Leveraging BCR Data To Improve Computational Design of Therapeutic Antibodies



Sarah A. Robinson
Wadham College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2022

Acknowledgements

During my DPhil I have benefited from the support of many people. I am enormously grateful to Professor Charlotte Deane for all of her advice (PhD-related and otherwise). I have become a much better researcher because of her guidance. I would also like to thank my industrial supervisor, James Snowden, and UCB Pharma, EPSRC, MRC and SABS CDT for their studentship funding.

I feel very lucky to have been part of the Oxford Protein Informatics Group, where I have been surrounded by truly brilliant people. I am deeply appreciative of the antibody wisdom received from Aleksander Kovaltsuk, Catherine Wong and Matt Raybould. I am indebted to Alissa Hummer, Brennan Abanades Kenyon, Leo Klarner, Lewis Chinery, Lucy Vost, Tobias Olsen, Tom Hadfield, and especially Eve Richardson, for all the gin, tea and acceptance of my office decorating efforts. It has been a pleasure to be an OPIGlet and I will miss 2.17 immensely.

I am grateful to the SABS 2018 cohort, particularly Evan Roberts, Maz Aspbury, Miriam Stricker and Solveig Van der Vegt for being excellent housemates. I have also benefited immeasurably from the coaches, support staff, athletes and donors of Oxford University Women's Lightweight Rowing Club.

I am incredibly thankful to Matt Kerin for his patience and kindness.

Finally, I would like to express my never-ending gratitude to my parents, my brothers, and our dogs, for all their love and encouragement.

Ownership of Work

I performed all of the work described in this thesis, unless otherwise stated. Where work was performed by other individuals, they have been credited by name.

List of Publications

This DPhil led to the following publications:

Wong, W.K., **Robinson, S.A.**, Bujotzek, A., Georges, G., Lewis, A.P., Shi, J., Snowden, J., Taddese, B. and Deane, C.M. (2021). Ab-Ligity: identifying sequence-dissimilar antibodies that bind to the same epitope. MAbs (Vol. 13).

Robinson, S.A., Raybould, M.I., Schneider, C., Wong, W.K., Marks, C., Deane, C.M. (2021). Epitope profiling using computational structural modelling demonstrated on coronavirus-binding antibodies. PLoS Computational Biology, 17(12).

Macpherson, A., Laabei, M., Ahdash, Z., Graewert, M.A., Birtley, J.R., Schulze, M.S.E., Crennell, S., **Robinson, S.A.**, Holmes, B., Oleinikovas, V. and Nilsson, P.H., (2021). The allosteric modulation of complement C5 by knob domain peptides. Elife, 10.

Khetan, R., Curtis, R., Deane, C.M., Hadsund, J.T., Kar, U., Krawczyk, K., Kuroda, D., **Robinson, S.A.**, Sormanni, P., Tsumoto, K. and Warwicker, J. (2022) Current advances in biopharmaceutical informatics: guidelines, impact and challenges in the computational developability assessment of antibody therapeutics. MAbs (Vol. 14).

Abstract

Antibodies are highly useful therapeutics but their development can be costly and resource intensive. As increasing quantities of high quality B cell receptor (BCR) data become available, the challenge of how to best utilise the data to improve the antibody development process remains. In the results chapters of this thesis, we begin by describing our novel approach to numbering antibody sequences, ANARCII. Our model, trained using deep learning-based sequence embeddings derived from BCR data, can correctly number antibody sequences that existing tools cannot. Next, we describe our development of existing antibody comparison tools Ab-Ligity and paratyping, to create computational pipelines for identification of antibodies likely to bind to a target antigen, with or without a known binding antibody. We compare antibodies based on sequence, paratope and structural similarity to predict similarly-binding antibodies from an antibody model library. We expand on the approach of identifying structurally similar, functionally similar, antibodies in our final results chapter where we describe our novel computational epitope binning method, SPACE. Our method of clustering antibody models based on predicted structural similarity can group sequence-dissimilar antibodies into epitope bins with high accuracy, as demonstrated on coronavirus-binding and malaria-binding antibodies. We conclude with potential future research directions, including how latest developments in antibody structure prediction and protein language models might be incorporated into our work.

Contents

1 Introduction **1**

1.1 Antibodies . . . 2

1.1.1 Antibody sequence and structure . . . 2

1.1.2 Antibody diversity . . . 3

1.1.3 B cell development . . . 5

1.1.4 Role of antibodies in the immune system . . . 5

1.1.4.1 Antigen recognition . . . 5

1.1.4.2 Affinity maturation . . . 6

1.1.4.3 Clonal selection . . . 7

1.1.4.4 Memory . . . 7

1.1.4.5 Functionality . . . 7

1.1.5 Antibodies as therapeutics . . . 8

1.1.5.1 Hit identification . . . 9

1.1.5.2 Hit expansion . . . 10

1.1.5.3 Lead optimisation . . . 10

1.2 BCR sequence data . . . 11

1.2.1 Generation of BCR datasets . . . 11

1.2.2 Paired sequences . . . 12

1.2.3 Linking sequences to antigen information . . . 12

1.2.4 Sequence annotation . . . 13

1.2.4.1 Germline alignment . . . 13

1.2.4.2 Numbering . . . 14

1.2.4.3 Antibody numbering tools . . . 16

1.2.5 Sequence embeddings . . . 18
1.3 Antibody structural data . . . 19
1.3.1 Methods for determining antibody structure . . . 19
1.3.2 Antibody modelling . . . 19
1.3.2.1 Homology modelling approaches . . . 19
1.3.2.2 Deep learning approaches . . . 21
1.3.2.3 Model quality metrics . . . 22
1.3.3 Repertoire structural analysis . . . 22
1.3.3.1 Repertoire modelling . . . 22
1.3.3.2 Repertoire structural annotation . . . 23
1.3.3.3 Repertoire structural profiling and antibody model libraries . . . 23
1.4 Experimental characterisation of antibody-antigen interactions . . . 24
1.4.1 Paratope-epitope interactions . . . 24
1.4.2 Competition epitope binning . . . 25
1.5 Computational methods for prediction of antibody-antigen binding . . . 26
1.5.1 Paratope prediction . . . 26
1.5.2 Epitope prediction . . . 27
1.5.3 Docking . . . 27
1.5.4 Computational epitope binning . . . 28
1.5.4.1 Sequence-based clustering . . . 28
1.5.4.2 Structure-informed clustering . . . 29
1.6 Impact of BCR data on therapeutic antibody development . . . 31
1.6.1 Immune system mechanisms . . . 32
1.6.2 Biotherapeutic discovery . . . 32
1.6.3 Antibody design against target epitope, given known binder . 33
1.6.4 Antibody design against target epitope, without known binder 34
1.7 Thesis overview . . . 36

2 ANARCII: A tool for numbering antibody sequences 38
2.1 Introduction . . . 38
2.2 Materials and methods . . . 41
2.2.1 Datasets . . . 41
2.2.1.1 Training data - heavy chain . . . 41

2.2.1.2 Training data - light chain . . . 43
2.2.1.3 Test data . . . 44
2.2.1.4 Nanobodies . . . 44
2.2.1.5 T cell receptors . . . 44
2.2.1.6 Non-immunoglobulin superfamily proteins . . . 45
2.2.2 Methods . . . 46
2.2.2.1 ANARCI . . . 46
2.2.2.2 Evolutionary scale modeling (ESM)-1b . . . 46
2.2.2.3 Antibody representations . . . 46
2.2.2.4 Multiclassification task . . . 47
2.2.2.5 Random forests . . . 47
2.2.2.6 XGBoost . . . 48
2.2.2.7 Model output processing . . . 48
2.2.2.8 Binary CDR/FWR classifier . . . 49
2.3 Results . . . 50
2.3.1 Antibody number labels are highly diverse . . . 50
2.3.2 Random forest models can accurately predict numbering for heavy, kappa and lambda chains . . . 53
2.3.3 XGBoost performance is inferior to RF . . . 57
2.3.4 Binary CDR prediction . . . 57
2.3.5 Nanobodies . . . 57
2.3.6 Distinguishing antibodies from non-antibodies . . . 58
2.4 Discussion and future work . . . 58

3 Computational Methods for Hit Identification and Expansion in Existing Antibody Datasets 62
3.1 Introduction . . . 62
3.2 Materials and methods . . . 65
3.2.1 Data . . . 65
3.2.1.1 SAbDab: antibody-antigen complexes . . . 65
3.2.1.2 Lysozyme-binding antibodies . . . 65
3.2.1.3 SAbDab case studies . . . 66
3.2.1.4 Public baseline AML . . . 66
3.2.2 Methods . . . 67
3.2.2.1 Clonotyping . . . 67
3.2.2.2 Paratyping . . . 68

3.2.2.3 Ab-Ligity 69
3.2.2.4 Sequence-agnostic Ab-Ligity 69
3.2.2.5 Definition of similar antibodies 69
3.2.2.6 Comparison of Ab-Ligity and paratyping 70
3.2.2.7 Identification of sequence-dissimilar binders given known binding antibody sequence 71
3.2.2.8 Identification of binders without a known binding antibody sequence 72
3.2.2.9 Construction and scoring of position specific scoring matrices 73
3.2.2.10 Incorporation of BLOSUM matrix to PSSM 75
3.3 Results 76
3.3.1 Comparison of Ab-Ligity and paratyping 76
3.3.1.1 Performance on antibody crystal structures 77
3.3.1.2 Performance on antibody models 78
3.3.1.3 Performance on heavy chain data only 78
3.3.1.4 Comparison of identified hits 78
3.3.2 Identification of binding antibodies to a target epitope, given a known binding sequence 80
3.3.3 Prediction of binding antibodies to a target epitope, without a known binding sequence 81
3.3.3.1 Incorporation of sequence-agnostic Ab-Ligity to identify potential binders 81
3.3.3.2 Comparison of sequence-specific and sequence-agnostic Ab-Ligity hit identification 83
3.3.3.3 Paratope similarity of AML sequences to target antibody 84
3.3.3.4 Clonotype diversity within AML families 85
3.3.4 Prediction of binding antibodies to a target epitope, without known binders - PB-AML2 86
3.3.4.1 Impact of AML architecture on similar structure identification 86
3.3.4.2 Alternative hit identification 87
3.3.4.3 PSSMs fail to distinguish hits from non-hits 88
3.3.4.4 Sequence-agnostic Ab-Ligity identifies additional hits 91
3.3.4.5 Ab-Ligity hits occupy different paratope clusters . . 91

3.3.4.6 Position-only paratyping . . . 92
3.4 Discussion and future work . . . 93

4 Structural Profiling of Antibodies to Cluster by Epitope 97
4.1 Introduction . . . 98
4.2 Material and methods . . . 101
4.2.1 Data . . . 101
4.2.1.1 Coronavirus antibody database (CoV-AbDab) . . . 101
4.2.1.2 Solved co-crystal structures . . . 102
4.2.1.3 Malaria antibodies . . . 102
4.2.2 Methods . . . 102
4.2.2.1 Numbering scheme and region definitions . . . 102
4.2.2.2 ABodyBuilder . . . 102
4.2.2.3 Structural clustering of homology-template models . 103
4.2.2.4 ABodyBuilder-ML . . . 104
4.2.2.5 Structural clustering of non-homology-template models 104
4.2.2.6 'Domain-consistent' structural cluster definitions . . 104
4.2.2.7 Clonotyping . . . 105
4.2.2.8 Paratyping . . . 105
4.2.2.9 Arpeggio . . . 106
4.3 Results . . . 106
4.3.1 Sequences and structures in CoV-Ab-Dab . . . 106
4.3.2 Analysis of SARS-CoV-2—antibody structural complexes . . . 106
4.3.2.1 Epitope binning . . . 107
4.3.2.2 Paratope analysis . . . 109
4.3.3 Structural convergence across CoV-AbDab antibodies . . . 113
4.3.3.1 Epitope binning . . . 115
4.3.3.2 Evolutionarily-conserved epitope topologies across coronaviruses . . . 116
4.3.3.3 Epitopes targetable by multiple species . . . 117
4.3.3.4 Structural clusters frequently span multiple clonal lineages . . . 117
4.3.3.5 'Public' response antibodies . . . 118
4.3.4 Updated CoV-AbDab analysis . . . 120
4.3.4.1 Epitope binning of homology modelled structures . . 120
4.3.4.2 Epitope binning of deep learning modelled structures 121

4.3.5 Malaria 122
4.3.5.1 Epitope binning of homology modelled structures . . 122
4.3.5.2 Epitope binning of deep learning modelled structures 123
4.3.5.3 Paratyping 124
4.4 Discussion and future work 125

5 Conclusions and Future Work **129**
5.1 Conclusions 129
5.2 Future work 130

References **132**

Appendix **155**

List of Figures

1.1 Schematic diagram of antibody structure 2
1.2 Labelled strands of the immunoglobulin fold 3
1.3 Antibody genes and VDJ recombination 4
1.4 Heavy chain sequence numbering according to IMGT, Kabat, Chothia, Martin and Aho schemes. 15
1.5 Illustration of HMM-based numbering approaches such as ANARCI . 17
1.6 Ab-Ligity method 31

2.1 IMGT/Collier de Perles representation numbering of an antibody sequence 39
2.2 Incorrect ANARCI numbering: long CDRH3 loop 40
2.3 Antibody sequence representations from the ESM-1b transformer . . 46
2.4 ANARCII pipeline for antibody numbering 49
2.5 Tools give different IMGT numbering 55
2.6 Altered sequence numbering with different versions of ANARCI . . . 55

3.1 13 antibodies binding to the same epitope on hen-egg lysozyme . . . 66
3.2 Repertoire structural profiling algorithm 68
3.3 Ab-Ligity and sequence-agnostic Ab-Ligity 70
3.4 Ab-Ligity pipeline: identifying sequence-dissimilar binders from an antibody dataset, given a known binding antibody sequence 71
3.5 Paratyping pipeline: identifying sequence-dissimilar binders from an antibody dataset, given a known binding antibody sequence 71
3.6 Proof of concept pipeline using a known set of binders 72

3.7 Pipeline for identifying potentially binding antibodies to a target epitope, without known binders 73
3.8 Clonotype frequency matrices for PSSM construction 75
3.9 Paratype frequency matrices for PSSM construction 76
3.10 CDRH3 sequence and paratope conservation across example AML family 90

4.1 Antibodies in the SARS-CoV-2-RBD 'neck' cluster 107
4.2 CDRH3 loops of nanobodies and antibodies binding the 'neck' and 'left flank' 109
4.3 CDR paratope conservation across antibodies in the 'neck' cluster . . 111
4.4 Structural epitope residue overlap between epitope group members . . 113
4.5 SARS-CoV-2 RBD-binding antibodies with similar predicted structure that span multiple clonotypes 120
4.6 SARS-CoV-2 RBD-binding antibodies with similar predicted ABB-ML structures that span multiple clonotypes 122
4.7 Differences in ABodyBuilder and ABodyBuilder-ML CDRH3 loops . 124

List of Tables

1.1 CDR definitions for IMGT numbered sequences 16

2.1 Heavy chain amino acid training datasets 42
2.2 Training datasets of amino acids from ESM-1b sequence representations 43
2.3 200 most common number labels in OAS 43
2.4 Test datasets from paired and from unpaired OAS data 45
2.5 Non-immunoglobulin superfamily proteins 45
2.6 Random forest models trained on heavy chain datasets 47
2.7 Insertion number labels with the greatest differences in frequency between kappa and lambda datasets 52
2.8 Overlap of ANARCII antibody numbering with ANARCI 53
2.9 ANARCII performance using random forest models trained on only the most common number labels in OAS 56
2.10 XGBoost model performance on kappa and lambda chains 57
2.11 Overlap of ANARCII nanobody numbering with ANARCI 58

3.1 Lysozyme-binding antibodies 65
3.2 Test case antibodies from SAbDab used as test cases 66
3.3 Precision-recall values for prediction of paratope similarity on crystals and models by Ab-Ligity and paratyping 77
3.4 Precision-recall values for prediction of paratope similarity by Ab-Ligity and paratyping, using only heavy chain information. 78
3.5 Precision-recall values for prediction of paratope similarity using sequence-agnostic Ab-Ligity 83

3.6 AML hits identified using sequence-agnostic Ab-Ligity and sequence-specific Ab-Ligity 84
3.7 Comparison of Ab-Ligity hits for SAbDab targets in PB-AML1 and PB-AML2 87
3.8 4i18 score against PSSMs of known Ab-Ligity AML hits and non-hits 89
3.9 Clonotyping AML hits to 4i18 90
3.10 Comparison of sequence-specific and sequence-agnostic Ab-Ligity hits for SAbDab targets in PB-AML2 91
3.11 Number of paratope clusters within 4i18 hit AML structural families. 92
3.12 Position-only paratope clustering of 4i18 hit AML structural families 93

4.1 Epitope size across epitope groups 112
4.2 Structural clustering of CoV-AbDab antibodies (time-stamped 26th July 2022) 121

A1 Antibodies with a solved crystal structures in complex with a SARS-CoV-2 antigen 158
A2 Nanobodies with a solved crystal structures in complex with a SARS-CoV-2 antigen 159
A3 Analysis of paratope-epitope interactions between antibodies and SARS-CoV-2 RBD in cocrystal structures 161

CHAPTER 1

Introduction

In this introduction I will give an overview of antibodies and their use as therapeutics. I will begin by describing antibody sequence and structural information, and how diversity arises and contributes to the role of antibodies in the immune system. I will summarise how antibody information can be obtained, including experimental and computational methods for characterising antibody structure and antibody-antigen interactions. Finally, I will discuss existing methods for utilising sequence and structural information to aid design and development of therapeutic antibodies.

The section describing the impact of BCR data on therapeutic antibody development is adapted from text I authored in the following published article on the design of antibody libraries with improved developability and on the role of antibody repertoires in target validation and drug discovery.

Khetan, R., Curtis, R., Deane, C.M., Hadsund, J.T., Kar, U., Krawczyk, K., Kuroda, D., Robinson, S.A., Sormanni, P., Tsumoto, K. and Warwicker, J. (2022) Current advances in biopharmaceutical informatics: guidelines, impact and challenges in the computational developability assessment of antibody therapeutics. MAbs (Vol. 14).

1.1 Antibodies

Antibodies are Y-shaped proteins involved in the immune response of jawed vertebrates. They recognise and bind to antigens - substances identified by the immune system as a threat. Antibodies are produced by B cells, lymphocytes critical to the adaptive immune response, and are secreted or displayed on the B cell membrane as B cell receptors (BCRs). In this thesis I use the terms 'BCR' and 'antibody' interchangeably.

1.1.1 Antibody sequence and structure

Antibodies are composed of four peptide chains: two identical heavy chains of ∼440 amino acids and two identical light chains of ∼220 amino acids, linked by disulphide bonds. Antibodies have a crystallisable region, Fc, and an antigen binding fragment, Fab (Figure 1.1).

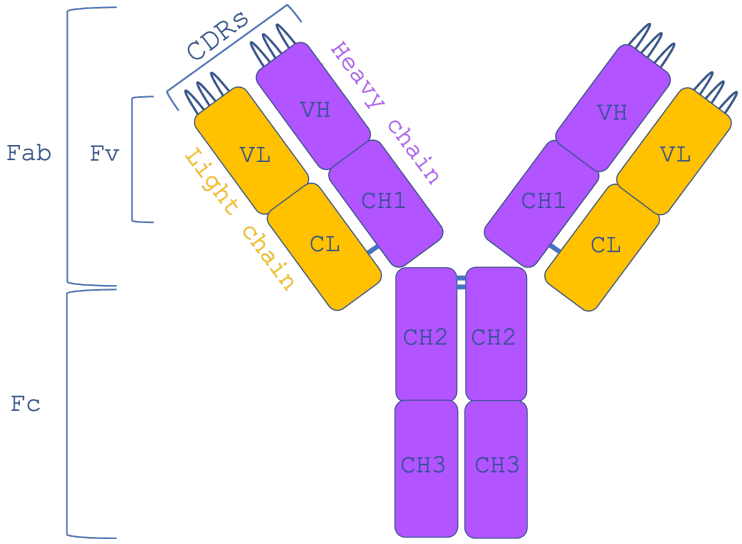


Figure 1.1: Schematic diagram of antibody structure including the crystallisable region (Fc) and the antigen binding fragment (Fab), which encompasses the variable region (Fv) with six CDR loops. Peptide chains are linked by disulphide bonds.

The Fc includes the C-terminal ends of the four chains and determines the effector functionality of the antibody. The Fab region encompasses the N-terminal ends and includes the variable region Fv, which is important for antigen specificity. The Fv includes six hypervariable loops known as complementarity determining regions (CDRs). The three CDR loops in the heavy chain (CDRH1, CDRH2 and CDRH3) and the three in the light chain (CDRL1, CDRL2 and CDRL3) are crucial for antigen

interaction and contain the majority of the binding region of the antibody (Kunik et al. 2012). The set of amino acids that comprise the antibody binding region are known as the paratope, while the antigen region it binds to is the epitope.

The CDR loops, other than CDRH3, are thought to adopt a limited number of protein backbone conformations known as canonical classes (Chothia and Lesk 1987). As more antibody structures have been made available, the number and definition of the classes have been revised (Adolf-Bryfogle et al. 2015; Nowak et al. 2016; Wong et al. 2019b). CDR conformation is influenced by loop length and the amino acids in the loop and in the framework regions either side of the loop. The framework regions surrounding CDRs primarily act as a scaffold for the loops to maintain structure, though in some cases they are thought to be directly involved in antigen-binding (Sela-Culang et al. 2013). For example, insertions in the DE loop in framework region 3 (between strands D and E of the antibody immunoglobulin fold, see Figure 1.2) have been associated with broadly neutralising antibodies binding to HIV-1, including via direct contact with the antigen (Kelow et al. 2020).

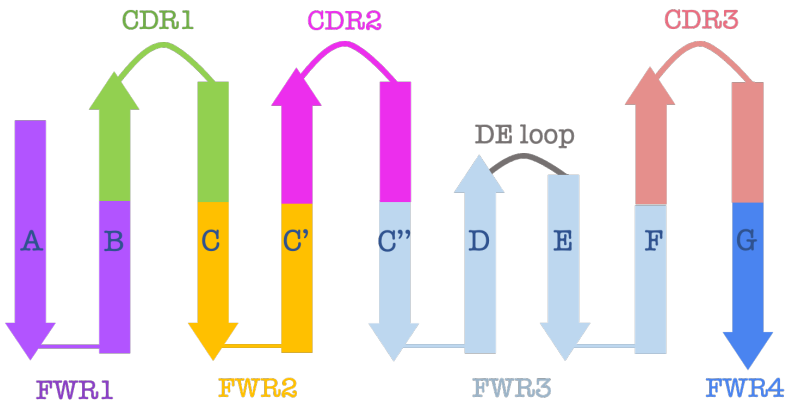


Figure 1.2: Representation of a single chain of an antibody, with labelled strands of the immunoglobulin fold (A-G), four framework regions (FWR1-4), three complementarity determining regions (CDR1-3) and the DE loop.

1.1.2 Antibody diversity

In humans, it is estimated that around 10^{13} unique antibodies can be generated by the immune system (Greiff et al. 2015a). Antibody diversity is generated through recombination of the immunoglobulin (Ig) locus - where the variable (V), diverse (D) and joining (J) genes reside. The heavy chain (VH) of an antibody is encoded by V,

D and J gene segments; the light chain (VL) is formed from V and J segments only (Figure 1.3). V genes encode CDR1 and 2, whilst CDR3 spans the V(D)J genes. Humans possess 44 V genes, 27 D genes and 6 J genes, with many different alleles (Yang et al. 2021). Heavy chains are encoded by the immunoglobulin heavy chain locus on chromosome 14 which contains multiple copies of V, D, and J gene segments. Light chains are encoded by kappa or lambda loci on chromosomes 2 and 22 respectively, which each have V and J gene segments.

To generate a unique antigen-binding region, each B cell undergoes VDJ somatic recombination; the rearrangement of V, D and J gene segments. In VDJ recombination, one of each type of gene segments is randomly selected and double stranded DNA breaks made so the segments can be joined (Shi et al. 2019). The genes are spliced together imperfectly with small insertions and deletions, resulting in a huge number of potential combinations and sequences that might be translated as antibodies (Li et al. 2004).

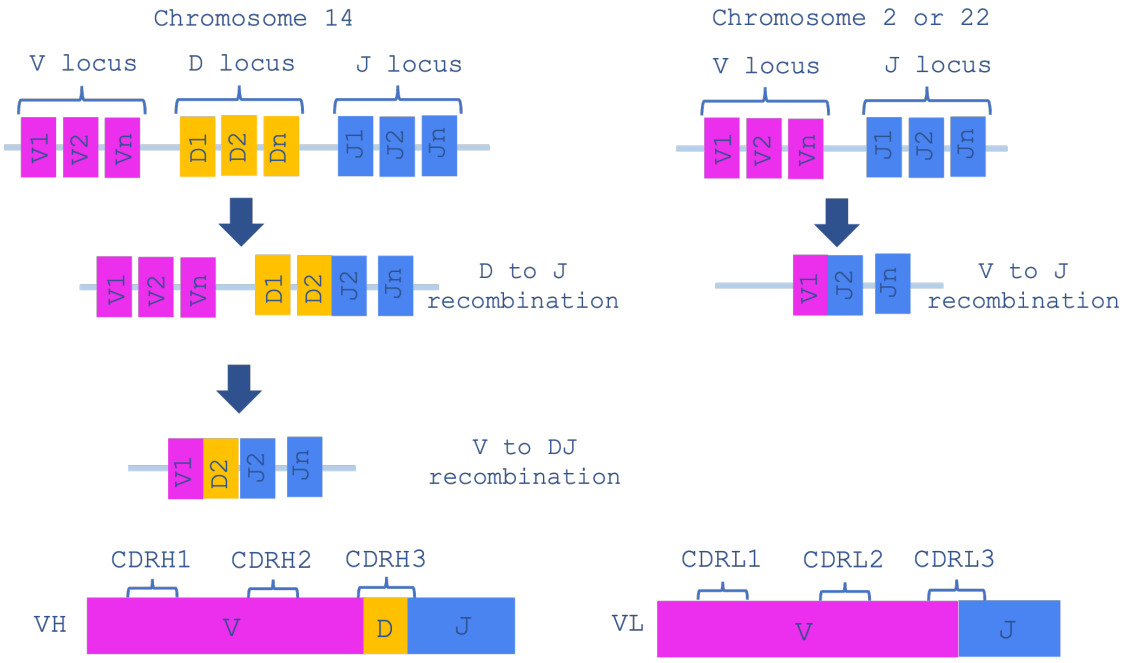


Figure 1.3: V, D and J gene segments encode VH chains; V and J encode VL chains. V, D and J gene segments are rearranged through VDJ recombination, first with D to J recombination (VH only), followed by V to (D)J recombination.

Additional diversity is introduced through somatic hypermutation (SHM). SHM occurs after B cell activation by an antigen, resulting in production of an antibody

with a unique antigen binding site (Zhou and Kleinstein 2020) (see Section 1.1.4). As the CDRH3 loop of the heavy chain is formed at the junction between the V, D and J gene segments (Figure 1.1), it is the most variable of the CDR loops in terms of length, sequence and 3D structure (Marks and Deane 2020; Morea et al. 1998; Regep et al. 2017). CDRH3 is also thought to be the most important region in determining antibody-antigen binding (Sela-Culang et al. 2013; Xu and Davis 2000).

1.1.3 B cell development

B cells originate from pluripotent hematopoietic stems cells (HPCs) in mammalian bone marrow after birth. HPCs differentiate into progenitor B cells and subsequently pre-B cells, immature B cells expressing IgM, and finally mature B cells which express IgM and IgD (Nagasawa 2006). Through allelic exclusion only one BCR is expressed per B cell by restricting rearrangement to only one chromosome and either the kappa or lambda loci (isotype exclusion), meaning only one allele encodes a functional heavy and light chain. However, a study has suggested multiple heavy chains per B cell in some cases (Shi et al. 2019).

Before leaving the bone marrow B cells are checked for affinity to host antigens (autoantigens) to reduce risk of autoimmunity; B cells expressing BCRs which bind to autoantigens are deleted. B cells which do not bind to autoantigens with high affinity exit the bone marrow and migrate to the spleen or lymph nodes, where they undergo B cell maturation and antigen-induced activation (Melchers 2015). An estimated 85% of immature B cells die in the bone marrow (Melchers 2015) and mature B cells circulating through the secondary lymphoid organs have a half life of only a few days if they do not bind to antigen (Nutt et al. 2015). Upon contact with an antigen, the naive B cells proliferate and mutate to optimise affinity.

1.1.4 Role of antibodies in the immune system

1.1.4.1 Antigen recognition

Antibodies form an integral part of the adaptive immune response and are able to bind to almost any antigen (Sela-Culang et al. 2013). Naive B cells (B cells which have not been exposed to an antigen) encounter antigens whilst in secondary lymphoid organs. Depending on the antigen type, T cells may be required for B cell antigen-activation. Antigens such as polysaccherides and unmethylated CpG DNA (from bacteria) are T-cell independent antigens and induce activation of B cells directly (Nutt et al.

2015). Protein antigens are bound and taken up by B cells and presented on the cell surface, where T helper cells recognise the antigen and activate the B cells to undergo clonal expansion (cell proliferation) (Nutt et al. 2015). After proliferation B cells may differentiate into short lived plasma cells, memory B cells, or may move to the germinal centre for further clonal expansion and affinity maturation.

1.1.4.2 Affinity maturation

Germinal centres are formed from B cells in secondary lymphoid organs during T-cell dependent antibody responses, and are organised into a dark zone and a light zone (MacLennan 1994). In the dark zone B cells undergo cellular division and mutation to improve binding affinity to the antigen through somatic hypermutation (SHM) and class switch recombination (CSR). Mutations are mediated by activation-induced deaminase (AID), which is expressed in high amounts in the dark zone. AID deaminates deooxycytidine to deoxyuridine, creating U:G mismatches in the V(D)J genes (Methot and Di Noia 2017). If the DNA is not repaired before cell replication, a C $\rightarrow$ T point mutation is created. DNA repair mechanisms are also error-prone which may lead to further mutations.

These DNA mutations, SHMs, may alter antibody binding affinity for an antigen. SHM is not uniform and depends on position and DNA sequence, with known mutation hotspot and coldspot motifs and increased frequency in CDR1 and CDR2 compared to the framework regions (Spisak et al. 2020; Tang et al. 2022; Yaari et al. 2013; Zhou and Kleinstein 2020). Levels of SHM also vary depending on the antigen and B cell, for example, antibodies known to bind to the receptor binding domain of SARS-CoV-2 can have low levels of SHM (Rogers et al. 2020; Yuan et al. 2020), whereas some broadly neutralising HIV antibodies have over 100 mutations (Klein et al. 2013).

In addition to point mutations in single DNA strands, double stranded DNA breaks are also induced by AID. This can lead to class switch recombination, where the antibody class, also known as isotype, changes. The isotype is determined by the Fc domain of the antibody, and can be IgA, IgD, IgE, IgG or IgM. Each isotype has differing functions depending on the type of Fc receptors (FcRs) that the Fc portion binds to, and the type of immune cells which express the specific FcRs (Lu et al. 2018). Naive B cells initially produce high molecular weight, low affinity IgM, but may class switch upon antigen stimulation to higher affinity, lower molecular weight

IgG, IgA, or IgE (Shi et al. 2019). Lack of the AID enzyme causes immunodeficiency by limiting the antibody repertoire to unmutated IgM (Methot and Di Noia 2017).

Whilst antibodies have generally been thought to bind specifically to one epitope, antibodies may be polyreactive (bind weakly to many targets) or polyspecific (discrete binding to different targets) (Cunningham et al. 2021; Kelly et al. 2017). In the natural antibody repertoire, polyreactivity means an antibody can recognise and bind multiple targets and therefore is useful as a low affinity, low specificity, first line of defence. Up to 50% of IgM in human serum is thought to be polyreactive and subsequent class-switching to IgG can increase affinity (Cunningham et al. 2021).

1.1.4.3 Clonal selection

B cells cycle between the dark and light zones of the germinal centre. The light zone contains T helper cells and antigen-presenting dendritic cells and is the site of clonal selection. After undergoing SHM and CSR in the dark zone, B cells move to the light zone and bind to antigen. B cells which do not bind the antigen with high affinity are destroyed, whereas B cells with high affinity are selected for and move back to the dark zone for further mutation, or exit the germinal centre (Bannard and Cyster 2017).

1.1.4.4 Memory

Following affinity maturation and selection in the germinal centre, memory B cells and long lived plasma cells are formed (Palm and Henry 2019). Upon re-exposure to an antigen, antibodies are secreted by long lived plasma cells created during the first antigen exposure. If this response is insufficient to clear the pathogen, memory B cells residing in the affected tissue, circulating, or in the secondary lymphoid organs, are activated. Depending on the type of memory cell, cells proliferate and differentiate into plasma cells for antibody secretion, or may form and re-enter germinal centres for further affinity maturation (Palm and Henry 2019).

1.1.4.5 Functionality

Antibodies can cause pathogen destruction in several ways, influenced by both the Fab and Fc domains of the antibody. Neutralising antibodies can directly bind to a virus and block entry into host cells, for example, by binding to a receptor-binding site required for virus internalisation into the host cell. Non-neutralising antibodies can recruit innate immune effectors via the Fc region. The functionality of the Fc

region depends on the isotype; different antibody isotypes, and isotype subclasses, have differing structures and functionality (Lu et al. 2018). These functions include complement-dependent cytotoxicity, antibody-dependent cell-mediated cytotoxicity (ADCC) and antibody-dependent cellular phagocytosis (ADCP), which are triggered by antibodies binding to immunoproteins and FcRs on immune cells (Tay et al. 2019). In ADCC antibodies bound to cell surface antigens trigger cell lysis via immune cells such as natural killer cells binding to the antibody Fc region and releasing cytotoxins (Zahavi et al. 2018). IgG and IgM antibodies can trigger complement-dependent cytotoxicity when bound to a target cell, where the classical complement pathway leads to cell lysis. In ADCP cells are phagocytosed (engulfed by the immune cell) rather than lysed (Tay et al. 2019).

In some cases, non-neutralising antibodies can have a negative effect on the immune response by actually promoting viral infection of host cells. Antibodies may bind to a virus (via the Fab) and also to cell FcRs (via the Fc), promoting cellular uptake of the virus. This is known as antibody dependent enhancement and occurs in viruses including dengue (Yang et al. 2022).

1.1.5 Antibodies as therapeutics

Antibodies have been licensed as therapeutics since the 1980s, with 126 therapies approved in the US and/or EU for commercial use as of 21st November 2022 (Kaplon et al. 2023). Most therapies are monoclonal antibodies (mAbs) indicated for treatment of cancer (45%), immune-mediated disorders (27%), infectious disease (8%) and cardiovascular/hemostasis disorders (7%) (Kaplon et al. 2022). Antibodies are useful as therapeutics due to their ability to bind to diverse targets with high affinity and specificity, and ability to be engineered for tailored functionality (Zhu et al. 2022). Antibodies can be used to bind receptors and extracellular proteins, to block or to induce certain pathways. IgG is the most common isotype used in therapeutic mAbs due to its long half life of 21 days (Jefferis 2012), compared to 3-6 days for other isotypes (Snapkov et al. 2022). Aside from the conventional mAb format, other antibody formats include antibody-drug conjugates to direct delivery of toxic small molecule drugs (Carter and Lazar 2018), bispecifics to bind two separate epitopes (Brinkmann and Kontermann 2021), and antibody fragments such as nanobodies.

Traditional antibody development pathways involve the generation and laboratory screening of many potential candidates. Multi-parameter optimisation of all

the characteristics needed to be a successful biotherapeutic, for example, specificity, potency, expression, solubility, aggregation, immunogenicity and stability, is highly challenging (Mason et al. 2021).

1.1.5.1 Hit identification

The antibody development process begins with antibody discovery, also known as hit identification, to find antigen-specific antibodies which bind to the target of interest. Antibody discovery may involve *in vivo* or *in vitro* approaches.

In vivo discovery typically includes immunisation of animals with a chosen antigen and identification of antigen-binding antibodies in the subsequent immune response. Antibodies can be identified directly from plasma or from isolated B cells; B cells can be immortalised by fusion with a myeloma cell to improve *ex vivo* cell lifespan for antibody production (Laustsen et al. 2021). B cells may also be isolated from convalescent humans, typically collected via blood samples. Antibodies or B cells which bind to the antigen of interest are identified through assays (see Section 1.4).

Antibodies from non-humans require humanisation to reduce the risk of immunogenicity, an immune response to proteins recognised as foreign (Marks et al. 2021). Animals can be engineered to have human immunoglobulin genes and therefore produce antibodies similar to those in humans, though these may still differ from typical human sequences and structures (Richardson et al. 2022). Antibodies identified from an *in vivo* immune response may be biased towards certain immunodominant epitopes and have low epitope diversity. However, the advantages of these antibodies include that they have undergone *in vivo* affinity maturation and tend to have decreased polyreactivity and hydrophobicity compared to *in vitro* approaches (Shehata et al. 2019).

In vitro approaches to antibody discovery include displaying recombinant antibody libraries via yeast or phage. Due to the size of full length antibodies, display is typically limited to antibody Fab fragments or single chain variable fragments which can be expressed on the cell surface. Libraries are typically incubated with the target antigen, and antibodies that bind (hits) are identified (Section 1.4). Multiple rounds of selection against the target antigen may occur, followed by sequencing of the identified hits to obtain the genetic information (Valldorf et al. 2022).

Advantages of *in vitro* approaches include identification of antibodies which bind to toxic or non-immunogenic targets, which would not be possible with animal immunisation (Laustsen et al. 2021). However, these approaches rely on having purified antigen suitable for use in large scale screening assays and the ability to use display libraries. Phage-derived antibodies have been linked to potential developability issues due to containing more aliphatic residues than antibodies derived from other approaches (Kaleli et al. 2019).

1.1.5.2 Hit expansion

When binding antibodies have been identified through initial library screening, immunised animals or convalescent patients, these hits may be expanded to identify more candidates to take forward in development. This typically involves looking for sequence- or paratope- similar antibodies within a library or dataset (see clonotyping and paratyping 1.5.4.1). One of the main purposes of hit expansion is to diversify candidates in case of developability issues or unfavourable activity associated with certain antibody sequences. For example, antibodies can have sequence motifs which lead to poor chemical stability via deamination, isomerisation or oxidation (Dashivets et al. 2016; Khetan et al. 2022; Sydow et al. 2014), or to poor physical stability through low solubility, high viscosity or aggregation (Agrawal et al. 2016; Chennamsetty et al. 2009; Nichols et al. 2015; Sormanni et al. 2017), making formulation challenging.

1.1.5.3 Lead optimisation

Optimisation of lead antibody candidates might include random and targeted mutagenesis for affinity maturation and to improve specificity and functionality. This may include creation and screening of phage display libraries with mutations in the CDRs of the lead candidates. Characteristics are also evaluated to flag potential developability issues and derisk the development of an antibody candidate (Khetan et al. 2022; Wang et al. 2021a). For example, properties can be predicted such as as immunogenicity (Marks et al. 2021; Prihoda et al. 2022), aggregation (Lai et al. 2021), post-translational modifications (Vatsa 2022) and chemical degradation (Yan et al. 2018). Sequences can also be compared to therapeutic antibodies which have already passed phase 1 clinical development (and are therefore unlikely have developability issues), for example, using metrics thought to be implicated in poor developability: CDR lengths, CDR positive and negative charges, surface hydrophobicity, and asymmetry in surface charges (Raybould et al. 2019a).

1.2 BCR sequence data

High-throughput, next generation sequencing methods have led to the generation of vast quantities of BCR sequence data. Billions of antibody sequences are publicly accessible in repositories of curated BCR sequence information, such as the Observed Antibody Space (OAS) (Olsen et al. 2022b), iReceptor (Corrie et al. 2018) and ImmuneDB (Rosenfeld et al. 2017). B cell repertoires are typically assessed through sequencing of the immunoglobulin genes in circulating B cells; a sequenced BCR repertoire therefore provides a snapshot of the BCRs present at one time. Given the number of circulating B cells in blood in humans is around 5×10^9, this represents a fraction of the total possible diversity of antibodies that can be generated (Briney et al. 2019). BCR repertoire data is also known as adaptive immune receptor repertoire sequence (AIRR-seq) data (Snapkov et al. 2022), repertoire sequence (Rep-seq) data and immunoglobulin sequence (Ig-seq) data.

1.2.1 Generation of BCR datasets

BCR sequence datasets are generated by isolating the desired B cells, obtaining their genetic material, preparing a library, and sequencing the library. Bulk BCR repertoire data is obtained from heavy and light chain messenger RNA (mRNA) fragments or from genomic DNA (gDNA) from lysed B cells (Ghraichy et al. 2020). The full antibody repertoire includes antibodies in different compartments, such as mucosal (gut and lung), systemic (in blood) and lymph nodes and spleen; however, most BCR datasets use B cells isolated from blood samples, typically using density gradient centrifugation (Ghraichy et al. 2020). mRNA-based approaches involve mRNA extraction from lysed B cells and synthesis of complementary DNA (cDNA) via reverse transcriptase. cDNA or gDNA is amplified using polymerase chain reaction (PCR) and sequenced (Ghraichy et al. 2020). To improve accuracy, unique molecular identifiers (UMIs) can be added to cDNA during reverse transcription prior to amplification (Wang et al. 2021b). UMIs are barcodes added to individual molecules of cDNA so that any errors introduced during amplification or sequencing are not counted as true variant sequences as they contain the same UMI as another sequence. mRNA approaches are more commonly used than gDNA and have the advantage of multiple copies per cell, compared to a single copy for DNA, making identification and resulting amplification easier (Georgiou et al. 2014). This may however lead to bias as quantities of BCR mRNA depend on cell type and activation state (Langerak 2022).

Proteomics-based analysis of serum antibodies has also been used, through its use is much less widespread than genomic assessment of BCR repertoires (Cheung et al. 2012; Snapkov et al. 2022). Antibodies are evaluated at the protein-level using liquid chromatography and mass spectroscopy, rather than genetic analysis of DNA or mRNA. The overlap between antibodies identified using these two approaches is debated (Snapkov et al. 2022). Additionally, there is discussion regarding whether BCR repertoires can be sufficiently characterised from only blood; proteomics analysis in rabbits showed 27% of heavy and 15% of light-chain clones were only found in bone marrow or spleen, and not in blood samples (Bonissone et al. 2020).

1.2.2 Paired sequences

When B cells are lysed in bulk to obtain mRNA or gDNA, native pairing information of heavy and light chains is lost (Kovaltsuk et al. 2017). Paired sequencing methods retain the native heavy and light chain pairings by sequencing single cells. Methods involve separating B cells into individual droplets so that the heavy and light sequences can be covalently linked together (DeKosky et al. 2013) or given the same unique barcode (Goldstein et al. 2019). With the 10xGenomics platform, each droplet includes a bead containing oligonucleotides with a unique barcode. As with UMIs, during reverse transcription of the mRNA to cDNA barcodes are added to the heavy and light sequences. In this way, the appropriate heavy and light chains can be paired up by matching the barcodes. Single cell methods are considerably more expensive and time consuming than bulk methods (Kim and Park 2019). As a result, most single-cell approaches use lower numbers of cells ($<2 \times 10^4$ cells, compared to up to hundreds of thousands for bulk methods) (Langerak 2022), though a recent study sequenced 1.2×10^6 single B cells (Jaffe et al. 2022).

1.2.3 Linking sequences to antigen information

BCR datasets typically lack specific antigen information; whilst a repertoire from an antigen-immunised animal likely contains antibodies binding to that antigen, actual binding information for individual antibody sequences is not known. Recent advances in sequencing have enabled the linking of BCR sequence data to antigen specificity (Setliff et al. 2019). The high-throughput LIBRA-seq (LInking B cell Receptor and Antigen through sequencing) platform links paired BCR sequence information to antigen binding through use of oligo-barcode labelled antigens. B cells are exposed to labelled antigens in a screening library, and a LIBRA-seq score is calculated based on

the number of antigen barcodes associated with the cell.

LIBRA-seq was experimentally optimised by Hurtado et al., to improve upon the initial sequence recovery rate of 6-22% to >25% and to use next-generation antigen barcodes which bind site-specifically (Hurtado et al. 2022). The original LIBRA-seq oligonucleotide conjugation was not site-specific so could lead to variation in the number of barcodes attached per antigen, interfering with the binding scoring. Epitopes might also be disrupted by the barcode, affecting binding capabilities (Hurtado et al. 2022).

LIBRA-seq can be combined with ligand blocking where ligand and antigen are each labelled with a barcode. LIBRA-seq scores are calculated for the antigen and for the ligand to assess whether the antibody is able to block the ligand binding to the antigen. This was used to identify SARS-CoV-2-binding antibodies which may block ACE-2-SARS-CoV-2 interactions and therefore have neutralisation capabilities (Shiakolas et al. 2022). LIBRA-seq has also been combined with a screening library of labelled mutated antigens to get residue-level epitope specificity for antibodies against HIV (Walker et al. 2022).

1.2.4 Sequence annotation

Antibody nucleotide or amino acid sequences can be annotated with their corresponding V(D)J genes, sequence numbering, and structural features such as the framework and CDR loops, for downstream analysis and comparison.

1.2.4.1 Germline alignment

Nucleotide or amino acid sequences can be mapped to their closest reference V(D)J genes or alleles using tools such as IgBlast (Ye et al. 2013) or MiXCR (Bolotin et al. 2015), which use germline sequence databases such as IMGT or Genbank (Lefranc et al. 2003). Mapping sequences to germlines can help to suggest antibody functionality as certain germlines have been linked to viral responses, such as V3-53/V3-66 for binding SARS-CoV-2 (Brouwer et al. 2020), VH1-69 for broadly-neutralising influenza-binding (Chen et al. 2019) and VH1-02 for HIV-1 CD4 binding (Navis et al. 2014).

1.2.4.2 Numbering

Antibody sequences can be annotated so that each amino acid in the Fv domain is numbered. The numbering of sequences allows for standardisation and comparison of antibodies; for example, the amino acids present at a given position in multiple antibodies can be compared. Numbering also allows for annotation of the CDR loops in an antibody sequence. There is no single standardised antibody numbering scheme; schemes exist with varied nomenclature and position labelling approaches for accounting for sequences with different lengths (Figure 1.4) and for defining CDR regions (Table 1.1).

The first numbering scheme, Kabat, was introduced in the 1980s (Kabat 1983). Antibody sequences available at the time were aligned and the most common CDR length regions were used to define numbering, with insertions for CDRs beyond these lengths indicated by lowercase letters such as 27a, 27b. Kabat is entirely sequence-based, so residues at the same Kabat-numbered position may not be structurally equivalent.

Chothia and Lesk were the first to describe a structure-based antibody numbering scheme (Chothia and Lesk 1987). Antibody crystal structures were aligned to identify the CDR loops and Kabat CDRH1/CDRL1 insertion points were amended to better reflect the structures (from H35 to H30 and L27 to L30). The CDRL1 insertion point was later revised to L31 (Chothia et al. 1989) before returning to the original L30 in later analysis (Al-Lazikani et al. 1997). Chothia and Lesk also identified canonical classes of CDR loops based on structural clustering (Chothia and Lesk 1987).

The Martin scheme modified and updated Chothia numbering, focusing on correct numbering of varied framework lengths (Abhinandan and Martin 2008). Insertion points and deletion points were again amended, e.g. H72 rather than H82 (Figure 1.4) and CDRL2 insertion point at L52, analogous to the CDRH2 insertion point at H52, rather than at L54 as in Kabat. Chothia numbering did not include a CDRL2 insertion point as all available structures had identical length CDRL2 (Abhinandan and Martin 2008).

The most widely used numbering scheme, IMGT, was introduced in 1997 as a method for numbering antibody and T cell receptor chains (Giudicelli et al. 1997). The method was based on germline V gene sequence alignment so originally only

	E	V	Q	L	Q	E	S	G	P	S	L	V	K	P	S	Q	T	L	S
IMGT	1	2	3	4	5	6	7	8	9	11	12	13	14	15	16	17	18	19	20
Kabat	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Chothia	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Martin	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Aho	1	2	3	4	5	6	7	8	9	11	12	13	14	15	16	17	18	19	20

	L	T	C	S	V	T	G	D	S	I	T	S	D	Y	W	S	W	I	R
IMGT	21	22	23	24	25	26	27	28	29	30	35	36	37	38	39	40	41	42	43
Kabat	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
Chothia	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
Martin	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
Aho	21	22	23	24	25	26	27	28	29	30	31	32	39	40	41	42	43	44	45

	K	F	P	G	N	R	L	E	Y	M	G	Y	V	S	Y	S	G	S	T
IMGT	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	63	64	65
Kabat	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57
Chothia	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57
Martin	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57
Aho	46	47	48	49	50	51	52	53	54	55	56	57	58	59	64	65	66	67	68

	A	Y	N	P	S	L	K	S	R	I	S	I	T	R	D	T	S	K	N
IMGT	66	67	68	69	70	71	72	74	75	76	77	78	79	80	81	82	83	84	85
Kabat	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76
Chothia	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76
Martin	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	72A	72B	72C	73
Aho	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87

	Q	Y	Y	L	D	L	N	S	V	T	T	E	D	T	A	T	Y	Y	C
IMGT	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104
Kabat	77	78	79	80	81	82	82A	82B	82C	83	84	85	86	87	88	89	90	91	92
Chothia	77	78	79	80	81	82	82A	82B	82C	83	84	85	86	87	88	89	90	91	92
Martin	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92
Aho	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106

	A	N	W	D	G	D	Y	W	G	Q	G	T	L	V	T	V	S	S
IMGT	105	106	107	108	115	116	117	118	119	120	121	122	123	124	125	126	127	128
Kabat	93	94	95	96	97	101	102	103	104	105	106	107	108	109	110	111	112	113
Chothia	93	94	95	96	97	101	102	103	104	105	106	107	108	109	110	111	112	113
Martin	93	94	95	96	97	101	102	103	104	105	106	107	108	109	110	111	112	113
Aho	107	108	109	110	111	137	138	139	140	141	142	143	144	145	146	147	148	149

Figure 1.4: Heavy chain antibody sequence numbered using ANARCI according to IMGT, Kabat, Chothia, Martin and Aho schemes.

numbered up to CDR3. Initially, CDR insertions and deletions were placed at the end of the loops, but this was later revised to place them in the centre of the CDRs to better represent loop structures. The IMGT CDR definitions are identical for heavy and light chains, unlike the North and Chothia definitions (Table 1.1). North CDR definitions are longer than those of Chothia and IMGT (aside from CDR3) and are based on structural clustering including consideration of anchor residues at both ends

of the loops (North et al. 2011).

The Honegger numbering scheme (AHo's) extended the original IMGT numbering to include CDR3 and FWR4 (Honegger and Plückthun 2001). As with Chothia, the numbering used structural alignments of the variable domains of antibodies and TCRs.

CDR definition	CDRH1	CDRH2	CDRH3	CDRL1	CDRL2	CDRL3
IMGT	27-38	56-65	105-117	27-38	56-65	105-117
North	24-40	55-66	105-117	24-40	55-69	105-117
Chothia	27-37	57-64	107-117	24-40	56-69	105-117

Table 1.1: CDR definitions for IMGT numbered sequences

1.2.4.3 Antibody numbering tools

Tools for applying the various numbering schemes to antibody sequences exist in the form of webservers and downloadable tools. Approaches rely on identifying similar sequences or sequence patterns to those which have already been numbered. This is achieved through sequence-alignment methods or hidden markov models (HMMs).

ANARCI (Dunbar and Deane 2016) uses a HMM-based approach. V and J gene sequences for heavy chains and for lambda and kappa light chains, for different species, were obtained from the IMGT database (Giudicelli et al. 2005). V and J genes were paired in differing combinations to form potential germline sequences, which were subsequently grouped by species and domain (heavy, kappa or lambda). Sequences were aligned using Muscle (Edgar 2004) and used to build HMMs. Query input sequences requiring numbering are aligned to each HMM using HMMER3 (Eddy 2011). The HMM with the highest alignment score to the input query sequence is used to number the query sequence. Insertions and deletions in the query sequence compared to the HMM family can be accounted for in the scoring, and included in numbering, so exact sequence motif matches are not required (Figure 1.5).

HMM-based approaches were used by other numbering tools prior to ANARCI, including PyIgClassify (Adolf-Bryfogle et al. 2015), ProABC (Olimpieri et al. 2013) and DIGIT (Chailyan et al. 2012). These tools were limited in the numbering schemes

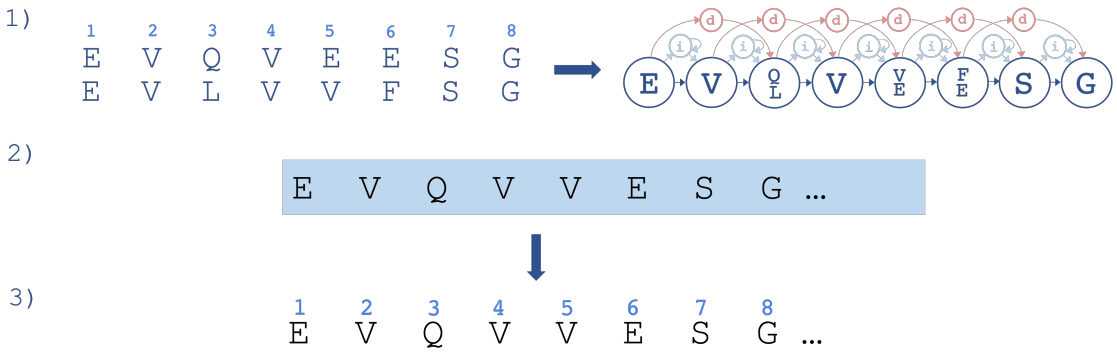


Figure 1.5: Illustration of HMM-based numbering approaches such as ANARCI. 1) HMM profiles are built from aligned germline sequences. 2) Input sequence is scored against HMM profiles, accounting for insertions (i) and deletions (d) at any position in the sequence. 3) Numbering is transferred from the highest scoring HMM to the input sequence.

they could apply, whereas ANARCI is able to apply IMGT, Chothia, Martin and Kabat schemes. Only ANARCI and PyIgClassify could be freely downloaded; ProABC and DIGIT webservers were formerly freely available, and ProABC-2 is currently freely available via a webserver (as of November 2022).

Alternative approaches to HMMs include AbNum (Abhinandan and Martin 2008), a method based on pattern matching to only certain areas of the sequence. Framework regions at the start and end of each CDR are used as anchor points, from which the numbering can be filled in. These anchor points rely on sequence conservation in the framework regions and consider the frequency of the amino acids found in six residues before and after the CDR loops. Profiles of these anchor points are created using multiple sequence alignment of known numbered antibody domains, and profiles are scored against to obtain the best pattern match and infer numbering. If these anchor points cannot be found in an input sequence, AbNum is unable to number the sequence.

AbRSA is a more recent method for antibody numbering based on sequence alignment which divides a sequence into segments of CDRs, framework, insertion positions and conserved positions. Different weightings are given to the different segment types, with higher weightings for more conserved regions including the framework. On a benchmark dataset of 983 heavy chains and 833 light chains, the AbRSA authors found that AbRSA could number all the sequences, ANARCI failed on 8 and AbNum

failed on 37 (Li et al. 2019).

In Chapter 2, I describe my development of an alternative approach to antibody numbering, which makes no use of sequence alignment or HMMs but instead uses sequence embeddings.

1.2.5 Sequence embeddings

Natural language processing approaches have been applied to proteins, where each amino acid in a protein sequence can be considered a word in a sentence (Alley et al. 2019; Rives et al. 2021). Protein language models can be used to obtain embeddings, where sequences are represented as vectors of numbers instead of strings of amino acids or nucleotides. Such representations can be used for tasks including structure and function prediction.

Evolutionary Scale Modeling (ESM-1b) is a transformer-based protein language model (Rives et al. 2021) pre-trained on 27.1 million protein sequences from UniRef (Suzek et al. 2015). Transformers can model long range dependencies, so capture interactions between residues not together in sequence space, for example, cysteines which form disulphide bonds and are close in structural space but not the in linear sequence. The ESM-1b model, of 33 layers and 650M parameters, represents each amino acid in a sequence by a vector of 1280 values.

Language models have also been trained exclusively on antibody sequences and can be used to predict B cell types, structures, and immunogenicity. DeepAb, an LSTM for antibody structure prediction, was pretrained on 100,000 paired antibody sequences from OAS (Ruffolo et al. 2022b), whilst humanisation tool Sapiens, consisting of two four-layer transformer models, was trained on 20 million heavy chain sequences and 19 million light chain sequences from OAS (Prihoda et al. 2022). AntiBERTa, a 12 layer transformer model, was pretrained on 42 million heavy chain sequences and 15 million light chain sequences (Leem et al. 2022) and used for antibody binding residue prediction. AbRep was developed to generate sequence embeddings for AbLang, to complete antibody sequences if regions have been missed in sequencing (Olsen et al. 2022c). With these models, sequences are represented differently; AntiBERTa and AbLang embeddings use 768 values per amino acid, compared to 128 for Sapiens.

1.3 Antibody structural data

1.3.1 Methods for determining antibody structure

The structural antibody database (SAbDab) (Dunbar et al. 2014; Schneider et al. 2022b) is regularly updated with antibody structures from the protein data bank (PDB) (Berman et al. 2000). SAbDab contains over 6700 solved antibody structures, including over 5250 solved antibody-antigen structures (as of November 2022). Antibody structures can be determined using X-ray crystallography, electron microscopy and nuclear magnetic resonance (Abbott et al. 2014). Whilst these methods may give high resolution structural information, they can be difficult and time-consuming. Accuracy depends on the experimental conditions and antibody properties (for instance, CDRH3 length) but is estimated at 0.6 Å for framework regions and 1 Å for CDR loops in crystal structures (Eyal et al. 2005). Given the difficulties in obtaining antibody structures, prediction of likely structures based on an antibody sequence is a vital tool.

1.3.2 Antibody modelling

1.3.2.1 Homology modelling approaches

Initial approaches to predicting antibody structures relied on homology modelling, where solved structures with sequence similarity to the target antibody are used as templates. Whilst the increasing numbers of available solved antibody structures have improved the accuracy of homology modelling methods, approaches are still limited by finding appropriate template structures, which is not always possible. Moreover, sequence-similar antibodies which are identified as templates may in fact have dissimilar structures. The general method for homology modelling of antibodies is as follows: templates are selected for the heavy and light chain framework regions and are orientated; templates are selected for the CDR loops and grafted onto the frameworks; finally, side chains are remodelled (to account for residue differences to the selected homology templates). Given the sequence and structural conservation of framework regions, these regions can typically be modelled well. CDRH3 is the trickiest CDR to model given its lack of canonical classes and diversity in sequence and structure, particularly for long length CDRH3s where fewer structures have been solved (Regep et al. 2017).

Tools for antibody homology-based modelling are available as webservers and/or downloadable tools, e.g. ABodyBuilder (Leem et al. 2016), AbPredict2 (Lapidoth et al. 2019), PIGS (Marcatili et al. 2008), RosettaAntibody (Weitzner et al. 2017). ABodyBuilder is currently one the fastest homology-based methods, taking around ∼30s per model (on one CPU) (Leem et al. 2016), and was used to build models of SAbDab antibodies in Chapter 3 and SARS-CoV-2-binding antibodies and malaria-binding antibodies in Chapter 4.

Following the general method for homology modelling described above, ABodyBuilder first selects sequence-similar antibody framework templates and orientates them. This orientation constrains the selection of CDR templates by FREAD (Choi and Deane 2010), which ranks loop templates using an environment-specific substitution score (ESS) based on loop length, sequence similarity, backbone dihedral angle compatibility and loop graftability onto the framework template. Loops are modelled in the order CDRL2, CDRH2, CDRL1, CDRH1, CDRL3, CDRH3, to prevent clashes. Following the grafting of the loops onto the frameworks, side chains may be remodelled using Position-dEpendent Antibody Rotamer Swapper (PEARS) (Leem et al. 2018). PEARS samples potential conformations for amino acid side chains from a library based on antibody residues observed at that IMGT position; other antibody modelling tools use rotamer libraries from general proteins, based on backbone geometry.

Homology modelling tools may use *ab initio* modelling of CDRH3 in cases where no suitable homology template is found, such as in ABodyBuilder (Leem et al. 2016), or as the default approach, as in RosettaAb (Weitzner et al. 2017). Most *ab initio* approaches generate decoy structures and rank them to obtain the most suitable option. If no CDRH3 templates score above the ESS threshold, ABodyBuilder can use Sphinx, a hybrid method which uses structural fragments adjusted to match the desired loop length (Marks et al. 2017). *Ab initio* modelling takes far longer than homology-based methods; using Sphinx increases model building time to 14 hours. RosettaAntibody uses *ab initio* modelling of the CDRH3 loop using kinematic loop closure (Stein and Kortemme 2013) and homology modelling for the other CDRs. This takes around 1,000 CPU hours for a full Fv antibody model (Weitzner et al. 2017).

1.3.2.2 Deep learning approaches

Protein structure prediction has progressed hugely in recent years with the advent of AlphaFold (Jumper et al. 2021). General protein modelling tools can be used for antibody structure prediction, with current best performance from protein complex structure prediction tool AlphaFold Multimer (AFM) (Evans et al. 2022). AFM uses a multiple sequence alignment module (MSA) to generate an embedding of a protein sequence based on similar sequences identified in a genetic database. The MSA embedding, plus identified structural templates, are passed through a transformer (Evoformer) to the structure module which generates the structure. The structure can be passed back to the Evoformer for iterative refinement of the modelled structure. Whilst AFM has shown high accuracy for antibody structure prediction, the latest antibody-specific modelling tools are far quicker and have slightly better accuracy (Abanades et al. 2022b).

CDR loop prediction in ABodyBuilder (either using homology templates or *ab initio*) can be replaced by the deep learning method ABlooper (Abanades et al. 2022a). ABlooper uses five equivariant graph neural networks (EGNNs) for CDR loop structure prediction. The five EGNNs separately predict loop structures, which are averaged to produce the final loop structure. This approach means a model quality confidence score can be estimated based on the diversity between the five predictions. ABlooper was trained on CDR loop backbone atoms, plus two anchor residues either side. Each atom is represented as a node and contains information including the one-hot encoded amino acid identity, atom type, CDR loop, and position to the anchor residues. Predicted loop structures can be relaxed using OpenMM (Eastman et al. 2017) to ensure the loops are physically plausible. ABlooper is fast, with 100 unrelaxed CDR loops predicted in <5 seconds on one GPU, or 30 seconds for an antibody structure with relaxation and side-chain grafting. ABodyBuilder with ABlooper was used to generate models of SARS-CoV-2-binding antibodies and malaria-binding antibodies in Chapter 4.

