Investigating B-cell repertoire data using deep learning approaches to aid in the development of antibody therapeutics

Tobias Hegelund Olsen

St. Anne's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2023

Acknowledgements

I would like to extend my deepest gratitude to my supervisor, Prof. Charlotte Deane MBE, for your invaluable guidance, consistent support, and constructive feedback. Your expertise and mentorship have been instrumental in shaping this work and my academic growth.

I feel incredible lucky to have been part of the stimulating, vibrant and fun environment at OPIG. This group has been a significant factor in making this PhD one of the most rewarding experiences of my life. A special thanks goes to the OPIGlets Alex Watson, Alissa Hummer, Broncio Aguilar-Sanjuan, Carlos Outeiral, Eve Richardson, Leo Klarner, Lewis Chinery, Martin Buttenschoen, Matt Raybould, and Sarah Robinson. I am also grateful to the amazing cube people Brennan Abanades, Gemma Gordon, Lucy Vost, and Marc Moesser, who made sure every day was fun.

Further, I thank my former supervisor at DTU, Prof. Paolo Marcatili, for introducing me to immunoinformatics, a field which I have come to love, and encouraging my pursuit of a PhD. I also extend thanks to Magnus Høie, a close friend and fellow PhD student, for the endless discussions about our research.

I would also like to thank my family for their unending love, encouragement, and belief in my abilities. Their unwavering support has been a pillar of strength throughout my academic endeavors.

Last but not least, heartfelt thanks go to my partner, Sarah Olafsson. It is inexpressible how much your emotional support, understanding, and countless sacrifices, including relocating to the UK, have meant to me. You always brighten up my day, and I cannot wait to explore the world with you.

The collective contributions of all mentioned have been invaluable in the completion of this Ph.D. thesis, and for that, I am eternally grateful.

Ownership of work

All research and writing presented in this thesis are solely my contributions, unless specifically indicated otherwise. The use of the first-person plural is a stylistic choice and should not be interpreted as indicating multiple authors. Any work conducted by others is properly attributed and acknowledged by name.

Publications

This DPhil led to the following research outputs:

Olsen T.H., Boyles F., and Deane C.M. (2022) OAS: A diverse database of cleaned, annotated and translated unpaired and paired antibody sequences. *Protein science*, 31(1):141-146.

Olsen T.H., Moal I.H., and Deane C.M. (2022) AbLang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*. 2(1):vbac046

Olsen T.H., Abanades B., Moal I.H., and Deane C.M. (2023) KA-Search, a method for rapid and exhaustive sequence identity search of known antibodies. *Scientific Reports*, 13:11612.

Abanades B., **Olsen T.H.**, Raybould M.I.J, Aguilar-Sanjuan B., Wong W.K., Georges G., Bujotzek A., and Deane C.M. (2024) The Patent and Literature Antibody Database (PLAbDab): an evolving reference set of functionally diverse, literature-annotated antibody sequences and structures. *Nucleic Acids Research*. 52(D1):D545–D551.

Olsen T.H., Moal I.H., and Deane C.M. (2024) Addressing the antibody germline bias and its effect on language models for improved antibody design. *bioRxiv*.

Abstract

Antibodies have become an invaluable form of biotherapeutics, with an increasing number of new antibody derived therapeutics being developed and marketed each year. Despite their success, the process of antibody discovery and design remains challenging. In this DPhil, we leverage publicly available antibody sequences to develop computational tools to aid in the design of therapeutic antibodies. We introduce two new databases; an updated and expanded Observed Antibody Space (OAS) and the Patent and Literature Antibody Database (PLAbDab). To investigate these databases, we developed KA-Search, a rapid and flexible tool for antibody sequence identity search, and demonstrated its use at mining the billions of sequences in OAS and obtaining new binding-specific insights with PLAbDAb.

Deep learning methods also benefits from the growth of available antibody data. General protein language models can effectively capture context-aware protein sequence representations, useful for state-of-the-art predictions. For antibody specific tasks, a language model trained solely on antibodies may be more powerful. We therefore developed AbLang, trained on OAS, and demonstrated how it learns inherent antibody patterns and can restore fragmented antibody sequences. However, we reveal how antibody sequences are considerably biased towards the germline, potentially limiting antibody-trained models ability to suggest relevant non-germline mutations. To overcome this, we introduced AbLang-2, a refined model capable of suggesting a diverse set of valid mutations with high cumulative probability.

Collectively, the insights, databases, and computational tools presented in this work enhance our computational capabilities in antibody design and opens the way for leveraging deep learning in therapeutic antibody discovery and design.

Contents

1 Introduction

Contents

1.1 Chapter Abstract

In this introduction we will give an overview of antibodies, their use as therapeutics and how machine learning, especially language models, can help their discovery and design. We will begin by summarising the sequence and structure of antibodies before describing how their diversity arises. This will be followed by how antibody repertoires within individuals can be used to explore vaccine and disease states, and as a source of functional antibodies. The focus will then be on therapeutic antibody development, both lead selection and developability issues. In the last part of the introduction, we will introduce transformers and how they have revolutionized language models for text understanding and generation. This will be followed by how language models can be adapted to protein sequences, and more specifically, antibody sequences. Finally, we will briefly discuss how language models can potentially be used to aid in the design and development of therapeutic antibodies.

Parts of the introduction are adapted from text I authored in the following published articles.

> **Olsen T.H.**, Boyles F., and Deane C.M. (2022) OAS: A diverse database of cleaned, annotated and translated unpaired and paired antibody sequences. *Protein science*, 31(1):141-146.
>
> **Olsen T.H.**, Abanades B., Moal I.H., and Deane C.M. (2023) KA-Search, a method for rapid and exhaustive sequence identity search of known antibodies. *Scientific Reports*, 13:11612.
>
> **Olsen T.H.**, Moal I.H., and Deane C.M. (2022) AbLang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*. 2(1):vbac046

1.2 Antibodies

Antibodies are proteins with a central role in the adaptive immune response due to their potential to bind and neutralise any pathogen (i.e. bacteria and viruses), by either blocking their function or marking them for removal [1–3]. Antibodies are produced by B-cells, with each B-cell either presenting on their membrane or

secreting a single unique antibody. Antibodies are also known as immunoglobulins (Ig), and when presented on the membrane, referred to as B-Cell receptors (BCR).

1.2.1 Antibody function and diversity

Antibodies are found, with some differences, across a variety of animals. In humans, antibodies are Y-shaped polymers composed of four protein chains, two identical larger chains and two identical shorter chains, called heavy and light, respectively, linked together with disulphide bonds [2]. Each chain has one variable (V) domain [4] and one or more conserved (C) domains [5], with each domain composed of ∼110 amino acids [6] (see Fig. 1.1). The C domains of the heavy chain can be of five different isotypes, IgA, IgD, IgE, IgG and IgM, differing the biological properties and functional location of the antibody [7–9]. Traditionally, antibodies have been characterised by their crystallisable fragment (Fc) and two antigen binding fragment (Fab) regions (see Fig. 1.1). Located in the Fabs are the heavy chain variable domain (VH) and the light chain variable domain (VL), which paired together form the variable fragment (Fv) region [2, 8].

1.2.1.1 Antibody binding and functionality

Antibodies can remove pathogens by either neutralising (neutralising antibodies) or recruiting innate immune effector cells (non-neutralising antibodies). Neutralising antibodies renders the pathogen harmless, for example by binding a receptor-binding site required for virus internalisation into the host cell [10]. The effect of non-neutralising antibodies depends on its isotype, but include complement-dependent cytotoxicity, antibody-dependent cell-mediated cytotoxicity and antibody-dependent cellular phagocytosis [11].

The antibody binding site is located at the tip of the Fv and is formed by residues from both the VH and VL. The exact residues in the binding site that facilitate binding are collectively known as the paratope [3]. The molecule bound

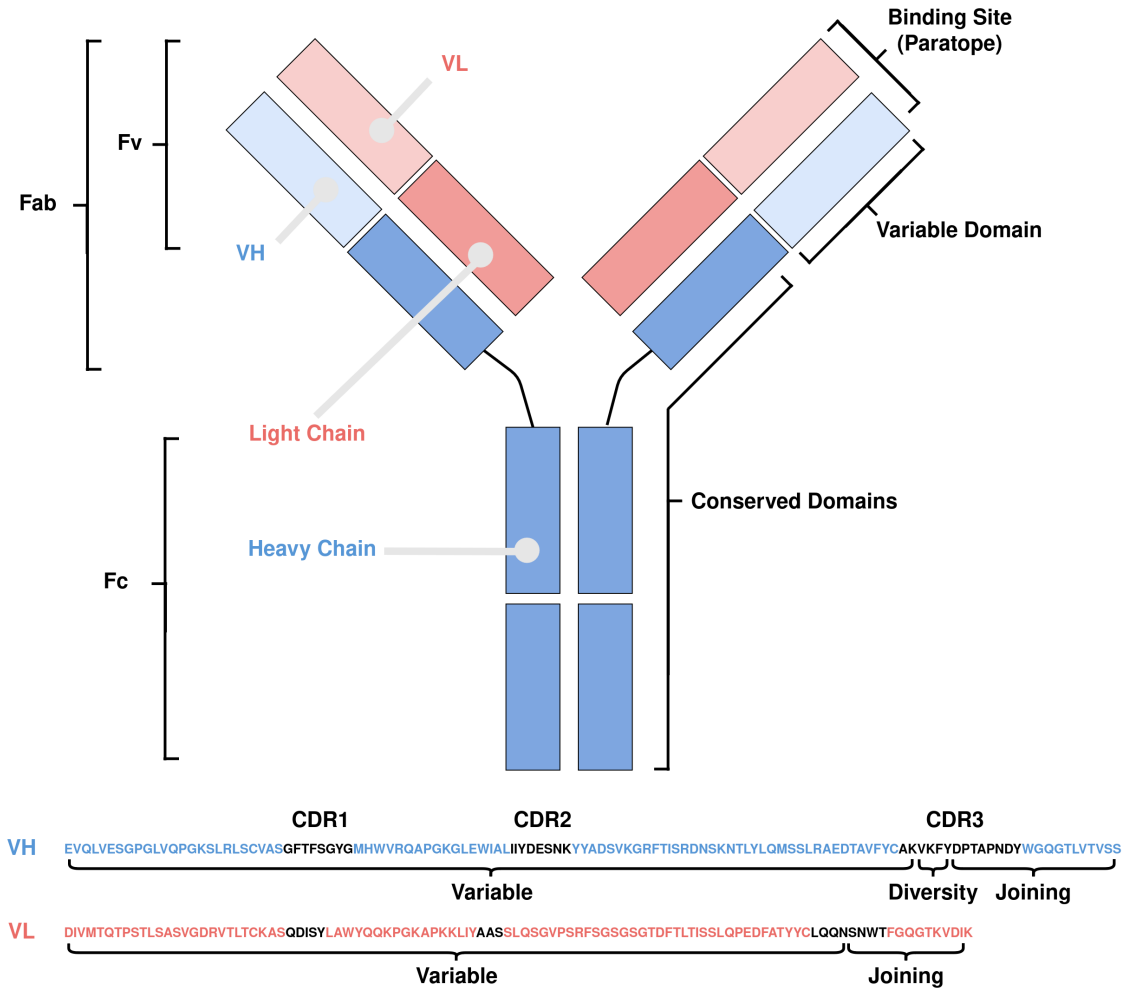

Figure 1.1: Overview of the sequence and structure of an antibody. An antibody has two heavy (blue) and two light (red) chains, with each chain separable into one or more conserved (C) and one variable (V) domain. The domains can also be grouped into the crystallisable fragment (Fc) and two antigen binding fragment (Fab). In the Fab, the Fv is the paired heavy and light V domains, annotated as VH and VL. The Fv contains the binding site, with the complementarity-determining region (CDR)1, CDR2 and CDR3 (black) constituting the majority of the binding site. The VH and VL are made up of multiple gene segments, called the variable, diversity and joining segment.

by an antibody is called the antigen, and the exact bound region is known as the epitope [3]. The majority of the antibody paratope is compromised of residues in three highly variable loops, called complementarity-determining region (CDR)1, CDR2 and CDR3, located in both the VH and VL (highlighted in black in Fig. 1.1). Among the CDR loops, the heavy chain CDR3 is the most diverse in sequence and structure and often the primary contributor to antigen binding [8, 12]. However, the other five CDR loops and the relative orientation of the heavy and light

chain also affect binding [13].

1.2.1.2 V(D)J rearrangement

The high variability of the Fv, especially across the paratope, is what enables antibodies to potentially bind any pathogen. The diversity of the paratope, can be partially explained by the possible combinations of gene segments making up the V domains. While the C domains are encoded by a single gene segment, the V domain is encoded by multiple gene segments rearranged together by a process known as V(D)J rearrangement [8]. As shown in Fig. 1.1, the VL is made up of a variable (V) segment and a joining (J) segment, while the VH also contains a diversity (D) segment. In humans, the segments forming a light chain comes from either the Igκ or -λ loci on chromosomes 2 and 22, respectively, while the segments for the heavy chain come from a single Ig heavy chain (IgH) locus on chromosome 14 [8]. The annotation of functional Ig genes are constantly being revised, but as of Sep. 2023, the human IgH locus has 46, 23, and 6, known functional V, D, and J genes, respectively, and 33 V and 5 J functional genes in both the Igλ and Igκ loci [14]. This is without taking into account allelic variation. The possible combinations of V(D)J segments therefore results in a diverse set of antibodies. Moreover, while CDR1 and CDR2 are both located within the V gene, the CDR3 spans over each of the V(D)J segments, making this loop particular variable in both length and amino acid composition.

V(D)J rearrangement happens in the bone marrow, when pluripotent hematopoietic stem cells differentiate along the B lineage pathway, a sequential set of events leading to the assembly and expression of functional BCRs [15]. The heavy chain undergoes V(D)J rearrangement first, in the pro-B stage, starting with D and J gene recombination followed by V and DJ recombination. Afterwards, in the pre-B stage, the light chain undergoes V and J recombination. The B-cell then differentiates into immature B-cells expressing surface bound IgMs [15]. During rearrangement, incorrect joining of gene segments can lead to junctional diversity. The segments can

be truncated or elongated, with elongation happening when the enzyme Terminal deoxynucleotidyl Transferase (TdT) adds short random stretches of 2-24 nucleotides known as nontemplated (N) regions. Junctional diversity therefore adds even more variability to the CDR3 length and amino acid composition (see Fig. 1.2) [16].

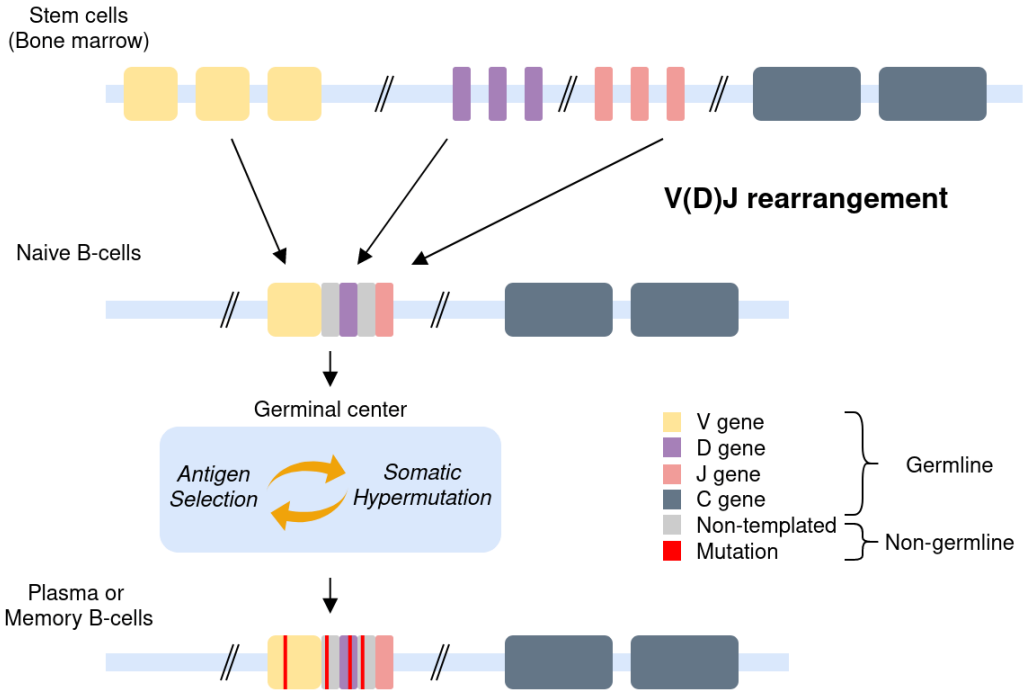

Figure 1.2: Overview of B-cell development and affinity maturation of an antibody heavy chain. In the bone marrow, pluripotent hematopoietic stem cells undergo B-cell development leading to differentiation into naive B-cells. The naive B-cells have undergone V(D)J rearrangement, where a V (yellow), D (purple) and J (pink) gene segment are recombined. Between the gene segments, random stretches of 2-24 nontemplated nucleotides (grey) might have been added [15, 16]. The functional B-cell recepters (BCRs) expressed by naive B-cells, then undergo affinity maturation in the germinal center in secondary lymphoid tissues. Here, non-germline mutations (red) are introduced via somatic hypermutation. The final affinity maturation, the B-cell has either become an antibody secreting plasma cell or a memory B-cell presenting cell surface bound BCRs [17].

Throughout B-cell development, a set of positive and negative selections help remove auto-reactive (BCRs binding host proteins) and non-functional BCRs [15]. Surviving B-cells then migrate to secondary lymphoid tissues, such as the spleen and lymph nodes, where they undergo another negative selection before differentiating into mature naive B-cells, now expressing either IgM or IgD surface bound functional

BCRs [15]. Although new B-cells are continuously generated, the turn-over rate is extremely high. In mice, it has been estimated that of the 10^7 daily generated B-cells in the bone marrow, only about 10% survive to become naive B-cells [18].

1.2.1.3 Germinal center and somatic hypermutation

While V(D)J rearrangement generates a lot of diversity, naive B-cells have had no previous exposure to an antigen and often exhibit a weak unspecific binding [17, 19]. A strong and specific BCR binding is developed during affinity maturation, where mutations away from the V, D and J germline are introduced (see Fig. 1.2). This happens in the secondary lymphoid organs, where B-cells form germinal centers (GCs) together with T helper cells and follicular dendritic cells (FDC). Individual early GCs are composed of tens to hundreds of distinct B-cells, all exposed to the same FDC-presented antigens. In the GC, the B-cells are organised into a dark zone (DZ) and a light zone (LZ), and because of the environment the B-cells have cell cycles as short as 5-6 hours. In the LZ, B-cells binding to the antigen presented by FDCs proliferate, resulting in a competitive expansion [17]. Antigen-activated B-cells then move to the DZ where they undergo somatic hypermutation (SHM). SHM is a process where the activation-induced deaminase (AID) enzyme deaminates cytosine to uracil directly on the DNA, creating U:G mismatches that upon repair lead to point mutations at roughly 1 million times the base mutation rate. AID also facilitate class switch recombination from IgM and IgD into IgA, IgE, and IgG [17]. The mutated B-cells then move back to the LZ, where B-cells with higher-affinity BCR mutations are positively selected via further proliferation followed by re-entry to the DZ. Auto-reactive or low-affinity binding B-cells undergo induced apoptosis, leaving only 10-30% of B-cells to re-enter the DZ [17].

GC B-cells which have achieved a high antigen-specific affinity are favoured for exiting the GC and differentiation into into plasma B-cells or memory B-cells. Short-lived plasma B-cells can secrete 10^3 high-affinity binding antibodies per

second, initiating a fast removal of the pathogen, but only circulate for 3-5 days [20]. Long-lived plasma B-cells also non-proliferating but with a life span of several months to lifetime. Long-lived plasma B-cells function as an immediate defense to re-exposure of an antigen, but if it is insufficient to clear the pathogen, memory B-cells can upon antigen activation differentiate into short-lived plasma B-cells or re-enter a new GC for further affinity maturation [17].

For functional antibodies, the majority of the SHMs occur in the CDRs, especially the CDR3, further diversifying the V domains [7, 8]. However, SHMs also occur outside the CDRs and can have crucial effects on binding [19]. At any given time, there is an estimated $\sim 5 * 10^9$ unique antibody sequences circulating the blood of a healthy human individual, highlighting the immune systems reliance on antibodies ability to neutralize pathogens for protection against diseases [21].

1.2.2 Antibody annotation

Proteins are often compared by aligning their sequences, typically using specially designed sequence alignment methods [22]. While these methods often create good alignments, the alignments are less reliable for less conserved regions. Antibodies have conserved framework regions which are easily aligned, but their variable CDRs have proved difficult to align using conventional methods [23].

Antibody numbering schemes were developed to address this challenge when aligning the variable domains of antibodies. Instead of aligning the whole VH or VL, these schemes use the conserved framework regions as reference points to number each residue in an antibody sequence. Several numbering schemes have been developed, all with their own definition of the framework and CDRs [5, 24, 25]. The International Immunogenetics Information System (IMGT) numbering scheme is widely used, numbering residues in both the VH and VL in a standardised way. The residues in the VH are numbered from 1-128 and in the VL from 1-127, and insertions are included with Latin letters [5]. Antibodies can readily be numbered

using various computational tools, like ANARCI [26], which uses hidden markov models trained on pre-numbered IMGT germline sequences to predict numbering for antibody VH or VL sequences.

1.3 Antibody Repertoires

The entire set of antibodies produced in an individual is called their antibody repertoire, also referred to as Ig repertoire or BCR repertoire [27, 28]. Antibodies against specific diseases should only be present in individuals that have been exposed to the disease, and only in high numbers if the exposure was recent [29]. The BCR repertoire can therefore represent the immune state of an individual [30], and can be used to investigate effects from diseases or vaccination responses and for finding and developing potential therapeutics [30, 31].

1.3.1 Repertoire sequencing methods

Immense effort has been put into developing methods for BCR repertoire sequencing (BCR-seq). However, with the large number of antibody sequences estimated to be in a human antibody repertoire, it is infeasible at present to sequence every single antibody in the repertoire [2, 3]. The constant generation and removal of B-cells, also makes BCR repertoires extremely variable, even for the same individual [18]. Nonetheless, several high-throughput BCR-seq techniques have been developed. This has allowed researchers to routinely to capture and study a snapshot of the immune state, with antibody sequence data currently available from various species [32–34] and from individuals with differing diseases [35, 36]. Here, we will introduce two commonly used approaches for sequencing single chains and paired VH-VL.

1.3.1.1 Unpaired VH and VL sequencing

Although it is currently impractical to sequence every antibody in the repertoire, BCR-seq approaches based on Next-Generation Sequencing (NGS) techniques can

sequence millions of single chain sequences per experiment, offering a snapshot into the antibody repertoire [21, 28, 37]. Before sequencing, a sample of B-cells needs to be prepared. B-cells are usually isolated from peripheral blood, as it is the least invasive source of B-cells, although the majority of disease related memory and plasma B-cells are expected to be in the primary and secondary lymphoid organs [17]. The B-cells are then lysed, resulting in a pool of unpaired antibody heavy and light chain mRNA fragments. The mRNA fragments are then reverse transcribed before the resulting complementary DNA is amplified using antibody germline-specific primers [28]. The sample is now ready for sequencing. A current limitation of the commonly used NGS platforms, including Illumina MiSeq, HiSeq, and Roche 454, for BCR-seq, is the possible length that can be sequenced. Depending on the platform, a maximum of 300–400 base pairs (bp) can be sequenced, but the variable domain alone is already 300-400 bp, meaning the full antibody cannot be reliably sequenced. Most studies therefore only sequence the VH and VL domains, or even only parts of these domains, instead of the full antibody sequence [21, 28, 37]. As most of the variability and the binding site is located in these domains, this is usually not a problem for most studies. The raw nucleotide sequences, stored in FASTQ files, can then be aligned to known germline sequences using computational tools, to obtain the final sequence for the V domains [38].

Sequencing errors can arise during amplification or from sequence misreads and misalignments, with error rates increasing the longer the sequence read is. To counter this, Illumina MiSeq provides reads from both ends, known as paired-end. When the reads are assembled, a quality score for each base is then used to correct disagreeing bases [28]. Errors can also be corrected using unique molecular identifiers (UMIs). Here, a UMI is added to each sequence during reverse transcription, which can then be used to group reads from the same antibody after sequencing and be used to correct for errors [28].

1.3.1.2 Paired VH-VL sequencing

Unpaired heavy and light chains can be used to investigate immune states, immune repertoire mining (see Section 1.3.2), and as source of functional antibodies for machine learning. However, to achieve a more accurate picture of the antibodies present in a repertoire, the VH-VL pairing information needs to be conserved. Over the last few years various sequencing techniques have been developed that focus on mapping sequenced genomic or transcriptomic data to individual cells, known as single-cell sequencing. Single-cell sequencing is similar to NGS, with the addition of barcoding on a cell level, allowing all genetic material from the same cell to be grouped together. For BCR-seq, the widely used 10X platform has recently been used to sequence ∼1.5 million paired VH-VL [39]. While still a far smaller snapshot of a repertoire than unpaired BCR-seq, the technique is continuously improving, making it an exiting area for the future.

1.3.2 Immune repertoire mining

A technique which has shown promise for utilizing the vast amounts of available BCR-seq data to investigate antibodies is immune repertoire mining [40–43]. In immune repertoire mining, an antibody of interest is compared against natural antibody repertoires to find identical or similar antibodies. This is useful for finding mutations which could improve binding affinity, as antibodies with few differences in the paratope, derived from patients with the disease of interest, might yield better binders. Finding antibodies in nature with an identical paratope is also a powerful method for getting insight into which mutations could improve an antibody's developability profile or reduce its immunogenicity, without changing its binding properties [40–44].

1.3.3 Repertoire sequence databases

Continuous improvements of BCR-seq methods and increased adoption by research labs means that the amount and diversity of public antibody sequences is rapidly

increasing [21, 37]. There are now almost 100 unpaired BCR-seq datasets available, which has grown from hundreds of thousands of unpaired sequences per sample [45, 46], to hundreds of millions of unpaired sequences [21, 47]. The data is also diverse, containing antibodies from a variety of different hosts, diseases, treatments and cell types [30]. Moreover, with the introduction of single-cell RNA sequencing techniques, paired VH-VL sequences are becoming more readily available [39, 48, 49].

The use of BCR repertoire data from multiple studies, allows larger and more complete analyses to be carried out, but is hindered by nonstandard data formats. To counter this, the Adaptive Immune Receptor Repertoire (AIRR) Community proposed the Minimal Information about AIRR (MiAIRR) standard [50], which outlines a set of minimal required reported information. However, even with MiAIRR, a wide range of different processing pipelines still exist, complicating direct comparisons [51]. Furthermore, the standard data release contains only the raw FASTQ files, therefore, to utilize the available data, it is often necessary to carry out extensive processing [51].

As a response to this, in 2018 Kovaltsuk *et al.* [52] created the Observed Antibody Space (OAS) database. This was a database of unpaired VH and VL antibody protein sequences, derived by processing 55 unpaired BCR-seq datasets containing 600 million sequences [52]. Related efforts include ImmuneAccess [53], PIRD [54] and RAPID [55], and the AIRR Data Commons [56] (ADC), a network of geographically distributed AIRR compliant repositories accessible through a single API. These databases are a good source of annotated antibodies and include access to antibody visualization and analysis tools. ImmuneAccess holds a large set of annotated CDR3 sequences and RAPID identically processed human antibodies. PIRD and ADC contain large collections of annotated BCR-seq, processed and deposited by independent researchers.

The gathering of large sets of antibody data has enabled us to use computational approaches, such as immune repertoire mining (see Chapter 3), to investigate and interrogate antibodies in new ways and improve our understanding of antibodies. Furthermore, it provides large set of functional antibodies, ideal for training deep learning models (Chapter 5-6).

1.4 Antibodies as Therapeutics

Antibodies with high specificity and affinity are a valuable tool in many areas of medical and scientific research. For example, antibodies are used routinely in diagnostic assays [57], and to better understand the effects of vaccination on the immune system [58]. Antibodies are also by far the most successful type of biotherapeutic, evident by the more than 170 antibody therapeutics that are in regulatory review or approved for clinical use to date [31, 59]. However, therapeutic antibody development is a complex task, traditionally taking years, with many requirements for the final antibody. An antibody needs to not only bind strongly and exclusively to the targeted disease, but it also needs to adhere to numerous developability requirements, such as being stable, soluble, having low viscosity, having a long half-life in the bloodstream, not trigger an immune response, and being resistant to aggregation, modification and degradation [60, 61]. Retaining function whilst removing these undesirable properties, collectively known as developability issues, is what makes it a costly and challenging process to successfully bring an antibody to the market. In this section we will introduce the development of therapeutic antibodies and highlight some of the experimental and computational methods used.

1.4.1 Affinity measurement

Affinity is the most important antibody property, being vital for using antibodies as therapeutics. There exists numerous methods for measuring binding affinity, with most methods trying to measure or estimate the equilibrium dissociation constant (K_D) or half maximal inhibitory concentration (IC50). K_D is the ratio between the

antibody dissociation rate and association rate to its antigen [62], while IC50 is the antibody concentration required to bind 50% of the antigen [63].

Qualitative methods like isothermal calorimetry [64] and surface plasmon resonance [62] offer precise affinity measurements, but are limited by their low-throughput nature. To overcome this when screening larger subsets of antibodies, high-throughput methods like enzyme-linked immunosorbent assay (ELISA) [62], bio-layer interferometry (BLI) [65], and deep mutational scanning (DMS) [66] have been developed. However, the increased throughput comes at the cost of less accurate affinity measurements.

1.4.2 Lead selection

Following the identification of a target of interest, the next step in antibody development is the antibody discovery or lead selection phase, with the purpose of finding target-specific antibodies. Although affordable high-throughput affinity measurement techniques are becoming more available, the vast number of antibodies still makes it impossible to test every antibody for binding. To counter this, different screening methods have been developed to more quickly identify subsets of possible hits. This smaller subset of hits can then be verified using the measuring methods mentioned in Section 1.4.1. These screening methods can be both experimental or computational, or a combination [66–69].

1.4.2.1 Experimental selection

Traditionally, antibody binders were found by animal immunization with an antigen. Antigen-specific binders can then be sourced and identified from the hosts plasma or isolated B-cells [67]. Alternatively, B-cells can also be isolated from humans following natural infection [70]. This approach benefit from *in vivo* affinity maturation, resulting in the generation of strong and specific binders. However, non-human antibodies often cause immunogenicity in humans and therefore needs to be further

engineered to be more human-like, also process also known as humanization [71]. To counter this, transgenic mice models, like XenoMouse [72], have been genetically modified to produce human-like antibodies, reducing the need for humanization [67]. Nonetheless, *in vivo* antibody lead selection heavily relies on the hosts antibody development and affinity maturation to generate optimal binders, and tend to produce antibodies favoring certain epitopes [67, 73].

Biopanning of phage displays is another method for finding antibody leads [68]. Instead of using a hosts immune system to generate strong binders, it screens a large antibody library for binders. It takes advantage of phages ability of displaying antibodies from inserted foreign DNA on their surface. Biopanning is then used to isolate phages presenting antigen-specific antibodies [68]. The libraries usually contains 10^8-10^{10} unique antibodies. Antibody phage display technologies have several advantages such as avoiding working with animals, only selecting human antibodies and making it possible to screen for antibodies against toxic antigens. However, phage displayed antibodies are not glycosylated, often have developability issues, and a weaker and less specific binding than *in vivo* derived antibodies [68]. Selected antibody leads therefore usually have to undergo further antibody engineering.

1.4.2.2 Computational selection

As a consequence of the expensive and time-consuming lead selection using *in vivo* and *in vitro* approaches, there has been an increasing focus on developing computational methodologies to help the discovery process [3]. For this, machine learning (ML) algorithms have gained traction, as they can learn complex underlying patterns from a given dataset and utilize them to predict with [74]. However, accurate prediction of which antibody will bind to which antigen is very challenging, due to limited datasets of known antibody-antigen pairings and the variability in both paratope and epitope [75]. Nonetheless, training ML models on a set of experimentally proven binders for a single target have seen some success [66, 69].

ML tools have also been used to predict antibody self-association and non-specific binding [76].

Immune repertoire mining can also be used as a screening approach. Richardson *et al.* [43] showed this by clustering antibodies binding the same antigen based on their paratope, called paratyping. The paratope used for clustering can be predicted using various ML tools [77–79]. Knowing a single binder therefore allows you screen for new potentially better antigen-binding antibodies, by utilizing the vast amounts of BCR repertoire data (see Chapter 3).

While antibodies discovered using computational tools still need to have their binding verified, they enable the screening of a much larger space of antibodies than possible using *in vivo* and *in vitro* approaches, with a fraction of the cost. Although antibodies selected from BCR repertoires have the benefit of being functional antibodies, computationally selected antibodies may also need to undergo optimization to solve developability issues.

1.4.3 Developability issues

After the selection of strong and specific antigen-binding antibodies, suboptimal properties, also known as developability issues, needs to be removed by mutating away the issues. Developability issues include undesirable post-translational modification sites [80], potential antibody aggregation [81], and the previously mentioned immunogenicity. However, given the huge space of antibody sequences and possible mutations, it is complicated to find the correct mutations which improve developability. Additionally, any mutation might give rise to other developability issues or reduce binding. Antibody optimization is therefore a multi-dimensional optimization problem. In this section, we will introduce some of the often encountered developability issues.

1.4.3.1 Immunogenicity

The immunogenicity of a foreign molecule, including non-human derived antibodies, is its ability to provoke an immune response. The immune response can potentially lead to a severe inflammatory reaction. Immunogenicity is therefore a common issue with antibodies isolated from immunized animals [82]. Immunogenicity can be reduced by making the antibody more human and therefore not recognised by the immune system. Chimerisation is a method where the Fv is grafted onto a human antibody, thereby reducing how much of the antibody is non-human, while keeping the binding site located in the Fv [83]. Another approach is humanisation, which focuses on making the antibody more human-like, by mutating to individual residues often seen in human sequences [71]. Several ML tools are available to predict the humanness of an antibody sequence and suggest mutations which can potentially reduce immunogenicity [71, 84].

