Structure-Aware Tools for the Development of Therapeutic Antibodies from Natural Immunoglobulins

Matthew I. J. Raybould

New College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2020

Acknowledgements

This thesis would not have been possible without the immeasurable support of my supervisor, Professor Charlotte Deane. Her sage advice and mentorship have taught me to be a clearer thinker, a better scientist, and a more effective communicator. I will be forever grateful for the diverse research experiences I enjoyed as a DPhil student in her group. My gratitude extends to Dr. Claire Marks, for her invaluable help throughout my DPhil (particularly with lunchtime crosswords!). Further thanks must go to my fantastic group of industrial collaborators for their supervision and additional funding throughout the project: Dr. Alan Lewis (GSK), Dr. Jiye Shi (UCB), Dr. Bruck Taddese (AstraZeneca), and Dr. Alexander Bujotzek (Roche); especially to Bruck and Alex for allowing me to experience industrial life at their respective companies. I am also grateful to the EPSRC, MRC, and the SABS CDT for supplying the *core* funding that kept the computers running — despite my best attempts to crash them.

I am extremely appreciative of Francesco Fiorentino and Anka Lucic, members of my SABS CDT cohort, for their friendship from the very first day of Postgraduate life. Being a member of OPIG for four years was a wonderful experience; thanks to all members of the group for the banter and camaraderie that alleviated even the most frustrating periods of the project. Special thanks go to Aleksandr Kovaltsuk for his patience and councel whenever I needed a sounding board for my ideas. Aleks and his girlfriend, Ruth Jones, have been a constant source of bonhomie and support, and were amazing housemates (I still miss those weekend pancakes...).

2020 has been a difficult year, with the pandemic leading to lockdowns and sustained lifestyle restrictions from March onwards. I am deeply grateful to my girlfriend, Lexi Miller, who — though we only met in January — stuck by me and gave me the love and encouragement I needed to finish my DPhil on time and in good spirits.

Above all, I would like to thank my parents, who have worked tirelessly and sacrificed much to ensure I have always had the best opportunities in life. My debt to you is immense and I wouldn't be where I am today without your enduring love and guidance.

Work Ownership

The use of the first personal plural in this thesis is stylistic — it can be assumed that I performed all the work described herein, unless otherwise stated. Where work was performed by other individuals, they have been credited by name.

COVID-19 Statement

In Chapter 5, we had originally intended to benchmark a tiered docking protocol that could identify genetically-diverse natural antibodies with similar binding site topologies that are complementary to the same predefined antigen epitope. This would have required travelling to Roche's lab in Switzerland to analyse IP-sensitive experimental validation data, which was not feasible due to the COVID-19 pandemic. A description of our intentions for this work are outlined in the Future Work section.

Instead, building off our expertise in curating Thera-SAbDab, we built the Coronavirus Antibody Database (CoV-AbDab) to document the molecular properties of all experimentally-confirmed coronavirus binders. We also collaborated with new academic and industrial partners, using the antibody sequences in CoV-AbDab as probes to functionally characterise convergent clonotypes from SARS-CoV-2 response repertoires; this was the first study to report strongly convergent SARS-CoV-2-specific antibody responses. As this work is of direct relevance to the theme of this thesis, we have included a detailed description in Chapter 5.

Thesis Abstract

In this thesis, we establish the foundations for structure-based antibody drug discovery from natural immunoglobulin sequences. This is achieved through two novel software packages: 'Repertoire Structural Profiling' (RSP), as a means of generating structurally-diverse virtual screening libraries, and the 'Therapeutic Antibody Profiler' (TAP), for rapid structure-aware developability assessment of candidates in early-stage antibody drug discovery. Our approaches are orthogonal to existing work in these fields and yield new insights into the diversity of the adaptive immune system and the physicochemical characteristics of therapeutic antibodies. We also describe a new database (Thera-SAbDab), which provides the increased therapeutic antibody sequence and structural data necessary to benchmark RSP and TAP, and to facilitate future investigations in the area. Finally, we report our work in the COVID-19 response effort, which, through a novel database of thousands of coronavirus antibody-binders (CoV-AbDab), has already contributed functional annotations to SARS-CoV-2 immune response repertoires.

This DPhil led to the following research outputs:

Publications

Matthew IJ Raybould, Claire Marks, Konrad Krawczyk, Bruck Taddese, Jaroslaw Nowak, Alan P Lewis, Alexander Bujotzek, Jiye Shi and Charlotte M Deane (2019) Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci. USA.* 116(10):4025-4030.

Matthew IJ Raybould[†], Wing Ki Wong[†] and Charlotte M Deane (2019) Antibody-antigen complex modelling in the era of immunoglobulin repertoire sequencing. *Mol. Syst. Des. Eng.* 4:679-688. ([†]Joint authorship)

Konrad Krawczyk, **Matthew IJ Raybould**, Aleksandr Kovaltsuk and Charlotte M Deane (2019) Looking for therapeutic antibodies in Next-Generation Sequencing repositories. *mAbs.* 11(7):1197-1205.

Matthew IJ Raybould, Claire Marks, Alan P Lewis, Jiye Shi, Alexander Bujotzek, Bruck Taddese and Charlotte M Deane (2020) Thera-SAbDab: the Therapeutic Structural Antibody Database. *Nucleic Acids Res.* 48(D1):D383-D388.

Aleksandr Kovaltsuk, **Matthew IJ Raybould**, Wing Ki Wong, Claire Marks, Sebastian Kelm, James Snowden, Johannes Trück and Charlotte M Deane (2020) Structural diversity of B-cell receptor repertoires along the B-cell differentiation axis in humans and mice. *PLoS Comput. Biol.* 16(2): e1007636.

Matthew IJ Raybould, Aleksandr Kovaltsuk, Claire Marks and Charlotte M Deane (2020) CoV-AbDab: the Coronavirus Antibody Database. *Bioinformatics.* doi: 10.1093/bioinformatics/btaa739.
Jacob D Galson, Sebastian Schaetzle, Rachael JM Bashford-Rogers, **Matthew IJ Raybould**, Aleksandr Kovaltsuk, Gavin J Kilpatrick, Ralph Minter, Donna K Finch, Jorge Dias, Louisa James, Gavin Thomas, Wing-Yiu Jason Lee, Jason Betley, Olivia Cavlan, Alex Leech, Charlotte M Deane, Joan Seoane, Carlos Caldas, Dan Pennington, Paul Pfeffer and Jane Osbourn (2020) Deep sequencing of B cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures. *Front. Immunol.* doi: 10.3389/fimmu.2020.605170

Preprints

Matthew IJ Raybould, Claire Marks, Aleksandr Kovaltsuk, Alan P. Lewis, Jiye Shi and Charlotte M. Deane (2020) Evidence of antibody repertoire functional convergence through 'Public Baseline' and 'Shared Response' structures. *bioRxiv.* doi: 10.1101/2020.03.17.993444.

Invited Articles

Matthew IJ Raybould and Charlotte M Deane (2019) Structural information to aid *in silico* therapeutic antibody design from Next-Generation Sequencing repertoires. *Am. Pharm. Rev.* 22(5):28-33.

Book Chapters

Matthew IJ Raybould and Charlotte M Deane (2020) The Therapeutic Antibody Profiler for Computational Developability Assessment in Houen, G. (ed.) Methods in Molecular Biology: Therapeutic Antibodies. Humana Press. [In Press]

Contents

Chapter 1

Introduction

1.1 Thesis Summary

In this thesis, we establish the foundations for structure-based antibody drug discovery from immunoglobulin gene sequencing (Ig-seq) samples of B-cell receptor (BCR) repertoires. We begin by collating all existing sequence and structural data on therapeutic antibodies into a database that enables the research described in this thesis and facilitates other investigations in the field ('Thera-SAbDab', Chapter 2). Based on these datapoints, we found high sequence identities between existing therapeutic antibodies and natural antibody/BCR sequences, indicating that Ig-seq datasets should represent a fertile pool in which to search for novel therapeutically-exploitable molecules (Chapter 2). By clustering repertoire sequences based on their predicted structural templates, we convert Ig-seq datasets into structurally-representative homology models ('Antibody Model Libraries', AMLs) that can be used as a diverse basis set for virtual screening ('Repertoire Structural Profiling', Chapter 3). We also derive five computational developability guidelines through comparisons between the physicochemical properties of therapeutic antibodies and natural antibodies (the 'Therapeutic Antibody Profiler', Chapter 4). These can be applied within, or independently of, this pipeline to rapidly highlight problematic candidates during early-stage development.

We conclude by describing our research in response to the SARS-CoV-2 pandemic, which yielded a novel database (CoV-AbDab) that we harnessed to add functional annotations to the BCR repertoires of patients infected with SARS-CoV-2 (Chapter 5). Analysis enabled by CoV-AbDab highlights strong SARS-CoV-2-specific antibody responses across UK COVID-19 patients, identifies potential pitfalls for SARS-CoV-2 antibody test diagnostics, and adds supporting evidence that our AMLs can capture therapeutically-relevant binding site topologies. In time, the CoV-AbDab molecular

probes will be used to profile COVID-19 vaccine-induced immune responses and to inform decisions in the development of anti-SARS-CoV-2 therapeutic antibodies.

1.2 Chapter Abstract

This chapter will cover the essential background knowledge underpinning the research described in this thesis. It starts with a brief overview of antibodies (also known as immunoglobulins), a class of protein native to jawed vertebrates and crucial to the adaptive immune response. We describe their sequence and structural properties, role in the immune system, and how biotechnology is currently applied to sample the antibody diversity expressed by an individual at a point in time. This leads into a section describing how the Pharmaceutical industry, since the late twentieth-century, have developed monoclonal antibodies, and related modular formats, into drugs. We then cover the two — often competing — factors in therapeutic antibody design: target complementarity and developability. Finally, we propose a novel structure-aware paradigm for therapeutic antibody discovery directly from sequencing samples of natural immunoglobulin repertoires. The tools described in this thesis can be inserted into the framework to bring our proposed pipeline closer to reality. We conclude with a high-level summary of each of the subsequent chapters.

This chapter contains reproduced material from the following review:
Raybould, M.I.J.[†], Wong, W.-K.[†], Deane, C.M. (2019) Antibody-Antigen Complex Modelling in the Era of Immunoglobulin Gene Sequencing. *Mol. Syst. Des. Eng.* 4:679-688. [1] [†]Joint Authorship

1.3 Antibodies

1.3.1 The Role of Antibodies in the Adaptive Immune System

Antibodies are B-cell-encoded globular proteins that help initiate the adaptive immune response in vertebrates by recognising pathogenic invaders and recruiting other components of the immune system to eliminate them. They achieve their function by being able to adopt a huge diversity of binding sites (see Section 1.3.2) that can be mutated to achieve high complementarity to almost any pathogen.

A circulating set of 'naïve' B-cells, each of which display a single antibody[1] (at this point membrane-bound to their parent B-cell, so usually termed a 'B-cell receptor', BCR) patrol the bloodstream for foreign entities. The diversity in this set of naïve BCRs is limited relative to theoretical binding site diversity ($\sim 10^{10}$ *vs.* $\sim 10^{16}+$ [3, 4]), and so initial pathogen binding is often only of low-moderate affinity. Once a BCR recognises a pathogen through a sufficiently complementary binding site, its parent B-cell migrates to the germinal centre of the lymph nodes. Here, intentional mutations are made to its DNA sequence to modify the amino acids in the BCR binding site (a process known as 'affinity maturation'). A chemical feedback signal selects for mutations that increase the affinity of the BCR for the pathogen, and so as the BCR 'matures', it becomes more 'specific' for its target. At this point, the BCRs can be secreted in their soluble (antibody) format in a range of different 'isotypes' with different effector functions (see Section 1.3.2). It can take several days-weeks for evolutionary pressure to optimise the antibodies such that they can completely neutralise the pathogen, accounting for the lag time between feeling unwell and beginning to recover. Once a pathogen has been neutralised, the B-cells that encode for the optimised antibodies are preserved in low concentration as long-lived 'memory B-cells'. Upon reinfection by the same pathogen, or one with a sufficiently similar binding site, these memory B-cells are reactivated and the optimised antibodies secreted. Vaccination harnesses this characteristic of the immune system, achieving immunological memory for a particular set of pathogen proteins that can be rapidly recalled upon reinfection.

1.3.2 Antibody Composition and Sequence Diversity

Native antibodies are Y-shaped homodimers, with each monomer possessing two chains, covalently bound by disulfide bridges, and labelled as either heavy (H) or light (L) according to their difference in molecular weight (Fig. 1.1). Each chain can be divided into certain key regions [5]. The Fragment crystallisable (Fc) region forms the stalk of the Y, whereas the Fragment antigen binding (Fab) region forms the arms. Each antibody isotype (IgA, IgD, IgE, IgG, or IgM) has a distinct Fc region, responsible for multimerisation or for recruiting additional components of the immune system through mechanisms such as opsonisation, neutralisation, agglutination or complement activation [6]. The Fab region is solely responsible for pathogen recognition and can be split up into a 'constant domain' and a more diverse 'variable

[1]This assumption is being questioned by recent 'single-cell' sequencing studies, although one antibody is usually dominant [2].

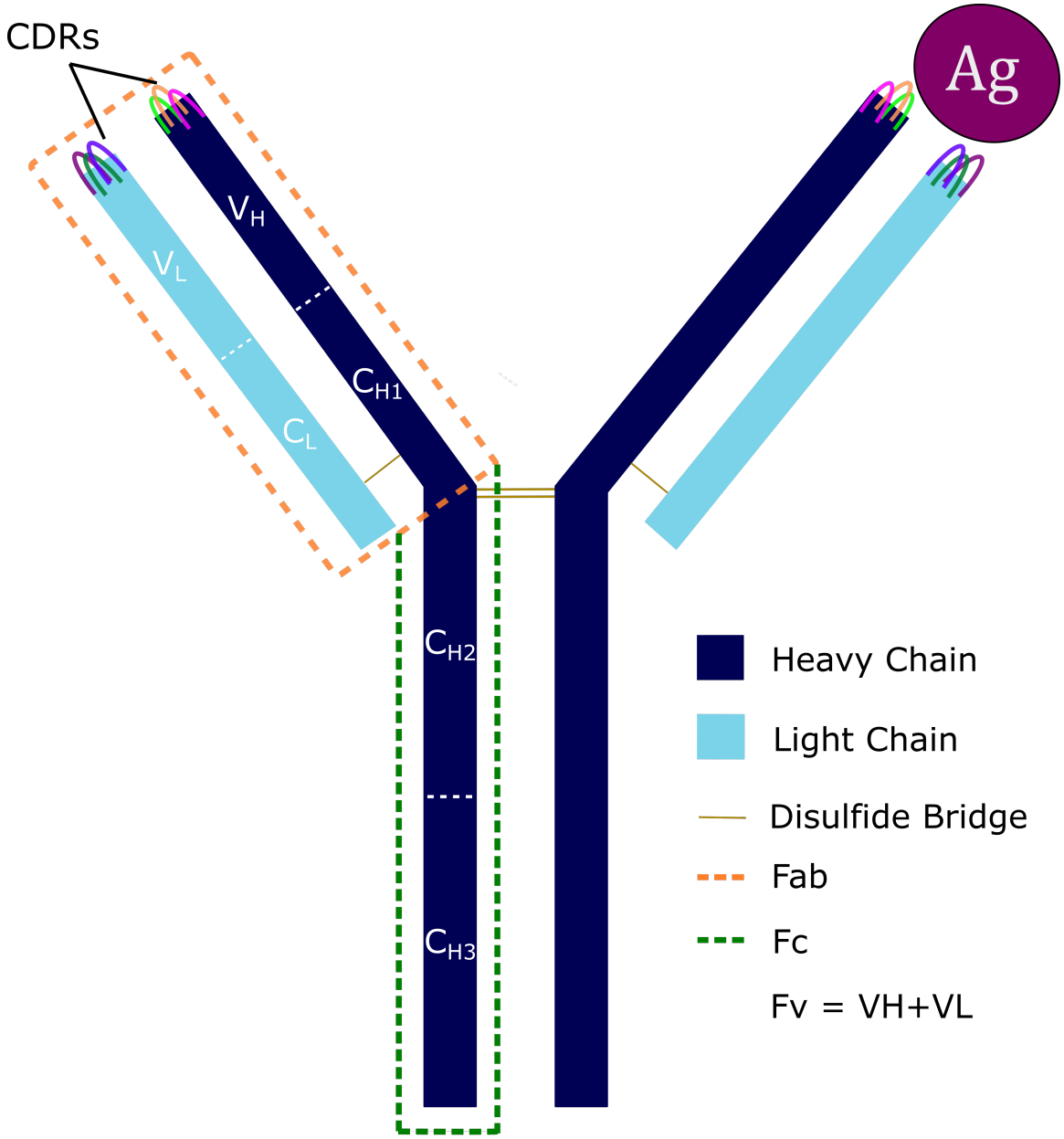


Figure 1.1: Antibody Composition. Antibodies consist of two heavy (dark blue) and light (light blue) chains, covalently held together by disulfide bonds (gold lines). The left-hand side monomer is labelled with its constituent domains. The Fab region comprises the heavy chain's variable (V_H) and first constant (C_{H1}) domains, as well as the light chain's variable (V_L) and constant (C_L) domains. The remaining heavy constant domains (C_{H2} and C_{H3}) make up the Fc region. Within each variable domain are three hypervariable loop regions collectively known as the Complementarity-Determining Regions (CDRs) that form most interactions to the antigen (Ag). The remainder of each Fv is known as the framework region.

(Fv) region' (also seen labelled as the V_H domain of the heavy chain and the V_L domain of the light chain, see Fig. 1.1).

The large diversity observed in the Fv sequence can largely be attributed to genetics. Rather than being entirely translated from a single gene, the antibody Fv is encoded across five distinct genes. Three (IGHV, IGHD, and IGHJ — VDJ for short) originate on chromosome 14 and recombine to form the VH region. VL is formed on recombination of a V and J gene, either from the 'kappa' (κ) loci on chromosome 2 (IGKV, IGKJ) or from the 'lambda' (λ) loci on chromosome 22 (IGLV, IGLJ). In humans, we currently know of 56 variable (V), 23 diversity (D), and 6 joining (J) heavy gene segments, 41 V and 5 J κ gene segments, and 33 V and 5 J λ gene segments[2] [7]. The first contribution towards antibody binding site diversity comes from the large number of potential distinct V(D)J recombinations within each chain. Additionally, before gene recombination can occur, each segment's DNA hairpin loop must be cleaved. Once cleaved, additional nucleotides can be inserted before the two genes are stitched together by the B-cell DNA repair machinery. This 'junctional diversity' adds further sequence and length variation to the antibody Fv region. On top of this, the process of somatic hypermutation during antibody maturation adds point nucleotide mutations throughout the Fv domain, which can lead to substantial divergence from the 'germline' (gene-encoded) amino acid sequence even in non-junctional regions. Finally, compatible VH and VL chains pair up to create many distinct antibodies; estimated at upwards of 10^{16} potential binding sites [4].

When the first antibody sequences were analysed, it became clear that amino acid variation was not uniform across the Fv region [8]. Instead, six 'complementarity-determining regions' (CDRs) were identified, three in the heavy chain and three in the light chain, that exhibit most of the sequence variation between antibodies (Fig. 1.1). This was further rationalised from a structural perspective, as these six CDRs form neighbouring loop structural elements while much of the remainder of the Fv domain (termed the 'framework' region) is locked in tight antiparallel beta sheets and so is less accommodating of amino acid mutation. Through solving antibody-antigen complex structures, it was confirmed that antibody residues in the CDR regions do indeed make almost all of the interactions to the cognate antigen [9].

Across the six CDR loops (CDRH1-3, CDRL1-3) it has been shown that, in many cases, the CDRH3 loop makes a disproportionately large contribution towards selectivity for any 'epitope' (region/set of functional groups on the antigen

[2]These numbers represent the 'functional' immunoglobulin gene loci and do not account for allelic variation. The numbers of productive genes are being continually revised upwards.

surface; the complementary set of functional groups on the antibody is known as the 'paratope') [10]. This is because the CDRH3 region falls at the junction of both IGHV-IGHD and IGHD-IGHJ, leading to a huge diversity in length, sequence, and 3D structure (see Section 1.3.4.5). CDRL3 also falls at a junction region on the light chain (IG[K/L]V-IG[K/L]J), but since it lacks the diversity gene, its length, sequence, and structural profile is much narrower than that of CDRH3. The CDR1 and CDR2 loops of both chains are germline-encoded on the V gene, further limiting their initial sequence and length diversity.

1.3.3 Antibody Numbering Schemes and Region Definitions

When comparing protein sequences of different lengths, it is typical to perform a 'sequence alignment', lining up the most conserved positions to identify and contextualise the length-variant regions. As antibody framework regions are sufficiently conserved, and almost all length variation occurs within the CDRs, several 'numbering schemes' can be applied directly to the antibody sequence to avoid the need for explicit alignment [8, 11–14]. All numbering schemes account for relative shifts by assigning insertions and deletions ('indels') to certain residue numbers. Within each numbering scheme, the boundaries between the framework and CDR regions are also debated — some define the CDRs as the regions of highest sequence variation [8], others as full structural loop elements [12, 13, 15], and still others attempt to find a unifying definition applicable to both VH and VL [11] or amongst classes of immune proteins [11, 14]. Several tools exist to apply numberings to antibody sequences [13, 16–18], including the ANARCI tool developed in our group [18]. ANARCI uses Hidden Markov Models built from reference pre-numbered IMGT germline sequences [11] to probabilistically assign residue numbers and indels to input antibody VH or VL chains.

Throughout this thesis, we use the International Immunogenetics Information System (IMGT) numbering scheme, as it is consistent between the heavy and light chains, and deletions/insertions are added to loops in a symmetrical fashion (Fig. 1.2A). IMGT numbering is thus equivalent to a combined sequence and structural alignment as residues with the same identifier are approximately structurally equivalent, enabling a rapid and meaningful comparison to be made between the amino acids seen at a given residue number. An example is shown in Fig. 1.2B, where a length 14 CDRH3 loop is compared to one of length 19. Using asymmetrical insertions, Adalimumab's central loop residues would be labelled as 111-111A-112, while Galiximab's would be labelled 111-111A-111B-111C-111D-111E-111F-112. Adalimumab's residue

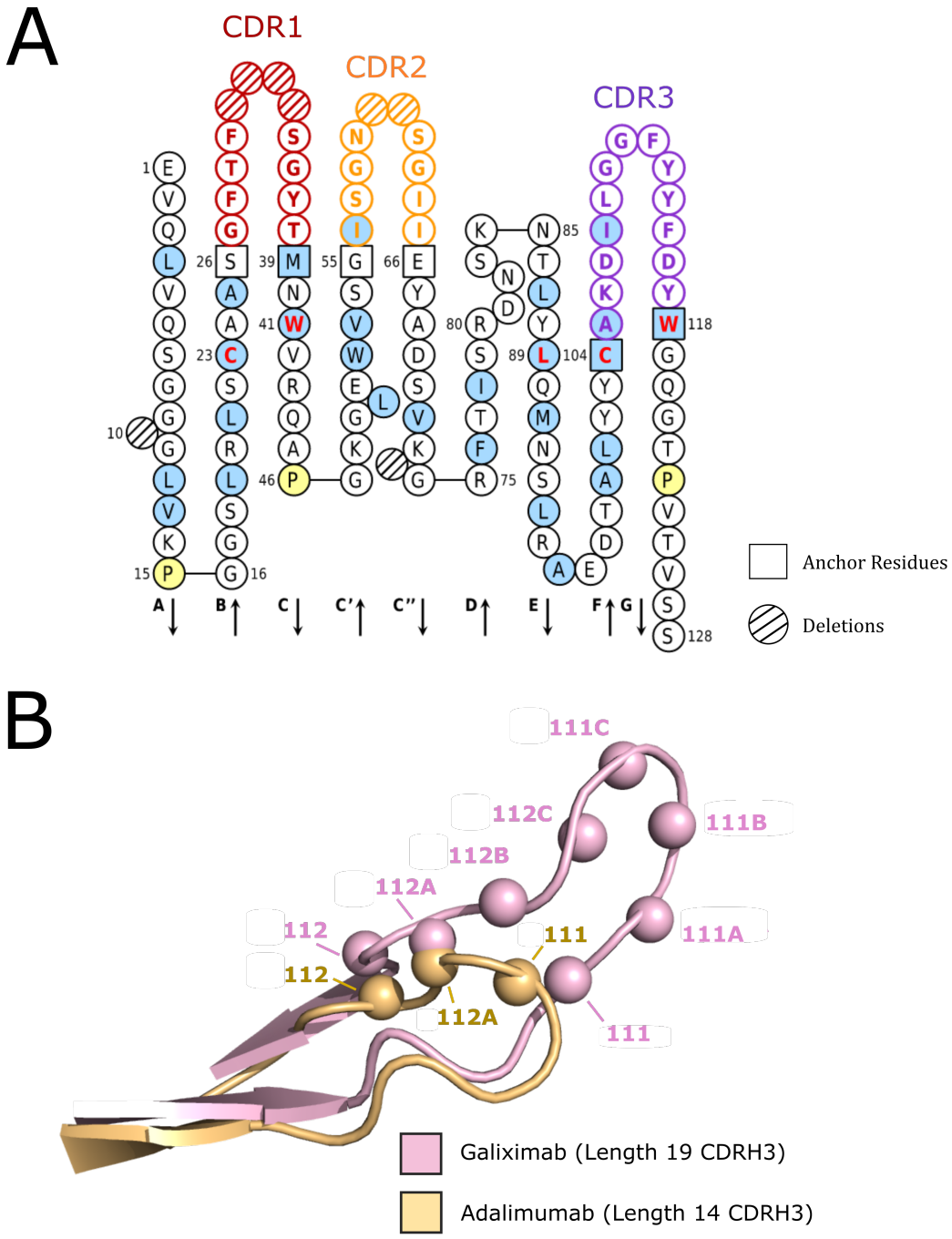


Figure 1.2: (A) A 'Collier-de-Perles' diagram showing a typical VH sequence in the IMGT numbering scheme with IMGT region definitions [11]. CDR1 lies between residues 27-38, CDR2 lie between residues 56-65, and CDR3 lies between residues 105-117. This numbering is consistent for VH and VL. (B) IMGT insertion codes in the center of the Galiximab and Adalimumab CDRH3s. The CDRH3 loops of the two antibodies are aligned and shown in cartoon representation, with backbone carbonyl carbon atoms highlighted by spheres. Residues with the same IMGT number are approximately topologically equivalent, while the longer loop is accommodated with symmetrical insertion codes.

'111A' is topologically equivalent to Galiximab's '111F', but these residue identifiers in no way reflect this. Instead, the IMGT numbering scheme assigns them both to identifier '112A'. Equally, shorter CDRH3/CDRL3 loops are symmetrically assigned deletions to ensure a consistent structural mapping. In terms of region definitions, we tend to use the IMGT definitions [11] if making explicit comparisons between heavy and light chain sequences, as the VH and VL CDR regions each lie between the same residue numbers. For structural modelling purposes (see Section 1.3.5), we use the North *et al.* set of definitions, since they better capture each CDR's full loop structure and so lead to improved framework grafting [15].

1.3.4 Antibody Structural Diversity and Structure Prediction

Some understanding of antibody structural diversity has been gained through analysis of the over 4,200 solved Fv structures in the Protein Data Bank [19], downloadable as a pre-numbered set from the Structural Antibody Database (SAbDab) [9]. However, solving an antibody structure remains a slow and laborious process, and so building upon our existing knowledge to predict as yet unsolved antibody structures is both necessary and informative. The following subsections describe the main sources of Fv structural diversity and how we can predict antibody structure from sequence.

1.3.4.1 The Framework Region

A large portion of the antibody framework region is locked in a tight anti-parallel beta sheet network that provides a structural scaffold for the CDR loops. This structural motif imposes tight restrictions on amino acid identity, as an incompatible residue (e.g. one with too bulky a side chain) would preclude folding leading to a non-functional antibody. However, one Fv structural property governed by the framework region is inherently variable — the relative orientation of the VH and VL domains. This orientation can be summarised in a set of six parameters (5 dihedral angles and one distance) and can significantly affect the 'bite angle' of the antibody paratope [20]. Statistical predictors based on framework amino acid sequence have been shown to regress to these parameters with relatively high accuracy [21].

1.3.4.2 The Canonical CDR Loops

Aligning the increasing number of solved antibody structures, researchers noticed that the CDRH1-2 and CDRL1-3 loops tended to adopt their own characteristic set

of backbone structural conformations [12, 15, 22–24]. These became known as the 'canonical forms', and the non-CDRH3 antibody loops were termed the 'canonical CDRs'. These canonical forms encompass between 81% and 98% of each CDR's currently-solved structural diversity and range from 1 to 12 clusters per CDR loop[3] (see Supporting Information of Wong *et al.*, 2018 [25]). Interestingly, the same canonical CDR loop structure can be adopted by sequences of different lengths [24, 25]. The SCALOP software, developed in our group in 2018, is able to rapidly and reliably classify loop canonical forms from sequence alone using 'Position-Specific Substitution Matrices' and a Random Forest machine learning architecture [25]. Position-Specific Substitution Matrices capture how often an amino acid can be substituted at a given position without changing the canonical form, based on inspection of the PDB. More refined predictions of canonical loop structure (not limited by clustering) can be made using homology modelling approaches such as FREAD [26, 27] (see Section 1.3.4.3).

1.3.4.3 CDRH3 Loop Structure

Owing to its extraordinary sequence diversity, attempts to derive similar canonical forms for the whole of the CDRH3 loop have been unsuccessful [15, 28]. CDRH3 loop structure is notoriously difficult to predict and has been the subject of much research over recent years [10]. Some approaches tackle the problem '*ab initio*' [29–35], a physics-based approach that seeks to convert the CDRH3 sequence into the corresponding structure with lowest free energy. *Ab initio* methods can yield very accurate representations, but require a very long run-time; for example, RosettaAntibody [35] takes ∼one hour over many Central Processing Units (CPUs) to model a typical antibody CDRH3 region, as it continually 'relaxes' the structure and re-refines its prediction using the KIC algorithm. A faster method, which relies on our existing knowledge of CDRH3 structures, is homology modelling [26, 27, 36–43]. These methods harness loop structure databases and a statistical framework to pick the best-fit structural template for a given sequence. For example, the FREAD software developed by Choi and Deane uses Environment-Specific Substitution Tables and anchor residue[4] graftability to rank potential CDRH3 structural templates for a given CDRH3 sequence and framework structure. Environment-Specific Substitution Tables consider, for a given loop, position, and set of backbone dihedral angles, how likely two amino acids are to interconvert without changing loop geometry. For most

[3]These clusters have changed over time with the addition of more structures to the PDB, and may eventually account for the currently unclusterable canonical loop structures.

[4]The anchor residues are the residues immediately before and after each CDR loop (see Fig. 1.2A).

lengths of CDRH3 loop (lengths 5 through 19), accuracy of homology modelling methods is now competitive with *ab initio* approaches and is set only to improve over time as more structural data becomes available [27, 44]. However, longer CDRH3 loops of 20+ residues have a huge potential conformational space which is poorly sampled in the PDB, and existing homology approaches can only generate low-confidence predictions. In this thesis, our perception of what constitutes a 'long' CDRH3 is thus tied to the lengths of loop for which there are inadequate structural samples in the PDB. Our notion of what constitutes a 'long CDRH3' has already changed as the PDB has grown and will likely change into the future.

Owing to speed constraints, homology modelling is currently the only practical way to incorporate structural analytics into high-throughput analysis pipelines. Alternative 'hybrid' approaches [10, 45–47], such as SPHINX [48], have been developed that merge homology modelling and *ab initio* optimisation, seeking to find a compromise between accuracy and speed. As of yet, while they significantly improve the accuracy of longer CDRH3 homology loop modelling, they remain too slow to run on thousands of input sequences.

1.3.4.4 Side Chain Conformations

Beyond backbone structural variability, each residue's side chain can also adopt distinct conformations, altering the surface chemical interaction profile. If modelling antibodies *ab initio*, all side chains must be assigned initial conformations and subsequently relaxed alongside the backbone structure (another contributor to long runtimes, as energetic convergence can be elusive). Conversely, homologous templates already come with a solved set of side chain coordinates. Opinions differ as to whether retaining the side chain conformations of matching target/template residues, or whether remodelling all side chains in the context of the new model gives the best performance[5]. Side chains can be modelled-in or remodelled using tools that combine an energy function with a rotamer library. Most tools were developed for general protein side chain modelling [49–52], while the PEARS tool is designed specifically to model antibody side chains [53]. PEARS uses Gaussian Mixture Models built from the PDB to assign the most probable of the three staggered-conformation rotamers to each freely rotatable side chain bond at each IMGT residue position.

[5]This can depend on whether the antibody is 'apo' (solved without a partner antigen) or in complex, as the free energy landscape is very likely to change as two proteins approach one another.

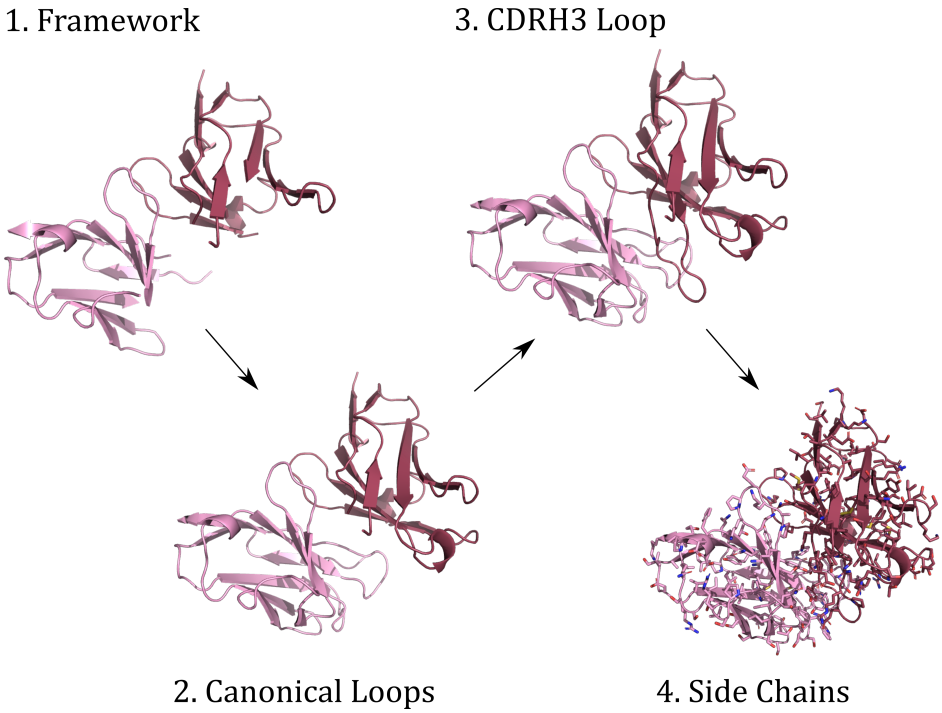


Figure 1.3: A typical antibody variable domain (Fv) modelling pipeline. First, a framework region is chosen for the target sequence. Secondly, the canonical loops are modelled onto the appropriate anchor residues. Thirdly, the CDRH3 loop is modelled in the context of the other CDRs. Finally, side chains are modelled-in or refined to yield an antibody Fv model in atomistic detail.

1.3.4.5 Modelling Variable Domain Structures

The above factors can be considered in a modular format to generate model structures of the entire antibody Fv domain to atomistic detail (Fig. 1.3). There are many software packages capable of performing this task [35, 44, 54–61]. Some tools are freely available to all users [44, 54–56], others are free to use under an academic license, but a paid subscription applies for commercial users [35], and still others require the purchase of either an academic or commercial license [57–61].

Most of these software packages reliably achieve close to sub-Ångström accuracy for the more sequence homogenous regions among antibodies. These include the four framework regions (FWR1-4) and the canonical CDR loops, and are rapidly solved by homology. Framework region templates are typically selected based solely on maximal sequence identity or similarity to the target. Either a single template antibody can be used for both chains, a separate template can be chosen for each chain, or templates can be chosen separately for each intra-chain framework region. Attention is then given to predicting a relative VH and VL domain orientation [20, 59, 62]. If all

templates come from the same parent antibody, then the parent's VH–VL orientation is usually assigned directly to the model structure, but if not then interface parameters must be assigned algorithmically [59]. In contrast to the FWRs, the CDRs are always considered separately, and any combination of loop length, canonical form, sequence similarity, dihedral compatability, and anchor residue distance is used to select the best template. The treatment of CDRH3 distinguishes different modelling tools, which use a variety of *ab initio*, homology, or hybrid algorithms (see Section 1.3.4.3). Side chains are usually refined after initial backbone structure prediction.

Throughout this thesis, we use the ABodyBuilder modelling software developed by Leem *et al.* [44]. Framework regions are selected by sequence identity, all loops are grafted onto the framework region by FREAD in the order L2-H2-L1-H1-L3-H3. Any loops that cannot be modelled by FREAD are either passed to Modeller [63], passed to Sphinx [48] (see Section 1.3.4.3), or the parent antibody is classified as unmodellable, depending on the use case. Target residues that match the template have their side chain conformations preserved while mismatched side chains are modelled-in using PEARS (see Section 1.3.4.4). On generating a model, ABodyBuilder returns a statistical estimate of regional model accuracy based on PDB benchmarking, which is helpful to guide future model refinement. ABodyBuilder has both the accuracy (shown by comparison to the AMA-II benchmarks [44]) and speed characteristics (30 seconds per model per CPU) to model the large number of antibodies required for a representative sample of the human antibody repertoire.

1.4 Sampling and Studying the Antibody Repertoire

1.4.1 Repertoire Sequencing Methods

The antibody/BCR repertoire is an enormously diverse and dynamic system. To obtain representative samples of the diversity of immunoglobulins being produced by an individual at any one time, high-throughput sequencing techniques are essential. Here, we briefly discuss the two main immunoglobulin gene sequencing (Ig-seq) platforms exploited to sample BCR repertoires — 'Next-Generation Sequencing' and 'Single-Cell Sequencing', how clinical blood-sampling experiments are typically designed, and the insights that can be obtained by analysing the resulting data.

1.4.1.1 Unpaired-chain ('Next-Generation') Sequencing

Next-Generation Sequencing (NGS) technologies, such as Illumina MiSeq and Roche 454, have the capacity to record millions of antibody chain sequences in a single experiment [64]. For example, Illumina MiSeq works by sequencing the complementary DNA (cDNA) of mRNA fragments from a large number of B-cells that have been pooled, reverse transcribed, and amplified. As a result, it samples the separate VH and/or VL chains being expressed in an individual at a given point in time, but loses the native Fv pairings of the chains as all B-cells are pooled together. Germline-specific primers are used to sequence ∼300 base pairs of either the heavy or light chain from both the start and the end of the sequence, reading ∼80 amino acid residues in each direction with high fidelity. The ends of the resulting reads are then computationally 'assembled' to deduce the entire sequence in the reading frame that most closely resembles a known germline.

The sequence data retrieved depends heavily on the nature of the primers used. For example, if researchers want to sort antibody sequences by isotype, they can chose appropriate primers that amplify enough of the CH1 region to capture this information. Data also depends heavily on the origin of isolated B-cells; peripheral blood is usually sampled as it is the least invasive source of patient B-cells, but immune response signals can be expected to be considerably weaker in peripheral blood than in the primary lymphoid organs. On top of sample variability, one should also consider sequence read accuracy. Sequencing error can come from a range of sources, for example through mutation during cDNA amplification or through sequence misreads or misalignments [65]. Illumina sequencing provides a 'Q-score' for each base in a read, which serves as a confidence metric for read quality. Unique Molecular Identifiers (UMIs) can also be used to assist in error-proofing read assembly as well as to identify over-amplification biases.

The major advantage of NGS sequencing is its ability to capture extremely deep (up to 10^8 [4]) samples of the VH or VL repertoire, although not capturing VH-VL pairings results in a loss of resolution equivalent to three CDR regions of every binding site. However, recently, single-cell RNA sequencing technologies have emerged that allow us to deduce entire binding site sequences [66–68].

1.4.1.2 Paired-chain ('Single-Cell') Sequencing

Over the past five years, a range of sequencing techniques have been developed that are able to capture genomic or transcriptomic data at the resolution of individual

cells [69]. Single-cell filtering leads to similar protocols as used in NGS sequencing, including molecular barcoding, amplification, and short-read sequencing followed by assembly. Amongst these new technologies are platforms such as 10X Genomics Chromium, which (with the appropriate bead/primer kit) can capture the IGHV/D/J and IG[K/L]V/J genes encoded by the individual B-cells present in a blood sample. Building on these advances are platforms such as Libra-seq and CelliGo, which have the potential to combine combine single-cell sequencing with high-throughput *in vitro* antigen specificity assessment (see Section 1.6.1) in a single pipeline [70, 71]. Freely available 'VDJ' (immune repertoire) single-cell sequencing data is now starting to emerge in a range of disease contexts [2, 70, 72–78] and exciting discoveries are already starting to be made, such as the fact that individual B-cells may be able to contemporaneously express different BCRs on their surface [2]. The current major limitations of VDJ single-cell sequencing are data availability, high expense, and that sampling is currently limited to around 10^4 B-cells, which may prove to be too small a number to be representative of an entire repertoire.

1.4.2 Latitudinal and Longitudinal Sequencing Regimens

There are two primary regimens for repertoire sequencing experiment design. Latitudinal studies focus on sequencing very large cohorts of volunteers at a single point in time. For example, they are used to sample the antibodies that typically exist in people's naïve repertoires when they are at a particular age [79], are apparently healthy [4, 80], or when they have a particular disease [81], chronic condition [82], or allergy [83]. Sampling many different individuals helps researchers to learn genuine immune repertoire features rather than personal gene expression preferences. In contrast, longitudinal studies focus on capturing immune system dynamics upon perturbation, and so usually focus on a narrower cohort of volunteers whose blood is taken at multiple time points. Experiments in this category include studying BCR repertoire changes in response to vaccination [73, 84], controlled disease exposure, or natural infection[6] [82]. Both regimens aim to identify commonalities in the antibody repertoires of different individuals, which, given the theoretical diversity of the immune repertoire, are statistically unlikely by chance and so should reflect underlying antigenic selection pressures.

[6]These studies usually occur when diseases are endemic to a population, so there is a high likelihood that a reasonable number of an initial cohort of healthy patients will become ill within a certain time frame.

1.4.3 Repertoire Sequence Databases

To unite these dispersed sequencing datasets into a single location with consistent annotation, several adaptive immune repertoire sequence databases have been developed [85–88]. Each repository has its own particular focus — for example, iReceptor seeks to capture the totality of freely available BCR and T-cell receptor (TCR)[7] sequence data [87], while the Observed Antibody Space (OAS) database focuses just on BCR/antibody Ig-seq studies with sufficient sequencing quality and annotates all sequences with potential liabilities and metadata [88]. As of September 2020, iReceptor contains $\sim$ 2.7Bn sequences, while OAS contains $\sim$1.65Bn (predominantly VH chain reads). The unifying features of all repertoire sequence database efforts are consistent formatting and free data accessibility, the two major obstacles that prevent impactful data analysis.

In this thesis, we use the OAS database as the source of Ig-seq data [88]. This is because it pre-filters datasets to only list sequences in which every CDR is resolved, pre-numbers every sequence in the IMGT numbering scheme, and supplies vital metadata such as antibody isotype, volunteer age, and disease state, allowing for the interrogation of specific sub-populations. A large portion of the Ig-seq datasets contain VH sequences only, as the established method of high-throughput sequencing (NGS sequencing, see Section 1.4.1.1) is unable to preserve native pairings, and the VH is prioritised as it contains the most diagnostic loop, CDRH3. Since we seek to recreate entire Fv binding sites, we have focussed instead on the smaller subset of unpaired sequencing studies that also supply VL data [83, 84, 89].

1.4.4 Repertoire Bioinformatic Analysis

Bioinformatic analysis of natural BCR repertoire data has the potential to resolve many important immunological questions. A new network of researchers (known as the 'Adaptive Immune Receptor Repertoire' [AIRR] community) has emerged, connecting the immunologists with the expertise to sequence antibody repertoires with the bioinformaticians with the statistical background and high-performance computing resources to study them [90].

AIRR community researchers, bolstered by increased sequence data accessibility and novel statistical approaches such as deep learning, have made significant strides forward that have improved our understanding of immunology [91]. To name just

[7] T-cell receptors are another crucial component of the adaptive immune response that recognise short linear antigen peptide fragments presented by the Major Histocompatibility Complex.

a few advances, far more is now known about the relatively small number of naïve antibodies able to initiate an immune response against a near endless array of possible pathogens [4, 80, 92, 93]. Improvements have also been made in our ability to distinguish the features of each species' antibodies [94–96], to track how the repertoire changes with age [79] or maturation state [96, 97], and to identify which similar antibodies tend to be raised by different individuals against the same pathogenic stimulus [3, 73, 84, 98].

Many advances in this latter category of investigation have relied on a clustering technique known as 'clonotyping'. Clonotyping harnesses the underlying genetics and biological selection of antibodies to group them into sets likely to have derived from similar ancestors. Numbered sequences are aligned to reference V and J genes (e.g. by ANARCI [18]) and the closest gene loci identified. Sequences with the same predicted V/J origins are then clustered by CDR3 sequence identity, at a threshold ranging from 80% [80] through to 100%[8] [4]. Clonotyping is usually performed only on the VH chain, assuming that most of the selection pressure in antigen recognition falls on the CDRH3 loop. It does not explicitly consider the structural component of complementarity, instead assuming that high sequence identity is a good proxy for structural similarity.

1.5 Therapeutic Antibodies

When the role of antibodies/BCRs to selectively identify and eliminate antigens became apparent, the Pharmaceutical industry began investing heavily into efforts to isolate and engineer soluble antibodies as therapeutic agents. Since 1990, over 600 therapeutic antibodies have been registered with the World Health Organisation [99]. A report for Fortune Business Insights in May 2019 valued the therapeutic antibody industry at around $123B and projected it to grow at a compound annual growth rate of 14% to reach a value of $350B in 2027[9]. This section first describes the dominant therapeutic format (the 'monoclonal antibody'), then related engineered formats, and finally walks through the 'three generations' of therapeutic antibody discovery.

[8]This is the most common way to perform clonotyping; occasionally D gene identity is incorporated into the VH clonotype definition, though accurate D gene assignment is challenging given the high levels of somatic hypermutation throughout the CDRH3 loop.

[9]See the report here: *https://www.fortunebusinessinsights.com/monoclonal-antibody-therapy-market-102734*, noting that the health/economic implications of the COVID-19 pandemic could significantly change these forecasts in either direction.

1.5.1 Monoclonal Antibodies (mAbs)

Monoclonal antibodies (mAbs) are equivalent to natural antibodies insofar as they are homodimeric complete immunoglobulins with two identical binding sites (Fig. 1.4). The first mAb, Muromonab, was approved in 1990[10] and was a murine antibody designed to prevent rejection in kidney transplantation and a host of other autoimmune conditions. MAbs, both in isolation and covalently linked to other drugs (antibody-drug conjugates, ADCs), are now in widespread use in the clinic [100, 101] for therapeutic use in a wide range of indications from cancers and viral infections to migraines and allergies [99]. To date (September 2020) there are 98 unique antibodies approved by the FDA or EU for therapeutic use, 25 of which were first approved in 2017-2019 alone, and 79 novel antibody drugs in the late stages (Phase-III+)[11] of clinical trials [101]. These biotherapeutics have a broad array of targets, from membrane proteins to cytokines and even amyloid peptide chains [102].

Therapeutic mAbs can be known by several different names, including a developmental name, a 'mab' name, and (if approved) a trade name. Throughout this thesis, therapeutic antibodies are consistently labelled by their 'mab' name, as the extended suffix has historically provided context on development methodology (*i.e.* whether the antibody is murine [-omab], chimeric [-ximab], humanised [-zumab], or fully human [-mumab]; see Section 1.5.3.1). It should be noted that this convention was terminated during the course of this DPhil and will not continue for future antibody therapies [103].

As the whole mAb format resembles that of naturally-secreted antibodies, it is expected to offer the highest probability of immune system tolerance, reducing immunogenicity (see Section 1.7.2) and achieving a long serum half-life necessary to stay within the 'therapeutic window'[12] between hospital visits. Preserving the Fc region also enables efficient activation of other components of the immune system against the bound target. The modular structure of mAbs means that their specificity for a single target can be coupled with the cytotoxicity of a small molecule through an engineered linker in the constant region ('antibody-drug conjugates'), or with a diagnostic radioactive element [104].

[10] It was later discontinued due to a poor safety profile.

[11] The core stages of therapeutic assessment are Preregistration, Phase-I, Phase-II, Phase-III, Preapproval, and Approval/Phase-IV. Phase-I testing is primarily to assess safety and dosage using healthy volunteers, then progressively larger cohorts of patients are recruited to access drug efficacy during Phases II and III. Phase-IV monitors the performance of the drug in the clinic after approval.

[12] The 'therapeutic window' refers to the concentration span within which a drug remains therapeutically active. At concentrations above the therapeutic window a drug causes adverse effects, while at lower concentrations it offers little to no therapeutic benefit.

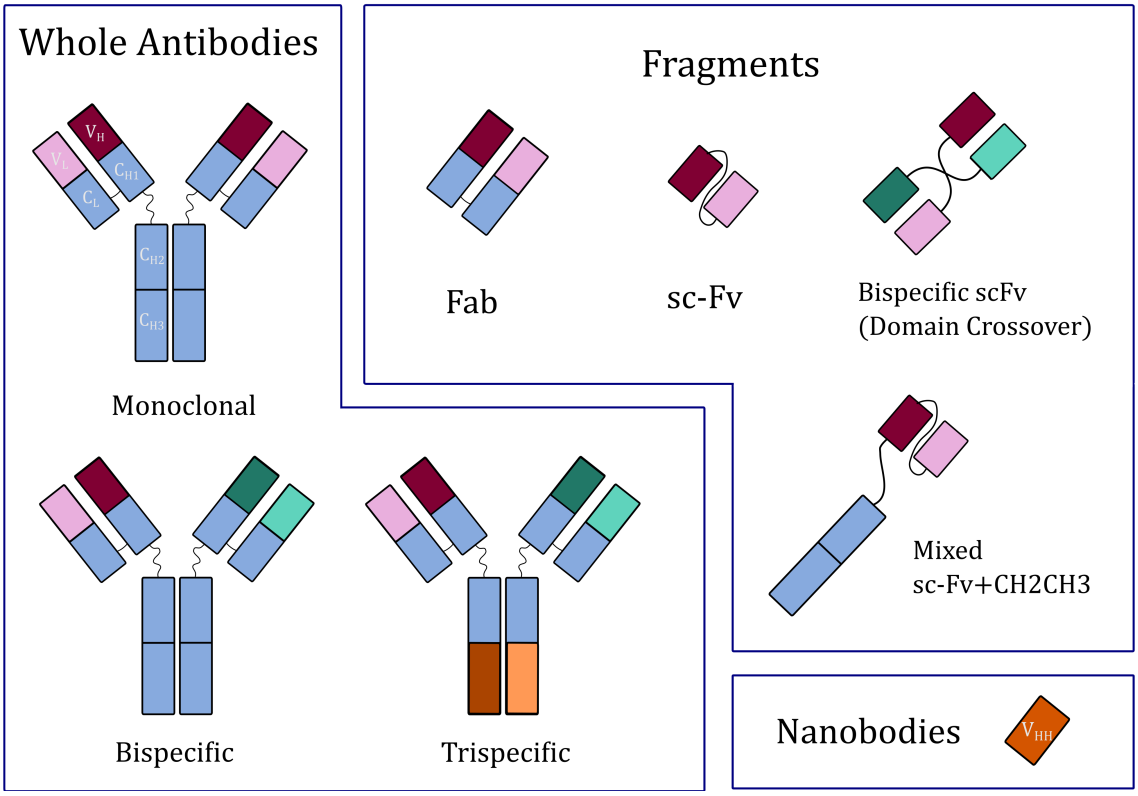


Figure 1.4: A selection of therapeutic immunoglobulin formats. Darker colours signify the VH domain, light colours signify the VL domain, and colour schemes (purple/pink, green/turquoise, brown/orange) signify different binding specificities.

However, the Fc also represents additional molecular weight. This means whole mAb therapies are not orally available, and instead rely on subcutaneous injection for their delivery. They are therefore delivered in a clinical setting infrequently and in high concentrations, which can lead to formulation issues (see Section 1.7). As a result, many alternative immunoglobulin-like formats have been explored that preserve the variable domain component (to selectively recognise antigen epitopes), but alter the constant regions to achieve particular biophysical profiles.

1.5.2 Immunoglobulin-like Formats

An enormous variety of immunoglobulin-like formats now exist, some natural and others engineered (Fig. 1.4). A list of the primary categories follows:

1.5.2.1 Constant Domain Truncation: Fabs and scFvs

Some of the first engineered antibody therapeutics were whole mAbs with sections of the constant domain removed to improve therapeutic pharmacokinetic properties,

enable deeper tumour penetration, and to access more cryptic antigen binding sites [105]. Fab-format therapeutics are created by excising the entire Fc domain from a whole antibody. This can be expected to dampen some immune system recruitment mechanisms, while others (such as neutralisation) can still occur so long as the variable domain retains high complementarity for the antigen active site. Avidity[13] can be retained by connecting two Fab regions with a flexible glycine/series linker.

This principle can be extended further to remove the CH1/CL domains, a format known as the single-chain Fv region (scFv), which this time requires a linker to replace the lost CH1-CL disulfide bridge [105]. Formats also exist where the VH and VL domains of the same peptide chain contribute to different binding sites; these are termed 'domain crossover' Fabs/scFvs.

More recently, engineered formats incorporating some regions of the constant domain have been proposed. These 'mixed' formats include scFv-CH2-CH3, scFv-CH3, and scFv-Fc.

1.5.2.2 Bispecific and Trispecific Antibodies

Another branch of antibody engineering involves incorporating two or more distinct antigen binding sites into the same molecule [106, 107]. Therapeutics with two binding site specificities are termed 'bispecific', and those with three are labelled 'trispecific'. These binding sites can be displayed on whole antibody, Fab, scFv, or mixed formats. The main goal of multispecific antibodies is either to target multiple epitopes on the same antigen (e.g. against a viral protein making mutational escape more challenging) or to deliberately draw together the targeted antigen and components of the immune system [106]. The approved trispecific therapeutic Catumaxomab, now withdrawn for financial reasons, targets EpCAM with its first Fab region, CD3 (presented on T-cells) with its second, and the Fc region recruits other cytotoxic immune cells [99].

A challenge of bispecific antibody engineering is ensuring intended binding site coupling (known as the 'chain-association' issue) [106]. This can be solved trivially (for example by designing both binding sites to use the same VL), or by engineering technologies such as 'knobs into holes' to ensure a heterodimeric product [108].

[13] Avidity is the compound of binding site affinity plus the additional binding energy that results from having multiple binding sites on the same protein scaffold. The expectation is that as one Fv domain is engaging with the antigen, the antigen is co-localised into the same neighbourhood as the other Fv domain, and that this pre-organisation lowers the entropic cost of binding.

1.5.2.3 Foreign Immune Proteins: Nanobodies

Unlike antibodies, which use homodimeric couplings of two heavy and light chains, nanobodies (Nbs/VHHs) are monomeric comprising a single contiguous protein chain weighing just 15kDa. Their lightness brings numerous pharmacodynamic and pharmacokinetic advantages, including the ability to penetrate even more deeply into tumours than engineered antibody formats [109] and also to target intracellular antigens, opening up entire new disease control strategies [110]. The first nanobody therapeutic, Caplacizumab, developed by AbLynx, was approved in 2019 for the treatment of thrombotic thrombocytopenic purpura.

Deriving from camelids and cartigenous fish, the nucleotide composition of the VHH gene loci is distinct from the human VH gene loci. However, the resulting pattern of diversity is similar, as V, D, and J gene fragments recombine to create three CDRs of which one is hypervariable (CDR3) since it lies across both gene junction points. The nanobody CDR3 loop is on average longer than the antibody CDRH3, and can access its own distinct array of conformations [111]. This extra CDRH3 length can be challenging from a developability perspective (see Section 1.7), but can also make accessible previously untargettable concave antigen binding sites, such as those of HIV viral antigens [112]. Recent engineering developments have enabled the development of libraries of synthetic nanobodies ('sybodies') with well-characterised binding site topographies (concave, flat, or convex). The primary uncertainty overshadowing nanobody therapeutics is their potentially high rate of clearance/degree of unintentional immunogenicity, given that heavy chain-only antibodies are not native to humans (see Section 1.7).

1.5.3 Therapeutic Antibody Discovery Platforms

Sormanni *et al.* in a 2018 review framed therapeutic antibody discovery as transitioning through three technological 'generations' [113]. Below, we summarise the principles of each one, from *in vivo* discovery ('1st Gen'), to *in vitro* discovery ('2nd Gen'), to *in silico* discovery ('3rd Gen').

1.5.3.1 *In vivo* Antibody Discovery

In vivo antibody discovery harnesses the power of a natural immune system to identify promising prophylactic antibodies. The technology was first developed in animal models (usually mice). In this context, a pathogen is subcutaneously injected to 'challenge' the murine immune system. After around two weeks, the B-cells are

harvested from the murine immune organs, immortalised using hybridoma technology, and are 'panned' for binding against the antigen of interest (see Section 1.6.1) [114]. Engineering is usually required to improve the longevity and tolerance of these murine antibodies within the human body (see Section 1.7.2): either the CDRs from the murine antibodies can be grafted onto a human constant region ('chimerisation'), or portions of the binding mouse Fv can be back-mutated to human germline residues ('humanisation'). Recently, *in vivo* antibody discovery has seen a revival in both the context of transgenic animals and human immune responses. Technology now exists to genetically incorporate human immune gene loci into animals (e.g. Kymab's Intelliselect® mice [115], or Open Monoclonal Tecnology's Omnirats® [95]), as well as to isolate blood serum from infected (or ideally convalescent) human patients and to rapidly pan it for antigen-binding antibodies [116]. Serum biopanning was pivotal in the rapid development of potential neutralising antibody therapies against SARS-CoV-2 (see Chapter 5).

1.5.3.2 *In vitro* Antibody Discovery

In vitro antibody discovery exploited advances in recombinant DNA technology to remove living organisms from the lead generation process. The phage display technology, discovered by George Smith and developed by Sir Gregory Winter[14], incorporates genes of proteins of interest into the gene encoding the cell wall of the M13 filamentous phage, allowing them to be 'displayed' on the surface [117]. In the context of antibodies, enormously diverse phage display libraries (on the order of 10^{10}+ unique antibodies) are typically generated by recombinantly inserting heavy and light chain genes into the phage DNA. Each distinct antibody binding site is presented on the surface of the phage as an scFv. Similar technology also exists to display antibodies on eukaryotic (e.g. yeast) cell surfaces [118].

These libraries then undergo several cycles of 'washing' over antigens of interest fixed to the wells of a 96-well plate, to identify which Fv binding sites stick most persistently [117]. These complementary Fvs are detected by fluorescent emission caused by the catalytic decomposition of an added substrate by horseradish peroxidase, which is conjugated to the surface of each phage.

Phage display libraries are reliable sources of sets of antibodies for further development. However, because these antibodies are selected *ex vivo* using recombinant DNA and without selection against native proteins, extensive engineering is usually

[14]Sir Gregory Winter and George Smith shared the Nobel Prize in Chemistry in 2018 for developing this technology.

required to convert them into promising therapeutics [113]. Display technologies are also currently inefficient, as many non-complementary candidates must be accessed to find the small complementary subset.

1.5.3.3 *In silico* Antibody Discovery

With recent improvements to computer processing speed, several computational methods have emerged seeking to streamline the process of antibody drug discovery [113]. These vary from tools to predict biophysical characteristics that affect developability (see Section 1.7) to engineering approaches that modularly combine a framework and set of CDRs into antibodies likely to be complementarity to given antigen surface. Many apply advanced statistical techniques to find correlations between the antibody sequence properties and antibody phenotype [119–122]. An increasing number, including OptMaven/OptCDR [123, 124], AbDesign [125], and RosettaAntibodyDesign [126], incorporate structural features to find complementary geometries for the predefined antigen binding site, with subsequent iterative mutation of the CDR residues to maximise attractive interaction energy according to an energy forcefield[15] (*in silico* affinity maturation [128]). Overall, these tools currently have a relatively low accuracy rate when experimentally validated on conformational epitopes (see Section 1.6.2.2), however are expected to propose a set of antibodies significantly enriched over the random selections generated by recombinant library generation techniques, improving efficiency [119]. A key target for the field is the ability to propose more diverse candidates against a given binding site, as design suggestions can easily get trapped in local minima that differ by a few trivial (e.g. non-binding) residues, rely on the optimisation of known binders, or focus only on engineering CDRH3 while holding the rest of the CDRs constant [121, 122].

In the next two sections, we expand upon the two crucial factors in therapeutic antibody design: measuring/estimating target complementarity and predicting antibody developability.

1.6 Target Complementarity

For a drug to be approved for clinical use, it must have a clear theorised mechanism of action (MOA). This almost always translates to evidence that the therapeutic binds to a native protein that is uncontrollably over-expressed or dangerously mutated, or to

[15]This often assumes the conformation does not change upon mutation, which — particularly for CDRH3 — can lead to inaccuracies [127].

a foreign pathogenic entity, with the implication that antibody binding will dampen the antigen's deleterious signals and return the body to homeostasis. It is therefore critical to prove that a therapeutic antibody binds selectively to its intended target with sufficient strength to account for the proposed MOA[16]. This section describes how this is done experimentally in a qualitative or quantitative manner, followed by the primary computational method for predicting complementarity to a particular antigen surface: antibody-antigen complex modelling.

1.6.1 Experimental Affinity Measurement

There are many experimental ways to qualitatively and quantitatively measure binding affinity, most of which either calculate or estimate the ratio between the mathematical product of the concentration of free antibody and antigen, and the concentration of the antibody-antigen complex (the 'dissociation constant', k_d), or the concentration of antibody required to bind 50% of the antigen substrate (the 'IC50' concentration)[17].

