophys Biomol Struct*, **29**(1), 291–325. (Cited on pages 37, 38, and 40.)

Martin, A. C. and Thornton, J. M. (1996). Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *J Mol Biol*, **263**(5), 800–815. (Cited on page 15.)

Marze, N. A., Lyskov, S., and Gray, J. J. (2016). Improved prediction of antibody VL–VH orientation. *Protein Eng Des Sel*, **29**(10), 409–418. (Cited on pages 21, 22, 133, and 173.)

Masuda, K., Sakamoto, K., Kojima, M., Aburatani, T., Ueda, T., and Ueda, H. (2006). The role of interface framework residues in determining antibody VH/VL interaction strength and antigen–binding affinity. *FEBS J*, **273**(10), 2184–2194. (Cited on pages 18 and 21.)

Mathonet, P. and Ullman, C. (2013). The Application of Next Generation Sequencing to the Understanding of Antibody Repertoires. *Front Immunol*, **4**, 265. (Cited on page 34.)

Matsuda, F., Ishii, K., Bourvagnet, P., Kuma, K.-i., Hayashida, H., Miyata, T., and Honjo, T. (1998). The Complete Nucleotide Sequence of the Human Immunoglobulin Heavy Chain Variable Region Locus. *J Exp Med*, **188**(11), 2151–2162. (Cited on page 27.)

May, A., Pool, R., van Dijk, E., Bijlard, J., Abeln, S., Heringa, J., and Feenstra, K. A. (2013). Coarse–grained versus atomistic simulations: realistic interaction free energies for real proteins. *Bioinformatics*, **30**(3), 326–334. (Cited on page 51.)

McCafferty, J., Griffiths, A. D., Winter, G., and Chiswell, D. J. (1990). Phage antibodies: filamentous phage displaying antibody variable domains. *Nature*, **348**(6301), 552–554. (Cited on pages 32 and 33.)

Melchers, F. (2005). The pre–B–cell receptor: selector of fitting immunoglobulin heavy chains for the B–cell repertoire. *Nat Rev Immunol*, **5**(7), 578–584. (Cited on pages 19, 78, 92, and 104.)

Messih, M. A., Lepore, R., Marcatili, P., and Tramontano, A. (2014). Improving the accuracy of the structure prediction of the third hypervariable loop of the heavy chains of antibodies. *Bioinformatics*, **30**(19), 2733–2740. (Cited on pages 37 and 42.)

Miao, Z., Cao, Y., and Jiang, T. (2011). RASP: rapid modeling of protein side chain conformations. *Bioinformatics*, **27**(22), 3117–3122. (Cited on pages 141, 154, 162, and 166.)

Miklos, A. E., Kluwe, C., Der, B. S., Pai, S., Sircar, A., Hughes, R. A., Berrondo, M., Xu, J., Codrea, V., Buckley, P. E., Calm, A. M., Welsh, H. S., Warner, C. R., Zacharko, M. A., Carney, J. P., Gray, J. J., Georgiou, G., Kuhlman, B., and Ellington, A. D. (2012). Structure–Based Design of Supercharged, Highly Thermoresistant Antibodies. *Chem Biol*, **19**(4), 449–455. (Cited on pages 35 and 108.)

Moal, I. H. and Fernandez-Recio, J. (2013). Intermolecular Contact Potentials for Protein–Protein Interactions Extracted from Binding Free Energy Changes upon Mutation. *J Chem Theory Comput*, **9**(8), 3715–3727. (Cited on pages 46, 53, and 54.)

Moal, I. H., Agius, R., and Bates, P. A. (2011). Protein–protein binding affinity prediction on a diverse set of structures. *Bioinformatics*, **27**(21), 3002–3009. (Cited on pages 52 and 55.)

Morstadt, L., Bohm, A., Yüksel, D., Kumar, K., Stollar, B. D., and Baleja, J. D. (2008). Engineering and characterization of a single chain surrogate light chain variable domain. *Prot Sci*, **17**(3), 458–465. (Cited on pages 80 and 103.)

Mullard, A. (2013). Maturing antibody–drug conjugate pipeline hits 30. *Nat Rev Drug Discov*, **12**(5), 329–332. (Cited on page 32.)

Mullard, A. (2015). 2014 FDA drug approvals. *Nat Rev Drug Discov*, **14**(2), 77–81. (Cited on page 31.)

Mullard, A. (2016). 2015 FDA drug approvals. *Nat Rev Drug Discov*, **15**(2), 73–76. (Cited on pages 31 and 32.)

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, **247**(4), 536–540. (Cited on page 4.)

Muyldermans, S., Cambillau, C., and Wyns, L. (2001). Recognition of antigens by single–domain antibody fragments: the superfluous luxury of paired domains. *Trends Biochem Sci*, **26**(4), 230–235. (Cited on page 78.)

Nakanishi, T., Tsumoto, K., Yokota, A., Kondo, H., and Kumagai, I. (2008). Critical contribution of VH–VL interaction to reshaping of an antibody: The case of humanization of anti–lysozyme antibody, HyHEL–10. *Prot Sci*, **17**(2), 261–270. (Cited on pages 18 and 21.)

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. *J Mol Biol*, **388**(5), 941–953. (Cited on page 21.)

Nelson, A. L., Dhimolea, E., and Reichert, J. M. (2010). Development trends for human monoclonal antibody therapeutics. *Nat Rev Drug Discov*, **9**(10), 767–774. (Cited on page 32.)

Nelson, D. L. and Cox, M. M. (2004). *Lehninger Principles of Biochemistry*. W. H. Freeman, 4th edition. (Cited on pages 3 and 192.)

Nemazee, D. (2006). Receptor editing in lymphocyte development and central tolerance. *Nat Rev Immunol*, **6**(10), 728–740. (Cited on pages 30, 78, 195, and 205.)

North, B., Lehmann, A., and Dunbrack, R. L. (2011). A new clustering of antibody CDR loop conformations. *J Mol Biol*, **406**(2), 228–256. (Cited on pages 10, 14, 15, 17, 109, 111, 125, and 127.)

Novotný, J. and Haber, E. (1985). Structural invariants of antigen binding: comparison of immunoglobulin VL–VH and VL–VL domain dimers. *Proc Natl Acad Sci USA*, **82**(14), 4592–4596. (Cited on pages 18 and 19.)

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), 751–760. (Cited on pages 15 and 18.)

Nutt, S. L., Hodgkin, P. D., Tarlinton, D. M., and Corcoran, L. M. (2015). The generation of antibody–secreting plasma cells. *Nat Rev Immunol*, **15**(3), 160–171. (Cited on page 30.)

Ohnishi, K. and Melchers, F. (2003). The nonimmunoglobulin portion of λ5 mediates cell–autonomous pre–B cell receptor signaling. *Nat Immunol*, **4**(9), 849–856. (Cited on page 103.)

Olimpieri, P. P., Chailyan, A., Tramontano, A., and Marcatili, P. (2013). Prediction of site–specific interactions in antibody–antigen complexes: The proABC method and server. *Bioinformatics*, **29**(18), 2285–2291. (Cited on page 23.)

Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994). Protein superfamilles and domain superfolds. *Nature*, **372**(6507), 631–634. (Cited on page 4.)

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). CATH – a hierarchic classification of protein domain structures. *Structure*, **5**(8), 1093–1109. (Cited on page 4.)