1.4.3.2 Aggregation

Antibody aggregation can be caused for a number of reasons. Hydrophobic regions exposed on the antibody surface can induce self-association and aggregation due to hydrophobic interactions. Imbalance in charge distribution or the ionic environment can also contribute to antibody aggregation. Increases in temperature can cause partial unfolding of antibodies with low thermal instability, leading to exposed hydrophobic areas and aggregation. Additionally, deviations from the optimal pH can induce conformational changes, also reducing antibody stability [85].

Various experiments can be used to screen for these issues, such as hydrophobic interaction chromatography (HIC) and standup monolayer adsorption chromatography (SMAC) for hydrophobicity, melting temperature (Tm) for thermal stability, and accelerated stability (AS) for aggregation [86]. Computational tools have also been developed to highlight potential issues, such as the therapeutic antibody profiler [87].

1.4.3.3 Post-translational modifications

Post-translational modifications (PTMs) are chemical changes to a protein after it has been expressed. Certain PTMs, such as isomerization, oxidation and deamidation, are irreversible and can severely change the behavior of antibodies, such as its binding when happening in the CDRs [88]. For instance, deamidation occurs when an asparagine is converted to aspartic acid or isoaspartic acid, amino acids with completely different properties [88]. It is therefore preferential to reduce PTMs by removing sites where they can occur [80]. Traditionally, specific motifs have been used to predict certain PTMs, however; ML tools trained for predicting PTMs have shown to achieve higher accuracy [80].

1.4.3.4 Manufacturability issues

Manufacturability issues are a subset of developability issues that focuses on issues related to an antibody's production and includes expression and purification issues [89]. Consistently expressing high-yields of antibodies is important in order to reduce production costs. Production yield is also affected by how easy the antibody can be purified from host cell proteins, endotoxins, and other contaminants without affecting the antibody [89].

1.4.4 Immunoglobulin-like therapeutics

Monoclonal antibodies are currently the most used antibody therapeutic, however; other variants with different benefits are also being explored [31]. Bispecific and trispecific antibodies are antibodies with multiple distinct antigen-binding sites on the same antibody, able to target multiple epitopes or bridge targeted antigens with the immune system [31]. Instead of using the whole antibody, individual Fabs can also be used by itself, allowing for a simpler therapeutic to produce. The single-chain Fv region (scFv) is even smaller, consisting only of the VH and VL with a linker [31]. Nanobodies are even smaller, consisting of only the VH, and can be derived from camelids and certain fish species. They have a unique sequence

composition, being distinct from human VH gene loci and having a CDR3 loop usually longer than human CDR3s [90].

1.5 Transformers

Machine learning (ML), a subfield of artificial intelligence, focuses on the development of algorithms that can learn to perform tasks autonomously. Since their introduction by Vaswani *et al.* [91], transformers, and variants hereof, have become central in many ML algorithms, achieving state-of-the-art (SOTA) in numerous fields like natural language processing (NLP), computer vision, and audio processing [92]. Transformers are further categorized as deep learning algorithms, because of their complexity and depth relative to other ML methods like Random Forest [93].

Deep learning algorithms are trained using backpropagation [94], which computes the gradients of the model's trainable parameters in relation to a loss function. During training, these parameters are then iteratively updated with gradient descent, using the model's learning rate and derived gradients, with the purpose of minimizing the loss between the output predictions and ground truth [94]. In this section, we will give an overview of the essential components of the transformer architecture, introduce transformer-based language models, and their present applications in protein and antibody design.

1.5.1 Transformer architecture

The original transformer, also known as the vanilla transformer, was initially introduced for sequence-to-sequence machine translation, performing better and faster than the otherwise used recurrent neural networks. It has an encoder-decoder architecture, with the encoder processing an input, i.e. English text, into a context-aware representation and the decoder using the representations to predict the output, i.e. German translation, in an autoregressive manner [91].

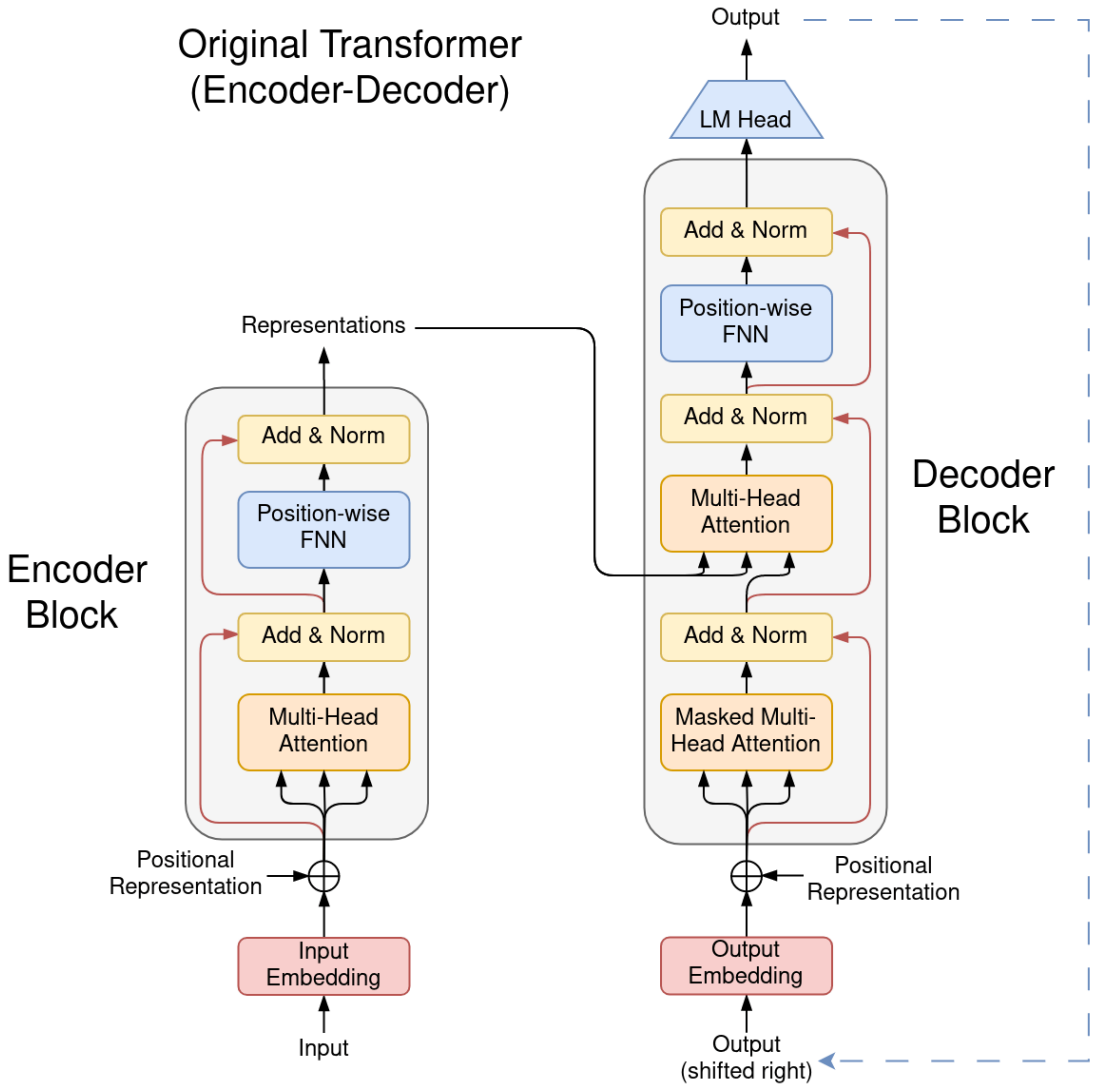

Figure 1.3: Overview of the original transformer by Vaswani *et al.* [91]. The model takes an input sentence and then autoregressively translates it into another language. The input is first embedded before position information is added to the embeddings. The embeddings then enter an encoder block. Encoder blocks consists of a multi-head attention (MHA) layer (orange) followed by a point-wise feed-forward neural network (FNN) layer (blue). After each layer, residual connections (red) are added before a layer normalisation (yellow). The input and output dimension of the encoder block is identical, meaning multiple encoder blocks can easily be stacked. The encoder block output is the learned representation, which is then used for the decoder block. The input to the decoder block is the previously predicted output, which is embedded before having position information added. The decoder block consists of a MHA layer followed by another MHA layer where the key and value are the encoder representations, and finally a point-wise FNN. Like the encoder blocks, the decoder blocks have residual connections and layer normalisation after each layer. The next token (word) is then predicted with a language model (LM) head, which consists of a linear layer followed by a softmax function.

The encoder and decoder consists of blocks, which can be stacked to increase

the size of the model. The blocks feature multi-head attention and a position-wise feed-forward network (FFN), with slight differences between the encoder and decoder block (see Fig. 1.3) [91].

1.5.1.1 Multi-head attention

Attention is the mechanism that allows the transformer to attend to different parts of the input data simultaneously. Attention takes as input three matrices, called query (Q), key (K) and value (V) (see Eq. 1.1), with each matrix consisting of a vector (embedding) for each word in the input sentence.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}}) \cdot V \tag{1.1}$$

The Q, K and V are first linearly transformed with trainable weight matrices. The matrix product of the Q and transposed K is then calculated and scaled down by the square root of the dimentionality of K, $\sqrt{d_k}$, to mitigate gradient vanishing issues when applying the softmax function. The derived attention weights are used to decide which values in V to attend to. The final attention score is therefore the matrix product of the attention weights and V [91].

Instead of a single set of attention weights, multi-head attention has several sets, allowing it to attend to different patterns simultaneously. This is done by mapping the word embeddings of Q, K and V into a smaller vectors using weight matrices before computing the attention (see Eq. 1.2) [91].

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{1.2}$$

The multi-head attention is then derived by concatenating the attention heads and transforming it with trainable output weights, W^O (see Eq. 1.3) [91].

$$MultiHeadAttention(Q, K, V) = concat(head_1, ..., head_h) \cdot W^O \tag{1.3}$$

It is generally desired to have the same dimension for the input and the resulting multi-head attention in order to conserve dimension in the transformer blocks,

allowing them to easily be stacked. The number of heads used therefore needs to be the word embedding dimension of Q divided by the dimension of the heads [91].

Attention is used slightly differently in the transformer encoder and decoder block. For the encoder, the Q, K and V are all derived from the same input embedding, and is referred to as self-attention. In the decoder block, the first attention layer is also self-attention, but includes masking of unseen output thereby enabling parallel training. The second attention layer is referred to as cross-attention, as it uses the output from the previous layer in the decoder as the Q and the encoder representation as the K and V (see Fig. 1.3) [91].

1.5.1.2 Transformer blocks

The encoder consists of a multi-head self-attention layer followed by a position-wise feed-forward neural network (FNN) which operates individually on each position, thereby treating them separately (see Fig. 1.3). The position-wise FNN has a linear layer that expands the word embedding dimension, usually to four times the input dimension, followed by an activation function like ReLU, and another linear layer reducing the dimensionality back to the initial input dimension [91]. To allow for a deep architecture it has residual connections (highlighted in red in Fig. 1.3) [95] and layer normalization [96] after each layer [91].

The decoder also starts with a multi-head self-attention layer followed by a multi-head cross-attention layer and a position-wise FNN. Like with the encoder, the decoder also has residual connections and layer normalization after each layer [91].

1.5.1.3 Tokenization and input embeddings

Tokenizing and embedding input text is a common practise in NLP that usually leads to improved results [97]. Each unique token is embedded with a unique learned embedding. Vaswani *et al.* [91] used a 512-dimensional vector, meaning each token is converted into 512 numerical values. Because of the number of existing words, their conjugations and the arrival of new words, it is not possible to treat

every unique word as an individual token. Sentences are therefore tokenized, by being split into a selected set of words, subwords and lastly characters. A common approach is to train a tokenizer using a method like byte pair encoding [98]. The set of tokens used for a model is called their vocabulary.

1.5.1.4 Position representations

Unlike recurrent and convolutional networks, transformers are invariant to sequential ordering. For data like text, order is extremely important for its context and positional information is therefore required to learn its context. In the original transformer, this was done by adding positional information to the input embeddings, thereby encoding the absolute position of each element in a sequence. These absolute position representations can be both static [91] or learned [99]. Positional information can also be incorporate into the attention layer. Relative position representations are embeddings added to the values and keys [100], or just the keys [101] in the attention layers, thereby emphasizing pairwise token relationships over individual positions. Rotary Positional Embedding (RoPE) is a position embedding that combines absolute and relative representations, and has been shown to often outperform other position representations [102, 103].

1.5.1.5 Language model head

The language model head is a linear layer followed by a softmax function. The purpose of the linear layer is to map the output embeddings from the transformer block back into the used vocabulary. The softmax function then produce a probability distribution over the vocabulary. The token with the highest probability is then selected as the prediction [91].

1.5.2 Transformers as pre-trained language models

Transformers have become indispensable in the NLP field. Their main use is as the fundamental building blocks of pre-trained language models (LMs), which themselves have become SOTA for language tasks, like language translation, question answering

and human-like text generation. An LM is a model designed to predict sequences of words, and it being pre-trained, means it was first trained on a problem with abundant data. The pre-trained LM can then be used as is or be further trained, also known as fine-tuned, for a specific related downstream task (see Fig. 1.4) [99]. This is known as transfer learning and has shown to work particularly well for tasks with few labelled data. It is however often a challenge to find a suitable related task with enough labelled data required for pre-training. A common solution is to instead train on vast amounts of unlabelled text, readily sourced from the internet, in an unsupervised manner [99].

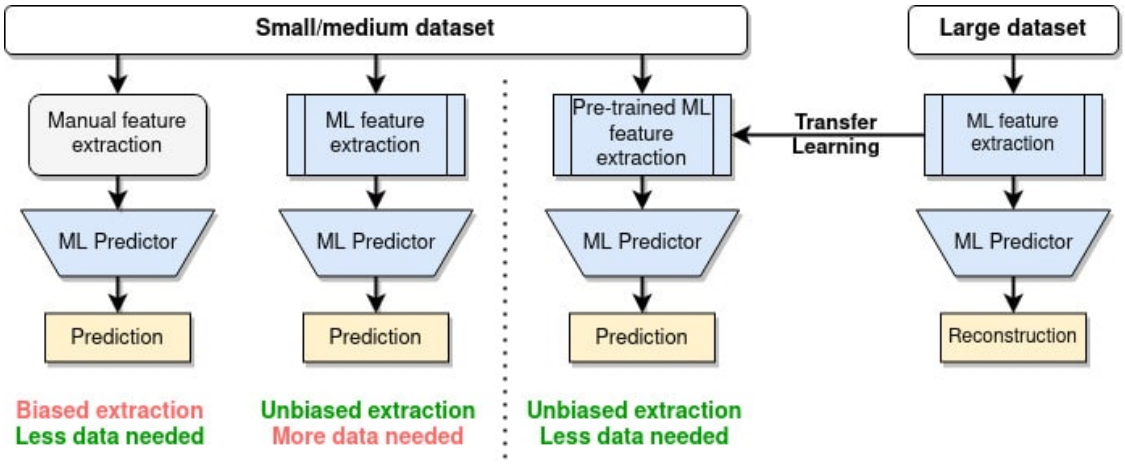

Figure 1.4: Overview of different approaches for training machine learning models to predict tasks with limited data available.

ELMo, based on bidirectional Long Short-Term Memory (LSTM) networks, was the first model pre-trained on a large corpus of unlabelled text with the purpose of learning context-aware word embeddings useful for other tasks [104]. The embeddings generated by ELMo achieved SOTA on many language tasks, sparking interest in pre-trained representations. LSTMs are recurrent networks and therefore process input sentences one word at a time. In contrast, transformers are parallelizable, allowing for larger and faster trained models. While ELMo consists of $\sim$100 million trainable parameters, "small" transformer-based LMs have $\sim$7 billion trainable parameters with the largest LMs having trillions [105].

LSTMs also struggle with long sequences, unable to keep context received early in a sentence for predictions towards the end of the sentence [106]. As attention layers attend to the whole sentence when predicting each word, they do not have the same issue [99]. The first transformer-based pre-trained LM for learning context-aware word embeddings was the Bidirectional Encoder Representations from Transformers (BERT) model introduced by Devlin *et al.* [99]. BERT only uses the encoder transformer blocks with a language model head (see Fig. 1.5) and is trained as an autoencoder using masked language method (MLM) and Next Sentence Prediction (NSP) training objectives. NSP is the prediction of whether two input sentences are connected, but was removed from later refined versions of BERT [107]. For MLM, a random set of input tokens are masked and the model is then tasked to predict the masked tokens, allowing the training data to be unlabelled. After pre-training, the language model head is removed and the representations are used as context-aware embeddings for downstream tasks. BERT showed large improvements over LSTM based models, and was therefore soon followed by a number of improved BERT variants, such as the Robustly Optimized BERT pre-training Approach (RoBERTa) [107].

Pre-trained LMs using the full transformer architecture, both encoder and decoder block, also exists, like BART [108] and T5 [109]. Notably, T5 introduced task-specific text prefixes to refine guidance for downstream tasks. However, architectures using either only the encoder or decoder has had more success. Decoder-only LMs like the Generative Pre-trained Transformer (GPT) models and variants (see Fig. 1.5) [105, 110], have revolutionized language generation. Instead of being trained with MLM, they are trained to predict the next word, similarly to the original transformer. Further training with reinforcement learning from human feedback (RLHF) has led to models like ChatGPT, known for their human-like responses [111]. While GPT-like models are ideal for text generation, BERT-like models excel in language understanding [99].

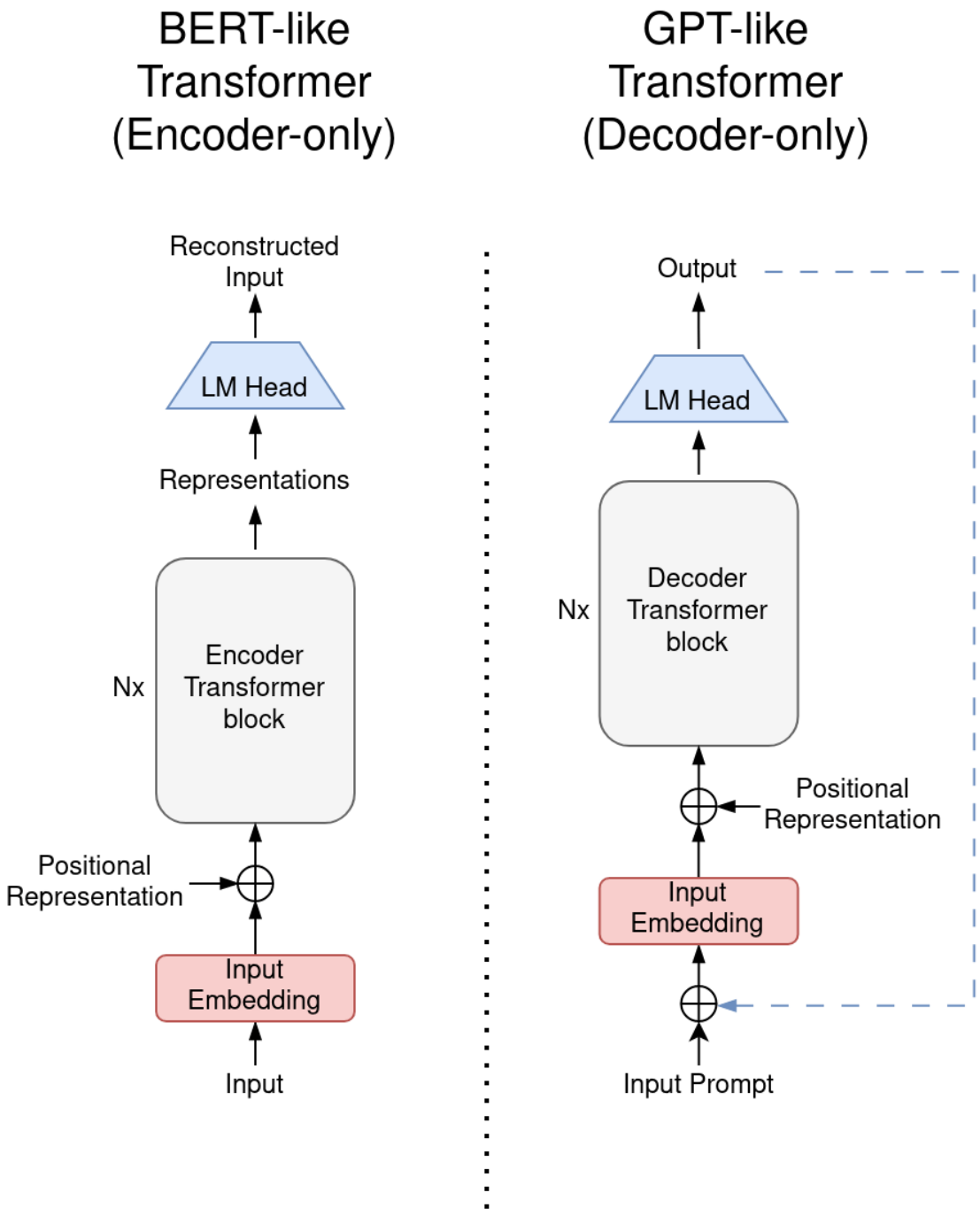

Figure 1.5: Overview of the BERT-like and GPT-like transformer variants. The BERT-like models are commonly trained using masked language modeling, with the purpose of creating a model which can generate context-aware embeddings for downstream tasks [99, 107]. GPT-like models are trained to predict the next word, resulting in models which can generate human-like responses [105, 111].

1.5.3 Protein language models

Protein sequences and natural language share many similarities, both comprised of basic units, in the form of amino acids and words, and inherent syntax. The meaning of a sentence is determined by the order of the words and the order of amino acids determines a proteins structure and function. This parallelism has motivated the use of protein LMs for improved prediction of protein tasks. In 2019, Alley *et al.* [112] introduced UniRep, an LSTM-based autoencoder trained to reconstruct protein sequences. They demonstrated the first use of a protein LM for improving performance on protein prediction problems where data was limited. However, LSTMs are known to struggle with long sequences, which is a common occurrence for proteins [106]. BERT-like models for proteins like ProtBert [113] and Meta's 650M parameter ESM-1b [114] therefore soon followed. ESM-1b, and its newest iteration ESM-2 [115], have proven that transformers and transfer learning can create SOTA predictions on various protein tasks [114, 115]. For protein sequence generation, GPT-like protein models, like ProtGPT2 [116] and ProGen2 [117], have shown the potential of generating novel functional protein sequences.

1.5.3.1 Protein representations

Protein representations from encoder-only protein LMs have been used successfully for many different tasks. Residue representations have been used for SOTA prediction of epitopes [118], secondary structure [119], protein structure [115] and various other protein tasks [120, 121]. Sequence representations have been used for protein clustering and annotation [122], and protein homology search [123].

1.5.3.2 Protein guided evolution

For BERT-like models trained on proteins, the head model generates likelihoods of each amino acid at each position. Hie *et al.* [124] proposed that these likelihoods could be used to visualize the evolution of proteins. New mutations often had a higher likelihood, so from the difference in likelihood between two variants, a mutation

direction could be determined. In another paper, it was explored how selecting mutations based on higher likelihoods could help better select mutations improving antibody binding [125]. Ideally, as protein LMs have learned the semantics of proteins, they will predict high likelihoods for mutations which are functional. This heavily reduces the possible mutational space, increasing the chances of selecting mutations which might also improve properties of interest.

1.5.4 Antibody language models

General protein LMs are possible because of the large amount of available protein sequences. Although, training an LM for a specific class of proteins could potentially further improve performance, it is usually not possible because of the required amount of training data. As the antibody community has worked at the sequencing and centralisation of the sequenced data into databases like the Observed Antibody Space (OAS) (see Chapter 2), enough antibody sequences might be available to train an antibody-specific model, creating representations that could potentially further improve antibody design [126]. This idea has led to immense effort in designing and pre-training both BERT-like [84, 127–130] and GPT-like [117, 131] antibody-specific LMs.

1.5.4.1 Antibody representations

Several antibody-specific LMs trained on the data within OAS have been published and demonstrated their use for antibody-specific tasks. Leem *et al.* [127] introduced AntiBERTa, a RoBERTa inspired model trained on data from OAS, and showed its ability at predicting the paratope. Similarly Ruffolo *et al.* [128] introduced AntiBERTy, and showed its use for predicting antibody binding, and in another paper, Ruffolo *et al.* [132] used the AntiBERTy embeddings to train IgFold, a model for rapid antibody structure prediction. AntiBERTa and AbLang (see Chapter 5) also showed how the antibody representations can be used to readily cluster antibody sequences based on germline. However, it is uncertain whether representations derived from general protein or antibody-specific LMs are most

suitable for predicting antibody tasks. Currently, different studies argue for either general protein [125] or antibody-specific [126, 130] LMs.

1.5.4.2 Antibody language model guided optimization

While antibody representations can be used to improve performance at predicting antibody-specific tasks, the produced likelihoods of each amino acid at each position quickly showed promise for antibody optimization. Prihoda *et al.* [84] introduced Sapiens, a BERT-like model trained only on human antibody sequences from OAS, and used it to humanize antibody sequences by substituting low likelihood amino acids to high likelihood ones. Similarly, AbLang, showed how antibody-specific LMs can correctly restore fragmented antibodies better than the standard baseline and the general protein LM, ESM-1b. Outperforming the germline-baseline, indicates antibody-specific LMs learn more than the underlining data distribution, and outperforming ESM-1b, supports the potential of training antibody-specific models for antibody design.

In this thesis, the key scientific challenges to be addressed are:

- Utilising the vast and continuously expanding publicly available antibody data to derive novel antibody insights.
- Leveraging the state-of-the-art deep learning algorithms from the natural language processing field to advance computational antibody therapeutic development.

In response to these challenges, this thesis aims to:

- Collect, prepare, and investigate relevant antibody sequence data for computational methods and the training of machine learning models.
- Explore the potential of BERT-like language models for designing better antibody therapeutics.

1.6 Thesis outline

In Chapter 2, we will start by describing our update and expansion to the Observed Antibody Space (OAS) database. This database is a collection of billions of unpaired and 1.5 million paired antibody sequences, as of Sep. 2023. Its creation has enabled the investigation of vast amounts of functional antibody data from various studies, exploring disease and vaccine states, extended immune repertoire mining and has enabled the training of antibody specific LMs.

In Chapter 3, we introduce KA-Search, a method for rapidly and exhaustively searching billions of antibody sequences for similarity. KA-Search allows us to readily perform immune repertoire mining on OAS, which we show can be used for finding new possible SARS-CoV-2 binders.

A limitation of OAS, is its lack of antigen-specific information. In an effort to address this, in Chapter 4 we introduce The Patent and Literature Antibody Database (PLAbDab), a collection of paired antibodies with antigen-specific information available. Searching PLAbDab with KA-Search can help discovery which antigen a given antibody might bind.

In Chapter 5, we utilize the vast amounts of functional unlabelled antibody sequences in OAS to train an antibody-specific BERT-like LM. We show how it can restore missing residues better than a general protein LM and better than restoring using a germline-baseline, indicating its use for antibody optimization.

In Chapter 6 we investigate how the germline-bias in OAS affects antibody-specific LMs, and explore how to we can overcome this problem, hopefully leading to antibody-specific LMs which suggest more relevant mutations.

Finally, we conclude this thesis in Chapter 7, summarizing the most relevant findings and future work.

2

Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences

Contents

2.1 Chapter abstract

The rapid emergence and widespread adoption of BCR-seq have resulted in the generation of massive volumes of sequencing data from numerous studies (see Introduction 1.3.3). However, significant challenges persist in accessing, processing, and comparing this data due to the nonstandard formats or raw FASTQ files in which it often exists (see Introduction 1.3.1.1). In an attempt to address this issue, Kovaltsuk et al created the Observed Antibody Space (OAS) database. Nevertheless,

with the subsequent surge in available data and the advent of paired (VH/VL) sequence data, there emerged an urgent need to further develop, update, and expand the functionality of the OAS database.

This chapter is based on material from the following paper, where we detailed our advancements made to OAS. All the work described in this chapter was carried out by me.

> **Olsen T.H.**, Boyles F., and Deane C.M. (2022) OAS: A diverse database of cleaned, annotated and translated unpaired and paired antibody sequences. *Protein science*, 31(1):141-146.

2.2 Introduction

The antibody repertoire represents the immunological condition of an individual, and can be used to investigate their immune states (effects from diseases or vaccination responses) and for finding and developing potential therapeutics [30, 133]. As mentioned in the Introduction 1.3.1, the development of BCR-seq techniques has allowed researchers to probe the diversity of the repertoire and the similarities and differences between individuals and different species [28]. With BCR-seq steadily improving [27], hundreds of millions of unpaired sequences are now sequenced for each study [21, 47], and techniques for sequencing paired VH and VL have been developed [48, 49].

The use of data from multiple studies, allows larger and more complete analyses to be carried out, however; this is hindered by nonstandard data formats. The MiAIRR standard (see Introduction 1.3.3) was introduced to counter this, however; direct comparisons are still complicated because of different processing pipelines or data often only released as raw FASTQ files, necessitating extensive processing [51].

As described in Introduction 1.3.3, the increase in available BCR-seq data, has led to a set of related efforts of collating the available data into a single antibody

database, including ImmuneAccess [53], PIRD [54] and RAPID [55], the AIRR Data Commons [56] (ADC), and the Observed Antibody Space (OAS) database [52]. OAS created in 2018 by Kovaltsuk et al. and was a database of unpaired VH and VL antibody protein sequences, derived by identically processing 55 BCR-seq datasets containing 600 million sequences [52]. Since the first publication of OAS, an increased volume of data has been published as well as the appearance of paired (VH/VL) sequence data, leading to a need to expand and overhaul OAS.

In this chapter, we present an updated and expanded OAS. The database contains, as of July 2023, 2.4 billion unpaired sequences from 90 studies, including recent studies featuring SARS-CoV-2 data, and paired sequencing data from eight studies. The new database also now provides the nucleotides for the VH and VL chains, in addition to amino acids. It also now contains additional sequence annotations, such as the antibodies junction sequence and whether it is a productive sequence, making the data MiAIRR-compliant. Comments on potential problems (e.g. lack of conserved cysteines or unusual insertions and deletions) with the sequence have also been added. OAS is accessible via a new web server (`http://opig.stats.ox.ac.uk/webapps/oas/`), which provides standardised search parameters and a new option to search for sequences with the same V and J genes as a query sequence, allowing for a fast initial query of 1000 antibody sequences similar to a given sequence of interest.

2.3 Methods

Accession numbers of publicly available BCR-seq datasets were extracted from the literature and NCBI's SRA Run Selector [134] was used to download the metadata of each run within a study. The metadata was then processed into a standardised format which contains the run accession number, the author and associated DOI, subject identifier, age of subject, species, B-cell source, B-cell type, vaccinations, diseases and any longitudinal information.

The FASTQ files of each run were then downloaded using SRA-Tools fastq-dump 2.9.1 [135]. For runs sequenced using a paired-end library layout, FLASH 1.2.11 [136] was used to merge the FASTQ files into a single file. Unpaired sequences were then converted from a FASTQ file into a FASTA file using fastq_to_fasta from the FASTX-Toolkit 0.0.14 [137]. Paired sequences were converted to a FASTA file using 10x Genomics Cell Ranger 6.0.2 [138]. The barcode in the resulting FASTA file was then used to map heavy and light chains that form a pair. The unpaired and paired datasets then follow the same processing steps. Fig. 2.1 shows an overview of the pipeline.

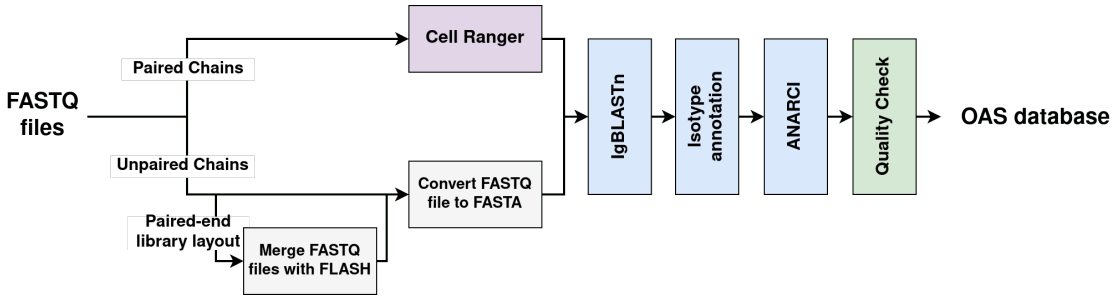

Figure 2.1: Overview of the OAS pipeline. FASTQ files with unpaired sequences are processed and converted to FASTA with FLASH [136] and FASTX-Toolkit [137]. Files with paired sequences are converted to FASTA files using 10x Genomics Cell Ranger [138]. Paired and unpaired sequences are then cleaned, annotated, and translated using the same steps.

The nucleotide-containing FASTA file was aligned to germlines and translated using IgBLASTn 1.17.1 [38]. For IgBLASTn, the VDJ germline databases for human, mouse, rat, rabbit and monkey were created using IMGT germline sequences derived from IMGT [139]. For camel sequences, the nearest relative alpaca was used. Sequences deemed unproductive by IgBLASTn or with no resulting aligned sequence were removed. IgBLASTn considers a sequence productive if the V(D)J rearrangement frame is in-frame, there is no internal frame shift in the V gene, and no stop codon is found. To determine the isotype of heavy chain sequences, the first 21 nucleotides of the constant heavy domain 1 (CH1), obtained from IgBLASTn, was aligned to a set of IMGT derived constant domain germlines with annotated

isotypes. The alignment was done using the same conservative alignment approach as used by Kovaltsuk *et al.* [52] and sequences with no high-confidence alignments or with missing CH1s had their isotype annotated as Bulk.

For each sequence, the IMGT numbering scheme was added using ANARCI 2020.04.23 [26]. Any sequence ANARCI could not process was removed. This step predominantly removes sequences that contain a stop codon. An ANARCI status highlighting potential problems for each sequence is retained in the database. This status contains comments regarding unusual residues, lack of conserved cysteines, deletions and insertions outside of the CDRs, truncation of frameworks 1 or 4, and if the CDR3 is longer than 37 residues. Lastly, sequences were grouped into units sharing the same metadata, the same chain (e.g. heavy, light or paired) and isotype.

