Computational approaches to therapeutic nanobody design



Gemma Gordon

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2025

Acknowledgements

This DPhil would not have been possible without the generous support of many people. Firstly, I would like to thank Professor Charlotte Deane for her invaluable guidance and expertise, as well as my industrial supervisors across the pond at Twist Bioscience. I would also like to acknowledge the EPSRC and the SABS CDT for providing the studentship funding that enabled this research.

It has been a huge privilege to be part of the Oxford Protein Informatics Group. I would like to thank everyone who has collaborated with me for their contributions to the work in this thesis. I am especially grateful to those I have shared an office with – Lucy Vost, Brennan Abanades, Tobias Olsen, Isaac Ellmen, and Henriette Capel – whose company and sense of humour have made this experience all the more enjoyable.

I am incredibly fortunate to have worked alongside many talented women who have been both close friends and role models as women in science, including (again) Lucy Vost and Henriette Capel, as well as Kate Fieseler, Anna Carbery, Alissa Hummer, Yael Ziv, Nele Quast, Eve Richardson, Sarah Robinson, and Steph Wills. My time in OPIG has also been enriched by the support of friends and colleagues such as Guy Durant, Ísak Valsson, Fabian Spoendlin, Ollie Turnbull, Lewis Chinery, Matt Raybould, Alex Greenshields-Watson, and Matteo Cagiada.

To Dylan Adlard, thank you for your kindness, patience and support – I am hugely grateful that we were able to go through the last few years together. I would also like to thank my family, who have always ensured I had the opportunities that made pursuing this degree possible, and the friends and supporters who have encouraged me throughout my many years as a student. Finally, a special thank you to Dr Janette Shepherd and Dr Jessica Martin, my school biology and chemistry teachers, whose own achievements and belief in my abilities helped shape my path toward becoming a scientist.

Ownership of work

The research presented in this thesis is solely my work, unless specifically indicated otherwise. The use of the first-person plural is a stylistic choice. Any work conducted by others is properly attributed and acknowledged by name.

Publications

This DPhil led to the following publications and preprints:

Gordon, G.L., Capel H.L., Guloglu, B., Richardson, E., Stafford, R.L., and Deane, C.M. A comparison of the binding sites of antibodies and single-domain antibodies. *Frontiers in Immunology* **14**, 1231623 (2023).

Gordon, G.L., Raybould, M.I.J., Wong, A., and Deane, C.M. Prospects for the computational humanization of antibodies and nanobodies. *Frontiers in Immunology* **15**, 1399438 (2024).

Gordon, G.L., Greenshields-Watson, A., Agarwal, P., Wong, A., Boyles, F., Hummer, A., Lujan Hernandez, A.G., and Deane, C.M. PLAbDab-nano: a database of camelid and shark nanobodies from patents and literature. *Nucleic Acids Research*, **53**, D535–D542 (2025).

Greenshields-Watson, A., Agarwal, P., Robinson, S.A, Williams, B.H., **Gordon, G.L.**, Capel, H.L., Li, Y., Spoendlin, F.C., Aguilar Sanjuan, B., Boyles, F., and Deane, C.M. ANARCII: A Generalised Language Model for Antigen Receptor Numbering. *bioRxiv* (2025).

Gordon, G.L., Gervasio, J., Souders, C., and Deane, C.M. The Therapeutic Nanobody Profiler: characterising and predicting nanobody developability to improve therapeutic design. *bioRxiv* (2025).

Abstract

Nanobodies, or single-domain antibodies, have favourable characteristics that make them a promising novel therapeutic format. However, their unique structural and functional features also present distinct challenges for optimisation in therapeutic applications. Whilst computational approaches have significantly advanced the rational design of conventional antibodies, similar efforts for nanobodies remain limited. To fully realise the therapeutic potential of nanobodies, there is a pressing need for *in silico* methods tailored to their specific properties and associated challenges. Progress in this area is currently constrained by an incomplete understanding of nanobody behaviour in therapeutic contexts and the relatively limited availability of relevant data. In this thesis, we develop computational approaches to support the design of therapeutic nanobodies. We focus on characterising the binding and developability properties of nanobodies to establish a foundation for building bespoke tools that can identify, evaluate, and optimise novel candidates.

Firstly, we compare the binding sites of conventional antibodies and nanobodies to understand the structural basis of antigen recognition. By identifying residues most critical for binding, we provide a basis for multi-parameter optimisation, helping to distinguish positions that can be modified to improve developability without compromising specificity or affinity. To address the limited availability of nanobody data, we created PLAbDab-nano, a curated database of nanobody sequences and modelled structures sourced from patents and academic literature. We envisage that this resource can be used as a foundation for designing diverse nanobody libraries, for benchmarking and improving existing scaffolds by comparison to therapeutically relevant or optimised sequences, and for the development of computational models trained specifically on nanobody data. The PLAbDab-nano database was used to characterise developability traits in nanobodies, leading to the development of the Therapeutic Nanobody Profiler, a computational tool that evaluates developability using sequence- and structure-based features tailored to the unique properties of nanobodies. Collectively, this work aims to enable rational design and multi-parameter optimisation of therapeutic nanobodies by combining structural analysis, improved data availability, and computational tools specific to this novel format.

Contents

List of Figures **xv**

List of Tables **xxi**

1 Introduction **1**

1.1 Abstract . . . 2
1.2 Immune proteins . . . 3
1.2.1 Antibodies . . . 3
1.2.2 Nanobodies . . . 9
1.3 Experimental approaches to therapeutic design . . . 14
1.3.1 Antibodies and nanobodies as therapeutics . . . 14
1.3.2 Identifying binders . . . 15
1.3.3 Optimising developability . . . 16
1.4 Computational approaches to therapeutic design . . . 20
1.4.1 Sequence data . . . 20
1.4.2 Structural data . . . 22
1.4.3 Databases . . . 23
1.4.4 Developability . . . 24
1.4.5 Multi-parameter optimisation . . . 28
1.5 Thesis outline . . . 30

2 A comparison of the binding sites of antibodies and camelid nanobodies **31**

2.1 Abstract . . . 32
2.2 Introduction . . . 33
2.3 Methods . . . 34
2.3.1 Sequence datasets . . . 34
2.3.2 Structure datasets . . . 35
2.3.3 Numbering schemes . . . 36
2.3.4 Binding site definitions . . . 36
2.3.5 Amino acid composition . . . 38
2.3.6 Epitope accessibility . . . 38

2.3.7 Canonical forms of the CDRs . . . 39
2.3.8 Structural clustering . . . 39
2.3.9 Orientation of CDR3 loops . . . 40
2.3.10 Statistical tests . . . 40
2.4 Results . . . 41
2.4.1 The nanobody CDR3 loop is longer than the antibody CDR-H3 loop . . . 41
2.4.2 Structural clustering shows a separation between antibody and nanobody CDR structures . . . 42
2.4.3 Nanobodies and antibodies have more identical CDR sequences than expected by chance . . . 44
2.4.4 Paratopes of nanobodies and antibodies show small differences in their amino acid compositions . . . 45
2.4.5 Nanobody paratopes are significantly smaller than those of antibodies . . . 45
2.4.6 Epitopes targeted by nanobodies and antibodies have similar amino acid compositions . . . 47
2.4.7 Epitopes targeted by antibodies are more linear than those targeted by nanobodies . . . 47
2.4.8 Epitopes targeted by nanobodies and antibodies are of comparable size . . . 49
2.4.9 Epitopes targeted by nanobodies and antibodies are of similar accessibility . . . 50
2.4.10 CDR3 loop length does not correlate with epitope accessibility 51
2.4.11 Antibodies and nanobodies target epitopes of similar accessibility due to packing of nanobody CDR3 loops against the VHH domain . . . 52
2.4.12 Nanobodies establish more interactions with their epitope per paratope residue than antibodies . . . 56
2.4.13 Hydrophobic interactions dominate both nanobody-antigen and antibody-antigen complexes . . . 57
2.4.14 CDR3 and framework residues are of increased importance for interactions in the nanobody-antigen complex . . . 58
2.4.15 Interacting framework residues are often conserved in nanobodies . . . 61
2.4.16 Antibodies and nanobodies can bind the same epitopes but interact with them differently . . . 62
2.5 Discussion . . . 64

3 PLAbDab-nano: a database of camelid and shark nanobodies from patents and literature 67
3.1 Abstract . . . 68
3.2 Introduction . . . 69
3.3 Methods . . . 72
3.3.1 Collecting nanobody sequences . . . 72
3.3.2 Annotation of VHH and VNAR sequences . . . 74
3.3.3 Searching PLAbDab-nano . . . 76
3.4 Results . . . 78
3.4.1 Database statistics . . . 78
3.4.2 VNAR data in PLAbDab-nano provides CDR3 diversity . . . 82
3.5 Discussion . . . 84

4 The Therapeutic Nanobody Profiler: characterising and predicting nanobody developability to improve therapeutic design 87
4.1 Abstract . . . 88
4.2 Introduction . . . 89
4.3 Methods . . . 91
4.3.1 Sequence data . . . 91
4.3.2 Structural data . . . 92
4.3.3 Numbering schemes . . . 93
4.3.4 Computed descriptors for developability properties . . . 93
4.3.5 Describing CDR3 loop conformations . . . 94
4.3.6 Interactions between the CDR3 loop and nanobody tetrad residues . . . 95
4.3.7 *In vitro* developability assays . . . 95
4.4 Results . . . 101
4.4.1 Distributions of the TAP metrics differ between antibody and nanobody clinical data . . . 103
4.4.2 Total CDR length does not sufficiently capture nanobody paratope shapes . . . 103
4.4.3 More compact CDR3 loops do not show significant conformational change upon binding . . . 106
4.4.4 Nanobody subtypes based on CDR3 loop conformations show associated tetrad motifs . . . 107
4.4.5 Hydrophobic interactions between aromatic residues often occur between nanobody tetrad residues and more compact CDR3 loops . . . 108
4.4.6 *In vitro* assays support the presence of two equally developable nanobody subtypes . . . 109

4.4.7 TNP properties 109
4.4.8 Developability guidelines for TNP are set by 36 clinical-stage nanobody sequences and predicted structures 110
4.4.9 *In vitro* and *in silico* methods provide complementary approximations of developability 110
4.4.10 Web application 114
4.5 Discussion 117

5 Conclusions and future work **121**
5.1 Conclusions 122
5.2 Future work 123

Appendices

A A comparison of the binding sites of antibodies and camelid nanobodies **127**

B PLAbDab-nano: a database of camelid and shark nanobodies from patents and literature **149**

C The Therapeutic Nanobody Profiler: characterising and predicting nanobody developability to improve therapeutic design **155**

References **177**

List of Figures

1.1 There are five main immunoglobulin isotypes present in mammals: IgG, IgM, IgA, IgE and IgD. 6
1.2 The general structure of a mammalian IgG antibody consists of four polypeptide chains arranged in a Y-shaped structure. 8
1.3 Convergent evolution of heavy-chain-only antibodies in camelid and shark species. 10
1.4 Nanobodies can be derived from the VHH domains of camelid heavy-chain antibodies. 12
1.5 Nanobodies can also be derived from the VNAR domains of shark IgNARs. 13
1.6 Humanisation methods aim to reduce non-humanness without compromising functionality. 19
1.7 The framework regions and CDR loops of an immunoglobulin variable domain annotated according to the IMGT numbering scheme. . . . 21

2.1 Visualising the differences between the SASA-defined, distance-defined and interactions-defined epitopes and paratopes. 37
2.2 Comparison of the distributions of CDR3 loop lengths of nanobodies and antibody VH domains with sequence and structural data 42
2.3 The paratopes of nanobodies tend to contain fewer residues than the paratopes of antibodies. 46
2.4 The epitopes targeted by antibodies are relatively more linear than the epitopes targeted by nanobodies. 48
2.5 Nanobodies are able to target epitopes of equal size (as defined by number of residues) to those targeted by conventional antibodies. . 49
2.6 Epitopes targeted by nanobodies are slightly less accessible than those targeted by antibodies. 50
2.7 There is no correlation between the length of the CDR3/CDR-H3 loop and the accessibility of the epitope surface for either nanobodies or antibodies. 52
2.8 The orientation of the CDR3 loops of nanobodies suggests why there is an overlap in the accessibility of the epitopes targeted by nanobodies and antibodies. 53

2.9 Relationships between spherical angles and compactness suggest that the paratope is stabilised by the CDR-H3 loop packing against VL domains in antibodies, or the CDR3 loop packing against the rest of the VHH domain in nanobodies. 55
2.10 The distributions of the number of interactions initiated by nanobody and antibody paratopes demonstrate that nanobody paratopes establish significantly more interactions per residue than antibody paratopes. 56
2.11 Hydrophobic interactions dominate across nanobody-antigen and antibody-antigen complexes. 58
2.12 Assessing the relative contributions of each CDR loop to the paratope shows that for both nanobodies and antibodies, the CDR3 and CDR-H3 loops rarely do not contribute interactions to the paratope. . . . 59
2.13 Distributions of the proportion of CDR residues in the paratope and the proportion of interactions from CDR residues across the whole paratope indicates that, in nanobodies, more framework residues are involved in binding the epitope. 60
2.14 Sequence logo plots of amino acid composition at framework positions often involved in the paratopes of antibodies and nanobodies suggest that framework residues identified to often occur in the paratope are highly conserved in nanobodies. 62
2.15 A nanobody and antibody are able to bind the SARS-CoV-2 RBD with overlapping epitopes. 63

3.1 Nanobodies are derived from the antigen-binding portion of heavy-chain antibodies (hcAbs), which lack the light chain pairing possessed by conventional antibodies 70
3.2 PLAbDab-nano is freely accessible via a web application. 72
3.3 Data is scraped and processed from GenBank, SAbDab, and TheraSAbDab to generate the PLAbDab-nano database 73
3.4 An example of the PLAbDab-nano sequence search functionality on the web application. 76
3.5 An example of the PLAbDab-nano text search functionality on the web application. 77
3.6 There is an increasing number of depositions of nanobody sequences per year to publicly available sources. 78
3.7 The distribution of classifier scores for sequences from patents and literature in the INDI database indicates that the majority of entries would not pass our filtering methods. 79

3.8 A breakdown of entries in PLAbDab-nano from different sources and species. 80
3.9 A comparison of CDR3 length distributions of VHH and VNAR sequences in PLAbDab-nano. 81
3.10 Low pairwise percentage sequence identities for VHH CDR3 loops versus VNAR CDR3 loops indicate that the two types of nanobody occupy different regions of sequence space. 82
3.11 VNARs and VHHs binding to overlapping epitopes on the receptor-binding domain of SARS-CoV-2 83

4.1 CDR3 conformations can be classified as either 'kinked' or 'extended'. 95
4.2 Distributions of the TAP metrics on our 8 datasets of nanobodies and antibody VH chains from different sources. 102
4.3 Loop compactness provides a continuous descriptor for CDR3 conformation. 104
4.4 Distributions of CDR3 length and CDR3 compactness on our 8 datasets of nanobodies and antibody VH chains from different sources. 105
4.5 Pairwise comparison of compactness scores for bound and unbound nanobody crystal structures 106
4.6 Amino acid frequencies across the nanobody tetrad positions show greater variation per position in nanobody loops of lower compactness versus higher compactness. 107
4.7 Hydrophobic interactions between aromatic residues in the CDR3 loop and the first nanobody tetrad position often occur in a more compact CDR3 loop which tends to be folded over the FR2 region. 108
4.8 Distributions of values for each TNP metric for nanobody clinical-stage and proprietary data, coloured by flags assigned according to the thresholds set by the clinical-stage data. 111
4.9 The web application for TNP allows users to input a VHH sequence to generate results for the metrics we have described 114
4.10 The main results page for the TNP web application. 115
4.11 The user can also view an interactive visualisation of the patches of surface hydrophobicity and charge on their modelled structure. . . . 116
4.12 The user is further provided with visualisations of the sequence liabilities and nanobody tetrad residues which can be used to inform design strategies. 117

A.1 An epitope identity cut-off of 75% was used for reintroducing complexes in the dataset. 130
A.2 Binding site definitions based on interacting residues are stricter. . . 131

A.3 A visualisation of the coordinate system used to determine the orientation of the heavy chain CDR3 loops. 132
A.4 Sequence logo plots of the amino acid composition at each position for each of the CDR loops shows similarities between nanobody sequences and antibody VH sequences from the IGHV3 germline. . 133
A.5 Distance-defined paratopes of nanobodies and antibodies show small differences in their amino acid compositions. 134
A.6 Interactions-defined paratopes of nanobodies and antibodies show small differences in their amino acid class compositions. 135
A.7 Comparing the distributions of paratope size for nanobodies and antibody complexes using the SASA-defined paratope shows that the nanobody paratopes are on average significantly smaller 136
A.8 Distance-defined epitopes of nanobodies and antibodies show similar amino acid compositions. 137
A.9 Interactions-defined epitopes of nanobodies and antibodies show similar amino acid class compositions. 138
A.10 The epitopes targeted by antibodies are relatively more linear than the epitopes targeted by nanobodies for distance-defined and interactions-defined epitopes. 139
A.11 The SASA-defined epitopes targeted by antibodies are relatively more linear than the SASA-defined epitopes targeted by nanobodies. . . . 140
A.12 Comparing the distributions of SASA-defined epitope size for nanobodies and antibodies shows that there is no significant difference between epitopes targeted by nanobodies and antibodies 141
A.13 The mean number of interactions per CDR loop, normalised by loop length per complex. 142
A.14 The CDR3 loop is of increased importance in nanobodies. 143
A.15 When comparing the nanobody paratope to the paratope residues from the antibody VH chains only, relatively fewer interactions are contributed by the CDRs in nanobodies, indicating that there is greater involvement of framework residues in the nanobody paratope. 144

B.1 A Hidden Markov Model built from VNAR sequences provides a method to distinguish between VNAR and non-VNAR sequences. . 150
B.2 Distributions of positive (VHH) and negative (VH) data were used to identify a rational threshold for the VH/VHH classifier, which was used to include or exclude entries from GenBank in PLAbDab-nano. 151
B.3 Distributions of classifier scores for sequences from the INDI database, split by source into data from patents, GenBank, structures, and literature. 152

C.1 Distributions of scores for patches of positive charge across all of our datasets, determined by sampling with replacement 156
C.2 Distributions of scores for patches of negative charge across all of our datasets, determined by sampling with replacement 157
C.3 Distributions of scores for patches of surface hydrophobicity across all of our datasets, determined by sampling with replacement 158
C.4 Distributions of total CDR length across all of our datasets, determined by sampling with replacement 159
C.5 Correlations between the reach and length of the CDR3 loop for the different datasets analysed in this study. This relationship can be described by loop compactness. 160
C.6 Predicted structures for the 36 clinical-stage nanobody sequences in order of increasing compactness, annotated with their nanobody tetrad motif 161
C.7 CDR3 conformations can be classified as kinked or extended according to the α and τ angles between backbone atoms in the CDR3 loop . 162
C.8 Assessing the agreement between compactness and the kinked-extended classifications 163
C.9 The bimodal distribution of compactness can be used to separate nanobodies into two structural subtypes. 164
C.10 Distributions of CDR3 length across all of our datasets, determined by sampling with replacement 165
C.11 Distributions of CDR3 compactness across all of our datasets, determined by sampling with replacement 166
C.12 Change in compactness is minimal between bound and unbound pairs of nanobody CDR3 loop crystal structures 167
C.13 Amino acid frequencies across the nanobody tetrad positions for nanobodies versus antibody VH chains shows that FERG is the most common nanobody motif, whilst VGLW is the most common in antibody VH chains. 168
C.14 Comparing the relative frequencies of atom-atom and atom-plane interactions between pairs of residues in the CDR3 loop and the nanobody tetrad positions shows that hydrophobic, aromatic and carbon-π interactions are the most common for our data. 169
C.15 The top 10 most common interacting CDR3-tetrad residue pairs in decreasing order of relative frequency, across all of our nanobody datasets, split by compactness. 170
C.16 The ranges of experimental assay values split into nanobodies with CDR3 loops of higher and lower compactness. 171

C.17 A heatmap of the Spearman's rank correlations calculated pairwise for our extended list of computational descriptors, comparing across all of our nanobody datasets combined. 172
C.18 A heatmap of the Spearman's rank correlations for each computational descriptor with each *in vitro* assay, for the 108 clinical-stage and proprietary nanobodies that were experimentally tested. 173

List of Tables

2.1 Clustering the structures of the CDR loops of nanobodies and antibody VH domains based on length and conformational similarity 43

4.1 Total sequences and modelled structures in each dataset. 92
4.2 Flagging thresholds for each of the TNP metrics are set according to the distributions of the 36 nanobody clinical-stage data-points. . . . 110
4.3 Comparison of proprietary candidates that were assigned red flags by TNP to their *in vitro* assay data demonstrates that *in silico* and *in vitro* measures provide different approximations for developability. 112

A.1 Filtering steps for the curation of the antibody and nanobody structural datasets 145
A.2 For the structural datasets, greater species variation is observed in antibodies than nanobodies. 146
A.3 The most common framework residues observed in the interactions-defined paratope for at least 10% of nanobodies. 147

B.1 Breakdown of total sequences in PLAbDab-nano by type of nanobody and source. 153
B.2 VH sequences collected from GenBank and patents for the INDI database may be misclassified as nanobodies. 154

C.1 The pairs of PDB IDs used to compare CDR3 compactness between bound and unbound nanobodies. 174
C.2 The additional *in silico* descriptors calculated for our nanobody datasets. 175
C.3 Flag assignment and raw values for the TNP metrics and five *in vitro* assays across the 36 clinical-stage nanobody sequences. 176

1 Introduction

Contents

1.1 Abstract 2
1.2 Immune proteins 3
1.2.1 Antibodies 3
1.2.2 Nanobodies 9
1.3 Experimental approaches to therapeutic design 14
1.3.1 Antibodies and nanobodies as therapeutics 14
1.3.2 Identifying binders 15
1.3.3 Optimising developability 16
1.4 Computational approaches to therapeutic design . . . 20
1.4.1 Sequence data 20
1.4.2 Structural data 22
1.4.3 Databases 23
1.4.4 Developability 24
1.4.5 Multi-parameter optimisation 28
1.5 Thesis outline 30

1.1 Abstract

This introduction provides an overview of the main concepts encountered in this thesis, summarising the basis for therapeutic nanobody design. Firstly, I will describe the role of antibodies in vertebrate immune systems and the evolution of heavy-chain-only antibodies, from which nanobodies can be derived. I will review how nanobody structures differ from antibodies and how the properties of immunoglobulins make them well-suited for use as biotherapeutics. Next, I will outline conventional experimental methods for the discovery and optimisation of therapeutic antibodies and nanobodies. Finally, I will introduce the developments in computational approaches that aim to accelerate this process and highlight the current progress and challenges for therapeutic nanobody design.

Sections of the text in this chapter are based on the following publications:

> **Gordon, G.L.**, Capel H.L., Guloglu, B., Richardson, E., Stafford, R.L., and Deane, C.M. A comparison of the binding sites of antibodies and single-domain antibodies. *Frontiers in Immunology* **14**, 1231623, (2023).
>
> **Gordon, G.L.**, Greenshields-Watson, A., Agarwal, P., Wong, A., Boyles, F., Hummer, A., Lujan Hernandez, A.G., and Deane, C.M. PLAbDab-nano: a database of camelid and shark nanobodies from patents and literature. *Nucleic Acids Research* **53**, D535–D542, (2025).
>
> **Gordon, G.L.**, Raybould, M.I.J., Wong, A., and Deane, C.M. Prospects for the computational humanization of antibodies and nanobodies. *Frontiers in Immunology* **15**, 1399438, (2024).
>
> **Gordon, G.L.**, Gervasio, J., Souders, C., and Deane, C.M. The Therapeutic Nanobody Profiler: characterising and predicting nanobody developability to improve therapeutic design. *bioRxiv* (2025).

1.2 Immune proteins

1.2.1 Antibodies

1.2.1.1 Adaptive immune systems

B-cells and antibody production

The immune response of jawed vertebrates comprises two main components: innate and adaptive immunity. These are activated in response to an antigen: a protein or molecule recognised by the immune system as foreign or pathogenic. Innate immunity provides immediate, non-specific defence against a broad spectrum of pathogens. Adaptive immunity mediates antigen-specific responses and supports the development of immunological memory, enabling faster and improved protection on re-exposure to the same pathogen [1]. Central to this system are lymphocytes, a subset of white blood cells derived from pluripotent haemopoietic stem cells. These differentiate into B-lymphocytes (B-cells), which develop and mature in the bone marrow, and T-lymphocytes (T-cells), which mature in the thymus [1]. The humoral arm of adaptive immunity is mediated by B-cells, which produce antibodies (immunoglobulins). These are Y-shaped globular proteins which can bind antigens with high specificity and affinity, promoting pathogen elimination through a range of effector mechanisms.

The B-cell repertoire and V(D)J recombination

Genetic recombination during B- and T-cell development in the bone marrow and thymus generates unique antigen receptors on each naïve lymphocyte [1]. The diversity of the antibody repertoire arises from recombination during B-cell development of variable (V), diversity (D), and joining (J) gene segments within the immunoglobulin (Ig) loci [1, 2]. These segments are assembled at three distinct loci: the immunoglobulin heavy chain (IGH) on chromosome 14, the kappa light chain (IGK) on chromosome 2, and the lambda light chain (IGL) on chromosome 22. Each locus also includes a set of constant (C) gene segments. During V(D)J recombination, one segment from each gene family (V, D, and J) is randomly

selected. Double-strand breaks are introduced adjacent to each segment. The segments are then ligated via the non-homologous end joining (NHEJ) pathway [3]. Junctional diversity is further expanded by the imperfect alignment of DNA ends during ligation introducing insertions and deletions (indels), broadening the antigen receptor repertoire beyond the combinatorial joining of gene segments [1, 3, 4]. As a result, the adaptive immune system has a highly diverse antibody repertoire, enabling specific recognition of virtually any antigen. The naïve B-cell repertoire in humans is incredibly diverse; it is estimated that it can generate at least 10^{13} unique antibody sequences [5, 6].

Clonal selection

The adaptive immune system is activated in the presence of an antigen. Immature dendritic cells, which function as antigen-presenting cells, ingest and process the antigen to display antigenic peptides on their cell surface. These dendritic cells then migrate via the lymphatic system to regional lymphoid tissues, such as the lymph nodes, spleen, or mucosa-associated lymphoid tissues. There, the antigen is recognised by naive T-cells expressing cognate antigen-specific receptors [1].

Following antigen recognition, T-cells then proliferate into antigen-specific effector cells, activating B-cells. These B-cells stop migrating and enlarge into plasmablasts. The plasmablasts proliferate into clone daughter cells, which differentiate into antibody-producing plasma cells [1, 7]. This is known as clonal selection and expansion: B-cells produce antibodies of a uniform clonotype, corresponding to the antigen specificity determined by T-cell activation.

Somatic hypermutation, affinity maturation and isotype switching

In response to activation by an antigen, some proliferating B-cells migrate to germinal centres within secondary lymphoid tissues, such as the lymph nodes or the spleen [1]. Germinal centres are micro-environments that form when B-cells are activated and migrate into secondary lymphoid follicles [8]. They proliferate to create localised regions of cell division surrounded by resting (mature, naïve, unactivated) B-cells.

In the germinal centre, somatic hypermutation introduces rapid variation in the immunoglobulin V genes [1, 9]. This process occurs at a much higher rate (approximately 10^{-3} mutations per base pair per generation) than typical spontaneous mutations in somatic cells [10]. This results in affinity maturation, where B-cells accumulate point mutations in the immunoglobulin genes. In a Darwinian process, only B-cells that produce high-affinity antibodies are selected for and retained [8]. B-cells in the germinal centres may differentiate into plasma cells that secrete antibodies, or memory B-cells which enable long-term immunological memory, ensuring a rapid response on subsequent antigen exposures [11].

Isotypes and effector mechanisms

Of the antibodies produced by the B-cells, there are several isotypes, which are structural subclasses determined by the constant region of the antibody heavy chain. In mammals there are five main isotypes: IgG, IgM, IgA, IgE and IgD [1] (**Figure 1.1**). The Fc region of the antibody differentiates these isotypes and defines their functional properties. Isotype switching, or class switch recombination, can occur in the germinal centres, allowing B-cells to switch from producing IgM – the first antibody to be produced during the humoral immune response – to another isotype [12].

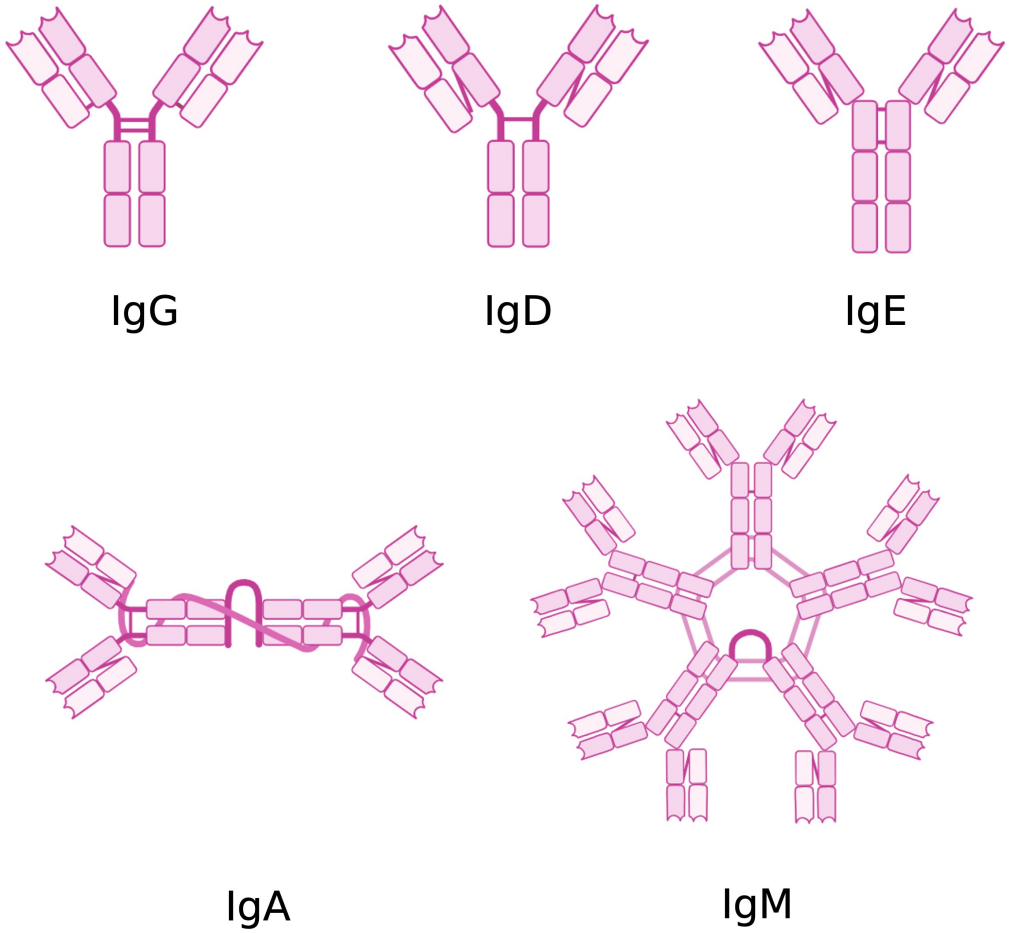


Figure 1.1: *There are five main immunoglobulin isotypes present in mammals: IgG, IgM, IgA, IgE and IgD. These can be differentiated by their Fc region and their effector mechanisms. IgG, IgD and IgE exist as monomers, IgA is most prevalent as a dimer and IgM as a pentamer.*

Different antibody isotypes are distributed across various compartments of the body and, whilst they do not alter antibody specificity, they trigger distinct effector mechanisms to eliminate pathogens upon antigen recognition [1]. These effector functions include the neutralisation of pathogens, such as blocking viral entry or inhibiting bacterial toxin binding to host cells [13]. Antibodies also promote opsonisation, coating pathogens to enhance their recognition by phagocytic cells (such as macrophages and neutrophils) via Fc receptors. Different Fc receptors, recognising different isotypes, mediate this process [1]. Additional mechanisms

include antibody-dependent cellular cytotoxicity, where natural killer (NK) cells kill antibody-coated target cells [14], or antibody-dependent cellular phagocytosis, facilitating pathogen engulfment by phagocytes [15]. Alternatively, antibodies can activate the classical complement pathway, leading to pathogen lysis via complement-dependent cytotoxicity [16].

1.2.1.2 Antibody structure and binding

IgG general structure

Of the immunoglobulin isotypes, IgG is the most abundant in humans [17]. There are four subclasses of IgG: IgG1, IgG2, IgG3 and IgG4. These have slight variations in their hinge regions and CH2 domains, but otherwise adhere to a common structure [17]. Typically, IgG antibodies have a molecular weight of approximately 150 kDa and consist of four polypeptide chains, arranged in a Y-shaped structure: two heavy chains and two light chains, each containing a variable and constant domain [1]. The heavy chains are further linked to the Fc (crystallisable fragment) region and are encoded by V, D, and J gene segments, while the light chains are encoded by only V and J segments. The variable and constant regions of both chains form the Fab (antigen binding fragment) region, while the Fv (variable fragment) region consists of only the variable domains of the heavy and light chains [1] (**Figure 1.2**).

Variable domains

The variable domains of both the heavy and light chains (VH and VL) contain four framework regions (FR1-4), which form a stable scaffold on which the hypervariable complementarity-determining region (CDR) loops are located (**Figure 1.2**). The VH and VL domains each contain three CDR loops: CDR-H1, CDR-H2 and CDR-H3, and CDR-L1, CDR-L2 and CDR-L3 respectively.

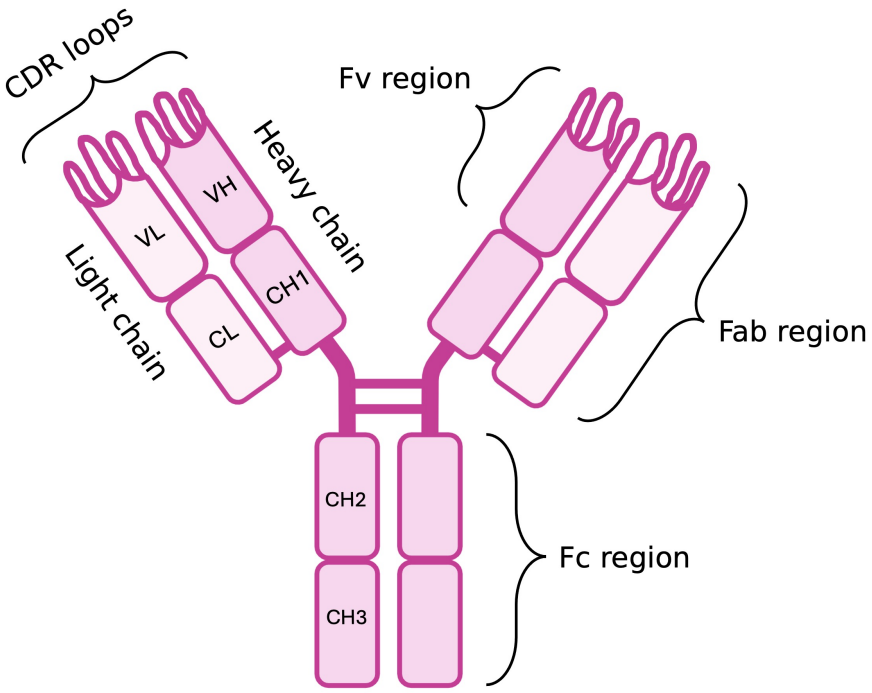


Figure 1.2: *The general structure of a mammalian IgG antibody consists of four polypeptide chains arranged in a Y-shaped structure. There are two heavy chains and two light chains, each containing a variable and constant domain. The heavy chains are further linked to the Fc region. The variable domains are composed of framework regions, which act as a scaffold for three CDR loops which determine binding affinity and specificity.*

The sequence diversity of the CDR loops is primarily generated through V(D)J recombination and somatic hypermutation [18]. The CDR-H3 loop sequence spans the junctions of the V, D, and J gene segments [19], and therefore is the most diverse in terms of amino acid composition and length. This allows the CDR-H3 to adopt a greater range of conformations [2, 20] than the heavy and light chain CDR1 and CDR2 loops. These are encoded by only the V gene segments, and alongside the CDR-L3, tend to fall within a more limited set of structural conformations known as canonical forms [21–24]. Due to the proximity of the six CDR loops at the interface of the VH and VL chains, the CDR-H3 conformation is constrained to some extent by the length and orientation of the light chain CDR loops [25].

Antibody binding sites and protein-protein interactions

The diversity of the CDR loops, especially CDR-H3, play an essential role in antibody specificity, as they make up the majority of the antibody binding site [26, 27]. The binding site on the antibody is known as the paratope. The paratope interacts with the epitope, the region on the antigen surface that is recognised by the antibody. The binding of the paratope and epitope is mediated by non-covalent interactions such as weak van der Waals forces, hydrophobic interactions, electrostatic attraction, hydrogen bonding or interactions between aromatic rings. The nature of these interactions depends on the properties of the amino acid side chains involved, which can be broadly classified as hydrophobic or hydrophilic based on their polarity, and as acidic or basic based on their charge.

1.2.2 Nanobodies

Convergent evolution of camelid and shark immune systems

The immune systems of camelid species and sharks have convergently evolved heavy-chain antibodies (hcAbs), which exist without a light-chain partner, in addition to conventional antibodies [28–31] (**Figure 1.3**). In sharks, these heavy-chain antibodies are called immunoglobulin new antigen receptors (IgNARs).

Nanobodies are derived from the antigen-binding fragment of heavy-chain antibodies. There are two varieties of nanobody: the VHH (the Variable Heavy domain of the Heavy chain) which is derived from the camelid hcAbs, and the VNAR (Variable New Antigen Receptor), which is derived from shark IgNARs. Due to their more recent shared evolutionary ancestry, camelid species such as llamas, alpacas, and camels produce VHH domains that retain homology with the VH domain of conventional IgG antibodies found in humans [32]. Sharks (and other cartilaginous fish) are the earliest jawed vertebrates to have evolved an adaptive immune system as observed in mammals. As such, the greater evolutionary distance between VNARs and

VHHs or the VH domains of conventional antibodies means that sequence and structural similarities are relatively reduced [33, 34].

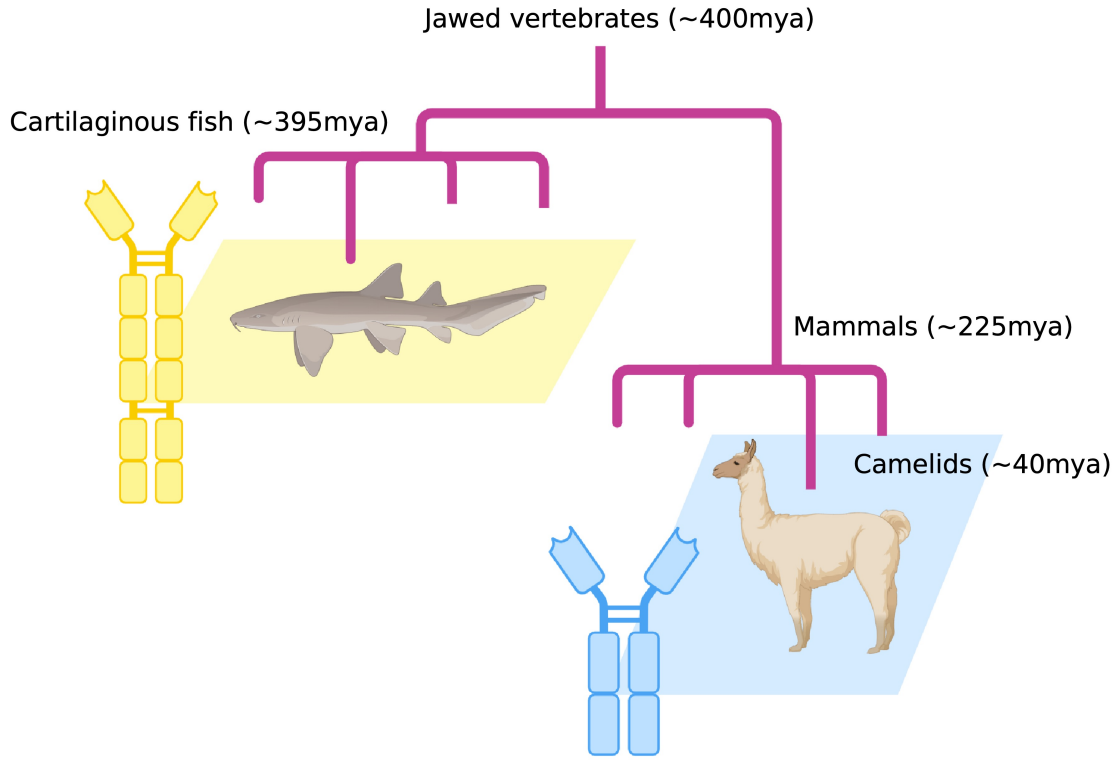


Figure 1.3: *The evolution of camelid and shark species occurred approximately 350 million years apart from each-other, however convergent evolution has led to them both having adaptive immune systems that produce heavy-chain-only antibodies.*

1.2.2.1 Nanobody sequence and structure

VHH and VNAR sequence diversity

As in conventional antibodies of jawed vertebrates, the diversity of VHH domains arises through the process of V(D)J recombination. VHHs are derived from the same gene loci as conventional VH domains, but there is a V(D)J gene usage bias towards V genes within the IGHV3 subgroup [35–38]. These particular gene segments have evolved to support the unique structure and function of heavy-chain-only antibodies: for example, they often show mutations at four positions in the framework-2 (FR2) region, known as hallmark residues or the nanobody tetrad, which are theorised to

confer improved solubility and stability [35, 37].

For shark IgNARs, generating diversity relies on a specialised form of V(D)J recombination. Unlike mammals, where V, D and J gene segments are drawn from large pools and recombined freely, sharks use a more compact, modular system. Their genome contains multiple independent IgNAR gene clusters, each made up of a V gene segment, three D gene segments and a J gene segment (and one set of C gene segments). These undergo four rearrangement events to form the VNAR domain [39, 40]. This concentrates much of the diversity within the CDR3 loop. Sharks compensate for the reduced combinatorial diversity through extensive somatic hypermutation [41].

VHH structure

VHH and VNAR domains exhibit distinct structural features compared to conventional antibodies. The VHH is a single domain with a molecular weight of approximately 15 kDa [42]. It forms an immunoglobulin fold with three CDR loops (**Figure 1.4**). The CDR3 of VHHs are on average significantly longer than that of conventional antibodies [43–47]. This extended length allows for a greater diversity of CDR3 conformations, as it is not constrained by the presence of a light chain [38, 48, 49]. These loop conformations generally fall into two categories: extended, where the CDR3 extends away from the framework, or kinked, where the loop folds downward over the FR2 region [21, 38, 50].

Most VHHs have a canonical disulfide bond between framework regions 1 and 3 (between cysteines at positions 23 and 104 according to IMGT numbering [51] - see section **1.4.1.1**). Some also feature a non-canonical disulfide bond between the CDR1 and CDR3 loops, which may enhance thermostability [36, 52]. Furthermore, VHHs often show hallmark residue mutations which typically have more hydrophilic properties, to replace hydrophobic residues which are buried by the light chain in antibodies but are now solvent-exposed [48, 53]. Though typically the three CDR

loops make up the majority of the binding site, framework residues on the surface have also been shown to be incorporated into the paratope more frequently than seen in conventional antibodies [54, 55].

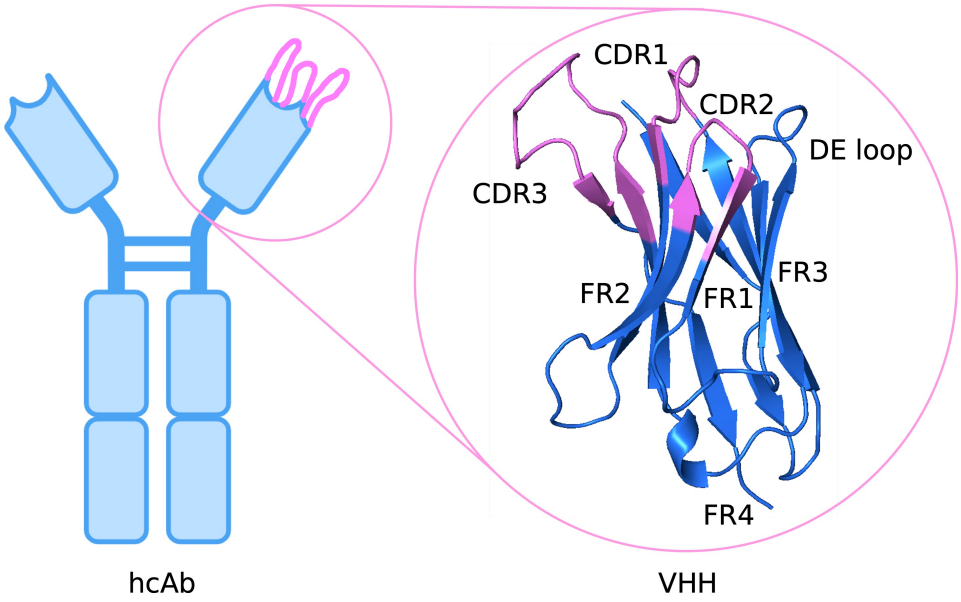


Figure 1.4: *Nanobodies can be derived from the VHH domains of camelid heavy-chain antibodies. They share structural homology with the VH domains of conventional antibodies, consisting of an immunoglobulin fold with four framework regions (shown in blue) and three CDR loops (shown in pink). The CDR3 loop tends to be relatively longer in VHH domains than the conventional antibody CDR-H3 and can adopt either kinked or extended conformations.*

VNAR structure

VNARs possess a structurally analogous scaffold to VHHs, characterised by an immunoglobulin-like fold with a truncated framework. Unlike conventional antibodies and VHHs, VNARs lack a CDR2 loop and instead contain only CDR1 and CDR3 regions [56–58]. Similar to VHHs, the CDR3 in VNARs is often relatively long compared to the CDR-H3 of conventional antibodies [56–58]. Additionally, VNARs contain two unique hypervariable regions, HV2 and HV4, that contribute to antigen binding and structural diversity [56–58] (**Figure 1.5**).

VNARs can be categorised into four structural types (I–IV) based on the number and location of cysteine residues, and the presence of canonical and non-canonical disulfide bonds [59, 60]. Type I contains one canonical and two non-canonical disulfide bonds, and may be specific to nurse sharks. Type II is the most common across studied shark species (making up ~57% of VNARs in nurse sharks [61]) and has four cysteines, with two extra cysteines in CDR1 and CDR3 that form a disulfide bond. Type III VNARs resemble Type II, but are uniquely present in neonatal sharks and differ in sequence composition [58–60]. Type IV only has one canonical disulfide bond contained within the framework regions; it tends to be the most rare type and the lack of non-canonical cysteines may confer greater binding flexibility [57, 58, 60]. Typically, the long CDR3 is dominant in the binding site of VNARs, and this is used in combination with the CDR1, HV2 and HV4 loops to interact with an antigen.

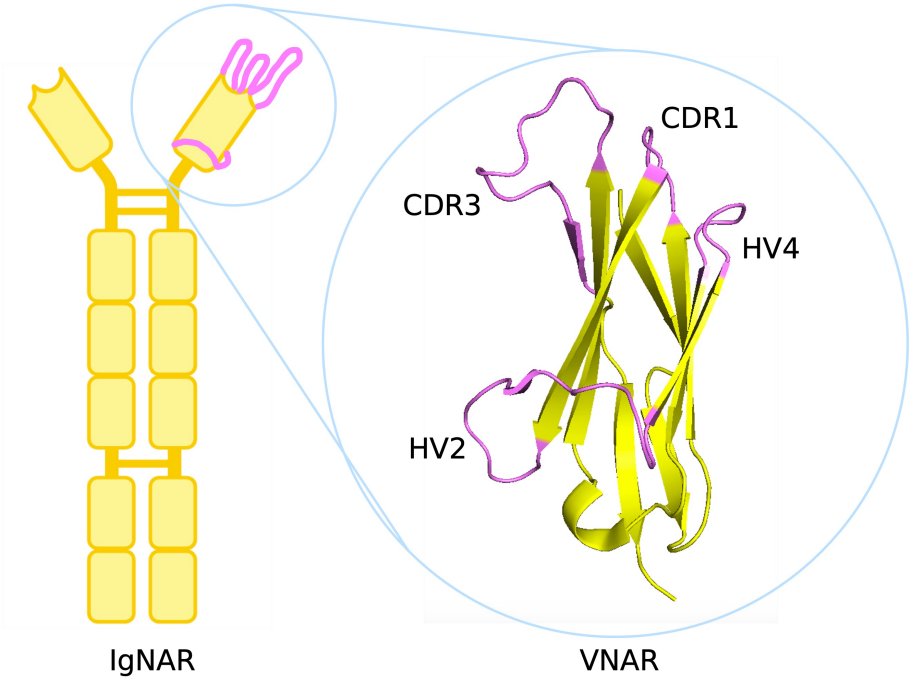


Figure 1.5: *Nanobodies can also be derived from the VNAR domains of shark IgNARs. Although they have a similar scaffold, VNARs have a truncated framework with only a CDR1 and CDR3 loop (shown in pink), the latter of which is relatively long (similar to VHHs) and a significant source of diversity. In addition, VNARs have hypervariable loops, HV2 and HV4 (shown in pink), which are used in binding antigens.*

1.3 Experimental approaches to therapeutic design

1.3.1 Antibodies and nanobodies as therapeutics

The ability of immunoglobulins to bind with high affinity to a wide diversity of antigens makes them an attractive modality for use as biotherapeutics. Currently, there are over 200 approved therapeutic antibodies and almost 1400 in total in various stages of clinical trials [62, 63]. These treat a range of cancers, viruses and autoimmune diseases, and cover a range of formats including monoclonal antibodies, single-chain variable fragments (scFvs) and multi-specifics, which can be used to target multiple epitopes.

However, the high molecular weight of monoclonal antibodies can cause high production costs as well as poor diffusion rates that limit tissue penetration [64–66]. These issues have led to an increased interest in recent years in smaller antibody fragments such as nanobodies. Research towards therapeutic nanobodies has focused predominantly on VHHs which have advantageous properties such as their relatively small size, high solubility and thermostability [65, 67, 68]. Given their smaller size, the structural diversity available to VHHs is significantly reduced compared to antibodies. However, VHHs have been shown to achieve comparable binding specificities and affinities [30, 54]. Furthermore, VHHs have shown higher blood clearance and tissue penetration than standard antibodies [65, 67, 68].

Although interest in VNARs as therapeutics has only emerged more recently, these domains are similarly small in size (~12kDa) and are able to adopt unique paratope shapes given their relatively long CDR3 [69]. In addition, VNARs have been shown to be highly thermostable [56, 70, 71]. VNARs are currently understudied in therapeutic applications relative to VHHs, however they have been evaluated against therapeutic targets for cancer and malarial proteins [72, 73]. Furthermore VNARs show promising results for use in multi-specifics to target the COVID-19

receptor-binding domain and PD-L1 receptor for cancer therapies [74, 75].

Although the development of VNARs is still in preclinical stages [76], to date, four VHH-derived therapeutics have been approved. These include caplacizumab (a humanised VHH to treat the blood disorder acquired thrombotic thrombocytopenic purpura (aTTP), approved in 2018 by the FDA and EU) [77], ciltacabtagene autoleucel (a VHH-based CAR-T therapy for multiple myeloma, approved in 2022 by the FDA and EU) [78], ozoralizumab (a trivalent bispecific, approved in Japan in 2022 to treat rheumatoid arthritis) [79] and envafolimab (a humanised VHH-antibody Fc fusion protein targeting PD-L1, approved in China in 2022 for the treatment of solid tumours) [80]. Though this number is small relative to conventional antibody therapeutics, there are many more VHHs progressing through the stages of clinical development [62, 81, 82].

1.3.2 Identifying binders

After selecting a target antigen, antibodies or nanobodies are screened to identify those that bind to a target with high specificity and affinity. Early experimental methods for antibody discovery involved immunising non-human species, such as mice or rabbits, to elicit an immune response and generate antigen-specific antibodies. For nanobody discovery, the same approach can be used by immunising camelids or sharks to produce heavy-chain-only antibodies.

Following immunisation, antibodies can be retrieved from blood serum or by isolating B-cells. Hybridoma technology can be used for antibody production *ex vivo* [83]. Current methods to retrieve VHH or VNAR domains involves isolating B-cells from immunised camelids or sharks, extracting RNA, and amplifying the sequences using polymerase chain reaction (PCR) techniques [84, 85].

The nanobody sequences can then be used to construct phage, yeast, or ribosome display libraries, which undergo panning to identify binders [84–86]. Additional

methods such as fluorescence-activated cell sorting (FACS) and surface plasmon resonance (SPR) can sort cells and identify binders with high affinity for the antigen [84, 86]. Although both nanobodies and antibody fragments (such as scFvs and Fabs) can be used in these platforms, the single-domain nature of nanobodies makes them particularly well-suited for display-based screening, as their small size, high solubility, and stable folding allow for efficient expression and binding selection in these systems.

1.3.3 Optimising developability

The success of a new biotherapeutic in clinical trials relies not only on its ability to bind to the target of interest, but also on developability. Developability is defined as the ability of a molecule to undergo manufacture, storage and transport without detriment to its activity or safety [62, 87, 88], and encompasses the physicochemical properties which must be optimised after hit identification. To maximise chances of success, it is important to identify and minimise these risks as early as possible in the discovery process.

1.3.3.1 Aggregation and stability

Aggregation refers to the irreversible self-association of proteins in a medium. This process is largely driven by hydrophobic patches or uneven charge distribution on the protein surface, where hydrophobic or electrostatic interactions can lead to non-specific binding [89, 90]. Patches of hydrophobicity may already be present on the protein surface, or hydrophobic regions may become exposed under temperature, pH, or mechanical stress during formulation or storage, where poor thermal or chemical stability causes unfolding of the protein.

Aggregation is a major developability concern, as it can compromise manufacturability by reducing purity and yield, thereby increasing production costs and limiting

scalability [90, 91]. Secondly it can increase the risk of an immunogenic response against the therapeutic, leading to adverse reactions in a patient [92]. Aggregation propensity may be assessed directly via cross-interaction chromatography, or indirectly by testing for sample purity (size exclusion chromatography), hydrophobic interactions (hydrophobic interaction chromatography) or thermostability (differential scanning fluorimetry) [93–96].

1.3.3.2 Solubility and viscosity

Surface hydrophobicity and charge also impacts the solubility of immunoglobulin domains. Poor solubility increases the likelihood of self-association at higher concentrations and results in high viscosity [97]. The formulation within which a therapeutic is delivered is often required to contain a high concentration of the protein and so poor solubility can complicate development. Self-association can be quantified using assays such as affinity-capture self-interaction nanoparticle spectroscopy (AC-SINS) as well as the aforementioned techniques to measure aggregation propensity [98, 99].

1.3.3.3 Sequence liabilities

Here, we define developability to include sequence liabilities, which are chemical modifications that can alter the properties of a protein. Since the diversity of the variable region in immunoglobulins enables them to bind specifically to a target, if modifications were to occur within the paratope residues, this could result in loss of antigen-binding ability [100]. Oxidation of methionine and tryptophan residues, deamidation of asparagine, and isomerisation of aspartic acid are examples of common post-translational modifications seen in monoclonal antibodies [100]. Oxidation in IgG1 antibodies has been shown to induce aggregation and immunogenic responses [101]. Further liabilities include unpaired cysteines, which are highly reactive [102], and N-linked glycosylation, which can lower thermal stability, affect solubility and increase aggregation propensity [101]. Motifs containing tryptophan, valine, glycine and arginine present in the CDR3 have been identified as drivers of polyreactivity, which can cause off-target effects [103]. Collectively, these modifications may

increase the risk of undesirable events such as aggregation, heterogeneity, and immunogenicity, and reduce efficacy of the therapeutic.

1.3.3.4 Immunogenicity and humanness

Immunoglobulins must also be tolerated by the human immune system to be successful in therapeutic applications. Rational approaches to reduce the risk of unwanted immunogenicity involve maximising the 'humanness' of the candidate drug. Humanness can be defined as similarity to antibodies raised naturally and tolerated by healthy human immune systems. This aspect of therapeutic design is particularly pertinent for the development of nanobody therapeutics, which are fundamentally non-human.

General principles for the experimental humanisation of nanobodies are akin to those for antibodies. The first broadly applicable strategy to increase the human portion of variable domains was chimerisation, the grafting of a non-human variable region onto human constant domains [104]. Methods to further increase the humanness of the antibody sequence include CDR-grafting, where only the CDR loops of a non-human antibody are transplanted onto a human framework region [105], and specificity-determining region (SDR) grafting, where only the binding site residues from the CDR loops are modified [106, 107] (**Figure 1.6**).