Padlan, E. A., Abergel, C., and Tipper, J. P. (1995). Identification of specificity–determining residues in antibodies. *FASEB J*, **9**(1), 133–139. (Cited on page 23.)

Pallara, C., Jiménez-García, B., Pérez-Cano, L., Romero-Durana, M., Solernou, A., Grosdidier, S., Pons, C., Moal, I. H., and Fernandez-Recio, J. (2013). Expanding the frontiers of protein–protein modeling: From docking and scoring to binding affinity predictions and other challenges. *Proteins*, **81**(12), 2192–2200. (Cited on page 51.)

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 Eng Des Sel*, **23**(11), 849–858. (Cited on page 35.)

Pei, X. Y., Holliger, P., Murzin, A. G., and Williams, R. L. (1997). The 2.0–Å resolution crystal structure of a trimeric antibody fragment with noncognate VH–VL domain pairs shows a rearrangement of VH CDR3. *Proc Natl Acad Sci USA*, **94**(18), 9637–9642. (Cited on page 90.)

Peled, J. U., Kuang, F. L., Iglesias-Ussel, M. D., Roa, S., Kalis, S. L., Goodman, M. F., and Scharff, M. D. (2008). The Biochemistry of Somatic Hypermutation. *Annu Rev Immunol*, **26**(1), 481–511. (Cited on pages 29 and 30.)

Peterson, L. X., Kang, X., and Kihara, D. (2014). Assessment of protein side–chain conformation prediction methods in different residue environments. *Proteins*, **82**(9), 1971–1984. (Cited on pages 140, 141, and 142.)

Pierce, B. and Weng, Z. (2007). ZRANK: Reranking protein docking predictions with an optimized energy function. *Proteins*, **67**(4), 1078–1086. (Cited on pages 53 and 57.)

Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., van der Spoel, D., Hess, B., and Lindahl, E. (2013). GROMACS 4.5: a high–throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, **29**(7), 845–854. (Cited on page 58.)

Qian, J., Liu, T., Yang, L., Daus, A., Crowley, R., and Zhou, Q. (2007). Structural characterization of N–linked oligosaccharides on monoclonal antibody cetuximab by the combination of orthogonal matrix–assisted laser desorption/ionization hybrid quadrupole–quadrupole time–of–flight tandem mass spectrometry and sequential enzymatic digestion. *Anal Biochem*, **364**(1), 8–18. (Cited on page 135.)

Raghunathan, G., Smart, J., Williams, J., and Almagro, J. C. (2012). Antigen–binding site anatomy and somatic mutations in antibodies that recognize different types of antigens. *J Mol Recogn*, **25**(3), 103–113. (Cited on pages 23 and 24.)

Robin, G., Sato, Y., Desplancq, D., Rochel, N., Weiss, E., and Martineau, P. (2014). Restricted diversity of antigen binding residues of antibodies revealed by computational alanine scanning of 227 Antibody–Antigen complexes. *J Mol Biol*, **426**(22), 3729–3743. (Cited on pages 54, 61, and 69.)

Robinson, W. H. (2015). Sequencing the functional antibody repertoire–diagnostic and therapeutic discovery. *Nat Rev Rheumatol*, **11**(3), 171–182. (Cited on pages 34, 109, and 130.)

Ross, G. A., Morris, G. M., and Biggin, P. C. (2013). One size does not fit all: The limits of structure–based models in drug discovery. *J Chem Theory Comput*, **9**(9), 4266–4274. (Cited on pages 46, 54, 71, and 141.)

Ruoslahti, E. (1996). RGD and other recognition sequences for integrins. *Annu Rev Cell Dev Biol*, **12**(1), 697–715. (Cited on page 116.)

Ryu, J., Lee, M., Cha, J., Laskowski, R. A., Ryu, S. E., and Kim, D.-S. (2016). BetaSCPWeb: side–chain prediction for protein structures using Voronoi diagrams and geometry prioritization. *Nucleic Acids Res*, **44**(W1), W416–W423. (Cited on page 142.)

Šali, A. and Blundell, T. L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J Mol Biol*, **234**(3), 779–815. (Cited on pages 37, 38, 113, 118, 120, 130, and 134.)

Samudrala, R. and Moult, J. (1998a). An all–atom distance–dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol*, **275**(5), 895–916. (Cited on pages 52, 53, 54, 59, and 170.)

Samudrala, R. and Moult, J. (1998b). Determinants of side chain conformational preferences in protein structures. *Protein Eng*, **11**(11), 991–997. (Cited on page 166.)

Sander, C. and Schneider, R. (1991). Database of homology–derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**(1), 56–68. (Cited on pages 4 and 5.)

Scarabelli, G. and Grant, B. J. (2013). Mapping the structural and dynamical features of kinesin motor domains. *PLOS Comput Biol*, **9**(11), e1003329. (Cited on pages 83 and 96.)

Schatz, D. G. and Ji, Y. (2011). Recombination centres and the orchestration of V(D)J recombination. *Nat Rev Immunol*, **11**(4), 251–263. (Cited on page 27.)

Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., and Platt, J. (2000). Support vector method for novelty detection. In *Advances in Neural Information Processing Systems 12*, pages 582–588, Cambridge, MA, USA. Max–Planck–Gesellschaft, MIT Press. (Cited on page 105.)

Schroeder, H. W. and Cavacini, L. (2010). Structure and function of immunoglobulins. *J Allergy Clin Immunol*, **125**, 41–52. (Cited on pages 8, 24, and 27.)

Schwede, T., Kopp, J., Guex, N., and Peitsch, M. C. (2003). SWISS–MODEL: an automated protein homology–modeling server. *Nucleic Acids Res*, **31**(13), 3381–3385. (Cited on page 38.)

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Res*, **33**, W382–W388. (Cited on pages 58, 61, 68, and 69.)

Seeliger, D., Schulz, P., Litzenburger, T., Spitz, J., Hoerer, S., Blech, M., Enenkel, B., Studts, J. M., Garidel, P., and Karow, A. R. (2015). Boosting antibody developability through rational sequence optimization. *mAbs*, **7**(3), 505–515. (Cited on pages 31, 108, and 115.)

Sela-Culang, I., Kunik, V., and Ofran, Y. (2013). The Structural Basis of Antibody–Antigen Recognition. *Front Immunol*, **4**, 302. (Cited on page 8.)

Shapovalov, M. V. and Dunbrack, R. L. (2011). A smoothed backbone–dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, **19**(6), 844–858. (Cited on pages 39, 138, 139, 140, 141, 142, 144, 154, 155, 158, and 165.)

Shetty, R. P., de Bakker, P. I. W., DePristo, M. A., and Blundell, T. L. (2003). Advantages of fine–grained side chain conformer libraries. *Protein Eng*, **16**(12), 963–969. (Cited on page 139.)

Shirai, H., Kidera, A., and Nakamura, H. (1999). H3–rules: identification of CDR–H3 structures in antibodies. *FEBS Lett*, **455**(1âĂŞ2), 188–197. (Cited on pages 18 and 43.)