2.4 Results

We have created an unpaired and paired OAS database using the approach given in the Methods section. Unpaired OAS consists of the 55 BCR-seq repertoires from the original OAS [52] and, as of the 30th September 2021, 25 new studies. From the 80 studies, a total of 3,549,291,485 sequences (3,453,356,223 VH and 95,935,262 VL) and 1,535,831,565 unique sequences (1,499,142,547 VH and 36,689,018 VL) were retrieved. Heavy chains with all five isotypes are present in the database. The majority are IgM (1,019,175,618), IgG (287,174,563) and undetermined Bulk (167,366,413). B-cell specific information is also included, with most of the sorted antibodies derived from naive B-cells (125,521,501) or plasma B-cells (40,126,155), and the bulk of the B-cells extracted from leukopaks/peripheral blood (1,268,447,750) or the spleen (121,114,530).

Information about the individuals from which the BCR-seq repertoires are taken was annotated where possible. The sequences can be mapped to 825 unique subjects and six different species, with the majority coming from humans (88%) and mice

(11%). If given, age is annotated as precisely as possible, either as the exact age or as an age group. Available data includes ages younger than 18 (253,056,094), between 18-70 (875,775,859) and over 70 (32,964,094). For any study where the same subject had a sample sequenced at different time points, the difference in time is annotated.

The data also covers a diverse set of immune states. 233,686,163 sequences are seen in individuals with 25 different diseases, including 61,730,169 and 47,853,000 sequences from SARS-CoV-2 and HIV infected patients respectively. Twenty different vaccine combinations are present in unpaired OAS, with the majority derived from species vaccinated with HepB (52,111,613) or OVA (40,743,076). Many sequences are also available for other well studied antigens, like the Flu (15,042,090), Ebola (11,903,910), and RSV (900,259).

Paired OAS contains 121,838 sequence pairs from five studies. Although less diverse, paired OAS already contains 11,388 pairs from SARS-CoV-2 infected patients and 17,913 pairs from Memory-B-Cells.

For every sequence in OAS, information about its nucleotide sequence, including the constant domain if given, amino acid sequence, chain, isotype and annotated germlines is specified. To make OAS MiAIRR compliant, information required by AIRR's rearrangement schema, such as whether a sequence is productive, is also included. The IMGT numberings and an ANARCI status has been added to highlight any deletions, insertions, missing conserved cysteines or truncated ends.

To allow for a simple extraction of the data present in OAS, or subsets of it, we have created the OAS website `http://opig.stats.ox.ac.uk/webapps/oas/`. In addition to the ability to download all of OAS, the data can be queried by all meta labels (e.g. study, species, chain and disease). As an example, to compare heavy chain sequences from different patients infected with SARS-CoV-2, one could go to

the unpaired data and select Chain: Heavy, Disease: SARS-COV-2 and Subject: Defined. The website will then provide links to download files containing 61 million unique sequences from five studies and 130 different SARS-CoV-2 patients (Fig. 2.2).

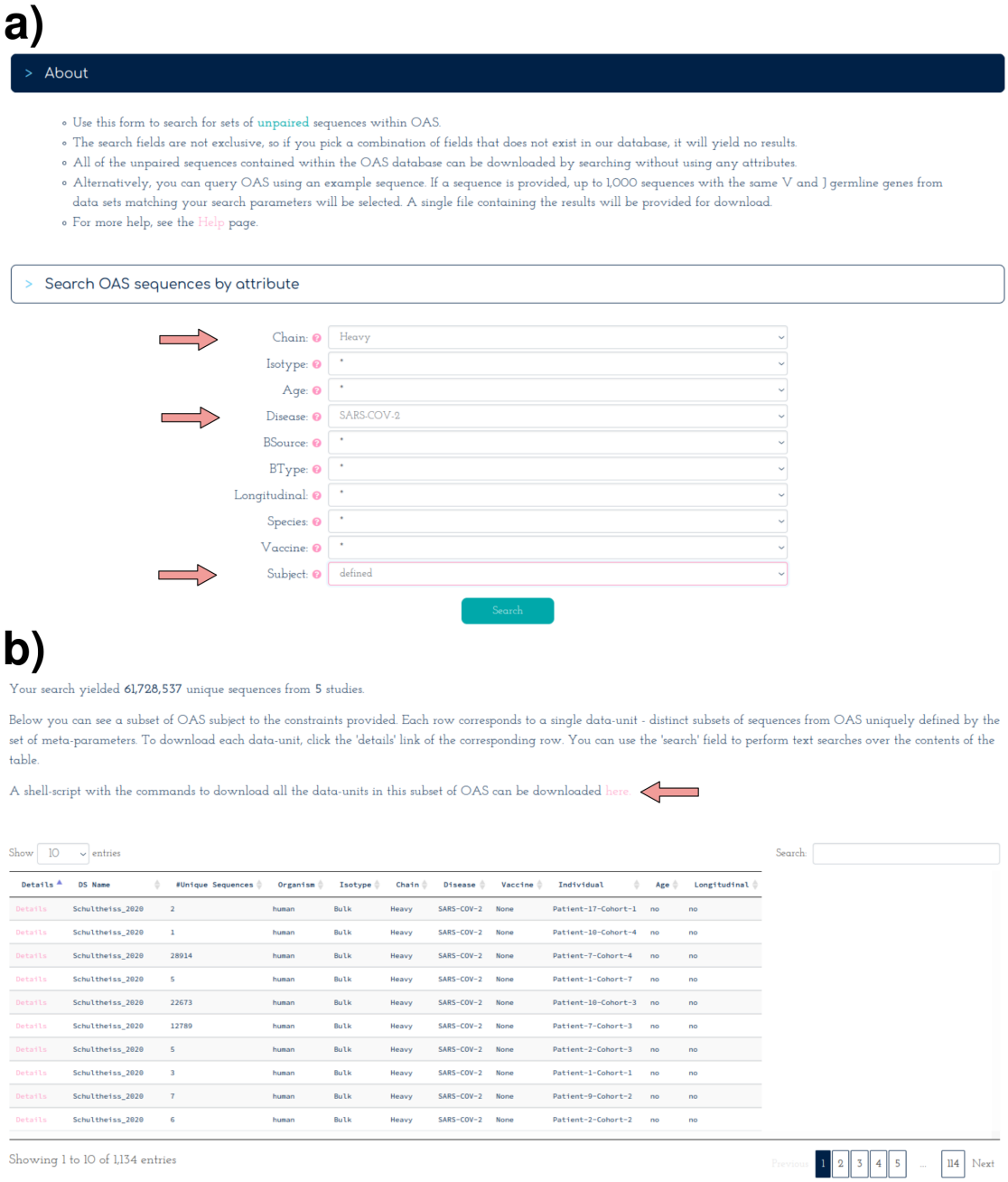

Figure 2.2: Downloading from OAS. **a**, shows the sequence search tab for unpaired sequences, with the search options filled for heavy chain sequences from SARS-CoV-2 infected patients (shown with red arrows). **b**, shows the search result, with each data unit matching the search and a downloadable link containing the links for the relevant data units (with a red arrow).

We have also added a sequence search option, allowing a user to select sequences in OAS that are similar to a query sequence. The search will return a sample of 1,000 sequences from the database with the same V and J germline genes as the query sequence, and can be combined with the meta label search to obtain, for example, a sample of 1,000 mouse heavy chain sequences with the same V and J gene as a given query.

Since its initial publication, OAS has been subject to continuous updates, reinforcing its status as an actively maintained and evolving resource. As of July 2023, OAS contains over 2.4 billion unpaired sequences and 1.5 million paired VH/VL sequences, illustrated in Fig. 2.3. This substantial increase in data over just a few years highlights the advancement and growing use of BCR-seq techniques. Notably, while OAS still predominantly contain heavy chains from humans and mice, the recent updates have led to a remarkable tenfold increase in both light chains (356 million) and paired antibodies. Consequently, studies delving into these previously underrepresented aspects of antibody repertoires are now becoming more feasible.

2.5 Discussion

Antibody repertoires have become crucial datasets for examining the immune response and the development of antibodies as therapeutics. Although large BCR-seq datasets from a variety of studies exist, data processing is often inconsistent between datasets, or the data is only publicly available as raw FASTQ files. This prompted us to create OAS, a resource of BCR-seq datasets all processed in a standard manner. In this chapter, we describe an updated version of OAS featuring an improved processing pipeline, additional sequence information, an increased amount of unpaired data and the inclusion of paired data.

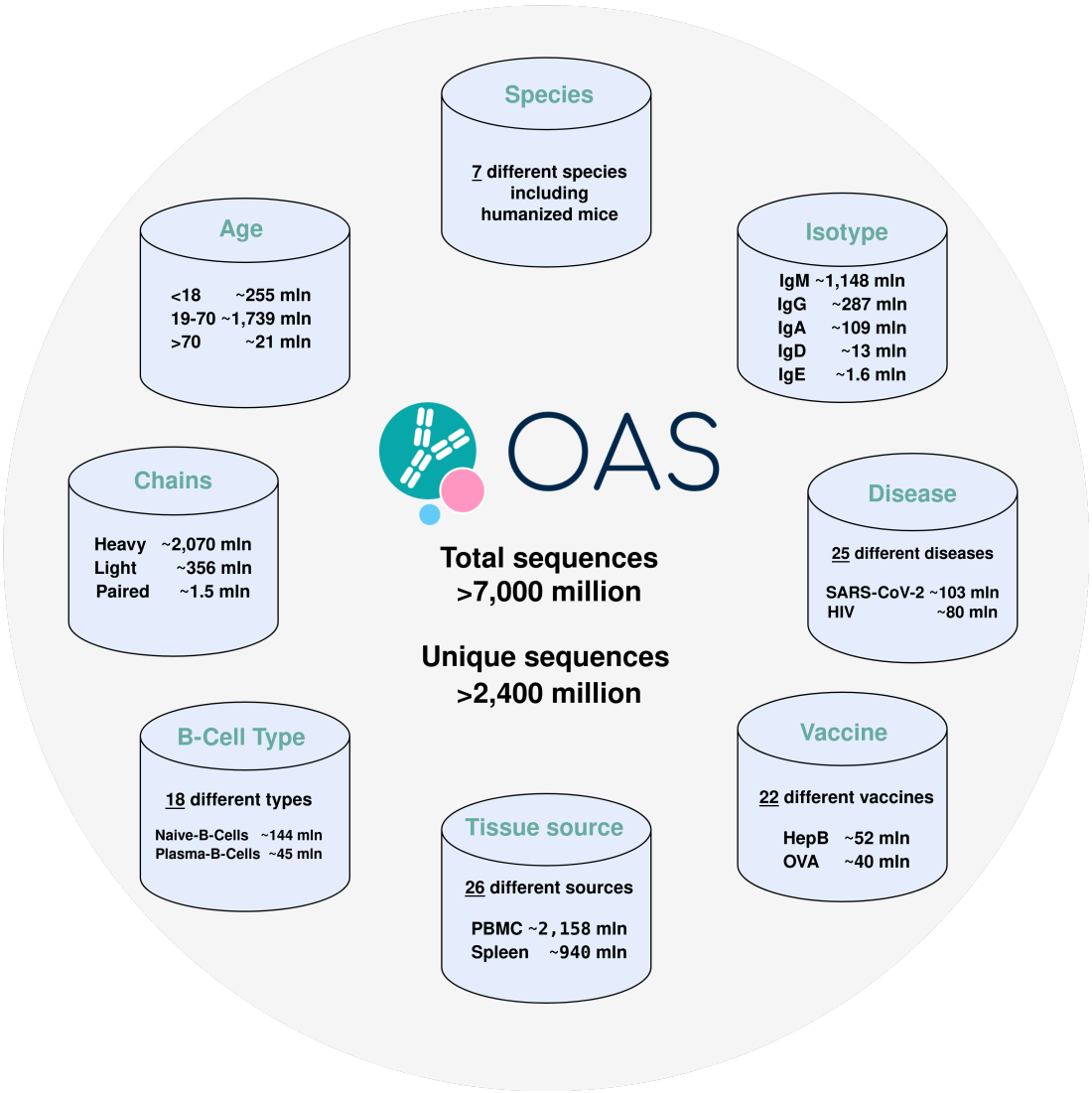

Figure 2.3: Overview of the OAS database size and diversity. OAS is continuously updated with the availability of new data and currently, as of July 2023, accommodate over 2.4 billion unpaired antibody sequences and 1.5 million antibody pairs from 90 and eight studies, respectively. The sequences are derived from a variety of different sources (i.e. species, tissue, B-cell and disease-state) and antibody type (i.e. chain and isotype), resulting in a diverse database. While still dominated by heavy chains, the latest OAS updates have led to a remarkable tenfold increase in both light chains and paired antibodies. This remarkable growth also highlights the continuous efforts to capture the full spectrum of antibody repertoire diversity, offering researchers a valuable resource for studying and understanding immune responses.

In the processing pipeline, germlines are now annotated using nucleotides, instead of amino acids, allowing for better annotations. Moreover, additional information, for example nucleotides and ANARCI status, and new data, e.g. SARS-CoV-2 infected patients, has been added. Most importantly, we also introduce paired OAS. Paired data will allow investigations of the whole binding site, instead of solely the heavy or light chain.

To make OAS more findable, accessible, interoperable, and reusable (FAIR) [140], OAS has been adapted to be compliant with MiAIRR. The first four sets of MiAIRR are associated with the creation of FASTQ files. As we cannot control the creation of the FASTQ files, we instead include the source in the metadata so a sequence can always be linked back to its original study and run. The 5th MiAIRR set encompasses data processing. Here, all sequences are processed as elaborated in the Methods section with the software version given for each tool used. Lastly, the 6th MiAIRR set is followed by including all the required elements in MiAIRR's rearrangement schema. By following this schema, OAS can be merged with data from other MiAIRR compliant sources, such as the ADC, and interoperate with the same tools. However, as data from the ADC is not processed using a single standardised processing pipeline, and those pipelines are also potentially different from OAS, this presents a challenge when reusing studies from different databases for direct comparisons, and for creating truly FAIR repertoire data.

To further improve the accessibility, OAS is freely available at `http://opig.stats.ox.ac.uk/webapps/oas/`. To ensure OAS is updated after the graduation of the first author, a new OAS representative is appointed within the group. The representative is selected based on using OAS for their own work, incentivising us to maintain and consistently add the newest unpaired and paired data to OAS. Subsets of interest can be queried and downloaded based on all meta labels from the website. Additionally, users can perform a search for sequences with

the same V and J gene allowing for a fast initial exploration of a sequence of interest.

This updated and expanded version of OAS is a versatile and valuable resource for a wide range of antibody research areas, including investigation of immune states, by enabling comparison of antibody repertoire data between studies, and therapeutic antibody design, by giving access to billions of cleaned, annotated and translated antibody sequences, paired VH/VL and facilitate a fast exploration of similar antibody sequences.

In subsequent Chapters, we will introduce novel tools designed to harness the full potential of OAS for antibody discovery and design. In the next chapter, we will detail the development of a method for rapid and exhaustive sequence identity search of known antibodies, named KA-Search. KA-Search's speed allows researchers to swiftly explore billions of antibodies based on amino acid sequence identity, enabling new approaches for gaining insights about specific antibodies of interest and facilitates the discovery of new antibodies with similar function. To illustrate this, we show how KA-Search's ability to find the most similar sequences among the 2.4 billion antibodies in OAS within 30 minutes can be used to find new potential SARS-CoV-2 binders.

3

KA-Search, a method for rapid and exhaustive sequence identity search of known antibodies

Contents

3.1 Chapter abstract

Antibodies with similar amino acid sequences often exhibit shared properties. Therefore, identifying highly similar antibodies among those naturally expressed in healthy or diseased repertoires is a powerful method for exploring possible

mutations that might retain certain properties while improving others. This type of search is called immune repertoire mining [40–43] (see Introduction 1.3.2). An instance of such mining could involve identifying mutations that improve binding affinity. For an antibody targeting a particular pathogen, more effective binders could be discovered by identifying similar antibodies with minor differences in the binding site, which are also derived from patients with the specific disease. Furthermore, since these mutations are found in naturally expressed antibodies, they are also functional. However, as the number of available antibody sequences now reaches into the billions and continues to grow, repertoire mining for similar sequences is becoming increasingly computationally intensive [141].

In this chapter we describe our work on building an antibody specific tool for rapid, exhaustive and flexible search of similar antibodies, to enable immune repertoire mining of all antibodies in OAS. This chapter is based on the following paper, where I am a co-first author alongside Brennan Abanades, and my contributions to the work include envisioning the project and co-writing the code alongside Brennan Abanades. Moreover, I prepared the OAS sequences for search, carried out the benchmarks, and undertook the immune repertoire mining case study with the COVOX-253 antibody.

> **Olsen T.H.**, Abanades B., Moal I.H., and Deane C.M. (2023) KA-Search, a method for rapid and exhaustive sequence identity search of known antibodies. *Scientific Reports*, 13:11612.

3.2 Introduction

Antibodies have become an invaluable form of therapeutics, with an increasing number of new antibody derived therapeutics being developed and marketed each year [142]. Despite their success, the process of antibody discovery and design is still challenging [87]. As mentioned in Introduction 1.3.2, immune repertoire mining has emerged as a promising technique to explore the vast antibody sequence and mutation space [40–44]. By comparing an antibody of interest against natural

antibody repertoires, immune repertoire mining identifies identical or highly similar antibodies, offering valuable insights into potential mutations to improve binding affinity, developability profile, or reduce immunogenicity without altering binding properties [40–44] (see Fig. 3.1a).

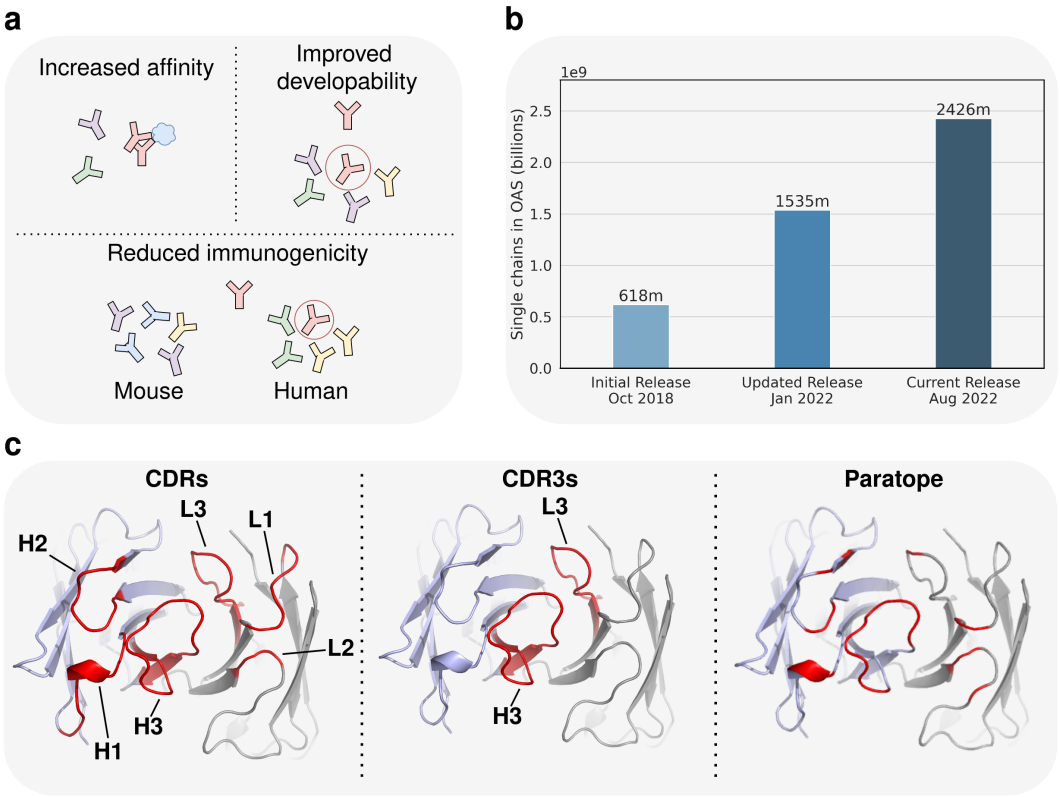

Figure 3.1: **a**, To correctly identify relevant mutations for optimising a given antibody for increased affinity, an improved developability profile or reduced immunogenicity, a huge space of possible mutations needs to be searched. As similar antibodies often bind to the same epitope but with different strengths, immune repertoire mining can be used to find similar antibodies with potentially better binding affinity. Immune repertoire mining can also be used to find antibodies with the same binding but with mutations improving their developability profile or reducing their immunogenicity. **b**, Overview of the available number of single chain antibody variable domain sequences in the Observed Antibody Space database over time. **c**, Highlight (red) of different specific search regions. The antibody variable domain is derived from PDB structure 7JOO and the CDRs are annotated using IMGT numberings [143]. Heavy chain complementarity-determining region (CDR)1, 2 and 3 are denoted as H1, H2 and H3, and light chain CDRs as L1, L2 and L3. The paratope was defined as any residue within 4.5Å of the antigen (in this case an inducible T-cell costimulator) and consists of IMGT position 35, 36, 57, 58, 64, 66 and 109-113 on the heavy chain and IMGT position 37, 38, 55, 56, 66 and 114 on the light chain.

Similarity between antibodies can be measured in different ways. The most common ones are via sequence identity or structural similarity [144, 145]. With a protein's function being preserved in the structure, structural similarity is often superior for finding proteins with analogous functions, such as antibodies binding the same epitope [44, 146]. However, with orders of magnitude more sequence data available than structural data, a sequence identity search enables the exploration of a much larger space. Available sequence data is also more diverse, as next generation sequencing of B-cell receptors (BCR) is routinely being applied to study adaptive immunity (see Chapter 2), generating sequences from a range of species [32–34] and from individuals with differing disease states [35, 36]. Furthermore, as seen in Chapter 2, continuous improvements in high-throughput sequencing methods and increased adoption by research labs means that the amount and diversity of sequence data is rapidly increasing [21, 37].

Whilst freely available, searching this immune repertoire data for similar antibody VH and VL protein sequences, still requires extensive post-processing of each source, such as translating the nucleotide sequences to protein sequences. A database providing a single entry to already processed antibody data to search against is therefore advantageous. In Chapter 2 we introduced one such effort, the Observed Antibody Space (OAS) [52, 141] database, which as of July 2023 contains sequences of the V domain for ∼2.4 billion unpaired heavy and light antibody chains. These sequences are derived mostly from humans and mice, but also include sequences from rabbits, rats, rhesus', camels and humanized mice. While the size of OAS is promising from a scientific perspective, its scale and continuous growth, visualized in Fig. 3.1b, make mining it effectively a challenge. Though sequence identity calculations are simple, without software specially optimised for the task, the computational cost of exhaustively searching OAS or any other large antibody sequence databases is becoming prohibitive. There is therefore a need for specialized tools to search this space now and in the future.

There exist many tools for searching large datasets of protein sequences for similar sequences, for example BLASTp [22], CD-HIT-2D [147], and newer methods such as MMseqs2 [148]. However, these tools are all designed around searching a diverse set of proteins and not specifically antibody sequences. BLASTp finds similar sequences by searching for high-scoring 3-mers for a query within a set of target sequences. This scoring is done using a substitution matrix, such as BLOSUM62 [149]. For target sequences with exact matched 3-mers, the alignment is then extended in both ends until the score decreases and ranked based on their expect value. BLASTp is much faster than performing a pairwise alignment with the Smith-Waterman algorithm between the query and each target sequence; however, BLASTp does not guarantee optimal alignments [22]. To further increase speed, both CD-HIT-2D and MMseqs2 use fast prefiltering steps to remove target sequences with low identity to the query, thereby reducing the number of pairwise alignments to make, a computational expensive step. CD-HIT-2D prefiltering removes target sequences that have an estimated similarity to the query below a specified threshold. Simplified, the estimation is based on two sequences of certain lengths requiring to share a minimum number of k-mers of different sizes, in order to be above a specific sequence identity. After prefiltering, pairwise alignment is performed on the remaining sequences [150]. Prefiltering with MMSeqs2 is also based on comparing k-mers between the query and target sequence. However, instead of exact matches like CD-HIT-2D, MMseqs2 uses a BLAST-like approach of matching k-mers with a BLOSUM62 score above a certain threshold. For a query and target sequence pair with two k-mer matches found on the same diagonal, an ungapped, and finally a gapped alignment, using Smith-Waterman [151], is performed [148].

While prefiltering greatly speeds up sequence search algorithms, it can cause issues when searching a set of closely related sequences, as is the case with antibodies, as the prefiltering step can remove good hits. Further, each tool uses an alignment method designed for general protein sequences, which can result in

unreliable antibody alignments, especially in the highly variable CDRs. Within the immunoinformatics field, this alignment problem is often overcome by using antibody specific numbering schemes, like the ImMunoGeneTics (IMGT) scheme [4, 143]. Another issue with non-antibody specific tools, is the lack of flexibility in their searches. These tools can only readily be used for searching against the whole antibody chain of target sequences and not for finding similar sequences based on specific subregions. Searching for identical regions at specific antibody positions, especially the CDRs, is often used when looking for similar binders [43]. With the majority of the residues involved in binding being located in the CDRs, the sequence identity over these regions is often more relevant than that of the whole antibody. For some applications, the exact set of residues involved in the paratope may be known. In these cases, searching based on the sequence identity of the paratope may be even more informative (see Fig. 3.1c). An antibody specific tool that utilizes antibody numbering schemes for better searches, without prefiltering for an exhaustive search, and with the ability to search user-defined continuous or non-continuous regions, such as the paratope, would improve our ability to make best use of the antibody sequence data available.

Recent efforts to create antibody specific searching tools include iReceptor [152], AbDiver [153] and CompAIRR [154]. iReceptor, only allows for a V-, D-, or J-gene search or an exact CDR3 match search. AbDiver uses an antibody numbering scheme to align sequences and allows for both CDR3 and whole V domain searches. AbDiver restricts CDR3 searches against CDR3s with a specified V gene and species of origin, and whole V domain searches against sequences with same length CDR1 and 2 and ± 1 length CDR3. These restrictions narrow and greatly speed up the search but can occasionally lead to it finding no matches. Further, both iReceptor and AbDiver are not open-source and are only freely available to use via their website, so can only be used against their own databases. While CompAIRR is designed for finding the overlap of CDR3's across different antibody repertoires, it can also only be used to search for either exact or similar CDR3's. However, like

iRecepter and AbDiver, the restriction of the search limits its use cases, for example none of the tools can search for exact or similar CDR1 or 2, or combinations of CDRs. There therefore exists the need for an open-source antibody specific tool not limited by either being low-throughput, non-exhaustive, or only searching against entire V domain sequences.

Here, we introduce Known Antibody Search (KA-Search), a tool that allows for rapid amino acid sequence identity search across the VH and VL domains of billions of unpaired antibody chains, across either the whole domain, the CDRs, or a user defined antibody region. We demonstrate KA-Search can be used to find the most similar sequences from the ∼2 billion heavy chain sequences in the OAS database within 30 minutes using 5 CPUs. We also show how KA-Search can be used for immune repertoire mining to obtain new insights about an antibody of interest. KA-Search is freely available at `https://github.com/oxpig/kasearch`.

3.3 Method

3.3.1 Data preprocessing

KA-Search pre-aligns the V domain of antibody sequences to a canonical alignment capable of accommodating the most common numbering positions. To do this, every sequence is first numbered with ANARCI [26] using the IMGT numbering scheme [4, 143] and then converted to a vector of the same length. As our canonical alignment, we use all of the 196 unique positions seen in at least 40,000 different sequences in OAS, as of May 2022, and four additional unique positions seen in therapeutics from Thera-SAbDab [59]. The exact unique positions are given in Supplementary Table A.1 and cover around ∼99.8% of sequences in OAS. The 0.2% of antibody sequences that contain a rare insertion in their V domain cannot be searched using KA-Search. This set of unusual sequences is provided together with the aligned sequences and can be searched using other methods. Every aligned sequence is accompanied by two index values which can be used to retrieve its

metadata.

All amino acid sequences of the antibody VH and VL domains (derived from the sequence_alignment_aa column) in OAS (September 2022) are pre-aligned using this method to generate a dataset ready to be used by KA-Search. This results in over 2,070 million heavy and 355 million light chain sequences. Sequences are split into heavy and light chains, and by species information, e.g. human, mouse, rabbit, rat, rhesus, camel and humanized, allowing for faster specific searches. We call this data set of heavy and light chains OAS-aligned. OAS-aligned also contains sequences that cannot be aligned in files labeled as unusual, for search using other methods. We also built a subset of the heavy chain dataset, OAS-aligned-small, that contains 118 million heavy chain sequences, which was generated by removing sequences containing ambiguous residues or seen less than five times. Further, a smaller subset of 10 million human heavy chain sequences, OAS-aligned-tiny, was built by removing any sequence in OAS-aligned-small not having a residue at position one and removing duplicate sequences. OAS-aligned, OAS-aligned-small and OAS-aligned-tiny, and the code to update the data sets or expand it with an in-house data set is made freely available with KA-Search (`https://github.com/oxpig/kasearch`).

3.3.2 Identity calculation

The identity between a region in the query and target sequence is computed as the percentage of identical residues across a specific region, including indels present in only one of the sequences. A region can be either the whole V domain or a set of antibody numbering positions, such as the CDR3. Length matched sequence identity is only calculated if the compared region has the same length in both the query and target sequence. The identity can also be calculated excluding missing residues at the ends of the sequences; however, the default is to include them. By converting the query and target sequences into fixed length vectors, their sequence identity can be calculated using matrix operations. For KA-Search, this is

implemented using the heavily optimised library JAX [155].

To search for the identity of a specific user-defined region, i.e. the CDRs, a list with the desired positions can be specified (see Fig. 3.2). These positions need to be one of the 200 unique positions in the canonical alignment. In default mode, KA-Search will search for similar whole V domains of variable length, and the three CDRs and CDR3 regions with exact length match. KA-Search returns for each target sequence, the sequence identity of the defined region and the target sequence's metadata, sorted by sequence identity. For the OAS derived data, the metadata includes each column from AIRR's rearrangement schema [50] and additional columns derived when preparing OAS [141] with the last column being the sequence identity.

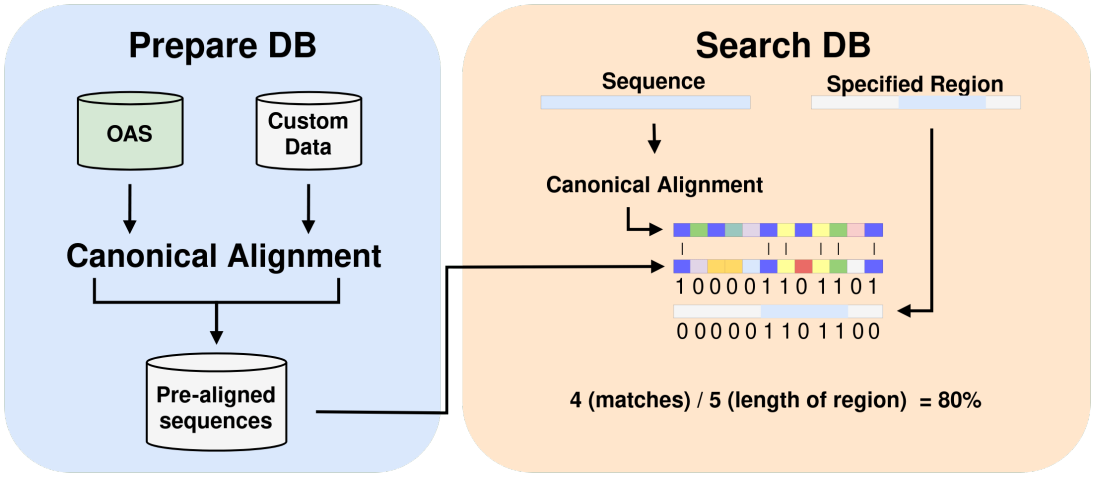

Figure 3.2: Overview of KA-Search. Before search, target sequences are pre-aligned using a canonical alignment of 200 unique antibody positions. Once an antibody has been entered into a pre-aligned database this calculation does not have to be repeated even when new data is added. For search, a query and the specific region to search, for example whole variable domain or CDRs, is specified. The query is then aligned using the canonical alignment and matched with each aligned target sequence. The specific region mask is then applied before calculating the sequence identity for the region. The exact method for sequence identity calculation is either with or without length match.

3.3.3 Sensitivity and speed comparison

KA-Search was compared to BLASTp (version 2.13.0), CD-Hit-2d (version 4.8.1) and MMseqs2 (version 13.45111), for sensitivity and speed at searching for the closest whole antibody chain match. A set of 100 randomly selected non-redundant heavy chains of therapeutics were used to search for the closest sequence within OAS-test, a set of 10 million heavy chain antibody sequences that could be aligned with the KA-Search canonical alignment. The sequences in OAS-test were randomly extracted from the full OAS database, cleaned and reduced as done in [129]. The 100 therapeutic heavy chains and OAS-test are available for download, see `https://doi.org/10.5281/zenodo.7561985`.

For BLASTp, we first pre-built a BLAST database of OAS-test using makeblastdb. BLASTp was then run using an expect value threshold of 10, word size of 3, BLOSUM62 as the substitution matrix, gap costs of 11 for existence and 1 for extension, with conditional compositional score matrix adjustment and no other filters or masks. For sensitivity comparisons we sorted hits by the expect value. With CD-Hit we searched using a sequence identity threshold of 70% for the global sequence identity and a word length of 5. For MMseqs2, we first precomputed a sequence database of OAS-test with the createdb and createindex modules. The prepared database was then searched using the easy-search workflow with the default arguments; sensitivity of 5.7, BLOSUM62 as the substitution matrix, gap costs of 11 for existence and 1 for extension. For sensitivity comparison, 300 sequences were allowed to pass prefiltering and were thereafter sorted by sequence identity. Lastly, for KA-Search we used a pre-aligned OAS-test aligned as described above. For sensitivity comparisons, we compared over the whole V domain without length matching. For each tool, speed was calculated using the same single CPU to search for one sequence against OAS-test and sensitivity by how well each tool found the closest sequence in OAS-test to each query. The closest sequence was defined by either having the highest KA-Search identity or the highest BLOSUM62 score among the top-100 returned from each method. The BLOSUM62

score was calculated after alignment with the Smith-Waterman algorithm, using BLOSUM62 as the substitution matrix, gap costs of 11 for existence and 1 for extension. All time measurements in this paper were performed using CPUs from an Intel Xeon Gold 6240 Processor.