To preserve favourable properties of the non-human antibody, humanisation has also been carried out using the most closely related human germline sequences or homologous framework regions as a template [108]. Alternative approaches include framework shuffling [109], or resurfacing, which preserves the non-human frameworks: only surface-exposed residues which differ from human are replaced, on the basis that buried residues will not impact the level of immunogenicity [110, 111].

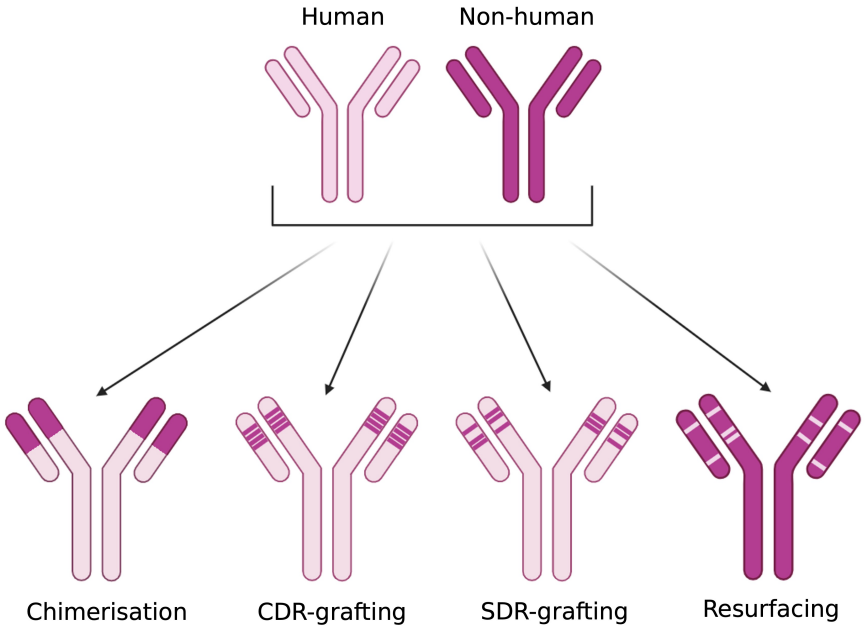


Figure 1.6: *Humanisation methods have evolved from initial chimerisation approaches, aiming to reduce non-humanness without compromising functionality. These include: CDR-grafting, where non-human CDR loops are grafted on to a human framework, SDR-grafting, where only the binding residues of the CDRs are grafted on to the framework, and resurfacing, where exposed non-human residues are replaced by human ones.*

There are additional considerations to make when optimising developability for nanobodies, since their structural features differ from antibodies. For example, mutating nanobody hallmark residues may increase their humanness, however this could also diminish the favourable properties of the nanobody, since these residues tend to be hydrophilic and are known to contribute to overall solubility [48, 53, 65]. Furthermore, mutations at these exposed positions to more VH-like hydrophobic residues will increase aggregation propensity, decreasing developability and increasing the likelihood of immunogenicity [112–114].

1.4 Computational approaches to therapeutic design

The optimisation of a biotherapeutic candidate via experimental methods can be inefficient and costly. Computational approaches to therapeutic design may offer a solution to accelerate the discovery and development pipeline and minimise failure rates by informing rational engineering strategies. Although *in silico* methods show promise, they rely on the quantity and quality of data available and still require robust experimental validation to be integrated into existing workflows with confidence. In this section, we will discuss current computational approaches to therapeutic design and the challenges in the application of these approaches for nanobodies as a novel format.

1.4.1 Sequence data

1.4.1.1 Numbering schemes

In this thesis, we analyse annotated amino acid sequences of antibodies and nanobodies to compare their framework regions, CDR loops, and amino acid compositions at specific positions. These comparisons are made possible by standardised antibody numbering schemes.

Sequence alignment methods are commonly used to compare protein or nucleotide sequences supported by algorithms such as Needleman–Wunsch [115], Smith–Waterman [116] or BLAST [117]. Whilst the conserved framework regions of antibodies are well-suited to sequence alignment, comparison of the highly variable CDR loops is more challenging [118]. To address this, several antibody numbering schemes have been developed that enable sequence comparison without the need for direct alignment. Numbers are assigned to each amino acid position relative to conserved framework residues.

Various numbering schemes have been proposed, including Kabat, Chothia, Martin, AHo, and the international ImMunoGeneTics information system (IMGT) [23, 51, 119–121]. Although broadly similar, they have minor differences in how the framework and CDR regions are defined (which may be sequence- or structure-based), and in how they accommodate insertions and deletions in the sequence [122]. In this thesis, we adopt the IMGT numbering scheme for the annotation of immunoglobulin domains due to its widespread use and structural basis (**Figure 1.7**).

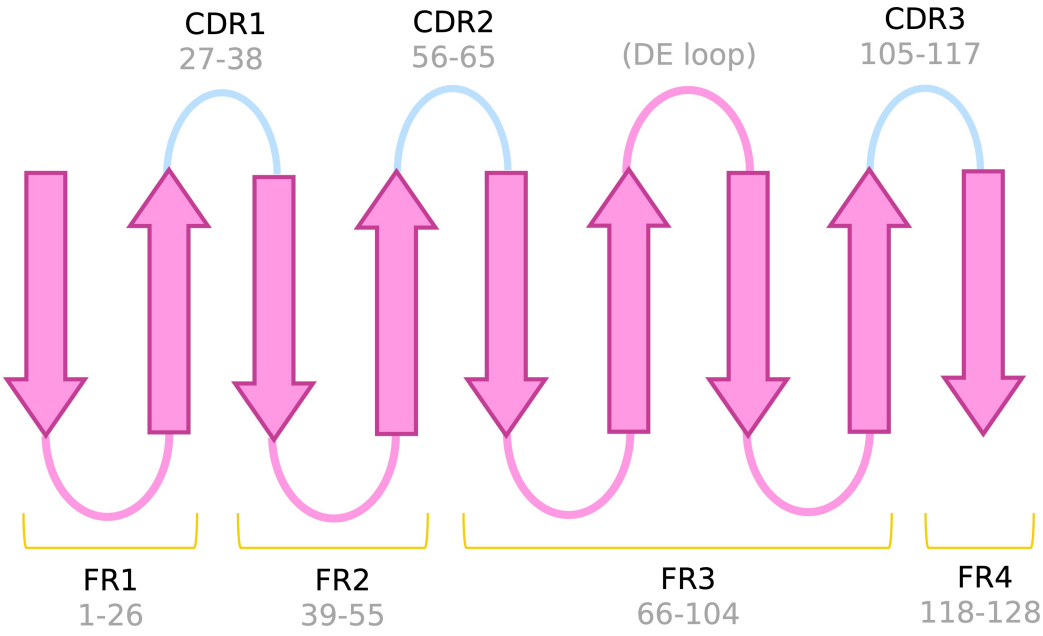


Figure 1.7: *The framework regions and CDR loops of an immunoglobulin variable domain annotated according to the IMGT numbering scheme. This numbering is consistent for both VH/VHH and VL domains.*

1.4.1.2 Numbering tools

Computational tools such as ANARCI [123] can be used to number antibody sequences at high throughput. ANARCI uses Hidden Markov Models (HMMs) to predict numbering. An HMM can be used to understand systems where 'hidden' states (variables within the system that we want to determine) produce observable outputs (the measurable data generated by those states). In the context of numbering sequences, the hidden states are the numbered positions which are

not known in advance for a new sequence. The observations are the amino acids that make up the antibody sequence. By training a model on previously annotated antibody sequences, an HMM can predict the numbered position for each amino acid. ANARCI annotates sequences based on a set of HMMs built from germline sequences from the IMGT database [124]. The HMM with the best alignment score compared with the input sequence is used for numbering. ANARCI HMMs are species-specific, with separate HMMs for each species with available germline data.

1.4.2 Structural data

1.4.2.1 Experimental structure determination

Protein structures can be determined experimentally through techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, or electron microscopy [125]. However, structure determination is often time-consuming and can be challenging for proteins with highly flexible regions, including loops. For the design of biotherapeutics, structural data is a critical component to guide rational engineering strategies. For example, it can be used to identify how to modulate binding specificity and affinity, or how to improve developability [27, 126].

1.4.2.2 Structure prediction

Structure prediction methods have been developed as a means to overcome the bottleneck of experimental structure determination. Earlier approaches such as homology modelling predict structures based on sequence similarity to known structures which are used as templates. For proteins in general, tools like MODELLER [127], SWISS-MODEL [128], and Phyre2 [129] were commonly used, while software such as ABodyBuilder [130], PIGS [131] or RosettaAntibody [132] catered specifically to antibody structure prediction. The highly conserved framework regions of antibodies and nanobodies are often easier to model using homology-based methods, however, the diverse CDRs, particularly CDR-H3, are more challenging to predict [27].

The release of AlphaFold2 [133] transformed structure prediction methods, combining a deep-learning approach with evolutionary information. Several machine-learning-based antibody-specific methods have since been developed, including IgFold [134], RosettaFold2 [135], and ImmuneBuilder [136]. NanoBodyBuilder2 (part of the ImmuneBuilder suite) shows superior accuracy for nanobodies compared to AlphaFold2, and was used to model nanobody structures throughout this thesis.

Predicting protein-protein complexes remains a significant challenge. Whilst methods have been developed for the prediction of protein complexes and protein-protein interactions [137, 138], models for general protein complexes have been shown not to translate well to specifically antibody-antigen interactions. Approaches tailored specifically to antibodies, and particularly to nanobodies, are still in the very early stages of development [139, 140]. One major limitation is the scarcity of high-quality training data, as far fewer crystal structures of complexes exist compared to monomeric proteins [141, 142]. As a result, the reliability of these methods does not yet match that of established approaches for monomeric structure prediction. Therefore, the structural data for antibody-antigen and nanobody-antigen complexes used in this thesis is restricted to experimentally-solved structures.

1.4.3 Databases

Databases provide centralised, accessible repositories which enable and accelerate large-scale analysis and aid the development of computational methods. Sequence and structural data for antibodies and nanobodies can be found in general biological databases such as the Protein Data Bank (PDB) [143], UniProt [144], and resources such as the NCBI [145]. However, a number of specialised databases have been developed to support immunoglobulin research. For large-scale sequence data, the Observed Antibody Space (OAS) [146, 147] and iReceptor [148] provide access to immune repertoire sequencing across various studies and species, including both antibodies and VHHs. Antigen- and disease-specific resources include CoV-AbDab [149], which focuses on antibodies and VHHs against coronaviruses, and

the Immune Epitope Database (IEDB), which catalogs antibody and T-cell epitope data [150, 151]. PLAbDab [152] collects antibody sequences from patents and publications, plus data from SAbDab [141, 142] and TheraSAbDab, which records clinical-stage therapeutic antibodies and VHHs, including associated structural data where available [153].

SAbDab serves as a central database for antibody structural data curated from the PDB [141]. SAbDab-nano is a nanobody-specific subset of SAbDab, containing all available solved VHH structures [142]. Additional structure-focused resources include IMGT/3Dstructure-DB [154] and AbYsis [155], which provide annotated antibody sequences and structures.

VHH-specific databases that have been developed include INDI [156], NanoLAS [157], and sdAb-DB [158], which aim to aggregate VHH sequence and structural data from various sources. Despite the growing number of resources, nanobody data, especially for VNARs, remains relatively scarce compared to antibody data, limiting the development of computational methods for these molecules.

1.4.4 Developability

The rapid expansion of sequence and structural data from both next-generation sequencing and structure prediction has enabled the advancement of computational methods for protein property prediction and antibody developability assessment.

1.4.4.1 Aggregation

Methods to predict aggregation propensity typically focus on identifying hydrophobic patches that are prone to causing self-interaction between proteins. Many of these approaches are applicable to proteins in general and are not specific to antibodies.

Sequence-based methods include Aggrescan [159], TANGO [160], Zyggregator [161], SALSA [162], and PASTA [163]. The Aggrescan algorithm was built on the basis that

specific linear peptide sequences are more aggregation prone [159]. Structure-based methods such as Aggrescan3D may overcome the limitations of sequence-methods such as Aggrescan by better accounting for aggregation-prone regions formed by non-contiguous sequences and solvent exposure [164]. Other structure-based methods include the Structural Aggregation Propensity (SAP) score, which uses molecular dynamics simulations to locate and characterise hydrophobic regions that are exposed, either in the native state or due to conformational changes [165]. AggScore predicts aggregation-prone regions from the distributions of hydrophobic and electrostatic surface patches and is applicable to globular proteins including antibodies [166].

In addition to purely computational approaches, aggregation propensity may also be predicted from experimental data. For example, some machine learning-based tools use antibody sequence features to predict aggregation risk based on hydrophobic interaction chromatography (HIC) column retention times [94, 167].

1.4.4.2 Thermostability

The thermostability of a protein can influence aggregation propensity and yield. Various computational tools specific to antibody thermostability have been developed. For example, AbMelt models antibody structural flexibility using molecular dynamics simulations and structure-based descriptors such as solvent-accessible surface area (SASA) to estimate thermostability [168]. Harmalkar *et al.* (2023) propose two machine learning approaches: a pre-trained language model leveraging sequence data, and a convolutional neural network (CNN) using structural data, to classify thermostable single-chain variable fragment (scFv) variants [169]. However, they highlight limitations in publicly available thermostability datasets such as ProTherm [170, 171], which lack sufficient coverage for specific formats like scFvs, restricting the generalisability of thermostability prediction tools to newer formats.

Similar challenges would exist for nanobodies as a novel format, however the recent release of NbThermo in 2023 [172], a nanobody-specific database of experimental thermostability data, has enabled the development of nanobody-specific thermostability predictors. For example, TEMPRO predicts nanobody melting temperature using a combination of language model embeddings and structural features and is validated on approximately 500 sequences from NbThermo [173]. Similarly, NanoMelt uses 640 melting temperature measurements, including data from NbThermo, and sequence embeddings to develop an ensemble learning model for nanobody thermostability prediction [174].

1.4.4.3 Solubility and viscosity

Poor solubility can contribute to aggregation and high viscosity, leading to challenges in formulation and delivery. Computational tools to predict and improve solubility include both general protein-level predictors and more specialised approaches for antibodies. Amongst early general tools for solubility prediction are sequence-based methods such as SOLpro, which uses a support vector machine (SVM) model trained on over 17,000 protein sequences [175], DeepSol, which employs a convolutional neural network (CNN) [176], and PaRSnIP, which uses gradient boosting [177]. SOLart uses structural features to train a random forest regression model for general protein solubility prediction [178].

Antibody-specific approaches include Sharma *et al.* (2014), who proposed a sequence-based method to predict viscosity in IgG1 monoclonal antibodies using descriptors for net Fv charge, charge symmetry, and hydrophobicity [179]. Agrawal *et al.* (2015) introduced a spatial charge map (SCM) score using homology modelling to quantify negative charge patches, which have been shown to correlate with high viscosity [180]. More recently, the SCM score was incorporated into DeepSCM, a CNN-based surrogate model trained on SCM scores for viscosity prediction directly from sequence data [181].

For nanobodies, solubility prediction remains relatively underdeveloped. General methods that have been applied to nanobodies include CamSol, which calculates a solubility score based on surface hydrophobicity and electrostatics, identifies unfavourable surface residues, then rapidly screens mutations to find variants with improved solubility [182].

1.4.4.4 Sequence liabilities

Sequence liabilities can affect the chemical stability of immunoglobulins. Many of these liabilities are associated with specific amino acid motifs [183–186], making them well-suited to detection through computational sequence-based methods such as string matching. However, not all sequence liabilities pose a real risk: only those exposed on the protein surface are likely to be chemically reactive. Structure-based computational approaches can be valuable for incorporating features like solvent-accessible surface area (SASA), secondary structure or conformational flexibility to better identify liabilities more likely to be problematic [184]. Several approaches combine both sequence and structural features [100, 187–189]. The Therapeutic Antibody Profiler (TAP) reports sequence liabilities and allows users to visualise liability motifs directly on antibody structures [62].

1.4.4.5 Immunogenicity and humanness

Computational methods have been developed to quantify the humanness of a sequence as a proxy metric for the likelihood of immunogenicity. Earlier methods introduced scores based on pairwise sequence identity such as the H-score [190], G-score [191], T20 score [192], and the human string content (HSC) score [193]. Other methods aimed to capture higher-order relationships between pairs of amino acids [194, 195], or determine similarity based on nucleotide or amino acid frequencies and predict possible mutations using scoring matrices [196, 197].

The growing availability of sequence data [146, 147] has made *in silico* humanness scoring and humanisation increasingly amenable to machine-learning methods. Methods including random forests [198], contrastive-learning [199], language models

[200–202] and neural networks [203, 204] have been used. Structure-guided methods for computational humanisation aim to balance stability, affinity and humanness such as CoDAH [205, 206], and CUMAb [207]. CUMAb simulates CDR-grafting onto human frameworks to select designs using an energy-based ranking. Nanobodies more often incorporate framework residues into their paratopes [54, 55]. This is critical to consider in applying computational methods based on protocols such as CDR grafting, where it may be assumed that only the CDRs contribute to antigen-binding [208].

The development of humanisation software dedicated to nanobodies is still a relatively new endeavour. The first was Llamanade [209], which avoids the humanisation of residues that are integral to nanobody physicochemical properties, such as conserved residues in the FR2 region, or those that contribute to CDR3 conformation. AbNatiV [210], a deep-learning-based pipeline for nanobody (and antibody) humanisation, also prioritises this balance between nanobody nativeness and humanness. Sequence data is used to train variational auto-encoders (VQ-VAE) models which quantify the similarity of a sequence to human VH or camelid VHH domains. This measure of nanobody or antibody nativeness is coupled with a sequence profile that can be used to inform engineering.

1.4.5 Multi-parameter optimisation

When designing a biotherapeutic, ideally, binding affinity, specificity, and developability factors should all be optimised. However in practice, it is often not feasible to co-optimise all of these, and so compromises are made. Experimental assays to determine binding and developability issues can be expensive and time-consuming to carry out, and can limit the number of candidates that can be investigated further. Using computational methods for the simultaneous optimisation or selection of antibodies or nanobodies with a general set of drug-like properties represents a compelling strategy towards reducing the failure rate of candidate therapeutics and

accelerating the therapeutic design pipeline.

There are several computational approaches that assess antibody developability properties in parallel, such as TAP [62] and the Developability Index [211], amongst others [212–214]. Methods for the co-optimisation of closely related properties such as affinity and specificity [215], affinity and stability [216] and stability and solubility [217] have also been developed in recent years. Towards multi-parameter optimisation, Bachas *et al.* (2022) [218] use language models to predict binding affinity alongside a 'naturalness' metric that relates to developability and immunogenicity.

There are also increasing examples of library generation designed with in-built favourable properties in a shift towards *de novo* design [219, 220]. Arras *et al.*. (2023) [221] developed a semi-synthetic method to generate humanised and developable nanobodies (VHH) from camelid immunisation. In subsequent work, they combined their library approach with a long short-term memory (LSTM) model, trained on sequences selected from the library, to design new humanised and developable nanobodies [222].

These approaches show that there is potential to achieve multi-parameter optimisation provided there is access to high-throughput and high-quality data for model training. In general, data is far more limited for nanobodies as a newer therapeutic format, hence why much of the work so far in this area has focused on conventional antibodies. As we have described in this chapter, nanobodies pose additional challenges, as their structural differences affect how they interact with their antigens. This complicates the simultaneous optimisation of binding and developability whilst maintaining the properties that make them attractive as a drug format. Therefore, it is likely that many *in silico* protocols designed for the selection or optimisation of conventional antibodies are not immediately applicable to nanobodies, and that there is a need to design protocols that are nanobody-specific.

1.5 Thesis outline

The second chapter of this thesis is entitled: 'A comparison of the binding sites of antibodies and camelid nanobodies'. In this chapter we present a comparative study of the binding sites of antibodies and VHH nanobodies, using sequence and structural data to compare the paratopes, epitopes and antigen interactions of antibody-antigen and nanobody-antigen complexes. Our results have implications for the engineering of therapeutic nanobodies and can be used to inform the selection of the best modality for targeting a particular epitope.

The third chapter is entitled: 'PLAbDab-nano: a database of camelid and shark nanobodies from patents and literature'. We describe the curation of PLAbDab-nano, a searchable database containing ~5000 annotated VHH and VNAR sequences, alongside modelled VHH structures. The data were drawn from the increasing volume of academic literature and patents related to nanobodies. In addition to it being the first nanobody database to include VNAR data, PLAbDab-nano also provides the key data to help design diverse libraries, as well as find sequences similar to known patented or therapeutic entries.

The fourth chapter, 'The Therapeutic Nanobody Profiler: characterising and predicting nanobody developability for therapeutic design', describes an open-source computational tool, TNP, for predicting nanobody developability relative to currently available clinical-stage data. Tailored specifically for nanobodies, it accounts for their unique properties compared to conventional antibodies for more efficient development of this novel therapeutic format. We calibrate the TNP metrics using clinical-stage nanobody sequences and *in vitro* developability data for clinical-stage and proprietary nanobodies.

Finally, chapter five presents the conclusions of this thesis and suggests directions for future work.

2

A comparison of the binding sites of antibodies and camelid nanobodies

Contents

2.1 Abstract 32
2.2 Introduction 33
2.3 Methods 34
2.4 Results 41
2.5 Discussion 64

2.1 Abstract

An understanding of the differences between antibody and nanobody structures, particularly the way in which these molecules interact with their antigens, is critical to the rational design of therapeutic nanobodies. Firstly, we can identify whether antibodies and nanobodies show preferences for different types of epitopes, informing the choice of modality when designing a therapeutic against a particular antigen. Secondly, characterising their interactions with an antigen can inform engineering strategies, as we can determine which residues or regions are more often integral to binding affinity or specificity. This approach offers applications in optimising developability and humanisation whilst ensuring that binding affinity is not compromised.

In this chapter, we use non-redundant sequence and structural data to conduct a comparative analysis of the binding sites of antibodies and camelid nanobodies (VHHs), investigating their paratopes, epitopes and antigen interactions. We demonstrate that even though camelid nanobodies have smaller paratopes, they target epitopes of equal size to those targeted by antibodies. To achieve this, their paratopes contribute more interactions per residue than the paratopes of antibodies. Additionally, we find that conserved framework residues are of increased importance in camelid nanobody paratopes, suggesting that they include non-specific interactions to achieve comparable affinity. Furthermore, the epitopes targeted by camelid nanobodies are only marginally less accessible than those of antibodies: we posit that this may be explained by differences in the orientation and compactness of the nanobody and antibody CDR3 loops.

The text in this chapter is based on the following publication:

> **Gordon, G.L.**, Capel H.L., Guloglu, B., Richardson, E., Stafford, R.L., and Deane, C.M. A comparison of the binding sites of antibodies and single-domain antibodies. *Frontiers in Immunology* **14**, 1231623, (2023).

References to nanobodies in this chapter should be interpreted as to the camelid (VHH) type only, unless specified otherwise.

Myself and Henriette Capel curated the sequence and structural datasets respectively and carried out the analyses of the paratopes, epitopes and antigen interactions. I investigated and implemented methods to characterise epitope shape and accessibility. Bora Guloglu devised the methods for the analysis of the CDR3 loop conformation.

2.2 Introduction

As described in **Chapter 1**, there are major structural differences between camelid nanobodies (VHHs) and antibodies. The most conspicuous of these is the lack of a light chain in nanobodies, which therefore have only three CDR loops. The CDR loops in both antibodies and nanobodies comprise the majority of the binding site [26, 27].

In previous work, Zavrtanik *et al.* (2018) suggested that nanobodies bind to more "rigid, concave, conserved and structured" epitopes [54]. This hypothesis that nanobodies can target epitopes that are inaccessible to antibodies is often linked to the fact that the CDR3 loops of nanobodies tend to be longer than the analogous CDR-H3 of conventional antibodies [43, 44]. Zavrtanik *et al.* (2018) [54] and Mitchell and Colwell (2018) [55] found an average difference in loop length of between three and four residues. Many papers have theorised that the longer CDR3 loops of nanobodies can protrude into concave spaces in a protein antigen surface that would be inaccessible to a conventional antibody with a shorter CDR-H3 loop [45–47]. However, as highlighted by Henry and Mackenzie (2018) [223], isolated case studies make up much of the supporting literature on this idea. They note that "the degree to which nanobodies bind cryptic epitopes *vs.* conventional antibody-accessible epitopes...remain[s] unknown".

Aside from differences in CDR3 loop length, previous comparisons of the paratopes of nanobodies and antibodies have shown that nanobody paratopes have a more hydrophobic character than antibody paratopes, but are similarly enriched in aromatic residues [54]. Furthermore, nanobodies tend to draw more residues from framework regions into the paratope, whereas antibodies are more reliant on the CDR loops to interact with an antigen [54, 55].

The previous studies of Zavrtanik *et al.* [54] and Mitchell and Colwell [55, 224] are limited by their relatively small datasets: Zavrtanik *et al.* analysed 105 nanobody-antigen crystal complexes, while Mitchell and Colwell compared sets of 90 nanobody-antigen and antibody-antigen crystal complexes [55], and in a later study, analysed 156 nanobody-antigen and antibody-antigen complexes [224].

As available nanobody datasets have increased in size in recent years [142], we have examined the binding sites of nanobodies and antibodies using non-redundant datasets of 892 antibody-antigen and 345 nanobody-antigen structural complexes, alongside non-redundant datasets of 1,614,526 human VH sequences (from Eliyahu *et al.* (2018) [225]) and 783,889 camel VHH sequences (from Li *et al.* (2016) [52]).

2.3 Methods

2.3.1 Sequence datasets

Non-redundant sequence datasets were obtained from the OAS database [147]. A set of 1,621,889 human VH sequences generated by Eliyahu *et al.* (2018) [225] and 784,291 camel VHH sequences generated by Li *et al.* (2016) [52], were filtered to remove duplicated sequences. Final datasets consist of 1,614,526 human VH sequences and 783,889 camel VHH sequences. These sequence datasets were used to compare the CDR lengths and the amino acid compositions of framework residues and CDR loops between antibodies and nanobodies.

2.3.2 Structure datasets

We created up-to-date, non-redundant datasets of both antibodies and nanobodies that were in complex with protein antigens. PDB entries for crystal structures were extracted from SAbDab [141, 142] and SAbDab-nano [142] on the 23rd February 2022. The datasets were extracted as follows:

1. Only antibody-antigen and nanobody-antigen complexes for which at least one of the CDR residues of the antibody is in close contact, defined as under 4.5 Å, to the antigen.

2. Only the antibodies and nanobodies identified as in a complex with a protein antigen ($>$ 50 residues), according to SAbDab annotations.

3. Only structures of complexes solved by X-ray crystallography to $\leq$ 3.0 Å resolution.

4. Antibodies and nanobodies were filtered separately to remove redundancy using a sequence identity cut-off of 95% across the IMGT-defined CDR residues using cd-hit [226].

5. A small number of complexes were reintroduced if their epitope identity score was less than 75% compared to any other complex, to include complexes containing similar CDRs but different epitopes. To calculate epitope identity, epitope sequences were first aligned using cd-hit. Based on the aligned positions, the epitope identity score was determined as the fraction of matching epitope residues (same amino acids and same aligned position) across the epitope residues of the two antigens. The epitope is defined by distance in this case (see **Section 2.3.4**).

The resulting nanobodies structural dataset consisted of 345 complexes, of which 309 had "unique" CDRs. The final antibodies structural dataset consisted of 892 complexes, of which 792 had "unique" CDRs.

Text A.1 and **Table A.1** give further detail on dataset curation and a breakdown of the number of complexes remaining at each filtering step. **Table A.2** shows the species variation for both structural datasets. **Figure A.1** shows the distributions of epitope identity across datasets.

2.3.3 Numbering schemes

The IMGT numbering scheme and CDR definitions were used throughout this work (CDR1: IMGT residues 27-38, CDR2: IMGT residues 56-65, CDR3: IMGT residues 105-117) [51]. ANARCI [123] was used to number all of the antibodies and nanobodies.

2.3.4 Binding site definitions

We describe the binding site using three definitions (**Figure 2.1**). As is used in most methods annotating and predicting paratopes or epitopes, we consider a distance definition, which includes all antibody residues which are in close contact with the antigen ($\leq$ 4.5 Å). A very similar result is achieved by defining the binding site using the solvent-accessible surface area (SASA), where residues are included in the paratope or epitope if they become buried upon complex formation. In our work, we focus on defining the binding site by the interactions occurring between pairs of residues, identified using Arpeggio [227] (**Figure A.2**).

Arpeggio determines interaction types based on distance, angle, and atom type. Each PDB file in both structure datasets was analysed with Arpeggio using a distance threshold of 4.5 Å. This generates a five-bit fingerprint for each pairwise interatomic contact which shows the type of interactions occurring. These include: van der Waals, steric clashes, covalent bonds, proximal interactions (defined as being within the cut-off distance but not representing a meaningful interaction) and specific interactions such as hydrogen bonds. This output was processed to exclude interactions with water molecules, and chains other than the antibody and antigen.

Heterogens (non-standard residues) were removed using BioPython [228].

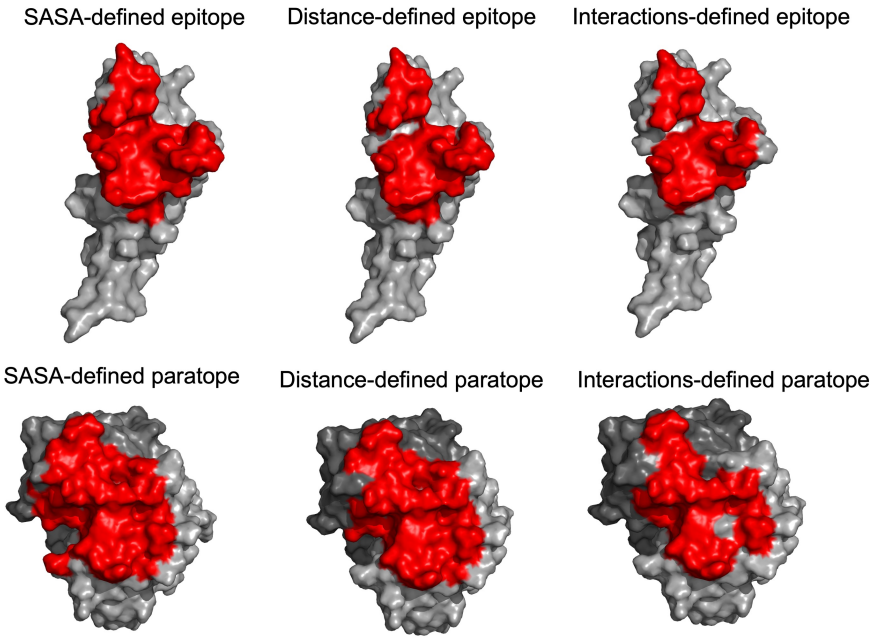


Figure 2.1: *Visualising the differences between the SASA-defined, distance-defined and interactions-defined epitopes and paratopes. PDB ID: 6AZZ [229]. For the epitope representations above, the epitope surface is red, whilst the antigen is grey. For the paratope representations below, the paratope surface is red, whilst the heavy chain of the antibody is dark grey and the light chain is light grey.*

For all remaining positions, the interatomic interactions were summarised per residue-residue pair. Residues were considered to interact if at least one of the atom-atom pairs in these residues established a van der Waals (vdW) bond or a specific interaction. Clashing vdW and proximal interactions were classified as contacts if no specific bonds were observed. We refer to this latter definition of the binding site as the 'interactions-defined paratope' and 'interactions-defined epitope'.

Interatomic interactions between the antibody-antigen and nanobody-antigen complexes were compared by counting the total number observed. If multiple interaction types were identified between a single pair of atoms, the interactions were counted individually. The mean and standard deviation of the observed interactions were

calculated by subsampling 10% of the total set of interactions 1000 times.

2.3.5 Amino acid composition

The sequence datasets were used to compare the amino acid compositions of the nanobody and antibody VH CDR loops. The nanobody and antibody sequence datasets were split by germline, and only those belonging to IGHV3 were compared: this includes all sequences for the nanobody dataset but reduces the antibody dataset to 761,235 sequences. Sequences were aligned using ANARCI numbering annotation [123]. The proportions of individual amino acids at each position in each CDR loop were determined. Positions were omitted where less than 5% of sequences had an amino acid at that position.

To assess the conservation of framework residues that appear in the paratope, firstly the structural datasets were used to determine which positions are often involved in the paratopes of nanobodies and antibodies. Framework residues were considered as important contributors to the paratope if they were observed in at least 10% of the complexes in our datasets. The amino acid compositions of these same positions were then obtained from the sequence datasets as a background for comparison.

2.3.6 Epitope accessibility

Multiple methods are available that describe the curvature of a surface. However, these methods struggle to successfully capture the complex nature of the epitope surface. Here, we have designed a simple metric using the solvent accessible surface area (SASA) to compare the accessibility of the epitopes targeted by nanobodies and antibodies.

We define ‘epitope accessibility’ as the SASA of the epitope surface, relative to the sum of the SASA values of the epitope residues as if they were isolated in

space. The function “get_sasa_relative" from the PyMOL cmd package [230] was used to calculate the SASA values, where residues with a value of 0 are considered completely buried, and those with a value of 1 are completely exposed. As such, the sum of the SASA of epitope residues were they to be isolated in space is equivalent to the total number of residues in the epitope: this is reflected in our implementation of the metric. Differences in the distributions of epitope accessibility for nanobodies and antibodies are determined via bootstrap resampling.

2.3.7 Canonical forms of the CDRs

The canonical forms of the CDR loops for the nanobody and antibody structures were identified using the PyIgClassify2 database [231].

2.3.8 Structural clustering

Antibody chains from the 345 nanobody-antigen and 892 antibody-antigen complexes were extracted, giving 301 and 838 unique nanobody and antibody structures (as some PDB entries include nanobodies or antibodies that form complexes with multiple antigens). A greedy clustering method was used to cluster each of the sets of antibody VH or nanobody CDR1, CDR2 and CDR3 loops based on their length and RMSD (root mean square deviation), with a cut-off of 1.5 Å. The number of clusters which contain both nanobodies and antibodies (overlap clusters) was determined. The expected number of overlap clusters was found by generating random clusters of matching size. Random clusters were generated 20 times from the original set of all antibody and nanobody structures and the mean and standard deviations for the number of overlap clusters was calculated.

2.3.9 Orientation of CDR3 loops

We analysed the general orientation of the antibody CDR-H3 and nanobody CDR3 loops by examining their centres of geometry in reference to an $\mathbb{R}^3$ coordinate system (see **Text A.2** and **Figure A.3** for methods). The dataset used for this analysis includes the structures of 388 bound nanobodies, 116 unbound nanobodies, 1977 bound antibodies and 862 unbound antibodies. Structures were downloaded from SAbDab [141, 142] on 8$^{\text{th}}$ August 2022 and were curated to be non-redundant at 95% sequence identity. Structures were numbered with the IMGT scheme using ANARCI [123] and CDR definitions used accordingly. Any structures with missing backbone atoms in CDR loops or anchors (three residues on either side of each loop) were also removed.

Using the spherical coordinates method, ρ (rho) describes the reach of the CDR3 loop away from the rest of the heavy chain domain. A CDR3 loop in an extended conformation will have a high value for ρ, whereas a loop of identical length that is folded down against the framework will have a lower value. ϕ (phi) gives an indication of whether the CDR3 loop is horizontally oriented towards the rest of the VH/VHH domain or away from it. In the case of antibody structures, a high ϕ value indicates packing against the VL domain. θ (theta) gives a measure of the elevation of the loop. A low value corresponds to a CDR3 that extends directly up and away from the rest of the VH/VHH domain, whereas a high value indicates that the loop is folding down. In the case of antibody structures, a high θ value corresponds to a loop that is packed into the groove created by the VH-VL interface. Lastly, we divide the loop length by ρ to give a measure of compactness. A loop with low compactness uses its entire length to reach away from the VH/VHH domain, whereas high compactness corresponds to a loop that is packed against the framework.

2.3.10 Statistical tests

As not all distributions followed the normal distribution, significant differences between the nanobodies and antibodies were tested by bootstrap resampling, in

which 5000 bootstrap samples were taken of size 300. The unpaired mean difference and the p-value of the two-sided permutation t-test are reported. Results are described as significant for p-value < 0.05.

2.4 Results

2.4.1 The nanobody CDR3 loop is longer than the antibody CDR-H3 loop

Previous work has shown that the CDR3 loops of nanobodies are longer than the equivalent antibody CDR-H3. The CDR loop lengths of nanobodies and the antibody VH domains were compared for both sequence and structural datasets, in terms of number of residues. When comparing the sequence datasets, we find that the CDR1 loops of antibodies are, on average, similar to nanobodies, with no statistical significance found for a difference of 0.06 residues ($p = 0.61$). Antibodies have on average longer CDR2 loops than nanobodies by 0.23 residues ($p = 0.0026$). The nanobody CDR3 loops are significantly longer than the antibody VH CDR3 by 3.2 residues on average ($p << 0.05$) (**Figure 2.2A**). The results from the structural dataset are mostly consistent with the trends observed for the sequence datasets: for the solved structures, bootstrap resampling shows that for CDR1, there is a significant difference between nanobodies and antibodies of 0.2 residues. For CDR2, we find that there is a difference of 0.08, however this was not significant (p-value = 0.12). For the structural datasets, the CDR3 loop is significantly longer in nanobodies than in antibodies by 1.6 residues (**Figure 2.2B**), consistent with previous studies [43–47].

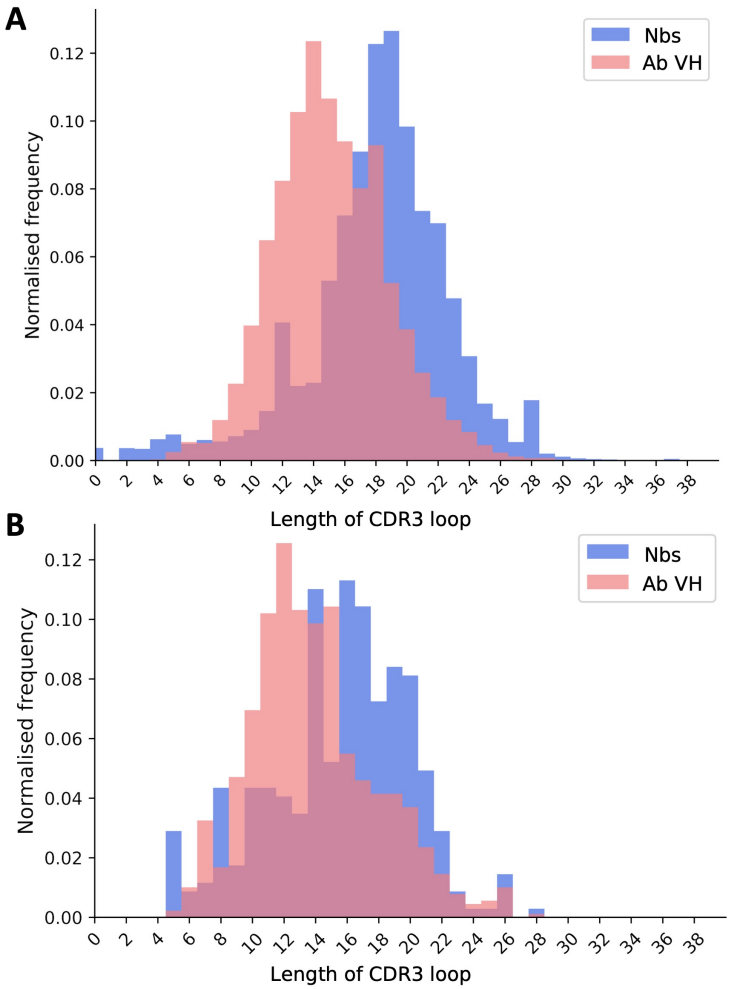


Figure 2.2: *The distributions of CDR3 loop length for (**A**) sequence data and (**B**) structural data both show that the CDR3 loops in nanobodies (Nbs) tend to be longer than the equivalent CDR-H3 in antibodies (Ab VH).*

2.4.2 Structural clustering shows a separation between antibody and nanobody CDR structures

Further to comparing the lengths of the CDR loops, we next structurally clustered them to determine whether loops from nanobodies and antibodies adopt distinct conformations and occupy different regions of structural space. If they were to adopt different paratope shapes, this would suggest that the epitopes they are able to bind to would differ.

Our initial approach was to assign canonical forms to each of the loop structures,

according to updated canonical forms from Kelow *et al.* (2022) [231]. However, for both antibodies and nanobodies, a significant percentage of CDR loops could not be assigned a canonical form. Therefore, CDR loops were clustered based on length and RMSD, with a cut-off of 1.5 Å. Clustering of the CDR loops of our 838 antibody VH domains and 301 nanobody structures collectively returned 168 clusters for CDR1, 94 clusters for CDR2, and 729 CDR3 clusters. As expected, given the differences in CDR3 length and the high variability of CDR3 in general, the number of clusters for CDR3 is much higher than for CDR1 or CDR2.

Table 2.1: *Clustering the structures of the CDR loops of nanobodies and antibody VH domains based on length and conformational similarity (measured by RMSD) shows that there is overlap between the shapes that the CDR loops of nanobodies and antibodies form. However, this overlap is within the range of that observed on random clustering, and as such suggests that nanobodies and antibodies adopt distinct CDR conformations. Values in the table show the number of structures within each cluster, with the number of clusters containing only a single structure shown in brackets.*

	VH/VHH CDR loop		
	CDR1	**CDR2**	**CDR3**
Antibodies only (single occupancy)	66 (48)	35 (21)	489 (383)
Nanobodies only (single occupancy)	84 (64)	36 (25)	230 (197)
Overlap	18	23	10
Total	168	94	729

The number of clusters containing both antibody VH and nanobody structures was determined and a mean and standard deviation for the expected number of overlap clusters (if random clustering had occurred) was calculated (**Table 2.1**). For CDR1, 18 clusters contained both antibody VH domains and nanobodies, compared to an expected value of 16.2 ± 1.29 for random clusters. For CDR2, 23 clusters contained both antibody VH domains and nanobodies, compared to an expected value of 22.3 ± 0.829. For CDR3, there were 10 overlap clusters compared to an expected value of 3.30 ± 1.55. Overall, we observe that for the CDR1 and CDR2 loops, the number of clusters we see with both antibody VH domains and nanobodies

occurring within them is within the range of what would be expected had the structures been clustered at random. This indicates that nanobodies and antibodies may assume distinct CDR conformations. As the CDR loops form the majority of the binding site, this suggests that antibodies and nanobodies may prefer to bind in different ways.

2.4.3 Nanobodies and antibodies have more identical CDR sequences than expected by chance

We next examined the CDR loop sequences of nanobodies and antibody VH domains belonging to IGHV3 germlines, taken from the sequence datasets. This reduced the size of the antibody dataset to 761,235 sequences (all 783,889 sequences in the nanobody sequence dataset belong to the IGHV3 germline). Sequences within each dataset were aligned via ANARCI annotation and the amino acid composition at each position in each loop was determined. Positions were omitted if less than 5% of sequences in a dataset had a residue at that position. **Figure A.4** shows sequence logo plots comparing the amino acid composition of the CDR loops of antibody VH domains and nanobodies.

Given the size of the sequence space, the probability of finding the same sequences in both antibody and nanobody CDR loops is low. The expected proportion of identical sequences between the two datasets was calculated for each loop and compared to the actual overlap. For CDR1, the expected overlap is 7.78×10^{-11} versus 0.013, for CDR2, 6.69×10^{-11} versus 0.016, and for CDR3, 4.08×10^{-22} versus 1.87×10^{-4}. As the actual number of identical sequences is greater than the expected number, this suggests that there are similarities in the amino acid compositions of nanobody and antibody CDR loops, which likely arise from their similar genetic background.

2.4.4 Paratopes of nanobodies and antibodies show small differences in their amino acid compositions

In addition to assessing differences in the CDR loops of antibodies and nanobodies, we considered whether there are overall differences in their respective paratopes by firstly comparing their amino acid composition. Due to the relatively small size of the structural datasets, following the work of Wong *et al.* (2021) [232], amino acid compositions for the paratopes were determined by classifying amino acids into seven classes (aliphatic, aromatic, sulfur, hydroxyl, basic, acidic and amine). For each paratope, the fraction of each observed class was determined and the distributions of amino acid types for paratopes of nanobodies and antibodies were compared.

Comparisons of the seven classes reveal that, for both distance-defined and interactions-defined paratopes, there are small increases in the proportions of aliphatic, sulfur and basic residues in nanobody paratopes (**Figures A.5**, **A.6**). We observe a decrease in aromatic residues in nanobody paratopes. There are no significant differences in the proportions of residues in the hydroxyl, acidic or amine classes.

2.4.5 Nanobody paratopes are significantly smaller than those of antibodies

Next, we compared the sizes of nanobody and antibody paratopes. Here, we define size by the number of residues in the paratope. Previous work has revealed that nanobodies can show comparable binding affinity to antibodies despite their smaller size [30, 54]. Given that nanobodies are missing the VL domain, and therefore half of a potential binding site, we would expect them to also have a smaller paratope. Using our non-redundant structural datasets, we compared the size of nanobody and antibody paratopes for each of the distance-defined, interactions-defined and SASA-defined paratopes. We found that for distance-defined paratopes, nanobody paratopes are significantly smaller than antibody paratopes by 3.6 residues and

for interactions-defined paratopes, nanobody paratopes are smaller than antibody paratopes by 2.6 residues (**Figure 2.3**). These results are consistent for the SASA-defined paratopes (**Figure A.7**). The differences found between the CDRs and the paratopes of the nanobodies and antibodies in our datasets suggest that these two modalities may target distinct epitopes.

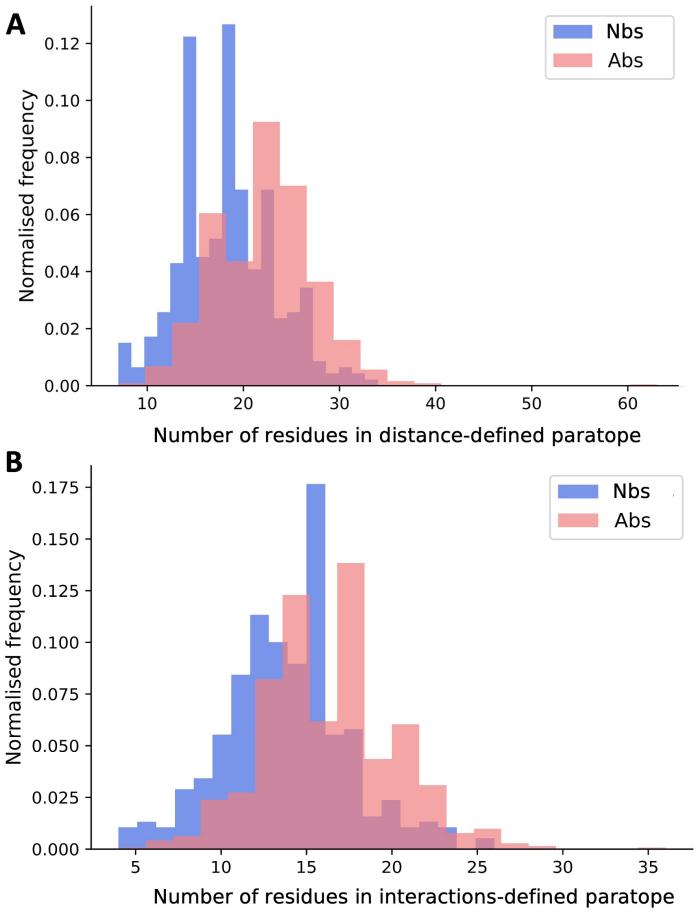


Figure 2.3: *The paratopes of nanobodies tend to contain fewer residues than the paratopes of antibodies. (**A**) Distributions of the number of residues in the distance-defined paratopes, where nanobody paratopes contain significantly fewer residues than antibody paratopes, by 3.6 residues on average. (**B**) Distributions of the number of residues in the interactions-defined paratopes, where nanobody paratopes contain significantly fewer residues than antibody paratopes, by 2.6 residues on average.*

2.4.6 Epitopes targeted by nanobodies and antibodies have similar amino acid compositions

We next assessed the epitopes of antibodies and nanobodies. One factor that may differ between nanobodies and antibodies is the amino acid compositions of their target epitopes. As for the paratope amino acid compositions, amino acid compositions for the epitopes were determined by classifying amino acids into seven classes (aliphatic, aromatic, sulfur, hydroxyl, basic, acidic and amine).

Comparisons of the seven classes for both distance-defined and interactions-defined epitopes show that for the epitopes targeted by nanobodies, there is a small but significant increase in the number of aromatic residues, and a significant decrease in the number of basic residues (**Figures A.8**, **A.9**). Given that antibodies and nanobodies are a highly similar class of molecules, it would be expected that differences in the epitope amino acid compositions would be minimal. Our results reflect this: significant differences are found but these are minor in the absolute sense. Thus, we conclude that the epitopes of nanobodies and antibodies are difficult to distinguish between based on their amino acid composition.

2.4.7 Epitopes targeted by antibodies are more linear than those targeted by nanobodies

Epitopes are often characterised by whether they are more linear or discontinuous in nature. A linear epitope is formed from amino acid residues that fall next to each other at the primary sequence level, whereas a discontinuous epitope is formed from residues that are not adjacent in the amino acid sequence but are pulled together upon folding [233, 234]. Here, we determined whether antibodies and nanobodies show distinct epitope preferences in terms of epitope continuity. We represent how continuous an epitope is by the number of contiguous residues in the epitope sequence.

For both the distance and interactions-based definitions, the epitopes targeted by antibodies tend to be slightly more linear than those targeted by nanobodies (**Figure 2.4**). Antibodies showed a significantly greater percentage of linear residues for both the distance-defined (4.6%) and interactions-defined (6.9%) epitopes. Similar results are observed when comparing the raw count of linear residues (**Figure A.10**). Results are replicated for the SASA-defined epitopes (**Figure A.11**). As the epitopes targeted by nanobodies and antibodies are of comparable size, the fact that antibodies have slightly more linear epitopes than nanobodies is not due to a difference in epitope size.

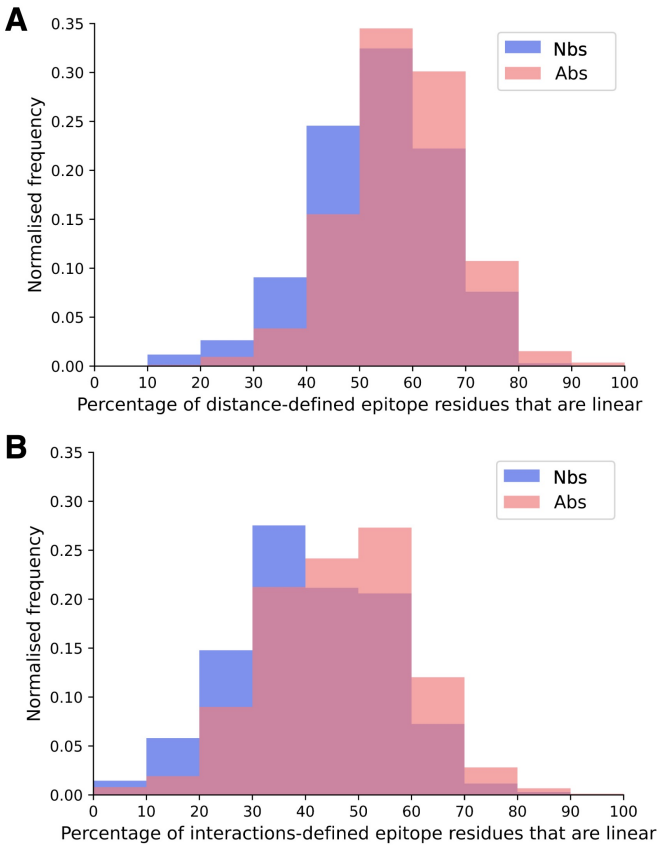


Figure 2.4: *The epitopes targeted by antibodies are relatively more linear than the epitopes targeted by nanobodies, as suggested by the distributions of percentage of linear residues for epitopes targeted by antibodies and nanobodies for the (**A**) distance-defined epitopes and (**B**) interactions-defined epitopes.*

2.4.8 Epitopes targeted by nanobodies and antibodies are of comparable size

When size is defined by the number of residues, the paratopes of nanobodies are smaller than those of antibodies, which suggests that nanobodies may be limited to binding smaller epitopes. Here, we determined the number of residues in the distance-defined epitopes, the SASA-defined epitopes and the interactions-defined epitopes for our non-redundant structural datasets. Our results show that for each of our epitope definitions, there is no significant difference between the size of the epitopes targeted by nanobodies and antibodies (**Figure 2.5**, **Figure A.12**).

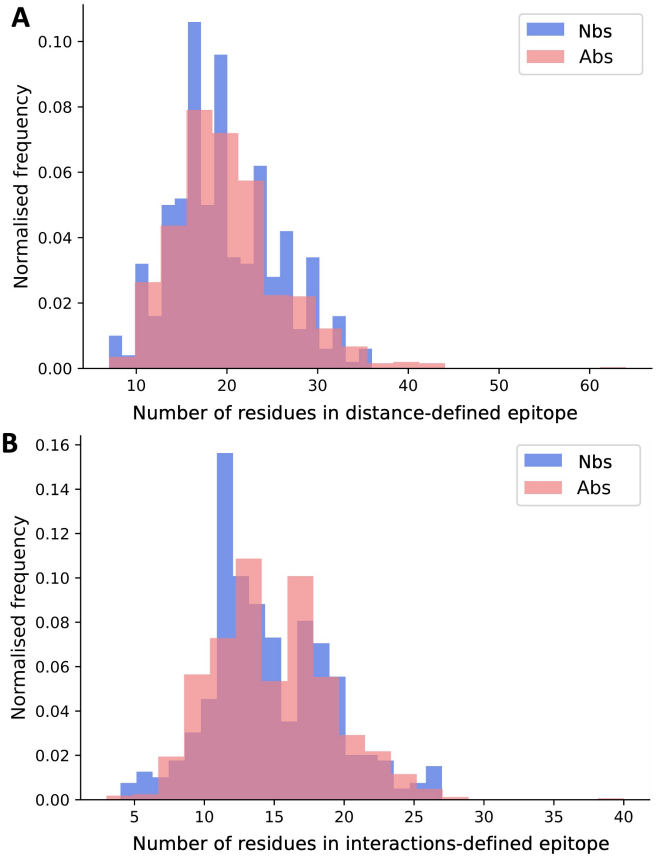


Figure 2.5: *Nanobodies are able to target epitopes of equal size (as defined by number of residues) to those targeted by conventional antibodies, as shown by the distributions of the number of residues in the (**A**) distance-defined epitopes for the antibody and nanobody structural datasets, where a mean difference of 0.59 is observed between nanobodies and antibodies (p-value = 0.22) and (**B**) interactions-defined epitopes, where a mean difference of 0.32 is observed between nanobodies and antibodies (p-value = 0.34).*

Despite their smaller paratope size, nanobodies target epitopes of equal size to those targeted by antibodies. This indicates that the paratopes of nanobodies must interact with their epitopes in a different way to that of the paratopes of antibodies.

2.4.9 Epitopes targeted by nanobodies and antibodies are of similar accessibility

In agreement with existing studies on smaller datasets, we found that nanobodies have longer CDR3 loops than the equivalent antibody CDR-H3 loop. Previous work has suggested that this facilitates interactions between nanobodies and epitopes that are less accessible to conventional antibodies [30, 44–47]. To assess whether the epitopes of nanobodies do indeed tend to be less accessible, the accessibility of all interactions-defined epitopes of nanobodies and antibodies was analysed.

We define epitope accessibility as the total relative SASA for the epitope surface, divided by the sum of the relative SASA values for each epitope residue if they were completely exposed (equivalent to the number of residues in the epitope).

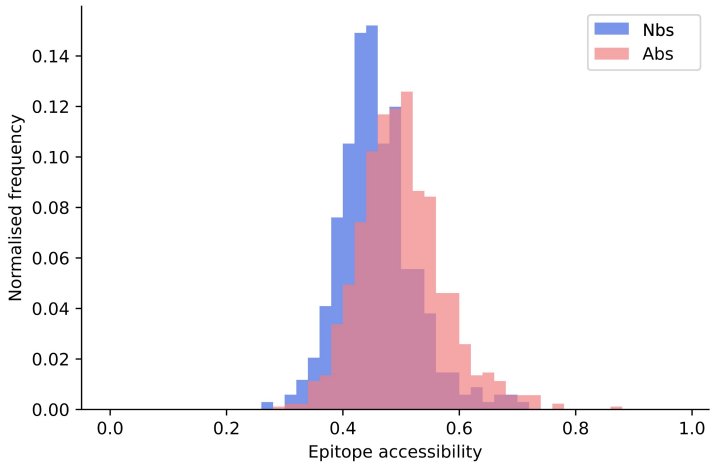


Figure 2.6: *Epitopes targeted by nanobodies are slightly less accessible than those targeted by antibodies. Distributions of epitope accessibility for the interactions-defined epitopes of nanobodies and antibodies were found to be significantly different, though the absolute difference is small: the unpaired mean difference between nanobodies and antibodies epitope accessibility was 0.046.*

We found that the epitope accessibility of nanobodies was significantly lower than that of antibodies: the unpaired mean difference between the epitope accessibility of nanobodies and antibodies was 0.046 (**Figure 2.6**). These results support previous studies that suggest that nanobodies are able to target epitopes that are inaccessible to antibodies [54]. However, there is also a large overlap in the distributions, and the absolute difference is small: this supports the suggestion from Henry and MacKenzie (2018) that there is likely overlap in the types of epitopes that nanobodies and antibodies target [223].