Shirai, H., Ikeda, K., Yamashita, K., Tsuchiya, Y., Sarmiento, J., Liang, S., Morokata, T., Mizuguchi, K., Higo, J., Standley, D. M., and Nakamura, H. (2014). High–resolution modeling of antibody structures by a combination of bioinformatics, expert knowledge, and molecular simulations. *Proteins*, **82**(8), 1624–1635. (Cited on pages 42, 43, 109, 129, and 133.)

Sippl, M. J. (1990). Calculation of Conformational Ensembles from Potentials of Mean Force. *J Mol Biol*, **213**(4), 859–883. (Cited on pages 63 and 73.)

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. (Cited on pages 35, 37, 42, 43, 108, 129, 133, and 172.)

Smith, B. P. and Roman, C. A. J. (2010). The unique and immunoglobulin–like regions of surrogate light chain component $\lambda 5$ differentially interrogate immunoglobulin heavy–chain structure. *Mol Immunol*, **47**(6), 1195–1206. (Cited on pages 21 and 104.)

Studer, G., Biasini, M., and Schwede, T. (2014). Assessing the local structural quality of transmembrane protein models using statistical potentials (QMEANBrane). *Bioinformatics*, **30**(17), i505–i511. (Cited on pages 63, 71, 73, and 170.)

Sutcliffe, M. J., Haneef, I., Carney, D., and Blundell, T. L. (1987). Knowledge based modelling of homologous proteins, part I: three–dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng*, **1**(5), 377–384. (Cited on page 38.)

Swindells, M. B., Porter, C. T., Couch, M., Hurst, J., Abhinandan, K. R., Nielsen, J. H., Macindoe, G., Hetherington, J., and Martin, A. C. R. (2016). abYsis: Integrated Antibody Sequence and Structure–Management, Analysis, and Prediction. *J Mol Biol*. (Cited on pages 81 and 110.)

Sydow, J. F., Lipsmeier, F., Larraillet, V., Hilger, M., Mautz, B., Mølhøj, M., Kuentzer, J., Klostermann, S., Schoch, J., Voelger, H. R., Regula, J. T., Cramer, P., Papadimitriou, A., and Kettenberger, H. (2014). Structure–Based Prediction of Asparagine and Aspartate Degradation Sites in Antibody Variable Regions. *PLOS ONE*, **9**(6), e100736. (Cited on pages 116 and 135.)

Teplyakov, A. and Gilliland, G. L. (2014). Canonical structures of short CDR–L3 in antibodies. *Proteins*, **82**(8), 1668–1673. (Cited on page 18.)

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), 1045–1063. (Cited on pages xvii, 78, 79, 81, 104, 172, and 202.)

Tiller, T., Schuster, I., Deppe, D., Siegers, K., Strohner, R., Herrmann, T., Berenguer, M., Poujol, D., Stehle, J., Stark, Y., Heßling, M., Daubert, D., Felderer, K., Kaden, S., Kölln, J., Enzelberger, M., and Urlinger, S. (2013). A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. *mAbs*, **5**(3), 445–470. (Cited on pages 34, 78, and 87.)

Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature*, **302**(5909), 575–581. (Cited on page 26.)

Townsend, C. L., Laffy, J. M. J., Wu, Y.-C. B., Silva O'Hare, J., Martin, V., Kipling, D., Fraternali, F., and Dunn-Walters, D. K. (2016). Significant Differences in Physicochemical Properties of Human Immunoglobulin Kappa and Lambda CDR3 Regions. *Front Immunol*, **7**, 388. (Cited on pages 18 and 79.)

Towse, C.-L., Rysavy, S., Vulovic, I., and Daggett, V. (2016). New Dynamic Rotamer Libraries: Data–Driven Analysis of Side–Chain Conformational Propensities. *Structure*, **24**(1), 187–199. (Cited on pages 39, 139, 140, 144, and 145.)

Tuffery, P., Etchebest, C., Hazout, S., and Lavery, R. (1991). A New Approach to the Rapid Determination of Protein Side Chain Conformations. *J Biomol Struct Dyn*, **8**(6), 1267–1289. (Cited on pages 139 and 140.)

Vangone, A. and Bonvin, A. M. J. J. (2015). Contacts–based prediction of binding affinity in proteinâĂŞprotein complexes. *eLife*, **4**, e07454. (Cited on pages 46, 52, 53, and 57.)

Vargas-Madrazo, E. and Paz-García, E. (2003). An improved model of association for VH–VL immunoglobulin domains: asymmetries between VH and VL in the packing of some interface residues. *J Mol Recogn*, **16**(3), 113–120. (Cited on page 18.)

Vargas-Madrazo, E., Lara-Ochoa, F., Ramirez-Benites, M. C., and Almagro, J. C. (1997). Evolution of the structural repertoire of the human V_H and $V\kappa$ germline genes. *Int Immunol*, **9**(12), 1801–1815. (Cited on page 27.)

Vlasak, J. and Ionescu, R. (2011). Fragmentation of monoclonal antibodies. *mAbs*, **3**(3), 253–263. (Cited on page 116.)

Vreven, T., Hwang, H., and Weng, Z. (2011). Integrating atom–based and residue–based scoring functions for protein–protein docking. *Prot Sci*, **20**(9), 1576–1586. (Cited on pages 52, 53, and 57.)

Vreven, T., Hwang, H., Pierce, B., and Weng, Z. (2012). Prediction of protein–protein binding free energies. *Prot Sci*, **21**(3), 396–404. (Cited on pages 53 and 65.)

Wan, S., Knapp, B., Wright, D. W., Deane, C. M., and Coveney, P. V. (2015). Rapid, Precise, and Reproducible Prediction of Peptide–MHC Binding Affinities from Molecular Dynamics That Correlate Well with Experiment. *J Chem Theory Comput*, **11**(7), 3346–3356. (Cited on page 51.)

Wang, B., Kluwe, C. A., Lungu, O. I., DeKosky, B. J., Kerr, S. A., Johnson, E. L., Jung, J., Rezigh, A. B., Carroll, S. M., Reyes, A. N., Bentz, J. R., Villanueva, I., Altman, A. L., Davey, R. A., Ellington, A. D., and Georgiou, G. (2015). Facile Discovery of a Diverse Panel of Anti–Ebola Virus Antibodies by Immune Repertoire Mining. *Sci Rep*, **5**, 13926. (Cited on page 34.)

Wang, F., Sen, S., Zhang, Y., Ahmad, I., Zhu, X., Wilson, I. A., Smider, V. V., Magliery, T. J., and Schultz, P. G. (2013). Somatic hypermutation maintains antibody thermodynamic stability during affinity maturation. *Proc Natl Acad Sci USA*, **110**(11), 4261–4266. (Cited on page 30.)

Wang, K., Fain, B., Levitt, M., and Samudrala, R. (2004). Improved protein structure selection using decoy–dependent discriminatory functions. *BMC Struct Biol*, **4**(8), 1 – 18. (Cited on pages 54, 61, 62, and 68.)

Weiner, G. J. (2015). Building better monoclonal antibody–based therapeutics. *Nat Rev Cancer*, **15**(6), 361–370. (Cited on page 25.)

Willis, J. R., Briney, B. S., DeLuca, S. L., Crowe, J. E., and Meiler, J. (2013). Human Germline Antibody Gene Segments Encode Polyspecific Antibodies. *PLOS Comput Biol*, **9**(4), e1003045. (Cited on pages 24 and 30.)