ABodyBuilder2 is one of the most recent deep learning approaches for antibody structure prediction (Abanades et al. 2022b). On a test set of unseen structures it had an average RMSD of 2.81 Å across CDRH3 (compared to 2.90 Å with AFM, 3.34 Å with ABlooper and 3.46 Å with ABodyBuilder) and takes 5 seconds per structure on a GPU. ABodyBuilder2 is adapted from the structure module of AFM, where protein embeddings are used to predict the 3D structure. Due to not requiring a MSA

component, the method is at least 100x faster than AFM (Abanades et al. 2022b). As with AFM, ABodyBuilder2 predicts all atom models, including side chains, unlike backbone-only methods such as ABlooper (which uses PEARS and OpenMM for side chains).

Another recent method for antibody structure prediction is IgFold (Ruffolo et al. 2022b). This method is based on embeddings of antibody sequence and structure, and graph neural networks. IgFold predicts backbone coordinates and uses Rosetta for side chain prediction (Leman et al. 2020). The method can predict CDRH3 with 3.34 Å RMSD, the same as ABlooper, but generated a number of cis-peptide bonds and steric clashes in an antibody test set (Abanades et al. 2022b).

1.3.2.3 Model quality metrics

When crystal structures are available, RMSD comparisons of a model to the solved structure are useful as an accuracy metric. When solved structures are not available, it is important to consider other metrics for model evaluation. Deep learning models in particular may predict conformations that are physically impossible (Abanades et al. 2022b). Models can be evaluated for physical bond lengths, steric clashes, chirality and stereochemistry; TopModel is a recent model quality assessment tool which will evaluate models for these characteristics (Fernández-Quintero et al. 2022).

1.3.3 Repertoire structural analysis

1.3.3.1 Repertoire modelling

Despite the advances in modelling antibody structures, current tools are still too slow for building models of entire BCR repertoires. The ability to model millions of antibodies would be useful to evaluate structural convergence within or across different repertoires, including in response to disease or vaccination, and to develop approaches for large-scale screening of antibodies. A homology-based method, RepertoireBuilder, was developed specifically for large-scale repertoire data analysis (Schritt et al. 2019). The method uses MSA to increase the speed of template selection; however, the method takes 10s per model per CPU (slower than many of the more recent antibody-specific deep learning methods) and is only available as a webserver, with jobs limited to 10,000 models.

1.3.3.2 Repertoire structural annotation

An alternative to modelling is to instead use structural annotation of antibody sequences to capture key information. For example, rapid prediction of likely CDR conformations, without building a full antibody model. The SAAB method mapped sequences to structures in the PDB based on sequence similarity (over whole sequences and across individual framework and CDR regions) (Krawczyk et al. 2018). SAAB used an adapted version of FREAD to identify the closest CDR loop template structures - essentially, the first step of homology modelling. SAAB+ built on this annotation approach to use CDRH3 FREAD-assigned templates (Choi and Deane 2010) clustered based on RMSD structural similarity and SCALOP-based CDR canonical class prediction (Wong et al. 2019a); using antibody sequences, SCALOP predicts the canonical class of CDRs by clustering with known antibody crystal structures. SAAB+ was used to evaluate the diversity of antibody repertoires in human and mouse (Kovaltsuk et al. 2020) and showed differentiation of B cell types by CDR template structural usage, with naive B cells using more conserved structures compared to antigen-exposed B cells, and different structures across human and mouse. SAAB+ can be used to compare repertoires to identify over-represented CDRH3 templates or clusters of templates which may represent a response to an antigen and therefore be useful in therapeutic antibody design.

1.3.3.3 Repertoire structural profiling and antibody model libraries

The repertoire structural profiling (RSP) pipeline uses full antibody models to understand the structural space of repertoires (Raybould et al. 2021b). BCR repertoire data can be clustered based on similar predicted structures using RSP to create antibody model libraries (AMLs). AMLs represent the structural diversity of a BCR repertoire, as a library of common antibody structures. Heavy and light chain antibody sequences are first numbered using ANARCI (Dunbar and Deane 2016) and assessed for modellability using SCALOP to predict canonical CDR forms (Wong et al. 2019a) and FREAD to identify appropriate CDR template structures for homology modelling (Choi and Deane 2010). If SCALOP cannot predict a canonical form, or FREAD cannot find a suitable homology template, modelling the antibody is likely to fail and/or the accuracy of the model is likely to be low. To reduce numbers (and corresponding computational resource requirements), similar sequences in each BCR repertoire can first be clustered using CD-HIT (Fu et al. 2012) before modelling assessment. To predict the modellability of orientations of the heavy and light

chains, the twenty most important residues involved in this interface are identified for pairs of VH and VL and compared to template antibodies. If this comparison suggests that they are modellable, the antibody structures can be clustered based on the VH-VL orientation template, FREAD-defined CDR templates and lengths of CDRs. Singleton clusters are designated distinct structures; other structures are greedily clustered in CDR-length groups if the VH-VL orientation RMSD <1.5 Å. An AML built using this approach was used in Chapter 3 and the RSP pipeline influenced the computational epitope binning method developed in Chapter 4.

1.4 Experimental characterisation of antibody-antigen interactions

Knowledge of the epitope being targeted by an antibody, and the residues involved in binding, is important in understanding antibody mechanisms of action and the immune response (Nilvebrant and Rockberg 2018). Antibodies which bind the same epitope often (though not always) have similar functionality (Mahita et al. 2022) and so epitope identification may be predictive of function. Antibodies with similar paratopes may bind to similar epitopes, therefore, identifying antibodies with similar paratopes may be indicative of function (Ghanbarpour et al. 2022; Mahita et al. 2022; Richardson et al. 2021; Wong et al. 2021).

Experimental methods to determine paratope-epitope interactions can be high resolution but are often expensive, laborious, and time consuming (Sela-Culang et al. 2013). As a result, these methods are primarily used later in therapeutic antibody development on a limited number of lead candidates to confirm binding interactions, rather than to influence early development and antibody selection. Faster, lower resolution experimental approaches, such as epitope binning, can be used to reduce the number of lead candidates ahead of higher resolution epitope mapping techniques (Matharu et al. 2021).

1.4.1 Paratope-epitope interactions

High throughput experimental methods for determining an antibody's epitope generally involve testing an antibody for binding against a set of antigens or against different parts of an antigen. For instance, using a peptide microarray where peptides are fixed onto a membrane and incubated with antibodies. A fluorescently labelled secondary antibody is applied and the resulting fluorescent signals used to

locate bound antibody-antigens, which can be sequenced (Henkel et al. 2016). A similar process can be achieved with expressing peptides on E.coli or phage display rather than fixing to a membrane (Gershoni et al. 2007). Surface plasmon resonance (SPR) approaches do not require the secondary antibody step described above; instead, antibodies are fixed and antibody-antigen binding is detected through changes in light (Matharu et al. 2021). Despite the methods being relatively high-throughput, creating a library of antigens or parts of antigens to test the antibodies against can be time consuming and often best suited to linear epitopes (sequential amino acids) rather than conformational epitopes (3D structure) (Gershoni et al. 2007).

Lower throughput technologies can provide more detailed information on epitopes and the specific residues involved in antibody-antigen binding. Such methods include structural determination of the antibody-antigen complex or site-specific mutagenesis of the epitope to assess the impact on antibody binding (Paes et al. 2009). These techniques require knowledge of the epitope and may be used at later stages in antibody drug development to confirm a lead candidate. If solved structures of antibody-antigen complexes are available, paratope and epitope residues can be identified based on proximity to one another. Binding interactions between these residues, typically involving the side chains, include electrostatic connections, hydrogen bonds, van der Waals bonds, and hydrophobic interactions (Peng et al. 2014; Ramaraj et al. 2012). Arpeggio (Jubb et al. 2017) can be used to classify paratope-epitope interactions in solved antibody-antigen structures into 15 subtypes based on atom type, distance and angle terms. Arpeggio was used in the analysis of SARS-CoV-2 binding antibodies in Chapter 4.

1.4.2 Competition epitope binning

Competition-based epitope binning experiments group antibodies which cannot simultaneously bind to the same epitope. This can be useful for identifying antibodies which target a range of epitopes (fall into separate epitope bins) and may have differing mechanisms of action, to create a diverse selection of candidates to take forward in drug development (Matharu et al. 2021). The main limitation of this approach is speed, as antibodies are tested in a pairwise manner and therefore the approach does not scale well; systems such as Carterra^{TM}, which can test 384 antibodies at once, are still time consuming if used on large datasets. Such methods may also not work well with certain antigens such as membrane receptor targets (Chan et al. 2018).

These methods do not provide any information on the exact epitope location, epitope residues or antibody binding mode, just that the two antibodies may overlap in binding sites (or may be non-overlapping but block each others binding in some way). Analysis of the resulting competition binding information, to separate antibodies in to epitope bins, can be complex.

1.5 Computational methods for prediction of antibody-antigen binding

Accurate prediction of which antibody will bind to which antigen, or the likelihood of a paratope binding to a given epitope, is very challenging. Computational methods, similarly to experimental approaches, may be high throughput but low resolution, or low throughput but higher resolution. Methods can be adapted from general protein-protein interaction tools, however, antibody-antigen interactions are generally more difficult to predict than general protein-protein interactions due to limited datasets of known antibody-antigen pairings, high diversity of antibody sequences (compared to other protein families), and variety in antigens and epitopes (Hummer et al. 2022).

1.5.1 Paratope prediction

Computational prediction of the antibody residues which form the paratope is typically done on an epitope-agnostic basis with a binary classification of whether a residue is likely involved in antigen binding, for example with tools such as iPatch (Krawczyk et al. 2013), proABC (Olimpieri et al. 2013), Parapred (Liberis et al. 2018) and Paragraph (Chinery et al. 2023). Original methods for paratope prediction focused only on sequence information but with advances in antibody modelling, approaches using structural information have been developed.

Parapred is a machine learning algorithm for paratope prediction, given antibody sequence data as input. The algorithm uses recurrent and convolutional neural networks trained on data from 277 antibody-antigen crystal structures to predict the probability of each amino acid residue in the CDR regions + 2 adjacent residues participating in binding to an antigen (Liberis et al. 2018). The algorithm was the best performing paratope predictor at the time of most of the research conducted in this thesis, and was used for paratope prediction in Chapter 3.

Paragraph is a recent structure-based antibody paratope predictor which utilises equivariant graph neural network layers and uses modelled structures antibodies (Chinery et al. 2023). Paragraph outperforms PECAN (Pittala and Bailey-Kellogg 2020), the previous best performing structure-based paratope predictor, and Parapred. Paragraph was used for paratope prediction in Chapter 4.

1.5.2 Epitope prediction

As with paratope prediction, epitope prediction can be sequence-based or also include structural information, with tools such as DiscoScope2.0 (Kringelum et al. 2012), BEpIpred-2.0 (Jespersen et al. 2017), DLBEpitope (Liu et al. 2020c) and EpiDope (Collatz et al. 2021). Antibody epitope prediction is more challenging than paratope prediction due to difficulties in differentiating epitope from non-epitope, given the ability of antibodies to bind to highly varied antigens and sites and as epitope residues may not be close in sequence space (but close in structural space) (Kringelum et al. 2012). Methods including antibody information appear to perform better than antibody-agnostic methods (Dai and Bailey-Kellogg 2021; Jespersen et al. 2019) but this limits the use of the methods as antibody information may not be known.

1.5.3 Docking

Even if the antibody paratope and antigen epitope have been predicted, this does not necessarily provide information on paratope-epitope binding or the relative orientations of the molecules. However, this information can be used to constrain molecular docking of models of the antibody and antigen. Molecular docking is a tool to predict the most likely binding mode of a protein and a ligand (or antibody and antigen), based on minimising the free energy of the bound complex. Docking approaches can be flexible, where antibody and antigen can change conformation during binding, or rigid-body, whether the molecules are treated as rigid objects which cannot change shape. The latter is much faster, but less accurate, with any inaccuracies in modelled antibody structures, or conformational change on antigen-binding, reducing the likely accuracy of docking results (Hummer et al. 2022). Docking is not suitable for large scale antibody datasets as it is very resource intensive (Hummer et al. 2022).

Docking can be combined with experimental approaches, for example, docking antibodies selected from clusters identified through experimental epitope binning, before conducting higher resolution examination of the epitope regions predicted through

docking using site-directed mutagenesis (Brooks et al. 2020). Experimental knowledge of the antibody-antigen binding interface or results from epitope mutagenesis experiments can increase the accuracy of the docked poses (Ambrosetti et al. 2023; Wollacott et al. 2019). However, methods requiring experimental steps, rather than being entirely *in silico*, have associated experimental and resource costs. Methods for improved the ranking of docked poses have also been developed to attempt to improve docking accuracy (Schneider et al. 2022a). Whilst docking can be a very useful addition to experimental work, it is not currently feasible to run for many antibodies early in drug discovery.

1.5.4 Computational epitope binning

Prediction of antibody-antigen binding can be approached as grouping antibodies based on similar properties which suggest they bind to the same epitope. In this way, if binding information of some antibodies is known, binding information can be inferred about any antibodies clustered with them. Sequence-based methods such as clonotyping (Galson et al. 2015; Greiff et al. 2015b; Hershberg and Luning Prak 2015; López-Santibáñez-Jácome et al. 2019; Miho et al. 2018) and paratyping (Richardson et al. 2021) can be applied to large number of antibodies but are limited by requiring sequence similarity or same CDR lengths. On the other hand, methods incorporating structural information can group sequence-dissimilar antibodies which might bind to the same target, such as Ab-Ligity (Wong et al. 2021), but require more time and computational resource to build antibody models.

Accurate computational epitope binning methods would be highly valuable for grouping large numbers of antibodies much faster than could be achieved experimentally. They would also be useful for challenging targets, such as membrane antigens and unstable reagents, where functional assays are challenging.

1.5.4.1 Sequence-based clustering

Clonotyping is the standard sequence-based approach for clustering antibodies into 'clones'. The aim of clonotyping is to identify sets of antibodies that are the progeny of a common B cell ancestor. These sets of sequences are then presumed to bind to the same epitope. A clonotype is typically defined as a set of sequences with identical V and J genes, identical CDR3 length and 80-100% CDR3 sequence identity (Galson et al. 2015; Greiff et al. 2015b; Hershberg and Luning Prak 2015; López-Santibáñez-Jácome et al. 2019; Miho et al. 2018). However, sequences which are similar may not

have the same function due to differences in shape and physicochemical properties, and vice versa where antibodies may have dissimilar sequences but similar functionality based on overall 3D structural similarity. Despite this, clonotyping can be useful in a hit expansion step (see Section 1.1.5.2) as it will always identify very similar sequences that are likely to share similar properties, but may miss out on the more sequence-diverse antibodies that bind to the same epitope.

Paratyping is a sequence-based method for clustering similar paratopes (Richardson et al. 2021). The method originally used Parapred for paratope prediction (Liberis et al. 2018) (see Section 1.5.1) but can be used with more recent paratope prediction tools such as Paragraph (Chinery et al. 2023). Paratyping clusters paratopes based on paratope amino acid identity and identical CDR lengths. The approach means antibodies with different lineages (with dissimilar CDR3 sequences, V and J genes) can be identified as similar and therefore as likely binding to the same epitope - a link which is not possible to make with clonotyping. Paratyping has been experimentally validated by predicting and testing antibody hits identified against pertussis toxoid (Richardson et al. 2021).

A recent method for paratope clustering used a fine-tuned protein language model for sequence-based antibody paratope prediction (Ghanbarpour et al. 2022). The language model embeddings for predicted paratope residues were used to cluster SARS-CoV-2-RBD binding antibodies, based on the Euclidean distance between their paratope vectors (Ghanbarpour et al. 2022).

1.5.4.2 Structure-informed clustering

Many standard methods for exploiting BCR data use only antibody sequence information; however, making paratope-epitope interaction predictions is challenging due to the lack of understanding of the three-dimensional shape of the antibody binding site which underpins its potential binding capabilities. Furthermore, existing sequence-only methods such as clonotyping typically only use part of the antibody sequence information (CDRH3 sequence, V and J gene assignment). If models can be built, structural clustering with known-function antibody datasets, such as CoV-AbDab (Raybould et al. 2021a), can predict the function of uncharacterised antibodies identified from BCR repertoires (Finn et al. 2020; Raybould et al. 2021c).

Docking information can be used for epitope clustering of antibody structures, through throughput is limited due to the molecular docking stage. For instance, docked models were used to group IL-6 binding repebodies (binding scaffolds of repeated leucine-rich modules) (Mahita et al. 2022). Repebodies were clustered based on sequence similarity, structural similarity and interaction similarity, derived from the docked models. 'Dock binning' utilised existing experimental epitope binning data to select epitope bin representatives to homology modelling and dock (Brooks et al. 2020). Clustering of the docked models was used to identify epitope regions on the herpes glycoprotein antigen structure.

Ab-Ligity is a method for identifying epitope or paratope similarity which incorporates structural information (Wong et al. 2021). The method is an antibody-specific development of Ligity (Ebejer et al. 2019), applying a hashing approach to compare pharmacophoric and structural profiles of antibodies using residue-based descriptors (Figure 1.6). Paratope residues are predicted, if comparing models of antibodies, or paratope and epitope residues can be identified through proximity to the antibody/antigen in the case of co-crystal antibody-antigen structures. All possible triplet combinations of the residues are taken and used to characterise the antibody via hashcodes. The residues are tokenised by classification into physicochemical groups, for example, acidic or hydrophobic. Every combination of two physicochemical tokens within a triplet has a different hashcode, for example, a pair of aliphatic residues are designated 'A' whereas an aliphatic-aromatic pair are 'B'. Distances are calculated between the centre of the C atoms of tokens and are put into bins of 1 Å in size to give a hashcode (e.g., 'a'). Each triplet is described by a hexamer hashcode, including the three hashcodes from the physicochemical token pairings, alphabetically, and the three bin lengths describing the distances between the token pairings. Frequency tables of the hexamer hashcodes representing all triplets of binding residues represent the antibody, and these are compared using a Tversky index to determine similarity:

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

S(X,Y) is the similarity between two antibody hash tables, X and Y. $|X\cup^{Y}|$ is the intersection (overlap of hashcodes in the hash tables) and $|X \cap {}^{Y}|$ is the union (total hashcodes included in either X or Y or both). |X - Y | and |Y - X| are the differences between the two hash tables; we used $\alpha = \beta = 0.5$.

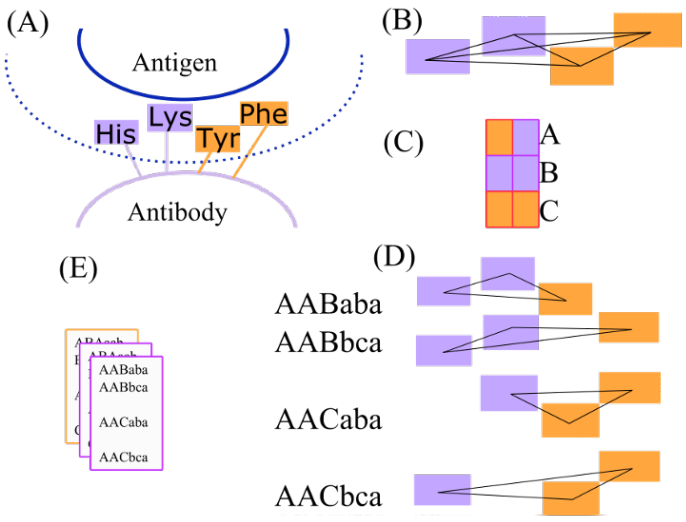


Figure 1.6: Ab-Ligity method for paratope comparison A) Paratope residues are tokenised based on physicochemical properties (histidine and lysine are basic, shown in purple; phenylalanine and tyrosine are aromatic, shown in orange). B) Residues are grouped into triplets and distances between residues are hashed into 1 Å bins. C) Pairs of residues within triplets are given a hashcode based on their physicochemical group. D) For each triplet of residues, physicochemical hashcodes are combined with distance hashcodes to generate a six-character hashcode. E) A paratope is represented by all of the hashcodes, stored in a frequency table. Figure reproduced from Wong et al. 2021.

I contributed to the development of Ab-Ligity by assessing the impact of adding a CDRH3 length constraint to the similarity assessment and assisted with the preparation of the Ab-Ligity paper, on which I am an author:

Wong W.K., Robinson S.A., Bujotzek A., Georges G., Lewis AP., Shi J., Snowden J., Taddese B., Deane C.M. (2021) Ab-Ligity: identifying sequence-dissimilar antibodies that bind to the same epitope. MAbs (Vol. 13(1)).

1.6 Impact of BCR data on therapeutic antibody development

Methods for obtaining BCR repertoires and binding affinities have been varied and lacking in standardised protocols and analysis pipelines, and many existing BCR sequencing datasets contain only heavy chain information. The development of high throughput methods for single cell sequencing and antigen specificity mapping (Setliff et al. 2019), and protocols and guidelines (Breden et al. 2017; Langerak 2022) might

enable new accurate and reliable computational methods for biotherapeutic discovery. As increasing quantities of high quality antibody sequencing data become available, the task becomes how to best utilise and learn from the data. The development of predictive computational tools trained on BCR data could optimise the therapeutic antibody development pipeline by reducing the number of antibodies requiring expression and experimental screening to only those predicted most likely to bind, and further our ability to design antibodies *de novo* against a specific target (Marks and Deane 2020).

1.6.1 Immune system mechanisms

Typical BCR repertoire profiling includes analysis of B cell isotypes, gene usage and clonality (see Section 1.5.4.1). The availability of large BCR datasets has been useful across many areas of the immune system underlying therapeutic discovery including estimating true antibody genetic diversity (Briney et al. 2019), VDJ recombination bias (Greiff et al. 2017) and characterising the response to antigens (Miho et al. 2018).

BCR datasets can be used to understand more about the immune response, for example, before and after vaccination (Galson et al. 2014) and during disease (e.g. autoimmunity (Bashford-Rogers et al. 2018) and Covid-19 (Montague et al. 2021)). By comparing healthy and disease-related repertoires, we can learn about the role of B cells and identify potential disease signatures (Bashford-Rogers et al. 2018; Ghraichy et al. 2018). The role of antibodies in disease pathogenesis, such as autoimmune disease, is also being understood (Sun et al. 2020). Such analyses may indicate targets for potential therapeutic intervention and, in the future, hopefully improve patient outcomes with better diagnostic tools, tracking disease progression and monitoring response to therapies.

1.6.2 Biotherapeutic discovery

BCR sequence repertoires can be used to suggest suitable candidates for drug development. In fact, 16 clinical stage therapeutic antibodies were found to share >95% identity with a BCR sequence in OAS (Krawczyk et al. 2019). Individuals are known to share sequence- and structural- antibody similarity, with sequence and structural convergence identifiable in unrelated individuals both in unstimulated repertoires (healthy individuals) (Briney et al. 2019; Raybould et al. 2021b), and in response to antigen stimulation (Trück et al. 2015). Convergent antibody responses in individuals

with the same disease or who have been have been exposed to the same antigen (either through infection or vaccination) can be a useful starting point for a potential therapeutic. Evidence to support this approach for drug discovery includes vaccine studies (Galson et al. 2014) and more recently in SARS-CoV-2 infected individuals, where convergent antibodies had sequence similarity with identified SARS-CoV-2-binding antibodies (Galson et al. 2020).

1.6.3 Antibody design against target epitope, given known binder

A number of approaches have been used to design new antibodies against a target epitope, if a binding antibody is already known. The binding antibody may be used as a template with the aim of improving binding, specificity or developability properties. In Chapter 3, I describe pipelines for identifying likely binders to a target epitope from an antibody dataset, given a known binder.

Liu et al. designed an epitope-specific antibody to Keap1 by grafting structural interaction patterns from the known Keap1 ligand onto various antibody scaffolds from the PDB (Liu et al. 2017). Whilst this approach developed new binding antibodies, it required a crystal structure and knowledge of the target antigen in complex with a known ligand.

Mason et al. used neural networks to predict likely binding antibodies from sequencing data, using the therapeutic antibody trastuzumab as a starting point (Mason et al. 2021). Mutagenesis libraries were generated from trastuzumab and used to train long short-term memory recurrent neural networks and convolutional neural networks. Of 30 predicted binders that were assessed, all were specific binders to the Her-2 receptor antigen which trastuzumab binds.

Makowski et al. improved the specificity and affinity of existing therapeutic antibody emibetuzumab, known to have some non-specific binding activity (Makowski et al. 2022). Mutagenesis libraries were experimentally generated and tested for affinity and specificity; machine learning models trained on the resulting library could predict the binding classification of sequences.

1.6.4 Antibody design against target epitope, without known binder

The task of *de novo* antibody design to a target epitope, without prior knowledge of a binding antibody, is highly challenging. I describe the development of pipelines for this in Chapter 3.

Akbar et al. aimed to investigate whether epitope-to-paratope translation could be achieved, using Neural Machine Translation models (Akbar et al. 2021). Their approach investigated whether paratope-epitope interactions can be predicted by breaking down the binding into subunits, the 'vocabulary', of structural motifs governing antibody-antigen interactions. 398 paratope and 1594 epitope motifs were identified from 825 antibody-antigen crystal structures, where each motif was represented only by the presence of any interacting residue as X and then the number of noninteracting residues, e.g. X2X, per CDR or framework region. Neural Machine Translation models trained on the 5,327 epitope-paratope motif pairs achieved correct motif prediction 70% of the time, with a median error of 30%; performance was considerably worse when attempting to predict actual sequences rather than motif representations. This method is currently not able to predict if a new antibody will bind to an epitope, but provides some evidence that it might be possible. However, whilst this approach provides evidence that the epitope motif determines the paratope interaction motif, and therefore may be a predictable interaction, an approach only using structure and not incorporating any physicochemical properties of the interacting residues may miss key information.

Baran et al. aimed to computationally mimic VDJ recombination to generate antibodies (Baran et al. 2017). Antibody sequences were designed based on the assembly of antibody fragments using the AbDesign algorithm. Sequences were then modelled and docked against two target antigens. Top docked poses were evaluated to suggest mutations to the antibodies, and 200 mutated antibodies were experimentally tested to yield 3 low-affinity binders that were then optimized using random libraries made with error-prone PCR. Whilst binders were ultimately identified, the computational cost for this approach was huge, at half a billion processing hours, alongside substantial experimental work.

Nimrod et al. successfully designed a binder to cytokine IL-7A using antibodies with some epitope similarity, but no binding, to IL-7A (Nimrod et al. 2018). This

process, termed 're-epitoping', involved first docking 501 unique antibodies from the PDB to the target antigen IL-7A, and then generating antibody libraries based on the top ranked docked antibodies, with up to four expression and selection rounds for each antibody. The libraries were designed by mutating amino acids at positions selected through binding and stability energy calculations, molecular dynamics simulations and visual inspection. Due to the docking and screening of multiple antibody libraries, this was not a quick approach but provides evidence for designing an antibody to a specific epitope using antibodies which bind to similar epitopes.

Aguilar Rangel et al. recently used known antibody-antigen structures to inform antibody design against target epitopes (Aguilar Rangel et al. 2022). A database of CDR structures and CDR-like fragments (linear motifs structurally compatible with antibody CDR loops), and their epitopes, was created from the PDB. Target epitopes were fragmented into linear and structural epitopes and compared to the epitopes in the database using sequence- and structural- similarity. Similar CDR or CDR-like fragments were ranked based on number of favourable interactions with the epitope and solubility score, and the top scoring CDRs grafted onto an antibody scaffold. Using this approach, binding antibodies were generated targeting human serum albumin, SARS-CoV-2 spike protein, and bovine trypsin.

The field of *de novo* protein design has been transformed with the development of novel machine learning and deep learning methods (Strokach and Kim 2022). Deep learning models aim to generate novel antibodies which retain selected properties of the training data, for example, developability or binding affinity. Generative models for antibody design, such as Generative Adversarial Networks (GANs), can be trained on natural human antibody repertoires and biased via transfer learning (further training on antibodies with known properties such as solubility, stability and predicted immunogenicity) to generate sequences predicted to have desired biophysical properties (Amimeur et al. 2020). However, more information is needed to understand how these properties influence the overall developability of the antibody. Additionally, experimental validation of the predicted properties is necessary, as has been done for proteins; a modified GAN trained on malate dehydrogenase enzyme sequences was able to generate functional and soluble enzyme sequences with 24% hit success (Repecka et al. 2021).

Remaining challenges also include the successful development of generalisable models for designing antibodies to a target epitope, without requiring training on datasets of existing binders. Latest generative deep learning approaches based on generative image approaches, diffusion models, appear highly promising, generalisable methods for designing protein sequences to target structures (Cao et al. 2022; Watson et al. 2022). For example, the RFdiffusion model was able to design high affinity binders to 4 targets with fewer than 100 designs per target, rather than thousands of sequences which then require testing (Watson et al. 2022).

1.7 Thesis overview

In Chapter 2, we describe our novel approach to antibody numbering, ANARCII. We use transformer-based sequence embeddings, derived from BCR-data in OAS, to train random forest models. ANARCII can be used to annotate antibody sequences with the IMGT numbering scheme and is able to number sequences which existing numbering tools get incorrect.

In Chapter 3, we explore the development of Ab-Ligity and paratyping to create computational pipelines for identification of antibodies likely to bind to a target antigen, with or without a known binding antibody. We compare antibodies based on sequence, paratope and structural similarity, to identify likely-binding antibodies from an antibody model library. Such a pipeline would be a useful tool in therapeutic antibody development by identifying antibody candidates likely to bind to the target epitope, which could be useful starting point for antibody therapeutic design or experimental screening.

In Chapter 4, we describe our novel computational epitope binning method, SPACE. By analysing X-ray crystal structures of antibodies bound to SARS-CoV-2 antigens, we found high levels of structural conservation across binders to the same epitope. Our method of clustering antibody models based on structural similarity can group sequence-dissimilar antibodies into epitope bins with high accuracy. We demonstrate our method on coronavirus-binding and malaria-binding antibodies.

Lastly, we describe our overall conclusions and suggestions for future work, based on the research presented in this thesis. We also include the following paper as an

appendix:

Macpherson, A., Laabei, M., Ahdash, Z., Graewert, M.A., Birtley, J.R., Schulze, M.S.E., Crennell, S., Robinson, S.A., Holmes, B., Oleinikovas, V. and Nilsson, P.H., (2021). The allosteric modulation of complement C5 by knob domain peptides. Elife, 10.

I am a coauthor on this paper and carried out analysis of the predicted paratopes of two bovine antibody knob domains with ultra long CDRH3s, including characterisation of paratope-epitope interactions using Arpeggio and comparison to antibodies in SAbDab using Ab-Ligity.

CHAPTER 2

ANARCII: A tool for numbering antibody sequences

2.1 Introduction

As described in the first chapter, annotation of protein sequences with numbers enables comparison of structurally-equivalent amino acid residues, without requiring multiple sequence alignment (Olsen et al. 2022a). This is highly useful for antibodies and forms the first step of many analyses, including definition of CDR loops, clonotyping and other sequence-based methods. Existing approaches to antibody numbering rely on identifying similar amino acid motifs to sequence patterns which have already been numbered. This is typically achieved through hidden markov models (HMMs) or, in some cases, sequence-alignment methods (see Introduction, Section 1.2.4.3). In this chapter, we focus on numbering according to the the IMGT numbering scheme (Figure 2.1).

Due to the reliance on amino-acid level pattern recognition, methods can struggle to number sequences which are dissimilar to sequences that have been previously numbered and can misidentify sequence patterns. This may lead to erroneous numbering and discarding of data. Given the increasing ability to engineer and design novel antibody sequences, and to identify rarer sequences within antibody repertoires, a more robust numbering tool would be highly beneficial.

Antibody Numbering and Antigen Receptor ClassIfication (ANARCI) is a tool for numbering antibody sequences and assigning germline sequences (Dunbar and Deane

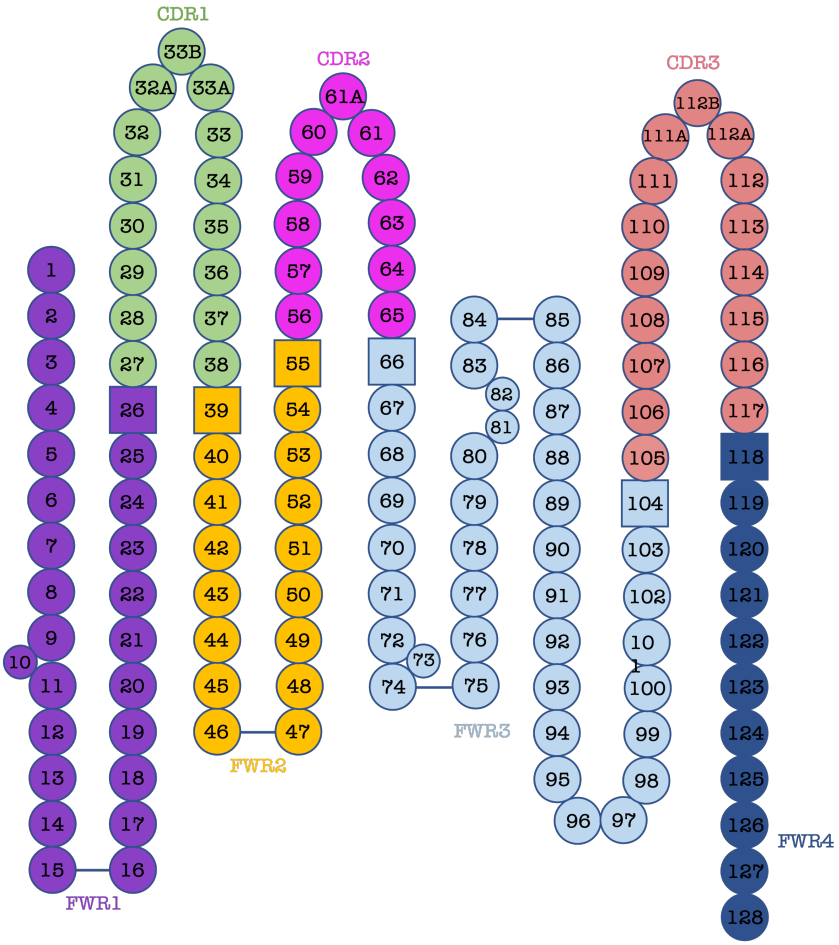


Figure 2.1: IMGT/Collier de Perles representation numbering of an antibody VH/VL sequence (Ehrenmann et al. 2011). IMGT numbering is shown including framework regions (FWR) and complementarity determining regions (CDR): FWR1 (1-26, purple), CDR1 (27-38, green), FWR2 (39-55, yellow), CDR2 (56-65, pink), FWR3 (66-104, blue), CDR3 (105-117, red) and FWR4 (118-128, navy). Framework residues 10, 73, 81, and 82 are absent in human VH germlines and are therefore shown in smaller circles, adjacent to the main numbering. Framework residues flanking the CDR loops are indicated in squares. Longer CDRs are incorporated with additional residues in the centre of the CDR loops (at positions 32/33, 60/61, 111/112), labelled symmetrically with letters A-Z; the illustration includes examples of these.

2016) (see Chapter 1, Section 1.2.4.3). ANARCI builds HMMs from a pre-numbered set of germline sequences, downloaded from IMGT (Giudicelli et al. 2005). Query sequences are aligned to 24 HMMs from six different species using HMMER3 (Eddy 2011). The HMM with the highest scoring alignment to the query sequence is used and the query sequence is annotated by transferring the numbering of the equivalent position from the known sequence in the HMM to the query sequence. HMMs do not

require exact sequence pattern matches and input sequences can be scored compared to the HMM profiles taking into consideration insertions and gaps. ANARCI was the first tool to use all numbering schemes and was able to number 99.5% of VH sequences (Dunbar and Deane 2016). ANARCI occasionally falsely identifies sequence patterns and assigns incorrect numbering, most commonly at the end of CDRH3 in antibodies with long CDRH3s, such as HIV-binding antibodies (Figure 2.2).

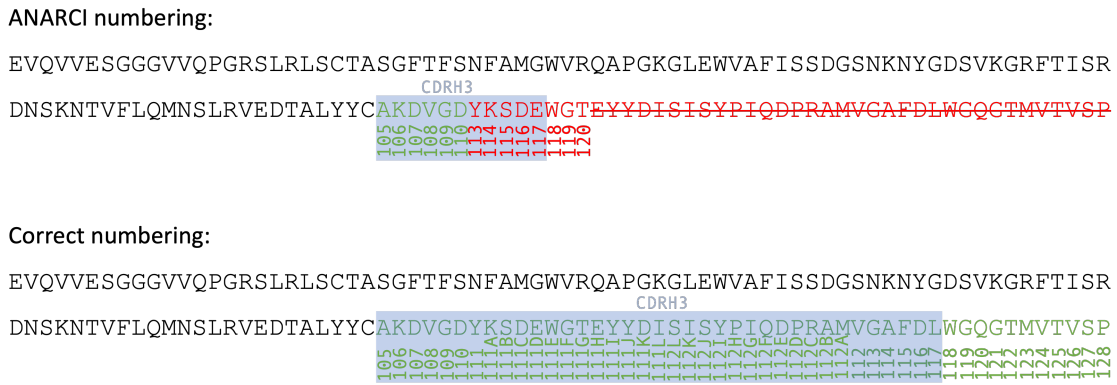


Figure 2.2: Incorrect numbering: ANARCI misidentifies the end of a long CDRH3 loop in an HIV-1 neutralising antibody (PDB code 4ocr). Residues numbered correctly by ANARCI are represented in green and incorrect are shown in red. ANARCI incorrectly numbers the CDRH3 (IMGT residues 105-117, shaded in blue) with a length of 11 by falsely identifying a 'DE' amino acid motif as the end of the CDRH3 loop, and discards the end of the antibody sequence after the incorrectly assigned number 120 (red strike-through). The correct CDRH3 is length 37 and includes number labels up to 111L and 112L.

Other antibody numbering tools include AbNum (Abhinandan and Martin 2008) and AbRSA (Li et al. 2019) (see Chapter 1, Section 1.2.4.3). AbNum uses pattern matching to specific areas of the sequence; framework regions at the start and end of each CDR are used as anchor points to fill in numbering. If conserved sequence anchor points cannot be identified, AbNum is unable to number the sequence. AbRSA is a sequence alignment based method and divides a sequence into segments of CDRs, framework, insertion positions and conserved positions. Different weightings are given to the different segment types, with higher weightings for more conserved regions including the framework.

Huge quantities of antibody sequencing data are now publicly available (Olsen et al. 2022b). Existing antibody numbering methods were not trained on large datasets,

for example, AbRSA was trained on only 503 heavy chains and 475 light chains from UniProt (Pundir et al. 2017). Advances in the development of protein language models have enabled vector representations of antibody sequences, which have been used to separate B cell types and predict structure (Lin et al. 2022; Olsen et al. 2022c) (see Chapter 1, Section 1.2.5). In this chapter, we describe ANARCII - an updated version of ANARCI. ANARCII is a random forest multi-label classifier for antibody numbering, trained on transformer-based representations of antibody sequences. Whilst training on the equivalent of only around 5,000 sequences, we show high accuracy with over 99% correct identification of CDR ends, including for some sequences which ANARCI and AbNum cannot number or number incorrectly.

2.2 Materials and methods

2.2.1 Datasets

Antibody sequence datasets were prepared from the Observed Antibody Space (OAS) database (Olsen et al. 2022b). Sequences available in OAS have been processed and numbered with ANARCI (IMGT numbering scheme). Sequences from all studies, including varying species, disease states and vaccination states, were considered for training and test datasets. Separate datasets were created for heavy (VH), kappa (Vκ) and lambda (Vλ) chains. Sequences shorter than 90 amino acids were excluded to ensure models were trained on whole antibody sequences, rather than sequence fragments.

2.2.1.1 Training data - heavy chain

The aim of training dataset curation was to include as many number labels as possible; most antibody sequences contain standard number labels 1-128 and 200 number labels are known to cover $\sim$99.8% of OAS (Olsen et al. 2022a). To create training datasets, sequences were selected from OAS. Data files from OAS were randomly ordered and sequences within a datafile were randomly shuffled. Sequences within a datafile were added to the sequence dataset if they occurred more than once within the dataunit (redundancy of ≥ 2), contained any insertion residues (not labels 1-128) which did not already feature in the training dataset 1000 times, and if the sequence was not already included in the training dataset. All 13,935 unpaired heavy chain sequence files were iterated through in order to obtain a sequence dataset. Given issues with ANARCI numbering long CDR3 loops, and to reduce sequence dataset size, the dataset was filtered to remove sequences with CDRH3 length >18, to produce a set

of 78,927 unique heavy chain sequences, with 886 unique number labels.

All antibody sequences used in this work were run through ESM-1b (see Methods: 2.2.2.2 and Figure 2.3). Datasets of specific amino acids were curated from the ESM-1b representations of the antibody sequences selected from OAS, rather than using the full sequence representations. Amino acid datasets were generated where up to 5000 instances of each standard number (1-128) and up to 1000 instances of each insertion residue were included (datasets were named 5000_1000 to reflect this). Given the rarity of some insertions, some number labels featured <10 times. The training dataset was curated in this way to avoid a highly unbalanced and large dataset; if whole sequences were used, the training dataset would be large and highly skewed towards number labels 1-128, with most insertion residues present very infrequently, if at all.

Training datasets with alternative balances of standard number labels and insertions were also explored (Table 2.1). Due to each amino acid being represented as a vector of 1280 features, training datasets were large in size for relatively few amino acids. The 5000_1000 dataset included 832127 amino acids from heavy chain sequences, with 886 unique number labels (of which 54 only appeared once) and insertion positions making up 23.9% of the data (n = 199103) (Table 2.2).

Dataset name	Max fr. of labels 1-128	Max fr. of insertion labels	Total no. of amino acids	Size (GB)
2000_1000	2000	1000	453415	2.2
5000_1000	5000	1000	832127	4.0
8000_1000	8000	1000	1202976	6.0

Table 2.1: Heavy chain amino acid training datasets with varying maximum frequencies (fr) of standard IMGT number labels 1-128 and insertion labels, and corresponding sizes when embedded by ESM-1b.

Given the knowledge that 200 number labels cover nearly all OAS sequences (Table 2.3), we also created datasets of ESM-1b representations covering only these 200 positions, with maximum frequency of 5000.

Training Dataset	Unique labels	Amino acids	Labels present once only	Labels present <10 times	Labels present 5000 times
VH	886	832127	54	225	124
Vκ	606	632385	84	251	109
Vλ	593	622469	95	242	112

Table 2.2: Training datasets of amino acids from ESM-1b sequence representations. Datasets contained up to 5000 instances of each number label 1-128 and up to 1000 instances of each insertion residue (IMGT numbering), dependent on number label frequency in OAS. The number of labels which appear fewer than 10 times in the dataset includes those which are only present once.

Region	Insertion Number Labels
FWR1	3A
CDR1	32A, 32B, 33C, 33B, 33A
FWR2	40A, 44A, 45A, 46A, 47A, 48A, 48B, 49A, 51A
CDR2	60A, 60B, 60C, 60D, 61E, 61D, 61C, 61B, 61A
FWR3	67A, 67B, 68A, 68B, 69A, 69B, 71A, 71B, 73A, 73B, 80A, 81A, 81B, 81C, 82A, 83A, 83B, 85A, 85B, 85C, 85D, 86A, 96A
CDR3	111A, 111B, 111C, 111D, 111E, 111F, 111G, 111H, 111I, 111J, 111K, 111L, 112L, 112K, 112J, 112I, 112H, 112G, 112F, 112E, 112D, 112C, 112B, 112A
FRW4	119A

Table 2.3: 200 number labels cover nearly all of the total number labels in OAS, identified by Olsen et al. 2022a. For comprehensibility, only insertion labels are shown here (standard integer number labels 1-128 are included in the 200 labels but not listed in the table)

2.2.1.2 Training data - light chain

OAS contains substantially fewer light chain sequences than heavy chain sequences, as a number of studies only report heavy chain data. Following the same method as described above for selecting heavy chain sequences, a dataset of 17916 kappa sequences was curated, with 606 unique number labels, 84 of which only appeared once (Table 2.2). In our 5000_1000 dataset, 2.4% of amino acids are insertions - significantly lower than for the heavy chain amino acid dataset curated in the same manner. For lambda chains, 18679 sequences were identified with 593 unique number labels, including 95 which appeared only once. Insertions made up 2.3% of the 5000_1000 dataset.

As with the heavy chain data, datasets of only the most common 200 number labels for kappa and lambda were also generated.

2.2.1.3 Test data

Test sets were created of heavy chains, kappa chains and lambda chains. Sets of 10000 unique sequences of length 90+ were randomly selected from unpaired and also from paired datasets within OAS (Table 2.4). Both unpaired and paired datasets were used due to differences in the sequences contained in each set. Heavy chain sequences in the paired dataset were on average 121 amino acids long, compared to 108 in the unpaired VH sequences, 110 for lambda sequences (both paired and unpaired datasets) and 107-108 (for paired and unpaired) for kappa sequences. Heavy chain test sets contained vastly more insertions than the light chains (200 - 1400x more). The unpaired heavy chain sequence test set contained the most insertions; despite this, these sequences were shorter than the paired sequences due to missing parts of FWR1, with only 29/10000 heavy chain sequences containing the number labels 1-5 and fewer than 7% of sequences containing number labels 7-15. This was not the case in the paired data, where 98% of heavy chain sequences included number label 1.

Heavy chain test sets included more unique number labels than light chain test sets. This is partially due to a few long sequences present in the test set; one heavy chain sequence had a very long CDRH3 of length 64, so number labels 112A-112Z and 111A-111Y were included. Thirty sequences included up to 112H. This contrasts with the light chain test set where the longest CDR3 included 112B for kappa chains and 112D for lambda (see Figure 2.1 for IMGT numbering scheme information).

Overlap between training and test datasets was not assessed, as the model training data consisted of individual amino acids rather than full sequences.

2.2.1.4 Nanobodies

All available nanobody structures were downloaded from SAbDab on September 30th 2022. The 656 non-redundant nanobody sequences were numbered using ANARCI and used to test ANARCII.

2.2.1.5 T cell receptors

T cell receptor (TCR) sequences were downloaded from the Immune Epitope Database (https://www.iedb.org/home_v3.php) on 21st December 2022. Non-redundant full

Test dataset	No. of unique number labels	Total no. of insertions	Labels present once	Labels present <10 times
VH - Paired	187	21502	33	43
VH - Unpaired	230	28271	54	83
Vκ - Paired	146	19	19	30
Vκ - Unpaired	165	53	27	43
Vλ - Paired	148	23	13	23
Vλ - Unpaired	147	89	15	23

Table 2.4: Composition of test datasets of 10000 antibody sequences from paired and unpaired OAS data. Total number (no.) of insertions in the dataset refers to the number of amino acids within the dataset which are not 1-128. The number of labels which appear fewer than 10 times in the dataset includes those which are only present once.

length sequences (>90 amino acids) were numbered by ANARCI, to create a dataset of 276 TCRα and 281 TCRβ sequences.

2.2.1.6 Non-immunoglobulin superfamily proteins

To assess the ability of ANARCII to differentiate non-immunoglobulin protein sequences from antibody sequences, ten protein antigens in complex with antibodies were randomly selected from SAbDab (Table 2.5). None of the proteins were from the immunoglobulin superfamily and none were assigned numbering by ANARCI.

PDB code	Protein antigen
1mhp	Chimeric alpha1 integrin I-domain
2atk	KcsA K+ channel
4dqo	HIV-1 V1V2 region
4m3k	Beta-lactamase BlaP from Bacillus licheniformis
6dc3	Respiratory syncytial virus prefusion F-protein
6gs7	Peptide transporter DtpA
6pwc	Beta-arrestin-1
7b18	SARS-CoV-2 spike protein
7su0	CTLA-4
7sue	SARS CoV-2 nucleocapsid

Table 2.5: Non-immunoglobulin superfamily proteins randomly selected from SAbDab

2.2.2 Methods

2.2.2.1 ANARCI

Sequences not already numbered by ANARCI (i.e. not downloaded from OAS) were numbered with the latest version of ANARCI (https://github.com/oxpig/ANARCI) using the IMGT numbering scheme (Giudicelli et al. 2005).

2.2.2.2 Evolutionary scale modeling (ESM)-1b

The evolutionary scale modeling (ESM)-1b protein language model from Facebook is a transformer of 33 layers, 650M parameters, trained on 27.1M amino acid sequences curated from UniRef (Rives et al. 2021). The model was trained on a variety of protein amino acid sequences and is not antibody-specific (see Chapter 1, Section 1.2.5). This unsupervised language model was found to encode structural properties within protein representations, with embeddings of 1280 dimensions per amino acid. The pretrained model was downloaded from Github (https://github.com/facebookresearch/esm).

2.2.2.3 Antibody representations

Antibody sequences from OAS were embedded using ESM-1b, with each amino acid in the sequence represented by 1280 features. From the resulting antibody representations, balanced datasets were curated by selecting amino acid representations corresponding to particular number labels (see Datasets).

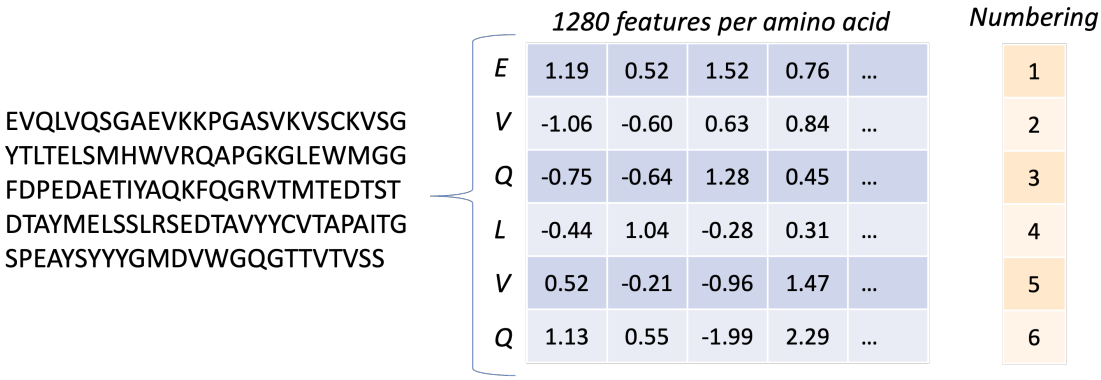


Figure 2.3: Antibody sequence representations from the ESM-1b transformer. Each amino acid is represented by 1280 features. Training datasets included amino acid representations and ANARCI number labels.

2.2.2.4 Multiclassification task

Classifier machine learning models were trained on ESM-1b representations of heavy, lambda and kappa antibody amino acid sequences, and ANARCI numbering. Given the large number of potential numbering labels, due to a high variety of insertions across all of the OAS data, all insertions (non 1-128 numbers) were designated the class 'X'. Further, if the probability of a label prediction was lower than a confidence threshold, the number label was designated as 'U' for uncertain.

2.2.2.5 Random forests

We trained separate random forest models on ESM-1b representations of heavy, lambda and kappa antibody amino acid chains and number labels from ANARCI as downloaded from OAS, using sci-kit learn (Pedregosa et al. 2011). Random forests are collections of decision trees and can be useful machine learning models for classification tasks. Decision trees are supervised learning tools where data is split based on features, to make a prediction. In a random forest, decision trees are built independently from one another and make their own classification predictions. Classification decisions are taken based on the average prediction of all trees in the forest.

Hyperparameter sweeping was conducted to evaluate the impact of number of trees, tree depth, bootstrapping, minimum number of samples required to split a node, minimum number of samples required at each leaf node and class weights. The effect of using the predicted probability of classification, based on the fraction of trees in the forest which predicted the modal label, as a confidence threshold was investigated, and a threshold of 0.6 was selected (unless otherwise stated). Models were large in size, increasing with the amount of training data and number of decision trees in the forest (Table 2.6).

Training dataset	Number of trees	Depth	Size (GB)
2000_1000	100	100	3.4
5000_1000	100	100	6
5000_1000	200	100	12
8000_1000	100	100	7.6

Table 2.6: Overview of random forest models trained on heavy chain datasets (see Table 2.1 for dataset descriptions).

2.2.2.6 XGBoost

We trained XGBoost models (Chen and Guestrin 2016, August 13) on ESM-1b encoded datasets of heavy, kappa and lambda sequences, with ANARCI numbering. XGBoost is a library for implementing gradient-boosted decision trees. Compared to random forests where all decision trees are built in parallel and independently, gradient-boosted decision trees are built sequentially with the accuracy of previous trees used to inform the building of the next tree (so the gradient of results is considered). As with the random forests, hyperparameter searching was used to optimise model performance including number of trees, tree depth and learning rate. Best performance was obtained with 200 tree models, with unlimited tree depth and learning rate of 0.1.

2.2.2.7 Model output processing

Number predictions made by the machine learning models were processed and in some cases corrected before the final numbering output was given (Figure 2.4). Given the approach to predict all insertions as 'X', numbered sequence predictions including 'X's were replaced with the correct number label, based on nearby number predictions. If numbers were predicted with low confidence and therefore classified as 'U', these were also replaced with the required number label. This was done according to the following steps:

- Replacement of non-label(s) between two known labels: depending on the number difference and number of X/U amino acids between the two known labels, integer number labels or insertion labels are assigned. For example:
  14 - U - 16, U = 15
  14 - X - 15, X = 14A
  14 - X - U - 16 → 14 - 14A - 15 - 16.

- Within CDR loops: X/Us replaced with labels following the IMGT numbering pattern, which differs from outside of the CDR loops. For example:
  111 - X - 112, X = 112A.
  108 - U - 115, U = 114.
  Insertion residues within the CDRs assigned 'symmetrically' as per IMGT, compared to linearly at other sites, e.g:
  14 - X - X - 15 → 14 - 14A - 14B - 15
  111 - X - X - 112 → 111 - 111A - 112A - 112.

- Sequence start: If the start of a sequence was not able to be numbered by the model, the first predicted number is used and labels were stepped backwards. If the first predicted number is after the start of CDR1 (position 27), identification of the conserved cysteine residue at position 23 was used.

- Sequence end: if no numbers were predicted towards the end of the sequence (CDR3, FWR4), pattern recognition was included for commonly observed amino acid motifs. For example, 'TVSS' is very commonly found at positions 125-128 of heavy chain sequences. 'W..G' is also a commonly found motif, with tryptophan highly conserved at position 118.

If number predictions for a sequence contained 35 sequential positions that were not numbered (predicted as 'U'), the sequence was designated as un-numberable by the model.

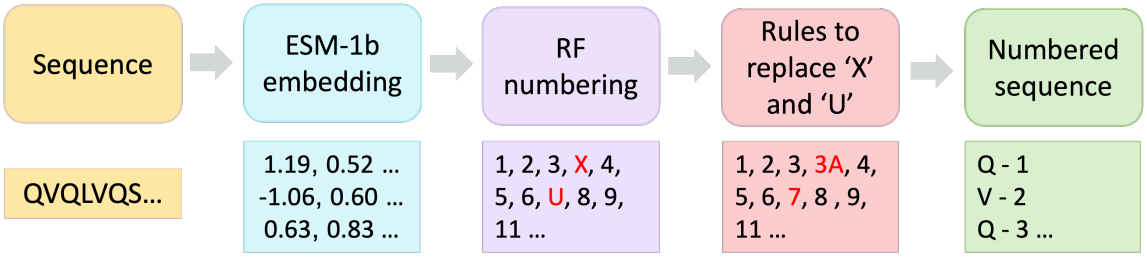


Figure 2.4: ANARCII pipeline for antibody numbering. An antibody amino acid sequence is fed through the ESM-1b transformer. The sequence embeddings are used by the random forest (RF) to predict numbering. The random forest output is processed to replace any insertion classifications (X) or uncertainties (U) to produce the final numbered sequence output.

2.2.2.8 Binary CDR/FWR classifier

In order to examine whether the ends of CDRs could be accurately predicted, to be used as anchor points and infer numbering, random forest models were trained to predict a binary outcome of whether a residue was within a CDR. The model was trained and tested on ESM-1b sequence representations of heavy chain sequences from OAS, and a binary yes/no of whether the residue was within a CDR (CDR1, CDR2 or CDR3). If a residue was not within a CDR, it was labelled as in the FWR. Best performance was achieved with a 200 tree, depth 100 model.

2.3 Results

2.3.1 Antibody number labels are highly diverse

When compiling training data from OAS, insertions were found at most non-CDR positions, making the multi-classification problem more complex than anticipated (additional residues within CDRs are always incorporated at the centre of the loops, rather than insertions at other positions within the CDRs). Frequencies of insertion number labels differed between the chain types, with the heavy chain data including longer CDRs and so more insertions within CDR1, CDR2 and CDR3. Considering all heavy chain sequences within OAS, without redundancy or length filters, over 1200 unique number labels were identified. This included sequences with very long CDRH3s; 143 sequences had CDRH3s $\geq$ 75 amino acids, including three sequences with a CDRH3 of length 83. We filtered our training dataset to remove sequences with CDRH3 length >18, resulting in 886 unique heavy chain number labels, 606 kappa number labels and 593 lambda number labels. As our models predict insertions as X, long CDRH3s are predicted as insertions (X) or uncertain (U) and numbering can be assigned based on nearby predicted labels (see Methods 2.2.2.7).

Excluding those in long CDRH3s, many other VH number labels were uncommon; 54 number labels appeared only once and 224 appeared fewer than 10 times in our OAS sequence datasets (Table 2.2). Rare labels which appeared only once included: 7B, 11G, 22D, 43J, 45K, 49Q, 70P, 73M, 74K, 79O, 84M, 89J, 91J, 93M, 98J, 99M, 119M, 120I, 121G, and 122E. Aside from these insertions, numbers at the sequence start were often infrequent, with residues numbered 1, 2 and 3 only found in around 30% of sequences used to our create training datasets (40% of sequences in OAS are missing the first 15 residues (Olsen et al. 2022c)). Some residues across CDRH1 and CDRH2 (positions 32, 33, 34, 60 and 61) were present in <19% of sequences, as some sequences have short CDRs and so do not have these assigned numbers, which are the centre of CDR loops and are the site of insertions/deletions. Position 73 is only present in 5% of our OAS VH sequences.

We originally considered training only two models: a heavy chain numbering model and a light chain model. However, the lambda and kappa light chains in our datasets contained differing number labels, with the frequency of some labels varying by over 100x between the two datasets (Table 2.7) and 43 number labels appearing in either only kappa or only lambda sequences. For instance, 44A, 73A, and 96A were

far more prevalent in lambda sequences than kappa, whereas 79A and 79B, 81C and 100A were far more common in kappa sequences. Some of these differences provide information about CDR lengths; for example, 33E was present in 12x more kappa sequences than lambda sequences indicating longer CDRL1 regions in kappa sequences, whereas 111E features in 12x more lambda sequences than kappa sequences, indicating longer CDRL3 regions. The kappa dataset included 606 unique number labels, 84 of which only appeared once, whilst the lambda dataset included 593 unique number labels with 95 only appearing once. Because of this, and the sequence differences between kappa and lambda chains, we trained separate numbering models for heavy, kappa and lambda sequences.

We initially trained random forest models to predict all the number labels observed in our training datasets. However, accurately and robustly predicting all potential number labels as 593, 606 or 886 separate classes (for lambda, kappa and heavy chain sequences, respectively) was challenging and misclassification was common between similar residues. For example, residues within the CDRH3 loop were commonly misclassified - likely due to sequence variation at these sites making model training more difficult than on the more conserved framework regions. We investigated whether using only 'anchor residues', a subset of positions consistently predicted correctly by the model, could be used instead of predicting all number labels. When using anchor residues, number labels between the anchors are annotated depending on the number of amino acids between the anchors. This could be achieved by using starts/ends of CDRs and other conserved residues such as Cys23, Trp41 and Cys104. Misclassifications between similar residues, for example within the CDRH3, could be avoided using this approach as these labels do not need to be assigned, i.e. good precision and recall is not required for every class. However, due to the potential presence of insertions at almost all positions throughout a sequence, it was not possible to identify enough consistently predicted anchor residues without losing accuracy of the location of insertion positions. For example, if there was an insertion present in the sequence at some point between anchors at the end of CDR2 (position 65) and Cys104, it would not be possible to precisely assign this insertion and the subsequent labels.

To reduce the complexity of classification, all insertion labels were assigned to one class, class 'X', to give only 129 possible classes (labels 1-128 plus 'X'). To further reduce the likelihood of misclassifications, a confidence threshold was set for the class prediction for each amino acid. If this threshold was not met by the model, the amino

Label	Vκ dataset fr.	Vλ dataset fr.	Difference between Vκ and Vλ (%)
16B	12	1	12.0
22B	16	1	16.0
33E	134	11	12.2
43C	14	221	15.8
43D	3	205	68.3
43E	1	184	184.0
44B	13	233	17.9
44C	8	148	18.5
44D	5	70	14.0
45C	25	264	10.6
45G	1	67	67.0
67D	1	14	14.0
68B	24	2	12.0
68D	14	1	14.0
70C	13	1	13.0
71A	16	172	10.8
71B	14	172	12.3
77A	319	10	31.9
77B	314	4	78.5
78G	24	2	12.0
78H	18	1	18.0
79H	40	2	20.0
80I	1	11	11.0
81G	550	43	12.8
81H	509	5	101.8
82I	1	46	46.0
86F	26	1	26.0
96A	49	898	18.3
96C	4	53	13.3
97A	18	436	24.2
97B	10	348	34.8
97C	6	285	47.5
97D	2	272	136.0
97E	2	74	37.0
97F	2	72	36.0
100A	746	40	18.7
111E	27	327	12.1

Table 2.7: Insertion number labels with the greatest differences in frequency (fr.) between kappa (Vκ) and lambda (Vλ) datasets. The thirty number labels appearing in both datasets with the biggest frequency differences are shown.

acid was designated class 'U' for uncertain.

2.3.2 Random forest models can accurately predict numbering for heavy, kappa and lambda chains

Random forests were trained on heavy, kappa, and lambda datasets. Sequence numbering accuracy was measured via the sequences being numbered identically to ANARCI (our ground truth). Accuracy in determining the ends of the CDRs, at positions 27, 38, 56, 65, 105 and 117, was also calculated to identify sequences which may have discrepancies with ANARCI in some of their numbering (for example, where an insertion occurs) but correctly identified the CDRs.