3.4 Results

Immune repertoire mining to find similar antibodies with shared properties is becoming increasingly computational expensive because of the increase in available antibody sequences. This is illustrated in Fig. 3.1b, which shows how publicly available sequences in OAS have increased by 1.8 billion in less than four years. Below we describe KA-Search, a freely available tool to search immune repertoires that is optimised to handle the vast amount of available data.

3.4.1 Computational speed of KA-Search

KA-Search's exact speed is dependent on the hardware used, the number of queries, number of output sequences desired and number of regions searched over. Fig. 3.3 shows a comparison between different KA-Search runs with different numbers of CPUs, when searching against the 2,070 million heavy chains in OAS-aligned. The number of closest matches returned has a minimal impact on speed, with returning the best or 10,000 best matches taking approximately the same time. Searching over multiple regions simultaneously slows the search but is faster than doing them individually.

When using a single CPU one region takes 43.01min ± 9s, three 60.85min ± 13s, and ten 158.45min ± 2s. The time required per query is reduced when searching with multiple queries at a time, as searching with a single query takes ∼43 minutes while searching with 100 queries takes ∼6.6 minutes per query. KA-Search is limited by loading data into memory when searching with few queries. The optimal use of KA-Search is therefore to search with many queries and multiple regions

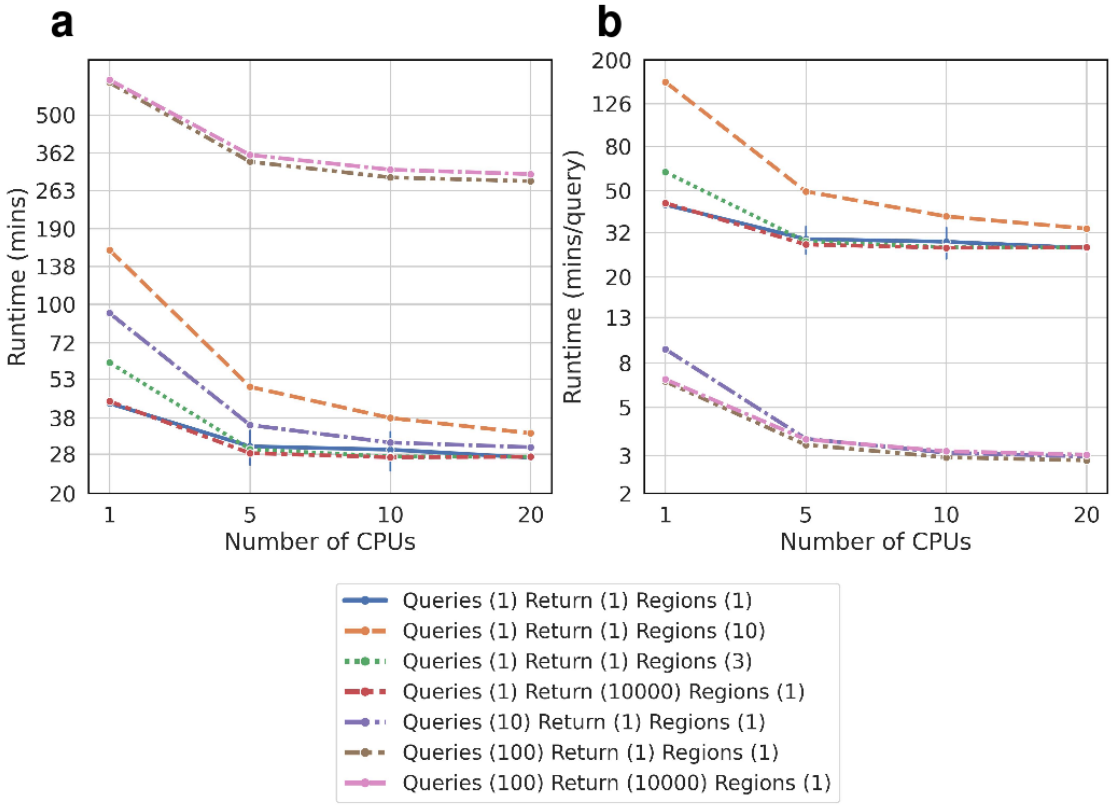

Figure 3.3: Runtime comparison of KA-Search with different numbers of queries, returned sequences and number of regions. Each search was done against the 2,070 million heavy chains in the Observed Antibody Space database. **a**, The runtime is minimally impacted by the number of closest matches returned, and increases when searching over multiple regions simultaneously or with multiple queries. **b**, Searching with multiple queries greatly reduces the runtime per query, up until 10 queries per search. It is therefore optimal to search with many queries simultaneously.

simultaneously using multiple CPUs.

3.4.2 Comparison with other common sequence identity search tools

To compare KA-Search with current freely available and downloadable protein sequence search tools, we selected the amino acid sequence of the VH domain for 100 non-redundant heavy chains of therapeutics, and searched for the most similar sequence within OAS-test, a set of 10 million VH antibody sequences (see methods), using BLASTp [22], CD-HIT-2D [147], MMseqs2 [148] and KA-Search. For each tool, the mean and standard deviation of their speed was calculated

based on seven runs (see Fig. 3.4). KA-Search takes 8.3s ± 22.7ms, which is far faster than BLASTp and CD-HIT-2D, 103s ± 75.9ms and 82s ± 55ms seconds respectively, but slower than MMseqs2 at 3.57s ± 78.8ms.

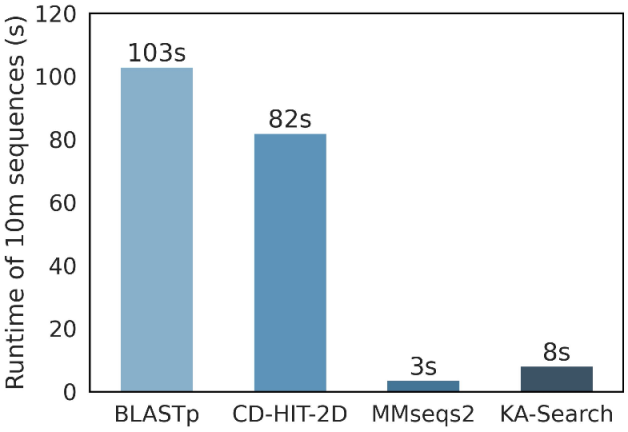

Figure 3.4: Speed comparison between KA-Search and the commonly used protein search tools BLASTp (version 2.13.0) [22], CD-Hit-2d (version 4.8.1) [147] and MMseqs2 (version 13.45111) [148]. The default settings were used for each tool, and each search was done against the whole variable domain. Runtime was calculated as the average time it took to search 10 million sequences (OAS-test) with a single query on a single CPU.

In terms of sensitivity, we examined the ability of these sequence search methods to identify the most similar antibody sequence in OAS-test as defined by either the Smith-Waterman aligned BLOSUM62 score used by BLASTp, or the KA-Search identity (see Fig. 3.5). Tools that use prefiltering struggle to find the exact closest match. CD-HIT-2D's highest ranked sequences matched the target sequence with the best BLOSUM62 score and KA-Search identity for only one and four out of the 100 sequences, respectively, and MMseqs2's highest ranked sequences matched none of the closest target sequences for either metric. When looking for the closest match within the top-100 highest ranked sequences, CD-HIT-2D found the closest match based on the BLOSUM62 score and KA-Search identity for 6 and 12 sequences, respectively, while MMseqs2 found none. BLASTp and KA-Search find the closest match as highest ranked based on the BLOSUM62 score for 54 and 27 sequences, respectively, and for 33 and 100 sequences based on the KA-Search identity. Within

the top-100 highest ranked sequences, BLASTp and KA-Search find the closest match for 96 and 83 sequences, respectively, and 90 and 100 sequences based on the KA-Search identity. The full sequence of the test queries and their respective top-1 sequences for each method can be found as Supplementary Data 1 online at `https://tinyurl.com/mwfdya9e`.

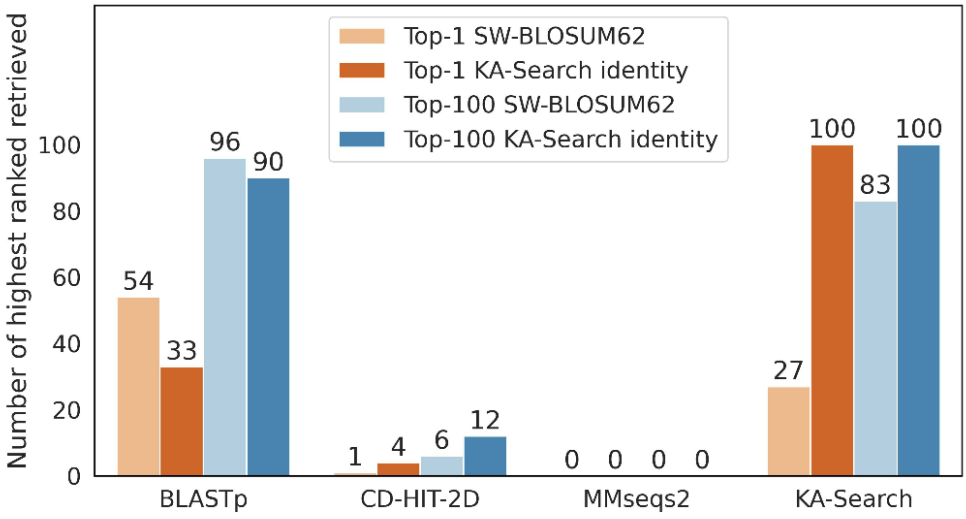

Figure 3.5: Sensitivity comparison between KA-Search and the commonly used protein search tools BLASTp (version 2.13.0) [22], CD-Hit-2d (version 4.8.1) [147] and MMseqs2 (version 13.45111) [148]. The default settings were used for each tool, and each search was done against the whole variable domain. BLOSUM62 scores were calculated by aligning with the Smith-Waterman algorithm using BLOSUM62 [149] as the substitution matrix and gap costs of 11 for existence and 1 for extension. Sensitivity was calculated by how often each tool returned the closest or the closest within the top-100 match for 100 heavy variable domains (test queries) against the same 10 million sequences. The closest match was defined using two different identity metrics, the highest BLOSUM62 score and KA-Search identity.

We further compared the average BLOSUM62 score between the query and highest ranked (see Fig. 3.6a). For BLASTp, CD-HIT-2D, MMseqs2 and KA-Search the score was for the highest ranked on average 537, 461, 507 and 533, respectively, and for best within top-100, on average 539, 480, 508 and 539. Furthermore, we compared the sensitivity of each method by calculating how similar the returned sequences are to the query. This comparison is shown in a density plot of the

difference in sequence identity, based on exact matches, between the query and highest ranked sequences (see Fig. 3.6b-c). For BLASTp, CD-HIT-2D, MMseqs2 and KA-Search this difference was on average 15.26%, 23.04%, 20.29% and 14.43% identity, respectively. The difference between the query and the best within the top-100 highest ranked sequences were 14.12%, 21.41%, 19.52% and 14.06%, respectively.

While CD-HIT-2D is better than MMseqs2 at finding the closest match, MMseqs2 returns on average better matches. The highest ranked from BLASTp and KA-Search are slightly biased towards their used metric; however, the closest match within the top-100 from both methods are very similar. Unlike all the other methods KA-Search is exhaustive, so it finds the exact closest match every time using the KA-Search identity.

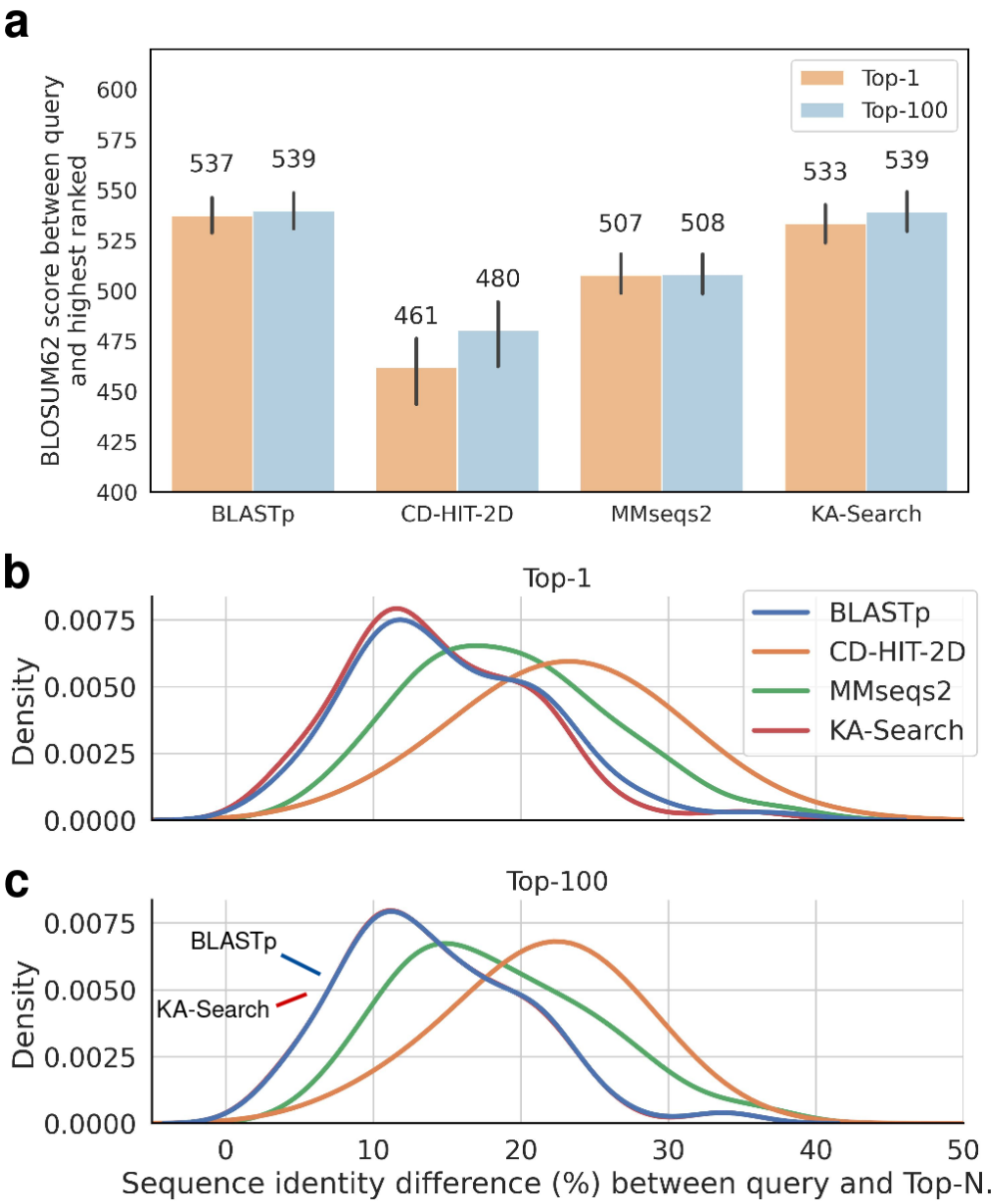

Figure 3.6: Highest similarity comparison between KA-Search and the commonly used protein search tools BLASTp (version 2.13.0) [22], CD-Hit-2d (version 4.8.1) [147] and MMseqs2 (version 13.45111) [148]. The default settings were used for each tool, and each search was done against the whole variable domain. BLOSUM62 scores were calculated by aligning with the Smith-Waterman algorithm using BLOSUM62 [149] as the substitution matrix and gap costs of 11 for existence and 1 for extension. **a**, Highest similarity, was first compared with the average BLOSUM62 score between the test queries and target sequences from the top-1 and the closest within the top-100 returned sequences from each search tool. Highest similarity was then compared using the density of sequence identity difference between the test queries and, **b**, the top-1 returned sequences and, **c**, the closest within the top-100 returned sequences from each search tool. For top-100, the density of BLASTp and KA-Search are nearly perfectly overlapping. Sequence identity was calculated as the percentage of exact matches after alignment with the Smith-Waterman algorithm using BLOSUM62 as the substitution matrix and gap costs of 11 for existence and 1 for extension.

3.4.3 Immune repertoire mining with the COVOX-253 antibody

COVOX-253 is an antibody which binds to the neck of SARS-CoV-2's Receptor-Binding Domain (RBD) [156]. Using KA-Search, up to the 1,000 closest sequences to the heavy chain of COVOX-253, with over 90% identity, were extracted for four different regions: the whole V domain, the three CDRs, the CDR3 and the paratope. The paratope was derived from the PDB structure 7BEN and defined as any residue in the antibody which was within 4.5Å of the RBD [79]. Fig. 3.7a shows the disease of the patient the antibody sequence found in OAS comes from and in Fig. 3.7b each antibody's combination of V and J genes. Most matched sequences are derived from the gene alleles IGHV1-58*01 and IGHJ3*02; however, COVOX-253's CDR3 is seen with six different V gene alleles, IGHV1-58*01, IGHV1-58*02, IGHV1-18*01, IGHV1-46*01, IGHV1-69*10 and IGHV1-69*13.

Searching for the closest match using the whole V domain returns 822 sequences from healthy individuals and 178 from patients with one of nine different diseases, eight which are SARS-CoV-2. Searching with the CDR positions returns one sequence from a healthy individual and 789 sequences from patients with SARS-CoV-2, while searching with the CDR3 or paratope positions returns 197 and 124 sequences, respectively, all from SARS-CoV-2 infected patients. The fact that OAS-aligned only contains ∼84 million heavy chains from patients with SARS-CoV-2, equivalent to ∼4% of all heavy chains in OAS-aligned, highlights the importance of being able to search over specific regions.

3.5 Discussion

Immune repertoire mining is a powerful method for identifying antibodies in nature which are similar to an antibody of interest and can help indicate likely specificity or immunogenicity. However, the number of available antibody sequences are now in the billions and is continuing to grow. Therefore, repertoire mining for highly similar sequences has become increasingly computationally expensive. Existing

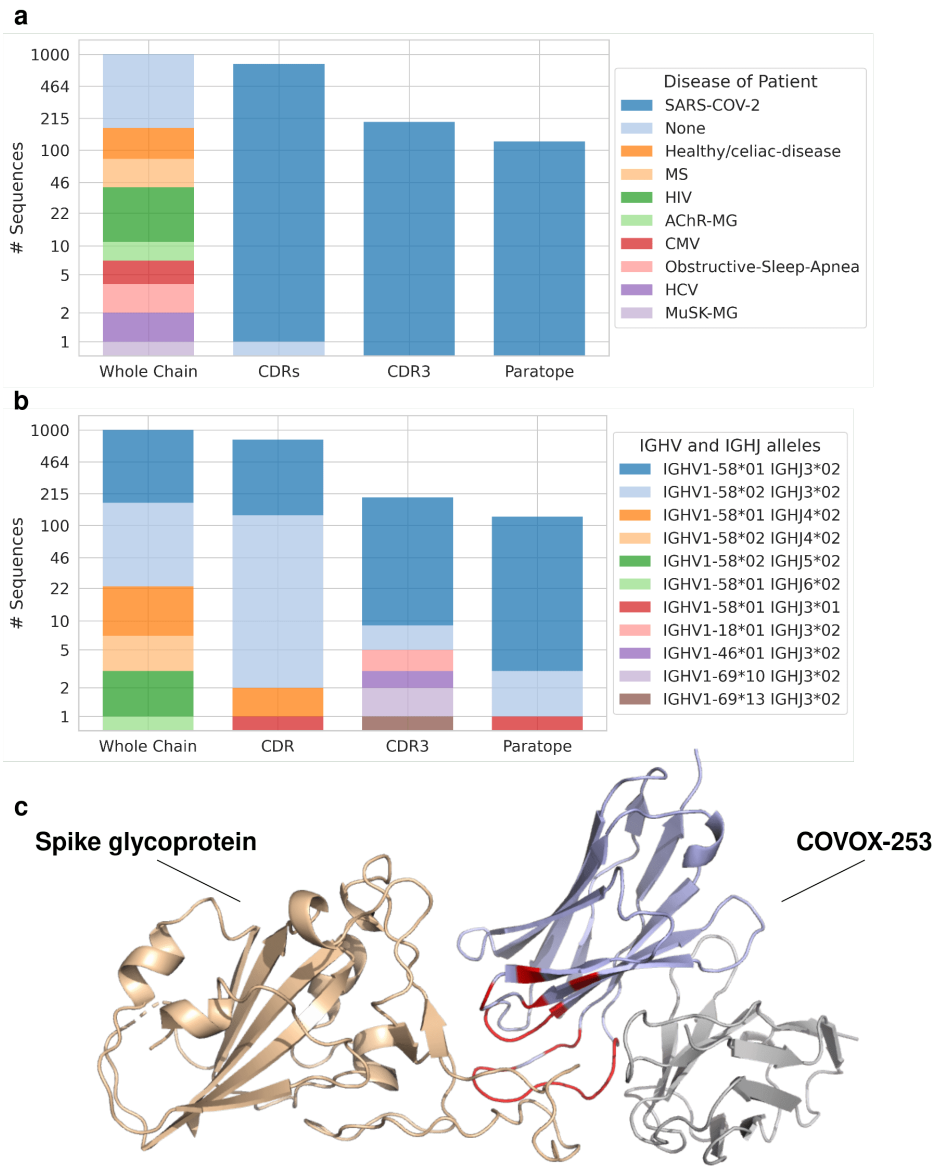

Figure 3.7: A KA-Search for sequence matches to the heavy variable domain of the SARS-CoV-2 RBD binding antibody COVOX-253. Returned antibodies with over 90% identity across four different regions were visualised based on, **a**, the disease state of the patient and, **b**, which V- and J-genes the antibody sequence is derived from. The y-axis is scaled logarithmically to better visualise the data. **c**, The variable domain of COVOX-253's heavy (purple) and light (grey) chain with the bound spike glycoprotein (beige), derived from PDB 7BEN. The paratope of the heavy chain, which was used to search with KA-Search, is shown in red.

approaches are limited by either being low-throughput, inaccessible for large scale searches, non-exhaustive, not antibody specific, or only searching against entire V domain sequences. There is therefore a need for a specialized tool, optimized for a rapid and exhaustive search of any antibody region against all known antibodies,

to better utilize the full number of available repertoire sequences.

In this paper, we introduce Known Antibody Search (KA-Search), a platform independent antibody search tool. KA-Search finds antibody sequences with an accuracy comparable to BLASTp, while being over an order of magnitude faster and allowing searches over specific regions. KA-Search exploits antibody numbering, allowing us to pre-align antibody sequences to a fixed-length vector. This circumvents pairwise alignment during search, an otherwise time-consuming step. This was done to keep the alignment short and increase speed. The increased speed allows KA-Search to avoid prefiltering and be exhaustive while still retaining a competitive speed. Avoiding prefiltering is crucial, as current prefiltering techniques greatly reduce sensitivity when searching highly related proteins, such as antibodies, where a single mutation can be of great importance. While pre-aligning the antibody sequences increases search speed, the initial pre-alignment is slow. We therefore provide a pre-aligned dataset of the current OAS, ready to use for searching. This dataset can be extended with future OAS updates or in-house data without the need to re-align the existing sequences. A guideline for preparing custom data for search with KA-Search is available at `https://github.com/oxpig/kasearch`.

Pre-aligning sequences also opens new use-cases. Instead of only searching against the whole antibody V domain, searches can now be focused on specific positions in the alignment. Searches can be specific for the CDRs or regions specific for individual antibodies, such as the paratope. This flexibility allows for studies which were previously difficult to execute. As an example, previously an extensive study was needed to search OAS across the whole V domain, CDRs and CDR3 for the closest match to a set of 242 therapeutics [144]. The same study can now be done on the 804 therapeutics within Thera-SAbDab [59] (as of August 2022) with KA-Search in less than two days compute and little configuration (see Supplementary Fig. A.1 and Supplementary Data 1). KA-Search can also

extract the metadata from OAS for the matched sequences, which can be used to obtain new insights about an antibody of interest. Using KA-Search to find the closest sequences with or above 90% identity across four different regions for the SARS-CoV-2 RBD binding COVOX-253 demonstrates the power of searching across particular regions. The closest matches from searching with the whole V domain comes from healthy patients or patients with a variety of diseases. However, binding region specific searches only return sequences found in SARS-CoV-2 infected patients. These sequences could therefore also have RBD binding properties and are possible candidates for further affinity studies. Sequences from patients not previously infected with SARS-CoV-2, but which have a high identity across the whole V domain, could also be derived from an immune response against one of the common cold coronaviruses. These coronaviruses are responsible for an estimated 10–20% of viral respiratory infections and are known to have some cross-reactivity with SARS-CoV-2 **??**. Investigating the genes of the closest sequences, also potentially indicates which other frameworks a region of interest could exist on. For COVOX-253, the CDR3 is seen in sequences with six different V genes, which each could be possible framework candidates for the CDR3. The ability to return high numbers of close matches without decreasing speed, also opens up KA-Search as a means for creating multiple sequence alignments of similar antibody sequences.

A limitation of KA-Search, is that the current version only searches across the V domain, disregarding the constant domain. This is driven by the current very limited number of available sequences of the constant domain. KA-Search can also only search with and against sequences which can be numbered and aligned with the 200 unique positions in the canonical alignment described in the methods. The 0.2% of sequences with rare insertions cannot be searched with KA-Search and are excluded from OAS-aligned. The rare insertions are seen across the whole V domain and while most are likely derived from sequencing or ANARCI numbering errors, e.g. the highly unlikely eight residue insertion giving position 81I, some rare insertions can be contributed to limited data of certain species within OAS, e.g.

position 112M seen in camel sequences with a CDR3 longer than 37 residues. As currently OAS mainly contains human and mouse sequences, the 200 positions cover those species well but may as described cover less of the sequences derived from other species. With OAS growing, the unique positions can be updated in the future to better handle other species. Sequences that failed to be aligned are also provided in the OAS-aligned download and can be searched using other methods if desired, for example if the sequence of interest contains rare insertions. Further, KA-Search currently only finds similar sequences using sequence identity. While this is sufficient for finding sequences with few mutations, calculating the sequence similarities using a substitution matrix would allow the exploration of more distant matches.

KA-Search can be run on any system with Python, requiring only 6GB RAM for searching OAS-aligned and 2GB RAM for OAS-aligned-tiny. This enables any researcher to readily search for similar sequences. Currently, when searching with few queries, the primary bottleneck related to the speed is the loading of data into memory to search against. For optimal use or large scale studies, KA-Search therefore benefits considerably from searching with many queries simultaneously on a high-performance computer using multiple CPUs.

KA-Search's speed, exhaustiveness and flexibility allows it to search the vast numbers of antibody sequences now available seamlessly, find viable mutations and gain new insight into antibodies of interest. We therefore believe KA-Search is a useful tool that will allow the antibody community to explore antibodies in new ways. To maximize KA-Search's possible contribution to the community, KA-Search is open source and freely available at `https://github.com/oxpig/kasearch`.

In the next chapter, we will describe the creation of the Patent and Literature Antibody Database (PLAbDab), a resource of functionally diverse, literature-annotated paired antibody sequences and structures. Additionally, we will discuss

various strategies for probing this data, including the utilization of KA-Search for investigating the potential properties of a given antibody.

4

The Patent and Literature Antibody Database (PLAbDab): an evolving reference set of functionally diverse, literature-annotated antibody sequences and structures

Contents

4.1 Chapter abstract

Antibody databases, derived from next generation sequencing (NGS) of B-cell receptors (see Chapter 2), accommodates vast amounts of functional antibodies. However, since the antibodies are sequenced without antigen context, these databases contain limited information about binding characteristics. In Chapter 3, we demonstrated the effective use of KA-Search for finding antibodies in OAS with a similar binding interface, thereby discovering potential similar binders. However, this method requires a known binder and does not guarantee true binders. Conversely, there exists a large body of academic literature and patents dedicated to their study and concomitant conversion into therapeutics, diagnostics, or reagents. These documents often contain extensive functional characterisations of the sets of antibodies they describe. However, leveraging these heterogeneous reports, for example to offer insights into the properties of query antibodies of interest, is currently challenging as there is no central repository through which this wide corpus can be mined by sequence or structure.

In this chapter, we present PLAbDab, our patent and literature antibody database, a collection of 150,000 paired antibody sequences from over 10,000 small scale studies. PLAbDab is self-updating and can be used for annotating query antibodies with potential antigen information from similar entries, analysing structural models of existing antibodies to identify modifications that could improve their properties, and compiling bespoke datasets of antibody sequences/structures known to bind to a specific antigen. This chapter is based on the below paper, where I am a co-first author alongside Brennan Abanades. I contributed to writing the paper and code, as well as conducting the case study using PLAbDab to generate new datasets.

> Abanades B., **Olsen T.H.**, Raybould M.I.J, Aguilar-Sanjuan B., Wong W.K., Georges G., Bujotzek A., and Deane C.M. (2023) The Patent and Literature Antibody Database (PLAbDab): an evolving reference set of functionally diverse, literature-annotated antibody sequences and structures. *BioRxiv*.

4.2 Introduction

With next generation sequencing (NGS) enabling researchers to take snapshots of the immune repertoire of an individual at a given point in time, vast amounts of single chain antibody sequences have been generated. In Chapter 2, we described how efforts to compile this data has led to the creation of datasets such as OAS [52] and iReceptor [152] which contain the sequence of the heavy or light variable domain for billions of antibodies. Conversely, paired VH/VL sequence data is more expensive to generate and currently just over a million paired antibody sequences can be found in OAS [141]. However, with the binding site in antibodies sitting across both chains, paired data gives a more complete picture of how and to what an antibody binds [39, 73].

Although OAS has proven invaluable as a source of functional antibodies and to compare repertoires between individuals, it provides little information on the functions of individual sequences within a repertoire. However, there also exists a large number of smaller scale studies, each one dedicated to investigating a small number of antibodies. When combined, the antibodies from these studies amount to a large number of sequences with rich metadata. There are a number of databases that aim to compile subsets of this data, for example SAbDab [157, 158] for antibodies with resolved crystal structures, Thera-SAbDab for antibody therapeutics [59], CoV-AbDab for COVID-19 binding antibodies [159], or PAD for unpaired antibody sequences from patents [160]. Paired antibody sequences with information on their epitope can also be obtained from IEDB [161].

Here we present PLAbDab (the Patent and Literature Antibody Database), a self-updating repository containing 150,000 paired antibody sequences, of which over 65,000 are unique, from over 10,000 small scale studies. PLAbDab is larger than any other non-NGS database of paired antibody sequences by at least an order of magnitude. We make the data freely available and provide methods to rapidly search it by either sequence identity using KA-Search [162], structural

similarity [44, 163], or by keywords in the title of the study. Each of the sequences comes with a direct link to its source material, making it easy to obtain additional information about any antibody of interest. PLAbDab is freely available via Github (`https://github.com/oxpig/PLAbDab`) and as a searchable webserver (`https://opig.stats.ox.ac.uk/webapps/plabdab/`).

4.3 Methods

4.3.1 Collecting unpaired antibody sequences

The majority of data in PLAbDab is extracted from the Protein database of NCBI [164]. The BioPython Entrez module [165] is used to query the database for entries containing the words "antibody" , "antibodies", "immunoglobulin", "scfv" or "bcr" in any of their fields. Due to the lack of any automated method to accurately differentiate antibody heavy chains from nanobodies, entries containing the words nanobody or nanobodies are removed at this stage. Entries with sequences longer than 1000 amino acids or shorter than 70 are also removed. Around 2.5 million entries were returned from this search.

Sequences for each of these 2.5 million entries were then searched for antibody variable domain sequences using ANARCI [26]. This resulted in around 530,000 potential antibody variable domain sequences from around 13,000 different sources.

4.3.2 Creating paired antibody sequences from the unpaired data

For a large number of entries the metadata provides enough information to pair the heavy and light variable domains. For entries from the same literature source, VH-VL pairing was attempted using the following heuristics:

1. "Same entry" - If there is only one VH and only one VL within a single entry, these are paired together.

2. “Unique word" - If there is a unique non-common word that is found in the description of exactly one VH and one VL, these are paired together. Uncommon words are defined as words found in the description of less than 20 entries in the unpaired database.

3. “Unique source" - Some entries contain information on the experimental source, such as the isolate or clone. If there is a unique VH and VL from the same experimental source they are paired together.

4. “Patent text" - Entries from patents have sequence IDs by which they are referred to in the patent text. If the patent text mentions a VH and VL within the same paragraph, these are paired together. If a paragraph in the patent text mentions the same number of VH and VL entries, these are paired in the same order as they are mentioned.

5. “Unique chain" - If there is a single VH (or VL) from one source, this entry is paired with all other VL (or VH) entries from that source.

6. “Ordered entries" - If there are the same number of VH and VL entries from one source, they are paired in order.

The described strategies for pairing VH-VL sequences have varying levels of accuracy. Twenty entries paired by each of the methods described above were randomly selected and manually checked to estimate the accuracy of each pairing method. Methods 1-4 achieved perfect accuracy on the twenty entry test set. We therefore refer to these methods as pairing with “high confidence”. Method 5 incorrectly paired one entry and Method 6 incorrectly paired two entries. Each paired entry is given a flag to indicate how it was paired, making it easy for users to filter for less accurate pairing methods. To further increase the coverage of the dataset, sequences from both SAbDab and Thera-SAbDab were also added to PLAbDab.

4.3.3 Labelling antibodies with potential antigen information

The target of the majority of antibodies in PLAbDab can be inferred from a manual review of their source literature. Although possible, this process can be time-consuming. To reduce the number of entries to review, we provide a “targets_mentioned" column that lists common antigens mentioned in the source material.

For patent entries, potential antigens were identified by searching the title, abstract, and claims sections for mentions of commonly recognised antigen names. When a term in the source literature was prefixed with “anti-", it was also included in the list of possible antigens. For non-patent entries, only the title was searched for potential targets.