2.4.10 CDR3 loop length does not correlate with epitope accessibility

The hypothesis that nanobodies are generally able to target epitopes that are less accessible to conventional antibodies derives from the finding that their CDR3 loops are longer than those of antibody heavy chains [43, 44]. However, for our datasets, there is no correlation between the length of the CDR3 loop and the epitope accessibility (**Figure 2.7**). For nanobodies, the Pearson correlation coefficient for epitope accessibility against the CDR3 loop length was -0.021. For antibodies, the Pearson correlation coefficient for epitope accessibility against the CDR-H3 loop length was -0.097. These results indicate that the length of the CDR3 loop alone does not influence the accessibility of the epitope targeted by either antibody type.

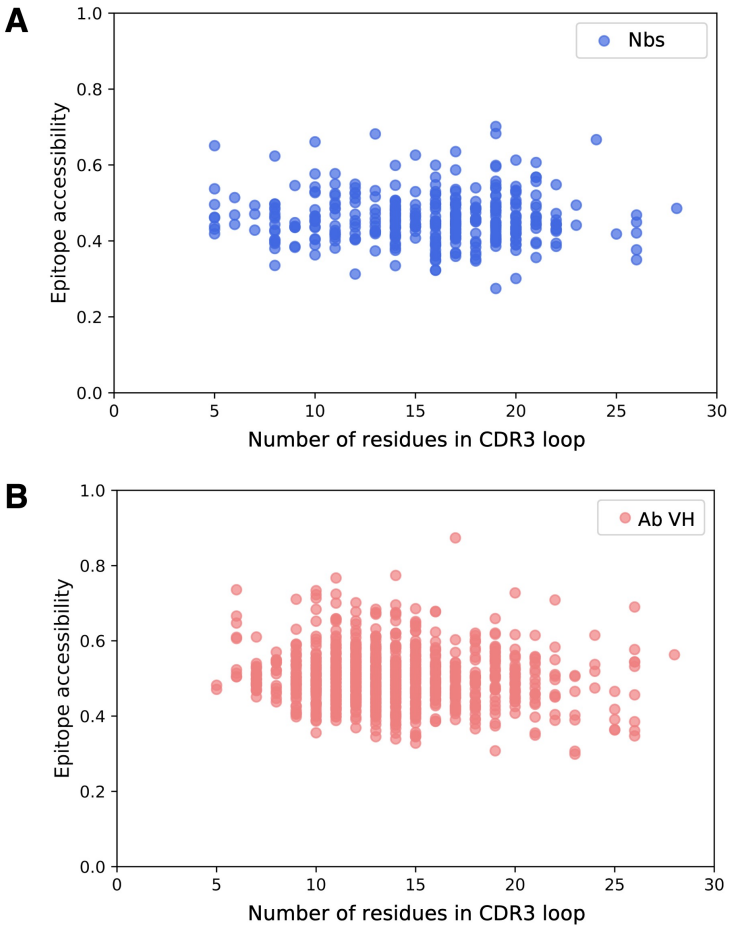


Figure 2.7: *There is no correlation between the length of the CDR3/CDR-H3 loop and the accessibility of the epitope surface for either nanobodies or antibodies. (**A**) Correlation between the accessibility of nanobody epitopes and the length of the CDR3 loop. (**B**) Correlation between the accessibility of the antibody epitopes and the length of the CDR-H3 loop.*

2.4.11 Antibodies and nanobodies target epitopes of similar accessibility due to packing of nanobody CDR3 loops against the VHH domain

In light of the finding that the length of the CDR3 loop does not dictate the accessibility of the epitope to which a paratope binds, we examined the differences in the orientation of antibody VH and nanobody CDR3 loops relative to the rest of the VH/VHH domain, to determine how the conformation of the CDR3 loop may affect epitope preference.

We use four descriptors to describe the orientation of the CDR3 loops (see **Methods**, **Text A.2** and **Figure A.3**). The parameter ρ (rho) represents the reach of the CDR3 loop away from the VH/VHH domain. ϕ (phi) describes the horizontal orientation of the CDR3 towards the rest of the VHH (for nanobodies), or against the VL domain (for antibodies). θ (theta) describes loop extension where a low value corresponds to a CDR3 extending up and away from the rest of the VH/VHH domain. Lastly, loop compactness is determined by dividing loop length by reach (ρ).

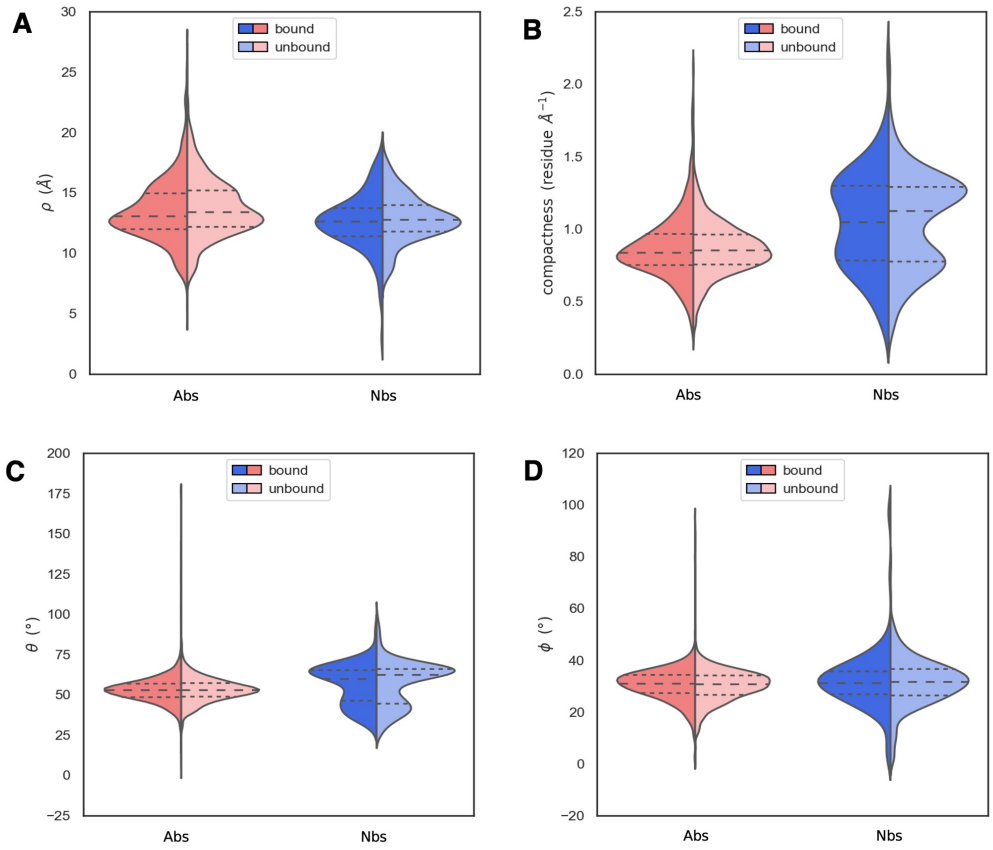


Figure 2.8: *The orientation of the CDR3 loops of nanobodies suggests why there is an overlap in the accessibility of the epitopes targeted by nanobodies and antibodies. (**A**) Distributions of ρ values show that nanobodies and antibodies have similar reach. (**B**) On average, nanobody CDR3 loops are more compact than antibody loops. (**C**) Distributions of θ values indicate that the majority of nanobody CDR3 loops do not extend upwards away from the VHH domain, but lie flat against it. (**D**) Distributions of ϕ values indicate that the majority of nanobody CDR3 loops pack against the VHH domain. In all figures, the results for the bound structures are shown in a darker shade on the left of the distributions, with the unbound structures in a lighter shade on the right.*

Near-identical distributions of ρ values suggests that antibody CDR-H3 and nanobody CDR3 loops have similar reach, indicating that nanobodies cannot necessarily provide extended paratopes via their CDR3 loops compared to antibodies (**Figure 2.8A**). A shoulder in the distribution of ρ values for antibodies above the median value suggests that antibodies may be more able to target deeper epitopes that require a longer reach.

The observation that nanobody CDR3 loops tend to be longer than the CDR-H3 antibodies, whilst having similar reach, may be explained by loop compaction. On average, nanobody CDR3 loops are much more compacted than antibody CDR-H3 loops (**Figure 2.8B**). The distribution of compactness scores for nanobodies is bimodal, with the first peak corresponding to the distribution found in antibodies. This suggests there is one population of nanobody CDR3 loops that behaves similarly to antibody CDR-H3 loops, with loops extending out from the framework, and another population with loops that tend to fold against the VHH domain. Nanobodies can either increase their reach with CDR3 length at a rate similar to antibodies, or their loops can remain in a more heavily compacted state (**Figure 2.9A**).

Compared to antibody CDR-H3 structures, nanobodies show a wider bimodal distribution of θ values, with the major peak corresponding to θ values in excess of those observed for antibody structures, and another minor peak below the antibody distribution (**Figure 2.8C**). This indicates that the majority of nanobody CDR3 loops lie flat against the rest of the VHH domain, therefore folding down. We observe a slight shift in θ in the distribution for bound nanobodies, but note that the position of the peaks still remains stable. We conclude that nanobodies generally do not extend their CDR3 loops upon binding, as has previously been hypothesised. Lastly, we find near-identical values of ϕ for nanobodies and antibodies, with nanobody ϕ values having a slightly wider distribution (**Figure 2.8D**).

To examine how loops pack against the VH or VL domains, we analysed the relationship between the spherical angles and compactness. Both nanobody CDR3 and antibody CDR-H3 loops become more compacted through an increase in θ – this can be interpreted as the packing of the loop down towards the rest of the VH/VHH domain decreases its reach (**Figure 2.9B**). We hypothesise that this is a mechanism to stabilise the paratope structure, by allowing the loop to pack against the rest of the VH/VHH domain. We also find an inverse relationship between compactness and ϕ for nanobodies and antibodies (**Figure 2.9C**). As ϕ increases (as the CDR3 loop is horizontally oriented away from the VH/VHH domain), nanobodies show an increase in compactness whereas the opposite is true for antibodies. For nanobodies, an increase in ϕ results in the CDR3 loop extending away into empty space, whereas in antibodies the CDR-H3 loop is positioned towards the VL domain. As the presence of the VL domain provides steric hindrance, the CDR-H3 loop is forced into a conformation that orients it away from the antibody, therefore reducing compactness and increasing reach.

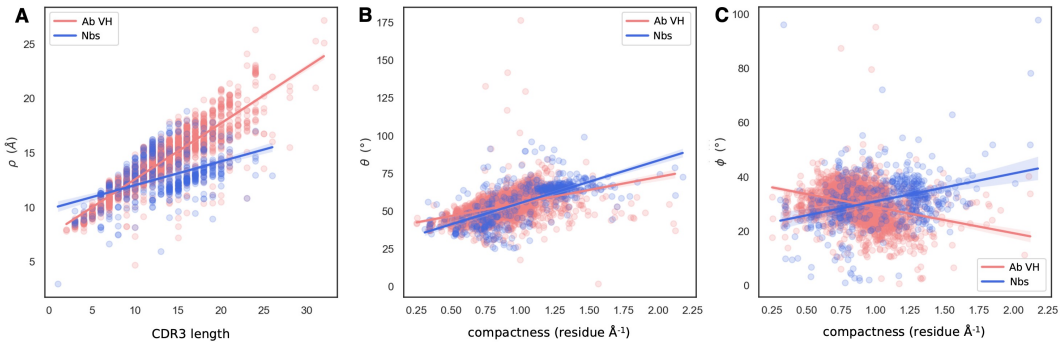


Figure 2.9: *Relationships between spherical angles and compactness suggest that the paratope is stabilised by the CDR3 loop packing against VL domains in antibodies, or the CDR3 loop packing against the the rest of the VHH domain in nanobodies. (**A**) Correlation between ρ and CDR3 length (**B**) Correlation between θ and compactness (**C**) Correlation between ϕ and compactness.*

2.4.12 Nanobodies establish more interactions with their epitope per paratope residue than antibodies

Our results thus far demonstrate that there are differences between the paratopes of nanobodies and antibodies. However, we find only limited differences between the epitopes targeted by these two modalities.

We have shown that for our datasets, antibodies and nanobodies are able to bind similarly-sized epitopes, despite the paratopes of nanobodies being smaller. In order to investigate how this is achieved, we compare the interactions observed within their binding sites. We find that, normalising for the size of the paratope, per paratope residue, nanobodies establish significantly more interactions than antibodies (**Figure 2.10**). This suggests that nanobodies establish a similar binding affinity to antibodies by each paratope residue having an increased number of interactions with the epitope.

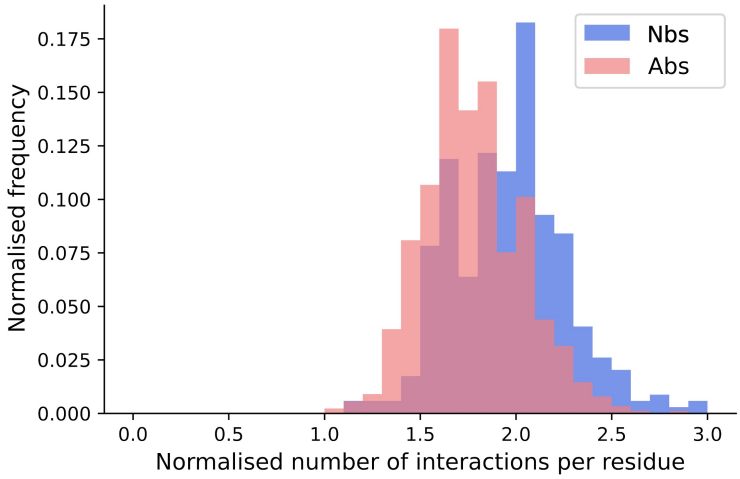


Figure 2.10: *The distributions of the number of interactions initiated by nanobody and antibody paratopes demonstrate that nanobody paratopes establish significantly more interactions per residue than antibody paratopes. Comparing the number of interactions from nanobodies to antibodies, normalised for paratope size, we find a mean increase of 0.19.*

2.4.13 Hydrophobic interactions dominate both nanobody-antigen and antibody-antigen complexes

As well as the number of interactions, the types of interactions established between nanobodies and antibodies and their antigens were compared. Here, the interactions-defined paratopes and epitopes are used. All interatomic interactions between the epitope and paratope residues were considered. Each type of interaction was counted individually if an atom-atom complex established more than one interaction type (see **Methods**).

In terms of interactions arising from the CDR loops, very similar types are observed (**Figure 2.11A**). For the framework regions involved in binding, we see an increase in hydrophobic interactions for nanobodies, compared to both antibodies and the VH domain of antibodies alone (**Figure 2.11B**).

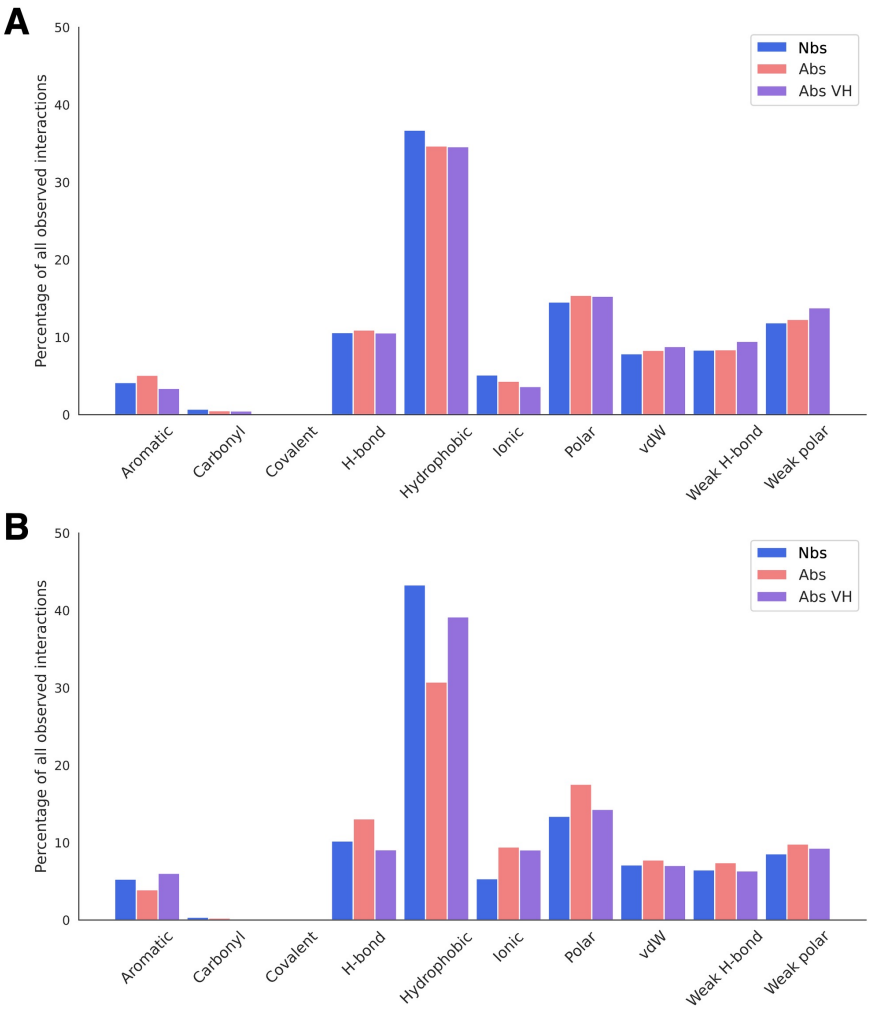


Figure 2.11: *Hydrophobic interactions dominate across nanobody-antigen and antibody-antigen complexes. Total occurrences as a percentage of all interaction types observed for the (**A**) CDR loops and (**B**) the framework regions. Results for nanobodies are shown in blue, for antibodies are shown in pink and for the antibody VH domains are shown in purple.*

2.4.14 CDR3 and framework residues are of increased importance for interactions in the nanobody-antigen complex

Next, we compared the relative contributions of the CDR loops to interactions within the binding site, including the mean number of interactions per loop (**Figure 2.12**, **Figure A.13**). In our data, we see the expected dominance of the CDR3 loop in binding. We found that there are significantly more interactions contributed

from the CDR3 in nanobodies than the CDR-H3 in antibodies (**Figure A.14A**) even after normalising for loop length (**Figure A.14B**) and that in nanobodies, there was a significantly greater contribution from the CDR3 residues both in terms of contributing residues to the paratope and contributing interactions (**Figure A.14C, A.14D**).

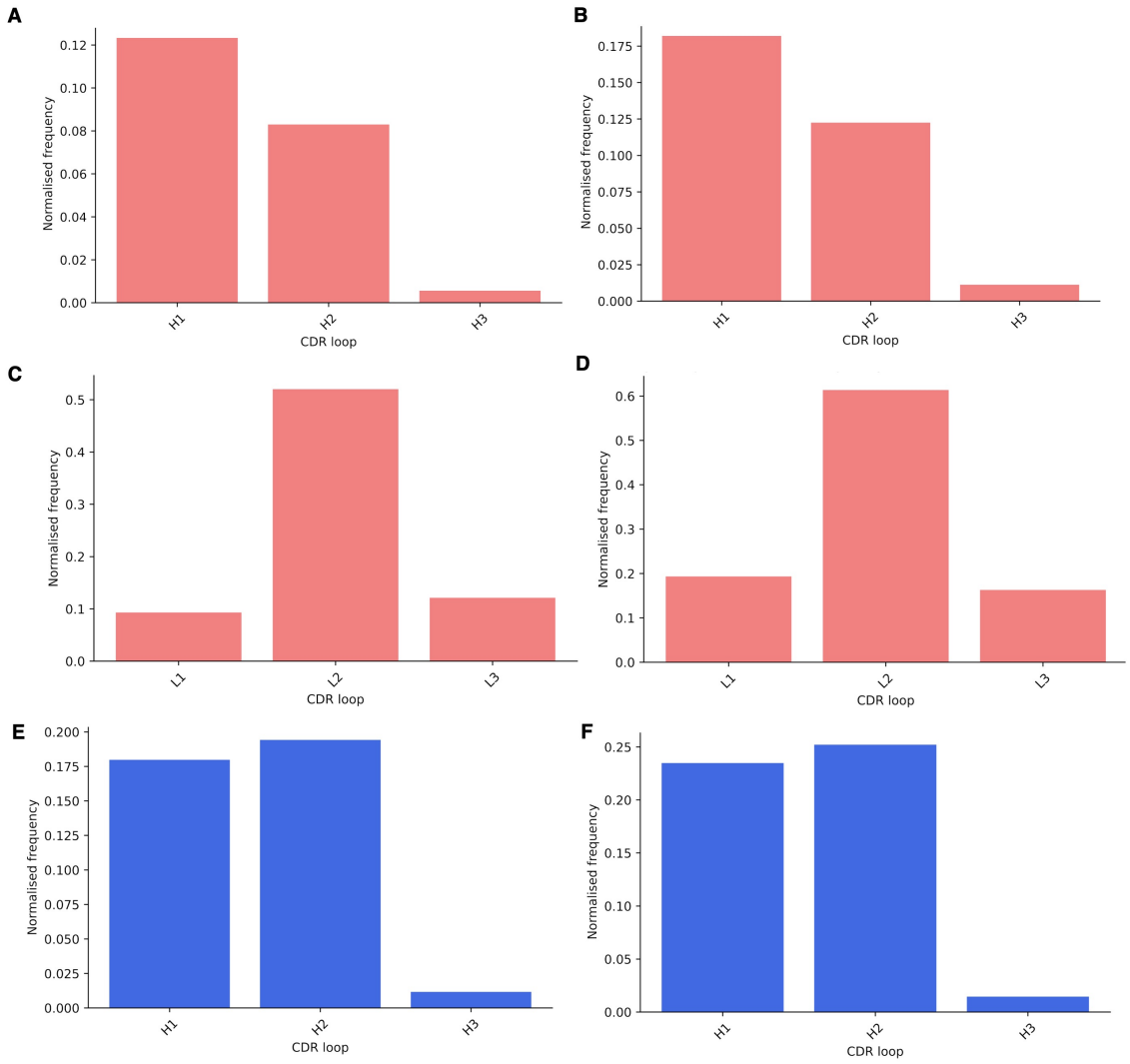


Figure 2.12: *Assessing the relative contributions of each CDR loop to the paratope shows that for both nanobodies and antibodies, the CDR3 and CDR-H3 loops rarely do not contribute interactions to the paratope. Bars show the number of times a CDR loop contributes zero interactions to a paratope as a proportion of all structures in that dataset for the distance-defined (**A**, **C**, **E**) and interactions-defined (**B**, **D**, **F**) paratopes for the antibody VH domains, antibody VL domains, and nanobodies respectively.*

When comparing the paratope of nanobodies only to the paratope residues from

the antibody VH domain, again significant differences are found (**Figure A.14E**, **A.14F**). We observe a minimal number of examples where the CDR3 loop contributes zero interactions (**Figure 2.12**). These results show that the highly variable CDR3 loop is even more dominant in nanobodies than the CDR-H3 in antibodies. This, however, is not the only difference: we also observe that the paratopes of nanobodies tend to contain a smaller proportion of CDR residues than antibodies (**Figure 2.13**, **Figure A.15**), from which we can infer that nanobodies show greater inclusion of framework residues in their paratopes than antibodies.

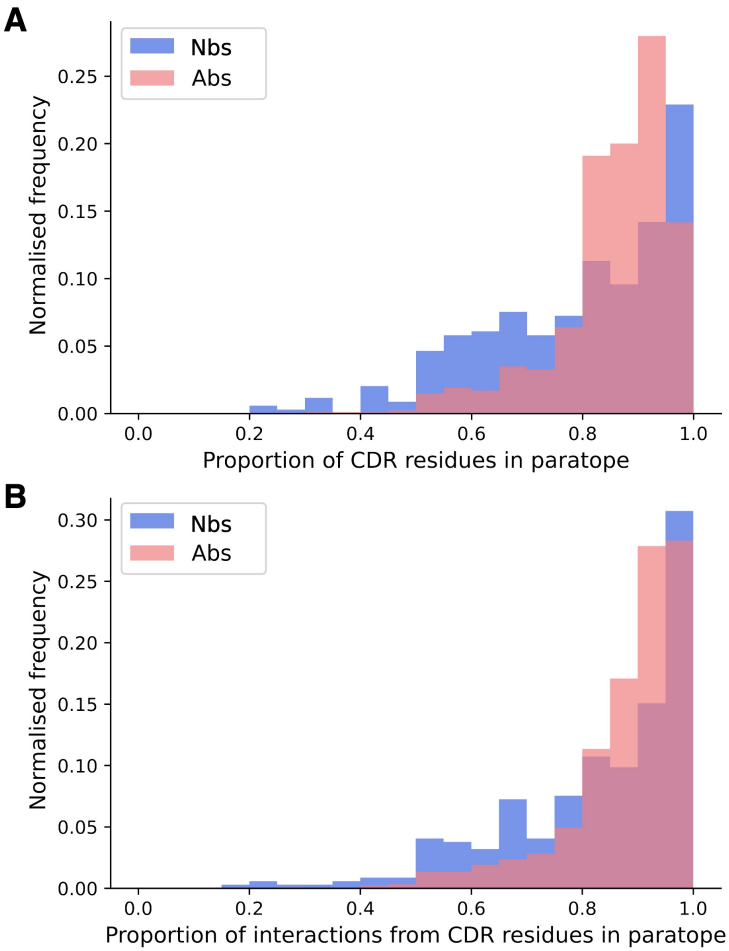


Figure 2.13: *Distributions of (**A**) the proportion of CDR residues in the paratope and (**B**) the proportion of interactions from CDR residues across the whole paratope, determined per complex in the nanobody and antibody structural datasets. Higher density on the lower end for the nanobody dataset, compared to the antibody dataset, indicates that more framework residues are involved in binding the epitope in nanobodies.*

2.4.15 Interacting framework residues are often conserved in nanobodies

Given we find that framework residues make up a larger proportion of the paratope in nanobodies than in antibodies (**Figure 2.13**), we next tested if these framework residues show high variability, undergoing somatic hypermutation to improve binding, or are conserved germline residues.

Framework residues observed in the interactions-defined paratope in at least 10% of the nanobody complexes were determined (**Table A.3**). In descending order of frequency, this includes IMGT positions 66 (50.4%), 52 (31.6%), 55 (27.2%), 42 (24.1%), 50 (17.4%), 118 (15.9%), 69 (12.8%), 67 (12.8%), 40 (10.4%), and 2 (10.1%).

The amino acid compositions of these identified framework positions were determined for both of the structural datasets and for the sequence datasets (**Figure 2.14**). Positions were not included if less than 5% of the structures or sequences had a residue at that position. We compare the positions found in the interactions-defined paratopes from the structural datasets to a background composition taken from the sequence datasets. The sequence logo plots show similarities between the paratope composition and background, particularly for positions 2, 50, 67, 69 and 118 in nanobodies. The low level of variation at these positions in nanobodies indicates that they are conserved, and suggests that they may not contribute to binding specificity.

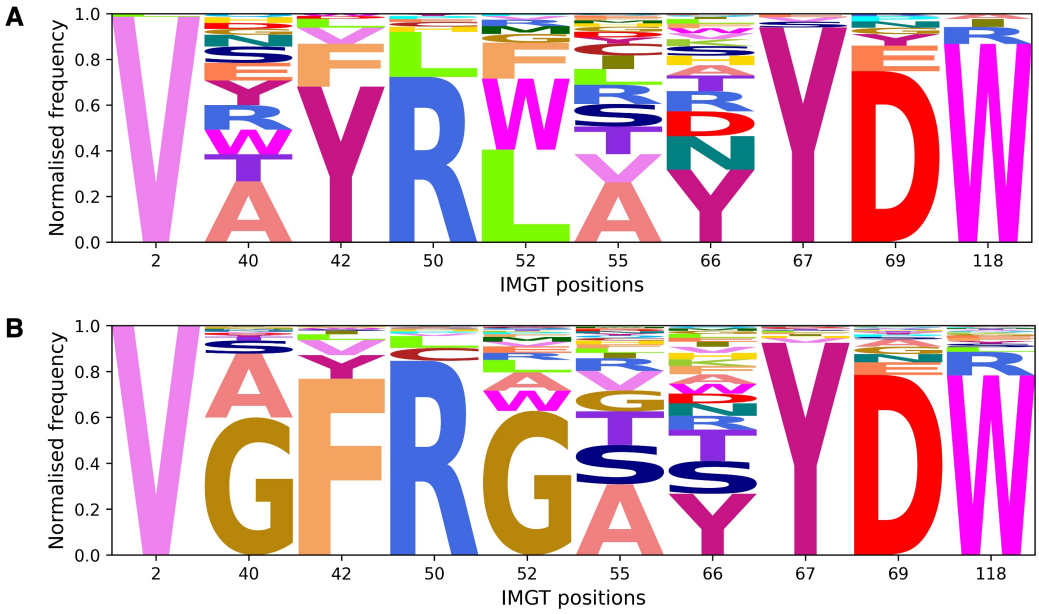


Figure 2.14: *Sequence logo plots of amino acid composition at framework positions often involved in the paratopes of antibodies and nanobodies suggest that framework residues identified to often occur in the paratope are highly conserved in nanobodies. (**A**) Amino acid compositions at positions found in at least 10% of nanobody paratopes in our nanobody structural dataset. (**B**) Background amino acid compositions in our nanobody sequence dataset for positions found in at least 10% of nanobody paratopes. Positions were not included if less than 5% of sequences had a residue at the given position.*

2.4.16 Antibodies and nanobodies can bind the same epitopes but interact with them differently

Our results suggest that antibodies and nanobodies can engage similar types of epitopes but use different mechanisms to do so. Here, we compare the features of an antibody (PDB ID: 6YLA [235]) and a nanobody (PDB ID: 6WAQ [236]) that both bind to the receptor-binding domain (RBD) of the SARS-CoV-2 spike protein, using the interactions-defined binding sites.

The nanobody has a longer CDR3 (18 residues) than the antibody CDR-H3 (12 residues) and the nanobody paratope is smaller than that of the antibody (15 compared to 26 residues). The nanobody paratope includes framework positions 66 and 69, both of which we found to be commonly part of nanobody paratopes. The antibody paratope includes framework positions 1 from the heavy chain and

position 68 from the light chain.

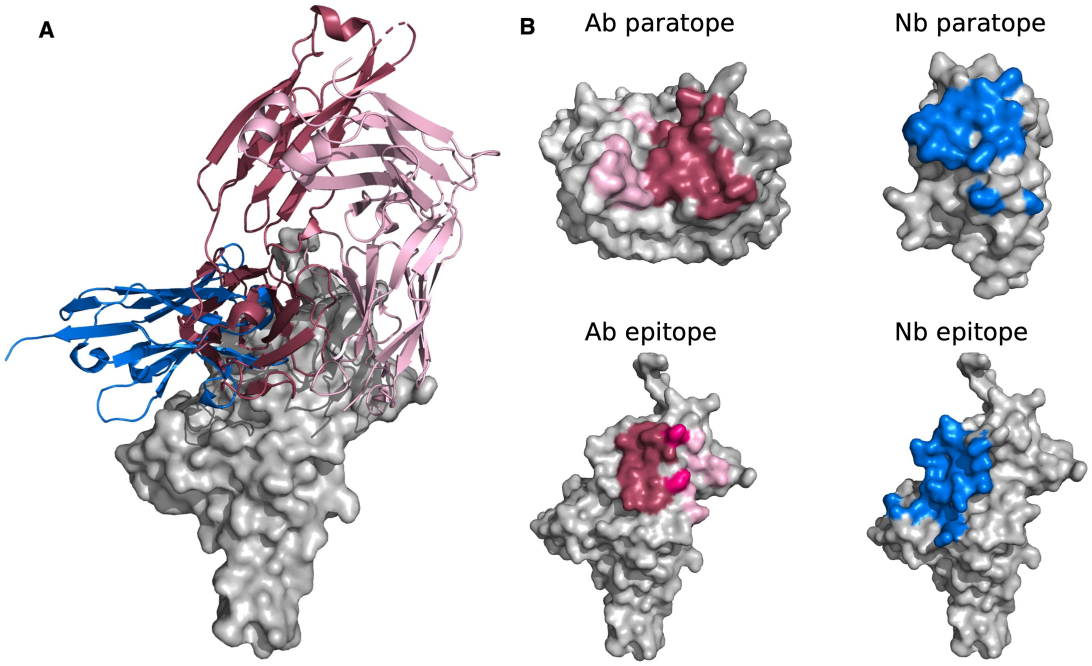


Figure 2.15: *(**A**) A nanobody (PDB ID: 6WAQ [236]) and antibody (PDB ID: 6YLA [235]) are able to bind the SARS-CoV-2 RBD with overlapping epitopes. Dark pink cartoon = antibody heavy chain, light pink cartoon = antibody light chain, blue cartoon = nanobody, grey = surface representation of the SARS-CoV-2 RBD. (**B**) Antibodies in general have larger paratopes than nanobodies, but nanobodies are able to bind similarly-sized epitopes as exemplified by structures 6YLA (antibody) and 6WAQ (nanobody). The surface of the antibody heavy chain is shown in dark grey and the light chain in light grey. The dark pink region represents paratope residues contributed by the VH and the light pink region represents paratope residues contributed by the VL. The surface of the nanobody is shown in light grey with the blue region representing the nanobody paratope residues. The surface of the SARS-CoV-2 antigen is shown in light grey for both the nanobody and antibody. The antibody epitope is coloured dark pink where it is targeted by the antibody VH, light pink where it is targeted by the antibody VL, and a medium pink where it is targeted by both chains. The nanobody epitope is shown in blue. The antigen structures from each PDB were merged to create a complete image of the antigen for the nanobody.*

Despite the differences in the nanobody and antibody paratopes, they are binding a very similar epitope (**Figure 2.15**). The epitopes on the RBD that these structures bind are of a similar size (15 residues for the antibody epitope and 18 residues for the nanobody epitope).

Thirty-one total interactions occur between the antibody paratope and the epitope that it targets, whilst there are twenty-nine for the nanobody binding site. However, when we consider the size of the paratope, this results in an average of 1.9 interactions per paratope residue for the nanobody, compared to 1.2 per antibody paratope residue. In addition, the CDR3 loop has increased importance for the nanobody binding activity. For the antibody, 6 out of the 26 residues in the paratope come from the CDR-H3 loop, whereas for the nanobody, it is 9 out of 15.

2.5 Discussion

In this chapter, we compared the binding sites of camelid nanobodies and antibodies, to assess whether these two modalities may be suited to different types of epitopes and understand the most important features involved in their antigen interactions. Overall, we conclude that the paratopes of nanobodies and antibodies have distinguishable characteristics. Paratopes of nanobodies tend to be smaller, the CDR conformations observed are different between nanobodies and antibodies, and nanobodies tend to have longer CDR3 loops than their antibody counterparts. These results are all consistent with previous studies on smaller datasets [54, 55, 224].

These differences in their paratopes led to the expectation that antibodies and nanobodies would bind distinct types of epitopes. However, we find that, apart from the epitopes of antibodies being slightly more linear than those of nanobodies, the epitopes targeted by nanobodies and antibodies cannot be easily distinguished. Nanobodies and antibodies target epitopes of similar size, similar amino acid compositions and similar accessibility.

There are several suggestions in the literature that the longer CDR3 loop of a nanobody means it can interact with epitopes that are less accessible to conventional antibodies by protruding into the cavity [45–47]. Henry and MacKenzie (2018) [223] stress that despite individual case studies supporting this hypothesis, the

evidence that nanobodies preferentially bind more cryptic epitopes is limited, and it is unknown whether this is a general trend across nanobodies. We find that overall, for our datasets, the epitopes targeted by nanobodies are slightly (but significantly) less accessible than epitopes targeted by antibodies. However, the absolute difference is small. Furthermore, we find no correlation between CDR3 loop length and epitope accessibility.

These results are supported by our finding that antibody CDR-H3 and nanobody CDR3 loops show differences in their orientation relative to the rest of the supporting VH/VL or VHH domain. We find that nanobody CDR3 loops are more compact than antibody CDR-H3 loops and are often found packed against the rest of the VHH domain. For antibodies, orientation of the CDR-H3 away from the VH domain leads to its positioning towards the VL domain. As the presence of the VL domain provides steric hindrance, the CDR-H3 loop is forced into a conformation that orients it away from the antibody, therefore reducing compactness and increasing reach. In contrast, for nanobodies, orientation of the CDR3 away from the VHH domain leads to positioning towards empty space and therefore packing against the rest of the VHH domain. These results offer a possible explanation for our observation that the longer CDR3 loops of nanobodies do not necessarily target deeper epitopes.

In addition, we more often observe framework residues in the paratopes of nanobodies. The importance of framework residues in nanobodies has been indicated in several studies [54, 55, 209, 237]. This increase in framework residues is likely related to their increased accessibility due to the lack of the VL domain. Indeed, our results show that most of the framework positions observed in more than 10% of the nanobodies paratopes are frequently observed in the VH-VL interface of antibodies [238]. Most of the framework positions commonly involved in binding in nanobodies belong to the FR2 region, which is identified by both Zavrtanik *et al.* (2018) and Mitchell and Colwell (2018) as an important region for antigen binding

[54, 55]. The majority of the potential paratope framework residues that we identify appear to be highly conserved. Our findings that nanobody CDR3 loops often pack against the VHH domain, and that FR2 residues are often conserved, is in agreement with that of Sang *et al.* (2022), who find that the longer CDR3 loops of nanobodies can fold back to interact with FR2 residues [209].

Finally, we also find that despite tending to have smaller paratopes, nanobodies are able to target similarly-sized epitopes to antibodies. This may be explained by our finding that the CDR3 loops of nanobodies make a significantly greater number of interactions with the epitope per loop residue than the CDR-H3 of antibodies, even after normalising for loop length. Given that these interactions may include those from conserved framework residues, which will contribute to binding affinity but not specificity, this raises important questions over the specificity of the nanobody binding site and has implications for the engineering of nanobody therapeutics.

Overall, this chapter highlights the structural characteristics of nanobodies pertinent to the design and engineering of nanobody-based therapeutics, in particular the use of framework regions in binding and the conformation of the CDR3 loop. These results call attention to the need for additional criteria, such as developability, when deciding on the best modality for a particular epitope, and strengthens the argument for the unique features of nanobodies to be taken into consideration when developing protocols and computational tools specifically for nanobody design. In the next chapter, we collate data from patents and the literature into a database, PLAbDab-nano. This central repository provides a curated and annotated set of nanobodies with which we can build nanobody-specific computational methods.

3

PLAbDab-nano: a database of camelid and shark nanobodies from patents and literature

Contents

3.1 Abstract **68**
3.2 Introduction **69**
3.3 Methods **72**
3.4 Results **78**
3.5 Discussion **84**

3.1 Abstract

In **Chapter 2**, we highlighted key differences between camelid nanobodies and conventional antibodies, demonstrating the need to account for their unique features when developing computational methods for nanobody design. A primary challenge in advancing these methods for nanobodies is the limited data availability: as described in **Chapter 1**, monoclonal antibodies are significantly more established as therapeutics than nanobodies, meaning datasets are more readily available. For example, there are far more publicly available antibody repertoire datasets than there is for nanobodies: at present, approximately ~0.8 million camel VHH sequences are deposited in the OAS database, compared to over 2 million paired and 2.4 billion unpaired antibody sequences [146, 147]. The amount of therapeutic data, in particular, for nanobodies is a fraction of that available for antibodies. Therefore, efforts thus far towards developing computational methods have been focused on conventional antibodies.

However, the volume of academic literature and patents related to nanobodies has risen significantly over the past decade. This includes both VHHs and VNARs, reflecting the rising interest in these alternative formats for therapeutic applications. To centralise the available data, we created a nanobody database, PLAbDab-nano, which collects sequences from patents and literature. The database is a searchable repository containing ~5000 annotated VHH and VNAR sequences, which can be easily updated via an automated pipeline, allowing us to benefit from the growing corpus of data over time. PLAbDab-nano could be used to design diverse libraries as well as find sequences similar to known patented or therapeutic entries.

The text in this chapter is based on material from the following publications:

> **Gordon, G.L.**, Greenshields-Watson, A., Agarwal, P., Wong, A., Boyles, F., Hummer, A., Lujan-Hernandez, A.G., and Deane, C.M. PLAbDab-nano: a database of camelid and shark nanobodies from patents and literature. *Nucleic Acids Research* **53**, D535–D542, (2025).
>
> Greenshields-Watson, A., Agarwal, P., Robinson, S.A, Williams, B.H., **Gordon, G.L.**, Capel, H.L., Li, Y., Spoendlin, F.C., Aguilar Sanjuan, B., Boyles, F., and Deane, C.M. ANARCII: A Generalised Language Model for Antigen Receptor Numbering. *bioRxiv* (2025).

I developed the main PLAbDab-nano web-scraping pipeline and curated the database, based on previous work by Brennan Abanades. Alex Greenshields-Watson and Parth Agarwal developed the language model used for VNAR CDR3 annotation; I curated the VNAR data used to refine the model. Ashley Wong and Alissa Hummer developed the random forest classifier model used in this work, which I retrained and implemented as part of the PLAbDab-nano pipeline. Fergus Boyles and I created the web application.

3.2 Introduction

The volume of academic literature and patents related to nanobodies has grown substantially over the past decade, driven by increasing interest in their therapeutic potential. As we described in **Chapter 1**, this not only concerns camelid-derived VHH nanobodies, but also extends to VNARs, a single-domain immunoglobulin derived from shark IgNARs [31, 56, 76]. VNARs, like VHHs, have a long CDR3 loop, but unlike VHHs they have only two CDR loops (CDR1 and CDR3) and two additional hypervariable loops (HV2 and HV4). This may broaden the range of epitopes that can be targeted by nanobody-based therapeutics [56, 57, 239] (**Figure 3.1**).

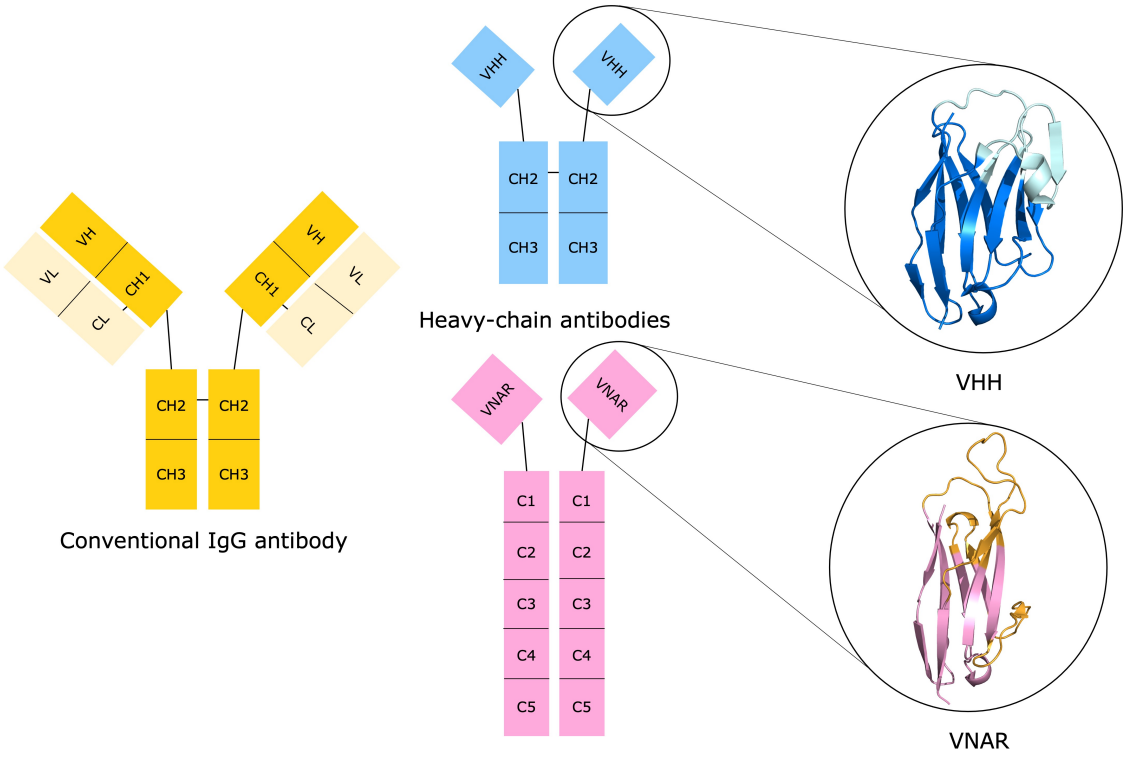


Figure 3.1: *Nanobodies are derived from the antigen-binding portion of heavy-chain antibodies (hcAbs), which lack the light chain pairing possessed by conventional antibodies (shown in yellow). There are two varieties of nanobody: the VHH from camelid species (shown in blue, with the CDR loops in light blue), and the VNAR from sharks (shown in pink, with the CDR loops and hypervariable regions shown in orange). PDB entries 8HR2 [240] and 7S83 [74] were used to create the VHH and VNAR figures respectively.*

Previous work has sought to collect nanobody data from many sources, but exclusively for VHHs. For example, the antibody sequence database OAS [146, 147] aims to collect immune repertoire data. At the time of writing, OAS contains ~0.8 million VHH sequences, but as these are from repertoire studies they do not have any functional annotation. The structural data for 1804 unique PDB IDs containing VHHs are available in SAbDab-nano [142]. TheraSAbDab is a collection of therapeutic antibodies, and currently contains 22 VHH or single-domain antibody entries, with 36 unique sequences [153]. CoV-AbDab, a database for antibodies targeting the COVID-19 virus, contains 801 nanobodies [149].

Other repositories dedicated solely to nanobodies have also been curated, drawing from multiple public sources, such as sdAb-DB [158], INDI [156], and most recently,

NanoLAS [157]. The sdAb-DB database contains only a small number of entries, with a total of 1446 sequences from publications, PDB entries and GenBank, suggesting it is not regularly updated. The INDI database contains approximately 11 million sequences from next-generation sequencing (NGS) studies, and ~21,000 sequences from patents and literature. The data is available to download, however our analysis (see **Results**) indicates that most of the sequences from patents and literature may be VH sequences and have not been verified as being single-domain antibodies. NanoLAS relies on INDI as a source for their ~20,000 sequences. Also, to our knowledge, none of these include VNAR data and thus they are missing an important component of the available nanobody data.

In this chapter, we describe the methods used to curate PLAbDab-nano, a database currently containing 4908 annotated VHH and VNAR sequences from 793 small scale studies. We offer open access to the data through a web application[1] (**Figure 3.2**) and a Python package[2], with functionality to search the database by sequence identity (using KA-search [241] and BLAST [117]), structural similarity, or keywords. Each sequence is accompanied by a direct link to its source material, facilitating access to further information on any nanobody of interest.

[1] *https://opig.stats.ox.ac.uk/webapps/plabdab-nano*
[2] *https://github.com/oxpig/PLAbDab_nano*

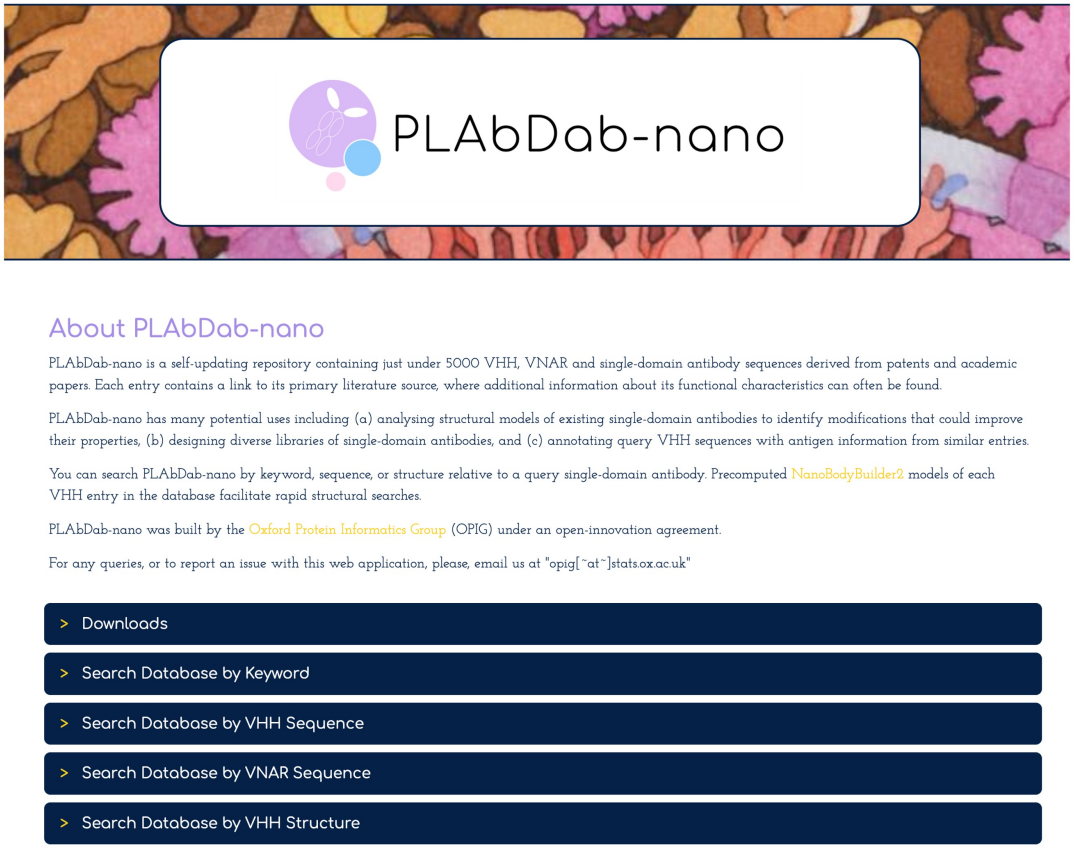


Figure 3.2: *PLAbDab-nano is freely accessible via a web application. Users can download the data and use the search functions to filter for nanobody data within chosen constraints.*

3.3 Methods

3.3.1 Collecting nanobody sequences

Sequences are drawn from the NCBI GenBank Protein database [242], SAbDab [141, 142], and TheraSAbDab [153] (**Figure 3.3**). Data is extracted from GenBank by querying the database with keywords, following the methods set out by Abanades *et al.* (2024) for the PLAbDab antibody database [152]. Here, we have used nanobody-specific search terms (*nanobody, nanobodies, VHH, VH, VNAR, shark novel antigen receptor, single chain antibody, single-chain antibody, single chain antibodies, single-chain antibodies*).

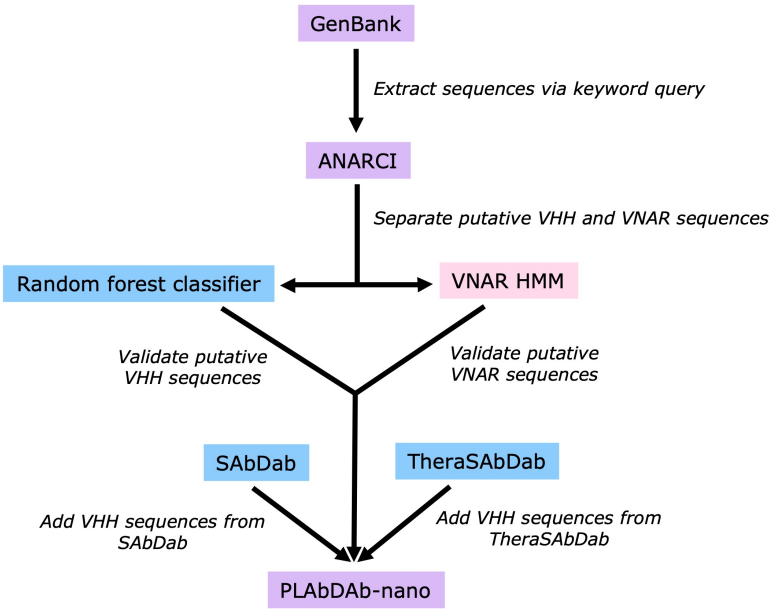


Figure 3.3: *Data is scraped and processed from GenBank [242], SAbDab [123, 142] and TheraSAbDab [153] to generate the PLAbDab-nano database.*

Entries resulting from this search are filtered using ANARCI [123] to identify antibody sequences, leveraging the Hidden Markov Models (HMMs) that support the numbering tool. Since ANARCI was not designed to handle VNAR sequences, separate HMMs were built with a manually curated set of 377 non-redundant VNAR sequences (sourced from GenBank) to mine VNARs from the search results. An HMM bitscore of 100 was used as a threshold above which sequences were classified as VNARs, based on the distributions of bitscores for a set of human VH and camelid VH/VHH sequences from SAbDab compared to the shark VNAR sequences (**Figure B.1**).

Subsequently, we trained a random forest classifier model on balanced datasets of 504,894 camel VHH sequences (from Li *et al.* (2016) [52]) and VH sequences. The VH sequences consisted of an equal mix of camel, human (IGHV3 only) and additional non-human sequences, curated by Marks *et al.* (2021) [198]. The data were divided into train, validation and test sets with a split of 80:10:10, achieving an accuracy of 91% and a ROC AUC of 0.95. This model was used to exclude VH sequences from the database using a threshold of 0.55, where 0 corresponds

to a VH sequence, and 1 a VHH sequence. This was calculated using Youden's J statistic, which optimises the balance between sensitivity and specificity, thus reducing misclassification of antibody VH sequences whilst accurately identifying nanobodies. For the first iteration of our database, we relaxed this threshold to 0.45, to minimise the loss of true nanobodies that share a high sequence similarity with VH sequences, such as those that have been humanised (**Figure B.2**). Updates to the database since have been conducted at a more conservative threshold of 0.6, to minimise the need for manual curation after the automated update.

Further nanobody entries were retrieved from SAbDab [141, 142] and TheraSAbDab [153]. TheraSAbDab was filtered for only nanobody entries, including multi-specific therapeutics (which are counted as separate entries for each unique sequence), and all stages of clinical trials.

3.3.2 Annotation of VHH and VNAR sequences

ANARCI [123] was used to annotate the CDR loops for all of the VHH sequences. VNARs diverge from the standard format of normal antibody heavy chains: they contain sequence features such as large gaps, an enrichment of cysteines, and they do not have the CDR2 loop seen in antibodies and camelid nanobodies [56, 58]. As such, tools such as ANARCI and AntPack [243] struggle to number such sequences, as their methodologies cannot generalise beyond the patterns captured by their reference sets or consensus sequence. When attempting to number VNARs, alignment-based methods tend to either fail to return any alignment, or else misnumber key conserved residues such as Cys104 (IMGT numbering). Therefore, for VNARs, we developed a bespoke method to annotate the CDR1 and CDR3 loops, using a large language model with a standard Seq2Seq transformer architecture [244].

The Seq2Seq model was trained on ~95.7M heavy, ~41.6M kappa and ~33.4M lambda non-redundant antibody sequences sampled from OAS [146, 147]. Sequences were clustered by identity and split into train, test, and validation sets (90:5:5). To

ensure they were representative and to reduce redundancy, up to 100,000 heavy and 500,000 light chain sequences were randomly sampled per file and duplicates were removed from the combined dataset. The number of sequences collected after this step were: ~95.6 M heavy and ~75.0 M light chains. All sequences were renumbered with ANARCI. Unsampled heavy sequences were used to enrich the numbers of rare insertions and non-human or mouse species contained in the dataset.

The model was used to number 456 VNAR sequences curated from GenBank [242] via our web-scraping pipeline. These numberings were compared to the IMGT germline reference for nurse shark V-genes (those with a large CDR2 deletion, for example IGHV2S1) to understand where the language model predictions differed. The model was then conditioned – defined as a simple correction at the time of autoregressive inference [245] – to correctly identify the CDR2 gap, GXGT motif (which denotes the end of the CDR3), and conserved residues in all cases.

The model was conditioned by inspecting numbering accuracy compared to the germline reference set, and identifying key patterns which preceded 'misnumbering' by the language model (according to IMGT germline annotations of a shark V gene). To correct the misidentification of the starting sequence, we identified the beginning pattern being missed and labelled the first residue according to IMGT numbering (often 1 or 2). Next, we used regular expressions to identify the sequence associated with the deleted CDR2 region (often flanked by a NEE motif), and labelled the subsequent sequence starting from residue 68 to preserve the large deletion characteristic of VNARs. The CDR3 annotation was completed by identifying the conserved cysteine at position 104 and the terminal GXGT motif at positions 119-122. Once these two patterns were found, then the intermediate sequence was renumbered according to standard rules.