Two test sets were created for each of heavy, lambda and kappa sequences - one set from paired data and another from unpaired data. Performance was slightly better on the sequences in the paired test sets, at 99.72% for heavy chains, 99.60% for lambda chains and 99.86% for kappa chains. The number of sequences with accurate CDRs predicted was high, at 99.90% of heavy chains, 99.83% of lambda chains and 99.94% of kappa chains (Table 2.8). Sequences were classified as 'failed' if the model uncertainty in numbering was high (35 sequential residues numbered with below 60% confidence), if duplicate residues were predicted, if a cysteine could not be found at position 104, or the numbering was otherwise flagged as incorrect by our internal quality checks before the final numbered sequence output.

Model	Test set	Numbered same as ANARCI (%)	Numbered differently to ANARCI (%)	Failed (%)	Correct CDR ends (%)
VH	Paired	99.72	0.27	0.01	99.90
VH	Unpaired	97.87	1.04	1.09	98.32
Vλ	Paired	99.60	0.23	0.16	99.63
Vλ	Unpaired	99.23	0.47	0.26	99.38
Vκ	Paired	99.86	0.11	0.03	99.94
Vκ	Unpaired	98.11	1.55	0.34	99.19

Table 2.8: Overlap of ANARCII antibody numbering with ANARCI. Random forest numbering models consisted of 100 trees and were trained on 5000_1000 datasets. Test sets of 10000 randomly selected sequences were sourced from paired or unpaired OAS data.

Sequences with different numbering to ANARCI may be due to incorrect predictions by the random forest or incorrect assignment of uncertain residues - particularly if sequences are missing residues or contain insertions. Some sequences may in fact be incorrectly numbered by ANARCI, though it is often difficult to say what the true numbering should be. For example, a sequence may be numbered differently by ANARCI, AbRSA and AbNum (see Figure 2.5). In this example, the ANARCII random forest does not predict two of the number labels, for residues S and P, at a confidence threshold $\geq$0.6; these U's are replaced in the ANARCII processing to labels '8' and '9' due to stepping backwards from the assigned 10 label, to give the same numbering as AbNum but which differs from ANARCI and AbRSA.

The lower overlap of identical ANARCI and ANARCII numbering in unpaired datasets, compared to the paired datasets, is likely due to higher variety in sequence length and insertions found in the unpaired test sets. Many sequences in the unpaired heavy chain test set are missing part of FWR1, causing ANARCII difficulty in correctly and confidently assigning the start of the numbering and thus leading to misnumbering. In addition, a number of the sequences may have incorrect ANARCI numberings; for example, six sequences in the unpaired heavy test set were numbered by ANARCI as missing residue 55, but this is likely an error (Figure 2.6). The ANARCI numbering from OAS also included cases of misassigning positions 68A and 73. Numbering these sequences on the ANARCI webserver gives different numbering, which agrees with ANARCII and which we believe to be correct. This suggests differences in ANARCI versions, germlines and/or species restriction (e.g. specifying only using human HMMs) leads to altered numbering (Tobias Olsen, personal correspondence). In ANARCI, HMM models are built from available germline data, therefore sequence numbering may change as a result of updates to germline datasets. The true accuracy of ANARCII may therefore be higher than suggested based on the overlap with ANARCI-numbered OAS sequences.

For some sequences, particularly those with unusual CDR3 lengths, pattern recognition was required to number the end of the sequence (see Methods). It was surprising that sequences containing these common patterns were not able to be numbered confidently by ANARCII. This may suggest the ESM-1b embeddings are weighted by surrounding residues more than the actual residue identity so if nearby residues are unusual such as in shorter or longer CDRH3s, numbering is not confidently assigned even if a common amino acid pattern is present (such as TVSS). Similarly, conserved

#IMGT	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ANARCI	I	Q	M	T	Q	S	P	—	S	L	S	A	S	V
AbRSA	I	Q	M	T	Q	S	P	S	—	L	S	A	S	V
AbNum	I	Q	M	T	Q	—	S	P	S	L	S	A	S	V
ANARCII	I	Q	M	T	Q	—	S	P	S	L	S	A	S	V

Figure 2.5: Tools give differing IMGT numbering. At the beginning of this sequence, ANARCI does not assign position 9, AbRSA does not assign position 10 and AbNum does not assign position 7. Our model, ANARCII, predicts two residues as uncertain (U) before they are processed to be labelled 8 and 9.

.

# IMGT	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69
ANARCI (OAS)	G	—	S	G	H	H	T	—	—	G	S	T	Y	Y	N	P
ANARCI (webserver)	G	S	G	H	H	T	—	—	—	G	S	T	Y	Y	N	P
AbRSA	G	S	G	H	—	—	—	H	T	G	S	T	Y	Y	N	P
AbNum	G	S	G	H	H	T	—	—	—	G	S	T	Y	Y	N	P
ANARCII	G	S	G	H	H	T	—	—	—	G	S	T	Y	Y	N	P

Figure 2.6: Altered sequence numbering with different versions of ANARCI, depending on the germlines used to build HMMs and species restriction. The ANARCI numbered OAS sequence misses label 55 and shifts subsequently labelling whereas webserver ANARCI assigns position 55.

residues did not always have confident number predictions, for example, Cys104 was not confidently predicted in 0.28% of test VH sequences, despite being present.

The choice of confidence threshold was assessed, with 0.6 proving the best value. Reducing the confidence threshold reduced the number of U's predicted and requiring processing (which can introduce numbering errors) but introduced more misclassifications and so lowered accuracy; for example, a threshold of 0.4 reduced numbering accuracy of unpaired kappa sequences from 98.11% to 95.20%, with 3.51% incorrect sequences. Given that the models predict the numbering of each amino acid independently from the numbering of the other amino acids in the sequence, we tested whether addition of the sequence index of a residue as an extra feature would impact numbering, but this did not alter accuracy. We also assessed combining multiple smaller random forests trained on differing subsets of the training data, but this did not improve performance beyond a single larger random forest.

A challenge with the random forests was the size of the models, which correlated with the size of the training datasets and number of classes to predict. Even a relatively small model of 100 trees was 12GB when predicting every possible VH number label (886 classes, rather than classifying insertions as 'X') and trained on a 8000_1000 dataset.

We evaluated whether models trained only on the 200 most common positions could achieve similar or greater accuracy with less training data and therefore be potentially smaller models (Olsen et al. 2022a). We found these models performed slightly better than the previous models, with 99.71% of heavy, 99.65% of lambda and 99.89% of kappa chains numbered identically to ANARCI, and accurate CDR predictions of 99.81%, 99.74% and 99.98%, respectively (Table 2.9). Models were trained on a maximum frequency of 5000 ESM-1b embeddings per number label (as was done with the previous models for labels 1-128). Given our training sequences downloaded from OAS were filtered by CDRH3 length, we trained on data from 181 unique number labels for heavy chain sequences, 188 for lambda and 187 for kappa (as not all CDRH3 residues included in the 200 most common number labels were present in the training sequences). We maintained the approach of predicting insertion residues as an 'X' class; if we tried to predict all number labels, performance decreased whilst the models increased in size (VH model 8.5GB vs 5.1GB, with accuracy 99.5% on paired data compared to 99.71%).

Model	Test set	Numbered same as ANARCI (%)	Numbered differently to ANARCI (%)	Failed (%)	Correct CDR ends (%)
VH	Paired	99.71	0.19	0.10	99.81
VH	Unpaired	98.28	0.78	0.94	98.60
Vλ	Paired	99.65	0.18	0.17	99.74
Vλ	Unpaired	99.20	0.45	0.35	99.65
Vκ	Paired	99.89	0.09	0.02	99.98
Vκ	Unpaired	98.47	1.06	0.47	99.27

Table 2.9: ANARCII performance using random forest models (100 trees of depth 100) trained on only the most common number labels in OAS.

2.3.3 XGBoost performance is inferior to RF

XGBoost models (200 trees) were trained on 5000_1000 datasets of heavy, kappa and lambda datasets. These models were much smaller than the random forests, at <40MB compared to 6-12GB; however, the models misnumbered positions, even at a high prediction confidence threshold of 95%. This most commonly occurred for residues within CDRH3 but also for other positions throughout the antibody sequence. As a result, many sequences failed our internal quality checks, or were numbered differently to ANARCI. The models also took substantially longer to train (days vs <2 hours).

Model	Numbered same as ANARCI (%)	Numbered differently to ANARCI (%)	Failed (%)
Vλ	93.82	1.14	5.05
Vκ	78.90	4.61	16.53

Table 2.10: Performance of 200 tree XGBoost models on paired test sets of 10000 lambda and kappa chains, at a prediction confidence threshold of 0.95.

2.3.4 Binary CDR prediction

For some sequences, the random forest models struggled to accuracy and confidently predict the ends of CDRs, particularly CDR2. We evaluated whether a model trained only to predict the binary outcome of whether a residue was in the CDR or in the framework could be used to accurately identify the ends of CDRs. If so, the model could be useful to combine with the number label classifier model in cases where number label predictions were of low confidence. However, we found that the binary classifier was unable to predict the ends of CDRs any better than the standard random forest predicting number labels, with 92% accuracy at predicting VH CDR ends.

2.3.5 Nanobodies

We tested the performance of ANARCII on nanobody sequences downloaded from SAbDab (see Datasets - Nanobodies). 95.7% of sequences were numbered identically to ANARCI using the ANARCII VH model (Table 2.11). 17 sequences were assigned different numbering to ANARCI, and numbering failed on 5 sequences. Those with different numbering to ANARCI were largely due to insertions or missing residues put at different positions (generally due to ANARCII having low confidence and not

assigning a number label, and so it was assigned by rules based on the nearest predicted residue). Lowering the confidence threshold for number assignment reduced the overall number of sequences numbered correctly.

Numbered same as ANARCI (%)	Numbered differently to ANARCI (%)	Failed (%)	Correct CDR ends (%)
96.60	2.59	0.76	97.71%

Table 2.11: Overlap of ANARCII nanobody numbering with ANARCI. Random forest numbering models consisted of 100 trees and were trained on 5000_1000 datasets. Test sets of 10000 randomly selected sequences were sourced from paired or unpaired OAS data.

2.3.6 Distinguishing antibodies from non-antibodies

ANARCI is able to number TCR sequences in addition to antibodies, as it can use HMMs built from T cell germline sequences. ANARCII is not trained on any TCR data and is unable to number sequences, though can predict a few number labels due to similarities with antibody sequences. For example, Cys104 was assigned in 83% of TCRβ sequences using the VH ANARCII model. Other residues which were predicted included Gly98 (47% of sequences), Tyr102 (59%) and Trp41 (42%). Most number labels were not assigned in any sequences; at most 10 amino acids were numbered in a TCRβ sequence. Similarly, some residues of TCRα sequences could be numbered using light chain ANARCII models, with a few conserved labels (100, 102, 104, 121, 124) assigned to $\geq$ 80% of sequences by lambda or by kappa (6, 89, 102, 103, 121) ANARCII models.

We also assessed the specificity of ANARCII using non-immunoglobulin sequences; an antibody numbering tool should not return any numbering for such sequences. We tested ANARCII on 10 protein antigen sequences from SAbDab. One protein sequence (7b18) was 1288 amino acids, which exceeds the ESM-1b length limit for sequence embeddings. The remaining nine sequences were embedded by ESM-1b but ANARCII did not assign any number labels.

2.4 Discussion and future work

Existing antibody numbering methods rely on amino acid pattern recognition to assign number labels. These methods can fail or provide incorrect numbering on

unfamiliar sequences which differ from the model training data. In addition, methods which rely on germline sequences are influenced by germline database alteration. This can cause difficulties and inconsistencies in numbering when there are updates to the germlines, and if input sequences differ substantially from those within germline databases.

The aim of the work described in this chapter was to develop a novel antibody numbering method. ANARCI is the current leading antibody numbering tool and is widely used as a webserver, command line tool and python module. ANARCI relies on downloading numbered germline sequences; the new approach described here does not require updating with data from external sources.

We explored several machine learning classifier methods and found a random forest produced the best results for antibody numbering. Whilst a relatively simple model, when trained using transformer-based sequence representations, the model has many features to use for classification. Framing antibody numbering as a multiclassification problem was made more difficult by the variety of different number labels which are now found in antibody sequences. As described in the results, we initially found over 1200 unique number labels in VH sequences in OAS, which we filtered by sequence frequency and CDRH3 length to 886 VH, 606 Vκ and 593 Vλ number labels. Given this large number of labels, we simplified the classification approach by assigning all insertion labels to an insertion class 'X', and subsequently processing the output predictions to replace Xs with genuine number labels. With this approach we were able to achieve agreement with the earlier version of ANARCI on our paired test datasets of over 99.6% for each chain type, and prediction of CDR ends at over 99.8%. We tested another simplification of our random forest model to train on only the most commonly observed number labels, as identified by Olsen et al. 2022a, and improved our ANARCI agreement slightly to over 99.65% whilst reducing model sizes.

Sequences which were classified as incorrectly labelled compared to ANARCI generally had uncertainty in the numbering within CDR2, likely due to variation in CDR2 and FWR3 lengths making classification more challenging. In addition, when ANARCI and ANARCII disagree on numbering it is unclear what the true numbering labels should be, particularly if insertions are present. In these cases, other numbering tools AbRSA and AbNum were often unable to number the sequence. The ANARCI/ANARCII overlap of CDR ends prediction is higher than that of the full

sequence, as incorrectly labelled sequences may be due to placing an insertion label at a different place to ANARCI, without affecting the determination of the CDR regions. We believe correct identification of the start and ends of the CDR regions is the most important accuracy measure for an antibody numbering tool, as this is the main purpose of antibody numbering for most researchers and would have the largest impact if incorrectly assigned. ANARCII can also number nanobody sequences but is unable to number TCRs or other protein sequences.

A limitation to our approach is that our models were trained using ANARCI numbering predictions as the ground truth. We know that ANARCI makes prediction errors, so a small proportion of the training data may be incorrect. However, no standardised numbered antibody sequence datasets exist, and ANARCI is the current leading method for antibody numbering. We hoped that by using sequence representations, our model would learn underlying relationships between amino acids within an antibody sequence that would result in more reliable numbering than purely sequence pattern recognition.

We were surprised that some conserved residues were not consistently predicted by ANARCII. For example, Cys23 and N-terminal TVSS motifs were not always labelled. This tended to occur if an antibody sequence had unusual lengths - for example, if the sequence was missing part of FWR1, had very short or long CDR3 loops, and/or had unusual amino acids neighbouring the conserved residues. Numbering across the end of CDR2 and FWR3 (IMGT residues 60 - 90) tended to be the most challenging to predict, for all chain types. These regions differed the most in length, with some sequences missing residues 81-82 and some sequences with insertions. This suggests the ESM-1b embeddings are impacted by these length differences. Training a binary random forest classifier to predict whether residues were located in the CDR or framework, to establish the end of CDR2, did not perform better than the number label random forest on the difficult sequences. Changing the training datasets to include a higher frequency of these challenging residues also did not improve overall performance.

Potential disadvantages of our approach are the requirement for pytorch and the ESMB-1b transformer, which would need to be installed by a user. The numbering model can currently only apply IMGT numbering, unlike ANARCI which can also

number with Chothia, Kabat, Martin and Aho schemes. Work to improve the antibody numbering model could involve further optimisation of the uncertainty thresholds and the training data beyond the existing efforts. Selection of an appropriate uncertainty threshold (implemented through the number of decision trees in the random forest which agree on the class prediction) was not clear cut, with overlap between probabilities given to correct and incorrect number labels. Whilst assigning a higher confidence threshold results in fewer prediction misclassifications, it also results in a greater number of residues needing to be labelled when processing the random forest output. This can lead to errors if there are many sequential uncertain residues, particularly if there are insertion residues or gaps, or if the uncertainty is around the starts and ends of CDR regions.

Our training data was selected from OAS, which is a rich source of curated antibody sequence data. We chose our training data by identifying sequences containing insertion residues, but other approaches could be taken. For example, we did not select for sequence diversity, which could be done by V or J gene assignment or amino acid sequence identity, or for species diversity.

The main limiting factor to the development of the random forest models was the size of the trained models. This impacted the amount of training data that could be used to produce the models, and the number of trees within the model. Training a model on a greater amount of data may result in better performance, though the volume of data we used is more than some existing antibody numbering tools such as AbRSA. If larger models are produced, they may be of more limited use to users to download and use.

CHAPTER 3

Computational Methods for Hit Identification and Expansion in Existing Antibody Datasets

This chapter includes evaluation and adaption of the method:

Wong, W.K., Robinson, S.A., Bujotzek, A., Georges, G., Lewis, A.P., Shi, J., Snowden, J., Taddese, B. and Deane, C.M. (2021). Ab-Ligity: identifying sequence-dissimilar antibodies that bind to the same epitope. MAbs (Vol. 13, No. 1, p. 1873478).

3.1 Introduction

Development of therapeutic antibodies is a complex and resource-intensive process. Challenges in the field include the prediction of the likelihood of an antibody paratope binding to a given antigen epitope, and the identification of similar antibodies that might bind to the same epitope. Robust and accurate methods to achieve this could reduce the number of therapeutic antibody candidates requiring expression and laboratory-based screening, by prioritising certain candidates with predicted binding. This could therefore optimise the speed and cost of the therapeutic antibody development pipeline.

During the hit identification stage of therapeutic antibody development, antibodies are typically expressed and screened for binding to a specific target antigen (see Chapter 1, Section 1.1.5.1). Antibodies which bind to the target may already be

known but be patented or unsuitable for therapeutic development, or may not be known. In hit expansion, which occurs after potential hits have been identified, iterations on the candidates are generated or identified from datasets. For example, antibodies with sequence differences to adjust their predicted binding or developability properties. These expanded hits will then be experimentally evaluated for their binding and functionality. In both hit identification and hit expansion, accurate prediction of antibodies most likely to bind to the target antigen would be hugely beneficial. In addition, identification of dissimilar antibodies able to bind to the same target epitope can diversify the candidates taken forward in development, in case of a particular sequence or structural motif leading to unfavourable properties such as aggregation, immunogenicity, or differing functionality.

Current high throughput computational methods in these areas often use antibody sequence information only; however, making paratope-epitope predictions is challenging due to the lack of understanding of the three-dimensional shape of the antibody binding site which underpins its potential binding capabilities. Further, existing sequence-only methods such as clonotyping typically only use part of the antibody sequence information: CDRH3 sequence, V and J gene assignment (Galson et al. 2015). The development of methods able to use information from the entire antibody sequence might better inform predictions of paratope-epitope interactions.

Paratyping and Ab-Ligity are tools which can identify sequence-diverse antibodies which bind to the same epitope, using antibody sequence information (Richardson et al. 2021; Wong et al. 2021). Paratyping is a sequence-based method for clustering similar paratopes (see Methods 3.2.2.2). Paratyping has comparable precision, accuracy and speed to clonotyping at identifying sequences that bind to the same epitope, achieving 84% precision and 76% recall for prediction of pertussis toxoid-binding, where clonotyping achieved 83% precision and 79% recall (Richardson et al. 2021). Paratyping and clonotyping identified non-overlapping sets of antibodies in this case study, with paratyping identifying more diverse antibody sequences.

Ab-Ligity is a method for comparing paratope or epitope similarity using solved structures or modelled antibodies (Wong et al. 2021). Predicted paratopes are represented by hash tables incorporating information on amino acid physicochemical groups and inter-residue distances, using modelled antibody structures (see Methods

3.2.2.3). When tested on a set of antibodies with diverse CDRH3 sequences, Ab-Ligity method achieved precision and recall values of 95% and 69% at identifying similar antibody paratopes, which would not be possible with clonotyping due to low CDRH3 similarity (Wong et al. 2021). As with paratyping, Ab-Ligity can be run on a single chain sequence of an antibody, or on both heavy and light chain sequences. Paratyping and Ab-Ligity allow antibodies with different lineages (with dissimilar CDRH3 sequences, V and J genes) to be identified as similar and therefore as likely binding to the same epitope - a link which is not possible to make with clonotyping. These methods could therefore could be used in hit expansion, in collaboration with clonotyping, to identify likely-binding antibodies within an antibody dataset or repertoire. I contributed to the development of Ab-Ligity by assessing the impact of adding a CDRH3 length constraint to the similarity assessment and assisted with the Ab-Ligity manuscript preparation.

Combining both sequence and structural information presents a promising approach to improve our understanding and accurate prediction of antibody binding. This may also lead to tools for *de novo* design of antibody sequences with known target binding affinity, and with appropriate physicochemical properties. In this chapter, we describe our work to develop a method for accurate and rapid identification of antibodies which bind to the same epitope, and our work towards epitope-to-paratope prediction. We began by comparing and applying existing tools for the prediction, comparison and grouping of similar antibody binding sites. We then developed these tools to create hit identification and expansion pipelines for the prediction of antibody-antigen binding given a target epitope structure, with and without prior knowledge of a binding antibody. Such pipelines would be a highly useful tool in therapeutic antibody development by identifying antibodies likely to bind to the target epitope, which could used as a starting point for therapeutic antibody design. We aimed to leverage existing knowledge of paratope-epitope pairs (using SAbDab) to identify antibodies with paratopes likely to bind to a target epitope, based on epitope similarity. We then used these antibodies to search in other datasets and create a shortlist of potentially binding antibodies. We evaluated use of antibody model libraries (AMLs) for efficient searching of large antibody sequence libraries based on shared predicted structures, and created an adapted version of Ab-Ligity to make comparisons using only predicted structural information.

3.2 Materials and methods

3.2.1 Data

3.2.1.1 SAbDab: antibody-antigen complexes

Structures of 2258 antibody-antigen co-crystal complexes involving a protein antigen were obtained from SAbDab (Dunbar et al. 2014) on 4th February 2020. A non-redundant set of 924 complexes was obtained by retaining only one copy of antibodies with 100% sequence identity, as used in Wong et al. 2021. Homology-based models were built from the antibody sequences using ABodyBuilder (Leem et al. 2016), which was prevented from using sequence-identical templates to build the models (see Chapter 1, Section 1.3.2.1). We used the SAbDab structures as a database of known epitope-paratope pairs.

3.2.1.2 Lysozyme-binding antibodies

Structures of 13 lysozyme-binding antibodies were extracted from the SAbDab dataset and corresponding models were built using ABodyBuilder. These antibodies are known to bind to the same epitope on hen-egg lysozyme (Figure 3.1); therefore, they can be a useful test set when attempting to extract similar binders from a larger dataset. These antibodies are murine and share high sequence similarity; across the 13 antibodies there are 5 unique VH sequences and 9 unique VL sequences, with only 2 unique CDRH3 and 6 unique CDRL3 sequences (Table 3.1).

PDB Code	CDRH3	CDRL3
1a2y	ARERDYRLDY	QHFWSTPRT
1bvk (BA chains)	ARERDYRLDY	QHFWSTPRT
1bvk (ED chains)	ARERDYRLDY	QHFWSTPRT
1fdl	ARERDYRLDY	QHFWSTPRT
1g7h	ARERDYRLDY	QHFASTPRT
1g7i	ARERDYRLDY	QHFFSTPRT
1g7j	ARERDYRLDY	QHFHSTPRT
1g7l	ARERDYRLDY	QHFSSTPRT
1g7m	ARERDYRLDY	QHFVSTPRT
1kip	ARERDYRLDY	QHFWSTPRT
1kiq	ARERDFRLDY	QHFWSTPRT
1kir	ARERDYRLDY	QHFWSTPRT
1vfb	ARERDYRLDY	QHFWSTPRT

Table 3.1: Hen-egg lysozyme-binding antibodies from SAbDab.



Figure 3.1: Hen-egg lysozyme (pink) with aligned structures of 13 antibodies binding to the same epitope (see Table 3.1 for PDB codes and CDRH3 and CDRL3 sequences).

3.2.1.3 SAbDab case studies

We selected antibodies from SAbDab with a variety of CDRH3 lengths and antigens to be used as test cases for our pipelines (Table 3.2). The antibodies were chosen based on predicted similarity with other antibodies in the Public Baseline AML.

Name	Description	CDRH3 Length
4i18	Mouse Fab against human prolactin receptor	18
5eu7	Human Fab against HIV-integrase	23
5u68	Human Ab against HIV/RSV	17
6apd	Infant human Ab against RSV F	15

Table 3.2: Antibodies from SAbDab used as test cases to compare pipeline methods.

3.2.1.4 Public baseline AML

Shared antibody structure model libraries (AML) were designed and generated by Matthew Raybould. The AML used at the beginning of this work was the first version of the Public Baseline AML (PB-AML1) (Raybould et al. 2021b), consisting of 23,700 antibody structures predicted to be shared by ten unrelated participants (Gidoni et al. 2019).

PB-AML1 was designed by predicting the shared structural similarity between individuals using the repertoire structural profiling algorithm (Figure 3.2). The structural similarity between two human antibody repertoires (Gidoni et al. 2019) was identified by using the modelled structures from one repertoire as cluster centres and clustering structures from the second repertoire if the RMSD to the cluster centre <1.5 Å. Centres of clusters containing predicted structures from both the first and second repertoires were then compared to another repertoire, and so on for a total of ten repertoires. The antibody models for these structural clusters were built by ABodyBuilder (Leem et al. 2016) (with FREAD template libraries dated February 2019). Each structural cluster corresponds to numerous antibody sequences predicted to adopt the structure, and sequences from participant 'S64' were assigned to represent each of the 23,700 structural clusters (Raybould et al. 2021b). PB-AML1 corresponds to 12,057,613 antibody sequences, with clusters ranging in size from 13 to 66,610 (average 509). The procedure for AML creation is described in detail in Raybould et al. 2021b. In the analysis presented here, all heavy chain sequences within a PB-AML1 structural family were given the same designated light chain sequence of the structural representation.

A revised AML, Public Baseline AML v2 (PB-AML2), was generated by Matthew Raybould in December 2020. Updated FREAD template libraries (dated November 2020) were used to build the ABodyBuilder models. Additionally, antibodies were clustered at 95% sequence identity and 90% interface identity (previously 90% and 85% respectively). This resulted in the new AML containing more structural clusters (37,533 vs 23,700) corresponding to more sequences (18,512,721 VH/VL pairs). Structural clusters contained an average of 493 sequences, though again with very large variation (range of 5 - 87830 sequences). Sequences from participant 'S64' were assigned as the cluster representative sequences, as with PB-AML1. The models from PB-AML2, are available for download from http://opig.stats.ox.ac.uk/resources.

3.2.2 Methods

3.2.2.1 Clonotyping

Clonotyping is a method for clustering of antibody sequences based on clonotype. Clonotyping was performed using Change-O (Gupta et al. 2015), with clonotype definition of identical V and J gene, and ≥90% CDR3 sequence identity (see Chapter 1, Section 1.5.4.1).

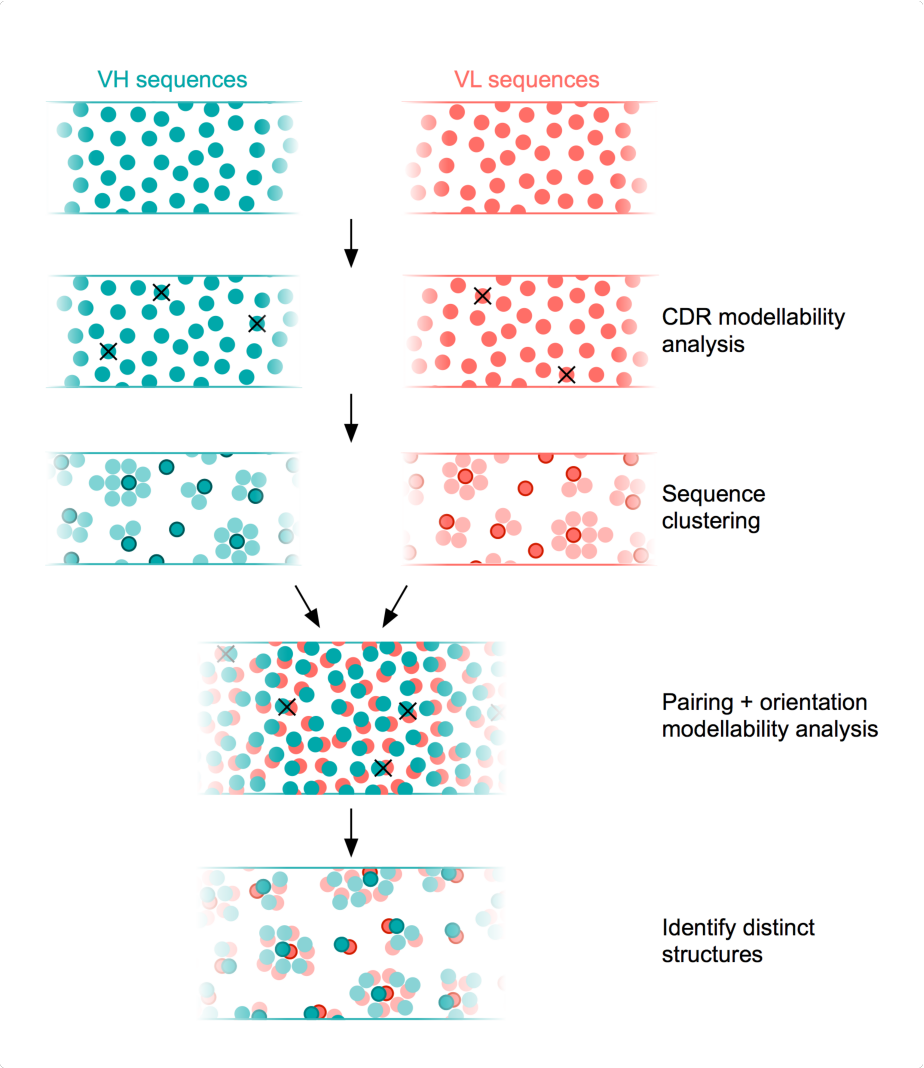


Figure 3.2: Repertoire structural profiling algorithm. Sequences from a BCR repertoire are assessed for their modellability, and sequences are clustered by sequence identity using CD-HIT (Fu et al. 2012) (unmodellable sequences are removed, shown by crosses). VH and VL chains are paired and their orientation predicted (unmodellable orientations are removed, shown by crosses). VH-VL pairings with identical combinations of CDR lengths are clustered based on their predicted structures to identify shared 'distinct structures'. Figure from Supplementary Materials, S4 (Raybould et al. 2021b).

3.2.2.2 Paratyping

Paratyping is a method for clustering paratope-similar antibody sequences with same length CDRs (Richardson et al. 2021). At the time of this research, Parapred was the best performing paratope predictor method; paratope residues were predicted from the antibody sequence using Parapred (Liberis et al. 2018). Parapred is a machine

learning algorithm which predicts the probability of each amino acid residue in the CDR regions, plus two adjacent residues, participating in binding to an antigen (see Chapter 1, Section 1.5.1). Residues were defined as being in the paratope if the Parapred probability was ≥ 0.67, as described in Richardson et al. 2021. Antibodies were clustered by the paratyping method if their paratopes shared $\geq 75\%$ identity.

3.2.2.3 Ab-Ligity

Ab-Ligity is a method for comparison of antibody epitopes or paratopes based on pharmacophoric and structural profiles (Figure 3.3) - see Chapter 1, Section 3.2.2.3) for full method description. Epitope and paratope residues are identified through proximity to the antibody/antigen in the case of co-crystal antibody-antigen structures, or paratope residues can be predicted using Parapred if using models of antibodies. All possible combinations of triplets of residues are evaluated based on distance between residues and physicochemical type. Each triplet is described by a hexamer hashcode, and frequency tables of the hashcodes can be assessed for similarity using a Tversky index.

3.2.2.4 Sequence-agnostic Ab-Ligity

A sequence-agnostic version of Ab-Ligity was developed to compare antibody paratopes based only on structure, ignoring the physicochemical properties of amino acid residues within the paratope. The paratope comparison is conducted identically to Ab-Ligity, with distances between paratope residues calculated and binned, and all possible triangles of residues represented by a hexamer hashcode. However, when paratope residues are tokenised to their relevant amino acid physicochemical groups, for example hydrophobic, all amino acids are assigned to the same group. This results in antibody binding sites being represented by hash tables of the distances between amino acids in the paratope, with no information on differing physicochemical groups (Figure 3.3). This method was developed for use with AMLs where antibody structures are shared by multiple antibody sequences.

3.2.2.5 Definition of similar antibodies

To assess method performance at identifying antibodies which bind to the same epitope, we defined pairs of antibodies as 'truly' similar if their epitope crystal structures scored ≥ 0.1 similarity by Ab-Ligity (using the antibody-antigen co-crystal structures from SAbDab; similarity threshold as used in Wong et al. 2021).

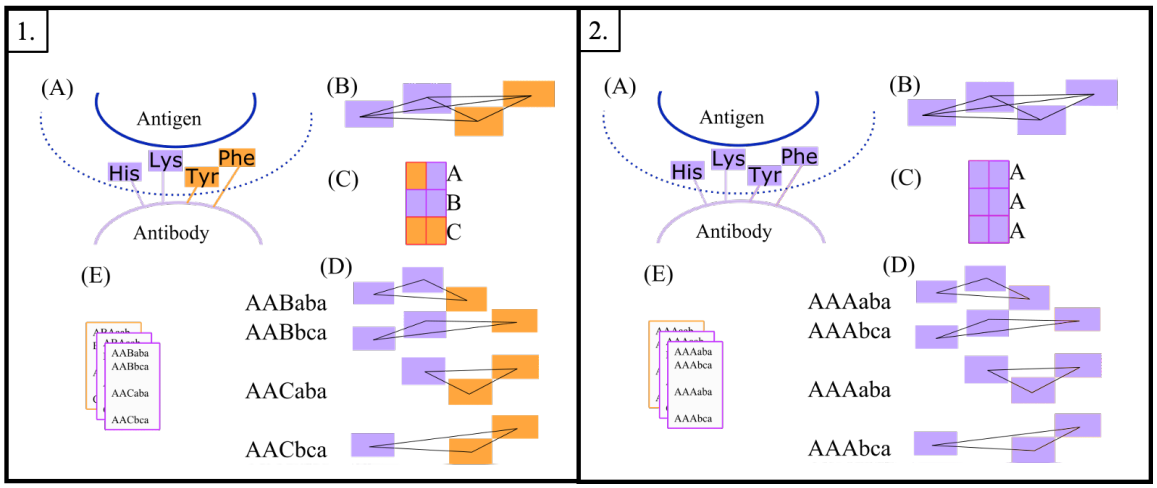


Figure 3.3: (1.) Ab-Ligity method. A) Paratope residues are tokenised based on physicochemical properties (histidine and lysine are basic, shown in purple, and phenylalanine and tyrosine are aromatic, shown in orange). (B) Distances between all triangles of paratope residues are hashed into 1.0 Å bins. (C) Pairs of residues within triplets are given a hashcode based on their physicochemical group. (D) For each triplet of residues, physicochemical hashcodes are combined with the distance hashcodes to generate a six-character hashcode. (E) A paratope is represented by all of the hashes, stored in a frequency table (2.) Sequence-agnostic Ab-Ligity is an adapted version of Ab-Ligity which ignores physicochemical information by assigning all paratope amino acids to the same physicochemical group (shown in purple). Hash tables of hexamers are created in the same way as in Ab-Ligity (B-E) but only the distances between the paratope triplets of paratope residues are variable. Figure adapted from Wong et al. 2021; (1.) is also included in Chapter 1.

3.2.2.6 Comparison of Ab-Ligity and paratyping

When comparing the performance of paratyping and Ab-Ligity at scoring the similarity of a pair of antibodies, if the antibody epitope similarity was ≥ 0.1 and their paratope similarity was above the method's paratope similarity threshold (0.75 for paratyping, 0.1 for Ab-Ligity), they were classed as a true positive (TP) pairing. If the paratope similarity calculated by paratyping or Ab-Ligity was lower than the threshold, they were classed as a false negative (FN) pairing. For pairs of antibodies where their epitope similarity was < 0.1, if their paratopes were predicted to be similar, they were declared a false positive (FP). If their paratopes were not predicted to be similar, they were a true negative (TN). CDR length matching on the heavy and light chains was used for both paratyping and Ab-Ligity, as it is required for the paratyping method (pairs of antibodies must have same length CDRH1, CDRH2, CDRH3, CDRL1, CDRL2 and CDRL3 for comparisons to be conducted).

Using these designations, precision and recall values were calculated using the formulae:

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN}$$

3.2.2.7 Identification of sequence-dissimilar binders given known binding antibody sequence

We created a pipeline to identify sequence-dissimilar antibodies likely to bind to the same target antigen, given the knowledge of a binding antibody sequence to the target. Using the known binding antibody sequence as an input, we look for other antibodies which potentially bind to the same epitope, using paratope comparisons. Using Ab-Ligity, this pipeline includes building structural models of the known binding antibody and of the antibody dataset to be searched for similar antibodies (Figure 3.4). Paratyping does not require generation of antibody models and instead can be used to conduct comparisons directly on the antibody sequence data (Figure 3.5). The output of the pipeline is a list of antibodies predicted to bind to the same epitope as the input antibody sequence.

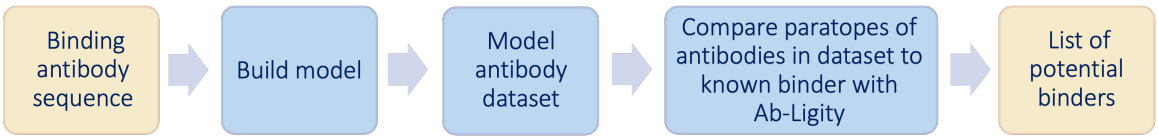


Figure 3.4: Ab-Ligity pipeline for identifying sequence-dissimilar binders from an antibody dataset, given a known binding antibody sequence. Ab-Ligity is used to compare the modelled paratope of a known binding antibody to antibodies in a dataset, to identify antibodies likely to bind to the same epitope.

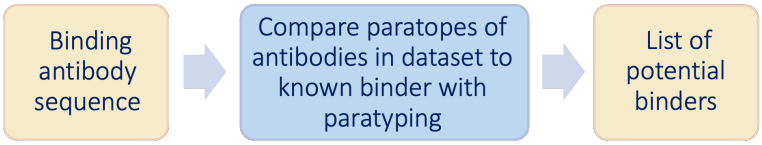


Figure 3.5: Paratyping pipeline for identifying sequence-dissimilar binders from an antibody dataset, given a known binding antibody sequence. Paratyping is used to compare the paratope of a known binding antibody to antibodies in a dataset, to identify antibodies likely to bind to the same epitope as the known binding antibody.

3.2.2.8 Identification of binders without a known binding antibody sequence

We created a pipeline to identify antibodies predicted to bind to a target crystal epitope without knowledge of a binding antibody sequence. The target epitope is compared to epitopes in SAbDab using Ab-Ligity. Epitopes similar to the target epitope are collated, and their corresponding binding antibodies (from the co-crystal structures) are identified. These antibodies are modelled using ABodyBuilder (Leem et al. 2016) and used to identify antibodies with similar paratopes in the AML, using Ab-Ligity calculations of paratope similarity (Figure 3.6). Modelled structures of the SAbDab antibodies are used to compare to the modelled structures in the AML, rather than using the available crystal structures, to avoid biases from comparing crystal- and model- derived paratopes. For proof of concept of our pipeline, we used the set of 13 known lysozyme-binding antibodies. We moved antibodies from this set from the SAbDab database into the AML, to evaluate whether they could be successfully identified.

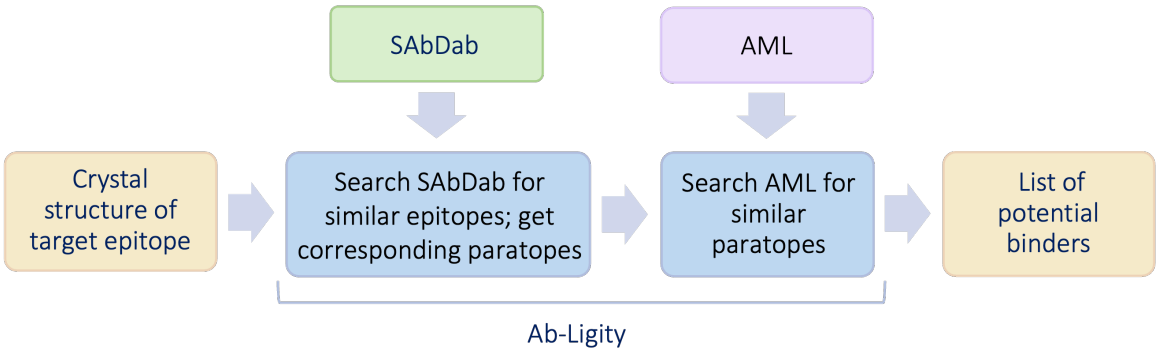


Figure 3.6: Proof of concept pipeline using a known set of binders. Using a crystal structure of a target epitope, SAbDab was searched for similar epitopes using Ab-Ligity. The corresponding paratopes for identified similar epitopes were used to identify similar paratopes in the AML, also using Ab-Ligity. Known binders were moved from SAbDab to the AML to determine whether the the full known set of binders could be identified from the AML with fewer similar epitopes in SAbDab.

Given that the single sequence assigned to each of the modelled structures in the AML does not necessarily reflect all of the sequences within that structural cluster predicted to adopt the structure, but is used by Ab-Ligity to calculate paratope similarity, we adapted the pipeline so that antibodies identified in SAbDab are used to find similar paratope structures in the AML using sequence-agnostic Ab-Ligity (Figure 3.7). AML antibody structures identified as structurally similar to the SAbDab

antibody paratope are shortlisted, and the sequences which map to these structures are identified. Ab-Ligity is used to calculate paratope similarity between these antibodies and those identified in SAbDab, to assess physicochemical paratope similarity in addition to the geometric paratope similarity. Antibodies which are similar to the SAbDab paratopes are identified as antibodies likely to bind to the target epitope.

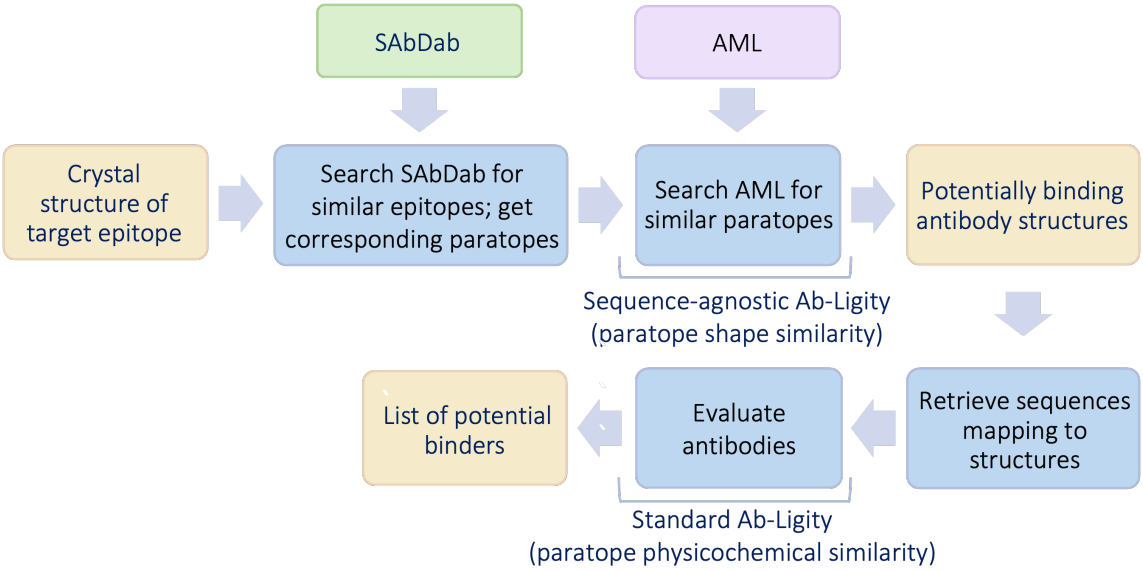


Figure 3.7: Pipeline for identifying potentially binding antibodies to a target epitope, without knowledge of binding antibodies. Using a crystal structure of a target epitope, SAbDab is searched for similar epitopes using standard Ab-Ligity or sequence-agnostic Ab-Ligity, and corresponding antibodies identified. These antibodies are used to identify paratopes with similar structures in the AML. Sequences mapping to the identified AML structures are retrieved and compared to the SAbDab paratopes using Ab-Ligity.

We explored other methods for evaluating the sequences retrieved from the AML structures, aside from Ab-Ligity, including clonotyping, paratyping and position-only paratyping.

3.2.2.9 Construction and scoring of position specific scoring matrices

As an alternative approach to using sequence-agnostic Ab-Ligity to search the AML for similar structures, we evaluated the use of different representations of the structural AML families for comparison with the SAbDab antibodies.

Position specific scoring matrices (PSSMs) were built as representations of antibodies, with one PSSM to represent each structural family in the AML. Target antibody PSSMs could then be assessed for similarity to AML family PSSMs. PSSMs were initially built using antibody CDR regions. The method in Wong et al. 2019a Supplementary Materials was used to generate and score PSSMs from the antibody sequences associated with each structure in the AML, based on the probability of encountering each amino acid at each position:

$$M_{k_j} = log_2(P_{k_j}/b_k)$$

where $M_{k,j}$ is the element score and $P_{k,j}$ is the probability of observing the amino acid k at the ANARCI-numbered position j in the sequence. Initially, $P_{k,j}$ was calculated through the proportional frequency of each amino acid at each position in the CDR regions of sequences in an AML family. This was possible without any sequence padding as all members of an AML family possess the same CDRH1, CDRH2 and CDRH3 lengths. b_k is the background probability of observing amino acid k at the position (1/20 for all amino acids). All elements with a value of zero, due to a particular amino acid not being observed at that position, were assigned a value of 0.001 to prevent computational errors.

The PSSM score for a target antibody, t_a, to an AML family is the sum of the element scores across all the positions:

$$t_a = \sum_{j=J0}^{J} M_{k,j}$$

If the score is greater than a selected threshold, the AML family was described as a potential hit to the target antibody.

PSSMs were also built using antibody clonotype and paratype (see below). All analysis was initially conducted using heavy chain data; PSSMs were also built using the light chain using the same method.

Clonotype PSSM Clonotype PSSMs were constructed using the CDRH3 proportional frequency matrices created as described above, plus the V and J gene assignments for each sequence (assigned using ANARCI (Dunbar and Deane 2016)), which were one hot encoded (see Figure 3.8).

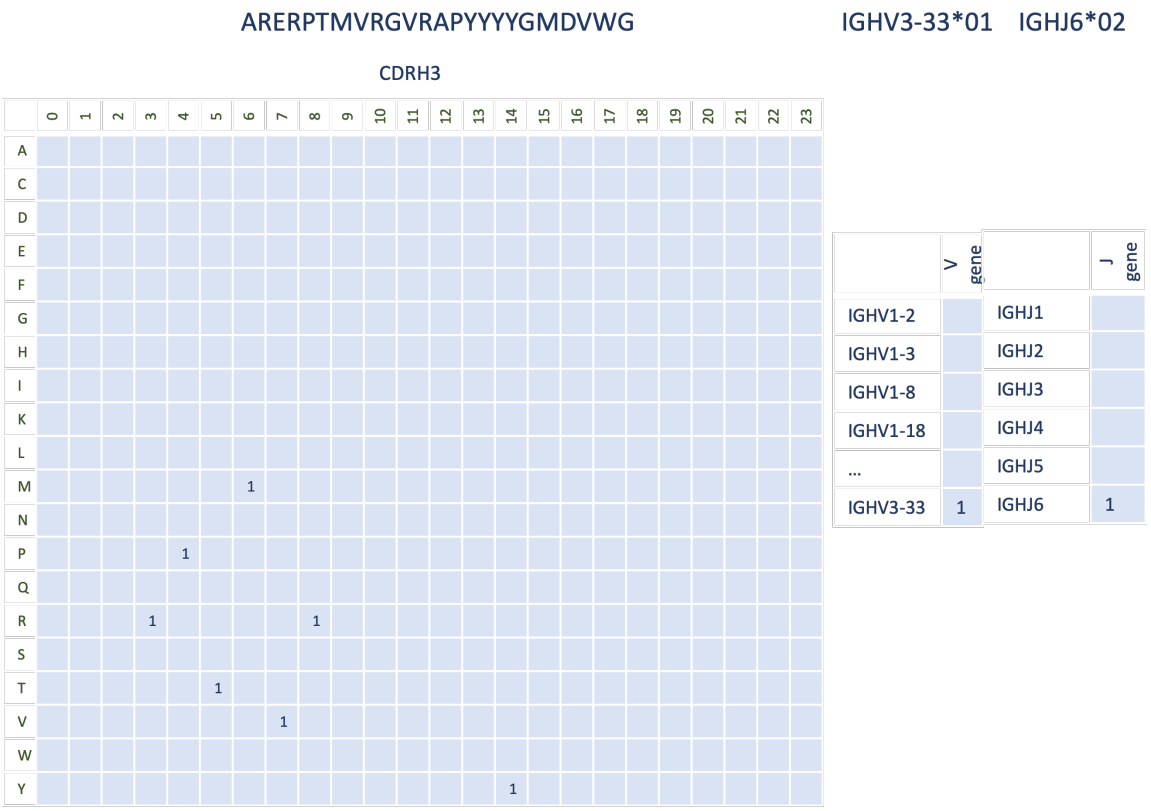


Figure 3.8: Example of frequency matrices for PSSM construction, using the clonotype of a heavy chain sequence from AML family H8641+L36964. Amino acids in the CDRH3 are one-hot encoded, as are V and J genes.

Paratype PSSM Paratype PSSMs were constructed using predicted paratope residues for each antibody sequence, identified using Parapred (Liberis et al. 2018) with a threshold of 0.67. Separate PSSMs were generated for each of CDRH1, CDRH2 and CDRH3, with only predicted paratope residues included rather than all amino acids in the CDRs (Figure 3.9).

3.2.2.10 Incorporation of BLOSUM matrix to PSSM

To account for similarities between non-identical amino acids, the BLOSUM matrix was included in PSSM scoring (Henikoff and Henikoff 1996). When comparing a target antibody PSSM to an AML family PSSM, at each position the BLOSUM score associated with substitution of the target amino acid was multiplied by the observed proportional frequency of amino acids ($P_{k,j}$).

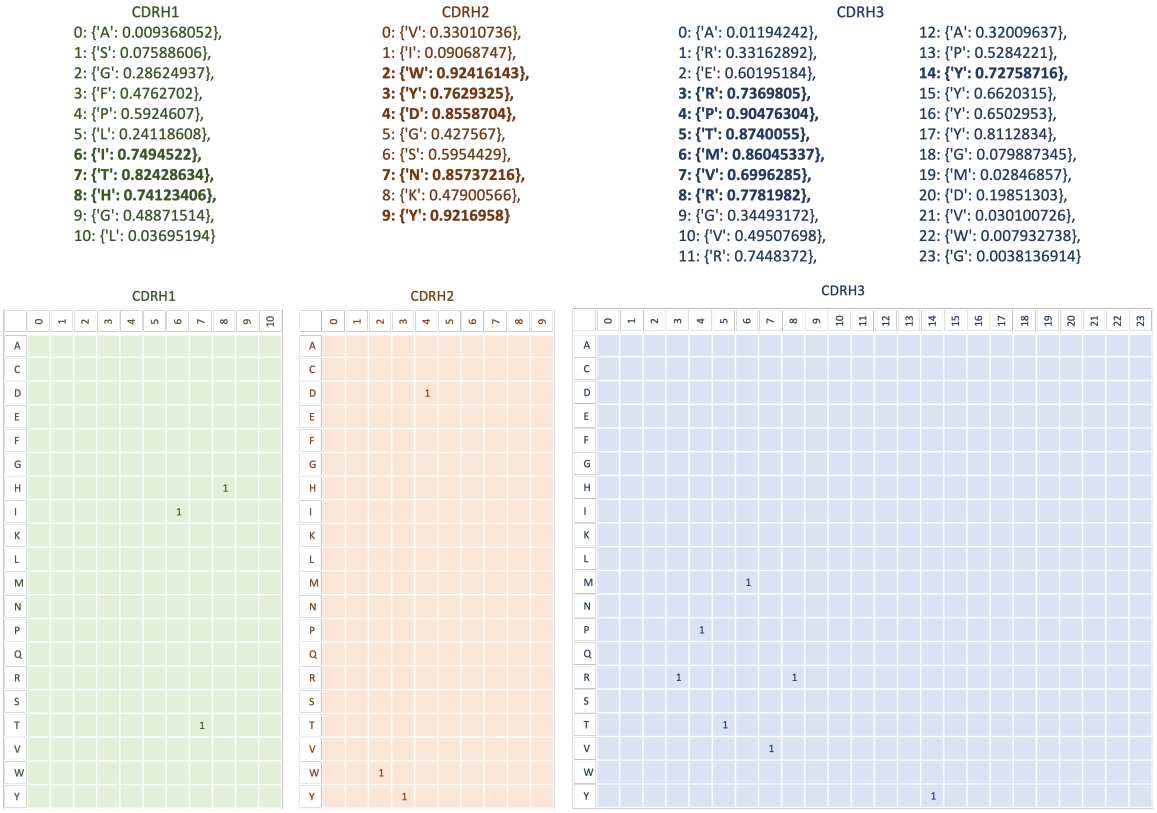


Figure 3.9: Example of frequency matrices for PSSM construction, using the paratype of a heavy chain sequence from AML family H8641+L36964. Paratope residues from the CDRs are predicted using Parapred and counted for each position; CDR residues for cluster sequence 57 are shown, with the corresponding Parapred scores. Residues above the threshold (in bold) are included in the predicted paratope.

3.3 Results

3.3.1 Comparison of Ab-Ligity and paratyping

Ab-Ligity and paratyping performance was compared on solved crystal structures and modelled structures from the non-redundant SAbDab dataset of 924 antibodies. This non-redundant set was used in order to compare the methods on a dataset less likely to be biased or skewed by over-representation of certain antibodies in SAbDab. Pairs of antibodies were predicted as similar if they achieved a paratope similarity score of ≥ 0.75 by paratyping or ≥ 0.1 by Ab-Ligity. Antibodies were classed as truly similar if their epitopes had Ab-Ligity similarity scores of ≥ 0.1 (see Methods 3.2.2.5). We were most concerned with the performance on modelled structures due to the lack of currently available crystal data and the length of time it takes to generate new crystal structures. Particularly, we were interested in the methods' performance on

the sequence diverse subset of structures, where the CDRH3 sequence identity of between pairs of antibodies is < 0.8, as these are hits that cannot be identified by existing clonotyping methods.

3.3.1.1 Performance on antibody crystal structures

Ab-Ligity and paratyping both performed well on the crystal structures of the antibody-antigen complexes, with precision of 0.98 and recall values of 0.96 and 0.91 respectively (Table 3.3). On pairs of sequence diverse antibodies (where CDRH3 identity < 0.8), performance of both methods decreased due to the higher proportion of false negatives (FN), meaning both methods failed to correctly identify some similar antibody pairs. This was not unexpected as both methods incorporate sequence information and if sequences are diverse they may be too different to be recognised as similar. This was particularly seen in paratyping which obtained a higher number of FN than Ab-Ligity, as the former method is entirely reliant on sequence information whereas Ab-Ligity can use structural similarities to identify some of the sequence-dissimilar antibodies missed by paratyping.

Performance on Crystal Structures						
Method	Antibody Set	Precision	Recall	TP	FP	FN
Ab-Ligity	Full	0.98	0.96	622	10	27
Paratyping	Full	0.98	0.91	590	14	59
Ab-Ligity	CDRH3 <0.8	0.99	0.90	193	1	21
Paratyping	CDRH3 <0.8	1.0	0.81	174	0	40
Both	CDRH3 <0.8	1.0	0.90	172	0	19
Performance on Modelled Structures						
Method	Antibody Set	Precision	Recall	TP	FP	FN
Ab-Ligity	Full	0.84	0.97	626	120	15
Paratyping	Full	0.98	0.91	590	14	59
Ab-Ligity	CDRH3 <0.8	0.66	0.93	192	100	15
Paratyping	CDRH3 <0.8	1.0	0.81	174	0	40
Both	CDRH3 <0.8	0.99	0.92	173	1	15

Table 3.3: Precision-recall values for prediction of paratope similarity for crystals and models of antibody pairs using Ab-Ligity and paratyping. Methods used both heavy and light chains, on all possible antibody pairs within non-redundant SAbDab or only on pairs of antibodies with diverse CDRH3 sequences (CDRH3 identity <0.8 antibody set). Ground truth of antibody similarity was established using Ab-Ligity similarity of crystal epitope structures. Numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) used in the precision and recall calculations are shown; see Section 3.2.2.5 for formulae.

3.3.1.2 Performance on antibody models

Paratyping performance is the same for modelled and crystal structures as the method uses only sequence information. The precision of Ab-Ligity decreased when using models, with many more false positives (FP) compared to using crystal structures (120 vs 10, Table 3.3), largely from the CDRH3 <0.8 subset. This suggested the models were not perfect representations of the crystal structures, which is not surprising. On the other hand, paratyping identified few FP but failed to detect a number of similar antibodies (59 compared to 15 by Ab-Ligity, on the full set). Antibody pairs which were predicted as similar by both Ab-Ligity and paratyping methods gave the best performance in terms of precision and recall; however, this does result in missing similar antibodies which Ab-Ligity alone identifies.

3.3.1.3 Performance on heavy chain data only

Light chain sequence information is not always available for antibodies, for example if bulk sequencing methods have been used (see Chapter 1, Section 1.2.1). Therefore, it is important to have tools that are still accurate when applied to the heavy chain only. Ab-Ligity and paratyping performance was assessed on the heavy-chain only non-redundant SAbDab dataset, and showed decreased precision and recall for both methods (Table 3.4). Ab-Ligity performance was more affected than paratyping, possibly due to inaccuracies when models are built using only the heavy chain. As was seen when using both heavy and light chains, Ab-Ligity identified a higher number of FP, but also more true positives (TP), than paratyping.

Method	Antibody Set	Precision	Recall	TP	FP	FN
Ab-Ligity	Full	0.75	0.89	588	191	76
Paratyping	Full	0.95	0.84	564	29	101
Ab-Ligity	CDRH3 <0.8	0.51	0.79	173	165	47
Paratyping	CDRH3 <0.8	0.97	0.74	163	5	57
Both	CDRH3 <0.8	0.97	0.78	160	5	44

Table 3.4: Precision-recall values for prediction of paratope similarity by Ab-Ligity and paratyping using heavy chain information only, on antibody pairs in non-redundant SAbDab.

3.3.1.4 Comparison of identified hits

To explore the difference in similarity predictions between Ab-Ligity and paratyping, pairs of antibody models identified as similar (referred to as 'hits' throughout

the remainder of this chapter) by only one method were evaluated. Here, we evaluated the methods on the redundant SAbDab set, as it contains a higher number of paratope-epitope pairs than the non-redundant set and was therefore used in the epitope-to-paratope pipeline development.

Ab-Ligity and paratyping both performed well at identifying similar antibodies, with high precision and recall values. The methods identified different hits; Ab-Ligity found 100 TP pairs which paratyping did not, and 3 TP pairs were found only by paratyping. This represented a small proportion of the total pairs of similar antibodies as 12,054 TP were found by both methods. Interestingly, the hits identified only by paratyping could be found by Ab-Ligity when the crystal structures of the antibodies were used, suggesting there may be inaccuracies within those antibody models. Many of the similar antibody pairs which only Ab-Ligity identified contained the antibody 1ndg. For these pairs, the paratyping score was not far below the similarity threshold, with values of 0.59-0.71. This suggested the antibodies may have different paratope amino acid identity but similar physicochemical properties, and/or similar binding site shape.

Without a CDR length matching constraint, Ab-Ligity identified an additional 275 TP antibody pairs. These were not identifiable by paratyping as the method requires the same length CDR regions. This is a potential advantage of Ab-Ligity, though Ab-Ligity is in principle more computationally expensive than paratyping due to the modelling stage (which takes around 30 seconds per structure with ABodyBuilder (Leem et al. 2016)). Ab-Ligity should be able to identify more paratope-similar sequence-diverse antibodies as the method compares binding site residue physicochemical group rather than exact residue identity, and it explicitly considers 3D location of the paratope residues rather than sequence alignment matching. In addition, Ab-Ligity does not require identical length CDR regions (which paratyping does), though we have generally used CDR length matching in the pipelines to improve the precision of hit identification; the drawback of such an approach is that it may miss some similar antibodies, as CDRH3 sequences with different lengths can engage the same epitope (Wong et al. 2021).

3.3.2 Identification of binding antibodies to a target epitope, given a known binding sequence

We first designed pipelines to identify antibodies likely to bind to the same epitope, based on paratope similarities to a known binding sequence. This is a hit expansion step which could be useful in early stage antibody development; if there is an antibody known to bind to the target antigen, the pipeline could predict binders from an antibody dataset based on similarities to the known binder, to take forward in development. Clonotyping is an established method to find other binding antibodies in hit expansion (see Chapter 1, Section 1.5.4.1), but is limited to identifying antibodies similar in sequence. Sequence-diverse antibodies may be preferential to identify to reduce the risk of all hits suffering from aggregation or immunogenicity issues, which could occur with similar sequences (Raybould et al. 2019b; Richardson et al. 2021).

A known group of 13 lysozyme-binding antibodies was selected for proof of concept (Table 3.1). These antibodies have very high antigen epitope similarity and antibody paratope similarity, as defined using Ab-Ligity. They are highly sequence similar, with the same V and J genes, and have an average epitope similarity score of 0.45 (range 0.21 - 0.77). These sequences therefore cannot reveal whether the pipeline can identify highly sequence dissimilar antibodies but offer a simple proof of principle. The lyosyzme-binding antibodies were chosen as they were the largest group of non-sequence-identical antibodies binding to the same epitope within SAbDab.

Using one of these lysozyme binders, 1a2y (Dall'Acqua et al. 1998), as the target antibody, the remaining 12 lysozyme binders were moved from SAbDab into the AML. Using the pipeline (Figure 3.4) we successfully identified the lysozyme binders in the AML as similar to our target antibody with Ab-Ligity, without predicting any of the other AML antibodies as similar. Whilst this example shows similar antibodies can be distinguished in the AML, this may be due to the lysozyme-binding examples being derived from mouse and having a length 10 CDRH3, which is shorter than most AML antibodies (only 380 shared structures in the AML have CDRH3 length 10). We expanded our set of test target antibodies beyond 1a2y in subsequent analysis (Table 3.2).