The use of the “targets_mentioned" label exhibits high recall: when an antibody’s antigen is known, it is often mentioned in the analysed text. However, its accuracy is moderate. Accurate labelling of each entry with antigen information requires manual inspection. Nonetheless, this approach offers a fast and straightforward method for generating an initial reference set.

4.3.4 Searching PLAbDab

To allow users to rapidly search the database for antibodies with a similar sequence, we implemented KA-Search [162]. KA-search is able to carry out rapid and exhaustive sequence identity searches, allowing users to find similar sequences over the whole variable domain, the CDR loops, the CDR-H3 or a user defined region. The entire database can be queried with KA-Search in under 5 seconds on 5 CPUs.

Structurally similar antibodies can have similar functions, even if they are distantly related in sequence [44, 163]. To enable CDR structure-based searching,

we model all paired sequences in the database using ABodyBuilder2 [166]. Entries missing more than eight residues at the start of the sequence were restored using AbLang (see Chapter 5) [129] prior to modelling. Antibody sequences with non-standard residues, or for which ABodyBuilder2 was not capable of generating a refined model, are labelled as such. Entries with the same sequence were only modelled once, in total 64,000 unique antibody models were generated.

To structurally search PLAbDab, query sequences are first modelled using ABodyBuilder2, without refinement. The refinement step was skipped as in order to boost speed with minimal compromise on backbone model quality. The framework of the predicted structure is then aligned to the structure of all other entries in PLAbDab with the same CDR loop lengths. Lastly, the carbon-alpha (C_α) root-mean squared deviation (RMSD) over all CDR residues is computed and used to rank entries. Searching the database in this way takes around 10 seconds on 5 CPUs.

Finally, as each sequence in PLAbDab is linked to a patent or publication, we provide users the ability to search fields, such as the title of the study, using regular expressions.

4.4 Results

4.4.1 Database statistics

Currently the total number of entries in PLAbDab is around 150,000, over 90% of which are paired with high confidence using methods 1-4, or comes from therapeutic or crystal structure entries. As can be seen in Fig. 4.1A, the number of antibody sequences that could be collected by the PLAbDab methodology has been steadily growing since the early 2000s, with somewhere between 10,000 and 30,000 new antibody sequences being published each year for the last five years.

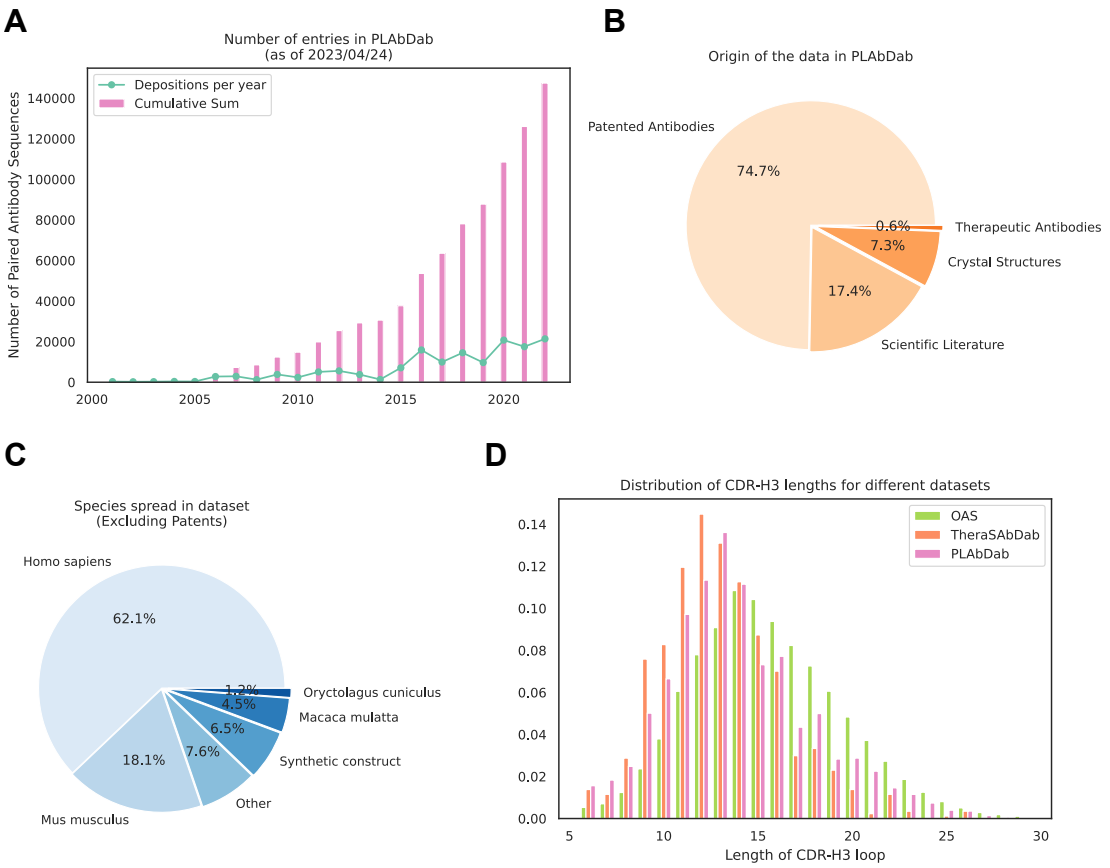

Figure 4.1: Summary statistics for PLAbDab. (A) Number of new antibody sequences each year and total number of antibody sequences that could be collected by the PLAbDab methodology. (B) The proportion of data in PLAbDab by type of source. (C) The proportion of antibodies from each species in non-patent entries. (D) Comparison of the distribution of CDR-H3 lengths in PLAbDab, Thera-SAbDab [59], and OAS [141].

Fig. 4.1B shows the distribution of PLAbDab entries by source, around three quarters of entries come from antibody sequences described in patents, while less than 20% are derived from the scientific literature. This may be due to patents following a more standardised way of depositing antibody sequences. The rest of the entries in PLAbDab come either from solved structures in SAbDab [157], or from additional sequences in Thera-SAbDab [59].

Patent applicants are not obligated to upload species information about their sequences to NCBI. This means that the majority of entries in PLAbDab lack a species annotation. Unfortunately, computationally predicting the species origin

for unannotated patent antibodies in a systematic way is likely to be ineffective due to the extent of non-natural engineering these sequences undergo. However, species information does exist for the majority of the entries sourced from the academic literature. Across the sequences with species information, most entries are labelled as human, with a smaller number of sequences annotated as mouse, macaque or rabbit (Fig. 4.1C).

Search method	Retrieved (cons.)	Sources (cons.)	Unique (cons.)
PD-1			
VH identity	576 (222)	180 (41)	258 (61)
VH+VL identity	155 (132)	39 (28)	22 (16)
CDR structure	227 (168)	60 (26)	46 (19)
CDR structure+identity	127 (127)	29 (29)	14 (14)
RSV			
VH identity	2 (2)	1 (1)	2 (2)
VH+VL identity	2 (2)	1 (1)	2 (2)
CDR structure	29 (29)	4 (4)	23 (23)
CDR structure+identity	4 (4)	1 (1)	4 (4)
CD200			
VH identity	13 (13)	4 (4)	2 (2)
VH+VL identity	13 (13)	4 (4)	2 (2)
CDR structure	1175 (95)	259 (15)	440 (50)
CDR structure+identity	37 (37)	5 (5)	5 (5)
CD3e			
VH identity	26 (15)	9 (4)	9 (2)
VH+VL identity	18 (15)	6 (4)	4 (2)
CDR structure	94 (51)	40 (15)	34 (14)
CDR structure+identity	15 (15)	4 (4)	2 (2)

Table 4.1: Results from searching PLAbDab with three different queries using four methods. For each method, the number of retrieved entries, the number of different sources the entries come from and the number of unique antibodies found is given. The value given in parenthesis is the number of entries or sources that are functionally consistent with the query. Details on the three query antibodies and the four searching methods are given in the main text.

It has been observed that therapeutic antibodies tend to have shorter CDR-H3 loops than those seen in large repertoire studies, an observation that may be due to longer CDR loops leading to issues during therapeutic antibody development [87, 167]. The average CDR-H3 loop length from antibodies in PLAbDab (around

∼14.0) falls somewhere between the average CDR-H3 loop length of OAS (around ∼15.6) and Thera-SAbDab (around ∼12.9) (Fig. 4.1D).

4.4.2 Searching PLAbDab

As described in the methods, PLAbDab allows users to search the database using a variety of methods. To demonstrate the sequence and structure search options, we compared the results of searching the database in four ways:

- For antibodies with a sequence identity of over 90% over the VH (VH identity).
- For antibodies with a sequence identity of over 90% for both the VH and VL (VH+VL identity).
- For antibodies with a C_α RMSD over the CDR loops of under 1.25Å to an ABodyBuilder2 model of the query (CDR structure).
- For antibodies with a C_α RMSD over the CDR loops of under 1.25Å to an ABodyBuilder2 model of the query and sequence identity over the CDR loops of over 80% (CDR structure+identity).

We performed the above searches for four antibodies each of which target a different antigen. Antibody one binds programmed cell death protein 1 (PD-1), and was taken from the patent "Anti-PD-1 antibodies and methods of use thereof."; antibody two binds Respiratory Syncytial Virus (RSV), sourced from the paper "Rapid profiling of RSV antibody repertoires from the memory B cells of naturally infected adult donors." [168]; antibody three is the therapeutic antibody Samalizumab, which binds the OX2 membrane glycoprotein (CD200); and antibody four is the therapeutic antibody Foralumab, that targets the CD3-epsilon peptide (CD3e). For each of the four queries we took the sequence and built an ABodyBuilder2 model for use in the structure searches. The result of using each search method described above to query PLAbDab with these antibodies is shown in Table 4.1.

The PD-1 case study shows the benefits of having the paired heavy and light chain variable domain sequence. While a sequence identity search over the VH finds antibodies binding to the same antigen around 25% (61 out of 258) of the time, including the VL improves the accuracy to around 75% (16 out of 22). The fact that there are many highly identical heavy chain sequences that bind to different antigens suggests that, for this antibody, the light chain may contribute significantly to binding. To validate this, we analysed the crystal structure of an antibody bound to RSV which is returned by all four search methods (PDB ID: 5GGR). According to Arpeggio [169], the paratope is evenly split between the heavy and light chain, with eight heavy chain and seven light chain residues found to interact with the antigen.

Searching PLAbDab for entries similar to our PD-1 target by CDR structure returns antibodies that target the same epitope around 40% of the time (19 out of 46), but many of them are from different studies than those found by sequence search. All the entries returned by a CDR structure plus sequence identity search target the same epitope as our query.

For the RSV binding antibody, there are only two entries in PLAbDab with greater than 90% sequence identity over the VH. Searching PLAbDab for entries with a similar CDR structure returns 23 unique antibodies which bind the same epitope. This suggests that this antibody may require a very specific CDR loop conformation to bind the RSV Fusion Glycoprotein [171] at that specific site. Two of the retrieved entries belong to antibodies with resolved crystal structures (PDB IDs: 7LUC and 7LUD), one of which is bound to the antigen. Fig. 4.2 shows the similarity of these structures to the ABodyBuilder2 predicted structure of the RSV binding antibody query.

Only two unique antibody sequences in PLAbDab have a VH sequence identity of over 90% to Samalizumab. A structure search using Samalizumab as a query

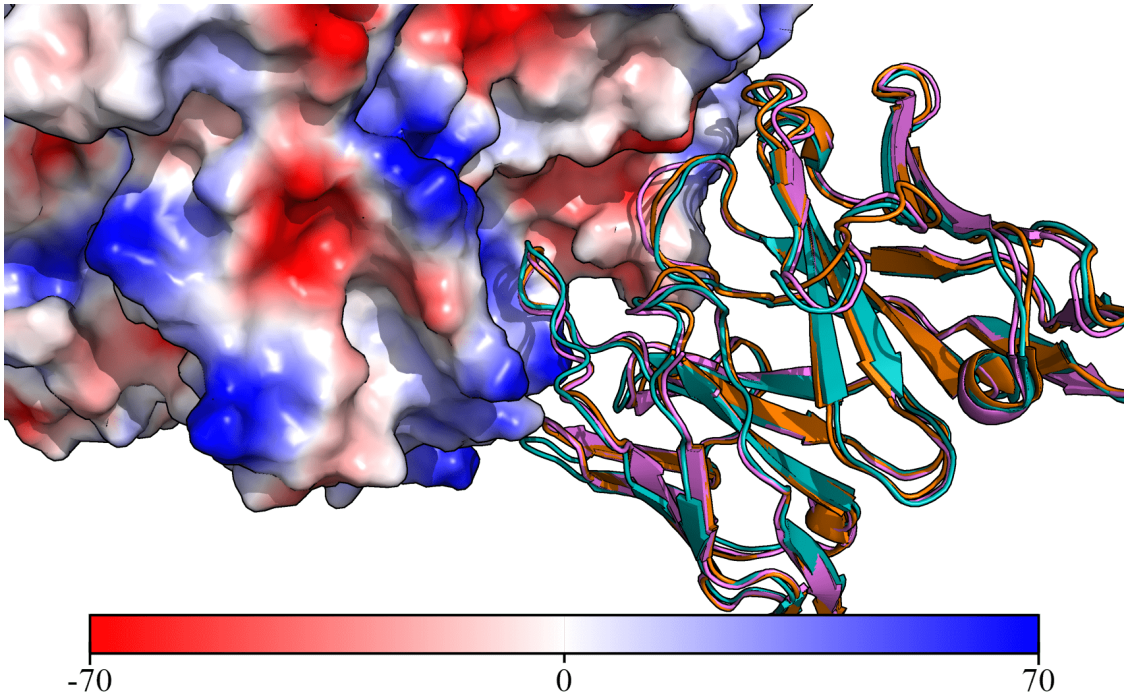

Figure 4.2: Crystal structures of anti-RSV antibodies aligned to the structural model of the query. A cartoon representation of the ABodyBuilder2 model for the query antibody is shown in purple with the crystal structure for 7LUC shown in blue and the crystal structure for 7LUD shown in orange. The surface of the RSV antigen is coloured based on electrostatics using PyMol [170].

results in over a thousand antibodies that bind a very diverse set of epitopes. In Fig. 4.3 we show a selection of antibodies with very similar CDR backbone structures to Samalizumab binding to four very different epitopes on different antigens. This indicates that the backbone structure of the CDR loops is likely not the primary driver of binding specificity for this antibody. However, when adding an 80% sequence identity cutoff over the CDR loops to the structure search, we are left with five unique antibody sequences all of which bind the same epitope.

For the Foralumab case study, only two out of nine sequences with a VH identity of over 90% were found to also bind CD3e. Adding an additional 90% sequence identity filter over the VL reduces the number of non-CD3e binders returned to two. Searching PLAbDab for entries that share a similar CDR structure to Foralumab expands the number of unique sequences retrieved to 34, out of which 14 were

found to bind CD3e. The CDR structure plus sequence identity search achieves perfect accuracy, but reduces the number of retrieved sequences to two.

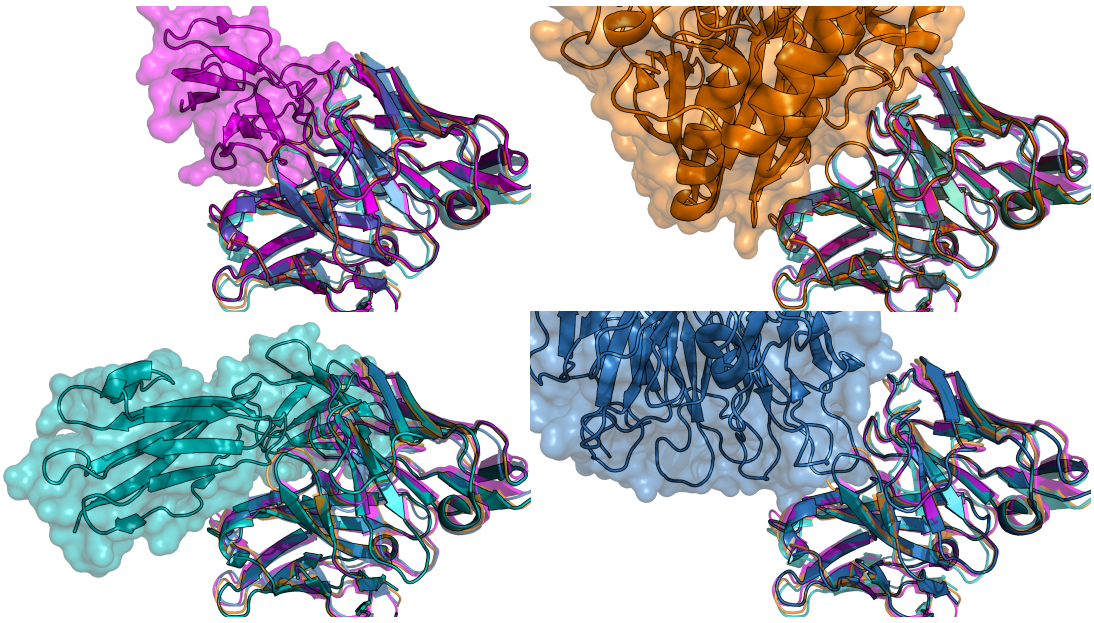

Figure 4.3: Crystal structure of four antibodies with very similar structure that bind very different antigens. The crystal structure of all four aligned antibodies is shown in each figure (4AL8 in purple, 7XY8 in green, 7LA4 in blue, and 7BBJ in orange). On the top left they are shown in complex with the antigen of 4AL8 (Dengue virus DII protein); top right 7BBJ (CD73); bottom left 7XY8 (Emmprin CD147); and bottom right 7LA4 (Integrin $\alpha_{\mathrm{IIb}}\beta_3$).

These four case studies highlight the utility of PLAbDab's data and various search methods to identify antibodies with similar function to a query. The case studies also highlight that any computational search should be further investigated via both the meta data in PLAbDab and, if possible, experimental means.

4.4.3 Using PLAbDab to Generate New Datasets

As PLAbDab contains a large number of antibody sequences targeting a diverse set of antigens, it can be used to facilitate the generation of antigen-specific antibody libraries. For example, searching PLAbDab for entries containing the keywords "ebov" or "ebola" in their title returns almost 1,500 unique antibody sequences from 56 sources. Using the keywords "hiv" or "immunodeficiency" finds over

6,200 entries with over 3,800 unique antibody sequences from over five hundred sources, while “autoimmune” and “autoantibody” returns over 400 unique antibody sequences from over 90 sources. Not every antibody sequence retrieved from a patent or paper will necessarily bind the searched target, but such prefiltered sets provide a very valuable starting point towards generating an antigen specific library to a target of interest.

To benchmark how effective searching the titles of the sources of PLAbDab entries by keywords is at retrieving relevant antibodies, we manually inspected the results of a search by expression “hiv|immunodeficiency". By manually inspecting the source literature from a random set of 100 antibodies from different studies, we found that 88 were true HIV binders. Of the non-HIV binders, four were anti-idiotype antibodies, three bound CD4, three bound other antigens and two were incorrectly paired. Using the search terms “covid|corona|sars”, we inspected another 100 randomly-selected antibodies from different studies. Of these, 98 targeted a coronavirus, one targeted ACE2, and the specificity of the last could not be verified. This highlights how PLAbDab can aid in the creation of antigen-specific antibody datasets, but also shows that manual inspection of the extracted sequences is still required to remove false positives.

4.5 Discussion

We present PLAbDab, a large database of paired antibody sequences from patents and papers. Each entry in the database contains a link to its original source material, which relates the paired sequences to useful functional information. PLAbDab can be automatically updated without any manual input, making it straightforward to keep up to date with the latest publications and patents.

To demonstrate the utility of PLAbDab, we explore different search methods for identifying antibodies with similar binding properties to a given antibody of interest. Our results demonstrate that for different antibodies, different types

of search yield the best results. In all three cases investigated, CDR structure alongside CDR sequence identity gave the most accurate results, but in most cases, this also missed functionally cognate antibodies. The inclusion of the light chain in any search always improved accuracy, in agreement with recent studies that have highlighted the importance of the light chain residue motifs for antigen binding [73], and restricted heavy and light chain gene pairs observed among functional antibodies [39]. In one case searching by structure provides a link that is not immediately apparent from sequence search alone. This finding is consistent with recent work which uses structural information for epitope binning [44, 163].

One of the main challenges in developing PLAbDab, and the area with most potential for improvement, is the pairing of heavy and light chains based on metadata. Although the strategy described in this paper accurately pairs a large number of antibody sequences, there are still many for which pairing was not possible. The PLAbDab generation code also relies on authors publishing their antibodies to databases such as NCBI [164] or the PDB [172], and although many authors do submit their sequences to these databases, there remains a large body of antibody sequence data shared in the text, figures, or tables of papers. Patented sequences are deposited directly into NCBI by the U.S. Patent and Trademark Office (USPTO), or indirectly through collaborators for the European Patent Office (EPO) and Japan Patent Office (JPO). Patents in NCBI are therefore limited by what gets deposited by each patent office, like USPTO only depositing granted patents [164].

Nevertheless, PLAbDab already contains over 60k unique annotated antibody sequences providing an invaluable resource to the antibody research community. PLAbDab can also be used to generate antigen specific libraries, and while it might be biased towards certain antigens, like cytokines and immune receptors, it can still be used to train novel machine learning models or as a starting point for the development of novel therapeutics. These libraries can also be used to

investigate the different epitopes and binding antibodies for popular targets, such as Tumour necrosis factor (TNF). Future work, could potentially also solve the problem of manually double checking the antigen label by using language models like ChatGPT, specifically designed to retrieve antigen information. Additionally, PLAbDab could also include nanobodies. However, this would require a robust *in-silico* tool for correctly separating heavy chains from nanobodies, which does not currently exist.

We have made PLAbDab freely available to download and query via a webserver (`https://opig.stats.ox.ac.uk/webapps/plabdab/`) and the code is available to download via github (`https://github.com/oxpig/PLAbDab`).

In the following chapter, we will describe the development of AbLang, an antibody-specific language model that leverages the immense amount of data available in OAS to offer a novel and potentially groundbreaking approach to antibody design and discovery. Unlike the KA-Search method, which can be used to discover similar functional antibodies with potentially similar properties within OAS, AbLang can learn the intrinsic patterns of all antibodies in OAS. This knowledge can then be used to guide the design of antibodies by predicting potential useful mutations. To showcase AbLang's usability, we visualise its antibody representations and, as a case study, demonstrate its use for restoring fragmented antibodies.

5

AbLang: an antibody language model for completing antibody sequences

Contents

5.1 Chapter abstract

General protein language models have been shown to summarize the semantics of protein sequences into representations that are useful for state-of-the-art predictive methods (see Introduction 1.5.3). However, for antibody specific problems, such as restoring residues lost due to sequencing errors, a model trained solely on antibodies may be more powerful. Antibodies are one of the few protein types where the volume of sequence data needed for such language models is available, such as in OAS (see Chapter 2).

In this chapter, we introduce AbLang, a language model trained on the antibody sequences in OAS. We demonstrate the power of AbLang by using it to restore missing residues in antibody sequence data, a key issue with B-cell receptor repertoire sequencing, as over 40% of OAS sequences are missing the first 15 amino acids. AbLang restores the missing residues of antibody sequences better than using IMGT germlines or the general protein language model ESM-1b. Further, AbLang does not require knowledge of the germline of the antibody and is seven times faster than ESM-1b. The following chapter is based on the below paper. I conceived the project, designed the model, wrote the code, trained the models and performed all tests.

> **Olsen T.H.**, Moal I.H., and Deane C.M. (2022) AbLang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*. 2(1):vbac046

5.2 Introduction

Recent progress within protein informatics has led to the development of pre-trained protein representations, derived from protein language models such as ESM-1b [114], which, as described in Introduction 1.5.3, have been used to perform state-of-the-art predictive tasks. Such protein language models require vast amounts of training data and so far have tended to use all protein sequences and therefore be general protein representations [112–114]. With the creation of the Observed Antibody Space (OAS) database [52] and our subsequent update [141] (see Chapter 2), enough curated antibody sequences are now available to train a language model specifically for antibodies. An antibody specific model that has learnt the semantics of their sequences would allow for more precise predictions of antibody properties and new use cases.

Over the last decade, billions of antibodies have been sequenced [27]. However, as we describe in Introduction 1.3.1.1, in some cases the sequenced antibodies are missing residues due either to sequencing errors, such as ambiguous bases [173], or

the limitations of the sequencing techniques used [28]. We find in OAS, that ∼80% of the sequences are missing more than one residue at the N-terminus and ∼43% are missing the first 15 positions, and ∼1% contain at least one ambiguous residue somewhere in the sequence.

The ability to accurately restore these missing residues would increase data availability and be of benefit to antibody drug discovery. Currently, sequence imputation can only be done by identifying the correct ImMunoGeneTics (IMGT) germlines from the IMGT/GENE-DB [174] and using the germline sequence to add the missing residues. This approach requires correctly determining the allele of the sequence, a process that can be time consuming and/or produce ambiguous results.

Here, we present AbLang, an antibody specific language model trained on either the heavy or light chain antibody sequences from OAS. While AbLang can be used to create representations for residue or sequence specific predictions and residue engineering, in this paper we focus on showing how AbLang can be used to restore missing residues in antibody sequences, more accurately than using IMGT germlines or a general protein model like ESM-1b.

5.3 Methods

Two AbLang models were trained, one for heavy and one for light chains. Each AbLang model consists of two parts: AbRep, which creates representations from antibody sequences, and AbHead, which uses the representations to predict the likelihood of each amino acid at each position (Fig. 5.1).

AbLang was implemented using PyTorch 1.8.1 and was inspired by HuggingFace's [175] Transformer 3.0.2 library. AbRep follows the architecture of RoBERTa [107], except it uses a learned positional embedding layer with a max length of 160. Each of its 12 transformer blocks has 12 attenuated heads, an inner hidden size of 3072 and a hidden size of 768. From AbRep, the res-codings (768 values for each residue)

are obtained. AbHead follows the design of RoBERTa's head model, with a hidden size of 768.

During training, between 1-25% of residues from each sequence were selected, and of these, 80% were masked, 10% randomly changed to another residue and 10% left unchanged. One AbLang model was trained on heavy chain sequences for 20 epochs with a batch size of 8,192, and another on light chain sequences for 40 epochs with a batch size of 4,096. Both models were optimized using an Adam optimizer with a linear warm-up period for 5% of the steps, a peak learning rate of 0.0002, a learning rate decrease following a cosine function, and a weight decay of 0.01. For every dropout and layer normalization, a 0.1 rate and $1e^{-12}$ epsilon was used. The hyperparameters were selected to be similar to those used in the RoBERTa paper [107].

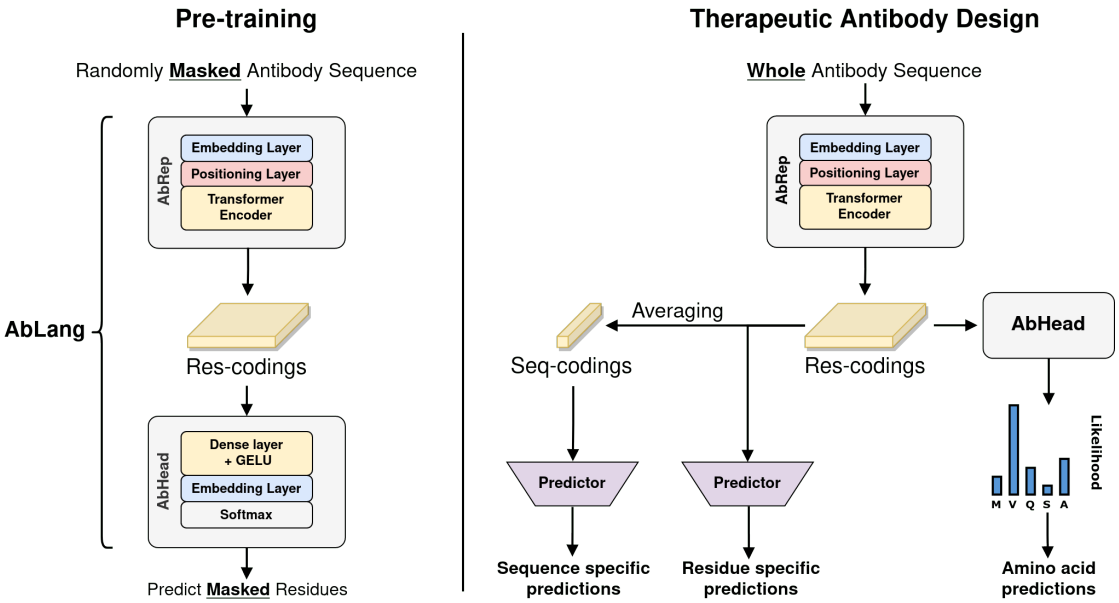

Figure 5.1: Overview of the architecture of AbLang, AbRep and AbHead, and examples of possible use cases. For pre-training, residues are randomly masked in each sequence and the masked residues are predicted and compared to the original residue. After pre-training, the model can, through different use cases, be used to improve therapeutic antibody design. AbHead can be removed and the res-codings from AbRep used for either residue or sequence specific predictions, or AbHead can be kept and used for restoring missing residues or exploring possible mutations.

5.4 Results

5.4.1 Data preparation

All antibody sequences seen three or more times in the OAS database as of Oct. 2021 were downloaded. The heavy and light chain sequences were then clustered separately based on identical CDR3s and thereafter clustered further by 70% identity over the whole sequence using Linclust [148], with the longest sequence selected from each cluster. The selected sequences were then randomly divided into training sets of 14,126,724 heavy and 187,068 light sequences, and two evaluation sets of 100,000 heavy and 50,000 light sequences. The training sets were then used to train AbLang as described in the Materials and methods section.

5.4.2 AbLang's antibody sequence representations

AbLang can be used to generate three different sets of antibody sequence representations. The first representation, the res-codings, consists of 768 values for each residue, useful for residue specific predictions. The second representation, the seq-codings, represent the whole sequence and is derived from the mean of all res-codings in a sequence. The seq-codings are 768 values for each sequence and are useful for sequence specific predictions. Additionally, they have the benefit of having the same length for each sequence, removing the need to align antibody sequences. Lastly, AbLang can be used to generate the likelihoods of each amino acid at each position in a given antibody sequence, useful for antibody engineering.

To investigate the sequence information extracted by AbLang and compare it to that of ESM-1b, we visualised the AbLang and ESM-1b sequence representations of 10,000 naïve and 10,000 memory B-cell sequences from Ghraichy et al (2021) with a t-SNE [176] plot (Fig. 5.2). t-SNE was used for dimensionality reduction, due to its emphasis on preserving small pairwise distances, in contrast to PCA [177], a commonly used method that focuses on maintaining large pairwise distances to

maximize variance.

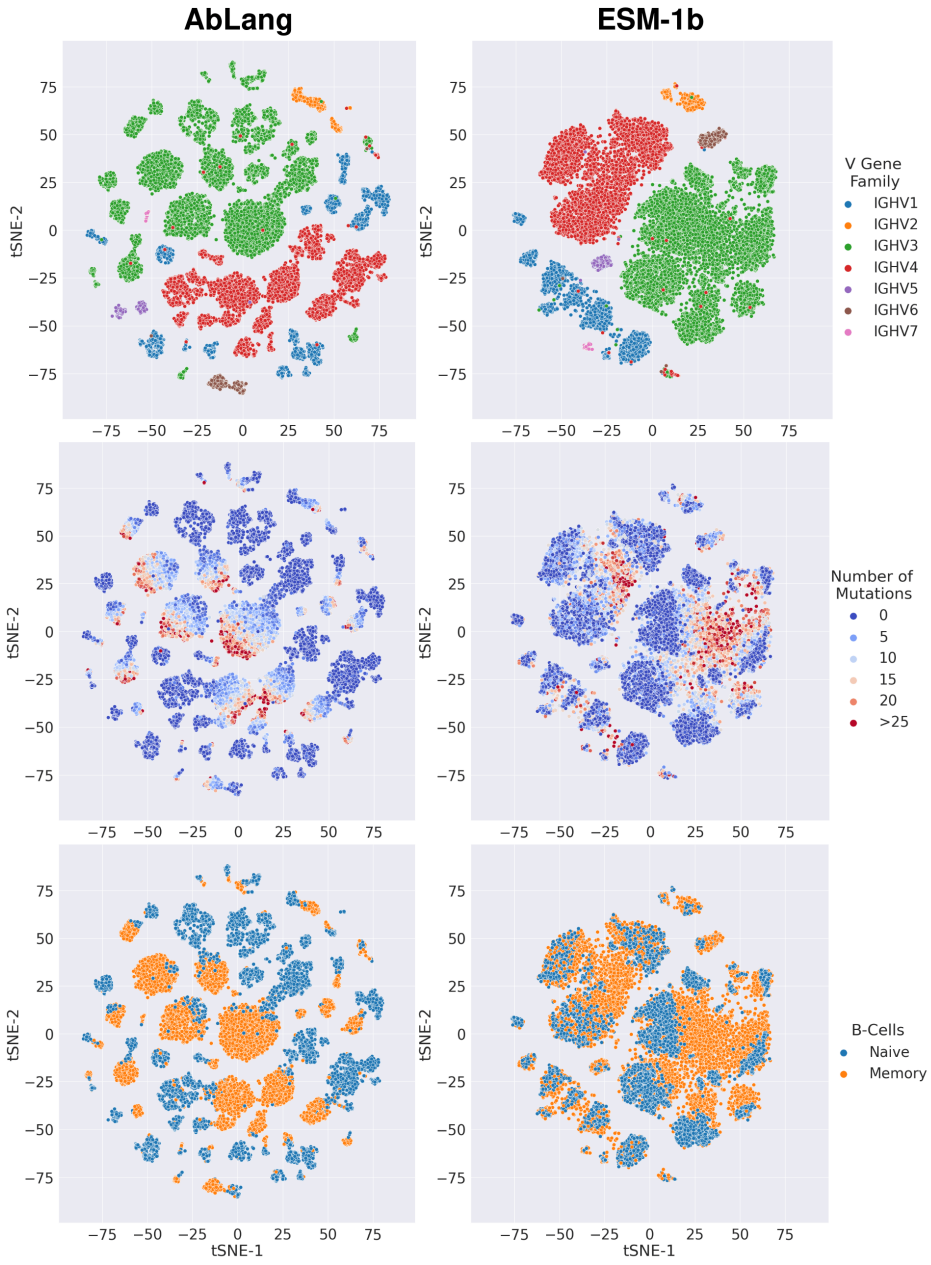

Figure 5.2: Comparison of AbLang and ESM-1b representations at clustering sequences based on their V-genes, originating cell type and number of mutations.