Intensity readouts from isothermal calorimetry [129] and surface plasmon resonance [130] experiments are correlated to binding strength and can be calibrated against reference complexes of known affinity. For high-throughput purposes, the enzyme-linked immunosorbent assay (ELISA) has emerged as a dominant semi-quantitative technique, where assessed antibodies are fixed in the wells of a plate and are washed repeatedly with antigen bound to a secondary enzyme-conjugated antibody [131]. A substrate is then added, which is broken down into a fluorescent compound by the enzyme, to identify which wells contain antigen-specific antibody. ELISA can also be run in a competitive sense to assess whether the tested antibody can outcompete the reference/binds to an overlapping epitope. Similar setups can be used in 'directed evolution' experiments, where only antibodies that bind the antigen survive and proceed to the next round of chemical modification [132].

Alternative techniques such as bio-layer interferometry [133] can offer a more quantitative measurement of binding affinity, but have not yet been applied to high-throughput binding assessment.

[16]Increasingly, binding kinetics (so called 'on/off rates') are being shown to be highly important in determining binding proclivity. Throughout this thesis, we focus on complementarity from the view of free energy (interaction profiles), but it should be noted that an antibody with good thermodynamic properties but a poor kinetic profile may not be efficacious *in vivo*.

[17]In both cases, lower values indicate stronger binding.

1.6.2 Computational Affinity Estimation

If sufficient quantitative experimental binding affinity data exists to the epitope of interest, machine learning protocols can capture the crucial sequence patterns that correlate with high affinity. This has already been implemented to classify modified Fv sequences as complementary to the same target [122]. However, this is only ever practical for the most intensely studied epitopes where sufficiently diverse antibodies have been investigated to avoid overfitting. In the absence of such data, the most popular option is antibody-antigen (Ab-Ag) complex modelling, a structure-based pipeline involving binding site prediction, binding pose generation ('docking'), and pose quality assessment ('scoring') [1] (Fig. 1.5). The following subsections outline the state of the art approaches in this field.

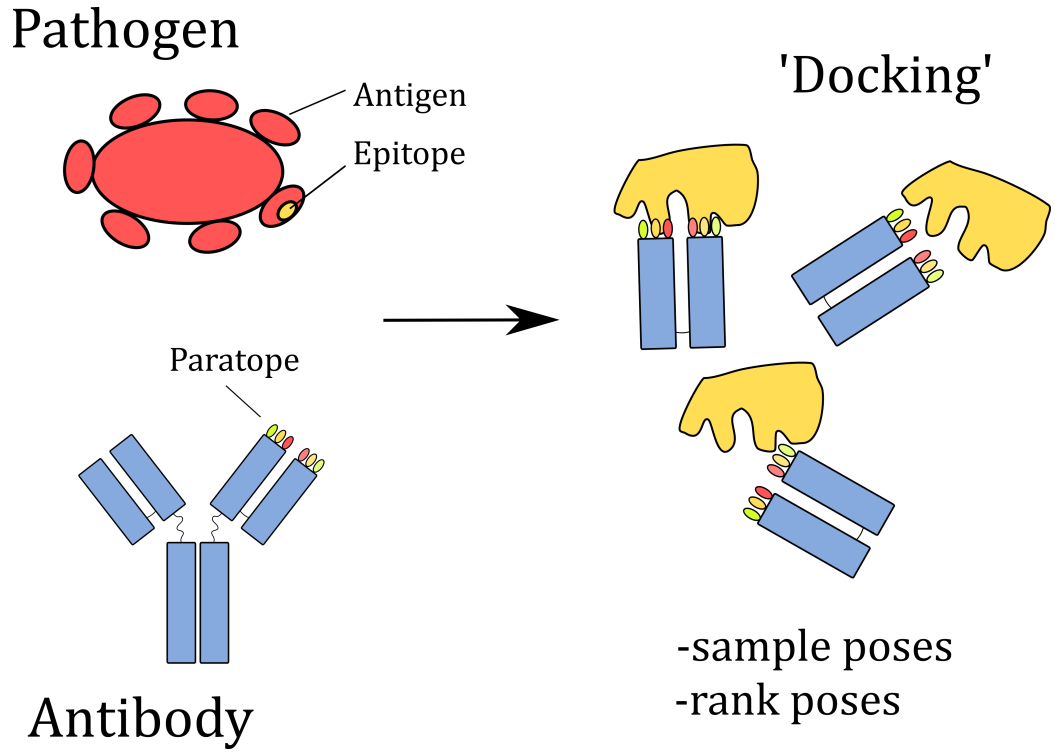


Figure 1.5: Antibody-antigen complex modelling, the current paradigm for *in silico* complementarity assessment. An epitope region on a pathogen's antigen is defined or computationally determined, while the antibody's most likely binding residues (paratope residues) are predicted. 'Docking' then occurs between the crystallographically-solved or structurally-modelled co-ordinates of the putative paratope and epitope pair, to determine if any generated pose meets the energetic threshold for binding.

1.6.2.1 Structural Data

The accuracy of Ab-Ag complex modelling ought to be optimal with solved crystal structures of the antibody and antigen. However, it is impractical to assume that solved structures exist for each antibody of interest in a drug discovery campaign, so high-quality antibody models are essential. Fortunately, the typical accuracy of antibody modelling (assuming short-moderate CDRH3 length) is commensurate with the experimental uncertainty (resolution) of a typical antibody crystal structure [44, 134]. However, sufficiently high quality antigen modelling currently relies on the existence of other homologous proteins [135], and so antigen crystal structures are typically a necessity. The following subchapters assume the comparison of high-quality antibody models and a solved antigen structure.

1.6.2.2 Antibody Paratope Prediction

A sensible first step of Ab-Ag complex modelling involves identifying which residues are likely to comprise the antibody paratope, as this guidance helps improve and accelerate pose generation (see Section 1.6.2.3). Antigen binding typically involves residues in the CDR loops; on average, across all definitions, CDRs capture over 80% of the antigen-binding residues [136, 137]. Several computational methods exist to accurately predict paratopes with sufficient speed for use in high-throughput contexts.

The majority of these methods take only a variable domain sequence as input. Kunik *et al.* [137] developed their Paratome software by harnessing sequence and structural data to estimate the energetic importance of each structurally-conserved antibody position to antigen binding. By incorporating binding residue patterns into a random forest model, proABC can also predict paratope residues on any given input antibody sequence [138]. Most recently, Liberis *et al.* [139] have built Parapred, which trains a neural network on non-redundant Ab-Ag complexes to predict paratope residues, achieving an impressive ROC AUC of 0.878 ± 0.004 across 10-fold cross validation. They also show that this improvement in prediction accuracy translated into better subsequent docking performance, strongly suggesting that further improvements in paratope prediction will return tangible benefits to the reliability of Ab-Ag complex modelling.

The Antibody i-Patch software [140] utilises structures of both the antibody and antigen to generate its paratope prediction. It assigns a binding likelihood score to each input antibody residue based on the frequency of triplets of binding residues observed across antibody-antigen crystal structure interfaces. As it takes into account

the structure of both partners, this tool returns more bespoke results for the particular cognate antigen of interest.

1.6.2.3 Antigen Epitope Prediction

An epitope of an antigen is defined as a subset of its surface residues to which an antibody can bind. Epitopes fall into two categories: linear and conformational.

Linear epitopes are contiguous polypeptide chains, and are relatively easy to predict through sequence analysis. Alignment or sliding windows can highlight residues likely to contribute to binding, distinguished by their predicted surface-exposure alongside their intrinsic chemical properties [141–146].

Conformational epitopes are collections of sequentially-discontinuous residues brought into close proximity by protein folding. Most antibodies target conformational epitopes on proteins, as residues across multiple CDRs engage different regions of the antigen surface to create a more specific, complementary interface [9].

Attempts to predict conformational epitopes began with generic protein-protein interface prediction algorithms (see review by Esmaielbeiki *et al.* [147]). However, the types of contacts found in Ab-Ag complexes were soon shown to be different from those found in general protein-protein interactions [148–150], displaying a unique pattern both in terms of amino acid usage and in binding site interactions [140, 148, 149]. More recent predictors seek improved performance by accounting for this specialised binding, for example by retraining existing protein-protein interaction predictors only on Ab-Ag structures [151–153]. While progress has been made, prediction precision is still close to random for most tools, meaning we remain unable to distinguish the region(s) of an antigen surface that are generally more prone to being bound by antibodies.

Increasingly proteins are found to have multiple epitopes, implying that many, often overlapping, surface patches on a protein antigen can engage an antibody [154, 155]. Epitope prediction algorithms have therefore evolved to incorporate properties of the partner antibody as inputs [156–162]. With this new perspective, both graph-based approaches [158–160] and neural networks [161] have demonstrated an improvement in epitope prediction.

Recently, Bourquard *et al.* [163] have released a pipeline that performs global molecular docking of an antibody into an antigen, and demonstrated the potential of *in silico* epitope mapping. Their algorithm, MAbTope, predicts epitope residues based on consensus epitopes shared by top-ranked poses [163]. This protocol is highly relevant if a number of antigen binders are known, but their respective epitopes remain

unclear. However, its lack of scalability severely limits its use in high-throughput modelling contexts.

1.6.2.4 Antibody-Antigen Docking and Scoring

With atomic representations of the antibody and antigen, docking proposes potential binding configurations of two molecular partners by assessing surface complementarity. The first docking methods were designed for use in small molecule drug discovery, predicting protein-ligand binding interfaces [164]. Since 2005 [165], molecular docking tools have been generalised to allow macromolecular docking, enabling the prediction of protein-protein binding interfaces. Docking algorithms typically survey the conformational space for many binding pose guesses ('decoys'), and then rank them based on a scoring function to highlight the most probable, low-energy configuration(s). In recent years, improvements have been made to both sampling and scoring.

As the initial positioning of binding partners heavily biases sampling towards a particular binding site, global docking algorithms were developed to offer an unbiased sampling of potential binding sites across the antigen surface [165–170]. They generate coarse representations of each complex structure, followed by an evaluation of the shape and physicochemical complementarity at the interaction site. These approaches are very computationally expensive, and so high-throughput modelling currently requires we accept the sampling bias resulting from predefined paratopes and epitopes.

Many docking algorithms are 'rigid-body', meaning that both binding partners are prevented from exploring conformational degrees of freedom during pose generation. The payoff for this is that these methods are very rapid, taking advantage of fast Fourier transform algorithms. However, some binding sites are not accessible through a 'lock and key' binding analysis, and removing these conformational constraints can improve binding site identification, as well as ranking [171]. For each pose, the backbone and side chain conformations at the interface can be optimised for interfacial energy and conformational entropy [124, 172–175]. Such approaches could be especially advantageous in Ab-Ag complex modelling, as compensation could be made for CDR loops with lower predicted accuracy [44], higher predicted flexibility [176], or to capture cooperative binding, in which the Ab, Ag, or Ab and Ag structures distort from their apo conformations into new co-complementary modes as they approach one another. Though knowledge-based constraints [172–174] and the advent of algorithms optimised for graphical processors [177] have improved their efficiency, flexible docking methods remain too slow to use on more than around a

hundred candidate antibody models. This means early stages of any high-throughput docking protocol would only be able to consider rigid-body complementarity, likely precluding the identification of partners that bind through an induced fit mechanism.

Bespoke Ab-Ag scoring functions have been built that take into account the unique binding tendencies of antibodies. For example, by comparing the epitope of the docked complex to its predicted likelihood of being the actual epitope, Krawczyk *et al.* [160] used their EpiPred score to improve the ranking of docked poses. Ramirez *et al.* [178] developed uniquely-weighted scoring schemes for several classes of protein-protein complex, including Ab-Ag complexes. Through their FRODOCK algorithm, they proved that these optimised weights can improve the scores of correct complexes in each class. However, there is still considerable room for improvement, as even class-specific ranking schemes have limited success in consistently recognising the near-native decoy as the 'top hit' [160, 178]. Until this is the case, it may be advantageous to examine the properties of several top-ranking decoys.

1.7 Developability Issues

Achieving the desired target affinity is just one of many characteristics that a therapeutic antibody must possess. This section discusses the many undesirable properties that should be avoided, termed 'Developability Issues' in the field. We cover their biological consequences, potential biophysical origins, and how researchers attempt to detect them experimentally and computationally.

1.7.1 Self-Association: Reversible (Viscosity) and Irreversible (Aggregation)

When antibodies in a medium do not distribute evenly as a colloid, they are likely to be self-associating. This can take the form of reversible self-association, which results in higher viscosity, or irreversible self-association ('aggregation'), which results in precipitation [179]. High viscosity affects drug delivery and pharmacodynamics, while aggregation not only reduces the efficacy of the antibody therapy, but can also trigger a strong anti-drug immune response. Owing to the serious phenotypic effects, most work has been done to find reliable methods of predicting antibody aggregation.

Aggregation commonly results when antibodies have many co-localised non-polar residues on their surface, where the hydrophobic effect drives self association [180]. It is also possible to trigger aggregation through partial unfolding to a non-native structure, as this can force once-buried hydrophobic residues onto the surface [181].

Charge/dipolar effects have also been implicated in both aggregation and viscosity [182–185].

A range of experimental *in vitro* methods have been developed to predict aggregation propensity, including Hydrophobic Interaction Chromatography, Size-Exclusion Chromatography, and Standup Monolayer Adsorption [186]. All create conditions in which aggregating particles are expected to take longer to progress through the experimental apparatus. Recently, a directed evolution approach has been used to filter out aggregating antibodies within the periplasm of E-coli [132] — the first to our knowledge to measure aggregation propensity *in vivo*. *In silico* aggregation prediction methods have historically focussed on identifying atypically large hydrophobic patches in solved structures [187, 188], or on finding correlations between sequence properties and *in vitro* assay values [186, 189].

1.7.2 Immunogenicity

Antibodies are immunogenic if, on injection into the body, a strong immune response is mounted against them[18]. These anti-drug antibodies severely reduce efficacy and the shock to the body can lead to an extremely serious condition known as a 'cytokine storm' [190]. An antibody can be intrinsically immunogenic, either through its binding site resembling a commonly-encountered antigen [191] or being able to interact with the Major Histocompatability Complex (MHC), or they can be immunogenic through colloidal instability (see Section 1.7.1). Older antibody therapies were highly immunogenic, as they were developed with hybridoma technologies (fusing affinity-matured mouse/rabbit B-cells with myeloma cells, see Section 1.5.3.1). Chimerisation and humanisation play a key role in reducing the immunogenicity of animal-expressed antibodies [191]. More recently, recombinant human antibody technologies and transgenic mice allow us to generate panels of 'fully human' antibodies, reducing — though not eliminating — immunogenicity risk[19].

Predicting human immunogenicity is notoriously difficult. Pre-clinical testing of *in vivo* immunogenicity always takes place in animal models rather than humans, while *in vitro* immunogenicity assessment is limited as it relies on biopanning against panels of common self-antigens [179]. Several different *in silico* approaches exist to

[18] 'Immunogenicity' as a developability issue often causes confusion, as antibodies are intentionally immunogenic against their antigen target. In this thesis, we always refer to 'immunogenic antibodies' as those that are recognised as non-self by the body and against which an immune response is triggered.

[19] Even some approved antibody therapeutics trigger anti-drug antibodies, but this is tolerated owing to the severity of the condition they are designed to treat [192].

estimate sequence 'humanness' and by extension the likelihood of an antibody being immunogenic [94, 193, 194].

1.7.3 Poor Expression

Therapeutic antibodies need to be readily expressable in industrial quantities to be commercially viable. Researchers routinely check the expression levels of their candidates in their chosen system, commonly Human Embryonic Kidney 293 (HEK293) or Chinese Hamster Ovary (CHO) cells [186]. Poor expression is likely to result from an inherent instability in the antibody structure. Computational methods have been used to detect rarely-used framework residues in a particular sequence context, which is linked to conformational instability [195], but CDR properties can also be responsible [196]. More generally, thermal stability is tested experimentally *via* Differential Scanning Fluorimetry (DSF) [197]. In this technique, the temperature is raised until the antibody begins to denature, revealing its once-buried hydrophobic residues and these cause the DSF reagent to fluoresce. The higher the temperature before denaturation begins, the more intrinsically stable the antibody.

1.7.4 Polyspecificity

If an antibody binds to multiple biological targets (*i.e.* is 'polyspecific'), it is likely to result in serious side effects, and much reduced efficacy against the intended target owing to its shortened half-life. Experimental methods to test for polyspecificity include biopanning for binding against 'poly-specificity reagent' or panels of common antigens [186], or cross-interaction chromatography [198]. *In silico* assessment of this property is not routine, but may theoretically be performed through comparison to solved complexes of antibodies targeting other antigens.

1.7.5 Chemical Instability

If an antibody residue is prone to chemical modification *in vivo*, then it leads to product heterogeneity, which reduces efficacy and could induce immunogenicity (see Section 1.7.2). This is especially the case if the post-translational modification results in fragmentation of the structure. Prediction of these hotspot residues is typically performed *in silico* through sequence motif similarity [179]. N-linked glycosylation and non-conserved cysteine residues are trivial to identify, while sites of lysine glycation, aspartate isomerisation, and asparagine deamidation require more complex statistical methodology to predict [199].

1.7.6 The Trade-off Between Affinity and Developability

Therapeutic discovery and development requires the simultaneous optimisation of affinity and all the aforementioned developability properties. This is extremely challenging, as frequently improvements in one property lead to a corresponding deterioration of others [113, 200, 201]. Several aspects of developability ought therefore to be actively considered at each step of an antibody development pipeline (as per the Modular method proposed by Sormanni *et al.* [113]), for example as part of a cost function in machine-learning-based affinity maturation algorithms, to yield the most amenable therapeutic candidates possible.

1.8 From Human Antibody Repertoires to Therapeutic Antibodies

1.8.1 Our Proposed Structure-Aware Approach

Here, we describe our proposed computational pipeline to identify natural antibodies that represent promising therapeutic leads against a predefined antigen epitope (Fig. 1.6). By uniting the principles of *in vivo* and *in silico* drug discovery, we believe this approach has the potential to dramatically accelerate drug development — particularly against previously uncharacterised epitopes — reducing resource expenditure and crucially leading to more human-compatible lead candidates.

The pipeline would work in several stages. First, an appropriate set of Ig-seq datasets should be selected. These may be chosen to reflect particular health states (e.g. healthy/convalescent) and/or B-cell types (e.g. naïve/mature/memory) of interest. If the data is unpaired, heavy and light chain pairings would need to be computationally assigned. If required, a subset of antibodies could be chosen from across the repertoires, using methods such as sequence and predicted structural clustering, to obtain a more computationally tractable number. Each representative antibody's full Fv structure should then be predicted using a sufficiently rapid and accurate homology modelling tool. At this point, *in silico* developability assessment should take place to highlight which antibodies may be particularly prone to developability issues. Separately, a structure of the antigen of interest should be collected (or accurately modelled), and an epitope of interest determined. Each antibody model should be marked up with their most probable paratope, docked, and scored for complementarity against the selected epitope. Multiple rounds of docking and scoring may be required to first hone in on the most promising binding site topologies, then

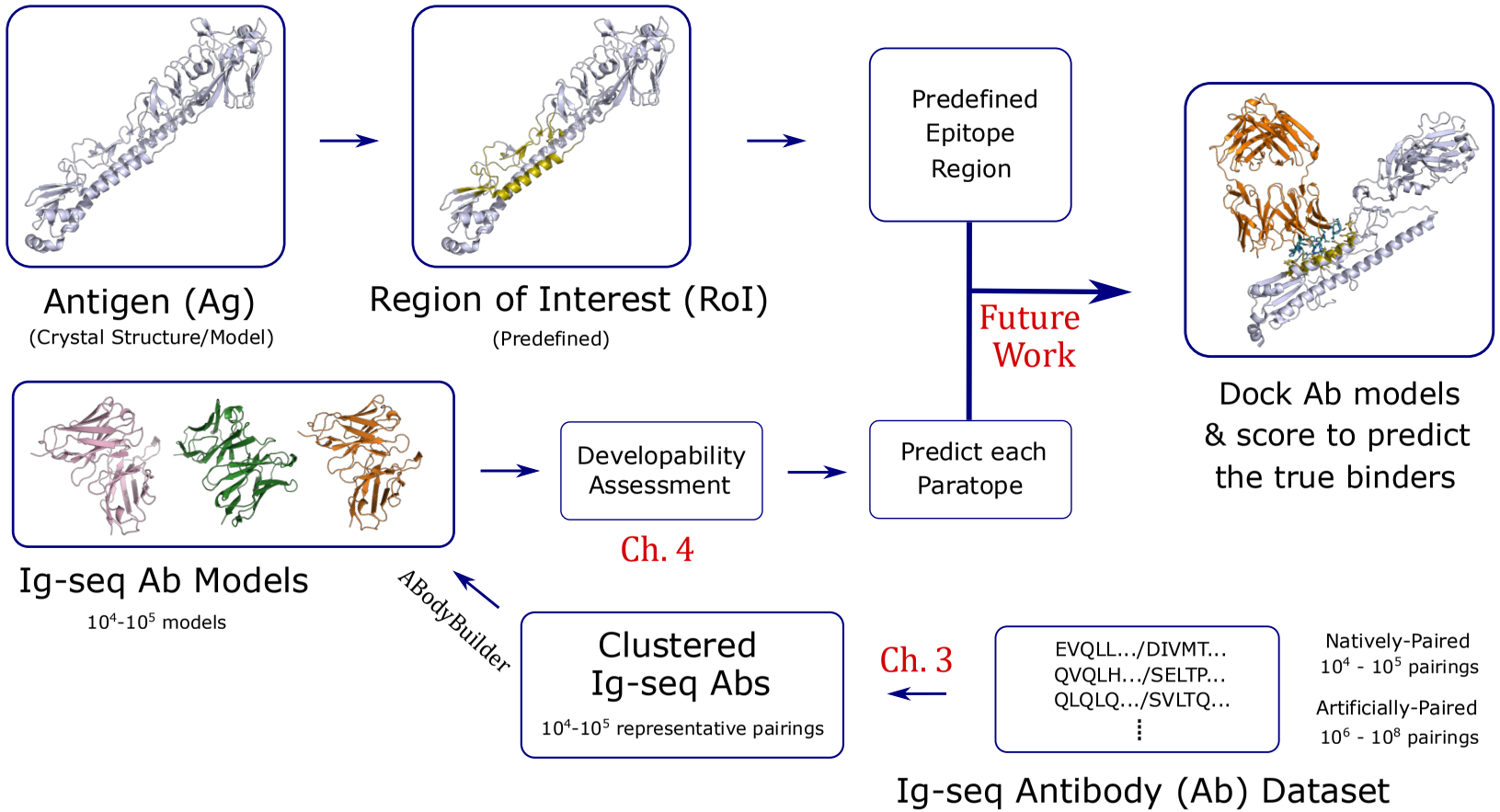


Figure 1.6: A schematic showing our proposed paradigm for high-throughput antibody-antigen (Ab-Ag) complex modelling of immunoglobulin gene sequencing (Ig-seq) data. This pipeline could act as a lead generation platform for therapeutic antibody design. Our novel algorithm for clustering Ig-seq Ab datasets into representative structures is described in Chapter 3 (Ch. 3), while our novel framework for early-stage *in silico* developability assessment is described in Chapter 4 (Ch. 4).

to reintroduce sequence diversity around these geometries, and finally to account for flexibility once candidates have been narrowed down even further.

Two of the software packages described in this thesis (Repertoire Structural Profiling, and the Therapeutic Antibody Profiler) can be applied in concert to convert Ig-seq samples of BCR repertoires into structurally representative sets of Fvs and assess them for their propensity for developability issues. Work is ongoing to develop the final tool to rapidly assess antibody-antigen complementarity (see Chapter 6). The two databases reported in this thesis (Thera-SAbDab and CoV-AbDab) represent ancillary resources with many applications, including providing training data for future docking scoring function development, or for the *in silico* functional annotation of BCR repertoire Ig-seq datasets.

1.8.2 Chapter Walkthrough

In Chapter 2, we describe our novel database — the Therapeutic Structural Antibody Database (Thera-SAbDab) — which documents the sequences, solved structures, and metadata of every therapeutic antibody recognised by the World Health Organisation. We use Thera-SAbDab to show that natural BCR repertoires contain near-therapeutic sequences and so ought to be represent promising starting points for future therapeutic development.

In Chapter 3, we describe Repertoire Structural Profiling, a new method we have developed to capture the maximum modellable binding site structural diversity within Ig-seq datasets, leading to 'Antibody Model Libraries'. We show how these Antibody Model Libraries can act as a basis set for designing novel screening libraries for *in vitro* and *in silico* therapeutic antibody drug discovery.

In Chapter 4, we describe our set of five computational developability guidelines derived from the physicochemical properties of therapeutic antibody models and BCR repertoire Antibody Model Libraries. We introduce the Therapeutic Antibody Profiler, a software package for early-stage drug discovery that flags therapeutic candidates likely to have disadvantageous biophysical properties.

In Chapter 5, we describe our contribution to the COVID-19 research effort, centered around our novel database (CoV-AbDab) that contains sequence and structural information on all antibodies empirically proven to bind coronaviruses. We describe how this database has already found use in SARS-CoV-2 response repertoire profiling.

Finally, we describe the logical next steps from our work, in particular those required to complete the hypothesised therapeutic design pipeline and to benchmark its performance.

References and an Appendix containing supplemental methods, tables, and figures are provided at the end of the thesis.

Chapter 2

The Therapeutic Structural Antibody Database: Implications for Drug Discovery from Natural Repertoires

2.1 Chapter Abstract

Since our long-term goal is to develop a novel structure-based pipeline for therapeutic antibody discovery, we conducted a thorough survey of all existing clinically-tested therapeutics with B-cell genetic origin, focussing on documenting their structural properties. We found that over three times the structural data was available for therapeutic antibodies/nanobodies than had been documented in existing databases. We collated this knowledge into a new database (Thera-SAbDab), made it freely downloadable as a single file, and added an array of search features not offered by other repositories.

Thera-SAbDab has played an important role in developing and benchmarking many of the methods described in this thesis; a study facilitated by Thera-SAbDab is described later in this chapter, in which we evaluated the proximity of advanced-stage therapeutic antibody sequences to natural B-cell receptor (BCR) repertoires sampled using immunoglobulin gene sequencing (Ig-seq). The results demonstrate the immense potential that natural repertoire sequencing holds for future therapeutic antibody design. The chapter concludes with an update on the current status of the database, alongside a discussion on how its contents may develop over time.

This chapter contains reproduced material from the following papers:
Raybould, M.I.J., Marks, C.M., Lewis, A.P., Bujotzek, A., Taddese, B., Deane, C.M. (2020) Thera-SAbDab: the Therapeutic Structural Antibody Database. *Nucleic Acids Res.* 48(D1):gkz827. [99]
Krawczyk, K., **Raybould, M.I.J.**, Kovaltsuk, A., Deane, C.M. (2019) Looking for Therapeutic Antibodies in Next-Generation Sequencing Repositories. *mAbs.* 11(7):1197-1205. [202]

2.2 Thera-SAbDab: the Therapeutic Structural Antibody Database

2.2.1 Introduction

Immunotherapeutics derived from B-cell genes are an increasingly successful and significant proportion of the global drugs market, designed to treat a wide range of diseases [100, 203, 204]. Whole monoclonal antibody (mAb) therapies dominate the industry — drugs that mimic natural antibodies by containing two identical variable domain structures with a particular specificity [100]. The broader class of monoclonal therapies also includes Fragment antigen binding (Fab) regions (a single arm of a whole antibody), single-chain Fv (scFv) regions (a heavy and light chain variable domain connected by an engineered glycine-rich linker), and single-domain variable fragments. These fragments can be expressed in dimeric form to improve avidity, or conjugated with polyethylene glycol ('pegylated') for slower clearance [205], with radioisotopes for diagnostic purposes [206], or with radioisotopes or noxious small molecules/peptides for cytotoxicity [207].

Recent developments in protein engineering have resulted in bispecific immunotherapies, where two distinct variable domain binding sites are incorporated into a single protein. As of June 2019, bispecific mAbs, linked Fabs, linked scFvs and linked single-domain variable fragments had all been assessed in clinical trials [106].

A primary source of information on immunotherapies is the World Health Organisation (WHO), which publishes biannual 'Proposed' [208] and 'Recommended' [209] International Nonproprietary Name (INN) lists. These INNs serve as globally-recognized generic names by which pharmaceuticals can be identified. To be granted an INN, applicants must include a full amino acid sequence, the closest V and J gene, the IG subclass, and the light chain type (*https://extranet.who.int/tools/inn_online_application*). This information, coupled with the $12,000 cost of application (as of

August 2019), makes INN lists a useful source of therapies that companies intend to carry forward into clinical trials.

Several databases already harvest this information; three non-commercial repositories that specifically record antibody data are the IMGT Monoclonal Antibody Database (IMGT mAb-DB) [210], the AntiBodies Chemically Defined (ABCD) database [211], and WHOINNIG (*http://www.bioinf.org.uk/abs/abybank/whoinnig*).

The Therapeutic Antibody Database (TABS; *https://tabs.craic.com*) is antibody-specific and commercial, also scraping patents for therapies. Other databases not specific to antibodies can also capture WHO information, such as ChEMBL (*https://www.ebi.ac.uk/chembl*), DrugBank (*https://www.drugbank.ca*) and KEGG DRUG (*https://www.genome.jp/kegg/drug*).

Most databases supply additional metadata for their therapeutic entries, such as clinical trial status, companies involved in development, target specificity, and alternative names. For example, the ABCD database provides antibody synonyms, antigen UniProt links and publication references [211]. However, while these repositories supply sequence information (either on individual summary pages or through reference to the primary literature), it is currently not possible to query them by sequence, nor to bulk-download relevant sets of therapeutic sequences for direct bioinformatic analysis. The largest previous effort to assimilate therapeutic antibody sequences for analysis was a set of 137 collated by Jain *et al.* in January 2017 [186].

Structural knowledge about both the intended target and the therapeutic lead compound is of high importance for rational drug discovery [1, 212]. For example, co-crystal complexes reveal where a drug binds to its target (the surface 'epitope'), and separately-solved structures enable more accurate docking experiments. It can also assist subsequent development and optimization, as homology models of mutants derived from a known structure are in general more accurate than those for which no close structural partner is available [213]. The Protein Data Bank [19] (PDB) now contains over 150 000 solved structures, and though it is highly biased towards certain protein classes, many diverse targets of pharmacological interest are represented. A significant fraction of these structures contain antibody variable domains, and these are recorded by the Structural Antibody Database (SAbDab [9]; 7184 variable domain structures over 3663 PDB entries as of 5$^{\text{th}}$ August 2019). Both IMGT mAb-DB and TABS report a set of known therapeutic structures in the PDB, but their reported structural coverage of therapeutic space is low. For example, neither database reports any known structural information for bispecific immunotherapeutics.

To address these deficiencies, we created the Therapeutic Structural Antibody Database (Thera-SAbDab; *http://opig.stats.ox.ac.uk/webapps/therasabdab*). We harvest sequences as they are released by the WHO, number them with ANARCI [18], and perform a weekly sequence alignment of all therapeutic variable domain sequences to the sequences of known structures stored in SAbDab. Structures with sequence identity matches of 100%, 99% and 95–98% are recorded and categorized, with alignments on each therapeutic summary page to show precisely where each near-identical structure differs from the therapeutic sequence.

Thera-SAbDab can be queried by INN, by a combination of metadata, such as INN proposal year, clinical trial status, or target, or by sequence (including over a specified region of the sequence). We make available all therapeutic sequences contained within Thera-SAbDab, alongside metadata, to facilitate further research.

2.2.2 Data Sources

2.2.2.1 Sequence Data

Proposed INN lists [208], published by the WHO, are the source of the majority of sequence information in Thera-SAbDab. These are released biannually (one in January/February and another in June/July) and—since list P95 in 2006—represent a reliable record of variable domain sequences for all antibody- and nanobody-related therapeutics granted a proposed INN. Of the 129 antibody-related therapeutics proposed before 2006, we were able to find sequence information for 47 (36.4%) through the IMGT mAb-DB (*http://www.imgt.org/mAb-DB/*). Although we continue to search, and joint academia-industry initiatives such as Abvance encourage their release (*https://www.pistoiaalliance.org/projects/abvance/*), sequences for the remaining 82 may never become public knowledge.

All sequences are then numbered by ANARCI [18], which uses Hidden Markov Models to align input sequences to pre-numbered germline sequences. Assigning a numbering allows users to more easily interpret the significance of mutations in near-identical sequence matches. For example, if the mismatch occurs in the extremities of the framework region, it may be judged to have minimal effect on binding site structure.

2.2.2.2 Structural Data

Thera-SAbDab compares all numbered therapeutic sequences to the structures in SAbDab [9], which pre-filters the PDB [19] for all structures whose sequences align to

B-cell germline genes. As all SAbDab structures are also pre-numbered, the comparison of therapeutics to public structural space is efficient. All the existing functionality of SAbDab (e.g. interactive molecular viewers and numbered structure downloads) is made easily accessible from Thera-SAbDab search results.

2.2.2.3 Therapeutic Metadata

Therapeutic metadata comprises a mixture of inherent characteristics and continually-changing status updates.

Certain static properties can be acquired automatically. For example, light chain type is identified through our ANARCI germline alignment [9], while isotype, INN Proposed and Recommended years, and intended target(s) can be harvested directly from the INN lists. Sequence comparison can also be used to identify where different INN names refer to identical variable domains. Other characteristics, such as which companies are involved in therapeutic development, must be manually curated at the time of deposition.

Time-dependent characteristics for new entries are also manually curated after sequence identification, and thereafter every 3 months. We obtain clinical trial information, developmental status, and investigated condition data from a range of sources including AdisInsight (*https://adisinsight.springer.com*), ClinicalTrials.gov (*https://clinicaltrials.gov*), and DrugBank (*https://www.drugbank.ca*). These websites are updated more regularly, and so are preferable sources for this time-sensitive metadata; we include these fields in Thera-SAbDab to allow for more pharmacologically relevant searches, as well as to identify all post Phase-I candidates for inclusion in our five updating developability guidelines [214] (described in Chapter 4). The SAbPred links provided allow users to build a homology model of a therapeutic antibody sequence with ABodyBuilder [44] using the latest loop template database from SAbDab [9] (updated weekly) and to evaluate length-independent canonical classes of the non-CDRH3 loops using the latest SCALOP [25] definitions (updated monthly).

2.2.3 Contents

As of 5$^{\text{th}}$ August 2019, Thera-SAbDab was tracking 558 INNs, representing 543 unique therapeutics. Of the 558 INN names, 473 could be mapped to variable domain sequences (87.1%), representing 461 unique therapeutics with sequence data. 436 were monoclonal therapies (three pairs of which share identical variable domains: avelumab & bintrafusp, losatuxizumab & serclutamab and radretumab & bifikafusp), and 25

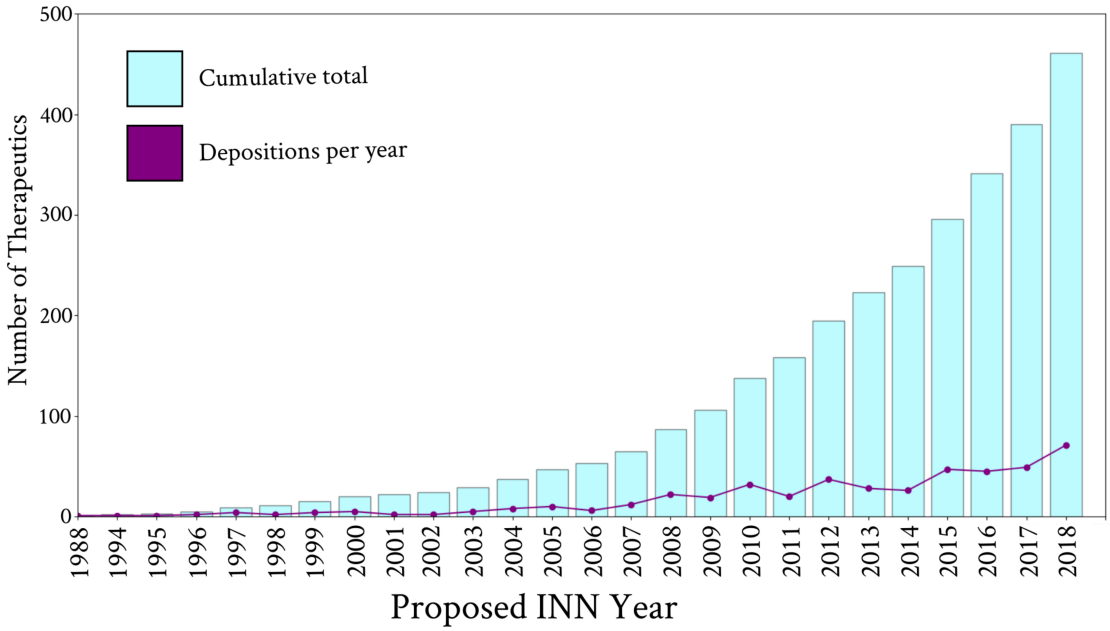


Figure 2.1: The number of antibody- and nanobody-related therapeutics assigned an International Nonproprietary Name (INN) by year. A record number of 72 of these therapeutics were recognised by the WHO in 2018.

were bispecific therapies. Plotting the cumulative sum of these unique therapeutics by year deposited in a WHO 'Proposed INN' list shows an exponential increase since the early 2000s (Fig. 2.1).

We searched the IMGT mAb-DB [210] and TABS databases (on 28$^{\text{th}}$ June 2019) for structures of these 461 therapeutics. IMGT mAb-DB identified 72 structures of therapeutic variable domains, across 36 different monoclonal therapeutics, while TABS reported 53 structures of therapeutic variable domains, across 32 different monoclonal therapeutics. In contrast, Thera-SAbDab (at the 100% sequence identical threshold) contained 152 therapeutic variable domain structures, across 84 distinct monoclonal therapeutics and 7 distinct bispecific therapeutics. A further 21 monoclonal therapeutics had maximum sequence identity matches of 99% (up to two mutations away from a publicly-available structure), and 13 monoclonals and 4 bispecifics had maximum sequence identity matches of 95–98%. We concluded that, at this time, around a quarter (27.1%) of WHO-recognized monoclonal therapeutics had exact or close ($\geq$ 95% sequence identity) structural coverage. 44.0% of bispecific therapeutics had at least one variable domain with exact or close structural coverage, and two had exact matches for both variable domains.

Thera-SAbDab contains structural information for even the most diversely formatted therapeutics. Ozoralizumab, a bispecific therapy in active Phase-III clinical trials for rheumatoid arthritis, has a VH(TNFF-α)–VH(ALB)–VH(TNFF-α) configuration, where VH(TNF-α) is a heavy chain designed to bind to Tumour Necrosis Factor *alpha* (TNF-α, an inflammatory cytokine), and VH(ALB) is another heavy chain designed to bind human serum albumin (ALB, an abundant human serum protein that can be targeted for an improved serum half-life). Thera-SAbDab has identified a structure for the TNF-α binding domain with sequence identity of 95.65% [5m2j; chain D]. Inspection of the sequence alignment shows that 5m2j has a 100% Chothia-defined CDRH3 sequence match to VH(TNF-α), and in fact only differs by one mutation across all Chothia-defined [23] CDRs: D31 in VH(TNF-α) is N31 in 5m2j. 5m2j is a VHH2 llama nanobody, suggesting that SAbDab's coverage of nanobody structural space will be increasingly highlighted by Thera-SAbDab as more single-chain therapies arrive in the clinic.

Therapeutically-relevant structures are continually being deposited in the PDB, even many years after initial development. For example, since 2009, the WHO have recorded nine antibody-related therapeutics against IL17A—seven monoclonals and two bispecifics. The first, secukinumab, was recognized in 2009, and since 2014 has been approved for use in certain types of arthritis, psoriasis, and spondylitis. As of early June 2019, there were no close structures for any of these IL17A-binders. However, on 19$^{\text{th}}$ June 2019, Eli Lilly deposited an exact variable domain structure for ixekizumab (an IL17A-targetting monoclonal antibody, 6nov) and a close structure for tibulizumab (an IL17A-binding and TNFSF13B-binding bispecific antibody, 6nou) in the PDB [215]. SAbDab detected and numbered them in its weekly update, making Thera-SAbDab the first antibody database to link to the structures of IL17A-binding therapeutic antibodies.

2.2.4 Usage

There are multiple ways to search Thera-SAbDab. Thera-SAbDab can be queried directly by INN if structural information about a particular therapeutic is needed. Alternatively a combination of metadata can be specified to identify structures for a particular subset of therapeutic space, for example binders to a particular antigen, or therapeutics at a particular stage of clinical trials (Fig. 2.2A). Results are returned in a table format, with links to each therapeutic summary page and a selected array of metadata (Fig. 2.2B).

A

B

8 therapeutic(s) match your criteria. Click on the therapeutic name to open a summary page.
[NFD = No Further Development, (w) = Withdrawn, Semicolons delimit separate variable domains for bispecifics]

Therapeutic	Format	Highest Clinical Trial (June '19)	Est. Status (June '19)	Target	Year Proposed	100% SI Struc.	99% SI Struc.	95-98% SI Struc.
disitamab	Whole mAb ADC	Phase-II	Active	ERBB2	2018	no	no	no
gancotamab	scFv	Phase-II	Discontinued	ERBB2	2018	no	no	no
margetuximab	Whole mAb	Phase-III	Active	ERBB2	2013	no	no	no
marstacimab	Whole mAb	Phase-II	Active	ERBB2	2013	no	no	no
pertuzumab	Whole mAb	Approved	Active	ERBB2	2003	YES	YES	YES
timigutuzumab	Whole mAb	Approved	NFD	ERBB2	1997	YES	YES	YES
trastuzumab	Whole mAb	Approved	NFD	ERBB2	1997	YES	YES	YES
zenocutuzumab	Bispecific mAb	Phase-II	Active	ERBB3;ERBB2	2017	YES;YES	no;no	no;no

Figure 2.2: Searching by Attribute (*http://opig.stats.ox.ac.uk/webapps/therasabdab*). **(A)** Here, we search for any therapeutic designed to bind to ERBB2 (often over-expressed in breast cancer). **(B)** Eight therapeutics are designed to bind to ERBB2, seven monoclonals and one bispecific. Four have exact structural information for the ERBB2 binding site. Click the therapeutic name to enter the therapeutic summary page.

Each therapeutic summary page lists a structural summary (including our database sequence), with links to relevant SAbDab entries (with PDB codes and chains), and alignment charts (if structures with 95–99% sequence identity are detected). Each SAbDab link redirects the user to the SAbDab summary page for the relevant PDB entry, where all existing functionality can be accessed. Links to appropriate SAbPred [216] informatics tools (such as ABodyBuilder [44] for variable domain structure modelling, and the Therapeutic Antibody Profiler [214] [see Chapter 4] for developability assessment) are also provided. Finally, we list all the remaining metadata that we have recorded for the therapeutic, ranging from records of investigated conditions, to which companies are developing the therapeutic, to its estimated

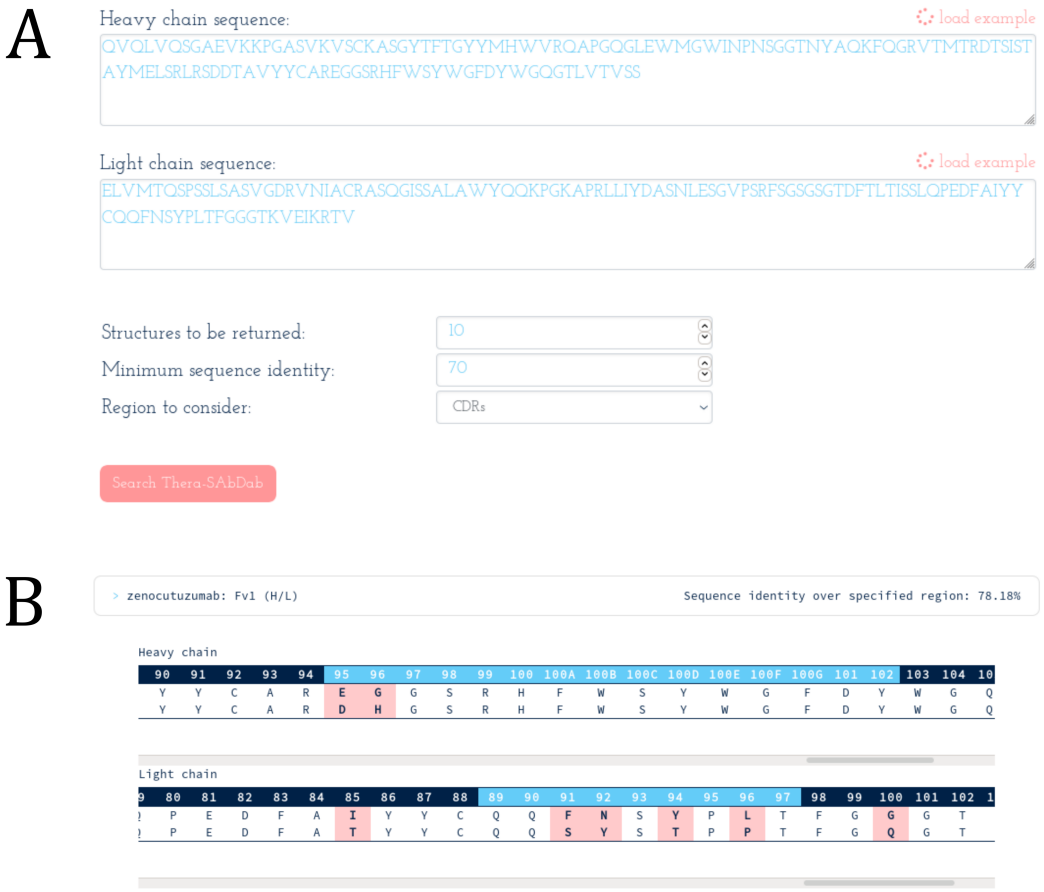


Figure 2.3: Searching by Sequence (*http://opig.stats.ox.ac.uk/webapps/therasabdab*). **(A)** Here, we search for therapeutics with at least 70% sequence identity across the heavy and light chain CDRs of the input sequence. **(B)** Any results are returned alongside sequence identity across the specified region. Alignments show any sequence mismatches across the variable domain sequence.

developmental status.

A third way to search Thera-SAbDab is by sequence (Fig. 2.3). This can be harnessed in numerous ways. For example, by querying with a known therapeutic sequence, researchers can look for sequence commonalities between therapeutics over any region of the variable domain. Alternatively, by querying with a developmental candidate sequence, researchers can search for similarity to any other therapeutic, or specifically to those designed to bind to the same target. This could identify potential patenting issues, highlight a risk of polyspecificity, or suggest a binding mode to the intended target.

A case study workflow harnessing Thera-SAbDab is available at *http://opig.stats.ox.ac.uk/webapps/therasabdab/about*.

2.2.5 Implementation

The original web application for Thera-SAbDab (released on paper submission, July 2019) used a combination of PHP and JavaScript to dynamically handle user search queries. For consistency with our new SAbBox Virtual Box suite, the latest version (with the assistance of Dr. Claire Marks) has been transferred to a Flask database framework (*https://github.com/pallets/flask*), which handles dynamic inputs using Python.

2.2.6 Conclusion

We have created Thera-SAbDab with the central aim of collating all public structural knowledge for WHO-recognized antibody- and nanobody-related therapeutic variable domains. Rather than relying on text-mining approaches, which can miss PDB depositions that omit reference to the structure's therapeutic relevance, Thera-SAbDab uses a systematic approach at the level of sequence identity to detect exact and close matches to our repository of therapeutic variable domains.

This approach has not only enabled us to identify over twice the number of monoclonal therapies with 100% sequence-identical structures in the PDB than in existing databases, but has also identified exact variable domain structures for several bispecific therapies. Our approach can also distinguish between PDB structures with 100%, 99%, and 95–98% sequence identity matches. Sequence alignments guide the interpretation of structures of near-identical sequence. Therapeutics without solved structures can be passed directly to our homology modeller, ABodyBuilder, to approximate their structure using a continually-updated loop template database [9, 26, 27].

Like IMGT-DB, Thera-SAbDab can be queried by metadata, but uniquely it can also be queried by variable domain sequence. This enables researchers to identify therapeutics proximal over any variable domain region to their query sequence.

As shown for IL17A-binding therapeutics, new clinically-relevant structures are continually being released. Accordingly, Thera-SAbDab checks SAbDab after each weekly update for new matches, ensuring that this data is rapidly captured. Thera-SAbDab's sequence database is updated with new sequence information twice per year, in line with the release of new WHO Proposed INN lists. An updated list of all therapeutic variable domain sequences with metadata is supplied as a single file to facilitate further analysis. We have already used the data pooled by Thera-SAbDab to probe how similar therapeutic sequences are to natural antibody sequences; this study is described in the next section.

2.3 Do Natural Repertoires Harbour Therapeutic Antibody Sequences?

2.3.1 Introduction

One of the major requirements for a therapeutic antibody is that it does not initiate harmful immunogenicity when delivered to the body. This is usually mediated by triggering anti-drug antibodies (ADAs), which are most likely raised in humans if the injected therapeutic has discernibly 'foreign' (*i.e.* non-human) characteristics or if it is self-reactive to native human antigens.

At the time of publication, the vast majority of later-stage (post-Phase I) therapeutic antibodies referenced in Thera-SAbDab [99] were developed from parent antibodies identified using established industrial protocols. These protocols fall broadly into three categories: animal immune challenge, transgenic animal immune challenge, or recombinant human antibody phage display panning[1]. While the latter two methods harness human V(D)J gene transcripts, none of the techniques guarantee that the parent antibody is naturally expressible in humans, nor that it is unable to bind to human auto-antigens[2]. In addition, subsequent engineering is usually required to optimise binding affinity or other biophysical properties, and this process can introduce issues not present in the parent mAb.

Nevertheless, regardless of developmental origin, advanced clinical-stage therapeutic antibodies could owe their progression to convergence on some characteristics of naturally expressed human antibodies, thus avoiding both detrimental properties that correlate with ADA production. To investigate the likelihood of this happening at the variable domain sequence level, we tested the sequence proximity of the clinical-stage therapeutic (CST) antibodies in Thera-SAbDab to a large number of natural human antibodies/BCRs identified in high-throughput repertoire sequencing studies.

2.3.2 Methods

By April 2019, the Observed Antibody Space (OAS) database, our source of BCR repertoire sequences, had been expanded by four datasets since its publication [88]. The largest of these was the Peripheral Blood Mononuclear Cell (PBMC) IgG/IgM

[1]Recent years have seen the development of many new antibody engineering and formatting methods, leading to modifications of the original INN antibody naming convention [217].

[2]One discovery method that favours both of these qualities is the direct 'peptide baiting' of therapeutic candidates from antigen-responding human antibody repertoires, see Section 2.4.

repertoire study by Briney *et al.* [4]. All the sequences in OAS were obtained using Illumina MiSeq or Roche 454 platforms, meaning heavy and light chains are sequenced separately and native pairings are not preserved (see Section 1.4.1.1).

The CST sequences were numbered using ANARCI [18] according to the IMGT [11] numbering scheme. Each CST was classified into four groups (chimeric, humanised, human, mouse), based on their International Non-proprietary Names [103, 218]. Sequences with names containing '-xizumab' or '-ximab' were labeled as 'chimeric'. Sequences not matching this criterion but containing '-zumab' in their name were classified as 'humanised'. Sequences that contained only '-umab' in their name were labeled as 'fully human'. Three mouse antibodies (muromonab, abagovomab and racotumomab), were labeled as 'mouse'.

We separately aligned the heavy chain, light chain, the combination of the three heavy or light chain IMGT-defined CDRs, and the IMGT-defined CDR-H3 of CSTs to each of the sequences in OAS [88]. We noted a match if an IMGT position in a 'query' CST is also found in a 'template' sequence from OAS, and they have the same amino acid residue. For the full sequence alignments, the number of matches was divided by the length of the query and by the length of the template, producing two sequence identities. The final sequence identity is the average between these two values; calculating the sequence identity in this way prevents the scenario where one sequence is a substring of another, creating an artificially high sequence identity with a large length discrepancy. The CDR alignments were performed only when the IMGT-defined loop lengths matched.

2.3.3 Results

All code comparing therapeutic antibody sequences to the Observed Antibody Space database was written and run by Dr. Konrad Krawczyk, who conceived this study. I analysed the resulting data and wrote the code to evaluate pairwise sequence identity between therapeutic sequences (Section 2.3.3.3).

2.3.3.1 Sequence Identity over Different Regions

We used a set of 242 clinical-stage therapeutic (CST) antibody sequences, all of which have completed Phase 1 clinical trials[3] (post-Phase-I therapeutics have the advantage that they have already been trailed in healthy human volunteers and found not to yield significant ADAs). We separately aligned the CST variable regions (VH or VL),

[3]These represented all the post Phase-I therapeutics identified in the prototype version of Thera-SAbDab [99].

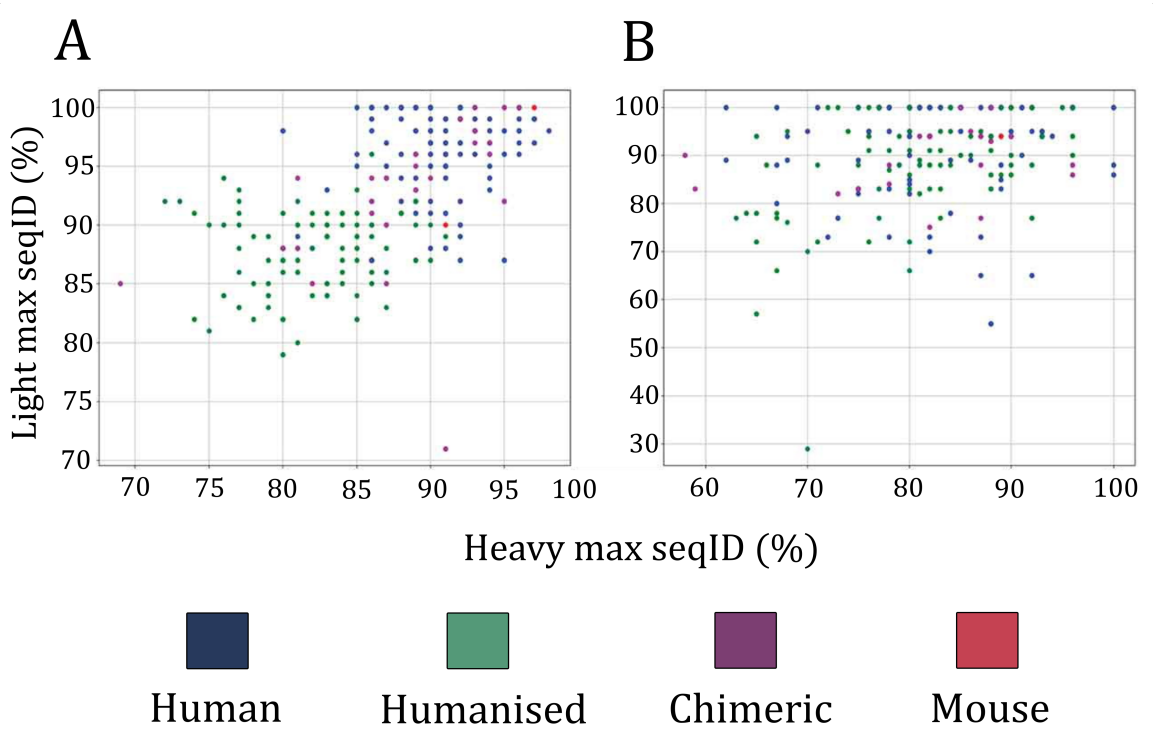


Figure 2.4: Comparing therapeutic sequences to OAS. (A) The highest therapeutic heavy (VH) and light (VL) variable domain sequence identity matches to the BCR repertoire sequences in OAS. (B) The highest therapeutic concatenated heavy CDR (H1-H2-H3) and light CDR (L1-L2-L3) sequence identity matches to the BCR repertoire sequences in OAS. seqID: sequence identity.

combination of the three CST complementarity-determining regions (CDRs) from VH or VL, and CST CDR-H3 sequences to all the sequences in the Observed Antibody Space (OAS) database of BCR repertoire sequencing studies [88] (see Section 2.3.2). We performed the search across all organisms, individuals and immune states to be comprehensive and to reflect the myriad antibody types, including fully human, humanised, chimeric or fully mouse [186]. The individual identities of the CSTs with respect to the closest match from OAS are given in Fig. 2.4 and their distributions are plotted in Fig. 2.5. The aligned sequences and raw sequence identity values are available in the Supplementary Information of Krawczyk *et al.* [202].

The highest sequence identity matches of CST variable regions to naturally sourced BCR repertoire datasets in OAS are given in Fig. 2.4A. Ninety (37.1%) CST heavy chains had matches within OAS of $\geq$ 90% sequence identity (seqID), with 18 (7.4%) $\geq$ 95% seqID. We found 158 (65.2%) therapeutic light chains with $\geq$ 90% seqID to an OAS sequence, with 96 (39.7%) $\geq$ 95% seqID, and 28 (11.5%) with 100% seqID.

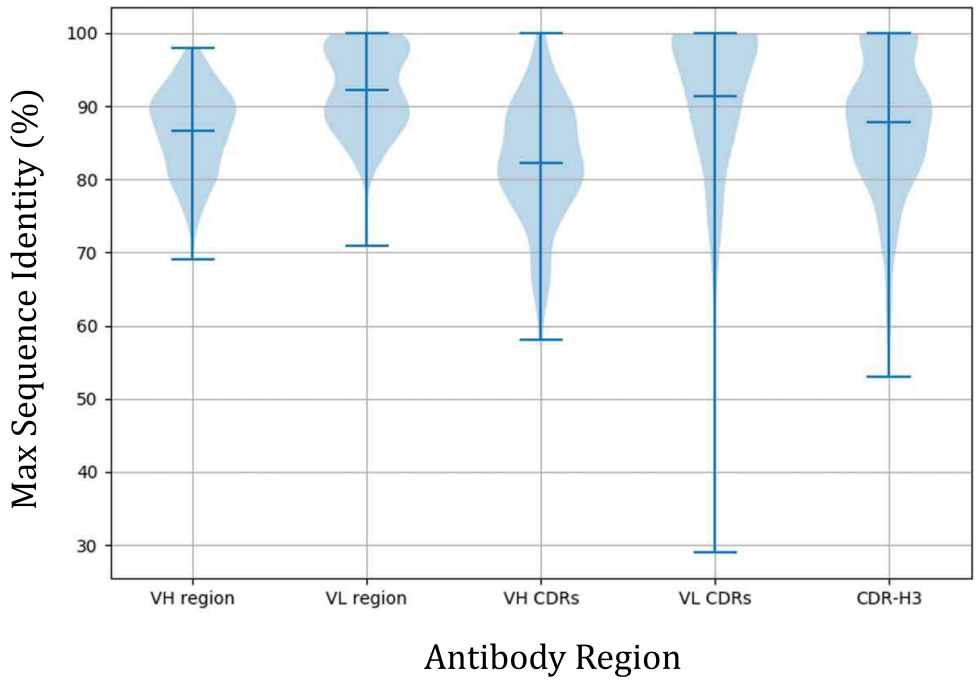


Figure 2.5: Violin plots showing the median, distribution, and ranges of the closest therapeutic sequence identity matches to OAS across different regions of the antibody sequence.

For 16 (6.6%) of the CSTs, we found both heavy and light chain matches $\geq$ 95% seqID. In the most extreme case, Enfortumab, we were able to find both heavy and light chain matches of 98% seqID (the differences are H38: N-S, H88: S-Y, L37: G-S, L52: F-L, where the first amino acid comes from Enfortumab and the second from an OAS sequence).

The largest discrepancy between the CSTs and OAS antibodies is typically concentrated in the CDR regions that determine antigen complementarity [218]. The next investigation measured the extent to which the highly mutable CDR loops of engineered therapeutics differ from those that are expressed naturally, by searching for the closest CST matches to the CDR regions in OAS. The sequence identity was calculated across the entire CDR region testing if all three CDR lengths matched between the CST and a BCR repertoire sequence. The search was performed using the international ImMunoGeneTics information system® (IMGT)-defined CDR triplets [11] from the heavy or light chain, disregarding the framework region (i.e., we concatenated sequences of the CDRH1-3 loops, or CDRL1-3 loops; Figs. 2.4B and

2.5). We found 46 (19.0%) of CST heavy chain CDR triplets to have matches to an OAS CDR triplet with $\geq$ 90% seqID, 15 (6.1%) with $\geq$ 95% seqID and 4 (1.6%) with 100% seqID. There were 156 (64.4%) CST light CDR triplets with $\geq$ 90% seqID to an OAS CDR triplet, with 110 (45.4%) $\geq$ 95% seqID, and 90 (37.1%) with 100% seqID. For Obiltoxaximab and Zanolimumab, we found BCR repertoire sequences where all three heavy and light chain CDRs were identical.

Of the six CDRs, CDR-H3 is the most sequence and structurally diverse [219, 220]. Due to its key role in binding, it is subjected to extensive antibody engineering [221, 222]. We checked how frequently CST-derived CDR-H3s were observed in naturally-sourced sequences by searching for the highest CDR-H3 sequence identity match between each CST and OAS, irrespective of framework region or canonical CDR identity (Fig. 2.5). Of our 242 CST CDR-H3s, we found 54 identical matches in OAS distributed among all antibody types (23 humanised, 22 fully human, 8 chimeric and 1 mouse; 21.9%, 22.0%, 22.8% and 50.0% of each category respectively).

These matches tended to be for shorter CDR-H3s, but some longer loops also recorded 100% sequence identity matches (see Fig. A2.1). We note that finding such close matches is highly unlikely by chance alone, accounting for the potential for sequencing errors throughout the CDRH3 sequence. This is illustrated in Table A2.1, where the probability of matching a single sequence with 960 million random samples is compared for each CDR length. This simple comparison does not account for underlying biases such as the existence of a D gene or inter-residue mutation probabilities, and so is likely to be an overestimate of the realistically-observable sequence space (accurate consideration of these biases is very difficult to achieve). However, particularly for longer CDRs in which D gene influence is diluted, it provides a rough order-of-magnitude contextualisation for the probability of the observed convergence by chance alone — even if understated by a significant factor, the probability of finding five independent 100% sequence identity length-13 CDRH3 matches would be minute.