3.3.3 Prediction of binding antibodies to a target epitope, without a known binding sequence

We also investigated identifying antibodies likely to bind the same epitope without relying on known binders. This epitope-to-paratope approach would be useful in hit generation against a target, and beneficial for early stage antibody discovery.

To gain proof of concept for the pipeline (Figure 3.7), we returned to the lysozyme binders. Using 1a2y as the target antibody, we moved the remaining 12 lysozyme binders one by one from the redundant SAbDab into the AML, to assess whether we could still identify the remaining lysozyme binders in the AML (Figure 3.6). Similar epitopes to 1a2y were identified in SAbDab, and then the corresponding antibodies to these epitopes were used to look for similar antibodies within the AML. Using Ab-Ligity, the lysozyme binders could be distinguished from the AML without identifying any other AML antibodies as similar, provided at least one known binder was still contained in SAbDab so that a good epitope similarity match could be identified. If no similar epitopes were identified within SAbDab, because the lysozyme binders had all been moved to the AML, performance was poor as no other similar epitopes in SAbDab could be identified. The non-lysozyme epitope with highest similarity scored 0.027, well below the Ab-Ligity similarity threshold of 0.1. Taking the corresponding antibody for this epitope forward to search for similar antibodies within the AML did not lead to identification of the known lysozyme binders. The lack of identification of a suitably similar epitope is likely common, as the redundant SAbDab database contained only 2258 paratope-epitope pairs.

3.3.3.1 Incorporation of sequence-agnostic Ab-Ligity to identify potential binders

The pipelines discussed above utilised an AML to search for similar antibodies, where many antibody sequences map onto the same structure. The sequences of the 23,700 antibody models in the AML are merely representative of the sequences in their structural cluster and do not correspond to a consensus sequence, therefore it may be beneficial to ignore the sequence and initially look only for structural similarity. We therefore developed a sequence-agnostic version of Ab-Ligity.

The sequence-agnostic method calculates the similarity of antibody binding sites solely based on the spatial information of the antibody paratope, ignoring physico-chemical amino acid information (see Methods 3.2.2.4). In incorporating this method

into the pipeline, we hoped to better identify antibodies in the AML with similar structures to our identified SAbDab paratopes. The method could also be used to identify SAbDab epitopes based only on structural similarity to the target epitope.

To determine an appropriate similarity threshold for sequence-agnostic Ab-Ligity, we looked at the correlation between Ab-Ligity similarity score and sequence-agnostic Ab-Ligity score for pairs of antibodies in SAbDab. It proved challenging to identify an obvious similarity threshold, as a large number of pairings had low sequence-specific scores (i.e. below the 0.1 Ab-Ligity similarity threshold and defined as not similar) but high sequence-agnostic scores. Comparison of antibodies in SAbDab to those in the AML showed the same trend, including pairs with a sequence-specific score of 0 and sequence-agnostic values of up to 0.75. We had expected to see higher similarity scores with the sequence-agnostic method compared to sequence-specific as antibodies may have very high binding site shape similarity but very low amino acid physicochemical similarity, and therefore score highly with sequence-agnostic but low with sequence-specific Ab-Ligity.

We focused on correct identification of all similar antibody pairs, to maximize the number of true positives whilst minimising the number of false positives. With the sequence-agnostic method we obtained huge numbers of false positives; for example, if a 0.6 threshold was used, 367 TP pairs were missed but 72603 FP pairs identified (Table 3.5). When searching the AML, we hope to identify structural families which might contain sequences of higher similarity to the target than the structural family representative sequence, therefore some FP are acceptable. Raising the similarity threshold improved precision by reducing the number of FP, but this also increased the number of FN as similar antibodies were missed. Sequence-agnostic Ab-Ligity may identify a higher number of similar antibodies, and more sequence-diverse similar antibodies, than would be obtained with sequence-specific Ab-Ligity.

As a test case, we used the same lysozyme target antibody, 1a2y, as used in the previous pipeline, and aimed to identify the other lysozyme binders from a combined dataset of SAbDab and the AML using sequence-agnostic Ab-Ligity. Using CDR length matching and a 0.4 paratope similarity cut off (to minimise the number of FN - see Table 3.5), we identified 23 hits from SAbDab and 99 hits from the AML. Of the 23 SAbDab hits, the 12 known lysozyme binders had the highest similarity scores (range 0.70 - 1.0). With a threshold of 0.6 (which may increase FN but reduce FP),

Threshold	Precision	Recall	TP	TN	FP	FN
0.1	0.10	1.00	12497	0	111602	0
0.2	0.10	1.00	12497	6	111596	0
0.3	0.10	1.00	12497	124	111478	0
0.4	0.10	1.00	12497	1548	110054	0
0.5	0.11	0.98	12253	7530	104072	244
0.6	0.14	0.97	12130	38999	72603	367
0.7	0.62	0.92	11483	104666	6936	1014
0.8	0.97	0.87	10843	111254	348	1654

Table 3.5: Precision-recall values for prediction of paratope similarity using sequence-agnostic Ab-Ligity performed on pairs of antibodies in SAbDab with varying similarity thresholds.

we identified 19 hits in SabDab (including all the lysozyme binders) and 66 AML hits. With a threshold of 0.7 (further increasing FN but reducing FP), we identified all the lysozyme binders and only 1 AML hit.

We have previously discussed that the lysozyme binding antibodies, whilst a useful example as a group of antibodies with known highly similar epitopes, may be biased due to the corresponding SAbDab antibodies being of mouse origin. We also tested our pipeline with a human antibody against a chimeric glycoprotein from respiratory syncytial virus (RSV) and human immunodeficiency virus (HIV), 5u68 (Wen et al. 2017), which has 8 sequence-specific Ab-Ligity hits in SAbDab. Using the sequence-agnostic method, we identified 54 SAbDab hits and 982 AML hits with a 0.6 similarity threshold, and 11 SAbDab hits and 250 AML hits with a 0.7 threshold.

3.3.3.2 Comparison of sequence-specific and sequence-agnostic Ab-Ligity hit identification

To assess the benefit of using sequence-agnostic Ab-Ligity, we ran comparisons of our pipeline with both sequence-specific and sequence-agnostic Ab-Ligity. We tested two antibodies previously discussed, lysozyme binder 1a2y and HIV/RSV binder 5u68, and a further three antibodies from SAbDab: 4i18 (Rizk et al. 2015), 5eu7 (Galilee et al. 2016) and 6apd (Goodwin et al. 2018) (see Table 3.2). We chose to use the antibodies directly, rather than their epitopes, for this comparison. These antibodies were selected to test the pipeline as they had high numbers of similar antibody paratopes identified in the AML by sequence-specific Ab-Ligity (699, 351 and 253

CDRH length matched hits respectively).

Using these five target antibodies, we identified their corresponding sequence-agnostic hits in SAbDab and subsequently the AML hits arising from these SAbDAb hits. We performed the same analysis using sequence-specific Ab-Ligity (see Table 3.6). These hits were then 'checked' by assessing their similarity to the original target antibody, using either sequence-specific or sequence-agnostic Ab-Ligity. Using a sequence-agnostic threshold of 0.6, we found more hits than with sequence-specific Ab-Ligity. With the sequence-agnostic method, 61-88% of the identified AML hits were above the similarity threshold to the starting antibody. With the sequence-specific method, this was the case for 48-72% of AML hits. This suggests the sequence-agnostic method could be particularly useful for instances where we identify no sequence-specific hits.

		Sequence-agnostic			Sequence-specific		
Name	CDRH3 length	SAbDAb hits	AML hits	Checked hits	SAbDAb hits	AML hits	Checked hits
1a2y	10	19	91	66	12	0	0
4i18	18	51	1664	1212	19	850	610
5eu7	23	3	840	743	0	0	0
5u68	17	54	1609	982	8	607	361
6apd	15	127	1383	1036	9	484	230

Table 3.6: AML and expanded AML hits identified using sequence-agnostic Ab-Ligity and sequence-specific Ab-Ligity. SAbDab was searched for paratopes similar to the the target (SAbDab hits). These were used to search for paratope-similar antibodies in the AML (AML hits). All paratope comparisons were length-matched across VH CDRs. A paratope similarity threshold of 0.6 was used for sequence-agnostic Ab-Ligity. AML hits were assessed for paratope similarity to the original (modelled) antibody, calculated using sequence-specific or sequence-agnostic Ab-Ligity; 'checked' hits exceeded the similarity threshold.

3.3.3.3 Paratope similarity of AML sequences to target antibody

To investigate whether the AML works as an useful dataset format to identify antibodies, the sequences corresponding to each AML structure identified as similar were retrieved and assessed for their similarity to the target antibody. We refer to these as the expanded sequences for a structural family. Due to the large numbers of sequences involved, an example subset of the top AML hits (as identified in Table 3.6, with sequence-specific Ab-Ligity similarity $\geq$0.2) for 4i18 was chosen, and the

similarity scores of the all sequences corresponding to these AML hits was calculated. This included 18 AML hits, corresponding to 1460 expanded sequences. Of these, 1310/1460 were sequence-specific Ab-Ligity hits to 4i18, a hit rate of 87.9%. This suggests that using the AML is a good way to screen a large amount of antibodies more quickly than standard BCR datasets, as we can run initial comparisons against tens of thousands of shared structures rather than millions of sequences.

3.3.3.4 Clonotype diversity within AML families

As the expanded sequences from the AML hits to target antibody 4i18 were predicted to also have similar paratopes to the target, we investigated the sequence diversity contained within an AML family to determine whether sequence-diverse hits were likely to be identified in this manner. We performed clonotyping on a number of AML families (clonotype definition of 90% CDRH3 sequence identity, identical V and J gene assignment, VH only given identical VL sequences within a family) using Change-O (Gupta et al. 2015). We found that the majority of sequences within an AML family fell into individual clonotypes. V gene diversity within a structural family contributed to this; there were an average of 8.93 different V genes per structural family. As an example, the 85 unique VH sequences associated with AML public structure H10481+L36568 fell into 83 clonotypes (10 different V genes, all identical J gene). Another structural family, H14206+L16211, included 40 unique sequences which fell into 40 distinct clonotypes (5 V genes, 3 J genes). This suggests that identification of a hit AML antibody structure could be used to identify paratope-similar, sequence-diverse antibodies.

We compared the sequence diversity within an AML family to sequence similarity across the whole of the AML. Clonotyping across the whole AML revealed clonotypes are shared across the 23,700 AML structural families, with 57,468 clonotype clusters (60,568 unique CDRH3, V + J gene combinations). The largest cluster contained 238 sequences. This revealed that similar sequences were spread across different structural families, suggesting that the method would provide different hits to those obtained by clonotyping, and combining these methods may be beneficial.

3.3.4 Prediction of binding antibodies to a target epitope, without known binders - PB-AML2

3.3.4.1 Impact of AML architecture on similar structure identification

An updated AML, PB-AML2, was built by Matthew Raybould with altered sequence processing, modelling and clustering (see Methods 3.2.1.4). We compared the AMLs using our five SAbDab targets (Table 3.2), to look directly for similar antibodies within the PB-AML2 using sequence-specific Ab-Ligity (rather than first looking for similar epitopes in SAbDab). Far fewer antibodies in the PB-AML2 were predicted to be similar to our SAbDab targets compared to the PB-AML1 (Table 3.7). We had been somewhat surprised at the high number of antibodies in PB-AML1 predicted as similar to our SAbDab targets, and hoped the reduction in number of similar antibodies might reflect higher accuracy of hit identification. The PB-AML2 contains more distinct structural clusters compared to the previous AML (37,533 structural families compared to 23,700) due to higher similarity requirements in the clustering method, therefore identified hits are likely to be more structurally distinct.

We found a striking difference in the number of AML structural families predicted as similar to any of the SAbDab antibodies; previously, 1383/2253 (61%) of SAbDab antibodies had no Ab-Ligity CDR length matched hits in the AML, whereas in PB-AML2 2209/2253 have no hits (98%). We would not expect every SAbDab antibody to have similarities to the healthy human Public Baseline AMLs, given differences in species and target epitopes. However, given the number of human antibodies in SAbDab, we anticipated greater than 2% to have AML hits. Interestingly, we found antibodies predicted to be similar to mouse Fab 4i18 (binding human prolactin receptor) in both AMLs (Table 3.7). No AML hits were identified for 6iec, which neutralises rift valley fever virus. This may be partially due to the short CDRH3 length of 6iec compared to antibodies in the AML; 6iec CDRH3 is only length 12, so with length matching constraints only 10 hits were identified in the PB-AML1. We did not identify any similar structures to 5u68 which was unexpected - as a human antibody binding to respiratory syncytial virus, we might expect to find similar structures in human repertoires (CDRH3 length 17).

We hypothesized that structures from PB-AML2 which were predicted as similar to a target may contain a higher proportion of sequences also predicted to be similar to the target (we refer to this as a higher hit rate within the expanded family).

Structural cluster family members were assessed for similarity to the structural cluster representative and were found to be more similar than was found in the previous AML. An average of 6.5% of antibodies from the 2 hit AML structural families were similar to target antibody 6apd (Table 3.7). The hit rate for 4i18 was similar at 5.5% across 6 structural families (hit rate per family ranged from 2.3 - 19.0%), whilst human antibody against HIV integrase, 5eu7, achieved the highest number of expanded hits at 20.5% across 25 hit structures.

		No length match		CDR length match		
Target	CDRH3 Length	Hits in AML1	Hits in AML2	Hits in AML1	Hits in AML2	Expanded hits from AML2
4i18	18	9091	184	699	6	63/1147
5u68	17	4506	3	434	0	-
5eu7	23	6396	282	351	25	1560/7593
6apd	15	3923	15	253	2	102/1579
6iec	12	196	0	10	0	-

Table 3.7: Comparison of the number of AML models identified as similar to the SAbDab target antibodies in PB-AML1 and PB-AML2. Performance was assessed with and without CDRH length matching requirements. Expanded hits were identified by retrieving sequences corresponding to the AML structures and assessing paratope similarity to the SAbDab target antibody. All similarity calculations were done using sequence-specific Ab-Ligity.

3.3.4.2 Alternative hit identification

As no similar antibodies were identified in the PB-AML2 for a high number of SAbDab targets, including those we expected to see some similarity to, we investigated whether the most similar AML structures (with the highest Ab-Ligity scores but below the similarity threshold) contained any sequences that would be classed as hits. For 6iec, the AML family with the highest Ab-Ligity score contained 1 hit antibody out of 287 sequences. Despite the low hit rate within the family, this suggested that if no AML families are above the similarity threshold, lowering the threshold might lead to identification of hits.

We also investigated returning to the pipeline using Ab-Ligity to first identify a similar epitope within SAbDab, then taking the corresponding SAbDab antibody which binds to that epitope and using it to search for similar paratopes in the AML. For 6iec we identified the 6fla epitope as similar, however, searching in the AML with

the 6fla antibody did not lead to identification of any similar antibodies (highest Ab-Ligity score of 0.03). We also tried looking directly for similar paratopes in SAbDab, to then search the AML with. 3r1g has the highest model paratope similarity to 6iec, but searching the AML revealed no similar antibodies to 3r1g above the 0.1 threshold. The AML structure with the highest similarity to 3r1g (Ab-Ligity score of 0.08) did not contain any other hits within the sequence family (highest similarity 0.04). However, assessment of the 3r1g crystal structure reveals a low epitope similarity to 6iec of 0.0095, so may not be a suitable choice regardless. This was a reminder that whilst our pipeline assumes that epitope similarity means paratope similarity, antibody paratope similarity does not guarantee epitope similarity.

Our analysis suggests Ab-Ligity is not sufficient to identify hits within the PB-AML2 for SAbDab targets which do not initially have hits, including for targets where we would expect to find similar antibodies in the AML. We therefore explored alternative methods and strategies for determining paratope similarity.

3.3.4.3 PSSMs fail to distinguish hits from non-hits

When potential hits within PB-AML2 were identified, the subsequent hit rate within the expanded families was often fairly low. Whilst we did not expect all expanded sequences to be similar to the target, due to sequence diversity within the structural families, for some families only the cluster representative sequence scored above the similarity threshold to the target. This suggests that the sequence assigned to each AML structural cluster - from participant 'S64' - may differ from the other sequences included within AML family structural family (see Methods 3.2.1.4 for details on the AML) and may not best reflect all the sequences mapped onto the structure. This also suggests that other AML families may contain potential hits, but are missed due to dissimilarity between the target and their sequence representative. We tested an alternative searching approach that would be less dependent on the one assigned sequence and therefore might be less skewed by a cluster representative sequence differing from the sequences within an AML family.

Position-specific sequence matrices (PSSMs) were generated for each of the AML structural families (see Methods 3.2.2.9). These were created using the frequencies of amino acids in the CDRs of antibodies within the structural family, using only the predicted paratope residues across the CDRs (using Parapred (Liberis et al. 2018)),

and also using only the predicted clonotype (Methods 3.2.2.9). A PSSM of the target SAbDab antibody was compared to PSSMs from all AML families. We also incorporated a BLOSUM substitution matrix into the scoring to account for similar amino acids within positions rather than requiring exact amino acid identity, and compared paratope residue physicochemical properties rather than amino acid identity. For SAbDab target 4i18, we found that only scoring AML families based on CDR PSSMs, without BLOSUM, generated scores which could separate hit families from non-hits (Table 3.8). Comparison of physicochemical groups and incorporating BLOSUM lead to varied scores, which overlapped between hit and non-hit families. These results were not consistent across different target antibodies, with amino acid PSSM comparison unable to identify hits to antibody 6iec.

AML structural family	Amino acids	Physicochemical groups	Amino acids + BLOSUM
H10243+L3513	-23.73	0.54	-1.5
H12545+L44982	-4.71	9.75	-6.57
H54157+L112494	-30.48	-12.91	2.1
H13063+L16916	-7.66	23.76	-6.1
H54157+L76109	-19.35	0.61	6.1
H1621+L94608	-31.87	19.56	-6.9
Non-hit families	-36.80 to -60.2	-26.30 to +17.03	-11.9 to +4.1

Table 3.8: Similarity of 4i18 to AML structural family hits (determined using Ab-Ligity scoring) and non-hits, using PSSM-based methods. PSSMs were based on amino acid composition of the CDRs, physicochemical groups of the amino acids in the CDRs, and incorporated amino acid substitutions using a BLOSUM matrix. Scores across four AML families with no or very few hits was also calculated.

To investigate why the PSSM approach did not enable better sorting of hits and non-hits, we analysed the diversity of sequences within the AML hits. Based on the clonotyping analysis of the previous AML, we knew that similar sequences are found in different AML structural families. Performing clonotyping on members of the six Ab-Ligity hit families to 4i18 in the PB-AML2, we found similar results; many clonotypes were found per family (see Table 3.9). There was some overlap across the different structural families, with 54 shared CDRH3 sequences out of 491 unique CDRH3s present across the six families. This may contribute to why the sequence-based representations of AML families, based on the average of a family, were not helpful; if families are diverse but share sequence overlap with other families, averaging across a family may make families of hits and non-hits more similar.

AML structural family	Number of sequences	Unique CDRH3s	Clonotypes (VH)	Clonotypes (VH+VL)
H10243+L3513	165	56	56	80
H12545+L44982	212	114	115	135
H54157+L112494	174	40	39	74
H13063+L16916	476	206	199	322
H54157+L76109	18	14	14	15
H1621+L94608	94	61	61	73

Table 3.9: Clonotyping within AML structural families that were hits to 4i18 (similar to 4i18 using Ab-Ligity). Clonotyping was performed using VH only, and VH and VL. Clonotypes were defined as identical V and J genes, 90% CDRH3 identity.

Looking at the sequence and paratope residue conservation (Figure 3.10) in one AML structural cluster family as an example, we see high sequence diversity across residues 106-112 in the CDRH3 (IMGT numbering, IMGT CDRH3 definition). This diversity may lead to PSSMs built on sequence information not being a useful representation of an AML family, and therefore not helpful in determining hits.



Figure 3.10: A) Sequence conservation and B) paratope conservation across the CDRH3 of an example AML family, H10243+L3513, which is a hit to 4i18 (produced using SeqLogo). Paratope residues were predicted using Parapred.

3.3.4.4 Sequence-agnostic Ab-Ligity identifies additional hits

Given the lack of success with various sequence and physicochemical property representations of the AML families, we investigated whether sequence-agnostic Ab-Ligity could be used to identify similar antibodies. The sequence-agnostic method identified hits in PB-AML2 for all five SAbDab targets (Table 3.10) at a similarity threshold of 0.45. Furthermore, expanded sequences associated with the structural hits were similar to the SAbDab target, as determined by sequence-specific Ab-Ligity (0.2 - 12.7% of sequences per structural family). Compared to sequence-specific Ab-Ligity, additional hits were identified for all SAbDab targets, including those where no hits had previously been identified (9 found for 6iec, 21 for 5u68). This suggests the sequence-agnostic method may be a useful way of identifying hits in the AML when no hits are identified by the more stringent sequence-specific approach. We also looked at the overlap of hits identified by the methods; 1 AML hit to 6apd and 5 to 5eu7 were found by both methods. These structural clusters both contained expanded hits (90 for 6apd, 1074 across the 5 structural families for 5eu7).

		Sequence-specific		Sequence-agnostic		Both
Target Antibody	CDRH3 Length	Hits in AML	Expanded hits	Hits in AML	Expanded hits	Expanded hits
4i18	18	6	63/1147	8	45/2474	0
5u68	17	0	-	2	21/3060	0
5eu7	23	25	1560/7593	24	1207/9478	1074
6apd	15	2	102/1579	119	718/100082	90
6iec	12	0	-	9	9/4466	0

Table 3.10: Comparison of numbers of sequence-specific and sequence-agnostic Ab-Ligity hits for SAbDab targets in PB-AML2. sequence-agnostic were hits determined using a lowered threshold of 0.45. Expanded sequence-agnostic hits were antibody sequences associated with the AML structure which were determined as similar to the target using sequence-specific Ab-Ligity. The number of expanded hits identified by both methods are listed in the final column.

Due to the number of AML hits, and subsequent high number of expanded sequences, we looked at potential methods for evaluating and screening the expanded sequences to only select the most likely-binding antibodies.

3.3.4.5 Ab-Ligity hits occupy different paratope clusters

We assessed whether the hit sequences within an AML family share paratope similarities, so that once an AML structure had been identified, we might perform paratyping

on the expanded sequences and pull out sequences with the most physicochemically similar paratope to our target. We found that the hit sequences were sometimes in the same paratype, though not consistently. We identified many different paratope clusters per AML family, across VH chain only and across both VH and VL chains. Within the six hit AML families similar to 4i18, up to 70% of sequences do not cluster with any other sequences based on predicted paratope (Table 3.3.4.5). For one AML hit family, the 9 Ab-Ligity hits fell into 9 separate paratope clusters, though for another, the 9 hits fell into 2 separate clusters.

Not all antibody sequences within the same paratype were Ab-Ligity hits. Generally, only one or two sequences in a paratype would have similarity above the Ab-Ligity threshold. We show an example of this for 4i18 but similar results were found for other SAbDab targets.

AML structural family	Number of sequences	VH paratope clusters	VH+VL paratope clusters	Clusters containing Ab-Ligity hits
H10243+L3513	165	51	54	7 (18 hits)
H12545+L44982	212	102	99	5 (7 hits)
H13063+L16916	476	186	207	7 (10 hits)
H1621+L94608	94	58	58	6 (10 hits)
H54157+L112494	174	31	39	2 (9 hits)
H54157+L76109	18	13	14	9 (9 hits)

Table 3.11: Number of paratope clusters within 4i18 hit AML structural families (similar to 4i18 using Ab-Ligity scoring). Clusters were obtained using paratyping, with a similarity threshold of 0.75. Paratyping was performed using heavy chain only, and heavy and light chain.

3.3.4.6 Position-only paratyping

Given the sequence diversity within AML families, we adapted paratyping to cluster based on similar paratope residue positions only, ignoring amino acid or physicochemical properties of the residues. This is similar to the adaptation of Ab-Ligity to sequence-agnostic Ab-Ligity. Clustering in this way drastically reduced the numbers of clusters within an AML family to 5-10% of the number of standard paratyping clusters (Table 3.12). Again, the Ab-Ligity hits often occupied different paratope clusters. For example, the 476 sequences within the H13063+L16916 structural family fell into 16 position-paratope clusters, compared to 207 standard paratyping clusters. The 10 Ab-Ligity hits within this family were found in 6/16 position-only paratyping

clusters, compared to 7/207 standard paratyping clusters. Thus, the paratope-based clustering of expanded sequences does not appear to be a valuable way to identify or stratify the identification of potentially-binding antibodies from this AML.

AML structural family	Position-only paratope clusters	Clusters containing Ab-Ligity hits
H10243+L3513	2	1 (18 hits)
H12545+L44982	5	3 (7 hits)
H13063+L16916	16	6 (10 hits)
H1621+L94608	5	3 (10 hits)
H54157+L112494	4	2 (9 hits)
H54157+L76109	2	2 (9 hits)

Table 3.12: Position-only paratope clustering of hit AML structural families to 4i18 (similar to 4i18 using Ab-Ligity scoring). Clustering encompassed heavy and light chains, with a similarity threshold of 0.75. The number of position-only paratope clusters containing hits to 4i18 based on Ab-Ligity similarity score from each structural family was noted (and the number of hits contained within these clusters).

3.4 Discussion and future work

In this chapter, we have demonstrated that the existing techniques of Ab-Ligity and paratyping can identify sequence-dissimilar antibodies binding to the same epitope, based on paratope similarity. Our analysis of the methods' performance on known paratope-epitope pairs in SAbDab showed that paratyping missed similar antibodies more frequently than Ab-Ligity, as the method relies entirely upon sequence information, whereas Ab-Ligity can identify structural similarities in sequence-dissimilar antibodies. The methods largely identified the same hits (antibodies which were similar), though with a small number of differences; inaccuracies within antibody models disrupted Ab-Ligity similarity scoring, while paratyping missed some antibodies with dissimilar paratope sequences which shared structural and physicochemical similarities.

We incorporated these techniques into a pipeline to identify antibodies likely to bind to a target epitope, with and without knowledge of a known binding antibody. Our aim was to develop a method which could identify likely binders within a database or repertoire of antibody sequences. We evaluated AMLs, where multiple sequences can map onto each public antibody structure. This allowed for screening of a greater

number of antibodies more quickly than if searching through all sequences individually (similarity calculations to thousands of structures vs millions of sequences). We tested our approach on two versions of an AML, which utilised differing cluster methods. We found substantial differences in the number of antibodies we predicted as similar to our target using the different AMLs, suggesting the methods for AML generation have a large impact on subsequent analyses.

For two antibodies to bind to the same epitope, they likely have similar paratope structure, to fit close enough to the epitope to form interactions, without clashes, and similar key paratope residues which can form interactions with the epitope. When comparing antibodies to those in the AML using methods which include sequence or physicochemical information, such as Ab-Ligity, we relied on the one sequence assigned to that AML modelled structure. We attempted to negotiate this using two approaches, developing a sequence-agnostic method of Ab-Ligity which looked at paratope structural similarity only, and also by representing AML families as PSSMs using average CDR sequence identity, clonotype and paratype for better sequence-based comparisons. We found that the PSSM method was unable to accurately and robustly identify the Ab-Ligity hits from non-hits. We believe this may be due to sequence diversity within structural families. Using the sequence-agnostic method meant that we could first identify structurally similar paratopes, and subsequently evaluate the sequences that mapped onto these structures. Despite assessment of differing similarity thresholds, we were unable to find a sequence-agnostic threshold with an appropriate balance of false positives and false negatives; the method could either predict many structures as similar, or fewer structures which then missed structures which Ab-Ligity identified as hits. If many structures were predicted as similar, we attempted to then further evaluate these structures and the sequences which mapped to these structures. We performed clonotyping, paratyping and position-only paratyping on the sequences within each AML structural families to see if we could successfully group the sequences predicted by Ab-Ligity to be similar to our target antibody. We found many unique clonotypes, paratypes and position-only paratypes within each structural family, with hits spread across different groups.

Whilst the sequence-agnostic Ab-Ligity could be used to identify structurally similar families, and then the expanded sequences evaluated using sequence-specific Ab-Ligity, this could still involve many thousands of similarity calculations, depending on the similarity threshold used for sequence-agnostic Ab-Ligity and the number of AML

structural hits. An alternative option to explore could be using sequence-specific AbLigity to identify AML hits but using, instead of the representative cluster sequence, an 'average' sequence - for example, based on the most common physicochemical groups at each position in the sequence from the family members.

Our approach aimed to use a database of known paratope-epitope pairs (SAbDab) to predict likely binding antibodies to a target crystal structure of an epitope. This could allow us to predict a paratope for a given epitope, which could then be used to inform therapeutic antibody design, either as a template or as a tool to then search for similar antibodies within databases, repertoires or AMLs. At the time of this research, SAbDab contained 2258 antibody-antigen co-crystal complexes involving a protein antigen, with only 924 non-redundant antibodies (based on 100% sequence identity). The biases within SAbDab and the PDB, and the limited number of these known paratope-epitope pairings, imposed limitations in the ability to successfully identify similar epitopes. Even when using some SAbDab epitopes as targets, we were unable to find similar epitopes within SAbDab. If a novel antigen had a structure solved, with a known epitope, it is unclear whether an appropriate similar epitope could be identified. We also assumed that antibodies with similar paratopes bind to the same, or similar, epitopes, which is not always true; antibodies may have similar paratopes, sequence and/or structure but bind to different epitopes and antigens.

Since our research, Aguilar Rangel et al. have used a similar approach to attempt epitope-to-paratope design, but focused on identifying likely binding antibody fragments rather than full paratopes. Instead of using known paratope-epitope pairs, the authors generated a database of CDRs and CDR-like proteins from SAbDab and the PDB, and their corresponding epitopes. They then compared fragments of a target epitope crystal structure or modelled structure to their database to identify potential CDRs that would bind to the epitope fragment (Aguilar Rangel et al. 2022). These CDRs could then be grafted onto an antibody scaffold. This approach increases the likelihood of finding an epitope fragment with similarity to the target epitope.

As more paratope-epitope data becomes available, we may be able use this information to understand more about the relationship between paratope-epitope pairings and use this to inform therapeutic antibody design. The number of antibody-antigen structures deposited in the PDB is increasing every year; as of November

2022, SAbDab contains 4112 antibody-antigen complexes involving a protein or peptide antigen (Schneider et al. 2022b). There have also been substantial developments in antibody modelling and in paratope prediction since this research was conducted (Abanades et al. 2022a; Chinery et al. 2023). We are now able to model antibodies more quickly and with higher accuracy. This means that tools relying on antibody models, such as Ab-Ligity, could be used on larger datasets where previously sequence-based methods such as paratyping may have been preferred. Revising our pipeline using improved antibody models, with better prediction of paratope residues, and a larger SAbDab database, may improve performance. Advances in structure-based representations of proteins as graphs and sequence based representations using natural language processing embeddings (see Chapter 1, Section 1.2.5) may also be useful in comparing both epitopes and antibody paratopes to assess similarity.

CHAPTER 4

Structural Profiling of Antibodies to Cluster by Epitope

This chapter contains material reproduced from:

Robinson, S.A., Raybould, M.I., Schneider, C., Wong, W.K., Marks, C., Deane, C.M. (2021). Epitope profiling using computational structural modelling demonstrated on coronavirus-binding antibodies. PLoS computational biology, 17(12), e1009675.

I am a co-first author on the paper alongside Dr. Matthew Raybould. I conducted the analysis of available cocrystal structures of coronavirus-binding antibodies to develop the understanding of the epitopes on the SARS-CoV-2 receptor binding domain (RBD). I assessed the sequence and structural diversity of antibodies within these epitopes, including agreement with published epitope group information and analysis of antibody interactions with epitope residues. I analysed the largest structurally-convergent cluster of SARS-CoV-2 RBD binding antibodies to assess the conservation of paratope residues. Following the paper publication, I performed updated analysis on the Coronavirus Antibody Database (CoV-AbDab) including use of more recent antibody modelling methods. I have also contributed to the curation of the Coronavirus Antibody Database, CoV-AbDab, with the addition and update of over 4000 entries.

This chapter also includes unpublished research on malaria antibodies which was a collaboration with Professor Simon Draper's group from the Department of Bio-

chemistry, University of Oxford - in particular with Dr Kirsty McHugh, Dr Francesca Donnellan and Jordan Barrett.

4.1 Introduction

Identifying the site which an antibody binds, the epitope, is a key step in understanding the function of an antibody and its potential as a therapeutic. Experimental determination of epitopes can be achieved through laboratory techniques to assess antibody-antigen binding (see Chapter 1, Section 1.4); however, these techniques are time and resource intensive and therefore not suitable if investigating thousands of antibodies. Computational methods to predict the epitope information, using available sequence or structural data, would be a useful tool to select antibodies for further therapeutic development.

The COVID-19 pandemic has generated worldwide efforts to isolate and characterise antibodies able to confer protection against SARS-CoV-2. Hundreds of studies have released data on diverse antibodies and nanobodies capable of binding at least one coronavirus antigen (Raybould et al. 2021a). Due to the huge number of individuals infected by SARS-CoV-2, most of the reported coronavirus-binding antibodies have been sourced directly from the blood of convalescent human patients. The primary technique used to identify such binders is 'serum baiting', where an extracellular coronavirus antigen (often the receptor binding domain (RBD)) is used to pan donated blood serum directly for complementary antibodies (Yan et al. 2020; Yang et al. 2020). Deep sequencing of the SARS-CoV-2 convalescent BCR repertoire can also be utilised, which can implicate particular expanded antibody lineages as important to adaptive immunity, without biasing towards a chosen antigen bait (de Assis et al. 2021; Galson et al. 2020; Zhao et al. 2005). Other antibody discovery methods have included mining surface display libraries, challenging and harvesting transgenic animals, and antibody engineering (Raybould et al. 2021a).

The Coronavirus Antibody Database (CoV-AbDab) includes sequence and, where possible, structural information of coronavirus-binding antibodies and nanobodies sourced from patents and the academic literature (Raybould et al. 2021a). One way to use this collated data is to look for similarities between binders. For instance, when a novel antibody is antigen-baited out of SARS-CoV-2 response serum, or is identified as an expanded clonal lineage post-SARS-CoV-2 infection, one can assess

whether it bears resemblance to any other antibody previously reported to bind a coronavirus. This resemblance can then be used to predict functional properties of the newly-isolated antibody, such as its epitope.

A common way to cluster antibodies into such functional groupings is clonotyping (see Chapter 1, Section 1.5.4.1). This can be performed in several different ways depending on the criteria for clustering. For example, strict Fv-clonotyping requires identical VH and VL V and J genes, identical CDRH3 and CDRL3 lengths and similar sequence identities (using a threshold close to 100% per CDR3 region) (López-Santibáñez-Jácome et al. 2019). This approach typically yields tight and functionally-significant clustering but is severely limited by its ability to bring together all antibodies able to engage a particular epitope (Kovaltsuk et al. 2017; Richardson et al. 2021; Wong et al. 2021). As a result, leniency is often introduced to the clonotyping protocol, by lowering the sequence identity threshold to 80% (Soto et al. 2019), ignoring J gene annotations, and/or only considering the heavy chain (VH-only clonotyping) (Galson et al. 2020). Convergent lenient VH-only clonotypes have been identified between multiple SARS-CoV-2 infected or convalescent individuals (Bullen et al. 2020; Kreye et al. 2020; Montague et al. 2021; Robbiani et al. 2020; Schultheiß et al. 2020; Sokal et al. 2021; Zhang et al. 2020) and across different studies, for example the overlap between the clonotypes found by Galson et al. 2020 and Nielsen et al. 2020. Several papers have compared BCR sequences from individuals to verified SARS-CoV-2 binders in CoV-Ab-Dab and identified clonal similarities, e.g. Bullen et al. 2020; Galson et al. 2020.

Despite this, the clonotypes found to be enriched or to bear similarity to CoV-AbDab antibodies post SARS-CoV-2 exposure often differ across studies (Ju et al. 2020; Xiang et al. 2022; Yuan et al. 2021). This may be partly due to the small sample sizes used in individual studies, and the intrinsic biases in individual VJ gene usage; Xiang *et al.* found a larger variation between individuals within a cohort (healthy, or three different severities of COVID symptoms) than between cohorts (Xiang et al. 2022). Another contributing factor is clonotyping itself. Levels of functional convergence may in fact be higher than implied even by lenient clonotyping, as antibodies that derive from different lineages can engage the same epitope. The evidence for this phenomenon has been growing over recent years (Kovaltsuk et al. 2020; Raybould et al. 2021b; Rees 2020; Richardson et al. 2021; Wong et al. 2021). For example, solved

structures of antibody-antigen complexes reveal pairs of antibodies with different genetic origins but sufficiently similar binding site geometry and paratope similarity that they bind to the same antigen with near-identical binding modes (Richardson et al. 2021; Wong et al. 2021). Furthermore, given that individuals' naive repertoires typically share very few clonotypes (Briney et al. 2019; Soto et al. 2019) and yet are often found to respond to similar 'immunodominant' epitopes, it follows that multiple evolutionary routes may lead from low-moderate affinity naive BCRs to high affinity antibodies against the same antigen surface region. This is supported by statistical arguments showing the implausability of a purely "random repertoire" for an efficient immune response (Arora and Arnaout 2020; Rees 2020). Epitope immunodominance could be rationalised *via* the existence of a more 'public' set of backbone structures in the BCR repertoire and the concept that BCRs with similar topologies and sufficient chemical complementarity engage the same epitope (Raybould et al. 2021b; Rees 2020).

Structural comparisons offer a way to analyse antibody data over and above clonotype-based approaches. Analysis of epitope regions using solved structures and competition assays of SARS-CoV-2—binding antibodies has already revealed discrete antibody binding sites (Dejnirattisai et al. 2021; Yin et al. 2021). However, such assessments are very biased towards RBD-binding, neutralising antibodies, whose therapeutic potential renders them worth the expense and effort of structure determination. The vast majority of datapoints in CoV-AbDab do not have solved structures and must instead be structurally modeled.

In this chapter, we describe our novel computational method for epitope profiling based on structural modelling and clustering, SPACE (Structural Profiling of Antibodies to Cluster by Epitope). We began by structurally analysing CoV-AbDab to enhance our functional understanding of coronavirus-binding antibodies. We analysed X-ray crystal structures of antibodies and nanobodies bound to SARS-CoV-2 antigens, showing both that structure is conserved more often than clonality across same-epitope binders, and that paratope profiles typically involve multiple regions of the antibody across both chains. This provided direct evidence that relatively sequence dissimilar coronavirus-binding antibodies with high variable domain (Fv) structural similarity are able to exhibit functional commonality. We then modelled and structurally clustered the thousands of antibody Fv sequences in CoV-AbDab

and show that 92% of multiple-occupancy structural clusters bin together antibodies that bind to consistent coronavirus antigens/domains. We also showed that, in accordance with our analysis of the SARS-CoV-2 X-ray co-crystal structures, the antibodies within these structural clusters frequently transcend clonal lineages. This not only demonstrates that our computational structural analysis pipeline provides orthogonal information to clonotyping to improve antibody functional profiling, but also that antibody immune responses to SARS-CoV-2 are likely to be even more convergent than currently understood. Given recent improvements in protein modelling, we re-performed our analysis using an improved antibody modelling tool, leveraging deep learning rather than homology modelling, and obtained similar results with 94% accuracy of epitope binning. We also applied our method to a set of malaria antibodies where we found 93% epitope consistency, including across sequence-dissimilar antibodies. Our results show that applying structural analytics to large class-specific antibody databases will enable high confidence structure-function relationships to be drawn, yielding new opportunities to identify functional convergence hitherto missed by sequence-only analysis.

4.2 Material and methods

4.2.1 Data

4.2.1.1 Coronavirus antibody database (CoV-AbDab)

The Coronavirus Antibody Database (CoV-AbDab) contains available information on coronavirus-binding antibodies and nanobodies identified in patents and the academic literature, including (where available) details on sequence, structure, epitope, binding capabilities, neutralising capabilities and origin (Raybould et al. 2021a).

Two CoV-AbDab datasets of were analysed: one dated 11th March 2021 and one dated 26th July 2022. The framework and 6 FREAD (Choi and Deane 2010; Choi and Deane 2011) CDR loop databases, which were used during structural modeling to find suitable homologous templates for each antibody region, were timestamped to contain the quality-filtered contents of SAbDab (Dunbar et al. 2014) on 11$^{\text{th}}$ March 2021. Quality filtering restricted templates to those solved by X-ray crystallography, with a resolution $\leq$ 2.5 Å and a B-factor <80. These framework and FREAD databases were used for the analysis of both March 2021 and July 2022 CoV-AbDab datasets.

4.2.1.2 Solved co-crystal structures

Sixty solved X-ray co-crystal structures of antibodies and nanobodies bound to SARS-CoV-2 were downloaded from SAbDab (11th March 2021). All antibodies were aligned based on the coordinates of the cognate SARS-CoV-2-RBD chain using PyMOL functions. Paratope residues were defined as any antibody residues with a heavy atom within 4.5 Å of an antigen heavy atom. We refer to antibodies and nanobodies by name as referenced in CoV-AbDab, using the nomenclature as set out in the literature from which the antibody was sourced (see Table A1 and Table A2 for the names and corresponding PDB codes of antibodies and nanobodies).

4.2.1.3 Malaria antibodies

257 paired Fv sequences of antibodies which bound to a specific malaria antigen were provided by Dr Kirsty McHugh, Dr Francesca Donnellan and Jordan Barrett from Professor Simon Draper's lab in the Department of Biochemistry, University of Oxford. Epitope community information was available for 240 of the antibodies, identified through competition epitope binning assays (epitope communities were defined as groups of antibodies which compete with one another). Ten shared epitope groups were identified, in addition to 21 antibodies binding independent sites. Four of the antibodies had published solved structures available.

4.2.2 Methods

4.2.2.1 Numbering scheme and region definitions

IMGT numbering (Lefranc et al. 2003) and CDR region definitions were used throughout the analysis (Chapter 1, Section 1.2.4.3). ABodyBuilder uses North CDR definitions in template selection. The CDR3 region is defined as the same residues by both North and IMGT definitions (IMGT residues 105 - 117) in both the heavy and light chains, meaning clonotype definitions are consistent regardless of region definition (see Methods - Clonotyping).

4.2.2.2 ABodyBuilder

At the time of method development, ABodyBuilder (Leem et al. 2016) was the best available antibody modelling tool. This method is described in full in Chapter 1, Section 1.3.2.1. Full variable domain (Fv) sequences from the three datasets - 11th March 2021 CoV-AbDab, 26th July 2022 CoV-AbDab and malaria - were submitted to the ABodyBuilder antibody modelling tool with default Environment Substitution

Score cutoffs. In the first instance, ABodyBuilder seeks to model antibody CDR regions entirely by homology; *i.e.* to use the FREAD software (Choi and Deane 2010; Choi and Deane 2011) to identify a CDR structural 'template' likely to be adopted by each of the submitted antibody's CDR sequences, considering backbone dihedral angle compatibility and loop graftability onto the framework template. If no suitable structural template can be found for a CDR sequence, *ab initio* or hybrid homology/*ab initio* approaches must be used to predict the loop structure, adding uncertainty to model quality (Marks et al. 2017). To ensure high model quality, only the models for which ABodyBuilder used FREAD to homology model all six CDR loops were carried forward for structural clustering. For further information on ABodyBuilder and antibody modelling tools, see Chapter 1, Section 1.3.2.1).

4.2.2.3 Structural clustering of homology-template models

We developed the Structural Profiling of Antibodies to Cluster by Epitope (SPACE) algorithm to structurally cluster antibody models.

Antibodies were first split by their combination of six CDR lengths. For each unique CDR length combination, the first antibody in the list was selected as a cluster centre and every subsequent antibody is fed into the following equation:

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

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 root-mean square deviation (RMSD) between the template used to model CDRX in antibody 1 and antibody 2 respectively. If this formula equates to a clustering threshold of ≤ 0.75 Å, the antibody is clustered with the first cluster centre, otherwise it is held out for the next round of clustering. Once all the antibodies have been considered relative to the first cluster centre, the algorithm progresses in a greedy fashion to select the next unclustered antibody as the second cluster centre. Greedy clustering was selected due to its simplicity, good performance, and speed, ensuring scalability across larger datasets. The result is a set of structural cluster centres and associated antibody Fv sequences, where each structural cluster only contains antibodies with 6 identical CDR lengths. This algorithm is adapted from the final step of Repertoire Structural Profiling (Raybould et al. 2021b).

4.2.2.4 ABodyBuilder-ML

Sequences from the 26th July 2022 CoV-AbDab and malaria datasets were modelled using the updated ABodyBuilder tool, ABodyBuilder-ML, featuring ABlooper (Abanades et al. 2022a). ABlooper is an equivariant deep-learning based tool for CDR loop structure prediction. This loop prediction replaces the step of the previous ABodyBuilder method where structural templates were used to homology model CDR loops with FREAD and Sphinx (See Chapter 1, Section 1.3.2.2).

4.2.2.5 Structural clustering of non-homology-template models

ABodyBuilder-ML models were clustered using the CDR structuring clustering tool (Brennan Abanades Kenyon, personal correspondence). Antibody models are first clustered by CDR length, then by structural similarity using pairwise comparisons and a greedy clustering algorithm, as with the original SPACE clustering described above. Structural similarity is assessed by RMSD between ABlooper modelled CDR loops, rather than the templates used to homology model CDRs (as no CDR templates are used with ABlooper). Clustering thresholds of ≤ 0.75 Å and ≤ 0.50 Å were utilised.

4.2.2.6 'Domain-consistent' structural cluster definitions

Multiple-occupancy structural clusters were each classified as 'domain-consistent' or 'domain-inconsistent' based on the CoV-AbDab binding metadata of their mapped antibody sequences. For instance, the following examples of structural clusters would each be considered as 'domain-consistent':

1. A structural cluster that only contains antibodies characterised as binding to the same antigen and domain (e.g. all shown to bind the SARS-CoV-2 RBD).

2. A structural cluster that only contains antibodies characterised as binding to internally-consistent domains (e.g. some antibodies labeled as spike N-terminal domain (NTD) binders and others labeled as S1 non-RBD binders; where the NTD is a subdomain of S1 non-RBD).

3. A structural cluster that contains some antibodies that are characterised as binding to the same domain, and others that bind to the same antigen without domain-level resolution (e.g. 4 antibodies shown to bind the RBD, and 2 antibodies

shown to bind to the full-length spike protein).

4. A structural cluster that only contains antibodies characterised to bind to the same antigen, but no antibody has domain-level resolution (e.g. 5 antibodies all shown to bind to the full-length spike protein, but none are localised to a particular domain).

The following structural clusters would both be considered 'domain-inconsistent':

1. A structural cluster that contains antibodies shown to bind to different antigens.

2. A structural cluster that contains antibodies that bind to the same antigen, but to inconsistent domains (e.g. 3 antibodies that have been shown to bind the spike RBD and 1 shown to bind the spike S2 domain).

4.2.2.7 Clonotyping

Clonotyping, clustering by clonotype, was performed using an in-house script. Our lenient VH-clonotyping protocol groups antibody sequences with matching IGHV genes, same length CDRH3, and $\geq 80\%$ CDRH3 sequence identity, while our lenient Fv-clonotyping protocol additionally requires cluster members to have a matching IG[K/L]V gene, same length CDRL3, and $\geq 80\%$ CDRL3 sequence identity. These are lenient clonotyping threshold conditions by community standards (López-Santibáñez-Jácome et al. 2019), as the CDR3 sequence identity threshold is set to its typical lower bound and there is no requirement for cluster members to map to the same J gene. We chose to use this definition in our analysis to ensure that antibodies from different lenient clonotypes would not be grouped together by standard clonotyping approaches.

4.2.2.8 Paratyping

Paratyping (Richardson et al. 2021), clustering by predicted paratope similarity, was performed using Paragraph (Chinery et al. 2023) (see Chapter 1, Section 3.2.2.2). Using modelled antibody structures, Paragraph predicts the probability of each residue (within CDR residues +2) being in the paratope, using graph neural networks. The standard paratope identity threshold of 0.734 was used. Antibodies were clustered if paratopes had $> 75\%$ similar identity.

4.2.2.9 Arpeggio

Arpeggio (Jubb et al. 2017) was used to analyse the solved crystal structures of SARS-CoV-2-binding antibodies. Epitope residues were defined as those within 4.5 Å of the antibody in the crystal structure. Analysis of the paratope-epitope interactions between antibodies and the SARS-CoV-2 RBD in the cocrystal structures was conducted, including overlap of epitope residues between clustered antibodies.

4.3 Results

4.3.1 Sequences and structures in CoV-Ab-Dab

The growth of coronavirus-binding antibody and nanobody data in CoV-AbDab since its public release on 7th May 2020 has been substantial. The availability of sequence data rose much more rapidly than structural data at the start of the pandemic, stabilising at a level roughly an order of magnitude higher. The availability of solved structures stood at 449 solved antibody and nanobody structures as of 26th July 2022 (across 9,579 entries with full sequence information), compared to 132 solved structures in March 2021 (across 2180 entries with full sequence information).

4.3.2 Analysis of SARS-CoV-2—antibody structural complexes

Experimentally-solved structures allow us to analyse the diversity of antibody geometries and paratopes that engage coronavirus antigens (Barnes et al. 2020; Dejnirattisai et al. 2021; Yin et al. 2021). At the time of the study (11st March 2021), CoV-AbDab (Raybould et al. 2021a) referenced 2,304 antibodies and 420 nanobodies able to bind to a coronavirus, with 132 having at least one solved structure. 111 of these solved structures were binders to SARS-CoV-2, of which 91 were solved in complex with the cognate antigen.

A total of 48 antibodies and 12 nanobodies had at least one published solved X-ray crystal structure in complex with SARS-CoV-2, all binding to the spike RBD (Bertoglio et al. 2021; Clark et al. 2021; Dejnirattisai et al. 2021; Du et al. 2020; Guo et al. 2021; Huo et al. 2020; Hurlburt et al. 2020; Jones et al. 2021; Ju et al. 2020; Kim et al. 2021; Koenig et al. 2021; Kreye et al. 2020; Liu et al. 2020a; Liu et al. 2020b; Piccoli et al. 2020; Rapp et al. 2021; Robbiani et al. 2020; Rogers et al. 2020; Rujas et al. 2021; Shi et al. 2020; Thomson et al. 2021; Wrapp et al. 2020; Wu et al. 2020a; Wu et al. 2020b; Wu et al. 2020c; Xiang et al. 2020; Yao et al. 2021; Zhou et al. 2020)

(see Table A1 and Table A2 for names and PDB codes). A further 31 antibodies and nanobodies were solely structurally characterised by cryo-EM (Raybould et al. 2021a). In our analysis we focused on the 60 crystal structures in order to determine more precise antibody binding site topologies and paratope profiles.

4.3.2.1 Epitope binning

Inspecting the 48 antibodies solved by X-ray crystallography in complex with the RBD, 46 appear to fall cleanly into the binding regions previously defined by Dejnirattisai et al. 2021 (only approximate as the original clustering was performed *via* competition assays (see Chapter 1, Section 1.4)). These binding regions are based on representation of the RBD as a human torso, with epitope regions at the 'neck', left and right 'shoulders', and left and right 'flanks'. The 2/48 remaining antibodies spanned the left and right 'shoulder' regions (Table A1). Interestingly, most of the nanobodies could also be assigned to these predefined regions (9/12, see Table A2).

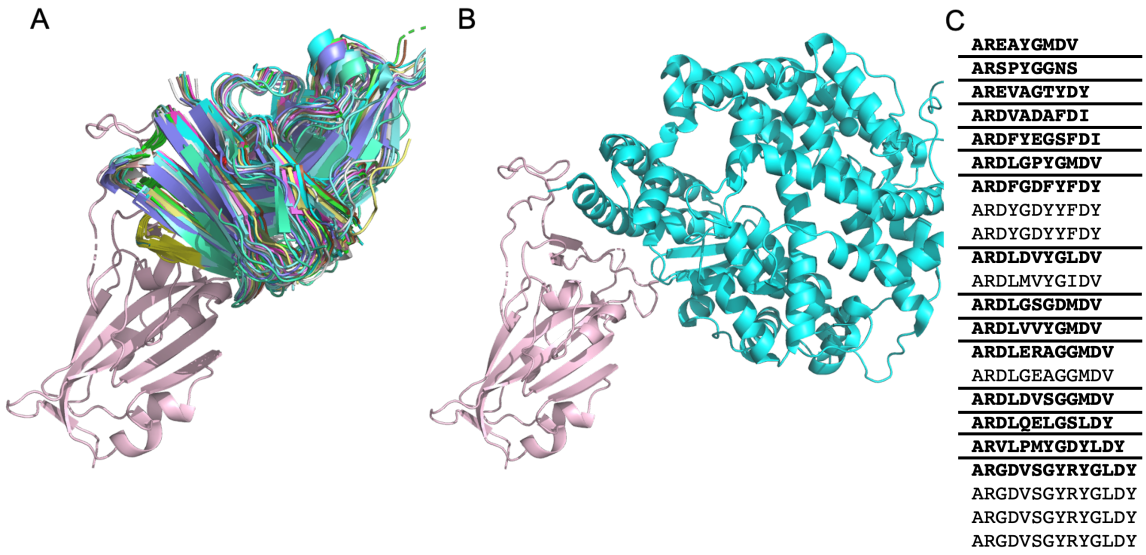


Figure 4.1: a) A cartoon representation of the 22 antibodies in the RBD 'neck' cluster binding to the SARS-CoV-2-RBD (salmon) [PDB code 6xc2] at a similar site to ACE-2. See Table A1 for PDB codes of the 22 antibodies. b) A cartoon representation of ACE-2 (cyan) binding to SARS-CoV-2 RBD (salmon) [PDB code 6vw1]. C) The CDRH3 sequences represented across the 22 RBD 'neck'-binding antibodies. Lenient VH-clonotypes are separated with solid lines, with the cluster representative highlighted in bold font.

Structural alignment of the RBD of all complexes reveals that the 22 antibodies

that bind to the 'neck' (as termed by Dejnirattisai et al. 2021) have high structural conservation (Figure 4.1A). These antibodies all compete for the ACE-2 receptor binding site (Figure 4.1B). Dejnirattisai *et al.* identified 13 IGHV3-53/IGHV3-66-derived antibodies engaging this binding site. Even from relatively early in the pandemic, it was clear that a disproportionate number of antibodies reported as able to block ACE-2 binding exploited the IGHV3-53/IGHV3-66 genes (Barnes et al. 2020; Dejnirattisai et al. 2021; Raybould et al. 2021a); IGHV3-53/IGHV3-66 antibodies appeared significantly more often as binders of this region than would be expected by their abundance in healthy antibody repertoires. Banach *et al.* realised that many of the coronavirus-binding antibodies deriving from the IGHV3-53/IGHV3-66 genes possess a conserved set of structural motifs that enable complementarity to a SARS-CoV-2 RBD epitope (Banach et al. 2021). Our analysis reiterates the importance of these V gene origins in engaging this highly conserved binding site: 19/22 (86%) of the antibodies align closest to the IGHV3-53 gene, while the remaining 3 align closest to IGHV3-66.

Despite the highly similar V genes, these 22 tightly structurally-clustered antibodies represent 15 different 'lenient VH-only clonotypes' (Figure 4.1C) - where clustered antibodies have the same CDRH3 length, 80% CDRH3 sequence identity, and identical heavy V gene (see Methods). This corresponds to 18 'lenient Fv-clonotypes' when the light chain is also considered (clustered antibodies must also have the same CDRL3 length, 80% CDRL3 sequence identity, and an identical light V gene, see Methods). The analysis of the co-complex structures of these 22 antibodies suggests highly similar functionality, which cannot be wholly identified through clonotyping. Even using lenient clonotype definitions, the antibodies would not be grouped together, so their similar binding mode and functional similarity would be missed.

Dejnirattisai and colleagues also highlighted an antibody binding cluster located away from the ACE-2 binding site, the 'left flank' (Dejnirattisai et al. 2021). This binding region contains four antibodies, EY6A, CR3022, S304 and COVA1-16, (see Methods and Table A1 for antibody naming conventions). As shown in Figure 4.2A, these antibodies appear to bind to slightly different areas. In our analysis, we identify one antibody (MW06) and four nanobodies (VHH_U, VHH_V, VHH_W and VHH-72) all able bind to the 'left flank' binding region (Figure 4.2A). We highlight the antibody pair of EY6A and S304, which are structurally similar and adopt a common binding mode to the same RBD epitope (Yin et al. 2021) but share only 43% CDRH3

sequence identity (Figure 4.2B) so could not have been identified as binding to the same site by sequence data alone.

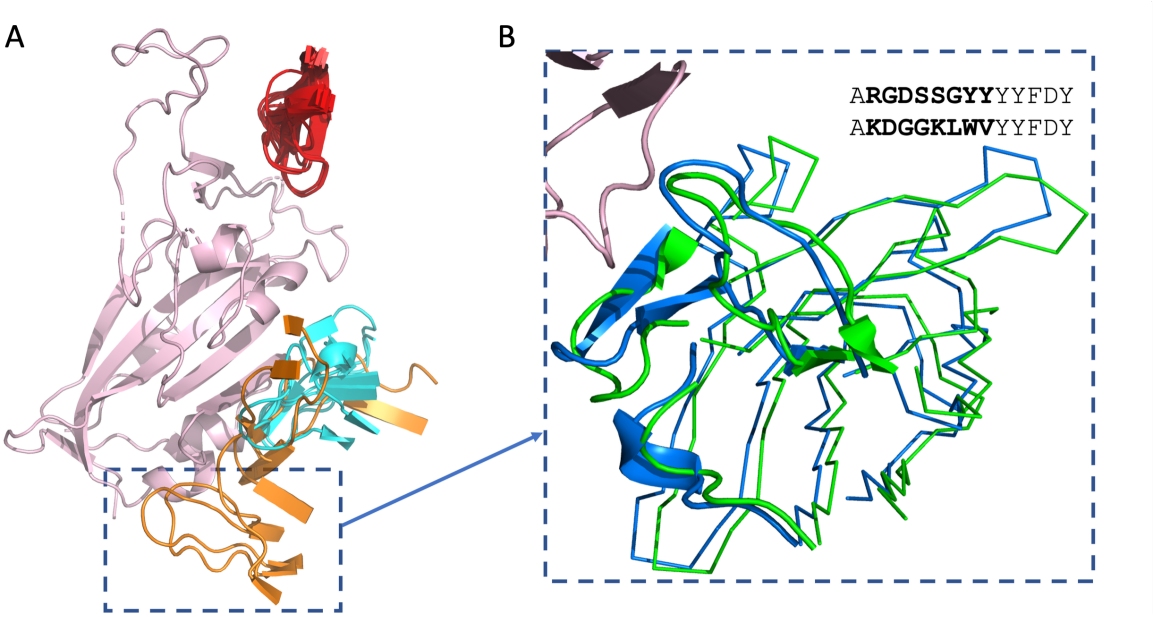


Figure 4.2: a) The CDRH3 loops of nanobodies and antibodies of two binding regions, the 'neck' and 'left flank', binding to SARS-CoV-2 RBD (salmon). The 'neck' cluster includes 22 antibodies (red). The 'left flank' region includes five antibodies (four antibodies identified by (Dejnirattisai et al. 2021) are in orange, antibody MW06 is in magenta) and four nanobodies (blue). MW06 and the nanobodies were not included in the analysis by (Dejnirattisai et al. 2021). See Table A1 and Table A2 for antibody and nanobody PDB codes respectively. b) A ribbon representation of sequence dissimilar antibodies S304 (blue) and EY6A (green) binding to the SARS-CoV-2-RBD (salmon). The CDRH1, CDRH2 and CDRH3 loops are illustrated in cartoon. The CDRH3 sequences of the two antibodies are shown, with dissimilar residues indicated in bold.

Even across the relatively small number of solved SARS-CoV-2—antibody structures, we can see numerous examples of functionally and structurally similar antibodies that would not be grouped by sequence clustering alone. Grouping antibodies into sets that have similar structures therefore represents an orthogonal and promising approach by which to highlight the potential functional commonalities of sequence-dissimilar antibodies.

4.3.2.2 Paratope analysis

We analysed the binding interfaces across the set of 48 antibodies, to investigate whether structure needs to be conserved across the whole Fv, or whether conserva-

tion over particular binding regions is typically sufficient.

An average of ∼67% of paratope residues were found to lie on the heavy chain while ∼33% resided on the light chain. The percentage of paratope residues donated by the most hypervariable region, the CDRH3 loop, varied from just 9% up to 58%. As expected, structures with longer CDRH3 regions tend to exhibit more CDRH3-dominated binding; the paratopes of the nine antibodies with CDRH3 length ≥ 19 on average comprised 41% CDRH3 residues. For the 22 highly structurally-conserved antibodies in the RBD neck epitope region, an average of 23% of the paratope residues originate from CDRH3 (range 8-34%).

These RBD neck-binding antibodies exhibit high levels of paratope residue conservation across the CDRH1 and CDRH2 (Figure 4.3), which is also seen at the sequence similarity level. This explains the predominance of the IGHV3-53/IGHV3-66 genes, as the residues and topologies pre-encoded by this germline play a key role in neck epitope complementarity. Paratope conservation is considerably less consistent across the CDRH3 region, accounting for the 18 unique CDRH3 sequences seen across the 22 antibodies. For such epitopes, a clonotyping framework (which conditions on high CDRH3 sequence identity being a pre-requisite for same-epitope binding), will clearly fail to capture the functional similarity of the spectrum of cognate antibodies.