Additional metadata, including antigen information, is added following the original PLAbDab database [152], with a small adjustment for species annotation:

GenBank species annotation was replaced by labels from SAbDab, to account for occasions on which non-nanobody chains within an entry caused mislabelling of the origin organism.

3.3.3 Searching PLAbDab-nano

KA-search [241], an antibody-specific rapid sequence search tool, can be used to carry out a sequence similarity search of VHH entries. Due to the dependency of KA-search on ANARCI, a separate search function for VNARs was built using BLAST [117]. Users can search over the whole sequence, all CDRs, or the CDR3 loop, with or without matching the length of the query, and a user-defined sequence identity threshold can be set (**Figure 3.4**).

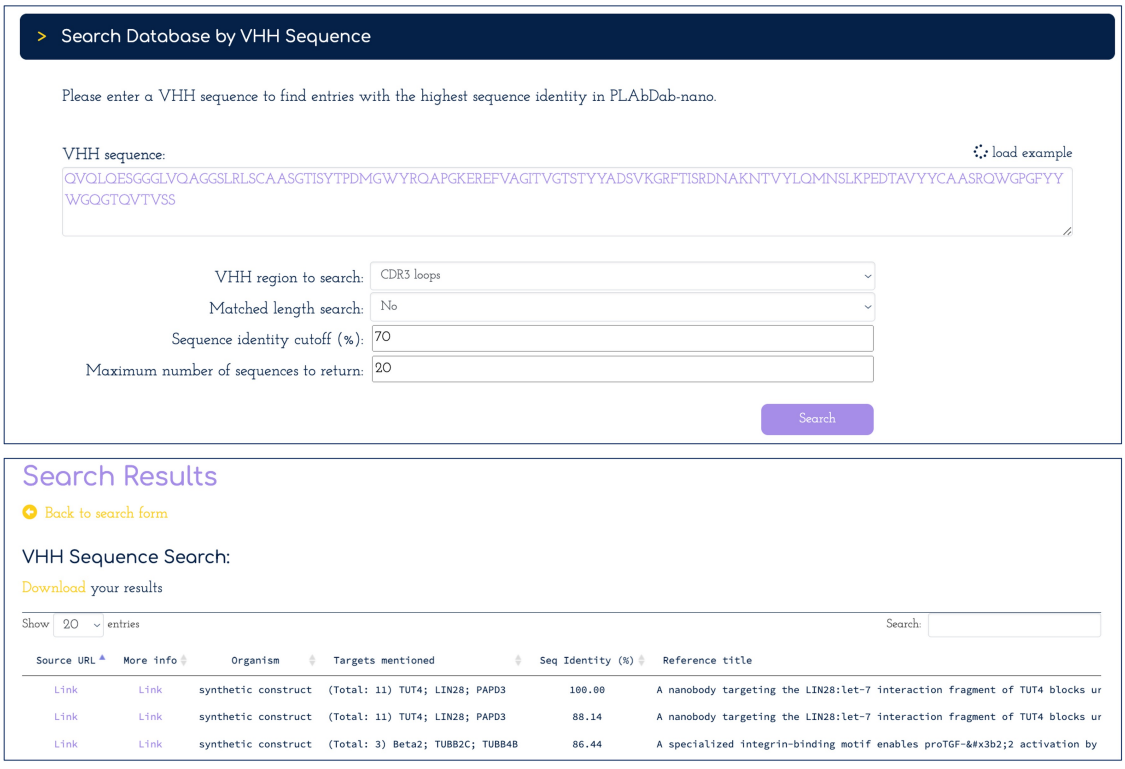


Figure 3.4: *An example of the PLAbDab-nano sequence search functionality on the web application. Users will input their sequence and specify their constraints; the web application will return sequences that match the query in order of decreasing sequence identity.*

The VHH sequences were modelled using NanoBodyBuilder2 [136], allowing users to search for similar structures. Sequences for which NanoBodyBuilder2 was not able to generate a model, including all VNAR sequences (as the NanoBodyBuilder2 model is not trained on VNARs), are labelled as 'FAILED' in the 'model' column of the database. Input query sequences are modelled without refinement and the framework is aligned to entries with equivalent CDR loop lengths. As in the original PLAbDab database [152], entries are ranked using the carbon-alpha (Cα) root-mean squared deviation (RMSD) over all CDR residues. Lastly, a text search is available for users to search the metadata by keyword (**Figure 3.5**).

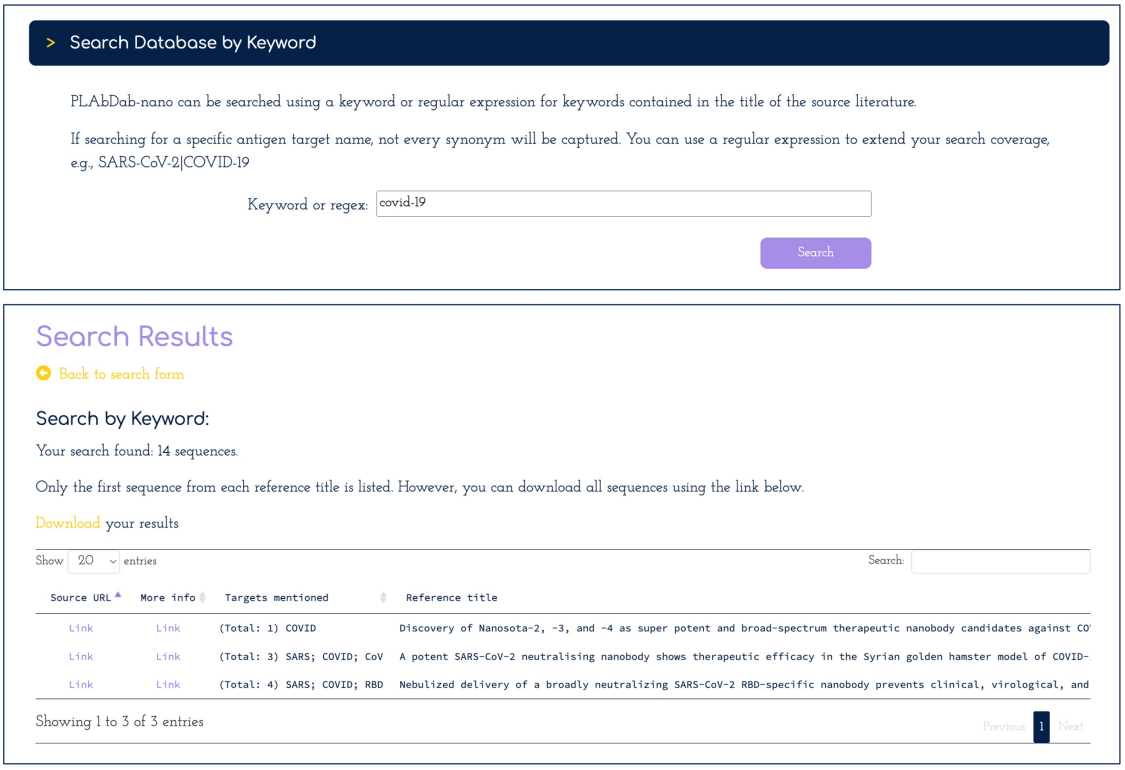


Figure 3.5: *An example of the PLAbDab-nano text search functionality on the web application. Users will input a keyword of interest, for example a particular antigen. The web application will return results with metadata that matches the keyword.*

3.4 Results

3.4.1 Database statistics

The cumulative number of publicly available sequences has increased substantially over the past decade (**Figure 3.6**). Currently, the total number of entries in PLAbDab-nano is 4908 (**Table B.1**).

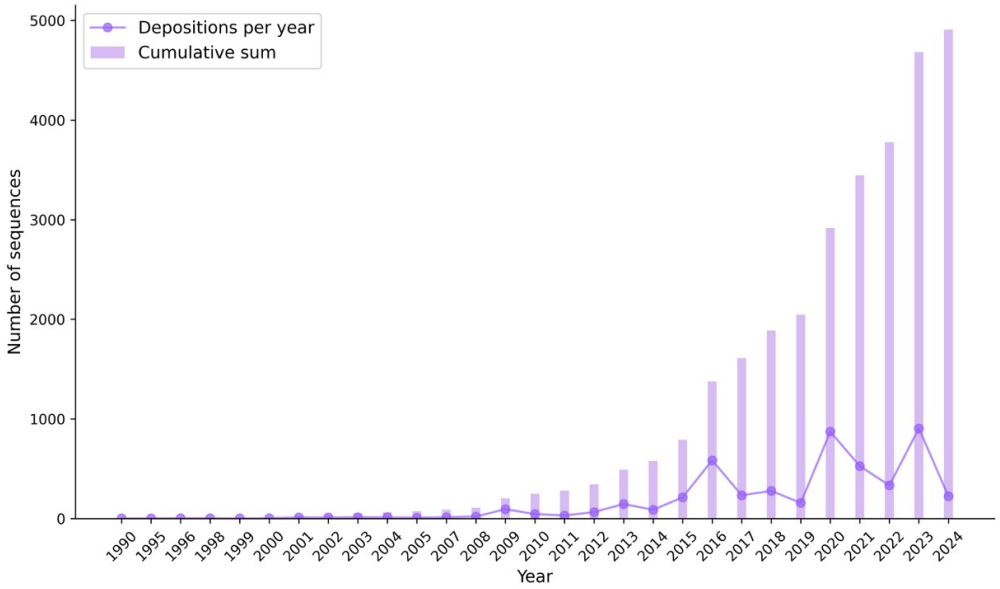


Figure 3.6: *There is an increasing number of depositions of nanobody sequences per year to publicly available sources.*

Compared to the INDI database [156], PLAbDab-nano has fewer entries overall. However, our methods more stringently filter data from GenBank in an attempt to include VHHs only. We exclude VH sequences since it is challenging to automatically verify whether these sequences originate from conventional antibodies, or are engineered single-domain antibodies. For our random forest classifier, a threshold of 0.55 was calculated using Youden's J statistic. Applying our model to INDI's sequences derived from the PDB, GenBank, patents and literature, indicates that, at the 0.55 threshold, almost 70% of their data are classified as VH sequences (**Figure 3.7**, **Figure B.3**). These have not been verified as being nanobodies before their inclusion in INDI (**Table B.2**).

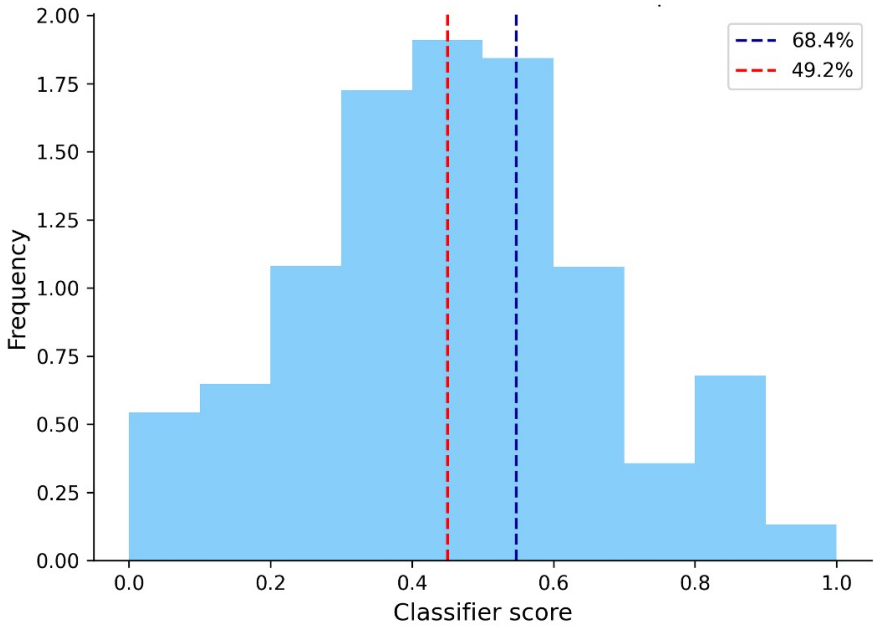


Figure 3.7: *The distribution of classifier scores for sequences from patents and literature in the INDI database [156] indicates that 68.4% of entries at our standard threshold of 0.55 (blue dashed line) and 49.2% at a relaxed threshold of 0.45 (red dashed line) would not pass our filtering methods, as they are classified as VH sequences by our random forest model.*

When accounting for redundant sequences, entries in PLAbDab-nano are sourced relatively evenly from the different sources: patents, literature, and crystal structures, with a minority from therapeutic data (**Figure 3.8A**). Excluding patents due to a lack of species annotation, for VHHs most entries are labelled as from llamas (*Lama glama*), followed by alpaca (*Vicugna pacos*) and camel (*Camelus bactrianus*, *Camelus dromedarius*) species (**Figure 3.8B**). For VNARs, entries are split between various shark species, but primarily derive from nurse sharks (*Ginglymostoma cirratum*) and the spotted wobbegong (*Orectolobus maculatus*) (**Figure 3.8C**). For both VHH and VNARs, approximately 10% of entries come from synthetic constructs.

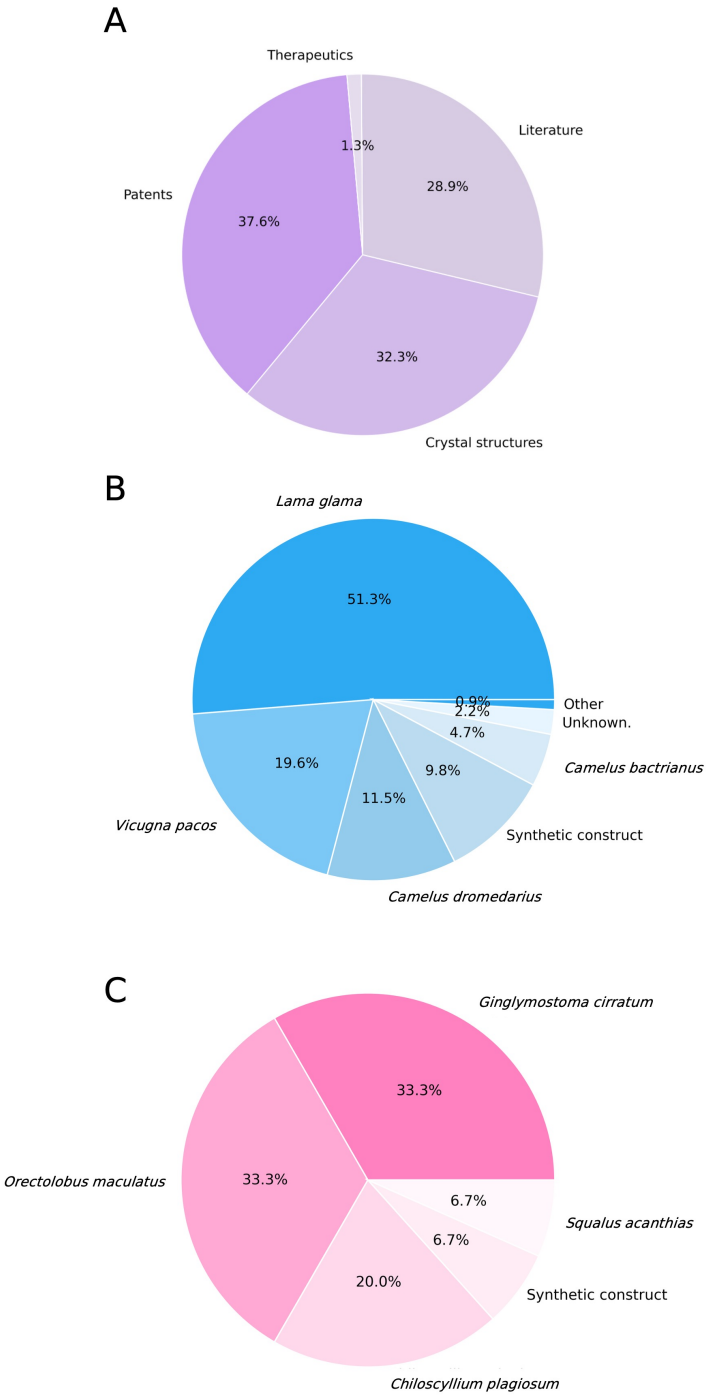


Figure 3.8: *A breakdown of entries in PLAbDab-nano from different sources and species. (**A**) Entries in PLAbDab-nano are sourced relatively evenly from the different sources: patents, literature and crystal structures, with a minority from therapeutic data. (**B**) Excluding patents due to a lack of species annotation, most VHH entries are sourced from camelid species and synthetic constructs. The total number of VHH sequences is 4452. (**C**) For VNARs, entries are split between various shark species and synthetic constructs. The total number of VNAR sequences is 456.*

The distribution of CDR3 loop lengths from VHH entries in PLAbDab-nano reflects those found in natural immune repertoires (**Figure 3.9A**), contrasting with the trend observed for natural versus engineered conventional antibodies, where the CDR-H3 in engineered antibodies tends to be shorter [152]. The CDR3 loops of VNARs tend to be longer than VHH CDR3 loops, based on our annotation (**Figure 3.9B**).

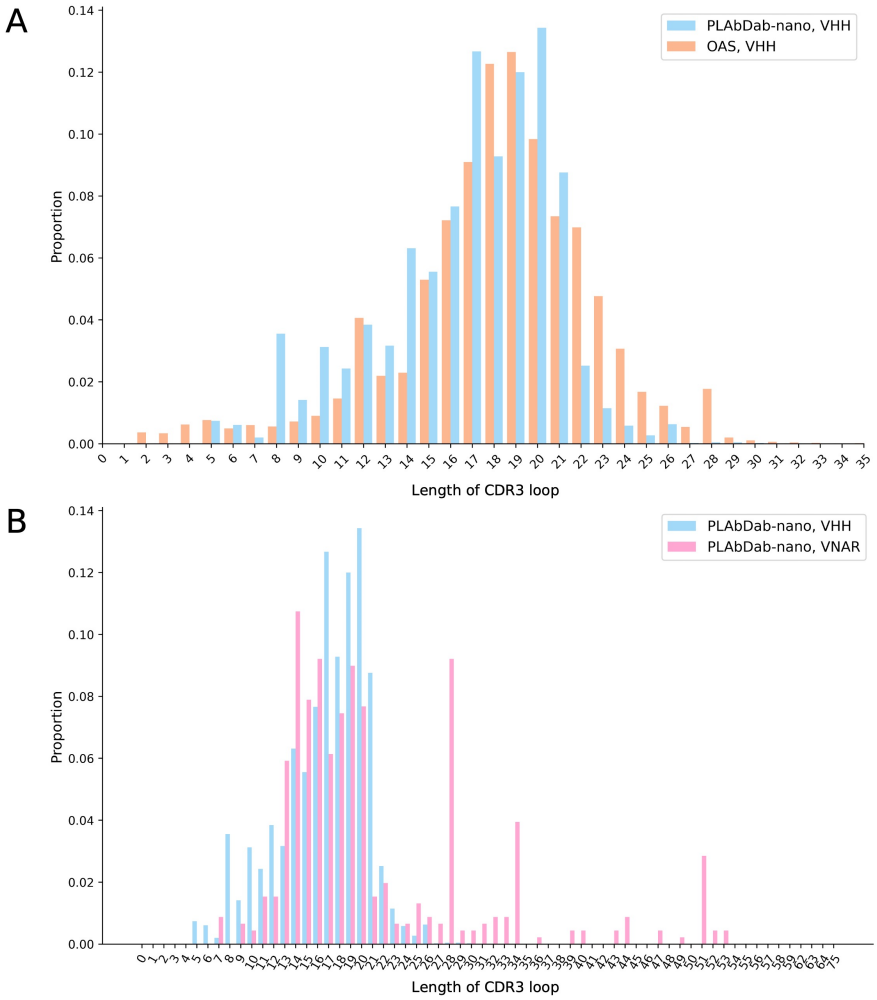


Figure 3.9: *A comparison of CDR3 length distributions of VHH and VNAR sequences in PLAbDab-nano. (**A**) The distribution of CDR3 loop lengths (by number of residues) from VHH entries in PLAbDab-nano closely matches that from natural VHH immune repertoire data from OAS [146, 147] (**B**) For entries included in PLAbDab-nano, the CDR3 loops of VNARs tend to be longer than VHH CDR3 loops.*

3.4.2 VNAR data in PLAbDab-nano provides CDR3 diversity

Previous studies have established that VHHs have longer CDR3 loops than those occurring in the VH chains of conventional antibodies [43–47, 246]. We find that according to our annotation, VNARs tend to exhibit even longer CDR3 loops than VHHs, further increasing the possible variation in the CDR3 which plays a significant role in binding. Here, we investigated the diversity of the CDR3 loops recorded in the PLAbDab-nano database.

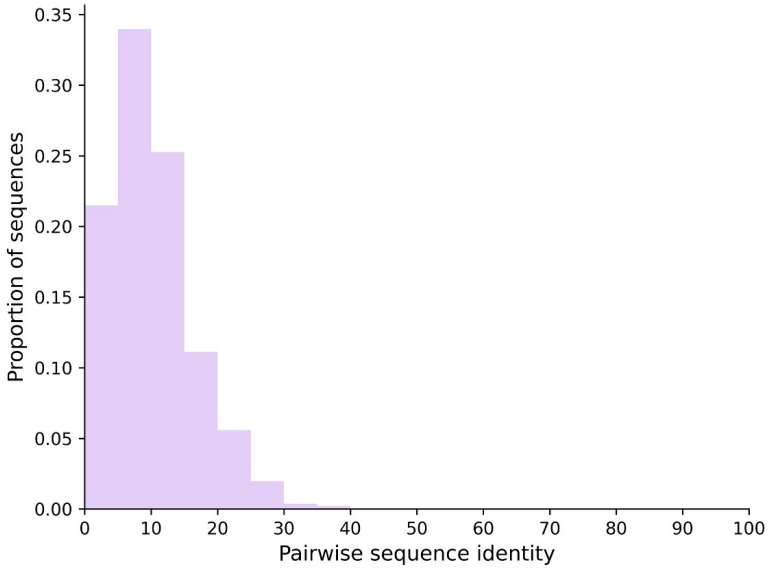


Figure 3.10: *Low pairwise percentage sequence identities for VHH CDR3 loops versus VNAR CDR3 loops indicate that the two types of nanobody occupy different regions of sequence space.*

A pairwise comparison of our VHH and VNAR CDR3 loops, including only those that matched by length (35,314 out of 537,915 possible combinations of non-redundant VHH/VNAR pairings) demonstrated low sequence identity between these two types of nanobody (**Figure 3.10**). This would indicate that the majority of VNAR CDR3 loops occupy a distinct sequence space compared to VHHs, and thus nanobodies from sharks provide additional diversity that could be exploited for therapeutic use.

We further assessed the CDR3 diversity from a structural perspective. Using our keyword search and querying for 'SARS-CoV-2' in the PLAbDab-nano database, we found 278 entries, 2 of which were VNARs and the remaining 276 VHHs, 1 of which was a therapeutic entry (Rimteravimab). Of the structures available, we found two pairs of VHHs and VNARs that bind at overlapping epitopes on the receptor-binding domain spike protein with very different CDR3 loop sequences (**Figure 3.11**). These structures all show the CDR3 loop as dominant in binding their target, but with differing orientations of the scaffold between the two epitopes. This illustrates that increased sequence diversity can aid the design of therapeutics targeting multiple epitopes, reinforcing ongoing interest in using the modularity of nanobodies for novel multi-specific antibodies [74, 75].

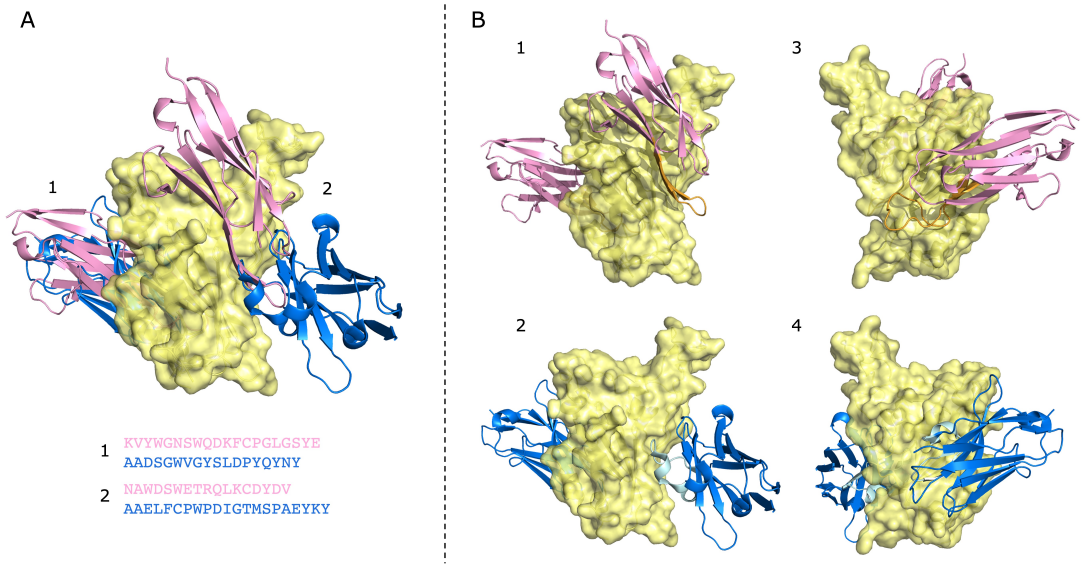


Figure 3.11: *(**A**) VNARs (pink) and VHHs (blue) binding to overlapping epitopes on the receptor-binding domain (RBD) of SARS-CoV-2 (yellow). The CDR3 sequences of the two VHHs and VNARs for the two different binding sites are labelled 1 and 2. (**B**) These nanobodies bind to their epitopes in differing orientations, with the CDR3 (shown in orange for the VNAR, and light blue for the VHH) dominant in binding in all cases. Structures 1 and 2 are rotated 180° across the y-axis to produce structures 3 and 4. Structures come from PDB entries 7S83 [74] for the VNARs and 8HR2 [240] for the VHHs.*

3.5 Discussion

In this chapter we described the curation of PLAbDab-nano, a database of nanobody sequences from patents and literature, including both VHHs and VNARs. PLAbDab-nano is the first database, to our knowledge, to collate VNAR data as well as VHH data into one central, nanobody-specific repository.

PLAbDab-nano can be automatically updated as new publications and patents arise to maintain an up-to-date record of nanobody data. We demonstrate the utility of PLAbDab-nano in providing a record of CDR diversity amongst nanobodies and expect that this could be most useful in designing nanobody libraries. Moreover, collating VNAR data in particular will drive forward our understanding of this nanobody variety and their unique characteristics.

The development of PLAbDab-nano involved the design of a numbering tool for VNARs to enable the annotation of their CDR loops, as no such tool existed at the time of publication. This work contributed to the development of ANARCII, a language model and numbering tool for the annotation of antigen receptors, including rare sequence types such as VNARs [247]. Although the strategy described in this paper draws out CDR loops that look reasonable for the structures available, there will be exceptions given the diversity of these regions, particularly for the CDR3 loop.

In addition, as was the case for PLAbDab [152], the resulting database generated by PLAbDab-nano relies on data being publicly available and in a format amenable to web-scraping. As such, since not all sequences are submitted to repositories such as the NCBI or PDB, the volume of data able to be collated is limited where sequences are deposited within the text of papers, or as images.

Given the high sequence similarity between VHHs and VHs from conventional antibodies, future work could be to further develop our classification methods to

better identify humanised VHHs, or else single-domain antibodies engineered from VH chains, so that they may be included in the database.

Overall, PLAbDab-nano addresses the issue of limited data availability by providing a growing dataset of nearly 5000 annotated nanobody sequences. This resource is especially valuable for VNAR studies, where large datasets for computational studies are scarce. PLAbDab-nano can be used to generate diverse libraries with properties based on patented and therapeutic nanobodies, to facilitate a greater depth of knowledge into nanobody properties, and to support the advancement of computational methods for nanobody design. Ultimately, this will help advance the development of novel nanobody therapeutics. In the next chapter, we use the data from PLAbDab-nano to explore the biophysical properties of nanobodies and develop a tool to evaluate nanobody developability relative to clinical-stage candidates.

4

The Therapeutic Nanobody Profiler: characterising and predicting nanobody developability to improve therapeutic design

Contents

4.1 Abstract 88
4.2 Introduction 89
4.3 Methods 91
4.4 Results 101
4.5 Discussion 117

4.1 Abstract

As we described in **Chapter 1**, developability optimisation is a critical aspect of therapeutic design. For monoclonal antibodies, developability is relatively well characterised. However, progress for novel biotherapeutics, such as nanobodies, is more limited. Differences in structural features between antibodies and nanobodies render current antibody computational methods unsuitable for direct application to nanobodies. Progress for the characterisation and prediction of nanobody developability has thus far been limited by poorer data availability. In **Chapter 3**, we collated nanobody data from a diverse range of sources into a database, PLAbDab-nano. In this chapter, we present the Therapeutic Nanobody Profiler (TNP), an open-source computational tool for predicting nanobody developability following the principles of the Therapeutic Antibody Profiler (TAP). Tailored specifically for nanobodies, it accounts for their unique properties compared to conventional antibodies for more efficient development of this novel therapeutic format. We calibrate TNP metrics using the 36 currently available clinical-stage nanobody sequences. We also collected experimental developability data for 108 nanobodies and examine how these results are related to the TNP guidelines.

The text in this chapter is based on the following preprint:

> **Gordon, G.L.**, Gervasio, J., Souders, C., and Deane, C.M. The Therapeutic Nanobody Profiler: characterising and predicting nanobody developability to improve therapeutic design. *bioRxiv* (2025).

References to nanobodies in this chapter should be interpreted as to the camelid (VHH) type only, unless specified otherwise.

I carried out the curation of the sequence and structure datasets and the analysis of this data. João Gervasio developed HeavyBuilder2, the modelling tool used to predict structures for antibody VH chains in the absence of a light chain partner. The TNP codebase is based on the original software for TAP, built by Matthew Raybould. I created the web application with assistance from Ben

Williams and Alex Greenshields-Watson. The *in vitro* data were generated from assays carried out by Twist Bioscience.

4.2 Introduction

The design of a successful biologic drug requires the optimisation of properties beyond the ability to bind to a target. Candidates must also demonstrate favourable developability: their capacity to undergo manufacture, storage and transport whilst retaining their function [62, 87].

Experimental methods are well-established for assessing developability in antibodies, which are the largest class of biologic drugs [248]. Techniques such as size exclusion chromatography (SEC), cross-interaction chromatography (CIC) or hydrophobic interaction chromatography (HIC) are used to assess the extent of aggregation or aggregation propensity, which may arise from non-specific binding or hydrophobicity [87, 93, 95]. Susceptibility to non-specific binding can be further tested via baculovirus (BVP) ELISA [87]. Light-based methods such as affinity-capture self-interaction nanoparticle spectroscopy (AC-SINS) and dynamic light scattering (DLS) serve to identify self-association and quantify polydispersity, which can lead to challenges in formulation [98, 99]. Melting and aggregation temperatures (Tm and Tagg) indicative of thermostability can be measured using differential scanning fluorimetry (DSF) [96].

Reliable *in silico* approaches have the potential to accelerate the time- and cost-intensive bench work involved in antibody development [249, 250]. **Chapter 1** provided an overview of the computational tools for antibody developability prediction that have emerged in recent years. These leverage sequence and structural data, often alongside *in vitro* assays [94, 179, 180, 211, 251–253]. This includes machine learning models, which have increasingly been applied to antibody developability prediction [254–256] and multi-parameter optimisation [215, 257–259].

Whilst the simultaneous optimisation of multiple developability properties remains challenging [215], several tools assess these properties in parallel. For example, the Therapeutic Antibody Profiler (TAP) scores developability based on five surface and structural features which are benchmarked against clinically-tested therapeutics [62]. These metrics include descriptors for surface hydrophobicity and charge, and characterisation of the CDR loops. Building on this, Ahmed *et al.* (2021) determined a physicochemical profile for the variable regions of marketed biotherapeutics and compared this profile with the characteristics of candidate antibodies [214]. Similarly, Park and Izadi (2024) have proposed six descriptors for electrostatic surface properties and hydrophobicity for antibody developability prediction, examining how these correlate with experimentally-determined properties [213]. Most recently, Bashour *et al.* (2024) extended this approach, computing a total of 86 sequence- and structure-based descriptors in order to map developability 'landscapes' for antibodies [260].

Existing studies on developability have focused primarily on monoclonal antibodies. As a result, efforts towards *in silico* prediction for nanobodies remain sparse, with the few examples centred around library design or optimisation of a single property [221, 222, 261]. Computational tools designed for conventional antibodies are usually not directly applicable to nanobodies due to their distinct structural features [250]. For example, nanobodies have no light chain, existing as a single domain. They show a tendency towards longer CDR3 loops, which allows for a broader range of possible conformations [48, 49]. In addition, the exposed FR2 region contains conserved residues which enhance solubility and stability, a feature not observed in conventional antibodies since this region is typically buried by the VL domain [48]. As we found in **Chapter 2**, these characteristics affect not only antigen binding but also the biophysical properties of the nanobody, and thus their developability.

In this chapter, we characterise the structural properties of nanobodies in the context of developability, conducting a comparative analysis with clinical-stage

nanobodies and experimental developability data. Our results reveal structural diversity among clinical-stage nanobodies, suggesting that developability is not constrained to a narrow subset of their conformational space. In order to further explore nanobody developability, we experimentally tested 108 clinical stage and proprietary nanobodies for properties such as aggregation, polyspecificity, self-association and thermostability. We assess the relationship between *in silico* and *in vitro* developability assessments.

Based on this analysis, we have built the Therapeutic Nanobody Profiler (TNP): a developability predictor tailored to nanobodies that follows the framework of the Therapeutic Antibody Profiler (TAP) [62]. TNP enables the comparison of predicted nanobody structures against 36 clinical-stage nanobody domains to inform the design of developable nanobodies. TNP will be updated as clinical-stage data grows over time, enabling increasingly robust evaluations of developability. Both a web application[1] and Python package[2] are publicly available to support users across computational and experimental disciplines.

4.3 Methods

4.3.1 Sequence data

Non-redundant sequence datasets were obtained from TheraSAbDab [153], Observed Antibody Space (OAS) [146, 147], PLAbDab-nano (described in **Chapter 3**) [262] and SAbDab [141, 142] to represent nanobody diversity across clinical-stage data, natural immune repertoires, patents, publications, and crystal structures (**Table 4.1**).

[1] *https://opig.stats.ox.ac.uk/webapps/tnp*
[2] *https://github.com/oxpig/TNP*

Dataset	Total
Clinical-stage Nbs	36
Clinical-stage antibody VH	919
Natural VHH	4059
Natural antibody VH	2047
Nb patents	563
Nb literature	410
Nb crystal structures	738
Nb proprietary	72

Table 4.1: *Total sequences and modelled structures in each dataset.*

Clinical-stage data was obtained from TheraSAbDab, including sequences at any trial stage and including bi- and multi-specifics, for which each unique sequence is counted separately. Natural sequences for VH and VHH respectively were obtained from studies by Jaffe *et al.* (2022) [263] and Li *et al.* (2016) [52] from OAS. To support structure prediction with a feasible dataset size, a single file was selected from these datasets and smaller representative sets were sampled from the non-redundant sequences. The data from PLAbDab-nano were split into three datasets, dividing sequences from patents, published literature and crystal structures (which are collected from SAbDab [141, 142]). The final dataset used in this work is comprised of 72 proprietary VHH sequences from Twist Bioscience.

4.3.2 Structural data

Predicted structures of the variable heavy domain were generated for all sequence datasets using NanoBodyBuilder2 [136] for nanobody data, and HeavyBuilder2, a modified version of ABodyBuilder2 (ABB2) [136], for antibody VH sequences.

HeavyBuilder2 uses the same architecture as ABB2 and a similar dataset of crystal structures for training curated from SAbDab [141, 142]. While the ABB2 training data ensures that structures had both VH and VL chains, for HeavyBuilder2 the VL was removed from the dataset. The validation set was the VH chain of 49 antibodies from the Rosetta Antibody Benchmark. The test set was the same as that used to benchmark ABB2: all Fv files added to SAbDab between the 1st of August

2021 and the 1st of June 2022. Three deep-learning models were independently trained for the prediction of antibody VH chains. The best prediction was selected by aligning predicted structures and choosing the one with the closest CDR-H3 conformation to the average of the three predictions. By default, HeavyBuilder2 passes the prediction through a structural refinement step.

For pairwise comparison of the CDR3 conformations of known crystal structures, 76 pairs of bound and unbound nanobodies were curated from SAbDab [141, 142]. A list of the PDB IDs for these pairs is given in (**Table C.1**).

4.3.3 Numbering schemes

The IMGT numbering scheme and CDR definitions were used throughout this work (CDR1: IMGT residues 27-38, CDR2: IMGT residues 56-65, CDR3: IMGT residues 105-117) [51]. The nanobody 'tetrad', or hallmark residues, refers to residues at positions 42, 49, 50 and 52 on the FR2 region. ANARCI [123] was used to number sequences for the annotation of CDR3 loops and tetrad residues.

4.3.4 Computed descriptors for developability properties

Patches of surface hydrophobicity and charge across the CDR vicinity were calculated following methods from TAP [62]. The CDR vicinity is defined using IMGT numbering and includes every surface-exposed CDR residue, anchor residue and any other residue with a heavy atom within a 4 Å radius. Surface-exposed residues are those with $\geq 7.5\%$ relative exposure across side-chain atoms [62, 264].

To calculate the score for patches of surface hydrophobicity (PSH), each amino acid is assigned a hydrophobicity score according to a given scale: by default, TNP uses the Kyte and Doolittle scale [265]. Salt bridges are assigned the same value as glycine. The PSH score quantifies the clustering of hydrophobic residues on the protein surface by summing the product of normalised (scaled between 1-2) hydrophobicity scores for all pairs of surface-exposed residues within 7.5 Å, divided

by the square of the distance between them. Charge is calculated similarly, but with the hydrophobicity score per residue substituted for the absolute charge value of the residue. The charges assigned are: aspartic acid, -1; glutamic acid, -1; lysine, +1; arginine, +1; and histidine, +0.1. Salt bridges are assigned a value of 0 [62].

A range of additional developability-related sequence and structural properties were calculated following methods from Bashour *et al.* (2024) [260], with a Python implementation of their associated code base which we make available at *github.com/oxpig/TNP*. A full list of the properties calculated, grouped by class according to Bashour *et al.* can be found in **Table C.2**.

4.3.5 Describing CDR3 loop conformations

CDR3 loop conformation was analysed according to a spherical coordinate system, following methods we set out in **Chapter 2** [246]. The CDR3 loop is described by its 'compactness', defined as the length of the CDR3 loop in number of residues divided by its reach away from the rest of the variable domain. A loop with lower compactness reaches away from the variable domain, whereas a loop of higher compactness tends to be folded down against the variable domain.

We compared this to the classification of CDR3 conformations as either 'kinked' or 'extended', following methods in Weitzner *et al.* (2015) and Bahrami-Dizicheh *et al.* (2023) [21, 38]. This classification is based on the τ (tau) pseudo-dihedral and α (alpha) angles between backbone Cα atoms in the CDR3 at positions 116, 117 and 118 (by IMGT numbering). Kinked loops have α and τ angles in the range: $0° < \alpha116 < 120°$, $85° < \tau116 < 130°$. Extended loops have α and τ angles in the range: $-100° > \alpha116$, $100° < \tau116 < 145°$ (**Figure 4.1**).

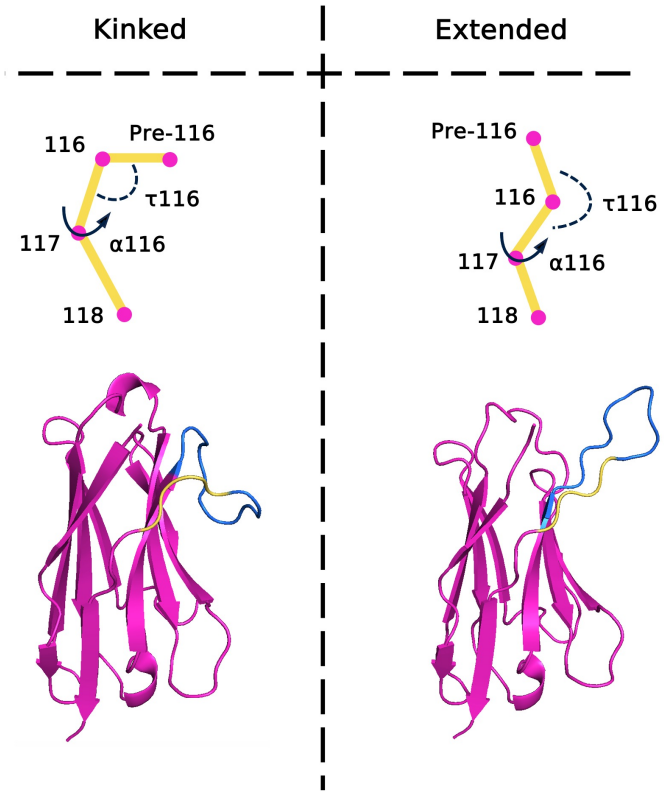


Figure 4.1: *CDR3 conformations can be classified as either 'kinked' or 'extended', following methods in Weitzner et al. (2015) and Bahrami-Dizicheh et al. (2023) [21, 38]. This classification is based on the τ pseudo-dihedral and α angles between backbone Cα atoms in the CDR3 at positions 116, 117 and 118 (by IMGT numbering). Kinked loops have α and τ angles in the range: $0° < \alpha116 < 120°$, $85° < \tau116 < 130°$. Extended loops have α and τ angles in the range: $-100° > \alpha116$, $100° < \tau116 < 145°$. Figure is adapted from Bahrami-Dizicheh et al. [38].*

4.3.6 Interactions between the CDR3 loop and nanobody tetrad residues

Arpeggio (v1.4.1) [227] was used to analyse interatomic interactions between pairs of residues from the CDR3 loop and nanobody tetrad. Structures were hydrogenated for this analysis using gemmi v0.6.7 [266], following methods in Bashour *et al.* (2024) [260].

4.3.7 *In vitro* developability assays

In vitro assays for developability were carried out on 108 nanobody sequences: 36 derived from clinical-stage data, with the remaining 72 from proprietary data from Twist Bioscience.

4.3.7.1 Analytical Size Exclusion Chromatography (aSEC)

SEC was used to evaluate the molecular size distribution and structural integrity of protein samples, including detection of monomer content, aggregation, and degradation. Approximately 5 µg of each protein sample was injected onto an Advance Bio SEC column (300 Å, 4.6 × 150 mm, 2.7 µm particle size; Cat.# PL1580-3301, Agilent Technologies) pre-equilibrated with phosphate-buffered saline (PBS) (Cat.#BP399-20; Fisher Scientific).

The separation was carried out on an Agilent 1260 Infinity II HPLC system using PBS as the mobile phase at a flow rate of 0.35 mL/min. Elution was monitored by UV absorbance at 280 nm. Since the separation is based on hydrodynamic volume, larger protein species elute earlier, while smaller species elute later.

Chromatograms were analysed using Agilent Chem Station Open Lab CDS software. Peak retention times and areas were used to assess the proportion of monomeric protein, as well as the presence of high molecular weight aggregates or low molecular weight degradation products.

4.3.7.2 Cross Interaction Chromatography (CIC)

The CIC assay was employed to assess the propensity of a test protein to bind non-specifically to immobilised proteins (commonly human IgG) which serves as an indicator of aggregation risk. In this HPLC-based method, as outlined previously [87, 93] the test proteins are passed through a column containing immobilised polyclonal human IgG, and their retention times are recorded.

To construct the CIC column, approximately 30 mg of polyclonal human antibodies (I4506; Sigma) are covalently bound to a 1-mL NHS-activated HiTrap column (17-0716-01; Cytiva), followed by quenching with ethanolamine (Cat # 104679; Carterra). For each run, around 5 µg of antibody is injected, and phosphate-buffered saline (PBS) is used as the mobile phase, flowing at 0.1 mL/min on an Agilent 1260

Infinity II HPLC system. Chromatographic data are acquired and analysed using Agilent Chem Station Open Lab CDS software.

Proteins that interact more strongly with the column will exhibit longer retention times. A high degree of interaction may suggest a risk of aggregation, making such proteins less suitable for therapeutic development.

4.3.7.3 Hydrophobic Interaction Chromatography (HIC)

The HIC assay was performed to evaluate the relative hydrophobicity of antibody samples, which can be indicative of developability risks such as aggregation and nonspecific interactions [87, 267]. The assay was conducted using a Proteomix HTC Butyl-NP5 column, 4.6 × 100 mm (Cat. #431NP5-4610; Sepax Technologies).

Antibody samples (5 µg at 1 mg/mL) were diluted with mobile phase A (1.8 M ammonium sulfate (Cat. # J64419.A3; Thermo Scientific) and 0.1 M sodium phosphate (Cat. #S373-500; Fisher Scientific, Cat. #S468-500; Fisher Scientific), pH 6.5) to achieve a final ammonium sulfate concentration of 1 M prior to injection. Chromatographic separation was performed using a linear gradient elution over 20 minutes, transitioning from mobile phase A to mobile phase B (0.1 M sodium phosphate, pH 6.5) at a flow rate of 1.0 mL/min. UV absorbance was monitored at 280 nm using an Agilent 1260 Infinity II HPLC system. Retention times were determined using Agilent Chem Station Open Lab CDS software.

Proteins with greater hydrophobicity exhibited longer retention times due to extended interaction with the hydrophobic stationary phase.

4.3.7.4 Dynamic Light Scattering (DLS)

Protein particle size distribution and solution homogeneity were characterised using dynamic light scattering (DLS). This approach offers high sensitivity to early aggregate formation, allowing detection of subtle variations in conformational

behavior and sample composition.

Each protein sample was loaded into standard capillaries (Cat.# PR-AC002; Nano Temper Technologies) with an approximate volume of 10 µL and analysed using the Prometheus Panta instrument. Sample-specific details including identity, buffer conditions, and concentration were entered into the software platform (PR Panta Analysis, Version 1.9) before measurements were initiated. DLS measurements yielded two principal values: the hydrodynamic radius (rH) and the polydispersity index (PDI). The rH reflects the average diffusion-based size of particles in solution, indicative of the protein's physical state.

The PDI is an indicator of size uniformity within the sample population; lower values suggest minimal heterogeneity and a well-behaved molecular species, while elevated values point to aggregation, misfolded structures, or the presence of particulate impurities.

4.3.7.5 Nano Differential Scanning Fluorimetry (nDSF)

Samples were introduced into the Prometheus Panta system (Nano Temper Technologies) using standard capillaries (Cat. # PR-AC002; Nano Temper Technologies) with a loading volume of approximately 10 µL. Prior to measurement and data acquisition, sample metadata including identification, concentration, and buffer composition were recorded using the system's control software (PR Panta Analysis, Version 1.9).

An initial discovery scan was conducted to optimise excitation parameters by determining the appropriate light intensity tailored to the optical characteristics of each sample. Following this calibration step, thermal unfolding experiments were performed by gradually increasing the temperature from 15°C to 95°C at a constant ramp rate of 1°C per minute. Fluorescence signals were captured simultaneously at 330 nm and 350 nm throughout the scan to monitor temperature-dependent

changes in protein conformation.

Protein thermal behavior was evaluated using nano differential scanning fluorimetry (nDSF), which detects alterations in intrinsic fluorescence signals as the sample undergoes a controlled temperature increase. Structural changes during heating lead to changes in the environment of aromatic residues, primarily tryptophans, causing a measurable shift in their emission profile.

Fluorescence was monitored at two wavelengths – 330 nm and 350 nm – where folded proteins preferentially emit at the shorter wavelength due to shielding of hydrophobic residues. As unfolding progresses and these residues become solvent-exposed, the emission shifts toward the longer wavelength. In rare instances where tryptophans are absent, tyrosine fluorescence contributed to the signal.

The ratio of fluorescence intensity at 350 nm to 330 nm (F350/F330) was plotted against temperature. Transitions in structural conformation were identified as inflection points in the curve, with the first deviation from baseline indicating the onset of unfolding (Tonset). The point of maximum slope, derived from the first derivative of the fluorescence ratio curve, was used to determine the thermal transition midpoint (Tm). These data were used to compare and prioritise protein variants based on their thermal stability characteristics.

4.3.7.6 Affinity-capture Self-interaction Nanoparticle Spectroscopy (AC-SINS)

To evaluate the self-interaction potential of protein candidates, the AC-SINS assay was carried out following previously established protocols [87]. This method enables assessment of solution-phase self-association by exploiting changes in optical properties of functionalised gold nanoparticles upon protein binding.

Gold nanoparticles (Cat. #15705, Ted Pella Inc.) were prepared by coating with a mixture of antibodies: 80% anti-human IgG Fc-specific goat antibody (Cat.

#109-005-098, Jackson ImmunoResearch) and 20% non-specific goat polyclonal antibody (Cat. #005-000-003, Jackson ImmunoResearch). These coated particles serve as a capture surface for test proteins via Fc-mediated interactions. Protein samples were incubated with the functionalised nanoparticles for 2 hours at room temperature. Following incubation, spectral absorbance was recorded using an Agilent Biotech Epoch Microplate Spectrophotometer reader, and the data were analysed. A shift in the peak absorbance wavelength relative to the mouse IgG or PBS control was interpreted as a signature of particle aggregation caused by antibody self-association. Larger wavelength shifts reflect stronger self-attractive interactions, indicating proteins that may be prone to high-concentration aggregation and therefore less favourable for formulation and development.

4.3.7.7 BVP ELISA

The assay procedure for BVP-ELISA was adapted from the protocol described by Jain *et al.* (2017) [87]. In summary, baculovirus particles (Cat.# E3001;Medina Scientific) were prepared by mixing viral stock with 50 mM sodium carbonate buffer (pH 9.6) in 1:200 dilution and applied to 384 well ELISA plates (Cat.#464718;Thermo Scientific). Plates were incubated overnight at 4°C to allow coating. The following day, unbound particles were removed by aspiration, and all subsequent steps were performed at room temperature. Wells were blocked with 75 µL of PBS containing 2% BSA for 1 hour, followed by three washes with PBS (75 µL per well). Test antibodies, diluted to 100 µg/mL in blocking buffer (25 µL), were then added to each well and incubated for 1 hour. Plates were washed six times with PBS before the addition of 25 µL of HRP-conjugated Goat Anti Human IgG, Fcy sp. secondary antibody (Cat.# 115-035-098; Jackson Immuno Research), followed by another 1-hour incubation. After a second series of six washes with PBS, 25 µL of TMB substrate (Cat.# 5120-0083; Sera Care) was added, and the reaction allowed to proceed for 10–15 minutes. The enzymatic reaction was stopped with 25 µL of 2 M sulfuric acid, and absorbance was measured at 450 nm using an Agilent

Biotech Epoch Microplate Spectrophotometer reader. To calculate the BVP score, absorbance values were normalised to wells containing no test antibody.

4.4 Results

We first examine the differences in features between antibody VH chains and nanobodies that are likely to impact developability, initially investigating the TAP metrics that can be directly applied to nanobodies (total CDR length and patches of surface hydrophobicity, positive charge, and negative charge across the CDR vicinity) and then more nanobody-specific features, including CDR3 loop conformation and the nanobody tetrad residues. This leads to a list of six metrics for use in the TNP. We next calibrate these metrics using the 36 currently available clinical-stage nanobody sequences, and a set of 72 nanobodies for which we collected *in vitro* developability data, including assessments for non-specific binding (using CIC), polyspecificity (BVP-ELISA), hydrophobicity (HIC), self-association (AC-SINS) and aggregation (DLS).

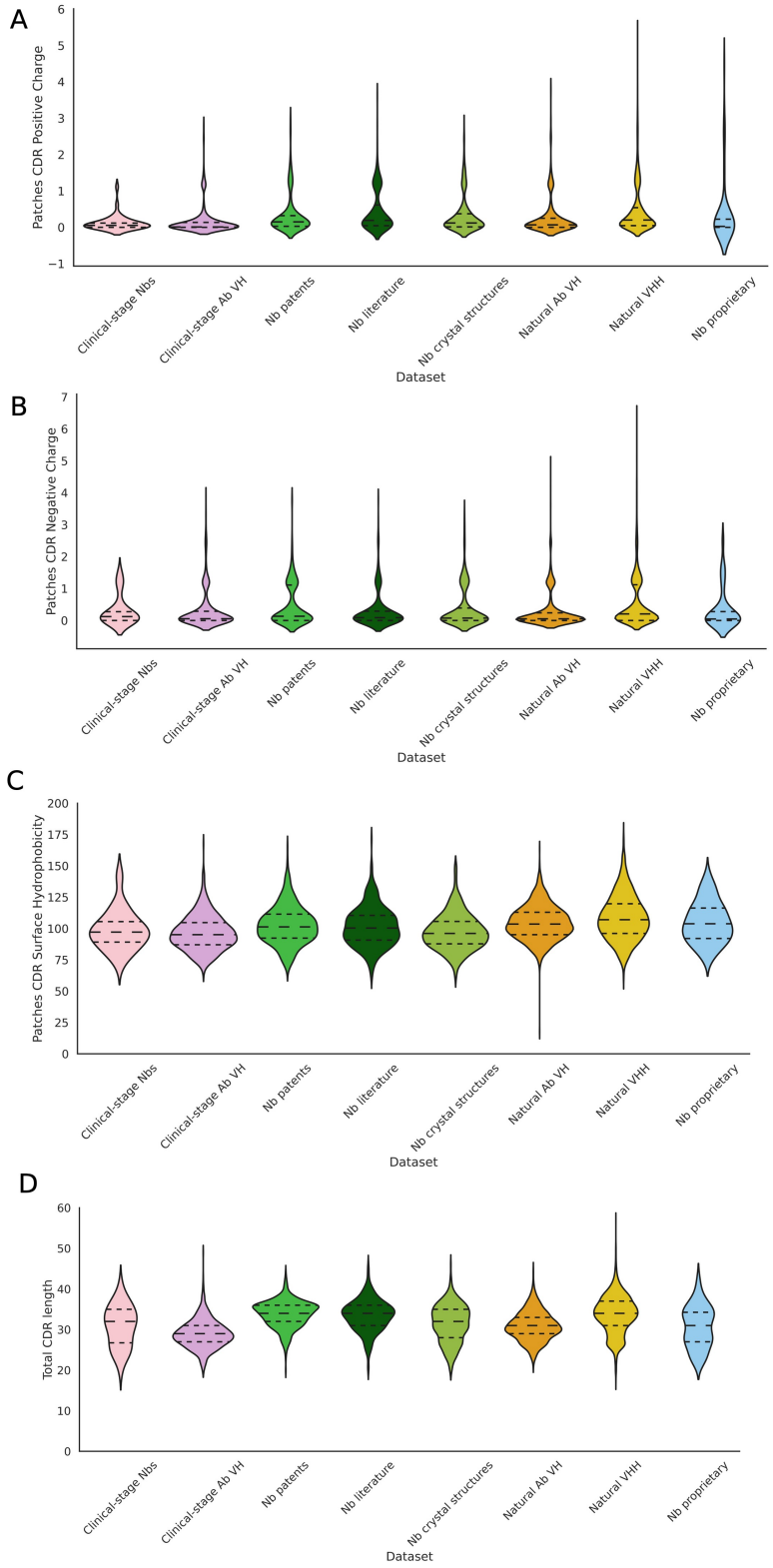


Figure 4.2: *Distributions of the TAP metrics on our 8 datasets of nanobodies and antibody VH chains from different sources. The TAP metrics score candidates on (**A**) patches of positive charge, (**B**) patches of negative charge, (**C**) patches of surface hydrophobicity and (**D**) total CDR length. The clinical-stage nanobody data are distinct from the antibody VH data and nanobodies from non-clinical sources.*

4.4.1 Distributions of the TAP metrics differ between antibody and nanobody clinical data

To assess whether clinical-stage nanobodies occupy distinct ranges across the applicable TAP metrics, the distributions of these metrics were computed across multiple datasets on sequences and predicted structures (**Figure 4.2**). We compared clinical-stage nanobodies to those from natural immune repertoire data, structures, publications and patents, as well as to antibody VH domains from both natural and clinical-stage sources.

Clinical-stage nanobodies have a more constrained range for patches of positive and negative charge, than any other dataset (**Figures 4.2A, 4.2B**; **Figures C.1, C.2**). For patches of surface hydrophobicity, the range for clinical-stage nanobodies largely overlaps with the other datasets (**Figure 4.2C**, **Figure C.3**). Total CDR length tends to be greater for all nanobody datasets than the antibody VH datasets, though the range for clinical-stage nanobodies is more limited than in the natural VHH data, and even the clinical-stage antibody VH data (**Figure 4.2D**, **Figure C.4**). The difference between the nanobody and antibody datasets is most likely due to the longer CDR3 in nanobodies.

4.4.2 Total CDR length does not sufficiently capture nanobody paratope shapes

In TAP, the total CDR length measure for antibodies is a strong indicator of both the length of the heavy chain CDR3 and the paratope shape. As nanobodies typically have a longer CDR3 than antibodies [43, 44, 54, 55, 246] and do not have a light chain, the diversity of CDR3 conformations is greater [44, 48, 49]. This means that the total CDR length measure may not fully capture CDR3 conformation in nanobodies. We investigate the relationship between CDR3 conformation (and therefore paratope shape) and loop length in nanobodies, in particular looking at whether CDR3 length or total CDR length are able to capture CDR3 compactness.