Twenty-nine therapeutic CDRH3s were found in just one recent deep sequencing study by Briney *et al.* [4], which sampled the diversity of the IgG/IgM human antibody gene repertoires of ten unrelated individuals at unprecedented depth. Forty-seven perfect matches were found in a version of OAS excluding Briney *et al.*, showing that artificial CDR-H3 sequences can be independently observed even in less deeply-sequenced BCR repertoires. Twenty-two of these matches were found in both Briney *et al.* and the other OAS datasets (9 humanised and 13 fully human CSTs).

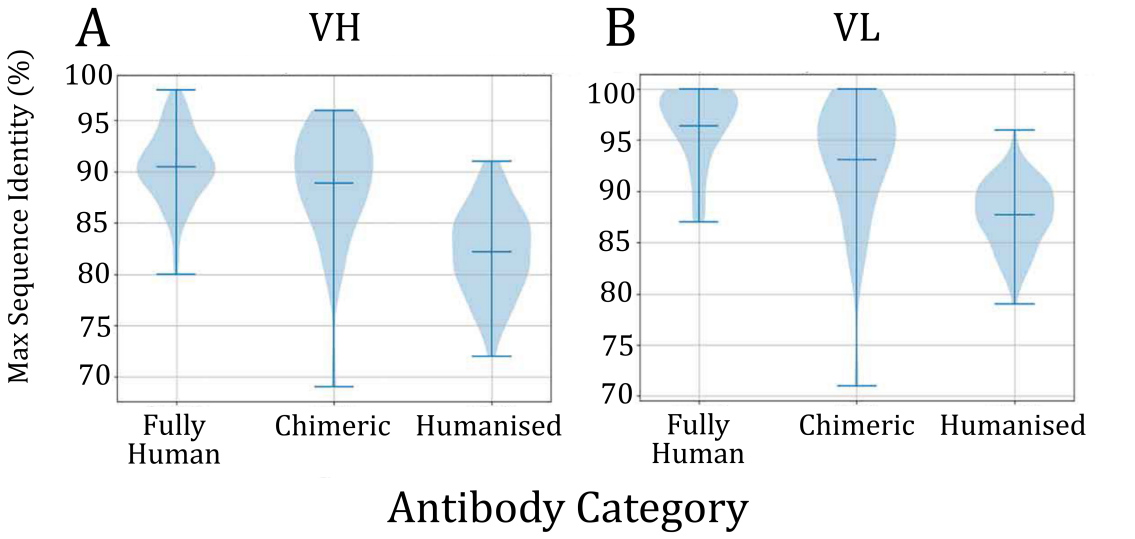


Figure 2.6: Violin plots showing the median, distribution, and ranges of the highest therapeutic (A) heavy chain (VH) sequence identity and (B) light chain (VL) sequence identity matches to OAS binned by developmental category.

These results indicate that, despite the enormous theoretical sequence space accessible to the CDR-H3 region [4], over 20% of proven therapeutically exploitable CDR-H3 loops have been seen in just ∼960 million heavy chain sequences from 60 BCR repertoire sequencing studies.

2.3.3.2 Dependence on Developmental Origin

The proximity of the closest CST variable region match to OAS appears to be highly dependent on the CST's discovery platform (Fig. 2.6). Antibodies produced *via* more artificial protocols such as humanization have lower variable region sequence identities to sequences in OAS from those of fully human molecules. For the majority of the fully human sequences we find matches of 90% seqID or better, whereas matches to the majority of humanised molecules fall below 90% seqID (Fig. 2.6). Chimeric antibodies appear to have seqID values intermediate between the two classes.

The CST developmental origin is usually consistent with the organism that produced the closest BCR repertoire seqID match. Of the 100 fully human CSTs, the 90 (90.0%) most similar heavy chains, 100 (100.0%) most similar light chains, and 55 (55.0%) most similar CDR-H3 loops come from human antibody repertoires. Of the 105 humanised antibodies, 82 (78.0%) of heavy chains, and 79 (75.2%) of light chains found closest matches in human repertoires, while 71 (67.6%) of the closest CDR-H3s

matches were identified in mouse repertoires. This further reflects the dominance of CDR-H3 in binding, as companies often graft this loop from binding mouse antibodies to transfer specificity and binding affinity. It also suggests that mining a dataset such as OAS could provide a more accurate measure of antibody 'humanness' than the state-of-the-art metrics as of April 2019 [193, 223].

2.3.3.3 Comparison to Inter-Therapeutic Similarity

It was previously suggested that finding a therapeutic antibody variable domain that closely overlaps a natural BCR repertoire sequence may cause issues during patenting [224]. To help quantify this risk, we calculated the pairwise similarity between every therapeutic VH and VL (Fig. 2.7) and compared these distributions to the similarities to BCR repertoire sequences reported in Fig. 2.4A.

The multimodal nature of these distributions can be explained by B-cell biology. The right-hand-most peak in both distributions can be attributed to within-germline pairs - *i.e.* the median similarity between two different chains from the same V gene lineage. This is centred at a higher sequence identity for light chains given their absence of a D gene, that they only have one rather than two junctional regions, and that they typically display lower levels of somatic hypermutation. The left-hand peak in the VH distribution and the central peak in the VL distribution can be attributed to comparisons between chains belonging to different V germlines from the same gene locus (IGH-IGH for VH, or 'kappa'-'kappa' [IGK-IGK]/'lambda-lambda' [IGL-IGL] for VL). Most pairwise comparisons of VHs or VLs will involve different V gene origins, accounting for why this this peak has the highest count in both Fig. 2.7A and Fig. 2.7B. Again, this VL peak is centred at a higher median sequence identity owing to the aforementioned genetic differences. Finally, the VL distribution has a third mode, centred at around 47% sequence identity. This reflects comparisons made between VLs originating from the two distinct light chain loci ('kappa-lambda', IGK-IGL), which have markedly different framework sequences. The same phenomenon is not present in the VH distribution, where all chains derive from the IGHV locus. The paucity of therapeutics with lambda light chains (see Chapter 4) explains why inter-local comparisons are of lower abundance than intra-local comparisons, dominated by IGK-IGK.

Only four pairs of CSTs have VH sequence identity matches of greater than 94% to each other. In three of the pairs, both sequences originate from the same company while the fourth is the original patent-expired antibody and its derivative. This compares to 18 therapeutic heavy chains with matches to an OAS sequence $\geq 95\%$.

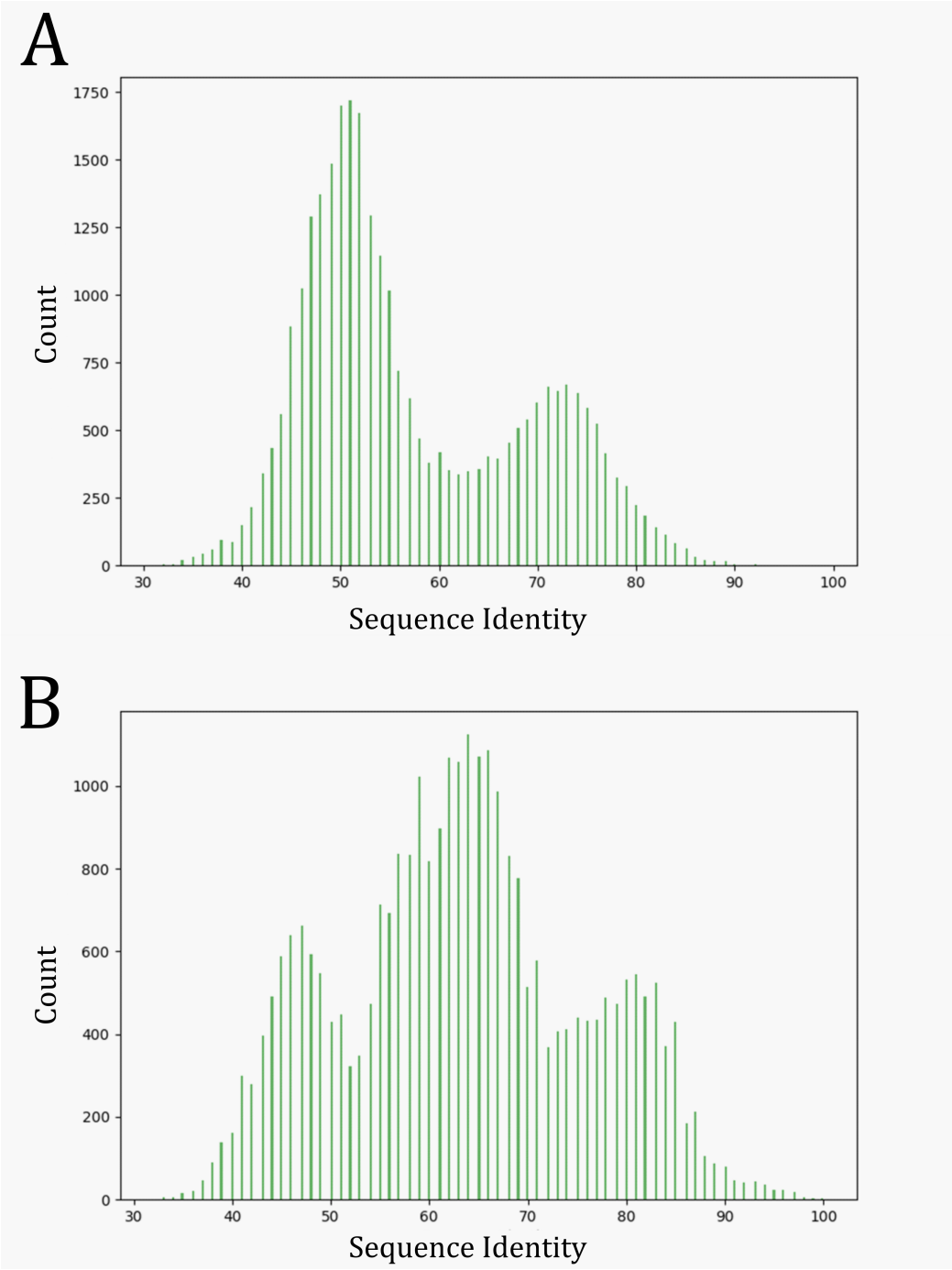


Figure 2.7: The pairwise sequence identities between (A) all therapeutic VH chains, and (B) all therapeutic VL chains. In only four cases is it possible to find VHs across two different therapeutics that are more than 94% sequence identical: Ravulizumab & Eculizumab (Alexion Pharmaceuticals), Ranibizumab & Bevacizumab (Genentech), and Palivizumab & Motavizumab (MedImmune). Tomuzotuximab (Glycotope), is based on Cetuximab (Bristol Myers Squibb), the patent on which expired several years ago.

This confirms that therapeutics have already strayed closer to natural sequence space than to each other, given existing commercial protections. In light of ongoing efforts to further consolidate antibody NGS data and make it more accessible [88, 90, 225], it follows that the risk of finding therapeutic candidate sequences in published BCR repertoire sequencing datasets will only become greater.

2.3.4 Discussion

These results demonstrate that, despite the theoretically large diversity accessible to antibodies, and developmental engineering [4, 226], there exists a nontrivial convergence between CSTs and BCR repertoire sequences. Closest matches were found in 48 of the 60 (80.0%) independent studies available in OAS, indicating that finding a close match to at least one CST is likely in most BCR repertoire datasets. Such a close overlap strongly suggests that it should be possible to data-mine BCR repertoire repositories for promising therapeutic leads.

A letter to Nature in 2018 [224] raised concern that the increasing number of BCR repertoire sequences becoming publicly available may be adjudged to represent 'prior art' that would be considered when assessing the novelty of a therapeutic antibody during patenting[4]. This would hinder progress towards the adoption of natural repertoire mining techniques to discover novel therapeutic agents. In this work, we provided a quantitative basis for discussions in this area. However, a recent article suggests that European Union directive (98/44/EC) and the European Patent Office's (EPO's) Guidelines for Examination leave room to commercialise some naturally occurring human proteins (*https://www.taylorwessing.com/synapse/ti-patenting-gene-sequences.html*). They cite the statement "the prior publication of a sequence of the human genome does not harm the novelty of the sequence claimed in an isolated state", alongside guidance from the EPO (*https://www.epo.org/law-practice/legal-texts/html/guidelines/e/g_ii_5_2.htm*). They also suggest protection in the USA is possible through the interpretation that sequenced complementary DNA (cDNA) does not constitute a naturally occurring product of nature. In addition, we note that several factors pertinent to the specific case of BCR repertoires. Isolation of an antibody sequence from a repertoire still requires molecular characterisation to prove a mechanism of action. Antibodies stimulated by synthetic vaccines may not be adjudged to have occurred 'naturally'. Most BCR repertoire sequencing technologies yield VH and VL chains unpaired, so the entire natural binding site can never be

[4]This was a broader interpretation of the original ruling against Myriad Genetics in 2013 (*https://www.supremecourt.gov/opinions/12pdf/12-398_1b7d.pdf*).

proven[5]. Finally, artificial nucleotide mutations can be introduced at random to antibody sequences by Next-Generation Sequencing techniques as well as during DNA sample preparation. Altogether, we conclude that pharmaceutical companies worldwide should be emboldened to start harnessing BCR repertoire data for therapeutic discovery.

Beyond directly mining repertoires for novel therapeutics, an increased appreciation of the relatedness between engineered antibodies and their naturally expressed counterparts should facilitate the selection of better candidate biotherapeutics, assuming that those that are more closely related have largely favorable biophysical properties [214]. This assertion could be explicitly tested by investigating the covariance of important clinical indicators, such as affinity, immunogenicity and solubility, with measures of similarity to naturally occurring antibodies.

2.4 Update and Chapter Conclusion

Thera-SAbDab continues to record the growing number of therapeutic antibodies and nanobodies entering clinical trials. Since publication, Thera-SAbDab has grown from 473 to 601 (+27.1%) therapeutic Fv/VHH sequences, of which 162 (∼27%) have been structurally characterised. We expect the degree of structural coverage to increase over the coming years, since a large portion of the database is in early-stage trials and structural data tends to be released later in the process. Due to the COVID-19 pandemic of 2019/2020, we also anticipate that SARS-CoV-2 will become the dominant antibody/nanobody target in the database by 2021; according to The Antibody Society, over 100 anti-SARS-CoV-2 biologics had entered development by early June 2020.

The need to rapidly develop safe anti-SARS-CoV-2 prophlyactic antibodies led many companies to attempt direct 'peptide baiting' of therapeutic candidates from the human antibody repertoires of convalescent patients [228] (see Chapter 5). The results of clinical trials on these candidates will be a pivotal moment; on top of the mounting body of computational evidence, success would represent direct empirical proof that BCR repertoires should play a central role in the future of antibody drug discovery.

[5]This is still true for the vast majority of publicly-available data, although this may soon change with the advent of single-cell sequencing platforms from 10X Genomics, Illumina, Takara Bio, and others [227].

In the next chapter, we will describe a novel algorithm ('Repertoire Structural Profiling') that transforms BCR repertoire sequencing datasets into 'Antibody Model Libraries', directly exploitable in drug discovery.

Chapter 3

Repertoire Structural Profiling: Implications for Immunology and Antibody Screening

3.1 Chapter Abstract

In Chapter 2, we outlined the evidence that natural B-cell receptor (BCR) repertoires contain variable domain sequences proximal to antibodies already identified as promising prophylactics. This convergence was particularly striking given that these therapeutics target a wide range of proteins, are likely to have been extensively engineered, and were derived agnostic to their theoretical expressibility in humans. We considered this a strong indication that future drug discovery campaigns could harness the power of human BCR repertoire sequencing to reliably generate human-compatible therapeutics against a broad array of antigens.

However, the diversity of the BCR repertoire in humans is enormous (estimated as on the order of 10^{16}-10^{18} unique variable domains), of which most deep BCR repertoire samples contain 10^{5}-10^{7} unpaired heavy (VH) and light (VL) chains. Given the expanse of natural sequence space, and the potential combinatorial diversity obtained by pairing these unpaired samples, how can researchers hope to efficiently pan the repertoires to find antibody variable domains with a desired complementarity? The solution is to develop methods that can effectively cluster immunoglobulin gene sequencing (Ig-seq) samples of BCR repertoires into sets of potentially functionally-equivalent antibodies for further investigation.

In this chapter, we describe our novel method, 'Repertoire Structural Profiling', which uses predicted variable domain (both VH and VL) binding site structure as a means of clustering repertoire samples. This is based on the theory that a highly

conserved property across same-epitope binding antibodies will be the adoption of similar structures, as this avoids prohibitive steric clash and enables chemically-complementary residues to be displayed within their interaction distance of the epitope. Our method does not impose any restrictions on predicted V or J gene origins, so is therefore able to pool together a much greater sequence diversity of antibodies than clonotyping[1]. Instead, the binding site structural topology of each putative VH:VL sequence pairing is first approximated by mapping each of its 6 CDR loop sequences onto a structural template from the PDB. Greedy 'structural clustering' then uses a precomputed matrix of root-mean-squared distances (RMSDs) between all CDR structural templates to evaluate whether two binding sites are likely to adopt 'distinct structures' (see Section 3.3). This allows us to approximate the structural diversity of an Ig-seq dataset without modelling each VH:VL pair, which is computationally intractable. Finally, a representative VH:VL sequence pairing from each predicted distinct structure is formally homology-modelled to yield a full 3D representation of the binding site for use in structure-based drug discovery.

The new representation of BCR repertoires we present here provides the first computational supporting evidence for baseline (naïve) repertoire functional commonality across many individuals and can detect convergent structural drifts in response to vaccination analogous to convergent clonotypes. We also demonstrate, by building an 'Antibody Model Library' from the baseline structures predicted to exist in ten unrelated individuals, that shared variable domain geometries are proximal to many solved therapeutic antibody structures against a large diversity of targets. This highlights Repertoire Structural Profiling's potential to capture the pluripotency of the naïve repertoire, allowing for the rational design of general *in vitro* or *in silico* screening libraries.

The chapter concludes with suggestions for further benchmarking and improvements to the technology.

This chapter contains reproduced material from the following preprint:
Raybould, M.I.J., Marks, C.M., Kovaltsuk, A.K., Lewis, A.P., Shi, J., Deane, C.M. (2020) Evidence of Antibody Repertoire Functional Convergence through Public Baseline and Shared Response Structures. *bioRxiv*. doi: 10.1101/2020.03.17.993444 [229]

[1]Clonotyping is an alternative and more established method for repertoire clustering based on matched genetic origins and high CDR3 sequence identity; see Section 1.4.4.

3.2 Introduction

A key component of the human immune system is the antibody/B-cell receptor (BCR) repertoire, a diverse array of immunoglobulins tasked with identifying pathogens and initiating the adaptive immune response. Broad pathogenic recognition is achieved through enormous variable domain sequence diversity, with an estimated 10^{10} unique heavy variable domains (VH) circulating at any one time from a theoretical set of 10^{12} (or 10^{16}-10^{18} full antibodies if light variable domain (VL) combinations are considered [4]).

On antigenic exposure, 'baseline' (resting-state) antibodies with sufficiently complementary binding sites to an antigen surface epitope are positively selected. The corresponding parent B-cells subsequently migrate to the marginal zone of the lymph nodes, where intentional mutations are introduced to their sequence and only the highest-affinity binders survive in the competition for cognate T-helper cells [2].

Therefore, sequencing antibody repertoires before and during an immune response (e.g. vaccination) can reveal how different people respond to the same antigenic challenge, and can both improve our understanding of immunology and inform future vaccine or therapeutic design [84, 230, 231]. Similarly, comparing the repertoires of healthy individuals against immunosuppressed (*e.g.* HIV) patients may also reveal the origins of increased disease susceptibility [232–234].

However, sequencing an entire antibody repertoire is challenging; they are so large that conventional sequencing techniques, such as Sanger sequencing, do not capture enough of the diversity to be informative. Instead, high-throughput Ig-seq technologies (e.g. Illumina MiSeq) are used (see Section 1.4.1.1). These methods create snapshots that are typically on the order of 10^6-10^7 VH and/or VL (unpaired) chains, up to a recent upper bound of around 10^9 [3, 4, 88]. Single-cell sequencing methods, capable of preserving VH-VL chain pairings, are now emerging, however their current throughput yields datasets that are too small to study entire repertoire diversity [70, 72, 235] (see Section 1.4.1.2).

Since most publicly-available BCR repertoire data covers only the VH domain [88], the vast majority of whole-repertoire analysis has been performed over this region alone [91]. The primary analytical method is currently 'clonotyping' [236–238]. Clonotyping is a computational technique used to sort sequencing datasets into sets of functionally similar chains based on sequence features, and can be performed in several ways. The most common implementation groups sequences with the same

predicted V and J gene transcript origins and above a certain percentage (same length) Complementarity-Determining Region H3 (CDRH3) sequence identity.

Such sequence-based approaches have contributed significantly to our knowledge of core immunology. For example, to estimate the true level of sequence similarity that exists across individuals, Briney *et al.* performed deep sequencing and clonotyping of the circulating baseline VH repertoires of ten volunteers [4]. They found that just 0.022% of observed clonotypes were 'public' (seen in everyone) and a similar study by Soto *et al.* found just ∼1% of clonotypes were public across three unrelated individuals. In a complementary approach, Greiff *et al.* trained a Support Vector Machine on public and private clonal sequences to identify their high-dimensional features, proving that they have distinct immunogenomic properties [92].

Clonotyping can also be used to detect antigen-specific immunoglobulins, through the identification of expanded clones after vaccination, or those present in unusually high proportions in individuals immune to certain diseases. Explorations of expanded lineages have yielded high-affinity antibodies and T cells against numerous pharmacologically interesting antigens, such as HIV proteins [232], cluster of differentiation proteins [239], botulinum neurotoxin serotype A [240], proteins implicated in type-1 diabetes [241], and many more.

However, clonotyping is only likely to identify a small subset of the true number of functionally equivalent antibodies. This is because it assumes that antibodies require a similar genetic background and high CDRH3 sequence identity to achieve complementarity to the same epitope. In reality, similar binding site structures and paratopes can be achieved from different genetic origins [242] and with surprisingly low CDRH3 sequence identity [127] (conversely, false positives can arise where antibodies with high CDRH3 sequence identity and the same genetic origins adopt markedly different binding site topologies [127]). It is also the case that not every epitope is naturally suited to CDRH3-dominated binding, instead preferring broader engagement by multiple CDRs [9], putting less selection pressure on CDRH3 sequence identity.

It is difficult to reliably pool together these hidden functionally equivalent antibodies within a clonotyping framework, as simply reducing the CDRH3 sequence identity threshold value lowers confidence in paratope residue similarity and increases the risk of grouping antibodies with fundamentally different binding site topologies. An alternative approach to relaxing the clustering criterion would be to initially ignore CDRH3 residue similarity, and instead to group antibodies with similar three-dimensional structures, as binders to a given epitope are likely to adopt a similar ge-

ometry. Geometrically-similar antibodies with sufficiently similar residue interaction profiles could then be capable of recapitulating key binding interactions at equivalent topological locations.

Experimental structure determination (e.g. by X-ray crystallography) remains too slow to solve representative portions of antibody repertoires [243]. However, structural annotation approaches are now fast enough to geometrically characterise the individual CDRs of millions of sequences a day with increasing accuracy [96, 244]. So far, these analyses have focussed solely on the VH chain, and none have considered the impact of VL on binding site configuration. This can most accurately be captured through variable domain (Fv) modelling, and recent developments have afforded homology approaches with sufficient throughput to analyse meaningful portions of the repertoire [44, 245]. For example, we developed a prototype structural profiling method that creates representative Fv model libraries from large repertoire snapshots, with applications in developability issue prediction (the Therapeutic Antibody Profiler [214], described in Chapter 4).

Here, we further refine this prototype and apply it to cluster antibody repertoires based on predicted binding site topology. We first analyse 41 naïve antibody repertoires from unrelated individuals, and find that the same representative ('distinct') binding site structures are predicted to appear across many individuals ('Public Baseline' structures). We also show, through the construction of 'Random Repertoires', that this level of structural sharing is far greater than would be expected by chance. Our data therefore represents the first supporting computational evidence that considerably more functional commonality than suggested by clonotyping could exist in the baseline repertoires of different people. We then implement the same pipeline on pre- and post-vaccination datasets from three unrelated individuals, detecting a significant increase in structural commonality, and identifying all convergent response structures that may recognise similar epitopes ('Public Response' structures). We built Antibody Model Libraries (AMLs) by homology modelling a VH-VL sequence pairing predicted to adopt each Public Baseline or Public Response structure. *In silico* analysis of these AMLs suggests that they represent a powerful geometric basis set of low-immunogenicity candidates exploitable for general or target-focused therapeutic antibody screening.

3.3 Methods

Immunoglobulin Gene Sequencing Datasets

The cleaned and translated antibody repertoire datasets [83, 84] were downloaded directly from the Observed Antibody Space (OAS) database [88]. For the naïve baseline repertoire study (Gidoni *et al.* [83]), only individuals for whom > 100,000 IgM VH and > 100,000 VL sequences were recorded were analysed. For the influenza vaccination response study (Gupta *et al.* [84]), we used all three individuals ('V1' = 'FV', 'V2' = 'GMC', and 'V3' = 'IB'). The 'Before Vaccination' data was defined as all VH and VL sequences recorded at 8 days, 2 days and 1 hour before vaccination. The 'After Vaccination' data was defined as all VH and VL sequences recorded at 1 week, 2 weeks, 3 weeks, and 4 weeks after vaccination. Sequences recorded 1 hour and 1 day after vaccination were discarded to avoid ambiguity. The 'Pure After Vaccination' data contained 'After Vaccination' sequences that did not fall into the structural clusters defined by each individual's 'Before Vaccination' repertoires. The seminal work in which 'FV', 'GMC', and 'IB' were vaccinated is detailed in Laserson *et al.* [231], however the data we use derives from Gupta *et al.* [84], who re-analysed each antibody repertoire snapshot with Illumina sequencing.

Repertoire Structural Profiling Pipeline

The described structural profiling pipeline was optimised from the protocol outlined in Chapter 4 [214] as follows:

Using FREAD to predict canonical loop modellability. CDRH1-2 and CDRL1-3 are termed the 'canonical' CDR loops. This is because, for each category of CDR, their backbone conformations can be broadly clustered into a relatively small number of distinct canonical forms (though a significant minority of uncluster-able structures remain [25]). In the first AML generation procedure [214], we used SCALOP, a machine learning algorithm that can reliably predict canonical forms from loop sequence alone [25]. We assumed that if SCALOP can assign the loop to a canonical form, it is likely to be 'FREAD-modellable'. The main benefit here was anticipated to be speed; using SCALOP as a pre-filter to FREAD would ensure that not every sequence has to pass through the more time-intensive processes of framework assignment and loop grafting.

However, not every SCALOP loop conformation prediction agrees with FREAD's predictions and imposing canonical form conformity on each sequence has the consequence of narrowing accessible extra-CDRH3 structural diversity. Additionally, as FREAD has to be run on all modellable VH/VL CDRs at some point in the algorithm, and as the number of loops trimmed out for canonical CDR non-modellability is relatively small, the time saved by running SCALOP was a small fraction of the overall algorithm run-time. The additional time taken to pass every chain through FREAD directly rather than *via* SCALOP was nearly entirely compensated by adapting an efficiently parallelised version of FREAD written by Jinwoo Leem for the *abodybuilder_parallel* module [44].

Deriving More Appropriate Environment-Specific Substitution Scores Thresholds for Ig-seq Data. The original antibody Environment-Specific Substitution (ESS) score threshold used by FREAD (25 for all loops, and loop lengths) was derived by Choi and Deane [26, 27] based on benchmarking of loop modelling performance using the Protein Data Bank (PDB). This involved first excising the CDRH3 loop from known PDB antibody structures, then feeding the framework structure and excised CDRH3 loop sequence into FREAD tasked with predicting the next-best (non-self) PDB CDRH3 template for modelling. The achieved root-mean-squared deviations (RMSDs) were averaged for different thresholds and an ESS cutoff of ≥ 25 proved to be the best compromise between accuracy and coverage (proportion of loops for which at least one template exceeded the threshold). However, the CDRH3 loops in Ig-seq samples of natural antibody repertoires deviate significantly in sequence from the heavily-engineered PDB, which has high redundancy and many closely-related structures that help to improve mean performance. To show this, we calculated the typical ESS value of the top-ranked FREAD CDRH3 template for an Ig-seq study of natural antibodies [84] (Fig. 3.1A), comparing it to the ESS of the top-ranked FREAD CDRH3 template obtained when modelling-in PDB CDRH3 loops (Fig. 3.1B), blinding access to the same structure as a template). The highest ranked template has the lowest anchor residue C_α RMSD, after surpassing the baseline threshold of ESS 25. The Ig-seq data benefits from far fewer high ESS scores, so application of the original thresholds on Ig-seq data would be expected to achieve below-headline performance. To both estimate our true performance on Ig-seq data, and derive new CDRH3 loop ESS score thresholds more appropriate for Ig-seq data, we subsampled the PDB comparison set to match the top-ranked ESS distributions seen for each length bin in the aforementioned Ig-seq sample. Based on this sample, we calculated that we should achieve a good accuracy (mean RMSD of 2.54Å over

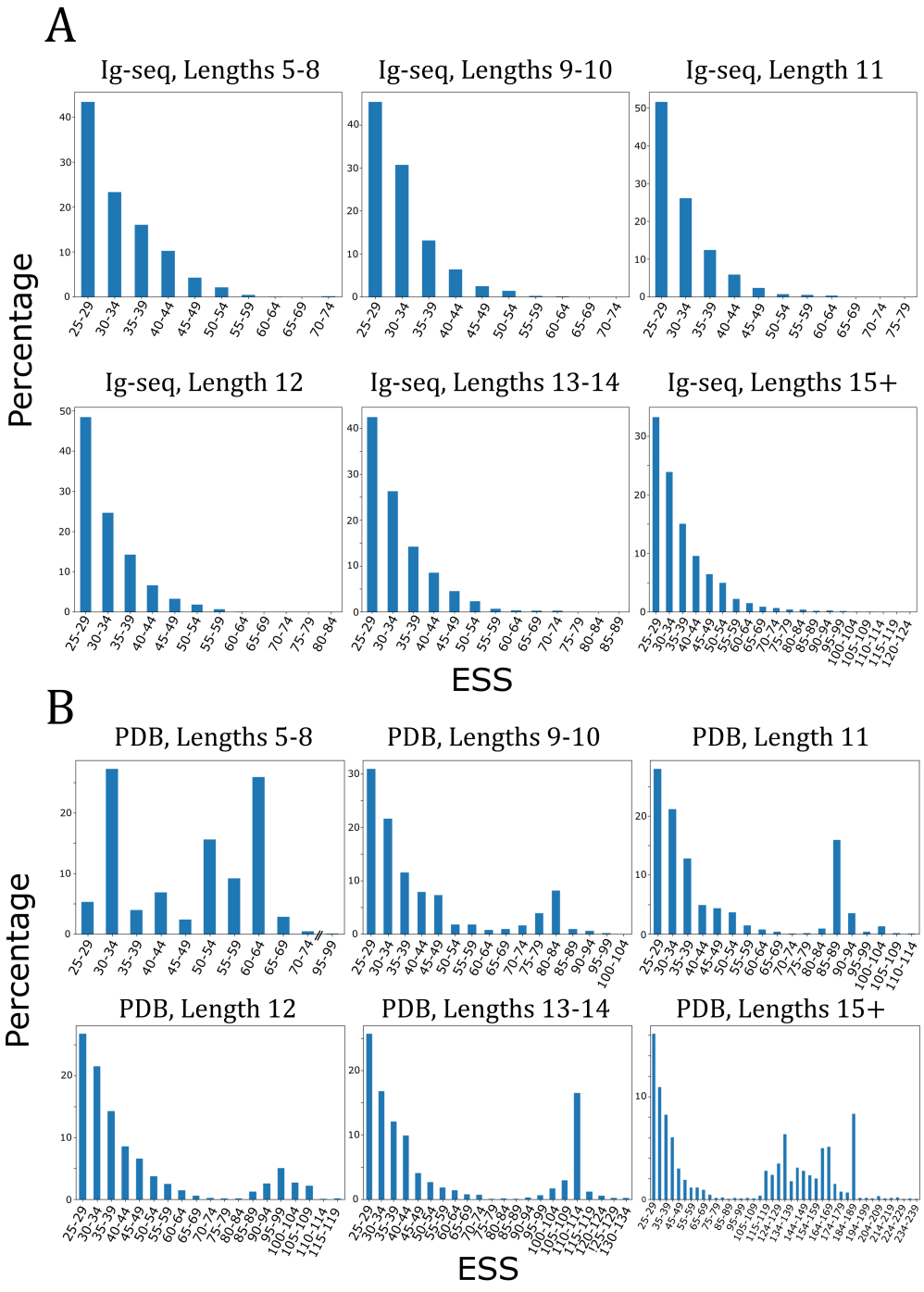


Figure 3.1: The percentage of each FREAD [26, 27] top-ranked CDRH3 templates with an Environment Specific Substitution (ESS) score within the labelled bin for (A) a typical Ig-seq dataset [84], and (B) the Protein Data Bank (blinded to self). The two sets have very different distributions; notably Ig-seq datasets rarely contain CDRH3 loops with extremely high ESS scores to dataset templates.

all lengths) with acceptable coverage on Ig-seq data using the following CDRH3 ESS cutoffs: Lengths 5-8, ESS $\geq$ 25; Lengths 9-10, ESS $\geq$ 35; Lengths 11+, ESS $\geq$ 40.

Determining the Most Important Residues for VH-VL Interface Orientation Prediction. Billions of comparisons are made between different VH and VL chains to determine which pairings have orientations that are considered modellable (*i.e.* have an interface residue identity greater than a threshold value to at least one reference Fv). Each pairwise comparison in the original version of this algorithm (see Section 4.3) considered all 120 residue positions found to be less solvent-exposed in the context of a complex than they were when the co-ordinates of their partner chain (VH if a light chain position, or VL if a heavy chain position) were deleted.

This interface comparison step was very time-intensive and so we sought to reduce the number of compared residue positions to a smaller set of key interface positions. We downloaded the 1,129 sequence non-redundant Fvs with resolution $\leq$ 2.5 Å from the SAbDab database [9] (12$^{\text{th}}$ February 2019) and identified all residues found to lie in the VH-VL interface. This was achieved by first calculating the relative solvent accessible surface area ($SASA_{rel}$, Shrake-Rupley Algorithm [246]) to determine the absolute SASA of each residue and dividing this number by the theoretical maximum SASA for that residue. The $SASA_{rel}$ of each position in the complex was then compared to the value for the equivalent position in the separated VH and VL chains (coordinates of the partner chain deleted), yielding the 120 positions that increase in surface-exposure. These were trimmed to 52 positions that appeared in at least 80% of complexes, and whose $SASA_{rel}$ was reduced by an average of at least 5% in at least 10% of those complexes. Their identities are listed in Fig. 3.3.

To reduce this number further, we performed a Random Forest regression analysis [standard scikit-learn implementation, 500 estimators] over these 52 residues to the 6 ABangle parameters [20] that have been shown to characterise VH-VL orientation. Firstly, we confirmed that the 1129 interfaces constituted a representative sample of all ABangle parameter space — essential to learn genuine residue to ABangle parameter responses. Each interface was then flattened and one-hot-encoded (21 columns for each position, for the 20 natural amino acids or a deletion/missing residue) to yield a 1129x1092 matrix which was separately regressed against each ABangle parameter (6 x (1129x1) vectors). Out-of-bag validation was used to estimate R^2 values, and showed predictive performance ranging from an estimated R^2 of 0.35-0.54. We calculated feature importance, and derived a new one-hot-encoded interface for each complex using only the 20 positions (Fig. 3.3) that were present in the top-5 most

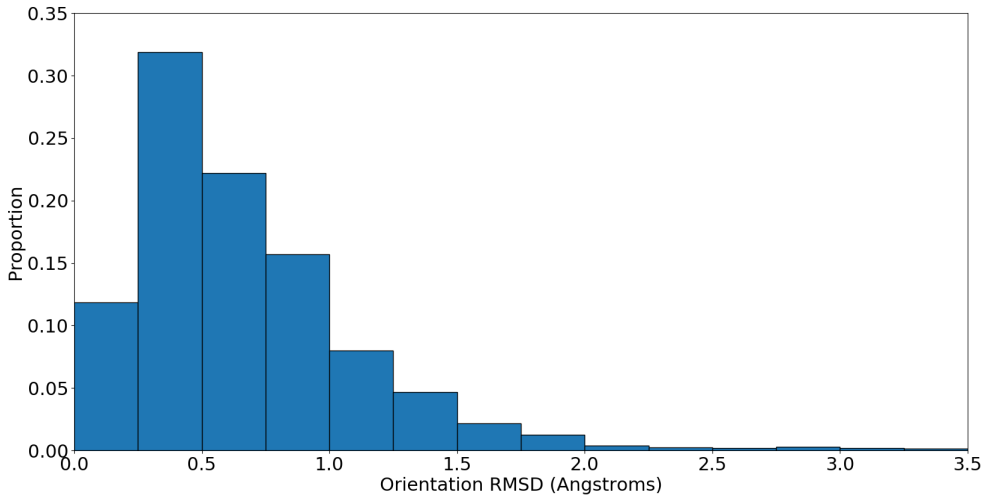


Figure 3.2: The distribution of orientation RMSDs observed between Fvs of identical heavy and light chain sequence. The vast majority (92%) have orientation RMSDs below 1.5Å.

important features across the 6 parameters. Predicted performance dropped by a mean estimated R^2 of only 0.05 across the six parameters (min: 0, max: -0.11).

The coverage and accuracy of the 52 and 20 residue interface definitions at predicting orientation RMSD was also assessed. Orientation RMSD between two complexes was measured by first aligning their VH domains and measuring the C_α distances between common VL positions, then by aligning their VL domains and measuring the C_α distances between common VH positions, and finally dividing by two. A 'correct/incorrect' orientation threshold RMSD of 1.5 Å was chosen by measuring the variation in pairwise orientation RMSD observed for sequence-identical SAbDab complexes, *i.e.* an estimate for the experimental limitation on what constitutes a 'correct' orientation RMSD. A threshold of 1.5 Å captures 92% of sequence identical Fvs (see Fig. 3.2). An orientation sequence identity threshold was then chosen for the 20-residue and 52-residue interface definitions that balanced acceptable coverage with a high proportion of Fvs within 1.5 Å above the threshold (see Fig. 3.4). We chose 82% for the 52-residue definition, and 85% for the 20-residue definition. Coverage was comparable and accuracy only slightly reduced (80.2% to 77.8%) on narrowing the interface definition to 20 residues.

Interface Residues	
Heavy Chain	Light Chain
44	36
47	37
48	38
49	40
50	42
52	44
55	47
57	48
66	49
68	**50**
69	52
70	55
101	**56**
103	66
107	68
108	**69**
109	**101**
113	**103**
114	105
115	107
116	108
117	**109**
118	114
119	**115**
120	**116**
	118
	120

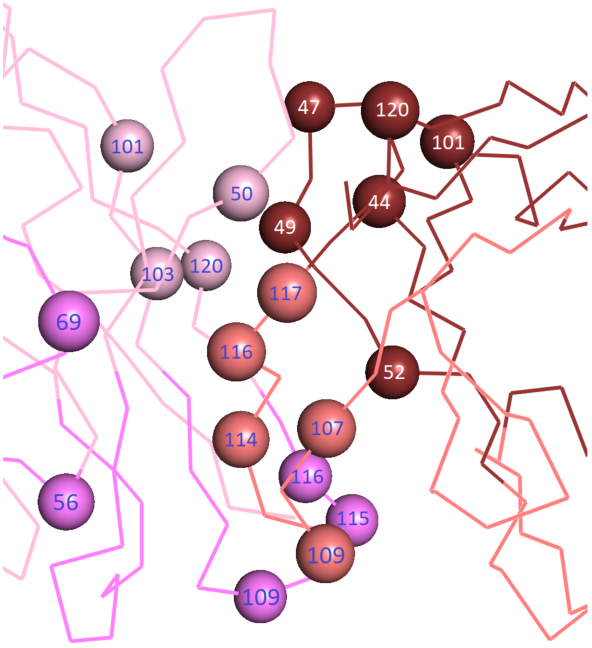


Figure 3.3: The 52 heavy and light chain IMGT-numbered [11] residues tending to lie in the heavy-light chain interface. Residue numbers in bold were determined to be amongst the five most important in the Random Forest regression model when predicting the six different ABangle parameters. A pymol molecular visualisation of a typical Fv region is provided, with the backbone chain displayed as a ribbon; the heavy chain is coloured dark red and the light chain is coloured pink. Heavy chain North-defined [15] CDR regions are highlighted in 'salmon red', while light chain North-defined CDR regions are highlighted in 'violet'. The C_α atoms of the 20 most important residues are highlighted by spheres and are labelled, to show their relative disposition within the VH:VL interface.

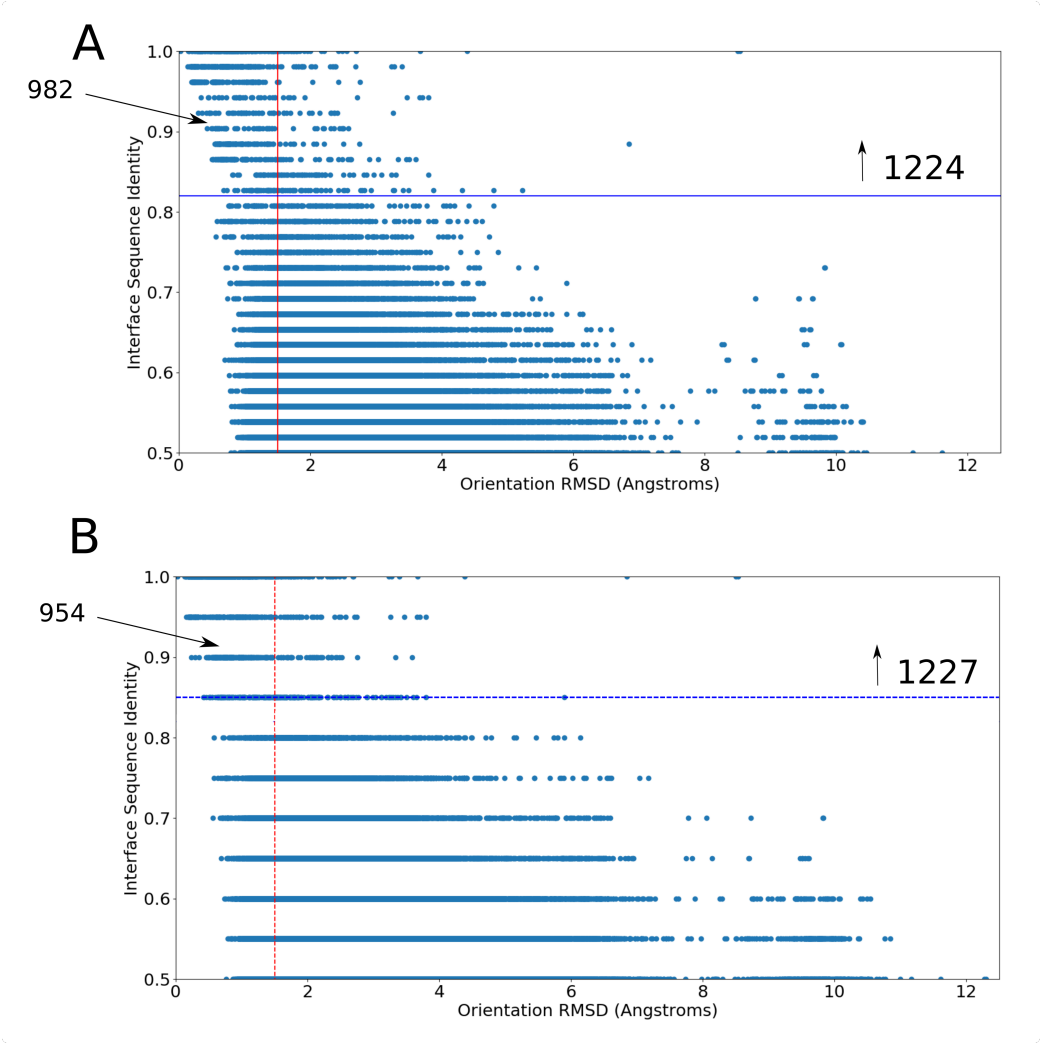


Figure 3.4: Graphs showing the orientation RMSD observed at each interface sequence identity value for (A) all 52 interface residues and (B) the 20 most important interface residues. The thresholds for (A) are set at 1.5Å and 82% sequence identity, while for (B) are set at 1.5Å and 85% sequence identity. The proportions above the sequence identity threshold and within 1.5Å orientation RMSD are 80.2% (982/1224) and 77.8% (954/1227) respectively.

Pipeline Summary

CDR Modellability Analysis. Each BCR repertoire sequence was first numbered using ANARCI [18] according to the IMGT numbering scheme [11], and the closest framework region (variable domain with North-defined CDRs [15] excised) in the SAbDab [9] database (12$^{\text{th}}$ February 2019) was identified. In the IMGT numbering scheme, the North CDRs lie between the following residue numbers: CDRH1: 24-40; CDRH2: 55-66; CDRH3: 105-117; CDRL1: 24-40; CDRL2: 55-69; CDRL3: 105-117.

FREAD [26, 27] was then used to attempt to map each BCR repertoire sequence to a length-matched North CDR template. The FREAD CDR databases contained all structures released up to 12$^{\text{th}}$ February 2019 that passed the resolution and B-factor quality controls [27], with the following numbers of templates per region - CDRH1: 2,526; CDRH2: 2,575; CDRH3: 2,502; CDRL1: 2,355; CDRL2: 2,373; CDRL3: 2,376. Templates were not restricted only to those with 'human' PDB organism assignments for multiple reasons. Antibodies in the PDB are highly engineered, both through point residue mutations and entire loop transplantation, meaning single organism origin labels are only accurate for a small number of entries. In addition, internal benchmarking of FREAD [26, 27] and ABodyBuilder [44] showed that including "non-human" templates in our FREAD loop databases (particularly the CDRH3 database) leads to greater structural coverage and a net improvement in CDR structure prediction accuracy. All loop templates contain the North-defined CDR loop and 5 'anchor residues' before and after the loop. Selection of CDRH3 templates was performed according to a bespoke set of Environment-Specific Substitution Score (ESS) thresholds established using Ig-seq data: Lengths 5-8, ESS $\geq$ 25; Lengths 9-10, ESS $\geq$ 35; Lengths 11+, ESS $\geq$ 40 (see above). Each template surpassing the threshold was subsequently grafted onto the corresponding framework anchor residues. The loop template with the lowest calculated C_α anchor RMSD was selected. Any sequences for which at least one loop could not be modelled were removed from the dataset.

Sequence Clustering. The modellable chains were then sequence clustered using CD-HIT [247] at a 90% sequence identity threshold and requiring all cluster members to be of the same length. CD-HIT is a greedy clustering algorithm that reduces the complexity of sequence clustering from the theoretical order of N^2 *via* 'short-word filtering', an algorithm which precludes unnecessary comparisons between chain sequences of very low sequence identity through substring analysis. Overall, this clustering step trims the number of VH-VL pairing comparisons to a computationally-tractable number.

Predicting Modellable VH-VL Orientations. The 20 most important VH-VL interface residues for orientation prediction were derived (see above); a sequence identity of 85% over these 20 residues resulted in an orientation RMSD of ≤ 1.5Å $\sim$ 80% of the time.

All remaining VH and VL domains after sequence clustering were paired together, and their 20 key interface residues were recorded. If the sequence identity over these residues was $\geq 85\%$ to at least one of 1,129 reference Fvs, the interface was classed as modellable — otherwise the VH-VL pairing was discarded. If multiple reference Fvs shared $\geq 85\%$ identity, the predicted modellable VH-VL pairing inherited the orientation parameters of the Fv reference with highest sequence identity.

Identifying Distinct Structures. At this stage, each predicted modellable VH-VL pairing (Fv) has eight associated parameters: its orientation template, its six CDR templates, and a length vector recording the combination of North CDR lengths [15] present in its binding site. Fvs were then structurally clustered to identify 'distinct structures' according to the following process. First, identically-predicted binding sites (for which the eight predicted parameters were the same) were identified. The retained pairing was randomly chosen, except in the overlap studies — if one of the pairings was present as a distinct structure of the first dataset, this pairing was selected and recorded as a shared structure across both repertoires.

Next, singleton length clusters (where a unique combination of 6 CDR lengths was observed in a modellable Fv) were identified and assigned as separate distinct structures, avoiding RMSD comparisons between loops of differing length. The remaining interfaces were split by their CDR length combinations, and were greedily clustered with all other pairings sharing the same length vector as follows:

1. Select the first pairing as a distinct structure (cluster centre).
2. Select the next pairing. If the orientation RMSD to all existing cluster centre orientation templates exceeds 1.5 Å, classify the new pairing as a distinct structure. Otherwise:
3. Calculate the RMSD between the CDR templates of the new pairing with those of all existing cluster centres using the formula:

$$\sqrt{\frac{\sum_{X}^{(H1-H3,L1-L3)} D_{X_{12}}^{2} L_{\mathrm{X}}}{\sum_{X}^{(H1-H3,L1-L3)} L_{\mathrm{X}}}}$$

where the sum over X refers to each of the six CDRs, $\mathrm{L_X}$ is the length of North CDRX, and $\mathrm{D}_{X_{12}}$ is the C_α RMSD between the CDRX in Fv 1 and Fv 2. If this

value exceeds 1 Å to all existing structural cluster centres, the pairing is assigned as a distinct structure. Otherwise the pairing is stripped from the dataset.
4. Return to step 2 until all pairings have been analysed.

Overlap Comparison

To identify shared structures between two BCR repertoire snapshots, the distinct structures from the first snapshot were listed followed by all predicted modellable Fvs of the second repertoire snapshot, as an input file to the clustering algorithm. The greedy clustering ensured that all distinct structures from the first dataset remained as cluster centres, and allowed for the identification of pairings in the second dataset that were also predicted to occupy the same structural neighbourhood.

'Random Repertoires'

To contextualise the structural diversity displayed in human antibody repertoires, we derived 'Random Repertoires' (RRs) according to the following method. First, a set of Modellable Repertoire Structures (MRS) was generated. When generating a structure, one of any of 663 orientation templates, 2,051 CDRH3 templates, and 2,125 CDRL3 templates previously assigned by FREAD to a human Fv/CDR sequence were available for selection. To mirror the genetics of the immune system (as they would be encoded on the same V gene transcript), CDR1 and CDR2 templates were restricted to being selected from the same SAbDab structure, limiting our choice to one of 789 CDRH1/2 templates and 912 CDRL1/2 templates, again all of which FREAD had previously assigned to human sequences. Each of these five sets was randomly sampled over 180 million times to create the MRS dataset. This was then filtered to create 41 Length-Accessible Repertoire Structure (LARS) datasets, containing only the combinations of CDR lengths observed in each baseline repertoire snapshot. Finally, RRs were created by sampling each LARS set the same number of times as the number of predicted modellable Fvs in the corresponding baseline repertoire snapshot. For example, 'RR_S64' was created by sampling the subset of the MRS with a CDR length combination seen in the natural S64 repertoire 6,420,211 times (the same as the number of modellable Fvs predicted during Repertoire Structural Profiling).

To obtain statistically expected values for structural overlap across individuals, the distinct structures from 'RR_S64' were randomly subsampled the same number of times as the number of distinct structures seen in 'S64' itself, yielding random

distinct structure samples occupying the same proportion of LARS-space. The process was repeated for each RR dataset, normalising to each respective baseline repertoire snapshot. Overlap comparison was then performed as described above, starting from the 'RR_S64' distinct structures, followed by all the pairings that fell into the selected 'RR_S57' distinct structures, *etc.*

Germline Proximity

The closest heavy and light V gene germline to each assessed antibody was evaluated using ANARCI [18] aligning to IMGT-supplied [11] germlines. Sequence identity to germline was then calculated at the amino acid level.

Clonotyping

Clonotyping was performed to group antibodies with the same closest V and J gene, and either identical CDRH3 sequences, as in Briney *et. al.* [4], or with CDRH3s within 80% sequence identity, as in Soto *et al.* [80].

Antibody Model Library Construction

Antibody model libraries (AMLs) were constructed with a parallel implementation of ABodyBuilder [44], using the FREAD [26, 27] Environment Specific Substitution Scores derived from Ig-seq benchmarking (see above). Some predicted modellable Fvs are not entirely homology modellable, as loop modellability is considered on a per-chain basis and does not take into account inter-chain loop clashes that become evident only upon full Fv homology modeling. Fvs that required any degree of *ab initio* modelling to fix such issues were trimmed out of the dataset.

Structural Comparison to Antibody Therapeutics

The set of 89 therapeutics with 100% sequence identical structures (as of November 2019) were retrieved from Thera-SAbDab [99]. A single structure was chosen for each therapeutic for the RMSD analysis; if multiple structures were available, we selected unbound structures with the best resolution. RMSD comparisons were only made between therapeutics and AML structures with matching combinations of CDR lengths. Fv RMSD was calculated over all C_α atoms after alignment of backbone atoms, using an in-house script.

3.4 Results

The 'Pipeline Summary' in Section 3.3 outlines our 'Repertoire Structural Profiling' algorithm in detail. As a brief overview (Fig. 3.5), Repertoire Structural Profiling takes as input an antibody/BCR repertoire snapshot containing heavy (VH) and light (VL) chain reads. It eliminates VH and VL chains for which not every CDR is modellable. All modellable VH and VL chains are then sequence clustered to reduce computational complexity. Surviving cluster centres are then paired together and the resulting Fvs that are likely to be successfully modelled are retained. Finally, predicted modellable Fvs with the same combinations of CDR lengths are structurally clustered based on the orientation and CDR loop templates forecast to be used during homology modelling. Antibody Model Libraries (AMLs) can then be built from these representative Fv sequences.

This study comprises two main investigations. Firstly, we analyse a Ig-seq study by Gidoni *et al.* [83] to investigate the degree of structural overlap in the circulating baseline naïve antibody repertoires of many unrelated individuals (Section 3.4.1). From this we derive an 'Antibody Model Library' based on the predicted structures common to ten individuals, and analyse its properties. Secondly, we analyse a longitudinal Ig-seq flu vaccination study by Gupta *et al.* [84] to measure three individuals' structural responses to exposure to a common antigen (Section 3.4.2).

3.4.1 Structurally Profiling the Baseline Immune Repertoire

3.4.1.1 Evaluating the Numbers of Distinct Structures in Real Repertoires

We first investigated the structural diversity present in the 41 selected Gidoni baseline repertoire datasets. Separately, each dataset was fed through our Repertoire Structural Profiling pipeline to identify the set of sequence diverse modellable VH and VL domains, then the number of predicted modellable Fvs, and finally the number of distinct structures in each dataset (Table 3.1, full table is available in the Appendix, Table A3.1).

The most structurally diverse dataset was 'S64' (209,394 distinct structures from ∼6.4M Fvs), and the least was 'S4' (78,588 distinct structures, from ∼750K Fvs). Datasets with a larger number of modellable sequence diverse VHs tended to result in a larger number of distinct structures. Datasets with a moderate/low number of modellable sequence diverse VHs but very large numbers of modellable sequence diverse VLs were amongst the least structurally diverse (*e.g.* 'S95'). This is consistent

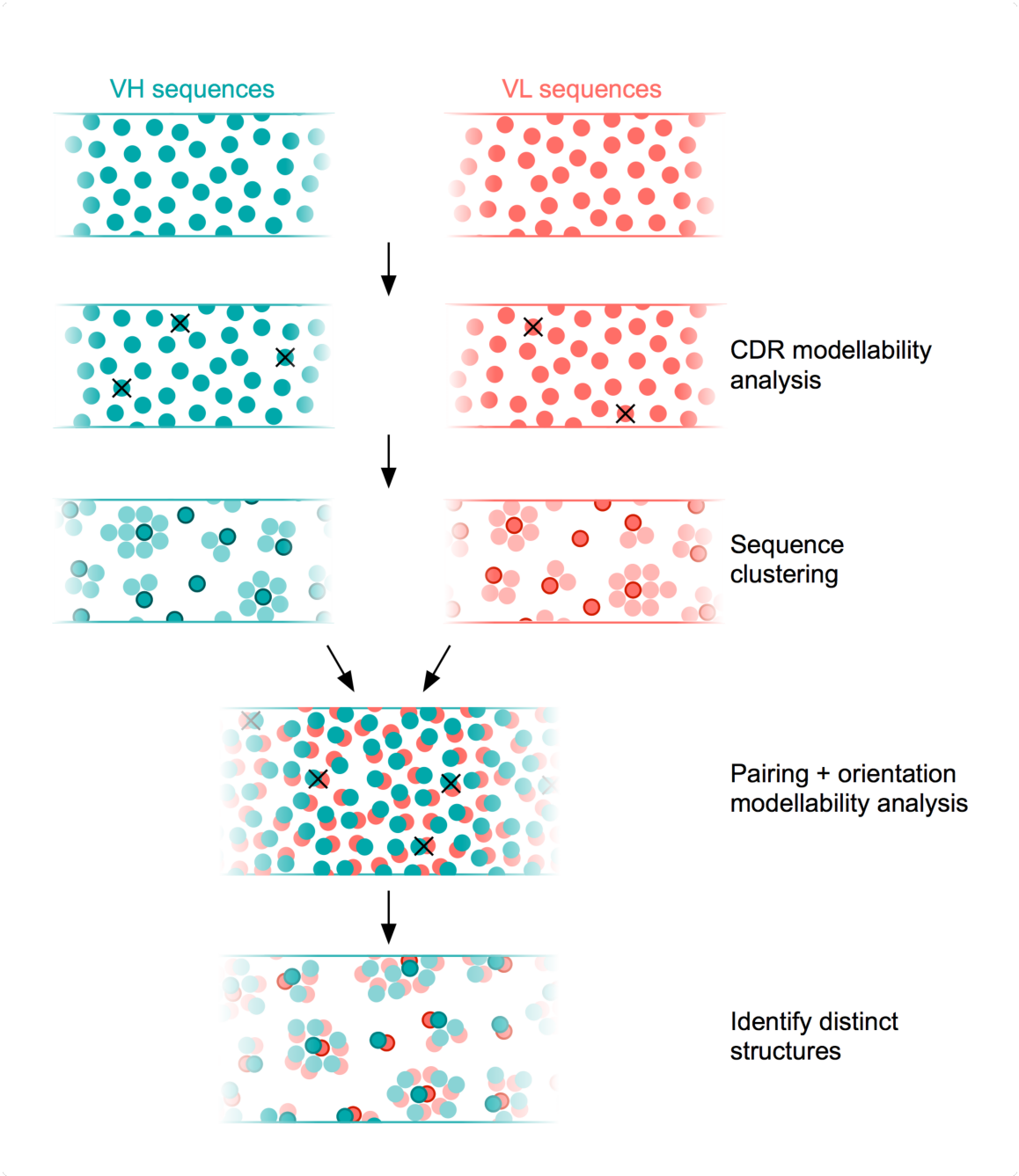


Figure 3.5: A schematic illustrating our Repertoire Structural Profiling algorithm. Heavy (VH) and light (VL) chain sequences from a repertoire snapshot are first analysed separately for their FREAD [26, 27] modellability (unmodellable chains are crossed out). Modellable VH and VL chains are clustered by sequence identity using CD-HIT [247] (90% threshold) for computational tractability. All VH and VL cluster centre chains are subsequently paired, and VH-VL orientations that cannot reliably modelled are removed (again shown by crosses). Finally, predicted modellable Fvs with identical combinations of CDR lengths are structurally clustered to identify 'distinct structures'.

Dataset	All VH	All VL	Modellable VH [90% SIC]	Modellable VL [90% SIC]	Predicted Modellable Fvs	Distinct Structures
1 (S64)	177,603	123,934	10,087	6,779	6,420,211	209,394
2 (S57)	169,805	118,020	9,860	7,922	7,225,630	201,039
3 (S5)	159,544	139,845	8,999	8,526	6,827,419	200,708
4 (S56)	162,446	136,874	9,309	7,168	6,628,683	195,061
5 (S83)	152,299	112,733	9,048	8,076	6,170,373	193,384
6 (S67)	173,722	120,237	9,349	6,424	5,544,952	193,061
7 (S84)	164,017	138,874	8,702	8,232	5,634,598	191,617
8 (S76)	148,180	126,713	8,778	7,047	5,856,150	191,162
9 (S54)	121,993	133,921	7,581	9,066	5,074,822	181,290
10 (S89)	152,710	144,340	8,923	9,293	5,414,820	177,829
...	...	...	...	...	...	...
39 (S95)	118,576	162,377	5,412	11,748	5,901,443	91,855
40 (S17)	102,405	111,669	5,310	7,945	2,690,081	91,229
41 (S4)	100,689	128,986	4,688	1,761	745,977	78,588

Table 3.1: Structurally profiling the baseline repertoire snapshots [83]. A full table containing the values for all 41 baseline datasets is available in the Appendix (Table A3.1). In order, the columns show: the dataset label, the number of VH and VL reads within each snapshot, the number of FREAD-modellable VH and VL reads (once clustered at 90% sequence identity), the number of predicted modellable Fvs resulting from these VH-VL pairings, and the number of distinct structures (cluster centres) identified in each dataset. SIC = Sequence Identity Clustered.

with our understanding of both length and structural variability in VH (particularly in CDRH3) relative to VL [15, 248, 249].

3.4.1.2 Expected Numbers of Distinct Structures *via.* 'Random Repertoires'

To contextualise the numbers of distinct structures observed for each baseline repertoire, we generated 'Random Repertoires' to obtain expected numbers of distinct structures assuming each genuine repertoire sampled randomly from modellable, accessible structure space. To achieve this, we derived:

(a) The *Modellable Repertoire Structures*: a sample of over 180 million structures built from a random combination of any orientation template, a CDR3 template, and a pair of CDR1/CDR2 templates from the same SAbDab entry (mimicking V gene-encoded predetermination). All CDR templates used had been previously assigned by FREAD to a human CDR, and all Fv templates used had been previous assigned by interface residue comparison to a human VH-VL pairing.

(b) The *Length-Accessible Repertoire Structures* for each baseline snapshot: the subset of the Modellable Repertoire Structures with a CDR length combination observed in that individual.

(c) A *'Random Repertoire'* for each baseline snapshot: the appropriate Length-Accessible Repertoire Structures dataset was sampled the same number of times as that individual's number of predicted modellable Fvs. Clustering these RRs then provided a reference number for the expected number of distinct structures per repertoire, given the depth of sampling in each dataset and assuming random sampling.