For those antibody paratopes that use CDRH3 sparsely, the paratope is mostly drawn from the CDRH1, CDRH2, CDRL1 and CDRL3 residues, with occasional amino acids provided by the framework regions (FWRs). In particular, a serine residue at position 83 and/or glycine at position 84 of the light chain (part of FWRL3) are disproportionately involved in binding. These residues were found in 18/48 (38%) of crystallised-antibody paratopes, of which 15 are from the 'neck' structural cluster. Using Arpeggio (Jubb et al. 2017) to identify the type of binding interactions these FWRL3 residues were involved in, 9 paratopes contained hydrogen bonds between L83 and the RBD, specifically to RBD residue 498 in 8/9 cases. 'Neck' antibodies were found to interact with more epitope residues than other groups (average of 31.7 epitope residues per antibody, compared to only 20.1 residues in the 'left shoulder' group) (Table 4.1). Variation in the number of epitope residues was seen amongst antibodies from the same epitope region; within the 'neck' group the number of epitope residues bound ranged from 17 to 38. Only 5 epitope residues were shared between all members of the group, though 23 residues were found in most (90%) of group

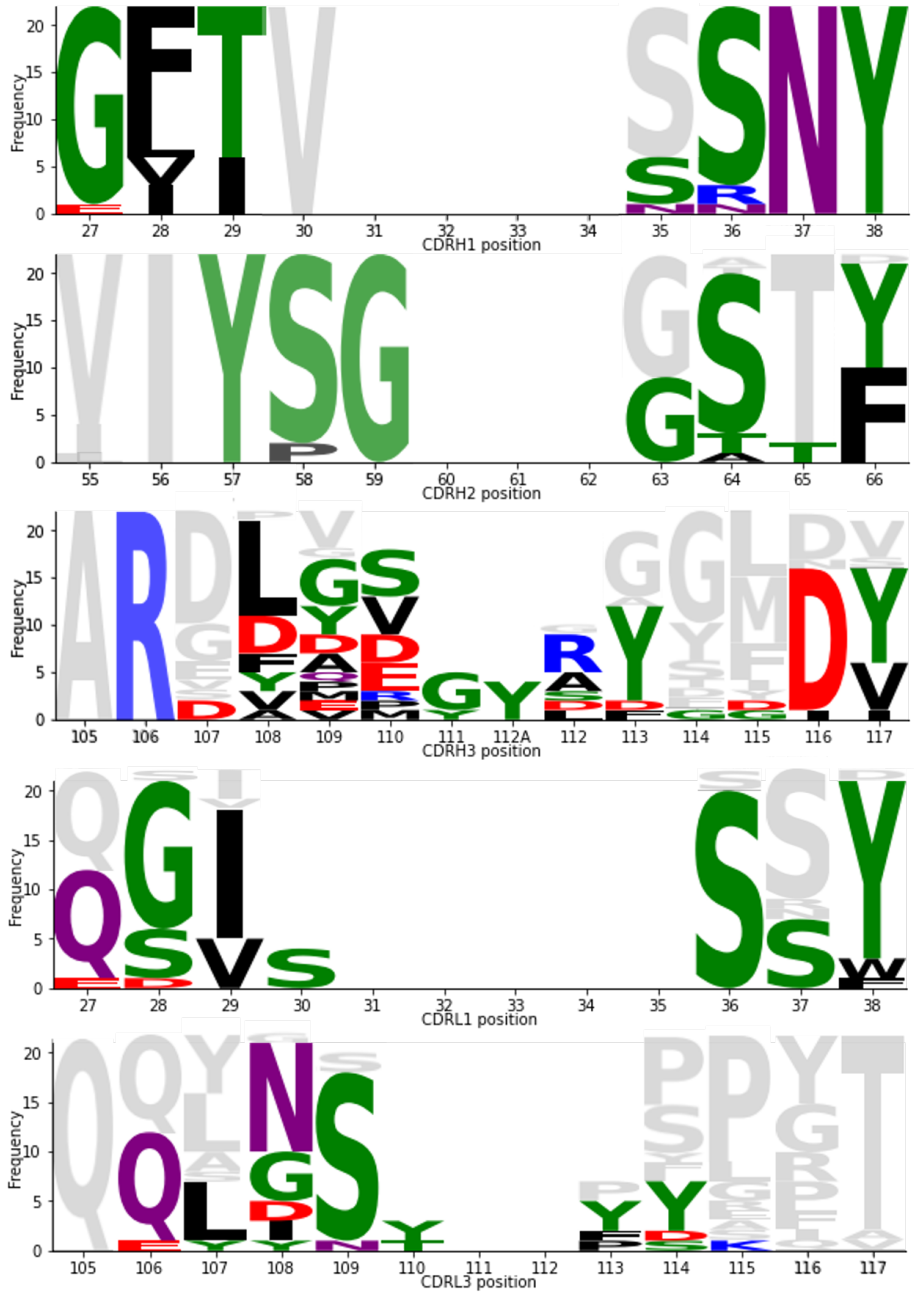


Figure 4.3: CDR paratope conservation and divergence across the 22 antibodies in the 'neck' cluster. Frequency indicates the number of times an amino acid was seen at each IMGT-defined CDR position (Lefranc et al. 2003). Paratope residues are coloured by side chain chemistry (black = hydrophobic, green = polar neutral, red = acidic, blue = basic, purple = amide). Grey indicates amino acids present at the positions but not in the paratope (within 4.5 Å of the SARS-CoV-2 antigen). CDRL2 is not shown as it was found not to contain paratope residues. Produced using Logomaker (Tareen and Kinney 2020).

Epitope group	Size	Average no. of antibody epitope residues (min - max)	Total epitope residues bound by group
Neck	23	31.7 (17 - 38)	47
Left Shoulder	10	20.1 (13-26)	67
Right Shoulder	6	21.2 (13 - 27)	54
Left Flank	5	24.6 (22-27)	58
Right Flank	2	22.5 (21 - 24)	39
Shoulder	2	23.0 (21 - 25)	29

Table 4.1: Analysis of structural epitope residues of antibodies by epitope group. The average number of epitope residues bound by an antibody in each group is shown, with the minimum and maximum number indicated in brackets, and the total number of antigen epitope residues bound by antibodies within each group. Epitope residues were defined as those within 4.5 of the antibody in the crystal structure. Analysis was conducted using Arpeggio (Jubb et al. 2017).

members and 32 residues were found in over half of the antibodies (Figure 4.4). The most variation was seen in the 'left shoulder' group, where no epitope residues were bound by all 10 antibodies and 67 unique epitope RBD residues were bound in total by all 10 antibodies.

Overall, current crystallographic knowledge of antibody-RBD co-crystal structures suggests a number of distinct epitopes and a larger patch of overlapping epitopes near the ACE-2 binding site. This dataset is biased towards neutralising antibodies; only one antibody (CR3022) in this set is non-neutralising towards SARS-CoV-2. As reported by Yuan and colleagues, consensus RBD mutations E484 and K417 impact the binding and neutralisation of IGHV3-53/3-66 and IGHV1-2 encoded antibodies, and do not impact CR3022 and S309 which bind elsewhere (Yuan et al. 2021). Many crystalised antibodies bind at epitopes containing RBD mutation sites E484, K417 and N501; 30/34 (88.3%) of crystallised antibodies had interactions with these residues.

Overall, we conclude that structural conservation is important across the entire Fv, to ensure that the paratope residues, which in most coronavirus-binding antibodies are spread evenly throughout the CDRs, are held in equivalent topological positions. We subsequently performed a comprehensive structure-based analysis on CoV-AbDab,

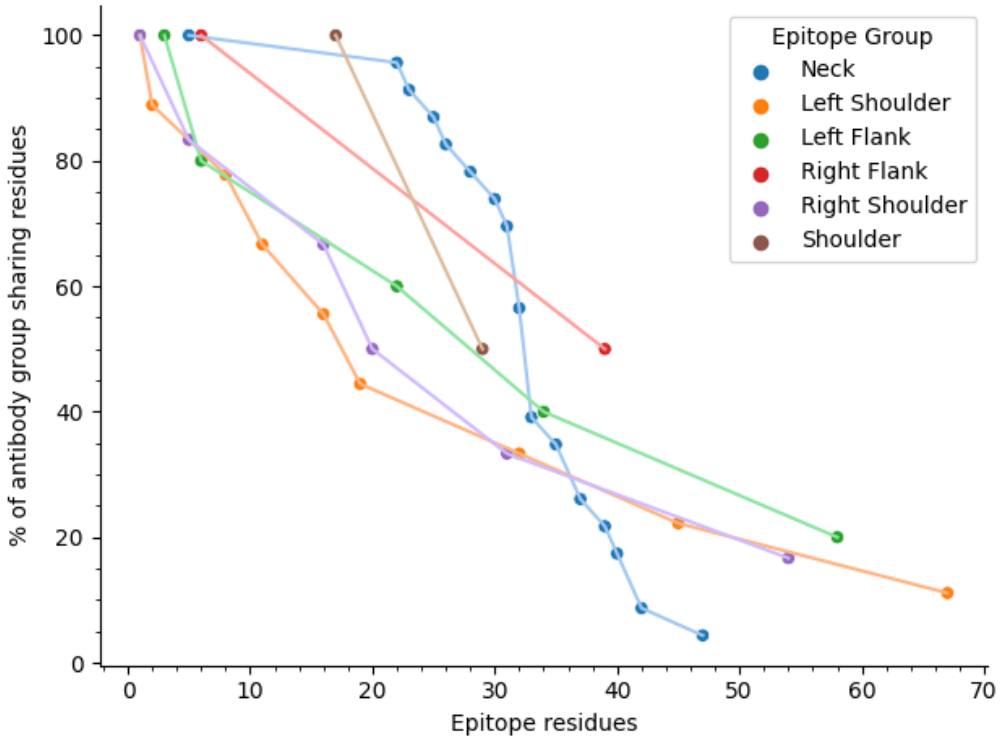


Figure 4.4: Analysis of structural epitope residue overlap between antibodies within an epitope group. The total number of epitope residues in a group is on the x axis, with the y axis showing the percentage of antibodies which bind to that number of the total epitope residues. For example, in the neck epitope group (blue): 5 residues are shared by all 23 antibodies (100%), 32 residues are each bound by 13 antibodies (57%) and 47 residues are in the epitopes of just a single antibody in the group (4%).

where 95% of SARS-CoV-2 binding antibodies had no solved structure and 27% completely lacked binding-domain annotation.

4.3.3 Structural convergence across CoV-AbDab antibodies

As of 11th March 2021, just ∼5% (113/2,304) of the antibodies in CoV-AbDab had at least one solved X-ray or cryo-EM structure, while ∼90% (2,063/2,304) of the antibodies had full Fv amino acid sequences. We used high-throughput homology modelling approaches to approximate and analyse the geometries of this much broader set of neutralising and non-neutralising antibodies able to bind to multiple coronaviruses, antigens, and domains.

We used ABodyBuilder (Leem et al. 2016) to homology model the 2,063 antibody entries in CoV-AbDab with full Fv sequences. This resulted in a total of 1,500 models

in which every loop was entirely FREAD-modellable without any need for *ab initio* loop modelling or backbone adjustment (see Methods); we focus on this subset of models as we have the highest confidence in their accuracy (Kovaltsuk et al. 2020; Leem et al. 2016; Raybould et al. 2021b). This represents 72.7% modellability across the set of Fv sequences, a remarkably high percentage relative to recent studies on both healthy and disease-related natural antibody datasets (Kovaltsuk et al. 2020; Raybould et al. 2021b). Typically, only $\sim 40\%$ of randomly-sampled human CDRH3s can be homology modelled by FREAD. The increase in modellability is likely to be related to the scientific effort that has gone into solving a large number of SARS-CoV-2-binding antibody structures within the first year of the pandemic. It also hints at a high degree of underlying structural convergence across the reported coronavirus-binding antibodies in CoV-AbDab.

After modelling, we performed structural clustering. Briefly, Fvs with 6 modellable CDRs were first clustered according to their combination of six CDR lengths, and then further structurally grouped by a greedy clustering algorithm using the pairwise structural RMSD between the selected FREAD template for each CDR region (see Methods for a full description). The result is a set of predicted 'structural clusters', each adopted by at least one Fv sequence. We term this algorithm Structural Profiling of Antibodies to Cluster by Epitope (SPACE).

The 1,500 homology modelled Fv regions fell into 1,159 structural clusters, of which 200 were adopted by more than one Fv sequence ('multiple-occupancy' structural clusters). In total, 541/1500 (36.1%) of the antibodies belonged to a multiple-occupancy structural cluster.

When applied to an entire antibody repertoire, the number of antibodies with similar structures but different functionalities is likely to be significant (Raybould et al. 2021b). However, on 'cleaner' datasets such as CoV-AbDab, where every antibody has been shown to bind a coronavirus antigen (and often a particular domain) *in vitro*, antibodies with similar Fv structures should have a high chance of binding to the same surface region and therefore being complementary to the same epitope, as demonstrated in our analysis of solved structures.

Nonetheless, inaccuracy in CDR structure prediction and/or the lack of consideration for paratope residues can lead to misleading clusters of antibodies that have

markedly different functionalities. Therefore, to assess (in the context of this large disease-specific antibody database) how predictive belonging to the same structural cluster is of engaging the same epitope, we estimated a pseudo-'true positive'/'false positive' ratio based on the consistency of the CoV-AbDab epitope metadata across antibodies grouped into the same structural cluster. 'Domain-consistent' structural clusters were classed as those only containing antibodies reported as binding to internally-consistent antigen domains (see Methods).

A total of 184/200 (92%) of our multiple-occupancy structural clusters were domain-consistent, indicating that structurally clustering with another member of CoV-AbDab is likely to be highly predictive of function. Amongst the 16 structural clusters considered to be 'false positives', the domain-inconsistent antibody within four structural clusters (SC2, SC84, SC96, SC136) bore significant similarities to at least one other antibody in the cluster, suggesting that some experimentally-deduced epitope labels may be inaccurate. The original papers detailing the binding of the antibodies support this possibility. For example, we classify SC2 as domain-inconsistent, as it contains 8 antibodies that bind to spike outside the RBD (Li et al. 2021a; Zost et al. 2020), and one (COVA2-32, Brouwer et al. 2020) labelled as binding to the RBD. However, COVA2-32 only marginally met the area under curve threshold to be classified as an RBD binder (see Figure 4 of Brouwer et al. 2020).

The remaining twelve 'false positive' structural clusters are likely to result from a combination of inaccurate structural modelling and/or the fact that bearing the same binding site structure does not guarantee functional commonality. We focus our analysis on the 92% of domain-consistent structural clusters.

4.3.3.1 Epitope binning

Some members of a structural cluster can have a lower resolution of functional characterisation than others, for example, they might be known to bind to the spike protein, rather than specifically the RBD. In these cases, functional properties of the less well-characterised antibodies can be inferred from other antibodies predicted to adopt the same structure.

Thirty-one antibodies experimentally shown only to bind to the whole spike protein, or to bind the spike protein but not the RBD, can be localised to a more precise epitope using our structural clusters. For example, three of the antibodies assigned

to SC11 that were shown to bind the full-length spike protein, but not a soluble RBD protein (Zost et al. 2020), can be inferred to bind to the S2 domain in the same way as cluster members DH1147 and DH1149 (Li et al. 2021a). Similarly, CC12.24, previously shown only to bind to the whole SARS-CoV-2 spike protein (Rogers et al. 2020), can be localised to the same binding site in the RBD as C139 (Robbiani et al. 2020) and COVOX-45 (Dejnirattisai et al. 2021) (SC57).

A further 62 antibodies fall into 19 structural clusters for which no antibody has been resolved as binding to a particular domain. For some structures, the selected FREAD templates could offer an indication of epitope specificity: the PDB structure 6nb8 (Pinto et al. 2020b, April 9) is used to model all three light chain CDRs in at least one antibody assigned to SC6, SC15, and SC71, while 7ben (Dejnirattisai et al. 2021) is used to model the CDRH3 in all four antibodies assigned to SC16. The antibody in both of these PDB structures engages the SARS-CoV-2 RBD, which could imply that the cluster groups RBD-complementary antibodies. Even ignoring these indications, just 19 binding characterisation experiments could lend functional annotations to over three-times the number of CoV-AbDab entries (62).

4.3.3.2 Evolutionarily-conserved epitope topologies across coronaviruses

Evolutionarily-conserved epitope topologies are implied by our structural clusters that contain antibodies able to bind multiple coronaviruses. As an extreme example, SC0 pools together a set of 13 IGHV1-69-derived antibodies, of which at least one which has been shown to engage each of SARS-CoV-1, SARS-CoV-2, HKU1, 229E, NL63, OC43, and MERS-CoV. More broadly, 69/184 (37.5%) of our structural clusters contain an antibody shown on the current levels of data to have cross-coronavirus binding potential. This number may be an underestimate, as several antibodies have only been tested against a single coronavirus strain *in vitro*.

These epitopes could represent sites of particular vulnerability across coronaviruses; antigen regions whose structure must be preserved for viral function. They therefore reflect the most promising regions against which to design pan-coronavirus neutralising antibodies, and could be exploited in epitope-focused vaccine design strategies (Correia et al. 2014) to achieve a more broadly neutralising response.

The epitope and corresponding paratope residues within these binding sites will differ between lineages, meaning that sequence-based clustering approaches would

struggle to spot their functional commonality. Structural clustering can see beyond paratope profiles to capture broader epitope topology conservation *via* the geometries of their cognate antibodies.

4.3.3.3 Epitopes targetable by multiple species

A recent study has shown that mouse and human antibodies use a different distribution of CDRH3 structures, which also varies by B-cell maturation stage (Kovaltsuk et al. 2020). This can be rationalised by their species-specific gene loci having different predetermined structural biases and the fact that negative selection occurs against different self-epitopes. Nevertheless, assuming that each CDR loop can only adopt a finite set of geometries imposed by the loop closure criterion, and that many of the same antigens would be considered pathogenic to both species, there ought to be some structural overlap between human and mouse antibodies and therefore the potential for some epitopes to be targetable by both species. These would be extremely hard to identify by sequence alone, as human and murine gene loci are highly sequence dissimilar.

We identified SC62, which groups human antibody Ab_511E7 (Han et al. 2021) alongside two murine antibodies DK4 & DK7 (patent CN111978395A), which have all been shown to bind the SARS-CoV-2 RBD - albeit only weakly in the case of Ab_511E7. Even more remarkably, SC174 pools antibodies from different species confirmed to bind to different coronaviruses; human antibody C131 (Robbiani et al. 2020) binds SARS-CoV-2 RBD and is grouped with murine antibody F26G18 (Gubbins et al. 2005) shown to bind the full-length SARS-CoV-1 spike protein. Should F26G18 be confirmed to engage the RBD, and the SC62 antibodies compete for the same epitope, this would show the ability of structural clustering to identify cross-coronavirus epitopes targetable by multiple organisms.

4.3.3.4 Structural clusters frequently span multiple clonal lineages

We analysed each of the 184 domain-consistent structural clusters to determine how often the antibodies clustered together belonged to multiple lenient Fv-clonotypes. A total of 88 (47.8%) contained at least one pair of antibodies from different lenient Fv-clonotypes and 73 (39.7%) of the structural clusters contain at least two lenient VH-only clonotypes. It is clear that antibodies with both heavy and light chains of differing clonality can frequently co-exist within our structural clusters.

Many structural clusters contain at least one pair of antibodies from the same clonotype; this is unsurprising since the 'near-identical sequence, similar function' assumption underpinning clonotyping experiments is often correct. However, the high frequency with which we group antibodies spanning several clonotypes into the same structural cluster recapitulates the findings of other papers (Kovaltsuk et al. 2017; Richardson et al. 2021; Wong et al. 2021), and our earlier analysis on solved structures, that clonotyping cannot group together all antibodies capable of same-epitope engagement.

In most cases where multiple clonotypes are found in the same structural cluster, it is due to significant differences in the CDRH3 sequence. However, some clusters such as SC134 (which pools COV2-2490 with H712061+K711727), contain many differences across the entirety of the sequence (26 differences across VH, 27 across VL) and align closest to different heavy V (IGHV3-7 vs. IGHV3-30) and light V (IGKV1-5 vs. IGKV1D-16) genes.

4.3.3.5 'Public' response antibodies

'Public' antibodies, those that are raised independently across multiple individuals against an immunodominant epitope, are of high interest to several fields of research, from vaccinology to drug discovery (Forgacs et al. 2021; Marks and Deane 2020; Setliff et al. 2018).

Several studies have already identified public SARS-CoV-2 response antibodies based on convergence towards particular clonotype lineages (Bullen et al. 2020; Kreye et al. 2020; Montague et al. 2021; Robbiani et al. 2020; Schultheiß et al. 2020; Sokal et al. 2021; Zhang et al. 2020) but none have yet considered the fact that antibodies from different lineages can exert similar functions. As demonstrated above, structural clustering enables us to group together clonally-distinct antibodies with a high chance of engaging the same epitope. We therefore examined our structural clusters to reveal functionally similar groups of antibodies from different genetic lineages that have been independently isolated across several different studies (*i.e.* 'public structures').

A striking example of a public structure that spans multiple clonotypes is SC3 (Figure 4.5). SC3 contains nine antibodies from five independent sources, spanning five lenient Fv-clonotypes. All antibodies align closest to the IGHV1-58/IGKV2-30

gene transcripts but have sequence-diverse CDRH3s all containing a common intra-loop disulfide bridge. The SC21 and SC24 clusters also map to these genes and contain a disulfide bridge; antibodies assigned to SC21 and SC3 have identical CDR lengths but are predicted to have different CDR structures, while SC24 is necessarily classed as a separate structural cluster to SC3 as its CDRL3 loop is of a different length (10 residues rather than 9). When aligned to the solved COVOX-253:RBD co-crystal structure, SC3, SC21, and SC24 appear to all be topologically complimentary to the same RBD epitope (Figure 4.5). The length-10, more protruding, CDRL3 loop of SC24 is accommodated by the small G485 residue on the RBD, while the CDRH3s across all three structural clusters protrude to a similar extent towards residues F456-N460 on the RBD. The antibodies mapped to clusters SC21 and SC24 (from an additional three independent sources) comprise an additional three lenient Fv-clonotypes, making a total of eight lenient Fv-clonotypes with potential same-epitope complementarity.

This set of similar structures has been observed across eight independent studies indicating that the corresponding epitope is immunodominant. Moreover, none of the antibodies directly engage the carboxylate group of residue E484 (5 Å for length-10 CDRL3s or 11 Å for length-9 CDRL3s, with too acute an angle for hydrogen bonding) nor the amide group of N501 ($\geq$12 Å for all antibody topologies). This should make them of particular interest as clones that might neutralise both wildtype SARS-CoV-2 and the E484K/N501Y-containing variants of concern.

Soon after we identified this broad structural cluster, a preprint was released by Schmitz *et al.* (Schmitz et al. 2021) showing that many IGHV1-58-encoded SARS-CoV-2 binding antibodies have highly similar residues at equivalent paratope positions (defined by the S2-E12 crystal structure (Tortorici et al. 2020)). On this occasion, the other CDRH3 sequence-diverse antibodies hypothesised to have similar function were shortlisted through inspection of their sequences; they all derive from IGHV1-58 germline and bear the -C$(X)_4$C- motif within their CDRH3 loop (X_4 representing four non-cysteine residues). We have shown that structural modelling and clustering supports the theory that these antibodies are functionally similar, and offers a systematic route to the identification of other sequence-diverse clusters of functionally-common antibodies that do not bear such clearly conserved motifs.

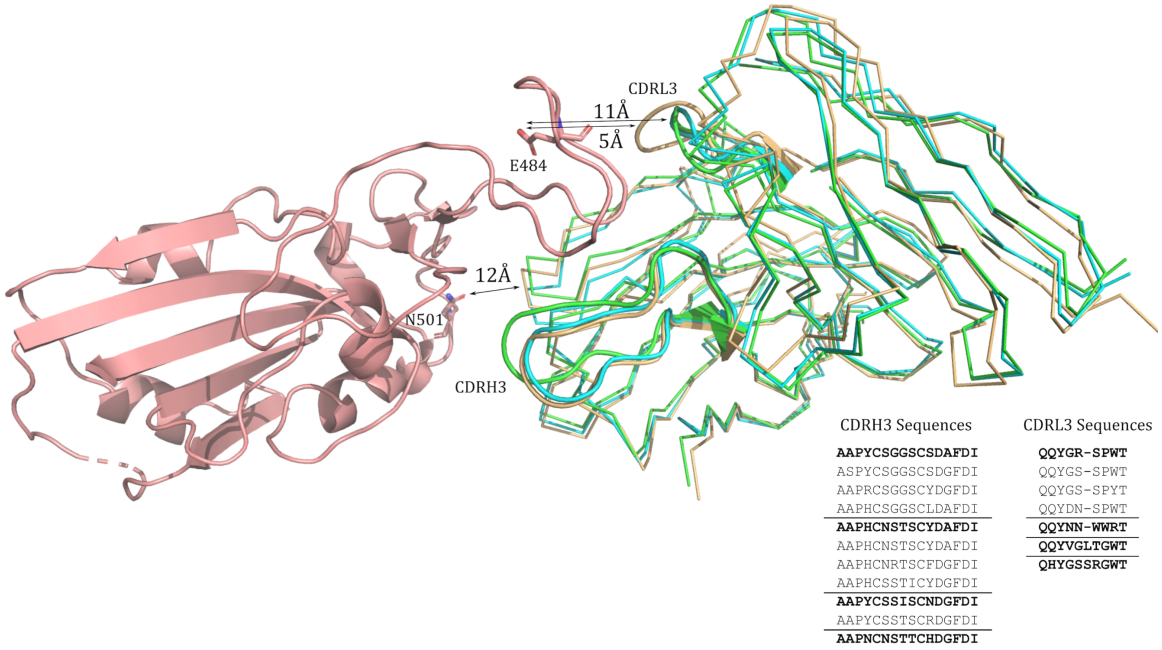


Figure 4.5: SARS-CoV-2 RBD-binding antibodies with similar predicted structure that span multiple clonotypes. A representative of structural cluster 3 (SC3, cyan), structural cluster 21 (SC21, orange), and structural cluster 24 (SC24, green) are aligned in the context of the COVOX-253:RBD co-crystal complex (RBD in salmon). The CDRH3 and CDRL3 structures are highlighted in cartoon representation. All 100% sequence non-redundant CDRH3 and CDRL3 sequences across the three structures are listed, grouped by lenient VH- or VL-clonotype, with the cluster representative in bold. The alignment shows the various CDRH3 and CDRL3 structures are likely to be topologically compatible with this RBD epitope. Residues E484 and N501, commonly mutated in SARS-CoV-2 variants of concern, are highlighted as sticks and coloured by default atom type. Closest heavy-atom distances between the functional group (carboyxlate/amide heteroatom) and the different structural classes of antibody are shown.

4.3.4 Updated CoV-AbDab analysis

4.3.4.1 Epitope binning of homology modelled structures

As of 26th July 2022, CoV-AbDab contained 8840 antibodies with full Fv sequences. As with the previous analysis of CoV-AbDab from 11th March 2021, we used ABodyBuilder (Leem et al. 2016) to homology model the sequences, resulting in 6639 models which passed our quality criteria (see Methods). This modellability is comparable to that achieved for the March 2021 dataset (75.1%, previously 72.7%). Using SPACE, we identified 879 'multiple-occupancy' structural clusters with 89.2% epitope consistency across clusters containing epitope information (Table 4.2), slightly lower than the 92% epitope consistency previously achieved for CoV-AbDab. 108 clusters did

not contain any specific epitope information.

	ABB models	ABB-ML models	
Clustering threshold	0.75	0.75	0.5
Antibodies successfully modelled	6639	8685	8685
Clusters with ≥ 2 members	879	1231	948
Epitope consistency	89.2%	87.5%	94.3%
Largest cluster size	55	205	52

Table 4.2: Structural clustering of CoV-AbDab antibodies (time-stamped 26th July 2022). Models were created with ABodyBuilder (ABB) models and ABodyBuilder-ML (ABB-ML). ABB models were clustered using RMSD across CDR templates used in homology modelling; ABB-ML models were clustered using RMSD across predicted CDR structures.

4.3.4.2 Epitope binning of deep learning modelled structures

Since our initial SPACE analysis of CoV-AbDab, there have been substantial advances in antibody modelling. We therefore also generated antibody models using the latest available tool, ABodyBuilder-ML (ABB-ML, see Methods) (Abanades et al. 2022a). ABodyBuilder-ML does not use templates for homology modelling of CDR loops, instead using deep learning.

Clustering the ABB-ML models with a 0.75 Å clustering threshold yielded 1231 multiple occupancy clusters, with 87.5% epitope consistency (117 clusters without epitope information, 141 epitope-inconsistent clusters) (Table 4.2). Analysis revealed one very large cluster of 205 antibodies, all RBD binders. This was 4x greater than the largest structural cluster when using the ABodyBuilder models (size 55). The second and third largest ABB-ML clusters were also very large, consisting of 148 members and 136 members. The 148 member-cluster was epitope-consistent with 139 RBD binders and 9 spike protein binders; the 136 member cluster was epitope-inconsistent with 6 RBD binders, 18 non-RBD binders, 39 S2 binders and 73 spike binders.

Subsequently, we ran our analysis using a narrower clustering threshold of 0.5 Å and achieved cluster sizes similar to those previously achieved with the ABB models (largest cluster size 52). With this approach, 948 multiple occupancy clusters were identified, with 94.4% epitope consistency (103 clusters with unknown epitopes, 48 epitope-inconsistent clusters). We continued to find examples of structural clustering of diverse sequences. For example, a structural cluster contained 11 antibodies from

V3-53/V3-66 germlines which form 8 lenient VH-clonotypes due to CDRH3 diversity (9 lenient VH+VL clonotypes). CDRH3 sequence identity varies substantially across the cluster, with pairwise CDRH3 identity ranging from 18-90% and several pairs sharing $< 28\%$ (Figure 4.6). The antibodies originate from 8 different sources, from SARS-CoV-2 recovered or vaccinated human subjects. Two of the 11 antibodies were present in the March 2021 dataset and were clustered together - the remaining 9 have been added to CoV-AbDab more recently.

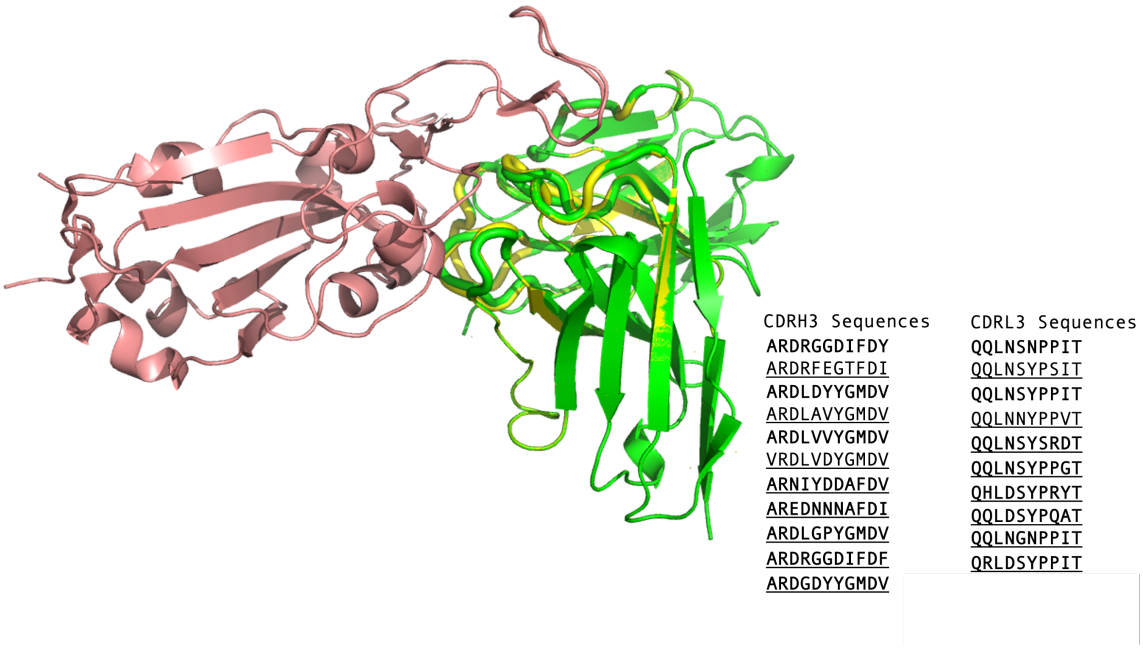


Figure 4.6: SARS-CoV-2 RBD-binding antibodies with similar predicted ABB-ML structures that span multiple clonotypes. Antibodies ADI-75774 (green) and R40-1C8 (yellow) are members of NSC32 and share only 18% CDRH3 identity. The antibodies are shown aligned to the RBD using solved structure of other cluster member CV2.2325, PDB 7wf0 (salmon). CDRH3 and CDRL3 sequences across the 11-member structural cluster are listed, grouped by lenient VH- or VL-clonotype, with the cluster representative in bold font.

4.3.5 Malaria

4.3.5.1 Epitope binning of homology modelled structures

We also tested our computational epitope binning method on a dataset of 257 antibody sequences against a malaria antigen, with epitope community information available for 240/257 (31 different epitopes, of which 10 are shared by multiple antibodies). 204 antibodies were successfully homology modelled by ABodyBuilder and

clustered with SPACE into 9 multiple occupancy structural clusters - all containing 2 members. 94 antibodies were not clustered due to having unique CDR length combinations (the remaining 92 antibodies have the same CDR lengths as other antibodies in the dataset but did not share structural similarity so were not clustered together). Due to the number of unique length CDR antibodies, we also performed structural clustering of non-CDR length matched antibodies using an adapted SPACE method and obtained a further 9 multiple occupancy clusters.

55.6% (5/9) of CDR length matched clusters and 71.4% (5/7) non-CDR length matched clusters were epitope consistent. Of the epitope consistent clusters, 2/5 CDR-length matched clusters and all non-CDR length matched clusters contained differing clonotypes and paratypes. One non-CDR length matched epitope-consistent cluster consisted of two antibodies with highly dissimilar sequences but bound to the same epitope; the antibodies shared the same heavy chain V gene assignment but different heavy chain J gene, light chain V and J genes, and CDR sequence identities - 42% CDRH3, 37.5% CDRH2, 75% CDRH1, 45% CDRL3, 33% CDRL3 and 62.5% CDRL1.

4.3.5.2 Epitope binning of deep learning modelled structures

We also modelled the malaria antibodies using ABB-ML and clustered the models using the 0.5 Å clustering threshold. A higher number of structural clusters were identified than with ABodyBuilder models, and with higher epitope consistency: 14 multiple occupancy CDR length matched clusters were obtained, with cluster sizes of 2-6 members. 92.9% of clusters (13/14) were epitope consistent. The one epitope-inconsistent cluster contained two antibodies with highly similar predicted structures and similar sequences aside from 36% CDRH3 sequence identity (100% sequence identity over CDRH2 and CDRL2, 70% CDRH1, 83% CDRL1 and 80% CDRL3).

Of the 14 clusters, fewer than half contained antibodies from the same lenient Fv-clonotype, which corresponded with lenient VH-only clonotype 6/14). Those that were from the same lenient Fv-clonotype were often highly sequence similar, with identical CDRH3 and CDRL3 regions. This was not the case for the other clusters; in one cluster with antibodies from different lenient clonotypes, the two members shared only 0.09% CDRH3 identity (though sequences were more highly conserved across the other CDRs: 75% CDRH2, 70% CDRH1, 75% CDRL3, 100% CDRL2,

80% CDRL1).

We found examples of antibodies which had been structurally clustered using ABodyBuilder, but were no longer predicted to share similar structures based on ABB-ML models. One epitope-inconsistent cluster was grouped based on ABodyBuilder model similarity but not based on ABB-ML model similarity, due to changes in the predicted structure of the CDRH3 loop (Figure 4.7). The CDRH3 loop is the most challenging antibody CDR to model due to the variability in the loop structure. ABB-ML CDRH3 prediction outperforms ABodyBuilder (ABB-ML achieved average CDRH3 prediction accuracy of 2.66 Å on a standard test set of SAbDab antibodies, compared to 3.25 Å for ABodyBuilder); therefore we can expect the ABB-ML models to be more accurate than the previous ABodyBuilder models (Abanades et al. 2022a). This is reflected in the improved epitope binning accuracy seen when clustering the ABB-ML models, compared to ABodyBuilder models.

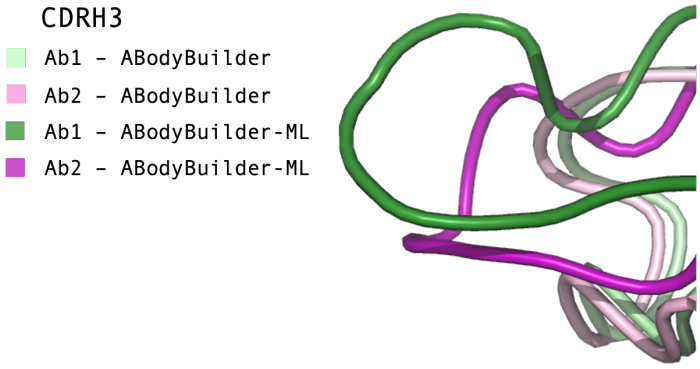


Figure 4.7: Aligned models of CDRH3 loops from two epitope-dissimilar antibodies (Ab1 and Ab2) clustered by SPACE based on similarity of ABodyBuilder structures (pale pink, pale green) but not based on ABodyBuilder-ML (ABlooper) modelled structures (dark green, dark pink).

4.3.5.3 Paratyping

We investigated the predicted paratope conservation across the malaria antibodies using paratyping (Richardson et al. 2021). Paragraph (Chinery et al. 2023) was used to predict which residues on the modelled structures would likely be involved in epitope binding, and paratopes were clustered based on amino acid position and identity. 9/10 paratype clusters identified were epitope consistent and 6/10 of these overlapped with the SPACE structural clusters. Both paratyping and SPACE identify the same 5 clusters which share lenient Fv clonotypes. The 3 epitope-consistent

clusters that are identified by paratyping and not SPACE have high paratope residue similarity but other parts of the sequence are less conserved, leading to dissimilar structure predictions.

4.4 Discussion and future work

We analysed solved X-ray crystal structures of antibodies and nanobodies bound to SARS-CoV-2 from the perspective of their structural and paratope conservation, and performed the first analysis of thousands of structural models of coronavirus-binding antibodies reported in over 100 independent literature sources. We updated the previously-reported sets of antibodies shown to bind to specific regions of the RBD (Dejnirattisai et al. 2021) and demonstrated that the antibodies within these clusters are often structurally similar but sequence dissimilar. For example, a cluster of 22 antibodies from 15 different lenient VH-clonotypes all approach the same SARS-CoV-2 RBD 'neck' epitope with a closely-related binding mode. The paratopes of these antibodies are highly conserved across the CDRH1 and CDRH2 regions (accounting for the strong bias towards IGHV3-53/IGHV3-66 gene origins) while the CDRH3 sequence can diverge substantially in sequence identity. The same phenomenon was observed for two antibodies solved engaging the RBD 'left flank' with a near-identical binding mode but just 43% CDRH3 sequence identity. These scenarios represent a problem for the functional interpretation of sequence-based clustering approaches such as clonotyping; antibodies that are functionally similar would be binned into different clusters. Inspired by the structural similarity of antibodies that bind to the same epitope, we predicted and clustered the structures of the broad set of antibodies documented in CoV-AbDab (Raybould et al. 2021a) using our SPACE algorithm. We first performed this with the available homology-based antibody modelling tool ABodyBuilder (Leem et al. 2016), and subsequently used a more recent deep-learning based method, ABodyBuilder-ML (Abanades et al. 2022a), which does not rely on CDR loop homology templates and has higher modelling accuracy. We also clustered a set of malaria-binding antibodies for which epitope information was available.

We found that structural clustering using both homology modelling and non-homology approaches achieves a very high accuracy of epitope binning. Our homology modelling analysis showed 92% of multiple-occupancy structural clusters grouped coronavirus-binding antibodies reported to bind to consistent domains, based on the available levels of epitope information. Using deep-learning based modelling, we

achieved 94% epitope consistency on coronavirus-binding antibodies and 93% for malaria-binding antibodies. This suggests that we can use our method to predict the epitopes of many as-yet uncharacterised binders, as well as prospectively to predict the epitopes of newly-isolated binding antibodies. These structural clusters also offer a unique perspective on the data that can not be identified through standard, sequence-based clonotyping. First, they can functionally-associate antibodies that derive from highly distinct clonal lineages; around 40% of our CoV-AbDab structural clusters and 50% of malaria structural clusters contain at least two antibodies from different lenient VH-clonotypes. Our CoV-AbDab clusters also functionally connect disease-response antibodies that originate from different species, which is of interest in the study of functional crossovers between immune repertoires that exploit different gene loci. Moreover, they can reveal which epitope topologies are likely to be conserved across coronavirus strains, helping to co-ordinate efforts to design prophylactics towards more fruitful sites for pan-coronavirus neutralisation.

Such powerful structure-function relationships are likely only possible due to the collation of clean, high confidence binding data. Our work demonstrates the clear value of building large class-specific databases of antibody-binders against extracellular disease-associated antigens. We hope to create similar datasets for other pathogen families and suspect others will do the same, enabling the validation of computational structural profiling in multiple different disease contexts. We were excited to find our method achieved high epitope binning accuracy on a smaller dataset of only 257 malaria-binding antibodies, though a smaller proportion of antibodies were able to be grouped into multiple occupancy clusters to infer properties compared to the larger CoV-AbDab datasets. Adjustment of the clustering similarity thresholds may allow a higher number of antibodies to be grouped, and potential functionality ascribed, but also leads to lower accuracy of epitope binning.

In addition to offering functional annotations to as-yet uncharacterised antibodies, these databases could also be used to identify important gaps in structural space. This would ensure that time-consuming experimental structure determination efforts are targeted towards sets of antibodies that yield maximal functional insight. For structural clusters with limited existing epitope knowledge, a central antibody could be selected for structural evaluation with the cognate antigen, enabling the functional annotation of many other antibodies in the same structural cluster. Similarly, structurally modelling the entire database reveals which antibodies cannot currently

be accurately modeled and should therefore be prioritised for experimental structure characterisation.

An open question remains as to how strictly an antibody's structure needs to be conserved to engage the same epitope. This is likely to be highly epitope dependent. For example, for epitopes naturally suited to VH-dominated engagement, less selection pressure would act upon light chain structure and vice versa, while some epitope topologies and environments (e.g. extent of glycosylation) may also exert different levels of pressure on complementary antibody geometries.

Overall, our results show that structural information *via* computational modeling enhances the picture of disease-characteristic convergence across SARS-CoV-2 response antibodies. It is clear that while clonotyping COVID-19 antibody repertoires can offer an indication for which epitopes are public in the response, they risk understating the true levels of same-epitope reactivity across individuals. Accurately capturing the functions of the antibodies raised during the immune response is critical when evaluating vaccine efficacy, both against the injected viral strain and against new variants of concern that may arise and alter the immunogenicity of certain epitopes. Structure prediction and clustering has a crucial role to play alongside clonotyping to yield the maximum functional inference from the vast amount of disease-specific antibody data available.

Developments of the SPACE method could include benchmarking the accuracy of clustering non-CDR length matched antibody sequences. Currently, the definition of the structural clusters includes only antibodies with the same set of 6 CDR lengths. Antibodies with differing CDR lengths can adopt highly similar structures and bind to the same epitope (Figure 4.5). Relaxation of this criteria would likely result in better recall of clustering antibodies which bind to the same epitope, but poorer precision. We performed non-CDR length matched SPACE on the homology models of malaria antibodies as a test example and achieved good performance. This may be a useful technique for smaller datasets with differing CDR lengths to obtain a higher number of multiple-occupancy clusters.

Recent developments in protein modelling techniques (Evans et al. 2022; Jumper et al. 2021) have lead to improvements in the speed and accuracy of antibody structure prediction (Abanades et al. 2022a; Ruffolo et al. 2022a). We expect further advances

in this area, and future work should consider assessment of model quality metrics and appropriate clustering methods and thresholds in greater detail than conducted here.

The information contained within CoV-AbDab has evolved over time as has data become available on antibody binding and neutralisation of different strains of SARS-CoV-2. Whilst this information is not known for every antibody, study of trends in structures correlating with strain-specific binding would be interesting, as antibodies developed against early strains of SARS-CoV-2 (both therapeutically developed and those found in convalescent patients) are often non-neutralising against the mutated epitopes of more recent strains such as BA.4 and BA.5 (Wang et al. 2022). Further, in our study, we focused on clustering binding antibodies rather than neutralising antibodies. Comparisons of structures of neutralising and non-neutralising antibodies could reveal insights into neutralising vs non-neutralising epitopes. Structures shared by both non-neutralising and neutralising antibodies may correspond to differences in paratope-epitope interactions due to differing amino acids and physicochemical properties. As more variants emerge, and more screening is conducted against multiple strains, CoV-AbDab will continue to be an incredibly valuable dataset for testing computational epitope binning methods and for furthering our understanding of antibody binding and neutralisation.

CHAPTER 5

Conclusions and Future Work

5.1 Conclusions

In this DPhil thesis we have investigated how antibody sequence data can be used as a powerful tool in the design of therapeutic antibodies. We have described ANARCII, a deep-learning approach to antibody numbering. ANARCII aims to improve upon existing methods by reducing reliance on prior annotation of closely related antibody sequences, resulting in fewer diverse sequences being discarded from subsequent analysis due to lack of numbering, or being erroneously numbered.

We have evaluated methods for predicting antibodies likely to bind to the same target epitope, for use in hit expansion and in hit identification, using antibody sequence data and antibody models built from sequences. We have developed an epitope-to-paratope pipeline leveraging information from known antibody-antigen pairs to identify antibodies likely to bind to a target antigen. By searching antibody model libraries with shared structures, we can more efficiently screen BCR repertoires or databases by reducing the number of comparison calculations required. Further, selection of shared human antibody structures as templates for therapeutic antibody design may prove useful for decreased risk of immunogenicity.

We also developed SPACE, which shows that structural clustering of antibody models groups antibodies which bind to the same epitope. We demonstrated its use on coronavirus-binding and malaria-binding antibodies. Such computational epitope

binning can reduce the need for experimental assays, and potentially identified conserved epitopes for vaccine and antibody design.

5.2 Future work

Further work is needed to make ANARCII available as a tool for others to use, such as webserver creation and development of a python package. ANARCII should also be expanded to use other numbering schemes, such as Chothia, as the existing ANARCI does.

There have been huge improvements in antibody modelling tools since the research described in this thesis began in 2019. As such, clear next steps for the epitope-to-paratope pipeline and for SPACE are to assess the methods when antibody models are built using the latest tools, such as ABodyBuilder2 (Abanades et al. 2022a). The resulting models should be quicker to generate and more accurate, and enable modelling of more structural space than was available through homology modelling. Advances in general protein modelling (Jumper et al. 2021), paratope prediction (Chinery et al. 2023) and antibody-antigen docking (Schneider et al. 2022a) may mean we can expand our database of known paratope-epitope pairs to include predicted information from docked models, rather than reliance on SAbDab. This could significantly increase the power of our proposed epitope-to-paratope pipeline, though would require high confidence in the docked structures.

The amount of data included in CoV-AbDab has increased dramatically since the original SPACE analysis, and since the updated analysis described in this thesis. Analysis of SPACE performance on the most recent datasets, now at over 11,000 sequences, including different clustering methods and similarity thresholds, should be performed. As binding/neutralisation information is available for different variants, it would be interesting to evaluate structural clusters containing antibodies able to bind multiple variants, as potential examples of conserved epitopes to target for therapeutic antibody design. For structural clusters containing antibodies where some variants can be neutralised but not others, investigation to see what the epitope change is between variants that affects the antibodies might be useful to assess whether the antibodies could be engineered to adapt to the mutation(s).

It is currently unclear how applicable SPACE is for datasets where the antibodies are known to bind multiple, diverse antigens (unlike CoV-AbDab or the malaria dataset). Just as sequence-similar antibodies may bind to different epitopes, structurally-similar antibodies may bind different epitopes if key paratope residues are dissimilar. A next step could be to evaluate SPACE on modelled SAbDab structures to assess whether antibodies binding to the same epitope can be identified, if the diversity of epitopes and antibodies means that few clusters are formed, or if antibodies which share structural similarities but in fact bind to different epitopes are grouped. In such scenarios, tightened similarity thresholds may help to distinguish antibodies most likely to bind to the same epitope.

Given our success with structure-based computational epitope binning, it would be interesting to assess whether clustering of sequence embeddings from a language model can be also used for computational epitope binning. This approach could be tested using ESM-1b embeddings, as used in ANARCI, or other language model embeddings, of antibody sequences from CoV-AbDab and the malaria dataset. These results could be compared to existing methods: SPACE, clonotyping, paratyping and Ab-Ligity.

References

Abanades, B., Georges, G., Bujotzek, A., & Deane, C. M. (2022a). ABlooper: Fast accurate antibody CDR loop structure prediction with accuracy estimation. *Bioinformatics*, *38*(7), 1877–1880.

Abanades, B., Wong, W. K., Boyles, F., Georges, G., Bujotzek, A., & Deane, C. M. (2022b). ImmuneBuilder: Deep-learning models for predicting the structures of immune proteins. *bioRxiv*.

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

Abhinandan, K., & Martin, A. C. (2008). Analysis and improvements to kabat and structurally correct numbering of antibody variable domains. *Molecular Immunology*, *45*(14), 3832–3839.

Adolf-Bryfogle, J., Xu, Q., North, B., Lehmann, A., & Dunbrack, R. L. (2015). PyIgClassify: A database of antibody CDR structural classifications. *Nucleic Acids Research*, *43*, D432–D438.

Agrawal, N. J., Helk, B., Kumar, S., Mody, N., Sathish, H. A., Samra, H. S., Buck, P. M., Li, L., & Trout, B. L. (2016). Computational tool for the early screening of monoclonal antibodies for their viscosities. *mAbs*, *8*(1), 43–48.

Aguilar Rangel, M., Bedwell, A., Costanzi, E., Taylor, R. J., Russo, R., Bernardes, G. J. L., Ricagno, S., Frydman, J., Vendruscolo, M., & Sormanni, P. (2022). Fragment-based computational design of antibodies targeting structured epitopes. *Science Advances*, *8*(45), 14.

Akbar, R., Robert, P. A., Pavlović, M., Jeliazkov, J. R., Snapkov, I., Slabodkin, A., Weber, C. R., Scheffer, L., Miho, E., & Haff, I. H. (2021). A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding [Publisher: Elsevier]. *Cell Reports*, *34*(11), 108856.

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

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., & Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation

learning [Number: 12 Publisher: Nature Publishing Group]. *Nature Methods*, *16*(12), 1315–1322.

Ambrosetti, F., Jandova, Z., & Bonvin, A. M. J. J. Information-driven antibody–antigen modelling with HADDOCK (K. Tsumoto & D. Kuroda, Eds.). In: *Computer-aided antibody design* (K. Tsumoto & D. Kuroda, Eds.). Ed. by Tsumoto, K., & Kuroda, D. Methods in Molecular Biology. Springer, 2023, pp. 267–282.

Amimeur, T., Shaver, J. M., Ketchem, R. R., Taylor, J. A., Clark, R. H., Smith, J., Van Citters, D., Siska, C. C., Smidt, P., Sprague, M., Kerwin, B. A., & Pettit, D. (2020). Designing feature-controlled humanoid antibody discovery libraries using generative adversarial networks. *bioRxiv*.

Arora, R., & Arnaout, R. (2020). Private antibody repertoires are public. *bioRxiv*.

Banach, B. B., Cerutti, G., Fahad, A. S., Shen, C.-H., Oliveira De Souza, M., Katsamba, P. S., Tsybovsky, Y., Wang, P., Nair, M. S., Huang, Y., Francino-Urdániz, I. M., Steiner, P. J., Gutiérrez-González, M., Liu, L., López Acevedo, S. N., Nazzari, A. F., Wolfe, J. R., Luo, Y., Olia, A. S., ... DeKosky, B. J. (2021). Paired heavy- and light-chain signatures contribute to potent SARS-CoV-2 neutralization in public antibody responses. *Cell Reports*, *37*(1), 109771.

Bannard, O., & Cyster, J. G. (2017). Germinal centers: Programmed for affinity maturation and antibody diversification. *Current Opinion in Immunology*, *45*, 21–30.

Baran, D., Pszolla, M. G., Lapidoth, G. D., Norn, C., Dym, O., Unger, T., Albeck, S., Tyka, M. D., & Fleishman, S. J. (2017). Principles for computational design of binding antibodies. *Proceedings of the National Academy of Sciences*, *114*(41), 10900–10905.

Barnes, C. O., Jette, C. A., Abernathy, M. E., Dam, K.-M. A., Esswein, S. R., Gristick, H. B., Malyutin, A. G., Sharaf, N. G., Huey-Tubman, K. E., Lee, Y. E., Robbiani, D. F., Nussenzweig, M. C., West, A. P., & Bjorkman, P. J. (2020). SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature*, *588*(7839), 682–687.

Bashford-Rogers, R. J. M., Smith, K. G. C., & Thomas, D. C. (2018). Antibody repertoire analysis in polygenic autoimmune diseases. *Immunology*, *155*(1), 3–17.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, *28*(1), 235–242.

Bertoglio, F., Meier, D., Langreder, N., Steinke, S., Rand, U., Simonelli, L., Heine, P. A., Ballmann, R., Schneider, K.-T., Roth, K. D. R., Ruschig, M., Riese, P., Eschke, K., Kim, Y., Schäckermann, D., Pedotti, M., Kuhn, P., Zock-Emmenthal, S., Wöhrle, J., ... Hust, M. (2021). SARS-CoV-2 neutralizing human recombinant antibodies selected from pre-pandemic healthy donors binding at RBD-ACE2 interface. *Nature Communications*, *12*(1), 1577.

Bolotin, D. A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I. Z., Putintseva, E. V., & Chudakov, D. M. (2015). MiXCR: Software for comprehensive adaptive immunity profiling. *Nature Methods*, *12*(5), 380–381.

Bonissone, S. R., Lima, T., Harris, K., Davison, L., Avanzino, B., Trinklein, N., Castellana, N., & Patel, A. (2020). Serum proteomics expands on high-affinity antibodies in immunized rabbits than deep b-cell repertoire sequencing alone.

Breden, F., Luning Prak, E. T., Peters, B., Rubelt, F., Schramm, C. A., Busse, C. E., Vander Heiden, J. A., Christley, S., Bukhari, S. A. C., Thorogood, A., Matsen IV, F. A., Wine, Y., Laserson, U., Klatzmann, D., Douek, D. C., Lefranc, M.-P., Collins, A. M., Bubela, T., Kleinstein, S. H., ... Kepler, T. B. (2017). Reproducibility and reuse of adaptive immune receptor repertoire data. *Frontiers in Immunology*, *8*.

Briney, B., Inderbitzin, A., Joyce, C., & Burton, D. R. (2019). Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*, *566*(7744), 393–397.

Brinkmann, U., & Kontermann, R. E. (2021). Bispecific antibodies [Publisher: American Association for the Advancement of Science]. *Science*, *372*(6545), 916–917.

Brooks, B. D., Closmore, A., Yang, J., Holland, M., Cairns, T., Cohen, G. H., & Bailey-Kellogg, C. (2020). Characterizing epitope binding regions of entire antibody panels by combining experimental and computational analysis of antibody: Antigen binding competition. *Molecules*, *25*(16), 3659.

Brouwer, P. J. M., Caniels, T. G., van der Straten, K., Snitselaar, J. L., Aldon, Y., Bangaru, S., Torres, J. L., Okba, N. M. A., Claireaux, M., Kerster, G., Bentlage, A. E. H., van Haaren, M. M., Guerra, D., Burger, J. A., Schermer, E. E., Verheul, K. D., van der Velde, N., van der Kooi, A., van Schooten, J., ... van Gils, M. J. (2020). Potent neutralizing antibodies from COVID-19 patients define multiple targets of vulnerability. *Science*, *369*(6504), 643–650.

Bullen, G., Galson, J. D., Villar, P., Moreels, L., Ledsgaard, L., Mattiuzzo, G., Hall, G., Bentley, E. M., Masters, E. W., Tang, D., Millett, S., Tongue, D., Brown, R., Diamantopoulos, I., Parthiban, K., Tebbutt, C., Leah, R., Chaitanya, K., Pazeraitis, D., ... Karatt-Vellatt, A. (2020). Deep mining of early antibody response in COVID-19 patients yields potent neutralisers and reveals high level of convergence. *bioRxiv*.

Cao, L., Coventry, B., Goreshnik, I., Huang, B., Sheffler, W., Park, J. S., Jude, K. M., Marković, I., Kadam, R. U., Verschueren, K. H. G., Verstraete, K., Walsh, S. T. R., Bennett, N., Phal, A., Yang, A., Kozodoy, L., DeWitt, M., Picton, L., Miller, L., ... Baker, D. (2022). Design of protein-binding proteins from the target structure alone [Number: 7910 Publisher: Nature Publishing Group]. *Nature*, *605*(7910), 551–560.

Carter, P. J., & Lazar, G. A. (2018). Next generation antibody drugs: Pursuit of the 'high-hanging fruit' [Number: 3 Publisher: Nature Publishing Group]. *Nature Reviews Drug Discovery*, *17*(3), 197–223.

Chailyan, A., Tramontano, A., & Marcatili, P. (2012). A database of immunoglobulins with integrated tools: DIGIT. *Nucleic Acids Research*, *40*, D1230–D1234.

Chan, B. M., Badh, A., Berry, K. A., Grauer, S. A., & King, C. T. (2018). Flow cytometry-based epitope binning using competitive binding profiles for the characterization of monoclonal antibodies against cellular and soluble protein

targets [Publisher: SAGE Publications Inc STM]. *SLAS DISCOVERY: Advancing the Science of Drug Discovery*, *23*(7), 613–623.

Chen, F., Tzarum, N., Wilson, I. A., & Law, M. (2019). VH1-69 antiviral broadly neutralizing antibodies: Genetics, structures, and relevance to rational vaccine design. *Current opinion in virology*, *34*, 149–159.

Chen, T., & Guestrin, C. XGBoost: A scalable tree boosting system. In: In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. KDD 2016: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM, 2016, August 13, 785–794.

Chennamsetty, N., Helk, B., Voynov, V., Kayser, V., & Trout, B. L. (2009). Aggregation-prone motifs in human immunoglobulin g. *Journal of Molecular Biology*, *391*(2), 404–413.

Cheung, W. C., Beausoleil, S. A., Zhang, X., Sato, S., Schieferl, S. M., Wieler, J. S., Beaudet, J. G., Ramenani, R. K., Popova, L., Comb, M. J., Rush, J., & Polakiewicz, R. D. (2012). A proteomics approach for the identification and cloning of monoclonal antibodies from serum [Number: 5 Publisher: Nature Publishing Group]. *Nature Biotechnology*, *30*(5), 447–452.

Chinery, L., Wahome, N., Moal, I., & Deane, C. M. (2023). Paragraph—antibody paratope prediction using graph neural networks with minimal feature vectors. *Bioinformatics*, *39*(1), btac732.

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

Choi, Y., & Deane, C. M. (2011). Predicting antibody complementarity determining region structures without classification. *Molecular BioSystems*, *7*(12), 3327–3334.

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

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., Colman, P. M., Spinelli, S., Alzari, P. M., & Poljak, R. J. (1989). Conformations of immunoglobulin hypervariable regions. *Nature*, *342*(6252), 877–883.

Clark, S. A., Clark, L. E., Pan, J., Coscia, A., McKay, L. G. A., Shankar, S., Johnson, R. I., Brusic, V., Choudhary, M. C., Regan, J., Li, J. Z., Griffiths, A., & Abraham, J. (2021). SARS-CoV-2 evolution in an immunocompromised host reveals shared neutralization escape mechanisms. *Cell*, *184*(10), 2605–2617.e18.

Collatz, M., Mock, F., Barth, E., Hölzer, M., Sachse, K., & Marz, M. (2021). EpiDope: A deep neural network for linear b-cell epitope prediction. *Bioinformatics*, *37*(4), 448–455.

Correia, B. E., Bates, J. T., Loomis, R. J., Baneyx, G., Carrico, C., Jardine, J. G., Rupert, P., Correnti, C., Kalyuzhniy, O., Vittal, V., Connell, M. J., Stevens, E., Schroeter, A., Chen, M., MacPherson, S., Serra, A. M., Adachi, Y., Holmes, M. A., Li, Y., ... Schief, W. R. (2014). Proof of principle for epitope-focused vaccine design. *Nature*, *507*(7491), 201–206.

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

Cunningham, O., Scott, M., Zhou, Z. S., & Finlay, W. J. (2021). Polyreactivity and polyspecificity in therapeutic antibody development: Risk factors for failure in preclinical and clinical development campaigns. *mAbs*, *13*(1), 1999195.

Dai, B., & Bailey-Kellogg, C. (2021). Protein interaction interface region prediction by geometric deep learning. *Bioinformatics*, *37*(17), 2580–2588.

Dall'Acqua, W., Goldman, E. R., Lin, W., Teng, C., Tsuchiya, D., Li, H., Ysern, X., Braden, B. C., Li, Y., Smith-Gill, S. J., & Mariuzza, R. A. (1998). A mutational analysis of binding interactions in an antigen-antibody protein-protein complex. *Biochemistry*, *37*(22), 7981–7991.

Dashivets, T., Stracke, J., Dengl, S., Knaupp, A., Pollmann, J., Buchner, J., & Schlothauer, T. (2016). Oxidation in the complementarity-determining regions differentially influences the properties of therapeutic antibodies. *mAbs*, *8*(8), 1525–1535.

de Assis, R. R., Jain, A., Nakajima, R., Jasinskas, A., Felgner, J., Obiero, J. M., Norris, P. J., Stone, M., Simmons, G., Bagri, A., Irsch, J., Schreiber, M., Buser, A., Holbro, A., Battegay, M., Hosimer, P., Noesen, C., Adenaiye, O., Tai, S., ... Khan, S. (2021). Analysis of SARS-CoV-2 antibodies in COVID-19 convalescent blood using a coronavirus antigen microarray. *Nature Communications*, *12*, 6.

Dejnirattisai, W., Zhou, D., Ginn, H. M., Duyvesteyn, H. M., Supasa, P., Case, J. B., Zhao, Y., Walter, T. S., Mentzer, A. J., Liu, C., Wang, B., Paesen, G. C., Slon-Campos, J., López-Camacho, C., Kafai, N. M., Bailey, A. L., Chen, R. E., Ying, B., Thompson, C., ... Screaton, G. R. (2021). The antigenic anatomy of SARS-CoV-2 receptor binding domain. *Cell*, *184*(8), 2183–2200.e22.

DeKosky, B. J., Ippolito, G. C., Deschner, R. P., Lavinder, J. J., Wine, Y., Rawlings, B. M., Varadarajan, N., Giesecke, C., Dörner, T., Andrews, S. F., Wilson, P. C., Hunicke-Smith, S. P., Willson, C. G., Ellington, A. D., & Georgiou, G. (2013). High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire [Number: 2 Publisher: Nature Publishing Group]. *Nature Biotechnology*, *31*(2), 166–169.