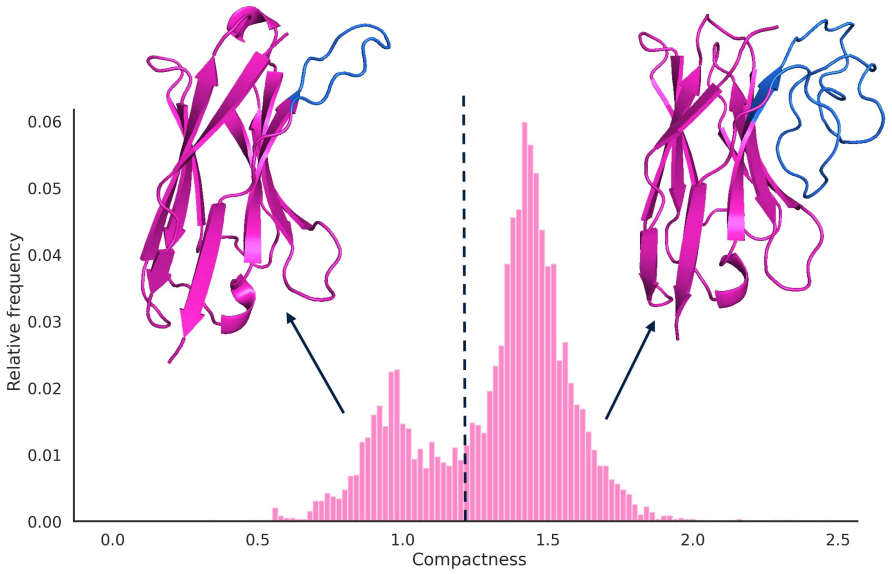


Figure 4.3: *Loop compactness provides a continuous descriptor for CDR3 conformation. The two peaks in the bimodal distribution of compactness for all of our nanobody datasets agree with previous classifications of loop conformations as either extended (lower compactness) or kinked (higher compactness).*

As we described in **Chapter 2**, compactness is a measure of loop length relative to loop reach (**Figure C.5**). We selected compactness as our descriptor for CDR3 conformation as it provides a continuous scale with which to describe kinked and extended CDR3 conformations (**Figure 4.3**, **Figure C.6**). Our continuous measure of compactness more precisely captures the diversity of loop conformations than previous methods [21, 38, 49], whilst remaining consistent with their overall classification system (**Figures C.7, C.8, C.9**).

Lower compactness corresponds to the CDR3 extending upward like an antibody CDR-H3 constrained by a VL chain, whereas higher compactness scores indicate that the loop folds down over the FR2 region. We have shown in **Chapter 2** that the CDR3 of nanobodies have a bimodal distribution of compactness, whilst antibody CDR-H3 loops do not [246], indicating that a compactness measure may be needed for a nanobody profiler.

In our datasets, clinical-stage nanobodies tend to show a relatively balanced

distribution of CDR3 length (in terms of number of residues), including shorter CDR3 loops not typically seen in natural, patented, or literature-derived nanobodies (**Figure 4.4A**, **Figure C.10**). The CDR3 length profiles of the clinical-stage, proprietary, and crystal structures sets of nanobodies align most closely with clinical-stage VH domains, suggesting the presence of more VH-like features.

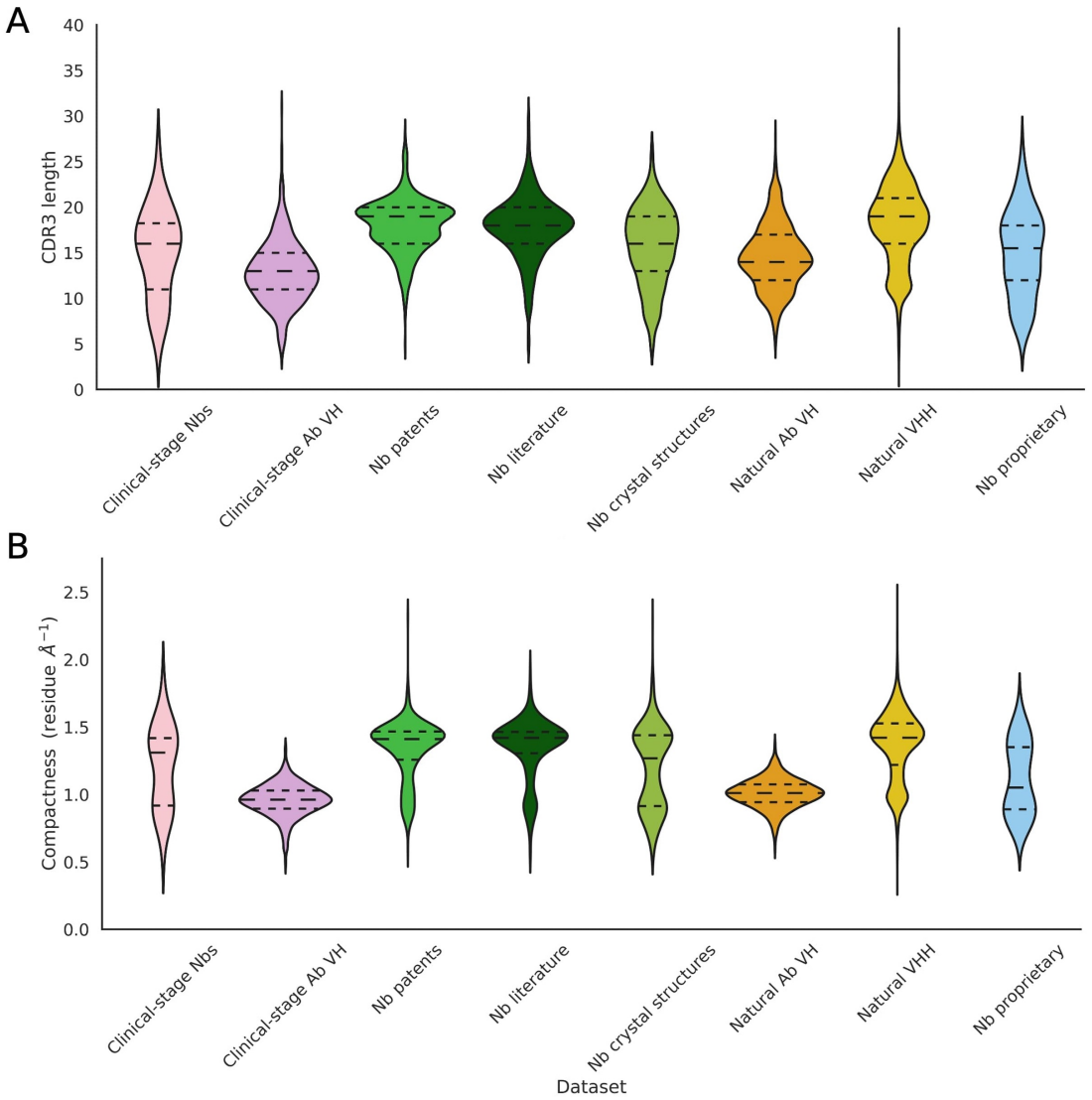


Figure 4.4: *Distributions of the CDR3 length and CDR3 compactness on our 8 datasets of nanobodies and antibody VH chains from different sources. (**A**) Clinical-stage nanobodies occupy a wide range of CDR3 lengths. (**B**) The CDR3 loop conformation can be described using compactness, a measure of the loop length relative to its reach. Nanobodies, including the clinical-stage data, show a bimodal distribution for compactness, representing two clusters of loop conformations.*

As expected, our nanobody datasets show a bimodal distribution of compactness, while the antibody VH data do not (**Figure 4.4B**, **Figure C.11**). Clinical-stage nanobodies appear to show the presence of two subpopulations of CDR3 conformations, which is not captured by total CDR length or CDR3 length alone, indicating that the inclusion of CDR3 compactness in the TNP metrics would help capture the nanobody paratope shape. These results also suggest that nanobody developability is not limited to either of the compactness subpopulations.

4.4.3 More compact CDR3 loops do not show significant conformational change upon binding

A large proportion of nanobodies with longer CDR3 loops show higher compactness scores (**Figure C.9**). To determine whether there is a conformational change happening upon binding to allow these compact loops to extend out, a set of 76 pairs of bound and unbound crystal structures were curated (see **Methods**, **Table C.1**) to allow a pairwise comparison of CDR3 compactness.

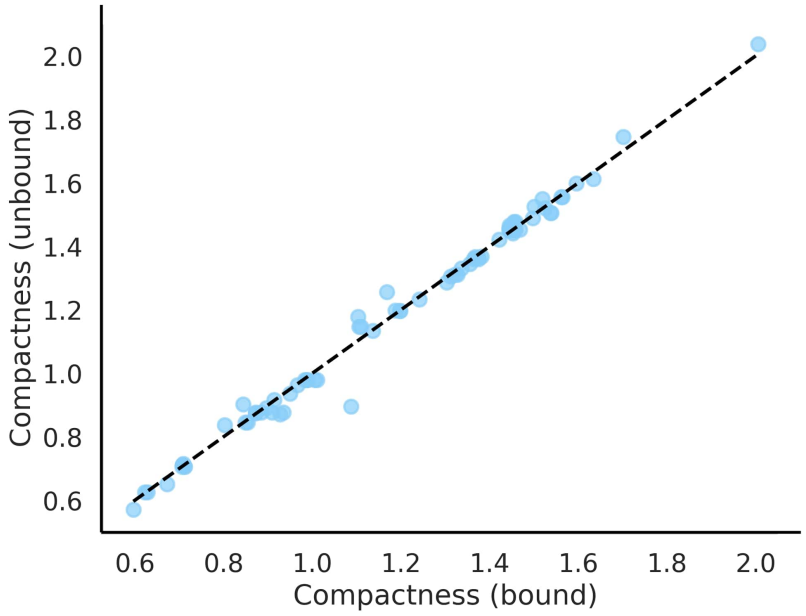


Figure 4.5: *Pairwise comparison of compactness scores for bound and unbound nanobody crystal structures suggests that the CDR3 loop does not undergo a significant conformational change upon binding.*

Compactness scores are consistent between bound and unbound nanobody structures (**Figure 4.5**). Any observed changes in compactness are minimal, and there is no correlation between the change in compactness and the root mean square deviation (RMSD) over the bound and unbound CDR3 loops (which indicates any difference between the crystal structures) (**Figure C.12**). This suggests that a conformational change, where a more compact loop unfolds and extends out from the FR2 region, is unusual.

4.4.4 Nanobody subtypes based on CDR3 loop conformations show associated tetrad motifs

Previous studies have described interactions between the CDR3 loop and nanobody tetrad residues present on the exposed FR2 surface [38, 50, 209, 268, 269]. These residues are critical to surface hydrophobicity and solubility, and thus affect developability [209].

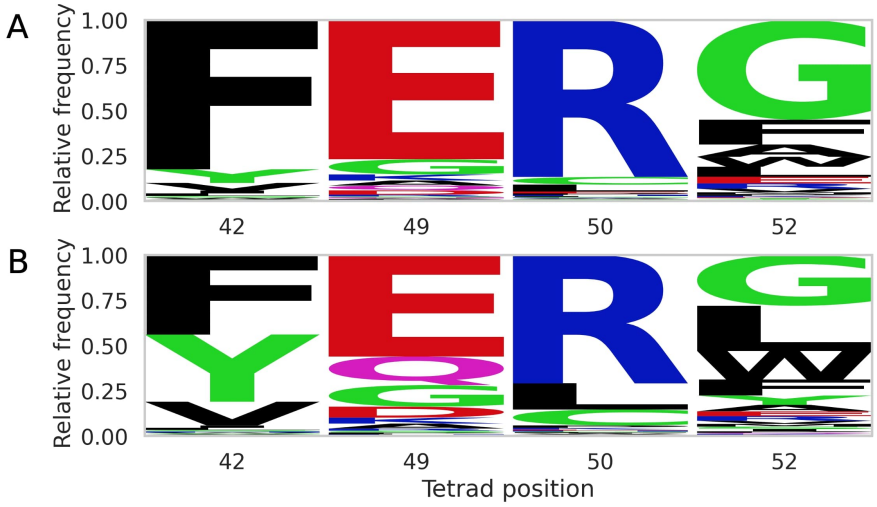


Figure 4.6: *Amino acid frequencies across the nanobody tetrad positions (42, 49, 50 and 52 by IMGT numbering) show greater variation per position in nanobody loops of lower compactness versus higher compactness. (**A**) Logo plot of relative amino acid frequencies in loops of higher compactness. (**B**) Logo plot of relative amino acid frequencies in loops of lower compactness.*

We observe that canonical nanobody tetrad motifs, such as FERG, are the predominant motif across nanobodies, while VGLW is most common in antibodies (**Figure C.13**), supporting the existing literature [50]. Among nanobodies, those with

higher compactness tend to retain these classically nanobody-like motifs. Loops with lower compactness display greater diversity in tetrad motifs (**Figure 4.6**). The tetrad motif is therefore a useful metric to include in the TNP, as it can help stratify nanobodies into distinct structural subtypes.

4.4.5 Hydrophobic interactions between aromatic residues often occur between nanobody tetrad residues and more compact CDR3 loops

Consistent with previous work, we find that hydrophobic, aromatic and carbon-π interactions are the most commonly occurring non-covalent interactions between pairs of CDR3 loop and nanobody tetrad residues (**Figure C.14**) [209]. For CDR3 loops of higher compactness, we found that the most common interacting amino acids are between aromatic residues such as phenylalanine, which is highly conserved at position 42 in the nanobody tetrad, and a second aromatic residue in the CDR3 loop (**Figure 4.7**, **Figure C.15**).

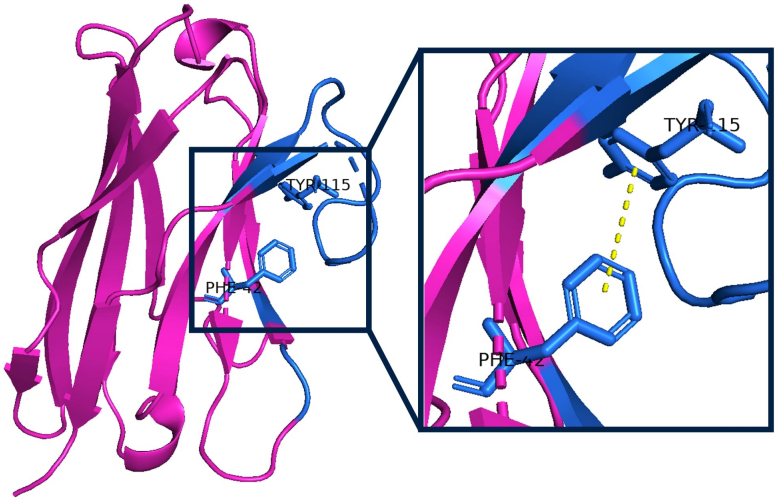


Figure 4.7: *Hydrophobic interactions between aromatic residues in the CDR3 loop and the first nanobody tetrad position (IMGT position 42), here between a tyrosine and phenylalanine, often occur in a more compact CDR3 loop which tends to be folded over the FR2 region. The framework region is shown in pink; the CDR loops and tetrad residues are shown in blue.*

The necessity of these interactions for structural stability in longer loops folded over the FR2 region makes the identities of these tetrad residues, alongside the CDR3

conformation, an important consideration for the design of developable molecules. This supports the inclusion of the tetrad motif and CDR3 compactness in the TNP, as the co-optimisation of these properties may lead to improved developability.

4.4.6 *In vitro* assays support the presence of two equally developable nanobody subtypes

Developability assays (including aSEC, CIC, HIC, BVP-ELISA, AC-SINS, DLS and DSF) were carried out on a set of 108 nanobody sequences, including 36 clinical-stage sequences, with the remainder being proprietary candidates from Twist Bioscience.

The data were divided into groups of lower and higher compactness (using a threshold score of 1.25 as per **Figure 4.3**), as a proxy to distinguish between the two subpopulations. Across all assay results, the range of values for nanobodies of lower and higher compactness are largely comparable (**Figure C.16**). Although some candidates within these ranges exhibit poorer scores and may be less developable, the primary conclusion is that there is no need to bias selection toward either nanobody subtype.

4.4.7 TNP properties

Based on the above analysis, we have selected six properties for use in TNP: total CDR length, CDR3 length, CDR3 compactness, and patches of surface hydrophobicity, positive charge and negative charge across the CDR vicinity. These are aimed at capturing the major sequence and structural features that will influence developability. In order to assess how well these features represent the data, we calculated a further 81 surface and structural properties for our nanobody datasets following the methods described in Bashour *et al.* (2024) [260]. We make this data publicly available at *github.com/oxpig/TNP*. These properties were correlated alongside the original TAP metrics and the additional descriptors introduced in TNP to assess their redundancy (**Figure C.17**). Overall, our results indicate that

our selected six metrics are a sensible small set of interpretable descriptors that avoid redundancy. However, there are many other choices that could be made.

4.4.8 Developability guidelines for TNP are set by 36 clinical-stage nanobody sequences and predicted structures

For our chosen metrics, clinical-stage nanobodies show distributions distinct from antibody clinical data or nanobody non-clinical data. Consequently, we defined thresholds for our developability guidelines based on the 36 clinical-stage nanobodies (**Table 4.2**). Following the TAP methodology, an amber flag is assigned where a nanobody scores within the lowest or highest 5% of a distribution, and a red flag is assigned where properties lie outside the current observed clinical range. These thresholds will be updated as new candidates progress through clinical trials.

Metric	Range	Amber flag region	Amber flag threshold	Red flag region	Red flag threshold
Total CDR length	20-39	Top 5%, bottom 5%	20-24, 37-39	Above or below	<20, >39
CDR3 length	5-23	Top 5%, bottom 5%	5-8, 21-23	Above or below	<5, >23
CDR3 compactness	0.56-1.61	Top 5%, bottom 5%	0.56-0.81, 1.57-1.61	Above or below	<0.56, >1.61
PSH	73.40-155.47	Top 5%, bottom 5%	73.40-79.59, 126.83-155.47	Above or below	<73.40, >155.47
PPC	0.00-1.18	Top 5%	0.39-1.18	Above	>1.18
PNC	0.00-1.88	Top 5%	1.47-1.88	Above	>1.88

Table 4.2: *Flagging thresholds for each of the TNP metrics are set according to the distributions of the 36 nanobody clinical-stage data-points.*

4.4.9 *In vitro* and *in silico* methods provide complementary approximations of developability

We performed *in vitro* assays (including aSEC, CIC, HIC, BVP-ELISA, AC-SINS, DLS and DSF – see **Methods**) for developability properties on a set of 108 nanobody sequences, comprising 36 clinical-stage sequences and 72 proprietary candidates from Twist Bioscience (for which one failed to express). We evaluated the proprietary sequences against our TNP metrics to determine their positioning relative to the clinical thresholds. These predictions were then compared against their *in vitro* assay data, in which five assays (CIC, HIC, BVP-ELISA, AC-SINS and DLS) were

scored using the same green–amber–red flagging system. Thresholds for these flags were derived based on the range of scores observed across all 108 VHHs in the panel.

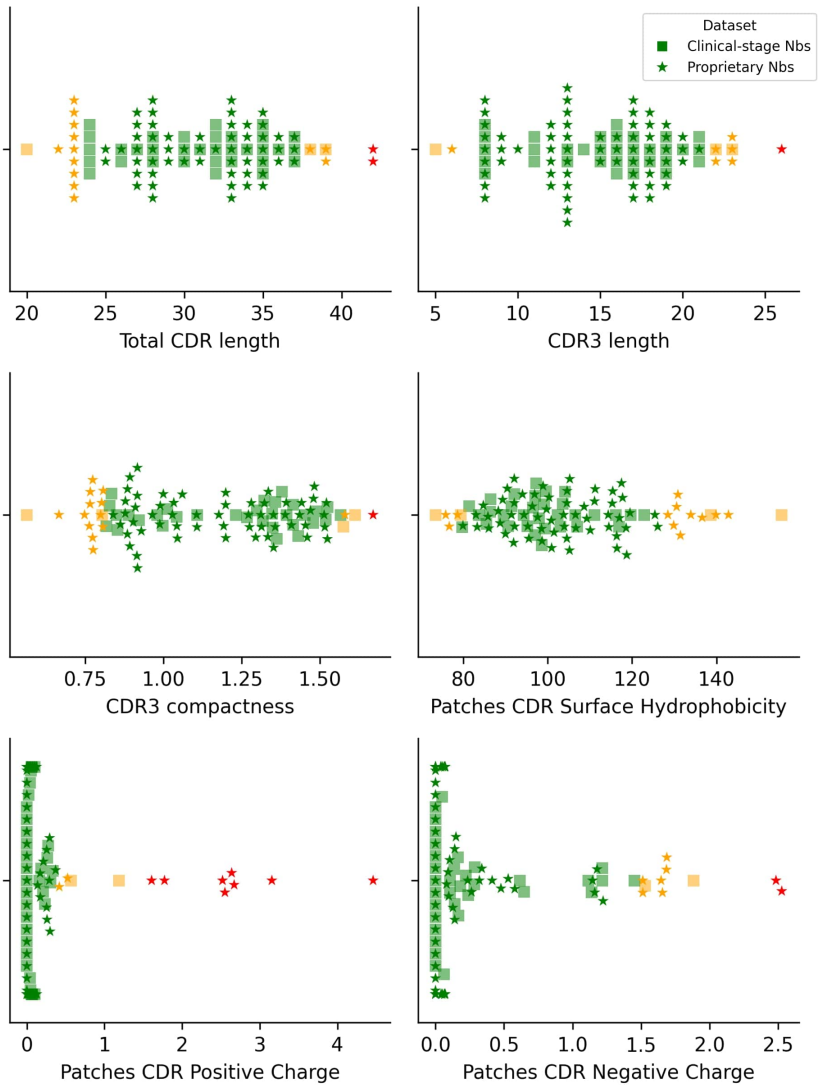


Figure 4.8: *Distributions of values for each TNP metric for nanobody clinical-stage and proprietary data, coloured by flags assigned according to the thresholds set by the clinical-stage data. Data points in red highlight proprietary candidates that fall outside of our current guidelines.*

The proprietary candidates predominantly fall within the range defined by the clinical-stage nanobodies (**Figure 4.8**). No red flags were observed for surface hydrophobicity patches, whilst they most frequently occurred for patches of positive charge. In total, 14 red flags were identified across 11 proprietary candidates (**Table 4.3**). Of these, only one candidate exhibited multiple red flags – the remaining ten

each had a single red flag.

For candidates flagged as high-risk by TNP, there is agreement with *in vitro* assay outcomes, where all those with red flags for at least one TNP metric also show at least one amber or red flag for the *in vitro* assays, with most showing multiple amber and red flags (**Table 4.3**).

Sample	Total CDR length	CDR3 length	CDR3 compactness	Patches CDR Surface Hydrophobicity	Patches CDR Positive Charge	Patches CDR Negative Charge	CIC Retention Time (min)	HIC Retention Time (min)	BVP-ELISA OD450 (a.u.)	AC-SINS Δλmax (nm)	DLS PDI
1	42	23	1.58	136.61	0.01	1.51	9.48	20.19	0.96	4.57	0.34
2	42	26	1.67	142.86	4.46	0.09	8.80	17.56	3.22	1.79	0.18
3	34	18	1.36	95.27	3.16	0.23	9.43	12.91	2.77	3.60	0.85
4	39	23	1.51	139.88	1.77	0.58	9.13	12.95	0.84	6.78	0.31
5	39	23	1.52	130.88	1.61	0.34	9.20	14.85	0.23	8.65	0.24
6	36	20	1.47	114.50	0.42	2.48	9.20	17.29	0.34	3.68	0.16
7	36	20	1.48	125.41	0.21	2.53	9.59	23.14	1.48	6.52	0.13
8	35	18	1.11	126.08	2.64	0.06	10.74	23.60	2.51	17.56	0.80
9	35	18	1.03	118.59	2.55	0.05	10.30	20.50	1.75	19.88	0.87
10	35	18	1.11	111.10	2.67	0.07	9.79	17.40	0.82	14.18	0.98
11	35	18	1.06	117.49	2.52	0.06	10.33	20.70	1.81	18.13	0.88

Table 4.3: *Comparison of proprietary candidates from Twist Bioscience that were assigned red flags by TNP to their in vitro assay data demonstrates that in silico and in vitro measures provide different approximations for developability. Values are coloured according to the TNP flags for the TNP metrics, derived from the ranges seen for these metrics across the 36 clinical-stage nanobodies. The in vitro assay values were assigned flags according to thresholds set by the ranges observed across all 108 nanobodies assayed.*

For example, Sample 2 was assigned red flags for total CDR length, CDR3 length, CDR3 compactness and patches of positive charge. This is accompanied by amber flags for patches of surface hydrophobicity and the HIC assay, as well as a red flag in the BVP-ELISA assay. Upon manual inspection, this structure has a very long, compact CDR3 loop, which can increase the likelihood of it containing larger, more hydrophobic residues to interact with the FR2 region [38] and may potentially lead to increased non-specific binding which is captured by the BVP-ELISA [89]. A similar pattern is observed for Sample 1, with consistent signals across both predicted and experimental metrics.

Some candidates display discrepancies between TNP predictions and *in vitro* assay results. Apart from red flags for patches of positive or negative charge, samples 7-11 receive green flags across all other TNP metrics. However, their *in vitro* assay data show multiple amber or red flags, often across all five assays, suggesting very poor developability profiles. A similar pattern is observed among the 36 clinical-stage nanobodies: although these molecules have advanced to clinical trials and therefore must to some extent achieve acceptable developability, many still receive amber or red flags in the *in vitro* assays (**Table C.3**). Relatively weak correlations between many of our calculated computational descriptors and the *in vitro* assay data suggests that there is limited agreement across all 108 experimentally-tested nanobodies, and that this is not specific to our chosen metrics (**Figure C.18**).

4.4.10 Web application

TNP is available as a web application at *opig.stats.ox.ac.uk/webapps/tnp*. Users can input a nanobody sequence to generate a developability profile of the six metrics we have described (**Figure 4.9**).

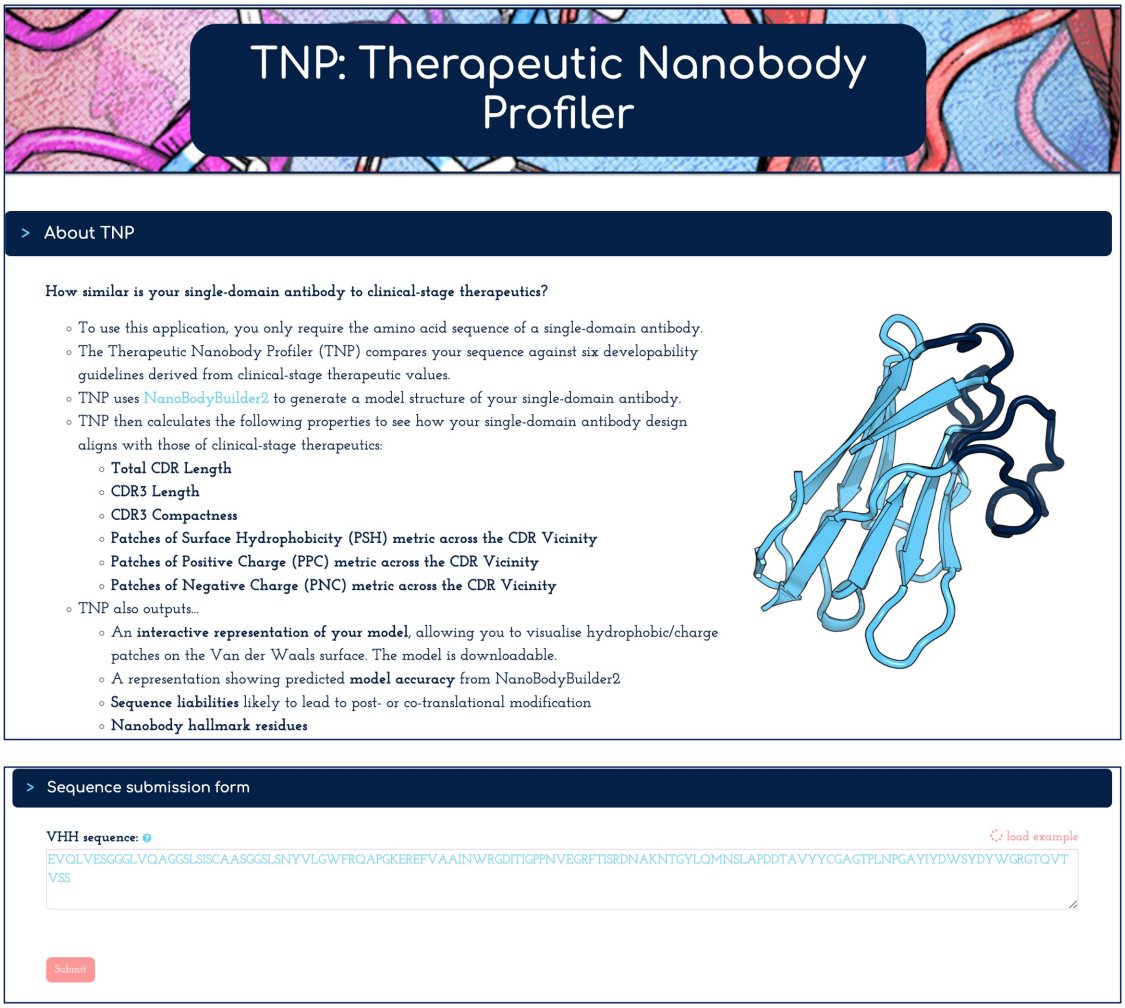


Figure 4.9: *The web application for TNP allows users to input a VHH sequence to generate results for the metrics we have described: total CDR length, CDR3 length, CDR3 compactness, and patches of surface hydrophobicity and charge across the CDR vicinity.*

The computed metrics include patch scores for surface hydrophobicity, positive charge, and negative charge, as well as total CDR length, CDR3 length, and CDR3 compactness. Each metric is benchmarked against the reference distribution derived from 36 clinical-stage nanobody sequences and predicted structures, with green, amber, or red flags assigned accordingly (**Figure 4.10**).



Figure 4.10: *The main results page for the TNP web application. The user is shown the different values for the TNP metrics for their sequence, which are then visualised on histogram plots to show how their sequence scores relative to the clinical-stage data.*

The web application enables interactive visualisation of the nanobody structure with mapped surface patches for hydrophobic and charged regions (**Figure 4.11**). Additionally, it provides annotations of tetrad residues and highlights potential sequence liabilities exposed on the nanobody surface (**Figure 4.12**).

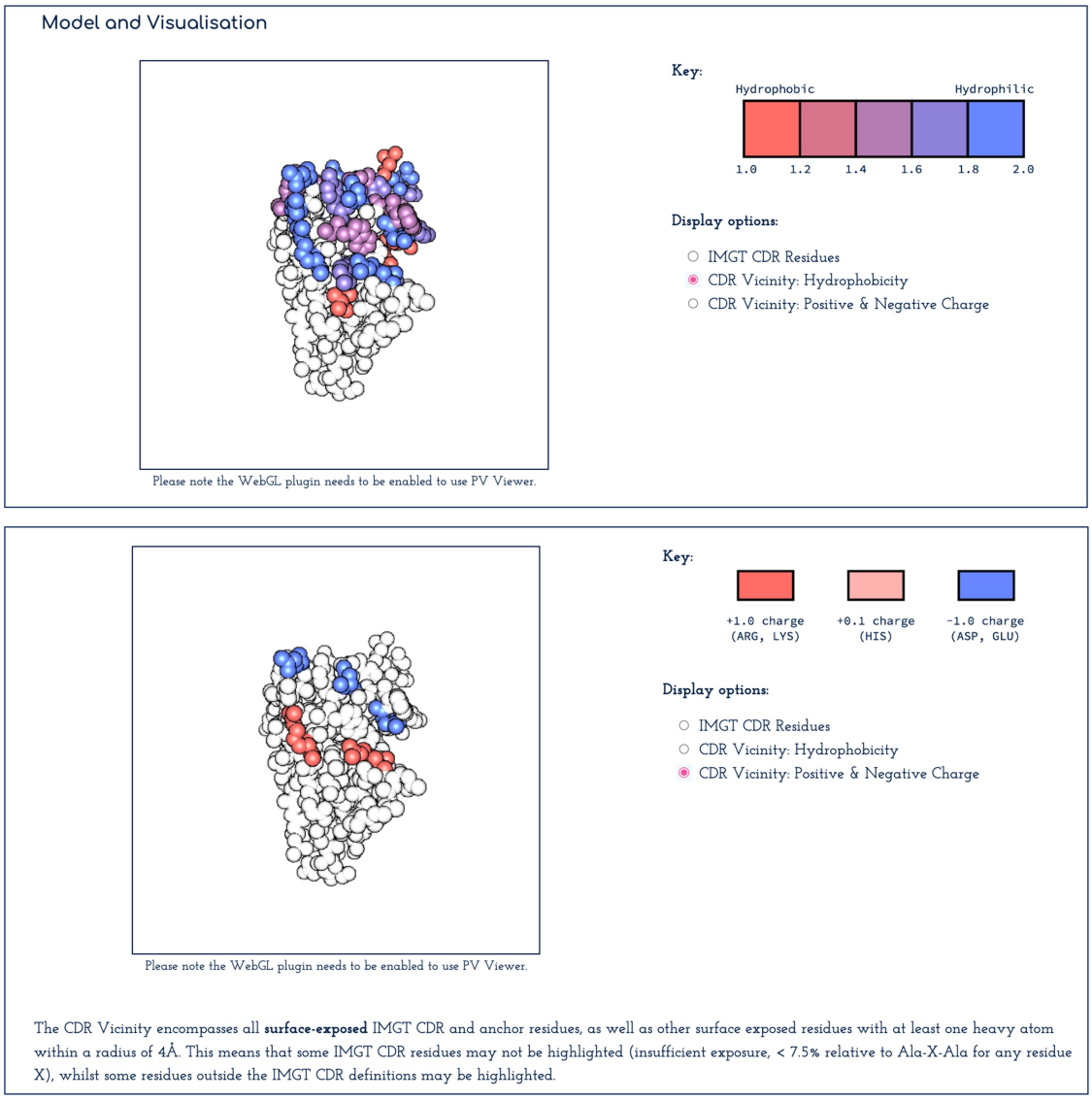


Figure 4.11: *The user can also view an interactive visualisation of the patches of surface hydrophobicity and charge on their modelled structure.*

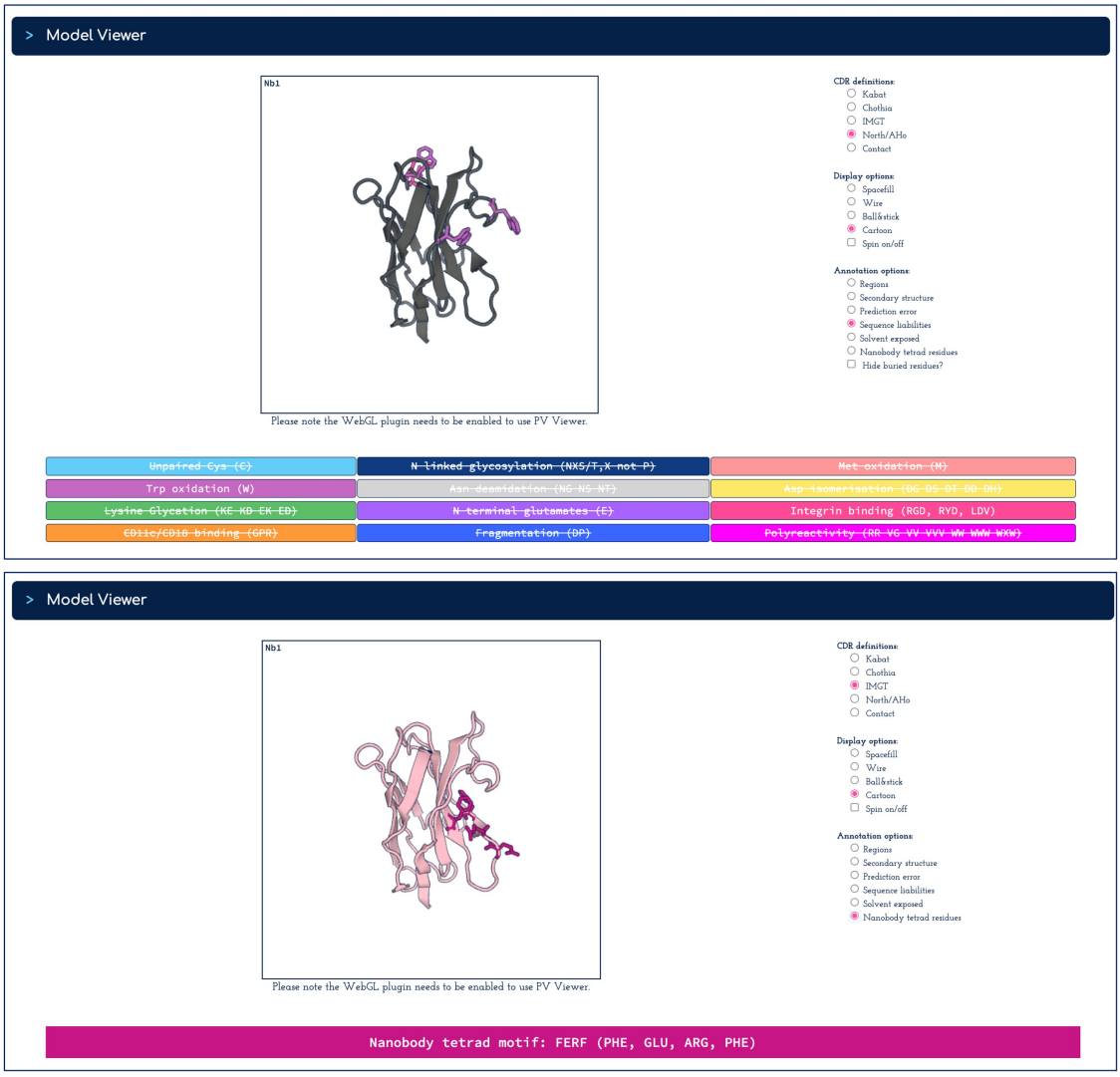


Figure 4.12: *The user is further provided with visualisations of the sequence liabilities and nanobody tetrad residues, which can be used to inform design strategies.*

4.5 Discussion

In this chapter, we introduced the Therapeutic Nanobody Profiler, a computational developability predictor specifically designed to account for the unique structural features of nanobodies, extending the methodology established by the Therapeutic Antibody Profiler (TAP) [62].

Our analysis demonstrates that clinical-stage nanobodies show distinct features compared to both clinical-stage and natural antibodies, supporting the need for a

nanobody-specific developability tool. We find that nanobody CDR3 loop diversity is less constrained than in antibody therapeutics, with clinical-stage nanobodies spanning the full range of CDR3 lengths, including longer loops not typically observed in the CDR-H3 of monoclonal antibodies [43, 44, 54, 55, 246].

Whilst these CDR3 loops adopt diverse conformations, they cluster into two structural subtypes based on compactness (our description of the loop length relative to its reach). The first subtype has generally longer, more compact loops (which fold over the FR2 region) and classical VHH-like tetrad motifs. Hydrophobic interactions between pairs of residues in the nanobody tetrad and CDR3 loop are common in these cases. The second subtype tends towards shorter, VH-like loops that extend away from the framework, with greater variation in the tetrad residue identities. Our findings are consistent with previous observations identifying distinct kinked and extended conformations in VHH CDR3 loops, and stabilising interactions between the CDR3 loop and nanobody tetrad residues in kinked conformations [38, 50, 209, 269].

Both of these subtypes of CDR3 compactness are present in the clinical-stage nanobody data, indicating that both are developable. These results suggest that therapeutic design need not be biased towards a particular conformational subtype. However, these subtypes may also impose design limitations: as longer CDR3 loops tend to fold over the FR2 region for stability, a compromise between CDR3 loop reach and length may be required when targeting less accessible epitopes. In the Therapeutic Antibody Profiler (TAP), total CDR length was used to capture both CDR3 length and paratope shape [62]. Total CDR length alone does not fully capture the more complex characteristics of the nanobody CDR3 loop, and so in the TNP, we include additional metrics for CDR3 length and compactness, to better describe the unique features we have observed.

We observed differences between the *in vitro* and *in silico* developability predictions for both clinical-stage and proprietary nanobody candidates. Although the 36 clinical-stage sequences must be developable to an extent since they have reached clinical trials, poor performance in the *in vitro* assays is not uncommon across the set. For our *in vitro* assays, developability flags were assigned based on thresholds which were set using the range of values from all 108 nanobodies assayed. However, since these are a small set relative to the entire nanobody sequence diversity, these thresholds may not accurately represent acceptable ranges for all nanobody constructs. This is further complicated by the exact format that molecules are assayed in: here, our nanobodies were expressed as part of an IgG construct, and the fusion to the Fc domain may affect biophysical measurements.

In comparison, TNP metrics are calculated on predicted structures of the VHH domain only. Previous studies have highlighted challenges in achieving reproducible results for both *in vitro* and *in silico* assays, noting that computational descriptors can vary significantly depending on the structural models and protocols used [213, 270, 271]. For example, hydrophobicity assessments may differ based on the choice of hydrophobicity scale [271]. Similarly, *in vitro* measurements can be influenced by assay-specific reagents and conditions and there can be limited agreement between different assays for the same property, further complicating direct comparisons between methods [87, 270]. Overall, current *in vitro* and *in silico* methods offer complementary but different approximations of developability.

As more nanobodies progress through clinical trials, the thresholds for the Therapeutic Nanobody Profiler will be adjusted, strengthening the reliability of developability predictions made using the software. Alongside the web application, we provide the Python package for high-throughput analyses, as well as the first publicly available set of nanobody developability assay data and extensive datasets of computational descriptors for the nanobodies analysed in this study. These resources aim to facilitate the development of more nanobody-specific computational tools and

support new hypotheses for the successful design of therapeutic nanobodies.

In the final chapter, we will summarise the main conclusions to be drawn from this thesis and suggest directions for future work.

5

Conclusions and future work

Contents

5.1 Conclusions 122
5.2 Future work 123

5.1 Conclusions

The aim of this DPhil was to improve understanding of the properties of nanobodies and develop computational methods to support their use in therapeutic applications. We have focused on characterising binding activity, addressing challenges in data availability, and contributing to current understanding of how nanobodies behave in clinical settings.

We conducted a comparative analysis of the binding sites of conventional antibodies and nanobodies using immune repertoire sequence data and crystal structures. By examining their paratopes, epitopes, and antigen interactions, we identified residues and regions essential for binding affinity and specificity, and highlighted the greater role that exposed framework residues and CDR3 conformation have in nanobody binding. These findings can be used to guide engineering strategies to optimise developability and humanisation without compromising binding, specifically accounting for the unique features of nanobodies.

A major bottleneck in developing nanobody-specific computational methods is the limited data. To that end, we created PLAbDab-nano, a searchable database of VHH and VNAR sequences and modelled structures collected from patents and academic literature. PLAbDab-nano allows users to compare input sequences against those in the database, which can help identify potential improvements by referencing against clinical-stage or patented sequences. Each entry is annotated with CDR loop information, which facilitates the design of nanobody libraries that capitalise on the CDR3 diversity we observe across our set of VHH and VNARs. To our knowledge, this is the first resource to both centrally collate VNAR data and provide automated sequence annotation using established numbering schemes.

VHH data from PLAbDab-nano was used to characterise developability traits in nanobodies, incorporating descriptors for CDR3 conformation developed as part

of our binding site analysis. Based on this, we built the Therapeutic Nanobody Profiler (TNP), a nanobody-specific computational tool that assesses multiple developability indicators in parallel by comparing candidates against clinical-stage nanobodies. TNP can be used at high throughput to screen the developability of new candidates or libraries. Alongside this, we have generated data for over 80 *in silico* developability-related sequence- and structure-based descriptors, as well as experimental developability data for 108 nanobodies. These resources are publicly available to support future computational approaches to therapeutic nanobody design.

5.2 Future work

Although therapeutic nanobody design encompasses both VHHs and VNARs, aside from the PLAbDab-nano database this work has primarily concentrated on VHHs. A promising direction for future work would be to extend the approaches we have described to VNARs. With the insights gained into VHH binding sites, it would be particularly valuable to investigate how VNARs differ in their epitope preferences and binding strategies, to inform computational approaches to their optimisation as therapeutics. Given the distinctive features of VNARs, they may offer access to novel epitopes that are not able to be targeted by conventional antibodies or VHHs, especially since we find a large overlap in the epitopes targeted by the latter. However, as reflected in the distribution of entries within PLAbDab-nano, available data for VNARs remains limited relative to VHHs. This currently constrains the scope of comparative and structure-based analyses. Nevertheless, the situation is likely to improve over time as more VNAR structures are solved and as structure prediction methods for antibody-antigen complexes continue to advance.

Improvements to the PLAbDab-nano pipeline could increase the volume of both VHH and VNAR data captured. We currently filter for VNAR sequences using an HMM, which may not recognise VNARs that deviate further from the training data. The ANARCII model could replace this method as it can both recognise and

annotate VNARs in one step. This would help to include more diverse sequences and streamline the database pipeline for more efficient data collection. For VHHs, the classification approach which filters VH sequences from VHHs could be refined to identify humanised VHH sequences, which may currently be misclassified as conventional VH domains. This would help recover additional VHH entries that are presently excluded, and increase the volume of nanobody data that can be used for larger scale analyses or model training.

Expanding the available VHH and VNAR data in PLAbDab-nano could enable more fine-grained analyses of nanobodies. For instance, stratifying sequences by species would allow for investigation of species-specific features that may influence binding properties and developability. This is relevant for both VHHs and VNARs, where certain subtypes appear to be species-specific. Such insights could inform experimental discovery efforts, especially when designing libraries enriched for particular sequence or structural features.

Finally, the development of the Therapeutic Nanobody Profiler involved generating new data from *in vitro* developability assays. While our dataset significantly contributes to current accessibility of nanobody experimental data, further work is needed to more accurately define acceptable clinical ranges of nanobody properties in these assays. In particular, aligning *in silico* predictions with *in vitro* observations remains a complex challenge. A more detailed analysis of the currently available data, combined with additional experimental datasets in the future, could help refine predictive models for nanobody developability. Ultimately, robust experimental validation remains essential to confirm the reliability of computational methods, and enable their integration into established therapeutic development pipelines.

Appendices

A

A comparison of the binding sites of antibodies and camelid nanobodies

Text A1

After creating the structural datasets as described in the **Methods** section, four nanobody complexes and seven antibody complexes were removed. The removed nanobody PDB entries are 7D30, 6UL6, 4HIJ and 5OMM. For the antibody dataset, the PDB entries are 2H32, 4NZR, 4NZT, and 1XIW, and the complexes including antibody chains KJ and ML in PDB entry 5MHR were removed. This was due to either erroneous annotation, lacking complete CDR loops, or structural anomalies.

Arpeggio failed to run for two PDB entries, even after cleaning the PDB files. These PDB entries are 4XWO and 5K9Q and contained antibodies. The antibody dataset containing the interactions-defined epitopes and paratopes therefore has a size of 890 complexes.

Text A2

We analysed the general orientation of CDR3 loops of nanobodies and antibody heavy chains by examining their centres of geometry in reference to an $\mathbb{R}^3$ coordinate system. We describe the centre of geometry of the CDR3 loop using spherical

coordinates. To orient the coordinate system identically in all examined structures, we used the following fitting procedure.

Let $\mathbf{c}$ be the centre of geometry of the backbone atoms of IMGT positions 102, 103, 118, and 119. Let $\mathbf{p}$ be the first principal component vector of the anchor positions which minimises the distance to the Cα atoms of IMGT position 103 when extended from $\mathbf{c}$. Let $\mathbf{a}$ be the centre of geometry of all VH/VHH anchor positions (we take the anchors to be the three residues on either side of the loop as defined by the IMGT numbering scheme.). Let $\mathbf{\Pi}$ be the plane that crosses $\mathbf{c}$ whose normal vector is $\mathbf{p}$. Therefore for all $\mathbf{r} = \langle x, y, z \rangle$,

$$\mathbf{p} \cdot (\mathbf{r} - \mathbf{c}) = 0 \tag{A.1}$$

$$\mathbf{p} \cdot \mathbf{r} = \mathbf{p} \cdot \mathbf{c} \, . \tag{A.2}$$

We calculate the shortest distance of any point $\mathbf{r}$ from $\mathbf{\Pi}$ using

$$d(\mathbf{r}) = \frac{\mathbf{p} \cdot \mathbf{r} - \mathbf{p} \cdot \mathbf{c}}{||\mathbf{p}||} \, . \tag{A.3}$$

Let $\mathbf{u} = \mathbf{a_\Pi} - \mathbf{c}$, where $\mathbf{a_\Pi}$ is the orthogonal projection of $\mathbf{a}$ onto $\mathbf{\Pi}$, given by

$$\mathbf{a_\Pi} = \mathbf{a} - d(\mathbf{a}) \frac{\mathbf{p}}{||\mathbf{p}||} \, . \tag{A.4}$$

Let $\mathbf{v}$ be a vector orthogonal to both $\mathbf{u}$ and $\mathbf{p}$, such that $\mathbf{v}$ the resultant coordinate system is left-handed,

$$\mathbf{v} = \underset{\mathbf{x} \in \{\mathbf{u} \times \mathbf{p}, -\mathbf{u} \times \mathbf{p}\}}{\operatorname{argmax}} \begin{vmatrix} x^{(1)} & u^{(1)} & p^{(1)} \\ x^{(2)} & u^{(2)} & p^{(2)} \\ x^{(3)} & u^{(3)} & p^{(3)} \end{vmatrix} \, . \tag{A.5}$$

Given the vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{p}$, such that $\mathbf{u} \perp \mathbf{v} \perp \mathbf{p}$, we aim to describe the CDR3 backbone atomic positions using spherical coordinates.

Let $\mathbf{h}$ be the centre of geometry of the CDR3 backbone positions as defined by the IMGT numbering scheme. To describe $\mathbf{h}$ relative to $\mathbf{c}$ using spherical coordinates, $\mathbf{h_c}(\rho, \theta, \phi)$, we define $\rho = ||\mathbf{h} - \mathbf{c}||$. θ describes the elevation or polar angle, and ϕ describes the azimuth (the horizontal angle). θ decreases as the CDR3 loop projects directly up and away from the VH/VHH domain. ϕ increases as the CDR3 loop points out from the VH/VHH domain. In the case of antibodies, this refers to a CDR-H3 loop that is oriented more towards the VL domain.

We use $\mathbf{p}$ as the polar reference vector. Thus,

$$\theta = \cos^{-1} \frac{(\mathbf{h} - \mathbf{c}) \cdot \mathbf{p}}{||\mathbf{h} - \mathbf{c}||\ ||\mathbf{p}||} \ . \tag{A.6}$$

Lastly, we use $\mathbf{u}$ as the azimuthal reference vector. Therefore, ϕ will be the angle between $\mathbf{u}$ and $\mathbf{h_\Pi} - \mathbf{c}$, where $\mathbf{h_\Pi}$ is the orthogonal projection of $\mathbf{h}$ onto $\mathbf{\Pi}$,

$$\mathbf{h_\Pi} = \mathbf{h} - d(\mathbf{h}) \frac{\mathbf{p}}{||\mathbf{p}||} \tag{A.7}$$

and

$$\phi = \cos^{-1} \frac{(\mathbf{h_\Pi} - \mathbf{c}) \cdot \mathbf{u}}{||\mathbf{h_\Pi} - \mathbf{c}||\ ||\mathbf{u}||} \ . \tag{A.8}$$

In order to consistently measure the azimuth angle, we calculate

$$\begin{vmatrix} h_\Pi^{(1)} & u^{(1)} & p^{(1)} \\ h_\Pi^{(2)} & u^{(2)} & p^{(2)} \\ h_\Pi^{(3)} & u^{(3)} & p^{(3)} \end{vmatrix} \ .$$

As we aim to measure the clockwise displacement angle using $\mathbf{p}$ as a reference, we use the calculated ϕ as the azimuth angle when the determinant is positive. Else we define the azimuth reference angle as $2\pi - \phi$.

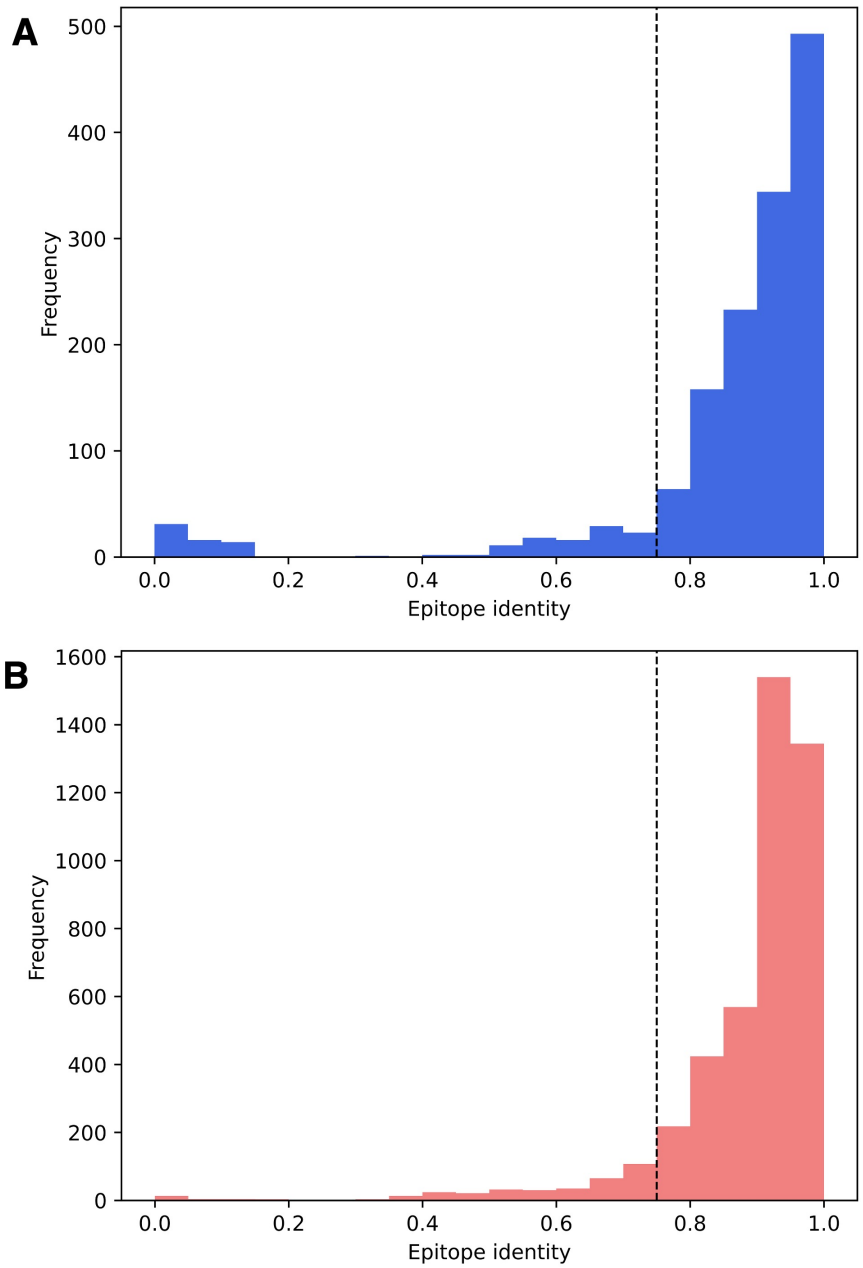


Figure A.1: An epitope identity cut-off of 75% (marked by the dotted line) was used for reintroducing complexes in the structural datasets. The epitope identity of all the complexes within each cluster determined by cd-hit is shown for (**A**) nanobodies and (**B**) antibodies.