To derive a set of Modellable Repertoire Structures, we took the same number of samples as the number of Fvs derived from all baseline repertoire snapshots (183,544,740, Table A3.1). Upon structural clustering, these samples yielded ∼24.4M distinct structures over ∼39.9K distinct combinations of CDR lengths, roughly 100x as many distinct structures as seen in any baseline repertoire sample. However, as each repertoire snapshot typically only contained between 2,000-3,500 different CDR length combinations, many of these 24.4M distinct structures could never be observed in the real data. Therefore, 41 'Length-Accessible Repertoire Structures' datasets were created, limiting the Modellable Repertoire Structures to the CDR length combinations seen in each snapshot. For example, considering only the 3,468 CDR length combinations observed in our most structurally diverse individual ('S64') reduced the Modellable Repertoire Structures to a Length-Accessible Repertoire Structures dataset of ∼154.5M structures. This clustered into ∼18.0M distinct structures (a 26.2% reduction from the Modellable Repertoire Structures, while the number of CDR length combinations dropped ∼91.3%), implying we have good structural sampling over the CDR length combinations typically seen in humans. Every Length-Accessible Repertoire Structures dataset contained a number of randomly-selected structures roughly 20-30 times larger than the number of predicted modellable Fvs observed in the corresponding baseline repertoire.

Finally, 41 separate 'Random Repertoires' were created to determine the expected number of distinct structures assuming random structural sampling and given the observed structural sampling depth. To do this, each individual's Length-Accessible Repertoire Structures were sampled randomly, without replacement, the same number of times as the number of predicted modellable Fvs (Table 3.2).

Again taking 'S64' as an example, the 6,420,211 samples comprising 'Random Repertoire S64' yielded 2,092,117 distinct structures, equating to an average of 3.07 Fvs per distinct structure, compared to 30.66 (9.99x more) Fvs per distinct structure in the genuine repertoire. This provides strong evidence that the modellable portions

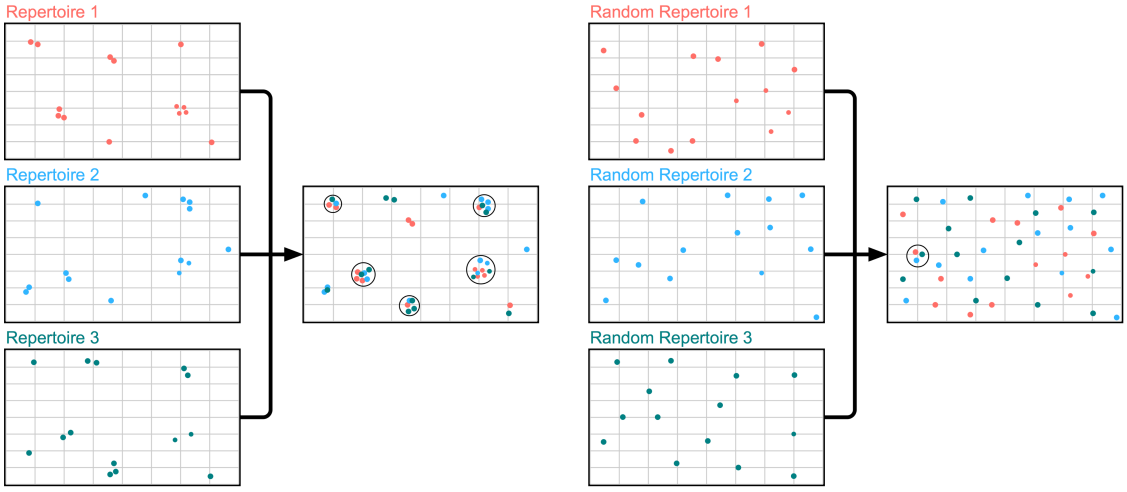


Figure 3.6: Comparing genuine repertoire snapshot to synthetic 'Random Repertoires' (RRs). Each dot represents a distinct structure mapped onto a two-dimensional representation of 'Length-Accessible Repertoire Structure' space. The genuine repertoire snapshots of all three individuals (red = repertoire 1, blue = repertoire 2, green = repertoire 3) exhibit focused structural sampling, covering ∼10% of the space as the corresponding RRs. Overlap analysis shows a high proportion of genuine repertoire distinct structures can characterise an Fv in all three individuals ('public structures', represented by black circles). When the same overlap analysis is performed on the equivalent 'Random Repertoires', far fewer public structures are observed.

of antibody repertoires occupy a highly focused region of modellable structure space - roughly 10% of the expected number given the sample size (Fig. 3.6), and 1% of a theoretical maximum estimate, across the same CDR length combinations.

3.4.1.3 Deriving 'Public Baseline' Structures in Unrelated Individuals

We next investigated whether structural commonality exists between baseline repertoire snapshots. This phenomenon would be statistically extremely unlikely by chance, given the focussed structural sampling observed in each repertoire. To do this, we performed structural clustering on pairs of repertoire snapshots, looking for evidence of structural overlap (i.e. distinct structures assigned to a predicted modellable Fv seen in both datasets, see Section 3.3 and Fig. 3.7).

Repertoire snapshots were ordered by their internal structural diversity ('S64' first, through to 'S4'). The 209,394 distinct structures of S64 act as a reference set of cluster centres. The 7,225,630 Fvs from snapshot S57 were then compared to these S64 cluster centres. Structures present in both S57 and S64 were termed public across two individuals, while S64 and S57 distinct structures unique to their own dataset

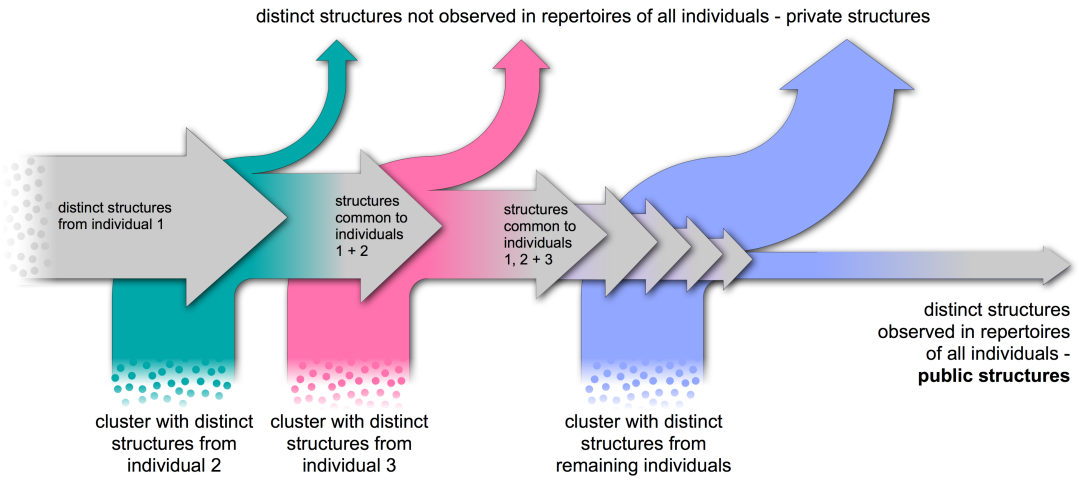


Figure 3.7: Structural overlap analysis. Datasets are arranged in order of their internal structural diversity (most diverse first). Distinct baseline structures from individual 1 are clustered sequentially with all other repertoire snapshots. Distinct structures present in every tested dataset are classed as 'public structures', whereas those that are absent in at least one individual are termed 'private structures'.

were termed private. Next, the 6,827,419 Fvs from S5 were compared to all public and private distinct structures observed in S64 and S57. We again evaluated the number of public structures, this time present in all three datasets. We repeated this analysis for all remaining baseline repertoire snapshots (first ten results in Table 3.2, all 41 results in the Appendix — Table A3.2).

To date, all *in silico* analysis of antibody repertoires has suggested that this number should drop rapidly towards 0. For example, a recent clonotype analysis of the baseline circulating repertoire estimated that only around 0.022% of clonotypes were public across ten unrelated individuals [83]. However, using our methodology, we found that the number of public distinct structures decreased at a far slower rate, still totalling 27,389 structures after ten unrelated individuals (Table 3.2). This represents 3.06% of all distinct structures observed up to that point, over 100 times the number of public clonotypes found by Briney *et al.* in their much deeper repertoire samples. Clonotyping our baseline snapshots, even at the lower 80% CDRH3 sequence identity threshold used by Soto *et al.*, revealed $< 0.01\%$ public clones after five individuals (Table A3.3).

To provide a statistical estimate for how many distinct structures would be expected to be shared across these ten baseline repertoires, the Random Repertoire distinct structures were subsampled to match the corresponding number of baseline

# Reps. (Dataset Added)	Predicted Modellable Fvs Added	Cumulative Public and Private Distinct Structures	Public Distinct Structures (Ovr. % Public)	Expected Public Distinct Structures (Ovr. % Public)
1 (S64)	6,420,211	209,394	209,394	209,394
2 (+S57)	7,225,630	340,915	100,824 (29.57%)	12,307 (3.10%)
3 (+S5)	6,827,419	445,045	71,743 (16.12%)	1,600 (0.28%)
4 (+S56)	6,628,683	527,668	58,043 (11.00%)	322 (0.06%)
5 (+S83)	6,170,373	604,124	48,703 (8.06%)	86 (< 0.01%)
6 (+S67)	5,544,952	670,833	42,277 (6.30%)	31 (< 0.01%)
7 (+S84)	5,624,598	734,374	37,151 (5.06%)	17 (< 0.01%)
8 (+S76)	5,856,150	793,831	33,572 (4.23%)	9 (< 0.01%)
9 (+S54)	5,074,822	846,670	30,474 (3.60%)	6 (< 0.01%)
10 (+S89)	5,414,820	896,328	27,389 (3.06%)	4 (< 0.01%)

Table 3.2: Public structure analysis across the ten most structurally diverse baseline repertoire snapshots. A table tracking the public structures across all datasets is available in the Appendix (Table A3.2). A statistical estimate for the number of public structures was derived by randomly sub-sampling each Random Repertoire to the yield the same number of distinct structures as its equivalent baseline repertoire snapshot. The 'Public Baseline' Antibody Model Library was derived from the 27,389 shared structures up to volunteer S89. Reps. = Repertoires; Ovr. = Overall.

repertoire distinct structures (see Section 3.3). In contrast to the genuine repertoires, the Random Repertoires overlapped sparsely, reaching $< 0.01\%$ public structures by just the fifth volunteer (Table 3.2).

We also tracked the cumulative number of public and private structures over all 41 baseline repertoire snapshots (Table A3.4). Even after the first few most diverse datasets, the deviation from an expected number of distinct structures (given the same ratio of distinct structures:modellable Fvs observed in S64) is quite substantial. This suggests that we might not expect much deviation from our observed fraction of public baseline distinct structures upon deeper repertoire sampling.

Finally, we tested whether the observed proportion of 'Public Baseline' structures would have been significantly different if the experiment had been run using an earlier FREAD database. We repeated Repertoire Structural Profiling for the two most structurally diverse datasets S64 and S57 removing any modellable Fv pairing whose best predicted template for any region was released by the PDB in 2018 or later. As expected, the number of predicted modellable Fv distinct structures in each sample fell from 209,394 and 201,039 to 186,677 and 179,763 respectively (a fall of around 10%). We then performed structural overlap analysis on these sets of distinct struc-

tures, finding a total of 305,948 distinct structures across both datasets, of which 87,920 were public to both S64 and S57. This degree of structural sharing (28.7%) is comparable to the degree observed with access to the entire FREAD database (29.6%).

The existence of such a large number of 'Public Baseline' structures would be statistically extremely unlikely without the presence of underlying functional commonality. Clonotyping is fundamentally unable to capture the same depth of signal, even on much deeper sequencing samples, as functional selection is occurring at the level of structural and paratopic similarity, which may not correspond with conservation of gene transcript origin or high CDRH3 sequence identity. Repertoire Structural Profiling is therefore the first computational method to provide supporting evidence that the naive repertoires of different individuals are structurally biased to target certain epitopes, as suggested by high levels of clonal convergence amongst individuals' affinity-matured antibodies.

3.4.1.4 Characterising the 'Public Baseline' Structures

CDR3 Length Usages. We compared the North-defined [15] CDRH3, CDRL3 and CDRH3+CDRL3 distributions of the S64 Fv sequences assigned to a 'Public Baseline' structure against those assigned to a 'Private Baseline' structure (Fig. A3.1). The CDRL3 and CDRH3+CDRL3 length usages demonstrate that 'Public Baseline' structures are not an artefact of using shorter CDR3 loops with more limited conformations. In fact, we find that modellability bias is likely to be overstating the proportion of 'Public Baseline' distinct structures with longer CDRH3 loop lengths (Fig. A3.1). The structural space available to long CDRH3 (20+) loops is enormous, and we have relatively poor template structural coverage. As a result, if an Fv containing a long CDRH3 loop is considered modellable, it is more likely to be assigned to a structural template further away from its true structure, thus artificially inflating the numbers of long CDRH3s that look structurally similar. These longer CDR length 'Public Baseline' structures should therefore be treated with caution and, as more templates of longer CDRH3 loops emerge improving CDRH3 modellability, we would expect their numbers to decrease to the public:private ratios seen at more moderate CDRH3 lengths.

Germline Proximity and Usages. We also investigated whether S64 Fv sequences assigned to 'Public Baseline' distinct structures were more proximal to germline than those assigned to 'Private Baseline' structures (see Section 3.3; Fig. A3.2). The germline proximity of both 'Public' and 'Private' Fvs to their closest IGHV

and IG[K/L]V genes is very similar, indicating that 'Public Baseline' structures are not solely an artefact of human V gene biases. Finally, we considered the constituent paired V genes across the 'Public Baseline' structures. As our pairing algorithm only predicts modellable Fv pairings based on PDB structures, we compared our IGHV/IG[K/L]V pairing frequencies with those observed in DeKosky *et al.*'s study of over 2000 natively-paired antibodies (Fig. A3.3) [235]. Our 'Public Baseline' gene pairing frequencies were very similar to DeKosky *et al.*'s native sample, with the IGHV1/IGKV1-4, IGHV1/IGLV1-3, IGHV3/IGKV1, IGHV3/IGKV3, and IGHV3/IGLV1-4 pairings the most abundant.

CDR Template Usages. We investigated the number of different structural templates that were assigned to each CDR in a 'Public Baseline' distinct structure (Table A3.5). As expected, the lowest median number of different templates per distinct structure was recorded for the CDRH3 loop (2 templates/structure), consistent with the large structural variation within the region driving the definition of distinct binding site structures. Collectively, the light chain CDRs recorded considerably more FREAD templates per structure (median of 20 templates/structure) than the heavy chain CDRs (median of 9 templates/structure).

3.4.1.5 Building and Characterising a 'Public Baseline' Antibody Model Library

We used ABodyBuilder [44] to construct an Antibody Model Library (AML) based on the 27,389 'S64' pairings predicted to adopt a 'Public Baseline' structure (as defined by the ten most structurally diverse repertoire snapshots). Some Fvs failed to be entirely homology modelled. For example, occasionally the CDRH3 template clashes irreparably with the CDRL3 template during construction of the full Fv model, necessitating *ab initio* treatment. Overall, 23,700 (86.53%) of 27,389 pairings were entirely homology modelled and comprise our 'Public Baseline AML' (downloadable from *http://opig.stats.ox.ac.uk/resources*).

Proximity to Therapeutics. Predicted structures shared between many individuals might represent good starting points for therapeutic development. Their widespread nature could point to their binding versatility, and also to broad immune system tolerance across many individuals, lowering the risk of drug immunogenicity. To test whether our 'Public Baseline AML' already contains binding site topologies proximal to known therapeutics, we mined Thera-SAbDab [99] for all 100% sequence identical structures of WHO-recognised therapeutics, selecting one per therapeutic (see Section 3.3). Of the 66 therapeutics with known structures that had at least

PB Structure	Therapeutic	RMSD (Å)	CDR Lengths	Antigen
H101/L64549	Ofatumumab	0.148	13-10-15-11-8-9	CD20
H11835/L101012	Durvalumab	0.162	13-10-14-12-8-9	CD274
H19709/L100051	Tanezumab	0.181	13-9-15-11-8-9	NGFB
H35853/L102278	Tremelimumab	0.255	13-10-18-11-8-9	CD152
H11488/L100048	Olokizumab	0.278	13-12-11-11-8-9	IL6
H10992/L14321	Tezepelumab	0.481	13-10-15-11-8-11	TSLP
H13677/L17885	Avelumab	0.525	13-10-13-14-8-10	CD274
H14012/L14649	Ustekinumab	0.638	13-10-12-11-8-9	IL12B
H38817/L68369	Aducanumab	0.681	13-10-17-11-8-9	APP
H26854/L108275	Imalumab	0.730	13-10-11-11-8-9	MIF
H30607/L11664	Quilizumab	0.730	13-10-10-16-8-9	IGHE

Table 3.3: The eleven clinical-stage therapeutic antibodies with a solved crystal structure within 0.75Å variable domain (Fv) root-mean-squared deviation (RMSD) of an antibody model structure from the Public Baseline Antibody Model Library (PB AML). The root-mean-squared deviation (RMSD) is calculated over all Fv residues and comparison were only made between AML structures and therapeutics with an identical combination of CDR lengths. This combination of North-defined [15] CDR lengths is provided in the table, listed in the order H1-H2-H3-L1-L2-L3. PB: Public Baseline.

Reference PDBs: Ofatumumab - 3giz (HL), Durvalumab - 5xj4 (HL), Tanezumab - 4edw (HL), Tremelimumab - 5ggu (HL), Olokizumab - 5tru (HL), Tezepelumab - 5j13 (CB), Avelumab - 4nki (HL), Ustekinumab - 3hmw (HL), Aducanumab - 6cnr, Imalumab - 6foe (HL), Quilizumab - 3hr5 (HL). Antigens: CD - Cluster of Differentiation protein, NGFB - Nerve Growth Factor B, IL - interleukin, TSLP - Thymic Stromal Lymphopoietin, APP - Amyloid Precursor Protein, MIF - Macrophage Migration Inhibitory Factor, IGHE - Immunoglobulin Heavy Constant Epsilon.

one antibody in our 'Public Baseline AML' with identical CDR lengths, all had a structural partner in the AML within a C_α Fv RMSD of 1.84Å, and 37 (56.1%) had a structural partner within 1.00Å Fv RMSD. Eleven therapeutic structures lay within 0.75Å Fv RMSD of a 'Public Baseline AML' structure (Table 3.3); these therapeutics spanned a wide range of targets and were primarily successful or promising drugs (4 approved, 5 active in Phase III, 1 active in Phase II, and 2 discontinued).

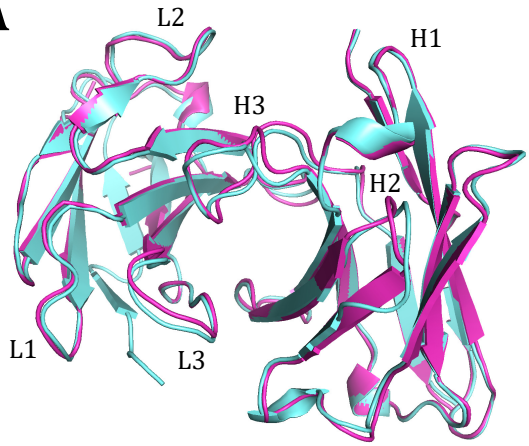
This result demonstrates that the antibody models within our 'Public Baseline AML', without any explicit design, can display high levels of geometric similarity to known therapeutics. To show that similar binding site residue profiles can also be found by Repertoire Structural Profiling, we examined 'Public Baseline' distinct structure 'H14012+L14649' as a case study (Fig. 3.8).

This structure lies within 0.64Å of the therapeutic Ustekinumab (Table 3.3). Examining the backbone-aligned structures shows this difference lies in slightly different CDR loop structures assigned to the CDRH2, CDRH3, and CDRL3 loops (Fig. 3.8A). We then examined all 4,911 Fv sequences assigned to this distinct structure across the ten individuals (S64 through S89), looking for the closest CDR sequence identity matches to Ustekinumab. The most similar of the 155 sequence-unique VH sequences assigned to this distinct structure is shown in Fig. 3.8B. While both the Ustekinumab and 'Public Baseline' VH sequences most closely aligned to the same V and J genes (IGHV5-51/IGHJ4), the CDRH3 sequences are only 66% sequence identical, and so would not have been assigned to the same VH clonotype (the typical minimum threshold is 80% identity, as used in Soto *et al.* [80]). This VH was observed coupled both with the VL sequence shown in Fig. 3.8B and with the VL sequence shown in Fig. 3.8C. The VL in Fig. 3.8B is more identical across the three CDRs (22/26, 85%), while the one in Fig. 3.8C is closer in CDRL3 identity but considerably less so in CDRL2 identity. Both these VLs derive from different IGKV germlines to the Ustekinumab VL (Ustekinumab: IGKV1D-16, Fig. 3.8B VL: IGHV1-9, Fig. 3.8C VL: IGKV3-15). Overall, the Fv described in Fig. 3.8B is 75% sequence identical to Ustekinumab across all 6 CDRs.

The degree of sequence and structural similarity between clinical-stage therapeutic antibodies and representatives of the 'Public Baseline' structural repertoire suggests that Repertoire Structural Profiling could prove an effective tool for designing general screening libraries containing promising drug leads.

VH Sequence Profiling the 'H14012+L14649' Distinct Structure. We performed clonotyping (80% sequence identity threshold [80]) on the 155 sequence non-redundant VH chains to determine the diversity of heavy chain clonotypes mapped

A



Ustekinumab

H14012+L14649

B

Ustekinumab, Gidoni

```
                                    CDRH1                                CDRH2
VH: EVQLVQSGAEVKKPGESLKISC KGSGYSFTTYWLG WVRQMPGKGLDW IGIMSPVDSDIR YSPSFQGQV
VH: -VQLVQSGAEVKTRGESLKLSC KGSGYSFTSYWTG GVRQMPGKGLEW TGIIYPGDSDTS NSPSFQGQV
                                    CDRH3
TMSVDKSITTAYLQWNSLKASDTAMYYC ARRRPGQGYFDF WGQGTLVTVSS
TISADKSSSTGYLQWSSLKASDTAMYYC ARREVGMGYFDY WGQGTLVTVSS
                                    CDRL1                                CDRL2
VL: DIQMTQSPSSLSASVGDRVTITC RASQGISSWLA WYQQKPEKAPKSLI YAASSLQS GVPSRFSGSGSG
VL: DIQLSQSPSFLSASVGDRVTITC RASQGISSSLA WYQQKQGKAPKVLI YAASTLQS GVPTRFRGSGSG
                          CDRL3
TDFTLTISSLQPEDFATYYC QQYNIYPYT FGQGTKLEIK
TEFTLTISSLQPEDFATYYC QQLNSCPPT FGQGTRLEIK
```

C

```
                                    CDRL1                                CDRL2
VL: DIQMTQSPSSLSASVGDRVTITC RASQGISSWLA WYQQKPEKAPKSLI YAASSLQS GVPSRFSGSGSG
VL: EIVMTQSPATLSVSPGARATLSC RASQSVNSNLA WYQQKPGQAPRLII FGSSTRST GIPVRFSGGGSG
                          CDRL3
TDFTLTISSLQPEDFATYYC QQYNIYPYT FGQGTKLEIK
TEFTLTISGLQSEDFEVYYC QQYNNWPYT FGQGTKLEIK
```

Figure 3.8: (A) Alignment of the solved Ustekinumab crystal structure (3hmw) and the closest Public Baseline AML structure (H14012+L14649). (B) Comparison of the Ustekinumab Fv sequence and a Gidoni *et al.* naïve Fv sequence assigned by Repertoire Structural Profiling to the H14012+L14649 Public Baseline structure. The North-defined CDR regions of each chain are highlighted in bold. (C) An alternative VL sequence coupled to the same Gidoni VH sequence. This sequence has a more sequence similar CDRL3 but a less similar CDRL2.

to the 'H14012+L14649' Public Baseline structure. The VH sequences clustered into 141 distinct clonotypes, whose germline gene combinations as assigned by ANARCI [18] are shown in Table A3.6. As clonotyping conditions on antibodies having the same V and J gene identities, it would never pool these VHs into a single category. Twelve of the 141 clonotypes have multiple occupancy (Table A3.7). Three clonotypes were found across multiple individuals:
V5-51+ARPYGSGSYSDY+J4: seen in S64, S54, and S76
V5-51+ARQGYGDYVTDY+J4: seen in S67 and S76
V5-51+ARMGARPGYFDY+J4: seen in S89 and S76
This shows how Repertoire Structural Profiling could be used in conjunction with clonotyping to add geometric support to convergent clones being functionally equivalent. Novel methods currently being developed that can predict paratope similarity across all six CDRs [250, 251] may be able to find considerably more antibodies within each distinct structure cluster with similar enough interaction profiles to be functionally equivalent.

3.4.2 Structurally Profiling a Flu Vaccine Response

3.4.2.1 Evaluating the Numbers of Distinct Structures Before and After Vaccination

Clonotyping is commonly used in antibody drug discovery to identify 'expanded clones' - novel genetic lineages present after vaccination/infection but that were absent, or low concentration, beforehand [236]. Often these expanded lineages are seen across many different individuals after vaccination, implying particular pathogenic epitopes are 'immunodominant' — more susceptible to immune recognition [252–254]. Here, we applied Repertoire Structural Profiling to investigate whether we could identify an analogous public structural response to vaccination.

To this end, we used a longitudinal 2009 seasonal flu vaccination study by Gupta *et al.* [84], in which three unrelated individuals ('V1-3') were sequenced at many time-points before and after vaccination (see Section 3.3). Using the same Repertoire Structural Profiling protocol as above, we calculated the number of distinct structures observed in each individual before and after vaccination (shown in Table 3.4).

To obtain an estimate for the degree of structural commonality pre- and post-vaccination, we again used a greedy clustering approach to evaluate the structural overlap between the 'Before Vaccination' datasets, and between the 'After Vaccination' datasets, separately (Fig. 3.9A; 3.9B). The first dataset in each overlap assess-

Rep.	All VH	All VL	Modellable VH [90% SIC]	Modellable VL [90% SIC]	Predicted Modellable Fvs	Distinct Structures
V1 Bef.	241,630	679,472	13,230	93,422	33,376,943	471,650
V2 Bef.	235,022	754,179	12,135	105,026	47,884,336	579,961
V3 Bef.	307,067	825,587	16,124	113,239	77,186,291	772,844
V1 Aft.	392,661	1,052,510	21,624	155,412	104,736,939	852,328
V2 Aft.	331,573	1,173,628	16,047	158,439	91,023,995	809,341
V3 Aft.	368,961	1,087,670	18,029	299,908	105,824,532	843,341

Table 3.4: Structurally profiling the 'Before Vaccination' (Before) and 'After Vaccination' (After) repertoire snapshots of three unrelated individuals (V1, V2, and V3) [84] . In order, the columns show: the dataset label, the number of VH and VL reads within each snapshot, the number of FREAD-modellable VH and VL reads (once clustered at 90% sequence identity), the number of predicted-modellable Fvs resulting from these VH-VL pairings, and the number of distinct structures (cluster centres) identified through greedy structural clustering. Rep: Repertoire; Bef: Before Vaccination, Aft: After Vaccination; SIC: Sequence Identity Clustered.

ment was the most structurally diverse (*i.e.* the 'V3' individual before vaccination, and 'V1' after vaccination).

Again, a significant number of public distinct structures were observed in 'V1','V2', and 'V3' ('Public Before Vaccination' structures, 17.78% (236,792/1,444,597) of all 'Before Vaccination' distinct structures). This indicates that the identification of 'Public Baseline' structures in the previous section was unlikely due to serendipitous Ig-seq amplification bias. For context, the proportion of all clonotypes that were public before vaccination was just 0.03% (Soto *et al.* definition [80], Table A3.8).

The degree of structural sharing appears to increase after vaccination, with 19.23% (350,710/1,823,648) public structures across the three volunteers. This is consistent with a degree of repertoire structural convergence driven by exposure to the same pathogenic epitopes and with an increase in the proportion of public clonotypes after vaccination to 0.13% (Table A3.8).

3.4.2.2 Deriving Convergent Structures that Only Occur After Vaccination

To derive these convergent structures, the structural overlap between each individual's 'Before Vaccination' and 'After Vaccination' datasets was measured, only retaining 'After Vaccination' pairings that could not be clustered into the same individual's 'Before Vaccination' distinct structures. 'V1' remained the most structurally diverse

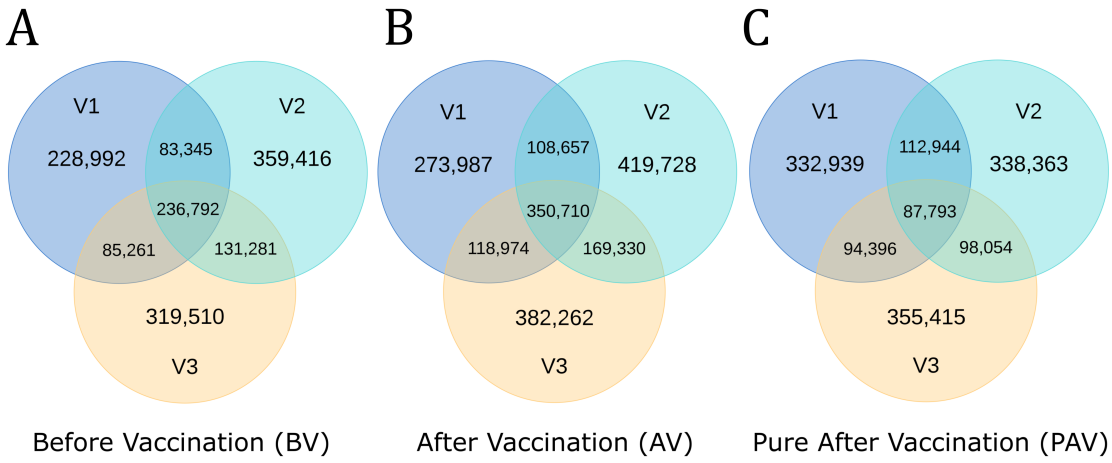


Figure 3.9: Venn diagrams showing the structural overlap between each individual's (A) 'Before Vaccination' dataset, (B) 'After Vaccination' dataset, and (C) 'Pure After Vaccination' dataset (distinct structures arising only after vaccination). Total distinct structures: Before Vaccination: 1,444,597; After Vaccination: 1,823,628; Pure After Vaccination: 1,419,904. V1-V3 = Volunteer 1-3.

dataset, with 628,072 'Pure After Vaccination' distinct structures. The overlap between these 'Pure After Vaccination' pairings (Fig. 3.9C) was then compared. This yielded a mixed picture of convergent and private vaccination response structures - 27.7% (393,187/1,419,904) of distinct structures were shared with at least one other individual, and 6.18% (87,793/1,419,904) were shared across all three individuals - which we term 'Public Response' structures.

There are two potential causes of overlap in the 'Pure After' vaccination set. One is a genuine common structural response to vaccination, while the other is that the initial baseline repertoire was under-sampled - *i.e.* the overlap reflects residual shared baseline structures. As a second test for baseline deviation, beyond absence before vaccination, we compared how many of the 27,389 'Public Baseline' distinct structures were within 1Å of a 'Public Before Vaccination' binding site, versus the number within 1Å of a 'Public Response' Structure binding site. We observed that 80.0% (21,922/27,389) of 'Public Baseline' structures were within 1Å of a 'Public Before Vaccination' structure, compared to just 24.2% (6,621/27,389) proximal to a 'Public Response' structure. This provides further evidence that a proportion of these convergent 'Public Response' structures reside in a distinct region of structural space and could harbour epitope-specific binding geometries. We have built a 'Public Response AML' based on these 87,793 shared structures, with 74,181 Fvs (84.4%) successfully homology modelled (downloadable from *http://opig.stats.ox.ac.uk/resources*).

3.5 Discussion

In this work, we have structurally profiled antibody repertoires to capture new insights into the baseline and antigen-responding immune system, and to create novel libraries of antibody model structures that could be exploited for immunotherapeutic discovery.

All of the structural analysis in this paper is limited to the antibody chains that are currently predicted to be modellable, and so there remain regions of natural structural space uninvestigated, and, once these become characterisable, the currently observed proportion of public structures may become diluted. Despite this, we show that antibody repertoires tend only to explore highly focused regions of currently-modellable structural space (∼10% of the diversity expected if templates were explored randomly across the same combinations of CDR lengths). Coupled with our experiment blinding Repertoire Structural Profiling to the most recent year's templates, this suggests that a large portion of structural commonality will remain across the currently unmodellable regions of structural space, although we might expect the number of 'Public Baseline' structures with long CDRH3 loops to fall as modellability could be increasing this figure.

The enormous sequence diversity exhibited across baseline antibody repertoires has long appeared to run contrary to the theory of baseline functional commonality. Here we have shown that, at least from a structural perspective, there is considerable opportunity for functional commonality across the circulating resting-state repertoires of unrelated individuals (∼3% of observed distinct structures are public across 10 individuals). The theoretical chemical diversity that could be displayed on each of these scaffolds is large, so many of these grouped binding sites will not be complementary to the same antigen epitope. However, there is good reason to believe that a certain proportion are, as geometric similarity is a likely prerequisite of functional commonality, and our structural clustering approach offers a route to detecting and analysing these antibodies. This knowledge could then be harnessed in vaccinology - for example, identifying an epitope targettable by a 'Public Baseline' structure may lead to a more reliable and convergent response. We note that some edge cases, such as geometric similarity obtained using loops of different lengths, or antibodies that can use different CDRs to fit the same epitope *via* an alternative binding mode, are currently undetectable using our framework.

Once grouped into public structures, Fvs can then be probed using an array of methods designed to measure binding residue similarity to identify the subset likely

to have common functionality. For example, finding convergent clonotypes within the public baseline structures may bolster confidence in their functionally convergent role. Alternatively, methods that do not condition on predicted antibody genetic origin, such as paratyping [250] or Ab-Ligity [251], could identify more genetically divergent antibodies capable of binding the same epitope. The public geometries themselves could also be harnessed in vaccinology, such as identifying an epitope targettable by a 'Public Baseline' structure which may lead to a more reliable and convergent response.

We also hypothesised that human 'Public Baseline' structures were more likely to display low levels of human immunogenicity and be versatile binders. Building full three-dimensional variable domain models of these distinct structures (an 'Antibody Model Library') produced geometries that were very close to several approved and late-stage active therapeutic antibodies targeting diverse antigens. To chemically elaborate this 'Public Baseline' structural basis set, a phage display library on the order of 10^6-10^7 sequence-unique human antibodies could be created from the many different Fv sequences predicted to adopt each public distinct structure.

Target-focused screening libraries against immunodominant epitopes are commonly derived through sequence analysis of longitudinal Ig-seq studies that track the immune response of many individuals to the same antigen. We show that when our methodology is applied to a longitudinal flu vaccination case study, we detect a higher level of structural convergence, commensurate with response to similar epitopes on the same antigen. We can also derive a large number of 'Public Response' structures, with divergent structural characteristics from the baseline repertoire. These could contain useful binding site structures exploitable for antigen-specific library design. One issue with this is that, currently, most disease-specific Ig-seq experiments only sample heavy chain diversity. However, light chain diversity is far lower than heavy chain diversity and may be able to be summarised by a representative set of sequences regardless of individual or disease state. This ought to be investigated, as such a set of representative chains would provide allow for the creation of considerably more disease-specific Antibody Model Libraries.

Whilst ever we must artificially pair Ig-seq datasets, we cannot conclusively prove that multiple individuals raised the same Fv binding site geometry in response to vaccination. This could soon be rectified with the advent of single-cell sequencing studies investigating vaccine response dynamics [73]. Repertoire Structural Profiling could readily be applied to such data by skipping the combinatorial pairing step, which would be expected to improve both speed and accuracy.

It is important to appreciate that there are biases inherent in structurally profiling human antibody repertoire data to suggest antibody leads for drug discovery. One such biased property is CDRH3 length: very short CDRH3 lengths will be under-sampled through their sparsity in natural human sequences [214], while very long CDRH3 lengths will be under-sampled because they are more difficult to homology model accurately. The former issue may be rectified by considering the public structures over fewer individuals, while the latter is simply insoluble until structures of more antibodies with longer CDRH3s are solved. Another consideration is that, while inherent immunogenicity should be diminished by virtue of exploiting naturally-expressed sequences, other developability issues may still arise, as not every human antibody has the biophysical properties ideal for large-scale manufacture and long-term storage [214]. Designing screening libraries around characteristics of the naïve antibody repertoire may lead to a trade-off between immunogenicity and promiscuity/polyspecificity.

Nevertheless, we believe that our approach should find immediate applicability in both *in silico* and *in vitro* screening. We have made available the 'Public Baseline' and 'Public Response' Antibody Model Libraries for further investigation, and will continue to build and share the Antibody Model Libraries derived from other unpaired and paired VH+VL datasets in the Observed Antibody Space database [88].

3.6 Update and Chapter Conclusion

Three months after our preprint was released, a group from Harvard preprinted a novel statistical method that identifies 'functionally-public' antibodies from the apparently sequence-private repertoire [93]. We expect that, through novel and creative ways of analysing Ig-seq data, computational evidence will continue to build demonstrating functional commonalities between the baseline repertoires of different individuals.

In light of our findings, we are encouraged to benchmark the performance of Public Baseline AMLs and Shared Response AMLs in a mock *in silico* drug discovery setting. Progress has been made by another group member (Constantin Schneider) towards an effective convolution neural network rigid docking rescoring function (we must initially use rigid docking when evaluating the order of 10^4 antibodies), but we still need to perform considerable amounts of benchmarking to determine a sufficiently accurate and rapid docking regime (see Chapter 6). In time, our goal is to be able to computationally propose a genetically-diverse set of antibodies that appear

complementary to a defined antigen epitope, and for this set to be considerably more enriched for binders than an equivalently-sized conventional screening library.

However, finding an antibody with good binding affinity to a target does not guarantee a successful therapeutic. Immunogenicity, instability, self-association, high viscosity, poor solubility, polyspecificity, or poor expression can all preclude an antibody from becoming a drug [214]. Early identification of these negative characteristics is essential.

In the next chapter, we compare the physicochemical properties of therapeutic antibodies and Antibody Model Libraries built from human Ig-seq data. In particular, we assay properties with intuitive or documented links to chemical, physical, or colloidal instability. This investigation led to a set of five computational antibody developability guidelines, akin to the Lipinski 'Rule of Five' in small molecule drug discovery, that can be applied in early-stage research to warn of candidates likely to possess poor developability [214].

Chapter 4

The Therapeutic Antibody Profiler: Developability Guidelines through Antibody Model Libraries

4.1 Chapter Abstract

We have previously described a database containing molecular information on therapeutic antibodies (Thera-SAbDab, Chapter 2) and a protocol for building Antibody Model Libraries (AMLs) that capture the maximal structural diversity of BCR repertoire sequencing datasets (Repertoire Structural Profiling, Chapter 3). This chapter brings these two concepts together to compare and contrast the typical physicochemical properties of clinical-stage therapeutics (CSTs) and natural antibodies. Led by these datasets, analysis of prior literature, and core chemical principles, we aimed to settle on molecular properties that ought to be held within a certain range to guard against problems such as molecular instability, self-association, high viscosity, polyspecificity, or poor expression ('Developability Issues', see Section 1.7).

The outcome of this chapter is a set of five computational developability guidelines that can be used in antibody drug discovery to prioritise candidates less likely to suffer with developability issues. Akin to the Lipinski 'Rule of Five', our guidelines can be rapidly and computationally calculated on any given antibody therapeutic sequence. We packaged the software into a web application, called the 'Therapeutic Antibody Profiler' (TAP), which is publicly available as part of the SAbPred suite of antibody informatics tools.

The chapter concludes with a discussion on how the guidelines may change in the future and how TAP has been used by the community since its release.

This chapter contains reproduced material from the following paper:
Raybould, M.I.J., Marks, C.M., Krawczyk, K., Taddese, B., Nowak, J., Lewis, A.P., Bujotzek, A., Shi, J., Deane, C.M. (2019) Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci. USA.* 106(10):4025-4030. [214]

4.2 Introduction

Monoclonal antibodies (mAbs) are increasingly used as therapeutics targeting a wide range of membrane-bound or soluble antigens - of the 73 antibody therapies approved by the EU or FDA since 1986 (valid as of June 12$^{\text{th}}$, 2018), 10 were first approved in 2017 [255]. There are many barriers to therapeutic mAb development, besides achieving the desired affinity to the antigen. These include intrinsic immunogenicity, chemical and conformational instability, self-association, high viscosity, polyspecificity, and poor expression. *In vitro* screening for these negative characteristics is now routine in industrial pipelines [179].

While some cases of poor developability are subtle in origin, others are less ambiguous. High levels of hydrophobicity, particularly in the highly variable complementarity-determining regions (CDRs), have repeatedly been implicated in aggregation, viscosity and polyspecificity [179, 183, 186–189, 256]. Asymmetry in the net charge of the heavy and light chain variable domains is also correlated with self-association and viscosity at high concentrations [182, 183]. High rates of clearance and poor expression levels have been associated with patches of positive [257] and negative [196] charge, while heterogeneity (e.g. through oxidation, isomerisation, or glycosylation) often results from specific sequence motifs liable to post- or co-translational modification.

An improved understanding of the factors governing these biophysical properties has enabled the development of *in silico* assays, which are much faster and cheaper than their experimental equivalents. Computational tools already facilitate the identification of sequence liabilities, *e.g.* sites of lysine glycation [258], aspartate isomerisation [259], asparagine deamidation [259], and the presence of non-canonical cysteines or N-linked glycosylation sites [260]. A primary focus in recent years has been in designing software that can better predict aggregation proclivity. Many algorithms designed for this purpose use only the antibody sequence [183, 186, 189], although some suggest an analogous equation to use if a structure is available [183]. One purely structure-based method is the structural aggregation propensity (SAP) metric [187], later included in the 'Developability Index' [188]. This has been shown to detect

aggregation-prone regions, such as surface patches [261], and to be able to rank candidates relative to a known antibody developability profile [196], using a very closely related antibody crystal structure. However, due to the laborious nature of X-ray crystallography, it is inefficient to solve the structures of the many lead antibodies implicated in the early stages of a drug discovery campaign. Therefore, to be of practical use, structure-based *in silico* approaches must prove they are informative when applied to antibody homology models, which can be generated in a high throughput manner only requiring knowledge of each antibody's amino acid sequence. Though proven on solved structures, an atomic-resolution analysis, such as that provided by SAP, may prove too sensitive when comparing homology models of diverse antibodies, judging by the current accuracy of structure prediction [134].

An alternative approach to predict antibodies likely to have poor developability profiles is to highlight those candidates whose characteristics differ greatly from clinically-tested therapeutic mAbs; a similar strategy in the field of pharmacokinetics led to the Lipinski rules for small molecule drug design [262]. Here, we build three-dimensional homology models of a large set of post-Phase I therapeutics and survey their sequence and structural properties. These values are then contextualised against human B-cell receptor (BCR) repertoire sequences and models, to see where therapeutics share and deviate from the properties of human antibodies.

Using the distributions of these properties, we built the Therapeutic Antibody Profiler (TAP), a computational tool that highlights antibodies with anomalous values compared to therapeutics. TAP constructs a downloadable structural model of an antibody variable domain sequence, and tests it against guideline thresholds of five calculated measures likely to be linked to poor developability. It also reports potential sequence liabilities and all non-CDRH3 loop canonical forms.

4.3 Methods

All 242 CST sequences are supplied in Dataset S2 of Raybould *et al.* [214] and the 551,193 heavy and 1,359,745 light chain non-redundant, 'healthy' human VdH Ig-seq sequences can be obtained from the Observed Antibody Space database [88]. The 4,587,907 heavy and 7,120,000 non-redundant human UCB Ig-seq sequences are available as separated CDR and framework regions at the OAS website [88]. Therapeutic models and human VdH Ig-seq models can be downloaded from *http://opig.stats.ox.ac.uk/resources*. Therapeutic antibody solved structures can be found through the Thera-SAbDab web application [99].

Human VdH Ig-seq Sequencing Techniques

The sequencing procedure followed to obtain the human VdH Ig-seq nucleotide reads is described in the manuscript and SI of Vander Heiden *et al.* [89].

Human UCB Ig-seq Sequencing Techniques

The experimental procedure performed by our UCB collaborators to obtain this dataset is described in the Appendix of this thesis.

Bioinformatic Annotation of Ig-seq Sequences

IgBLAST 1.4.0 [263] was used with Human V, D & J germline reference sets from IMGT for both heavy and light chains to germline annotate the full length reads. A custom Java pipeline was used to process the IgBLAST output and identify high quality sequences. The criteria for this were as follows: identified germline V & J genes; full length variable chain sequence (1-2 bp missing at 5' & 3' end was permitted); absence of stop codons or ambiguous nucleotide calls. Sequences successfully extracted were saved into flat files together with the identifiers of their assigned germline sequences. This Ig-seq dataset was translated, then IMGT-numbered and filtered by ANARCI to remove poor-quality reads [264]. This parsing removes sequences that do not align, have IMGT CDRH3 lengths ≥ 37, possess indels in the canonical CDRs or framework regions, start at IMGT position 24 or later, or have a J gene with sequence identity less than 50% to known IMGT germlines. It is at this stage that all Ig-seq datasets in the Observed Antibody Space database are supplied to researchers [88].

Preparation of the Human VdH Ig-seq Dataset for Analysis[1]

For structural comparisons, where heavy and light chains must be paired and then modelled, computational expense required us to reduce the size of the dataset, while still retaining as true an indication as possible of the sequence and structural variation inherent within the immunoglobulin repertoire. The chains were therefore filtered and paired according to the following protocol, closely related to that described in Chapter 3. The ANARCI-filtered dataset was trimmed to remove heavy and light chain sequences that contained IMGT CDRs [23] (CDRH1-2, CDRL1-3) for which SCALOP [25] could not predict a canonical form (inability to predict a canonical form is highly linked to an inability to model the sequence). Then FREAD [26, 27, 36]

[1]This procedure for capturing the structural properties of the Ig-seq datasets was a preliminary version of the Repertoire Structural Profiling protocol described in Chapter 3.

was run on each surviving heavy chain's CDRH3 loop to remove chains for which no viable CDRH3 template can be found. On this reduced number of heavy and light chains, the full version of FREAD was then run to assign templates to all loops. The V_H and V_L sequences were then greedily clustered with a 90% sequence identity threshold, and with a restriction that each cluster must only contain sequences of identical length. Clusters with fewer than 10 members were discarded to remove erroneous reads. From the surviving clusters, only the sequence with the lowest median sequence identity to the rest of the cluster members was retained. The chains were then paired by assigning the V_H-V_L angle from a diverse set of 989 complete antibodies from SAbDab [9] (selected in May 2016) to the heavy-light chain pair, if the sequence identity across the heavy-light chain interface residues exceeds 0.82 [44]. If multiple viable templates exist, the one with highest sequence identity was chosen. Low-resolution structural clustering of CDR binding sites was then performed in a two-step process. If the orientation RMSD between the orientation template of the newly-considered antibody and the templates of all other previously-considered antibodies exceeds 1.5 Å, the binding site was classified as distinct based on V_H-V_L angle alone, and the antibody was retained. Otherwise, CDR distance was evaluated as:

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

where the summation is over all CDRs, $DTW_{CDR\text{-}X}$ is the Dynamic Time Warping [24] distance between CDR templates, and $max(L_{CDR\text{-}X,1}, L_{CDR\text{-}X,2})$ is the maximum length of the two loops. The DTW algorithm is conventionally applied to identify similarities between two objects that differ in speed; here it is applied to identify similarities in loop structures that differ in sequence length (in other words, the position in the peptide sequence is analogous to moving forward one quantised unit in the time domain). The new antibody was retained if this equation returned a value greater than 1.0 Å to all other previously-considered antibodies. The surviving Fvs were then homology modelled using ABodyBuilder [44], and these 14,072 models became the Human VdH Ig-seq models dataset. The resulting total CDR length distribution of this set of models proved to be similar to that of the set of 137 CSTs (Fig. A4.6A).

Preparation of the Human UCB Ig-seq Dataset for Analysis

The UCB dataset was prepared in an analogous manner to the VdH Ig-seq dataset, but with a few modifications. As it was prepared before SCALOP [25] had been

developed, FREAD was run on all chains at the beginning to remove chains without three Chothia CDR loop templates. Being far larger than the VdH Ig-seq dataset (about 10x the number of heavy and light chain sequences), around 66,000 models were obtained after DTW clustering and ABodyBuilder modelling. To reduce this number further, higher-resolution structural clustering was performed on the CDR coordinates. The RMSD between common CDR residues in each pair of models, after Kabsch alignment [265], was evaluated and stored in a matrix, which was clustered using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) hierarchical method [266], with a 1.0 Å cutoff. A representative of each cluster was selected and the resulting 19,019 models became the Human UCB Ig-seq models dataset. The resulting total CDR length distribution of this set of models also proved to be similar to that of the set of 137 CSTs (Fig. A4.6B).

Antibody Structures and Models

Models of all antibodies were generated using the latest implementation of ABodyBuilder [44], used as published, except for the inclusion of PEARS [53] to model side chains. For all antibodies, ABodyBuilder was prevented from using sequence-identical templates, to ensure all structures were genuine models. The 31 unbound and 25 bound Fab structures used to estimate therapeutic model accuracy were X-ray crystal structures deposited in the PDB [19] (before May 4$^{\text{th}}$, 2018). Bound structures were used only if no unbound equivalent existed. If multiple unbound structures were available, the lowest resolution crystal structure was selected. If multiple bound structures were available, the crystals containing the native antigen were prioritised over those containing synthetic constructs, and then the lowest resolution crystal was chosen. Accuracy was measured using an in-house backbone RMSD calculator, with framework RMSDs obtained after aligning all Fv C_α atoms, and CDR RMSDs calculated after aligning only framework C_α atoms - a standard protocol in the field [134]. Case study proprietary sequences were modelled on-site at MedImmune for IP reasons. All MedImmune validation mAbs were models, despite ABodyBuilder having access to additional in-house structural information and their setup permitting sequence identical templates. The web version of TAP does not permit sequence-identical templates in the ABodyBuilder modelling protocol.

Canonical Forms

A length-independent canonical form clustering protocol [24] was run on the North-defined [15] CDR loops of a SAbDab [9] snapshot from 26th September 2017. Model

loops were inferred to have identical canonical forms to the template used by ABodyBuilder [44].

Surface-exposed Residues

Residues defined as 'surface-exposed' have > 7.5% relative exposure [267] across side chain atoms, compared to the open-chain form Alanine-R-Alanine, as calculated with the Shrake & Rupley algorithm [246].

CDR Vicinity

The 'CDR vicinity' comprises every surface-exposed IMGT-defined CDR and anchor residue, and all other surface-exposed residues with a heavy atom within a 4 Å radius.

Salt Bridges

Salt bridges were defined as pairs of lysines/arginines and aspartic acids/glutamic acids with a N^+-O^- distance $\leq$ 3.2 Å.

Hydrophobicity

Where R_1 and R_2 are two surface-exposed residues with a closest heavy atom distance, r_{12}, < 7.5 Å and H(R,S) is the normalised hydrophobicity score (between 1 and 2) for residue R in scheme S, the Patches of Surface Hydrophobicity (PSH) metric can be calculated as: $\sum_{R_1 R_2} \frac{H(R_1,S)H(R_2,S)}{r_{12}^2}$. The hydrophobicity scales tested were: Kyte & Doolittle [268], Wimley & White [269], Hessa *et al.* [270], Eisenberg & McLachlan [271] and Black & Mould [272]. Salt bridge residues were assigned the same value as glycine in each hydrophobicity scale.

Charge

The following charges were assigned by sequence: Aspartic acid: -1, Glutamic acid: -1, Lysine: +1, Arginine: +1, Histidine: +0.1 (Henderson-Hasselbalch equation applied: pk_a 6, pH 7.4, and rounded-up to one decimal place). Tyrosine hydroxyl deprotonation was not considered. Salt bridge residues were assigned a charge of 0. The Patches of Positive Charge (PPC) and Patches of Negative Charge (PNC) metrics are analogous in form to PSH, with H(R,S) substituted for |Q(R)|, the absolute value of the charge assigned to residue R. Structural Fv Charge Symmetry Parameter (SFvCSP) values were calculated as: $\left[\sum_{R_H} Q(R_H)\right]\left[\sum_{R_L} Q(R_L)\right]$, where R_H, R_L are surface-exposed V_H, V_L residues respectively.

Statistical Sampling for the TAP Metrics

We performed statistical sampling over all metric distributions to add error bars to the threshold values. Across 1000 repeats, we randomly selected 200 of the 242 CSTs, and calculated their amber (5th and/or 95th percentile value, depending on the metric) and red (0th and/or 100th percentile value, depending on the metric) thresholds. The results are presented in Table A4.4 and show that our threshold values are relatively robust to selecting different therapeutics.

Webapp Implementation

The original web application for TAP (released on paper submission, June 2018) used a combination of PHP and JavaScript to dynamically handle user sequence inputs. For consistency with our new SAbBox Virtual Box suite, the latest version (with the assistance of Dr. Claire Marks) has been transferred to a Flask database framework (*https://github.com/pallets/flask*), which handles dynamic inputs using Python.

4.4 Results

4.4.1 Sequence Data

As an initial dataset of mAbs unlikely to suffer with developability issues, we used the variable domain heavy and light chain sequences of 137 post Phase-I clinical-stage antibody therapeutics ('137 CSTs') [273][2]. To contextualise the properties of the CST set, we retrieved Vander Heiden's recent snapshot of the human antibody repertoire from the Observed Antibody Space database [88, 89] ('human VdH Ig-seq'). We also used a larger proprietary dataset procured by UCB Pharma Ltd. ('human UCB Ig-seq'). All comparisons in this chapter are made to the Vander Heiden data, with UCB comparisons available in the Appendix. Each human Ig-seq dataset was analyzed as a set of non-redundant heavy or light chains ('human Ig-seq non-redundant chains'), and as a set of non-redundant CDR sequences ('human Ig-seq non-redundant CDRs'). We chose these Ig-seq datasets as they contain simultaneously sequenced heavy and light chains, and so are a promising starting point for realistic *in silico* pairing, required to make sets of Ig-seq Fv models (a.k.a. Antibody Model Libraries, see Chapter 3).

[2]Before Thera-SAbDab, this was the largest collation of therapeutic antibody sequence data. For metric testing, we used the additional post Phase-I therapeutic sequences from the prototype version of Thera-SAbDab.

4.4.2 Model Structures

High quality structural information is critical to accurately predict the surface properties of antibodies. However, 'gold-standard' representations such as solved X-ray crystal structures are almost always unavailable in early-stage drug discovery. All our comparisons are therefore made between models, proving applicability to a real-world scenario. We build models even when a solved crystal structures is available, to account for systematic biases associated with the modelling process (e.g. higher values for the Patches of Surface Hydrophobicity metric; see Figs. A4.1 and A4.2, and Section 4.4.4).

ABodyBuilder [44] was run on the 56 CSTs with a reference PDB [19] structure (as of May 4$^{\text{th}}$, 2018). Sequence identical templates were not included, and each resulting model was aligned to its reference to evaluate the backbone root-mean-square deviation (RMSD) across all IMGT regions (see Methods). The mean framework and CDR RMSDs (Table A4.1) were commensurate with the current state of the art [134]. For our structural property calculations, we class surface-exposed residues as having a side chain with relative accessible surface area ($\text{ASA}_{\text{rel,X}}$) $\geq 7.5\%$, compared to Alanine-X-Alanine for each residue X [246, 267]. Using this definition, we identified all exposed residues in the models and PDB structures. Of the 7,057 exposed crystal structure residues, only 265 (3.76%) were wrongly assigned as buried in the models.

As these results suggest that ABodyBuilder models are accurate enough for our analysis, we used this software to model all 137 CSTs ("137 CST models") and diverse subsets of paired human VdH Ig-seq chains (14,072 'human VdH Ig-seq models') and paired human UCB Ig-seq chains (19,019 'human UCB Ig-seq models'). The pairing and modelling protocol was designed to capture the sequence and structural diversity in each dataset, within the constraints of modellability and computational expense (see Section 4.3; this was an early version of the Antibody Model Library generation protocol from Chapter 3). We then performed a series of *in silico* assays to determine the typical physicochemical properties of CSTs and natural human antibodies, which we then use to determine the TAP developability guidelines.

4.4.3 Physicochemical Properties

4.4.3.1 CDR Lengths

Loop length has a major impact on the nature of antigen binding. For example, if an antibody has a long CDRH3 loop, it tends to form most of the interactions with

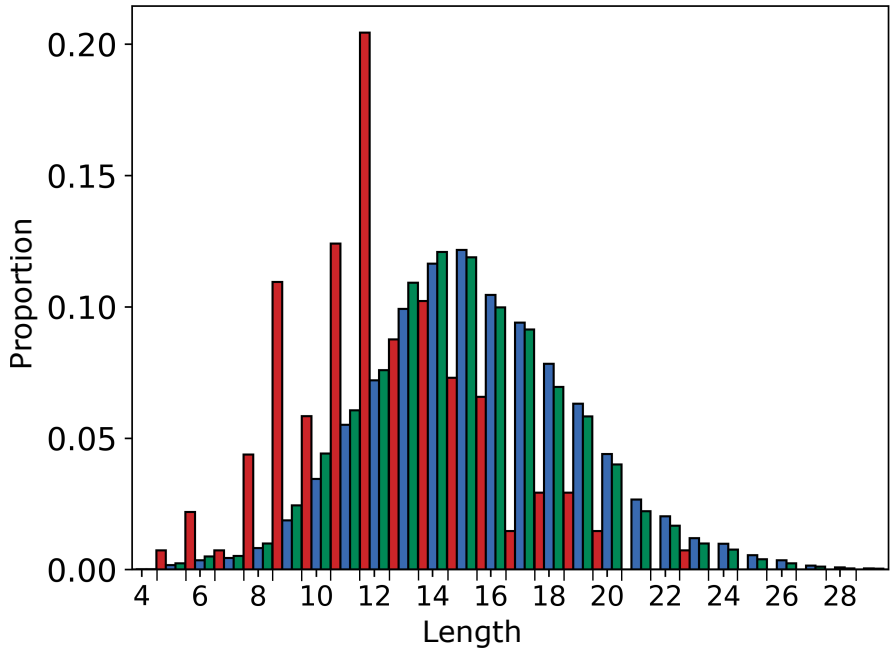


Figure 4.1: Comparing the CDRH3 length distributions of the 137 CSTs (red), 105,458 human VdH Ig-seq non-redundant CDRH3s (blue), and 551,193 human Ig-seq non-redundant heavy chains (green). The CSTs have a lower median CDRH3 length.

an antigen, while shorter CDRH3 loops tend to be part of concave binding sites with other CDRs often assisting in binding [274].

The 137 CST and human Ig-seq sequences were IMGT-numbered [11], and IMGT CDR definitions were used to split the sequences by region. The 137 CST CDRH3 loops had a median length of 12, compared to 15 for the human VdH Ig-seq dataset (Fig. 4.1). In the case of CDRL3 the distributions were closer, with a median length of 9 for the 137 CSTs and the human VdH Ig-seq data (Fig. A4.3E). The other CDR loops, whose length diversities are restricted by V-gene encoding, displayed similar ranges for both CST and Ig-seq datasets. An unexpected observation was a bimodal distribution in CDRL1 length for the CSTs (Fig. A4.3C). No correlation was found between shorter/longer CDRL1 length and antigen size (hapten/cytokine/receptor) or kappa *versus* light chain usage.

To test whether hybridomal development might account for CST *versus* Ig-seq CDR length discrepancies — as it is known that mouse antibodies tend to have different CDR length properties to human antibodies (e.g. shorter CDRH3 loops [275]) — we split the 137 CST dataset by developmental origin (Fig. A4.5). Fully human therapeutics were disproportionately represented at longer CDRH3s (mean: 13.21, median:

12), compared to chimeric, humanised, or fully murine therapeutics (mean: 11.91, median: 12). However, both therapeutic subsets still have shorter CDRH3s than human-expressed antibodies. On the other hand, the observed bimodal CDRL1 distribution across all CSTs appears to be strongly driven by the humanised/chimeric/murine subset (Fig. A4.5D).

The combined length of all CDRs for each antibody in the 137 CST dataset had a median value 48 (Fig. A4.6). The 137 CST total CDR length was highly correlated to CDRH3 length (Pearson's correlation coefficient of +0.77, with a two-tailed p-value of $2.44e^{-28}$). This significant co-variance means that total CDR length can be used as a reliable proxy for CDRH3 length (which is typically shorter in CSTs than human Ig-seq heavy chains). As it can also capture global binding site topology, which may be relevant to polyspecificity, total CDR length was selected for inclusion in the final five TAP guidelines.

Neither human Ig-seq dataset is natively paired, so as described in the Methods (Section 4.2) we generated artificially paired human Ig-seq models for subsequent structure-dependent metric evaluation. These models displayed a similar range of total CDR lengths to the CSTs, which a moderate bias towards longer lengths (Fig. A4.6), meaning this property is unlikely to be a trivial determinant of differences across structure-dependent metrics. The total CDR length difference is less profound than one might expect given the difference in raw sequence CDRH3 lengths, as the paucity of long CDRH3 structural templates frequently renders antibodies containing longer CDRH3s unmodellable, imposing a compensatory negative selection pressure against longer total CDR lengths in the models.

4.4.3.2 Canonical Forms

In natural antibodies, all CDR loops, apart from CDRH3, are thought to fall into structural classes known as canonical forms [12, 22]. It may be advantageous from an immunogenicity perspective to ensure that CST CDRs adopt canonical forms previously observed in natural human antibodies.