Wlodawer, A., Minor, W., Dauter, Z., and Jaskolski, M. (2013). Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination. *FEBS J*, **280**(22), 5705–5736. (Cited on page 5.)

Wong, S. E., Sellers, B. D., and Jacobson, M. P. (2011). Effects of somatic mutations on CDR loop flexibility during affinity maturation. *Proteins*, **79**(3), 821–829. (Cited on page 30.)

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. *J Exp Med*, **132**(2), 211–250. (Cited on pages 8, 9, and 13.)

Wüthrich, K. (1990). Protein structure determination in solution by NMR spectroscopy. *J Biol Chem*, **265**(36), 22059–62. (Cited on page 5.)

Xiang, Z. and Honig, B. (2001). Extending the accuracy limits of prediction for side–chain conformations. *J Mol Biol*, **311**(2), 421–430. (Cited on page 139.)

Yamashita, K., Ikeda, K., Amada, K., Liang, S., Tsuchiya, Y., Nakamura, H., Shirai, H., and Standley, D. M. (2014). Kotai Antibody Builder: Automated high–resolution structural modeling of antibodies. *Bioinformatics*, **30**(22), 3279–3280. (Cited on page 43.)

Yan, Z. and Wang, J. (2013). Optimizing Scoring Function of Protein–Nucleic Acid Interactions with Both Affinity and Specificity. *PLOS ONE*, **8**(9), e74443. (Cited on page 54.)

Yang, Y. and Zhou, Y. (2008). Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins*, **72**(2), 793–803. (Cited on page 57.)

Yugandhar, K. and Gromiha, M. M. (2014). Feature selection and classification of protein–protein complexes based on their binding affinities using machine learning approaches. *Proteins*, **82**(9), 2088–2096. (Cited on page 52.)

Zemla, A. (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res*, **31**(13), 3370–3374. (Cited on page 39.)

Zhang, Y. and Skolnick, J. (2004a). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci USA*, **101**(20), 7594–7599. (Cited on page 38.)

Zhang, Y. and Skolnick, J. (2004b). Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**(4), 702–710. (Cited on page 39.)

Zhou, H. and Skolnick, J. (2011). GOAP: A Generalized Orientation–Dependent, All–Atom Statistical Potential for Protein Structure Prediction. *Biophys J*, **101**(8), 2043 – 2052. (Cited on page 57.)

Zhu, J., Ofek, G., Yang, Y., Zhang, B., Louder, M. K., Lu, G., McKee, K., Pancera, M., Skinner, J., Zhang, Z., Parks, R., Eudailey, J., Lloyd, K. E., Blinn, J., Alam, S. M., Haynes, B. F., Simek, M., Burton, D. R., Koff, W. C., Program, N. C. S., Mullikin, J. C., Mascola, J. R., Shapiro, L., and Kwong, P. D. (2013). Mining the antibodyome for HIV–1–neutralizing antibodies with next–generation sequencing and phylogenetic pairing of heavy/light chains. *Proc Natl Acad Sci USA*, **110**(16), 6470–6475. (Cited on page 34.)

Zhu, K. and Day, T. (2013). Ab initio structure prediction of the antibody hypervariable H3 loop. *Proteins*, **81**(6), 1081–1089. (Cited on page 42.)

Zhu, K., Day, T., Warshaviak, D., Murrett, C., Friesner, R., and Pearlman, D. (2014). Antibody structure determination using a combination of homology modeling, energy–based refinement, and loop prediction. *Proteins*, **82**(8), 1646–1655. (Cited on pages 42, 44, and 128.)

Appendices

The greatest value of a picture is when it forces us to notice what we never expected to see.

— John Tukey

A

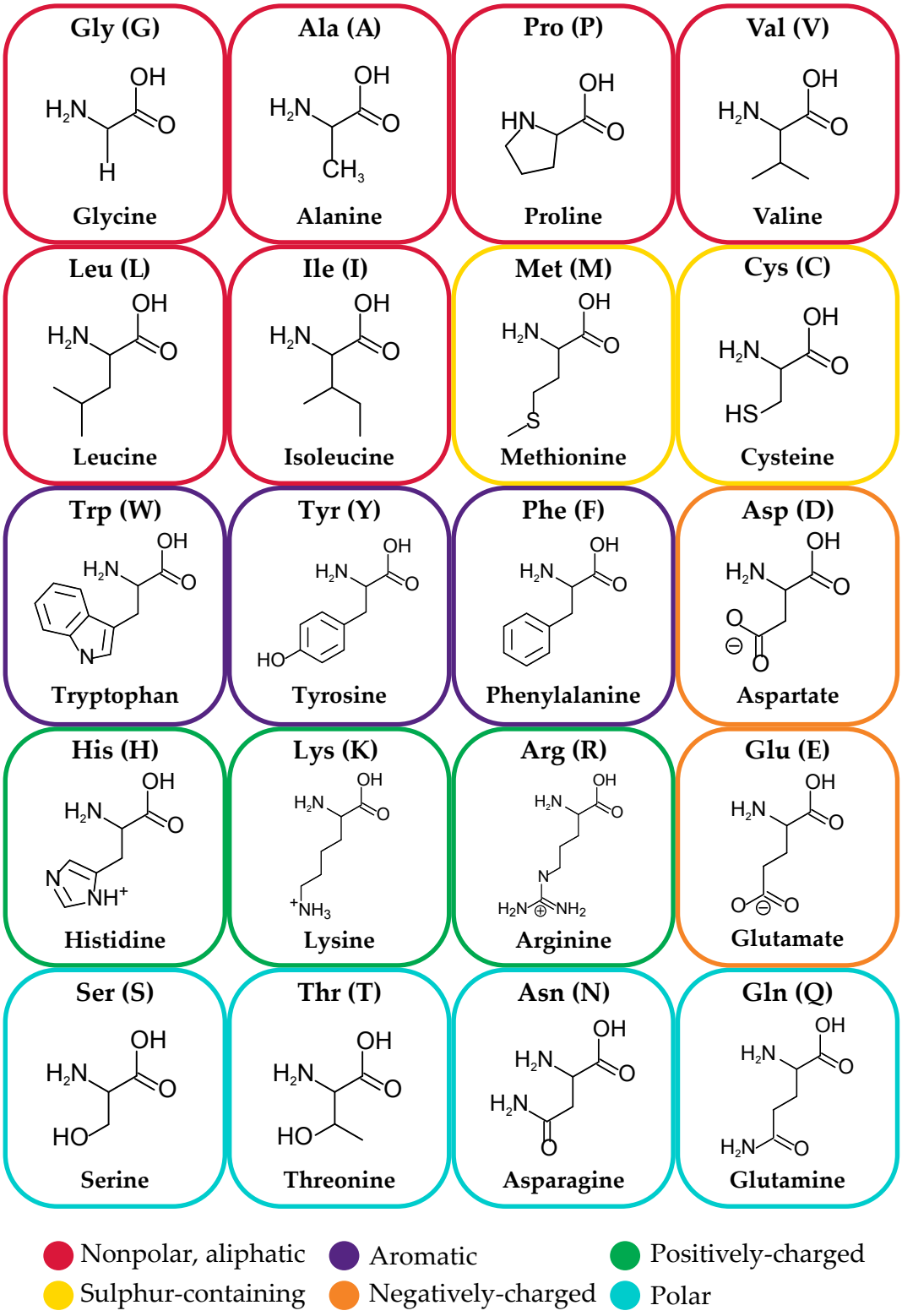


Figure A.1: The structures of 20 amino acids that are typically used in protein synthesis. Amino acids have been grouped according to chemical properties (Nelson and Cox, 2004). Throughout the thesis, the three–letter format of amino acids is used.