Du, S., Cao, Y., Zhu, Q., Yu, P., Qi, F., Wang, G., Du, X., Bao, L., Deng, W., Zhu, H., Liu, J., Nie, J., Zheng, Y., Liang, H., Liu, R., Gong, S., Xu, H., Yisimayi, A., Lv, Q., ... Qin, C. (2020). Structurally resolved SARS-CoV-2 antibody shows high efficacy in severely infected hamsters and provides a potent cocktail pairing strategy. *Cell*, *183*(4), 1013–1023.e13.

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

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

Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., & Pande, V. S. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics [Publisher: Public Library of Science]. *PLOS Computational Biology*, *13*(7), e1005659.

Ebejer, J.-P., Finn, P. W., Wong, W. K., Deane, C. M., & Morris, G. M. (2019). Ligity: A non-superpositional, knowledge-based approach to virtual screening. *Journal of Chemical Information and Modeling*, *59*(6), 2600–2616.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, *7*(10), e1002195.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797.

Ehrenmann, F., Giudicelli, V., Duroux, P., & Lefranc, M.-P. (2011). IMGT/collier de perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains) [Publisher: Cold Spring Harbor Laboratory Press]. *Cold Spring Harbor Protocols*, (6).

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., ... Hassabis, D. (2022). Protein complex prediction with AlphaFold-multimer. *bioRxiv*.

Eyal, E., Gerzon, S., Potapov, V., Edelman, M., & Sobolev, V. (2005). The limit of accuracy of protein modeling: Influence of crystal packing on protein structure. *Journal of Molecular Biology*, *351*(2), 431–442.

Fernández-Quintero, M. L., Kokot, J., Waibl, F., Fischer, A.-L. M., Quoika, P. K., Deane, C. M., & Liedl, K. R. (2022). Challenges in antibody structure prediction. *bioRxiv*.

Finn, J. A., Dong, J., Sevy, A. M., Parrish, E., Gilchuk, I., Nargi, R., Scarlett-Jones, M., Reichard, W., Bombardi, R., Voss, T. G., Meiler, J., & Crowe, J. E. (2020). Identification of structurally related antibodies in antibody sequence databases using rosetta-derived position-specific scoring. *Structure*, *28*(10), 1124–1130.e5.

Forgacs, D., Abreu, R. B., Sautto, G. A., Kirchenbaum, G. A., Drabek, E., Williamson, K. S., Kim, D., Emerling, D. E., & Ross, T. M. (2021). Convergent antibody evolution and clonotype expansion following influenza virus vaccination. *PLoS ONE*, *16*(2), e0247253.

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.

Galilee, M., Britan-Rosich, E., Griner, S. L., Uysal, S., Baumgärtel, V., Lamb, D. C., Kossiakoff, A. A., Kotler, M., Stroud, R. M., Marx, A., & Alian, A. (2016). The preserved HTH-docking cleft of HIV-1 integrase is functionally critical. *Structure*, *24*(11), 1936–1946.

Galson, J. D., Pollard, A. J., Trück, J., & Kelly, D. F. (2014). Studying the antibody repertoire after vaccination: Practical applications. *Trends in Immunology*, *35*(7), 319–331.

Galson, J. D., Schaetzle, S., Bashford-Rogers, R. J. M., Raybould, M. I. J., Kovaltsuk, A., Kilpatrick, G. J., Minter, R., Finch, D. K., Dias, J., James, L. K., Thomas, G., Lee, W.-Y. J., Betley, J., Cavlan, O., Leech, A., Deane, C. M., Seoane, J., Caldas, C., Pennington, D. J., ... Osbourn, J. (2020). Deep sequencing of b cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures. *Frontiers in Immunology*, *11*, 605170.

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

Ge, J., Wang, R., Ju, B., Zhang, Q., Sun, J., Chen, P., Zhang, S., Tian, Y., Shan, S., Cheng, L., Zhou, B., Song, S., Zhao, J., Wang, H., Shi, X., Ding, Q., Liu, L., Zhao, J., Zhang, Z., ... Zhang, L. (2021). Antibody neutralization of SARS-CoV-2 through ACE2 receptor mimicry. *Nature Communications*, *12*, 250.

Georgiou, G., Ippolito, G. C., Beausang, J., Busse, C. E., Wardemann, H., & Quake, S. R. (2014). The promise and challenge of high-throughput sequencing of the antibody repertoire [Number: 2 Publisher: Nature Publishing Group]. *Nature Biotechnology*, *32*(2), 158–168.

Gershoni, J. M., Roitburd-Berman, A., Siman-Tov, D. D., Freund, N. T., & Weiss, Y. (2007). Epitope mapping. *BioDrugs*, *21*(3), 145–156.

Ghanbarpour, A., Jiang, M., Foster, D., & Chai, Q. (2022). Structure-free antibody paratope similarity prediction for *in silico* epitope binning via protein language models. *SSRN Electronic Journal*.

Ghraichy, M., Galson, J. D., Kelly, D. F., & Trück, J. (2018). B-cell receptor repertoire sequencing in patients with primary immunodeficiency: A review. *Immunology*, *153*(2), 145–160.

Ghraichy, M., Galson, J. D., Kovaltsuk, A., von Niederhäusern, V., Pachlopnik Schmid, J., Recher, M., Jauch, A. J., Miho, E., Kelly, D. F., Deane, C. M., & Trück, J. (2020). Maturation of the human immunoglobulin heavy chain repertoire with age. *Frontiers in Immunology*, *11*.

Gidoni, M., Snir, O., Peres, A., Polak, P., Lindeman, I., Mikocziova, I., Sarna, V. K., Lundin, K. E. A., Clouser, C., Vigneault, F., Collins, A. M., Sollid, L. M., & Yaari, G. (2019). Mosaic deletion patterns of the human antibody heavy chain gene locus shown by bayesian haplotyping [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *10*(1), 628.

Giudicelli, V., Chaume, D., Bodmer, J., Müller, W., Busin, C., Marsh, S., Bontrop, R., Marc, L., Malik, A., & Lefranc, M.-P. (1997). IMGT, the international ImMunoGeneTics database. *Nucleic Acids Research*, *25*(1), 206–211.

Giudicelli, V., Chaume, D., & Lefranc, M.-P. (2005). IMGT/GENE-DB: A comprehensive database for human and mouse immunoglobulin and t cell receptor genes. *Nucleic Acids Research*, *33*, D256–D261.

Goldstein, L. D., Chen, Y.-J. J., Wu, J., Chaudhuri, S., Hsiao, Y.-C., Schneider, K., Hoi, K. H., Lin, Z., Guerrero, S., Jaiswal, B. S., Stinson, J., Antony, A., Pahuja, K. B., Seshasayee, D., Modrusan, Z., Hötzel, I., & Seshagiri, S. (2019).

Massively parallel single-cell b-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. *Communications Biology*, *2*(1), 304.

Goodwin, E., Gilman, M. S. A., Wrapp, D., Chen, M., Ngwuta, J. O., Moin, S. M., Bai, P., Sivasubramanian, A., Connor, R. I., Wright, P. F., Graham, B. S., McLellan, J. S., & Walker, L. M. (2018). Infants infected with respiratory syncytial virus generate potent neutralizing antibodies that lack somatic hypermutation. *Immunity*, *48*(2), 339–349.e5.

Greiff, V., Bhat, P., Cook, S. C., Menzel, U., Kang, W., & Reddy, S. T. (2015a). A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status [Publisher: BioMed Central]. *Genome Medicine*, *7*(1), 1–15.

Greiff, V., Miho, E., Menzel, U., & Reddy, S. T. (2015b). Bioinformatic and statistical analysis of adaptive immune repertoires [Publisher: Elsevier]. *Trends in Immunology*, *36*(11), 738–749.

Greiff, V., Weber, C. R., Palme, J., Bodenhofer, U., Miho, E., Menzel, U., & Reddy, S. T. (2017). Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires [Publisher: Am Assoc Immnol]. *The Journal of Immunology*, *199*(8), 2985–2997.

Gubbins, M. J., Plummer, F. A., Yuan, X. Y., Johnstone, D., Drebot, M., Andonova, M., Andonov, A., & Berry, J. D. (2005). Molecular characterization of a panel of murine monoclonal antibodies specific for the SARS-coronavirus. *Molecular Immunology*, *42*(1), 125–136.

Guo, Y., Huang, L., Zhang, G., Yao, Y., Zhou, H., Shen, S., Shen, B., Li, B., Li, X., Zhang, Q., Chen, M., Chen, D., Wu, J., Fu, D., Zeng, X., Feng, M., Pi, C., Wang, Y., Zhou, X., ... Rao, Z. (2021). A SARS-CoV-2 neutralizing antibody with extensive spike binding coverage and modified for optimal therapeutic outcomes. *Nature Communications*, *12*, 2623.

Gupta, N. T., Vander Heiden, J. A., Uduman, M., Gadala-Maria, D., Yaari, G., & Kleinstein, S. H. (2015). Change-o: A toolkit for analyzing large-scale b cell immunoglobulin repertoire sequencing data. *Bioinformatics*, *31*(20), 3356–3358.

Han, X., Wang, Y., Li, S., Hu, C., Li, T., Gu, C., Wang, K., Shen, M., Wang, J., Hu, J., Wu, R., Mu, S., Gong, F., Chen, Q., Gao, F., Huang, J., Long, Y., Luo, F., Song, S., ... Jin, A. (2021). A rapid and efficient screening system for neutralizing antibodies and its application for SARS-CoV-2. *Frontiers in Immunology*, *12*.

Henikoff, J. G., & Henikoff, S. (1996). Using substitution probabilities to improve position-specific scoring matrices. *Bioinformatics*, *12*(2), 135–143.

Henkel, S., Wellhausen, R., Woitalla, D., Marcus, K., & May, C. Epitope mapping using peptide microarray in autoantibody profiling (P. C. Li, A. Sedighi, & L. Wang, Eds.). In: *Microarray technology: Methods and applications* (P. C. Li, A. Sedighi, & L. Wang, Eds.). Ed. by Li, P. C., Sedighi, A., & Wang, L. Methods in Molecular Biology. Springer, 2016, pp. 209–224.

Hershberg, U., & Luning Prak, E. T. (2015). The analysis of clonal expansions in normal and autoimmune b cell repertoires. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *370*(1676), 20140239.

Honegger, A., & Plückthun, A. (2001). Yet another numbering scheme for immunoglobulin variable domains: An automatic modeling and analysis tool. *Journal of Molecular Biology*, *309*(3), 657–670.

Hummer, A. M., Abanades, B., & Deane, C. M. (2022). Advances in computational structure-based antibody design. *Current Opinion in Structural Biology*, *74*, 102379.

Huo, J., Le Bas, A., Ruza, R. R., Duyvesteyn, H. M. E., Mikolajek, H., Malinauskas, T., Tan, T. K., Rijal, P., Dumoux, M., Ward, P. N., Ren, J., Zhou, D., Harrison, P. J., Weckener, M., Clare, D. K., Vogirala, V. K., Radecke, J., Moynié, L., Zhao, Y., … Naismith, J. H. (2020). Neutralizing nanobodies bind SARS-CoV-2 spike RBD and block interaction with ACE2. *Nature Structural & Molecular Biology*, *27*(9), 846–854.

Hurlburt, N. K., Seydoux, E., Wan, Y.-H., Edara, V. V., Stuart, A. B., Feng, J., Suthar, M. S., McGuire, A. T., Stamatatos, L., & Pancera, M. (2020). Structural basis for potent neutralization of SARS-CoV-2 and role of antibody affinity maturation. *Nature Communications*, *11*, 5413.

Hurtado, J., Flynn, C., Lee, J. H., Salcedo, E., Cottrell, A, Skog, P. D., Nemazee, D., Schief, W. R., Landais, E., Sok, D., & Briney, B. (2022). Efficient isolation of rare b cells using next-generation antigen barcoding. *bioRxiv*.

Jaffe, D. B., Shahi, P., Adams, B. A., Chrisman, A. M., Finnegan, P. M., Raman, N., Royall, A. E., Tsai, F., Vollbrecht, T., Reyes, D. S., Hepler, N. L., & McDonnell, W. J. (2022). Functional antibodies exhibit light chain coherence. *Nature*, *611*(7935), 352–357.

Jefferis, R. (2012). Isotype and glycoform selection for antibody therapeutics. *Archives of Biochemistry and Biophysics*, *526*(2), 159–166.

Jespersen, M. C., Mahajan, S., Peters, B., Nielsen, M., & Marcatili, P. (2019). Antibody specific b-cell epitope predictions: Leveraging information from antibody-antigen protein complexes. *Frontiers in Immunology*, *10*.

Jespersen, M. C., Peters, B., Nielsen, M., & Marcatili, P. (2017). BepiPred-2.0: Improving sequence-based b-cell epitope prediction using conformational epitopes. *Nucleic Acids Research*, *45*, W24–W29.

Jones, B. E., Brown-Augsburger, P. L., Corbett, K. S., Westendorf, K., Davies, J., Cujec, T. P., Wiethoff, C. M., Blackbourne, J. L., Heinz, B. A., Foster, D., Higgs, R. E., Balasubramaniam, D., Wang, L., Zhang, Y., Yang, E. S., Bidshahri, R., Kraft, L., Hwang, Y., Žentelis, S., … Falconer, E. (2021). The neutralizing antibody, LY-CoV555, protects against SARS-CoV-2 infection in nonhuman primates. *Science Translational Medicine*, *13*(593), eabf1906.

Ju, B., Zhang, Q., Ge, J., Wang, R., Sun, J., Ge, X., Yu, J., Shan, S., Zhou, B., Song, S., Tang, X., Yu, J., Lan, J., Yuan, J., Wang, H., Zhao, J., Zhang, S., Wang, Y., Shi, X., … Zhang, L. (2020). Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature*, *584*(7819), 115–119.

Jubb, H. C., Higueruelo, A. P., Ochoa-Montaño, B., Pitt, W. R., Ascher, D. B., & Blundell, T. L. (2017). Arpeggio: A web server for calculating and visualising interatomic interactions in protein structures. *Journal of Molecular Biology*, *429*(3), 365–371.

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

Kabat, E. A. (1983). *Sequences of proteins of immunological interest*. National Institutes of Health.

Kaleli, N. E., Karadag, M., & Kalyoncu, S. (2019). Phage display derived therapeutic antibodies have enriched aliphatic content: Insights for developability issues. *Proteins: Structure, Function, and Bioinformatics*, *87*(7), 607–618.

Kaplon, H., Chenoweth, A., Crescioli, S., & Reichert, J. M. (2022). Antibodies to watch in 2022. *mAbs*, *14*(1), 2014296.

Kaplon, H., Crescioli, S., Chenoweth, A., Visweswaraiah, J., & Reichert, J. M. (2023). Antibodies to watch in 2023. *mAbs*, *15*(1), 2153410.

Kelly, R. L., Zhao, J., Le, D., & Wittrup, K. D. (2017). Nonspecificity in a nonimmune human scFv repertoire. *mAbs*, *9*(7), 1029–1035.

Kelow, S. P., Adolf-Bryfogle, J., & Dunbrack, R. L. (2020). Hiding in plain sight: Structure and sequence analysis reveals the importance of the antibody DE loop for antibody-antigen binding. *mAbs*, *12*(1), 1840005.

Khetan, R., Curtis, R., Deane, C. M., Hadsund, J. T., Kar, U., Krawczyk, K., Kuroda, D., Robinson, S. A., Sormanni, P., Tsumoto, K., Warwicker, J., & Martin, A. C. (2022). Current advances in biopharmaceutical informatics: Guidelines, impact and challenges in the computational developability assessment of antibody therapeutics. *mAbs*, *14*(1), 2020082.

Kim, C., Ryu, D.-K., Lee, J., Kim, Y.-I., Seo, J.-M., Kim, Y.-G., Jeong, J.-H., Kim, M., Kim, J.-I., Kim, P., Bae, J. S., Shim, E. Y., Lee, M. S., Kim, M. S., Noh, H., Park, G.-S., Park, J. S., Son, D., An, Y., ... Lee, S.-Y. (2021). A therapeutic neutralizing antibody targeting receptor binding domain of SARS-CoV-2 spike protein. *Nature Communications*, *12*, 288.

Kim, D., & Park, D. (2019). Deep sequencing of b cell receptor repertoire. *BMB Reports*, *52*(9), 540–547.

Klein, F., Disking, R., Scheid, J. F., Gaebler, C., Mouquet, H., Georgiev, I., Pancera, M., Zhou, T., Incesu, R.-B., Fu, B. Z., Gnanapragasam, P. N. P., Oliveira, T. Y., Seaman, M. S., Kwong, P. D., Bjorkman, P. J., & Nussenzweig, M. C. (2013). Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. *Cell*, *153*(1), 126–138.

Koenig, P.-A., Das, H., Liu, H., Kümmerer, B. M., Gohr, F. N., Jenster, L.-M., Schiffelers, L. D. J., Tesfamariam, Y. M., Uchima, M., Wuerth, J. D., Gatterdam, K., Ruetalo, N., Christensen, M. H., Fandrey, C. I., Normann, S., Tödtmann, J. M. P., Pritzl, S., Hanke, L., Boos, J., ... Schmidt, F. I. (2021). Structure-

guided multivalent nanobodies block SARS-CoV-2 infection and suppress mutational escape. *Science*, *371*(6530).

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

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

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

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

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.

Kreye, J., Reincke, S. M., Kornau, H.-C., Sánchez-Sendin, E., Corman, V. M., Liu, H., Yuan, M., Wu, N. C., Zhu, X., Lee, C.-C. D., Trimpert, J., Höltje, M., Dietert, K., Stöffler, L., von Wardenburg, N., van Hoof, S., Homeyer, M. A., Hoffmann, J., Abdelgawad, A., … Prüss, H. (2020). A therapeutic non-self-reactive SARS-CoV-2 antibody protects from lung pathology in a COVID-19 hamster model. *Cell*, *183*(4), 1058–1069.e19.

Kringelum, J. V., Lundegaard, C., Lund, O., & Nielsen, M. (2012). Reliable b cell epitope predictions: Impacts of method development and improved benchmarking [Publisher: Public Library of Science]. *PLOS Computational Biology*, *8*(12), e1002829.

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

Lai, P.-K., Fernando, A., Cloutier, T. K., Kingsbury, J. S., Gokarn, Y., Halloran, K. T., Calero-Rubio, C., & Trout, B. L. (2021). Machine learning feature selection for predicting high concentration therapeutic antibody aggregation. *Journal of Pharmaceutical Sciences*, *110*(4), 1583–1591.

Langerak, A. W., ed. *Immunogenetics: Methods and protocols*. Vol. 2453. Methods in Molecular Biology. New York, NY: Springer US, 2022.

Lapidoth, G., Parker, J., Prilusky, J., & Fleishman, S. J. (2019). AbPredict 2: A server for accurate and unstrained structure prediction of antibody variable domains. *Bioinformatics*, *35*(9), 1591–1593.

Laustsen, A. H., Greiff, V., Karatt-Vellatt, A., Muyldermans, S., & Jenkins, T. P. (2021). Animal immunization, in vitro display technologies, and machine learning for antibody discovery. *Trends in Biotechnology*, *39*(12), 1263–1273.

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

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

Leem, J., Mitchell, L. S., Farmery, J. H., Barton, J., & Galson, J. D. (2022). Deciphering the language of antibodies using self-supervised learning. *Patterns*, *3*(7), 100513.

Lefranc, M.-P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V., & Lefranc, G. (2003). IMGT unique numbering for immunoglobulin and t cell receptor variable domains and ig superfamily v-like domains. *Developmental & Comparative Immunology*, *27*(1), 55–77.

Leman, J. K., Weitzner, B. D., Lewis, S. M., Adolf-Bryfogle, J., Alam, N., Alford, R. F., Aprahamian, M., Baker, D., Barlow, K. A., Barth, P., Basanta, B., Bender, B. J., Blacklock, K., Bonet, J., Boyken, S. E., Bradley, P., Bystroff, C., Conway, P., Cooper, S., ... Bonneau, R. (2020). Macromolecular modeling and design in rosetta: Recent methods and frameworks. *Nature Methods*, *17*(7), 665–680.

Li, A., Rue, M., Zhou, J., Wang, H., Goldwasser, M. A., Neuberg, D., Dalton, V., Zuckerman, D., Lyons, C., Silverman, L. B., Sallan, S. E., Gribben, J. G., & for the Dana-Farber Cancer Institute ALL Consortium. (2004). Utilization of ig heavy chain variable, diversity, and joining gene segments in children with b-lineage acute lymphoblastic leukemia: Implications for the mechanisms of VDJ recombination and for pathogenesis. *Blood*, *103*(12), 4602–4609.

Li, D., Edwards, R. J., Manne, K., Martinez, D. R., Schäfer, A., Alam, S. M., Wiehe, K., Lu, X., Parks, R., Sutherland, L. L., Oguin, T. H., McDanal, C., Perez, L. G., Mansouri, K., Gobeil, S. M., Janowska, K., Stalls, V., Kopp, M., Cai, F., ... Saunders, K. O. (2021a). In vitro and in vivo functions of SARS-CoV-2 infection-enhancing and neutralizing antibodies. *Cell*, *184*(16), 4203–4219.e32.

Li, L., Chen, S., Miao, Z., Liu, Y., Liu, X., Xiao, Z., & Cao, Y. (2019). AbRSA: A robust tool for antibody numbering. *Protein Science*, *28*(8), 1524–1531.

Li, T., Cai, H., Yao, H., Zhou, B., Zhang, N., van Vlissingen, M. F., Kuiken, T., Han, W., GeurtsvanKessel, C. H., Gong, Y., Zhao, Y., Shen, Q., Qin, W., Tian, X.-X., Peng, C., Lai, Y., Wang, Y., Hutter, C. A. J., Kuo, S.-M., ... Li, D. (2021b). A synthetic nanobody targeting RBD protects hamsters from SARS-CoV-2 infection. *Nature Communications*, *12*, 4635.

Liberis, E., Veličković, P., Sormanni, P., Vendruscolo, M., & Liò, P. (2018). Parapred: Antibody paratope prediction using convolutional and recurrent neural networks (J. Hancock, Ed.). *Bioinformatics*, *34*(17), 2944–2950.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Santos Costa, A. d., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2022). Language models of protein sequences at the scale of evolution enable accurate structure prediction.

Liu, H., Wu, N. C., Yuan, M., Bangaru, S., Torres, J. L., Caniels, T. G., van Schooten, J., Zhu, X., Lee, C.-C. D., Brouwer, P. J., van Gils, M. J., Sanders, R. W.,

Ward, A. B., & Wilson, I. A. (2020a). Cross-neutralization of a SARS-CoV-2 antibody to a functionally conserved site is mediated by avidity. *Immunity*, *53*(6), 1272–1280.e5.

Liu, L., Wang, P., Nair, M. S., Yu, J., Rapp, M., Wang, Q., Luo, Y., Chan, J. F.-W., Sahi, V., Figueroa, A., Guo, X. V., Cerutti, G., Bimela, J., Gorman, J., Zhou, T., Chen, Z., Yuen, K.-Y., Kwong, P. D., Sodroski, J. G., ... Ho, D. D. (2020b). Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature*, *584*(7821), 450–456.

Liu, T., Shi, K., & Li, W. (2020c). Deep learning methods improve linear b-cell epitope prediction. *BioData Mining*, *13*(1), 1.

Liu, X., Taylor, R. D., Griffin, L., Coker, S.-F., Adams, R., Ceska, T., Shi, J., Lawson, A. D. G., & Baker, T. (2017). Computational design of an epitope-specific keap1 binding antibody using hotspot residues grafting and CDR loop swapping. *Scientific Reports*, *7*(1), 41306.

Lu, L. L., Suscovich, T. J., Fortune, S. M., & Alter, G. (2018). Beyond binding: Antibody effector functions in infectious diseases [Number: 1 Publisher: Nature Publishing Group]. *Nature Reviews Immunology*, *18*(1), 46–61.

López-Santibáñez-Jácome, L., Avendaño-Vázquez, S. E., & Flores-Jasso, C. F. (2019). The pipeline repertoire for ig-seq analysis. *Frontiers in Immunology*, *10*, 899.

MacLennan, I. C. (1994). Germinal centers. *Annual Review of Immunology*, *12*, 117–139.

Mahita, J., Kim, D.-G., Son, S., Choi, Y., Kim, H.-S., & Bailey-Kellogg, C. (2022). Computational epitope binning reveals functional equivalence of sequence-divergent paratopes. *Computational and Structural Biotechnology Journal*, *20*, 2169–2180.

Makowski, E. K., Kinnunen, P. C., Huang, J., Wu, L., Smith, M. D., Wang, T., Desai, A. A., Streu, C. N., Zhang, Y., Zupancic, J. M., Schardt, J. S., Linderman, J. J., & Tessier, P. M. (2022). Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nature Communications*, *13*(1), 3788.

Marcatili, P., Rosi, A., & Tramontano, A. (2008). PIGS: Automatic prediction of antibody structures. *Bioinformatics*, *24*(17), 1953–1954.

Marks, C., & Deane, C. M. (2020). How repertoire data are changing antibody science. *Journal of Biological Chemistry*, *295*(29), 9823–9837.

Marks, C., Hummer, A. M., Chin, M., & Deane, C. M. (2021). Humanization of antibodies using a machine learning approach on large-scale repertoire data (P. L. Martelli, Ed.). *Bioinformatics*, *37*(22), 4041–4047.

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

Mason, D. M., Friedensohn, S., Weber, C. R., Jordi, C., Wagner, B., Meng, S. M., Ehling, R. A., Bonati, L., Dahinden, J., Gainza, P., Correia, B. E., & Reddy, S. T. (2021). Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nature Biomedical Engineering*, *5*(6), 600–612.

Matharu, Z., Bee, C., Schwarz, F., Chen, H., Tomlinson, M., Wu, G., Rakestraw, G., Hornsby, M., Drake, A., Strop, P., Rajpal, A., & Dollinger, G. (2021). High-throughput surface plasmon resonance biosensors for identifying diverse therapeutic monoclonal antibodies [Publisher: American Chemical Society]. *Analytical Chemistry*, *93*(49), 16474–16480.

Melchers, F. (2015). Checkpoints that control b cell development [Publisher: American Society for Clinical Investigation]. *The Journal of Clinical Investigation*, *125*(6), 2203–2210.

Methot, S., & Di Noia, J. Molecular mechanisms of somatic hypermutation and class switch recombination. In: In *Advances in immunology*. Vol. 133. Elsevier, 2017, pp. 37–87.

Miersch, S., Li, Z., Saberianfar, R., Ustav, M., Brett Case, J., Blazer, L., Chen, C., Ye, W., Pavlenco, A., Gorelik, M., Garcia Perez, J., Subramania, S., Singh, S., Ploder, L., Ganaie, S., Chen, R. E., Leung, D. W., Pandolfi, P. P., Novelli, G., ... Sidhu, S. S. (2021). Tetravalent SARS-CoV-2 neutralizing antibodies show enhanced potency and resistance to escape mutations. *Journal of Molecular Biology*, *433*(19), 167177.

Miho, E., Yermanos, A., Weber, C. R., Berger, C. T., Reddy, S. T., & Greiff, V. (2018). Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. *Frontiers in Immunology*, *9*, 224.

Montague, Z., Lv, H., Otwinowski, J., DeWitt, W. S., Isacchini, G., Yip, G. K., Ng, W. W., Tsang, O. T.-Y., Yuan, M., Liu, H., Wilson, I. A., Peiris, J. S. M., Wu, N. C., Nourmohammad, A., & Mok, C. K. P. (2021). Dynamics of b cell repertoires and emergence of cross-reactive responses in patients with different severities of COVID-19. *Cell Reports*, *35*(8).

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

Nagasawa, T. (2006). Microenvironmental niches in the bone marrow required for b-cell development. *Nature Reviews. Immunology*, *6*(2), 107–116.

Navis, M., Tran, K., Bale, S., Phad, G. E., Guenaga, J., Wilson, R., Soldemo, M., McKee, K., Sundling, C., Mascola, J., Li, Y., Wyatt, R. T., & Karlsson Hedestam, G. B. (2014). HIV-1 receptor binding site-directed antibodies using a VH1-2 gene segment orthologue are activated by env trimer immunization. *PLoS Pathogens*, *10*(8), e1004337.

Nichols, P., Li, L., Kumar, S., Buck, P. M., Singh, S. K., Goswami, S., Balthazor, B., Conley, T. R., Sek, D., & Allen, M. J. (2015). Rational design of viscosity reducing mutants of a monoclonal antibody: Hydrophobic versus electrostatic inter-molecular interactions. *mAbs*, *7*(1), 212–230.

Nielsen, S. C., Yang, F., Jackson, K. J., Hoh, R. A., Röltgen, K., Jean, G. H., Stevens, B. A., Lee, J.-Y., Rustagi, A., Rogers, A. J., Powell, A. E., Hunter, M., Najeeb, J., Otrelo-Cardoso, A. R., Yost, K. E., Daniel, B., Nadeau, K. C., Chang, H. Y., Satpathy, A. T., ... Boyd, S. D. (2020). Human b cell clonal expansion and convergent antibody responses to SARS-CoV-2. *Cell Host & Microbe*, *28*(4), 516–525.e5.

Nilvebrant, J., & Rockberg, J. An introduction to epitope mapping (J. Rockberg & J. Nilvebrant, Eds.). In: *Epitope mapping protocols* (J. Rockberg & J. Nilvebrant, Eds.). Ed. by Rockberg, J., & Nilvebrant, J. Methods in Molecular Biology. New York, NY: Springer, 2018, pp. 1–10.

Nimrod, G., Fischman, S., Austin, M., Herman, A., Keyes, F., Leiderman, O., Hargreaves, D., Strajbl, M., Breed, J., Klompus, S., Minton, K., Spooner, J., Buchanan, A., Vaughan, T. J., & Ofran, Y. (2018). Computational design of epitope-specific functional antibodies. *Cell Reports*, *25*(8), 2121–2131.e5.

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

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

Nutt, S. L., Hodgkin, P. D., Tarlinton, D. M., & Corcoran, L. M. (2015). The generation of antibody-secreting plasma cells [Number: 3 Publisher: Nature Publishing Group]. *Nature Reviews Immunology*, *15*(3), 160–171.

Olimpieri, P. P., 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.

Olsen, T. H., Abanades, B., Moal, I. H., & Deane, C. M. (2022a). KA-search: Rapid and exhaustive sequence identity search of known antibodies. *bioRxiv*.

Olsen, T. H., Boyles, F., & Deane, C. M. (2022b). Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, *31*(1), 141–146.

Olsen, T. H., Moal, I. H., & Deane, C. M. (2022c). AbLang: An antibody language model for completing antibody sequences. *Bioinformatics Advances*, *2*(1).

Paes, C., Ingalls, J., Kampani, K., Sulli, C., Kakkar, E., Murray, M., Kotelnikov, V., Greene, T. A., Rucker, J. B., & Doranz, B. J. (2009). Atomic-level mapping of antibody epitopes on a GPCR. *Journal of the American Chemical Society*, *131*(20), 6952–6954.

Palm, A.-K. E., & Henry, C. (2019). Remembrance of things past: Long-term b cell memory after infection and vaccination. *Frontiers in Immunology*, *10*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, (12), 2825–2830.

Peng, H.-P., Lee, K. H., Jian, J.-W., & Yang, A.-S. (2014). Origins of specificity and affinity in antibody-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(26), E2656–2665.

Piccoli, L., Park, Y.-J., Tortorici, M. A., Czudnochowski, N., Walls, A. C., Beltramello, M., Silacci-Fregni, C., Pinto, D., Rosen, L. E., Bowen, J. E., Acton, O. J., Jaconi, S., Guarino, B., Minola, A., Zatta, F., Sprugasci, N., Bassi, J., Peter, A., De Marco, A., ... Veesler, D. (2020). Mapping neutralizing and immunodominant sites on the SARS-CoV-2 spike receptor-binding domain by structure-guided high-resolution serology. *Cell*, *183*(4), 1024–1042.e21.

Pinto, D., Park, Y.-J., Beltramello, M., Walls, A. C., Tortorici, M. A., Bianchi, S., Jaconi, S., Culap, K., Zatta, F., De Marco, A., Peter, A., Guarino, B., Spreafico, R., Cameroni, E., Case, J. B., Chen, R. E., Havenar-Daughton, C., Snell, G., Telenti, A., … Corti, D. (2020a). Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. *Nature*, *583*(7815), 290–295.

Pinto, D., Park, Y.-J., Beltramello, M., Walls, A. C., Tortorici, M. A., Bianchi, S., Jaconi, S., Culap, K., Zatta, F., De Marco, A., Peter, A., Guarino, B., Spreafico, R., Cameroni, E., Case, J. B., Chen, R. E., Havenar-Daughton, C., Snell, G., Telenti, A., … Corti, D. (2020b, April 9). Structural and functional analysis of a potent sarbecovirus neutralizing antibody.

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

Prihoda, D., Maamary, J., Waight, A., Juan, V., Fayadat-Dilman, L., Svozil, D., & Bitton, D. A. (2022). BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *mAbs*, *14*(1), 2020203.

Pundir, S., Martin, M. J., & O'Donovan, C. UniProt protein knowledgebase (C. H. Wu, C. N. Arighi, & K. E. Ross, Eds.). In: *Protein bioinformatics: From protein modifications and networks to proteomics* (C. H. Wu, C. N. Arighi, & K. E. Ross, Eds.). Ed. by Wu, C. H., Arighi, C. N., & Ross, K. E. Methods in Molecular Biology. Springer, 2017, pp. 41–55.

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

Rapp, M., Guo, Y., Reddem, E. R., Yu, J., Liu, L., Wang, P., Cerutti, G., Katsamba, P., Bimela, J. S., Bahna, F. A., Mannepalli, S. M., Zhang, B., Kwong, P. D., Huang, Y., Ho, D. D., Shapiro, L., & Sheng, Z. (2021). Modular basis for potent SARS-CoV-2 neutralization by a prevalent VH1-2-derived antibody class. *Cell Reports*, *35*(1).

Raybould, M. I. J., Kovaltsuk, A., Marks, C., & Deane, C. M. (2021a). CoV-AbDab: The coronavirus antibody database. *Bioinformatics*, *37*(5), 734–735.

Raybould, M. I. J., Marks, C., Kovaltsuk, A., Lewis, A. P., Shi, J., & Deane, C. M. (2021b). Public baseline and shared response structures support the theory of antibody repertoire functional commonality. *PLOS Computational Biology*, *17*(3), e1008781.

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

Raybould, M. I. J., Rees, A. R., & Deane, C. M. (2021c). Current strategies for detecting functional convergence across b-cell receptor repertoires. *mAbs*, *13*(1), 1996732.

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

Rees, A. R. (2020). Understanding the human antibody repertoire. *mAbs*, *12*(1), 1729683.

Regep, C., Georges, G., Shi, J., Popovic, B., & Deane, C. M. (2017). The h3 loop of antibodies shows unique structural characteristics. *Proteins: Structure, Function, and Bioinformatics*, *85*(7), 1311–1318.

Repecka, D., Jauniskis, V., Karpus, L., Rembeza, E., Rokaitis, I., Zrimec, J., Poviloniene, S., Laurynenas, A., Viknander, S., Abuajwa, W., Savolainen, O., Meskys, R., Engqvist, M. K. M., & Zelezniak, A. (2021). Expanding functional protein sequence spaces using generative adversarial networks. *Nature Machine Intelligence*, *3*(4), 324–333.

Richardson, E, Binter, Kosmac, M, Ghraichy, M, von Niederhausern, V, Kovaltsuk, A, Galson, J, Trück, J, Kelly, D., Deane, C., Kellam, P, & Watson, S. (2022). Characterisation of the immune repertoire of a humanised transgenic mouse through immunophenotyping and high-throughput sequencing. *bioRxiv*.

Richardson, E., Galson, J. D., Kellam, P., Kelly, D. F., Smith, S. E., Palser, A., Watson, S., & Deane, C. M. (2021). A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-pertussis toxoid antibodies. *mAbs*, *13*(1), 1869406.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, *118*(15), e2016239118.

Rizk, S. S., Kouadio, J.-L. K., Szymborska, A., Duguid, E. M., Mukherjee, S., Zheng, J., Clevenger, C. V., & Kossiakoff, A. A. (2015). Engineering synthetic antibody binders for allosteric inhibition of prolactin receptor signaling. *Cell Communication and Signaling : CCS*, *13*, 1.

Robbiani, D. F., Gaebler, C., Muecksch, F., Lorenzi, J. C. C., Wang, Z., Cho, A., Agudelo, M., Barnes, C. O., Gazumyan, A., Finkin, S., Hägglöf, T., Oliveira, T. Y., Viant, C., Hurley, A., Hoffmann, H.-H., Millard, K. G., Kost, R. G., Cipolla, M., Gordon, K., ... Nussenzweig, M. C. (2020). Convergent antibody responses to SARS-CoV-2 in convalescent individuals. *Nature*, *584*(7821), 437–442.

Rogers, T. F., Zhao, F., Huang, D., Beutler, N., Burns, A., He, W.-t., Limbo, O., Smith, C., Song, G., Woehl, J., Yang, L., Abbott, R. K., Callaghan, S., Garcia, E., Hurtado, J., Parren, M., Peng, L., Ramirez, S., Ricketts, J., ... Burton, D. R. (2020). Isolation of potent SARS-CoV-2 neutralizing antibodies and protection from disease in a small animal model. *Science*, *369*(6506), 956–963.

Rosenfeld, A. M., Meng, W., Luning Prak, E. T., & Hershberg, U. (2017). ImmuneDB: A system for the analysis and exploration of high-throughput adaptive immune receptor sequencing data. *Bioinformatics*, *33*(2), 292–293.

Ruffolo, J. A., Chu, L.-S., Mahajan, S. P., & Gray, J. J. (2022a). Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *bioRxiv*.

Ruffolo, J. A., Sulam, J., & Gray, J. J. (2022b). Antibody structure prediction using interpretable deep learning. *Patterns*, *3*(2), 100406.

Rujas, E., Kucharska, I., Tan, Y. Z., Benlekbir, S., Cui, H., Zhao, T., Wasney, G. A., Budylowski, P., Guvenc, F., Newton, J. C., Sicard, T., Semesi, A., Muthuraman, K., Nouanesengsy, A., Aschner, C. B., Prieto, K., Bueler, S. A., Youssef, S., Liao-Chan, S., ... Julien, J.-P. (2021). Multivalency transforms SARS-CoV-2 antibodies into ultrapotent neutralizers. *Nature Communications*, *12*(1), 3661.

Schmitz, A. J., Turner, J. S., Liu, Z., Zhou, J. Q., Aziati, I. D., Chen, R. E., Joshi, A., Bricker, T. L., Darling, T. L., Adelsberg, D. C., Altomare, C. G., Alsoussi, W. B., Case, J. B., VanBlargan, L. A., Lei, T., Thapa, M., Amanat, F., Jeevan, T., Fabrizio, T., ... Ellebedy, A. H. (2021). A vaccine-induced public antibody protects against SARS-CoV-2 and emerging variants. *Immunity*, *54*(9), 2159–2166.e6.

Schneider, C., Buchanan, A., Taddese, B., & Deane, C. M. (2022a). DLAB: Deep learning methods for structure-based virtual screening of antibodies. *Bioinformatics*, *38*(2), 377–383.

Schneider, C., Raybould, M. I. J., & Deane, C. M. (2022b). SAbDab in the age of biotherapeutics: Updates including SAbDab-nano, the nanobody structure tracker. *Nucleic Acids Research*, *50*, D1368–D1372.

Schritt, D., Li, S., Rozewicki, J., Katoh, K., Yamashita, K., Volkmuth, W., Cavet, G., & Standley, D. M. (2019). Repertoire builder: High-throughput structural modeling of b and t cell receptors. *Molecular Systems Design & Engineering*, *4*(4), 761–768.

Schultheiß, C., Paschold, L., Simnica, D., Mohme, M., Willscher, E., von Wenserski, L., Scholz, R., Wieters, I., Dahlke, C., Tolosa, E., Sedding, D. G., Ciesek, S., Addo, M., & Binder, M. (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), 442–455.e4.

Sela-Culang, I., Kunik, V., & Ofran, Y. (2013). The structural basis of antibody-antigen recognition. *Frontiers in Immunology*, *4*, 302.

Setliff, I., McDonnell, W. J., Raju, N., Bombardi, R. G., Murji, A. A., Scheepers, C., Ziki, R., Mynhardt, C., Shepherd, B. E., Mamchak, A. A., Garrett, N., Karim, S. A., Mallal, S. A., Crowe, J. E., Morris, L., & Georgiev, I. S. (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.e6.

Setliff, I., Shiakolas, A. R., Pilewski, K. A., Murji, A. A., Mapengo, R. E., Janowska, K., Richardson, S., Oosthuysen, C., Raju, N., Ronsard, L., Kanekiyo, M., Qin, J. S., Kramer, K. J., Greenplate, A. R., McDonnell, W. J., Graham, B. S., Connors, M., Lingwood, D., Acharya, P., ... Georgiev, I. S. (2019). High-

throughput mapping of b cell receptor sequences to antigen specificity. *Cell*, *179*(7), 1636–1646.e15.

Seydoux, E., Homad, L. J., MacCamy, A. J., Parks, K. R., Hurlburt, N. K., Jennewein, M. F., Akins, N. R., Stuart, A. B., Wan, Y.-H., Feng, J., Whaley, R. E., Singh, S., Boeckh, M., Cohen, K. W., McElrath, M. J., Englund, J. A., Chu, H. Y., Pancera, M., McGuire, A. T., & Stamatatos, L. (2020). Analysis of a SARS-CoV-2-infected individual reveals development of potent neutralizing antibodies with limited somatic mutation. *Immunity*, *53*(1), 98–105.e5.

Shehata, L., Maurer, D. P., Wec, A. Z., Lilov, A., Champney, E., Sun, T., Archambault, K., Burnina, I., Lynaugh, H., Zhi, X., Xu, Y., & Walker, L. M. (2019). Affinity maturation enhances antibody specificity but compromises conformational stability. *Cell Reports*, *28*(13), 3300–3308.e4.

Shi, R., Shan, C., Duan, X., Chen, Z., Liu, P., Song, J., Song, T., Bi, X., Han, C., Wu, L., Gao, G., Hu, X., Zhang, Y., Tong, Z., Huang, W., Liu, W. J., Wu, G., Zhang, B., Wang, L., ... Yan, J. (2020). A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. *Nature*, *584*(7819), 120–124.

Shi, Z., Zhang, Q., Yan, H., Yang, Y., Wang, P., Zhang, Y., Deng, Z., Yu, M., Zhou, W., Wang, Q., Yang, X., Mo, X., Zhang, C., Huang, J., Dai, H., Sun, B., Zhao, Y., Zhang, L., Yang, Y.-G., & Qiu, X. (2019). More than one antibody of individual b cells revealed by single-cell immune profiling. *Cell Discovery*, *5*, 64.

Shiakolas, A. R., Kramer, K. J., Johnson, N. V., Wall, S. C., Suryadevara, N., Wrapp, D., Periasamy, S., Pilewski, K. A., Raju, N., Nargi, R., Sutton, R. E., Walker, L. M., Setliff, I., Crowe, J. E., Bukreyev, A., Carnahan, R. H., McLellan, J. S., & Georgiev, I. S. (2022). Efficient discovery of SARS-CoV-2-neutralizing antibodies via b cell receptor sequencing and ligand blocking [Number: 8 Publisher: Nature Publishing Group]. *Nature Biotechnology*, *40*(8), 1270–1275.

Snapkov, I., Chernigovskaya, M., Sinitcyn, P., Lê Quý, K., Nyman, T. A., & Greiff, V. (2022). Progress and challenges in mass spectrometry-based analysis of antibody repertoires. *Trends in Biotechnology*, *40*(4), 463–481.

Sokal, A., Chappert, P., Barba-Spaeth, G., Roeser, A., Fourati, S., Azzaoui, I., Vandenberghe, A., Fernandez, I., Meola, A., Bouvier-Alias, M., Crickx, E., Beldi-Ferchiou, A., Hue, S., Languille, L., Michel, M., Baloul, S., Noizat-Pirenne, F., Luka, M., Mégret, J., ... Mahévas, M. (2021). Maturation and persistence of the anti-SARS-CoV-2 memory b cell response. *Cell*, *184*(5), 1201–1213.e14.

Sormanni, P., Amery, L., Ekizoglou, S., Vendruscolo, M., & Popovic, B. (2017). Rapid and accurate in silico solubility screening of a monoclonal antibody library [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *7*(1), 8200.

Soto, C., Bombardi, R. G., Branchizio, A., Kose, N., Matta, P., Sevy, A. M., Sinkovits, R. S., Gilchuk, P., Finn, J. A., & Crowe, J. E. (2019). High frequency of shared clonotypes in human b cell receptor repertoires. *Nature*, *566*(7744), 398–402.

Spisak, N., Walczak, A. M., & Mora, T. (2020). Learning the heterogeneous hypermutation landscape of immunoglobulins from high-throughput repertoire data. *Nucleic Acids Research*, *48*(19), 10702–10712.

Stein, A., & Kortemme, T. (2013). Improvements to robotics-inspired conformational sampling in rosetta. *PloS One*, *8*(5), e63090.

Strokach, A., & Kim, P. M. (2022). Deep generative modeling for protein design. *Current Opinion in Structural Biology*, *72*, 226–236.

Sun, B., Ramberger, M., O'Connor, K. C., Bashford-Rogers, R. J. M., & Irani, S. R. (2020). The b cell immunobiology that underlies CNS autoantibody-mediated diseases. *Nature Reviews Neurology*, *16*(9), 481–492.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., & the UniProt Consortium. (2015). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, *31*(6), 926–932.

Sydow, J. F., Lipsmeier, F., Larraillet, V., Hilger, M., Mautz, B., Mølhøj, M., Kuentzer, J., Klostermann, S., Schoch, J., Voelger, H. R., Regula, J. T., Cramer, P., Papadimitriou, A., & Kettenberger, H. (2014). Structure-based prediction of asparagine and aspartate degradation sites in antibody variable regions. *PLoS ONE*, *9*(6), e100736.

Tang, C., Krantsevich, A., & MacCarthy, T. (2022). Deep learning model of somatic hypermutation reveals importance of sequence context beyond hotspot targeting. *iScience*, *25*(1), 103668.

Tareen, A., & Kinney, J. B. (2020). Logomaker: Beautiful sequence logos in python. *Bioinformatics*, *36*(7), 2272–2274.

Tay, M. Z., Wiehe, K., & Pollara, J. (2019). Antibody-dependent cellular phagocytosis in antiviral immune responses. *Frontiers in Immunology*, *10*.

ter Meulen, J., van den Brink, E. N., Poon, L. L. M., Marissen, W. E., Leung, C. S. W., Cox, F., Cheung, C. Y., Bakker, A. Q., Bogaards, J. A., van Deventer, E., Preiser, W., Doerr, H. W., Chow, V. T., de Kruif, J., Peiris, J. S. M., & Goudsmit, J. (2006). Human monoclonal antibody combination against SARS coronavirus: Synergy and coverage of escape mutants (D. R. Burton, Ed.). *PLoS Medicine*, *3*(7), e237.

Thomson, E. C., Rosen, L. E., Shepherd, J. G., Spreafico, R., da Silva Filipe, A., Wojcechowskyj, J. A., Davis, C., Piccoli, L., Pascall, D. J., Dillen, J., Lytras, S., Czudnochowski, N., Shah, R., Meury, M., Jesudason, N., De Marco, A., Li, K., Bassi, J., O'Toole, A., ... Snell, G. (2021). Circulating SARS-CoV-2 spike n439k variants maintain fitness while evading antibody-mediated immunity. *Cell*, *184*(5), 1171–1187.e20.

Tortorici, M. A., Beltramello, M., Lempp, F. A., Pinto, D., Dang, H. V., Rosen, L. E., McCallum, M., Bowen, J., Minola, A., Jaconi, S., Zatta, F., De Marco, A., Guarino, B., Bianchi, S., Lauron, E. J., Tucker, H., Zhou, J., Peter, A., Havenar-Daughton, C., ... Veesler, D. (2020). Ultrapotent human antibodies protect against SARS-CoV-2 challenge via multiple mechanisms. *Science*, *370*(6519), 950–957.

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

Valldorf, B., Hinz, S. C., Russo, G., Pekar, L., Mohr, L., Klemm, J., Doerner, A., Krah, S., Hust, M., & Zielonka, S. (2022). Antibody display technologies: Selecting the cream of the crop [Publisher: De Gruyter]. *Biological Chemistry*, *403*(5), 455–477.

Vatsa, S. (2022). In silico prediction of post-translational modifications in therapeutic antibodies. *mAbs*, *14*(1), 2023938.

Walker, L. M., Shiakolas, A. R., Venkat, R., Liu, Z. A., Wall, S., Raju, N., Pilewski, K. A., Setliff, I., Murji, A. A., Gillespie, R., Makoah, N. A., Kanekiyo, M., Connors, M., Morris, L., & Georgiev, I. S. (2022). High-throughput b cell epitope determination by next-generation sequencing. *Frontiers in Immunology*, *13*, 855772.

Wang, B., Gallolu Kankanamalage, S., Dong, J., & Liu, Y. (2021a). Optimization of therapeutic antibodies. *Antibody Therapeutics*, *4*(1), 45–54.

Wang, Q., Guo, Y., Iketani, S., Nair, M. S., Li, Z., Mohri, H., Wang, M., Yu, J., Bowen, A. D., Chang, J. Y., Shah, J. G., Nguyen, N., Chen, Z., Meyers, K., Yin, M. T., Sobieszczyk, M. E., Sheng, Z., Huang, Y., Liu, L., & Ho, D. D. (2022). Antibody evasion by SARS-CoV-2 omicron subvariants BA.2.12.1, BA.4 and BA.5. *Nature*, *608*(7923), 603–608.

Wang, Q., Zeng, H., Zhu, Y., Wang, M., Zhang, Y., Yang, X., Tang, H., Li, H., Chen, Y., Ma, C., Lan, C., Liu, B., Yang, W., Yu, X., & Zhang, Z. (2021b). Dual UMIs and dual barcodes with minimal PCR amplification removes artifacts and acquires accurate antibody repertoire. *Frontiers in Immunology*, *12*, 778298.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., ... Baker, D. (2022). Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv*.

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

Wen, X., Mousa, J. J., Bates, J. T., Lamb, R. A., Crowe, J. E., & Jardetzky, T. S. (2017). Structural basis for antibody cross-neutralization of respiratory syncytial virus and human metapneumovirus. *Nature Microbiology*, *2*, 16272.

Wollacott, A. M., Robinson, L. N., Ramakrishnan, B., Tissire, H., Viswanathan, K., Shriver, Z., & Babcock, G. J. (2019). Structural prediction of antibody-APRIL complexes by computational docking constrained by antigen saturation mutagenesis library data. *Journal of Molecular Recognition*, *32*(7), e2778.

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

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

Wong, W. K., Robinson, S. A., Bujotzek, A., Georges, G., Lewis, A. P., Shi, J., Snowden, J., Taddese, B., & Deane, C. M. (2021). Ab-ligity: Identifying sequence-dissimilar antibodies that bind to the same epitope. *mAbs*, *13*(1), 1873478.

Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C.-L., Abiona, O., Graham, B. S., & McLellan, J. S. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, *367*(6483), 1260–1263.

Wu, N. C., Yuan, M., Bangaru, S., Huang, D., Zhu, X., Lee, C.-C. D., Turner, H. L., Peng, L., Yang, L., Burton, D. R., Nemazee, D., Ward, A. B., & Wilson, I. A. (2020a). A natural mutation between SARS-CoV-2 and SARS-CoV determines neutralization by a cross-reactive antibody. *PLoS Pathogens*, *16*(12), e1009089.

Wu, N. C., Yuan, M., Liu, H., Lee, C.-C. D., Zhu, X., Bangaru, S., Torres, J. L., Caniels, T. G., Brouwer, P. J., van Gils, M. J., Sanders, R. W., Ward, A. B., & Wilson, I. A. (2020b). An alternative binding mode of IGHV3-53 antibodies to the SARS-CoV-2 receptor binding domain. *Cell Reports*, *33*(3), 108274.

Wu, Y., Wang, F., Shen, C., Peng, W., Li, D., Zhao, C., Li, Z., Li, S., Bi, Y., Yang, Y., Gong, Y., Xiao, H., Fan, Z., Tan, S., Wu, G., Tan, W., Lu, X., Fan, C., Wang, Q., ... Liu, L. (2020c). A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. *Science*, *368*(6496), 1274–1278.

Xiang, H., Zhao, Y., Li, X., Liu, P., Wang, L., Wang, M., Tian, L., Sun, H.-X., Zhang, W., Xu, Z., Ye, B., Yuan, X., Wang, P., Zhang, N., Gong, Y., Bian, C., Wang, Z., Yu, L., Yan, J., ... Yang, P. (2022). Landscapes and dynamic diversifications of b-cell receptor repertoires in COVID-19 patients. *Human Immunology*, *83*(2), 119–129.

Xiang, Y., Nambulli, S., Xiao, Z., Liu, H., Sang, Z., Duprex, W. P., Schneidman-Duhovny, D., Zhang, C., & Shi, Y. (2020). Versatile and multivalent nanobodies efficiently neutralize SARS-CoV-2. *Science*, *370*(6523), 1479–1484.

Xu, J. L., & Davis, M. M. (2000). Diversity in the CDR3 region of VH is sufficient for most antibody specificities. *Immunity*, *13*(1), 37–45.

Yaari, G., Vander Heiden, J. A., Uduman, M., Gadala-Maria, D., Gupta, N., Stern, J. N. H., O'Connor, K. C., Hafler, D. A., Laserson, U., Vigneault, F., & Kleinstein, S. H. (2013). Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Frontiers in Immunology*, *4*.

Yan, Q., Huang, M., Lewis, M. J., & Hu, P. (2018). Structure based prediction of asparagine deamidation propensity in monoclonal antibodies. *mAbs*, *10*(6), 901–912.

Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., & Zhou, Q. (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science*, *367*(6485), 1444–1448.

Yang, J., Petitjean, S. J. L., Koehler, M., Zhang, Q., Dumitru, A. C., Chen, W., Derclaye, S., Vincent, S. P., Soumillion, P., & Alsteens, D. (2020). Molecular interaction and inhibition of SARS-CoV-2 binding to the ACE2 receptor. *Nature Communications*, *11*, 4541.

Yang, X., Zhang, X., Zhao, X., Yuan, M., Zhang, K., Dai, J., Guan, X., Qiu, H.-J., & Li, Y. (2022). Antibody-dependent enhancement: evil antibodies favorable for viral infections. *Viruses*, *14*(8), 1739.

Yang, X., Zhu, Y., Chen, S., Zeng, H., Guan, J., Wang, Q., Lan, C., Sun, D., Yu, X., & Zhang, Z. (2021). Novel allele detection tool benchmark and application with antibody repertoire sequencing dataset. *Frontiers in Immunology*, *12*, 739179.

Yao, H., Cai, H., Li, T., Zhou, B., Qin, W., Lavillette, D., & Li, D. (2021). A high-affinity RBD-targeting nanobody improves fusion partner's potency against SARS-CoV-2. *PLOS Pathogens*, *17*(3), e1009328.

Ye, J., Ma, N., Madden, T. L., & Ostell, J. M. (2013). IgBLAST: An immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research*, *41*, W34–40.

Yin, R., Guest, J. D., Taherzadeh, G., Gowthaman, R., Mittra, I., Quackenbush, J., & Pierce, B. G. (2021). Structural and energetic profiling of SARS-CoV-2 receptor binding domain antibody recognition and the impact of circulating variants. *PLOS Computational Biology*, *17*(9), e1009380.

Yuan, M., Huang, D., Lee, C.-C. D., Wu, N. C., Jackson, A. M., Zhu, X., Liu, H., Peng, L., van Gils, M. J., Sanders, R. W., Burton, D. R., Reincke, S. M., Prüss, H., Kreye, J., Nemazee, D., Ward, A. B., & Wilson, I. A. (2021). Structural and functional ramifications of antigenic drift in recent SARS-CoV-2 variants. *Science*, *373*(6556), 818–823.

Yuan, M., Liu, H., Wu, N. C., Lee, C.-C. D., Zhu, X., Zhao, F., Huang, D., Yu, W., Hua, Y., Tien, H., Rogers, T. F., Landais, E., Sok, D., Jardine, J. G., Burton, D. R., & Wilson, I. A. (2020). Structural basis of a shared antibody response to SARS-CoV-2 [Publisher: American Association for the Advancement of Science]. *Science*, *369*(6507), 1119–1123.

Zahavi, D., AlDeghaither, D., O'Connell, A., & Weiner, L. M. (2018). Enhancing antibody-dependent cell-mediated cytotoxicity: A strategy for improving antibody-based immunotherapy. *Antibody Therapeutics*, *1*(1), 7–12.

Zhang, F., Gan, R., Zhen, Z., Hu, X., Li, X., Zhou, F., Liu, Y., Chen, C., Xie, S., Zhang, B., Wu, X., & Huang, Z. (2020). Adaptive immune responses to SARS-CoV-2 infection in severe versus mild individuals. *Signal Transduction and Targeted Therapy*, *5*(1), 156.

Zhao, P., Cao, J., Zhao, L.-J., Qin, Z.-L., Ke, J.-S., Pan, W., Ren, H., Yu, J.-G., & Qi, Z.-T. (2005). Immune responses against SARS-coronavirus nucleocapsid protein induced by DNA vaccine. *Virology*, *331*(1), 128–135.

Zhou, D., Duyvesteyn, H. M. E., Chen, C.-P., Huang, C.-G., Chen, T.-H., Shih, S.-R., Lin, Y.-C., Cheng, C.-Y., Cheng, S.-H., Huang, Y.-C., Lin, T.-Y., Ma, C., Huo, J., Carrique, L., Malinauskas, T., Ruza, R. R., Shah, P. N. M., Tan, T. K., Rijal, P., … Huang, K.-Y. A. (2020). Structural basis for the neutralization of SARS-CoV-2 by an antibody from a convalescent patient. *Nature Structural & Molecular Biology*, *27*(10), 950–958.

Zhou, J. Q., & Kleinstein, S. H. (2020). Position-dependent differential targeting of somatic hypermutation. *The Journal of Immunology*, *205*(12), 3468–3479.

Zhu, Y., Wang, S. S., Zhou, Z. S., & Ho, M. (2022). The emergence of AntibodyPlus: The future trend of antibody-based therapeutics. *Antibody Therapeutics*, *5*(4), 280–287.

Zost, S. J., Gilchuk, P., Chen, R. E., Case, J. B., Reidy, J. X., Trivette, A., Nargi, R. S., Sutton, R. E., Suryadevara, N., Chen, E. C., Binshtein, E., Shrihari, S., Ostrowski, M., Chu, H. Y., Didier, J. E., MacRenaris, K. W., Jones, T., Day, S., Myers, L., ... Crowe, J. E. (2020). Rapid isolation and profiling of a diverse panel of human monoclonal antibodies targeting the SARS-CoV-2 spike protein. *Nature Medicine*, *26*(9), 1422–1427.

Appendices

Appendix: Tables

Antibody	PDB Code	Citation	Epitope Cluster
B38	7bz5	Wu et al. 2020c	Neck
Bamlanivimab	7kmg	Jones et al. 2021	*Left Shoulder
BD-236	7chb	Du et al. 2020	Neck
BD-368-2	7chc	Du et al. 2020	Left Shoulder
BD-604	7chf	Du et al. 2020	Neck
BD-629	7ch5	Du et al. 2020	Neck
C102	7k8m	Robbiani et al. 2020	*Neck
C1A-B3	7kfw	Clark et al. 2021	*Neck
C1A-B12	7kfv	Clark et al. 2021	*Neck
C1A-C2	7kfx	Clark et al. 2021	*Neck
C1A-F10	7kfy	Clark et al. 2021	*Neck
CB6	7c01	Shi et al. 2020	Neck
CC12.1	6xc2	Rogers et al. 2020	Neck
CC12.3	6xc4	Rogers et al. 2020	Neck
COVA1-16	7jmw	Brouwer et al. 2020; Liu et al. 2020a	Left Flank
COVA2-04	7jmo	Brouwer et al. 2020; Wu et al. 2020b	Neck
COVA2-39	7jmp	Brouwer et al. 2020; Wu et al. 2020b	Right Shoulder
COVOX-45	7bel	Dejnirattisai et al. 2021	Right Flank
COVOX-75	7beo	Dejnirattisai et al. 2021	Right Shoulder
COVOX-88	7bel	Dejnirattisai et al. 2021	Left Shoulder
COVOX-150	7bei	Dejnirattisai et al. 2021	Neck
COVOX-158	7bej	Dejnirattisai et al. 2021	Neck
COVOX-253	7ben	Dejnirattisai et al. 2021	Left Shoulder
COVOX-253H55L	7beo	Dejnirattisai et al. 2021	Left Shoulder
COVOX-269	7bem	Dejnirattisai et al. 2021	Neck
COVOX-316	7beh	Dejnirattisai et al. 2021	Left Shoulder
COVOX-384	7bep	Dejnirattisai et al. 2021	Left Shoulder
CR3022	7jn5	ter Meulen et al. 2006; Wu et al. 2020b	Left Flank
CV07-250	6xkq	Kreye et al. 2020	Right Shoulder
CV07-270	6xkp	Kreye et al. 2020	Left Shoulder
CV30	6xe1	Hurlburt et al. 2020; Seydoux et al. 2020	Neck
EY6A	6zer	Zhou et al. 2020	Left Flank

Table continued on next page.

Table continued from previous page.

Antibody	PDB Code	Citation	Epitope Cluster
Fab_15033-7	7klh	Miersch et al. 2021	*Right Shoulder
Fab-52	7k9z	Rujas et al. 2021	*Left Shoulder
Fab-298	7k9z	Rujas et al. 2021	*Right Shoulder
Fab2-15	7l5b	Liu et al. 2020b; Rapp et al. 2021	*Shoulder
LY-CoV481	7kmi	Jones et al. 2021	Neck
LY-CoV488	7kmh	Jones et al. 2021	Neck
MW06	7dpm	**jiangCharacterizationMW06Human**	*Left Flank
P4A1	7cjf	Guo et al. 2021	Neck
P2B-2F6	7bwj	Ju et al. 2020	Left Shoulder
P2C-1F11	7cdi	Ge et al. 2021	*Neck
P2C-1A3	7cdj	Ge et al. 2021	*Shoulder
Regdanvimab	7cm4	Kim et al. 2021	*Right Shoulder
S2H14	7jx3	Piccoli et al. 2020	Right Shoulder
S304	7l0n	Piccoli et al. 2020; Pinto et al. 2020a	Left Flank
S309	7jx3	Piccoli et al. 2020; Pinto et al. 2020a	Right Flank
STE90-C11	7b30	Bertoglio et al. 2021	*Neck

Table A1: Antibodies with a solved crystal structures in complex with a SARS-CoV-2 antigen. Antibodies were assigned to their nearest epitope cluster, as defined by Dejnirattisai et al. 2021. Antibodies whose epitope cluster is labelled with a star (*) were not included in the previous analysis by Dejnirattisai et al. 2021. The new cluster label 'Shoulder' indicates that antibody's epitope lies between the left and right shoulder clusters.