We assigned canonical forms (see Section 4.3) to the 137 CST and human Ig-seq model CDRs. All assignable CST model CDRs were labelled with a canonical form also present in at least one human VdH Ig-seq model dataset (Figs. 4.2 and A4.7). Fewer than 19% of CST CDRs were unassignable in each loop region, suggesting that, despite engineering, a clear majority of non-CDRH3 CST CDR loops adopt well-characterised canonical forms. While not an explicit TAP metric, we decided to

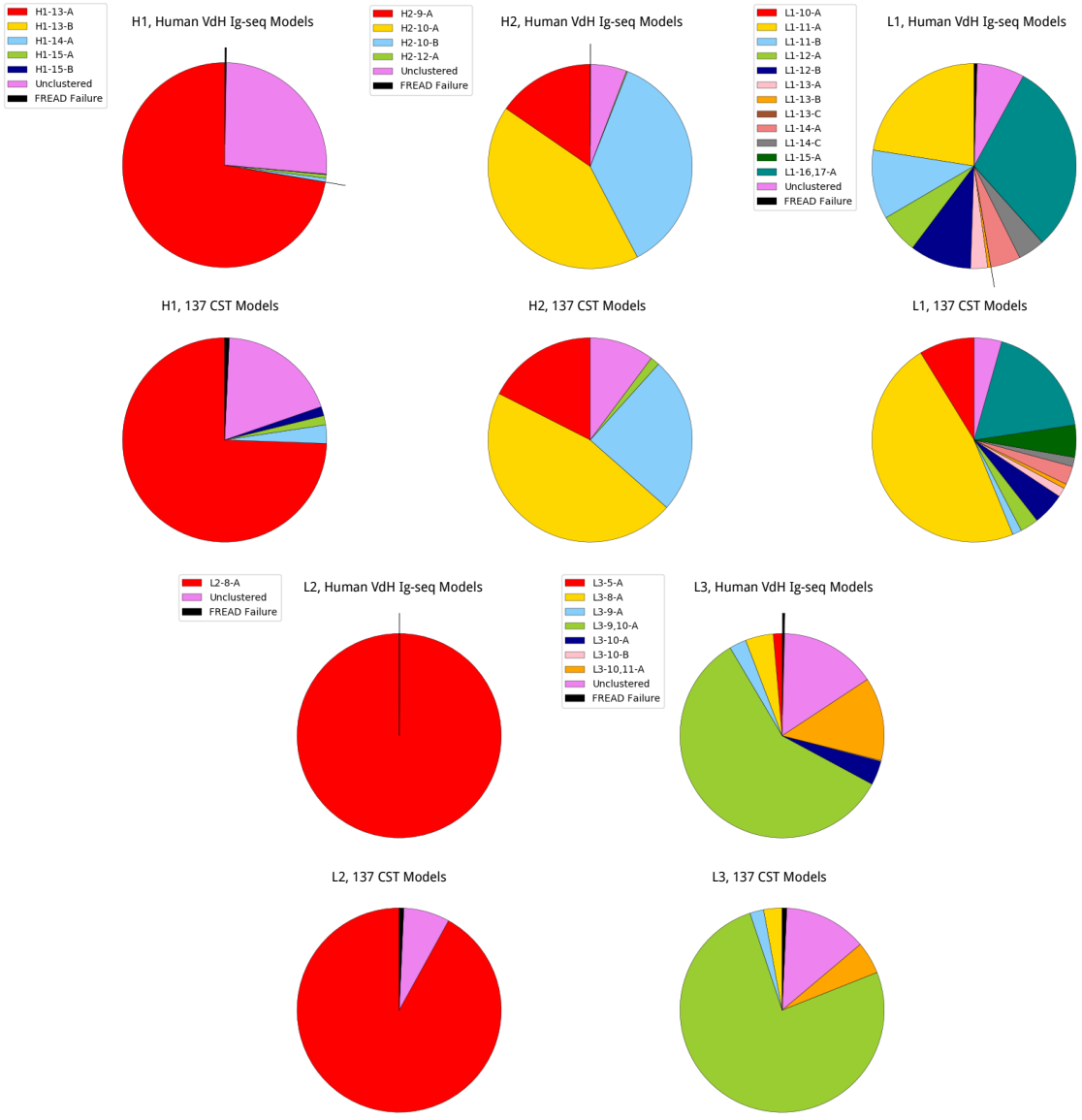


Figure 4.2: Length-independent canonical form assignments for the 137 CST and human VdH Ig-seq models.

report the canonical form of each modelled loop, highlighting any CDR that cannot be assigned as a potential developability liability.

4.4.3.3 Hydrophobicity

Hydrophobicity in the CDR regions has been repeatedly linked to aggregation propensity in mAbs [179, 186, 188, 189]. Using our homology models, we estimated the effective hydrophobicity of each residue by considering not only its degree of apolarity, but also whether or not it is solvent-exposed (side chain $ASA_{rel} > 7.5\%$ [246, 267]). As the energy of the hydrophobic effect is approximately proportional to the interface area [276], we developed a metric (Patches of Surface Hydrophobicity, PSH, see Section 4.3) that yields higher scores if hydrophobic residues tend to neighbor one another in a region, rather than being evenly separated. We evaluated PSH for the 137 CST and human Ig-seq models across two regions (the CDR vicinity (see Section 4.3) and the entire variable (Fv) region), and with five different hydrophobicity scales [268–272].

The results of all hydrophobicity scales were highly correlated (e.g. $R^2 \geq 0.91$ between all scales in the CDR vicinity), and so we use the Kyte & Doolittle [268] scale for all subsequent comparisons. The mean CDR vicinity PSH values for the CST and human VdH Ig-seq distributions were 123.30 ± 16.60 and 133.76 ± 21.08 respectively (Fig. 4.3A). CSTs were noticeably underrepresented at higher CDR PSH values; Galiximab is a rare example of a therapeutic antibody with a high value (Fig. 4.3B). A similar divergence occurred across the entire Fv region, with mean values of 357.69 ± 22.95 and 370.56 ± 24.45 respectively (Fig. A4.8), implying the primary difference occurs within the CDRs. This supports the theory that the high concentration conditions under which therapeutics are stored may render them less tolerant of large patches of hydrophobicity in the highly-exposed CDR vicinity, and also suggests that a subset of natural human antibodies would be unsuitable therapeutic candidates. We therefore included the CDR vicinity PSH score as a TAP metric.

4.4.3.4 Charge

Surface patches of positive or negative charge have also been linked to negative biophysical characteristics [196, 257]. We calculated two metrics designed to highlight regions of dense charge: the Patches of Positive Charge (PPC) and Patches of Negative Charge (PNC) measures (see Section 4.3). All surface residues were initially assigned the appropriate charge for their averaged pk_a values, as neighboring residues

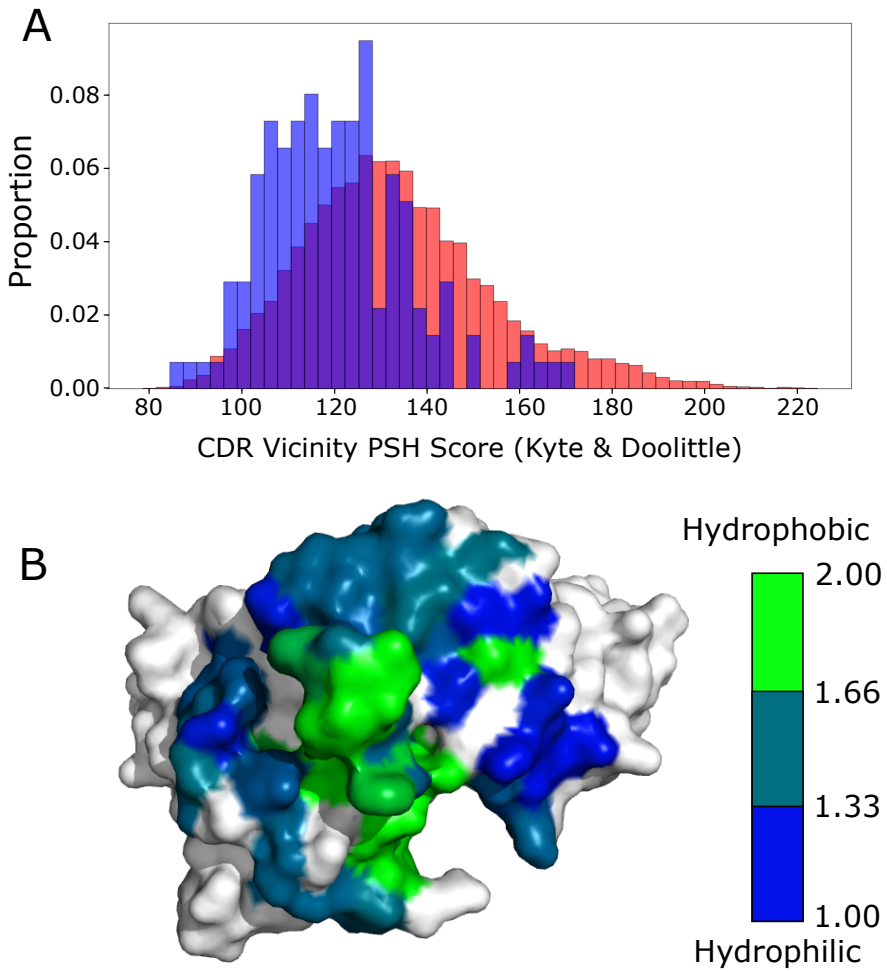


Figure 4.3: (A) CDR vicinity Patches of Surface Hydrophobicity (PSH) scores across the 137 CST (blue) and human VdH Ig-seq (red) models. The CSTs are underrepresented at higher PSH values. (B) Galiximab (Kyte & Doolittle CDR vicinity PSH score of 167.89) has a large surface-exposed patch of hydrophobicity in its CDRH3 loop. Heavy and light chain surfaces outside the CDR vicinity are coloured in white.

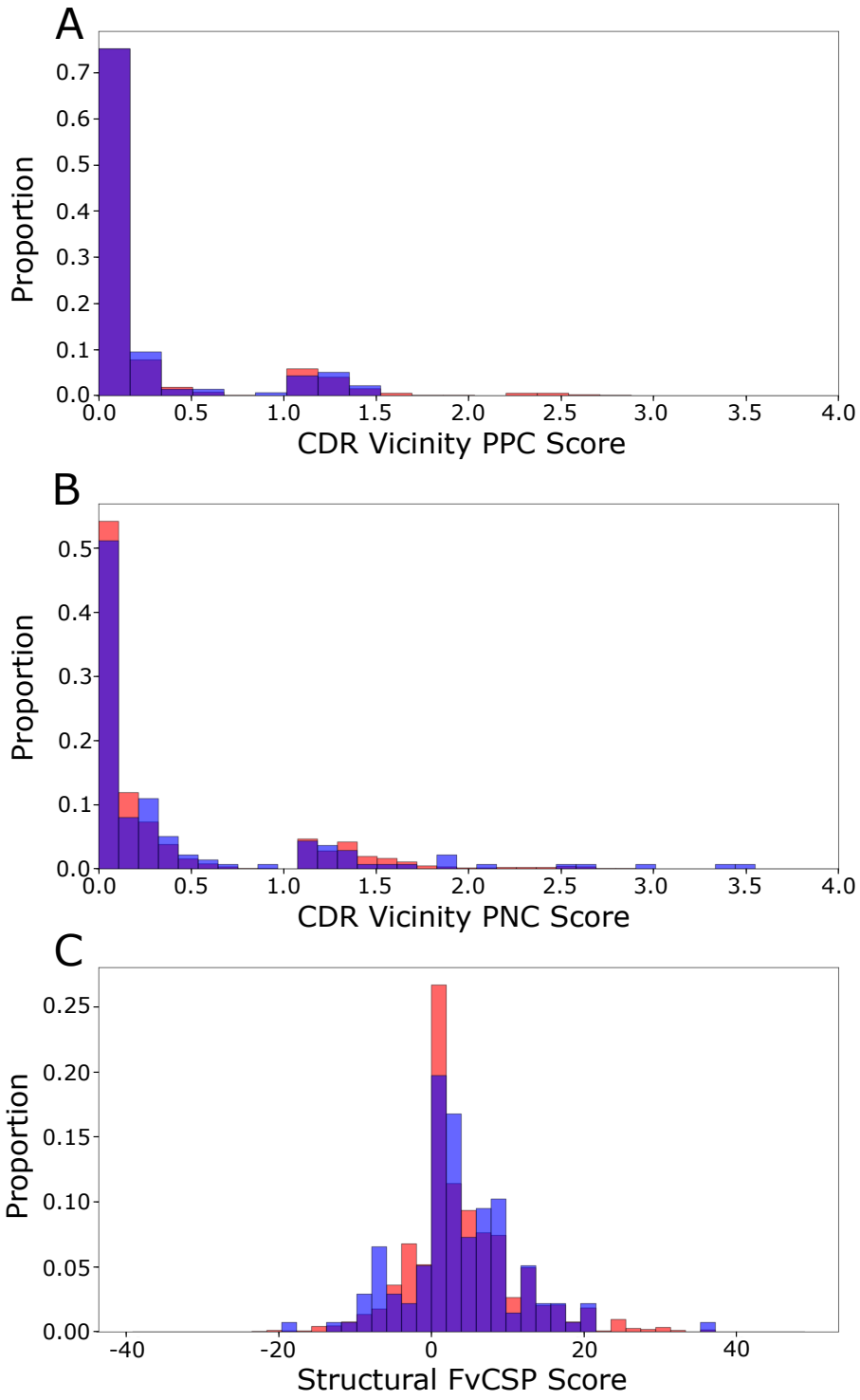


Figure 4.4: Histograms of 137 CST (blue) and human VdH Ig-seq model (red) values for the (A) Patches of Positive Charge (PPC) and (B) Patches of Negative Charge (PNC) metrics in the CDR vicinity. In both measures, the datasets are biased away from higher scores. (C) Histogram of Structural Fv Charge Symmetry Parameter values. Both datasets show a bias away from negative values.

appear to have a limited effect at pH 7.4 [183]. The charge of residues found to be engaging in salt bridges was then revised to zero.

The 137 CST models tend to avoid patches of charge in their CDR vicinities, with 88.32% and 80.30% having PPC (Fig. 4.4A) and PNC (Fig. 4.4B) values below 1, respectively. The human VdH and UCB Ig-seq models displayed similar PPC and PNC distributions. Both PPC and PNC assays were carried forward as TAP metrics.

When mAbs have oppositely charged V_H and V_L chains, they typically have higher *in vitro* viscosity values [183]. This aggregate-inducing electrostatic attraction is captured at the sequence level by the Fv Charge Symmetry Parameter (FvCSP) metric - the mAb tends to be more viscous if the product of net V_H and V_L charges is negative [183]. Harnessing our structural models, we calculated a variant (the Structural Fv Charge Symmetry Parameter, SFvCSP), which only includes residues that are surface-exposed, and not locked in salt bridges, in the evaluation of net charge. In Galiximab, for example, we 'correct' the charge of arginine H108 and aspartic acid L56 to 0, as the model indicates that they form a salt bridge. The charges of the glutamic acid at position H6, the aspartic acids at positions H107, L98, and L108, and the histidine at position L40 are ignored as their side chains are buried. The FvCSP score for this antibody would be 0 (net heavy chain charge of 0, net light chain charge of -2.9), while the SFvCSP score is +2.0 (net heavy chain charge of +2, net light chain charge of +1). A similarly low percentage of CST models (21.9%) and human VdH Ig-seq models (20.8%) had negative SFvCSP scores (Fig. 4.4C), with mean values of 3.34 ± 7.44 and 3.67 ± 7.40 respectively. With such a bias away from negative products in both natural and therapeutic antibodies, we chose the SFvCSP as our final TAP property.

4.4.4 The Effect of Modeling

When comparing the TAP metric values obtained for the 56 CST structures and their corresponding models, we saw positive correlations across all metrics, though PSH correlation was weaker than the other three (Fig. A4.2). The positive trends indicate that calculations performed on ABodyBuilder models are typically predictive of the results that would be obtained from a crystal structure, and therefore that threshold values derived from models are sufficiently informative across these metrics. The weaker correlation observed in the PSH score, which depends on the proximity of many residues' side chains, exemplifies how more fine-grained analyses may be expected to add noise rather than meaning to metric evaluation and highlights why we

decided not to consider atom-by-atom contributions as per the SAP hydrophobicity metric.

To confirm that the PSH values of natural and therapeutic antibodies differ even without modelling, we mined SAbDab [9] to find all the human, non-engineered, non-redundant (at 100% sequence identity) X-ray crystal structures in the PDB [19] (June 2018). We found only 33 such mAbs (identities listed in Dataset S1 of Raybould *et al.* [214]), as most human mAb PDB entries involve some degree of engineering. We found approximately the same difference in mean CDR Vicinity PSH score between therapeutic and human crystal structures as we did between therapeutic and human VdH Ig-seq models (-9.69 and -10.46 respectively, see Table A4.2). We note that if we had compared human *structures* to therapeutic models, we would not have detected a significant difference (therapeutic models: 123.30 ± 16.60; human structures: 124.61 ± 16.54). This systematic bias towards higher PSH values in models is seen most clearly when comparing the values for CST crystal structures with CST models (Fig. A4.1). Through inspection, we observed that this bias derives from our modelling software frequently underestimating the tightness of intramolecular residue-residue packing, leading to more residues being classified as surface-exposed and contributing to the overall metric value. This effect is particularly amplified for the PSH metric, as any surface-exposed residue can contribute to the score, whereas in the PPC, PNC, and SFvCSP metrics, only charged residues can contribute.

4.4.5 Developability Guidelines

While CSTs predictably share many features in common with human antibodies, our CDR length and hydrophobicity distributions imply that not every human antibody would make a good therapeutic. Consequently, our developability guidelines were set solely by CST values across the five selected metrics (Table 4.1); an amber flag indicates that the antibody lies within the extremes of the distribution, whereas a red flag indicates a previously unobserved value for that property. Total CDR Length and PSH amber and red flags are set at the 5^{th} and 95^{th}, 0^{th} and 100^{th} percentiles respectively, capturing unusually flat/concave binding sites and regions of unusually high polarity, as well as longer protruding CDRH3s and regions of high hydrophobicity. The PPC and PNC metrics only have upper-bound flags (amber: 95^{th} percentile, red: 100^{th} percentile), as low values are not intuitively linked to instability. Similarly, the SFvCSP metric only has lower-bound flagging (amber: 5^{th} percentile, red: 0^{th} percentile), as strong inter-domain repulsion ought to aid colloidal stability. These flagging percentiles were chosen to include the long-tail regions of each CST

Metric	Amber Flag Region	Red Flag Region
1. Total CDR Length	Bottom 5%, Top 5%	Above or Below
2. PSH, CDR Vicinity	Bottom 5%, Top 5%	Above or Below
3. PPC, CDR Vicinity	Top 5%	Above
4. PNC, CDR Vicinity	Top 5%	Above
5. SFvCSP	Bottom 5%	Below

Table 4.1: TAP amber and red flag cut-off thresholds, with respect to the clinical-stage therapeutic distributions.

Metric	Amber Flag Region	Red Flag Region
Total CDR Length	$39 \leq L \leq 43$	$L < 39$
	$54 \leq L \leq 60$	$L > 60$
PSH, CDR Vicinity	$83.84 \leq PSH \leq 100.71$	$PSH < 83.84$
	$156.200 \leq PSH \leq 173.850$	$PSH > 173.850$
PPC, CDR Vicinity	$1.25 \leq PPC \leq 3.16$	$PPC > 3.16$
PNC, CDR Vicinity	$1.84 \leq PNC \leq 3.50$	$PNC > 3.50$
SFvCSP	$-20.40 \leq SFvCSP \leq -6.30$	$SFvCSP < -20.40$

Table 4.2: TAP amber and red flag regions, as defined by the entire set of 242 CSTs. PSH score is calculated with the Kyte & Doolittle hydrophobicity scale. L = Length.

distribution (but may change over time with industrial feedback) and allow a user to know whether their tested antibody sits in the extremities of a CST physicochemical property.

To confirm that these threshold definitions do not typically flag mAbs without developability issues, we identified a further 105 mAb therapies ('105 CSTs', listed in Dataset S2 of Raybould *et al.* [214]), not included in the 137 CST dataset, that had advanced to at least Phase II in clinical development.

Only eight of this set (7.69%) were assigned a red developability flag according to the boundaries set by the 137 CSTs, an average of 0.08 red flags per newly-tested therapeutic (Table A4.3). Erenumab received the most red flags - for total CDR length (60), CDR vicinity PSH (173.85), and CDR vicinity PPC (1.53). All other red-flagged therapeutics received only one: rafivirumab for total CDR length (60), intetumumab for CDR PSH (83.84), adacanumab, derlotuximab, lanadelumab and teprotumumab for CDR PPC (2.67, 2.66, 2.48, and 3.16 respectively), and quilizumab for Fv charge asymmetry (-20.40). That some of these 105 therapeutics were red-flagged is not surprising for several reasons. Firstly, our threshold percentiles have not been calibrated and may well change over time. Secondly, the TAP metrics reflect average properties across all CSTs and it is entirely possible for individual therapeutics to

stray into extreme physicochemical space with compensatory measures (see Section 4.4.6 for an example where an amber-flagged PNC therapeutic candidate was rescued by engineering to compensate for the necessary charge patch). Finally, monoclonal antibodies are not delivered in isolation but rather in formulation, and advances in this technology may also compensate for previously deleterious physicochemical characteristics. Nevertheless, the low red-flagging rate confirms that the ranges of the physicochemical distributions on which the TAP guidelines were based are broadly representative of therapeutic-like antibodies.

Incorporating both sets of CSTs into a larger dataset ('242 CSTs') led to the new guideline values shown in Table 4.2. While most metrics were only slightly adjusted, the PPC thresholds changed quite significantly, showing the importance of regularly updating the metric distributions as more CSTs are developed. As a result, we performed statistical sampling over our TAP metric distributions to give a sense of the error that might be inherent in these current threshold values (see Section 4.3; Table A4.4). All 242 CST TAP metric values are listed in Dataset S3 of Raybould *et al.* [214].

4.4.6 Case Studies

We tested whether these updated guideline values could highlight candidates with developability problems by building models and running TAP on two datasets supplied by MedImmune (Fig. 4.5). A lead anti-NGF antibody, MEDI-578, showed minor aggregation issues during *in vitro* testing, of a level usually rectifiable in development, whereas the affinity matured version, MEDI-1912, exhibited unrectifiably high levels of aggregation [180]. This observation was rationalised through SAP score [188] values, indicating that a large hydrophobic patch on the surface was responsible. TAP assigns MEDI-578 an amber flag, and MEDI-1912 a red flag - by a large margin - in the CDR vicinity PSH metric (Fig. 4.5A). The paper describes how back-mutation of three hydrophobic residues in MEDI-1912 to those of MEDI-578 led to MEDI-1912STT, fixing the aggregation issue while maintaining potency. TAP assigns MEDI-1912STT no developability flags (Fig. 4.5A).

A lead anti-IL13 candidate, AB008, had no developability issues, but the affinity-matured version, AB001, had very poor levels of expression (seven times lower than AB008) [196]. The authors highlighted the role of four consecutive negatively charged residues in the L2 loop – mutation of the fourth negatively charged residue to neutral asparagine (AB001DDEN) was able to stabilise the loop backbone, mitigating the ionic repulsion of the DDE motif, and returning acceptable levels of expression. TAP

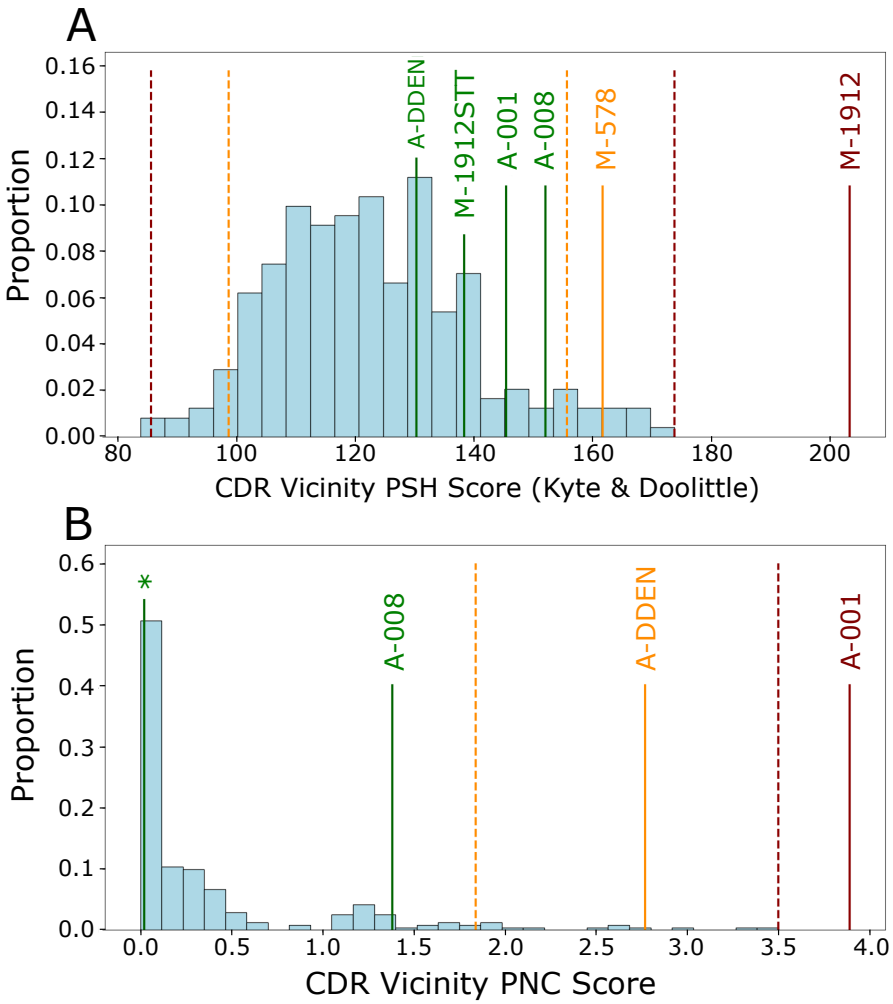


Figure 4.5: The (A) CDR vicinity PSH and (B) CDR vicinity PNC metrics for the combined set of 242 CSTs (light blue), and MedImmune case studies (coloured by assigned flag). MEDI-578, MEDI-1912, and MEDI-1912STT all have the CDR vicinity PNC value labelled by an asterisk. Amber and red dashed lines delineate the 242 CST guideline thresholds. Case studies with prohibitive developability issues (MEDI-1912, AB001) are red-flagged for the PSH and PNC metrics respectively. Engineered versions without developability issues (MEDI-1912STT, AB001DDEN) return to the range of values previously seen in CSTs for all metrics. MEDI-578/1912/1912STT are labelled as M-578/1912/1912STT, and AB-001/008/001DDEN are labelled as A-001/008/DDEN for legibility.

assigns no developability flags to AB008, but a red flag to AB001, and an amber flag to AB001DDEN for its CDR vicinity PNC metric (Fig. 4.5B), again red-flagging the candidate with prohibitive developability issues. Both AB001 and AB008, confirmed monomers in solution [196], did not flag for CDR vicinity PSH score (Fig. 4.5A).

4.4.7 TAP Web Application

We have packaged TAP into a web application, available at *http://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/TAP.php*. TAP only requires the heavy and light chain variable domain sequences as an input, returning a detailed profile of an antibody with a typical run-time of less than 30 seconds. Flags (green, amber, or red) are assigned to each of the TAP metrics, with accompanying histograms. An interactive molecular viewer allows the user to visualise hydrophobicity (Fig. 4.6A), charge, and probable sequence liabilities on the antibody model surface. Estimated model quality can be easily accessed to help guide interpretation of the results (Fig. 4.6B). Finally, canonical forms are assigned to each non-CDRH3 loop. A full sample output is shown in Fig. 4.7.

4.5 Discussion

We have analyzed several properties linked to poor developability across 242 post-Phase I therapeutics, with the assumption that mAbs that have reached this stage of clinical trials have characteristics amenable to therapeutic development.

By analyzing these properties, we have found evidence that suggests that not every human antibody would make a good therapeutic. This is somewhat intuitive, as therapeutics suffer a range of stresses during development (including variation in pH and temperature, sheer forces, and high concentration storage conditions) that human-expressed antibodies are not exposed to. The TAP metrics therefore depend on the values seen across CSTs alone.

Our simple TAP guidelines will not capture the whole spectrum of developability issues. For example, they will not detect sources of immunogenicity, nor more subtle mechanisms that lead to poor stability. Nevertheless, we have shown that the TAP guidelines can selectively highlight antibodies with expression or aggregation issues [180, 196].

We will regularly recalculate the threshold values to include new mAbs that have entered Phase-II of clinical trials. This will also allow for the inevitable fluctuation in

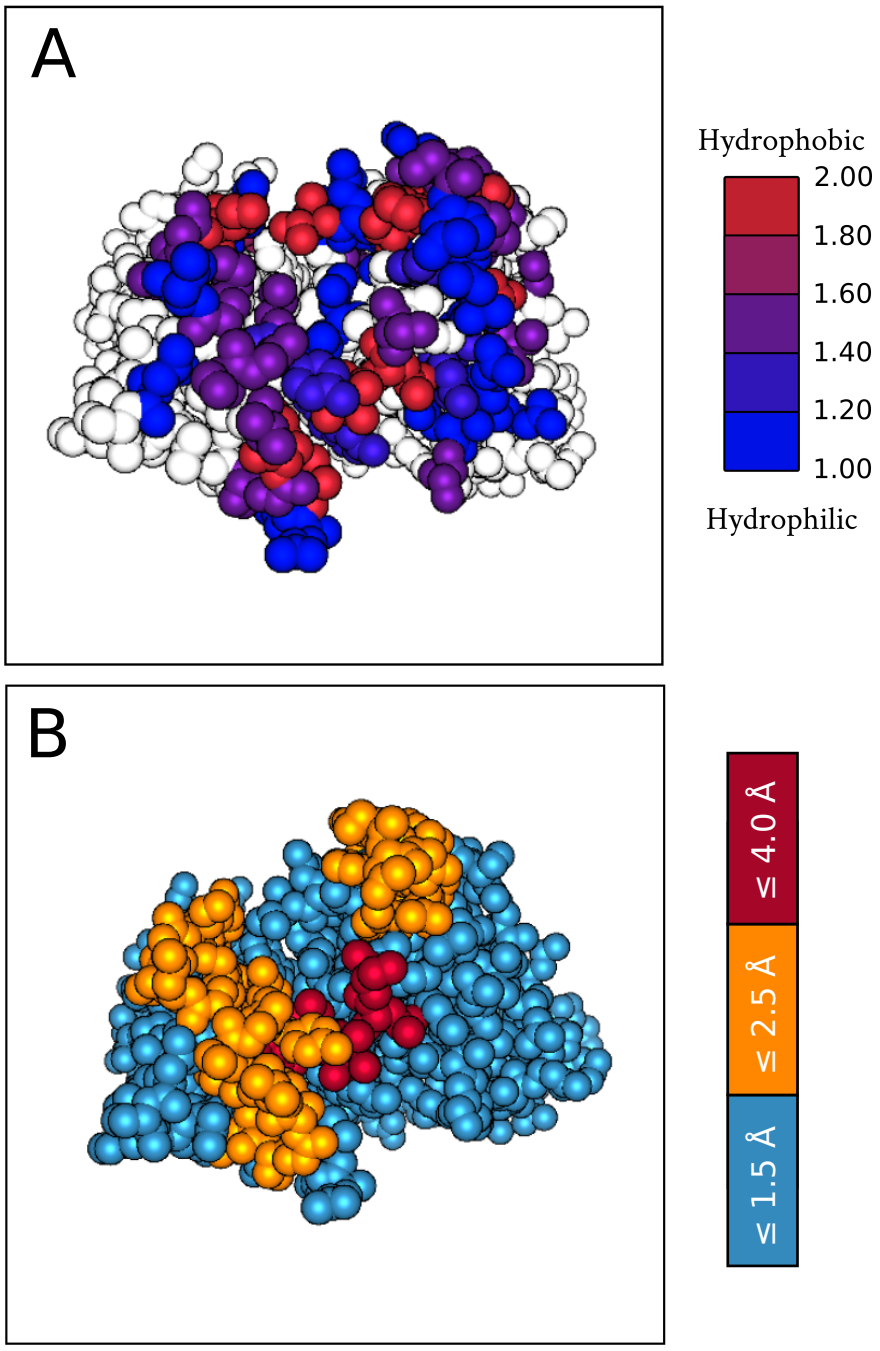


Figure 4.6: (A) An example TAP web application output showing the heavy atoms of an antibody as spheres coloured by the hydrophobicity (Kyte & Doolittle scale, normalised between 1 and 2) of each residue in the CDR vicinity. (B) The ABodyBuilder predicted model accuracy assignments [44] for each IMGT region, with heavy atoms shown as spheres. These are coloured according to three backbone RMSD thresholds at a 75% confidence interval (both thresholds and confidence intervals can be modified in the web application). Better quality models will yield more reliable TAP metric values.

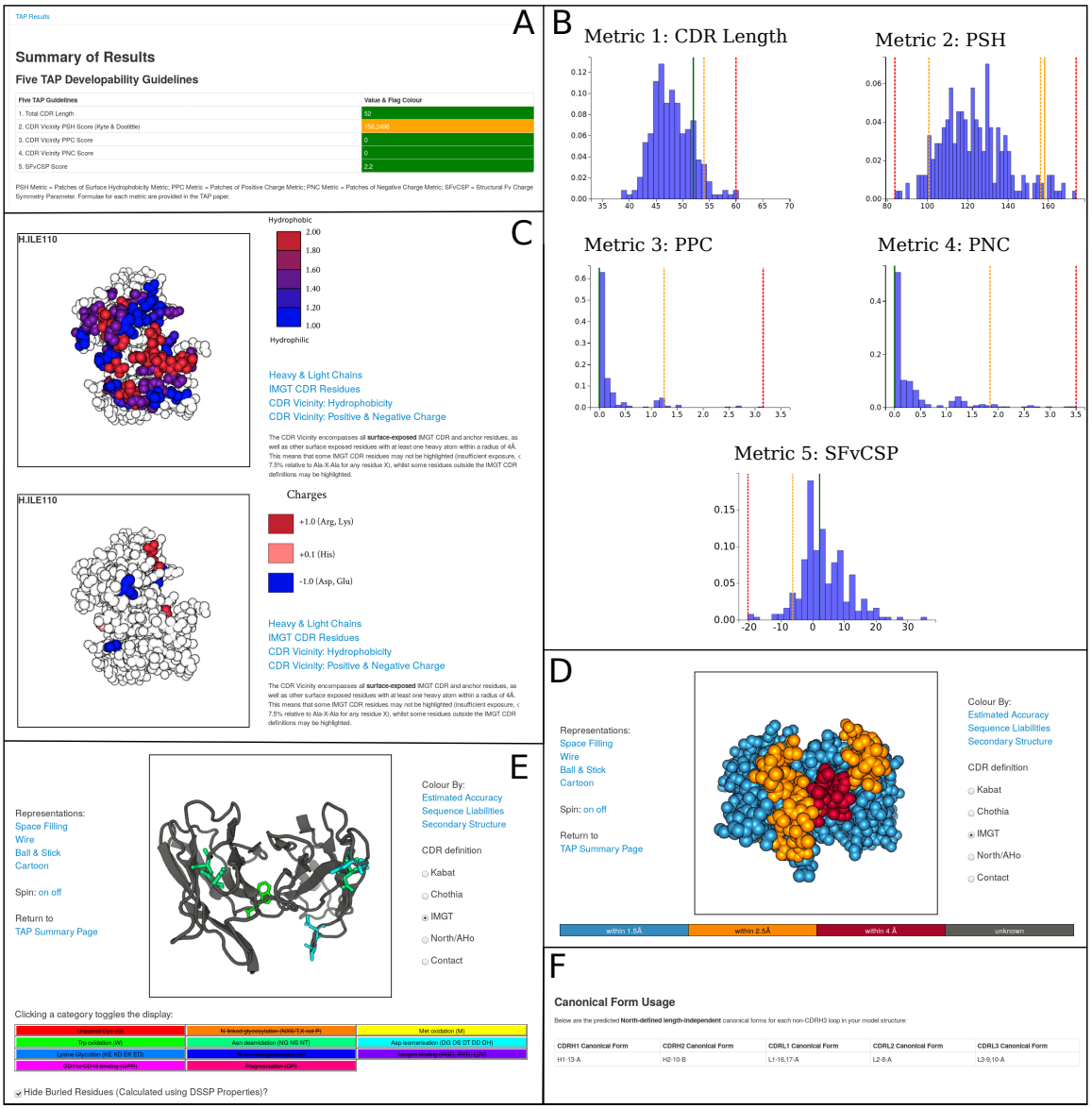


Figure 4.7: A sample TAP web application output. (A) The five TAP metric values are reported in a table, whose cells are coloured by assigned flag. (B) The metric values are also shown against the distributions of the 242 CSTs. Amber and red dashed lines indicate the amber and red flag threshold values, and the solid lines the value for the inputted antibody, coloured by assigned flag. (C) A molecular viewer allows for visualization of hydrophobicity and charge in the CDR vicinity. In this antibody, the charge is evenly spread across the CDR vicinity, while there is a large patch of hydrophobicity spanning the CDRH1 and CDRH3 loops. (D) Predicted model accuracy can be visualised across each IMGT region. The confidence interval and all RMSD cutoffs can be altered. (E) Potential sequence liabilities are shown on a cartoon representation of the antibody. A check box is available to hide buried residues. (F) North-defined, length-independent canonical forms are reported for each CDR loop in the antibody.

PSH, PPC, PNC, and SFvCSP values returned by CSTs, as ABodyBuilder models improve in accordance with the growing number of antibodies in SAbDab [9].

When enough CSTs are available, it may be possible to stratify the therapeutic guidelines into subclasses. For example, separate thresholds could be considered for mAbs involving kappa or lambda light chains. Lambda light chains tend to contribute to higher average CDR Vicinity PSH values across our 242 CST, 14,072 human VdH Ig-seq, and 19,019 human UCB Ig-seq models (Table A4.5); DeKosky *et al.* [277] also found, across their 2,000 natively-paired models of mAbs, that lambda CDRL3 loops are significantly more hydrophobic than their kappa equivalents. As only 25 of the 242 CSTs contain lambda light chains, we do yet not have enough data to safely determine a guideline threshold. Nevertheless, as around 90% of post-Phase I CSTs are derived from kappa light chains, this could suggest that hydrophobicity-driven developability issues are far more prevalent when using leads containing lambda light chains.

Other subclasses could include clinical trial progression, active/discontinued trial status, or therapeutic species origin. At this stage, neither splitting by clinical progression (Table A4.6) nor drug campaign status (Table A4.7) leads to significant differences in mean metric values. Human and humanised mAbs have noticeably higher mean PSH values than chimeric or mouse mAbs (Table A4.8) – with the caveat that there are only 36 mAbs in the latter category.

As with the Lipinski rule of five, the thresholds themselves should not be interpreted as hard-and-fast rules, and the distance of red-flagged candidates outside the previously-observed bounds should be taken into consideration. Advances in process development and formulation may soon redefine the limits of permissible values [273].

4.6 Update and Chapter Conclusion

Since TAP was developed, we have found sequences of an additional 135 post Phase-I CSTs through our Thera-SAbDab database [99] (see Chapter 2). Together with the ever-growing number of loop structural templates from the PDB, this has resulted in revised guidelines (Table 4.3). Most properties have seen only a modest adjustment of their amber and red flag threshold values, remaining within 5% of their original values. The exception to this is a marked increase in the red flag threshold for Patches of Positive Charge (PPC) score, from 3.16 to 3.76 (+18.8%), overtaking the maximum value seen for Patches of Negative Charge (PNC, 3.47). The amber threshold (95^{th} percentile) for the PPC metric actually reduced slightly from 1.25

Metric	Amber Flag Region	Red Flag Region
Total CDR Length	$39 \leq L \leq 42$ $55 \leq L \leq 60$	$L < 39$ $L > 60$
PSH, CDR Vicinity	$88.29 \leq PSH \leq 100.79$ $158.47 \leq PSH \leq 179.18$	$PSH < 88.29$ $PSH > 179.18$
PPC, CDR Vicinity	$1.24 \leq PPC \leq 3.76$	$PPC > 3.76$
PNC, CDR Vicinity	$1.83 \leq PNC \leq 3.47$	$PNC > 3.47$
SFvCSP	$-19.50 \leq SFvCSP \leq -6.02$	$SFvCSP < -19.50$

Table 4.3: TAP amber and red flag regions, as defined by all post Phase-I CSTs in Thera-SAbDab as of July 2020. PSH score is calculated with the Kyte & Doolittle hydrophobicity scale. L = Length.

to 1.24 (-1.04%), suggesting the red threshold increase is due to outlying mAbs. This could be indicative of poor ABodyBuilder models, or could meaningfully reflect improvements in formulation technology/antibody engineering that compensate for larger charge patches (such as designed-in backbone hydrogen bond stabilisation as seen in A-DDEN, Fig. 4.5).

As a component of the SAbBox suite of informatics tools, TAP is currently installed at eleven pharmaceutical companies and, thanks to our new academic SAbBox license, will increasingly be used in an academic context. We hope that this will lead to further refinements in metric formulation and threshold values, narrowing down on optimal properties to hold constant during mAb drug discovery. As it stands today, the TAP web application has already found use in drug discovery efforts against Severe-Acute Respiratory Syndrome-related Coronavirus-2 (SARS-CoV-2) [278, 279], filial antigen proteins [280], and the Human Papilloma Virus [281].

In the final research chapter of this thesis, we had intended to couple the Public Antibody Model Libraries derived in Chapter 3 with the TAP guidelines described here to run a mock *in silico* drug discovery experiment. This would involve benchmarking of an effective tiered docking strategy that could handle the large number of structural variants in each AML and rank binders reliably enough for positive experimental validation, to be provided by Roche. However, in early 2020, a novel highly infectious betacoronavirus began to spread around the world. The disease associated with viral infection, COVID-19, quickly became associated with thousands of premature deaths and Government lockdowns imposed to regain control of the pandemic threw the global economy into turmoil. We therefore adjusted our plans, instead performing research to assist in the COVID-19 response effort.

Chapter 5

COVID-19 Research

5.1 Chapter Abstract

In early 2020, reports emerged from Wuhan, China of a novel, highly infectious betacoronavirus ('SARS n-CoV19', later 'SARS-CoV-2'). This virus soon reached pandemic proportions around the world, infecting over 27.7 million people and being associated with over one million deaths by October 2020.

Since the start of the pandemic, researchers have been working to better understand the new virus and disease, including characterising its induced B-cell response and understanding the properties of neutralising antibodies. Soon after the start of the pandemic, the University of Oxford became a global hub of coronavirus research, not least because of its role in developing the first coronavirus vaccine ('ChAdOx1-nCoV19'). The following chapter describes our work in the COVID-19 response effort, centred around the creation of a novel repository, the Coronavirus Antibody Database (CoV-AbDab).

Building off the expertise we gained in building the Therapeutic Structural Antibody Database (Chapter 2), we built CoV-AbDab to document pertinent sequence and structural information on all existing published, preprinted, or patented coronavirus binders. In this chapter, we describe our analysis of the contents of the database observing trends in developmental origins, targets, and heavy V gene family usages. We then show how CoV-AbDab can already be harnessed to better understand the functional properties of B-cell receptor (BCR) repertoire samples from COVID-19 patients. Finally, we discuss how CoV-AbDab could prove useful in the future.

This chapter contains reproduced material from the following papers:
Raybould, M.I.J., Kovaltsuk, A.K., Marks, C.M., Deane, C.M. (2020) CoV-AbDab: the Coronavirus Antibody Database *Bioinformatics*. ():btaa739. [228]

Galson, J.D., Schaetzle, S., Bashford-Rogers, R., **Raybould, M.I.J.**, *et al.* (2020) Deep sequencing of B cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures. *Front. Immunol.* doi: 10.3389/fimmu.2020.605170 [81]

5.2 The Coronavirus Antibody Database

5.2.1 Introduction

To respond effectively to the recent Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) pandemic, it is essential to understand the molecular basis for a successful immune response to coronavirus infection [282]. In particular, characterising the B-cell response is important as the identification of potent neutralising antibodies could pave the way for effective treatments, aid in prior exposure diagnosis, or assist in predicting vaccine efficacy [81, 98, 283–286]. Molecular characterisations of binding/neutralising antibodies to SARS-CoV-2 antigens began to emerge in May 2020.

Early studies suggested that a large proportion of the SARS-CoV-2 neutralising BCR response is SARS-CoV-2 specific, although some SARS-CoV-2/SARS-CoV-1 (the virus responsible for the 2003 epidemic) cross-neutralising antibodies were found to exist [287–289]. This is possible because many of the SARS-CoV-1 and SARS-CoV-2 spike protein domains, including the Receptor Binding Domain (RBD), share high sequence and structural similarity [283]. Other SARS-CoV-2 surface proteins also display high sequence similarity to more distantly related betacoronaviruses, such as the Middle East Respiratory Syndrome coronavirus (MERS-CoV). Therefore, molecular knowledge of any antibody able to bind a betacoronavirus antigen could be relevant in treating SARS-CoV-2 infection.

In addition to the large existing body of work studying SARS-CoV-1 and MERS-CoV, the number of investigations into SARS-CoV-2 is extremely high. As an indication, 1,134 SARS-CoV-2 preprints were uploaded to bioRxiv and medRxiv between January-March 2020, while 6,039 were uploaded between April and July. We therefore built the Coronavirus Antibody Database (CoV-AbDab) to collate molecular information (*i.e.* sequence and structure) and metadata on all preprinted, published, and patented anti-coronavirus antibodies. It will save valuable time in the fight against COVID-19 and act as a central hub to consolidate knowledge and coordinate efforts to identify novel antibodies that neutralise SARS-CoV-2. Researchers can use CoV-AbDab to yield new insights, including deriving crucial sequence/structural patterns

that distinguish neutralising from non-neutralising SARS-CoV-2 binders [282], or deducing independent neutralising epitopes exploitable by combination therapies [290].

5.2.2 Data Sources

Academic papers and patents containing coronavirus-binding antibodies were sourced by querying PubMed, bioRxiv, medRxiv, GenBank, and Google Patents with relevant search terms. A full list of references is provided in Section 5.2.3. ANARCI [18] was used to number amino acid sequences in the IMGT [11] numbering scheme, and to assign V and J gene origins. SAbDab [9], which tracks all antibody structures submitted to the PDB [19], was mined to identify relevant solved structures. Our antibody/nanobody homology modelling tool, ABodyBuilder [44], was used to generate full Fv region structural models.

5.2.3 Database Contents

Where possible, the following information is documented for each entry:

1. The published name of the antibody/nanobody
2. Antigens that the antibody/nanobody has been proven to bind and/or neutralise.
3. The protein domain targeted by the antibody/nanobody (e.g. spike protein receptor binding domain)
4. The developmental origin of the antibody/nanobody (e.g. engineered/ naturally raised, species information, *etc.*)
5. Sequence information including: (a) the entire variable domain sequence for the antibody/nanobody, highlighting the CDR3 regions, and (b) V and J gene germline assignments.
6. Links to any available structures involving the antibody/nanobody
7. (If Fv/VHH sequence available) A homology model of the antibody/nanobody
8. References to the primary literature on the antibody/nanobody
9. Timestamps to show when the antibody/nanobody was added and last updated
10. Any steps we are taking to follow up on the entry (e.g. to source its sequence and/or add further metadata)

As of 5$^{\text{th}}$ August 2020, CoV-AbDab contains 1,402 entries. Of these entries, 147 bind to MERS-CoV, 483 bind to SARS-CoV-1, and 1,131 bind to SARS-CoV-2 (each entry may be tested against multiple coronaviruses). All entries contain a minimum

of germline information, the CDR3 sequence, or Material Transfer Agreement contact information (a complete list of metadata is provided in the Supporting Information). Currently, 1,303/1,402 entries (92.9%) contain full variable domain antibody (Fv) or nanobody (VHH) sequences.

The CoV-AbDab entries originate from 67 papers [76, 78, 144, 287–350] and 21 patents (CN1664100, CN1903878, CN100374464, CN104447986, CN106380517, EP1857116, EP2112164, JP2018203632, KR101828794, KR101969696, KR20190122283, KR20200020411, US7396914, WO2005/012360, WO2005/054469, WO2005/060520, WO2006/095180, WO2008/035894, WO2015/179535, WO2016/138160, and WO2019/039891). Several reviews were helpful for tracing relevant literature [351–355]. Overall, the numbers of binders to each epidemic coronavirus are as follows:

Number of SARS-CoV-2 Binders: 1,131
...of which neutralising: 271 [24.0%]

Number of SARS-CoV-1 Binders: 483
...of which neutralising: 95 [19.7%]

Number of MERS-CoV Binders: 147
...of which neutralising: 95 [64.6%]

As of 5th August 2020, CoV-AbDab linked to 84 relevant CoV-antibody structures across 40 distinct antibodies/nanobodies, with structures of a further 10 antibodies/nanobodies anticipated based on preprints. These solved structures indicate that many coronavirus binding antibodies use both their heavy and light chain complementarity-determining regions to engage the RBD (Fig. 5.1), highlighting the importance of capturing full Fv information [254, 279, 287, 329, 334, 343, 356, 357].

Our database does not contain immunoglobulin gene sequencing (Ig-seq) samples of SARS-CoV-2-responding BCR repertoires [81, 98, 289]; case studies from such studies will only been included if binding/neutralising activity has been confirmed experimentally. We are making such Ig-seq data available from our Observed Antibody Space (OAS) database [88].

We have reached out to authors of new studies characterising coronavirus binding antibodies to send us their data in Excel or CSV format (*opig@stats.ox.ac.uk*). We require a minimum of the antibody/nanobody clonotype (closest heavy and light V and J gene transcripts plus the CDR3 amino acid sequences) but ideally seek the full

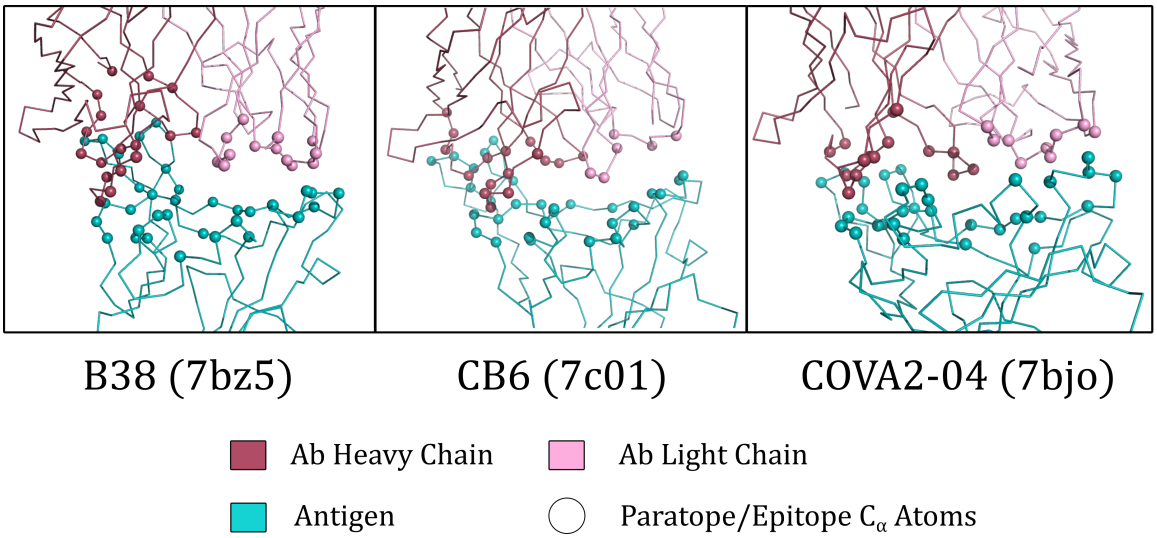


Figure 5.1: Pymol representations of the binding interfaces of three SARS-CoV-2 neutralising antibodies (B38 [343], CB6 [329], COVA2-04 [254]) that bind to the Receptor Binding Domain. Paratope residues are defined as those with a heavy atom within 4.5Å of an antigen heavy atom, while epitope residues are defined as those with a heavy atom within 4.5Å of an antibody heavy atom. C_α atoms of paratope and epitope residues are marked with spheres. Raspberry: Heavy chain, Pink: Light Chain, Teal: Antigen.

antibody variable domain [Fv] or full nanobody [VHH] sequence. We also require binding data against at least one specified coronavirus protein reported in a preprint, publication, or patent. Through these submissions and our own efforts to track the scientific literature, we hope to provide a central community resource for coronavirus antibody sequence and structural information.

5.2.4 Database Analysis

In this section, we describe a preliminary analysis of the contents of CoV-AbDab.

5.2.4.1 Biological/Synthetic Origins

First we investigated the reported biological/synthetic origins of each known binder to SARS-CoV-1, MERS-CoV, and SARS-CoV-2 (Fig. 5.2).

An immediate observation was the increase in the role of nanobodies being used to target MERS-CoV and SARS-CoV-2 relative to SARS-CoV-1. Nanobodies used as MERS-CoV therapies tend to have natural sources (e.g. infected camels, reflecting the geographical region of the epidemic), whereas the advance in 'sybody' (synthetic

nanobody) technology has led to phage display becoming the primary origin of anti-SARS-CoV-2 nanobodies.

The origins of antibody coronavirus binders have also significantly changed over time, with a clear move away from harvesting animal (mouse, chicken, rhesus) antibody response repertoires. Coupled with this, there has been a decrease in the use of phage display to isolate human antibody binders, with antigen baiting technologies becoming dominant. Human B-cells are the biological origin of 96% of the SARS-CoV-2 antibodies, compared to 42% of SARS-CoV-1 antibodies (excluding cross-reactive SARS-CoV-1/SARS-CoV-2 antibodies isolated in 2020, this number falls to just 23%).

Together these results indicate that there have been discernible shifts in the methods used to isolate binding antibodies/nanobodies since 2003. These trends should be carefully considered when directly comparing the nature of published binders to different coronaviruses.

5.2.4.2 Target Antigens

We next investigated the protein target for each documented antibody/nanobody coronavirus binder, to the domain resolution if possible (Fig. 5.3). Note that many of these categories are not theoretically mutually exclusive (e.g. an S; N-Terminal Domain [NTD] binder would also fit into the category of S; S1 non-Receptor Binding Domain [RBD], however there is experimental evidence to assign it specifically to the NTD domain so we have classified it as such).

For all three coronaviruses, the RBD is the target for most binding antibodies and nanobodies. It is unsurprising that the RBD is the most investigated domain, as a well-understood mode for viral neutralisation involves direct competition for the native human receptor (Angiotensin-converting Enzyme-2 (ACE-2) for SARS-CoV, or Dipeptidyl Peptidase-4 (DPP4) for MERS-CoV). A future bias towards disproportionately identifying RBD binders could also result from the use of RBD baits in addition to entire spike protein baits to isolate potential prophylactic antibodies from human serum.

5.2.4.3 Heavy V-Gene Germline Origins

We then compared the heavy V-gene (IGHV) germline origins of human antibody binders to all targets of SARS-CoV-2, SARS-CoV-1, and MERS-CoV (Table 5.1).

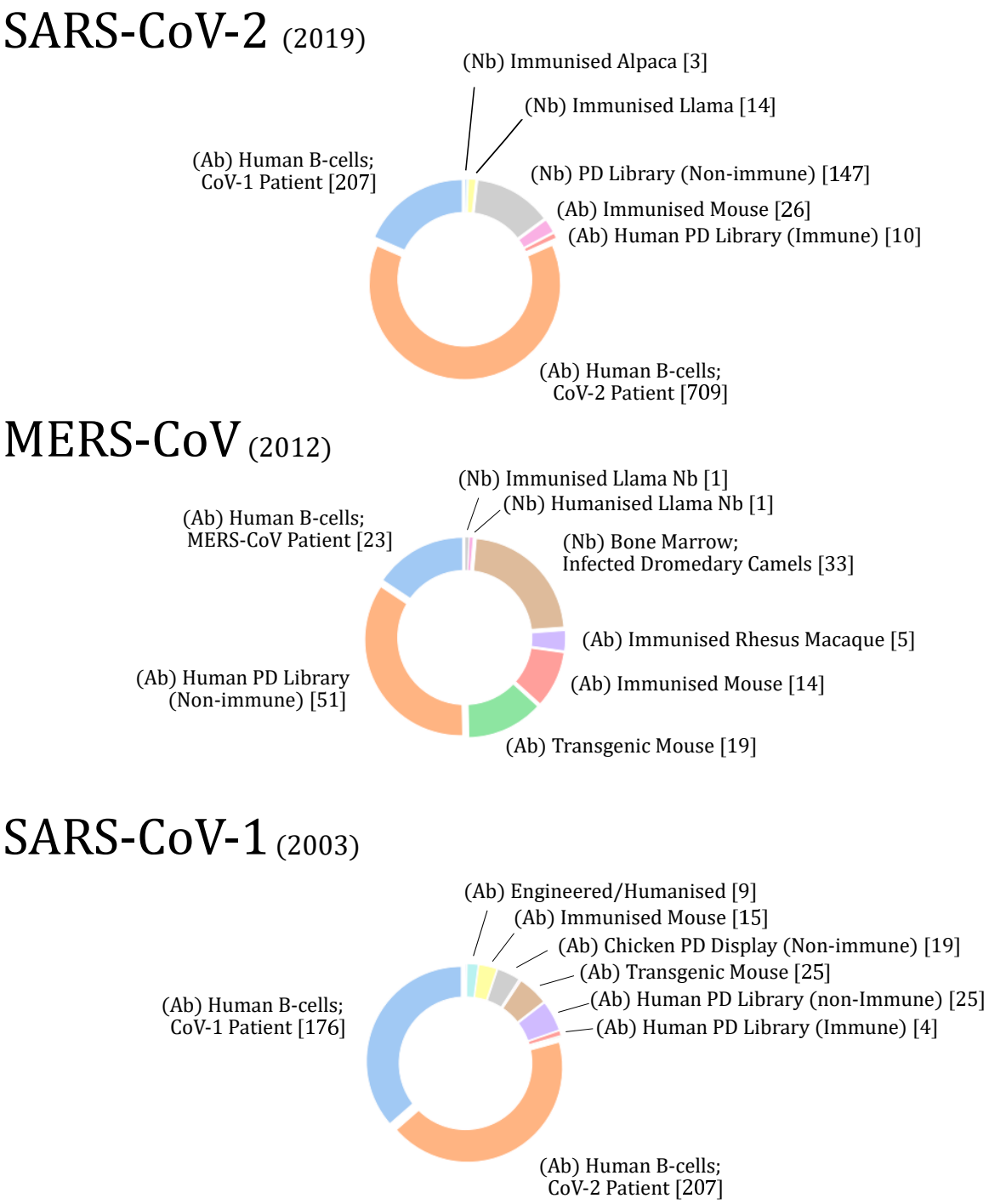


Figure 5.2: Donut plots comparing the biological/synthetic origin of documented binders to SARS-CoV-2 (2019 pandemic), MERS-CoV (2012 epidemic), and SARS-CoV-1 (2003 epidemic). Absolute numbers of binders in brackets (as of 5th August 2020). Ab: Antibody; Nb: Nanobody; PD: Phage Display.

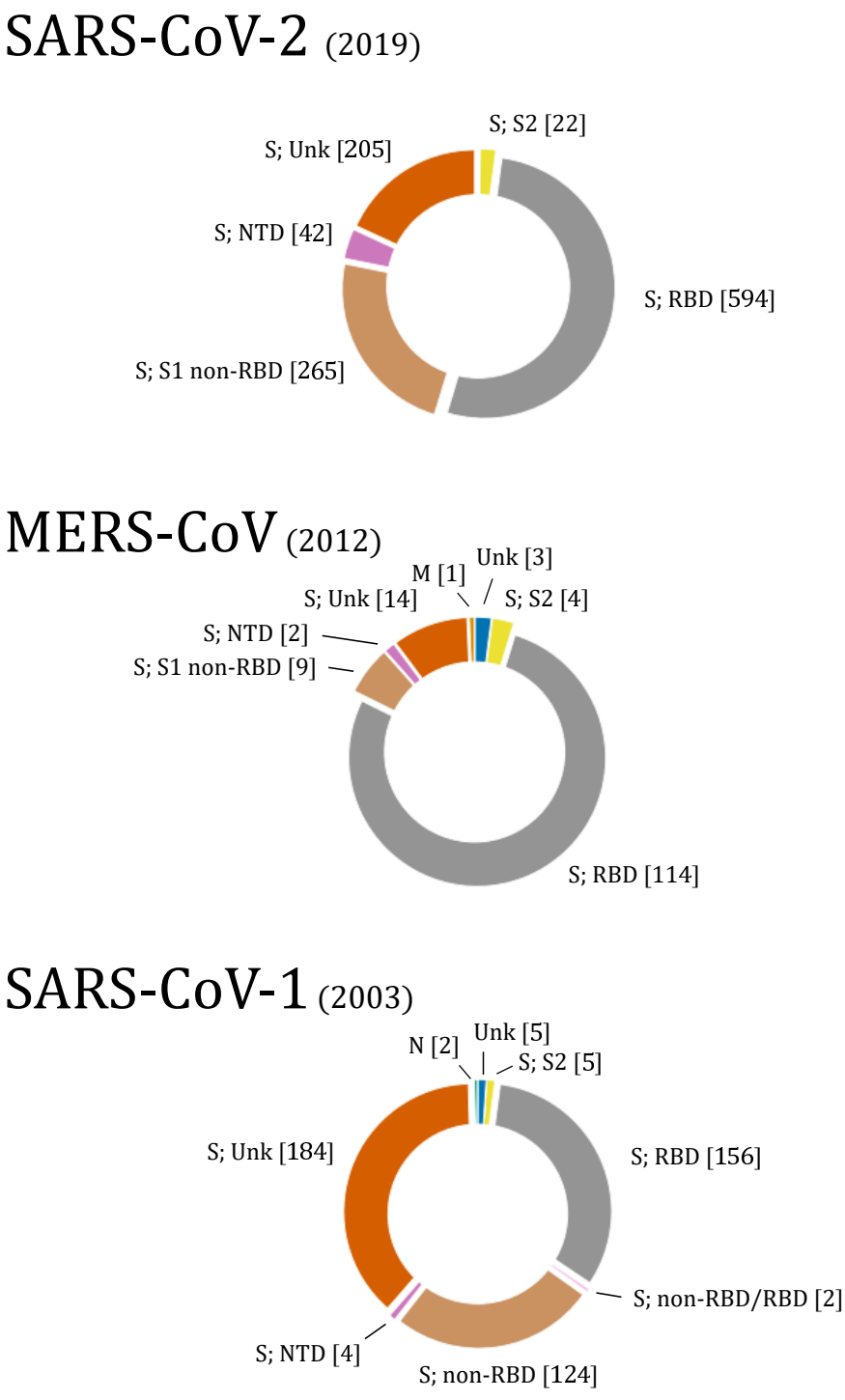


Figure 5.3: Donut plots comparing the protein epitopes of documented binders to SARS-CoV-2 (2019 pandemic), MERS-CoV (2012 epidemic), and SARS-CoV-1 (2003 epidemic). Absolute numbers of binders in brackets (as of 5th August 2020). M = Membrane Protein; N = Nucleocapsid Protein; NTD = N-Terminal Domain; RBD = Receptor Binding Domain; S = Spike protein; Unk = Unknown.