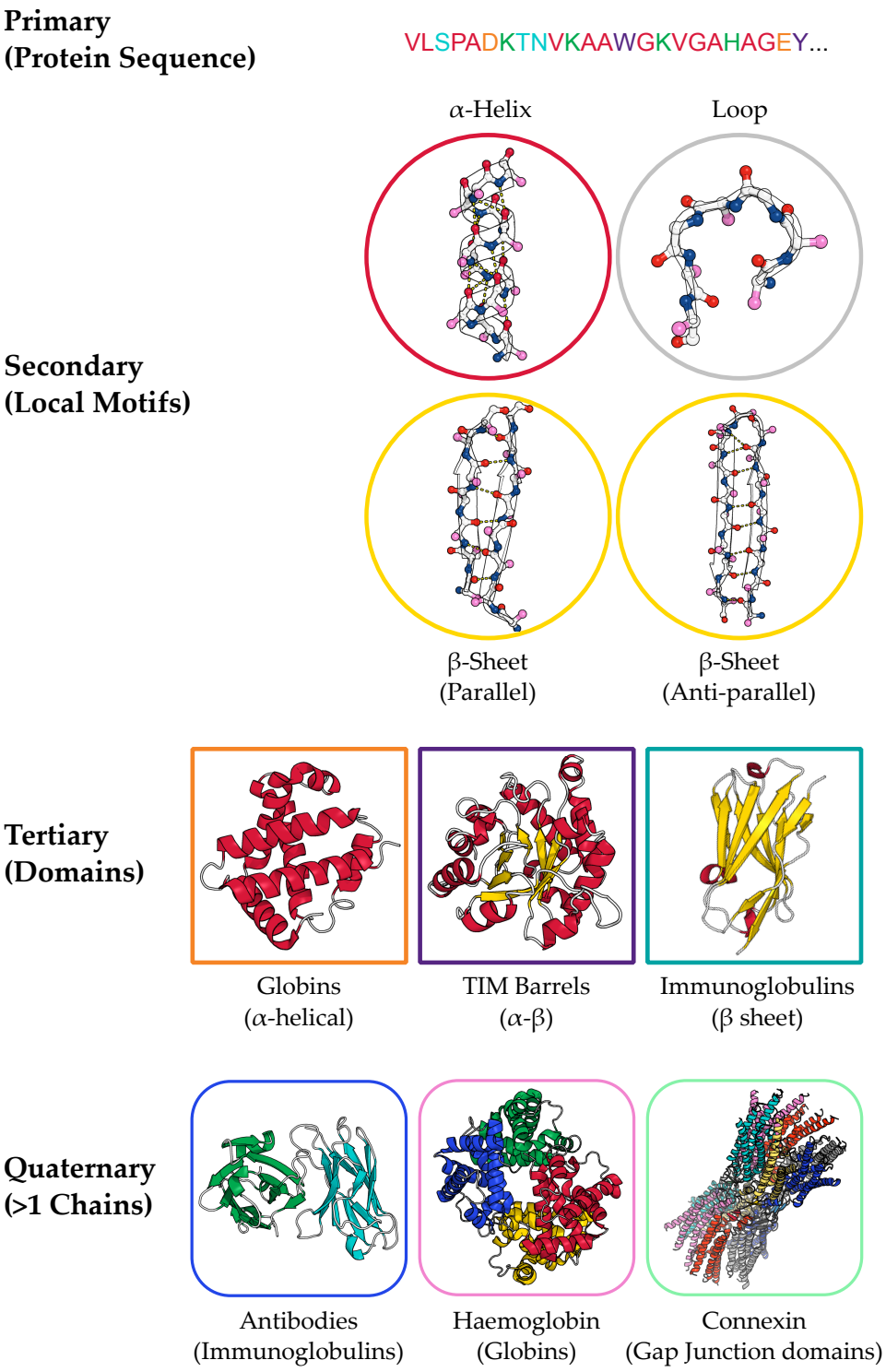


Figure A.2: Levels of protein structures. The protein sequence, or the 'primary' structure, folds into local components (secondary structures). Secondary structures are formed by hydrogen bonds between atoms in the protein backbone. The arrangement of secondary structures form a tertiary structure. Each protein chain has one or more folded units known as 'domains' in its tertiary structure. Multiple chains can assemble together to form quaternary structures.

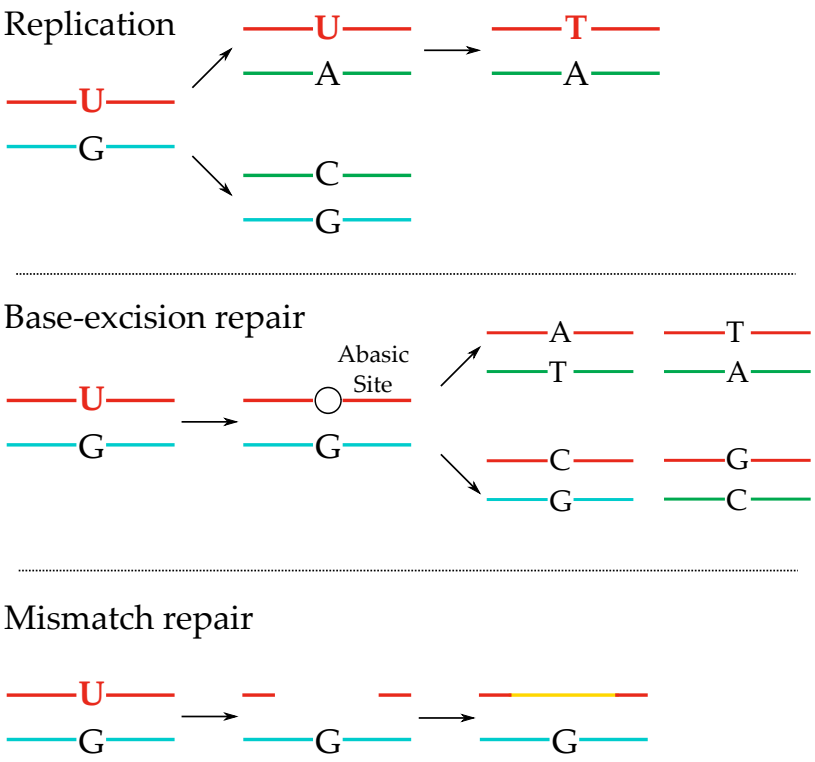


Figure A.3: DNA repair mechanisms following AID deamination. **A.** The mutant DNA strand can be replicated, thus propagating the mutation to daughter DNA strands. **B.** In base–excision repair, the DNA base is removed from the uracil by uracil DNA glycosylase. The abasic site is then filled with any of the four DNA nucleotides, and propagated for repair. **C.** In mismatch repair, a segment of DNA is excised and replaced by an error–prone DNA polymerase, namely DNA polymerase η.