As Fig. 5.2 shows, AbLang and ESM-1b both separate the sequences based on their V-gene families, however, AbLang separates the V-genes into smaller clusters. These smaller clusters can partly be attributed to AbLang's finer separation of V-genes (see Supplementary Fig. B.1). Within the AbLang clusters, a clearer

separation can be seen between naive B-cells and memory B-cells than with ESM-1b's clusters. Further, the memory B-cells, in AbLang's clusters, appear to be ordered based on a gradual increase in mutations. This potentially indicates that AbLang representations contain information about the order of antibody mutations.

5.4.3 AbLang for restoring missing residues

AbLang's representations can be used for a plethora of antibody design applications. As an example, we use AbLang to restore missing residues in antibody sequences. Fig. 5.3 demonstrates the need for such a tool, showing how over 40% of the sequences in OAS are missing the first 15 residues and ∼80% of the sequences are missing more than one residue at the N-terminus.

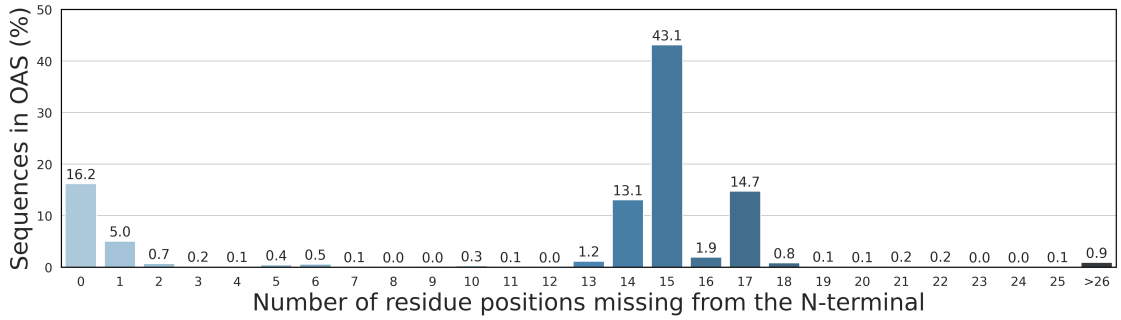

Figure 5.3: Overview of the antibody sequences in OAS, showing the percentage of sequences and number of residues they are missing from the N-terminus. Over 40% of the sequences in OAS are missing the first 15 residues.

The input to AbLang for sequence restoration is an antibody sequence with asterisks for unknown residues (Fig. 5.4). AbLang restore the missing residues by predicting the likelihood of each amino acid at the marked positions, with the amino acid with the highest likelihood then selected as the prediction.

We tested the ability of AbLang to restore both the N-terminus of antibody sequences and missing residues randomly scattered throughout the sequence. From the evaluation sets, 100 complete sequences for each of the 20 heavy and 42 light human V alleles seen in the evaluation set were randomly selected. These 2000

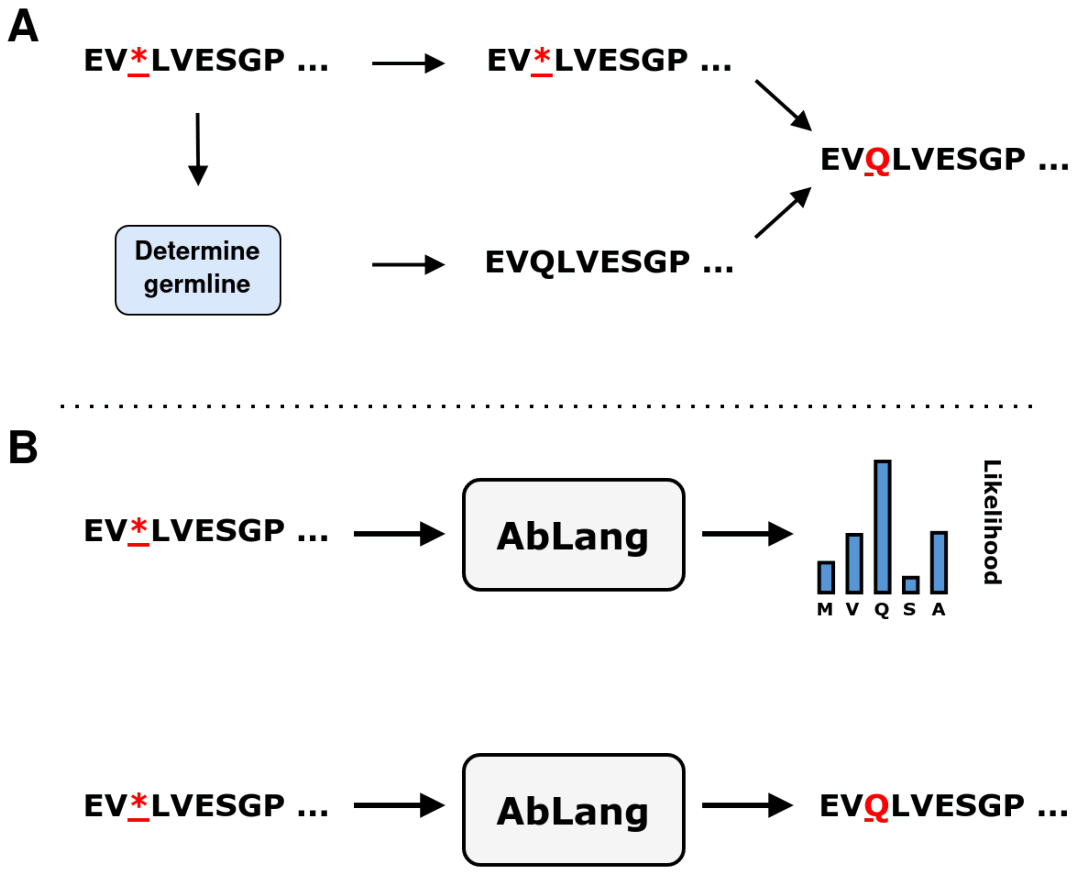

Figure 5.4: Illustration of how the IMGT-germline and AbLang method restores missing residues. A, from an input sequence the germline is determined and then used to restore missing residues. B, an input sequence has asterisks at the positions to predict. AbLang predicts the amino acid with the highest likelihood.

heavy and 4200 light sequences were used as the test set.

Fig. 5.5 shows a comparison of AbLang, to the general protein model ESM-1b and to the use of germline residues for the prediction of missing residues in an antibody sequence. Sequences were numbered in the IMGT scheme using ANARCI [26] and positions from 1 up to 30 were masked and then restored using the three different methods. The accuracy of this restoration was measured as the percentage of correctly predicted amino acids. IMGT germlines and AbLang achieve comparable accuracy, both restore missing N-terminus residues with accuracies of around 96% and 98% for the first 15 positions of the light and heavy chain, respectively. ESM-1b has far poorer performance achieving accuracies of

54% and 64%. The performance of IMGT germlines and AbLang are very similar, but the IMGT germline method requires knowledge of or accurate prediction of the germline, while AbLang can be rapidly deployed without any additional information.

In some cases, sequencing errors can result in residues being unknown at random sites throughout the antibody sequence. The ability of AbLang, IMGT germlines and ESM-1b to predict residues at randomly selected positions was also compared. Using the same test set as above, one, five or ten residues were randomly masked in each sequence's V domain. AbLang is more accurate at this task than both IMGT germlines and ESM-1b for both heavy and light chains. AbLang is also the fastest of the three methods, able to process 100 sequences in 6.5 seconds to ESM-1b's 44.9 seconds, using 4 cores on an Intel Core i7-10700.

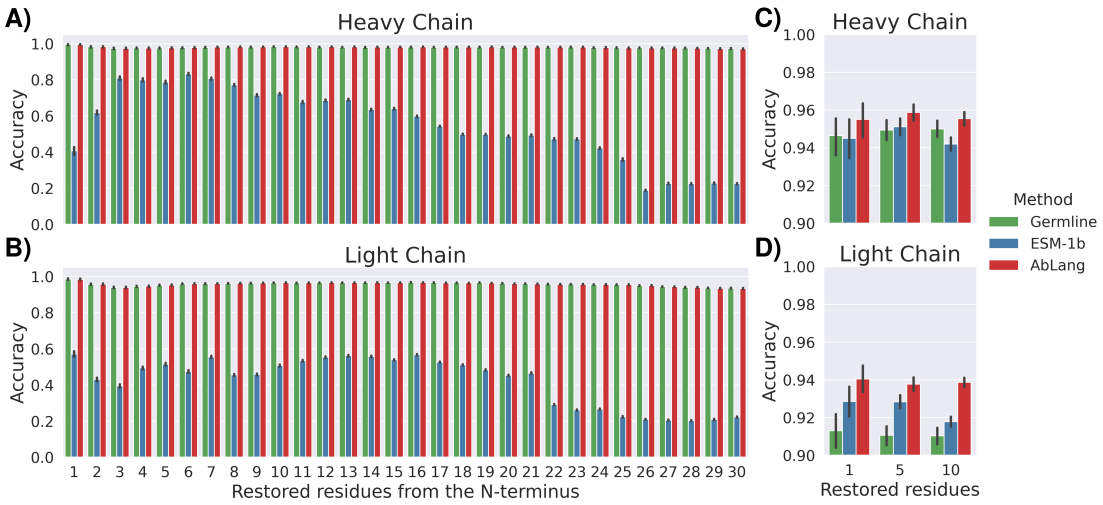

Figure 5.5: Antibody sequence restoration using IMGT germline sequences (green), a general protein language model ESM-1b (blue), and the antibody specific language model AbLang (red). A-B, shows the restoration of sequences missing up to 30 residues of the N-terminus, and C-D, the restoration of sequences with a random set (one, five or ten) of missing residues.

Often the number of missing residues at the N-terminus is unknown. To overcome this problem, we tested the use of ANARCI numberings and AbLang's predicted likelihood of the first residue together to determine the correct number of

missing residues.

The ANARCI numbering of antibody sequences gives an initial reasonable approximation of the number of residues missing from the N-terminus. However, because of possible indels and the variable length of CDR1, the ANARCI numbering alone is unable to determine the correct number of residues missing from the N-terminus. We observed that AbLang's predicted likelihood of the first residue in a sequence, was a good approximation of whether a sequence is the whole variable domain, see Supplementary Fig. B.2. We therefore used the likelihood of the first residue to identify if a sequence has been restored with the correct number of residues at its N-terminus.

We tested N-terminus lengths between eight residues shorter and up to two residues longer than the standard length given by ANARCI. This takes into account possible indels and a CDR1 region containing 5-12 residues. This process can be repeated and we found that this often improves the results, especially for heavy chains.

Fig. 5.6 compares the standard length given by ANARCI (green) with the ability of AbLang to restore the correct number of missing N-terminus residues, by either restoring once (blue) or twice (red). If the first 15 positions are missing, the ANARCI given length is correct for only one heavy chain sequence and 21.3% of the light chains, while restoring once with AbLang leads to the correct number of missing residues for 98.7% and 97.6% of the light and heavy chains, respectively. For improved performance, the restored sequences can go through the process again. This increases the restoration of the correct number of missing residues to 99.1% and 99.9% for light and heavy chains, respectively. ANARCI's inability to account for indels, such as the common deletion at position 10, can be seen in Fig. 5.6, where the ANARCI given length is highly inaccurate when nine or more residues are missing.

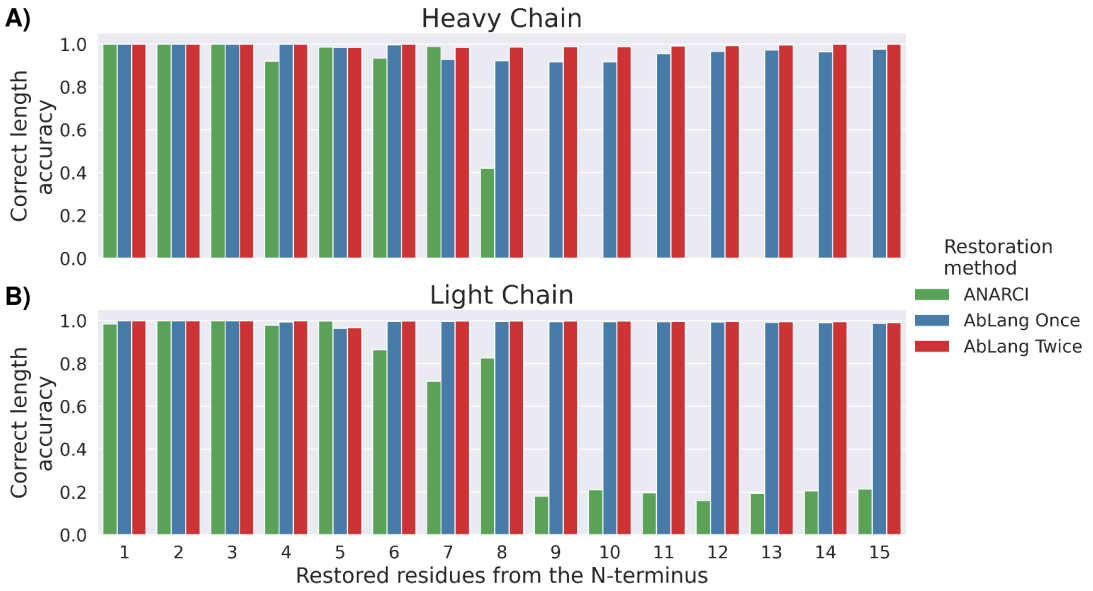

Figure 5.6: Comparison of the restoration of antibody sequences missing unknown numbers of residues at the N-terminus, using the standard length given by ANARCI (green) or the AbLang predicted likelihood of the first residue to determine the correct length. With AbLang, sequences were restored once (blue) or twice (red). The accuracy of selecting the correct number of residues at the N-terminus to restore for light (A) and heavy (B) sequences missing up to 30 residues of the N-terminus is shown.

5.5 Discussion

A language model specific for antibodies, should learn a deeper understanding of the semantics of antibodies than current general protein language models. In this work we present AbLang, a language model trained on a large dataset of antibody sequences from OAS.

AbLang was designed using a BERT-like architecture, unlike GPT-like antibody-specific language models such as IgLM [131], making it more suitable for antibody optimisation rather than *de novo* antibody generation. In comparison with other BERT-like antibody-specific language models, such as AntiBERTy [128], differences lie in training data preparation (e.g., extensive cleaning versus solely removing redundant sequences), training approach (e.g., using linear warm-up and varying batch sizes), model size, and minor architectural choices (e.g., using learnable or fixed positional embeddings).

AbLang can be used to derive three different sets of antibody representations, each with different possible use cases, as illustrated in Fig. 5.1. The res-codings and seq-encodings can be used for residue or sequence-specific representations or predictions. For instance, Yuan *et al.* used AbLang res-codings when training a model for antigen–antibody affinity prediction [178]. Similarly, AbLang's representations can also be used for antibody structure prediction, as demonstrated with other antibody-specific language models [132]. The last representation set, the likelihoods from AbHead, can be used to predict new mutations at each position in a given antibody sequence, as demonstrated by Hie *et al.* in their work on improving antibody affinity. For convenience, all three sets of representations are easily obtainable using the freely available AbLang python package at `https://github.com/oxpig/AbLang`.

To demonstrate that AbLang has learnt a useful sequence representation of antibodies, we show how AbLang's seq-codings contain knowledge of the germlines, originating cell type and number of mutations (see Fig. 5.2). However, these t-SNE visualizations are only indicative, and future work could explore these observations further.

To showcase AbLang's usefulness for antibody design, we explored its ability to restore missing residues in an antibody sequence. As shown in Fig. 5.3, for 80% of available antibody sequence data at least one residue is missing from the N-terminus. Sequences with missing residues are usually discarded, significantly diminishing available data. Accurate restoration of the N-terminus therefore allows the available data for further analysis to be more than doubled.

We demonstrate the use of AbLang to restore missing residues in antibody sequences, and show how AbLang performs on par or better than using IMGT germlines, but without the need to have knowledge of the germline. Further, we

describe a method for using AbLang to restore N-terminus regions with unknown arbitrary lengths.

The baseline IMGT germline method represents predicting the unmutated sequence. A better accuracy than this method therefore implies predictions which are not just the most often seen amino acid at a position and instead are specific to the input sequence. Further, we show how AbLang restores residues more accurately and faster than a current state-of-the-art protein language model ESM-1b, emphasising the benefits and potential of an antibody specific language model.

Overall the work shows the possibility of using protein language models for restoring residues in protein sequences, a crucial problem not only for antibody sequences but also for protein sequences in general. Though ESM-1b struggles with restoring longer end regions, it outperforms the IMGT germline baseline when restoring randomly placed residues in antibody sequences. ESM-1b might therefore be a useful tool for restoring a few missing residues in proteins, but less useful at restoring the ends of sequences and longer regions. The fact that ESM-1b struggles to restore residues at the N-terminus compared to single randomly distributed residues, could be because longer regions give rise to higher combinations of possible residues, and as ESM-1b does not have antibody specific context, it is unable to make accurate predictions.

In this work we give an example of how our antibody specific language model AbLang, can be used to create state-of-the-art solutions for antibody design. However, AbLang could be used for a wide range of other antibody discovery and design problems, which we hope by making it available, we and others can explore in future work.

Given that each antibody sequence predominantly comprises of germline residues, the data used to train antibody-specific models such as AbLang, inherently possess a considerable bias towards the germline. In the next chapter, we explore the implications of this bias and how it also affects general protein language models. We then focus on the prediction of non-germline residues, and subsequently train novel models specifically optimized for this task. Finally, we investigate and compare the cumulative probability of valid proposed mutations from clonotypes from current models and our new model, AbLang-2, hoping to provide a deeper understanding of how these models can be improved and effectively applied in the field of antibody design and discovery.

6

Addressing the antibody germline bias and its effect on language models for improved antibody design

Contents

6.1 Chapter abstract

The extreme variability of antibody sequences enables them to potentially bind to any pathogen. However, while there is an estimated 10^{16} to 10^{18} circulating unique antibody sequences in a human repertoire [21], this variability primarily arises from V(D)J recombination, mutations within the CDRs, or from a small

number of mutations away from the germline outside the CDRs. Consequently, a significant portion of the variable domain remains germline [179]. This affects the pre-training of antibody-specific language models like AbLang (see Chapter 5), where the sequence data introduces a prevailing bias towards germline residues. This poses a challenge, as mutations away from the germline are often more useful for generating more specific and potent binding to a target or improve antibody properties, and thus more desirable [179].

This chapter explores the implications of the germline bias, examining its impact on both general protein language models and antibody-specific language models, such as AbLang. To overcome this issue, we develop and train new antibody-specific language models optimized for predicting non-germline residues. Finally, we compare the cumulative probability of valid mutations from clonotypes from current models and our new model, AbLang-2. The following chapter is based on work in-progress, where I conceived the project, designed the models, wrote the code, trained the models and performed all tests.

6.2 Introduction

A lot of recent effort has been put into developing and training antibody-specific language models (LMs) to improve antibody discovery and design (see Introduction 1.5.4). Most LMs are pre-trained on the vast amounts of unlabelled antibody data in databases like OAS. The antibodies in these databases come from extremely diverse sources, with antibodies from patients with many different disease and vaccine states (see Chapter 2). However, antibody sequences are extremely germline-conserved with few mutations outside the CDR3 [179]. Moreover, although blood samples contain fewer affinity matured antibody producing B-cells than other tissues, BCR-seq are most commonly done on blood samples, because blood samples are less invasive to obtain compared to other samples (See Introduction 1.3.1.1). Databases like OAS are therefore most likely highly biased towards the

germline.

LMs are known to reproduce and even amplifying bias's in their training data [180]. For natural language LMs, efforts to reduce bias's have included pre-processing training data [180] or de-biasing with fine-tuning [181]. The germline-bias in antibody sequences can also be viewed as an imbalance problem. When predicting randomly selected masked residues, it is rarely a non-germline (NGL) residue which needs to be restored. The imbalance problem is well-known and many solutions have been proposed, like up or down-sampling [182] and focal loss [183]. Focal loss is a special loss function that down-weights the loss of well predicted labels. As rare labels, such as NGL residues, are usually poorly predicted, it results in an increased focus on these labels during training.

For antibodies, there are usually only a few non-germline mutations outside the CDR3 on affinity matured antibodies. However, there are often more than for naive B-cells, indicating a few mutations are required for a specific and high-affinity binding [179]. It is therefore important to understand if and how the germline-bias affects antibody-specific pre-trained LMs, especially their ability to suggest relevant NGL mutations. Correctly selecting relevant NGLs might result in the design and optimization of better therapeutic antibodies.

Here, we explore the germline-bias in antibody sequences, both from paired OAS and a set of therapeutic antibodies. We then investigate how the bias has affected BERT-like LMs, such as the general protein LM, ESM-2, and various antibody-specific LMs, ability to predict NGL residues. We then iteratively train and improve a new antibody-specific LM specifically for non-germline prediction. We then shown how our final model, AbLang-2, is able to more accurately suggest a diverse set of relevant mutations compared to previous models.

6.3 Methods

6.3.1 Dataset preparation

The training and test sets were derived from OAS. Antibody sequences were downloaded from the OAS database on Nov. 2022, yielding 2,072M heavy chains, 357M light chains and 1.57M paired antibodies. The sequences were then filtered by first removing duplicates. Then, sequences missing conserved cysteines or heavily fragmented (missing more than 16 residues from the N-terminus or 7 residues from the C-terminus) were removed. Unpaired sequences were additionally filtered for sequences only seen once. Finally, any amino acids other than the standard 20 were changed to X.

Redundancy was then reduced with clustering. Unpaired sequences were clustered first based on identical CDR3s and thereafter by 95% identity over the whole sequence using Linclust [148]. Paired sequences had first their heavy and light chain clustered individually as done for unpaired sequences. The paired sequences were then clustered by having the same heavy and light chain cluster. The longest sequence or sequence pair from each cluster was kept.

The paired antibodies were then randomly split into a train, validation and test set of 1.26M, 100k and 100k, respectively. The paired validation and test sets were clustered together with the reduced unpaired set by 95% identity over the whole sequence using Linclust [148]. Any unpaired sequences clustered with a heavy or light chain from the paired data, was removed.

This resulted in training sets with 27.5M heavy chains, 11.1M light chains and 1.26M paired antibodies, and validation and test set of 100k paired antibodies each.

The therapeutic sequences used in this study were sourced from Thera-SAbDab [59] (as of Feb. 2023) and were screened to discard therapeutics with either chain missing, resulting in 735 viable therapeutics.

6.3.1.1 Germline and non-germline residues

For OAS derived sequences, germline and NGL residues were determined with IgBLAST [38]. IgBLASTn uses the nucleotide sequences to predict each antibody's germlines, including non-templated regions within the CDR3, which was then used to label each residue. For the 735 therapeutic sequences, germlines were predicted with ANARCI [26] from the protein sequence, before being labelled as done for OAS derived sequences.

The standard, germline and NGL residue test sets, were generated from sequences in the validation set (see Methods 6.3.1). The standard test set represents how perplexity is usually measured, and is a random selection of 20,000 residues. For the germline test set, we sampled 20,000 known germline residues. For the NGL test set, we used all 475,000 NGL residues outside the CDR3 found in the validation set, and a random selection of 9,000 NGL residues within the CDR3.

6.3.1.2 Perplexity

Perplexity measures a model's uncertainty when predicting an amino acid at a given position and is commonly used for performance comparison. For BERT-like LMs, sequence perplexity can be derived by first computing the negative log-likelihood loss for each masked residue individually. The final perplexity is then the exponential of the mean of these losses [184]. The perplexity of a test set is then the average sequence perplexity. This is also how pseudo-perplexity, which they refer to as perplexity, is calculated by Lin *et al.* when evaluating ESM-2. In our work, instead of measuring perplexity based on every residue in a sequence, we only measure perplexity based on a subset of the residues within each sequence. Equation 6.1

shows how we define perplexity, where M is a set of residues within sequence x. This subset can be only NGL or germline residues, or all residues as done by Lin *et al.*.

$$Perplexity(x) = \exp(-\frac{1}{M}\sum_{i=1}^{M}\log p(x_i|x_{j\neq i})) \tag{6.1}$$

For consistency, the same residues are used to asses the perplexity for each model.

6.3.2 Architecture and training

A series of models were iteratively improved and trained using the training sets (see Methods 6.3.1). The models were implemented in PyTorch 2.0.1 and trained using the PyTorch-Lightning framework. The initial model (Ab-Unpaired) was based on the architecture of a 6-layered ESM-2 model with SwiGLU as its activation function, and trained on single chains from the paired training set. The model was optimized with an Adam optimizer. For stabilizing and enhancing training, we used a linear warm-up for 1k steps, a peak learning rate of 0.0004, a cosine-function rate decrease over 9k steps, and a weight decay of 0.01. An effective batch size of 8192 was used during the training, together with a dropout and layer normalization with a rate of 0.1 and an epsilon of $1e^{-12}$.

The model was then further improved over several iterations, with each new model being an expansion of the previous one:

- Ab-Paired: The input was modified to also handle paired antibodies, by separating true heavy-light chain pairs with a separator token. The model was then trained with unpaired heavy and light chains, and paired chains, from the paired training set.

- Ab-FL: Instead of the conventional cross-entropy loss function, we used focal loss. The purpose of this loss function, is to better address the challenge of imbalanced or sparse datasets (see Introduction).

- Ab-ModMask: The standard masking approach was modified to include two alternative masking methods; short 3-5 segment masking and singular large segment masking, both inspired by Tay *et al.* [185]. For each batch, a masking method (the two new masking methods and standard shotgun masking) is then selected uniformly. The proportion of masked residues was also changed to a dynamic value, selected uniformly between 10% and 40%.

- Ab-FT: The model was initially pre-trained exclusively on the unpaired sequences (see Methods 6.3.1) for 10,000 steps. This was followed by fine-tuning on paired sequences for an additional 1,000 steps and a peak learning rate of 0.0001.

- AbLang-2: The architecture was scaled up to 12 layers and an embedding size of 480. The model was then pre-trained on unpaired sequences for 200,000 steps and subsequent fine-tuned for 10,000 steps on paired sequences.

6.4 Results

6.4.1 Germline bias in antibody sequence data

NGL residues outside the CDR3 were found from paired sequences within the OAS database and subsequently visualized. A significant portion of antibodies (42%) originate from naive B-cells, followed by unsorted B-cells (39%), and memory B-cells (17%). The last 1% of antibodies are derived from other cells like plasma B-cells (see Fig. 6.1a).

When observing the distribution of NGL residues outside the CDR3 (see Fig. 6.1b), it is evident that the majority reside in the framework 3 region. This region also spans a greater length than both the CDR1 and CDR2.

Fig. 6.1c shows the distribution of NGL residues per sequence, for each cell source. As expected, the majority of antibodies from naive B-cells lack NGLs, while those from memory B-cells exhibits more NGLs, averaging $\sim$10 and $\sim$5.3 in the

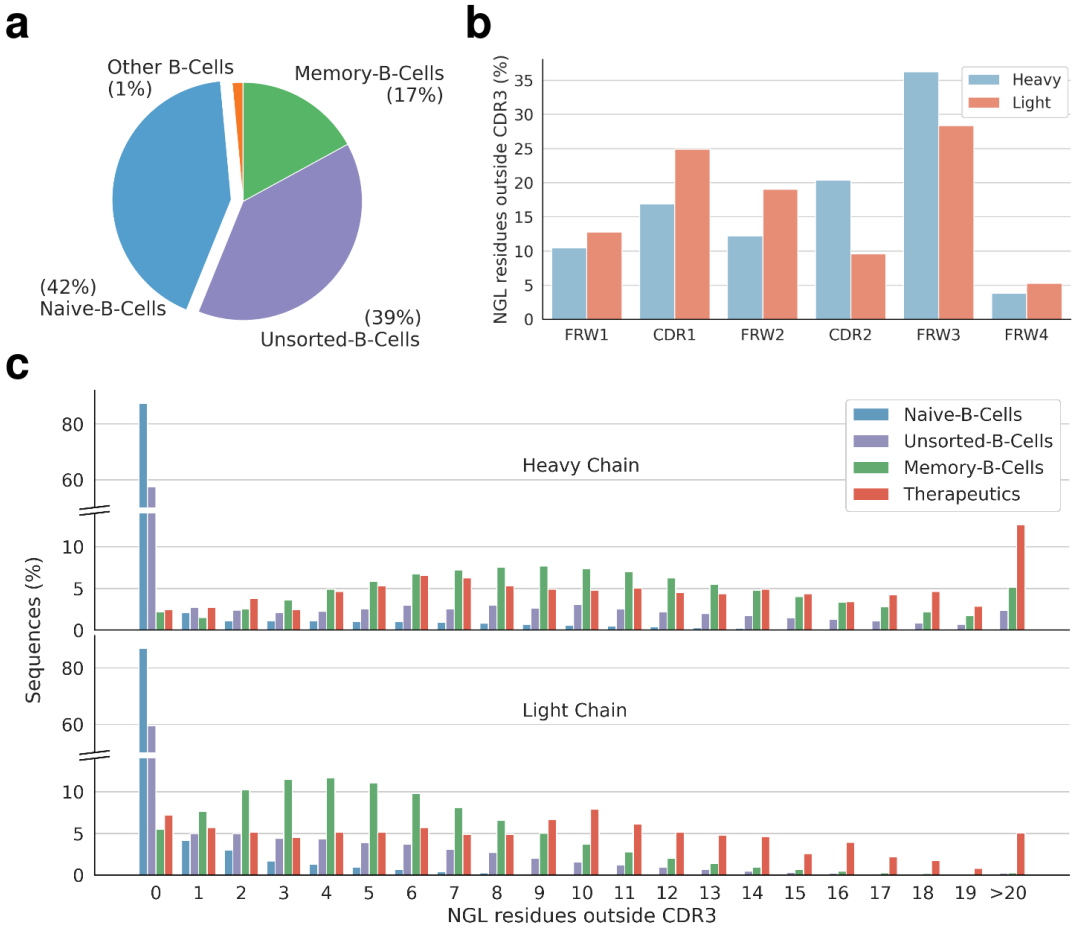

Figure 6.1: Overview of non-germline (NGL) residues from paired antibody sequences in the OAS database. **a**, Distribution of antibody origins, showing naive B-cells as the predominant source (42%), followed by unsorted B-cells (39%) and memory B-cells (17%). **b**, Distribution of NGL residues across different regions. **c**, Distribution of NGL residues per sequence by source. Naive B-cell derived antibodies predominantly lack NGLs, while memory B-cell derived antibodies display an average of ∼10 and ∼5.3 NGLs in the heavy and light chain, respectively. Therapeutic antibodies exhibit averages of ∼11.5 and ∼8.8 NGLs for the heavy and light chains. Supplementary Fig. C.1 provides an extended view of the distribution across both chains, with memory B-cell and therapeutic antibodies averaging ∼15.3 and ∼20.3 NGLs, respectively.

heavy and light chain, respectively. For comparison, a marginally higher count of NGLs was observed for antibody therapeutics (see Methods 6.3.1.1), averaging ∼11.5 and ∼8.8 in the heavy and light chain, respectively. Supplementary Fig. C.1 further shows the distribution across both chains. Here, memory B-cell derived antibodies averaged ∼15.3 NGLs, while therapeutic antibodies showed an average of ∼20.3 NGLs.

6.4.2 Germline bias in pre-trained language models

The impact of the germline bias in antibody sequences on pre-trained LMs was investigated for a set of general protein (ESM-2 [115]) and antibody-specific (Sapiens [84], AntiBERTy [128] and AbLang [129]) LMs. The ESM-2 models were trained on the UniRef50 [186] dataset, comprising approximately 60M protein sequences, of which a few hundreds are antibody sequences. The largest ESM-2 model have 3B parameters, however; in this work we use the 650M-parameter ESM-2 model. In contrast, the antibody-specific models are trained solely on unpaired antibody sequences from the OAS database.

The impact was first examined by investigating how often the germline is predicted for masked NGLs. For this we used the NGLs outside of the CDR3, derived as described in Methods 6.3.1.1, and predicted the masked residues with the four LMs. Fig. 6.2 shows the results for both the heavy and light chain, and with sequences grouped by the number of NGLs. The antibody-specific models of Sapiens, AntiBERTy, and AbLang-1, predicted the germline with frequencies of 87.6%, 86.7%, and 84.9%, respectively. ESM-2, despite its limited exposure to antibody sequences, still predicted the germline at a rate of 49.6%. While it remains unclear if models less germline-biased are better at predicting NGLs, it is clear that all the models tested preferentially suggest mutations towards the germline.

To investigate whether the models predicted NGLs more frequent for sequences further from the germline, we grouped the results by the number of NGLs per sequence. However, there does not appear to be a clear correlation (see Fig. 6.2).