A

IGHV Gene	Number of SARS-CoV-2 Binders [Percentage]
3-30	118 [12.9%]
1-69	115 [12.5%]
3-53	55 [6.0%]
1-2	35 [3.8%]
3-23, 3-66, 4-39	34 [3.7%]
3-9, 3-30-3	32 [3.5%]
1-46, 5-51	29 [3.2%]
4-4, 4-34	24 [2.6%]
3-7, 3-33	23 [2.5%]
1-18	22 [2.4%]
4-59	20 [2.2%]
3-48, 7-4-1	19 [2.1%]
1-8, 3-21	17 [1.9%]
1-24, 2-5	15 [1.6%]
1-58	14 [1.5%]
3-13	13 [1.4%]
3-11, 3-64(D), 4-31, 4-61	12 [1.3%]
3-20	9 [1.0%]
3-49	8 [0.9%]
2-70, 3-15	7 [0.8%]
2-26	5 [0.5%]
3-74, 4-38-2	4 [0.4%]
1-3, 4-30-4	3 [0.3%]
3-43, 5-10-1	2 [0.2%]
3-23, 3-72, 4-38	1 [0.1%]

B

IGHV Gene	Number of SARS-CoV-1 Binders [Percentage]
3-30	77 [16.7%]
1-69	75 [16.3%]
3-23	38 [8.2%]
1-18	19 [4.1%]
1-2, 3-30-3	18 [3.9%]
3-7	17 [3.7%]
3-33, 5-51	12 [2.6%]
1-46, 3-9, 3-11, 4-4	11 [2.4%]
4-39	10 [2.2%]
4-34, 4-59	9 [2.0%]
2-5, 4-61	8 [1.7%]
3-13, 3-48	7 [1.5%]
1-8, 3-20, 3-21, 3-53, 7-4-1	6 [1.3%]
3-49, 3-66, 4-31	5 [1.1%]
3-64(D), 3-72	4 [0.9%]
1-24, 4-38-2	3 [0.7%]
1-58, 2-26, 3-43, 3-64, 6-1	2 [0.4%]
1-3, 3-15, 3-23, 3-74	1 [0.2%]

C

IGHV Gene	Number of MERS-CoV Binders [Percentage]
1-69	37 [39.4%]
3-23	20 [21.3%]
4-39	6 [6.4%]
2-5	5 [5.3%]
3-30	4 [4.3%]
4-59	3 [3.2%]
1-3, 1-18, 3-11, 3-15, 3-21	2 [2.1%]
1-2, 1-46, 3-9, 3-30-3, 4-4, 4-34, 6-1	1 [1.1%]

Table 5.1: Comparing the human antibody heavy chain V-gene origin of documented binders to (A) SARS-CoV-2, (B) SARS-CoV-1, and (C) MERS-CoV. Numbers/percentages correct as of 5$^{\text{th}}$ August 2020.

Certain germlines were expanded in both SARS-CoV-2 and SARS-CoV-1 human antibody binders, for example the IGHV3-30 gene is represented in 13% of SARS-CoV-2 human antibodies over at least 7 studies [254, 287, 289, 290, 295, 327, 328] and 17% of SARS-CoV-1 human antibodies over at least 10 studies ([144, 254, 287, 289, 298, 306, 311, 328, 330] and patent CN1903878), but only in 4% of MERS-CoV human antibodies over at least 4 studies ([307, 331, 346] and patent WO2105179535).

In contrast, the IGHV3-53 gene represented 6% of SARS-CoV-2 human antibody binders, yet appeared in just 1.3% SARS-CoV-1 human antibody binders and no IGHV3-53 MERS-CoV binder has yet been reported. The SARS-CoV-2 IGHV3-53 epitopes have been explored structurally by Yuan *et al.* [358] and Wu *et al.* [357].

The IGHV1-69 gene was notably under-represented in reported SARS-CoV-2 binders until the study by Wec *et al.* [341], in which SARS-CoV-2 spike protein baits were used to fish cross-reactive binders from an uninfected human who caught and recovered from SARS-CoV-1 in 2003. The researchers found the VH1-69/VK2-30 germline gene pair to be highly over-represented (used by around 30% of all detected SARS-CoV-2 reactive antibodies) and displayed high levels of somatic hypermutation, implying substantial affinity maturation. That the IGHV1-69 gene family was associated with SARS-coronavirus cross-reactivity is striking evidence towards this heavy germline gene providing stable framework for long-CDRH3 broadly-neutralising antibodies, as has previously been reported against influenza hemaggluttin. Wec *et al.* are carrying forward the most promising candidates for clinical trials to investigate whether these SARS-CoV-1-raised antibodies can confer protection against SARS-CoV-2 infection. The results of these experiments will be profoundly useful for those investigating the plausibility of the 'herd immunity' public health strategy and, if found to be protective, will likely influence future vaccine design.

5.2.5 Web Application

CoV-AbDab is located at *http://opig.stats.ox.ac.uk/webapps/coronavirus* (Fig. 5.4A). Users can download the entire database or search-filtered results as a CSV file and bulk download all ANARCI numberings, IMGT-numbered antibody- or nanobody-CoV structure files, and IMGT-numbered antibody/nanobody homology models (Fig. 5.4B). A summary of tracked (but not included) antibodies and coronavirus studies is also provided.

The database can be queried by variable domain sequence (Fig. 5.4C) or by attribute (Fig. 5.4D). Attribute search results can be further filtered by a search term and ordered by any metadata field for maximum interpretability (Fig. 5.4E).

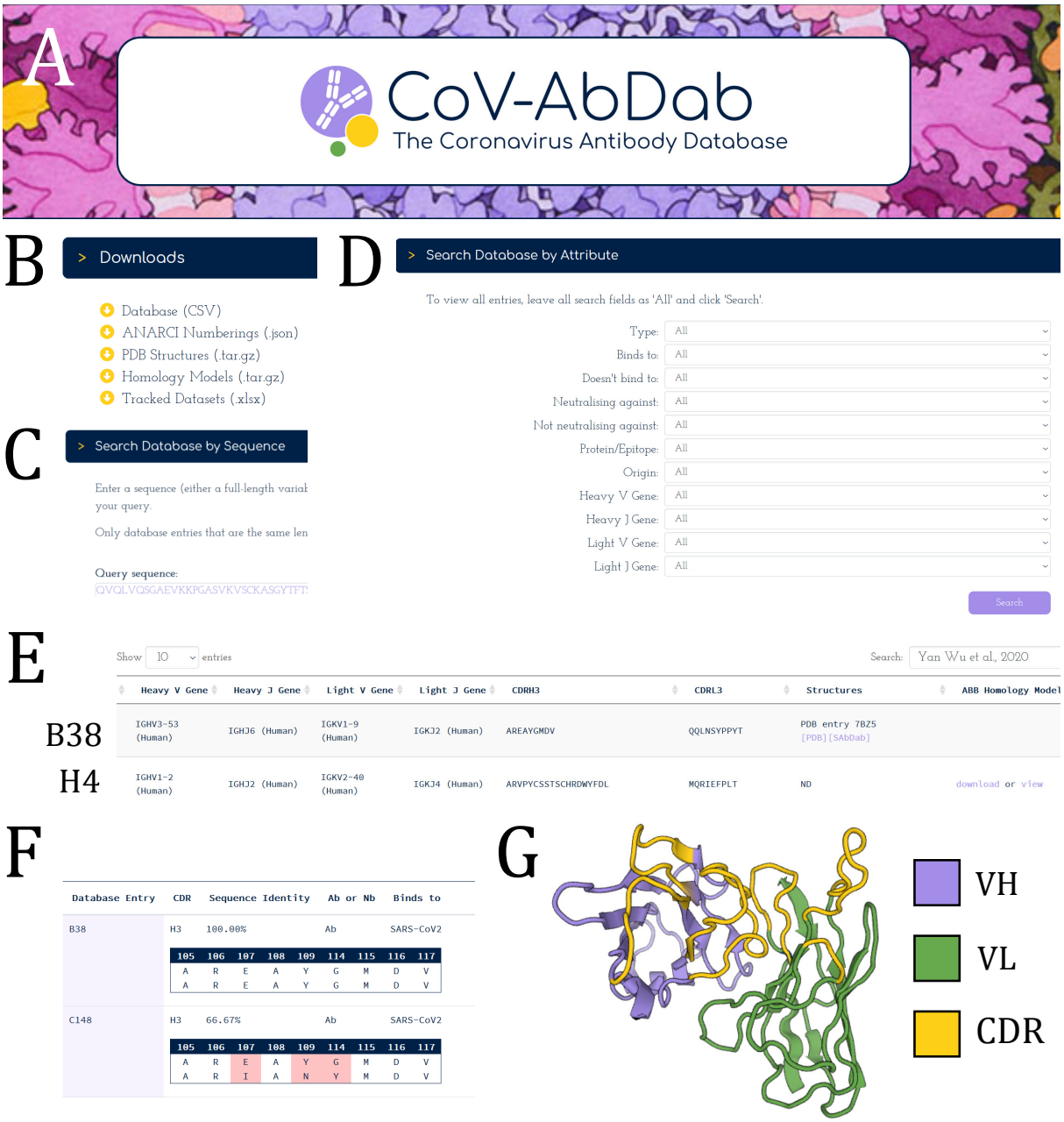


Figure 5.4: The CoV-AbDab website (*http://opig.stats.ox.ac.uk/webapps/coronavirus*). (A) CoV-AbDab homepage logo (background image credit: David Goodsell). (B) All CoV-AbDab data can be downloaded. (C) The database can be queried by sequence (full chain or CDRH3). (D) The database can be queried by attribute (neutralisation profile, developmental origin, germlines, *etc.*). (E) All results filtered by a particular study [343]. Shown is the information required for clonotyping, a solved structure for B38, and a homology model for H4. (F) The result of using the sequence search feature from panel 'C' with the CDRH3 from B38. Alignments and metadata are given for the top ten closest matches. (G) The in-browser viewer displaying the homology model of H4.

Searching by sequence returns the top-10 same-length whole chain and/or CDR3 sequence identities to the query. Query-target alignments are displayed (Fig. 5.4F). Any entry with a homology model can be viewed in-browser using our native molecular viewer (Fig. 5.4G).

5.3 B-Cell Receptor Repertoire Profiling

The following investigations are examples of how CoV-AbDab can be used to add putative functional annotations to BCR repertoire sequences and derived structures. The first study analyses UK COVID-19 patients to derive B-cell clones found to be uniquely convergent in response to SARS-CoV-2 (*i.e.* absent in a set of healthy and influenza-response control repertoires). The second study compares all CoV-AbDab probes to over 1 billion human VH chains from the OAS database [88], comprising a further five SARS-CoV-2 studies (not including our UK dataset) and over 50 healthy/unrelated disease studies from before the start of the COVID-19 pandemic. Finally, we show how the solved antibody-antigen structures in CoV-AbDab can be used to 'fish' for similar topologies in the 'Public Baseline' Antibody Model Library generated in Chapter 3.

5.3.1 Methods

Comparing Convergent Clones from SARS-CoV-2 Repertoires to CoV-AbDab

A series of experimental and computational protocols performed by our collaborators at Barts Healthy NHS Trust, Illumina, and Alchemab led to a total of 1,254 SARS-CoV-2 convergent clones that were unobserved in healthy or influenza control groups (full details in the Appendix). The CDRH3 sequences from the convergent clones were clustered alongside all the non-redundant CDRH3 sequences in CoV-AbDab [228] (5^{th} August version) using the cd-hit-2d package from CD-HIT [247], selecting an 80% sequence identity threshold and allowing at most a CDRH3 length mismatch of 1 amino acid (indels penalised as a mismatch). Cluster centres containing at least one CoV-AbDab CDRH3 and one convergent clone CDRH3 were further investigated for clonotype matches (*i.e.* where the probe sequence and the convergent SARS-CoV-2 clonotype share the same V and J gene origins). Due to their extremely similar amino acid sequence identity, the following V gene pairs were permitted within the same clone: IGHV3-30/IGHV3-30-3 and IGHV3-53/IGHV3-66.

Proximity of CoV-AbDab Clonotypes to SARS-CoV-2 and pre-COVID-19 BCR Repertoires

The 905 non-redundant human antibody CDRH3 sequences in CoV-AbDab (at the level of 100% sequence identity, as of 9th September 2020) were pooled as a set of reference probes (their corresponding V and J genes were also recorded). Separately, the OAS database was mined for the ∼1 billion VH sequences belonging to all human antibody Ig-seq studies from before November 2019 as well as five SARS-CoV-2 Ig-seq studies conducted since the start of the pandemic [98, 309, 359–361]. We assume that the vast majority of the pre-November 2019 sampled population is serologically naïve to both SARS-CoV-1 and SARS-CoV-2, given the estimated start date of the pandemic and the relatively low case numbers of SARS-CoV-1 recorded in 2003. We ensured no CoV-AbDab probes derived directly from the assessed Ig-seq studies to avoid trivial hits.

For each CoV-AbDab probe, we searched for the set of human sequences assigned to the same closest V and J genes, and thereafter recorded the maximum CDRH3 sequence identity seen within these common VJ pairings. Two different clonotype definitions were investigated: 'Briney V3J' clonotype matches require the same V and J gene identity and 100% CDRH3 sequence identity (as per Briney *et al.* [4]), while 'Soto V3J' clonotype matches require the same V and J gene identity and to be within 80% CDRH3 sequence identity (as in Soto *et al.* [80]).

Using a SARS-CoV-2 Complex Structure to Interrogate the Public Baseline Antibody Model Library

We downloaded the IMGT-numbered [11] solved structure for the SARS-CoV-2 RBD-binding antibody, B38 [357], from CoV-AbDab (original PDB code '7bz5'). Separately, we filtered the 'Public Baseline' AML for the subset of structures with the same combination of six North-defined [15] CDR lengths as B38. The SVDSuperimposer package from the Biopython suite [362] was then used to align the backbone atoms of each model to the B38 crystal structure. This package finds the ideal rotation/translation matrices to align two sets of points together, such that the RMSD between the two structures is minimised. The high structural homology in the Fv framework region ensures SVDsuperimposer generates reliable, consistent alignments between Fvs and provides a good means for judging structural similarity between antibody CDR regions. The Fv root-mean-squared deviation (RMSD) between each AML model and B38 was calculated using the same in-house script as used in the comparison to therapeutic structures (see Section 3.3). Repertoire Structural Profiling

and AML homology modelling (Chapter 3) were performed using FREAD databases compiled in February 2019, and so neither tool had access to the framework/CDR templates from the SARS-CoV-2 binding antibodies isolated during the pandemic.

On finding the closest model to B38 (H32304+L112151), the 105 Fv sequences predicted to lie on this model structure were transplanted onto the model backbone and the PEARS software (see Seciton 1.3.4.4) was used to model-in all mismatching residue side chains.

As a measure of paratope similarity to B38, we used the Ab-Ligity tool, recently developed in our group. Ab-Ligity converts antibody binding sites into hash table representations that capture both structural (*via* relative Cartesian coordinates) and chemical (by grouping residues based on interaction characteristics) features of paratopes. Here, paratope residues were not algorithmically predicted, instead the coordinates of the solved 7bz5 complex were used to mark up the set of antibody residues within 4.5Å of the cognate antigen. Through comparison of SAbDab paratopes that bind to the same/highly similar epitopes, the authors determined a threshold Ab-Ligity score (Tversky index) of ≥ 0.1 as having optimal precision-recall in distinguishing paratopes that are likely to bind the same epitope from those that are not.

Ab-Ligity was run on each of the 105 side chain-remodelled derivatives to create a hash table for each model paratope across the set of actual paratope positions. The hash table for each of these variants was compared to the hash table for the true B38 complex, and the predefined Ab-Ligity threshold score of 0.1 was used to distinguish paratopes with similar binding characteristics from those that are dissimilar.

5.3.2 Results

5.3.2.1 Comparing Convergent Clones from SARS-CoV-2 Repertoires to CoV-AbDab

Thirty-one clusters involving a total of 50 CoV-AbDab reference CDRH3 probes and 39 SARS-CoV-2 convergent clonotypes were recorded. Twelve (38.7%) of these clusters contained a CoV-AbDab probe VH and a SARS-CoV-2 convergent clone VH that align closest to the same V and J gene; these are recorded in Table 5.2 with probe sequences spanning 7 independent SARS-CoV-2 studies [254, 287, 289, 290, 327, 328, 350].

The metadata collated in CoV-AbDab for each of these reference hits can then be used for further stratification of consensus clonotypes. For example, three of the matched CoV-AbDab probes (COV2-2790, REGN10977, and COVA2-04) have been

Name	CDRH3	IGHV	IGHJ	Binds	Doesn't Bind	Neutralises	Doesn't Neutralise	Epitope	Source
C154	AKQAGPYCSGGSCYSAPFDY	3-30	4	SC1, SC2		SC2 (weak)		S1, RBD	[289]
ALC_3983948	**AKVSGPYCSGGSCYSFYFDY**	**3-30**	**4**						
COV2-2790	ARSYDILTGYRDAFDI	3-53	3	SC2	SC1	SC2	SC1	S1, RBD	[350]
ALC_2318471	**VRNYDILTGYSDAFDI**	**3-53**	**3**						
COV2-2007	AKVSATYYYYYYGMDV	3-30	6	SC1, SC2			SC2	S1, RBD	[350]
CC12.17	AKSSGSYYYYYYGMDV	3-30	6	SC2	SC1	SC2 (weak)	SC1	S1, RBD	[326]
ALC_2318830	**AKVMTTYYYYYYGMDV**	**3-30**	**6**						
REGN10977	ARTPFYYDSSGYYLDY	1-69	4	SC2		SC2		S1, RBD	[289]
COVA2-14	ARVR-YYDSSGYYEDY	1-69	4	SC1, SC2			SC1,SC2	S1, RBD	[253]
ALC_1780442	**ARYD-YYDSSGYYLDY**	**1-69**	**4**						
COV2-2270	AITYYYDSSGYWWDD	1-69	4	SC1, SC2			SC2	S1, not RBD	[350]
ALC_1781971	**ASTYYYDSSGYWFDY**	**1-69**	**4**						
COVA2-40	AGRYCSGGRCGWFDP	4-4	5	SC2	SC1		SC1, SC2	S1, not RBD	[253]
ALC_1784026	**ESRYCSGGSCGWFDP**	**4-4**	**5**						
COV2-2147	ARSTSGSYYYGMDV	3-30-3	6	SC1, SC2			SC2	S1, not RBD	[350]
CV34	ARSYGGSYYYGMDV	3-30-3	6	SC2	SC1		SC1,SC2	S1, not RBD	[327]
ALC_1249094	**ARGTRGSYYYGMDV**	**3-30-3**	**6**						
COV2-2639	ARAGGGSYRGPFDY	3-30	4	SC1, SC2			SC2	S2	[350]
2M-14E5	ARSGGGSYRGPFDY	3-30	4	SC2			SC2	S2	
ALC_1245591	**ARVIGGSYRGPFDY**	**3-30-3**	**4**						
COV2-2006	ARPQSGGYYAPLDY	3-30-3	4	SC1, SC2			SC2	S1, not RBD	[350]
ALC_1246650	**ARPYSGSYYAPLDY**	**3-30-3**	**4**						
S304	ARGDSSGYYYYFDY	3-13	4	SC1, SC2		SC1, SC2 (weak)		S; RBD	[286]
ALC_1245048	**ARGYSSGYYYYFDY**	**3-13**	**4**						
COV2-2027	AIYGYYYGLDV	3-30	6	SC2	SC1		SC1, SC2	S1; not RBD	[350]
ALC_480504	**AVYGYYYGMDV**	**3-30**	**6**						
ALC_481016	**ASYGYYYGMDV**	**3-30-3**	**6**						
COVA2-04	ARDLERAGGMDV	3-53	6	SC2	SC1	SC2	SC1	S; RBD	[253]
ALC_498298	**ARDLEAAGGMDV**	**3-66**	**6**						

Table 5.2: The twelve clusters where a CoV-AbDab reference VH (blue font) aligns to the same V and J gene and is within 80% sequence identity CDRH3 of an Alchemab convergent SARS-CoV-2 clone (black font). CDRH3 amino acids are coloured green for matches and red for mismatches, according to the reference with the highest CDRH3 sequence identity match. CDRH3 lengths were permitted to deviate by a single amino acid; insertions were scored as mismatches. Owing to their close sequence identity, IGHV3-30 and IGHV3-30-3, and IGHV3-53 and IGHV3-66, were considered gene matches. RBD: Receptor-binding domain, S: Spike (protein), SC1: SARS-CoV-1, SC2: SARS-CoV-2.

shown to strongly neutralise SARS-CoV-2 pseudovirus, and so their corresponding convergent clonotypes (ALC_2318471, ALC_1780442, and ALC_498298) or combinations thereof, might be chosen for further therapeutic development against SARS-CoV-2. Alternatively, if more broadly neutralising candidates are required, the convergent clones that map to CoV-AbDab references that bind and/or neutralise both SARS-CoV-1 and SARS-CoV-2 could be investigated. Clinical prognosis data could also be combined with the mapped CoV-AbDab probe neutralisation profiles to identify which of the convergent response antibodies may lead to the best outcomes at a phenotypic level.

Matches to CoV-AbDab can also be used to learn more about the immune response to SARS-CoV-2 infection. For example, five of the twelve convergent clones (42%) map to 'S1, not Receptor Binding Domain (RBD)' binders, despite binders to this epitope only constituting ∼23% of CoV-AbDab probes, and one cluster mapped to S2, despite the fact that only 22/1128 (2%) of the probes in CoV-AbDab bind this region of the Spike protein. This indicates that natural infection can lead to a convergent response against many different epitopes on the spike protein besides the RBD. More studies into binders to other SARS-CoV-2 surface proteins such as the Nucleocapsid and Envelope proteins are required to detect whether they too elicit a convergent response.

5.3.2.2 Proximity of CoV-AbDab clonotypes to SARS-CoV-2 and pre-COVID-19 BCR repertoires

In this work, we compared the CoV-AbDab reference sequences with human Ig-seq datasets recorded in the Observed Antibody Space (OAS) database [88] to profile which of the SARS-CoV-2 binding clonotypes are most diagnostic of a SARS-CoV-2 response. We first plotted the maximal CDRH3 sequence identity recorded between a CoV-AbDab probe sequence and the OAS datasets, conditioning on matches having the same V and J gene origins (Fig. 5.5; mean 76.9%, max 100%, min 30.4%). We observed 39/930 (4.2%) coronavirus binding antibodies with a 'Briney V3J' clonotype match to OAS, while 426/930 (45.8%) shared a 'Soto V3J' clonotype (see Section 5.3.1 for definitions). Conversely, 92 CoV-AbDab VJ matches could only reach peak CDRH3 sequence identities between 30%-60% when compared to the OAS VHs. These coronavirus binders are distal in identity both to SARS-CoV-2 and non-SARS-CoV-2 antibody repertoires, suggesting a significant proportion of the SARS-CoV-2 binding antibodies in CoV-AbDab (which were largely baited from convalescent human serum) represent part of a highly 'private' (bespoke to a single individual)

CoV-AbDab CDRH3 Sequence [Blue: SARS-CoV-2 Binder]	Epitope	CDRH3 Matches (Studies)	SC2 Rep. Matches (Studies)	Isotypes [B: Bulk]
ARDGYGSGSDYYYYYYMDV	S RBD	1 (1)	0%	B
ARDSVAGIYYYYGMDV	S n-RBD	1 (1)	0%	M
AAPYCSRTSCHDAFDI	S RBD	1 (1)	100% (1)	G
ARVNSGSYYSYFDY	S n-RBD	1 (1)	0%	M
ARRGDGLYYYGMDV	S Unk	1 (1)	0%	M
ARGVVAATPGWFDP	S n-RBD	1 (1)	0%	M
ARGPAATYYYYMDV	S n-RBD	2 (1)	100% (1)	M
AREDYYDSSGSFDY	S RBD	6 (1)	0%	M
ARAQGGNYYYGMDV	S n-RBD	250 (1)	100% (1)	G
AKDGSGSYYGWFDP	S n-RBD	2 (2)	0%	M
ARVGGYYYYYMDV	S RBD	14 (4)	0%	D,M,B
ASSSWLRGAFDI	S RBD	1 (1)	100% (1)	G
ARVGSSSWYFDY	S n-RBD	573 (7)	98.3% (4)	A,G,M,B
ARLGGSSWHFDY	S n-RBD	117 (1)	100% (1)	G
ARDLDVSGGMDV	S RBD	135 (1)	100% (1)	G
ARDGELLGWFDP	S RBD	3 (1)	0%	M
AKEIAVAGCFDY	S n-RBD	1 (1)	0%	B
AKATTVTTYFDY	S Unk	7 (3)	14.3% (1)	M
ARSGDSDAFDI	S RBD	1 (1)	0%	M
ARDYGDYYFDY	S RBD	13 (7)	38.5% (2)	D,G,M,B
ARDYGDLYFDY	S RBD	3 (3)	66.7% (2)	G,M
ARDYGDFYFDY	S RBD	75 (4)	96% (3)	A,G,M,B
ARDRVVYGMDV	S RBD	6 (3)	0%	D,M,B
ARDLYYYGMDV	S RBD	13 (4)	7.7% (1)	A,M,B
ARDLVVYGMDV	S RBD	26 (5)	84.6% (2)	A,G,M,B
ARDLSEGGMDV	S RBD	2 (1)	0%	M
ARDLGPYGMDV	S RBD	12 (4)	83.3% (2)	G,M
ARDFGEFYFDY	S RBD	1 (1)	100% (1)	A
VRDTDWAFDS	S Unk	3 (2)	0%	A,G
ARDRSYYLDY	S RBD	2 (1)	0%	M
ASGPNYFDY	S Unk	6 (2)	0%	A,G
ARRDTDFDY	S n-RBD	2 (1)	0%	M
ARGAASFDY	S n-RBD	1 (1)	0%	M
AREAYGMDV	S RBD	8 (4)	12.5% (1)	G,M
ARDLPPLDY	S Unk	988 (9)	2.3% (1)	A,D,G,M,B
ARGWYFDY	S NTD	27 (5)	3.7% (1)	G,M
AKPYGMDV	S n-RBD	110 (11)	7.3% (2)	A,D,G,M,B
ARGFDY	S n-RBD	20283 (19)	98.0% (5)	All
AGNDY	S RBD	6 (4)	0%	M,B

Table 5.3: The CDRH3s and metadata of all 'Briney V3J' clonotypes shared between CoV-AbDab and OAS repertoires [88]. NTD = N-Terminal Domain, (n-)RBD = (not) Receptor-Binding Domain, S = Spike, SC2 Rep. = SARS-CoV-2 Repertoire.

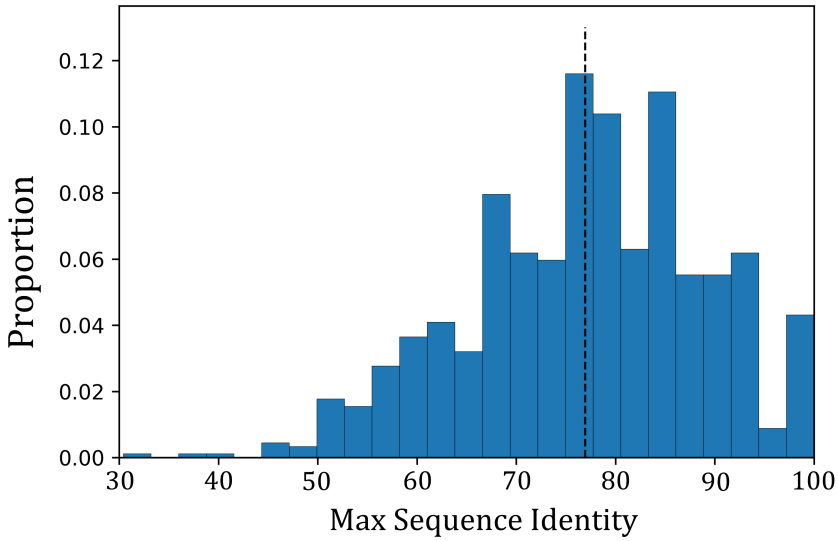


Figure 5.5: The highest CDRH3 sequence identity recorded between a CoV-AbDab VH and the selected set of OAS VHs, conditioning on both sequences having the same V and J gene origins. The dashed black line indicates a mean value of 76.9% maximum sequence identity.

antigen response. Their rarity in the wider population could make them extremely diagnostic of SARS-CoV-2 exposure, particularly considering the sampling depth of pre-COVID19 repertoires (both healthy and diseased) in this investigation.

We then recorded the OAS metadata for the 39 datasets that shared a 'Briney V3J' clonotype with CoV-AbDab, presented in Table 5.3. The matched CDRH3s spanned many lengths (5 through to 19). Several closest CDRH3 sequence identity clonotype matches occurred to pre-COVID-19 BCR repertoires rather than SARS-CoV-2 BCR repertoires, showing the presence of a CoV-AbDab clonotype match in itself is not necessarily diagnostic of a SARS-CoV-2 response. This is particularly evident for shorter CDRH3 loops. However, some clonotype matches do appear to be more characteristic of SARS-CoV-2 repertoires — seven clonotypes appeared only in SARS-CoV-2 repertoires and in no others. In six out of these seven cases, the OAS metadata shows evidence of maturation through class switching (five to IgG, one to IgA), which may be diagnostically relevant. Another observation is that several of these matches occurred many times but only to a single SARS-CoV-2 study, which suggests that protocol-specific biases may be expected to detect particular clonal matches and miss others. Finally, the number of pre-COVID IgM repertoire clonotype matches to CoV-AbDab is consistent with most humans being able to initiate an antibody immune

response to several SARS-CoV-2 antigens, a finding which is in accordance with the growing number of serum characterisation studies [98, 310, 358, 359, 363].

5.3.2.3 Using a SARS-CoV-2 Complex Structure to Interrogate the Public Baseline Antibody Model Library

One such study by Yuan *et al.* [358] identified a structural basis behind a public set of IGHV3-53 and IGHV3-66 antibodies with potential to neutralise SARS-CoV-2 by binding to the same RBD epitope. Here, we interrogate our 'Public Baseline' Antibody Model Library (AML, see Chapter 3) with a solved IGHV3-53-derived RBD binder ('B38' [357]) to see whether this complementary geometry is represented across the naïve repertoires of all ten healthy individuals, and if so, to compare the associated baseline paratopes to the proven SARS-CoV-2 binder.

The closest 'Public Baseline' AML geometry to the B38 solved structure had an Fv RMSD of 1.07Å; an aligned structure is shown in Fig. 5.6A. This distance is comparable to the mean closest geometry between the AML to solved therapeutic structures (though if ABodyBuilder had been able to choose FREAD templates from solved SARS-CoV-2 structures — as it was able to select fragments from most solved therapeutic structures, so long as the resolution was $\leq$ 2.5Å — an even closer AML binding site geometry might have been found).

Tracing back to find the Fv sequences predicted to adopt this structure across ten individuals, we found 105 unique sequence pairings. These various VH and VL sequence side-chains were then modelled onto the H32304+L112151 model backbone and Ab-Ligity was run on each resulting structure to predict the number of shared paratopes (see Section 5.3.1). A total of 86/105 (81.9%) models recorded an Ab-Ligity score of $\geq$ 0.1, indicating that they may have enough chemical and structural similarity to bind to the same epitope as B38 (Fig. 5.6B). The top 18 Ab-Ligity scores (0.268-0.459) belong to Fvs with an IGHV3-53 or IGHV3-66 heavy V gene family origin, consistent with B38. However, several Fvs with other heavy germline V genes yielded maximum Ab-Ligity scores significantly above the 0.1 threshold, including IGHV3-13 (0.260), IGHV4-59 (0.257), IGHV4-34 (0.252), and IGHV4-4 (0.252). These germlines may represent alternative evolutionary maturation pathways to engage the same RBD epitope as B38. From an industrial perspective, they may also present opportunities for 'germline hopping' to expand lead compound diversity, though their persistently lower Ab-Ligity scores suggest they might require targeted mutations to improve affinity to the desired level.

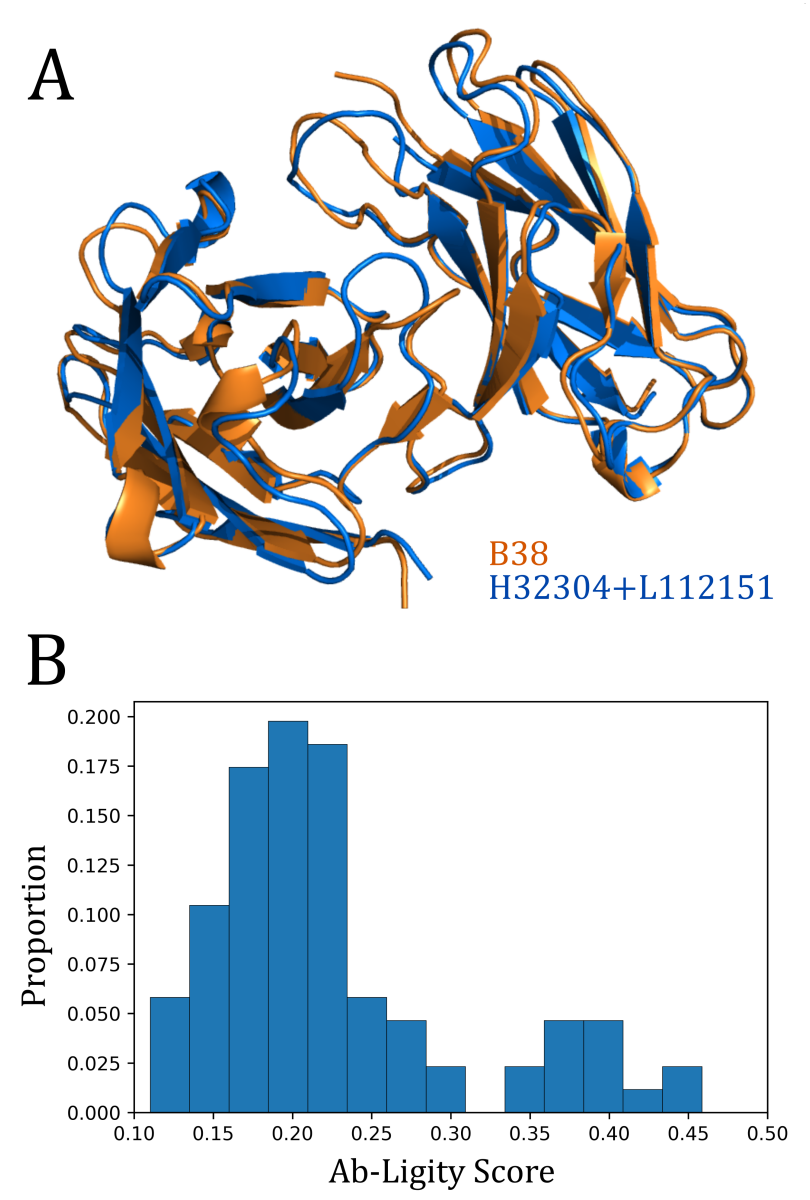


Figure 5.6: (A) The backbone structural alignment between the solved B38 crystal structure and the closest 'Public Baseline' Antibody Model Library model, H32304+L112151. (B) The distribution of Ab-Ligity scores assigned to the 86/105 Fvs that surpassed the 0.1 threshold.

5.4 Update and Chapter Conclusion

At the time of writing (2nd October, 2020), CoV-AbDab contains 1,545 entries and we are tracking at least 25 preprints likely to release sequence information upon publication. As the global pandemic has accelerated throughout 2020, it seems likely that many more immunological studies will be performed into SARS-CoV-2 infection

and that CoV-AbDab will grow accordingly.

Debate still rages as to whether the B-cell or T-cell compartment of the immune response confers most protection against the virus [364], and also as to the duration of effective immunological memory [365, 366]. It has also been hypothesised that some antibody responses might be associated with worse prognosis (*via.* mechanisms as antibody-dependent enhancement) while others may be efficacious in accelerating viral clearance from the body [81, 367–369]. At a general level, high titres of spike-binding antibodies have been correlated with more severe clinical symptoms in natural SARS-CoV-2 infections [370, 371]. Databases will play a key role in assimilating the clinical findings of different hospitals around the world, to observe which characterised antibodies are statistically significantly associated with various disease severities.

It is our hope that Ig-seq experiments, coupled with repositories like CoV-AbDab and high-throughput repertoire functional characterisation experiments [70, 71], will be harnessed to characterise vaccine responses to molecular detail. The level of understanding provided by such a framework has the potential to feedback into vaccine design and yield a step-change improvement in both efficacy and safety.

Finally, we note that a rare positive consequence of the coronavirus pandemic will be an unprecedented amount of publicly available antibody binding data against a wide diversity of antigen epitopes. It is therefore likely that CoV-AbDab will find use beyond the pandemic in providing training data and case studies for future bioinformatics software packages.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This DPhil project has made significant advances towards a structure-based paradigm for antibody drug discovery from natural repertoire sequences. We propose 'Repertoire Structural Profiling' (RSP) as a computationally-tractable means of generating structurally diverse natural antibody libraries for virtual screening. These representations of human antibody repertoires also hint at a set of 'public' naive binding site topologies, which could be used to further understand epitope immunodominance or to cluster together more sequence-diverse antibodies with same-epitope reactivity. Coupled with RSP, our 'Therapeutic Antibody Profiler' (TAP) shows how statistical and biochemical principles can be combined to prioritise more developable antibody lead candidates in early-stage antibody development. These fully-*in silico* approaches fit within our broader vision of informatics-driven antibody discovery. The next stage for RSP and TAP is benchmarking within industrial pipelines, where sufficient data can be retrieved to improve algorithmic performance and yield real-world impact.

Underpinning these discoveries is the groundwork performed in collating the Therapeutic Structural Antibody Database (Thera-SAbDab). Once at critical mass, this molecular information on clinically-investigated antibodies will be exploitable to discern further general features that distinguish successful from unsuccessful antibody drugs. We applied similar principles in compiling the Coronavirus Antibody Database (CoV-AbDab) during the early stages of the SARS-CoV-2 pandemic. The over 1,500 datapoints in CoV-AbDab are being used to guide prophylactic antibody design and to better understand the protective capacity of both natural infection and vaccine response B-cell receptor repertoires. The impact of this database, visited by over 3,000 independent users worldwide at the time of writing, builds a strong case for a

pandemic preparedness strategy that includes advance curation of similar libraries of immunoglobulins that can bind to all known infectious agents.

6.2 Future Work

The obvious next-step for this project is to complete the proposed pipeline by designing a docking protocol that can both cope with the computational complexity of assessment thousands of Antibody Model Library (AML) antibodies while simultaneously providing sufficiently accurate results that can stand up to experimental scrutiny. This is likely to require a tiered process of multiple docking rounds, starting with rapid rigid-body docking and adding more consideration for flexibility and reintroducing sequence diversity around promising scaffolds as the most obvious non-complementary AML topologies are weeded out.

A major challenge of molecular docking is accurate pose scoring. In recent years, three-dimensional convolutional neural network (3D CNN) architectures, where inputs are voxelised binding site co-ordinates and outputs can be interpreted as complementarity probabilities, have significantly improved small molecule pose ranking. Consequently, another group member (Constantin Schneider) has developed a bespoke 3D-CNN antibody-antigen pose scoring function, trained on PDB-derived antibody model-antigen crystal structure complexes. This work has delivered promising results and is nearing completion, so could be incorporated into this docking protocol within months.

To achieve maximal impact on future antibody drug discovery, we must be able to show that our methodology can computationally identify true antigen binders from AMLs with considerably higher enrichment than existing *in vitro* technologies and with shorter lead-times. We have therefore negotiated a Confidential Disclosure Agreement with Hoffmann-La Roche both for access to their database of binding affinity data against a range of antigen epitopes and use of their in-house antibody expression and binding assessment technology for experimental assessment of as yet untested antibody lineages and antigen targets. Their existing database will be useful for 'post-diction' studies, where we see how well a 'blindfolded' version of our algorithm can identify known binders. It will also provide extra training data for our CNN scoring function, which can learn from a greater diversity of true positives, improving performance while not compromising Intellectual Property. Having access to their antibody expression technology will be pivotal in benchmarking a subsequent computational affinity maturation protocol, which can suggest mutations in a

developability-aware manner, to improve binding strength to the levels required by industry.

A further investigation would be to analyse the effect of the light chain on observed AML structural diversity. Many publicly-available BCR repertoire samples only contain heavy chain information, and so cannot currently be analysed. Although deeper paired-chain sequencing data is on the horizon, it is likely to take several years for the full range of diseases analysed by Next-Generation Sequencing to be investigated at the single-cell level. Therefore, the identification of a representative structural subset of light chains to pair with any heavy-only dataset would enable the creation of a greater diversity of disease-specific AMLs in the nearer-term.

Reference List

[1] Raybould MIJ, Wong WK, Deane CM (2019) Antibody–antigen complex modelling in the era of immunoglobulin repertoire sequencing. *Mol. Syst. Des. Eng.* 4:679–688.

[2] Shi Z, et al. (2019) More than one antibody of individual B cells revealed by single-cell immune profiling. *Cell Discov.* 5:64.

[3] Greiff V, et al. (2017) Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. *Cell Rep.* 19(7):1467–1478.

[4] Briney B, Inderbitzin A, Joyce C, Burton DR (2019) Commonality despite exceptional diversity in the baseline antibody repertoire. *Nature* 566:393–397.

[5] Narciso JET, et al. (2011) Analysis of the antibody structure based on high-resolution crystallographic studies. *N. Biotechnol.* 28(5):435–447.

[6] Murphy K, Weaver C (2016) *Janeway's Immunobiology, 9th Edition.* (Garland Science).

[7] Giudicelli V, Chaume D, Lefranc MP (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.* 33:D256–D261.

[8] Kabat EA, Wu TT, Perry HM, Foeller C, Gottesman KS (1983) *Sequences of Proteins of Immunological Interest, 5th Edition.* (National Institutes of Health).

[9] Dunbar J, et al. (2014) SAbDab: The Structural Antibody Database. *Nucleic Acids Res.* 42(D1):1140–1146.

[10] Marks C, Deane CM (2017) Antibody H3 Structure Prediction. *Comput. Struct. Biotechnol. J.* 15:222–231.

[11] Lefranc MP, et al. (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.

[12] Chothia C, Lesk AM (1987) Canonical Structures for the Hypervariable Regions of Immunoglobulins. *J. Mol. Biol.* 196(4):901–917.

[13] Abhinandan K, Martin AC (2008) Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Mol. Immunol.* 45(14):3832–3839.

[14] Honegger A, Plückthun AP (2001) Yet Another Numbering Scheme for Immunoglobulin Variable Domains: An Automatic Modeling and Analysis Tool. *J. Mol. Biol.* 309(3):657–670.

[15] North B, Lehmann A, Dunbrack Jr. RL (2011) A new clustering of antibody CDR loop conformations. *J. Mol. Biol.* 406(2):228–256.

[16] Ehrenmann F, Kaas Q, Lefranc MP (2009) IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res.* 38(11):D301–D307.

[17] Adolf-Bryfogle J, Xu Q, North B, Lehmann A, Dunbrack, Roland L. J (2014) PyIgClassify: a database of antibody CDR structural classifications. *Nucleic Acids Res.* 43(D1):D432–D438.

[18] Dunbar J, Deane CM (2016) ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics* 32(2):298–300.

[19] Berman HM, et al. (2000) The Protein Data Bank. *Nucleic Acids Res.* 28(1):235–242.

[20] Dunbar J, Fuchs A, Shi J, Deane CM (2013) ABangle: characterising the VH–VL orientation in antibodies. *Protein Eng. Des. Sel.* 26(10):611–620.

[21] Bujoztek A, et al. (2015) Prediction of VH–VL domain orientation for antibody variable domain modeling. *Proteins* 83(4):681–695.

[22] Chothia C, et al. (1989) Conformations of immunoglobulin hypervariable regions. *Nature* 342(6252):877–883.

[23] Al-Lazikani B, Lesk AM, Chothia C (1997) Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.* 273(4):927–948.

[24] Nowak J, et al. (2016) Length-independent structural similarities enrich the antibody CDR canonical class model. *mAbs* 8(4):751–760.

[25] Wong WK, et al. (2018) SCALOP: sequence-based antibody canonical loop structure annotation. *Bioinformatics* 35(10):1774–1776.

[26] Choi Y, Deane CM (2010) FREAD revisited: Accurate loop structure prediction using a database search algorithm. *Proteins* 78(6):1431–1440.

[27] Choi Y, Deane CM (2011) Predicting antibody complementarity determining region structures without classification. *Mol. Biosyst.* 7(12):3327–3334.

[28] Regep C, Georges G, Shi J, Popovic B, Deane CM (2017) The H3 loop of antibodies shows unique structural characteristics. *Proteins* 85(7):1311–1318.

[29] Stein A, Kortemme T (2013) Improvements to Robotics-Inspired Conformational Sampling in Rosetta. *PLoS One* 8(5):1–13.

[30] Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B (2008) Loop modeling: Sampling, filtering, and scoring. *Proteins* 70(3):834–843.

[31] Liang S, Zhang C, Zhou Y (2013) LEAP: Highly accurate prediction of protein loop conformations by integrating coarse-grained sampling and optimized energy scores with all-atom refinement of backbone and side chains. *J. Comput. Chem.* 35(4):335–341.

[32] Jacobsen MP, et al. (2004) A hierarchical approach to all-atom protein loop prediction. *Proteins* 55(2):351–367.

[33] Fiser A, Do RKG, Šali A (2008) Modeling of loops in protein structures. *Protein Sci.* 9(9):1753–1773.

[34] Wang G, Dunbrack Jr RL (2003) PISCES: a protein sequence culling server. *Bioinformatics* 19(12):1589–1591.

[35] Weitzner BD, et al. (2017) Modeling and docking of antibody structures with Rosetta. *Nat. Protoc.* 12(2):401–416.

[36] Deane CM, Blundell TL (2001) CODA: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Sci.* 10(3):599–612.

[37] Karami Y, Guyon F, De Vries S, Tufféry P (2018) DaReUS-Loop: accurate loop modeling using fragments from remote or unrelated proteins. *Sci. Rep.* 8:13673.

[38] Messih MA, Lepore R, Tramontano A (2015) LoopIng: a template-based tool for predicting the structure of protein loops. *Bioinformatics* 31(23):3767–3772.

[39] Hildebrand PW, et al. (2009) SuperLooper—a prediction server for the modeling of loops in globular and membrane proteins. *Nucleic Acids Res.* 37:W571–W574.

[40] Fernandez-Fuentes N, Zhai J, Fiser A (2006) ArchPRED: a template based loop structure prediction server. *Nucleic Acids Res.* 34:W173–W176.

[41] Michalsky E, Goede A, Preissner R (2003) Loops In Proteins (LIP)—a comprehensive loop database for homology modelling. *Protein Eng. Des. Sel.* 16(12):979–985.

[42] Holtby D, Li SC, Li M (2013) LoopWeaver: Loop Modeling by the Weighted Scaling of Verified Proteins. *J. Comput. Biol.* 20(3):212–223.

[43] Marcatili P, Olimpieri PP, Chailyan A, Tramontano A (2014) Antibody modeling using the Prediction of ImmunoGlobulin Structure (PIGS) web server. *Nat. Protoc.* 9(12):2771–2784.

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

[45] Fasnacht M, et al. (2014) Automated antibody structure prediction using Accelrys tools: Results and best practices. *Proteins* 82(8):1583–1598.

[46] Martin AC, Cheetham JC, Rees AR (1989) Modeling antibody hypervariable loops: a combined algorithm. *Proc. Natl. Acad. Sci. USA* 86(23):9268–9272.

[47] Whitelegg NRJ, Rees AR (2000) WAM: an improved algorithm for modelling antibodies on the WEB. *Protein Eng. Des. Sel.* 13(12):819–824.

[48] Marks C, et al. (2017) Sphinx: Merging knowledge-based and ab initio approaches to improve protein loop prediction. *Bioinformatics* 33(9):1346–1353.

[49] Krivov GG, Shapovalov MV, Dunbrack Jr. RL (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* 11(4):778–795.

[50] Miao Z, Cao Y, Jiang T (2011) RASP: rapid modeling of protein side chain conformations. *Bioinformatics* 27(22):3117–3122.

[51] Nagata K, Randall A, Baldi P (2012) SIDEpro: a novel machine learning approach for the fast and accurate prediction of side chain conformations. *Proteins* 80(1):142–153.

[52] Wood CW, et al. (2014) CCBuilder: an interactive web-based tool for building, designing and assessing coiled-coil protein assemblies. *Bioinformatics* 30(21):3029–3035.

[53] Leem J, Georges G, Shi J, Deane CM (2018) Antibody side chain conformations are position-dependent. *Proteins* 86(4):383–392.

[54] Yamashita K, et al. (2014) Kotai Antibody Builder: automated high-resolution structural modeling of antibodies. *Bioinformatics* 30(22):3279–3280.

[55] Klausen MS, Anderson MV, Jesperson MC, Nielsen M, Marcatili P (2015) LYRA, a webserver for lymphocyte receptor structural modeling. *Nucleic Acids Res.* 43:W349–W355.

[56] Marcatili P, Rosi A, Tramontano A (2008) PIGS: automatic prediction of antibody structures. *Bioinformatics* 24(17):1953–1954.

[57] Kemmish H, Fasnacht M, Yan L (2017) Fully automated antibody structure prediction using BIOVIA tools: Validation study. *PLoS One* 12(5):e0177923.

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

[59] Bujotzek A, et al. (2015) MoFvAb: Modeling the Fv region of antibodies. *mAbs* 7(5):838–852.

[60] Zhu K, et al. (2014) Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction. *Proteins* 82(8):1646–1655.

[61] Berrondo M, Kaufmann S, Berrondo M (2014) Automated Aufbau of antibody structures from given sequences using Macromoltek's SmrtMolAntibody. *Proteins* 82:1636–1645.

[62] Abhinandan KR, Martin ACR (2010) Analysis and prediction of VH/VL packing in antibodies. *Protein Eng. Des. Sel.* 23(9):689–697.

[63] Sali A, Blundell TL (1993) Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* 234(3):779–815.

[64] Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17(6):333–351.

[65] Ma X, et al. (2019) Analysis of error profiles in deep next-generation sequencing data. *Genome Biol.* 20:50.

[66] Stubbington MJT, Rozenblatt-Rosen O, Regev A, Sarah A (2017) Single cell transcriptomics to explore the immune system in health and disease. *Science* 358(6359):58–63.

[67] Hedlund E, Deng Q (2018) Single-cell RNA sequencing: Technical advancements and biological applications. *Mol. Aspects Med.* 59:36–46.

[68] Rizzetto S, et al. (2018) B-cell receptor reconstruction from single-cell RNA-seq with VDJPuzzle. *Bioinformatics* 34(16):2846–2847.

[69] Stuart T, Satija R (2019) Integrative single-cell analysis. *Nat. Rev. Genet.* 20:257–272.

[70] Setliff I, et al. (2019) High-Throughput Mapping of B Cell Receptor Sequences to Antigen Specificity. *Cell* 179(7):1636–1646.

[71] Gérard A, et al. (2020) High-throughput single-cell activity-based screening and sequencing of antibodies using droplet microfluidics. *Nat. Biotechnol.* 38:715–721.

[72] Goldstein LD, et al. (2019) Massively parallel single-cell B-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. *Commun. Biol.* 2:304.

[73] Horns F, Dekker CL, Quake SR (2020) Memory B Cell Activation, Broad Anti-influenza Antibodies, and Bystander Activation Revealed by Single-Cell Transcriptomics. *Cell Rep.* 30(3):905–913.

[74] Adamo L, et al. (2020) Myocardial B cells are a subset of circulating lymphocytes with delayed transit through the heart. *JCI Insight* 5(3).

[75] Eccles JD, et al. (2020) T-bet+ Memory B Cells Link to Local Cross-Reactive IgG upon Human Rhinovirus Infection. *Cell Rep.* 30(2):351–366.

[76] Alsoussi WB, et al. (2020) A Potently Neutralizing Antibody Protects Mice against SARS-CoV-2 Infection. *J. Immunol.*

[77] King HW, et al. (2020) Antibody repertoire and gene expression dynamics of diverse human B cell states during affinity maturation. *bioRxiv.*

[78] Chen H, et al. (2020) BCR selection and affinity maturation in Peyer's patch germinal centres. *Nature* 582:421–425.

[79] Ghraichy M, et al. (2020) Maturation of the Human Immunoglobulin Heavy Chain Repertoire With Age. *Front. Immunol.* 11:1734.

[80] Soto C, et al. (2019) High frequency of shared clonotypes in human B cell receptor repertoires. *Nature* 566:398–402.

[81] Galson JD, et al. (2020) Deep sequencing of B cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures. *Front. Immunol.*

[82] Setliff I, et al. (2018) Multi-Donor Longitudinal Antibody Repertoire Sequencing Reveals the Existence of Public Antibody Clonotypes in HIV-1 Infection. *Cell Host Microbe* 23(6):845–854.

[83] Gidoni M, et al. (2019) Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. *Nat. Commun.* 10:628.

[84] Gupta NT, et al. (2017) Hierarchical clustering can identify B cell clones with high confidence in Ig repertoire sequencing data. *J. Immunol.* 198(6):2489–2499.

[85] Bhattacharya S, et al. (2018) ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci. Data* 5:180015.

[86] Christley S, et al. (2018) VDJServer: A Cloud-Based Analysis Portal and Data Commons for Immune Repertoire Sequences and Rearrangements. *Front. Immunol.* 9:976.

[87] Corrie BD, et al. (2018) iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunol. Rev.* 284(1):24–41.

[88] Kovaltsuk A, et al. (2018) Observed Antibody Space: a resource for data mining next generation sequencing antibody repertoires. *J. Immunol.* 201(8):2502–2509.

[89] Vander Heiden JA, et al. (2017) Dysregulation of B Cell Repertoire Formation in Myasthenia Gravis Patients Revealed through Deep Sequencing. *J. Immunol.* 198:1460–1473.

[90] Breden F, et al. (2017) Reproducibility and Reuse of Adaptive Immune Receptor Repertoire Data. *Front. Immunol.* 8:1418.

[91] Marks C, Deane CM (2020) How repertoire data is changing antibody science. *J. Biol. Chem.*

[92] Greiff V, et al. (2017) Learning the High-Dimensional Immunogenomic Features That Predict Public and Private Antibody Repertoires. *J. Immunol.* 199(8):2985–2997.

[93] Arora R, Arnaout R (2020) Private Antibody Repertoires Are Public. *bioRxiv.*

[94] Clavero-Álvarez A, Di Mambro T, Perez-Gaviro S, Magnani M, Bruscolini P (2018) Humanization of Antibodies using a Statistical Inference Approach. *Sci. Rep.* 8:14820.

[95] Joyce C, Burton DR, Briney B (2020) Comparisons of the antibody repertoires of a humanized rodent and humans by high throughput sequencing. *Sci. Rep.* 10:1120.

[96] Kovaltsuk A, et al. (2020) Structural diversity of B-cell receptor repertoires along the B-cell differentiation axis in humans and mice. *PLoS Comput. Biol.* 16(2):1–20.

[97] Fernàndez-Quintero ML, et al. (2020) Local and Global Rigidification Upon Antibody Affinity Maturation. *Front. Mol. Biosci.* 7:182.

[98] Nielsen SCA, et al. (2020) Human B cell clonal expansion and convergent antibody responses to SARS-CoV-2. *Cell Host Microbe.*

[99] Raybould MIJ, et al. (2019) Thera-SAbDab: the Therapeutic Structural Antibody Database. *Nucleic Acids Res.* 48(D1):D383–D388.

[100] Kaplon H, Reichert J (2019) Antibodies to watch in 2019. *mAbs* 11(2):219–238.

[101] Kaplon H, Muralidharan M, Schneider Z, Reichert J (2020) Antibodies to watch in 2020. *mAbs* 12(1):1703531.

[102] Sevigny J, et al. (2016) The antibody aducanumab reduces Aβ plaques in Alzheimer's disease. *Nature* 537(7618):50–56.

[103] Jones TD, et al. (2016) The INNs and outs of antibody nonproprietary names. *mAbs* 8(1):1–9.

[104] Carter PJ, Lazar GA (2018) Next generation antibody drugs: Pursuit of the 'high-hanging fruit'. *Nat. Rev. Drug Discov.* 17(3):197–223.

[105] Liu H, Saxena A, Sidhu SS, Wu D (2017) Fc Engineering for Developing Therapeutic Bispecific Antibodies and Novel Scaffolds. *Front. Immunol.* 8:38.

[106] Labrijn AF, Janmaat ML, Reichert JM, Parren PWHI (2019) Bispecific antibodies: a mechanistic review of the pipeline. *Nat. Rev. Drug Discov.* 18:585–608.

[107] Suurs FV, Lub-de Hooge MN, de Vries EGE, de Groot DJA (2019) A review of bispecific antibodies and antibody constructs in oncology and clinical challenges. *Pharmacol. Ther.* 201:103–119.

[108] Ridgway JBB, Presta LG, Carter P (1996) 'Knobs-into-holes' engineering of antibody CH3 domains for heavy chain heterodimerization. *Protein Eng. Des. Sel.* 9(7):617–621.

[109] Yang EY, Shah K (2020) Nanobodies: Next Generation of Cancer Diagnostics and Therapeutics. *Front. Oncol.* 10:1182.

[110] Messer A, Butler DC (2020) Optimizing intracellular antibodies (intrabodies/nanobodies) to treat neurodegenerative disorders. *Neurobiol. Dis.* 134:104619.

[111] Mitchell LS, Colwell LJ (2018) Comparative analysis of nanobody sequence and structure data. *Proteins* 86(7):697–706.

[112] Yu L, Guan Y (2014) Immunologic basis for long HCDR3s in broadly neutralizing antibodies against HIV-1. *Front. Immunol.* 5:250.

[113] Sormanni P, Aprile FA, Vendruscolo M (2018) Third generation antibody discovery methods: in silico rational design. *Chem. Soc. Rev.* 47(24):9137–9157.

[114] Liu JHL (2014) The history of monoclonal antibody development – Progress, remaining challenges and future innovations. *Ann. Med. Surg.* 3(4):113–116.

[115] Lee ECL, et al. (2014) Complete humanization of the mouse immunoglobulin loci enables efficient therapeutic antibody discovery. *Nat. Biotechnol.* 32:356–363.

[116] Frietze KM, Pascale JM, Moreno B, Chackerian B, Peabody DS (2017) Pathogen-specific deep sequence-coupled biopanning: A method for surveying human antibody responses. *PLoS ONE* 12(2):e0171511.

[117] Almagro JC, Martha PE, Arrieta HI, Pèrez-Tapia SM (2019) Phage Display Libraries for Antibody Therapeutic Discovery and Development. *Antibodies* 8(3):44.

[118] Rosowski S, et al. (2018) A novel one-step approach for the construction of yeast surface display Fab antibody libraries. *Microb. Cell. Fact.* 17:3.

[119] Fleishman SJ, et al. (2011) Rosettascripts: A scripting language interface to the Rosetta Macromolecular modeling suite. *PLoS ONE* 6(6):1–10.

[120] Liu S, et al. (2007) Nonnatural protein–protein interaction-pair design by key residues grafting. *Proc. Natl. Acad. Sci. USA* 104(13):5330–5335.

[121] Liu X, et al. (2017) Computational design of an epitope-specific Keap1 binding antibody using hotspot residues grafting and CDR loop swapping. *Sci. Rep.* 7:41306.

[122] Mason DM, et al. (2019) Deep learning enables therapeutic antibody optimization in mammalian cells by deciphering high-dimensional protein sequence space. *bioRxiv*.

[123] Pantazes RJ, Maranas CD (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.

[124] Li T, Pantazes RJ, Maranas CD (2014) OptMAVEn - A new framework for the de novo design of antibody variable region models targeting specific antigen epitopes. *PLoS One* 9(8):e105954.

[125] Lapidoth GD, et al. (2015) AbDesign: An algorithm for combinatorial backbone design guided by natural conformations and sequences. *Proteins* 83(8):1385–1406.

[126] Adolf-Bryfogle J, et al. (2018) RosettaAntibodyDesign (RAbD): A general framework for computational antibody design. *PLoS Comput. Biol.* 14(4).

[127] Kovaltsuk A, et al. (2017) How B-Cell Receptor Repertoire Sequencing Can Be Enriched with Structural Antibody Data. *Front. Immunol.* 8:1753.

[128] Barderas R, Desmet J, Timmerman P, Meloen R, Casal JI (2008) Affinity maturation of antibodies assisted by in silico modeling. *Proc. Natl. Acad. Sci. USA* 105(26):9029–9034.

[129] Livingstone JR (1996) Antibody characterization by isothermal titration calorimetry. *Nature* 384:491–492.

[130] Vincke C, et al. (2012) Generation of single domain antibody fragments derived from camelids and generation of manifold constructs. *Methods Mol. Biol.* 907:145–176.

[131] Aydin S (2015) A short history, principles, and types of ELISA, and our laboratory experience with peptide/protein analyses using ELISA. *Peptides* 72:4–15.

[132] Ebo J, et al. (2020) An in vivo platform to select and evolve aggregation-resistant proteins. *Nat. Commun.* 11:1816.

[133] Abdiche Y, Malashock D, Pinkerton A, Pons J (2008) Determining kinetics and affinities of protein interactions using a parallel real-time label-free biosensor, the Octet. *Anal. Biochem.* 377(2):209–217.

[134] Almagro JC, et al. (2014) Second Antibody Modeling Assessment (AMA-II). *Proteins* 82(8):1553–1562.

[135] Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J (2019) Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins* 87(12):1011–1020.

[136] Kunik V, Peters B, Ofran Y (2012) Structural consensus among antibodies defines the antigen binding site. *PLoS Comput. Biol.* 8(2):e1002388.