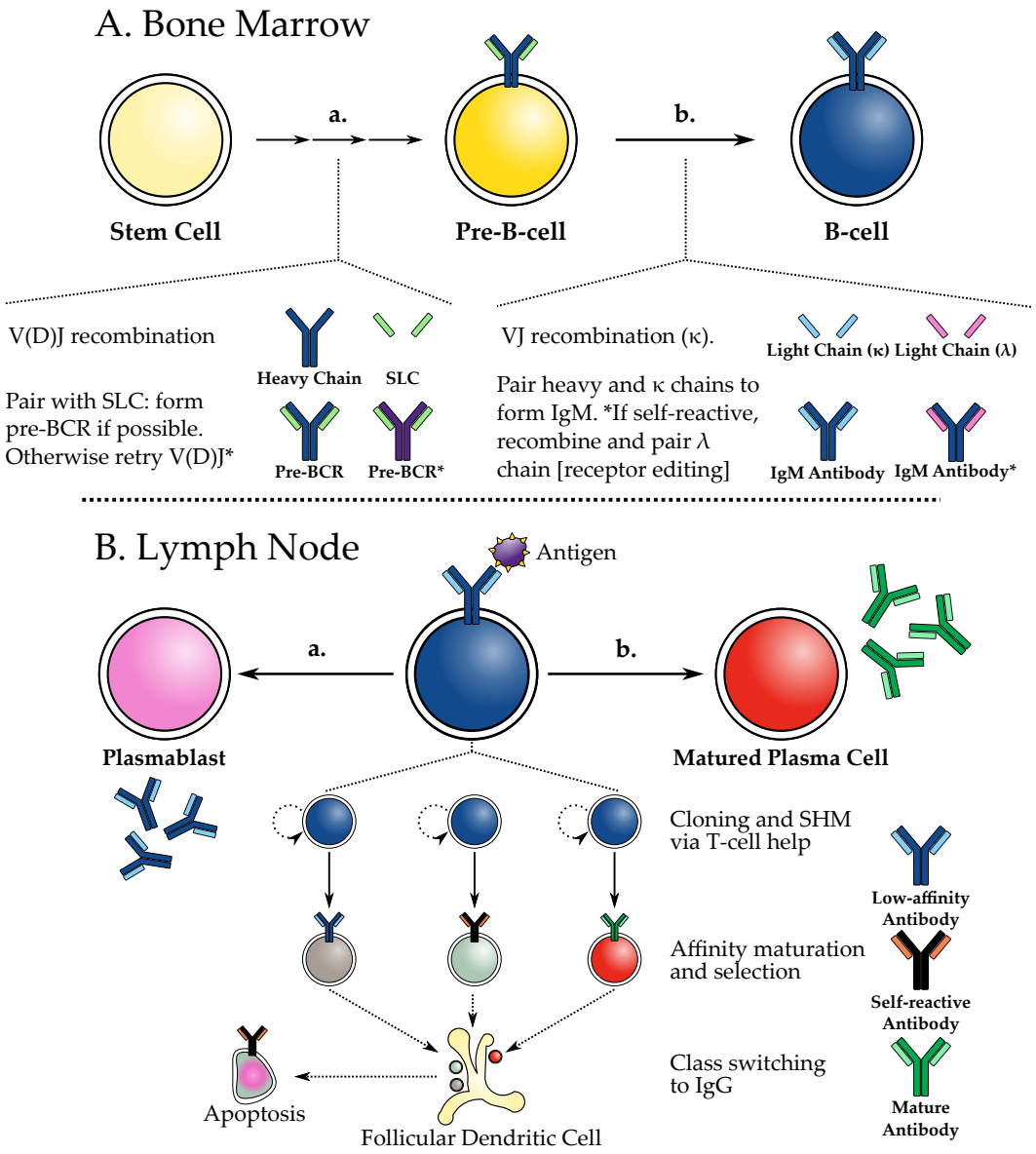


Figure A.4: B–cell maturation. **Aa.** In the bone marrow, a stem cell rearranges heavy chain gene segments via V(D)J recombination (Section 1.3.2.1). The newly–formed heavy chain pairs with the SLC to form the pre–BCR, and thus the pre–B–cell. **Ab.** The pre–BCR signals for light chain VJ recombination to form a light chain. The heavy and light chains pair to form an IgM antibody, and thus, a B–cell. If the IgM is self–reactive, VJ recombination is retried (*i.e.* receptor editing). **Ba.** An antigen–activated B–cell can clone and form plasmablasts, which secrete early–stage, low–affinity antibodies. **Bb.** Antigen–bound B–cells can migrate to the B–cell follicle to form GCs and proliferate rapdidly. The B–cell clones then mature the affinities of their antibodies SHM (Section 1.3.2.3). Clones are iteratively selected to favour the survival of a B–cell clone producing a high–affinity antibody. The surviving clone then differentiates into plasma cells or memory B–cells. SLC: surrogate light chain; BCR: B–cell receptor; SHM: somatic hypermutation; GC: germinal centre. (Janeway *et al.*, 2001; Nemazee, 2006; De Silva and Klein, 2015)