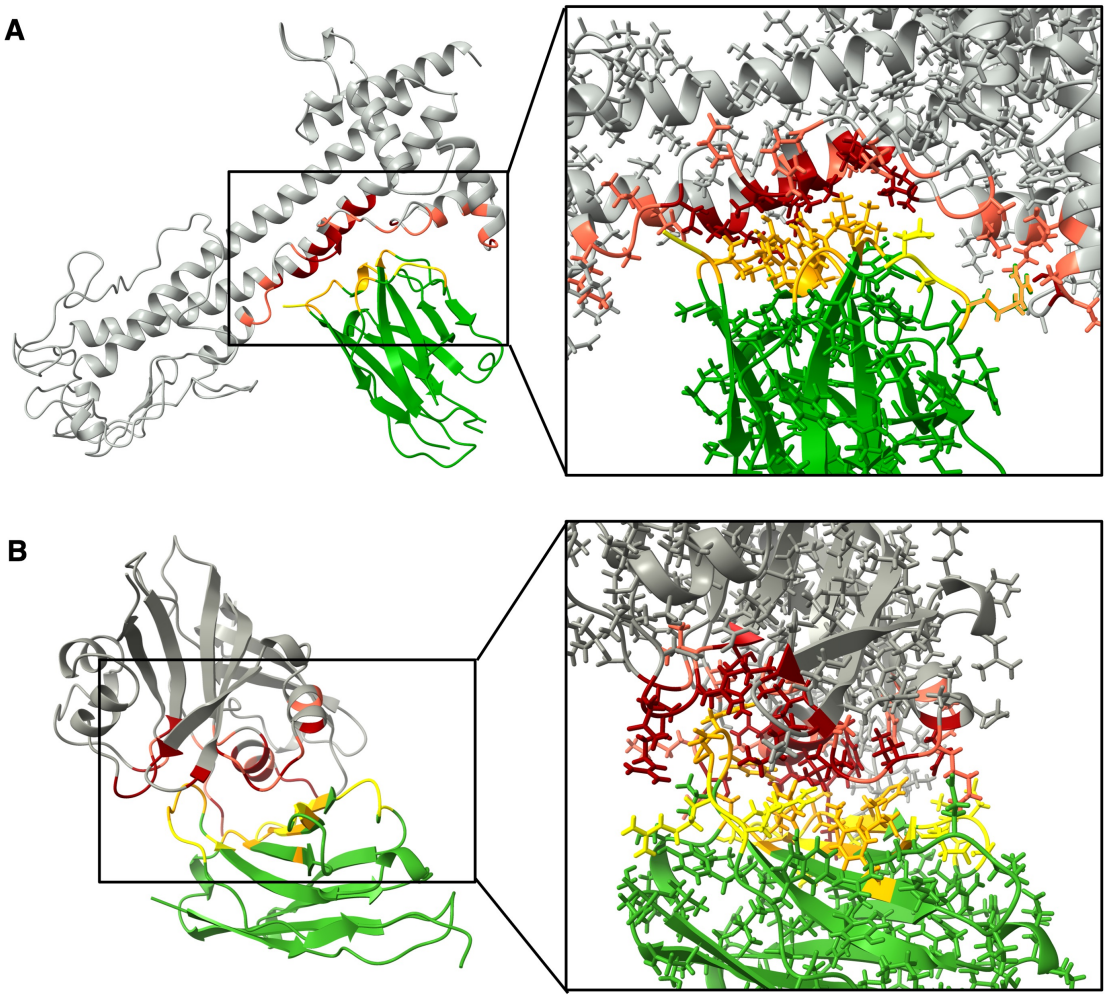


Figure A.2: Binding site definitions based on interacting residues are stricter. (**A**) Complex between a nanobody (chain C) and an antigen (chain A) corresponding to PDB entry 7AQX [272]. (**B**) Complex between a nanobody (chain B) and an antigen (chain A) corresponding to PDB entry 6QUP [273]. The antigen is coloured grey, the positions exclusively annotated by distance-defined epitope are pink, and positions in both the distance-defined epitope and interactions-defined epitope are dark red. The antibody is coloured green, the positions exclusively annotated by the distance-defined paratope are yellow, and positions in both the distance-defined and interactions-defined paratopes are orange. The pink and yellow residues are the contacts. The dark red and orange residues are identified as real interactions by Arpeggio.

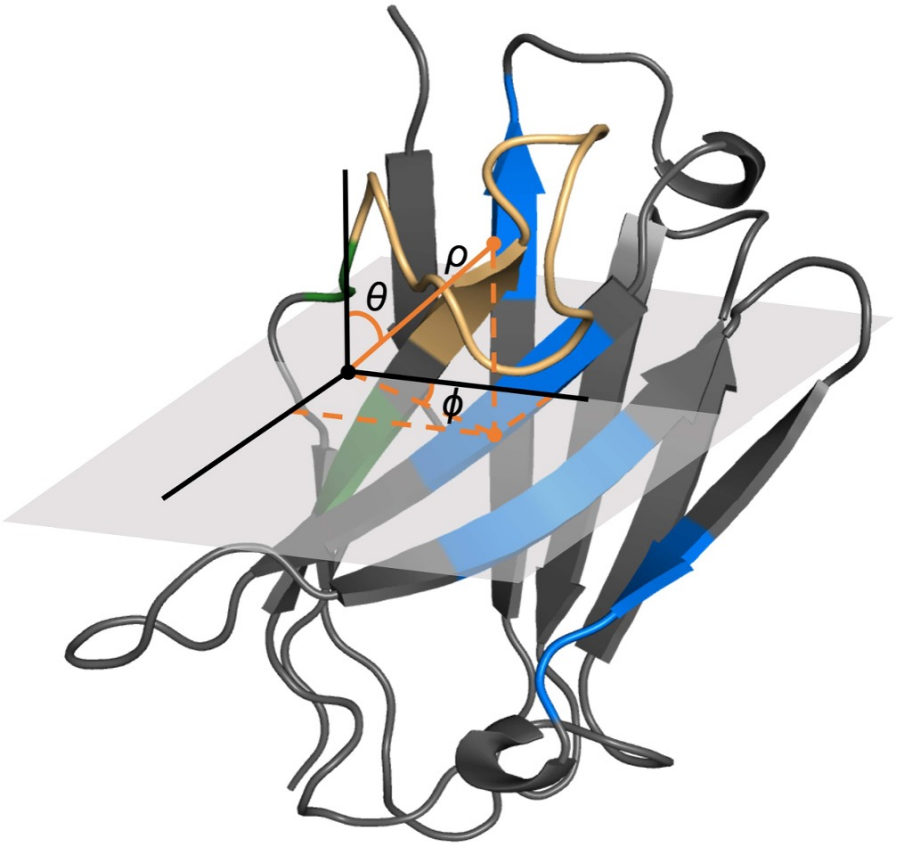


Figure A.3: A visualisation of the coordinate system used to determine the orientation of the heavy chain CDR3 loops, where ρ describes the reach of the CDR3 loop away from the rest of the VH/VHH domain, ϕ gives an indication of whether the CDR3 loop is horizontally oriented towards the rest of the VH/VHH domain or away from it, and θ gives a measure of the elevation of the loop.

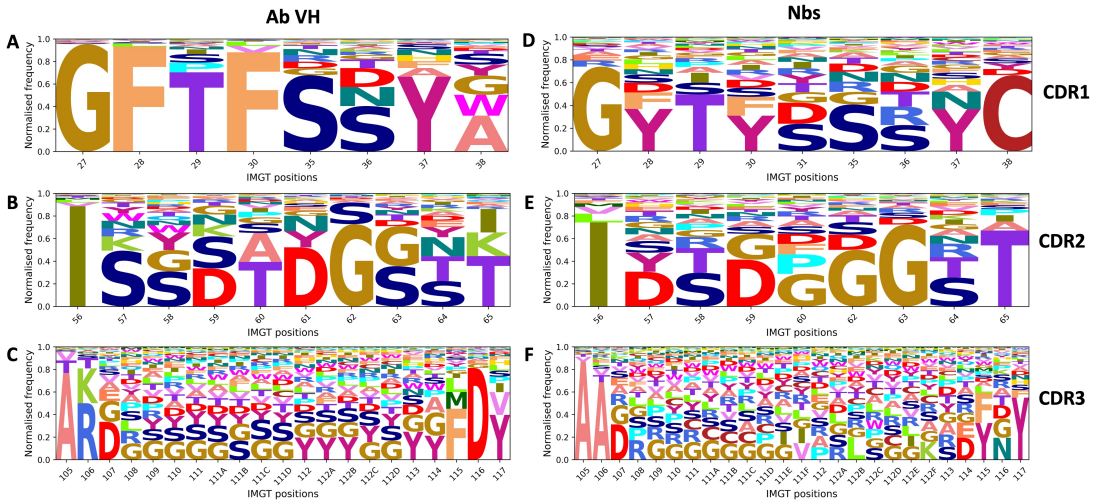


Figure A.4: Sequence logo plots of the amino acid composition at each position for each of the CDR loops CDR1 (**A, D**), CDR2 (**B, E**) and CDR3 (**C, F**) shows similarities between nanobody sequences and antibody VH sequences from the IGHV3 germline. Sequence logo plots were generated by determining the proportions of each amino acid in each position across all sequences. All positions that occurred in less than 5% of sequences have been removed.

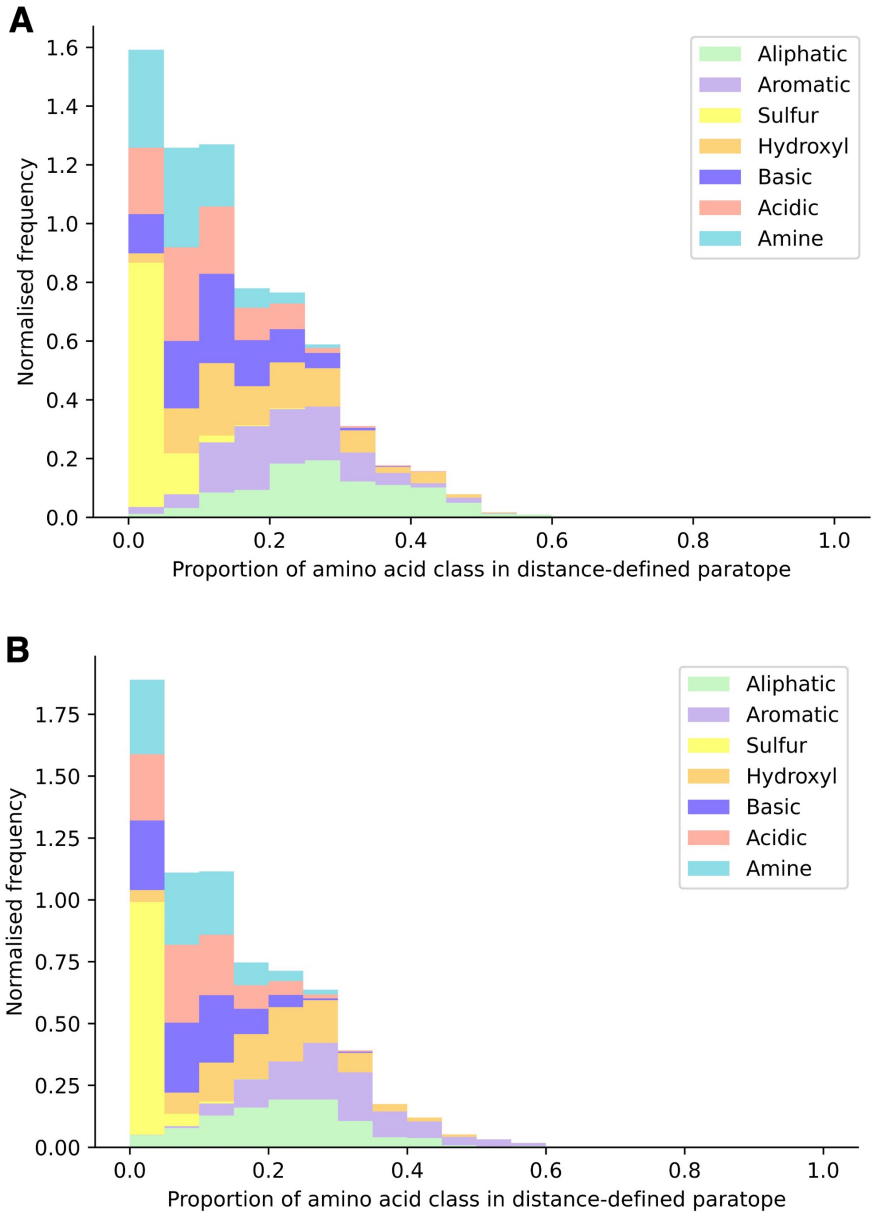


Figure A.5: Distance-defined paratopes of (**A**) nanobodies and (**B**) antibodies show small differences in their amino acid compositions. Unpaired mean differences between nanobodies and antibodies in the proportions of the different classes are: -0.059 (aliphatic), 0.078 (aromatic), -0.012 (sulfur), 0.0087 (hydroxyl, p-value = 0.29), -0.037 (basic), 0.00029 (acidic, p-value = 0.96), 0.0084 (amine, p-value = 0.10).

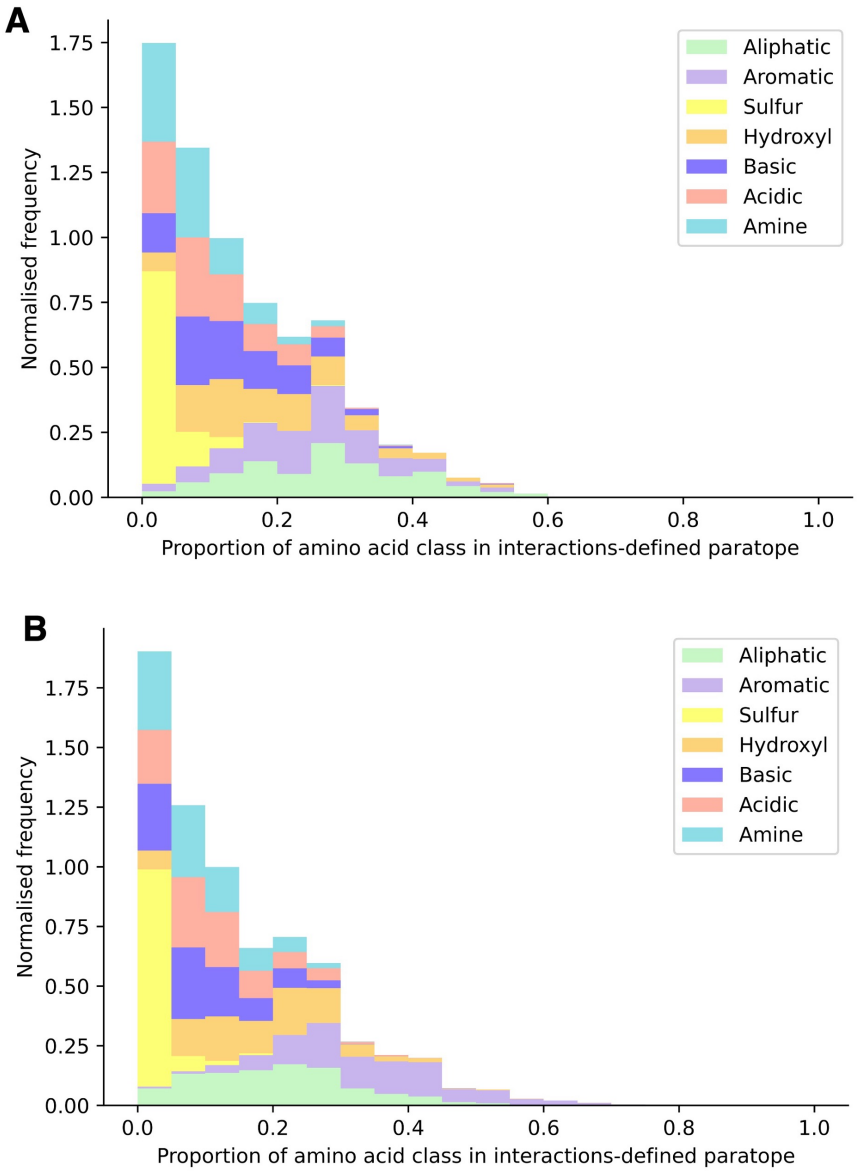


Figure A.6: Interactions-defined paratopes of (**A**) nanobodies and (**B**) antibodies show small differences in their amino acid class compositions. Unpaired mean differences between nanobodies and antibodies in the proportions of the different classes are: -0.053 (aliphatic), 0.094 (aromatic), -0.0082 (sulfur), 0.0027 (hydroxyl, p-value = 0.77), -0.036 (basic), 0.011 (acidic, p-value = 0.11), 0.0078 (amine, p-value = 0.19).

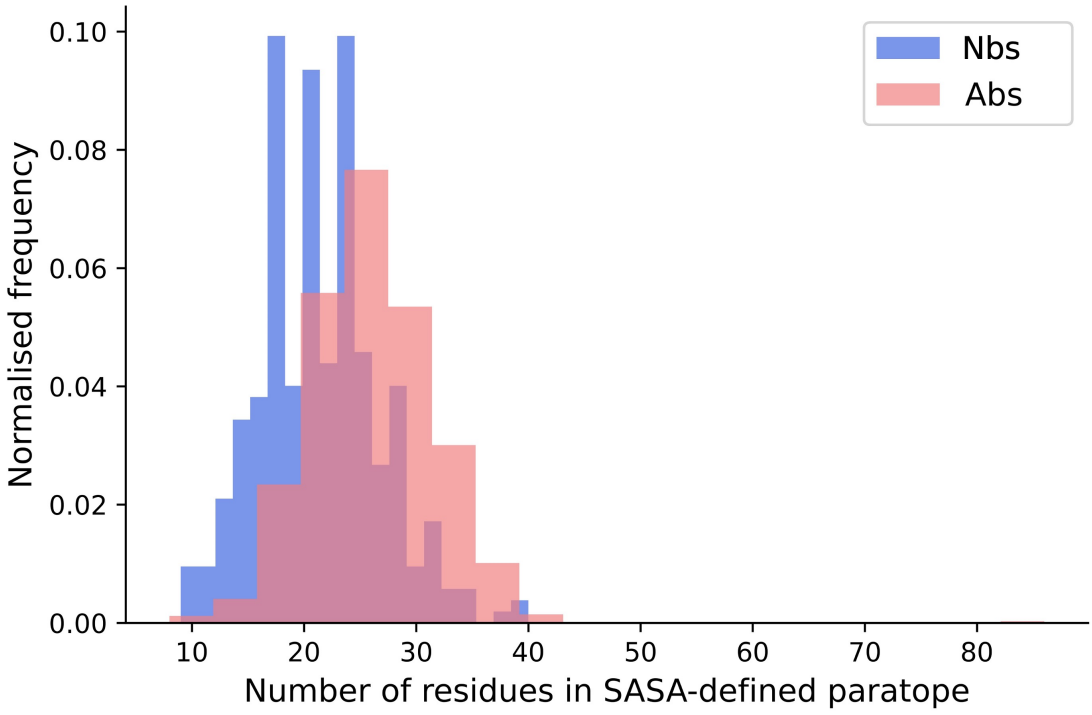


Figure A.7: Comparing the distributions of paratope size (as defined by number of residues) for nanobodies and antibody complexes using the SASA-defined paratope shows that the nanobody paratopes are significantly smaller by 4.9 residues on average.

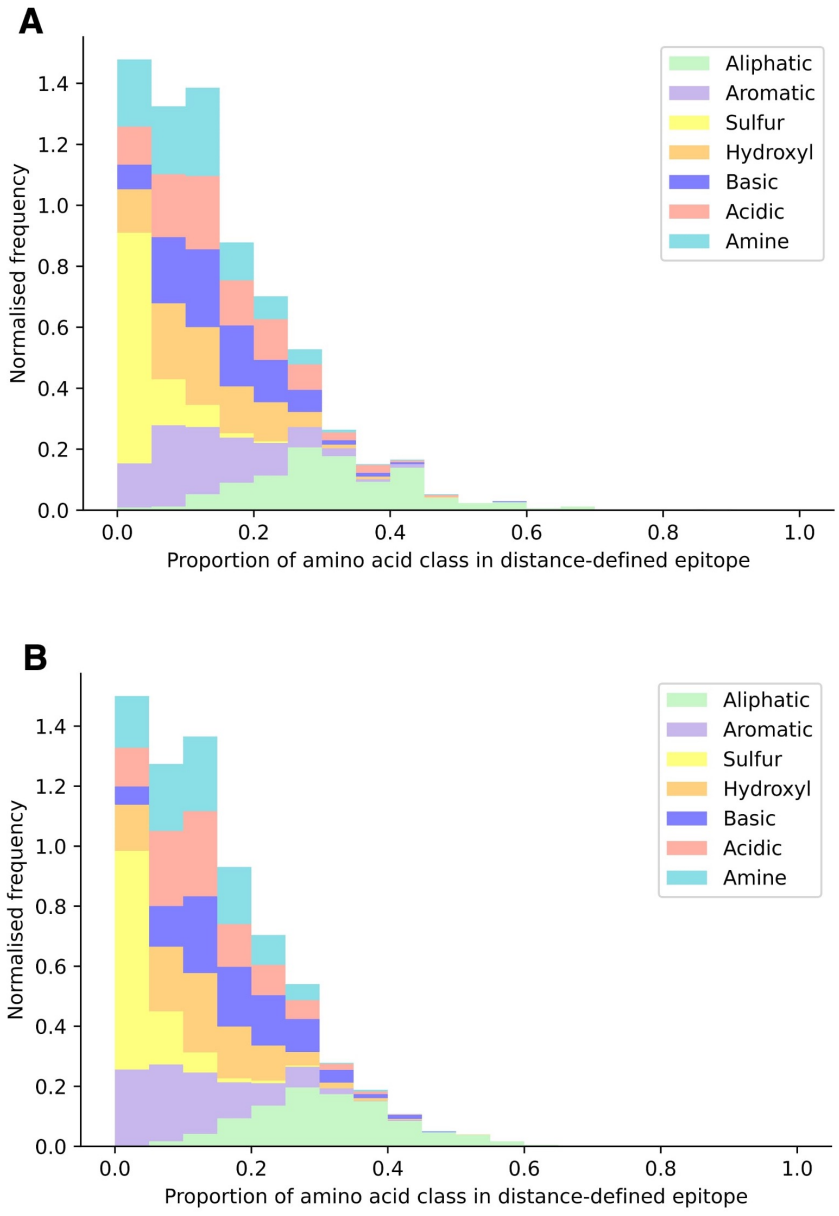


Figure A.8: Distance-defined epitopes of (**A**) nanobodies and (**B**) antibodies show similar amino acid compositions. The seven classes are: aliphatic, aromatic, sulfur, hydroxyl, basic, acidic and amine. The p-values obtained by the two-sided permutation t-test after bootstrap resampling on every class respectively are: 0.63, 0.015, 0.64, 0.76, 0.00060, 0.080 and 0.25.

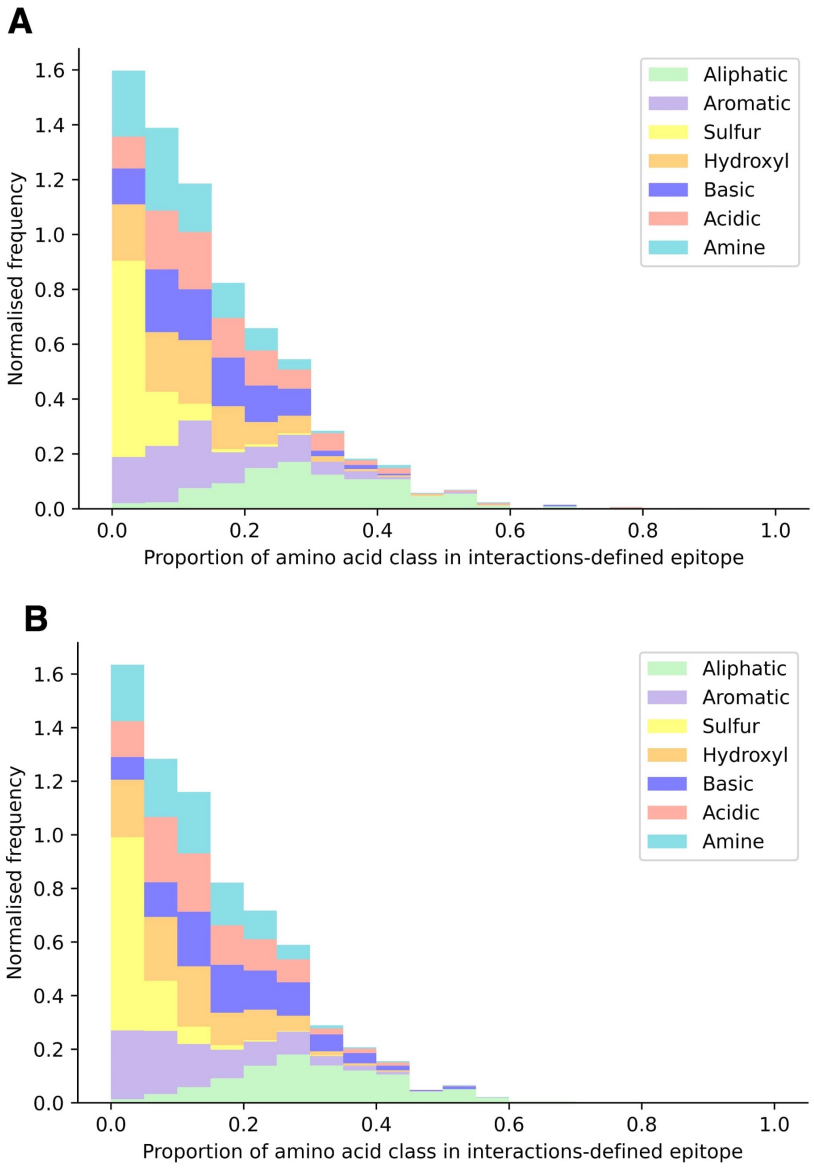


Figure A.9: Interactions-defined epitopes of (**A**) nanobodies and (**B**) antibodies show similar amino acid class compositions. The seven classes are: aliphatic, aromatic, sulfur, hydroxyl, basic, acidic and amine. The p-values obtained by the two-sided permutation t-test after bootstrap resampling on every class respectively are: 0.11, 0.027, 0.93, 0.31, 0.0014, 0.058 and 0.74.

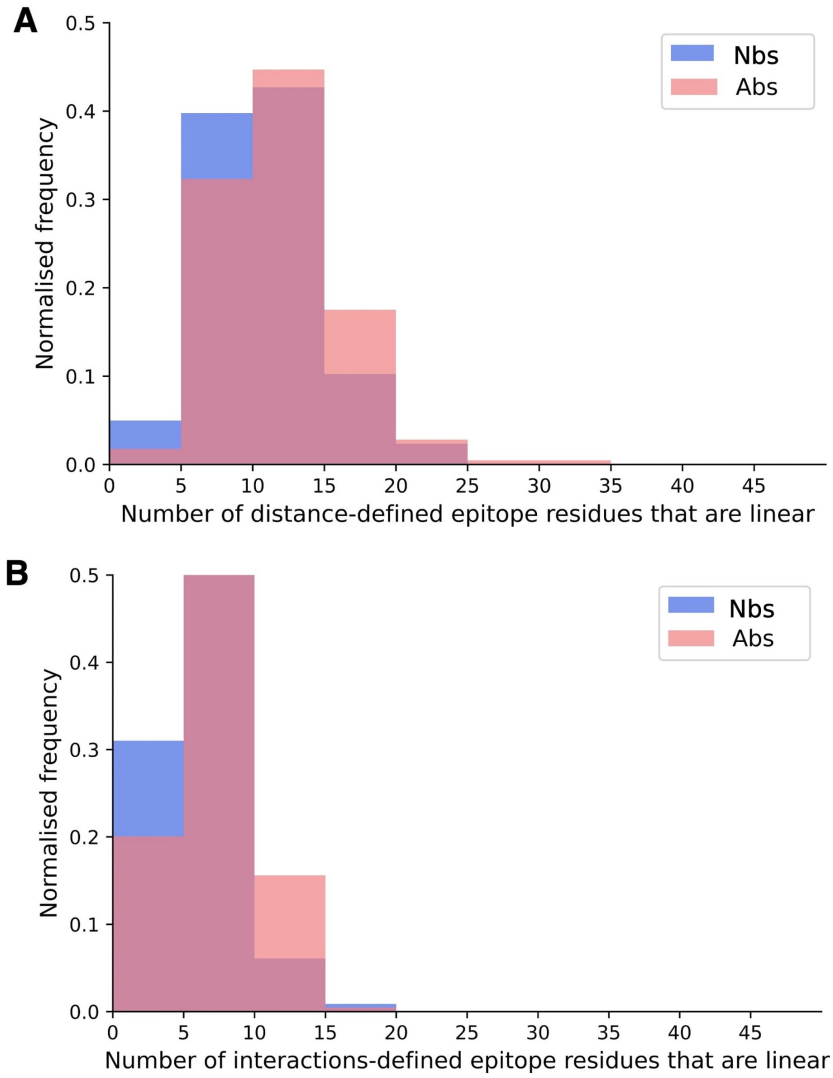


Figure A.10: The epitopes targeted by antibodies are relatively more linear than the epitopes targeted by nanobodies. Distributions of raw counts of linear residues for epitopes targeted by antibodies and nanobodies for the (**A**) distance-defined epitopes and (**B**) interactions-defined epitopes.

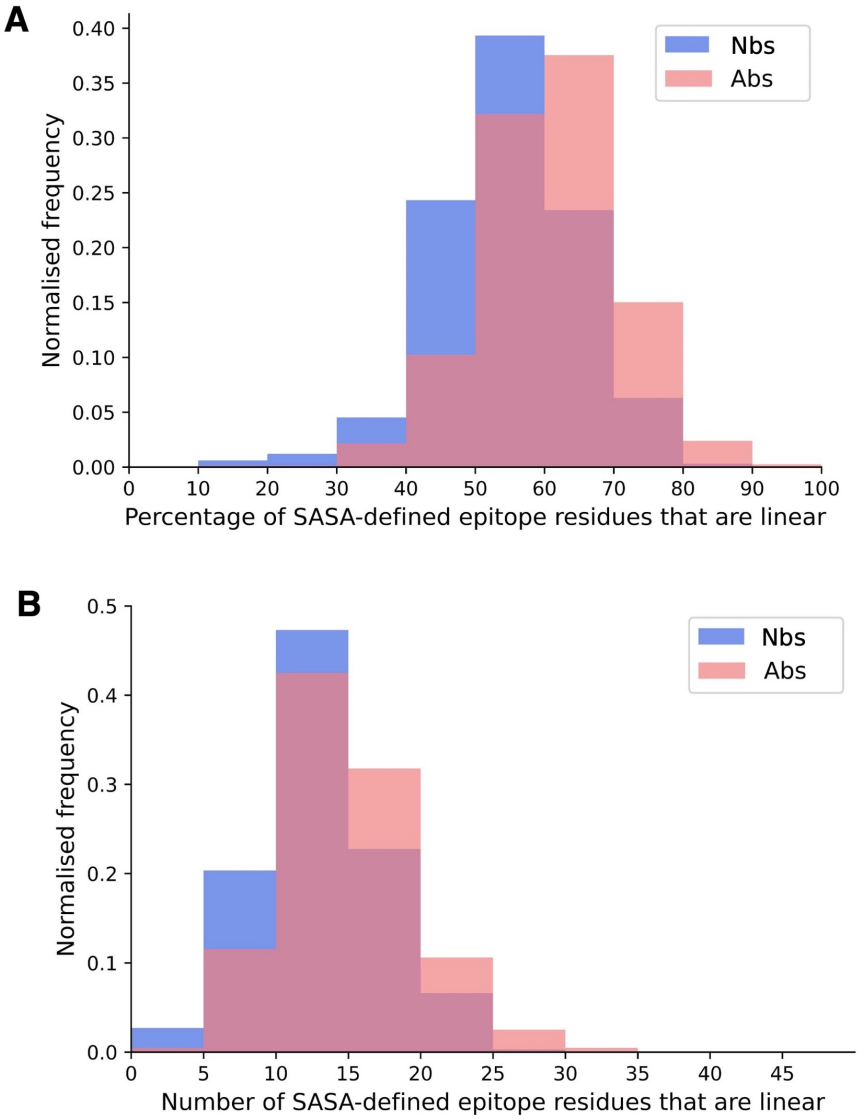


Figure A.11: The SASA-defined epitopes targeted by antibodies are relatively more linear than the SASA-defined epitopes targeted by nanobodies. (**A**) Distributions of percentage of linear residues for epitopes targeted by antibodies and nanobodies. The unpaired mean difference between nanobodies and antibodies was 6.3% (p-value $<<$ 0.01). (**B**) Distributions of raw counts of linear residues for epitopes targeted by antibodies and nanobodies. The unpaired mean difference between nanobodies and antibodies was 1.6 residues (p-value $<<$ 0.01).

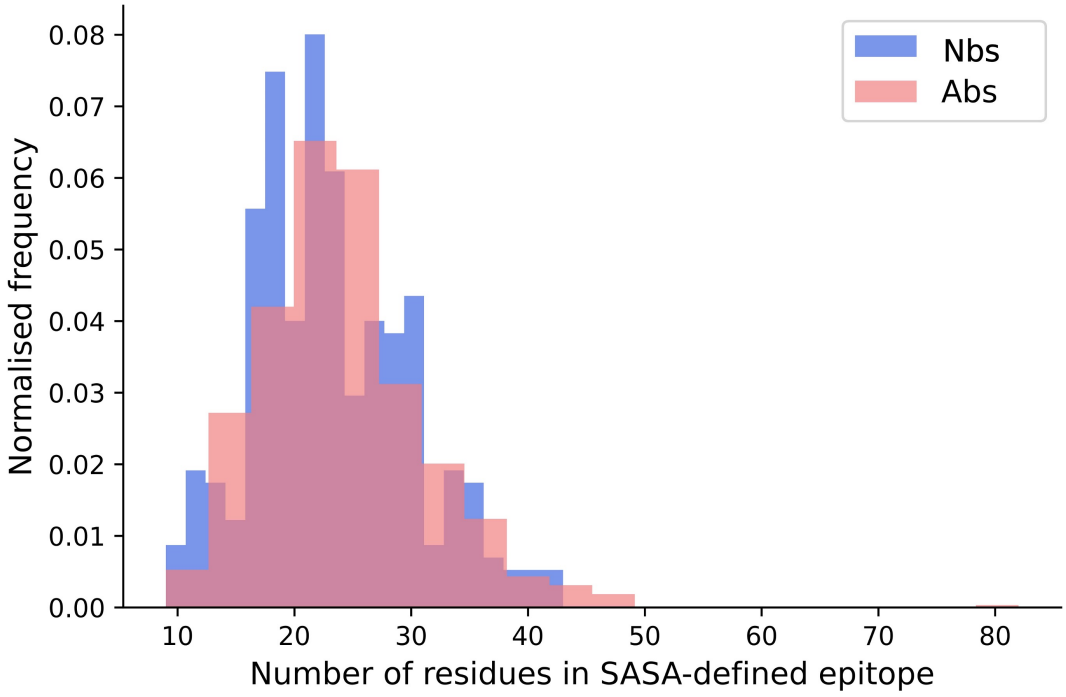


Figure A.12: Comparing the distributions of epitope size, as defined by number of residues, for nanobodies and antibodies using the SASA-defined epitope, suggests that there is no significant difference in size between the epitopes targeted by nanobodies and antibodies.

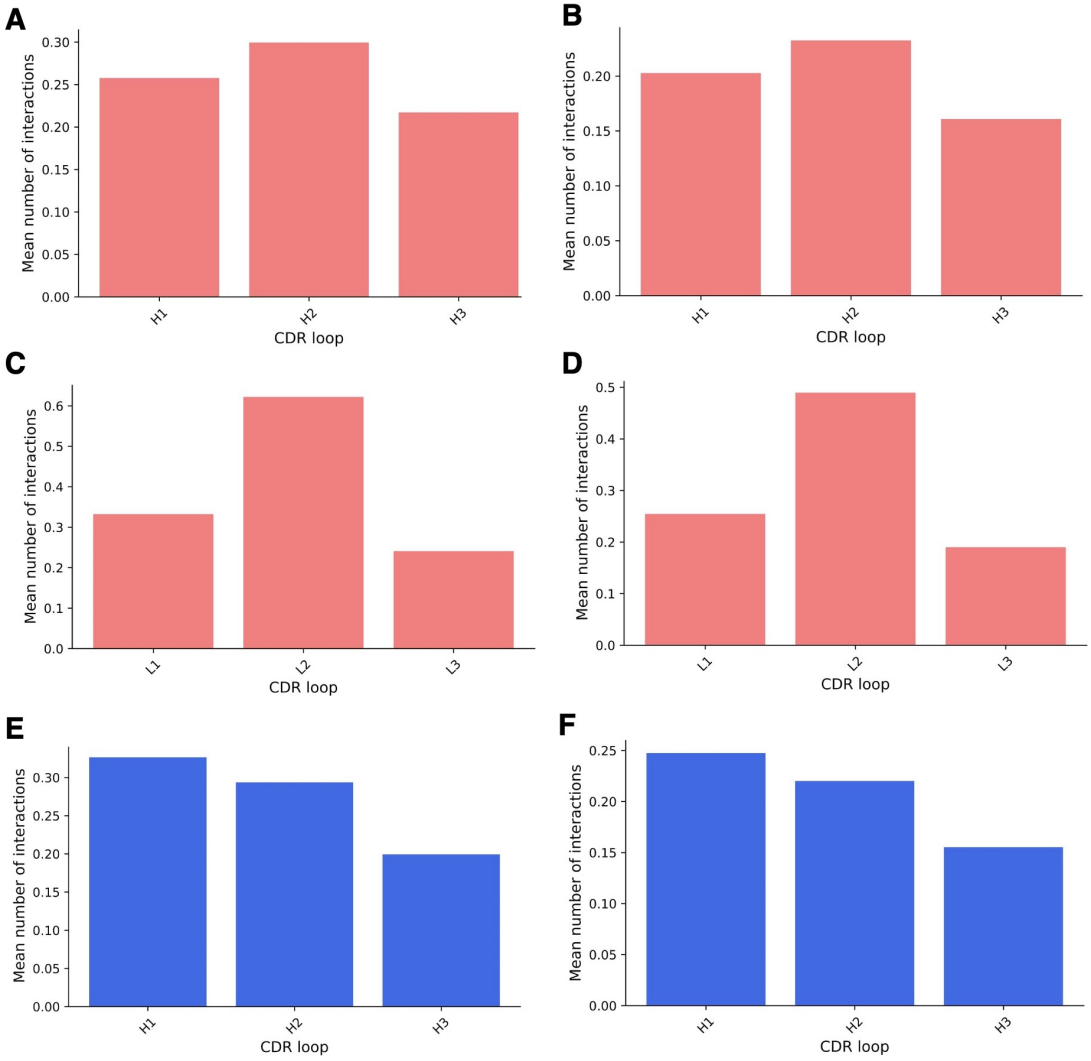


Figure A.13: The mean number of interactions per CDR loop, normalised by loop length per complex. (**A**), (**C**) and (**E**) show these results for the distance-defined paratopes for the antibody VH, antibodies VL and nanobodies respectively, whilst (**B**), (**D**) and (**F**) show the same but for the interactions-defined paratopes.

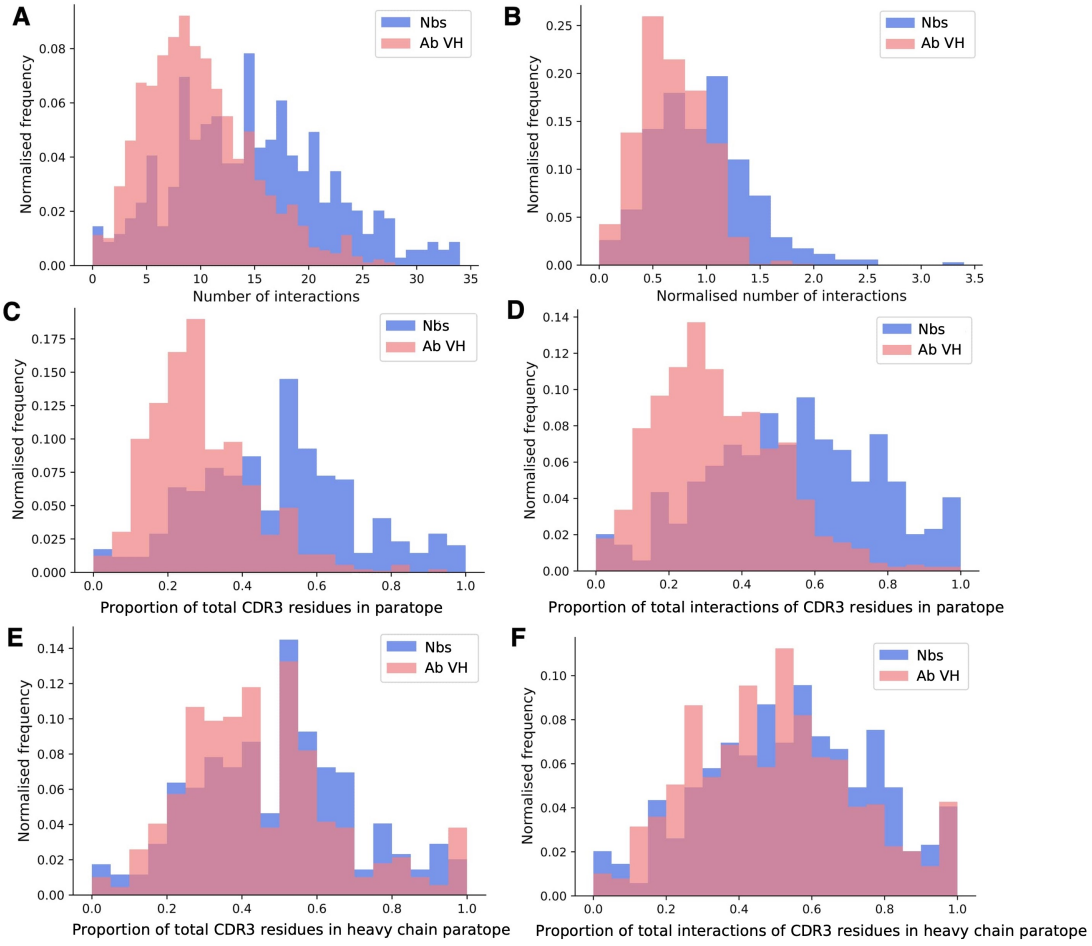


Figure A.14: The CDR3 loop is of increased importance in nanobodies. (**A**) Distribution of the number of interactions established by CDR3 and CDR-H3 residues per complex in the nanobody and antibody structure datasets. The unpaired mean difference between nanobodies and antibodies is -4.8, p-value << 0.01. (**B**) Normalised distribution for the number of interactions by the size (number of residues) of the CDR3/CDR-H3. The unpaired mean difference between nanobodies and antibodies is -0.27, p-value << 0.01. (**C**) Comparison of the proportion of residues contributed by the CDR3 loops for nanobodies and the CDR-H3 for antibodies. There is an unpaired mean difference of -0.19 between nanobodies and antibodies with a p-value << 0.01. (**D**) Comparison of the proportion of interactions contributed by the CDR3 loops for nanobodies and the CDR-H3 loops for antibodies. There is an unpaired mean difference of -0.21 between nanobodies and antibodies with a p-value << 0.01. (**E**) Comparison of the proportion of residues contributed by the CDR3 loops for nanobodies and CDR-H3 loops for antibodies, for the antibody VH only. There is an unpaired mean difference of -0.075 between nanobodies and antibody VH with a p-value << 0.01. (**F**) Comparison of the proportion of interactions contributed by the CDR3 loops for nanobodies and the CDR-H3 loops for antibodies, for the antibody VH only. There is an unpaired mean difference of -0.052 between nanobodies and antibodies with a p-value = 0.006.

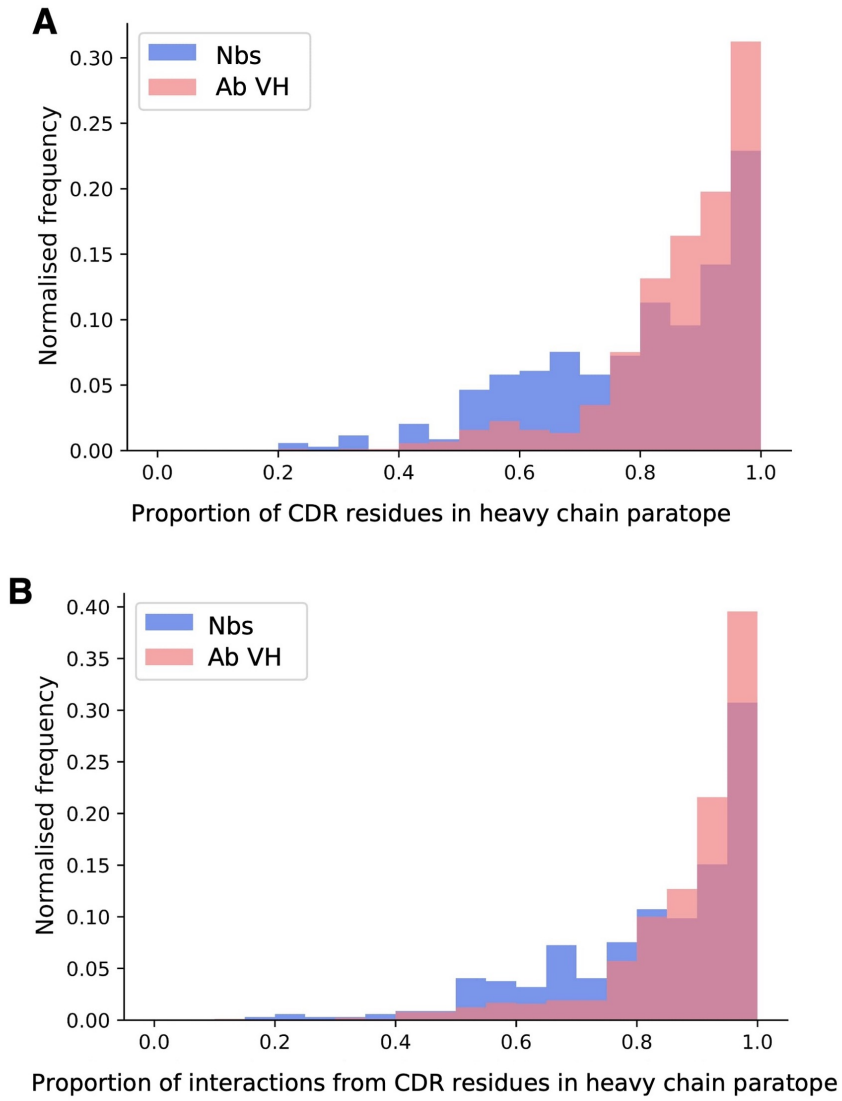


Figure A.15: When comparing the nanobody paratope to the paratope residues from the antibody VH chains only, we observe that relatively fewer interactions are contributed by the CDRs in nanobodies, indicating that there is greater involvement of framework residues in the nanobody paratope. (**A**) Comparison of the proportion of residues contributed by the CDR loops for nanobodies and antibody VH chains. There is an unpaired mean difference of 0.077 between nanobodies and antibodies with a p-value $<<$ 0.01. (**B**) Comparison of the proportion of interactions contributed by the CDR loops for nanobodies and antibody VH chains. There is an unpaired mean difference of 0.055 between nanobodies and antibodies with a p-value $<<$ 0.01.

Table A.1: The number of complexes remaining in the nanobody structural dataset and the antibody structural dataset are shown after performing each filtering step. Complexes with an epitope and/or paratope size smaller than seven residues were removed.

Filtering steps	**Total nanobodies**	**Total antibodies**
Selected PDB entries	411	1260
Binding CDR residue	1044	1751
Removed outliers	1038	1045
Filtered SAbDab	816	1703
Filtered binding site size	781	1699
Filtered CDR sequence identity	309	792
Reintroduced non-identical epitopes	345	892

Table A.2: For the structural datasets, greater species variation is observed in antibodies than nanobodies. This table shows the species assigned in the nanobody and antibody structural datasets. Species are derived from the heavy chain domain as stored in the SAbDab database [141, 142].

Species	**Nanobodies**	**Antibodies**
Lama glama	152	4
Vicugna pacos	94	0
Homo sapiens	19	557
Camelus dromedarius	26	0
Camelidae	10	0
None	9	15
Synthetic construct	16	6
Mus musculus	1	250
Lama pacos	1	0
Camelidae mixed library	5	0
Lama	5	0
Camelus bactrianus	5	0
Norovirus	1	0
Unidentified	1	2
Rattus norbegicus	0	14
Oryctolagus cuniculus	0	11
Pan troglodytes	0	1
Synthetic	0	1
Gallus gallus	0	2
Mus	0	2
Mus musculus, Homo sapiens	0	8
Influenza A virus	0	2
Homo sapiens, synthetic construct	0	2
Rattus	0	2
Macaca mulatta	0	7
Cricetulus migratorius	0	3
Rattus rattus	0	2
Other sequences	0	1

Table A.3: The most common framework residues observed in the paratope for at least 10% of nanobodies (IMGT numbering). Only positions 66 and 55 are observed in the paratope for more than 10% of the antibodies. Positions in bold tend to lie in the VH-VL interface [238].

Position	**Region**	**Observed percentage Nbs**	**Observed percentage Abs**
66	FR3	50.4	49.0
52	FR2	32.6	6.00
55	FR2	27.2	34.0
42	FR2	24.1	0.00
50	FR2	17.4	0.00
118	FR4	15.9	0.10
69	FR3	12.8	5.50
67	FR3	12.8	3.40
40	FR2	10.4	4.80
2	FR1	10.1	4.80

B

PLAbDab-nano: a database of camelid and shark nanobodies from patents and literature

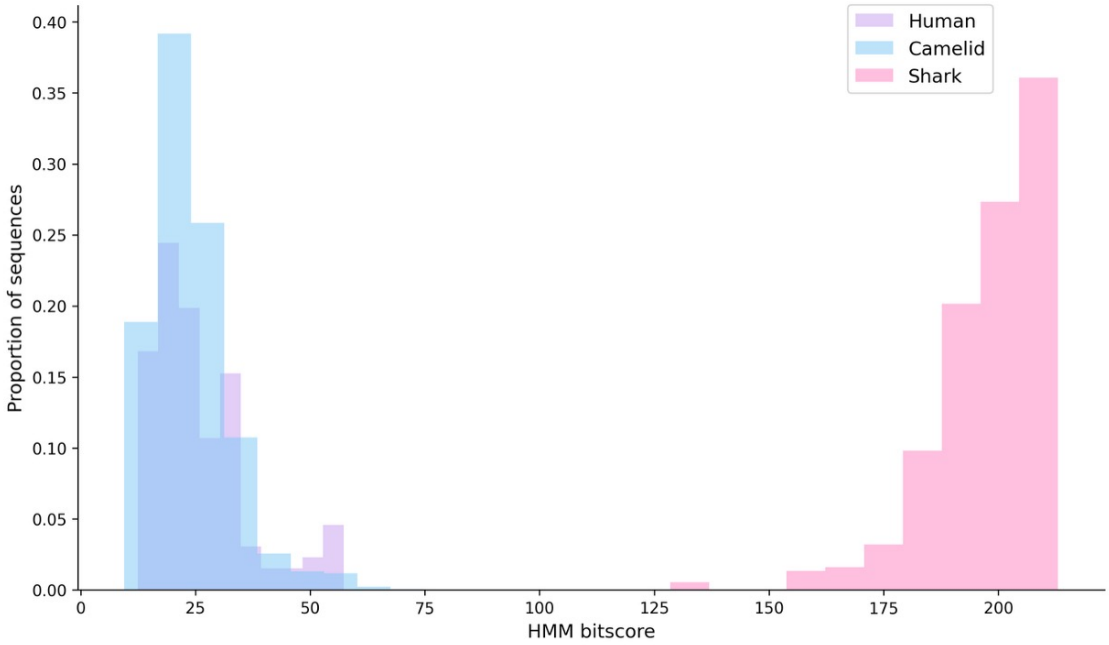


Figure B.1: A Hidden Markov Model (HMM) built from VNAR sequences provides a method to distinguish between VNAR and non-VNAR sequences. A bitscore threshold of 100 was chosen above which sequences are classified as VNARs. This threshold is based on the above distributions of HMM bitscores for non-redundant sets of the 377 shark VNAR sequences used to build the HMM, versus 131 human VH and 3182 camelid VH/VHH sequences.

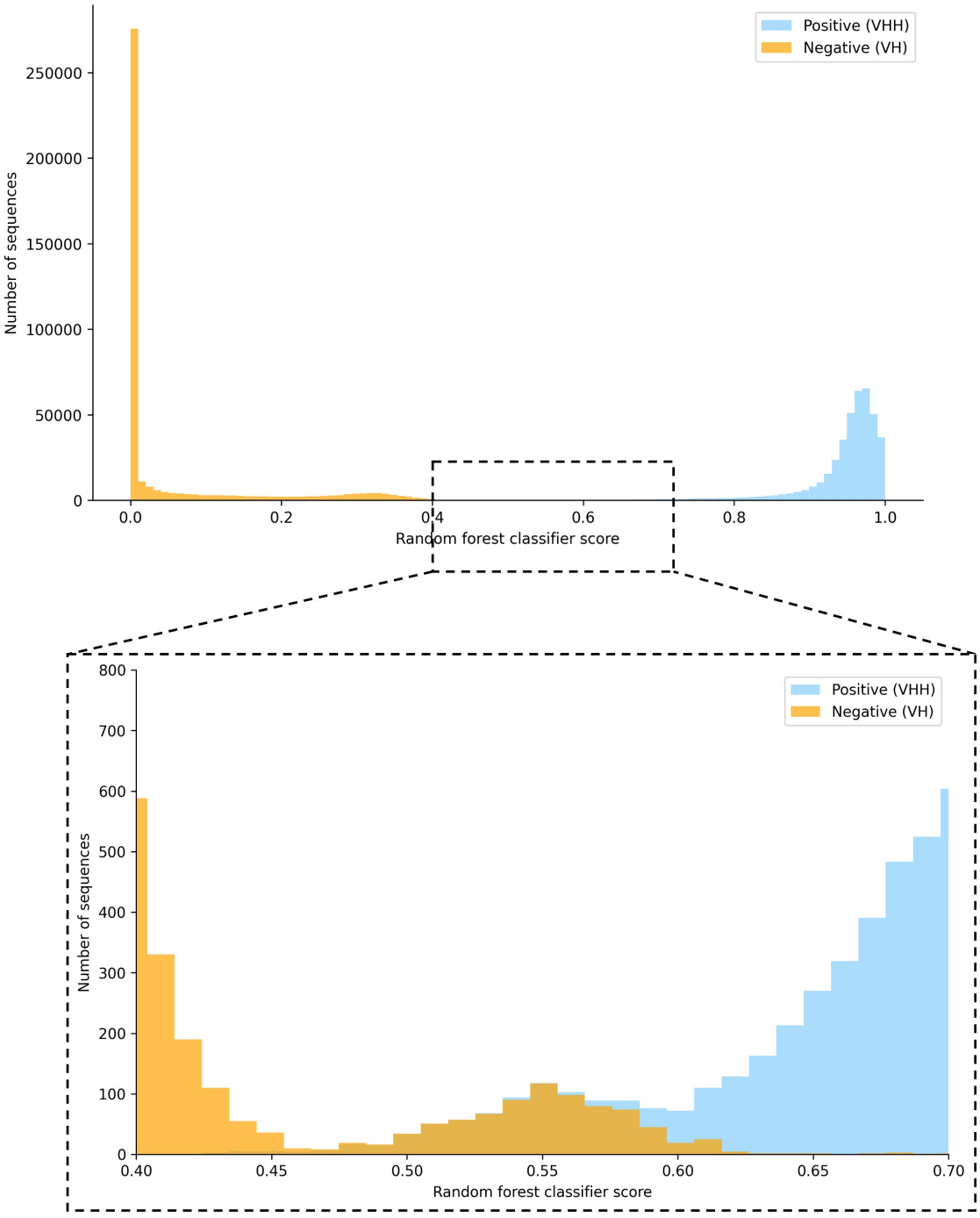


Figure B.2: Distributions of positive (VHH) and negative (VH) data were used to identify a rational threshold for the VH/VHH classifier, which was used to include or exclude entries from GenBank in PLAbDab-nano. To generate the first iteration of our database, the threshold was set at 0.45. This was to minimise the loss of true nanobodies that share a high sequence similarity with VH sequences, found in the overlap between the positive and negative data distributions. Updates will be conducted at a more conservative threshold of 0.6 to minimise the need for manual curation after running the automated update pipeline.