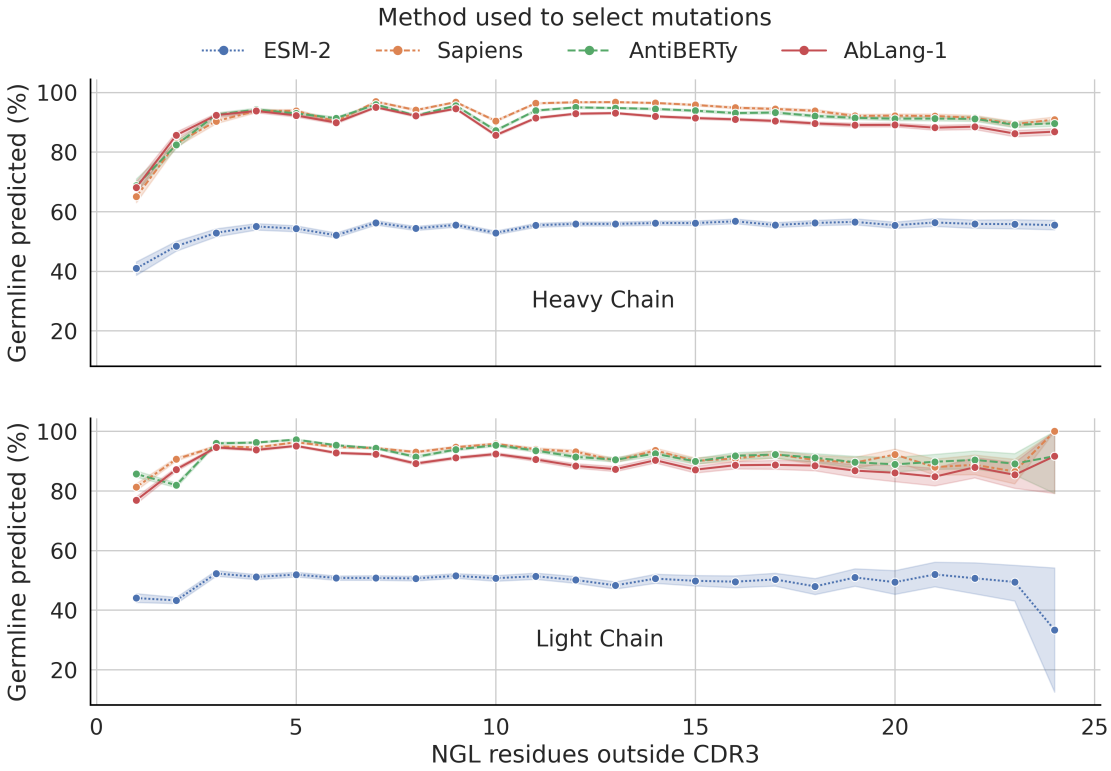

Figure 6.2: Germline prediction of masked non-germline (NGL) residues for four pre-trained language models (LMs). Results for the heavy and light chain are visualized separately, with predicted residues grouped by the number of NGLs outside the CDR3 in their sequence. The antibody-specific models of Sapiens [84], AntiBERTy [128], and AbLang-1 [129] predict the germline 87.6%, 86.7%, and 84.9% of the time, respectively. ESM-2 [115], trained with few antibody sequences, predicts the germline 49.6% of the time. This clearly shows a trend of each LM preferentially suggesting mutations to the germline.

To better understand what these models have learned, we evaluated and compared their perplexity when predicting masked residues on three different sets. A set of random residues, representing the standard approach for calculating perplexity, a set of germline residues, and a set of NGL residues (see Methods 6.3.1.1). We calculated the perplexity for each set for ESM-2, AntiBERTy, and AbLang-1 (see Fig. 6.3). We left out Sapiens, as their predictions are similar to

both AntiBERTy and AbLang-1. The perplexity metric spans from 1, denoting a perfect prediction, to positive infinity, representing zero probability for a correct prediction. A random prediction will result in a perplexity of 20, and predictions worse than random will result in values above 20.

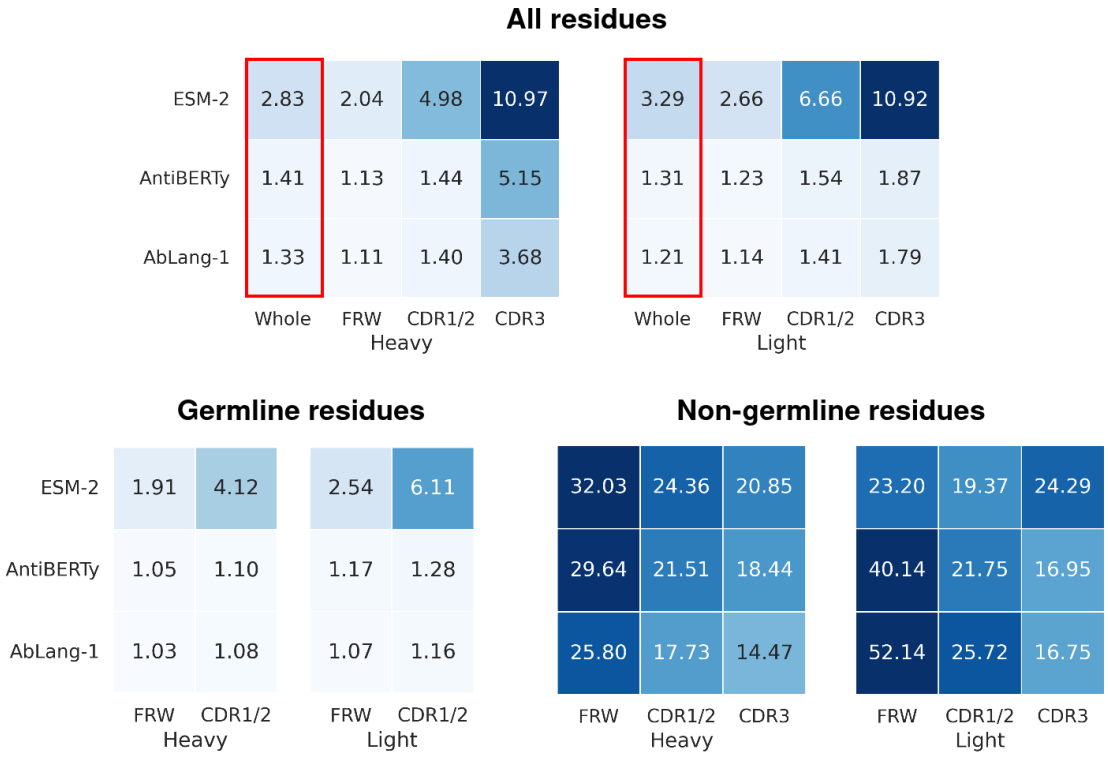

Figure 6.3: Perplexity comparison between the general protein language model (LM) ESM-2 [115], and the antibody-specific LMs AntiBERTy [128] and AbLang-1 [129]. Perplexity was calculated on a set of randomly selected residues, only germline residues, and only non-germline residues. Normally, perplexity is calculated for any residue across the whole sequence (red square), however; because of the short length of the difficult to predict CDRs and the germline bias, the actual performance is hidden.

Perplexity is normally calculated for all residues across the whole sequence or a random subset. With this standard approach, all models show a good performance (highlighted by the red square). However, when evaluating specific regions, the performance on the more variable CDRs, especially CDR3, is considerably worse. The CDRs are shorter in length, which hides some of their poor performance when using the standard approach for calculating perplexity.

When further splitting residues into known germline and NGL residues, it becomes clear that the standard perplexity is heavily dominated by accurately predicting the germline and not the NGLs. In fact, all the models perform poorly at predicting NGLs, while the antibody-specific models almost perfectly predict the germline. The germline bias also affects general protein language models, as seen with ESM-2's poor NGL prediction. Standard perplexity is not only skewed by the short CDR lengths, which hides the more difficult regions to predict, but also by the germline bias.

6.4.3 Reducing the germline bias

To limit the germline bias and improve NGL prediction, we trained several models while optimizing for NGL perplexity (see Methods 6.3.2). Starting from the architecture of a small ESM-2 model, each model introduced a new design choice. Fig. 6.4 shows each model and their incremental perplexity improvements. Like ESM-2, AntiBERTy, and AbLang-1, both Ab-Unpaired and Ab-Paired struggle with NGL prediction. Switching to focal loss significantly improves NGL predictions without compromising germline accuracy. Using a modified masking technique slightly improved NGL perplexity in the framework and CDR1/2, but also mildly reduced performance in the CDR3. Although pretraining on the large amounts of unpaired sequences (see Methods 6.3.1) before fine-tuning on paired sequences was expected to improve performance, it led to a small dip in perplexity. As a final effort to optimize performance, we scaled the model from 6 to 12 layers and extended its training duration. This resulted in the best performing model, AbLang-2.

6.4.4 Clonotype mutations

To better verify the suggested mutations, we examined positions in clonotypes with three or more NGLs. The clonotypes were created by grouping antibodies, from

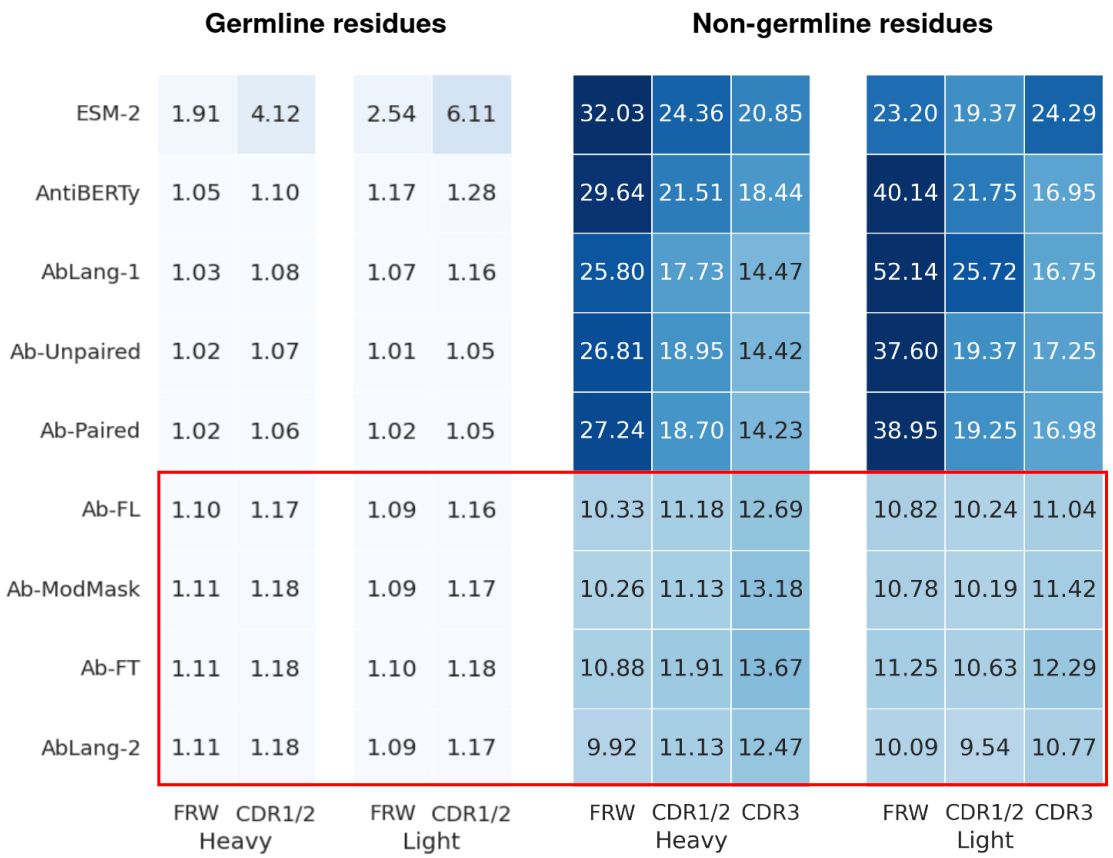

Figure 6.4: Perplexity comparison between the general protein language model (LM) ESM-2 [115], the antibody-specific LMs AntiBERTy [128] and AbLang-1 [129], and our new selection of antibody-specific LMs (see Methods 6.3.2). While most of the models are near perfect at predicting masked germline residues, predictions for non-germline (NGL) residues show significantly higher perplexities. For each modification, the best improvement for NGL prediction came from using focal loss (red square). Another notable improvement was scaling up the model, as seen by AbLang-2's performances compared to Ab-FT.

the validation set, based on identical source, V/J genes, and CDR3 length for both chains. This yielded 101 clonotypes, containing 226 and 60 sites with a minimum of three known NGLs in heavy and light chains, respectively. For each clonotype, a representative germline sequence was then generated by reverting NGLs outside of the CDR3 back to the germline for the sequence with the fewest NGLs.

For each site, the position was masked in the representative germline sequence and predicted using ESM-2, AntiBERTy, AbLang-1 and AbLang-2. The cumulative probability for known NGLs at the site was then compared across the models, see

Fig. 6.5a. For the heavy chain, both AntiBERTy and AbLang-1 have an average cumulative probability below 2%, in contrast to ESM-2's 20% and AbLang-2's 15%. Similarly, for the light chain, AntiBERTy and AbLang-1 have an average cumulative probability of 3% and 8%, respectively, while ESM-2 and AbLang-2 have 23% and 14%, respectively.

When the germline is included, see Fig. 6.5b, the cumulative probability for AntiBERTy, AbLang-1 and AbLang-2 hovers around 90-100%, underscoring how these models have a high probability of suggesting valid amino acids, i.e., mutations observed in the clonotypes and therefore known to be true. In contrast, ESM-2 has a cumulative probability of 52% and 66% for the heavy and light chain, implying that ESM-2 potentially suggests invalid amino acids with probabilities of 48% and 34%, respectively, as these mutations have not been observed in the clonotypes.

6.5 Discussion

Antibody sequences are predominantly composed of germline residues. Even those antibodies that are highly matured or have been optimized through extensive drug design campaigns, have only approximately 15 and 20 NGL residues, respectively, outside their CDR3s. Concurrently, over 93% of memory B-cells and 94% of Therapeutics have five or more NGL residues. This suggests that while an extensive number of mutations away from the germline is rare, a select few are common for effective antibodies. While being able to suggest these mutations is vital for the design of therapeutic antibodies, identifying these specific mutations remains a significant challenge.

With pre-trained LMs like ESM-2, Sapiens, AntiBERTy, and AbLang being used to suggest potentially property enhancing mutations [125], the effects of the germline bias present in antibody sequences becomes relevant, as it limits a models' ability to suggest relevant mutations that deviate from the germline. As seen in Fig. 6.2, these models predominantly suggest germline residues and, as seen in

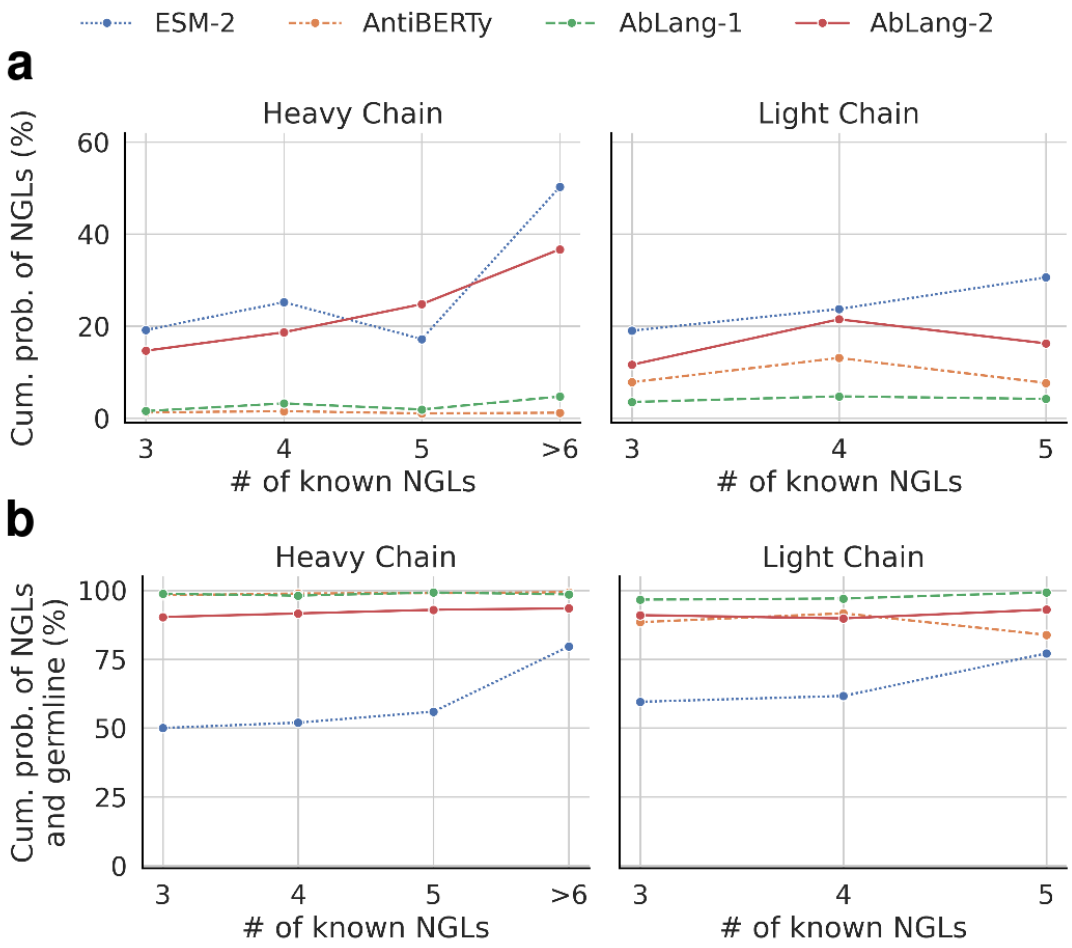

Figure 6.5: Comparison of cumulative probabilities of valid residues for the general protein language model (LM) ESM-2 [115] and the antibody-specific LMs AntiBERTy [128], AbLang-1 [129] and AbLang-2. Clonotypes were formed by grouping antibodies by source, V/J genes, and CDR3 length, yielding 226 and 60 sites in heavy and light chains, respectively, with at least three NGLs. **a**, Cumulative probabilities for known NGLs. AntiBERTy and AbLang-1 show <2% for the heavy chain, while ESM-2 and AbLang-2 display 20% and 15%. For the light chain, values are 3% and 8% for AntiBERTy and AbLang-1, and 23% and 14% for ESM-2 and AbLang-2. **b**, Cumulative probabilities for known NGLs and the germline. AntiBERTy, AbLang-1, and AbLang-2 demonstrate around 90-100% cumulative probabilities. ESM-2 presents 52% and 66% for the heavy and light chains, with the remaining probabilities suggesting amino acids different from the germline and at least three known NGLs.

Fig. 6.3, are poor predictors of NGL residues. Notably, even ESM-2, which was trained on fewer than 100 antibody sequences, is affected similarly. For reference, ESM-2 achieved a perplexity of 6.95 on a validation set of protein sequences [115]. The underperformance of these models in predicting NGL residues underscores the challenge of suggesting relevant changes to the germline.

"Pre-trained LMs like ESM-2, Sapiens, AntiBERTy, and AbLang are utilized for suggesting potentially property-enhancing mutations [125]. The impact of germline bias in antibody sequences becomes pertinent, constraining models' ability to propose relevant mutations deviating from the germline. As illustrated in Fig. 6.2, these models predominantly recommend germline residues and exhibit poor performance in predicting NGL residues, as depicted in Fig. 6.3. Notably, even ESM-2, trained on fewer than 100 sequences, is affected similarly. For reference, ESM-2 achieves a perplexity of 6.95 on a validation set of protein sequences [115]. The suboptimal performance of these models in predicting NGL residues highlights the challenge of suggesting pertinent alterations to the germline."

To try and design an antibody specific LM better capable of suggesting relevant NGL residues, we began with a small sized model with the same architecture as ESM-2 and iteratively improved it. First, the input was expanded to handle both unpaired and paired sequences. Then, focal loss was used during training, directing the model's attention to the less represented NGL residues, resulting in a much improved performance of predicting NGL residues, as seen in Fig. 6.4. Drawing inspiration from the training of other LMs, we then modified the masking approach. To broaden the exposure to more diverse data, we first pretrained the model on unpaired sequences before fine-tuning it on paired sequences. This allowed us to utilize the vast number of unpaired sequences, but still focus the model on handling paired data. For the final model, AbLang-2, we scaled up the size and training time.

A problem with perplexity is its presumption of a single correct prediction. However, for protein sequences, multiple mutations can be viable. While natural language has a similar problem, natural LMs typically select from tens of thousands of unique tokens [187]. In contrast, protein-specific LMs choose from just 20 amino acids and a few extra tokens, with up to half sometimes being valid predictions. To better assess the models' capacity to suggest valid mutations, we evaluated them on a dataset of same position mutations within clonotypes. AntiBERTy and AbLang-1 predicts a known valid amino acid with >90% accuracy, however; they only predict the germline (see Fig. 6.5). This limits their use for suggesting new relevant mutations. ESM-2 assigns higher probability to NGLs, however; it also tends to predict mutations other than the known valid mutations (34%-48% of the time), casting uncertainty over the validity of these suggestions. AbLang-2 exhibits a high cumulative probability for NGLs and simultaneously maintain a high probability for predicting known valid mutations. In other words, AbLang-2 suggests with high probability, a diverse set of valid amino acids.

It is worth highlighting that we are only aware of a subset of the valid mutations and we do not weigh the potential importance of certain mutations over others. Moreover, as we predict masked residues from a representative germline, some NGLs might not be viable within this sequence.

Correctly identifying every single residue as germline or NGL is not feasible with our current approach. The uncertainty in estimating the D germline complicates the identification of germline residues and the non-templated regions within the CDR3. For the CDR3, we therefore only measure the NGL perplexity, by using the estimated D germline to filter away likely germline residues. For a better separation of germline and non-germline residues, we focus on mutations outside of the CDR3. While estimating the V and J gene using IgBLAST is relatively reliable, it depends on a database of known V and J genes. The potential lack of certain alleles in this database, could result in the misclassification of some germline residues

as non-germline. Nonetheless, with most obvious germline residues filtered out, this approach still results in a much more challenging dataset than random selection.

In this work, we demonstrate the germline bias, a bias stemming from the low ratio of non-germline mutations in both naturally occurring antibodies as well as highly optimized therapeutic antibodies. We then show its effect on pre-trained LMs, especially how it affects their ability to suggest mutations away from the germline. In order to overcome this problem, we designed and pre-trained several antibody-specific LMs, with the final, AbLang-2, able to suggest a diverse set of valid mutations with high cumulative probability.

This work should facilitate the better design of therapeutic antibodies. For broader community engagement and research, we plan to make AbLang-2 easily accessible via freely available AbLang2 python package.

7 Conclusions and Future Work

7.1 Conclusions

The aim of this DPhil thesis was to collect, prepare, and investigate relevant antibody sequence data for computational methods and machine learning model training, as well as to explore the potential of BERT-like language models for improved therapeutic antibody design. This work thereby addresses the scientific challenges of utilising publicly available antibody data to derive novel insights and the application of state-of-the-art deep learning algorithms from the natural language processing field to advance computational antibody therapeutic development.

The advancements made in this thesis strengthens the foundation for future deep learning research in therapeutic antibody discovery and design, particularly through the contribution of key insights, new databases and computational tools. We updated and expanded OAS (Chapter 2), a valuable resource of billions of functional antibody sequences, enabling immune repertoire mining across various studies and the training of antibody-specific deep learning models. Despite OAS's utility, it lacks detailed antibody annotations, a gap we addressed by introducing PLAbDab (Chapter 4), a large curated collection of functionally

diverse, literature-annotated paired antibody sequences and structures, which can be searched for similar antibodies or used to generate antigen-specific libraries.

To enable efficient mining of databases like OAS or PLAbDab, we created KA-Search (Chapter 3), a rapid and exhaustive sequence identity search of known antibodies. Its capabilities were demonstrated by identifying antibodies in OAS with similar paratopes, thereby potentially discovering similar binders.

We also developed AbLang (Chapter 5), an antibody-specific LM. Trained on the extensive OAS, AbLang can learn the inherent patterns of functional antibody sequences, enabling the design of novel antibodies and the identification of beneficial mutations. We specifically demonstrated its utility in reconstructing fragmented antibodies.

The data used to train antibody-specific models, including AbLang, inherently possess a considerable bias towards the germline. We demonstrated this bias and illustrated its impact on pre-trained models, particularly their limitations in suggesting non-germline mutations. To overcome this, we introduced AbLang-2 (Chapter 6), a refined model capable of suggesting a diverse set of valid mutations with high cumulative probability.

The popularisation of ChatGPT has increased the concerns about the risks and ethics of AI models, and governments around the world are now discussing AI safety and regulation [188]. It is therefore relevant to also mention risks, like biases and malicious use, in relation to the models trained in this thesis. Although unlikely, the models could be used to design malicious antibodies. Also, as demonstrated in Chapter 6, pre-trained protein language models are prone to biases within their training data. An ethical concern can be how their predictions could be affected, and in turn benefit, certain demographics from which we have more data. For instance, suggested mutations could reduce or maintain the immunogenicity of an

antibody in well-represented demographics but increase it in underrepresented ones. This can however be solved by actively sequencing a more diverse population.

In conclusion, this thesis introduces new insights, databases and computational tools providing a foundation for future work using deep learning in the aid of therapeutic antibody design and discovery.

7.2 Future work

The design and discovery of antibodies using deep learning is still in its early stages, with advancements in every area like data generation, model design and antibody-specific insights. OAS already hosts ∼1.5 million paired antibody sequences, but with the continued BCR-seq advancements, available paired data will likely increase dramatically. The increasingly diverse data will permit the investigation of complex immunological questions, such as clonal diversity and antibody maturation pathways, which were previously challenging due to scarcity of paired data. Additionally, while current BCR-seq approaches focus on sequencing the variable domain, future techniques could extend to include the conserved domains. Although less variable, the domains are usually engineered in therapeutic antibodies, making their study critical for informed antibody design.

Current antibody-specific LMs are relatively modest in size when compared to natural language LMs. Scaling these models to similar sizes therefore has the potential of significantly improving their ability to capture the nuanced sequence patterns and inter-residue interactions, enabling more precise design and optimization of therapeutic antibodies. Further, the deep learning field is still relatively new, and new architectures are constantly being introduced. Transformer-based LMs might therefore be superseded by other architectures in the future.

Antibody-specific LMs are currently heavily inspired by natural language LMs. Introducing antibody insights, like the variable CDRs and the germline

bias, into the model design might help develop better models. While we only discuss sequence-based LMs in our work, LMs can be trained on other data, like structures, or as a multimodal model. Databases like PLAbDab can be used to incorporate other types of data, such as structural or functional annotations, into model training, potentially enabling more holistic insights into antibody function and design.

Future experimental advancements will also be crucial for the development of antibody-specific LMs. Currently, most human BCR-seq studies are performed on blood samples with low numbers of memory and plasma B-cells. New techniques enabling convenient sample extraction from tissues with more memory and plasma B-cells, like bone marrow, would expand the number of disease-related antibodies and enrich our antibody datasets. Additionally, current BCR-seq studies lack antibody-antigen information. To allow us to better model the antibody-antigen relationship, future experiments would need to provide this key information for each antibody. Techniques like LIBRA-seq [189], which uses DNA-barcoded antigens to recover antigen information when performing single-cell BCR-seq, have already allowed us to gain some antibody-antigen information for a set of antigens. In the future, we might be able to screen thousands of known antigens to reveal poly-specificity and identify antibodies with unknown antigens, knowledge which would help develop more informed models.

Appendices

A

Appendix Chapter 3

	The canonical alignment's unique positions
FRW1	1, 2, 3, 3A, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26
CDR1	27, 28, 29, 30, 31, 32, 32A, 32B, 33C, 33B, 33A, 33, 34, 35, 36, 37, 38
FRW2	39, 40, 40A, 41, 42, 43, 44, 44A, 45, 45A, 46, 46A, 47, 47A, 48, 48A, 48B, 49, 49A, 50, 51, 51A, 52, 53, 54, 55
CDR2	56, 57, 58, 59, 60, 60A, 60B, 60C, 60D, 61E, 61D, 61C, 61B, 61A, 61, 62, 63, 64, 65
FRW3	66, 67, 67A, 67B, 68, 68A, 68B, 69, 69A, 69B, 70, 71, 71A, 71B, 72, 73, 73A, 73B, 74, 75, 76, 77, 78, 79, 80, 80A, 81, 81A, 81B, 81C, 82, 82A, 83, 83A, 83B, 84, 85, 85A, 85B, 85C, 85D, 86, 86A, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 96A, 97, 98, 99, 100, 101, 102, 103, 104
CDR3	105, 106, 107, 108, 109, 110, 111, 111A, 111B, 111C, 111D, 111E, 111F, 111G, 111H, 111I, 111J, 111K, 111L, 112L, 112K, 112J, 112I, 112H, 112G, 112F, 112E, 112D, 112C, 112B, 112A, 112, 113, 114, 115, 116, 117
FRW4	118, 119, 119A, 120, 121, 122, 123, 124, 125, 126, 127, 128

Table A.1: Overview of the 200 unique positions in our canonical alignment. The positions are based on IMGT numbering of the variable domain [4, 143], however, instead of representing CDR3 gaps with numbers (i.e. 112.1) we use letters (i.e. 112A). We choose all 196 unique positions seen in at least 40.000 different sequences in OAS, as of May 2022, and four additional unique positions seen in therapeutics from Thera-SAbDab [59]. The four additional positions are 3A, 51A, 85C and 85D.

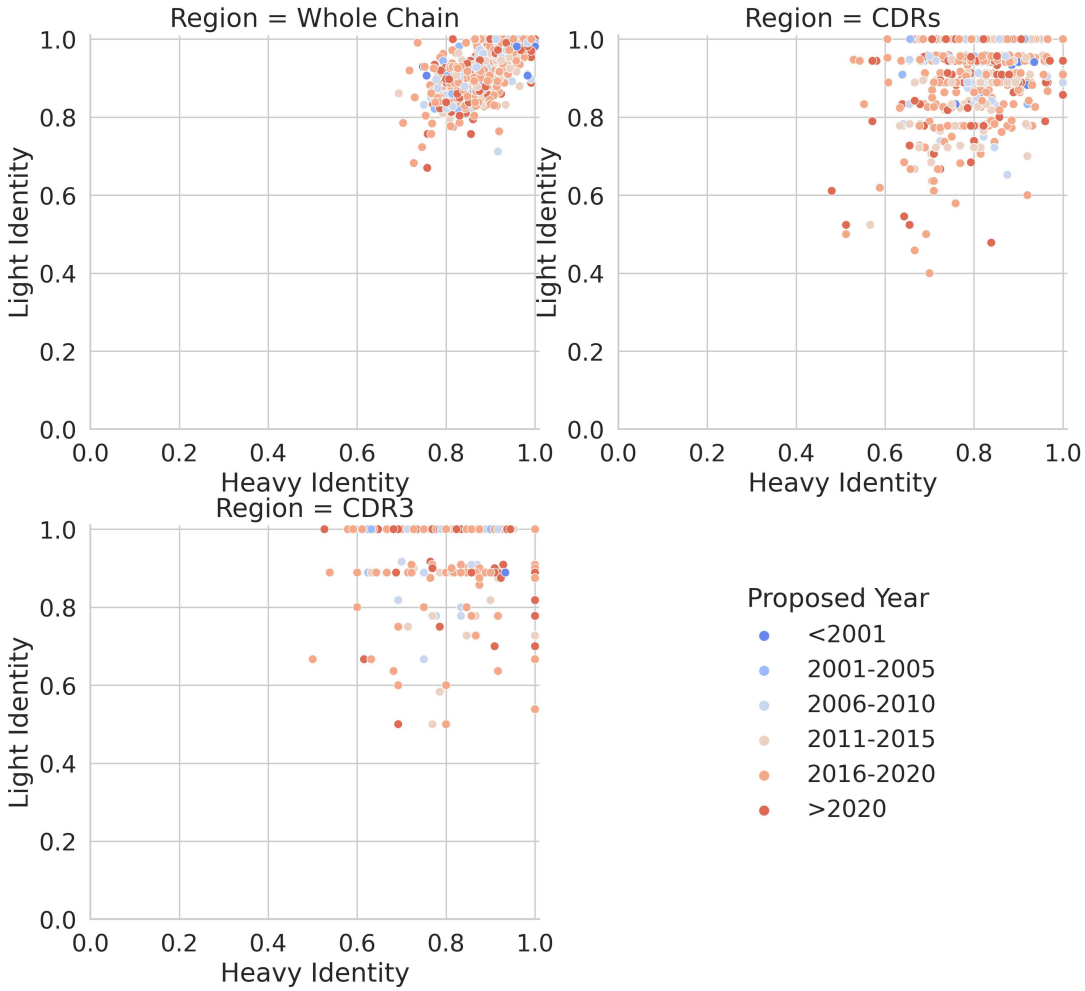

Figure A.1: KA-Search was used to find the closest matches in OAS to 804 therapeutics extracted in August 2022 from Thera-SAbDab [59]. Closest matches was found across the whole variable domain, the three CDRs and the CDR3. Each point is colored by the year they were proposed.