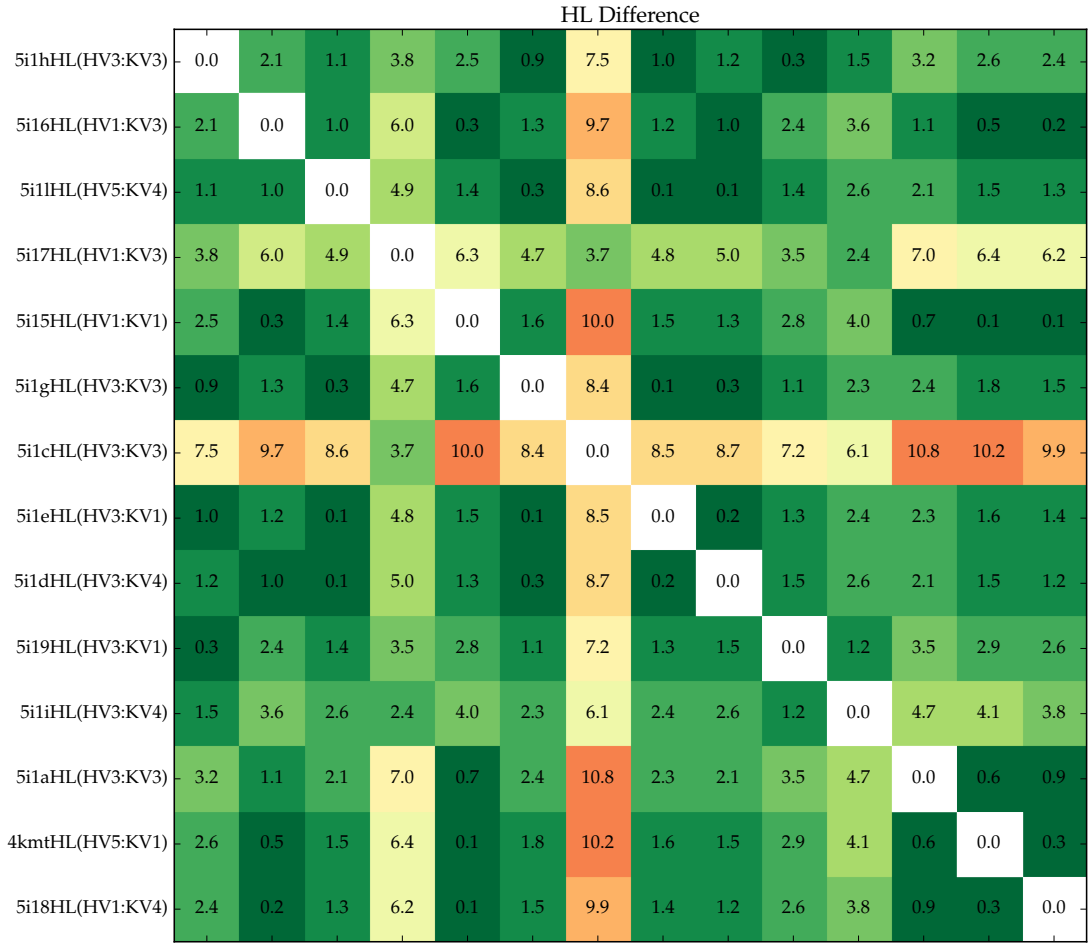


Figure A.5: Pairwise absolute difference of the HL angle between antibodies in the Teplyakov set. 5i17:HL shows the second largest deviation in the HL angle in comparison to other antibodies, yet the largest difference in the HL angle was observed in 5i1c:HL, which as a kinked CDRH3 loop. Other ABangles showed no difference (*e.g.* HC1), or were not correlated to a kinked or extended conformation.

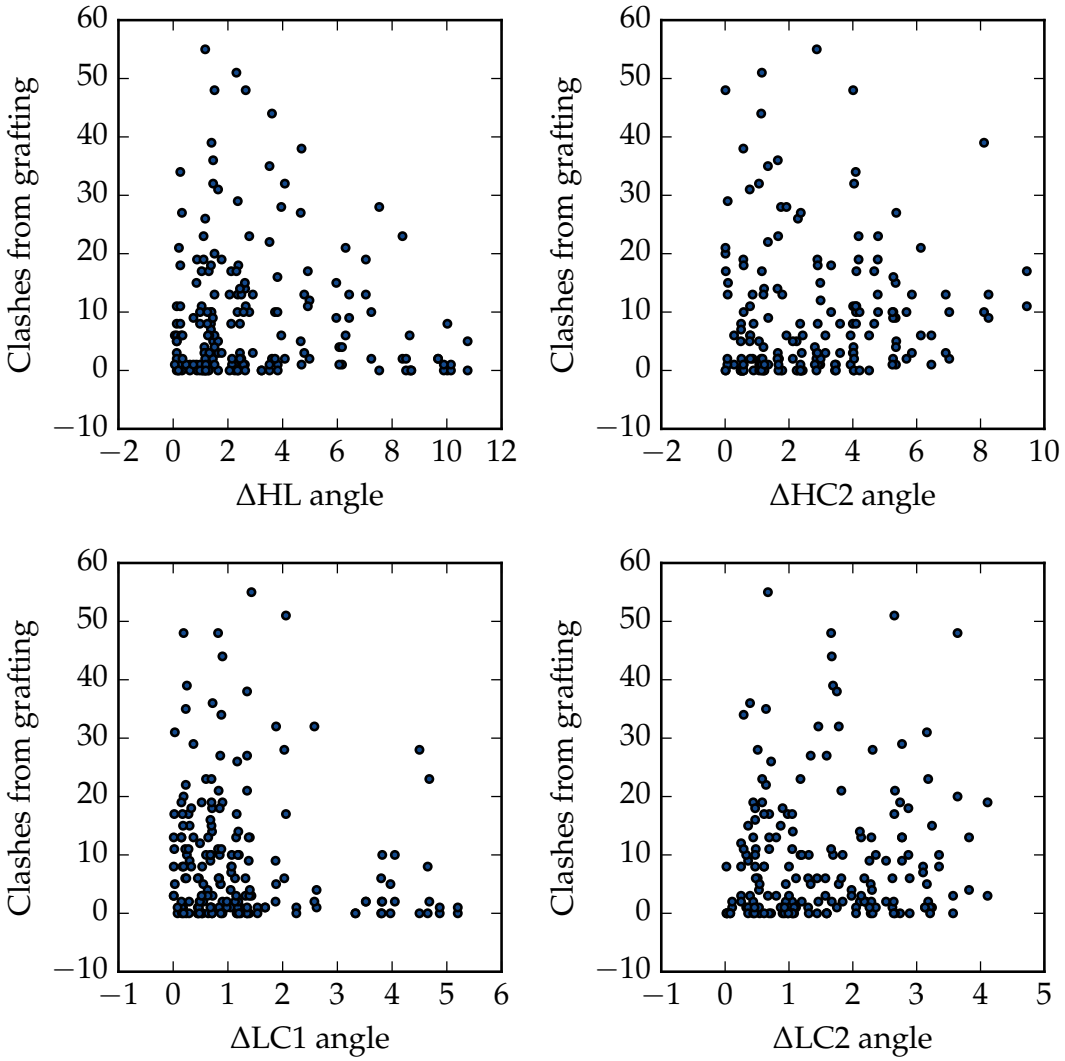


Figure A.6: There is no relationship between the difference in the ABangle parameters (*e.g.*Δ *HL*) and the number of clashes from grafting. Orientation differences were calculated between the 'source' antibody (*i.e.* antibody donating the CDRH3 loop) and the 'target' antibody (*i.e.* antibody framework for grafting the loop).

If the statistics are boring, then you've got the wrong numbers.

— Edward Tufte

B

Table B.1: Atom types used for building CAPTAIN.

Residue	Atom							
ALA	Backbone	CB						
ARG	Backbone	CB	CG	CD	NE	CZ	NH1	NH2	OXT
ASN	Backbone	CB	CG	OD1	ND2			
ASP	Backbone	CB	CG	OD1	OD2			
CYS	Backbone	CB	SG					
GLN	Backbone	CB	CG	CD	OE1	NE2	OXT	
GLY	Backbone	OXT						
GLU	Backbone	CB	CG	CD	OE1	OE2	OXT	
HIS	Backbone	CB	CG	CE1	CD2	ND1	NE2	
ILE	Backbone	CB	CG1	CG2	OXT			
LEU	Backone	CB	CG	CD1	CD2	OXT		
LYS	Backbone	CB	CG	CD	CE	NZ	C	OXT
MET	Backbone	CB	CG	SD	CE	C		
PHE	Backbone	CB	CG	CD1	CD2	CE1	CE2	CZ	C
PRO	Backbone	CB	CG	CD				
SER	Backbone	CB	OG					
THR	Backbone	CB	OG1	CG2	C			
TRP	Backbone	CB	CG	CD1	CD2	CE2	CE3	CZ2	CZ3
	CH2	NE1						
TYR	Backbone	CB	CG	CD1	CD2	CE1	CE2	CZ	OH
	OXT							
VAL	Backbone	CB	CG1	CG2				
HYP	Backbone	CB	CG	CD	OD1			
PCA	Backbone	CB						
SEP	Backbone	CB	OG	O1P	O2P	O3P	C	P	
TPO	Backbone	CB	OG1	CG2	O1P	O3P	C	P	

'Backbone' refers to the backbone atoms, N, CA, C, O, which is common to every residue type. HYP: hydroxyproline; PCA: pyroglutamate; SEP: phosphoserine, TPO: phosphothreonine.