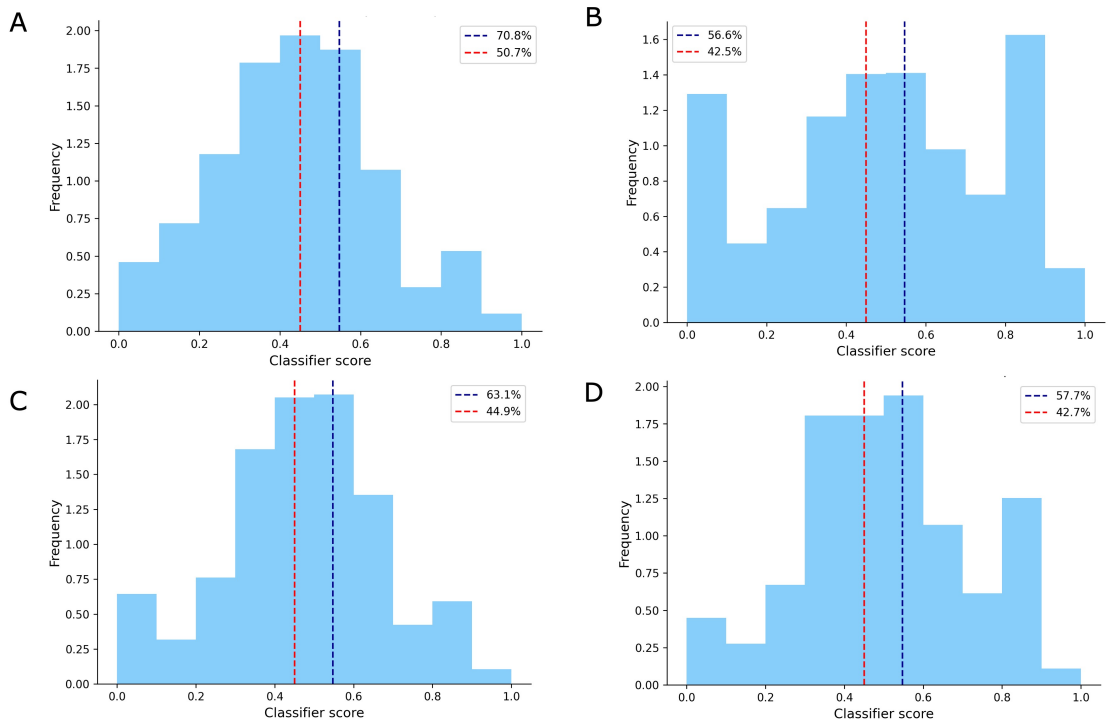


Figure B.3: Distributions of classifier scores for sequences from the INDI database [156], split by source into data from (**A**) patents, (**B**) GenBank, (**C**) structures, and (**D**) literature. Entries with scores to the left of the dark blue dashed lines (threshold = 0.55, calculated using Youden's J-statistic) and red dashed lines (where the threshold was relaxed to 0.45 to account for high VH and VHH sequence similarity) would not pass the filtering methods used to generate PLAbDab-nano, but are included in the INDI database. The values given in the figure legends next to the dashed lines indicate the percentage of sequences that would fail at that threshold.

Table B.1: Breakdown of total sequences in PLAbDab-nano by type of nanobody and source.

Type	GenBank patents	GenBank literature	SAbDab	TheraSAbDab	Total
VHH	1334	682	2395	41	4452
VNAR	416	40	0	0	456
All	1750	722	2395	41	4908

Table B.2: VH sequences collected from GenBank and patents for the INDI database [156] may be misclassified as nanobodies. Sources described in rows 1-7 do not contain nanobodies. Row 8 describes an example where VH sequences are given and these are able to function as nanobodies.

	GenBank/patent ID	**Number of sequences**	**Title**
1	HE653038-HE653050, HE653053, HE653054, HE653057, HE653059-HE653070, HE653072, HE653078, HE653079, HE653083, HE653085, HE653086, HE653090, HE653091	35	Analysis of heavy and light chain sequences of conventional camelid antibodies from Camelus dromedarius and Camelus bactrianus species
2	AJ245151, AJ245168-AJ245173, AJ245178-AJ245198	25	Camel heavy-chain antibodies: diverse germline VHH and specific mechanisms enlarge the antigen-binding repertoire
3	US11530271	13	CD70 Combination Therapy
4	WO2013064701	68	Specific antibodies and their isolation with anti-idiotypic antibodies
5	US20220073604	105	IL-6 binding molecules
6	US20220177604	20	Anti-IgE antibodies
7	EP3061768	27	Camelid derived antigen binding polypeptides comprising a VH and VL chain
8	AF442089, AF442907-AF442929, AF442931-AF442934, AF442936, AF442938-AF442944, AF442946, AF442947, AF442949	38	Selection by phage display of llama conventional V(H) fragments with heavy chain antibody V(H)H properties

C

The Therapeutic Nanobody Profiler: characterising and predicting nanobody developability to improve therapeutic design

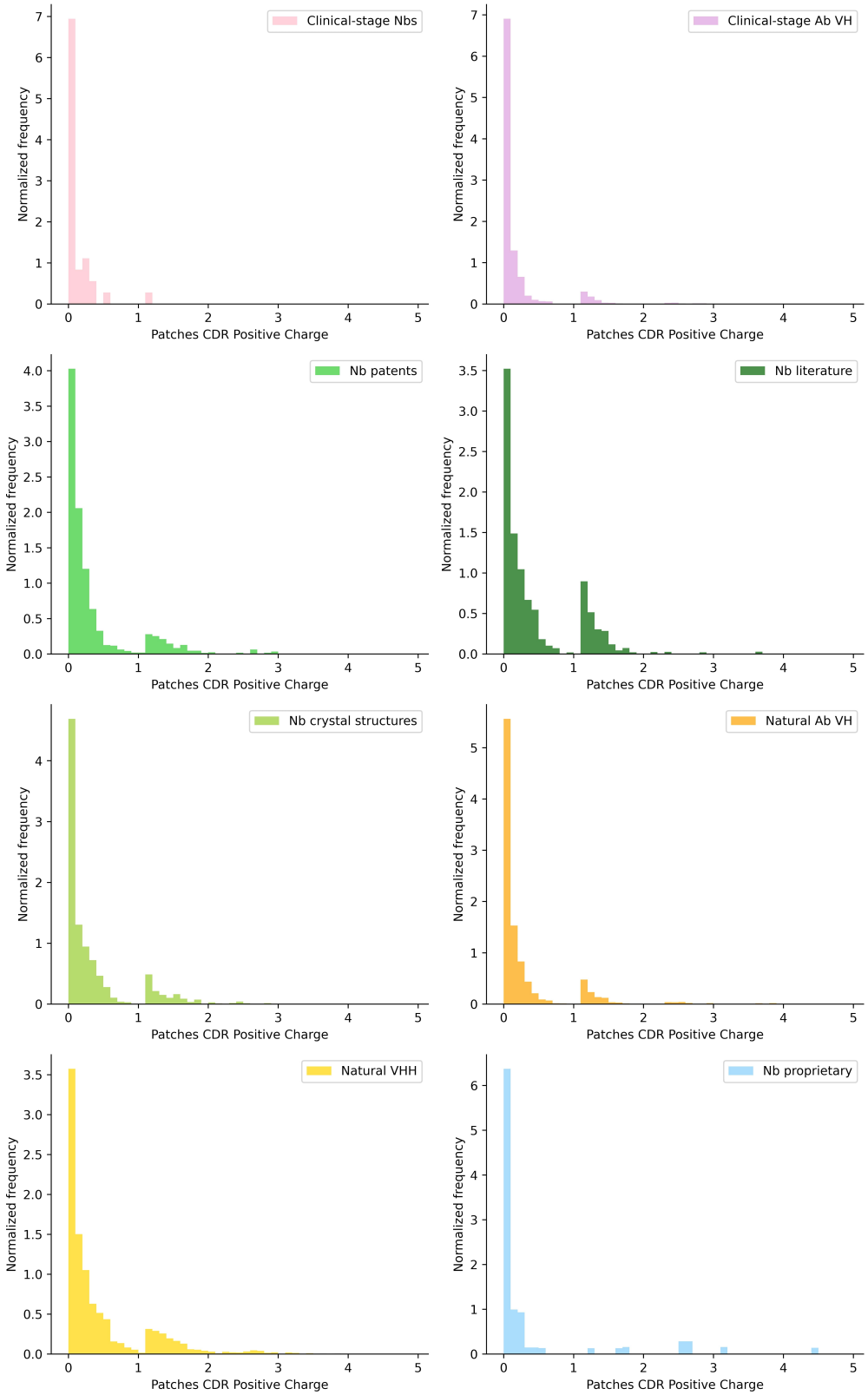


Figure C.1: Distributions of scores for patches of positive charge across all of our datasets, determined by sampling with replacement all datasets (aside from the 36 clinical-stage nanobodies), taking 300 samples with a sample size of 36.

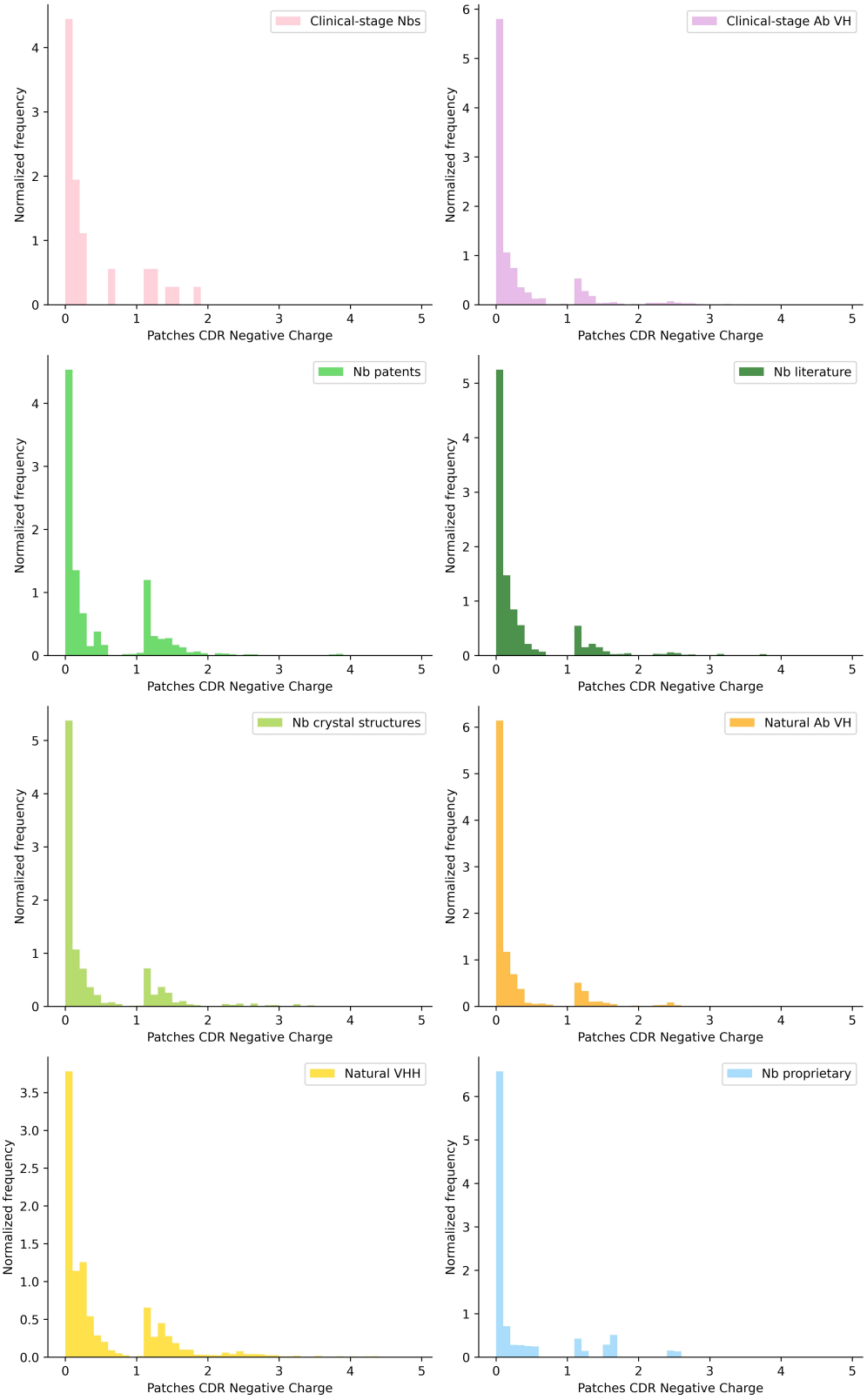


Figure C.2: Distributions of scores for patches of negative charge across all of our datasets, determined by sampling with replacement all datasets (aside from the 36 clinical-stage nanobodies), taking 300 samples with a sample size of 36.

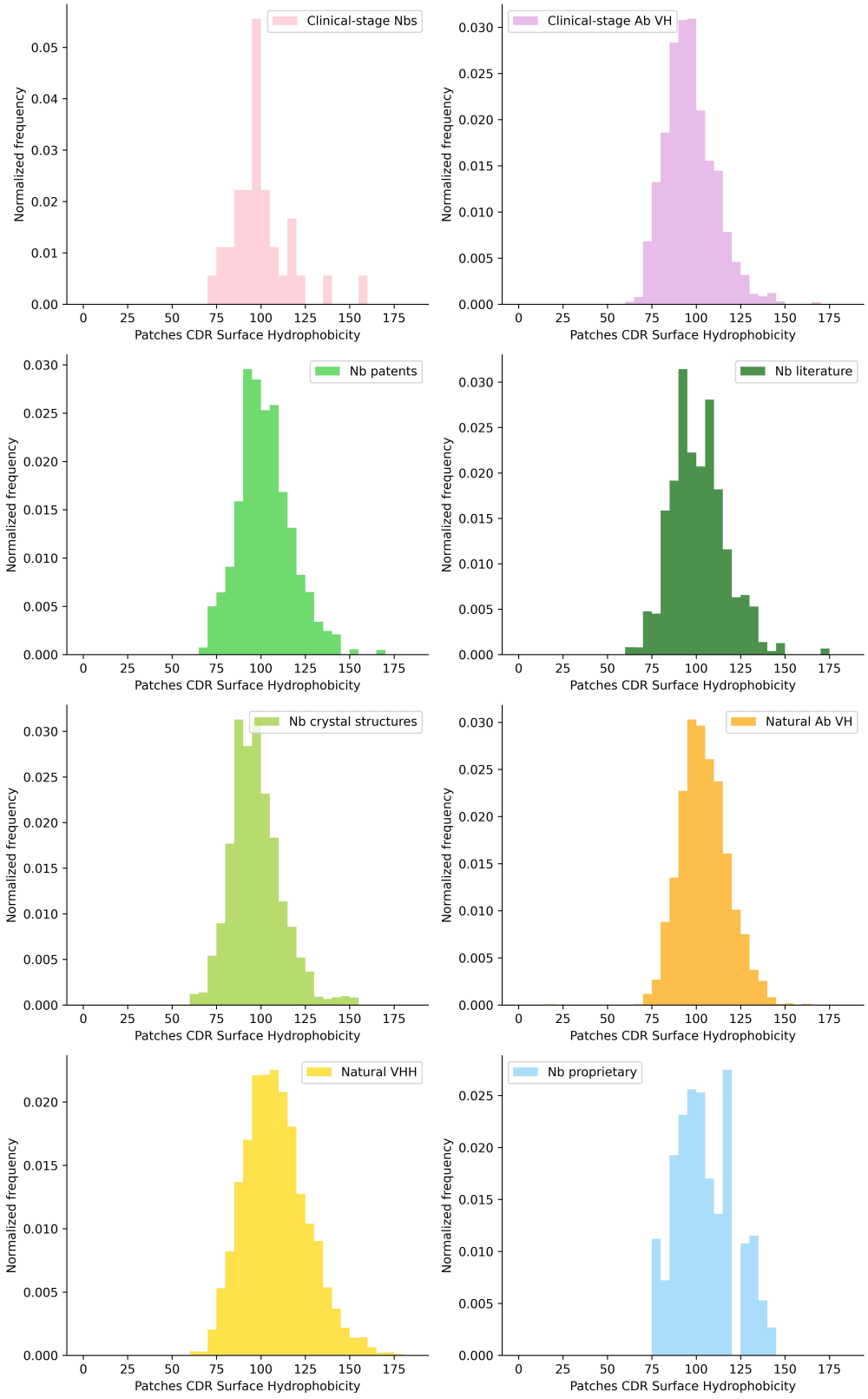


Figure C.3: Distributions of scores for patches of surface hydrophobicity across all of our datasets, determined by sampling with replacement all datasets (aside from the 36 clinical-stage nanobodies), taking 300 samples with a sample size of 36.

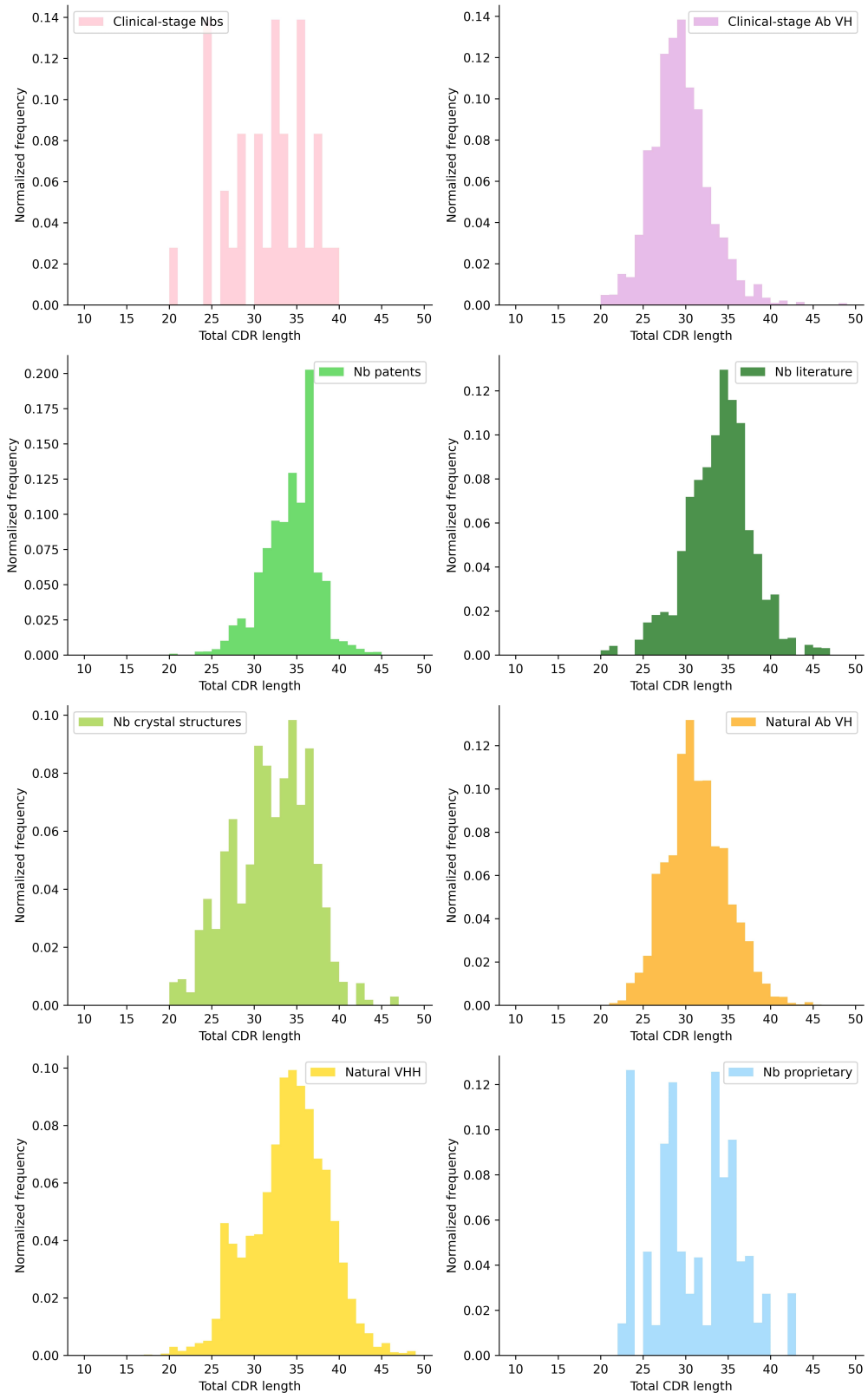


Figure C.4: Distributions of total CDR length across all of our datasets, determined by sampling with replacement all datasets (aside from the 36 clinical-stage nanobodies), taking 300 samples with a sample size of 36.

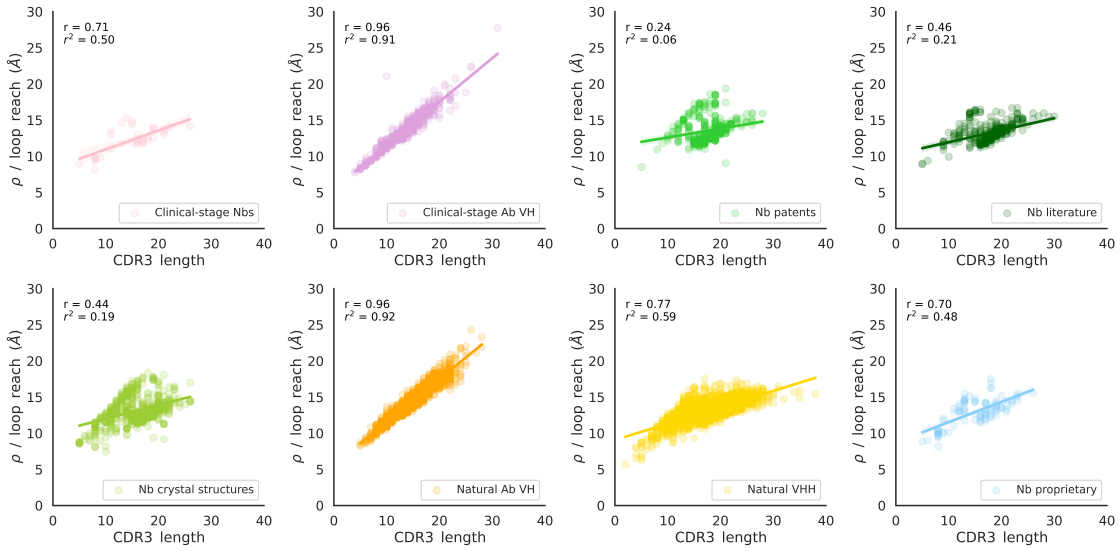


Figure C.5: Correlations between the reach and length of the CDR3 loop for the different datasets analysed in this study show divergent trajectories for nanobody data, where an increase in length does not translate to greater reach. This relationship can be described by loop compactness.

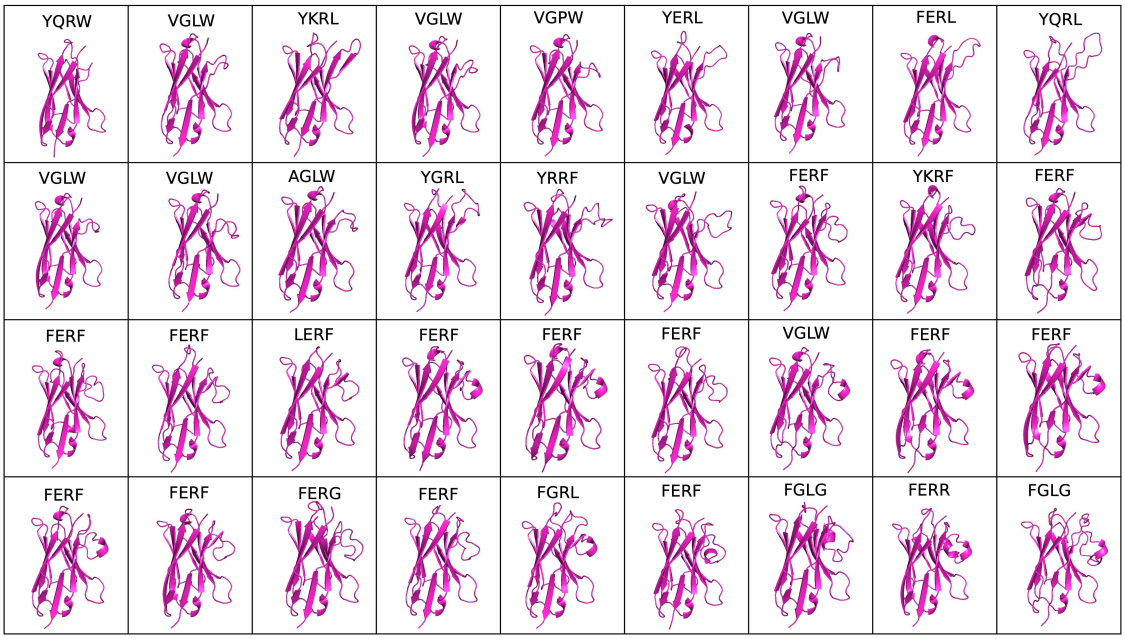


Figure C.6: Predicted structures for the 36 clinical-stage nanobody sequences in order of increasing compactness (top to bottom, left to right), annotated with their nanobody tetrad motif (amino acid identities at IMGT positions 42, 49, 50 and 52). The CDR3 loop is positioned in the top right-hand corner for each structure. As compactness increases, so does CDR3 length and conformational complexity.

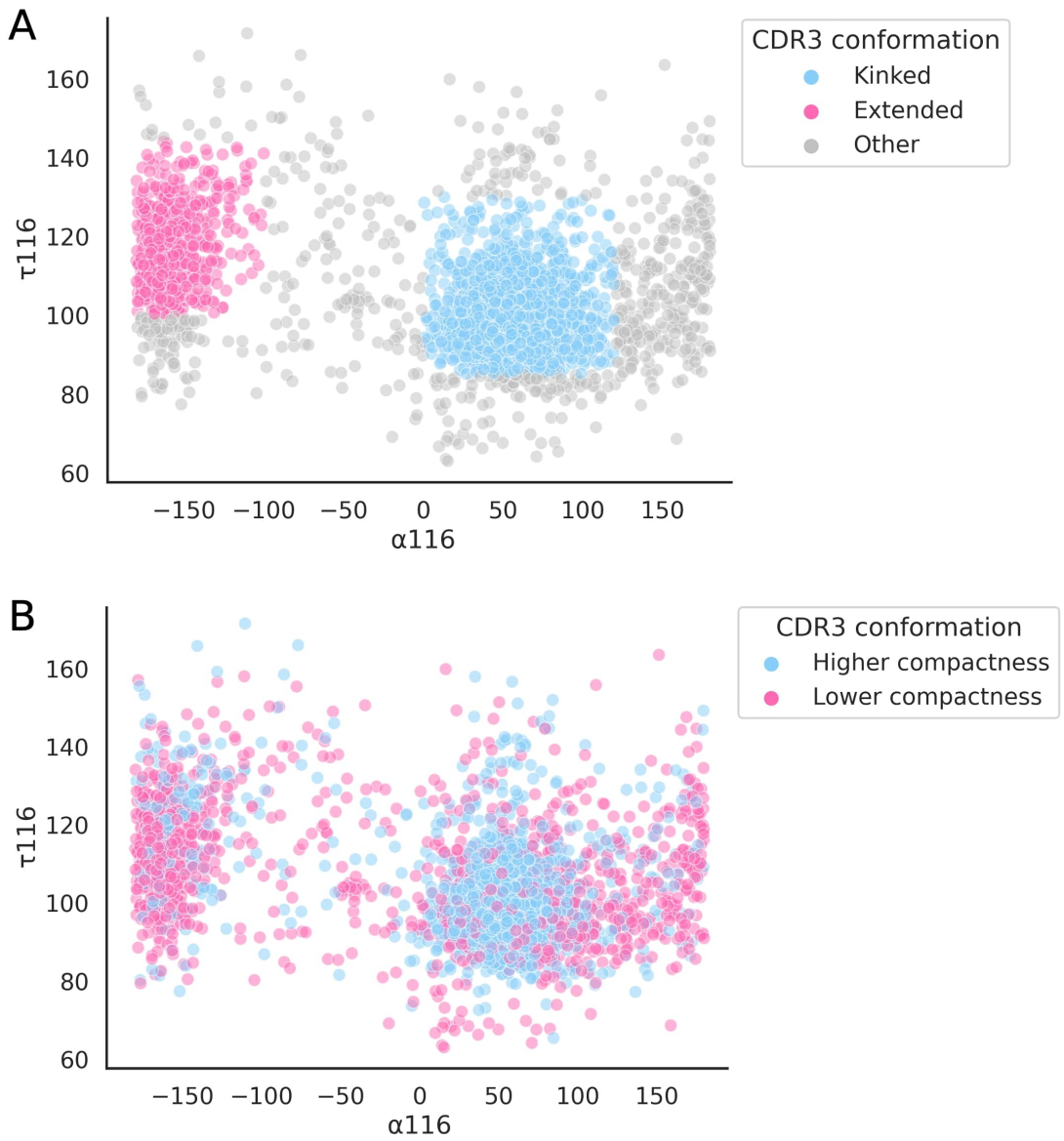


Figure C.7: CDR3 conformations can be classified as kinked or extended according to the α and τ angles between backbone atoms around IMGT position 116 in the CDR3 loop [21, 38]. (**A**) Classification of the CDR3 conformation for all nanobodies from our datasets combined shows that many do not fall into either category using this method. (**B**) Alternatively, the CDR3 conformation may be described using compactness. Using the same method to cluster the nanobodies on CDR3 angles, divided into groups of higher and lower compactness (using a threshold score of 1.25 as per **Figure 4.3**), shows the limitations of the classification method given the broad conformational diversity.

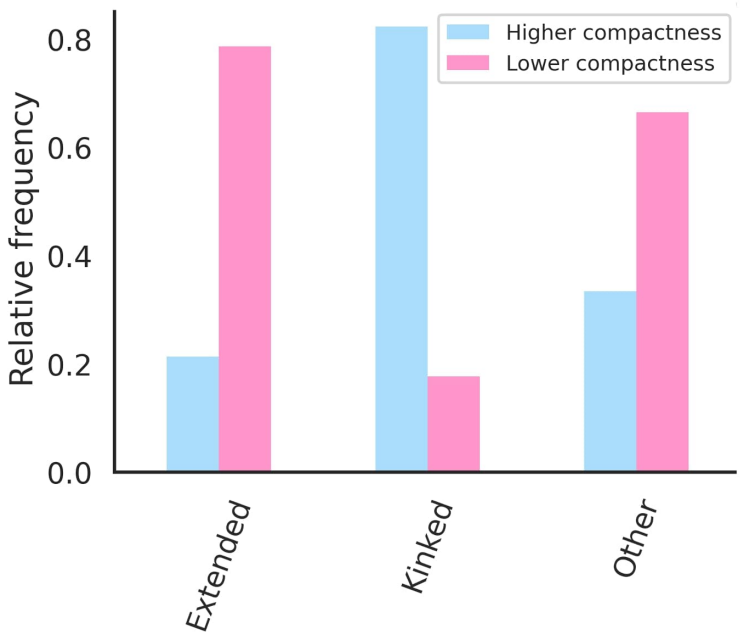


Figure C.8: To assess the agreement between compactness and the kinked/extended classifications, the data was split into loops of lower versus higher compactness, again using a threshold of 1.25 (**Figure 4.3**). Loops with lower compactness closely align with those classified as extended, while those with higher compactness correspond to kinked loops, showing consistency between the compactness measure and the classification method in describing CDR3 conformation.

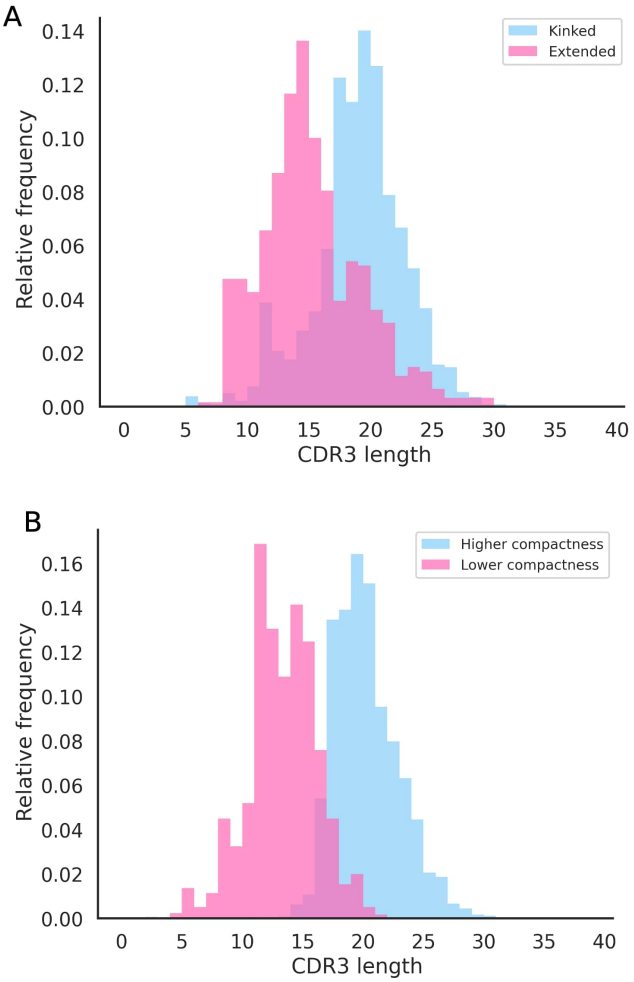


Figure C.9: The bimodal distribution of compactness can be used to separate nanobodies into two structural subtypes. We find that these demonstrate distinguishable profiles of secondary features associated with the CDR3 loop structure, which are consistent across descriptors for CDR3 conformation. When comparing CDR3 loop length, (**A**) CDR3 loops with a kinked conformation tend to be longer in terms of number of residues than extended conformations and (**B**) CDR3 loops with a higher compactness score tend to be longer than loops of lower compactness.

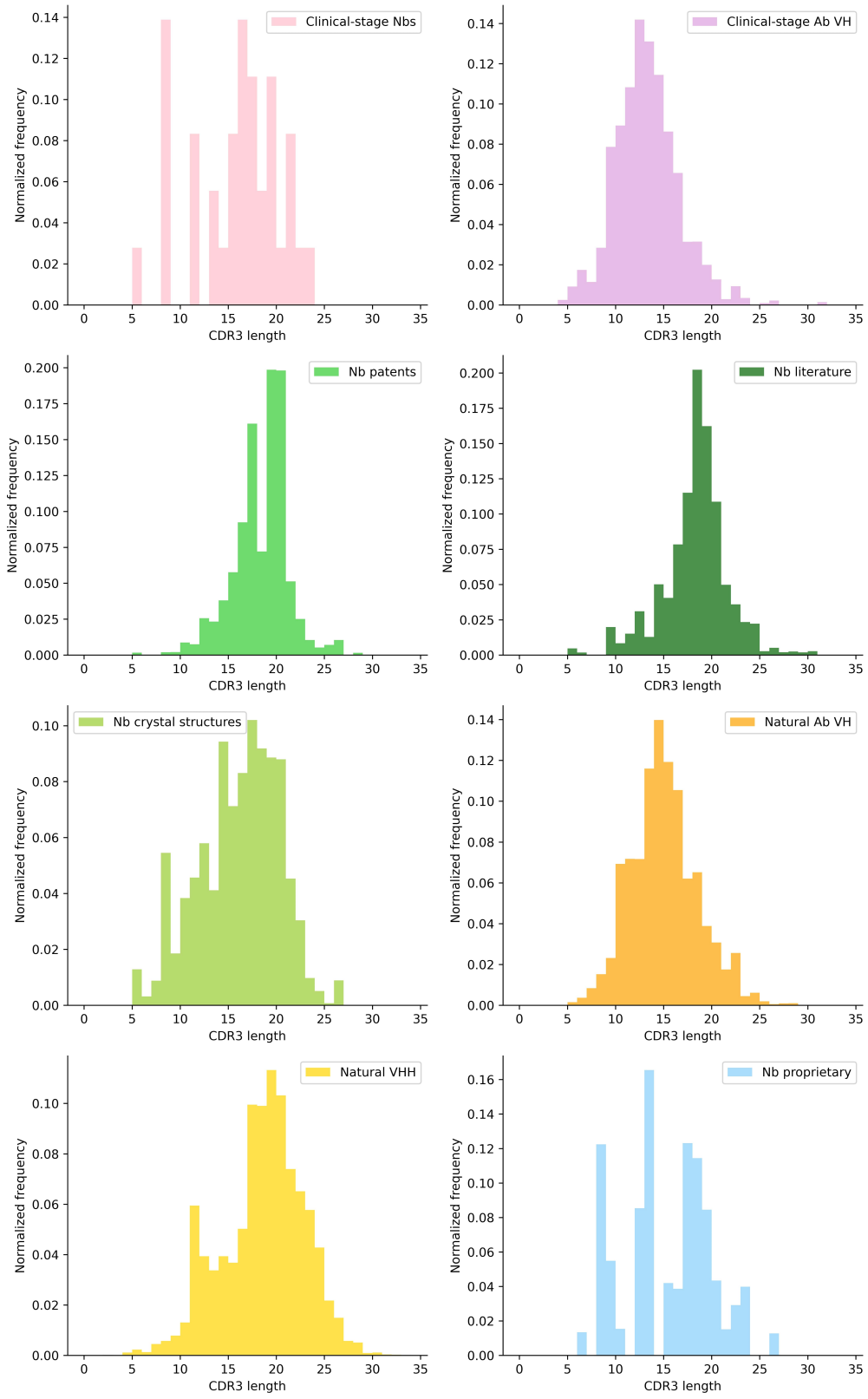


Figure C.10: Distributions of CDR3 length across all of our datasets, determined by sampling with replacement all datasets (aside from the 36 clinical-stage nanobodies), taking 300 samples with a sample size of 36.

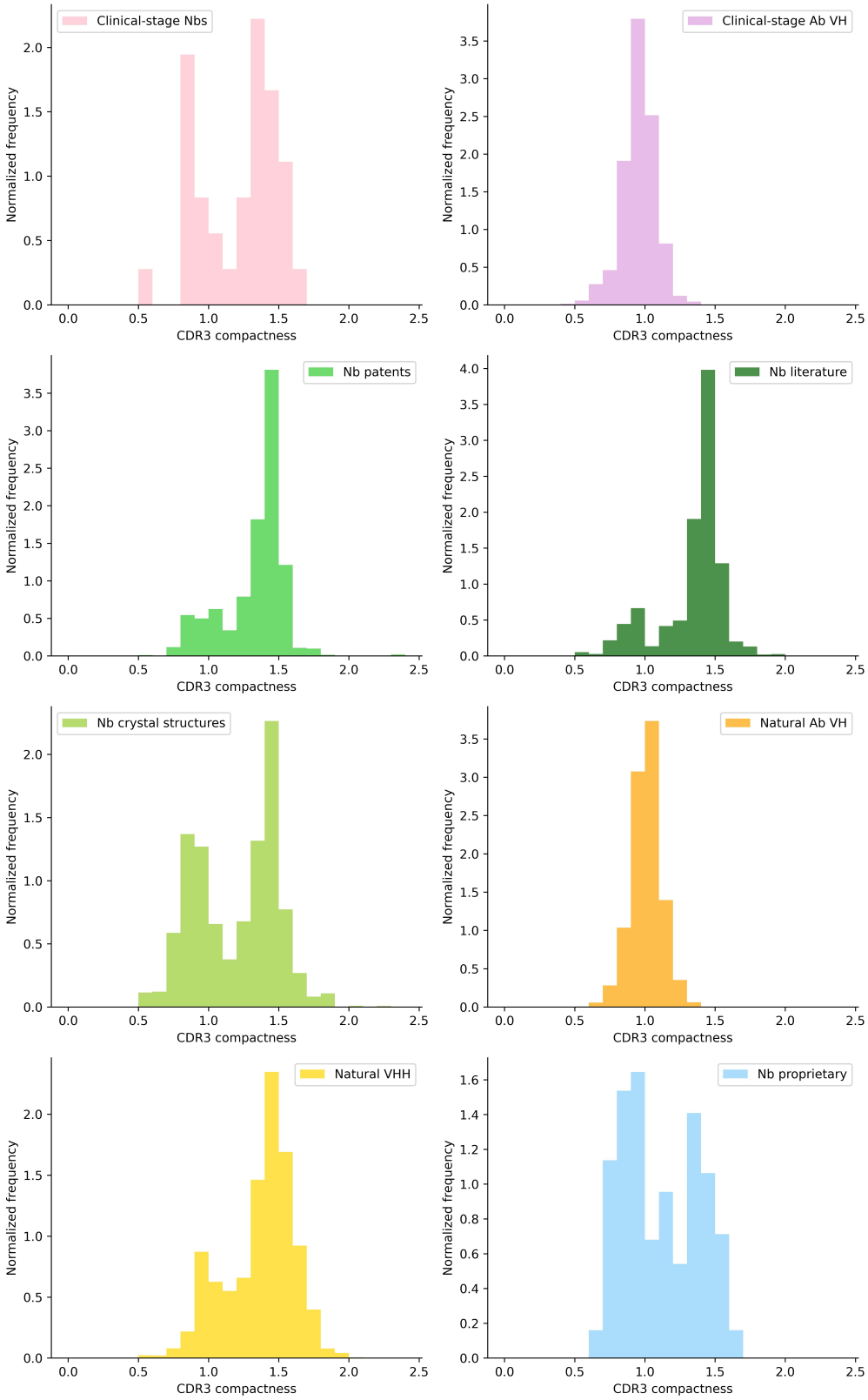


Figure C.11: Distributions of CDR3 compactness across all of our datasets, determined by sampling with replacement all datasets (aside from the 36 clinical-stage nanobodies), taking 300 samples with a sample size of 36.

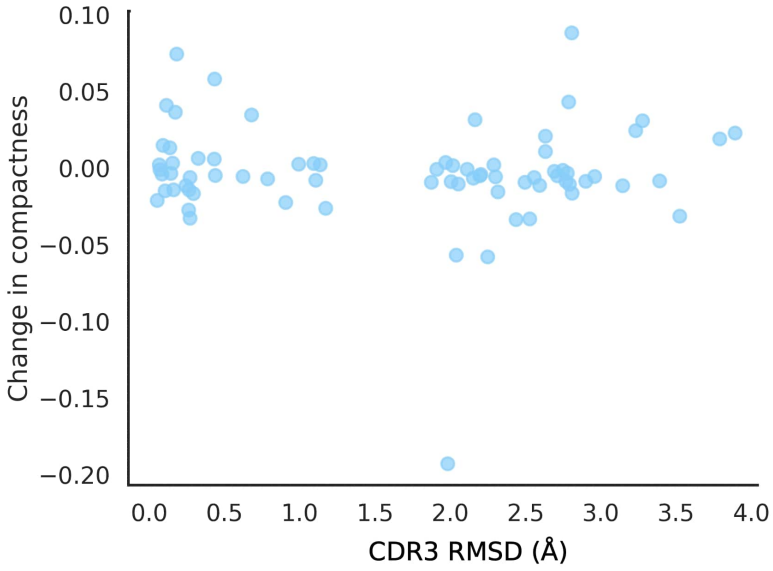


Figure C.12: Change in compactness is minimal between bound and unbound pairs of nanobody CDR3 loop crystal structures, and does not show a positive correlation with the RMSD between the bound and unbound CDR3 loops.

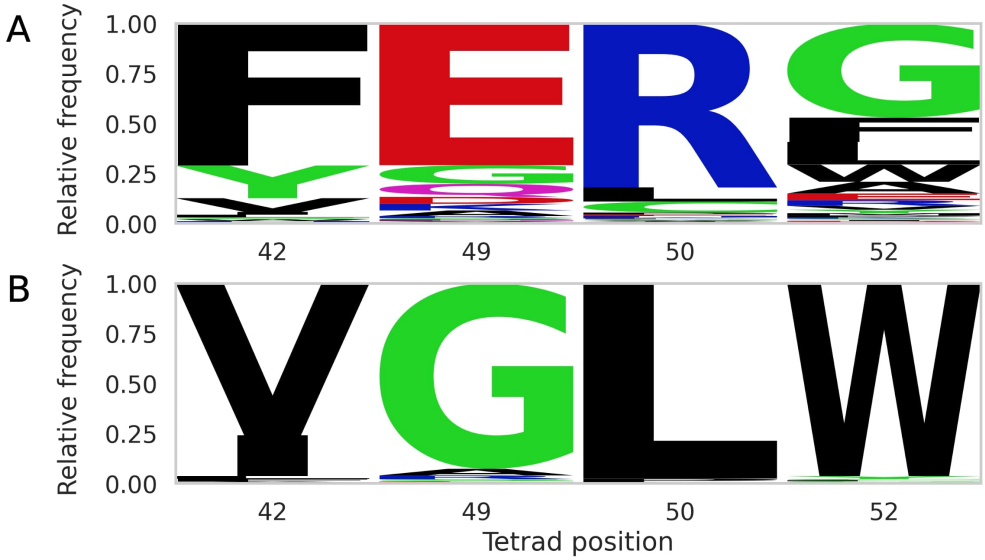


Figure C.13: Amino acid frequencies across the nanobody tetrad (positions 42, 49, 50 and 52 by IMGT numbering) in our (**A**) nanobody versus (**B**) antibody VH datasets shows that FERG is the most common nanobody motif, whilst VGLW is the most common in antibody VH chains.

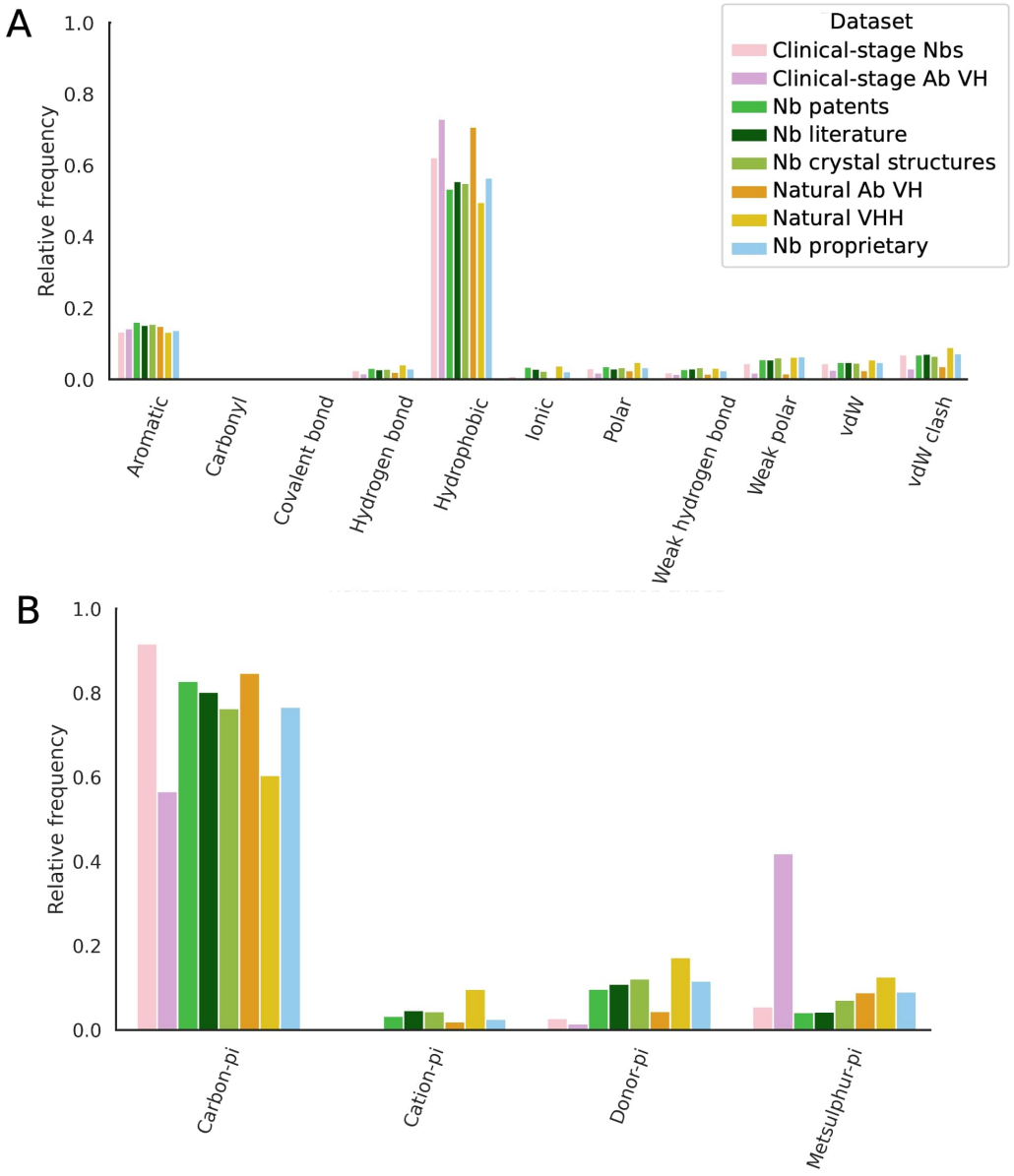


Figure C.14: Comparing the relative frequencies of atom-atom and atom-plane interactions between pairs of residues in the CDR3 loop and the nanobody tetrad positions shows that for (**A**) atom-atom interactions, hydrophobic, and aromatic interactions were the most common, and for (**B**) atom-plane interactions, carbon-π interactions were the most common.

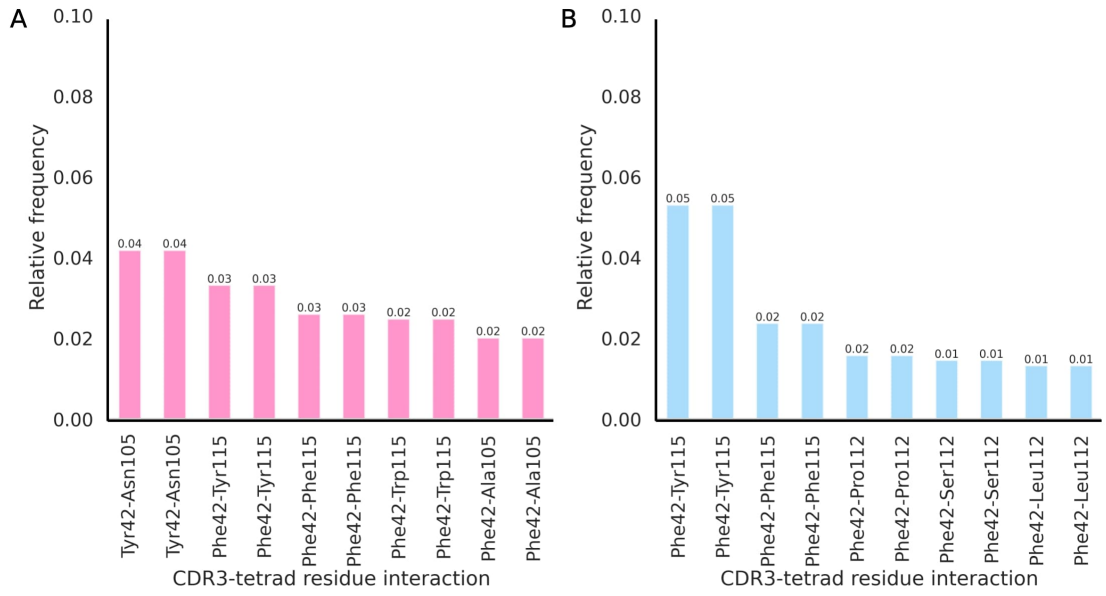


Figure C.15: The top 10 most common interacting CDR3-tetrad residue pairs in decreasing order of relative frequency, across all of our nanobody datasets for (**A**) those with CDR3 loops of lower compactness (that tend to extend out from the framework) and (**B**) loops of higher compactness (that tend to fold over the FR2 region).

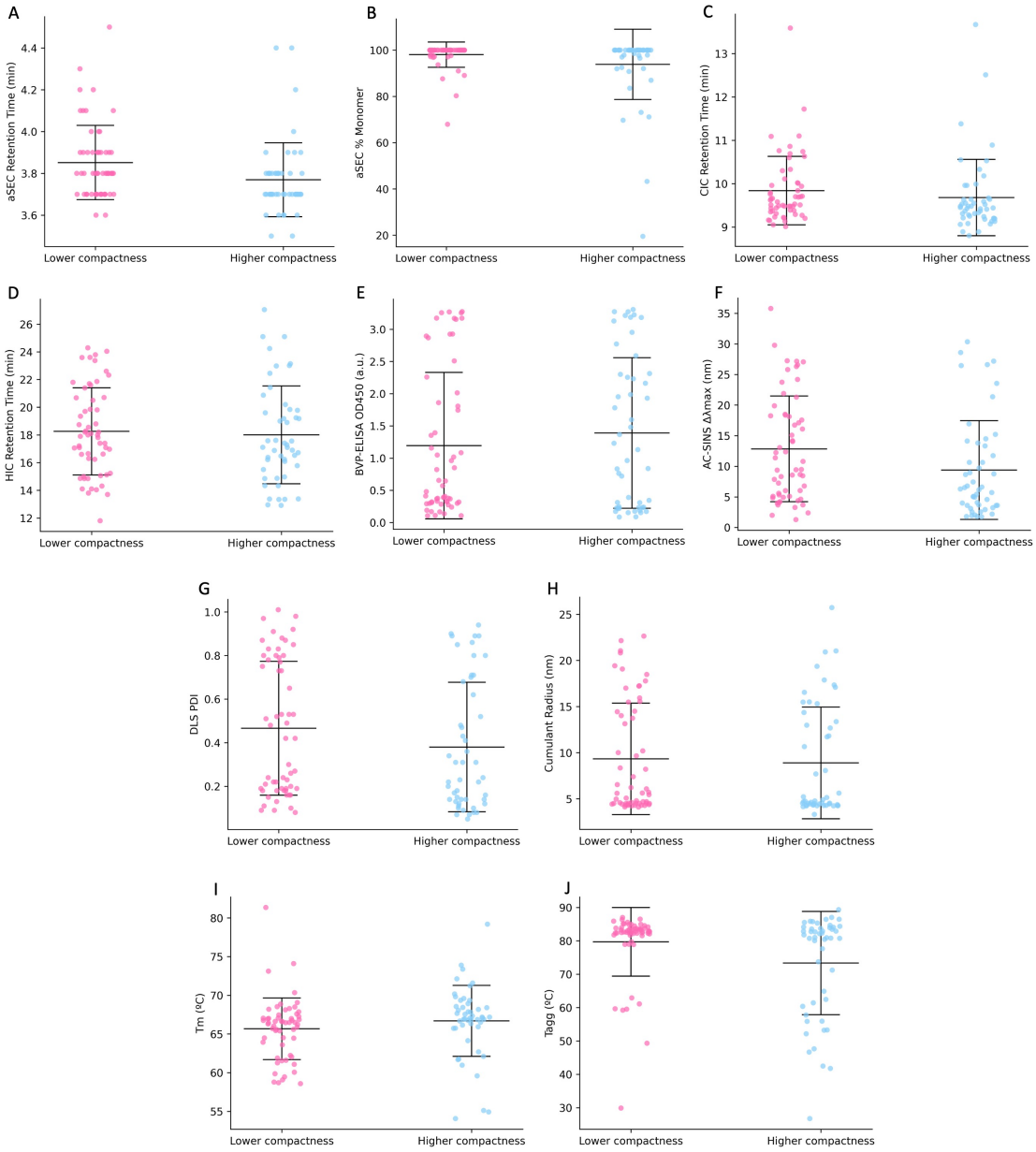


Figure C.16: The ranges of experimental assay values split into nanobodies with CDR3 loops of higher and lower compactness at a compactness score threshold of 1.25 as per **Figure 4.3**. (**A**) aSEC retention time (**B**) aSEC % monomer (**C**) CIC retention time (**D**) HIC retention time (**E**) BVP-ELISA OD450 (**F**) AC-SINS $\Delta\lambda$max (**G**) DLS PDI (**H**) Cumulant radius (**I**) Tm (melting temperature) (**J**) Tagg (aggregation temperature).

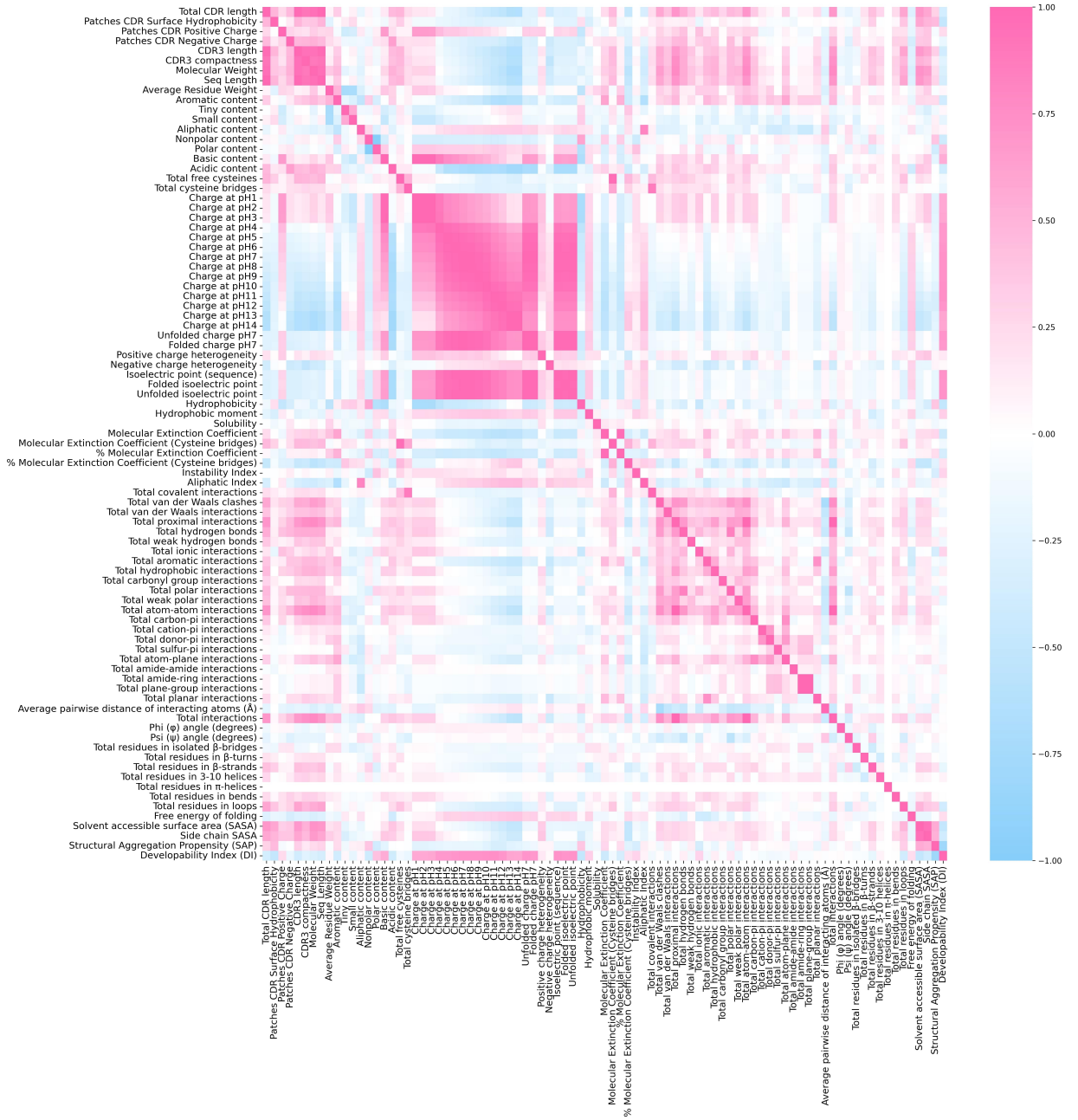


Figure C.17: A heatmap of the Spearman's rank correlations calculated pairwise for our extended list of computational descriptors, comparing across all of our nanobody datasets combined. A deeper pink shows a strong positive correlation, and a deeper blue shows a strong negative correlation.