B

Appendix Chapter 5

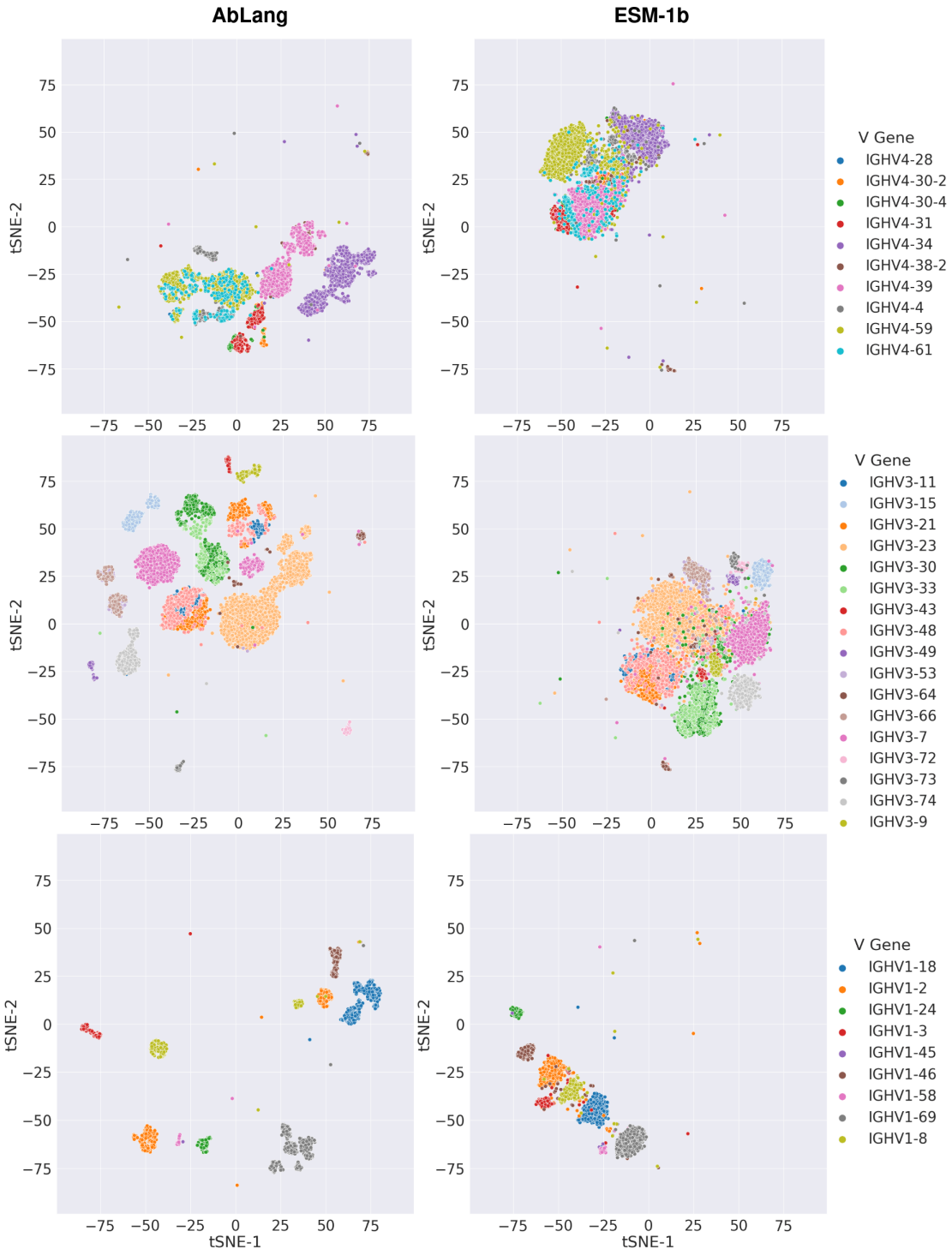

Figure B.1: Comparison of AbLang and ESM-1b representations at clustering sequences based on their V-genes. The figure compares the three most common heavy chain V-gene families in our dataset, IGHV1, IGHV3 and IGHV4.

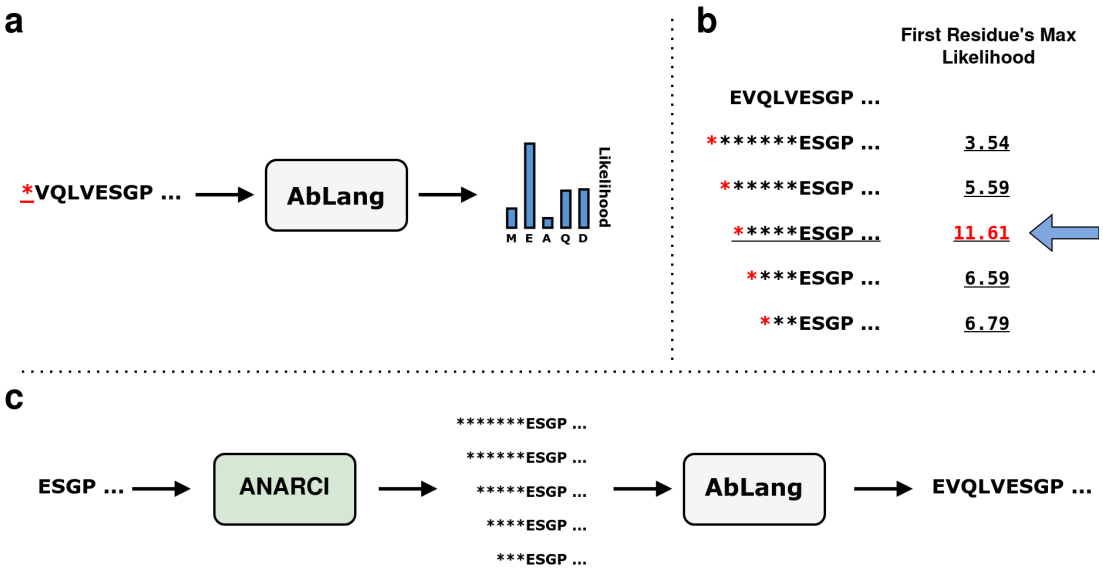

Figure B.2: Approach for restoring antibody sequences with an unknown number of missing residues. **a**, mask tokens are added at the termini of a fragmented antibody sequence, with their likelihoods subsequently predicted with AbLang. **b**, for the correct number of missing residues, the first mask token yields the highest likelihood. **c**, to restore an unknown length of missing residues, a sequence is first numbered using ANARCI to estimate the potential number of missing residues. From the estimate, sequences with between eight masks shorter and up to two masks longer are predicted with AbLang, and the sequence with the highest likelihood for the first residue is selected and returned.

C

Appendix Chapter 6

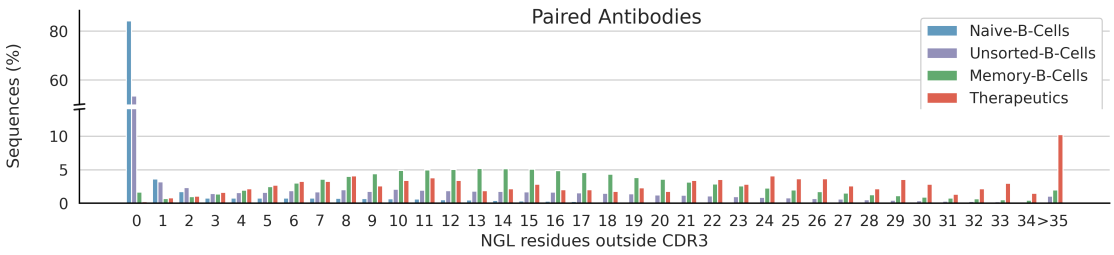

Figure C.1: Distribution of NGL residues per VH-VL domain by source. Naive B-cell derived antibodies predominantly lack NGLs, while memory B-cell derived antibodies display an average of ∼15.3. Therapeutic antibodies exhibit an average of ∼20.3 NGLs.

References

1. Lu, L. L., Suscovich, T. J., Fortune, S. M. & Alter, G. Beyond binding: antibody effector functions in infectious diseases. *Nature Reviews Immunology* **18,** 46–61 (2018).
2. Marks, C. & Deane, C. M. *How repertoire data are changing antibody science* 2020.
3. Norman, R. A., Ambrosetti, F., Bonvin, A. M. J. J., *et al.* Computational approaches to therapeutic antibody design: established methods and emerging trends. *Briefings in Bioinformatics* **21,** 1549–1567 (2019).
4. Lefranc, M.-P., Pommié, C., Ruiz, M., *et al.* IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Developmental and comparative immunology* **27,** 55–77 (2003).
5. Lefranc, M.-P., Pommié, C., Kaas, Q., *et al.* IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Developmental and comparative immunology* **29,** 185–203 (2005).
6. Janeway, C. *Immunobiology 5 : the immune system in health and disease* English (Garland Pub., New York, 2001).
7. Chiu, M. L., Goulet, D. R., Teplyakov, A. & Gilliland, G. L. Antibody Structure and Function: The Basis for Engineering Therapeutics. *Antibodies* **8,** 55 (2019).
8. Schroeder Jr., H. W. & Cavacini, L. Structure and function of immunoglobulins. *Journal of Allergy and Clinical Immunology* **125,** S41–S52 (2010).
9. Jacofsky, D., Jacofsky, E. M. & Jacofsky, M. Understanding Antibody Testing for COVID-19. *The Journal of arthroplasty* **35,** S74–S81 (2020).
10. Pang, N. Y.-L., Pang, A. S.-R., Chow, V. T. & Wang, D.-Y. Understanding neutralising antibodies against SARS-CoV-2 and their implications in clinical practice. *Military Medical Research* **8,** 47 (2021).
11. Sedova, E. S., Scherbinin, D. N., Lysenko, A. A., *et al.* Non-neutralizing Antibodies Directed at Conservative Influenza Antigens. *Acta naturae* **11,** 22–32 (2019).
12. Regep, C., Georges, G., Shi, J., Popovic, B. & Deane, C. M. The H3 loop of antibodies shows unique structural characteristics. *Proteins* **85,** 1311–1318 (2017).
13. Gordon, G. L., Capel, H. L., Guloglu, B., *et al.* A comparison of the binding sites of antibodies and single-domain antibodies. *Frontiers in immunology* (2023).
14. Lefranc, M.-P. & Lefranc, G. *The Immunoglobulin Factsbook* (Academic Press, 2014).
15. Melchers, F. Checkpoints that control B cell development. *The Journal of Clinical Investigation* **125,** 2203–2210 (2015).

16. Funck, T., Barnkob, M. B., Holm, N., *et al.* Nucleotide Composition of Human Ig Nontemplated Regions Depends on Trimming of the Flanking Gene Segments, and Terminal Deoxynucleotidyl Transferase Favors Adding Cytosine, Not Guanosine, in Most VDJ Rearrangements. *Journal of immunology (Baltimore, Md. : 1950)* **201,** 1765–1774 (2018).

17. Victora, G. D. & Nussenzweig, M. C. Germinal Centers. *Annual Review of Immunology* **40,** 413–442 (2022).

18. Rolink, A. G., Schaniel, C., Andersson, J. & Melchers, F. Selection events operating at various stages in B cell development. *Current Opinion in Immunology* **13,** 202–207 (2001).

19. Willis, J. R., Briney, B. S., DeLuca, S. L., Crowe, J. E. J. & Meiler, J. Human germline antibody gene segments encode polyspecific antibodies. *PLoS computational biology* **9,** e1003045 (2013).

20. Khodadadi, L., Cheng, Q., Radbruch, A. & Hiepe, F. The Maintenance of Memory Plasma Cells. *Frontiers in Immunology* **10,** 1664–3224 (2019).

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

22. Camacho, C., Coulouris, G., Avagyan, V., *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).

23. Dondelinger, M., Filée, P., Sauvage, E., *et al.* Understanding the Significance and Implications of Antibody Numbering and Antigen-Binding Surface/Residue Definition. *Frontiers in Immunology* **9** (2018).

24. Chothia, C. & Lesk, A. M. Canonical structures for the hypervariable regions of immunoglobulins. eng. *Journal of molecular biology* **196,** 901–917 (1987).

25. North, B., Lehmann, A. & Dunbrack Jr, R. L. A new clustering of antibody CDR loop conformations. *Journal of molecular biology* **406,** 228–256 (2011).

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

27. Chaudhary, N. & Wesemann, D. R. Analyzing Immunoglobulin Repertoires. *Frontiers in Immunology* **9,** 462 (2018).

28. Kim, D. & Park, D. Deep sequencing of B cell receptor repertoire. *BMB reports* **52,** 540–547 (2019).

29. Pantazes, R. J., Reifert, J., Bozekowski, J., *et al.* Identification of disease-specific motifs in the antibody specificity repertoire via next-generation sequencing. *Scientific Reports* **6,** 30312 (2016).

30. Georgiou, G., Ippolito, G. C., Beausang, J., *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnology* **32,** 158–168 (2014).

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

32. Li, X., Duan, X., Yang, K., *et al.* Comparative Analysis of Immune Repertoires between Bactrian Camel's Conventional and Heavy-Chain Antibodies. *PLOS ONE* **11,** e0161801 (2016).

33. Corcoran, M. M., Phad, G. E., Bernat, N. V., *et al.* Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nature Communications* **7,** 13642 (2016).

34. Cui, A., Di Niro, R., Vander Heiden, J. A., *et al.* A Model of Somatic Hypermutation Targeting in Mice Based on High-Throughput Ig Sequencing Data. *The Journal of Immunology* **197,** 3566–3574 (2016).

35. Johnson, E. L., Doria-Rose, N. A., Gorman, J., *et al.* Sequencing HIV-neutralizing antibody exons and introns reveals detailed aspects of lineage maturation. *Nature Communications* **9,** 4136 (2018).

36. Bernardes, J. P., Mishra, N., Tran, F., *et al.* Longitudinal Multi-omics Analyses Identify Responses of Megakaryocytes, Erythroid Cells, and Plasmablasts as Hallmarks of Severe COVID-19. *Immunity* **53,** 1296–1314 (2020).

37. Soto, C., Bombardi, R. G., Branchizio, A., *et al.* High frequency of shared clonotypes in human B cell receptor repertoires. *Nature* **566,** 398–402 (2019).

38. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic acids research* **41,** 34–40 (2013).

39. Jaffe, D. B., Shahi, P., Adams, B. A., *et al.* Functional antibodies exhibit light chain coherence. *Nature* **611,** 352–357 (2022).

40. Wang, B., Kluwe, C. A., Lungu, O. I., *et al.* Facile Discovery of a Diverse Panel of Anti-Ebola Virus Antibodies by Immune Repertoire Mining. *Scientific Reports* **5,** 13926 (2015).

41. Tian, X., Li, C., Wu, Y. & Ying, T. Deep Mining of Human Antibody Repertoires: Concepts, Methodologies, and Applications. *Small Methods* **4,** 2000451 (2020).

42. Hsiao, Y.-C., Shang, Y., DiCara, D. M., *et al.* Immune repertoire mining for rapid affinity optimization of mouse monoclonal antibodies. *mAbs* **11,** 735–746 (2019).

43. Richardson, E., Galson, J. D., Kellam, P., *et al.* A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-pertussis toxoid antibodies. *mAbs* **13,** 1869406 (2021).

44. Robinson, S. A., Raybould, M. I. J., Schneider, C., *et al.* Epitope profiling using computational structural modelling demonstrated on coronavirus-binding antibodies. *PLOS Computational Biology* **17,** 1–20 (2021).

45. Ota, M., Duong, B. H., Torkamani, A., *et al.* Regulation of the B Cell Receptor Repertoire and Self-Reactivity by BAFF. *The Journal of Immunology* **185,** 4128–4136 (2010).

46. Wu, X., Zhou, T., Zhu, J., *et al.* Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* **333,** 1593–1602 (2011).

47. Greiff, V., Menzel, U., Miho, E., *et al.* Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. *Cell Reports* **19,** 1467–1478 (2017).

48. Devulapally, P. R., Bürger, J., Mielke, T., *et al.* Simple paired heavy- and light-chain antibody repertoire sequencing using endoplasmic reticulum microsomes. *Genome Medicine* **10,** 34 (2018).

49. Goldstein, L. D., Chen, Y.-J. J., Wu, J., *et al.* Massively parallel single-cell B-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. *Communications Biology* **2,** 304 (2019).

50. Rubelt, F., Busse, C. E., Bukhari, S. A. C., *et al.* Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nature immunology* **18,** 1274–1278 (2017).

51. López-Santibáñez-Jácome, L., Avendaño-Vázquez, S. E. & Flores-Jasso, C. F. The Pipeline Repertoire for Ig-Seq Analysis. *Frontiers in Immunology* **10,** 899 (2019).

52. Kovaltsuk, A., Leem, J., Kelm, S., *et al.* Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. *The Journal of Immunology* **201,** 2502–2509 (2018).

53. AdaptiveBiotechnologies. *immuneACCESS*

54. Zhang, W., Wang, L., Liu, K., *et al.* PIRD: Pan Immune Repertoire Database. *Bioinformatics* **36,** 897–903 (2019).

55. Zhang, Y., Chen, T., Zeng, H., *et al.* RAPID: A Rep-Seq Dataset Analysis Platform With an Integrated Antibody Database. *Frontiers in Immunology* **12,** 3228 (2021).

56. Christley, S., Aguiar, A., Blanck, G., *et al.* The ADC API: A Web API for the Programmatic Query of the AIRR Data Commons. *Frontiers in Big Data* **3,** 22 (2020).

57. Espejo, A. P., Akgun, Y., Al Mana, A. F., *et al.* Review of current advances in serologic testing for COVID-19. *Am J Clin Pathol* **154,** 293–304 (2020).

58. Pollard, A. J. & Bijker, E. M. A guide to vaccinology: from basic principles to new developments. *Nat Rev Immunol* **21,** 83–100 (2021).

59. Raybould, M. I. J., Marks, C., Lewis, A. P., *et al.* Thera-SAbDab: the Therapeutic Structural Antibody Database. *Nucleic Acids Research* **48,** D383–D388 (2020).

60. Lu, R.-M., Hwang, Y.-C., Liu, I.-J., *et al.* Development of therapeutic antibodies for the treatment of diseases. *Journal of biomedical science* **27,** 1–30 (2020).

61. Kelley, B. Developing therapeutic monoclonal antibodies at pandemic pace. *Nature Biotechnology* **38,** 540–545 (2020).

62. Beeg, M., Nobili, A., Orsini, B., *et al.* A Surface Plasmon Resonance-based assay to measure serum concentrations of therapeutic antibodies and anti-drug antibodies. *Scientific Reports* **9,** 2064 (2019).

63. Aykul, S. & Martinez-Hackert, E. Determination of half-maximal inhibitory concentration using biosensor-based protein interaction analysis. *Analytical biochemistry* **508,** 97–103 (2016).

64. Dam, T. K., Torres, M., Brewer, C. F. & Casadevall, A. Isothermal titration calorimetry reveals differential binding thermodynamics of variable region-identical antibodies differing in constant region for a univalent ligand. *The Journal of biological chemistry* **283,** 31366–31370 (2008).

65. Dzimianski, J. V., Lorig-Roach, N., O'Rourke, S. M., *et al.* Rapid and sensitive detection of SARS-CoV-2 antibodies by biolayer interferometry. *Scientific Reports* **10,** 21738 (2020).

66. Hanning, K. R., Minot, M., Warrender, A. K., Kelton, W. & Reddy, S. T. Deep mutational scanning for therapeutic antibody engineering. *Trends in pharmacological sciences* **43,** 123–135 (2022).

67. Chen, W. C. & Murawsky, C. M. Strategies for Generating Diverse Antibody Repertoires Using Transgenic Animals Expressing Human Antibodies. *Frontiers in Immunology* **9** (2018).

68. Alfaleh, M. A., Alsaab, H. O., Mahmoud, A. B., *et al.* Phage Display Derived Monoclonal Antibodies: From Bench to Bedside. *Frontiers in Immunology* **11** (2020).

69. Shan, S., Luo, S., Yang, Z., *et al.* Deep learning guided optimization of human antibody against SARS-CoV-2 variants with broad neutralization. *Proceedings of the National Academy of Sciences* **119,** e2122954119 (2022).

70. Zost, S. J., Gilchuk, P., Chen, R. E., *et al.* Rapid isolation and profiling of a diverse panel of human monoclonal antibodies targeting the SARS-CoV-2 spike protein. *Nature Medicine* **26,** 1422–1427 (2020).

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

72. Jakobovits, A., Amado, R. G., Yang, X., Roskos, L. & Schwab, G. From XenoMouse technology to panitumumab, the first fully human antibody product from transgenic mice. *Nature Biotechnology* **25,** 1134–1143 (2007).

73. Shrock, E. L., Timms, R. T., Kula, T., *et al.* Germline-encoded amino acid–binding motifs drive immunodominant public antibody responses. *Science* **380,** eadc9498 (2023).

74. Liu, B. BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Briefings in bioinformatics* **20,** 1280–1294 (2019).

75. Hummer, A. M., Schneider, C., Chinery, L. & Deane, C. M. Investigating the Volume and Diversity of Data Needed for Generalizable Antibody-Antigen DDG Prediction. *bioRxiv,* 2023.05.17.541222 (2023).

76. Makowski, E. K., Wang, T., Zupancic, J. M., *et al.* Optimization of therapeutic antibodies for reduced self-association and non-specific binding via interpretable machine learning. *Nature Biomedical Engineering* (2023).

77. Liberis, E., Velickovic, P., Sormanni, P., Vendruscolo, M. & Lio, P. Parapred: antibody paratope prediction using convolutional and recurrent neural networks. *Bioinformatics* **34,** 2944–2950 (2018).

78. Ambrosetti, F., Olsen, T. H., Olimpieri, P. P., *et al.* proABC-2: PRediction of AntiBody contacts v2 and its application to information-driven docking. *Bioinformatics* **36,** 5107–5108 (2020).

79. Chinery, L., Wahome, N., Moal, I. & Deane, C. M. Paragraph - antibody paratope prediction using graph neural networks with minimal feature vectors. *Bioinformatics* (2022).

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

81. Van der Kant, R., Karow-Zwick, A. R., Van Durme, J., *et al.* Prediction and Reduction of the Aggregation of Monoclonal Antibodies. *Journal of molecular biology* **429,** 1244–1261 (2017).

82. Baker, M. P., Reynolds, H. M., Lumicisi, B. & Bryson, C. J. Immunogenicity of protein therapeutics: The key causes, consequences and challenges. *Self/nonself* **1,** 314–322 (2010).

83. Kaluza, B., Betzl, G., Shao, H., Diamantstein, T. & Weidle, U. H. A general method for chimerization of monoclonal antibodies by inverse polymerase chain reaction which conserves authentic N-terminal sequences. *Gene* **122,** 321–328 (1992).

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

85. Li, W., Prabakaran, P., Chen, W., *et al.* Antibody Aggregation: Insights from Sequence and Structure. *Antibodies* **5** (2016).

86. Jain, T., Sun, T., Durand, S., *et al.* Biophysical properties of the clinical-stage antibody landscape. *Proceedings of the National Academy of Sciences* **114,** 944–949 (2017).

87. Raybould, M. I., Marks, C., Krawczyk, K., *et al.* Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences of the United States of America* **116,** 4025–4030 (2019).

88. Lu, X., Nobrega, R. P., Lynaugh, H., *et al.* Deamidation and isomerization liability analysis of 131 clinical-stage antibodies. *mAbs* **11,** 45–57 (2019).

89. Wijesuriya, S. D., Pongo, E., Tomic, M., *et al.* Antibody engineering to improve manufacturability. *Protein Expression and Purification* **149,** 75–83 (2018).

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

91. Vaswani, A., Shazeer, N., Parmar, N., *et al.* Attention is All you Need. *ArXiv* **abs/1706.0** (2017).

92. Lin, T., Wang, Y., Liu, X. & Qiu, X. A survey of transformers. *AI Open* **3,** 111–132 (2022).

93. Couronné, R., Probst, P. & Boulesteix, A.-L. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics* **19,** 270 (2018).

94. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323,** 533–536 (1986).

95. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *CoRR* **abs/1512.0** (2015).

96. Ba, J. L., Kiros, J. R. & Hinton, G. E. *Layer Normalization* 2016.

97. Toraman, C., Yilmaz, E. H., Sahinuc, F. & Ozcelik, O. Impact of Tokenization on Language Models: An Analysis for Turkish. *ACM Transactions on Asian and Low-Resource Language Information Processing* **22,** 1–21 (2023).

98. Sennrich, R., Haddow, B. & Birch, A. Neural Machine Translation of Rare Words with Subword Units. *CoRR* **abs/1508.0** (2015).

99. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* **abs/1810.0** (2018).

100. Shaw, P., Uszkoreit, J. & Vaswani, A. Self-Attention with Relative Position Representations. *CoRR* **abs/1803.0** (2018).

101. Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., *et al.* An Improved Relative Self-Attention Mechanism for Transformer with Application to Music Generation. *CoRR* **abs/1809.0** (2018).

102. Su, J., Lu, Y., Pan, S., Wen, B. & Liu, Y. RoFormer: Enhanced Transformer with Rotary Position Embedding. *CoRR* **abs/2104.0** (2021).

103. Biderman, S., Black, S., Foster, C., *et al. Rotary Embeddings: A Relative Revolution*

104. Peters, M. E., Neumann, M., Iyyer, M., *et al. Deep Contextualized Word Representations* in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2018), 2227–2237.

105. Brown, T. B., Mann, B., Ryder, N., *et al.* Language Models are Few-Shot Learners. *CoRR* **abs/2005.1** (2020).

106. Trinh, T. H., Dai, A. M., Luong, T. & Le, Q. V. Learning Longer-term Dependencies in RNNs with Auxiliary Losses. *CoRR* **abs/1803.0** (2018).

107. Liu, Y., Ott, M., Goyal, N., *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* **abs/1907.1** (2019).

108. Lewis, M., Liu, Y., Goyal, N., *et al.* BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *CoRR* **abs/1910.1** (2019).

109. Raffel, C., Shazeer, N., Roberts, A., *et al.* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *CoRR* **abs/1910.1** (2019).

110. Radford, A., Wu, J., Child, R., *et al. Language Models are Unsupervised Multitask Learners* in (2019).

111. Ziegler, D. M., Stiennon, N., Wu, J., *et al.* Fine-Tuning Language Models from Human Preferences. *CoRR* **abs/1909.0** (2019).

112. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods* **16,** 1315–1322 (2019).

113. Elnaggar, A., Heinzinger, M., Dallago, C., *et al.* ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 1 (2021).

114. Rives, A., Meier, J., Sercu, T., *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118** (2021).

115. Lin, Z., Akin, H., Rao, R., *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379,** 1123–1130 (2023).

116. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications* **13,** 4348 (2022).

117. Nijkamp, E., Ruffolo, J., Weinstein, E. N., Naik, N. & Madani, A. *ProGen2: Exploring the Boundaries of Protein Language Models* 2022.

118. Clifford, J. N., Høie, M. H., Deleuran, S., *et al.* BepiPred-3.0: Improved B-cell epitope prediction using protein language models. *Protein Science* **31,** e4497 (2022).

119. Høie, M. H., Kiehl, E. N., Petersen, B., *et al.* NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Research* **50,** W510–W515 (2022).

120. Kroll, A., Ranjan, S., Engqvist, M. K. M. & Lercher, M. J. A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nature Communications* **14,** 2787 (2023).

121. Teufel, F., Almagro Armenteros, J. J., Johansen, A. R., *et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology* **40,** 1023–1025 (2022).

122. Littmann, M., Bordin, N., Heinzinger, M., *et al.* Clustering FunFams using sequence embeddings improves EC purity. *Bioinformatics* **37,** 3449–3455 (2021).

123. Kilinc, M., Jia, K. & Jernigan, R. L. Improved global protein homolog detection with major gains in function identification. *Proceedings of the National Academy of Sciences* **120,** e2211823120 (2023).

124. Hie, B. L., Yang, K. K. & Kim, P. S. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Systems* **13,** 274–285 (2022).

125. Hie, B. L., Shanker, V. R., Xu, D., *et al.* Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology* (2023).

126. Wang, D., Ye, F. & Zhou, H. *On Pre-trained Language Models for Antibody* 2023.

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

128. Ruffolo, J. A., Gray, J. J. & Sulam, J. *Deciphering antibody affinity maturation with language models and weakly supervised learning* 2021.

129. Olsen, T. H., Moal, I. H. & Deane, C. M. AbLang: an antibody language model for completing antibody sequences. *Bioinformatics Advances* **2,** vbac046 (2022).

130. Jing, H., Gao, Z., Xu, S., *et al.* Accurate Prediction of Antibody Function and Structure Using Bio-Inspired Antibody Language Model. *bioRxiv,* 2008–2023 (2023).

131. Shuai, R. W., Ruffolo, J. A. & Gray, J. J. Generative language modeling for antibody design. *bioRxiv,* 2021.12.13.472419 (2022).

132. Ruffolo, J. A., Chu, L.-S., Mahajan, S. P. & Gray, J. J. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature Communications* **14,** 2389 (2023).

133. Kaplon, H. & Reichert, J. Antibodies to watch in 2021. *mAbs* **13,** 1860476 (2021).

134. Coordinators, N. R. Database resources of the National Center for Biotechnology Information. *Nucleic acids research* **44,** D7–D19 (2016).

135. Coordinators, N. R. *The SRA Toolkit* 2021.

136. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27,** 2957–2963 (2011).

137. Hannonlab. *FASTX-Toolkit* 2014.

138. Zheng, G. X. Y., Terry, J. M., Belgrader, P., *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8,** 14049 (2017).

139. Giudicelli, V., Brochet, X. & Lefranc, M.-P. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harbor protocols* **2011,** 695–715 (2011).

140. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3,** 160018 (2016).

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

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

143. Lefranc, M.-P. Unique database numberings system for immunogenetic analysis. *Immunology Today* **18,** 509 (1997).

144. Krawczyk, K., Raybould, M. I. J., Kovaltsuk, A. & Deane, C. M. Looking for therapeutic antibodies in next-generation sequencing repositories. *mAbs* **11,** 1197–1205 (2019).

145. Krawczyk, K., Kelm, S., Kovaltsuk, A., *et al.* Structurally Mapping Antibody Repertoires. *Frontiers in immunology* **9,** 1698 (2018).

146. Van Kempen, M., Kim, S. S., Tumescheit, C., *et al.* Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* (2023).

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

148. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nature Communications* **9,** 2542 (2018).

149. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89,** 10915–10919 (1992).

150. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17,** 282–283 (2001).

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

152. Corrie, B. D., Marthandan, N., Zimonja, B., *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).

153. Młokosiewicz, J., Deszyński, P., Wilman, W., *et al.* AbDiver: a tool to explore the natural antibody landscape to aid therapeutic design. *Bioinformatics* **38,** 2628–2630 (2022).

154. Rognes, T., Scheffer, L., Greiff, V. & Sandve, G. K. CompAIRR: ultra-fast comparison of adaptive immune receptor repertoires by exact and approximate sequence matching. *Bioinformatics* **38,** 4230–4232 (2022).

155. Frostig, R., Johnson, M. & Leary, C. *Compiling machine learning programs via high-level tracing* in (2018).

156. Dejnirattisai, W., Zhou, D., Ginn, H. M., *et al.* The antigenic anatomy of SARS-CoV-2 receptor binding domain. *Cell* **184,** 2183–2200 (2021).

157. Dunbar, J., Krawczyk, K., Leem, J., *et al.* SAbDab: the structural antibody database. *Nucleic Acids Res* **42,** D1140–D1146 (2014).

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

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

160. Krawczyk, K., Buchanan, A. & Marcatili, P. Data mining patented antibody sequences. *mAbs* **13,** 1892366 (2021).

161. Fleri, W., Paul, S., Dhanda, S. K., *et al.* The immune epitope database and analysis resource in epitope discovery and synthetic vaccine design. *Front Immunol* **8,** 278 (2017).

162. Olsen, T. H., Abanades, B., Moal, I. H. & Deane, C. M. KA-Search, a method for rapid and exhaustive sequence identity search of known antibodies. *Scientific Reports* **13,** 11612 (2023).

163. Spoendlin, F. C., Abanades, B., Raybould, M. I. J., *et al.* Improved computational epitope profiling using structural models identifies a broader diversity of antibodies that bind the same epitope. *bioRxiv* (2023).

164. Sayers, E. W., Bolton, E. E., Brister, J. R., *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res* **50,** D20–D26 (2021).

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

166. Abanades, B., Wong, W. K., Boyles, F., *et al.* ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. *Commun Biol* **6,** 575 (2023).

167. Raybould, M. I. J., Turnbull, O. M., Suter, A., Guloglu, B. & Deane, C. M. Contextualising the developability risk of antibodies with lambda light chains using enhanced therapeutic antibody profiling. *bioRxiv* (2023).

168. Gilman, M. S. A., Castellanos, C. A., Chen, M., *et al.* Rapid profiling of RSV antibody repertoires from the memory B cells of naturally infected adult donors. *Sci Immunol* **1,** eaaj1879 (2016).

169. Jubb, H. C., Higueruelo, A. P., Ochoa-Montaño, B., *et al.* Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J Mol Biol* **429,** 365–371 (2017).

170. DeLano, W. L. *et al.* Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr* **40,** 82–92 (2002).

171. Mukhamedova, M., Wrapp, D., Shen, C.-H., *et al.* Vaccination with prefusion-stabilized respiratory syncytial virus fusion protein induces genetically and antigenically diverse antibody responses. *Immunity* **54,** 769–780 (2021).

172. Berman, H. M., Westbrook, J., Feng, Z., *et al.* The protein data bank. *Nucleic Acids Res* **28,** 235–242 (2000).

173. Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. & Welch, D. M. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome biology* **8,** R143–R143 (2007).

174. Giudicelli, V., Chaume, D. & Lefranc, M.-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic acids research* **33,** 256–61 (2005).

175. Wolf, T., Debut, L., Sanh, V., *et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing* 2019.

176. Van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9,** 2579–2605 (2008).

177. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. eng. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* **374,** 20150202 (2016).

178. Yuan, Y., Chen, Q., Mao, J., Li, G. & Pan, X. DG-Affinity: predicting antigen–antibody affinity with language models from sequences. *BMC Bioinformatics* **24,** 430 (2023).

179. Kitaura, K., Yamashita, H., Ayabe, H., *et al.* Different Somatic Hypermutation Levels among Antibody Subclasses Disclosed by a New Next-Generation Sequencing-Based Antibody Repertoire Analysis. *Frontiers in Immunology* **8** (2017).

180. Sun, T., Gaut, A., Tang, S., *et al. Mitigating Gender Bias in Natural Language Processing: Literature Review* in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Florence, Italy, 2019), 1630–1640.

181. Gira, M., Zhang, R. & Lee, K. *Debiasing Pre-Trained Language Models via Efficient Fine-Tuning* in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (Association for Computational Linguistics, Dublin, Ireland, 2022), 59–69.

182. Branco, P., Torgo, L. & Ribeiro, R. P. A Survey of Predictive Modelling under Imbalanced Distributions. *CoRR* **abs/1505.0** (2015).

183. Lin, T.-Y., Goyal, P., Girshick, R. B., He, K. & Dollár, P. Focal Loss for Dense Object Detection. *CoRR* **abs/1708.0** (2017).

184. Salazar, J., Liang, D., Nguyen, T. Q. & Kirchhoff, K. Pseudolikelihood Reranking with Masked Language Models. *CoRR* **abs/1910.1** (2019).

185. Tay, Y., Dehghani, M., Tran, V. Q., *et al. UL2: Unifying Language Learning Paradigms* 2023.

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

187. Zheng, B., Dong, L., Huang, S., *et al.* Allocating Large Vocabulary Capacity for Cross-lingual Language Model Pre-training. *CoRR* **abs/2109.0** (2021).

188. Why the UK-led global AI summit is missing the point. *Nature* **623** (2023).

189. Setliff, I., Shiakolas, A. R., Pilewski, K. A., *et al.* High-Throughput Mapping of B Cell Receptor Sequences to Antigen Specificity. *Cell* **179,** 1636–1646 (2019).