Nanobody	PDB Code	Citation	Epitope Cluster
H11-D4	6yz5	Huo et al. 2020	*Left Shoulder
H11-H4	6zbp	Huo et al. 2020	*Left Shoulder
MR17	7c8w	Li et al. 2021b; Yao et al. 2021	*Shoulder
MR17_K99Y	7can	Li et al. 2021b	*Shoulder
Nb20	7jvb	Xiang et al. 2020	*Shoulder
SR4	7c8v	Li et al. 2021b	*Left Shoulder
SR31	7d2z	Li et al. 2021b	*Left Shoulder
VHH-72	6waq	Wrapp et al. 2020	*Left Flank
VHH_E	7kn5	Koenig et al. 2021	Left Shoulder
VHH_U	7kn5	Koenig et al. 2021	*Left Flank
VHH_V	7kn6	Koenig et al. 2021	*Left Flank
VHH_W	7kn7	Koenig et al. 2021	*Left Flank

Table A2: Nanobodies with a solved crystal structures in complex with a SARS-CoV-2 antigen. nanobodies were assigned to their nearest epitope cluster, as defined by Dejnirattisai et al. 2021. Nanobodies whose epitope cluster is labelled with a star (*) were not included in the previous analysis by Dejnirattisai et al. 2021. The new cluster label 'Shoulder' indicates that antibody's epitope lies between the left and right shoulder clusters.

Name	Van der Waals	H Bond	Weak H Bond	Ionic	Aromatic	Hydrophobic	Carbonyl	Polar	Weak Polar
B38	9	20	13	0	3	34	1	33	13
Bamlanivimab	5	11	3	7	10	59	1	10	3
BD-236	9	13	13	1	1	27	0	22	16
BD-368-2	5	14	13	0	5	26	1	25	16
BD-604	6	15	14	0	0	27	1	23	9
BD-629	5	13	9	0	7	25	1	23	15
C102	9	13	6	0	16	33	0	20	18
C1A-B3	9	16	14	1	4	26	0	26	12
C1A-B12	5	17	12	1	4	36	1	23	14
C1A-C2	13	22	15	1	8	36	1	29	14
C1A-F10	11	17	11	1	6	33	1	23	16
CB6	11	15	10	2	7	38	0	19	14
CC12.1	19	22	18	2	7	38	0	30	26
CC12.3	7	11	15	0	8	33	0	18	13
COVA1-16	9	8	8	0	0	23	1	12	17
COVA2-04	10	14	14	1	4	31	0	20	12
COVA2-39	8	7	8	0	10	48	0	13	18
COVOX-150	11	21	12	4	5	39	2	33	17
COVOX-158	10	16	15	0	5	19	0	21	13
COVOX-253	3	8	5	0	8	36	1	9	4
COVOX-253H55L	8	7	6	0	11	39	1	9	6
COVOX-269	17	20	13	2	6	36	2	30	17
COVOX-316	4	4	3	0	11	52	0	11	5
COVOX-384	4	7	3	4	3	38	0	9	5
COVOX-45	6	16	11	8	5	39	1	20	19
COVOX-75	13	9	9	0	1	36	0	11	10
COVOX-88	10	9	1	4	10	37	0	12	13

Table continued on next page

Table continued from previous page

Name	Van der Waals	H Bond	Weak H Bond	Ionic	Aromatic	Hydrophobic	Carbonyl	Polar	Weak Polar
CR3022	9	13	13	10	0	35	0	14	11
CV07-250	3	14	9	0	11	34	0	16	9
CV07-270	9	12	5	9	6	37	1	15	9
CV30	10	14	11	0	0	23	0	22	11
EY6A	11	7	8	1	3	34	1	11	15
Fab_15033-7	8	10	9	4	11	36	0	13	15
Fab-298	1	6	2	0	1	20	0	10	6
Fab-52	6	3	5	2	2	34	0	5	10
Fab2-15	6	4	5	0	10	27	0	6	16
LY-CoV481	14	13	15	0	2	21	0	21	13
LY-CoV488	8	19	9	3	3	22	1	29	15
MW06	7	9	15	0	3	17	4	14	18
P2B-2F6	4	4	3	2	2	16	0	7	8
P2C-1A3	16	8	8	0	13	52	0	11	15
P2C-1F11	8	11	11	0	2	27	0	21	14
P4A1	16	20	13	1	4	33	1	28	16
Regdanvimab	9	8	5	7	15	57	0	11	3
S2H14	8	13	13	0	0	2	0	17	17
S304	9	10	12	0	3	36	1	16	12
S309	6	12	7	2	0	43	1	14	14
STE90-C11	5	14	10	1	7	38	0	22	14
Average	*8.5*	*12.3*	*9.6*	*1.7*	*5.5*	*33.1*	*0.5*	*17.9*	*12.8*

Table A3: Analysis of paratope-epitope interactions between antibodies and SARS-CoV-2 RBD in cocrystal structures, determined using Arpeggio (Jubb et al. 2017). H bond = hydrogen bond

Appendix: The allosteric modulation of complement C5 by knob domain peptides

RESEARCH ARTICLE

The allosteric modulation of complement C5 by knob domain peptides

Alex Macpherson[1,2]*, Maisem Laabei[2], Zainab Ahdash[1], Melissa A Graewert[3], James R Birtley[1], Monika-Sarah ED Schulze[1], Susan Crennell[2], Sarah A Robinson[4], Ben Holmes[1], Vladas Oleinikovas[1], Per H Nilsson[1,5,6], James Snowden[1], Victoria Ellis[1], Tom Eirik Mollnes[6,7,8], Charlotte M Deane[4], Dmitri Svergun[3], Alastair DG Lawson[1], Jean MH van den Elsen[2,9]*

[1]UCB, Slough, United Kingdom; [2]Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom; [3]European Molecular Biology Laboratory, Hamburg Unit, Hamburg, Germany; [4]Department of Statistics, University of Oxford, Oxford, United Kingdom; [5]Department of Chemistry and Biomedicine, Linnaeus University, Kalmar, Sweden; [6]Department of Immunology, Oslo University Hospital, University of Oslo, Oslo, Norway; [7]Research Laboratory, Bodø Hospital, K.G. Jebsen TREC, University of Tromsø, Tromsø, Norway; [8]Centre of Molecular Inflammation Research, Norwegian University of Science and Technology, Trondheim, Norway; [9]Centre for Therapeutic Innovation, University of Bath, Bath, United Kingdom

Abstract Bovines have evolved a subset of antibodies with ultra-long heavy chain complementarity determining regions that harbour cysteine-rich knob domains. To produce high-affinity peptides, we previously isolated autonomous 3–6 kDa knob domains from bovine antibodies. Here, we show that binding of four knob domain peptides elicits a range of effects on the clinically validated drug target complement C5. Allosteric mechanisms predominated, with one peptide selectively inhibiting C5 cleavage by the alternative pathway C5 convertase, revealing a targetable mechanistic difference between the classical and alternative pathway C5 convertases. Taking a hybrid biophysical approach, we present C5-knob domain co-crystal structures and, by solution methods, observed allosteric effects propagating >50 Å from the binding sites. This study expands the therapeutic scope of C5, presents new inhibitors, and introduces knob domains as new, low molecular weight antibody fragments, with therapeutic potential.

***For correspondence:**
Alex.MacPherson@ucb.com (AM);
bssjmhve@bath.ac.uk (JMHE)

Competing interest: See page 22

Funding: See page 22

Received: 29 September 2020
Accepted: 11 February 2021
Published: 11 February 2021

Reviewing editor: John Kuriyan, University of California, Berkeley, United States

© Copyright Macpherson et al. This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

Introduction

By the end of 2019, over 60 peptide drugs have received regulatory approval, with an estimated 400 more in active development globally (***Lau and Dunn, 2018***; ***Lee et al., 2019***). As a potential route to discover therapeutic peptides, we previously reported a method for deriving peptides from the ultra-long heavy chain complementarity determining region 3 (ul-CDRH3), which are unique to a subset of bovine antibodies (***Macpherson et al., 2020***). We have shown that knob domains, a cysteine-rich mini-domain common to all ul-CDRH3, can bind antigen autonomously when removed from the antibody scaffold (***Macpherson et al., 2020***). This allows peptide affinity maturation to be performed in vivo, harnessing the cow's immune system to produce peptides with complex stabilising networks of disulphide bonds.

For the discovery of knob domain peptides, immunisation of cattle is followed by cell sorting of B-cells using fluorescently labelled antigen. A library of antigen-specific CDRH3 sequences is created by performing a reverse transcription polymerase chain reaction (RT PCR) on the B-cell lysate,

eLife digest Antibodies are proteins produced by the immune system that can selectively bind to other molecules and modify their behaviour. Cows are highly equipped at fighting-off disease-causing microbes due to the unique shape of some of their antibodies. Unlike other jawed vertebrates, cows' antibodies contain an ultra-long loop region that contains a 'knob domain' which sticks out from the rest of the antibody. Recent research has shown that when detached, the knob domain behaves like an antibody fragment, and can independently bind to a range of different proteins.

Antibody fragments are commonly developed in the laboratory to target proteins associated with certain diseases, such as arthritis and cancer. But it was unclear whether the knob domains from cows' antibodies could also have therapeutic potential. To investigate this, Macpherson et al. studied how knob domains attach to complement C5, a protein in the inflammatory pathway which is a drug target for various diseases, including severe COVID-19.

The experiments identified various knob domains that bind to complement C5 and inhibits its activity by altering its structure or movement. Further tests studying the structure of these interactions, led to the discovery of a common mechanism by which inhibitors can modify the behaviour of this inflammatory protein.

Complement C5 is involved in numerous molecular pathways in the immune system, which means many of the drugs developed to inhibit its activity can also leave patients vulnerable to infection. However, one of the knob domains identified by Macpherson et al. was found to reduce the activity of complement C5 in some pathways, whilst leaving other pathways intact. This could potentially reduce the risk of bacterial infections which sometimes arise following treatment with these types of inhibitors.

These findings highlight a new approach for developing drug inhibitors for complement C5. Furthermore, the ability of knob domains to bind to multiple sites of complement C5 suggests that this fragment could be used to target proteins associated with other diseases.

followed by a PCR using primers specific to the conserved framework regions which flank CDRH3 (***Macpherson et al., 2020***). Upon sequencing, ul-CDRH3s are immediately evident and the knob domains can be expressed recombinantly as cleavable fusion proteins (***Macpherson et al., 2020***).

This method for the discovery of knob domain peptides was established using complement component C5, and we reported peptides which bound C5 with affinities in the pM–low nM range (***Macpherson et al., 2020***). Herein, we use these novel peptides to probe the structural and functional aspects of C5 activation.

C5 is the éminence grise of the complement cascade's druggable proteins, and the target of effective therapies for diseases with pathogenic complement dysregulation, of which paroxysmal nocturnal haemoglobinuria (***Rother et al., 2007***) and atypical haemolytic uraemic syndrome (***Nürnberger et al., 2009***) are notable examples. Six monoclonal antibodies targeting C5 have reached, or are entering, clinical trials, closely followed by C5-targeting immune evasion molecules (***Romay-Penabad et al., 2014***), aptamers (***Biesecker et al., 1999***), cyclic peptides (***Ricardo et al., 2014***), interfering RNA (***Borodovsky et al., 2014***), and small molecules (***Jendza et al., 2019***). Currently, C5 inhibitors are being trialled for the treatment of acute respiratory distress syndrome arising from severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection (***Smith et al., 2020***; ***Wilkinson et al., 2020***; ***Zelek et al., 2020***) and for the neuromuscular disease myasthenia gravis (***Albazli et al., 2020***).

C5 is the principal effector of the terminal portion of the complement cascade. At high local C3b concentrations, arising from activation of either or both of the classical (CP) and mannose binding lectin (LP) pathways, aided by the amplificatory alternative pathway (AP), C5 is cleaved into two moieties with distinct biological functions. Cleavage is performed by two convertases: C4bC2aC3b, formed in response to CP or LP activation (***Takata et al., 1987***) (henceforth the CP C5 convertase), and C3bBbC3b, formed in response to AP activation (***DiScipio, 1981***) (henceforth the AP C5 convertase). Although the constitutive components of the C5 convertases differ, they are thought to be

mechanistically identical. Once cleaved, the C5a fragment is the most proinflammatory anaphylatoxin derived from the complement cascade. When signalling through C5aR1 and C5aR2, C5a is a strong chemoattractant recruiting neutrophils, eosinophils, monocytes, and T lymphocytes to sites of complement activation, whereupon it activates phagocytic cells, prompting degranulation. C5b, meanwhile, interacts with C6, recruiting C7–C9 to form the terminal C5b-9 complement complex or TCC (**Lachmann and Thompson, 1970**). Once inserted into a cell membrane, the TCC is referred to as the membrane attack complex (MAC), a membrane-spanning pore which can lyse sensitive cells (**Götze and Müller-Eberhard, 1970**).

Aspects of the structural biology of C5 are well understood due to a crystal structure of the apo form (**Fredslund et al., 2008**) and a number of co-crystal structures of C5 with various modulators. By virtue of its constitutive role in the terminal pathway, C5 is a recurrent target for immune evasion molecules and structures have been determined of C5 in complex with an inhibitory molecule derived from *Staphylococcus aureus*, SSL-7 (**Laursen et al., 2010**), as well as several structurally distinct examples from ticks: OmCI (**Jore et al., 2016**), RaCI (**Jore et al., 2016**), and Cirp-T (**Reichhardt et al., 2020**). Additionally, the structures of C5 with the inhibitory monoclonal antibody (mAb) eculizumab (**Schatz-Jakobsen et al., 2016**), of C5 with a small molecule inhibitor (**Jendza et al., 2019**), and of C5 with the complement-depleting agent cobra venom factor (CVF) (**Laursen et al., 2011**) have all been determined.

Here, we probe C5 with knob domain peptides and explore the molecular processes which underpin allosteric modulation of this important drug target. This study is the first to investigate the molecular mechanisms and pharmacology of this recently isolated class of peptide.

Results

Bovine knob domain peptides as potential C5 inhibitors

We have previously shown that antigen-specific, disulphide-rich knob domain peptides derived from bovine antibodies have great potential for therapeutic utility. Using this approach, we obtained four knob domain peptides: K8, K57, K92, and K149, which we have shown to display tight binding to human C5. Previously we reported equilibrium dissociation constants of 17.8 nM for K8, 1.4 nM for K57, <0.6 nM for K92, and 15.5 nM for K149 (**Macpherson et al., 2020**).

Functional characterisation of anti-C5 bovine knob domain peptides

For functional characterisation of the peptides, we performed complement assays for CP and AP activation in human serum, assessing C5b neo-epitope formation and C5a release (schematically presented in **Figure 1A, B**), in combination with orthogonal ELISAs, measuring C3b and C9 deposition. Here, we show that K57 was a potent and fully efficacious inhibitor of C5 activation, preventing release of C5a, and deposition of C5b and C9. As expected, there was no effect on C3b, which is upstream of C5 (**Figure 1C, D**). In contrast, K149 was a high-affinity silent binder with no discernible effect on C5a release, formation of C5b neo-epitope or C9 deposition, even at peptide concentrations in excess of 100 × K_D (see **Supplementary file 1** Section 1).

K8 and K92 exerted more nuanced allosteric effects on C5 (**Figure 1C, D**). By ELISA, K92 partially prevented C5 activation by the AP, but, intriguingly, no effect was observed in CP assays, suggesting K92 selectively inhibits C5 activation by the AP C5 convertase, but not the CP C5 convertase. Partial antagonists, where the degree of inhibition for the asymptotic concentrations of a dose–response curve (E_{max}) is below 100%, are an impossible mode of pharmacology for orthosteric antagonists (**Klein et al., 2013**), and we therefore propose that K92 operates by a non-steric mechanism. K8 was also demonstrably allosteric, partially inhibiting both the AP and CP in ELISA experiments. For K8 and K92, no effect on C3b deposition was detected.

When tested in CP and AP haemolysis assays (**Figure 1E, F**), K57 was a potent and fully efficacious inhibitor of complement-mediated cell lysis. Consistent with the ELISA data, K92 was active solely in the AP-driven haemolysis assay, achieving E_{max} values of 30–40%; while K8 was efficacious in the CP assay but did not show activity in the AP assay below 10 μM, potentially a consequence of the increased serum concentration and stringency of the haemolysis endpoint.

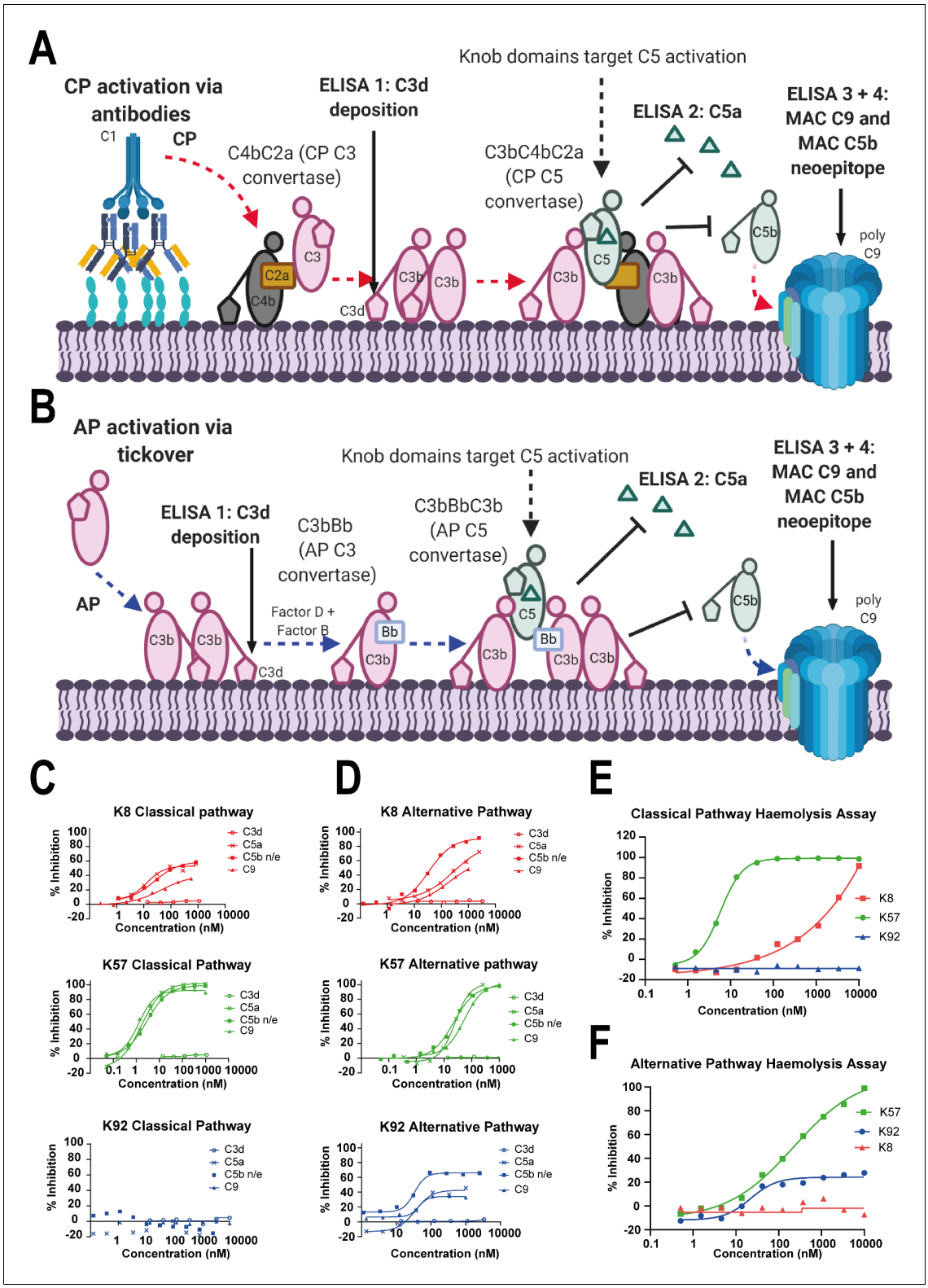


Figure 1. Functional modulation of C5 via knob domain peptides. (**A**) shows an abridged schematic for classical pathway (CP) activation. Following activation of C1q via antibody Fc, C4 and C2 are cleaved and form C4bC2a (the CP C3 convertase) which cleaves C3 into C3a (not shown) and C3b. At high C3b concentrations, C4bC2aC3b (the CP C3 convertase) forms and cleaves C5 into C5a and C5b. C5b associates with C6 and forms the membrane attack complex (MAC) with C7, C8, and multiple copies of C9. (**B**) shows an abridged schematic for surface phase alternative pathway (AP)

Figure 1 continued on next page

Figure 1 continued

activation in assays (where generation of C3b from the CP/LP is virtually excluded); tick-over of C3 generates C3a (not shown) and C3b. In the presence of factor B and factor D, C3bBb (the AP C3 convertase) generates additional C3b, prompting formation of C3bBbC3b (the AP C5 convertase), which cleaves C5 into C5a and C5b, driving MAC formation. CP-driven ELISAs (**C**) and AP-driven ELISAs (**D**) are shown. For both pathways, the inhibition of C3d (the surface-associated domain of C3b, which is upstream of C5 inhibition), C5a release, and C5b neo-epitope formation and C9 deposition were tracked within the MAC. Haemolysis assays with sheep erythrocytes, for the CP (**E**), and rabbit erythrocytes, for the AP (**F**), show that K57 is a potent and efficacious inhibitor of both pathways. K92 is selective, partial antagonist of the AP, while K8 is a weak antagonist of the CP but did not show efficacy in the AP haemolysis assay, below 10 μM. For the AP assays, 5% serum (v/v) gives a putative C5 concentration of 20 nM. For the CP assays, 1% serum (v/v) gives a putative C5 concentration of 4 nM, based on a reported C5 serum concentration of 397 nM/75 μg/mL (***Sjöholm, 1975***).

Cooperativity in C5 binding by knob domain peptides

To test for cross blocking, arising from overlapping epitopes, or cooperativity between knob domains, we performed a surface plasmon resonance (SPR) cross blocking experiment, where, using a Biacore 8K, we saturated a C5-coated sensor chip with two 20 μM injections of knob domain peptide before injecting a different peptide at 20 μM to assess its capacity to bind. This provides a qualitative measure of cross blocking, whereby an increase in response units (RUs) indicates ternary complex formation, stoichiometries, or kinetics cannot be reliably derived with concurrent dissociation of both peptides (***Figure 2***).

Saturation of C5 with K8, K57, or K92 did not prevent subsequent binding of the non-functional K149 (***Figure 2D***), suggesting K149 does not share an epitope with the other ligands, nor does it significantly perturb C5 such that the other binding sites are affected.

We detected negative cooperativity between K8 and K92, whereby saturation of C5 with K8 entirely prevented binding of K92. When the order of addition was changed and C5 was saturated with K92, K8 was still able to bind, albeit to a lesser degree (***Figure 2B***). Saturation of C5 with K8 also entirely eliminated binding of K57, with a similar order of addition effect, whereby K8 could still partially bind to the C5-K57 complex (***Figure 2A***). When C5 was saturated with K92 or K57, only very small amounts of subsequent binding of either peptide were observed by SPR (***Figure 2C***), suggesting that the epitopes do not overlap but that considerable negative cooperativity exists.

Structural analysis of C5-knob domain complexes

Crystal structure of the C5-K8 peptide complex

To elucidate the structural basis for the allosteric modulation of C5, we determined the crystal structure of the C5-K8 complex at a resolution of 2.3 Å (see ***Supplementary file 1,*** Table 2.1 for data collection and structure refinement statistics). The structure of the C5-K8 complex shows the K8 peptide binding to a previously unrecognised regulatory site on C5; the macroglobulin (MG) 8 domain of the α-chain (***Figure 3A***). K8 adopts a cysteine knot-like configuration, where a flattened 3-strand β-sheet topology is constrained by three disulphide bonds (***Figure 3A*** and ***Figure 3—figure supplement 1A***). Analysis of the K8-C5 complex with the macromolecular interfaces analysis tool PDBePISA (***Krissinel and Henrick, 2007***) reveals a large interaction surface (total buried surface area in complex: 1642 $Å^2$; with 852 $Å^2$ contributed by K8 and 790 $Å^2$ by C5), comparable to those seen in Fab-antigen complexes (***Ramaraj et al., 2012***), stabilised by an extensive network of 18 hydrogen bonds between K8 and the MG8 domain (***Figure 3A*** and ***Supplementary file 1***), dominated by arginine residues $R23_{K8}$, $R32_{K8}$, and $R45_{K8}$. The extensive H-bond network is further bolstered by several ionic interactions, between $R32_{K8}$ and $D1471_{C5}$ (C5 numbering based on mature sequence), $D25_{K8}$ and $K1409_{C5}$, and $H36_{K8}$ and $D1382_{C5}$ (***Figure 3A*** and ***Supplementary file 1*** Table 2.3). The opposing face of K8 was fortuitously stabilised by a substantial, 1275 $Å^2$, crystal contact with the C5d domain of a symmetry-related C5 molecule (***Figure 3—figure supplement 1B***), ensuring low relative B-factor values (K8: 58 $Å^2$, C5-K8 complex: 65 $Å^2$) (see also ***Figure 3—figure supplement 1C***) and clear and continuous electron density, enunciating the unique disulphide bond arrangement of the knob domain peptide and the backbone and side chains interactions with C5 (***Figure 3—figure supplement 2A*** shows a mFo-DFc simulated annealing OMIT map of the C5-K8 complex). Despite the overall resolution of the dataset comparing favourably with other C5 structures in the PDB (***Schatz-Jakobsen et al., 2016***; ***Laursen et al., 2010***; ***Jore et al., 2016***; ***Fredslund et al., 2008***), density for the C345c domain was largely absent due to this flexible domain occupying a solvent channel. This flexible attachment of the C345c domain to the α-chain of C5 is

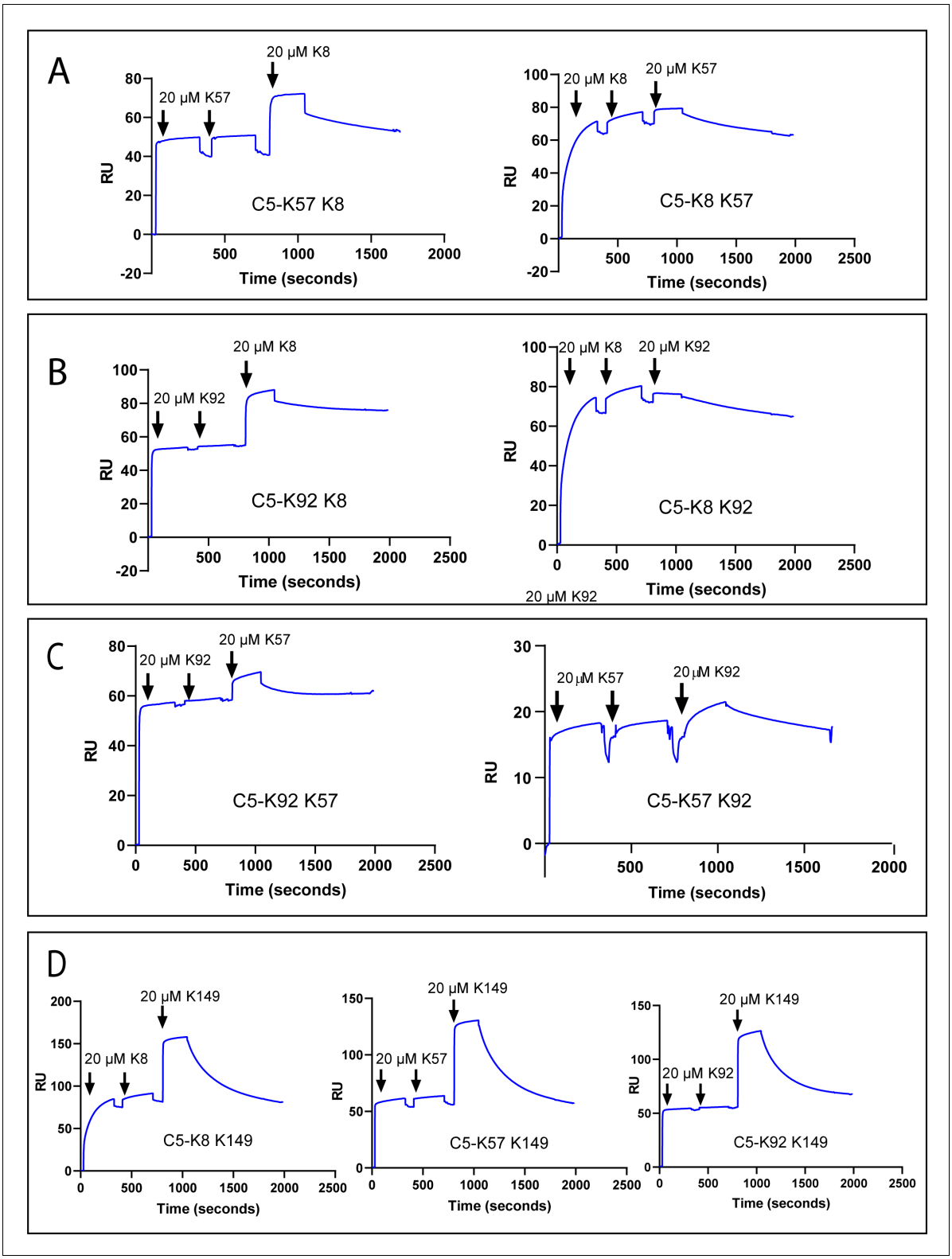


Figure 2. Surface plasmon resonance peptide cross blocking. (**A**), (**B**), and (**C**) highlight negative cooperativity between the K8, K92, and K57 peptides, respectively. Neither K57 or K92 can bind to the C5-K8 complex but K8 can bind, albeit at a lower level, to C5-K57 and C5-K92. We could not detect any negative cooperativity between K8, K57, or K92 with the silent binder K149, shown in (**D**). RU: response unit.

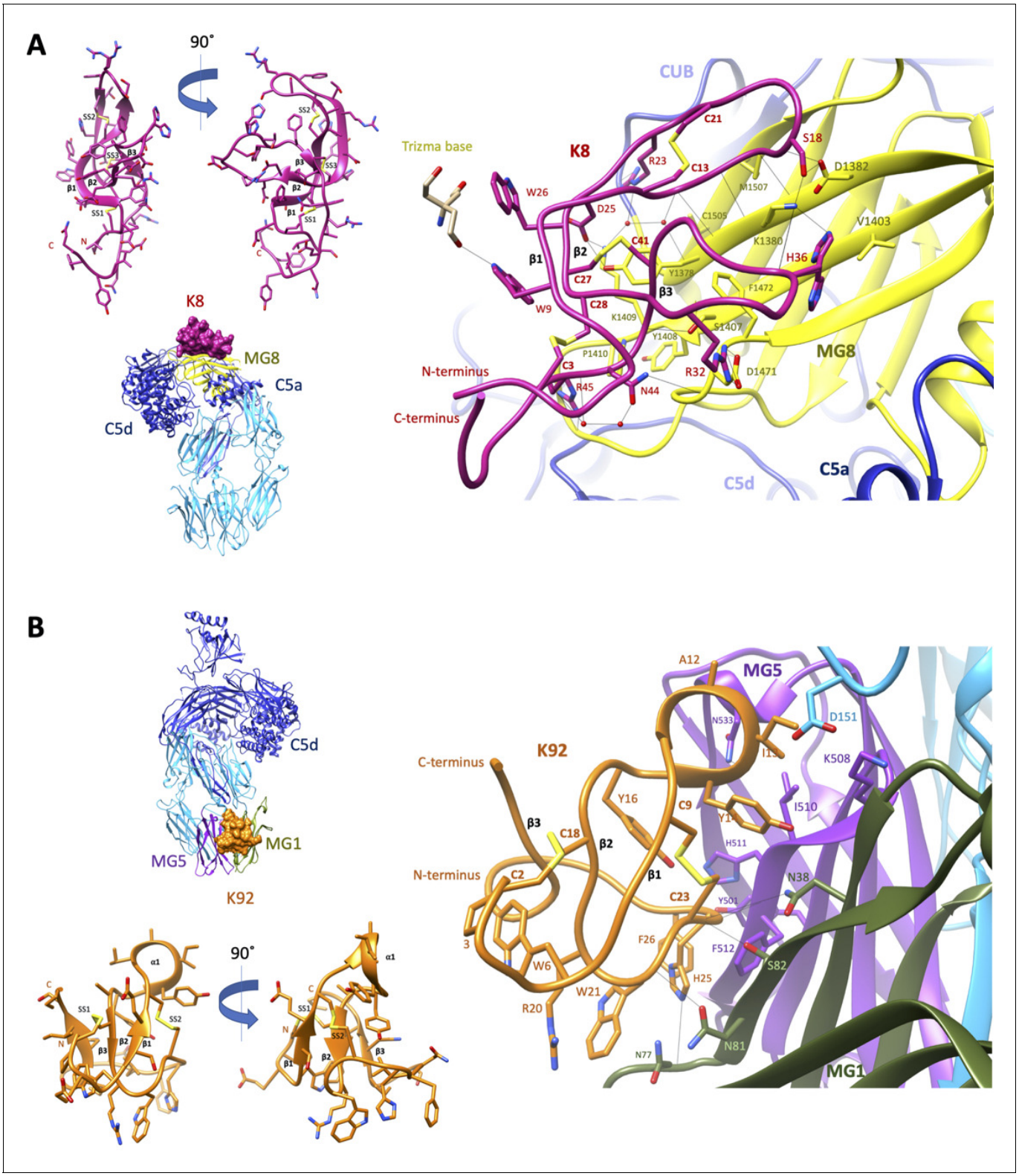


Figure 3. Crystal structures of C5-knob domain complexes. (**A**) and (**B**) show the crystal structures of C5 in complex with the K8 and K92 knob domain peptides, respectively. The binding site for the K8 peptide (**A**, shown in red) is located on a previously unreported ligand binding site on the macroglobulin (MG) 8 domain (shown in yellow) of C5. The binding site for K92 (**B**, shown in orange) is located between the MG1 and MG5 domains (shown in green and magenta, respectively).

Figure 3 continued on next page

Figure 3 continued

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Structural analysis of the C5-K8 complex.
Figure supplement 2. Simulated annealing OMIT maps of the C5-knob domain peptide complexes.
Figure supplement 3. Structural analysis of the C5-K92 complex.
Figure supplement 4. Paratope analysis of K8 and K92 peptides.

observed consistently across the C5 structures (***Schatz-Jakobsen et al., 2016***; ***Laursen et al., 2010***; ***Jore et al., 2016***; ***Fredslund et al., 2008***).

Crystal structure of the C5-K92 complex

We also present a crystal structure of the C5-K92 complex at a resolution of 2.75 Å (see ***Supplementary file 1*** for data collection and structure refinement statistics). Continuous electron density for the flexible C345c domain of C5 was observed due to it being stabilised in an upward pose by crystal contacts, akin to the C5-RaCI-OmCI ternary complex structures (Protein Data Bank [PDB] accession codes 5HCC, 5HCD, and 5HCE; ***Jore et al., 2016***). Despite displaying higher affinity binding to C5 than K8, electron density for K92 was less well defined as it occupies a solvent channel, and stabilising crystal packing interactions are absent (***Figure 3—figure supplement 3B***). A mFo-DFc simulated annealing OMIT map of the C5-K92 complex is displayed in ***Figure 3—figure supplement 2B***, showing clear but sparse electron density for the peptide at 1.3 σ. Correspondingly, only a small increase in R_{free}, of 25.35–25.53, is observed when the peptide is removed during refinement compared to an increase in R_{free} of 23.36–27.08 upon removal of the K8 peptide, potentially indicating that the occupancy is significantly below 1. This is also reflected in the high relative B-factor values for K92 (182.5 $Å^2$) compared to that of the complex (100.5 $Å^2$) (see also ***Figure 3—figure supplement 3C***). Model building of the K92 peptide was aided by disulphide mapping using mass spectrometry. The disulphide map of K92 identified formation of disulphide bonds between $C9_{K92}$ and $C23_{K92}$ and between $C2_{K92}$ and $C18_{K92}$ (***Supplementary file 1,*** Table 2.4), enabling completion of the model.

Similar to K8, K92 adopts a 3-strand β-sheet topology (***Figure 3B*** and ***Figure 3—figure supplement 3A***) but stapled with only two disulphide bonds. With shorter β-strands and longer connecting loop regions, K92 exhibits a more compact, globular arrangement (***Figure 3B***). Two extended loop regions interact with C5, including an α-helix containing loop between β-strands 1 and 2, occupying a cleft between the MG1 and MG5 domains of the β-chain of C5. The interaction surface (total buried surface area: 1365 $Å^2$; with 750 $Å^2$ contributed by K92 and 615 $Å^2$ by C5) is sustained via a sparse set of eight H-bonds (***Supplementary file 1,*** Table 2.5). A series of π–π and aliphatic–aromatic stacking interactions spans K92, encompassing $F26_{K92}$, $H25_{K92}$, $W21_{K92}$, $W6_{K92}$, and $P3_{K92}$ (***Figure 3B***). From within this hydrophobic patch, H-bonds occur between $H25_{K92}$ and the backbone carbonyls of $N77_{C5}$ and $N81_{C5}$ on the MG1 domain.

Validation of the observed C5-peptide complexes

To validate the observed K8 and K92 C5-binding modes observed in our crystal structures, we assessed the binding properties for a number of alanine mutants of K8 and K92. For K8, R23A and R32A mutants targeted the two salt-bridge interactions with C5. While for K92, where there were few electrostatic interactions mediated by side chains, we targeted a hydrogen bond, sustained by H25, and important hydrophobic interactions with C5, involving neighbouring aromatics W21 and F26. While the K92 H25A mutant could not be expressed, the other mutants were tested, alongside unmodified K8 and K92, in SPR multi-cycle kinetics experiments (n = 3). For K8, the R23A resulted in modest twofold decrease in affinity, but R32A was markedly more attenuating, with a 715-fold drop in affinity (***Supplementary file 1,*** Table 2.6). For K92, the loss of hydrophobic interactions with C5 in W21A and F26A mutants markedly abridged affinity with a 1209.2-fold and 45.7-fold drop in affinity, respectively (***Supplementary file 1,*** Table 2.6).

To analyse the interfaces observed in the structures, we performed binding pose metadynamics (***Clark et al., 2016***), an analysis typically employed to computationally evaluate the binding stability of chemical ligands (***Fusani et al., 2020***). This in silico analysis suggested that both

K8 and K92's binding poses were exceptionally stable, with the interface maintaining the key interactions in spite of applied force (***Supplementary file 1,*** Tables 2.7 and 2.8). This, in conjunction with earlier kinetic studies (***Macpherson et al., 2020***), highlights the stability of the interactions made by both knob domains.

Cysteines participate in inter- and intra-paratope interactions

In the near absence of secondary structure, disulphide bonds appear to act as sources of stability for both peptides. For K92, both the backbone amide and carbonyl of $C23_{K92}$ participate in H-bonds with the side chain of $S82_{C5}$ (***Figure 3B***). For K8, an interchain sulphur–π stack between the $C27_{K8}$-$C41_{K8}$ disulphide bond and the aromatic of $Y1378_{C5}$ positions the hydroxyl group of $Y1378_{C5}$ to make a H-bond with $D25_{K8}$ (***Figure 3A***). While for K92, an intra-chain sulphur–π stack between the $C9_{K92}$-$C23_{K92}$ disulphide bond and the aromatic of $Y14_{K92}$ orientates $Y14_{K92}$, such that its hydroxyl group participates in an interchain H-bond with $N38_{C5}$.

Comparison to known antibody paratopes

Although antibody-derived, K8 and K92 are structurally unique variable regions. We compared the K8 and K92 knob domains to a non-redundant set of 924 non-identical sequences of paired antibody–protein antigen structures from SAbDab (***Dunbar et al., 2014***). Paratopes were defined as any antibody residues within 4.5 Å of the antigen in the structure. The paratopes of K8 and K92 contain 18 and 10 residues, respectively, which are within the typical range of antibody paratope sizes (***Figure 3—figure supplement 4A***). Given this similarity in size, we searched for structurally and physicochemically similar antibody paratopes from the 924 antibody complexes but no similar paratope sites were found (***Wong et al., 2020***). While the limited examples preclude firm conclusions, this lack of similarity could be due to either the unusual fold of the knob domains or the differences in paratope amino acid composition.

In terms of residue usage, one difference in paratope composition that is potentially universal is the presence of cysteine in the knob domains (***Figure 3—figure supplement 4B***) which is uncommon in most antibody paratopes, with the exception of the CDR1-CDR3 disulphides, which have been described in camelid VHH (***Govaert et al., 2012***), and in broadly neutralising antibodies; those which cross react with several strains of a virus, and for which a disulphide bond in CDRH3 has been described in antibodies against HIV-1 (***Hutchinson et al., 2019***) and hepatitis C (***Flyak et al., 2018***). Using Arpeggio (***Jubb, 2015***) to identify inter- (antigen contacting) and intra-paratope interactions (hydrogen bond, polar, ionic, and hydrophobic) revealed that, on average, antibodies have 16 intra-paratope and 17 inter-paratope interactions; K8 is very close to this, with 15 intra-paratope and 17 inter-paratope interactions, whereas K92 paratope has fewer, with 9 intra-paratope and 10 inter-paratope interactions. A bovine Fab with an ul-CDRH3 was recently crystallised in complex with antigen, in this case a soluble portion of the HIV envelope (***Stanfield et al., 2020***). While the low resolution of the crystal structure hindered analysis, a casual inspection of the paratope suggests that 10 intra-paratope and 10 inter-paratope interactions are sustained by the knob domain, comparable to K92.

A search for structurally homologous proteins, using the DALI protein structure comparison server (***Holm, 2020***), did not find any 3D structures similar to K8 or K92, including the 14 known structures of bovine Fabs with ul-CDRH3 in the PDB. These results highlight the heterogeneity of these structural elements of the bovine immune system which likely arise through selection against a specific antigen/epitope. We next looked at homology with cyclic peptides. A recent review summarised the interactions mediated by cyclic peptides bound to proteins, across 65 co-crystal structures in the PDB (***Malde et al., 2019***). This revealed that cyclic peptides on average sustain eight electrostatic interactions with their protein target, with a range of 1–20. When we consider K8, its 19 inter-paratope interactions are comparatively high for a peptide, while the seven inter-paratope interactions of K92 are far more typical (***Figure 3—figure supplement 4C***).

The structural basis for allosteric inhibition of C5 by K8 and K92

When compared to the binding sites of other C5 modulators (***Figure 4A***), it can be observed that the epitope for K92 is entirely contained within the binding interface of a previously reported immune evasion molecule, the 23 kDa SSL7 protein from *S. aureus* (***Figure 4B***). While the C5-SSL7 structure reveals a shallow binding site involving a series of five H-bonds between SSL7 and a region

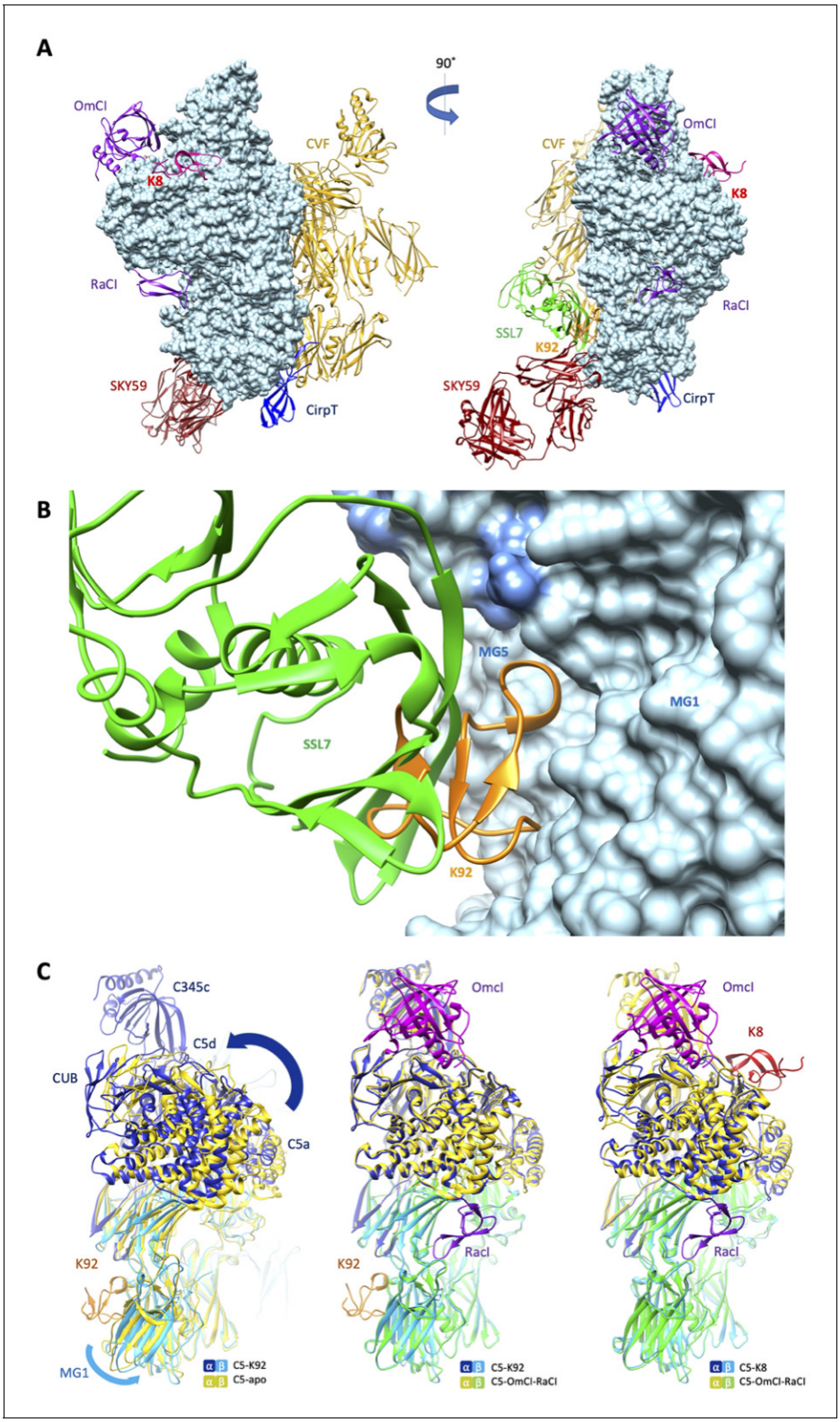


Figure 4. Comparison of the K8 and K92 binding sites with known C5 inhibitor complexes. Structural alignment of the complexes of C5 with the K8 and K92 knob domain peptides with the known structures for OmCI and RaCI (Protein Data Bank [PDB] accession code 5HCC; ***Jore et al., 2016***), SSL7 and cobra venom factor (CVF) (PDB accession code 3PRX; ***Laursen et al., 2011***), Cirp-T (PDB accession code 6RPT; ***Reichhardt et al., 2020***), and SKY59 (PDB accession code 5B71; ***Fukuzawa et al., 2017***) using UCSF Chimera (***Pettersen et al., 2004***). Alignments have been performed globally except for

Figure 4 continued on next page

Figure 4 continued

instances where the inhibitor has been crystallised bound to a single domain of C5. (**A**) shows two views of the superimposed C5-inhibitor complexes, differing by a 90° rotation. C5 is shown in molecular surface rendering, with ribbon representations of OmCI and RaCI in purple, SSL7 in green, CVF in gold, SKY59 in dark red, K8 in bright red, and K92 in orange. (**B**) shows a close-up view of the K92 binding site with that of SSL7 superimposed, for comparison. In contrast with the superficial binding mode of SSL7, K92 is wedged between the macroglobulin (MG)1 and MG5 domains of C5. (**C**) (left) shows that the interaction between K92 and C5 induces a slight separation of the MG1 and MG5 domains, resulting in a significant rotational movement of the C5a, C5d, and CUB domains, when compared to the C5-apo structure (PDB accession code 3CU7; ***Fredslund et al., 2008***). (**C**) also shows that the complex with OmCI and RaCI (PDB accession code 5HCC; ***Jore et al., 2016***) stabilises a similar conformation in C5 (**C**, middle) to that of K92, as well as K8 (**C**, right). For this structural comparison, the C5 MG5 domains of the complexes were superimposed.
The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. Structural comparison of C5-K92 and C5-CVF complexes.

of β-sheet on the MG5 domain, spanning $H511_{C5}$-$E516_{C5}$ (***Laursen et al., 2010***), here we show that K92 is wedged between the MG1 and MG5 domains, inducing a re-orientation of the side chain of $H511_{C5}$ and forming a backbone H-bond with $F510_{C5}$. When comparing K92 and SSL7, the small changes observed in the binding pose achieve different allosteric effects; SSL7, either in isolation or in complex with its second ligand IgA, is full, or occasional partial, antagonist of both the AP and CP (***Bestebroer et al., 2010***; ***Laursen et al., 2010***), while K92 is a selective partial antagonist of the AP.

Inspection of the C5a anaphylatoxin domain reveals that the C-terminus of the C5a domain in the C5-K92 complex adopts a helical conformation, which is analogous to the C5-OmCI-RaCI complex, burying the Bb-cleavage site (R751). In other C5 structures (including C5-apo and C5-CVF), this linker adopts an extended conformation following an unstructured loop and only sparse continuous electron density was observed for the linker extending from MG6 to C5a in the C5-K8 complex, possibly suggesting its R751 scissile bond is more exposed.

When the MG5 domains in the C5-K92 complex and the C5-apo structure are superimposed (***Figure 4C***), a slight twist can be observed in the MG1 domain, caused by the binding of K92 and resulting in a significant rotational movement of the C5 α-chain. A similar conformational change results from the binding of OmCI and RaCI, and to some extent K8, as these structures are virtually superimposable. CVF can form a highly stable C3/C5 convertase, following incubation with factor D and factor B in the presence of Mg^{2+} (***Vogel and Müller-Eberhard, 1982***), which may offer a surrogate model for C5 convertase (***Laursen et al., 2011***). When superimposing C5-K92 and C5-CVF (PDB accession code 3PVM) complexes, C5 does not adopt a similar conformation as when bound by K92 and K8 (***Figure 4—figure supplement 1***), potentially indicating both knob domains stabilise a different C5 conformation than binds the C5 convertase.

When considering the organisation of the C5 convertases, the C5-CVF crystal structure reveals that CVF and C5 align perfectly to create a mirror image complex, with a conformational change in the C5 convertase site at arginine 751, potentially placing C5a within range of the catalytic unit of the MG7-associated convertase complex, offering a surrogate model for C5 convertase activation (***Laursen et al., 2011***). We have shown that K92 is not an orthosteric inhibitor of either the CP or the AP convertase, thereby precluding binding of the convertase to a cleft between the MG1–MG5 domains. As the K92 epitope is entirely contained within the SSL7 binding site, this is compatible with the CVF model for C5 activation, with a co-crystal structure of the ternary complex of C5, CVF, and SSL7 (PDB accession code 3PRX6), demonstrating that the CVF and SSL7 binding sites do not cross block. Also consistent with the CVF model for C5 activation, binding of K8 to the MG8 domain would not appear to sterically block the catalytic unit. We therefore sought to further explore the apparent conformational changes in our structures.

Solution techniques reveal allosteric networks

To validate the apparent conformational changes occurring in C5 due to the binding of K8 and K92 as revealed by our crystal structures, we analysed the C5-knob domain complexes by two-solution biophysical techniques – small-angle X-ray scattering (SAXS) and hydrogen-deuterium exchange mass spectrometry (HDX-MS).

SEC-SAXS, where size exclusion chromatography (SEC) immediately precedes the solution X-ray experiment ensuring a monodispersed sample, was performed in concert with SEC multi-angle laser

light scattering (SEC-MALLS). Data were collected for C5 and the C5-K8, C5-K57, C5-K92, and C5-K149 complexes (***Figure 5A–C***). SEC-MALLS confirmed that the increases in molecular weight of the complexes were consistent throughout the elution peaks (***Supplementary file 1,*** Table 3.1 and ***Figure 5—figure supplement 1A***). While SEC-SAXS elution profiles gave stable estimates of the radius of gyration (R_G) across the tip of the peak, frames (scattering curves collected during the elution of the sample) from the descending elution peaks show lower R_G values, suggesting the presence of unbound C5.

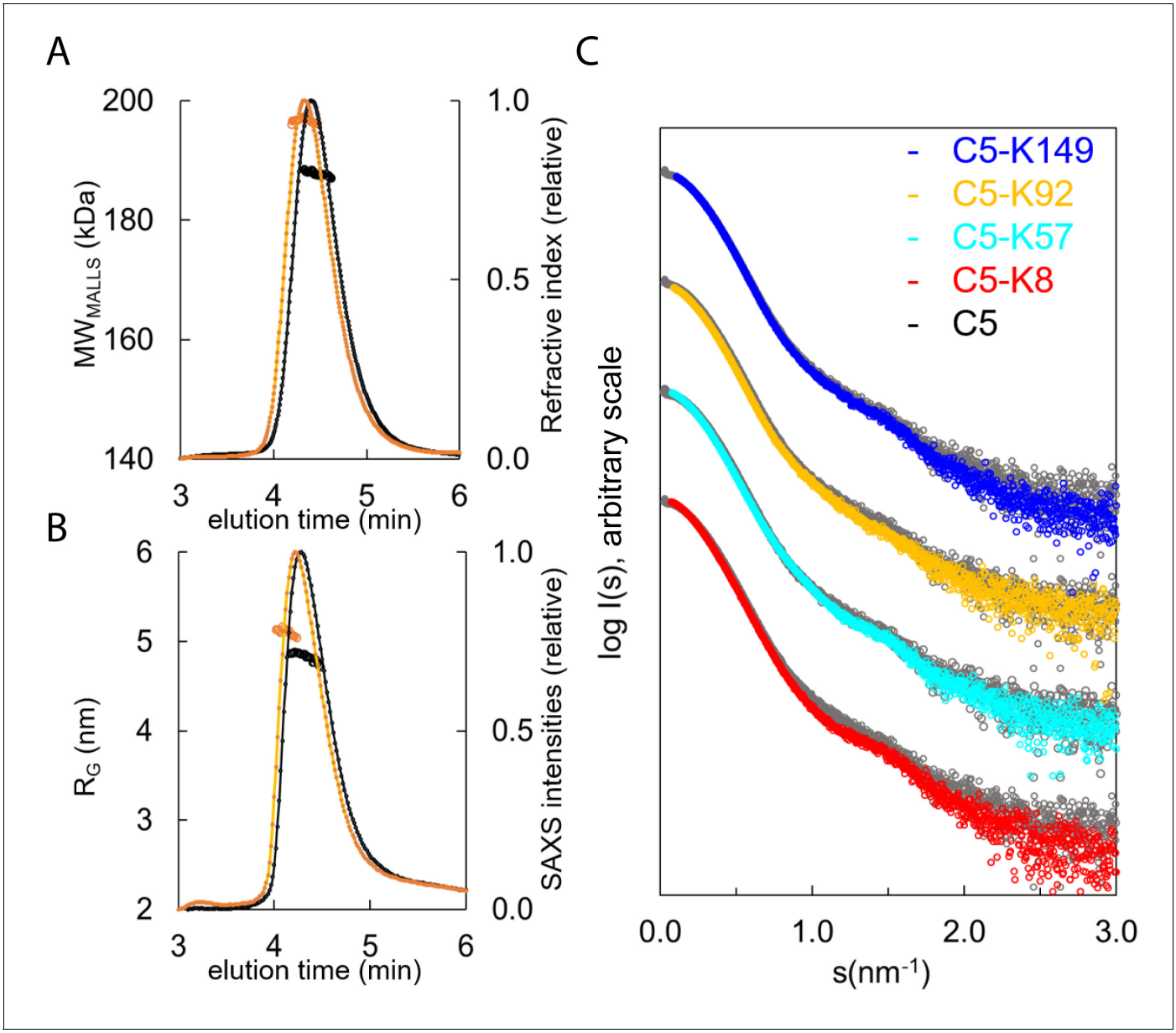


Figure 5. Hydrodynamic properties and solution conformation of C5 and C5-knob domain complexes by small-angle X-ray scattering (SAXS). Size exclusion chromatography multi-angle laser light scattering (SEC-MALLS) chromatograms (**A**) for apo C5 (black) and C5-K92 (orange) show a homogenous molecular weight increase across the C5-K92 elution peak. The SEC-SAXS elution profile collected under identical experimental conditions (**B**) shows an increase in radius of gyration (R_G) for the C5-K92 complex. Scattering curves of all C5-knob domains are shown (**C**); the C5-knob domain complexes are shown against apo C5 (in grey), and for ease of viewing, the curves are arbitrarily shifted in the Y axis.
The online version of this article includes the following figure supplement(s) for figure 5:

Figure supplement 1. SAXS analyses of the C5-knob domain peptide complexes.

Frames corresponding to the tip of the peak were averaged and submitted for full SAXS analysis. For the complexes, the scattering curves showed slight increases in both the R_G and solute volume (***Supplementary file 1,*** Table 3.1), with the C5-K8 complex showing the largest change and C5-K149 the smallest change, corresponding with the absence of function and suggesting K149 binds peripherally to a conformation closely resembling C5-apo. For K92 and K57, the discrepancies observed in the mid s range indicate an overall change in flexibility of C5 upon binding of these peptides, and this tuning of dynamics may contribute to their mechanism.

Consistent with earlier observations (***Fredslund et al., 2008***), comparison of the C5-apo experimental data with the theoretical scattering curve revealed discrepancies in the lowest angle range, indicating C5 adopts a more elongated conformation in solution than the crystal structure would suggest (***Figure 5—figure supplement 1B***). To better approximate C5 in solution, we performed a normal mode model analysis (NMA) using SREFLEX (***Panjkovich and Svergun, 2016***) and found that elongation of the C5 model improved the χ^2 from >13 to 1.55. The fit of the C5-K92 complex was also markedly improved by the NMA, whereby elongation and incorporation of the peptide improved the model from an initial χ^2 of >20, to 2.5 (with an overall root mean square [RMS] of 3.8 in both cases).

When using the C5-K8 co-crystal structure for fitting of the C5-K8 SAXS data, the absence of the C345c domain was problematic. The generation of a hybrid model where the C345c domain was reinstated initially produced a poor fit (χ^2=75). A restrained rigid body analysis of this model followed by NMA refinement allowed us to significantly reduce the discrepancy to χ^2=4.1 indicating an overall acceptable fit. The χ^2 value is still somewhat larger than those observed for the other complexes, which may suggest an increased flexibility around the C345c linker. This result correlates with the absence of clear electron density for the C345c domain in the crystal structure. The latter may be a consequence of K8 inducing additional flexibility to this region, which again could contribute to the efficacy of the peptide.

The discrepancies between the crystal structures and the solution scattering data indicate that while permitting elucidation of the molecular interaction of the epitopes, the constraints of the crystal lattice may impede the detection of more subtle, global changes, leading to underestimation of the conformational changes induced by the peptide.

To further explore such effects in solution, we used HDX-MS to provide molecular-level information on local protein structure and dynamics. HDX-MS measures the exchange of backbone amide hydrogen to deuterium in the solvent, with the rate of HDX determined by solvent accessibility, protein flexibility, and hydrogen bonding. To interpret the impact of peptide binding on C5 structural dynamics, we performed differential HDX (ΔHDX) analysis, comparing C5-knob domain complexes to apo C5, where shielding of C5 residues through participation in a binding interface will prevent deuteration, while conformational changes may increase or decrease deuterium uptake, in relation to the degree of solvent exposure.

For C5-K8, the sole protected region of C5 corresponded to the epitope on the MG8 domain ($L1380_{C5}$-$E1387_{C5}$), although the interface was not entirely defined (***Figure 6A***, see also ***Figure 6—figure supplement 1*** and ***Supplementary file 1,*** Table 3.2). Additional conformational changes were observed in the neighbouring C5d domain which becomes more solvent exposed, suggesting K8 is affecting the dynamics of this domain.

For the C5-K92 complex, consistent with the crystal structure, there was protection of the C5 residues located in the epitope between the MG1 and MG5 domains ($H70_{C5}$-$L85_{C5}$), shown in ***Figure 6C***. There were also effects distal to the K92 binding site, notably in C5d ($I1169_{C5}$-$F1227_{C5}$) and neighbouring CUB domain ($L1303_{C5}$-$L1346_{C5}$), indicating a K92-induced conformational change. Interestingly, the allosteric network can be visualised by changes in solvent exposure which propagate from the K92 epitope through MG2 domain ($L126_{C5}$-$V145_{C5}$) and into the C5d and CUB domains. For the C5-K57 complex, the absence of a co-crystal structure meant we had no prior knowledge of the K57 epitope. However, clear protection was observed in the MG5 domain, immediately adjacent to the K92 epitope ($N483_{C5}$-$L540_{C5}$), with sparse areas of increased solvent exposure located in the MG6 ($Q572_{C5}$-$L590_{C5}$), MG8 ($L1379_{C5}$-$A1388_{C5}$), and C5d ($K1048_{C5}$-$Y1064_{C5}$) domains (***Figure 6B***). A single protected peptide was also present in the CUB domain ($G951_{C5}$-$L967_{C5}$), suggesting the K57 epitope may be on either the MG5 or CUB domains.