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

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

[139] Liberis E, Veličković P, Sormanni P, Vendruscolo M, Liò P (2018) Parapred: antibody paratope prediction using convolutional and recurrent neural networks. *Bioinformatics* 34(17):2944–2950.

[140] Krawczyk K, Baker T, Shi J, Deane CM (2013) Antibody i-Patch prediction of the antibody binding site improves rigid local antibody-antigen docking. *Protein Eng. Des. Sel.* 26(10):621–629.

[141] Emini EA, Hughes JV, Perlow DS, Boger J (1985) Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *J. Virol.* 55(3):836–839.

[142] Karplus PA, Schulz GE (1985) Prediction of chain flexibility in proteins. *Naturwissenschaften* 72(4):212–213.

[143] Yao B, Zhang L, Liang S, Zhang C (2012) SVMTriP: A Method to Predict Antigenic Epitopes Using Support Vector Machine to Integrate Tri-Peptide Similarity and Propensity. *PLoS One* 7(9):e45152.

[144] Chen W, Guo WW, Huang Y, Ma Z (2012) Pepmapper: A collaborative web tool for mapping epitopes from affinity-selected peptides. *PLoS One* 7(5):e37869.

[145] Gao J, Faraggi E, Zhou Y, Ruan J, Kurgan L (2012) BEST: Improved prediction of B-cell epitopes from antigen sequences. *PLoS One* 7(6):e40104.

[146] Singh H, Ansari HR, Raghava GPS (2013) Improved method for linear b-cell epitope prediction using antigen's primary sequence. *PLoS One* 8(5):e62216.

[147] Esmaielbeiki R, Krawczyk K, Knapp B, Nebel JC, Deane CM (2015) Progress and challenges in predicting protein interfaces. *Brief. Bioinformatics* 17(1):117–131.

[148] Ramaraj T, Angel T, Dratz EA, Jesaitis AJ, Mumey B (2012) Antigen–antibody interface properties: composition, residue interactions, and features of 53 non-redundant structures. *Biochim. Biophys. Acta.* 1824(3):520–532.

[149] Peng HP, Lee KH, Jian JW, Yang AS (2014) Origins of specificity and affinity in antibody–protein interactions. *Proc. Natl. Acad. Sci. USA* 111(26):E2656–E2665.

[150] Kuroda D, Gray JJ (2016) Shape complementarity and hydrogen bond preferences in protein-protein interfaces: Implications for antibody modeling and protein-protein docking. *Bioinformatics* 32(16):2451–2456.

[151] Dalkas GA, Rooman M (2017) SEPIa, a knowledge-driven algorithm for predicting conformational B-cell epitopes from the amino acid sequence. *BMC Bioinformatics* 18(1):95.

[152] Jespersen MC, Peters B, Nielsen M, Marcatili P (2017) Bepipred-2.0: improving sequence-based b-cell epitope prediction using conformational epitopes. *Nucleic Acids Res.* 45(W1):W24–W29.

[153] Zhao L, et al. (2018) Novel overlapping subgraph clustering for the detection of antigen epitopes. *Bioinformatics* 34(12):2061–2068.

[154] Greenbaum JA, et al. (2007) Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *J. Mol. Recognit.* 20(2):75–82.

[155] Kunik V, Ofran Y (2013) The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops. *Protein Eng. Des. Sel.* 26(10):599–609.

[156] Rapberger R, Lukas A, Mayer B (2007) Identification of discontinuous antigenic determinants on proteins based on shape complementarities. *J. Mol. Recognit.* 20(2):113–121.

[157] Soga S, Kuroda D, Shirai H, Kobori M, Hirayama N (2010) Use of amino acid composition to predict epitope residues of individual antibodies. *Protein Eng. Des. Sel.* 23(6):441–448.

[158] Zhao L, Li J (2010) Mining for the antibody-antigen interacting associations that predict the B cell epitopes. *BMC Struct. Biol.* 10:S6.

[159] Zhao L, Wong L, Li J (2011) Antibody-specified B-cell epitope prediction in line with the principle of context-awareness. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 8(6):1483–1494.

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

[161] Ahmad S, Mizuguchi K (2011) Partner-aware prediction of interacting residues in protein-protein complexes from sequence data. *PLoS One* 6(12):e29104.

[162] Sela-Culang I, Ashkenazi S, Peters B, Ofran Y (2014) Pease: predicting b-cell epitopes utilizing antibody sequence. *Bioinformatics* 31(8):1313–1315.

[163] Bourquard T, et al. (2018) MAbTope: A Method for Improved Epitope Mapping. *J. Immunol.* 201(10):ji1701722.

[164] Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* 161(2):269–288.

[165] Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.* 33:W363–W367.

[166] Kozakov D, Brenke R, Comeau SR, Vajda S (2006) Piper: an fft-based protein docking program with pairwise potentials. *Proteins* 65(2):392–406.

[167] Tovchigrechko A, Vakser IA (2006) GRAMM-X public web server for protein–protein docking. *Nucleic Acids Res.* 34:W310–W314.

[168] Pierce BG, et al. (2014) ZDOCK server: Interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics* 30(12):1771–1773.

[169] Shimba N, Kamiya N, Nakamura H (2016) Model Building of Antibody–Antigen Complex Structures Using GBSA Scores. *J. Chem. Inf. Model* 56(10):2005–2012.

[170] Kozakov D, et al. (2017) The ClusPro web server for protein–protein docking. *Nat. Protoc.* 12(2):255.

[171] Pagadala NS, Syed K, Tuszynski J (2017) Software for molecular docking: a review. *Biophys. Rev.* 9(2):91–102.

[172] Sircar A, Gray JJ (2010) SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Comput. Biol.* 6(1):e1000644.

[173] Torchala M, Moal IH, Chaleil RA, Fernandez-Recio J, Bates PA (2013) SwarmDock: a server for flexible protein–protein docking. *Bioinformatics* 29(6):807–809.

[174] Van Zundert GCP, et al. (2016) The HADDOCK2. 2 web server: user-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol.* 428(4):720–725.

[175] Marze NA, Roy Burman SS, Sheffler W, Gray JJ (2018) Efficient Flexible Backbone Protein-Protein Docking for Challenging Targets. *Bioinformatics* 34(20):3461–3469.

[176] Marks C, Shi J, Deane CM (2018) Predicting loop conformational ensembles. *Bioinformatics* 34(6):949–956.

[177] Macindoe G, Mavridis L, Venkatraman V, Devignes MD, Ritchie DW (2010) Hexserver: an fft-based protein docking server powered by graphics processors. *Nucleic Acids Res.* 38:W445–W449.

[178] Ramírez-Aportela E, López-Blanco JR, Chacón P (2016) Frodock 2.0: fast protein–protein docking server. *Bioinformatics* 32(15):2386–2388.

[179] Jarasch A, et al. (2015) Developability assessment during the selection of novel therapeutic antibodies. *J. Pharm. Sci.* 104(6):1885–1898.

[180] Dobson CL, et al. (2016) Engineering the surface properties of a human monoclonal antibody prevents self-association and rapid clearance *in vivo*. *Sci. Rep.* 6:1–14.

[181] Mehta SB, Bee JS, Randolph TW, Carpenter JF (2014) Partial unfolding of a monoclonal antibody: Role of a single domain in driving protein aggregation. *Biochemistry* 53(20):3367–3377.

[182] Yadav S, Laue TM, Kalonia DS, Singh SN, Shire SJ (2012) The influence of charge distribution on self-association and viscosity behavior of monoclonal antibody solutions. *Mol. Pharm.* 9(4):791–802.

[183] Sharma VK, et al. (2014) In silico selection of therapeutic antibodies for development: Viscosity, clearance, and chemical stability. *Proc. Natl. Acad. Sci. USA* 111(52):18601–18606.

[184] Ferreira GM, Shahfar H, Sathish HA, Remmele Jr. RL, Roberts CJ (2019) Identifying Key Residues That Drive Strong Electrostatic Attractions between Therapeutic Antibodies. *J. Phys. Chem. B* 123(50):10642–10653.

[185] Hebditch M, Warwicker J (2019) Charge and hydrophobicity are key features in sequence-trained machine learning models for predicting the biophysical properties of clinical-stage antibodies. *PeerJ* 7:e8199.

[186] Jain T, et al. (2017) Prediction of delayed retention of antibodies in hydrophobic interaction chromatography from sequence using machine learning. *Bioinformatics* 33:3758–3766.

[187] Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL (2009) Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci. USA* 106(29):11937–11942.

[188] Lauer TM, et al. (2012) Developability index: A rapid in silico tool for the screening of antibody aggregation propensity. *J. Pharm. Sci.* 101(1):102–115.

[189] Obrezanova O, et al. (2015) Aggregation risk prediction for antibodies and its application to biotherapeutic development. *mAbs* 7(2):352–363.

[190] Tisoncik JR, Korth MJ, Simmons CP, Farrar J, Martin, Thomas R. Katze MG (2012) Into the Eye of the Cytokine Storm. *Microbiol. Mol. Biol. Rev.* 76(1):16–32.

[191] Harding FA, Stickler MM, Razo J, Dubridge RB (2010) The immunogenicity of humanized and fully human antibodies Residual immunogenicity resides in the CDR regions. *mAbs* 2(3):256–265.

[192] Vaisman-Mentesh A, Gutierrez-Gonzalez M, DeKosky BJ, Wine Y (2020) The Molecular Mechanisms That Underlie the Immune Biology of Anti-drug Antibody Formation Following Treatment With Monoclonal Antibodies. *Front. Immunol.*

[193] Abhinandan KR, Martin ACR (2007) Analyzing the "Degree of Humanness" of Antibody Sequences. *J. Mol. Biol.* 369(3):852–862.

[194] King C, et al. (2014) Removing T-cell epitopes with computational protein design. *Proc. Natl. Acad. Sci. USA* 111(23):8577–8582.

[195] Seeliger D, et al. (2015) Boosting antibody developability through rational sequence optimization. *mAbs* 7(3):505–515.

[196] Popovic B, et al. (2017) Engineering the expression of an anti-interleukin-13 antibody through rational design and mutagenesis. *Protein Eng. Des. Sel.* 30(4):303–311.

[197] Jiskoot W, C. BE, de Koning A. A., Herron JNCDJ (1990) Analytical approaches to the study of monoclonal antibody stability. *Pharm Res.* 7(12):1234–1241.

[198] Hedberg SHM, Rapley J, Haigh JM, Williams DR (2018) Cross-interaction chromatography as a rapid screening technique to identify the stability of new antibody therapeutics. *Eur. J. Pharm. Biopharm.* 133:131–137.

[199] Delmar JA, Wang J, Choi SW, Martins JA, Mikhail JP (2019) Machine Learning Enables Accurate Prediction of Asparagine Deamidation Probability and Rate. *Mol. Ther. Methods Clin. Dev.* 15:264–274.

[200] Rabia LA, Desai AA, Jhajj HS, Tessier PMT (2018) Understanding and overcoming trade-offs between antibody affinity, specificity, stability and solubility. *Biochem. Eng. J.* 137:365–374.

[201] Shehata L, et al. (2019) Affinity Maturation Enhances Antibody Specificity but Compromises Conformational Stability. *Cell Rep.* 28(13):3300–3308.

[202] Krawczyk K, Raybould MIJ, Kovaltsuk A, Deane CM (2019) Looking for therapeutic antibodies in next-generation sequencing repositories. *mAbs* 11(7):1197–1205.

[203] Grilo AL, Mantalaris A (2019) The Increasingly Human and Profitable Monoclonal Antibody Market. *Trends Biotechnol.* 37(1):9–16.

[204] Steeland S, Vandenbroucke RE, Libert C (2016) Nanobodies as therapeutics: big opportunities for small antibodies. *Drug Discov. Today* 21(7):1076–1113.

[205] Jevševar S, Mateja K, Kenig M (2012) PEGylation of Antibody Fragments for Half-Life Extension. *Methods Mol. Biol.* 901:233–246.

[206] Steiner M, Neri D (2011) Antibody-Radionuclide Conjugates for Cancer Therapy: Historical Considerations and New Trends. *Clin. Cancer Res.* 17(20):6406–6416.

[207] Beck A, Goetsch L, Dumontet C, Corvaïa N (2017) Strategies and challenges for the next generation of antibody–drug conjugates. *Nat. Rev. Drug Discov.* 16:315–337.

[208] (2018) WHO Proposed International Nonproprietary Names (INN) List 120. *WHO Drug Inf.* 32:559–689.

[209] (2019) WHO Recommended International Nonproprietary Names (INN) List 81. *WHO Drug Inf.* 33:59–134.

[210] Poiron C, et al. (2010) IMGT/mAb-DB: the IMGT database for therapeutic monoclonal antibodies. *JOBIM* 13:470b.

[211] Lima WC, et al. (2019) The ABCD database: a repository for chemically defined antibodies. *Nucleic Acids Res.* 48:gkz714.

[212] van Montfort RLM, Workman P (2017) Structure-based drug design: aiming for a perfect fit. *Essays Biochem.* 61(5):431–437.

[213] Muhammed MT, Aki-Yalcin E (2018) Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chem. Biol. Drug. Des.* 93(1):12–20.

[214] Raybould MIJ, et al. (2019) Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci. USA* 116(10):4025–4030.

[215] Benschop RJ, et al. (2019) Development of tibulizumab, a tetravalent bispecific antibody targeting BAFF and IL-17A for the treatment of autoimmune disease. *mAbs* 11(6):1175–1190.

[216] Dunbar J, et al. (2016) SAbPred: a structure-based antibody prediction server. *Nucleic Acids Res.* 44(W1):W474–W478.

[217] Robertson JS, et al. (2019) The INN global nomenclature of biological medicines: A continuous challenge. *Biologicals* 60:15–23.

[218] Sela-Culang I, Kunik V, Ofran Y (2013) The Structural Basis of Antibody-Antigen Recognition. *Front. Immunol.* 4:302.

[219] MacCallum RM, Martin ACRM, Thornton JM (1996) Antibody-antigen Interactions: Contact Analysis and Binding Site Topography. *J. Mol. Biol.* 262(5):732–745.

[220] Krawczyk K, Dunbar J, Deane CM (2017) *Computational Tools for Aiding Rational Antibody Design IN Samish I. (eds) Computational Protein Design. Methods in Molecular Biology.* (Humana Press, New York, NY.) Vol. 1529.

[221] de Kruif J, Boel E, Logtenberg T (1995) Selection and Application of Human Single Chain Fv Antibody Fragments from a Semi-synthetic Phage Antibody Display Library with Designed CDR3 Regions. *J. Mol. Biol.* 248(1):97–105.

[222] Knappik A, et al. (2000) Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J. Mol. Biol.* 296(1):57–86.

[223] Choi Y, Hua C, Sentman CL, Ackerman ME, Bailey-Kellogg C (2015) Antibody humanization by structure-based computational protein design. *mAbs* 7(6):1045–1057.

[224] Ponraj P (2018) Next-generation sequencing may challenge antibody patent claims. *Nature* 557:116.

[225] Rubelt F, et al. (2017) Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nat. Immunol.* 18:1274–1278.

[226] Glanville J, et al. (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.

[227] Eisenstein M (2020) Single-cell RNA-seq analysis software providers scramble to offer solutions. *Nat. Biotechnol.* 38:254–257.

[228] Raybould MIJ, Kovaltsuk A, Marks C, Deane CM (2020) CoV-AbDab: the Coronavirus Antibody Database. *Bioinformatics*.

[229] Raybould MIJ, et al. (2020) Evidence of Antibody Repertoire Functional Convergence through Public Baseline and Shared Response Structures. *bioRxiv*.

[230] Galson JD, et al. (2015) Analysis of B cell repertoire dynamics following hepatitis B vaccination in humans, and enrichment of vaccine-specific antibody sequences. *EBioMedicine* 2(12):2070–2079.

[231] Laseron U, et al. (2014) High-resolution antibody dynamics of vaccine-induced immune responses. *Proc. Natl. Acad. Sci. USA* 111(13):4928–4933.

[232] Wu X, et al. (2011) Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* 333(6049):1593–1602.

[233] Schanz M, et al. (2014) High-throughput sequencing of human immunoglobulin variable regions with subtype identification. *PLoS One* 9(11):e111726.

[234] Zhu J, et al. (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.

[235] DeKosky BJ, et al. (2013) High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* 31:166–169.

[236] López-Santibáñez Jácome L, Avendaño Vásquez SE, Flores-Jasso CF (2019) The Pipeline Repertoire for Ig-seq Analysis. *Front. Immunol.* 10:899.

[237] Hershberg U, Luning Prak ET (2015) The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Philos. Trans. R. Soc. B Biol. Sci.* 370:1676.

[238] Yaari G, Kleinstein SH (2015) Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med.* 7:121.

[239] Ljungars A, et al. (2019) Deep Mining of Complex Antibody Phage Pools Generated by Cell Panning Enables Discovery of Rare Antibodies Binding New Targets and Epitopes. *Front. Immunol.* 10:847.

[240] Yu R, et al. (2009) Neutralizing antibodies of botulinum neurotoxin serotype A screened from a fully synthetic human antibody phage display library. *J. Biomed. Screen.* 14(8):991–998.

[241] Cerosaletti K, et al. (2017) Single-cell RNA-seq reveals expanded clones of islet antigen-reactive CD4+ T cells in peripheral blood of subjects with type 1 diabetes. *J. Immunol.* 199(1):323–335.

[242] Mitsunaga EM, Snyder MP (2019) Characterization of the Human Antibody Response to Natural Infection Using Longitudinal Immune Repertoire Sequencing. *Mol. Cell. Proteom.*

[243] Lin Y (2018) What's happened over the last five years with high-throughput protein crystallization screening? *Expert Opin. Drug Dis.* 13(8):691–695.

[244] Krawczyk K, et al. (2018) Structurally Mapping Antibody Repertoires. *Front. Immunol.* 9:1698.

[245] Schritt D, et al. (2019) Repertoire Builder: high-throughput structural modeling of B and T cell receptors. *Mol. Syst. Des. Eng.* 4:761–768.

[246] Shrake A, Rupley JA (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* 79(2):351–364.

[247] Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23):3150–3152.

[248] Morea V, Tramontano A, Rustici M, Chothia C, Lesk AM (1998) Conformations of the third hypervariable region in the VH domain of immunoglobulins. *J. Mol. Biol.* 275(2):269–294.

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

[250] Richardson E, et al. (2020) A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-Pertussis toxoid antibodies. *bioRxiv*.

[251] Wong WK, et al. (2020) Ab-Ligity: Identifying sequence-dissimilar antibodies that bind to the same epitope. *bioRxiv*.

[252] Mordasini F, et al. (2006) Analysis of the Antibody Response to an Immunodominant Epitope of the Envelope Glycoprotein of a Lentivirus and Its Diagnostic Potential. *Journal of Clinical Microbiology* 44(3):981–991.

[253] Mukherjee S, Tworowski D, Detroja R, Mukherjee SB, Frenkel-Morgenstern M (2020) Immunoinformatics and Structural Analysis for Identification of Immunodominant Epitopes in SARS-CoV-2 as Potential Vaccine Targets. *Vaccines* 8(2):290.

[254] Brouwer PJM, et al. (2020) Potent neutralizing antibodies from COVID-19 patients define multiple targets of vulnerability. *Science* 369(6504):643–650.

[255] (2018) Approved antibodies. Published by The Antibody Society. Available to members at: https://www.antibodysociety.org/news/approved-antibodies/ [Accessed June 12, 2018].

[256] Xu Y, et al. (2013) Addressing polyspecificity of antibodies selected from an in vitro yeast presentation system: A FACS-based, high-throughput selection and analytical tool. *Protein Eng. Des. Sel.* 26(10):663–670.

[257] Datta-Mannan A, et al. (2015) Balancing charge in the complementarity- determining regions of humanized mAbs without affecting pI reduces non-specific binding and improves the pharmacokinetics. *mAbs* 7(3):483–493.

[258] Haberger M, et al. (2014) Assessment of chemical modifications of sites in the CDRs of recombinant antibodies. *mAbs* 6(2):327–339.

[259] Sydow JF, et al. (2014) Structure-based prediction of asparagine and aspartate degradation sites in antibody variable regions. *PLoS ONE* 9(6):e100736.

[260] Petrescu AJ, Milac AL, Petrescu SM, Dwek RA, Wormald MR (2004) Statistical analysis of the protein environment of N-glycosylation sites: Implications for occupancy, structure, and folding. *Glycobiology* 14(2):103–114.

[261] Courtois F, Agrawal NJ, Lauer TM, Trout BL (2016) Rational design of therapeutic mAbs against aggregation through protein engineering and incorporation of glycosylation motifs applied to bevacizumab. *mAbs* 8(1):99–112.

[262] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Del. Rev.* 23(1-3):3–25.

[263] Ye J, Ma N, Madden TL, Ostell JM (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* 41:34–40.

[264] Dunbar J, Deane CM (2015) ANARCI: Antigen receptor numbering and receptor classification. *Bioinformatics* 32(2):298–300.

[265] Kabsch W (1976) A solution for the best rotation to relate two sets of vectors. *Acta Cryst. A* 32(5):922–923.

[266] Sokal RR, Michener CD (1958) A Statistical Method for Evaluating Systematic Relationships. *Univ. Kansas Sci. Bull.* 38:1409–1438.

[267] Zhu ZY, Blundell TL (1996) The use of amino acid patterns of classified helices and strands in secondary structure prediction. *J. Mol. Biol.* 260(2):261–276.

[268] Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157(1):105–132.

[269] Wimley WC, White SH (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.* 3(10):842–848.

[270] Hessa T, et al. (2005) Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* 433(7024):377–381.

[271] Eisenberg D, McLachlan AD (1986) Solvation energy in protein folding and binding. *Nature* 319(6050):199–203.

[272] Black SD, Mould DR (1991) Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. *Anal. Biochem.* 193(1):72–82.

[273] Jain T, et al. (2017) Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl. Acad. Sci. USA* 114(5):944–949.

[274] Tsuchiya Y, Mizuguchi K (2016) The diversity of H3 loops determines the antigen-binding tendencies of antibody CDR loops. *Protein Sci.* 25(4):815–825.

[275] Shi B, et al. (2014) Comparative analysis of human and mouse immunoglobulin variable heavy regions from IMGT / LIGM-DB with IMGT / HighV-QUEST. *Theor. Biol. Med. Model* 11:30.

[276] Reynolds JA, Gilbert DB, Tanford C (1974) Empirical Correlation Between Hydrophobic Free Energy and Aqueous Cavity Surface Area. *Proc. Natl. Acad. Sci. USA* 71(8):2925–2927.

[277] DeKosky BJ, et al. (2016) Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc. Natl. Acad. Sci. USA* 113(19):E2636–E2645.

[278] Li W, et al. (2020) Rapid selection of a human monoclonal antibody that potently neutralizes SARS-CoV-2 in two animal models. *bioRxiv*.

[279] Desautels T, Zemla A, Lau E, Franco M, Faissol D (2020) Rapid in silico design of antibodies targeting SARS-CoV-2 using machine learning and supercomputing. *bioRxiv*.

[280] Rahumatullah A, Yunus MH, Tye GJ, Noordin R (2020) Applications of Recombinant Monoclonal Antibodies against Filarial Antigen Proteins. *Am. J. Trop. Med. Hyg.* 102(3):578–781.

[281] Amici C, Donà MG, Chirullo B, Di Bonito P, Accardi L (2020) Epitope Mapping and Computational Analysis of Anti-HPV16 E6 and E7 Antibodies in Single-Chain Format for Clinical Development as Antitumor Drugs. *Cancers* 12:1803.

[282] Iwasaki A, Yang Y (2020) The potential danger of suboptimal antibody responses in COVID-19. *Nat. Rev.* 20:339–341.

[283] Tay MZ, Poh CM, Rénia L, MacAry PA, Ng LFP (2020) The trinity of COVID-19: immunity, inflammation and intervention. *Nat. Rev. Immunol.* 20:363–374.

[284] Liao M, et al. (2020) Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* 26:842–844.

[285] Wen W, et al. (2020) Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing. *Cell Discov.* 6:31.

[286] Zhao J, et al. (2020) Antibody responses to SARS-CoV-2 in patients of novel coronavirus disease 2019. *Clin. Infect. Dis.*

[287] Pinto D, et al. (2020) Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. *Nature* 583:290–295.

[288] Yuan M, et al. (2020) A highly conserved cryptic epitope in the receptor-binding domains of SARS-CoV-2 and SARS-CoV. *Science* 368(6491):630–633.

[289] Robbiani DF, et al. (2020) Convergent antibody responses to SARS-CoV-2 in convalescent individuals. *Nature* 584:437–442.

[290] Hansen J, et al. (2020) Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. *Science* 369(6506):1010–1014.

[291] Acharya P, et al. (2020) A glycan cluster on the SARS-CoV-2 spike ectodomain is recognized by Fab-dimerized glycan-reactive antibodies. *bioRxiv*.

[292] Cao Y, et al. (2020) Potent neutralizing antibodies against SARS-CoV-2 identified by high-throughput single-cell sequencing of convalescent patients' B cells. *Cell* 182(1):73–86.

[293] Chen Z, et al. (2017) Human neutralizing monoclonal antibody inhibition of Middle East Respiratory Syndrome coronavirus replication in the common marmoset. *J. Inf. Dis.* 215(12):1807–1815.

[294] Cheng M, et al. (2005) Cross-reactivity of antibody against SARS-coronavirus nucleocapsid protein with IL-11. *Biochem. Biophys. Res. Commun.* 338(3):1654–1660.

[295] Chi X, et al. (2020) A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science* 369(6504):650–655.

[296] Choi JH, et al. (2020) Characterization of a human monoclonal antibody generated from a B-cell specific for a prefusion-stabilized spike protein of Middle East respiratory syndrome coronavirus. *PLoS ONE* 15(5):1–19.

[297] Corti D, et al. (2015) Prophylactic and postexposure efficacy of a potent human monoclonal antibody against MERS coronavirus. *Proc. Natl. Acad. Sci. USA* 112(33):10473–10478.

[298] Coughlin M, et al. (2007) Generation and characterization of human monoclonal neutralizing antibodies with distinct binding and sequence features against SARS coronavirus using XenoMouse. *J. Virol.* 361(1):93–102.

[299] Custódio TF, et al. (2020) Selection, biophysical and structural analysis of synthetic nanobodies that effectively neutralize SARS-CoV-2. *bioRxiv*.

[300] Du L, et al. (2014) A conformation-dependent neutralizing monoclonal antibody specifically targeting receptor-binding domain in Middle East respiratory syndrome coronavirus spike protein. *J. Virol.* 88(12):7045–7053.

[301] Du S, et al. (2020) Structures of potent and convergent neutralizing antibodies bound to the SARS-CoV-2 spike unveil a unique epitope responsible for exceptional potency. *bioRxiv*.

[302] Duan J, et al. (2006) A human neutralizing antibody against a conformational epitope shared by oligomeric SARS S1 protein. *Antivir. Ther.* 11(1):117–123.

[303] Duan J, et al. (2005) A human SARS-CoV neutralizing antibody against epitope on S2 protein. *Biochem. Biophys. Res. Commun.* 333(1):186–193.

[304] Esparza TJ, Brody DL (2020) High Affinity Nanobodies Block SARS-CoV-2 Spike Receptor Binding Domain Interaction with Human Angiotensin Converting Enzyme. *bioRxiv*.

[305] Gubbins MJ, et al. (2005) Molecular characterization of a panel of murine monoclonal antibodies specific for the SARS-coronavirus. *Mol. Immunol.* 42(1):125–136.

[306] Hwang WC, et al. (2006) Structural Basis of Neutralization by a Human Anti-severe Acute Respiratory Syndrome Spike Protein Antibody, 80R. *J. Biol. Sci.* 281(45):34610–34616.

[307] Jiang L, et al. (2014) Potent neutralization of MERS-CoV by human neutralizing monoclonal antibodies to the viral spike glycoprotein. *Sci. Transl. Med.* 6(234):234ra59.

[308] Kang X, et al. (2006) Human Neutralizing Fab Molecules against Severe Acute Respiratory Syndrome Coronavirus Generated by Phage Display. *Clin. Vaccine Immunol.* 13(8):953–957.

[309] Kim SI, et al. (2020) Stereotypic Neutralizing VH Clonotypes Against SARS-CoV-2 RBD in COVID-19 Patients and the Healthy Population. *bioRxiv*.

[310] Kreer C, et al. (2020) Longitudinal Isolation of Potent Near-Germline SARS-CoV-2-Neutralizing Antibodies from COVID-19 Patients. *Cell* 182(4):843–854.

[311] Lee YC, et al. (2007) Chicken single-chain variable fragments against the SARS-CoV spike protein. *J. Virol. Methods* 146(1):104–111.

[312] Lee YC, et al. (2007) A dominant antigenic epitope on SARS-CoV spike protein identified by an avian single-chain variable fragment (scFv)-expressing phage. *Vet. Immunol. Immunopathol.* 117(1):75–85.

[313] Liu J, et al. (2006) Production of an Anti-Severe Acute Respiratory Syndrome (SARS) Coronavirus Human Monoclonal Antibody Fab Fragment by Using a Combinatorial Immunoglobulin Gene Library Derived from Patients Who Recovered from SARS. *Clin. Vaccine Immunol.* 13(5):594–597.

[314] Liu L, et al. (2020) Potent Neutralizing Monoclonal Antibodies Directed to Multiple Epitopes on the SARS-CoV-2 Spike. *Nature* 584:450–456.

[315] Liu ZX, et al. (2005) Identification of single-chain antibody fragments specific against SARS-associated coronavirus from phage-displayed antibody library. *Biochem. Biophys. Res. Commun.* 329(2):437–444.

[316] Lv Z, et al. (2020) Structural basis for neutralization of SARS-CoV-2 and SARS-CoV by a potent therapeutic antibody. *Science*.

[317] Niu P, et al. (2018) Ultrapotent human neutralizing antibody repertoires against Middle East respiratory syndrome coronavirus from a recovered patient. *J. Inf. Dis.* 218(8):1249–1260.

[318] Noy-Porat T, et al. (2020) A panel of human neutralizing mAbs targeting SARS-CoV-2 spike at multiple epitopes. *Nat. Commun.* 11:4303.

[319] Pak JE, et al. (2009) Structural Insights into Immune Recognition of the Severe Acute Respiratory Syndrome Coronavirus S Protein Receptor Binding Domain. *J. Mol. Biol.* 388(4).

[320] Pallesen J, et al. (2017) Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. *Proc. Natl. Acad. Sci. USA* 114(35):E7348–E7357.

[321] Pascal KE, et al. (2015) Pre- and postexposure efficacy of fully human antibodies against Spike protein in a novel humanized mouse model of MERS-CoV infection. *Proc. Natl. Acad. Sci. USA* 112(28):8738–8743.

[322] Prabakaran P, et al. (2006) Structure of Severe Acute Respiratory Syndrome Coronavirus Receptor-binding Domain Complexed with Neutralizing Antibody. *J. Biol. Chem.* 281(23):15829–15836.

[323] Raj VS, et al. (2018) Chimeric camel/human heavy-chain antibodies protect against MERS-CoV infection. *Sci. Adv.* 4(8):eaas9667.

[324] Reguera J, et al. (2012) Structural bases of coronavirus attachment to host aminopeptidase N and its inhibition by neutralizing antibodies. *PLoS Pathog.* 8(8):e1002859.

[325] Roberts A, et al. (2006) Therapy with a Severe Acute Respiratory Syndrome–Associated Coronavirus–Neutralizing Human Monoclonal Antibody Reduces Disease Severity and Viral Burden in Golden Syrian Hamsters. *J. Inf. Dis.* 193(5):685–692.

[326] Rockx B, et al. (2008) Structural Basis for Potent Cross-Neutralizing Human Monoclonal Antibody Protection against Lethal Human and Zoonotic Severe Acute Respiratory Syndrome Coronavirus Challenge. *J. Virol.* 82(7):3220–3235.

[327] Rogers TF, et al. (2020) Isolation of potent SARS-CoV-2 neutralizing antibodies and protection from disease in a small animal model. *Science* 369(6506):956–963.

[328] Seydoux E, et al. (2020) Analysis of a SARS-CoV-2-Infected Individual Reveals Development of Potent Neutralizing Antibodies with Limited Somatic Mutation. *Immunity* 53(1):98–105.

[329] Shi R, et al. (2020) A human neutralizing antibody targets the receptor binding site of SARS-CoV-2. *Nature* 584:120–124.

[330] Sui J, et al. (2008) Broadening of Neutralization Activity to Directly Block a Dominant Antibody-Driven SARS-Coronavirus Evolution Pathway. *PLoS Pathog.* 4(11):1–14.

[331] Tang XC, et al. (2014) Identification of human neutralizing antibodies against MERS-CoV and their role in virus adaptive evolution. *Proc. Natl. Acad. Sci. USA* 111(19):E2018–E2026.

[332] ter Meulen J, et al. (2006) Human Monoclonal Antibody Combination against SARS Coronavirus: Synergy and Coverage of Escape Mutants. *PLoS Med.* 3(7):e237.

[333] van den Brink EN, et al. (2005) Molecular and Biological Characterization of Human Monoclonal Antibodies Binding to the Spike and Nucleocapsid Proteins of Severe Acute Respiratory Syndrome Coronavirus. *J. Virol.* 79(3):1635–1644.

[334] Walls AC, et al. (2019) Unexpected receptor functional mimicry elucidates activation of coronavirus fusion. *Cell* 176(5):1026–1039.

[335] Walter JD, et al. (2020) Sybodies targeting the SARS-CoV-2 receptor-binding domain. *bioRxiv*.

[336] Wan J, et al. (2020) Human-IgG-Neutralizing Monoclonal Antibodies Block the SARS-CoV-2 Infection. *Cell Rep.* 32(3):107918.

[337] Wang C, et al. (2020) A human monoclonal antibody blocking SARS-CoV-2 infection. *Nat. Commun.* 11:2251.

[338] Wang L, et al. (2018) Importance of neutralizing monoclonal antibodies targeting multiple antigenic sites on the Middle East respiratory syndrome coronavirus spike glycoprotein to avoid neutralization escape. *J. Virol.* 92(10):e02002–e02017.

[339] Wang L, et al. (2015) Evaluation of candidate vaccine approaches for MERS-CoV. *Nat. Commun.* 6(1):1–11.

[340] Wang N, et al. (2019) Structural Definition of a Neutralization-sensitive Epitope on the MERS-CoV S1-NTD. *Cell Rep.* 28(13):3395–3405.

[341] Wec AZ, et al. (2020) Broad neutralization of SARS-related viruses by human monoclonal antibodies. *Science* 369(6504):731–736.

[342] Wrapp D, et al. (2020) Structural Basis for Potent Neutralization of Betacoronaviruses by Single-domain Camelid Antibodies. *Cell* 181(5):1004–1015.e15.

[343] Wu Y, et al. (2020) A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. *Science* 368(6496):1274–1278.

[344] Ying T, et al. (2014) Exceptionally potent neutralization of Middle East respiratory syndrome coronavirus by human monoclonal antibodies. *J. Virol.* 88(14):7796–7805.

[345] Ying T, et al. (2015) Junctional and allele-specific residues are critical for MERS-CoV neutralization by an exceptionally potent germline-like antibody. *Nat. Commun.* 6(1):1–10.

[346] Zhang S, et al. (2018) Structural definition of a unique neutralization epitope on the receptor-binding domain of MERS-CoV spike glycoprotein. *Cell Rep.* 24(2):441–452.

[347] Zhao A, et al. (2007) Isolation and identification of an scFv antibody against nucleocapsid protein of SARS-CoV. *Microb. Infect.* 9(8):1026–1033.

[348] Zhou D, et al. (2020) Structural basis for the neutralization of SARS-CoV-2 by an antibody from a convalescent patient. *Nat. Struct. Mol. Biol.*

[349] Zhou H, et al. (2019) Structural definition of a neutralization epitope on the N-terminal domain of MERS-CoV spike glycoprotein. *Nat. Commun.* 10(1):1–13.

[350] Zost SJ, et al. (2020) Rapid isolation and profiling of a diverse panel of human monoclonal antibodies targeting the SARS-CoV-2 spike protein. *Nat. Med.*

[351] Coughlin MM, Prabhakar BS (2012) Neutralizing human monoclonal antibodies to severe acute respiratory syndrome coronavirus: target, mechanism of action, and therapeutic potential. *Rev. Med. Virol.* 22:2–17.

[352] Du L, et al. (2016) MERS-CoV spike protein: a key target for antivirals. *Expert Opin. Ther. Targets* 21(2):131–143.

[353] Jiang S, Hillyer C, Du L (2020) Neutralizing Antibodies against SARS-CoV-2 and Other Human Coronaviruses. *Trends Immunol.* 41(5):355–359.

[354] Shanmugaraj B, Siriwattananon K, Wangkanont K, Phoolcharoen W (2020) Perspectives on monoclonal antibody therapy as potential therapeutic intervention for Coronavirus disease-19 (COVID-19). *Asian Pac. J. Allergy* 38:10–18.

[355] Zhou Y, Yang Y, Huang J, Jiang S, Du L (2019) Advances in MERS-CoV Vaccines and Therapeutics Based on the Receptor-Binding Domain. *Viruses* 11(1):60.

[356] Ju B, et al. (2020) Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature* 584:115–119.

[357] Wu NC, et al. (2020) An alternative binding mode of IGHV3-53 antibodies to the SARS-CoV-2 receptor binding domain. *bioRxiv.*

[358] Yuan M, et al. (2020) Structural basis of a public antibody response to SARS-CoV-2. *Science* 369(6507):1119–1123.

[359] Kuri-Cervantes L, et al. (2020) Comprehensive mapping of immune perturbations associated with severe COVID-19. *Sci. Immunol.* 5(49).

[360] Schultheiß C, et al. (2020) Next-Generation Sequencing of T and B Cell Receptor Repertoires from COVID-19 Patients Showed Signatures Associated with Severity of Disease. *Immunity* 53(2):9442–455.

[361] Montague Z, et al. (2020) Dynamics of B-cell repertoires and emergence of cross-reactive responses in COVID-19 patients with different disease severity. *medRxiv.*

[362] Cock PJA, et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11):1422–1423.

[363] Zhang Y, et al. (2020) SARS-Cov-2-, HIV-1-, Ebola-neutralizing and anti-PD1 clones are predisposed. *bioRxiv.*

[364] Sewell HF, Agius RM, Stewart M, Kendrick D (2020) Cellular immune responses to COVID-19. *BMJ* 370.

[365] Vabret N, et al. (2020) Immunology of COVID-19: Current State of the Science. *Immunity* 52(6):910–941.

[366] Rodda LB, et al. (2020) Functional SARS-CoV-2-specific immune memory persists after mild COVID-19. *medRxiv*.

[367] Arvin AM, et al. (2020) A perspective on potential antibody-dependent enhancement of SARS-CoV-2. *Nature* 584:353–363.

[368] Halstead SB, Katzelnick L (2020) COVID 19 Vaccines: Should we fear ADE? *J. Inf. Dis.*

[369] Fierz W, Walz B (2020) Antibody Dependent Enhancement Due to Original Antigenic Sin and the Development of SARS. *Front. Immunol.* 11:1120.

[370] Long QX, et al. (2020) Antibody responses to SARS-CoV-2 in patients with COVID-19. *Nat. Med.* 26:845–848.

[371] Kowitdamrong E, et al. (2020) Antibody Responses to SARS-CoV-2 in Coronavirus Diseases 2019 Patients with Different Severity. *medRxiv*.

Appendix A

Supplementary Methods, Tables, and Figures

Chapter 2

CDR-H3 length	Number of possible sequences	Length frequency in CSTs	Perfect matches	Probability of finding a single sequence in 960m samples
5	3,200,000	3	3	1
6	64,000,000	5	3	1
7	1,280,000,000	2	0	0.75
8	25,600,000,000	9	3	0.0375
9	512,000,000,000	22	14	0.00187
10	10,240,000,000,000	21	7	9.375e-05
11	204,800,000,000,000	29	6	4.6875e-06
12	4,096,000,000,000,000	43	7	2.34375e-07
13	81,920,000,000,000,000	28	5	1.171875e-08
14	1,638,400,000,000,000,000	23	3	5.859375e-10
15	32,768,000,000,000,000,000	18	1	2.9296875e-11
16	655,360,000,000,000,000,000	14	1	1.46484375e-12
17	13,107,200,000,000,000,000,000	5	1	7.32421875e-14
18	262,144,000,000,000,000,000,000	7	0	3.662109375e-15
19	5,242,880,000,000,000,000,000,000	8	0	1.8310546875e-16
20	104,857,600,000,000,000,000,000,000	3	0	9.1552734375e-18
23	838,860,800,000,000,000,000,000,000,000	2	0	1.14440917969e-21

Table A2.1. The first two columns show the theoretical number of amino acid sequences for each CDR length, assuming any amino acid can be chosen at any position. The next two columns compare the number of times a tested therapeutic had a CDR-H3 of that length, along with the number of therapeutic CDR-H3s that recorded a perfect match to a CDR-H3 in the Observed Antibody Space database. For context, the final column shows the probability of observing one perfect match at each length, given 960 million random samples of the theoretical sequence space.

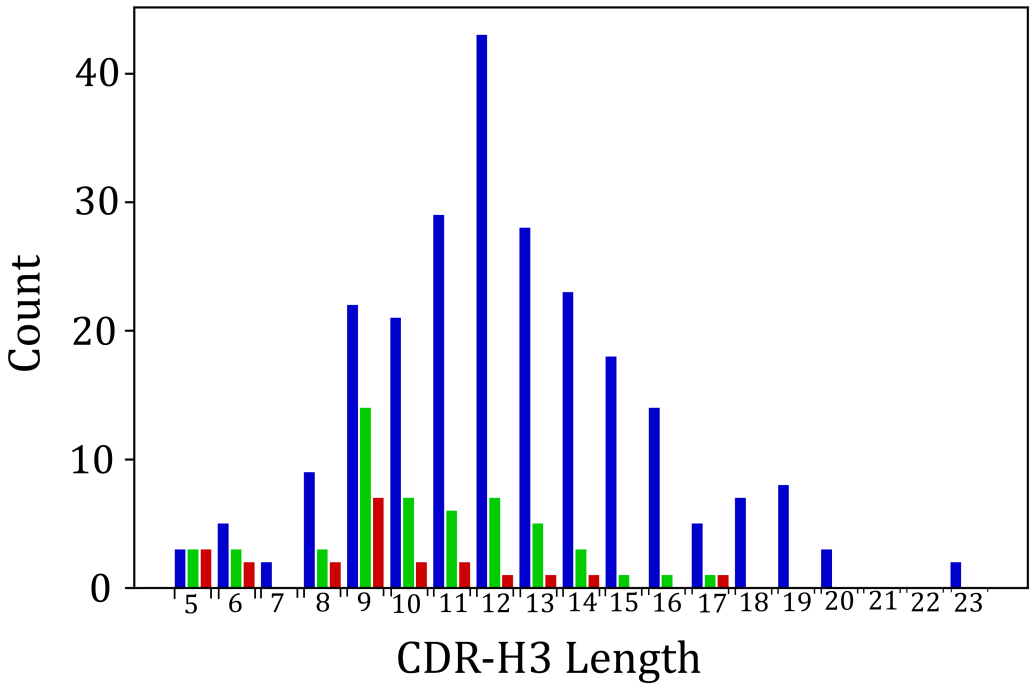


Figure A2.1. Bar charts showing (blue) the distribution of CDR lengths across all 242 tested therapeutics, (green) the distribution of CDR lengths across the 54 therapeutics with a perfect match to OAS, and (red) the distribution of CDR lengths across the 22 therapeutics with a perfect match to OAS with and without the Briney *et al.* dataset.

Chapter 3

Dataset	All VH	All VL	Modellable VH [90% SIC]	Modellable VL [90% SIC]	Predicted Modellable Fvs	Distinct Structures
1 (S64)	177,603	123,934	10,087	6,779	6,420,211	209,394
2 (S57)	169,805	118,020	9,860	7,922	7,225,630	201,039
3 (S5)	159,544	139,845	8,999	8,526	6,827,419	200,708
4 (S56)	162,446	136,874	9,309	7,168	6,628,683	195,061
5 (S83)	152,299	112,733	9,048	8,076	6,170,373	193,384
6 (S67)	173,722	120,237	9,349	6,424	5,544,952	193,061
7 (S84)	164,017	138,874	8,702	8,232	5,634,598	191,617
8 (S76)	148,180	126,713	8,778	7,047	5,856,150	191,162
9 (S54)	121,993	133,921	7,581	9,066	5,074,822	181,290
10 (S89)	152,710	144,340	8,923	9,293	5,414,820	177,829
11 (S13)	127,321	134,485	7,276	8,654	5,314,377	174,105
12 (S93)	110,676	110,904	7,260	7,528	4,799,497	173,006
13 (S87)	120,424	104,440	7,569	6,145	4,043,317	172,200
14 (S86)	156,096	132,580	8,411	7,475	5,130,237	171,940
15 (S10)	109,552	134,816	6,816	8,351	5,152,331	167,153
16 (S50)	157,428	109,437	8,614	5,533	4,556,841	162,663
17 (S75)	105,099	119,470	6,576	7,711	4,174,078	156,510
18 (S8)	150,763	112,479	8,241	6,604	4,305,148	156,044
19 (S37)	137,951	103,825	7,815	5,983	3,565,942	156,034
20 (S59)	111,842	125,621	6,598	8,117	4,807,933	150,425
21 (S22)	113,023	145,227	6,220	8,781	4,108,518	149,575
22 (S77)	118,326	114,384	7,407	7,195	3,821,870	148,950
23 (S58)	114,817	101,696	7,663	7,088	4,388,190	148,753
24 (S30)	112,518	107,459	6,383	6,842	3,667,996	148,691
25 (S27)	125,328	122,194	7,347	7,305	3,989,334	147,654
26 (S74)	119,371	104,430	6,375	5,753	2,961,937	146,049
27 (S12)	117,699	127,682	6,762	6,860	3,388,406	144,916
28 (S19)	111,012	102,905	6,595	6,064	3,471,725	143,027
29 (S52)	109,826	113,067	6,450	5,760	3,397,333	142,110
30 (S60)	112,802	108,429	6,513	5,790	3,310,781	140,525
31 (S11)	102,847	114,304	5,610	6,873	3,495,280	139,094
32 (S51)	117,156	122,970	6,187	6,652	3,828,676	139,024
33 (S25)	108,425	101,548	6,159	6,125	3,176,321	137,503
34 (S34)	114,748	174,784	6,195	11,937	6,404,104	136,973
35 (S29)	100,214	143,027	5,369	9,501	3,819,414	136,811
36 (S72)	118,995	111,611	6,360	6,152	2,627,255	136,161
37 (S91)	114,196	118,422	6,175	8,304	4,256,118	132,851
38 (S65)	134,646	124,116	7,008	8,584	3,446,622	108,021
39 (S95)	118,576	162,377	5,412	11,748	5,901,443	91,855
40 (S17)	102,405	111,669	5,310	7,945	2,690,081	91,229
41 (S4)	100,689	128,986	4,688	1,761	745,977	78,588
Overall					*183,544,740*	

Table A3.1. Structurally profiling the baseline repertoire snapshots of 41 unrelated individuals. In order, the columns show: the dataset label, the number of VH and VL reads within each snapshot, the number of FREAD-modellable VH and VL reads (once clustered at 90% sequence identity), the number of predicted modellable Fvs resulting from these VH-VL pairings, and the number of distinct structures (cluster centres) identified in each dataset. SIC = Sequence Identity Clustered.

# of Repertoires (Dataset Added)	Modellable Fvs Added	Cumulative Public & Private Distinct Structures	Public Distinct Structures (Overall % Public)
1 (S64)	6,420,211	209,394	209,394
2 (+S57)	7,225,630	340,915	100,824 (29.57)
3 (+S5)	6,827,419	445,045	71,743 (16.12)
4 (+S56)	6,628,683	527,668	58,043 (11.00)
5 (+S83)	6,170,373	604,124	48,703 (8.06)
6 (+S67)	5,544,952	670,833	42,277 (6.30)
7 (+S84)	5,624,598	734,374	37,151 (5.06)
8 (+S76)	5,856,150	793,831	33,572 (4.23)
9 (+S54)	5,074,822	846,670	30,474 (3.60)
10 (+S89)	5,414,820	896,328	27,389 (3.06)
11 (+S13)	5,314,377	940,957	25,621 (2.72)
12 (+S93)	4,799,497	980,905	24,015 (2.45)
13 (+S87)	4,043,317	1,023,105	22,052 (2.16)
14 (+S86)	5,130,237	1,061,003	20,867 (1.97)
15 (+S10)	5,152,331	1,100,394	19,468 (1.77)
16 (+S50)	4,556,841	1,130,974	18,421 (1.63)
17 (+S75)	4,174,078	1,161,111	17,157 (1.48)
18 (+S8)	4,305,148	1,188,715	16,071 (1.35)
19 (+S37)	3,565,942	1,218,071	15,302 (1.26)
20 (+S59)	4,807,933	1,243,044	14,669 (1.18)
21 (+S22)	4,108,518	1,269,972	13,992 (1.11)
22 (+S77)	3,821,870	1,294,338	13,380 (1.03)
23 (+S58)	4,388,190	1,316,084	12,953 (0.98)
24 (+S30)	3,667,996	1,342,141	12,542 (0.93)
25 (+S27)	3,989,334	1,365,687	12,017 (0.88)
26 (+S74)	2,961,937	1,387,177	11,482 (0.83)
27 (+S12)	3,388,406	1,406,918	11,050 (0.79)
28 (+S19)	3,471,725	1,426,572	10,732 (0.75)
29 (+S52)	3,397,333	1,446,838	10,319 (0.71)
30 (+S60)	3,310,781	1,466,071	10,078 (0.69)
31 (+S11)	3,495,280	1,486,113	9,714 (0.65)
32 (+S51)	3,828,676	1,504,633	9,478 (0.63)
33 (+S25)	3,176,321	1,522,848	9,141 (0.60)
34 (+S34)	6,404,104	1,544,438	8,615 (0.56)
35 (+S29)	3,819,414	1,564,003	8,226 (0.53)
36 (+S72)	2,627,255	1,581,699	7,966 (0.50)
37 (+S91)	4,256,118	1,598,785	7,818 (0.49)
38 (+S65)	3,446,622	1,614,661	6,891 (0.43)
39 (+S95)	5,901,443	1,629,262	5,935 (0.41)
40 (+S17)	2,690,081	1,642,531	5,110 (0.31)
41 (+S4)	745,977	1,650,922	4,573 (0.28)

Table A3.2. Evaluating the number of public distinct structures seen across multiple baseline repertoire snapshots. In order, the columns show: the number of repertoires compared (in brackets the identifier of the last dataset added), the number of predicted modellable variable domains (Fvs) added by the last dataset, the number of distinct structures added by the last dataset, the (cumulative) number of public and private distinct structures across all compared repertoires, and the number of proportion of these structures that are public. The sharp drop-off in the proportion of public structures in the final four repertoire snapshots can be rationalised by their substantially lower internal structural diversity (see Table A3.1).

Dataset	Public Clonotypes (Briney Def'n, 100% seqID)	Public Clonotypes (Soto Def'n, 80% seqID)
Reference Clonotypes (S64)	70,255 [100%]	68,672 [100%]
Shared with S57	11 [< 0.01%]	1,648 [1.30%]
And S5	0 [0%]	358 [0.14%]
And S56	0 [0%]	194 [0.04%]
And S83	0 [0%]	78 [< 0.01%]
And S67	0 [0%]	49 [< 0.01%]
And S84	0 [0%]	23 [< 0.01%]
And S76	0 [0%]	18 [< 0.01%]
And S54	0 [0%]	11 [< 0.01%]
And S89	0 [0%]	8 [< 0.01%]

Table A3.3. Tracking the number of public clonotypes shared across all naive baseline datasets analysed up to that point (e.g. 358 clonotypes are present in S64, S57, and S5 according to the Soto V3J definition). The 'Briney Definition' clusters VHs with the same V/J genes and 100% CDRH3 sequence identity, while the 'Soto Definition' clusters VHs with the same V/J genes and within 80% sequence identity.

# of Repertoires (Dataset Added)	Cumulative Modellable Fvs	Cumulative Distinct Structures	Expected Cumulative Distinct Structures
1 (S64)	6,420,211	209,394	209,394
2 (+S57)	13,645,841	340,915	445,057
3 (+S5)	20,473,260	445,045	667,732
4 (+S56)	27,101,943	527,668	883,925
5 (+S83)	33,272,316	604,124	1,085,170
6 (+S67)	38,817,268	670,833	1,266,018
7 (+S84)	44,451,866	734,374	1,449,463
8 (+S76)	50,308,016	793,831	1,640,461
9 (+S54)	55,382,838	846,670	1,805,975
10 (+S89)	60,797,658	896,328	1,982,578
11 (+S13)	66,112,035	940,957	2,156,232
12 (+S93)	70,911,532	980,905	2,312,767
13 (+S87)	74,954,849	1,023,105	2,444,639
14 (+S86)	80,085,086	1,061,003	2,611,960
15 (+S10)	85,237,417	1,100,394	2,780,003
16 (+S50)	89,794,258	1,130,974	2,928,623
17 (+S75)	93,968,336	1,161,111	3,064,760
18 (+S8)	98,273,484	1,188,715	3,205,172
19 (+S37)	101,839,426	1,218,071	3,321,474
20 (+S59)	106,647,359	1,243,044	3,478,284
21 (+S22)	110,755,877	1,269,972	3,612,283
22 (+S77)	114,577,747	1,294,338	3,736,932
23 (+S58)	118,965,937	1,316,084	3,880,052
24 (+S30)	122,633,933	1,342,141	3,999,683
25 (+S27)	126,623,267	1,365,687	4,129,795
26 (+S74)	129,585,204	1,387,177	4,226,398
27 (+S12)	132,973,610	1,406,918	4,336,910
28 (+S19)	136,445,335	1,426,572	4,450,139
29 (+S52)	139,842,668	1,446,838	4,560,943
30 (+S60)	143,153,449	1,466,071	4,668,923
31 (+S11)	146,648,729	1,486,113	4,782,921
32 (+S51)	150,477,405	1,504,633	4,907,793
33 (+S25)	153,653,726	1,522,848	5,011,389
34 (+S34)	160,057,830	1,544,438	5,220,257
35 (+S29)	163,877,244	1,564,003	5,344,827
36 (+S72)	166,504,499	1,581,699	5,430,514
37 (+S91)	170,760,617	1,598,785	5,569,326
38 (+S65)	174,207,239	1,614,661	5,681,737
39 (+S95)	180,108,682	1,629,262	5,874,212
40 (+S17)	182,798,763	1,642,531	5,961,948
41 (+S4)	183,544,740	1,650,922	5,986,278

Table A3.4. Tracking the total number of public and private distinct structures seen across multiple baseline repertoire snapshots. In order, the columns show: the number of repertoires compared (in brackets the identifier of the last dataset added), the cumulative number of predicted modellable Fvs, the number of public and private distinct structures seen across all compared repertoires, and the expected number of cumulative public and private distinct structures if new distinct structures were observed at the same rate per modellable Fv as seen in S64.

CDR	Templates/'Public Baseline' Structure (Median)
CDRH1	3
CDRH2	4
CDRH3	2
CDRL1	5
CDRL2	8
CDRL3	7

Table A3.5. The median numbers of unique FREAD templates assigned to each CDR within a 'Public Baseline' distinct structure.

V Gene	J Gene	Count
V5-51	J4	83
V5-51	J3	18
V5-51	J6	3
V5-51	J5	2
V1-8	J5	9
V1-8	J6	3
V5-10-1	J4	8
V5-10-1	J5	1
V1-2	J4	7
V1-2	J5	1
V1-3	J4	4
V1-46	J3	1
V1-69-2	J4	1

Table A3.6. The diversity of IGHV/IGHJ gene combinations represented across the 141 VH clonotypes assigned by Repertoire Structural Profiling to the H14012+L14649 'Public Baseline' distinct structure.

V Gene	J Gene	CDRH3 (individual)	Overall Occupancy
V5-51	J4	**ARPYGSGSYSDY** (S64) ARHDGSGSYSDY (S54) ARQYVSGSYSDY (S76)	3
V5-51	J4	**ARQGYGDYVTDY** (S67) ARQDYGDYVVDY (S76)	3
V5-51	J6	**ARLGCYYYGMDV** (S5) ARQGYYYYGMDV (S5)	2
V5-51	J4	**ARQSSNWNGVDY** (S89) ARQSSNWNGGDY (S89)	2
V1-69-2	J4	ATSRDGYNNLDY (S5) ATARDGYNKLDY (S5)	2
V5-51	J4	**AGQDDYGDYVDY** (S89) ARQGDYGDYVDY (S89)	2
V5-51	J4	**AVGGGWYAVGDY** (S5) AVGGGWYGGGDY (S5)	2
V5-51	J4	**ARMGARPGYFDY** (S89) ARPGEDGLEFDY (S89)	2
V5-51	J4	**AREETIARASDY** (S76)	2
V5-51	J4	**ARPDYAPSGIDY** (S64)	2
V5-51	J4	**ARLKKKENWFDP** (S67)	2
V5-51	J4	**ARRPMSYPEFDY** (S67)	2

Table A3.7. The 12 multiple-occupancy VH clonotypes assigned by Repertoire Structural Profiling to the H14012+L14649 'Public Baseline' distinct structure.

Dataset	Public Clonotypes, Briney Def'n (% Public)	Public Clonotypes, Soto Def'n (% Public)
V1 Before (Reference Clonotypes)	48,459 [100%]	43,166 [100%]
Shared with V2 Before	310 [0.32%]	262 [0.34%]
And V3 Before	38 [0.03%]	32 [0.03%]
V1 After (Reference Clonotypes)	86,837 [100%]	75,927 [100%]
Shared with V2 After	444 [0.27%]	1,508 [1.02%]
And V3 After	33 [0.02%]	272 [0.13%]

Table A3.8. Tracking the number of public clonotypes shared across all 'Before Vaccination' (Before) datasets and all 1After Vaccination' (After) analysed up to that point (e.g. 272 clonotypes are public across V1, V2, and V3 After Vaccination according to the Soto V3J definition). The 'Briney Definition' clusters VHs with the same V/J genes and 100% CDRH3 sequence identity, while the 'Soto Definition' clusters VHs with the same V/J genes and within 80% sequence identity.

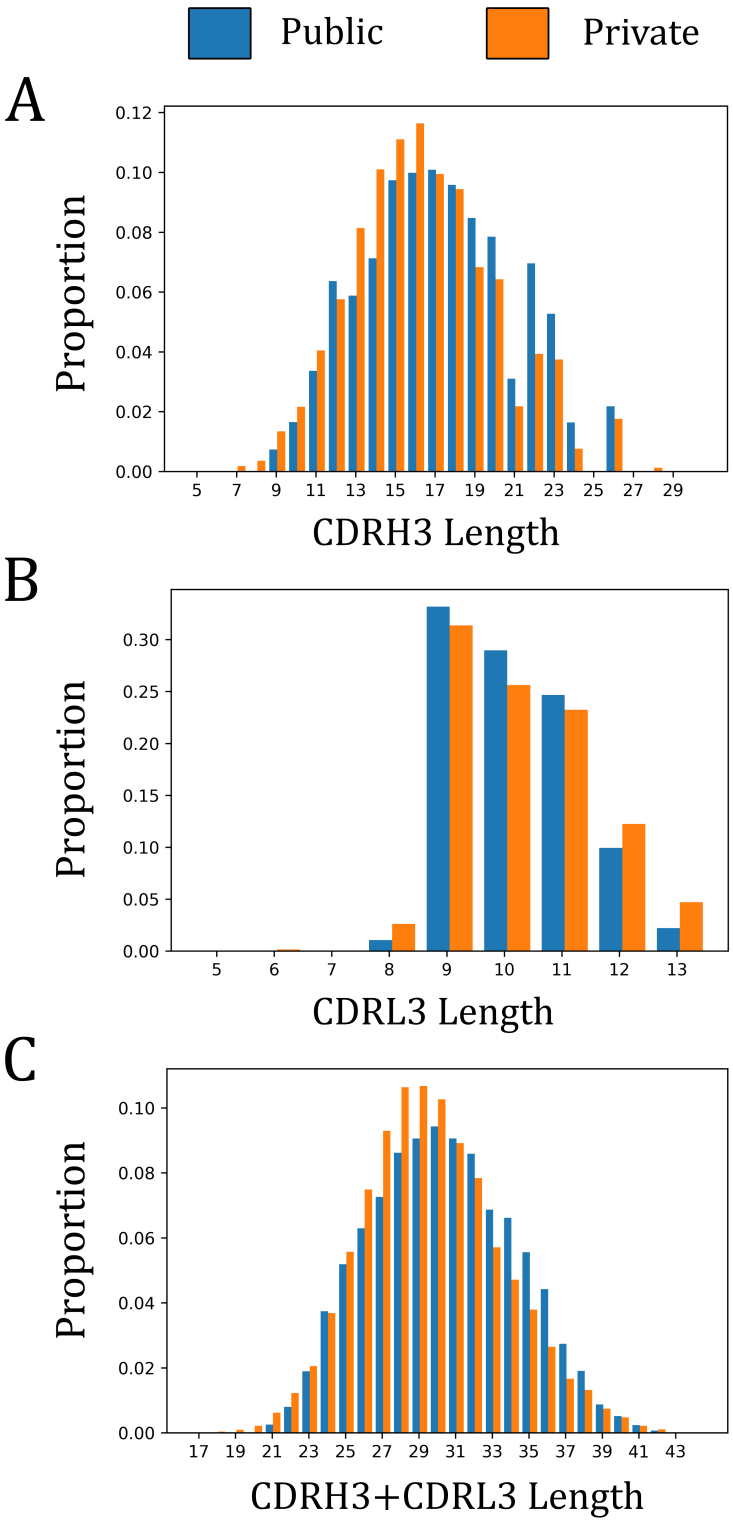


Figure A3.1. Bar charts comparing the (A) CDRH3 lengths, (B) CDRL3 lengths, and (C) Combination of CDRH3+CDRL3 lengths of S64 sequences assigned to 'Public Baseline' (blue) and 'Private Baseline' structures (orange).