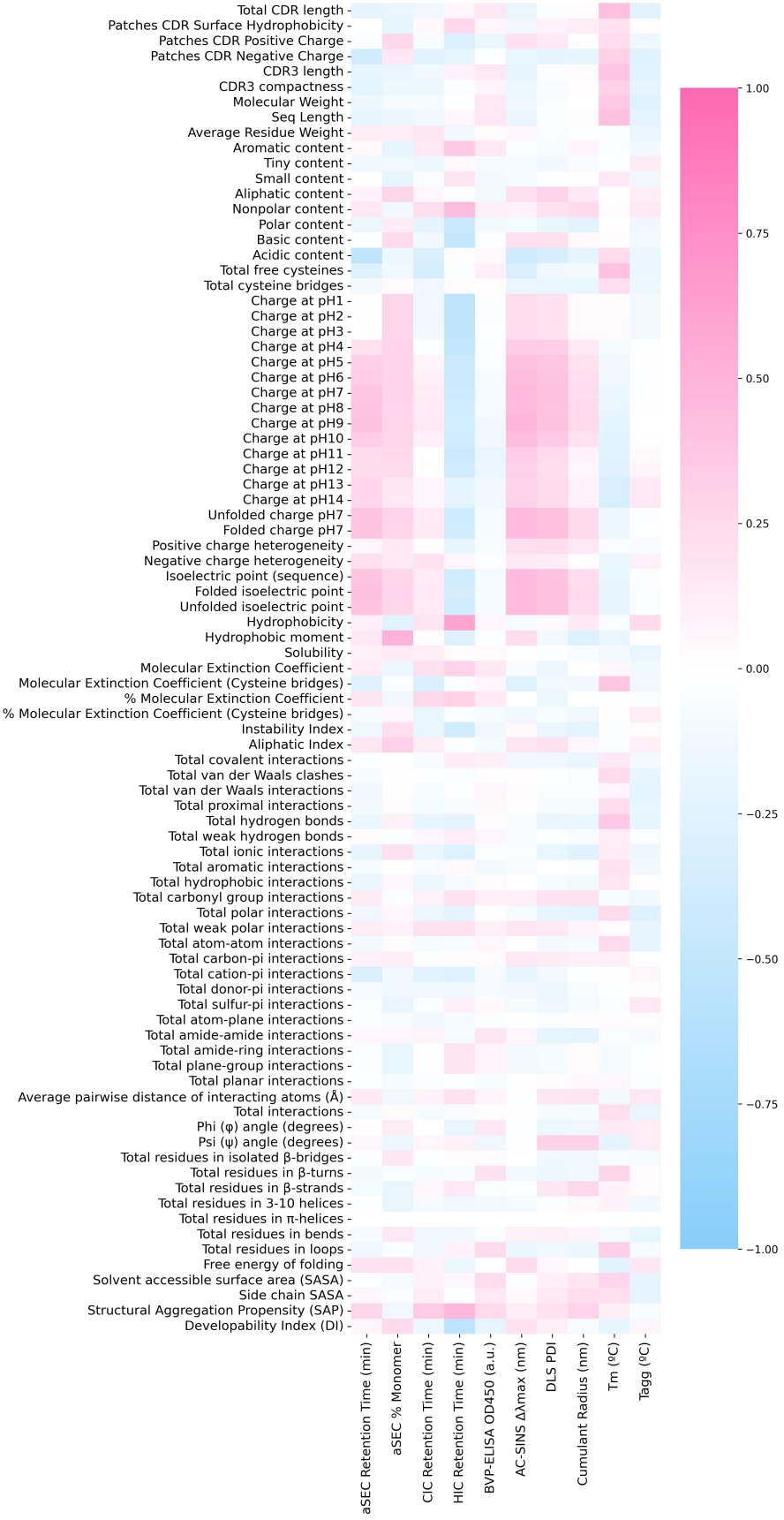


Figure C.18: A heatmap of the Spearman's rank correlations for each computational descriptor with each *in vitro* assay, for the 108 clinical-stage and proprietary nanobodies that were experimentally tested. A deeper pink shows a strong positive correlation, and a deeper blue shows a strong negative correlation.

Table C.1: The pairs of PDB IDs used to compare CDR3 compactness between bound and unbound nanobodies.

	Unbound PDB	Bound PDB		Unbound PDB	Bound PDB		Unbound PDB	Bound PDB
1	8f8v	8f8w	26	5tp3	5toj	51	7np9	7p2d
2	8f8v	8f8x	27	4krn	4kro	52	6dbg	6dbf
3	8gz6	8gz5	28	5h8d	5h8o	53	6dba	6dbg
4	7sah	6lr7	29	6y0e	6xzf	54	6dba	6dbf
5	7vpy	7vq0	30	3r0m	7ri2	55	6apq	4w2o
6	4b50	7aej	31	3r0m	3rjq	56	5wb1	4xt1
7	6u53	6u54	32	5sv4	5sv3	57	5wb2	4xt1
8	4w2p	4w2q	33	4s11	4s10	58	5hdo	5hgg
9	5vm4	5vm6	34	4s11	6h1f	59	5lz0	6ey0
10	7kgl	7kgk	35	4qgy	4y7m	60	6u50	6u51
11	7mfv	7kgk	36	4m3j	4m3k	61	6u50	6u52
12	7lvu	7lvw	37	5m2w	5mwn	62	7om5	7om4
13	5e7b	5e7f	38	5m2w	5m30	63	7omt	7omm
14	6fys	6fyw	39	7kkj	7kkl	64	2xa3	7r73
15	6ksn	7d5p	40	6ck8	6fyt	65	6cwk	6cwg
16	6ksn	7d5q	41	6z1z	6z20	66	6u14	6u12
17	3qxu	3qxt	42	5vlv	5vm0	67	6ru3	6ruv
18	7v0v	3k1k	43	3qxw	7rga	68	6ru3	7noz
19	7v0v	3ogo	44	3qxw	7rg7	69	2xxc	2xt1
20	1f2x	1g6v	45	3qxw	3qxv	70	2xxc	2xv6
21	6t2j	6sc5	46	5jmr	5jmo	71	2xxc	2xxm
22	6t2j	6sc9	47	6o89	6o8d	72	6jb9	1zvy
23	6t2j	6sc7	48	5vnv	5vnw	73	6jb9	6jb8
24	6t2j	6sc8	49	4dka	4dk6	74	5e03	5e5m
25	6t2j	6sc6	50	4dka	4dk3	75	5ivo	5ivn
						76	6apo	6app

Table C.2: The additional *in silico* descriptors calculated for our nanobody datasets, from Bashour *et al.* (2024) [260].

	Molecular				
1	Molecular Weight	29	Folded charge pH7	56	Total atom-atom interactions
2	Sequence Length	30	Positive charge heterogeneity	57	Total carbon-pi interactions
3	Average Residue Weight	31	Negative charge heterogeneity	58	Total cation-pi interactions
	Amino acid composition	32	Isoelectric point (sequence)	59	Total donor-pi interactions
4	Aromatic content	33	Folded isoelectric point	60	Total sulfur-pi interactions
5	Tiny content	34	Unfolded isoelectric point	61	Total atom-plane interactions
6	Small content	35	Hydrophobicity	62	Total amide-amide interactions
7	Aliphatic content	36	Hydrophobic moment	63	Total amide-ring interactions
8	Nonpolar content	37	Solubility	64	Total plane-group interactions
9	Polar content		Photochemical	65	Total planar interactions
10	Basic content	38	Molecular Extinction Coefficient	66	Average pairwise distance of interacting atoms (Å)
11	Acidic content	39	Molecular Extinction Coefficient (Cysteine bridges)	67	Total interactions
12	Total free cysteines	40	% Molecular Extinction Coefficient		Secondary structure
13	Total cysteine bridges	41	% Molecular Extinction Coefficient (Cysteine bridges)	68	Phi (ϕ) angle (degrees)
	Electrochemical		Stability	69	Psi (ψ) angle (degrees)
14	Charge at pH1	42	Instability Index	70	Total residues in isolated β-bridges
15	Charge at pH2	43	Aliphatic Index	71	Total residues in β-turns
16	Charge at pH3		Interactions	72	Total residues in β-strands
17	Charge at pH4	44	Total covalent interactions	73	Total residues in 3-10 helices
18	Charge at pH5	45	Total van der Waals clashes	74	Total residues in α-helices
19	Charge at pH6	46	Total van der Waals interactions	75	Total residues in bends
20	Charge at pH7	47	Total proximal interactions	76	Total residues in loops
21	Charge at pH8	48	Total hydrogen bonds		Thermodynamic
22	Charge at pH9	49	Total weak hydrogen bonds	77	Free energy of folding
23	Charge at pH10	50	Total ionic interactions		Solvent accessibility
24	Charge at pH11	51	Total aromatic interactions	78	Solvent accessible surface area (SASA)
25	Charge at pH12	52	Total hydrophobic interactions	79	Side chain SASA
26	Charge at pH13	53	Total carbonyl group interactions		Druggability
27	Charge at pH14	54	Total polar interactions	80	Structural Aggregation Propensity (SAP)
28	Unfolded charge pH7	55	Total weak polar interactions	81	Developability Index (DI)

Table C.3: Flag assignment and raw values for the TNP metrics and five *in vitro* assays across the 36 clinical-stage nanobody sequences.

Samples	CIC Retention Time (min)	HIC Retention Time (min)	BVP-ELISA OD450 (a.u.)	AC-SINS Δλmax (nm)	DLS PDI	Total CDR length	CDR3 length	CDR3 compactness	Patches CDR Surface Hydrophobicity	Patches CDR Positive Charge	Patches CDR Negative Charge
Brivekimig1_VHH1	9.25	13.35	1.95	26.45	0.41	33	16	1.35	81.36	0.26	0.22
Brivekimig1_VHH2	9.24	13.37	0.23	27.18	0.36	33	16	1.38	94.24	0.06	0.17
Brivekimig2_VHH1	9.38	16.60	3.16	7.25	0.09	24	8	0.83	106.99	0.00	0.00
Caplacizumab_VHH1	9.33	13.34	0.22	26.61	0.62	37	21	1.53	92.06	0.57	1.22
Enristomig_VHH1	9.23	15.23	0.11	4.38	0.26	20	5	0.56	79.36	0.05	0.00
Enristomig_VHH2	9.39	11.80	1.02	26.58	0.97	30	15	1.04	106.50	0.06	0.00
Envafolimab_VHH1	9.28	14.97	1.13	6.61	0.80	37	21	1.61	102.37	0.05	1.14
Erfonrilimab_VHH2	9.48	20.87	3.31	5.03	0.22	37	21	1.57	102.99	0.00	0.00
Gacovetug_VHH1	9.08	14.34	3.18	2.54	0.22	30	15	1.27	97.05	0.00	0.17
Gefurulimab_VHH1	9.67	15.50	0.24	28.59	0.89	32	17	1.42	97.04	0.06	0.18
Gefurulimab_VHH2	9.53	16.24	0.09	7.44	0.23	35	19	1.52	122.87	0.18	0.06
Gocatamig2_VHH1	9.70	18.74	0.29	9.43	0.73	26	11	0.82	138.68	0.00	1.11
Gontivimab_VHH1	9.37	19.24	1.93	6.15	0.17	35	19	1.36	99.40	0.04	0.11
Gontivimab_VHH2	9.37	19.18	0.72	5.71	0.14	35	19	1.36	91.42	0.04	0.11
Isecarosmab_VHH1	9.18	14.33	0.23	4.95	0.12	33	17	1.45	86.44	0.36	1.22
Isecarosmab_VHH2	9.51	17.81	0.33	6.03	0.16	24	8	0.80	96.49	0.00	0.00
Letolizumab_VHH1	9.49	16.97	1.39	9.48	0.87	27	11	1.00	95.13	0.27	0.28
Lofacimig_VHH1	9.31	18.90	3.27	2.92	0.10	39	23	1.49	155.47	0.00	0.00
Lofacimig_VHH2	8.89	19.84	2.30	1.79	0.31	35	20	1.32	119.68	0.00	1.53
Lunsekimig1_VHH1	10.63	17.06	3.25	27.24	0.51	32	16	1.23	73.40	0.00	0.00
Lunsekimig1_VHH2	9.96	16.88	3.23	15.21	0.10	35	19	1.39	116.89	0.00	0.00
Lunsekimig2_VHH1	9.20	18.28	0.24	3.97	0.18	26	11	0.85	116.19	0.00	0.05
Lunsekimig3_VHH1	8.89	13.38	0.17	2.04	0.14	36	18	1.48	97.22	0.10	1.88
Ozekibart_VHH1	9.44	15.78	0.17	14.49	0.86	32	16	1.29	79.67	0.21	0.00
Ozoralizumab_VHH1	9.69	14.04	0.27	29.78	0.08	24	8	0.98	83.65	0.10	0.00
Ozoralizumab_VHH2	9.65	17.21	0.19	5.09	0.21	24	8	0.85	111.02	0.00	0.00
Podentamig1_VHH1	9.31	15.12	3.28	4.21	0.53	28	14	0.92	86.31	0.02	0.00
Podentamig1_VHH2	9.43	17.09	0.11	18.5	0.13	24	8	0.83	98.59	0.00	0.00
Porustobart_VHH1	9.76	16.62	0.11	18.36	0.85	28	13	1.11	103.61	0.16	0.00
Produvofusp_VHH1	9.07	17.10	1.23	3.61	0.12	38	22	1.57	90.60	0.00	0.29
Rimteravimab_VHH1	9.06	16.9	0.15	3.99	0.14	34	18	1.43	89.51	0.00	0.65
Sonelokimab1_VHH1	9.41	17.11	0.09	11.74	0.71	32	16	1.32	104.10	0.07	0.24
Sonelokimab2_VHH1	9.55	16.38	0.83	3.23	0.05	31	17	1.35	97.97	0.23	1.45
Tarperprumig_VHH1	9.96	16.72	0.22	30.35	0.94	32	17	1.42	99.13	0.06	0.14
Tarperprumig_VHH2	9.22	14.86	0.17	3.68	0.24	28	13	0.89	99.60	0.30	0.14
Vobarilizumab_VHH1	9.48	17.09	3.17	35.79	0.42	30	15	1.01	85.49	1.19	0.62

References

[1] Janeway, C.A. et al. *Immunobiology*. 6th edition. Garland Science, 2006.

[2] Tonegawa, S. Somatic generation of antibody diversity. *Nature* **302**, 575–581 (1983).

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

[4] Chi, X., Li, Y., and Qiu, X. V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. *Immunology* **160**, 233–247 (2020).

[5] Briney, B.S. et al. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **566**, 393–397 (2019).

[6] Greiff, V. et al. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Medicine* **7**, 49 (2015).

[7] Nutt, S.L. et al. The generation of antibody-secreting plasma cells. *Nature Reviews. Immunology* **15**, 160–171 (2015).

[8] Victora, G.D. and Nussenzweig, M.C. Germinal Centers. *Annual Review of Immunology* **40**, 413–442 (2022).

[9] Young, C. and Brink, R. The unique biology of germinal center B cells. *Immunity* **54**, 1652–1664 (2021).

[10] Odegard, V.H. and Schatz, D.G. Targeting of somatic hypermutation. *Nature Reviews. Immunology* **6**, 573–583 (2006).

[11] Mikocziova, I., Greiff, V., and Sollid, L.M. Immunoglobulin germline gene variation and its impact on human disease. *Genes & Immunity* **22**, 205–217 (2021).

[12] Chaudhuri, J. and Alt, F.W. Class-switch recombination: interplay of transcription, DNA deamination and DNA repair. *Nature Reviews Immunology* **4**, 541–552 (2004).

[13] Lu, L.L. et al. Beyond binding: antibody effector functions in infectious diseases. *Nature Reviews Immunology* **18**, 46–61 (2018).

[14] Lo Nigro, C. et al. NK-mediated antibody-dependent cell-mediated cytotoxicity in solid tumors: biological evidence and clinical perspectives. *Annals of Translational Medicine* **7**, 105 (2019).

[15] Kamen, L. et al. A novel method for determining antibody-dependent cellular phagocytosis. *Journal of Immunological Methods* **468**, 55–60 (2019).

[16] Wang, B. et al. Regulation of antibody-mediated complement-dependent cytotoxicity by modulating the intrinsic affinity and binding valency of IgG for target antigen. *mAbs* **12**, 1690959 (2019).

[17] Vidarsson, G., Dekkers, G., and Rispens, T. IgG Subclasses and Allotypes: From Structure to Effector Functions. *Frontiers in Immunology* **5**, 520 (2014).

[18] Briney, B.S. and Crowe, J.E. Secondary mechanisms of diversification in the human antibody repertoire. *Frontiers in Immunology* **4**, 42 (2013).

[19] Arnaout, R.A. Specificity and overlap in gene segment-defined antibody repertoires. *BMC Genomics* **6**, 148 (2005).

[20] Regep, C. et al. The H3 loop of antibodies shows unique structural characteristics. *Proteins* **85**, 1311–1318 (2017).

[21] Weitzner, B.D., Dunbrack, R.L., and Gray, J.J. The Origin of CDR H3 Structural Diversity. *Structure* **23**, 302–311 (2015).

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

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

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

[25] Guloglu, B. and Deane, C.M. Specific attributes of the VL domain influence both the structure and structural variability of CDR-H3 through steric effects. *Frontiers in Immunology* **14**, 1223802 (2023).

[26] Wong, W.K., Leem, J., and Deane, C.M. Comparative Analysis of the CDR Loops of Antigen Receptors. *Frontiers in Immunology* **10**, 2454 (2019).

[27] Marks, C. and Deane, C.M. Antibody H3 Structure Prediction. *Computational and Structural Biotechnology Journal* **15**, 222–231 (2017).

[28] Greenberg, A.S. et al. A new antigen receptor gene family that undergoes rearrangement and extensive somatic diversification in sharks. *Nature* **374**, 168–173 (1995).

[29] Hamers-Casterman, C. et al. Naturally occurring antibodies devoid of light chains. *Nature* **363**, 446–448 (1993).

[30] Muyldermans, S. Nanobodies: Natural Single-Domain Antibodies. *Annual Review of Biochemistry* **82**, 775–797 (2013).

[31] Flajnik, M.F., Deschacht, N., and Muyldermans, S. A Case Of Convergence: Why Did a Simple Alternative to Canonical Antibodies Arise in Sharks and Camels? *PLoS Biology* **9**, e1001120 (2011).

[32] Klarenbeek, A. et al. Camelid Ig V genes reveal significant human homology not seen in therapeutic target genes, providing for a powerful therapeutic antibody platform. *mAbs* **7**, 693–706 (2015).

[33] Flajnik, M.F. and Kasahara, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nature Reviews Genetics* **11**, 47–59 (2010).

[34] Zhang, C., Du Pasquier L, and Hsu, E. Shark IgW C region diversification through RNA processing and isotype switching. *Journal of immunology (Baltimore, Md. : 1950)* **191**, 3410–8 (2013).

[35] Nguyen, V.K., Muyldermans, S., and Hamers, R. The specific variable domain of camel heavy-chain antibodies is encoded in the germline. *Journal of Molecular Biology* **275**, 413–418 (1998).

[36] Nguyen, V.K. et al. Camel heavy-chain antibodies: diverse germline VHH and specific mechanisms enlarge the antigen-binding repertoire. *The EMBO Journal* **19**, 921–930 (2000).

[37] Tu, Z. et al. Landscape of variable domain of heavy-chain-only antibody repertoire from alpaca. *Immunology* **161**, 53–65 (2020).

[38] Bahrami-Dizicheh, Z., Chen, I., and Koenig, P. VHH CDR-H3 conformation is determined by VH germline usage. *Communications Biology* **6**, 864 (2023).

[39] Zielonka, S. et al. Structural insights and biomedical potential of IgNAR scaffolds from sharks. *mAbs* **7**, 15–25 (2014).

[40] Khalid, Z. et al. IgNAR antibody: Structural features, diversity and applications. *Fish & Shellfish Immunology* **121**, 467–477 (2022).

[41] Stanfield, R.L. et al. Maturation of Shark Single-domain (IgNAR) Antibodies: Evidence for Induced-fit Binding. *Journal of Molecular Biology* **367**, 358–372 (2007).

[42] Muyldermans, S. Single domain camel antibodies: current status. *Reviews in Molecular Biotechnology* **74**, 277–302 (2001).

[43] Muyldermans, S. et al. Camelid immunoglobulins and nanobody technology. *Veterinary Immunology and Immunopathology* **128**, 178–183 (2009).

[44] Sircar, A. et al. Analysis and Modeling of the Variable Region of Camelid Single-Domain Antibodies. *The Journal of Immunology* **186**, 6357–6367 (2011).

[45] Genst, E.D. et al. Molecular basis for the preferential cleft recognition by dromedary heavy-chain antibodies. *Proceedings of the National Academy of Sciences* **103**, 4586–4591 (2006).

[46] Wesolowski, J. et al. Single domain antibodies: promising experimental and therapeutic tools in infection and immunity. *Medical Microbiology and Immunology* **198**, 157–174 (2009).

[47] Desmyter, A. et al. Camelid nanobodies: killing two birds with one stone. *Current Opinion in Structural Biology* **32**, 1–8 (2015).

[48] Conrath, K. et al. Antigen Binding and Solubility Effects upon the Veneering of a Camel VHH in Framework-2 to Mimic a VH. *Journal of Molecular Biology* **350**, 112–125 (2005).

[49] Eshak, F. and Goupil-Lamy, A. Advancements in Nanobody Epitope Prediction: A Comparative Study of AlphaFold2Multimer vs AlphaFold3. *Journal of Chemical Information and Modeling* **65**, 1782–1797 (2025).

[50] Kuroda, D. and Tsumoto, K. "Structural Classification of CDR-H3 in Single-Domain VHH Antibodies". *Computer-Aided Antibody Design*. 61–79, (Springer US, 2022).

[51] Lefranc, M. et al. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Developmental & Comparative Immunology* **27**, 55–77 (2003).

[52] Li, X. et al. Comparative Analysis of Immune Repertoires between Bactrian Camel's Conventional and Heavy-Chain Antibodies. *PLoS ONE* **11**, e0161801 (2016).

[53] Vincke, C. et al. General Strategy to Humanize a Camelid Single-domain Antibody and Identification of a Universal Humanized Nanobody Scaffold. *Journal of Biological Chemistry* **284**, 3273–3284 (2009).

[54] Zavrtanik, U. et al. Structural Basis of Epitope Recognition by Heavy-Chain Camelid Antibodies. *Journal of Molecular Biology* **430**, 4369–4386 (2018).

[55] Mitchell, L.S. and Colwell, L.J. Comparative analysis of nanobody sequence and structure data. *Proteins: Structure, Function, and Bioinformatics* **86**, 697–706 (2018).

[56] Cheong, W.S. et al. Diagnostic and therapeutic potential of shark variable new antigen receptor (VNAR) single domain antibody. *International Journal of Biological Macromolecules* **147**, 369–375 (2020).

[57] Juma, S.N. et al. Shark New Antigen Receptor (IgNAR): Structure, Characteristics and Potential Biomedical Applications. *Cells* **10**, 1140 (2021).

[58] Fernández-Quintero, M.L. et al. The influence of antibody humanization on shark variable domain (VNAR) binding site ensembles. *Frontiers in Immunology* **13**, 953917 (2022).

[59] Wen, J. et al. Identification and characterization of IgNAR and VNAR repertoire from the ocellate spot skate (Okamejei kenojei). *Frontiers in Marine Science* **10**, 1183744 (2023).

[60] Kovalenko, O.V. et al. Atypical Antigen Recognition Mode of a Shark Immunoglobulin New Antigen Receptor (IgNAR) Variable Domain Characterized by Humanization and Structural Analysis. *The Journal of Biological Chemistry* **288**, 17408–17419 (2013).

[61] English, H., Hong, J., and Ho, M. Ancient species offers contemporary therapeutics: an update on shark VNAR single domain antibody sequences, phage libraries and potential clinical applications. *Antibody Therapeutics* **3**, 1–9 (2020).

[62] Raybould, M.I.J. et al. Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences* **116**, 4025–4030 (2019).

[63] Crescioli, S. et al. Antibodies to watch in 2025. *mAbs* **17**, 2443538 (2025).

[64] Chames, P. et al. Therapeutic antibodies: successes, limitations and hopes for the future. *British Journal of Pharmacology* **157**, 220–233 (2009).

[65] Krah, S. et al. Single-domain antibodies for biomedical applications. *Immunopharmacology and Immunotoxicology* **38**, 21–28 (2015).

[66] Yang, E.Y. and Shah, K. Nanobodies: Next Generation of Cancer Diagnostics and Therapeutics. *Frontiers in Oncology* **10**, 1182 (2020).

[67] Bannas, P., Hambach, J., and Koch-Nolte, F. Nanobodies and Nanobody-Based Human Heavy Chain Antibodies As Antitumor Therapeutics. *Frontiers in Immunology* **8**, 1603 (2017).

[68] Czajka, T.F., Vance, D.J., and Mantis, N.J. Slaying SARS-CoV-2 One (Single-domain) Antibody at a Time. *Trends in Microbiology* **29**, 195–203 (2021).

[69] Liu, C. et al. Research progress on unique paratope structure, antigen binding modes, and systematic mutagenesis strategies of single-domain antibodies. *Frontiers in Immunology* **13**, 1059771 (2022).

[70] Griffiths, K. et al. Shark Variable New Antigen Receptor (VNAR) Single Domain Antibody Fragments: Stability and Diagnostic Applications. *Antibodies* **2**, 66–81 (2013).

[71] Chen, Y. et al. Screening and Characterization of Shark-Derived VNARs against SARS-CoV-2 Spike RBD Protein. *International Journal of Molecular Sciences* **23**, 10904 (2022).

[72] Zielonka, S. et al. Shark Attack: high affinity binding proteins derived from shark vNAR domains by stepwise in vitro affinity maturation. *Journal of Biotechnology* **191**, 236–245 (2014).

[73] Könning, D. et al. Semi-synthetic vNAR libraries screened against therapeutic antibodies primarily deliver anti-idiotypic binders. *Scientific Reports* **7**, 9676 (2017).

[74] Chen, W. et al. Shark nanobodies with potent SARS-CoV-2 neutralizing activity and broad sarbecovirus reactivity. *Nature Communications* **14**, 580 (2023).

[75] Ludwig, S.D. et al. Multiparatopic antibodies induce targeted downregulation of programmed death-ligand 1. *Cell Chemical Biology* **31**, 1023 (2024).

[76] Könning, D. et al. Camelid and shark single domain antibodies: structural features and therapeutic potential. *Current Opinion in Structural Biology* **45**, 10–16 (2017).

[77] Scully, M. et al. Caplacizumab Treatment for Acquired Thrombotic Thrombocytopenic Purpura. *New England Journal of Medicine* **380**, 335–346 (2019).

[78] Natrajan, K. et al. FDA Approval Summary: Ciltacabtagene Autoleucel for Relapsed or Refractory Multiple Myeloma. *Clinical Cancer Research* **30**, 2865–2871 (2024).

[79] Keam, S.J. Ozoralizumab: First Approval. *Drugs* **83**, 87–92 (2023).

[80] Markham, A. Envafolimab: First Approval. *Drugs* **82**, 235–240 (2022).

[81] De Pauw, T. et al. Current status and future expectations of nanobodies in oncology trials. *Expert Opinion on Investigational Drugs* **32**, 705–721 (2023).

[82] Minatel, V.M. et al. Nanobodies: a promising approach to treatment of viral diseases. *Frontiers in Immunology* **14**, 1303353 (2024).

[83] Laustsen, A.H. et al. Animal Immunization, in Vitro Display Technologies, and Machine Learning for Antibody Discovery. *Trends in Biotechnology* **39**, 1263–1273 (2021).

[84] Alexander, E. and Leong, K.W. Discovery of nanobodies: a comprehensive review of their applications and potential over the past five years. *Journal of Nanobiotechnology* **22**, 661 (2024).

[85] Kolmar, H. et al. "Construction of Semisynthetic Shark vNAR Yeast Surface Display Antibody Libraries". *Phage Display: Methods and Protocols*. 227–243, (Springer US, 2023).

[86] Takeda, H. et al. VNAR development through antigen immunization of Japanese topeshark (Hemitriakis japanica). *Frontiers in Bioengineering and Biotechnology* **11**, 1265582 (2023).

[87] Jain, T. et al. Biophysical properties of the clinical-stage antibody landscape. *Proceedings of the National Academy of Sciences* **114**, 944–949 (2017).

[88] Yang, R. et al. Rapid assessment of oxidation via middle-down LCMS correlates with methionine side-chain solvent-accessible surface area for 121 clinical stage monoclonal antibodies. *mAbs* **9**, 646–653 (2017).

[89] Ausserwöger, H. et al. Non-specificity as the sticky problem in therapeutic antibody development. *Nature Reviews Chemistry* **6**, 844–861 (2022).

[90] Zhang, W. et al. Developability assessment at early-stage discovery to enable development of antibody-derived therapeutics. *Antibody Therapeutics* **6**, 13–29 (2022).

[91] Mieczkowski, C. et al. Blueprint for antibody biologics developability. *mAbs* **15**, 2185924 (2023).

[92] Lundahl, M.L.E. et al. Aggregation of protein therapeutics enhances their immunogenicity: causes and mitigation strategies. *RSC Chemical Biology* **2**, 1004–1020 (2021).

[93] Jacobs, S.A. et al. Cross-Interaction Chromatography: A Rapid Method to Identify Highly Soluble Monoclonal Antibody Candidates. *Pharmaceutical Research* **27**, 65–71 (2009).

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

[95] Fernández-Quintero, M.L. et al. Assessing developability early in the discovery process for novel biologics. *mAbs* **15**, (2023).

[96] Khera, H.K. et al. Nano-differential scanning fluorimetry as a tool for the assessment of refolded antibody fragments: A case study for anti-Pfs25 single-chain antibodies. *Biochemical Engineering Journal* **206**, 109287 (2024).

[97] Zarzar, J. et al. High concentration formulation developability approaches and considerations. *mAbs* **15**, 2211185 (2023).

[98] Liu, Y. et al. High-throughput screening for developability during early-stage antibody discovery using self-interaction nanoparticle spectroscopy. *mAbs* **6**, 483–492 (2013).

[99] Starr, C.G. et al. Ultradilute Measurements of Self-Association for the Identification of Antibodies with Favorable High-Concentration Solution Properties. *Molecular Pharmaceutics* **18**, 2744–2753 (2021).

[100] Sankar, K. et al. Prediction of methionine oxidation risk in monoclonal antibodies using a machine learning method. *mAbs* **10**, 1281–1290 (2018).

[101] Houde, D. et al. Post-translational Modifications Differentially Affect IgG1 Conformation and Receptor Binding. *Molecular & Cellular Proteomics* **9**, 1716–1728 (2010).

[102] Buchanan, A. et al. Engineering a therapeutic IgG molecule to address cysteinylation, aggregation and enhance thermal stability and expression. *mAbs* **5**, 255–262 (2013).

[103] Kelly, R.L. et al. Reduction of Nonspecificity Motifs in Synthetic Antibody Libraries. *Journal of Molecular Biology* **430**, 119–130 (2018).

[104] Morrison, S.L. et al. Chimeric human antibody molecules: mouse antigen-binding domains with human constant region domains. *Proceedings of the National Academy of Sciences* **81**, 6851–6855 (1984).

[105] Jones, P.T. et al. Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature* **321**, 522–525 (1986).

[106] De Pascalis, R. et al. Grafting of "Abbreviated" Complementarity-Determining Regions Containing Specificity-Determining Residues Essential for Ligand Contact to Engineer a Less Immunogenic Humanized Monoclonal Antibody. *The Journal of Immunology* **169**, 3076–3084 (2002).

[107] Kashmiri, S.V.S. et al. SDR grafting—a new approach to antibody humanization. *Methods* **36**, 25–34 (2005).

[108] Safdari, Y. et al. Antibody humanization methods – a review and update. *Biotechnology and Genetic Engineering Reviews* **29**, 175–186 (2013).

[109] Dall'Acqua, W.F. et al. Antibody humanization by framework shuffling. *Methods* **36**, 43–60 (2005).

[110] Padlan, E.A. A possible procedure for reducing the immunogenicity of antibody variable domains while preserving their ligand-binding properties. *Molecular Immunology* **28**, 489–498 (1991).

[111] Roguska, M.A. et al. Humanization of murine monoclonal antibodies through variable domain resurfacing. *Proceedings of the National Academy of Sciences* **91**, 969–973 (1994).

[112] Gentiluomo, L. et al. Characterization of Native Reversible Self-Association of a Monoclonal Antibody Mediated by Fab-Fab Interaction. *Journal of Pharmaceutical Sciences* **109**, 443–451 (2020).

[113] Voynov, V. et al. Predictive tools for stabilization of therapeutic proteins. *mAbs* **1**, 580–582 (2009).

[114] Waibl, F. et al. Conformational Ensembles of Antibodies Determine Their Hydrophobicity. *Biophysical Journal* **120**, 143–157 (2021).

[115] Needleman, S.B. and Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**, 443–453 (1970).

[116] Smith, T.F. and Waterman, M.S. Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195–197 (1981).

[117] Altschul, S.F. et al. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

[118] Dondelinger, M. et al. Understanding the Significance and Implications of Antibody Numbering and Antigen-Binding Surface/Residue Definition. *Frontiers in Immunology* **9**, 2278 (2018).

[119] Wu, T.T. and Kabat, E.A. An analysis of the sequences of the variable regions of Bence Jones proteins and their implications for antibody complementarity. *Journal of Experimental Medicine* **132**, 211–250 (1970).

[120] Abhinandan, K.R. and Martin, A.C.R. Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Molecular Immunology* **45**, 3832–3839 (2008).

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

[122] Zhu, Z. et al. Antibody numbering schemes: advances, comparisons and tools for antibody engineering. *Protein Engineering, Design and Selection* **38**, gzaf005 (2025).

[123] Dunbar, J. and Deane, C.M. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, btv552 (2015).

[124] Lefranc, M. et al. IMGT, the international ImMunoGeneTics information system ®. *Nucleic Acids Research* **33**, D593–D597 (2005).

[125] Alberts, B. et al. "Analyzing Protein Structure and Function". *Molecular Biology of the Cell*. 4th edition. Garland Science, 2002.

[126] Khetan, R. et al. Current advances in biopharmaceutical informatics: guidelines, impact and challenges in the computational developability assessment of antibody therapeutics. *mAbs* **14**, 2020082 (2022).

[127] Webb, B. and Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics* **54**, 5.6.1–5.6.37 (2016).

[128] Waterhouse, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research* **46**, W296–W303 (2018).

[129] Kelley, L.A. et al. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* **10**, 845–858 (2015).

[130] Leem, J. et al. ABodyBuilder: Automated antibody structure prediction with data–driven accuracy estimation. *mAbs* **8**, 1259–1268 (2016).

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

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

[133] Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

[134] Ruffolo, J.A. et al. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature Communications* **14**, 2389 (2023).

[135] Baek, M. et al. Efficient and accurate prediction of protein structure using RoseTTAFold2. *bioRxiv*, 2023.05.24.542179 (2023).

[136] Abanades, B. et al. ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. *Communications Biology* **6**, 575 (2023).

[137] Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

[138] Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021.10.04.463034 (2022).

[139] Giulini, M. et al. Towards the accurate modelling of antibodyantigen complexes from sequence using machine learning and information-driven docking. *Bioinformatics* **40**, btae583 (2024).

[140] Shrestha, P. et al. NanoBinder: a machine learning assisted nanobody binding prediction tool using Rosetta energy scores. *Journal of Cheminformatics* **17**, 96 (2025).

[141] Dunbar, J. et al. SAbDab: the structural antibody database. *Nucleic Acids Research* **42**, D1140–D1146 (2013).

[142] Schneider, C., Raybould, M.I.J., and Deane, C.M. SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. *Nucleic Acids Research* **50**, D1368–D1372 (2021).

[143] Berman, H.M. et al. The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).

[144] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research* **53**, D609–D617 (2025).

[145] Sayers, E.W. et al. Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Research* **53**, D20–D29 (2024).

[146] Kovaltsuk, A. et al. Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. *J. Immunol* **201**, 2502–2509 (2018).

[147] Olsen, T.H., Boyles, F., and Deane, C.M. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science* **31**, 141–146 (2021).

[148] Corrie, B.D. et al. iReceptor: a platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunological Reviews* **284**, 24–41 (2018).

[149] Raybould, M.I.J. et al. CoV-AbDab: the coronavirus antibody database. *Bioinformatics* **37**, 734–735 (2020).

[150] Vita, R. et al. The immune epitope database (IEDB) 3.0. *Nucleic Acids Research* **43**, D405–D412 (2015).

[151] Vita, R. et al. The Immune Epitope Database (IEDB): 2024 update. *Nucleic Acids Research* **53**, D436–D443 (2025).

[152] Abanades, B. et al. The Patent and Literature Antibody Database (PLAbDab): an evolving reference set of functionally diverse, literature-annotated antibody sequences and structures. *Nucleic Acids Research* **52**, D545–D551 (2024).

[153] Raybould, M.I.J. et al. Thera-SAbDab: the Therapeutic Structural Antibody Database. *Nucleic Acids Research* **48**, D383–D388 (2020).

[154] Kaas, Q., Ruiz, M., and Lefranc, M. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Research* **32**, D208–D210 (2004).

[155] Swindells, M.B. et al. abYsis: Integrated Antibody Sequence and Structure—Management, Analysis, and Prediction. *Journal of Molecular Biology* **429**, 356–364 (2017).

[156] Deszyński, P. et al. INDI—integrated nanobody database for immunoinformatics. *Nucleic Acids Research* **50**, D1273–D1281 (2021).

[157] Xiong, S. et al. NanoLAS: a comprehensive nanobody database with data integration, consolidation and application. *Database* **2024**, baae003 (2024).

[158] Wilton, E.E. et al. sdAb-DB: The Single Domain Antibody Database. *ACS Synthetic Biology* **7**, 2480–2484 (2018).

[159] Conchillo-Solé, O. et al. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics* **8**, 65 (2007).

[160] Fernandez-Escamilla, A. et al. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology* **22**, 1302–1306 (2004).

[161] Tartaglia, G.G. and Vendruscolo, M. The Zyggregator method for predicting protein aggregation propensities. *Chemical Society Reviews* **37**, 1395–1401 (2008).

[162] Zibaee, S. et al. A simple algorithm locates -strands in the amyloid fibril core of -synuclein, A, and tau using the amino acid sequence alone. *Protein Science* **16**, 906–918 (2007).

[163] Walsh, I. et al. PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Research* **42**, W301–W307 (2014).

[164] Kuriata, A. et al. Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility. *Nucleic Acids Research* **47**, W300–W307 (2019).

[165] Chennamsetty, N. et al. Design of therapeutic proteins with enhanced stability. *Proceedings of the National Academy of Sciences* **106**, 11937–11942 (2009).

[166] Sankar, K. et al. AggScore: Prediction of aggregation-prone regions in proteins based on the distribution of surface patches. *Proteins: Structure, Function, and Bioinformatics* **86**, 1147–1156 (2018).

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

[168] Rollins, Z.A. et al. AbMelt: Learning antibody thermostability from molecular dynamics. *Biophysical Journal* **123**, 2921–2933 (2024).

[169] Harmalkar, A. et al. Toward generalizable prediction of antibody thermostability using machine learning on sequence and structure features. *mAbs* **15**, 2163584 (2023).

[170] Kumar, M.D.S. et al. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Research* **34**, D204–206 (2006).

[171] Nikam, R. et al. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Research* **49**, D420–D424 (2021).

[172] Valdés-Tresanco, M.S. et al. NbThermo: a new thermostability database for nanobodies. *Database* **2023**, baad021 (2023).

[173] Alvarez, J.A.E. and Dean, S.N. TEMPRO: nanobody melting temperature estimation model using protein embeddings. *Scientific Reports* **14**, 19074 (2024).

[174] Ramon, A. et al. Prediction of protein biophysical traits from limited data: a case study on nanobody thermostability through NanoMelt. *mAbs* **17**, 2442750 (2025).

[175] Magnan, C.N., Randall, A., and Baldi, P. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics* **25**, 2200–2207 (2009).

[176] Khurana, S. et al. DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* **34**, 2605–2613 (2018).

[177] Rawi, R. et al. PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics* **34**, 1092–1098 (2018).

[178] Hou, Q. et al. SOLart: a structure-based method to predict protein solubility and aggregation. *Bioinformatics* **36**, 1445–1452 (2020).

[179] Sharma, V.K. et al. In silico selection of therapeutic antibodies for development: Viscosity, clearance, and chemical stability. *Proceedings of the National Academy of Sciences* **111**, 18601–18606 (2014).

[180] Agrawal, N.J. et al. Computational tool for the early screening of monoclonal antibodies for their viscosities. *mAbs* **8**, 43–48 (2015).

[181] Lai, P. DeepSCM: An efficient convolutional neural network surrogate model for the screening of therapeutic antibody viscosity. *Computational and Structural Biotechnology Journal* **20**, 2143–2152 (2022).

[182] Sormanni, P., Aprile, F.A., and Vendruscolo, M. The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *Journal of Molecular Biology* **427**, 478–490 (2015).

[183] Gavel, Y. and von Heijne, G. Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. *Protein Engineering* **3**, 433–442 (1990).

[184] Sydow, J.F. et al. Structure-Based Prediction of Asparagine and Aspartate Degradation Sites in Antibody Variable Regions. *PLoS One* **9**, e100736 (2014).

[185] Kelly, R.L. et al. Reduction of Nonspecificity Motifs in Synthetic Antibody Libraries. *Journal of Molecular Biology* **430**, 119–130 (2018).

[186] Ruoslahti, E. RGD and other recognition sequences for integrins. *Annual Review of Cell and Developmental Biology* **12**, 697–715 (1996).

[187] Robinson, N.E. and Robinson, A.B. Prediction of protein deamidation rates from primary and three-dimensional structure. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 4367–4372 (2001).

[188] Wakankar, A.A. and Borchardt, R.T. Formulation considerations for proteins susceptible to asparagine deamidation and aspartate isomerization. *Journal of Pharmaceutical Sciences* **95**, 2321–2336 (2006).

[189] Agrawal, N.J. et al. Prediction of the Hydrogen Peroxide-Induced Methionine Oxidation Propensity in Monoclonal Antibodies. *Journal of Pharmaceutical Sciences* **107**, 1282–1289 (2018).

[190] Abhinandan, K.R. and Martin, A.C.R. Analyzing the "Degree of Humanness" of Antibody Sequences. *Journal of Molecular Biology* **369**, 852–862 (2007).

[191] Thullier, P. et al. The Humanness of Macaque Antibody Sequences. *Journal of Molecular Biology* **396**, 1439–1450 (2010).

[192] Gao, S.H. et al. Monoclonal antibody humanness score and its applications. *BMC Biotechnology* **13**, 55 (2013).

[193] Lazar, G.A. et al. A molecular immunology approach to antibody humanization and functional optimization. *Molecular Immunology* **44**, 1986–1998 (2007).

[194] Seeliger, D. Development of Scoring Functions for Antibody Sequence Assessment and Optimization. *PLoS One* **8**, e76909 (2013).

[195] Clavero-Álvarez, A. et al. Humanization of Antibodies using a Statistical Inference Approach. *Scientific Reports* **8**, 14820 (2018).

[196] Schmitz, S. et al. Human-likeness of antibody biologics determined by back-translation and comparison with large antibody variable gene repertoires. *mAbs* **12**, 1758291 (2020).

[197] Petersen, B.M. et al. Regulatory Approved Monoclonal Antibodies Contain Framework Mutations Predicted From Human Antibody Repertoires. *Frontiers in Immunology* **12**, (2021).

[198] Marks, C. et al. Humanization of antibodies using a machine learning approach on large-scale repertoire data. *Bioinformatics* **37**, 4041–4047 (2021).

[199] Ucar, T. et al. Improving Antibody Humanness Prediction using Patent Data. *arXiv*, 2401.14442 (2024).

[200] Prihoda, D. et al. BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *mAbs* **14**, 2020203 (2022).

[201] Vashchenko, D. et al. AbBERT: Learning Antibody Humanness via Masked Language Modeling. *bioRxiv*, 2022.08.02.502236 (2022).

[202] Zou, H. et al. Antibody Humanization via Protein Language Model and Neighbor Retrieval. *bioRxiv*, 2023.09.04.556278 (2023).

[203] Wollacott, A.M. et al. Quantifying the nativeness of antibody sequences using long short-term memory networks. *Protein Engineering, Design and Selection* **32**, 347–354 (2019).

[204] Chinery, L., Jeliazkov, J.R., and Deane, C.M. Humatch - fast, gene-specific joint humanisation of antibody heavy and light chains. *mAbs* **16**, 2434121 (2024).

[205] Choi, Y. et al. Antibody humanization by structure-based computational protein design. *mAbs* **7**, 1045–1057 (2015).

[206] Choi, Y. et al. Computationally driven antibody engineering enables simultaneous humanization and thermostabilization. *Protein Engineering, Design and Selection* **29**, 419–426 (2016).

[207] Tennenhouse, A. et al. Computational optimization of antibody humanness and stability by systematic energy-based ranking. *Nature Biomedical Engineering* **8**, 30–44 (2024).

[208] Hummer, A.M. and Deane, C.M. Designing stable humanized antibodies. *Nature Biomedical Engineering* **8**, 3–4 (2024).

[209] Sang, Z. et al. Llamanade: An open-source computational pipeline for robust nanobody humanization. *Structure* **30**, 418–429.e3 (2022).

[210] Ramon, A. et al. Assessing antibody and nanobody nativeness for hit selection and humanization with AbNatiV. *Nature Machine Intelligence* **6**, 74–91 (2024).

[211] Lauer, T.M. et al. Developability Index: A Rapid In Silico Tool for the Screening of Antibody Aggregation Propensity. *Journal of Pharmaceutical Sciences* **101**, 102–115 (2012).

[212] Evers, A. et al. SUMO: In Silico Sequence Assessment Using Multiple Optimization Parameters. *Methods in Molecular Biology* **2681**, 383–398 (2023).

[213] Park, E. and Izadi, S. Molecular surface descriptors to predict antibody developability: sensitivity to parameters, structure models, and conformational sampling. *mAbs* **16**, (2024).

[214] Ahmed, L. et al. Intrinsic physicochemical profile of marketed antibody-based biotherapeutics. *Proceedings of the National Academy of Sciences* **118**, (2021).

[215] Makowski, E.K. et al. Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nature Communications* **13**, (2022).

[216] Warszawski, S. et al. Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces. *PLoS Computational Biology* **15**, e1007207 (2019).

[217] Rosace, A. et al. Automated optimisation of solubility and conformational stability of antibodies and proteins. *Nature Communications* **14**, 1937 (2023).

[218] Bachas, S. et al. Antibody optimization enabled by artificial intelligence predictions of binding affinity and naturalness. *bioRxiv*, 2022.08.16.504181 (2022).

[219] Liu, G. et al. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics* **36**, 2126–2133 (2020).

[220] Saka, K. et al. Antibody design using LSTM based deep generative model from phage display library for affinity maturation. *Scientific Reports* **11**, 5852 (2021).

[221] Arras, P. et al. A library approach for the de novo high-throughput isolation of humanized VHH domains with favorable developability properties following camelid immunization. *mAbs* **15**, (2023).

[222] Arras, P. et al. AI/ML combined with next-generation sequencing of VHH immune repertoires enables the rapid identification of de novo humanized and sequence-optimized single domain antibodies: a prospective case study. *Frontiers in Molecular Biosciences* **10**, (2023).

[223] Henry, K.A. and MacKenzie, C.R. Antigen recognition by single-domain antibodies: structural latitudes and constraints. *mAbs* **10**, 815–826 (2018).

[224] Mitchell, L.S. and Colwell, L.J. Analysis of nanobody paratopes reveals greater diversity than classical antibodies. *Protein Engineering, Design and Selection* **31**, 267–275 (2018).

[225] Eliyahu, S. et al. Antibody Repertoire Analysis of Hepatitis C Virus Infections Identifies Immune Signatures Associated With Spontaneous Clearance. *Frontiers in Immunology* **9**, (2018).

[226] Li, W. and Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

[227] Jubb, H.C. et al. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of Molecular Biology* **429**, 365–371 (2017).

[228] Cock, P.J.A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

[229] Scally, S.W. et al. Molecular definition of multiple sites of antibody inhibition of malaria transmission-blocking vaccine antigen Pfs25. *Nature Communications* **8**, 1568 (2017).

[230] Schrödinger, L.L.C. *The PyMOL Molecular Graphics System, Version 1.8*. 2015.

[231] Kelow, S. et al. A penultimate classification of canonical antibody CDR conformations. *bioRxiv*, 2022.10.12.511988 (2022).

[232] Wong, W.K. et al. Ab-Ligity: identifying sequence-dissimilar antibodies that bind to the same epitope. *mAbs* **13**, 1873478 (2021).

[233] Sela, M. et al. Antibodies to Sequential and Conformational Determinants. *Cold Spring Harbor Symposia on Quantitative Biology* **32**, 537–545 (1967).

[234] Forsström, B. et al. Dissecting Antibodies with Regards to Linear and Conformational Epitopes. *PLoS ONE* **10**, e0121673 (2015).

[235] Huo, J. et al. Neutralization of SARS-CoV-2 by Destruction of the Prefusion Spike. *Cell Host & Microbe* **28**, 445–454.e6 (2020).

[236] Wrapp, D. et al. Structural Basis for Potent Neutralization of Betacoronaviruses by Single-Domain Camelid Antibodies. *Cell* **181**, 1004–1015.e15 (2020).

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

[238] Raybould, M.I.J. et al. Public Baseline and shared response structures support the theory of antibody repertoire functional commonality. *PLoS Computational Biology* **17**, e1008781 (2021).

[239] Matz, H. and Dooley, H. Shark IgNAR-derived binding domains as potential diagnostic and therapeutic agents. *Developmental & Comparative Immunology* **90**, 100–107 (2019).

[240] Sun, Z. et al. Structure basis of two nanobodies neutralizing SARS-CoV-2 Omicron variant by targeting ultra-conservative epitopes. *Journal of Structural Biology* **215**, 107996 (2023).

[241] Olsen, T.H. et al. KA-Search, a method for rapid and exhaustive sequence identity search of known antibodies. *Scientific Reports* **13**, 11612 (2023).

[242] Clark, K. et al. GenBank. *Nucleic Acids Research* **44**, D67–D72 (2015).

[243] Parkinson, J. and Wang, W. For antibody sequence generative modeling, mixture models may be all you need. *Bioinformatics* **40**, btae278 (2024).

[244] Vaswani, A. et al. Attention Is All You Need. *arXiv*, 1706.03762 (2017).

[245] Ziv, Y. et al. MolSnapper: Conditioning Diffusion for Structure-Based Drug Design. *Journal of Chemical Information and Modeling* **65**, 4263–4273 (2025).

[246] Gordon, G.L. et al. A comparison of the binding sites of antibodies and single-domain antibodies. *Frontiers in Immunology* **14**, 1231623 (2023).

[247] Greenshields-Watson, A. et al. ANARCII: A Generalised Language Model for Antigen Receptor Numbering. *bioRxiv*, 2025.04.16.648720 (2025).

[248] Mullard, A. FDA approves 100th monoclonal antibody product. *Nature Reviews Drug Discovery* **20**, 491–495 (2021).

[249] Bauer, J. et al. How can we discover developable antibody-based biotherapeutics? *Frontiers in Molecular Biosciences* **10**, (2023).

[250] Gordon, G.L. et al. Prospects for the computational humanization of antibodies and nanobodies. *Frontiers in Immunology* **15**, (2024).

[251] Chennamsetty, N. et al. Design of therapeutic proteins with enhanced stability. *Proceedings of the National Academy of Sciences* **106**, 11937–11942 (2009).

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

[253] Sydow, J.F. et al. Structure-Based Prediction of Asparagine and Aspartate Degradation Sites in Antibody Variable Regions. *PLoS ONE* **9**, e100736 (2014).

[254] Lai, P. DeepSCM: An efficient convolutional neural network surrogate model for the screening of therapeutic antibody viscosity. *Computational and Structural Biotechnology Journal* **20**, 2143–2152 (2022).

[255] Kalejaye, L. et al. DeepSP: Deep learning-based spatial properties to predict monoclonal antibody stability. *Computational and Structural Biotechnology Journal* **23**, 2220–2229 (2024).

[256] Kalejaye, L.A. et al. Accelerating high-concentration monoclonal antibody development with large-scale viscosity data and ensemble deep learning. *mAbs* **17**, (2025).

[257] Waight, A.B. et al. A machine learning strategy for the identification of key in silico descriptors and prediction models for IgG monoclonal antibody developability properties. *mAbs* **15**, (2023).

[258] Makowski, E.K. et al. Optimization of therapeutic antibodies for reduced self-association and non-specific binding via interpretable machine learning. *Nature Biomedical Engineering* **8**, 45–56 (2023).

[259] Dreyer, F.A. et al. Computational design of therapeutic antibodies with improved developability: efficient traversal of binder landscapes and rescue of escape mutations. *mAbs* **17**, 2511220 (2025).

[260] Bashour, H. et al. Biophysical cartography of the native and human-engineered antibody landscapes quantifies the plasticity of antibody developability. *Communications Biology* **7**, 922 (2024).

[261] Tomimoto, Y., Yamazaki, R., and Shirai, H. Increasing the melting temperature of VHH with the in silico free energy score. *Scientific Reports* **13**, (2023).

[262] Gordon, G.L. et al. PLAbDab-nano: a database of camelid and shark nanobodies from patents and literature. *Nucleic Acids Research* **53**, D535–D542 (2025).

[263] Jaffe, D.B. et al. Functional antibodies exhibit light chain coherence. *Nature* **611**, 352–357 (2022).

[264] Zhu, Z.Y. and Blundell, T.L. The use of amino acid patterns of classified helices and strands in secondary structure prediction. *Journal of Molecular Biology* **260**, 261–276 (1996).

[265] Kyte, J. and Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157**, 105–132 (1982).

[266] Wojdyr, M. GEMMI: A library for structural biology. *Journal of Open Source Software* **7**, 4200 (2022).

[267] Estep, P. et al. An alternative assay to hydrophobic interaction chromatography for high-throughput characterization of monoclonal antibodies. *mAbs* **7**, 553–561 (2015).

[268] Soler, M.A. et al. Effect of Humanizing Mutations on the Stability of the Llama Single-Domain Variable Region. *Biomolecules* **11**, 163 (2021).

[269] Kinoshita, S. et al. Molecular basis for thermal stability and affinity in a VHH: Contribution of the framework region and its influence in the conformation of the CDR3. *Protein Science* **31**, e4450 (2022).

[270] Jain, T., Boland, T., and Vásquez, M. Identifying developability risks for clinical progression of antibodies using high-throughput in vitro and in silico approaches. *mAbs* **15**, 2200540 (2023).

[271] Waibl, F. et al. Comparison of hydrophobicity scales for predicting biophysical properties of antibodies. *Frontiers in Molecular Biosciences* **9**, 960194 (2022).

[272] Hempelmann, A. et al. Nanobody-mediated macromolecular crowding induces membrane fission and remodeling in the African trypanosome. *Cell Reports* **37**, 109923 (2021).

[273] Wong, J.E.M.M. et al. Structural signatures in EPR3 define a unique class of plant carbohydrate receptors. *Nature Communications* **11**, 3797 (2020).