There was little protection or deprotection of proteolytic fragments of the C5a domain in any of the complexes; we therefore propose that the knob domain peptides do not act by inducing

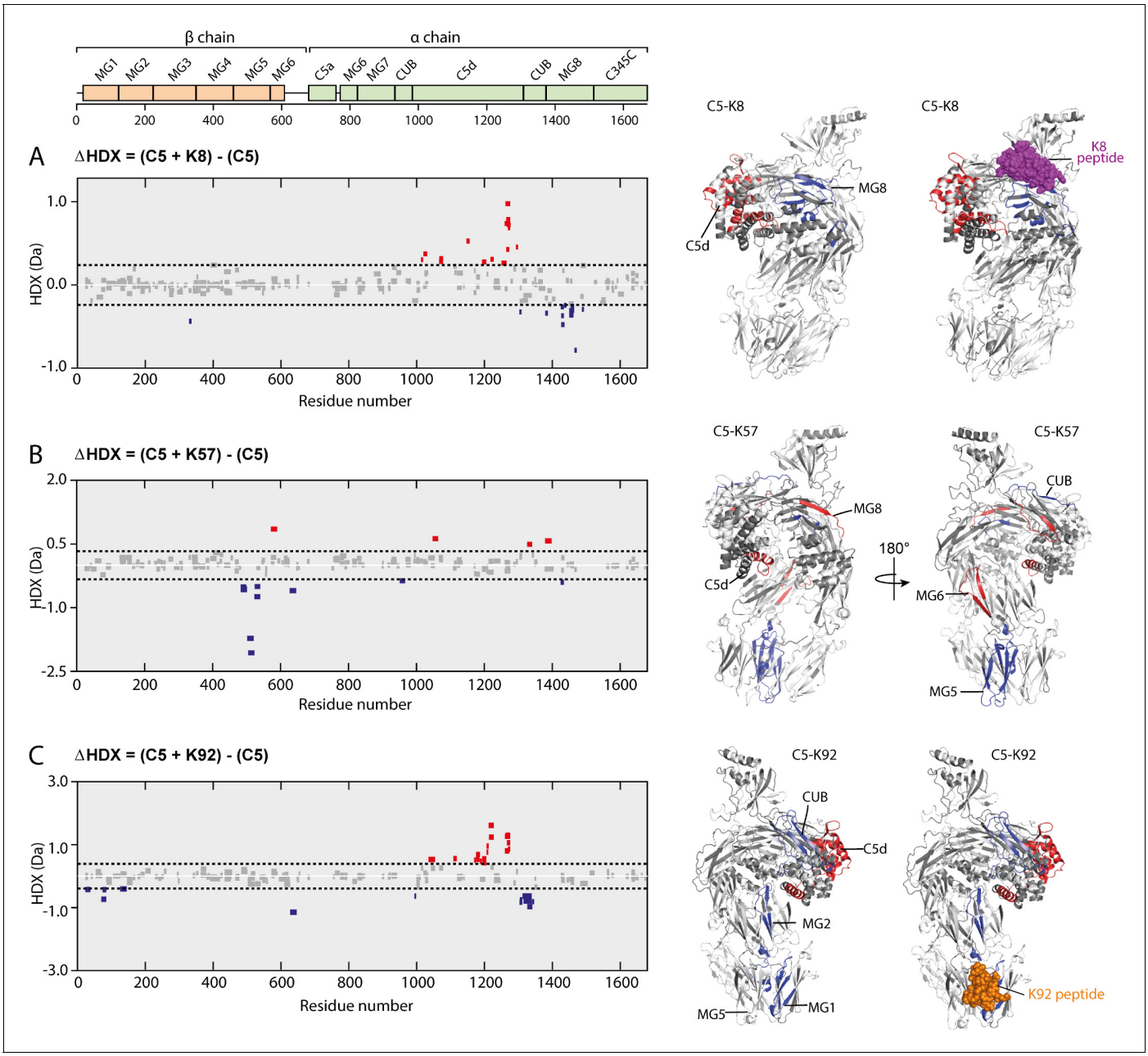


Figure 6. Impact of knob domain binding on the structural dynamics and conformation of C5. Differential hydrogen-deuterium exchange (ΔHDX) plots for C5 in complex with knob domains (**A**) K8, (**B**) K57, and (**C**) K92 at 1 hr of deuterium exposure. Blue denotes peptides with decreased HDX (backbone H-bond stabilisation), and red denotes peptides with increased HDX (backbone H-bond destabilisation). 98% confidence intervals are shown as dotted lines. Peptides in grey have insignificant ΔHDX. Measurements were performed in triplicate, and all HDX-MS peptide data are detailed in ***Supplementary file 1*** Table 3.2. ΔHDX for C5 + K8, C5 + K57, and C5 + K92 are coloured onto C5 (Protein Data Bank accession code 5HCC, minus OmCI and RaCI).

The online version of this article includes the following figure supplement(s) for figure 6:

Figure supplement 1. HDX analyses of the C5-knob domain peptide complexes.

conformational changes which shield the scissile arginine bond. Although in the structure of the C5-K92 complex the Bb-cleavage site is more buried compared to that in the C5-K8 complex. Taken in the context of the other changes, notably in the C5d and CUB domains, it is more probable that they affect more global changes in C5 which lower the affinity for C3b or the C5 convertases. The

HDX-MS data are in good agreement with our crystallographic data, with the K8 and K92 epitopes defined as clear areas of solvent protection. The conformational change in the C5d domain and significant rotational moment of the C5 α-chain, which were evident upon alignment of the MG1 domain of apo C5 with the C5-K8 and C5-K92 co-crystal structures, also appears to manifest in solution in response to binding of the knob domains.

Putative K57 binding site

To further home in on the K57 binding site, we measured binding to C5b and C5b-6 in an SPR single-cycle kinetics experiment (***Supplementary file 1,*** Table 4.1). Upon cleavage of C5a, the remaining domains of the α-chain undergo a substantial conformational change, mediated by rearrangement of the MG8, CUB, and C5d domains (***Hadders et al., 2012***; ***Aleshin et al., 2012***). The resulting C5b subunit is metastable and prone to aggregation and decay, which leaves it unable to bind C6 or form the MAC. By SPR, K8 did not bind C5b but could C5b-6. However, K57, K92, and K149 all bound C5b and C5b-6 (***Supplementary file 1*** Table 4.2). For C5b, this was within two-fold of their previously published affinities for C5 (17.8 nM for K8, 1.4 nM for K57, and <0.6 nM for K92; ***Macpherson et al., 2020***), except for K149, which displayed threefold higher affinity for both C5b and C5b-6 than previously reported for C5 (15.5 nM; ***Macpherson et al., 2020***). K92 did exhibit lower affinity for C5b-6 complex than C5, binding the complex at 6.7 nM, relative to <0.6 nM for C5 alone. As the CUB domain is significantly altered in C5b, this increases the likelihood that, of the two protected regions identified by HDX-MS, the K57 epitope is on the MG5 domain.

Discussion

We present a new family of peptides, bovine antibody-derived knob domains, and show that, upon binding of the therapeutic target complement component C5 with high affinity, they display a range of inhibitory mechanisms. Knob domain peptides rely on the bovine immune system to achieve high-affinity binding through optimisation of amino acid composition, 3D structure, and disulphide bond network. They have only been recently isolated as a practicable antibody fragment (***Macpherson et al., 2020***), and therefore, structural and functional analysis of the complexes with their clinical targets will greatly aid their development as therapeutics.

Function

Functional characterisation at the level of individual complement pathways identified K57 as a novel C5 inhibitor, which is a fully efficacious inhibitor of the terminal pathway in response to both CP and AP activation, and a potential therapeutic candidate for complement-mediated disorders, such as paroxysmal nocturnal haemoglobinuria and atypical haemolytic uraemic syndrome. Additionally, the discovery of K149 as a 'silent binder' of C5 may be of considerable value as a non-inhibitory reagent for the detection of native C5 (***Figure 1***). K8 and K92 both displayed allosteric inhibitory activity against C5. K92 achieved selective inhibition of the AP through a non-competitive mechanism (***Figure 1***). To our knowledge, this is the first reported example of complement pathway-specific inhibition through C5 and the first experimental evidence reported for mechanistic differences between the AP and CP C5 convertases. This suggests an expanded therapeutic scope for C5, whereby tuning of the conformational ensemble or dynamics with allosteric compounds can bias activation to leave certain complement pathways intact. Complete inhibition of the terminal pathway has been shown to increase the susceptibility of eculizumab patients to *Neisseria meningitidis* infections (***McNamara et al., 2017***). Selective inhibition of C5-cleavage by the AP C5-convertase, and not the CP C5-convertase, may partially preserve serum bactericidal activity, thereby lowering the risk of meningococcal disease.

Structure

Structural analyses, utilising X-ray crystallography, revealed the unique topologies of knob domain peptides K8 and K92 and their distinctive binding modes in C5. Due to the apparent structural homology of knob domains with certain venomous peptides, of which conotoxins and spider venoms are examples, it has been proposed that the knob domains of ul-CDRH3 might be similarly predisposed to target the concave epitopes of ion channels. Likewise, structural homology with defensin peptides has garnered hypotheses regarding an improved ability to bind viral capsid coats. Indeed,

bovine antibodies with ul-CDRH3 have been raised against the viral capsid of HIV with exceptional efficiency, given the challenging nature of the antigen (**Sok et al., 2017**; **Stanfield et al., 2020**). However, the study presented here shows that, in the case of C5, concave epitopes are not the knob domain's sole preserve. Notably, the MG8 domain epitope of K8 offers a planar pharmacophore and, while the K92 epitope is more undulating, casual inspection of the C5 structure reveals numerous deeper cavities available (**Figure 3**). This may mean that knob domains can be raised to inhibit flat surfaces involved in protein–protein interactions, which might not offer binding sites for orthosteric small molecules.

We note that the structural architecture of the knob domains varies for the epitope. Their immune derivation means that, unlike cysteine-rich peptides derived from other natural sources, such as venoms, the bovine immune system can be used to define specificity for any antigen. Comparative structural analysis suggests that knob domain paratopes are differentiated from conventional antibodies and the structures of known cyclic peptides, offering a different binding architecture to other small antibody fragments, such as the camelid VHH. While firm conclusions are hampered by limited examples, the number of interactions does not seem dissimilar from cyclic peptides or mAbs, both of which have been successfully applied to tackling high-affinity protein–protein interactions.

Our structures demonstrate that the importance of the network of disulphide bonds goes beyond a stabilising role. An apparent paucity of secondary structure would suggest that while stabilisation of the domain is indeed critical, disulphide bonds also participate in sulphur–π interactions to sustain intra- and inter-chain interactions. The structures of knob domains bound to their target antigen demonstrate both the diversity and versatility afforded to the bovine immune repertoire by these sequences.

Mechanism

Structural alignment of the C5-K92 co-crystal structure with the apo C5 structure (**Figure 4**), using the MG5 domain, revealed a rotational movement of the MG1 causing the α-chain to adopt a twisted conformation accompanied by a rotational movement in the C5d domain, in response to knob domain binding. Comparison of the C5-K8 structure with the apo C5 structure revealed a similar conformation but with less movement in the C5d domain. The helical C5d domain is the target of two immune evasion molecules which have evolved in ticks, OmCI and RaCI, both of which inhibit C5 by crosslinking C5d to neighbouring domains (**Jore et al., 2016**). Additionally, it has been shown that polyclonal antibodies raised against C5d inhibit binding of C5 to C3b (**DiScipio, 1981**). The binding site of OmCI is contained within the CUB and C5d domains, with only a single, non-bonded interaction to the C345c domain visible in the crystal structure (**Jore et al., 2016**), which appears mediated by crystal contacts. Interestingly, the C5-OmCI-RaCI crystal structure (PDB accession code 5HCC) reveals similar conformational changes in C5d, relative to the apo C5 structure. We therefore propose that rearrangement of C5d can lower the affinity, or preclude the interaction, of C5 for the convertases and that this may be a common inhibitory mechanism for OmCI, RaCI, K8, and K92. Should K92 and K8 inhibit C5 by modulating the C5d domain, in the case of K92, this occurs at a range of over 50 Å. Such remote effects are not unprecedented; allosteric structural changes can be propagated at over 150 Å in response to drug binding (**Haselbach et al., 2017**).

Subsequent solution biophysics methods substantiate our crystallographic observations. HDX-MS analysis revealed areas of solvent protection changes in the MG8 domain, resulting from the binding of K8, and in the MG1 and MG5 domains of the C5-K92 complex (**Figure 6**), corresponding to their respective epitopes, as identified with X-ray crystallography (**Figure 3**) and confirmed by site-directed mutagenesis analysis. Changes in solvent exposure were also observed in the α-chain for C5-K92, providing a route to visualise the allosteric network. As similar conformational or dynamic changes occur both in solution and in the crystal structure, this suggests that the effects are ligand induced and are not the result of crystal packing interactions. SAXS analysis also suggests that K8 and K92 increase the flexibility of C5 and effects on dynamics may be a contributing factor in realising efficacy (**Figure 5**).

For K57, which had an E_{max} of 100% for both pathways and was not demonstrably allosteric, HDX-MS and biacore experiments with the metastable C5b suggested a putative epitope on the MG5 domain. This could support an orthosteric mechanism of action as CVF, which can form a stable C5 convertase, contacts the MG5 domain in the C5-CVF co-crystal structure (PDB accession code 3PVM). However, by SAXS, similar changes in conformation and/or dynamics to the C5-K8 and C5-

K92 complexes were apparent and an allosteric network, including changes in the C5d domain, was observed in the C5-K92 complex by HDX-MS. These observations could support an allosteric mechanism for K57.

Importantly, we also saw a high degree of negative cooperativity between the different knob domains by SPR (***Figure 2***), suggesting that all the functional knob domains perturb the conformational state or dynamics of C5. K57 also showed cooperativity with other functional knob domains, suggesting providing further evidence that it stabilises a conformation of C5 that is less energetically favourable for binding of the other ligands.

Our observations with K92 suggest that further work may be required to elucidate the mechanism of action of another binder of the MG1 and MG5 domains, SSL7. Given that SSL7 can be a partial inhibitor (***Laursen et al., 2010***), even with co-binding of IgA, this precludes a steric mechanism and invites biophysical studies in solution. Additionally, another tick-derived inhibitor, Cirp-T, was also recently reported as predominantly binding to the MG4 domain, with an orthosteric mechanism of action attributed. However, we note that published data only showed an E_{max} of <90% in AP-driven assays (***Reichhardt et al., 2020***), indicating that it is an allosteric C5 inhibitor for the AP, and potentially also the CP, C5 convertase, which may merit further investigation.

In conclusion, we introduce knob domains as a new peptide modality, with unexplored therapeutic potential for the modulation of proteins and protein–protein interactions. This study is the first application of knob domain peptides and reveals an unexpectedly high incidence of allosteric modulators of complement C5, expanding its scope for complement-targeted therapies and providing important mechanistic tools for the study of C5 convertases. Knob domains can offer a range of advantages over the current macromolecular C5 inhibitors, including their use as peptide therapeutics, while grafting knob domains into the CDRH3 of well-characterised Fabs or using Fc tags could provide routes to extend half-life in vivo by attenuating renal clearance.

Materials and methods

Complement proteins

Human C5 was affinity purified using an E141A, H164A OmCI column (***Macpherson et al., 2018***). Briefly, human serum (TCS Biosciences, Botolph Claydon, UK) was diluted 1:1 (v/v) with phosphate buffered saline (PBS), 20 mM ethylenediaminetetraacetic acid (EDTA), and applied to a 5 mL Hi-Trap NHS column (GE Healthcare, Amersham UK), which contained 20 mg of E141A H164A OmCI protein, at a rate of 1 mL/min. The column was washed with 5× column volumes (CV) of PBS, C5 was then eluted using 2 M $MgCl_2$ and immediately dialysed into PBS. C5b was prepared from human C5 by incubating C5 with CVF, factor B and factor D, at a 1:10 molar ratio, as previously described (***Jore et al., 2016***). C5a was removed using a spin column with 30 kDa cut-off (Thermo Fisher Scientific, Horsham, UK).

Knob domain peptide production

Knob domain peptides were expressed fused to the CDRH3 of the PGT-121 Fab, as previously described (***Macpherson et al., 2020***). Plasmid DNA for each construct was amplified using QIAGEN Plasmid Plus Giga Kits. Expi293F cultures were transfected with Expifectamine 293 Transfection kits (Invitrogen, Renfrew, UK) as per the manufacturer's instructions. The cells were cultured for 4 days and supernatants harvested by centrifugation at 4000 rpm for 1 hr. Harvested supernatants were applied to a Hi-Trap Nickel excel columns (GE Healthcare, Amersham, UK) using an Akta pure (GE Healthcare, Amersham, UK). Cell supernatants were loaded at 2.5 mL/min, followed by a wash of 7× CV of PBS, 0.5 M NaCl. A second wash with 7× CV of buffer A (0.5 M NaCl, 0.02 M Imidazole, PBS pH 7.3) was performed, and samples were eluted by isocratic elution with 10× CV of buffer B (0.5 M NaCl, 0.25 M) Imidazole, PBS (pH 7.3). Post elution, the protein-containing fractions were pooled and buffer exchanged into PBS using dialysis cassettes (Thermo Fisher Scientific, Horsham, UK).

For isolation of the knob domain peptide, PGT-121 Fab-knob peptide fusion proteins were incubated with tobacco etch virus protease, at a ratio of 100:1 (w/w), for a minimum of 2 hr at room temperature. Peptides were purified using a Waters UV-directed FractionLynx system with a Waters XBridge Protein BEH C4 OBD Prep Column (300 Å, 5 µm, 19 × 100 mm, Waters Corp., Milford, MA). An aqueous solvent of water, 0.1% trifluoroacetic acid (TFA), and an organic solvent of 100%

MeCN was used. The column was run at 20 mL/min at 40˚C with a gradient of 5–50% organic solvent, over 11 min. Fractions containing knob peptide were pooled and lyophilised using a Labconco Freezone freeze drier.

Complement activation assays

For the C3 and C9 ELISAs, microtiter plates (MaxiSorp; Nunc) were incubated overnight at 4˚C with 50 µL of a solution of in 75 mM sodium carbonate (pH 9.6) containing either 2.5 µg/mL aggregated human IgG (Sigma-Aldrich, Gillingham, UK) for CP or 20 µg/mL zymosan (Sigma-Aldrich, Gillingham, UK) for AP. As a negative control, wells were coated with 1% (w/v) bovine serum albumin (BSA)/PBS. Microtiter plates were washed four times with 250 µL of wash buffer (50 mM Tris-HCl), 150 mM NaCl and 0.1% Tween 20 (pH 8) between each step of the procedure. Wells were blocked using 250 µL of 1% (w/v) BSA/PBS for 2 hr at room temperature. Normal human serum was diluted in either gelatin veronal buffer with calcium and magnesium (GVB^{++}: 0.1% gelatin, 5 mM Veronal buffer, 145 mM NaCl, 0.025% NaN_3, 0.15 mM calcium chloride, 1 mM magnesium chloride, pH 7.3; for CP) or Mg-ethylene glycol tetraacetic acid (Mg-EGTA) (2.5 mM veronal buffer [pH 7.3] containing 70 mM of NaCl, 140 mM of glucose, 0.1% gelatin, 7 mM of MgCl2, and 10 mM of EGTA; for AP). Serum was used at a concentration of 1% in CP or 5% in AP and was mixed with serially diluted concentrations of peptides (16 µM – 15.6 nM) in GVB^{++} or Mg-EGTA buffer, and preincubated on ice for 30 min. Peptide–serum solutions were then incubated in the wells of microtiter plates for 35 min for CP assays (both C3b and C9 detection) or 35 min for AP (C3b) or 60 min for AP (C9), at 37˚C. Complement activation was assessed through detection of deposited complement activation factors using specific antibodies against C3b (rat anti-human C3d HM2198, Hycult, Uden, The Netherlands) and C9 (goat anti-human C9, A226, Complement Technologies Tyler, TX) at a 1:1000 dilution. Bound primary antibodies were detected with horse rdischHRP-conjugated goat anti-rat (ab97057, Abcam, Cambridge, UK) or rabbit anti-goat (P0449, Agilent Dako, Santa Clara, CA) secondary antibodies at a 1:1000 dilution. Bound HRP-conjugated antibodies were detected using TMB One solution (Eco-TEK – manufactured by Bio-TEK, Winooski, VT) with absorbance measured at 450 nm.

For the C5b ELISA, assays were run using the CP and AP Complement functional ELISA kits (Svar Life Science, Malmö, Sweden). For sample preparation, serum was diluted as per the respective protocol for the CP and AP assays. Serial dilutions of peptides were prepared and allowed to incubate with serum for 15 min at room temperature prior to plating.

For the C5a ELISA, assays were run using the Complement C5a Human ELISA Kit (Invitrogen, Renfrew, UK). For sample preparation, at the end of the 37˚C incubation of the serum/peptide samples on the C5b ELISA plate, 50 µL of the diluted, activated serum was transferred to a C5a ELISA plate containing 50 µL/well of assay buffer. All subsequent experimental steps were performed as described in the protocol.

Haemolysis assays

GVB^{++} or Mg EGTA buffers, which had been supplemented with 2.5% glucose (w/v), were used for the CP and AP assays, respectively. For the AP, 150 µL of rabbit erythrocytes (TCS Biosciences, Botolph Claydon, UK) were washed twice, by addition of 1 mL of buffer and centrifugation at 800 $\times g$ for 1 min, and finally resuspended in 500 µL of buffer. For the CP, 150 µL sheep erythrocytes (TCS Biosciences, Botolph Claydon, UK) were washed twice with 1 mL of buffer and sensitised with a 1/1000 dilution of rabbit anti-sheep red blood cell stroma antibody (S1389, Sigma Aldrich, Gillingham, UK). After a 30˚C/30 min incubation, with shaking, the cells were rewashed and resuspended with 500 µL of buffer. Serial dilutions of peptide were prepared in the respective buffers and normal human serum was added at 1% for the CP and 4.5% for the AP (corresponding to CH50 of the serum). Also, 90 µL of peptide–serum mixtures were plated into a V-bottom 96-well microtiter plate (Corning) and 10 µL of erythrocytes were added. Plates were incubated for 30 min at 37˚C, with shaking. Finally, 50 µL of buffer was added, the plates centrifuged at 800 $\times g$, and 80 µL of supernatant was transferred to an ELISA plate (Nunc) and absorbance measured at 405 nm.

Crystallography and structure determination

6.1 mg/mL C5 (20 mM Tris-HCl, 75 mM NaCl, pH 7.35) was mixed at a 1:1 molar ratio with either the K8 or K92 peptides. Crystallisation trials were initiated by the vapor diffusion method at 18˚C with a 1:1 mixture of mother liquor (v/v). C5-K8 crystals were grown in a mother liquor of 0.1 M N-(2-acetamido)iminodiacetic A, 14% ethanol (v/v), pH 6.0. For C5-K92 crystals, the mother liquor was 0.1 M bicine/Trizma (pH 8.5), 10% (w/v) polyethylene glycol 8000, 20% (v/v) ethylene glycol, 30 mM sodium fluoride, 30 mM sodium bromide, and 30 mM sodium iodide (**Gorrec, 2009**). Prior to flash freezing in liquid nitrogen, C5-K8 crystals were cryoprotected in mother liquor with 30% 2-methyl-2,4-pentanediol (v/v). C5-K92 crystals were frozen without additional cryoprotection.

Data were collected at the Diamond Light Source (Harwell, UK), on beamline I03, at a wavelength of 0.9762 Å. The C5-K8 structure was solved using the automated molecular replacement pipeline Balbes (**Long et al., 2008**) using the apo C5 structure (PDB accession code 3CU7), minus the C345c domain. The C5-K8 complex crystallised in space group $P2_12_12_1$ with one molecule in the asymmetric unit. A backbone model of the K8 peptide was produced using ARP-wARP (**Langer et al., 2008**) which informed manual model building in Coot (**Emsley et al., 2010**), within the CCP4 suite (**Winn et al., 2011**). The model was subjected to multiple rounds of refinement in Refmac (**Murshudov et al., 1997**) and Phenix (**Adams et al., 2010**). The overall geometry in the final structure of the C5-K8 complex is good, with 97.2% of residues in favoured regions of the Ramachandran plot and no outliers.

The C5-K92 complex crystallised in space group C2 with one molecule in the asymmetric unit. C5 was solved by molecular replacement with Phaser (**McCoy et al., 2007**) using the C5-OmCI-RaCI structure (PDB accession code 5HCC), with OmCI and RaCI removed. Manual building of the K92 peptide in Coot was greatly informed by mass spectroscopy disulphide mapping experiments. The model was subjected to multiple rounds of manual rebuilding in Coot and refinement in Phenix (**Adams et al., 2010**). The overall geometry in the final structure of the C5-K92 complex is good, with 95.2% of residues in favoured regions of the Ramachandran plot and no outliers. Structure factors and coordinates for both C5-knob domain peptide complexes have been deposited in the PDB (PDB accession codes: 7AD6 (C5-K92 complex) and 7AD7 (C5-K8 complex)). Crystal trials were also performed with the C5-K57 and C5-K149 complexes, but the resulting crystals diffracted poorly.

Disulphide mapping of K92 peptide

A 250 µL K92 peptide at 1 mg/mL was alkylated with addition of 18 µL of 2-Iodoacetamide (Thermo Fisher Scientific, Horsham, UK) at room temperature for 30 min. Overnight dialysis into assay buffer (7.5 mM Tris-HCl, 1.5 mM $CaCl_2$, pH 7.9) was performed using 2 kDa slide-a-lyzer cassettes (Thermo Fisher Scientific, Horsham, UK). Chymotrypsin (sequencing grade, Roche Applied Sciences) was reconstituted to 1 µg/µL in assay buffer, and 5 µL of reconstituted enzyme was added to 80 µL of sample. Once mixed, the sample was incubated at 37˚C for 1.5 hr before being quenched with 5 µL of 1% TFA. Samples were diluted 1 in 10 and 5 µL was loaded onto the analytical column.

Liquid chromatography electrospray ionisation mass spectrometry was acquired using an Ultimate 3000 UHPLC system (Thermo Fisher Scientific, Horsham, UK) coupled with a Q-Exactive Plus Orbitrap (Thermo Fisher Scientific, Horsham, UK). Separations were performed using gradient elution (A: 0.1% formic acid; B: 0.1% formic acid in acetonitrile) on an Acquity UPLC BEH C18 column (130 Å, 1.7 µm, 2.1 × 150 mm; Waters Corp., Milford, MA) with the column temperature maintained at 40˚C.

The following analytical gradient at 0.2 mL/min was used: 1% B was held for 2 min, 1–36% B over 28 min, 36–50% over 5 min, and 50–99% B over 0.5 min. There were sequential wash steps with changes in gradient of 99%–1% B over 0.5 min (at a higher flow rate of 0.5 mL/min) before equilibration at 1% B for 6.5 min (at the original 0.2 mL/min).

A full MS/dd-MS2 (Top 5) scan was run in positive mode. Full MS: scan range was 200–2000 m/z with 70,000 resolution (at 200 m/z) and a 3×10^6 AGC target (the maximum target capacity of the C-trap), 100 ms maximum Injection time. The dd-MS2: 2.0 m/z isolation window, CID fragmentation (NCE 28) with fixed first mass of 140.0 m/z, with a 17,500 resolution (at 200 m/z), 1×10^5 AGC target, 200 ms maximum injection time. The source conditions of the MS were capillary voltage, 3 kV;

S-lens RF level, 50; sheath gas flow rate, 25; auxiliary gas flow rate, 10; auxiliary gas heater temperature, 150˚C; and the MS inlet capillary was maintained at 320˚C.

Data were acquired using XcaliburTM 4.0 software (Thermo Fisher Scientific, Horsham, UK), and raw files were analysed by peptide mapping analysis using Biopharma Finder 2.0 software (Thermo Fisher Scientific, Horsham, UK) by performing a disulphide bond search with a chymotrypsin (medium specificity) digest against the K92 peptide sequence. Assignments and integrations from Biopharma Finder were filtered to include only peptides identified as containing a single disulphide bond and with an experimental mass within |5| ppm of the theoretical mass. Intensities for all peptides containing the same cysteines pairing were summed and percentages were obtained from the summed against total intensities.

SPR multicycle kinetics

Kinetics were measured using a Biacore 8K (GE Healthcare, Amersham, UK) with a CM5 chip, which was prepared as follows: 1-ethyl-3-(−3-dimethylaminopropyl) carbodiimide hydrochloride (EDC)/N-hydroxysuccinimide (NHS) was mixed at 1:1 ratio (flow rate, 10 µL/min; contact time, 30 s), and human C5 at 1 µg/mL in pH 4.5 sodium acetate buffer was injected over flow cell one only (flow rate, 10 µL/min; contact time, 60 s). Final immobilisation levels in the range of 2000–3000 RUs were obtained to yield theoretical Rmax values of ~50–60 RU. Serial dilutions of K8 and K92 knob domains, and various mutants, were prepared in HBS-EP (0.01 M HEPES pH 7.4, 0.15 M NaCl, 3 mM EDTA, 0.005% v/v Surfactant P20) buffer and injected (flow rate, 30 µL/min; contact time, 240 s; dissociation time, 6000 s). After each injection, the surface was regenerated with two sequential injections of 2 M $MgCl_2$ (flow rate, 30 µL/min; contact time, 30 s). Binding to the reference surface was subtracted, and the data were fitted to a single-site binding model using Biacore evaluation software.

Binding pose metadynamics

Simulation structures were prepared using Schrodinger's Maestro Protein preparation wizard. The molecular dynamics runs were performed using the Schrodinger's default implementation of the binding pose metadynamics with the peptide chain considered in place of a ligand. Additional RMSD calculations for the peptide internal structure assessment in the last 20% of the dynamics were performed relative to the starting structures.

Small-angle X-ray scattering

Data was collected at the EMBL P12 beam line (PETRA III, DESY Hamburg, Germany; ***Blanchet et al., 2015***). Data was collected with inline SEC mode using the Agilent 1260 Infinity II Bio-inert LC. Also, 50 µL of complement component C5 at 31.6 µM (5.96 mg/mL) was injected onto a Superdex 200 Increase 5/150 column (GE Healthcare, Amersham, UK) at a flow rate of 0.35 mL/min. The mobile phase comprised 20 mM Tris pH 7.35, 75 mM NaCl, and 3% glycerol. The column elute was directly streamed to the SAXS capillary cell, and throughout the 15-min run, 900 frames of 1 s exposure were collected. After data reduction and radial averaging, the program CHROMIXS (***Panjkovich and Svergun, 2018***) was employed. Around 100 statistically similar buffer frames were selected and used for background subtraction of the sample frames from the chromatographic peak. This results in the final I(s) vs s scattering profiles, where $s = 4\pi\sin\theta/\lambda$, 2θ is the scattering angle, and λ = 1.24 Å. The scattering data in the momentum transfer range $0.05 < s < 0.32$ nm^{-1} were collected with a PILATUS 6M pixel detector at a distance of 3.1 m from the sample.

ATSAS 2.8 (***Franke et al., 2017***) was employed for further data analysis and modelling. The program PRIMUS (***Konarev et al., 2003***) was used to perform Guinier analysis (ln(I(s)) versus s^2) from which the radius of gyration, R_G, was determined. Distance probability functions, p(r), were calculated using the inverse Fourier transformation method implemented in GNOM (***Svergun, 1992***) that provided the maximum particle dimension, D_{max}. The concentration-independent molecular weight estimate, MW_{VC}, is based on the volume of correlation (***Rambo and Tainer, 2013***). The values are reported in ***Supplementary file 1,*** Table 3.1.

Theoretical scattering profiles were computed from X-ray coordinates using Crysol (***Svergun et al., 1995***), and SREFLEX (***Panjkovich and Svergun, 2016***) was used to refine the

models. For this, the program partitions the structure into pseudo-domains and hierarchically employs NMA to find the domain rearrangements minimising the discrepancy χ^2 between the SAXS curve computed from the refined model and the experimental data. Because of the absence of electron density for the C345c domain in the C5-K8 complex structure, we included a round of restrained rigid body refinement followed by NMA to obtain an improved fit.

On the same day, MALLS data were collected with a separate SEC run under the same experimental conditions (set-up, buffer, run parameters, etc.). For this, a Wyatt Technologies miniDAWN TREOS MALLS detector coupled to an OptiLab T-Rex differential refractometer for protein concentration determination (dn/dc was taken as 0.185) was used. The MALLS system was calibrated relative to the scattering from toluene. The MWMALLS distribution of species eluting from the SEC column were determined with the Wyatt ASTRA7 software package.

The experimental SAXS data and the models derived from them were deposited to the Small Angle Scattering Biological Data Bank (SASBDB accession number SASDJA6).

Hydrogen/deuterium mass spectrometry

A 6 µM of C5 was incubated with 10 µM of peptide (K8, K92, or K57) to achieve complex during deuterium exchange conditions. Then, 4 µL of C5 or the C5-peptide complex were diluted into 57 µL of 10 mM phosphate in H_2O (pH 7.0) or into 10 mM phosphate in D_2O (pD 7.0) at 25˚C. The deuterated samples were then incubated for 0.5, 2, 15, and 60 min at 25˚C. After the reaction, all samples were quenched by mixing at 1:1 (v/v) with a quench buffer (4 M guanidine hydrochloride, 250 mM Tris (2-carboxyethyl) phosphine hydrochloride, 100 mM phosphate) at 1˚C. The final mixed solution was pH 2.5. The mixture was then immediately injected into the nanoAcquity HDX module (Waters Corp., Milford, MA) for peptic digest using an enzymatic online digestion column (Waters Corp., Milford, MA) in 0.2% formic acid in water at 20˚C and with a flow rate of 100 µL/min. All deuterated time points and undeuterated controls were carried out in triplicate with blanks run between each data point.

Peptide fragments were then trapped using an Acquity BEH C18 1.7 µM VANGUARD chilled precolumn for 3 min. Peptides were eluted into a chilled Acquity UPLC BEH C18 1.7 µm 1.0 × 100 using the following gradient: 0 min, 5% B; 6 min, 35% B; 7 min, 40% B; 8 min, 95% B, 11 min, 5% B; 12 min, 95% B; 13 min, 5% B; 14 min, 9 5% B; and 15 min, 5% B (A: 0.2% HCOOH in H2O; B: 0.2% HCOOH) in acetonitrile. The trap and UPLC columns were both maintained at 0˚C. Peptide fragments were ionised by positive electrospray into a Synapt G2-Si mass spectrometer (Waters Corp., Milford, MA). Data acquisition was run in ToF-only mode over an m/z range of 50–2000 Th using an MSE method (low collision energy, 4 V; high collision energy: ramp from 18 V to 40 V). Glu-1-Fibrinopeptide B peptide was used for internal lock mass correction. To avoid significant peptide carry-over between runs, the on-line Enzymate pepsin column (Waters Corp., Milford, MA) was washed three times with pepsin wash (0.8% formic acid, 1.5 M Gu-HCl, 4% MeOH) and a blank run was performed between each sample run.

MSE data from undeuterated samples of C5 were used for sequence identification using the Waters Protein Lynx Global Server 2.5.1 (PLGS). Ion accounting files for the three control samples were combined into a peptide list imported into DynamX v3.0 software (Waters Corp., Milford, MA). The output peptides were subjected to further filtering in DynamX. Filtering parameters used were minimum and maximum peptide sequence length of 4 and 25, respectively, minimum intensity of 1000, minimum MS/MS products of 2, minimum products per amino acid of 0.2, and a maximum MH + error threshold of 10 ppm. DynamX was used to quantify the isotopic envelopes resulting from deuterium uptake for each peptide at each time point. Furthermore, all the spectra were examined and checked visually to ensure correct assignment of m/z peaks and only peptides with a high signal to noise ratios were used for HDX-MS analysis.

Following manual filtration in DynamX, confidence intervals for differential HDX-MS (ΔHDX) measurements of individual time point were calculated using Deuteros (***Lau et al., 2019***) software. Only peptides which satisfied a ΔHDX confidence interval of 98% were considered significant. The ΔHDX was then plotted onto the C5 structure in Pymol.

Surface plasmon resonance, single-cycle kinetics

On a Biacore 8K (GE Healthcare, Amersham, UK), human C5b was immobilised on a CM5 chip (GE Healthcare, Amersham, UK). Flow cells were activated using a standard immobilisation protocol: EDC/NHS was mixed at 1:1 ratio (flow rate, 10 µL/min; contact time, 30 s). C5b, at 1 µg/mL, or C5b-6 (Complement Technologies, Tyler, TX), at 2 µg/mL, in pH 4.5 sodium acetate buffer, were immobilised on flow cell two only (flow rate, 10 µL/min; contact time, 420 s). Finally, ethanolamine was applied to both flow cells (flow rate, 10 µL/min; contact time, 420 s). A final immobilisation level of 500–700 RUs was obtained for C5b and 1000–1150 RUs were obtained for C5b-6. Single-cycle kinetics were measured using a seven-point, threefold serial dilution (spanning a range of 1 µM to 1.4 nM) in HBS-EP buffer (GE Healthcare, Amersham, UK). A high flow rate of 40 µL/min was used, with a contact time of 300 s and a dissociation time of 2700 s. Binding to the reference surface was subtracted, and the data were fitted to a single-site binding model using Biacore evaluation software. All sensorgrams were inspected for evidence of mass transport limitation using the flow rate-independent component of the mass transfer constant (tc).

Surface plasmon resonance, cross blocking

On a Biacore 8K (GE Healthcare, Amersham, UK), human C5 was amine coupled to a CM5 chip using the same protocol as for C5b. A final immobilisation level of approximately 1000–2000 RUs was obtained. For cross blocking, the surface was saturated with two sequential injections of a 20 µM knob domain solution in HBS-EP buffer (GE Healthcare, Amersham, UK) using a flow rate of 30 µL/min and contact time of 300 s. This was immediately followed with an injection of a second knob domain peptide, again at 20 µM in HBS-EP, with a flow rate of 30 µL/min and a contact time of 270 s, and the dissociation phase was measured for 600 s. Binding to the reference surface was subtracted, and sensorgrams were plotted in GraphPad Prism (GraphPad Software, San Diego, California USA, www.graphpad.com).

Acknowledgements

We thank John Cashman for his help during crystallography screening. We are grateful to the three reviewers and editor for their valuable comments that helped to improve our manuscript.

Additional information

Competing interests

Charlotte M Deane: Reviewing editor, *eLife*. Alex Macpherson, Zainab Ahdash, James R Birtley, Monika-Sarah ED Schulze, Ben Holmes, Vladas Oleinikovas, James Snowden, Victoria Ellis, Alastair DG Lawson: employee of UCB and may hold shares and/or stock options. Tom Eirik Mollnes: T.E.M is a Board member of Ra Pharmaceuticals, Inc. The other authors declare that no competing interests exist.

Funding

No external funding was received for this work.

Author contributions

Alex Macpherson, Conceptualization, Data curation, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing - original draft, Writing - review and editing; Maisem Laabei, Data curation, Formal analysis, Investigation; Zainab Ahdash, Formal analysis, Investigation, Methodology; Melissa A Graewert, Sarah A Robinson, Ben Holmes, Vladas Oleinikovas, Per H Nilsson, James Snowden, Victoria Ellis, Tom Eirik Mollnes, Dmitri Svergun, Investigation, Methodology; James R Birtley, Monika-Sarah ED Schulze, Susan Crennell, Formal analysis, Validation, Investigation, Methodology; Charlotte M Deane, Formal analysis, Investigation, Methodology, Writing - original draft; Alastair DG Lawson, Conceptualization, Resources, Formal analysis, Supervision, Validation, Investigation, Methodology, Writing - original draft, Project administration, Writing - review and editing; Jean MH van den Elsen, Conceptualization, Resources, Data curation, Formal analysis, Supervision, Funding

acquisition, Validation, Investigation, Visualization, Methodology, Writing - original draft, Project administration, Writing - review and editing

Author ORCIDs

Alex Macpherson http://orcid.org/0000-0002-4508-5322
Maisem Laabei http://orcid.org/0000-0002-8425-3704
Zainab Ahdash http://orcid.org/0000-0002-4495-8689
James Snowden http://orcid.org/0000-0003-4855-7329
Jean MH van den Elsen https://orcid.org/0000-0002-0367-1956

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.63586.sa1
Author response https://doi.org/10.7554/eLife.63586.sa2

Additional files

Supplementary files

- Supplementary file 1. Section 1. Functional analysis. Table 1.1. Classical pathway C5b deposition ELISA; Table 1.2. Alternative pathway C5b deposition ELISA; Table 1.3. Inhibition of classical pathway-mediated C5a release; Table 1.4. Inhibition of alternative pathway-mediated C5a release; Table 1.5. Inhibition of classical pathway-mediated C9 deposition; Table 1.6. Inhibition of alternative pathway-mediated C9 deposition; Table 1.7. Inhibition of classical pathway haemolysis; Table 1.8. Inhibition of alternative pathway haemolysis. Section 2. Structural analysis. Table 2.1. Data collection and refinement statistics (molecular replacement); Table 2.2. Hydrogen bond interactions between K8 and C5; Table 2.3. Ionic interactions between K8 and C5; Table 2.4. Disulphide mapping of the K92 peptide; Table 2.5. Hydrogen bond interactions between K92 and C5; Table 2.6. Validation of molecular interactions by peptide mutagenesis analysis; Table 2.7. Individual, total, and average hydrogen bond persistence in a binding pose metadynamics simulation of the K8-C5 complex; Table 2.8. Individual, total, and average hydrogen bond persistence in a binding pose metadynamics simulation of the K92-C5 complex. Section 3. Solution structure analysis. Table 3.1. SAXS Summary data; Table 3.2. ΔHDX summary data. Section 4. Additional functional analyses. Table 4.1. SPR single-cycle kinetics of knob domains binding to human C5b; Table 4.2. SPR single-cycle kinetics of knob domains binding to human C5b-6.
- Transparent reporting form

Data availability

Structural datasets presented in this study have been made publicly available in the Protein Data Bank (PDB) and Small Angle Scattering Biological Data Bank (SASBDB).

The following datasets were generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Macpherson A, Elsen JM	2021	Crystal structure C5-K8 complex	https://www.rcsb.org/structure/7AD7	RCSB Protein Data Bank, 7AD7
Macpherson A, Elsen JM	2021	Crystal structure of C5-K92 complex	https://www.rcsb.org/structure/7AD6	RCSB Protein Data Bank, 7AD6
Macpherson A, Elsen JM, Graewert MA, Svergun D	2020	SAXS data and models of C5-bovine knob domain peptides	https://www.sasbdb.org/data/SASDJA6/	Small Angle Scattering Biological Data Bank, SASDJA6

References

Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH. 2010. *PHENIX*: a comprehensive Python-based system for macromolecular structure solution. *Acta*

Crystallographica Section D Biological Crystallography **66**:213–221. DOI: https://doi.org/10.1107/S0907444909052925, PMID: 20124702

Albazli K, Kaminski HJ, Howard JF. 2020. Complement inhibitor therapy for myasthenia gravis. *Frontiers in Immunology* **11**:917. DOI: https://doi.org/10.3389/fimmu.2020.00917, PMID: 32582144

Aleshin AE, DiScipio RG, Stec B, Liddington RC. 2012. Crystal structure of C5b-6 suggests structural basis for priming assembly of the membrane attack complex. *Journal of Biological Chemistry* **287**:19642–19652. DOI: https://doi.org/10.1074/jbc.M112.361121

Bestebroer J, Aerts PC, Rooijakkers SH, Pandey MK, Köhl J, van Strijp JA, de Haas CJ. 2010. Functional basis for complement evasion by staphylococcal superantigen-like 7. *Cellular Microbiology* **12**:1506–1516. DOI: https://doi.org/10.1111/j.1462-5822.2010.01486.x, PMID: 20545943

Biesecker G, Dihel L, Enney K, Bendele RA. 1999. Derivation of RNA aptamer inhibitors of human complement C5. *Immunopharmacology* **42**:219–230. DOI: https://doi.org/10.1016/S0162-3109(99)00020-X, PMID: 10408383

Blanchet CE, Spilotros A, Schwemmer F, Graewert MA, Kikhney A, Jeffries CM, Franke D, Mark D, Zengerle R, Cipriani F, Fiedler S, Roessle M, Svergun DI. 2015. Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA III, DESY). *Journal of Applied Crystallography* **48**:431–443. DOI: https://doi.org/10.1107/S160057671500254X, PMID: 25844078

Borodovsky A, Yucius K, Sprague A, Banda NK, Holers VM, Vaishnaw A, Maier M, Kallanthottathil R, Charisse K, Kuchimanchi S, Manoharan M, Salant DJ, Fitzgerald K, Meyers R, Sorensen B. 2014. Aln-CC5, an investigational RNAi therapeutic targeting C5 for complement inhibition. *Blood* **124**:1606–. DOI: https://doi.org/10.1182/blood.V124.21.1606.1606

Clark AJ, Tiwary P, Borrelli K, Feng S, Miller EB, Abel R, Friesner RA, Berne BJ. 2016. Prediction of Protein-Ligand binding poses via a combination of induced fit docking and metadynamics simulations. *Journal of Chemical Theory and Computation* **12**:2990–2998. DOI: https://doi.org/10.1021/acs.jctc.6b00201, PMID: 27145262

DiScipio RG. 1981. The conversion of human complement component C5 into fragment C5b by the alternative-pathway C5 convertase. *Biochemical Journal* **199**:497–504. DOI: https://doi.org/10.1042/bj1990497

Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, Shi J, Deane CM. 2014. SAbDab: the structural antibody database. *Nucleic Acids Research* **42**:D1140–D1146. DOI: https://doi.org/10.1093/nar/gkt1043

Emsley P, Lohkamp B, Scott WG, Cowtan K. 2010. Features and development of *coot*. *Acta Crystallographica. Section D, Biological Crystallography* **66**:486–501. DOI: https://doi.org/10.1107/S0907444910007493, PMID: 20383002

Flyak AI, Ruiz S, Colbert MD, Luong T, Crowe JE, Bailey JR, Bjorkman PJ. 2018. HCV broadly neutralizing antibodies use a CDRH3 disulfide motif to recognize an E2 glycoprotein site that can be targeted for vaccine design. *Cell Host & Microbe* **24**:703–716. DOI: https://doi.org/10.1016/j.chom.2018.10.009, PMID: 30439340

Franke D, Petoukhov MV, Konarev PV, Panjkovich A, Tuukkanen A, Mertens HDT, Kikhney AG, Hajizadeh NR, Franklin JM, Jeffries CM, Svergun DI. 2017. *ATSAS 2.8*: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *Journal of Applied Crystallography* **50**:1212–1225. DOI: https://doi.org/10.1107/S1600576717007786, PMID: 28808438

Fredslund F, Laursen NS, Roversi P, Jenner L, Oliveira CL, Pedersen JS, Nunn MA, Lea SM, Discipio R, Sottrup-Jensen L, Andersen GR. 2008. Structure of and influence of a tick complement inhibitor on human complement component 5. *Nature Immunology* **9**:753–760. DOI: https://doi.org/10.1038/ni.1625, PMID: 18536718

Fukuzawa T, Sampei Z, Haraya K, Ruike Y, Shida-Kawazoe M, Shimizu Y, Gan SW, Irie M, Tsuboi Y, Tai H, Sakiyama T, Sakamoto A, Ishii S, Maeda A, Iwayanagi Y, Shibahara N, Shibuya M, Nakamura G, Nambu T, Hayasaka A, et al. 2017. Long lasting neutralization of C5 by SKY59, a novel recycling antibody, is a potential therapy for complement-mediated diseases. *Scientific Reports* **7**:1080. DOI: https://doi.org/10.1038/s41598-017-01087-7, PMID: 28439081

Fusani L, Palmer DS, Somers DO, Wall ID. 2020. Exploring ligand stability in protein crystal structures using binding pose metadynamics. *Journal of Chemical Information and Modeling* **60**:1528–1539. DOI: https://doi.org/10.1021/acs.jcim.9b00843, PMID: 31910338

Gorrec F. 2009. The MORPHEUS protein crystallization screen. *Journal of Applied Crystallography* **42**:1035–1042. DOI: https://doi.org/10.1107/S0021889809042022, PMID: 22477774

Götze O, Müller-Eberhard HJ. 1970. Lysis of erythrocytes by complement in the absence of antibody. *Journal of Experimental Medicine* **132**:898–915. DOI: https://doi.org/10.1084/jem.132.5.898

Govaert J, Pellis M, Deschacht N, Vincke C, Conrath K, Muyldermans S, Saerens D. 2012. Dual beneficial effect of interloop disulfide bond for single domain antibody fragments*. *Journal of Biological Chemistry* **287**:1970–1979. DOI: https://doi.org/10.1074/jbc.M111.242818

Hadders MA, Bubeck D, Roversi P, Hakobyan S, Forneris F, Morgan BP, Pangburn MK, Llorca O, Lea SM, Gros P. 2012. Assembly and regulation of the membrane attack complex based on structures of C5b6 and sC5b9. *Cell Reports* **1**:200–207. DOI: https://doi.org/10.1016/j.celrep.2012.02.003, PMID: 22832194

Haselbach D, Schrader J, Lambrecht F, Henneberg F, Chari A, Stark H. 2017. Long-range allosteric regulation of the human 26S proteasome by 20S proteasome-targeting Cancer drugs. *Nature Communications* **8**:15578. DOI: https://doi.org/10.1038/ncomms15578, PMID: 28541292

Holm L. 2020. DALI and the persistence of protein shape. *Protein Science* **29**:128–140. DOI: https://doi.org/10.1002/pro.3749, PMID: 31606894

Hutchinson JM, Mesa KA, Alexander DL, Yu B, O'Rourke SM, Limoli KL, Wrin T, Deeks SG, Berman PW. 2019. Unusual cysteine content in V1 region of gp120 from an elite suppressor that produces broadly neutralizing

antibodies. *Frontiers in Immunology* **10**:1021. DOI: https://doi.org/10.3389/fimmu.2019.01021, PMID: 31156622

Jendza K, Kato M, Salcius M, Srinivas H, De Erkenez A, Nguyen A, McLaughlin D, Be C, Wiesmann C, Murphy J, Bolduc P, Mogi M, Duca J, Namil A, Capparelli M, Darsigny V, Meredith E, Tichkule R, Ferrara L, Heyder J, et al. 2019. A small-molecule inhibitor of C5 complement protein. *Nature Chemical Biology* **15**:666–668. DOI: https://doi.org/10.1038/s41589-019-0303-9, PMID: 31209353

Jore MM, Johnson S, Sheppard D, Barber NM, Li YI, Nunn MA, Elmlund H, Lea SM. 2016. Structural basis for therapeutic inhibition of complement C5. *Nature Structural & Molecular Biology* **23**:378–386. DOI: https://doi.org/10.1038/nsmb.3196, PMID: 27018802

Jubb H. 2015. Arpeggio: An interatomic interaction calculation web server. *ARPEGGIO*. 1. http://structure.bioc.cam.ac.uk/arpeggio

Klein MT, Vinson PN, Niswender CM. 2013. Approaches for probing allosteric interactions at 7 transmembrane spanning receptors. *Progress in Molecular Biology and Translational Science* **115**:1–59. DOI: https://doi.org/10.1016/B978-0-12-394587-7.00001-4, PMID: 23415091

Konarev PV, Volkov VV, Sokolova AV, Koch MHJ, Svergun DI. 2003. *PRIMUS* : a Windows PC-based system for small-angle scattering data analysis . *Journal of Applied Crystallography* **36**:1277–1282. DOI: https://doi.org/10.1107/S0021889803012779

Krissinel E, Henrick K. 2007. Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology* **372**:774–797. DOI: https://doi.org/10.1016/j.jmb.2007.05.022, PMID: 17681537

Lachmann PJ, Thompson RA. 1970. Reactive lysis: the complement-mediated lysis of unsensitized cells. II. the characterization of activated reactor as C56 and the participation of C8 and C9. *The Journal of Experimental Medicine* **131**:643–657. DOI: https://doi.org/10.1084/jem.131.4.643, PMID: 4193935

Langer G, Cohen SX, Lamzin VS, Perrakis A. 2008. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature Protocols* **3**:1171–1179. DOI: https://doi.org/10.1038/nprot.2008.91, PMID: 18600222

Lau AMC, Ahdash Z, Martens C, Politis A. 2019. Deuteros: software for rapid analysis and visualization of data from differential hydrogen deuterium exchange-mass spectrometry. *Bioinformatics* **35**:3171–3173. DOI: https://doi.org/10.1093/bioinformatics/btz022, PMID: 30649183

Lau JL, Dunn MK. 2018. Therapeutic peptides: historical perspectives, current development trends, and future directions. *Bioorganic & Medicinal Chemistry* **26**:2700–2707. DOI: https://doi.org/10.1016/j.bmc.2017.06.052, PMID: 28720325

Laursen NS, Gordon N, Hermans S, Lorenz N, Jackson N, Wines B, Spillner E, Christensen JB, Jensen M, Fredslund F, Bjerre M, Sottrup-Jensen L, Fraser JD, Andersen GR. 2010. Structural basis for inhibition of complement C5 by the SSL7 protein from *Staphylococcus aureus*. *PNAS* **107**:3681–3686. DOI: https://doi.org/10.1073/pnas.0910565107, PMID: 20133685

Laursen NS, Andersen KR, Braren I, Spillner E, Sottrup-Jensen L, Andersen GR. 2011. Substrate recognition by complement convertases revealed in the C5-cobra venom factor complex. *The EMBO Journal* **30**:606–616. DOI: https://doi.org/10.1038/emboj.2010.341, PMID: 21217642

Lee AC-L, Harris JL, Khanna KK, Hong J-H. 2019. A comprehensive review on current advances in peptide drug development and design. *International Journal of Molecular Sciences* **20**:2383. DOI: https://doi.org/10.3390/ijms20102383

Long F, Vagin AA, Young P, Murshudov GN. 2008. *BALBES*: a molecular-replacement pipeline. *Acta Crystallographica. Section D, Biological Crystallography* **64**:125–132. DOI: https://doi.org/10.1107/S0907444907050172, PMID: 18094476

Macpherson A, Liu X, Dedi N, Kennedy J, Carrington B, Durrant O, Heywood S, van den Elsen J, Lawson ADG. 2018. The rational design of affinity-attenuated OmCI for the purification of complement C5. *Journal of Biological Chemistry* **293**:14112–14121. DOI: https://doi.org/10.1074/jbc.RA118.004043

Macpherson A, Scott-Tucker A, Spiliotopoulos A, Simpson C, Staniforth J, Hold A, Snowden J, Manning L, van den Elsen J, Lawson ADG. 2020. Isolation of antigen-specific, disulphide-rich knob domain peptides from bovine antibodies. *PLOS Biology* **18**:e3000821. DOI: https://doi.org/10.1371/journal.pbio.3000821, PMID: 32886672

Malde AK, Hill TA, Iyer A, Fairlie DP. 2019. Crystal structures of Protein-Bound cyclic peptides. *Chemical Reviews* **119**:9861–9914. DOI: https://doi.org/10.1021/acs.chemrev.8b00807, PMID: 31046237

McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. 2007. *Phaser* crystallographic software. *Journal of Applied Crystallography* **40**:658–674. DOI: https://doi.org/10.1107/S0021889807021206, PMID: 19461840

McNamara LA, Topaz N, Wang X, Hariri S, Fox L, MacNeil JR. 2017. High risk for invasive meningococcal disease among patients receiving eculizumab (Soliris) Despite receipt of meningococcal vaccine. *MMWR. Morbidity and Mortality Weekly Report* **66**:734–737. DOI: https://doi.org/10.15585/mmwr.mm6627e1, PMID: 28704351

Murshudov GN, Vagin AA, Dodson EJ. 1997. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallographica Section D Biological Crystallography* **53**:240–255. DOI: https://doi.org/10.1107/S0907444996012255, PMID: 15299926

Nürnberger J, Philipp T, Witzke O, Saez AO, Vester U, Baba HA, Kribben A, Zimmerhackl LB, Janecke AR, Nagel M, Kirschfink M. 2009. Eculizumab for atypical Hemolytic–Uremic Syndrome. *New England Journal of Medicine* **360**:542–544. DOI: https://doi.org/10.1056/NEJMc0808527

Panjkovich A, Svergun DI. 2016. Deciphering conformational transitions of proteins by small angle X-ray scattering and normal mode analysis. *Physical Chemistry Chemical Physics* **18**:5707–5719. DOI: https://doi.org/10.1039/C5CP04540A, PMID: 26611321

Panjkovich A, Svergun DI. 2018. CHROMIXS: automatic and interactive analysis of chromatography-coupled small-angle X-ray scattering data. *Bioinformatics* **34**:1944–1946. DOI: https://doi.org/10.1093/bioinformatics/btx846, PMID: 29300836

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF chimera–a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**:1605–1612. DOI: https://doi.org/10.1002/jcc.20084, PMID: 15264254

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. *Biochimica Et Biophysica Acta (BBA) - Proteins and Proteomics* **1824**:520–532. DOI: https://doi.org/10.1016/j.bbapap.2011.12.007

Rambo RP, Tainer JA. 2013. Accurate assessment of mass, models and resolution by small-angle scattering. *Nature* **496**:477–481. DOI: https://doi.org/10.1038/nature12070, PMID: 23619693

Reichhardt MP, Johnson S, Tang T, Morgan T, Tebeka N, Popitsch N, Deme JC, Jore MM, Lea SM. 2020. An inhibitor of complement C5 provides structural insights into activation. *PNAS* **117**:362–370. DOI: https://doi.org/10.1073/pnas.1909973116, PMID: 31871188

Ricardo A, Arata M, DeMarco SJ, Dhamnaskar K, Hammer R, Josephson K, Seyb K, Tobe S, Wang Z, Zheng H, Treco D. 2014. Development of RA101348, a potent cyclic peptide inhibitor of C5 for Complement-Mediated diseases. *Blood* **124**:2936. DOI: https://doi.org/10.1182/blood.V124.21.2936.2936

Romay-Penabad Z, Carrera Marin AL, Willis R, Weston-Davies W, Machin S, Cohen H, Brasier A, Gonzalez EB. 2014. Complement C5-inhibitor rEV576 (coversin) ameliorates in-vivo effects of antiphospholipid antibodies. *Lupus* **23**:1324–1326. DOI: https://doi.org/10.1177/0961203314546022, PMID: 25228739

Rother RP, Rollins SA, Mojcik CF, Brodsky RA, Bell L. 2007. Discovery and development of the complement inhibitor eculizumab for the treatment of paroxysmal nocturnal hemoglobinuria. *Nature Biotechnology* **25**:1256–1264. DOI: https://doi.org/10.1038/nbt1344, PMID: 17989688

Schatz-Jakobsen JA, Zhang Y, Johnson K, Neill A, Sheridan D, Andersen GR. 2016. Structural basis for Eculizumab-Mediated inhibition of the complement terminal pathway. *The Journal of Immunology* **197**:337–344. DOI: https://doi.org/10.4049/jimmunol.1600280, PMID: 27194791

Sjöholm AG. 1975. Complement components in normal serum and plasma quantitated by electroimmunoassay. *Scandinavian Journal of Immunology* **4**:25–30. DOI: https://doi.org/10.1111/j.1365-3083.1975.tb02596.x, PMID: 49076

Smith K, Pace A, Ortiz S, Kazani S, Rottinghaus S. 2020. A phase 3 Open-label, randomized, controlled study to evaluate the efficacy and safety of intravenously administered ravulizumab compared with best supportive care in patients with COVID-19 severe pneumonia, acute lung injury, or acute respiratory distress syndrome: a structured summary of a study protocol for a randomised controlled trial. *Trials* **21**:639. DOI: https://doi.org/10.1186/s13063-020-04548-z, PMID: 32660611

Sok D, Le KM, Vadnais M, Saye-Francisco KL, Jardine JG, Torres JL, Berndsen ZT, Kong L, Stanfield R, Ruiz J, Ramos A, Liang CH, Chen PL, Criscitiello MF, Mwangi W, Wilson IA, Ward AB, Smider VV, Burton DR. 2017. Rapid elicitation of broadly neutralizing antibodies to HIV by immunization in cows. *Nature* **548**:108–111. DOI: https://doi.org/10.1038/nature23301, PMID: 28726771

Stanfield RL, Berndsen ZT, Huang R, Sok D, Warner G, Torres JL, Burton DR, Ward AB, Wilson IA, Smider VV. 2020. Structural basis of broad HIV neutralization by a vaccine-induced cow antibody. *Science Advances* **6**:eaba0468. DOI: https://doi.org/10.1126/sciadv.aba0468

Svergun DI. 1992. Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *Journal of Applied Crystallography* **25**:495–503. DOI: https://doi.org/10.1107/S0021889892001663

Svergun D, Barberato C, Koch MHJ. 1995. *CRYSOL* – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates . *Journal of Applied Crystallography* **28**:768–773. DOI: https://doi.org/10.1107/S0021889895007047

Takata Y, Kinoshita T, Kozono H, Takeda J, Tanaka E, Hong K, Inoue K. 1987. Covalent association of C3b with C4b within C5 convertase of the classical complement pathway. *Journal of Experimental Medicine* **165**:1494–1507. DOI: https://doi.org/10.1084/jem.165.6.1494

Vogel CW, Müller-Eberhard HJ. 1982. The cobra venom factor-dependent C3 convertase of human complement a kinetic and thermodynamic analysis of a protease acting on its natural high molecular weight substrate. *Journal of Biological Chemistry* **257**:8292–8299. DOI: https://doi.org/10.1016/S0021-9258(18)34330-8

Wilkinson T, Dixon R, Page C, Carroll M, Griffiths G, Ho LP, De Soyza A, Felton T, Lewis KE, Phekoo K, Chalmers JD, Gordon A, McGarvey L, Doherty J, Read RC, Shankar-Hari M, Martinez-Alier N, O'Kelly M, Duncan G, Walles R, et al. 2020. ACCORD: a multicentre, seamless, phase 2 adaptive randomisation platform study to assess the efficacy and safety of multiple candidate agents for the treatment of COVID-19 in hospitalised patients: a structured summary of a study protocol for a randomised controlled trial. *Trials* **21**:691. DOI: https://doi.org/10.1186/s13063-020-04584-9, PMID: 32736596

Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AG, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A, Wilson KS. 2011. Overview of the CCP4 suite and current developments. *Acta Crystallographica. Section D, Biological Crystallography* **67**:235–242. DOI: https://doi.org/10.1107/S0907444910045749, PMID: 21460441

Wong WK, Robinson SA, Bujotzek A, Georges G, Lewis AP, Shi J, Snowden J, Taddese B, Deane CM. 2020. Ab-Ligity: identifying sequence-dissimilar antibodies that bind to the same epitope. *bioRxiv*. DOI: https://doi.org/10.1101/2020.03.24.004051

Zelek WM, Cole J, Ponsford MJ, Harrison RA, Schroeder BE, Webb N, Jolles S, Fegan C, Morgan M, Wise MP, Morgan BP. 2020. Complement inhibition with the C5 blocker LFG316 in severe COVID-19. *American Journal of Respiratory and Critical Care Medicine* **202**:1304–1308. DOI: https://doi.org/10.1164/rccm.202007-2778LE, PMID: 32897730