Table B.2: PDB codes of antibody–antigen complexes in the affinity prediction test set.

1bj1	1bvk	1dqj	1dzb†	1eo8	1f90†	1i8k†	1kxq†	1kxt
1mlc	1nma†	1op9	1p2c	1p4b†	1pz5	1ri8†	1tet†	1tzh
1tzi	1uwx†	1vfb†	1yy9†	1za3	1zmy†	1zv5†	2a6i†	2eh8†
2fjg	2fjh	2fx7†	2fx8†	2fx9†	2hfg†	2hkf†	2hrp†	2j4w†
2j5l†	2nz9†	2oqj†	2or9†	2p42	2p43	2p44	2p49	2p4a
2qhr†	2qr0	2r0k	2r0z†	2vir	2vis	2wub†	2wuc†	2xra
2xtj†	2zpk†	3a67	3a6b	3a6c	3b2u†	3b2v†	3bdy	3be1†
3bgf	3bky†	3c09†	3e8u†	3eoa	3eob	3eyf†	3eys†	3eyu†
3g5v†	3g5y†	3ggw†	3ghe†	3hae†	3hi1	3ifl†	3ifo†	3ifp†
3iu3	3k2u	3l5w	3l5y†	3ma9	3mac†	3o6l†	3o6m†	3qsk†

†: PDB codes that were not processed by Gromacs, as described in Section 2.2.3.4.

Table B.3: Sixteen structures from (Teplyakov *et al.*, 2016).

PDB Code	Fab chains	V_H Germline	V_L Germline
5i19	HL	IGHV3–23	IGKV1–39
5i1a	HL, BA+	IGHV3–23	IGKV3–11
5i1c	HL	IGHV3–23	IGKV3–20
5i1d	HL, BA*	IGHV3–23	IGKV4–1
5i1e	HL	IGHV3–53	IGKV1–39
5i1g	HL	IGHV3–53	IGKV3–11
5i1h	HL	IGHV3–53	IGKV3–20
5i1i	HL	IGHV3–53	IGKV4–1
4kmt	HL	IGHV5–51	IGKV1–39
5i1j	HL*	IGHV5–51	IGKV3–11
5i1k	HL*	IGHV5–51	IGKV3–20
5i1l	HL, BA+	IGHV5–51	IGKV4–1
5i15	HL	IGHV1–69	IGKV1–39
5i16	HL, BA+	IGHV1–69	IGKV3–11
5i17	HL, BA*	IGHV1–69	IGKV3–20
5i18	HL	IGHV1–69	IGKV4–1

*: structure removed due to missing coordinates; +: structure removed due to a duplicate structure in the same PDB file.

Table B.4: Atom types used for calculating χ angles.

Angle type	Atom types	Residue types
χ_1	N, Cα, Cβ, Cγ	Arg, Asn, Asp, Gln, Glu His, Ile, Leu, Lys, Met Phe, Pro, Trp, Tyr
	N, Cα, Cβ, Sγ	Cys
	N, Cα, Cβ, Oγ	Ser, Thr
χ_2	Cα, Cβ, Cγ, Cδ	Arg, Gln, Glu, Ile, Leu Lys, Phe, Trp, Tyr
	Cα, Cβ, Cγ, Oδ	Asn, Asp
	Cα, Cβ, Cγ, Nδ	His
	Cα, Cβ, Cγ, Sδ	Met
χ_3	Cα, Cβ, Cγ, Nϵ	Arg
	Cα, Cβ, Cγ, Oϵ	Gln, Glu
	Cβ, Cγ, Cδ, Cϵ	Lys
	Cβ, Cγ, Sδ, Cϵ	Met
χ_4	Cγ, Cδ, Nϵ, Cζ	Arg
	Cγ, Cδ, Cϵ, Nζ	Lys

This [science of operations] constitutes the language through which alone we can adequately express the great facts of the natural world, and those unceasing changes of mutual relationship which, visibly or invisibly, consciously or unconsciously to our immediate physical perceptions, are interminably going on in the agencies of the creation we live amidst.

— Ada Lovelace

C

C.1 Receptor editing

Receptor editing refers to the process where light chains are VJ–recombined until a non–self–reactive antibody is produced (Nemazee, 2006). Pre–B–cells first recombine the κ chain's V and J gene segments to form an IgM antibody; if this fails, λ light chains are VJ–recombined until success (Appendix Figure A.4).

C.2 Quantifying affinity

The affinity of an antibody–antigen interaction can be measured in terms of the dissociation constant, K_D. The K_D represents the ratio of the unbound antibody and antigen with respect to the bound antibody–antigen complex. K_D is represented as a molar unit, and a lower K_D value corresponds to higher affinity. For example, an antibody–antigen complex with $K_D = 2.5nM$ has a stronger affinity than a complex with $K_D = 2.5\mu M$. Though less common, the affinity of the complex can also be represented as the free energy of binding, ΔG, where

$$\Delta G = -RT \ln K_D.$$

R is the ideal gas constant (–8.314 $\mathrm{J{\cdot}K^{-1}{\cdot}mol^{-1}}$), and T is the temperature, *e.g.* room temperature, which is $298K$. ΔG is measured in joules; converting to calories uses the conversion factor of $1\mathrm{J} = 0.239\mathrm{cal}$.

Differences in affinity are often represented as $\Delta\Delta G$, or the ratio of the wild–type and mutant antibodies' K_D values. For reference, a difference in 1kcal at $298K$ corresponds to a difference of 4184.1J, which is a 5.4–fold change in K_D;

$$\begin{aligned} \Delta\Delta G &= \Delta G_{\text{Mutant}} - \Delta G_{\text{Wild–type}} \\ 4184.1J &= -RT \ln K_{D,\text{Mutant}} - \left(-RT \ln K_{D,\text{Wild–type}}\right) \\ \frac{4184.1J}{-RT} &= \ln\left(\frac{K_{D,\text{Mutant}}}{K_{D,\text{Wild–type}}}\right) \\ &\approx -5.4 \end{aligned}$$

C.3 Calculation of RMSD

The RMSD is a value that is calculated between two protein structures A and B. It represents the square root of the average distance between the atoms of A and the atoms of B. RMSD does not follow the rules of triangle inequality; for example, if two structures A and B have an RMSD of 2Å, and structures B and C have an RMSD of 1Å, this does not necessarily mean that A and C will have an RMSD of 3Å.

In order to calculate the RMSD between a structure A with respect to structure B, common atoms must first be superimposed. This is typically done by Kabsch's algorithm (Kabsch, 1978). Briefly, a set of 'equivalent' atoms is extracted from A and B, for instance, the Cα atoms from both A and B. For each Cα atom, there are three coordinates: the x, y, and z coordinates. This leads to an $n \times 3$ matrix where n is the number of common Cα atoms between A and B. The Cα coordinates from A are then superimposed onto the coordinates of B by finding the rotation and translation matrices that minimise the RMSD between A and B. The coordinates of A are then transformed to yield A'; thus, the RMSD between A' and B is

$$RMSD(A', B) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((a'_{ix} - b_{ix})^2 + (a'_{iy} - b_{iy})^2 + (a'_{iz} - b_{iz})^2)}$$