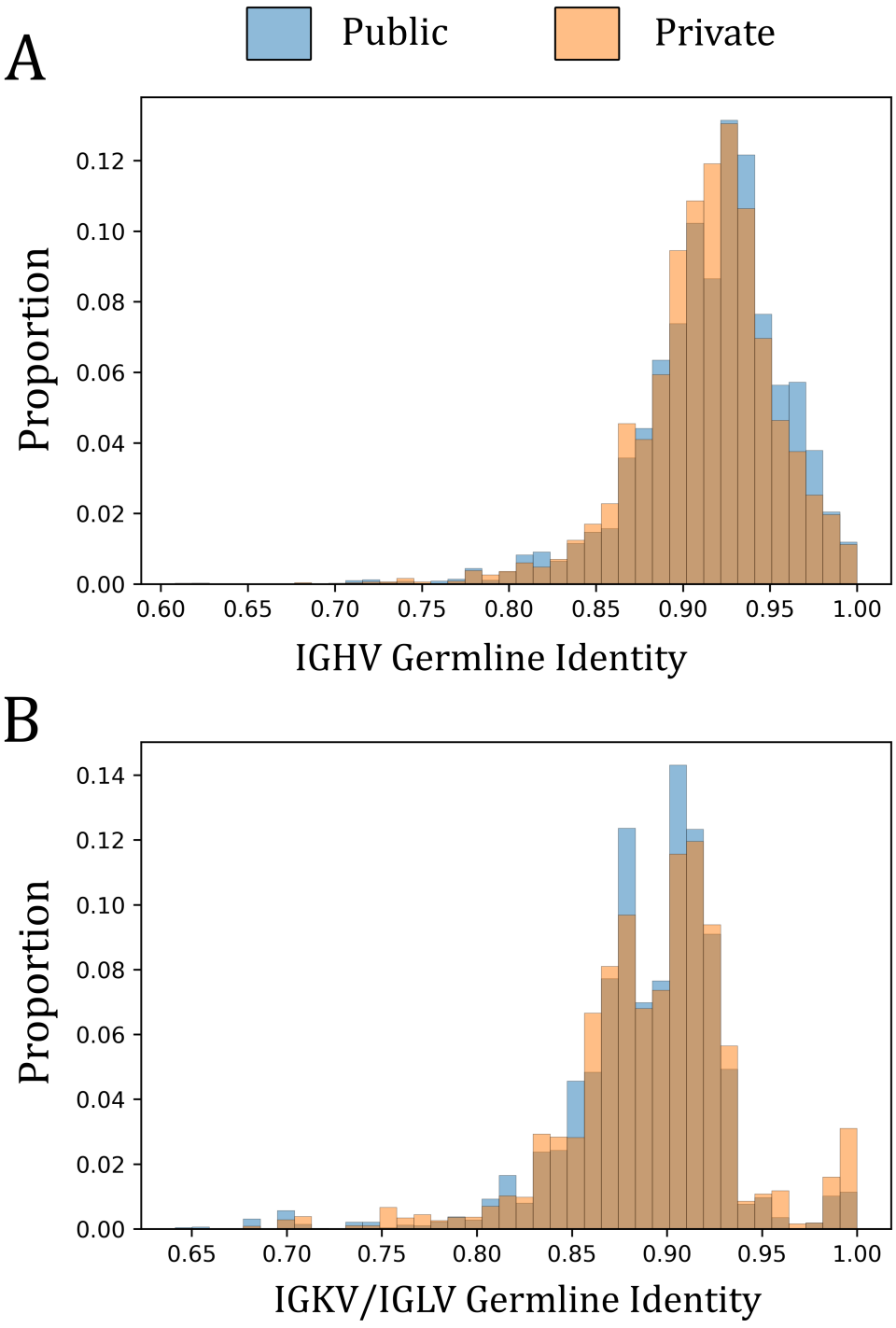


Figure A3.2. Histograms comparing the (A) closest IGHV germline sequence identity, and (B) closest IGKV/IGLV germline sequence identity of S64 sequences assigned to 'Public Baseline' structures (blue) against those assigned to 'Private Baseline' structures (orange).

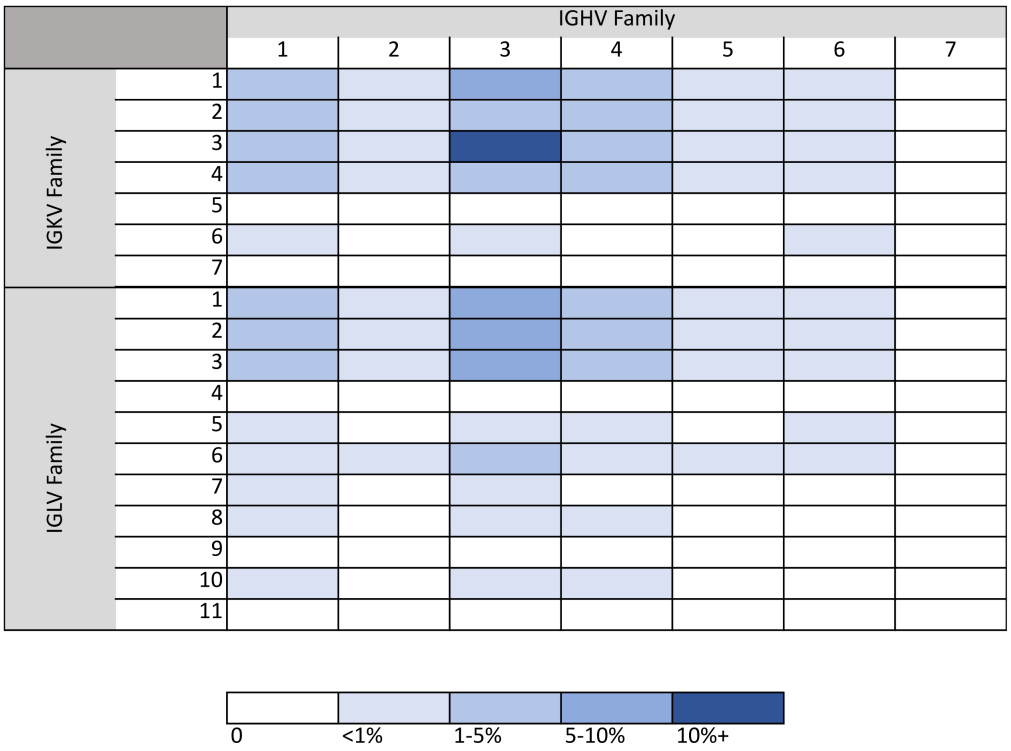


Figure A3.3. A heatmap showing IGHV:IGKV/IGLV gene family pairings across the 'Public Baseline' structures. The usage trends are consistent with the natural pairings observed in DeKosky *et al.* (10.1038/nbt.2492).

Chapter 4

Human UCB Ig-seq Techniques

1. Commercially Sourced RNA Samples. Three different total RNA samples at 1mug/mul were sourced from Clontech Laboratories comprising 50mug prepared from normal human spleen (SP), pooled from 12 male/female Caucasians aged between 18 and 54 years, 10mug from normal bone marrow (BM) from 56 Asian males aged between 22 and 85 and 10mug from normal human peripheral leukocytes (PBL) from 426 male/female Asians aged between 18 and 54 (PBL).

2. Reverse Transcription and C-region Specific Oligonucleotides. Multiple reverse transcription (RT) reactions were done for each RNA source (SP: 12xVH, 6xVK, 6xVL BM: 6xVH, 2xVK, 2xVL PBL: 6xVH, 2xVK, 2xVL) using human antibody constant region reverse oligonucleotides specific for the 3' ends of the CH1 gene of IgM and the 3' ends of the C-kappa and C-lambda genes. After an initial denaturation at 65°C for 5 min in the presence of oligonucleotide and dNTPs multiple 20mul reactions each containing 1mug RNA, 200U Superscript III (Life Technologies), 20U RNasin (Promega), 5mM DTT, 500muM dNTPs and 1muM oligonucleotide were incubated for 60 min at 50°C and 15 min at 70°C before being frozen at -20°C.

3. Primary PCRs. Twelve family-restricted primary PCRs (5xVH, 4xVK, 3xVL families) were done on each of the three cDNA template samples SP, BM and PBL. Where a V-gene family required more than one oligonucleotide these were mixed in proportions equivalent to estimates of sequence frequency. A total of 304 individual (192xVH, 64xVK and 48xVL) 25μl buffered reactions, each with 1μl of cDNA as template, consisted of 1mM dNTPs, 1.5mM MgSO4, 4μM forward and 4μM reverse oligonucleotides and 0.5U KOD hot start DNA polymerase (Merck Millipore). After an initial denaturation step of 96°C for 2 min, PCR cycling conditions for all reactions were 96°C for 15s, 55°C for 15s, 72°C for 15s for 40 cycles followed by a final extension step for 5 min at 72°C.

4. Secondary PCRs. An equivalent 304 (192xVH, 64xVK, 48xVL) 50μl individual secondary PCRs were done, keeping the DNA samples from the primary reactions separate in order to maximise V-gene diversity. The reactions, each with 2μl of primary PCR as template, had matched components and cycling conditions to the primary reactions except the cycle number was reduced to 30. Once again V-region family oligonucleotide sets were kept separate and members within each family were mixed at the pre-determined proportions.

5. Sample Preparation for the Oxford Sequencing Centre. The secondary PCR products for each of the specific V-gene family (VH1-6, VK1-4, VL1-2) were pooled, giving 12 samples. Approximately 1μg from each pool was analysed by agarose gel electrophoresis (Invitrogen UltraPureTM agarose) and the DNA of approximately 400bp was excised, gel extracted (Qiagen) and eluted into 50μl of water at a final concentrations of between 10-75ng/μl to be analysed by paired-end next generation sequencing on an Illumina MiSeq machine at the Oxford Genomics Centre (OGC) at the Wellcome Trust Centre for Human Genetics, Oxford.

Therapeutic	Unbound/Bound (PDB Code)	Framework (Å)	CDRH1 (Å)	CDRH2 (Å)	CDRH3 (Å)	CDRL1 (Å)	CDRL2 (Å)	CDRL3 (Å)
1. Abituzumab	Bound (4O02)	1.357	1.101	1.494	2.516	0.919	1.307	1.143
2. Adalimumab	Unbound (4NYL)	0.711	0.646	0.539	2.871	0.922	0.435	0.932
3. Alemtuzumab	Unbound (1BEY)	0.701	1.766	1.143	3.086	0.822	1.011	1.122
4. Anifrolumab	Unbound (4QXG)	0.575	0.674	0.864	1.492	2.305	0.746	1.275
5. Atezolizumab	Bound (5XXY)	0.734	1.146	1.785	3.403	1.481	0.552	1.451
6. Bapineuzumab	Bound (4OJF)	1.357	1.101	1.494	2.516	0.919	1.307	1.143
7. Basiliximab	Unbound (1MIM)	0.761	0.71	0.739	1.067	0.593	0.428	1.127
8. Belimumab	Unbound (5Y9K)	0.714	3.499	1.012	3.940	4.354	0.809	2.451
9. Bimagrumab	Unbound (5NHW)	0.819	1.842	0.858	0.553	1.027	0.606	4.669
10. Briakinumab	Unbound (5N2K)	0.756	0.752	0.529	2.645	1.655	0.686	5.077
11. Canakinumab	Unbound (4G5Z)	0.824	0.789	0.784	2.001	1.296	0.712	0.990
12. Carlumab	Unbound (4DN3)	0.692	2.816	2.702	1.780	1.707	0.416	1.640
13. Certolizumab	Unbound (5WUV)	0.578	1.215	1.253	3.648	0.412	0.195	2.362
14. Cetuximab	Unbound (1YY8)	0.810	0.251	0.302	0.348	0.236	0.139	0.401
15. Crenezumab	Unbound (5KMV)	0.658	0.829	0.618	0.733	1.127	0.579	0.405
16. Daclizumab	Unbound (3NFS)	0.682	0.672	0.925	1.310	1.267	0.683	0.849
17. Drozitumab	Bound (4OD2)	0.840	0.530	1.134	4.999	2.073	2.118	2.265
18. Eculizumab	Bound (5I5K)	1.752	1.028	1.077	2.006	1.114	0.589	1.345
19. Efalizumab	Unbound (3EO9)	0.740	1.049	2.069	2.038	0.617	0.278	1.123
20. Epratuzumab	Unbound (5VKK)	-	-	2.260	-	0.995	0.496	1.651
21. Fresolimumab	Unbound (3EO0)	0.652	3.679	1.193	8.882	1.829	0.173	0.729
22. Gantenerumab	Bound (5CSZ)	0.535	0.679	2.712	3.197	2.165	0.418	1.112
23. Gevokizumab	Unbound (4G6K)	1.014	2.592	1.219	1.332	0.917	0.447	0.932
24. Guselkumab	Unbound (4M6N)	1.290	0.431	0.485	3.180	1.371	0.518	1.490
25. Ibalizumab	Bound (3O2D)	1.132	0.828	1.627	3.933	0.957	0.409	0.920
26. Infliximab	Unbound (5VH3)	0.640	1.049	1.379	2.732	0.636	0.176	0.330
27. Ipilimumab	Bound (5TRU)	0.568	0.649	0.816	2.099	0.865	0.279	0.870
28. Lampalizumab	Bound (4D9Q)	1.008	0.387	0.549	3.792	0.547	0.662	0.873
29. Lebrikizumab	Bound (4I77)	0.503	1.575	1.663	2.006	0.645	0.464	1.051
30. Matuzumab	Unbound (3C08)	1.287	-	1.255	9.966	-	0.547	2.780
31. Motavizumab	Bound (3QWO)	1.231	3.299	1.684	3.827	1.949	0.504	3.314
32. Muromonab	Bound (1SY6)	0.193	0.441	0.553	2.101	0.469	0.333	1.119
33. Natalizumab	Bound (4IRZ)	0.838	3.180	0.874	4.652	1.161	0.512	1.476
34. Necitumumab	Bound (3B2U)	0.696	1.472	1.372	5.869	0.648	0.733	0.744
35. Nimotuzumab	Unbound (3GKW)	1.846	4.046	1.737	7.573	1.049	0.424	1.291
36. Nivolumab	Unbound (5GGQ)	0.847	0.597	0.851	2.393	3.280	0.390	0.851
37. Obinutuzumab	Unbound (3PP3)	0.792	0.746	0.368	5.147	2.972	0.726	0.804
38. Ofatumumab	Unbound (3GIZ)	0.909	0.584	2.226	3.581	0.410	0.292	1.119
39. Olokizumab	Bound (4CNI)	1.185	0.518	0.756	1.271	0.348	0.523	0.898
40. Omalizumab	Unbound (4X7S)	1.180	1.760	0.896	3.177	0.807	0.341	1.287
41. Onartuzumab	Bound (4K3J)	1.228	1.655	1.732	3.016	0.937	0.410	1.028
42. Palivizumab	Unbound (2HWZ)	0.645	0.766	0.392	1.573	0.672	0.528	2.111
43. Panitumumab	Bound (5SX4)	0.910	1.312	1.567	1.645	0.273	0.781	0.506
44. Pembrolizumab	Unbound (5DK3)	0.755	0.424	0.466	3.536	1.175	0.553	1.099
45. Pertuzumab	Bound (5JXE)	0.970	3.007	1.696	2.920	2.053	0.383	0.985
46. Pinatuzumab	Bound (6AND)	0.909	1.643	3.983	3.554	2.754	0.422	0.441
47. Ponezumab	Bound (3U0T)	1.438	1.837	0.753	4.425	1.882	0.339	0.758
48. Ramucirumab	Unbound (3S34)	0.718	0.419	0.962	0.910	0.370	0.271	0.576
49. Ranibizumab	Bound (1CZ8)	0.559	1.085	1.245	4.065	0.546	0.702	0.936
50. Rituximab	Unbound (4KAQ)	0.942	0.359	0.804	4.815	0.477	0.373	0.693
51. Sifalimumab	Bound (4YPG)	0.790	0.782	0.633	3.819	3.670	1.055	2.911
52. Tanezumab	Bound (4EDW)	0.919	1.828	0.898	2.065	0.352	0.792	0.594
53. Tralokinumab	Bound (5L6Y)	0.403	0.698	0.630	4.073	1.054	0.271	2.526
54. Trastuzumab	Bound (4HKZ)	0.595	0.449	0.814	3.226	0.278	0.192	0.537
55. Tremelimumab	Unbound (5GGU)	0.696	0.526	0.700	4.912	0.525	0.214	0.957
56. Ustekinumab	Unbound (3HMW)	0.495	0.411	1.054	2.363	0.254	0.254	0.592
MEAN VALUES		0.831	1.181	1.176	3.093	1.144	0.517	1.288

Table A4.1. Backbone Root-Mean-Square Deviation (RMSD) across each region for the 56 of 137 CSTs with unbound/bound PDB reference structures. All models were made with ABodyBuilder, without using sequence identical templates. Gaps are assigned if the PDB structure has missing residues, precluding an accurate RMSD assignment.

Metric	242 CST Models	14,072 VdH Ig-seq Models	56 CST Structures	33 NE Human Structures
Total CDR Length (L)	48.02 ± 3.77	49.75 ± 3.49	47.64 ± 3.20	51.03 ± 4.35
PSH, CDR Vicinity (Kyte)	123.30 ± 16.60	133.76 ± 21.08	114.92 ± 14.00	124.61 ± 16.54
PPC, CDR Vicinity	0.24 ± 0.49	0.25 ± 0.52	0.19 ± 0.36	0.44 ± 0.73
PNC, CDR Vicinity	0.41 ± 0.66	0.38 ± 0.62	0.35 ± 0.60	0.70 ± 1.09
SFvCSP	3.34 ± 7.44	3.67 ± 7.40	3.81 ± 6.87	3.44 ± 7.56

Table A4.2. Average TAP Metric Values for the 242 CST models, 14,072 Human VdH Ig-seq models, 56 CST crystal structures, and 33 Human non-engineered (NE), non-redundant crystal structures. Results are reported as mean values ± one standard deviation. The identities of the 33 human, non-redundant, non-engineered structures are listed in Dataset S1 of Raybould *et al.* (10.1073/pnas.1810576116) and the 56 therapeutic structures are listed in Table A4.1..

Metric	137 CST Amber Flag Region	Number Amber Flagged	137 CST Red Flag Region	Number Red Flagged
Total CDR Length (L)	$54 \leq L \leq 59$	6	$L > 59$	2*
PSH, CDR Vicinity (Kyte)	$85.65 \leq PSH \leq 98.74$	2	$PSH < 85.65$	1
	$155.76 \leq PSH \leq 171.91$	5	$PSH > 171.91$	1*
PPC, CDR Vicinity	$1.23 \leq PPC \leq 1.51$	1	PPC > 1.51	5*
PNC, CDR Vicinity	$1.90 \leq PNC \leq 3.50$	4	PNC > 3.50	0
SFvCSP	$-19.50 \leq SFvCSP \leq -9.00$	1	SFvCSP < -19.50	1

Table A4.3. The numbers of a test set of 105 CSTs that were assigned amber or red flags across the five TAP guideline metrics (flagging thresholds set by the 137 CST dataset).

Metric	Amber Flag Threshold Value	Red Flag Threshold Value
Total CDR Length (L)	54.13 ± 0.32	59.97 ± 0.21
PSH, CDR Vicinity (Kyte)	99.64 ± 0.76	84.28 ± 1.20
	156.48 ± 1.57	172.84 ± 2.18
PPC, CDR Vicinity	1.24 ± 0.02	3.07 ± 0.20
PNC, CDR Vicinity	1.83 ± 0.08	3.47 ± 0.10
SFvCSP	-6.49 ± 0.56	-20.19 ± 0.65

Table A4.4. Statistical Sampling of the TAP Metrics. Results are presented as the mean value ± one standard deviation, calculated over 1,000 repeats of randomly sampling 200 CSTs.

Dataset	TAP Metric	Kappa Subset ($\mu \pm \sigma$)	Lambda Subset ($\mu \pm \sigma$)
242 CST Models	PSH	120.89 ± 15.10	142.03 ± 19.09
	PPC	0.21 ± 0.47	0.53 ± 0.56
	PNC	0.38 ± 0.64	0.60 ± 0.77
	SFvCSP	3.82 ± 7.38	1.67 ± 7.87
14,072 VdH Ig-seq Models	PSH	131.27 ± 21.41	141.68 ± 17.82
	PPC	0.17 ± 0.40	0.52 ± 0.73
	PNC	0.27 ± 0.48	0.74 ± 0.83
	SFvCSP	4.56 ± 7.44	0.84 ± 6.48
19,019 UCB Ig-seq Models	PSH	125.40 ± 18.56	139.66 ± 17.88
	PPC	0.11 ± 0.31	0.31 ± 0.53
	PNC	0.22 ± 0.40	0.65 ± 0.88
	SFvCSP	3.67 ± 5.30	0.12 ± 5.24

Table A4.5. Therapeutic Antibody Profiler metric values across the 242 clinical-stage therapeutics and both repertoire sequencing datasets split by kappa/lambda light chain identity.

TAP Metric	117 Phase II ($\mu \pm \sigma$)	55 Phase III ($\mu \pm \sigma$)	69 Approved/Pre-registration ($\mu \pm \sigma$)
Total CDR Length	48.12 ± 3.90	47.55 ± 3.30	48.23 ± 3.86
PSH	122.82 ± 17.08	122.60 ± 15.50	123.88 ± 17.40
PPC	0.24 ± 0.51	0.26 ± 0.55	0.24 ± 0.40
PNC	0.38 ± 0.61	0.58 ± 0.81	0.30 ± 0.58
SFvCSP	3.03 ± 7.00	4.59 ± 8.74	3.75 ± 7.03

Table A4.6. Therapeutic Antibody Profiler metric values for the 242 clinical-stage therapeutics split by clinical development.

TAP Metric	178 Active/Approved ($\mu \pm \sigma$)	59 Discontinued ($\mu \pm \sigma$)
Total CDR Length	47.83 ± 3.59	48.64 ± 4.32
PSH	122.65 ± 16.57	124.19 ± 17.81
PPC	0.25 ± 0.52	0.20 ± 0.36
PNC	0.44 ± 0.71	0.31 ± 0.47
SFvCSP	3.05 ± 7.20	5.12 ± 7.84

Table A4.7. Therapeutic Antibody Profiler metric values for the 242 clinical-stage therapeutics split by developmental progression.

TAP Metric	101 Human ($\mu \pm \sigma$)	108 Humanized ($\mu \pm \sigma$)	30 Chimeric ($\mu \pm \sigma$)	3 Mouse ($\mu \pm \sigma$)
Total CDR Length	48.68 ± 4.09	47.80 ± 3.42	46.77 ± 3.55	46.33 ± 1.25
PSH	127.76 ± 18.56	120.90 ± 14.20	115.73 ± 15.58	117.26 ± 9.44
PPC	0.29 ± 0.58	0.20 ± 0.36	0.26 ± 0.55	0.05 ± 0.06
PNC	0.34 ± 0.56	0.50 ± 0.75	0.30 ± 0.63	0.50 ± 0.50
SFvCSP	4.06 ± 7.44	3.13 ± 7.80	3.29 ± 5.99	7.58 ± 6.75

Table A4.8. Therapeutic Antibody Profiler metric values for the 242 clinical-stage therapeutics split by species or engineering protocol.

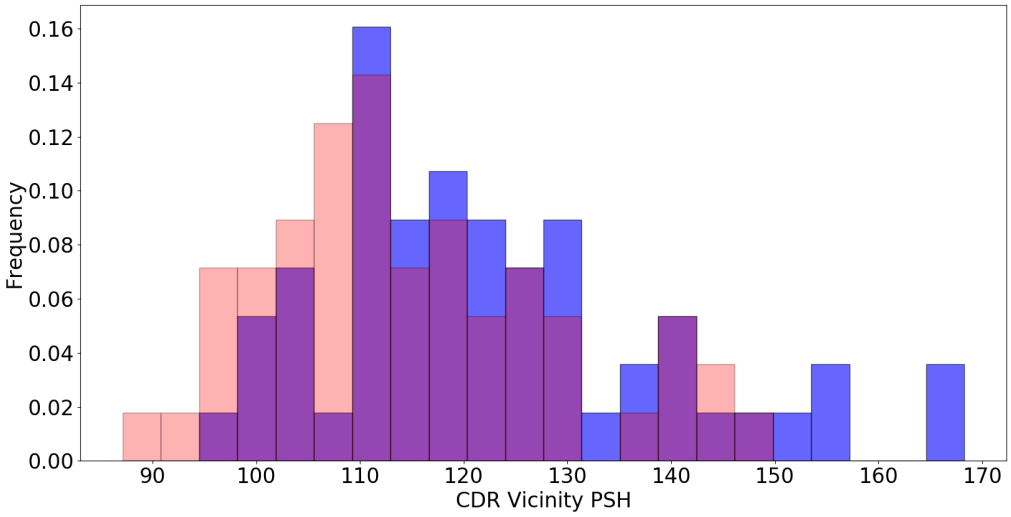


Figure A4.1. A comparison between CDR Vicinity PSH scores for 56 CST crystal structures (red) and their models (blue). Models tend to result in higher PSH scores (a mean bias of +7.96) than structures.

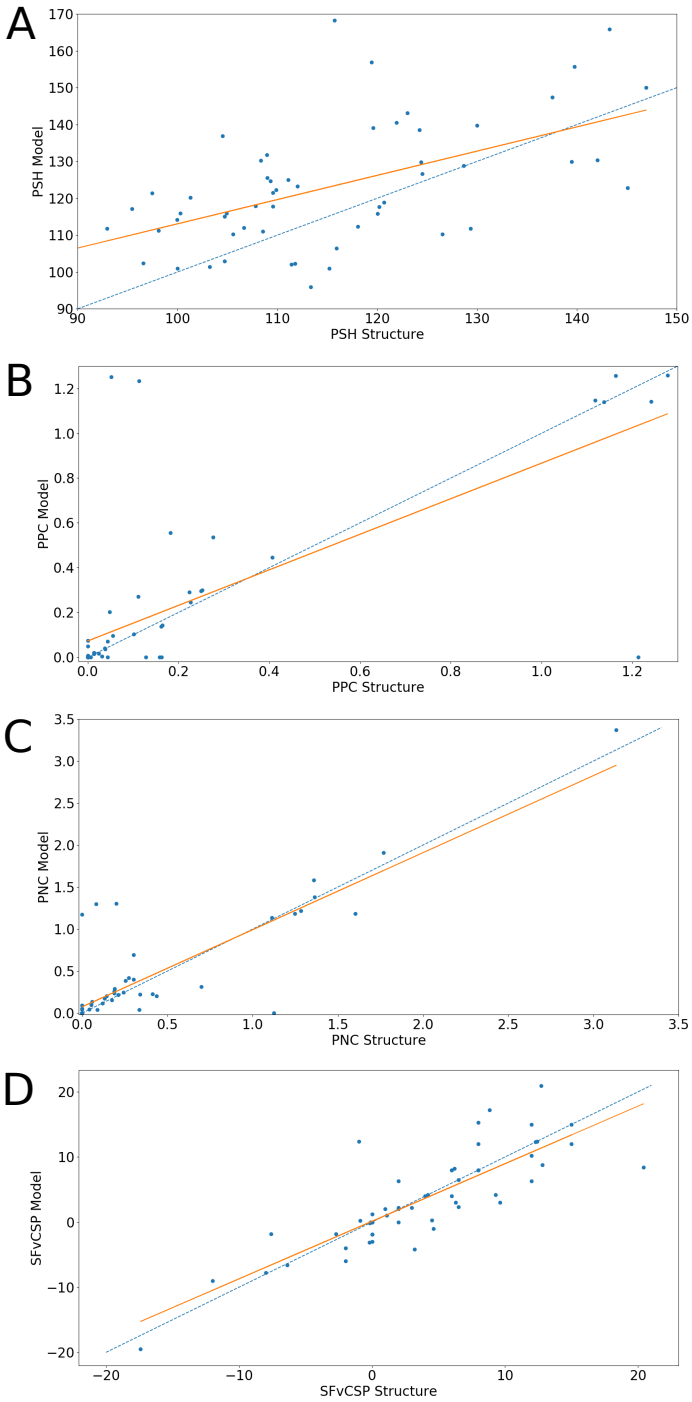


Figure A4.2. Plots of therapeutic crystal structure values (x-axis) against therapeutic model values (y-axis) for the CDR Vicinity (A) PSH, (B) PPC, (C) PNC metrics and the (D) SFvCSP metric. The identity line (blue dashes) and a line of best fit is plotted for each metric. Pearson correlation coefficients and p-values are as follows. PSH: $\rho = 0.558$, p = 7.98 x 10^{-6}; PPC: $\rho = 0.723$, p = 3.04 x 10^{-10}; PNC: $\rho = 0.858$, p = 2.78 x 10^{-17}; SFvCSP: $\rho = 0.835$, p = 1.27 x 10^{-15}.

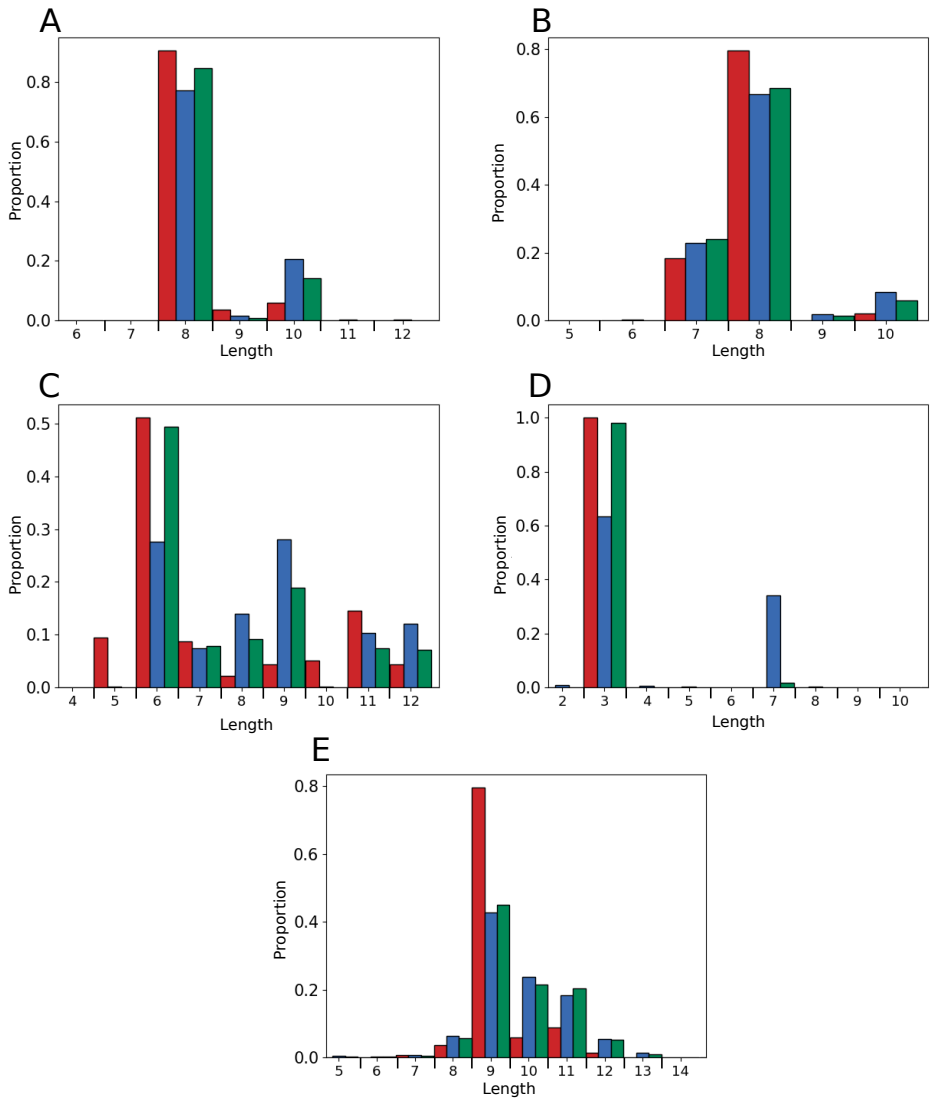


Figure A4.3. Comparing the (*A*) CDRH1, (*B*) CDRH2, (*C*) CDRL1, (*D*) CDRL2, and (*E*) CDRL3 length distributions of the 137 CST dataset (red), human VdH Ig-seq non-redundant CDRs (blue), and human VdH Ig-seq non-redundant chains (green). The VdH Ig-seq dataset contains 551,193 non-redundant heavy chains, 1,359,745 non-redundant light chains, and the following numbers of non-redundant CDR sequences: 86,345 CDRH1s, 39,449 CDRH2s, 105,458 CDRH3s, 107,721 CDRL1s, 5,276 CDRL2s, and 235,372 CDRL3s. Lengths which occur very rarely on the scale of non-redundant chains can appear much more often on the scale of non-redundant CDRs (e.g. CDRL2). This is because there are far fewer non-redundant CDR sequences for each CDR type than there are non-redundant chains, and non-redundant chains can contain the same CDR sequence (so frequently expressed CDRs of a certain length dominate the distribution).

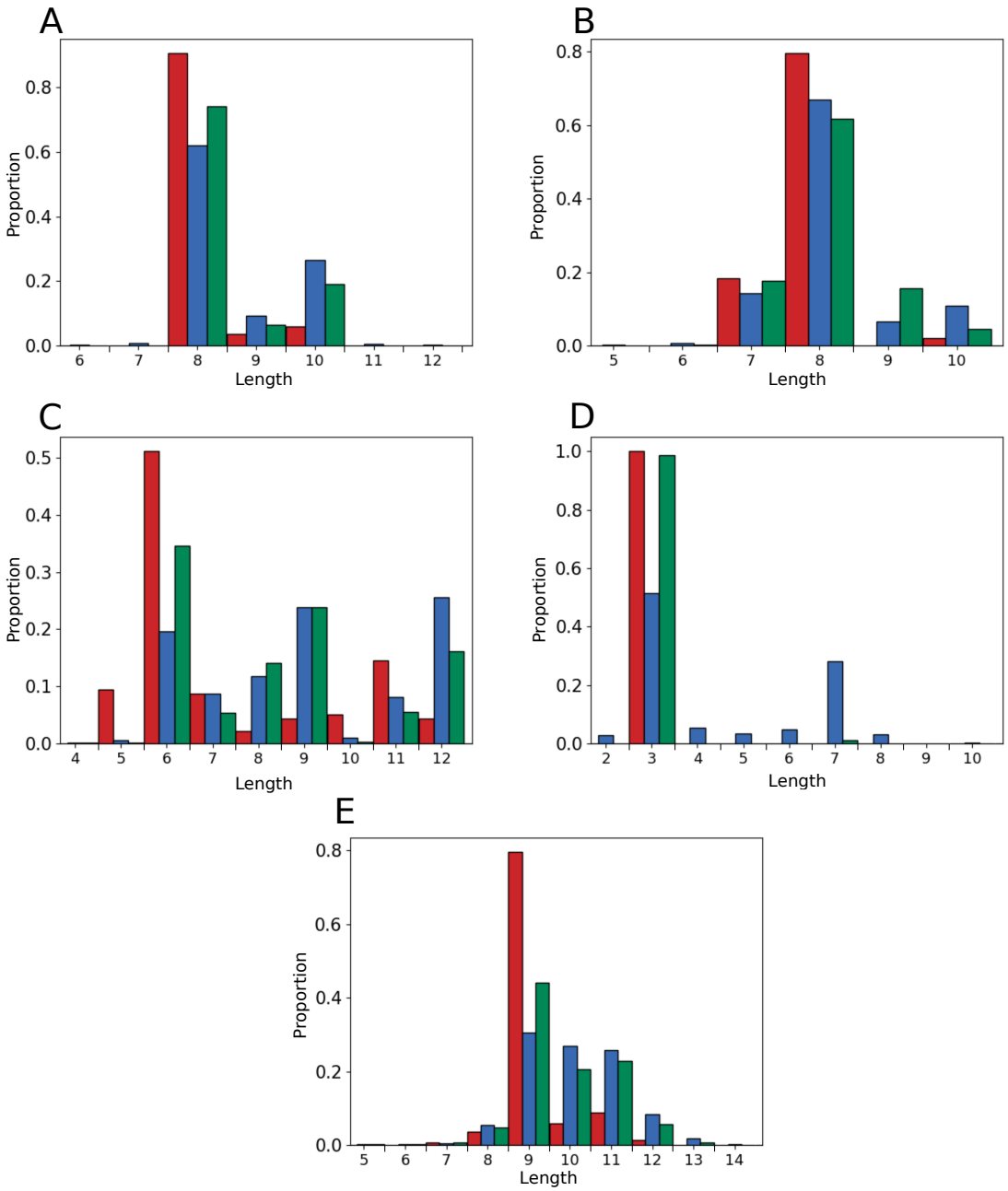


Figure A4.4. Comparing the (*A*) CDRH1, (*B*) CDRH2, (*C*) CDRL1, (*D*) CDRL2, and (*E*) CDRL3 length distributions of the 137 CST dataset (red), human UCB Ig-seq non-redundant CDRs (blue), and human UCB Ig-seq non-redundant chains (green). The UCB Ig-seq dataset contains 4,587,907 non-redundant heavy chains, 7,120,100 non-redundant light chains, and the following numbers of non-redundant CDR sequences: 174,490 CDRH1s, 279,873 CDRH2s, 1,696,918 CDRH3s, 455,125 CDRL1s, 8,708 CDRL2s, and 980,158 CDRL3s.

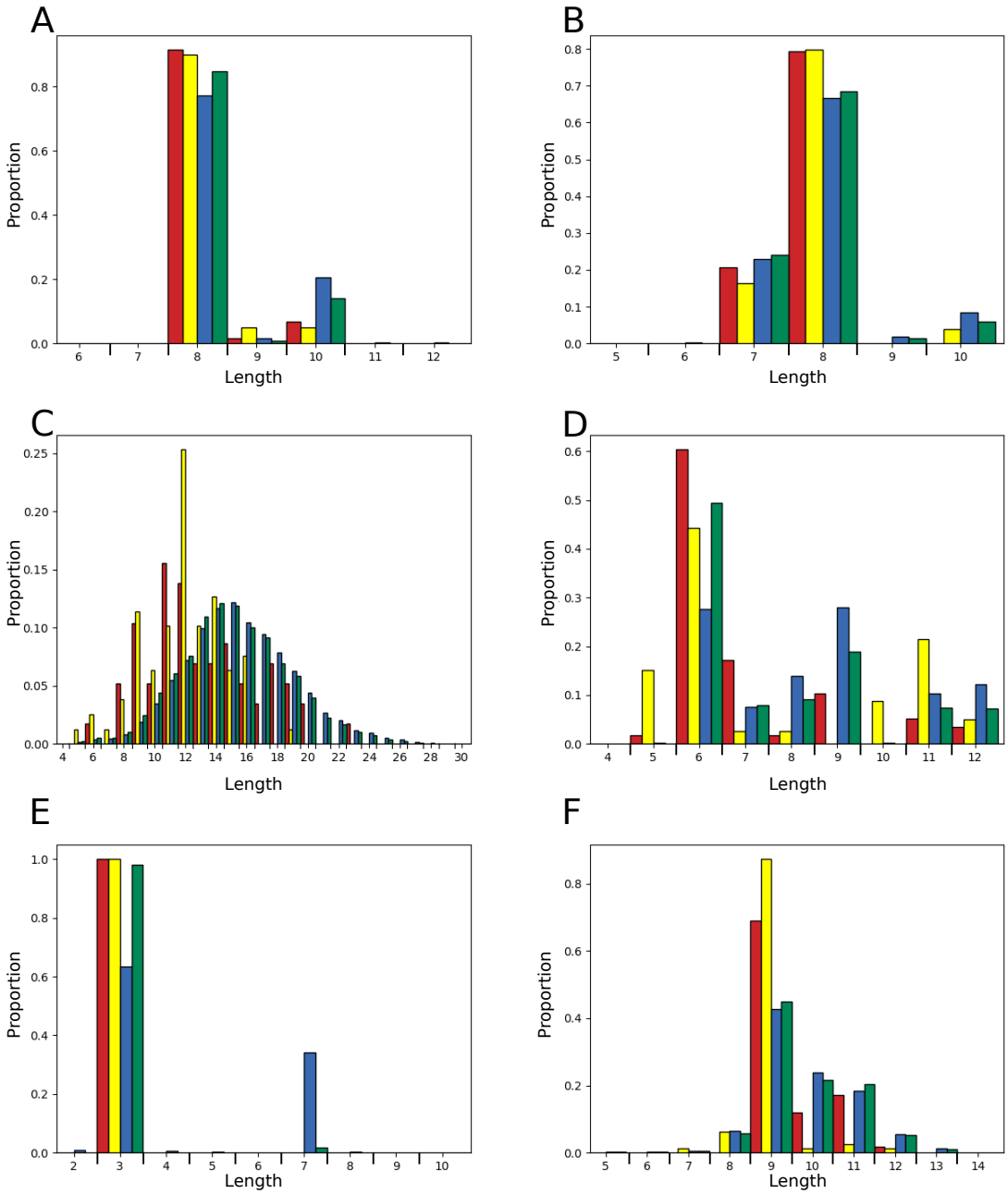


Figure A4.5. The (*A*) CDRH1, (*B*) CDRH2, (*C*) CDRH3, (*D*) CDRL1, (*E*) CDRL2, and (*F*) CDRL3 length distributions of the 58 human CSTs (red), 79 humanized, chimeric, or mouse CSTs (yellow), human VdH Ig-seq non-redundant CDRs (blue), and human VdH Ig-seq non-redundant chains (green).

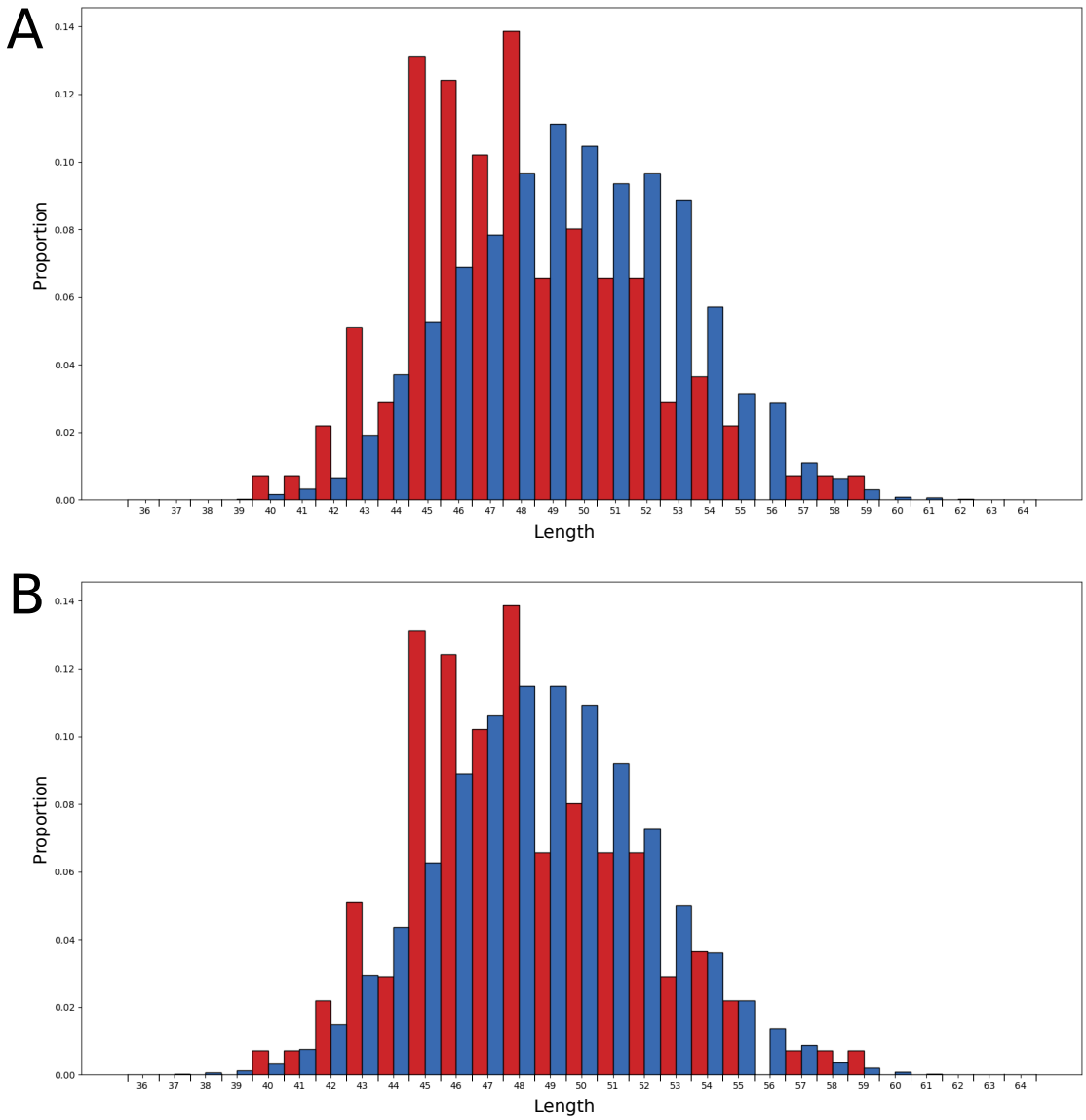


Figure A4.6. The total CDR length distributions for (A) the 137 CST (red) and the human VdH Ig-seq models (blue), and for (B) the 137 CST (red) and the human UCB Ig-seq models (blue).

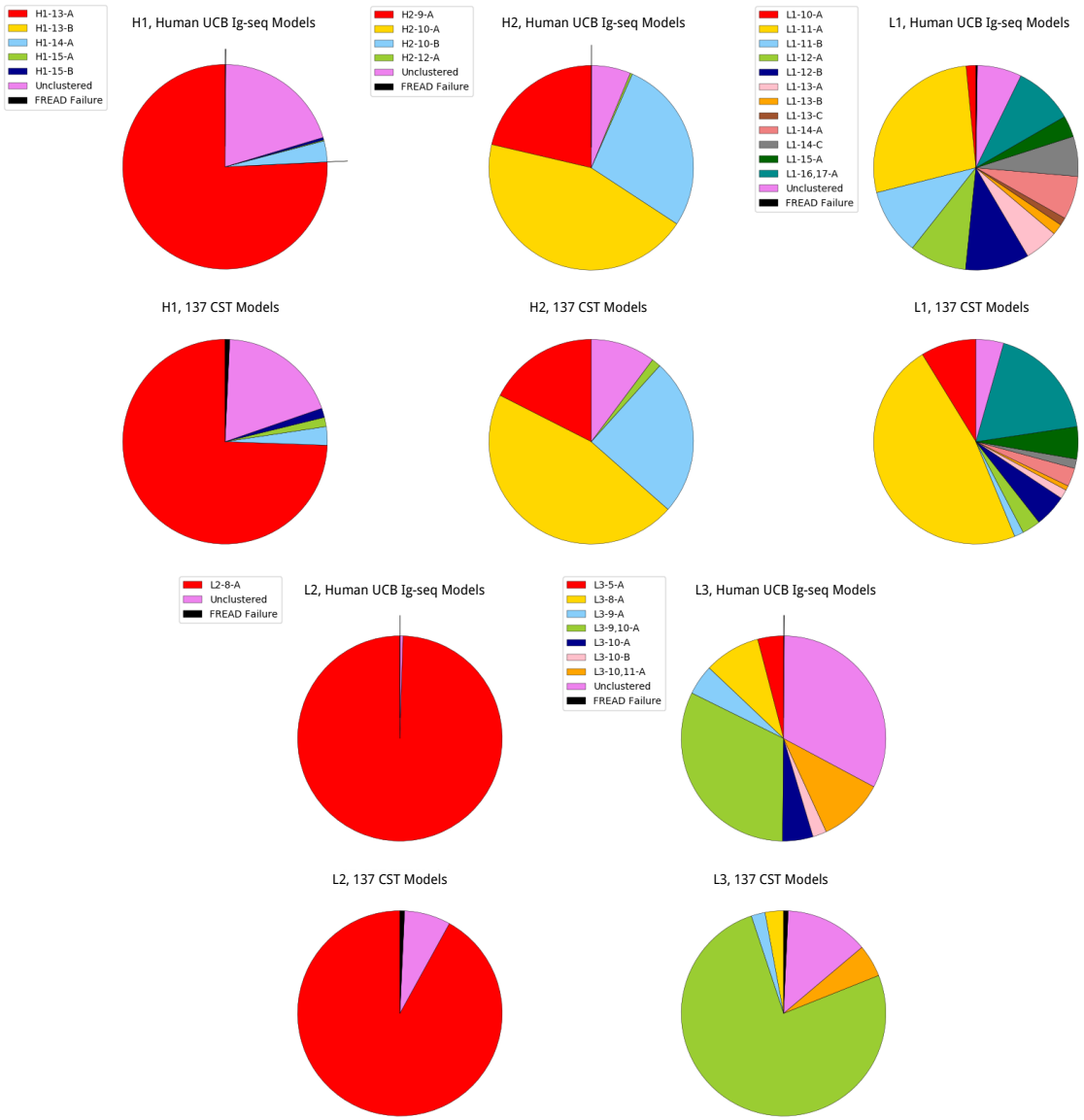


Figure A4.7. Length-independent canonical form assignments for the 137 CST and human UCB Ig-seq models.

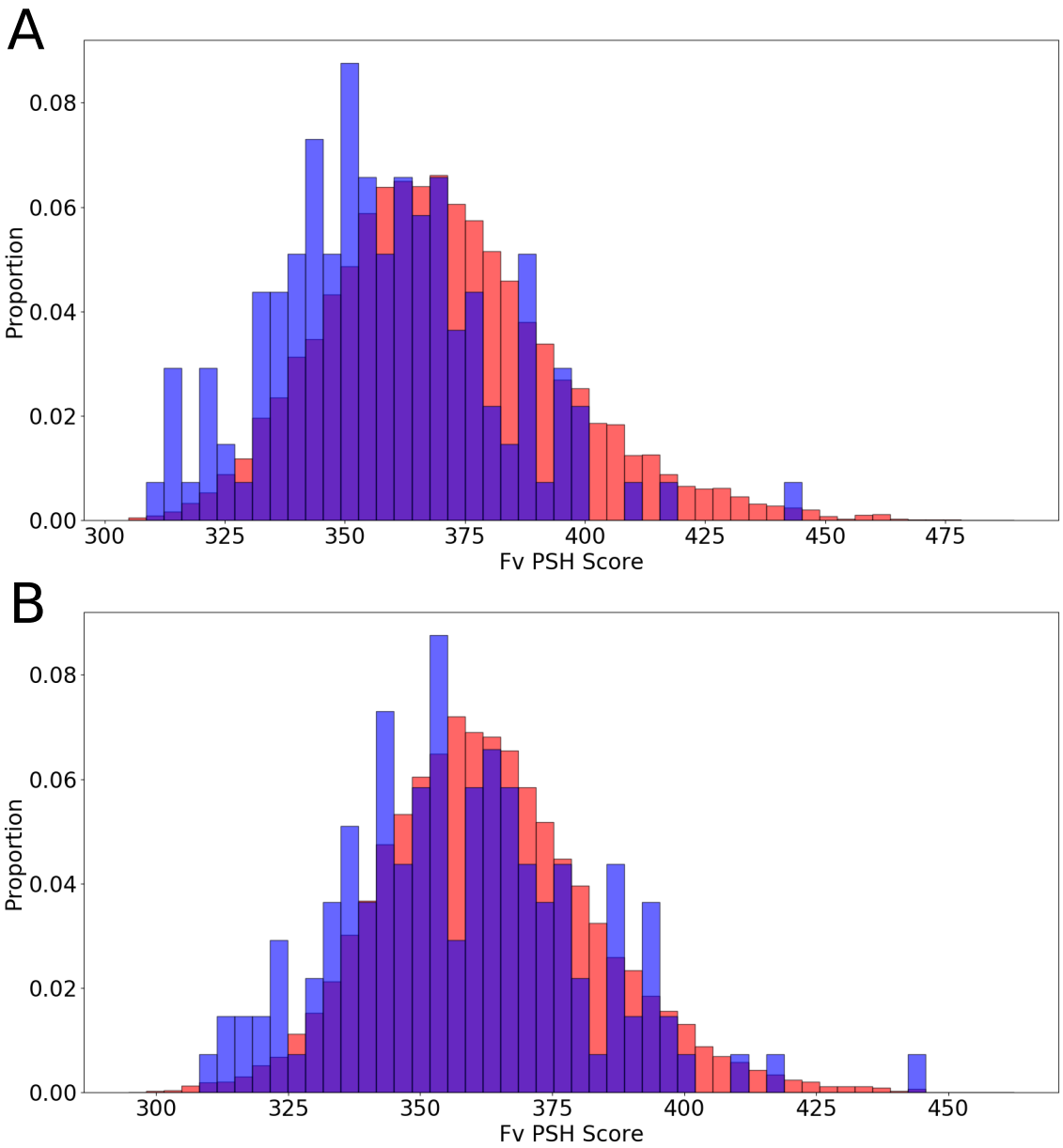


Figure A4.8. Fv Region PSH scores (A) across the 137 CST (blue) and human VdH Ig-seq (red) model datasets, and (B) across the 137 CST (blue) and human UCB Ig-seq (red) model datasets (Kyte & Doolittle hydrophobicity scale). The mean value for the human VdH Ig-seq model dataset was 370.56 ± 24.45, while for the human UCB Ig-seq model dataset was 363.13 ± 20.64.

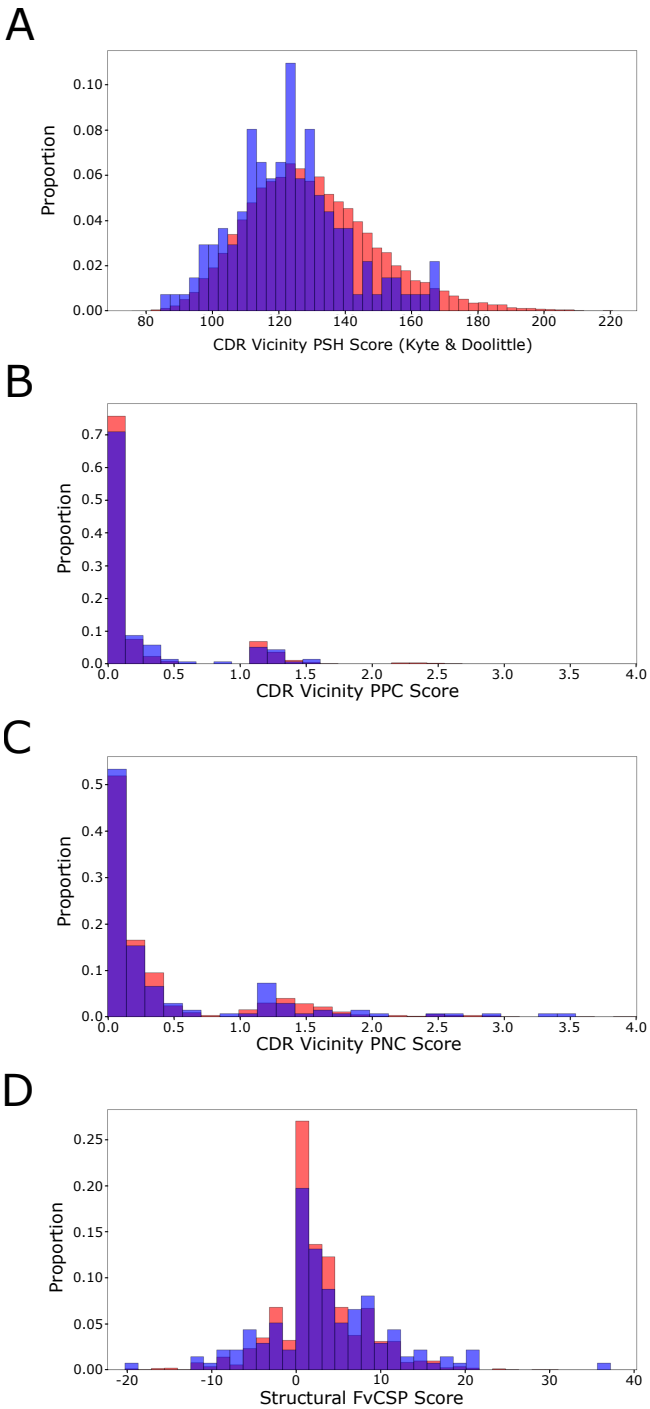


Figure A4.9. Distributions for the 137 CSTs (red) and the 19,019 human UCB Ig-seq models (blue) for the (A) CDR Vicinity PSH, (B) CDR Vicinity PPC, (C) CDR Vicinity PNC, and (D) SFvCSP metrics. The mean values for the human UCB Ig-seq models (for comparison with Table S2) are 130.10 ± 19.53, 0.18 ± 0.41, 0.36 ± 0.63, and 2.52 ± 5.54 respectively.

Chapter 5

Comparing Convergent Clones from SARS-CoV-2 Repertoires to CoV-AbDab

The following protocols were performed by our collaborators at Barts Healthy NHS Trust, Illumina, and Alchemab.

B-cell Sampling and Processing. Peripheral blood was obtained from patients admitted with acute COVID-19 pneumonia to medical wards at Barts Health NHS Trust, London, UK, after informed consent by the direct care team (NHS HRA RES Ethics 19/SC/0361). These samples were centrifuged at 150 xg for 15 minutes at room temperature to separate plasma. The cell pellet was resuspended with phosphate-buffered saline (PBS without calcium and magnesium, Sigma) to 20 ml, layered onto 15 ml Ficoll-Paque Plus (GE Healthcare) and then centrifuged at 400 xg for 30 minutes at room temperature without brake. Mononuclear cells (PBMCs) were extracted from the buffy coat and washed twice with PBS at 300 xg for 8 min. PBMCs were counted with Trypan blue (Sigma) and viability of >96% was observed. 5x106 PBMCs were resuspended in RLT (Qiagen) and incubated at room temperature for 10 min prior to storage at –80 °C.

BCR Sequencing. Total RNA from 5x106 PBMCs was isolated using RNeasy kits (Qiagen). First-strand cDNA was generated from total RNA using SuperScript RT IV (Invitrogen) and IgA and IgG isotype specific primers 32 including UMIs at 50 °C for 45 min (inactivation at 80 °C for 10 min).

The resulting cDNA was used as template for High Fidelity PCR amplification (KAPA, Roche) using a set of 6 FR1-specific forward primers (vanDongen *et al.*, 10.1038/sj.leu.2403202) including sample-specific barcode sequences (6bp) and a reverse primer specific to the RT primer (initial denaturation at 95 °C for 3 min, 25 cycles at 98 °C for 20 sec, 60 °C for 30 sec, 72 °C for 1 min and final extension at 72 °C for 7 min). The amount of Ig amplicons (∼450bp) was quantified by TapeStation (Beckman Coulter) and gel-purified.

Dual-indexed sequencing adapters (KAPA) were ligated onto 500ng amplicons per patient using the HyperPrep library construction kit (KAPA) and the adapter-ligated libraries were finally PCR-amplified for 3 cycles (98 °C for 15 sec, 60 °C for 30 sec, 72 °C for 30 sec, final extension at 72 °C for 1min). Pools of 10, 9 and 12 libraries were sequenced across three runs on an Illumina MiSeq using 2x300 bp chemistry.

Sequence Processing. The Immcantation framework (docker container v3.0.0) was used for sequence processing (Vander Heiden *et al.*, 10.1093/bioinformatics/btu138;

Gupta *et al.*, 10.1093/bioinformatics/btv359). Briefly, paired-end reads were joined based on a minimum overlap of 20 nt, and a max error of 0.2, and reads with a mean phred score below 20 were removed. Primer regions, including UMIs and sample barcodes, were then identified within each read, and trimmed. Together, the sample barcode, UMI, and constant region primer were used to assign molecular groupings for each read. Within each grouping, usearch (Edgar *et al.*, 10.1093/bioinformatics/btq461) was used to subdivide the grouping, with a cutoff of 80% nucleotide identity, to account for randomly overlapping UMIs. Each of the resulting groupings is assumed to represent reads arising from a single RNA. Reads within each grouping were then aligned, and a consensus sequence determined. Finally, duplicate reads were collapsed into a single processed sequence for analysis. Collapsing duplicate reads ensures that each processed sequence represents a sequence from a single B cell and our analysis is not confounded by expression level.

For each processed sequence, IgBlast (Ye *et al.*, 10.1093/nar/gkt382) was used to determine V, D and J gene segments, and locations of the CDRs and FWRs. Isotype was determined based on comparison to germline constant region sequences. Sequences annotated as unproductive by IgBlast were removed. The number of mutations within each sequence was determined using the shazam R package (Gupta *et al.*, 10.1093/bioinformatics/btv359).

Public BCR Sequence Data Processing. As a control set of resting-state repertoires, we used a healthy BCR sequence dataset previously described in Ghraichy *et al.* (10.3389/fimmu.2020.01734). This dataset was chosen as it was prepared using the same primer set as the SARS-CoV-2 samples, mitigating the risk of protocol-specific biases. The subset of samples from the peripheral blood of participants aged 10 years or older were used. This resulted in samples from 40 different participants, with a mean age of 28 (range: 11-51). Furthermore, only class-switched sequences were considered.

As a control set of repertoires responding to an unrelated virus, we use influenza vaccine datasets from two different studies across 6 different participants (Tipton *et al.*, 10.1038/ni.3175; Gupta *et al.*, 10.4049/jimmunol.1601850). All participants were administered a seasonal influenza vaccine, and peripheral blood was taken for BCR sequencing 6-9 days following vaccination. The processed data from these studies was obtained directly from the Observed Antibody Space database (Kovaltsuk *et al.*, 10.4049/jimmunol.1800708), and again only the class-switched sequences were considered.

Clonotyping. BCR sequences were clustered to identify those most likely to arise from clonally related B cells; a process termed 'clonotyping'. Clonotyping was performed using a previously described algorithm (Galson *et al.*; 10.1038/icb.2015.57); briefly, clonotype members must have identical V and J gene segment usage, identical CDRH3 length, and are allowed 1 amino acid mismatch for every 10 amino acids within the CDRH3. Cluster centers were defined as the most common sequence within the cluster.

Additionally, the datasets containing the 31 COVID-19 patients, the 6 influenza vaccine participants, and the 40 healthy controls were each internally clustered to identify the convergent clonotypes within each study. Convergent healthy and SARS-CoV-2 clonotypes were defined as those present in at least four patients, while convergent influenza clonotypes were defined as those present in at least two patients (owing to a smaller number of participating individuals).