Computational Developability Triaging of Antibody Libraries for Discovery of New Therapeutics

James Sweet-Jones

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Institute of Structural and Molecular Biology, Faculty of Life Science

University College London

December 31, 2024

I, James Sweet-Jones, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Therapeutic monoclonal antibodies (mAbs) are a successful class of biologics in the treatment of cancers, autoimmune diseases and others. However since the first antibody treatments were developed in 1986, only about 144 mAbs have gained clinical approval from regulatory bodies at the time of writing. A contributing factor to this is that their discovery pipelines and clinical trials can be subject to late-stage failures due to developability issues, which, in brief are a mAb's intrinsic ability to be produced at scale and tolerated by the patient. Concurrently, libraries of paired heavy and light chain antibody sequences have been collected from next generation sequencing of human, or genetically engineered mice, repertoires. This thesis hypothesises that by using antibody language models, the developability features of clinically approved mAbs and library antibodies can be compared and used to screen library antibodies for new therapeutics.

As a result, a triaging pipeline has been constructed where a library of antibody sequences are input, and undergo the following processes: physiochemical feature triaging based on sequence statistics; unsupervised learning to identify antibodies with similar features to clinical mAbs; prediction of physiochemical properties using linear models and supervised learning to predict which would pass clinical

trials, and which are therefore good candidates for new therapeutics. This pipeline hopes to improve the chances of a given mAb to reach the clinic through identifying candidates with good developability profiles early in the selection process.

Furthermore this thesis has worked to develop upon antibody annotation languages where a graphical drawing program for multispecific antibodies was written in order to encourage their development and improve the consistency of cataloguing their formats. This work has already been employed by the World Health Organisation International Non-proprietary Names Committee and is applied to new and historic applicants to clinical trials.

Impact Statement

Therapeutic antibodies have demonstrated their utility in treating many diseases which threaten human health organisations including cancers, auto-immune diseases, and recently have been used as emergency treatments for Covid-19. Antibodies can be raised against antigens of interest, making them useful for targeting any part of a disease pathway. However what becomes important is to screen candidates for developability issues which can hinder a drug's ability to be tolerated by the patient or produced at industrial scale. Furthermore, overcoming these developability issues during candidate selection does not guarantee that a drug will be successful at clinical trials, as 75% of antibody therapeutics fail at this stage . The cost of entering a new biologic to the clinic is estimated at $2.6 billion (USD) and so late-stage failures result in lost time and expenditure which could have been invested in a more successful candidate.

Consequently, developability prediction has been used to avert these issues, and bioinformatic pipelines have been published and can be licensed to pharmaceutical companies in stages of lead candidate selection. The shortcomings of these current pipelines is that their statistics are either based on simple sequence statistics which are poorly informative, or computationally expensive structural modelling,

which is unsuitable for high-throughput screening and can only be applied to lead candidates. This thesis aims to address the problem of high-throughput developability prediction by using antibody language models which can encode sequence and structural information learned through training on millions of sequences quickly to find features correlated with previously successful antibody drugs. The advantage of high-throughput screening for developability prediction is that developability can be considered at a stage earlier than lead candidate selection, and can be applied to libraries generated from immunization campaigns and phage display, to identify the most suitable candidates based on developability.

The work has potential to impact a number of stakeholders. Firstly, it will benefit research into antibody developability, demonstrating that language models can be used to identify antibodies with favourable developability profiles. Secondly the pharmaceutical companies aiming to bring new therapeutics to the clinic will benefit by this pipeline de-risking the road to the clinic and causing fewer expensive late-stage failures. Thirdly, if this pipeline is successful in bringing more therapeutics to market, it has large ramifications for patient outcomes if improved therapeutics which are better tolerated are introduced.

Acknowledgements

Thank you to Andrew for his continued guidance over the past four years, and for being the first LIDO tutor to reach out to me.

Thank you to Alex, Chris and all of the friends and family who have helped me through this process.

Thank you to all of the members of the Martin and Fraternali labs for your support during this process. Particularly to Chu'Nan Liu, Lillian Denzler and Veronica Boron.

Furthermore, I would like to acknowledge Dr. Denis Larkin, Prof. Rob Fowkes and Prof. Imelda McGonnell of The Royal Veterinary College for supporting my original application, without whom, I would not have pursued the doctoral degree.

For assistance throughout the project, additional thanks goes to to Okan Aydin at ENPICOM for working with me using the IGX Platform and Chris Ford at GenScriptBio for providing antibody expression quotes.

Contents

Data files are stored in an online repository

List of Figures

List of Tables

Abbreviations

ΔG Gibbs' Free Energy of Unfolding.

C_H Constant Domain Heavy.

C_L Constant Domain Light.

Fc Fragment Crystallizable.

Fv Fragment Variable.

V_H Variable Heavy Domain.

V_{HH} Variable Heavy Domain of Heavy chain.

V_L Variable Light Domain.

AAR Anti-Antibody Reaction.

AbML AntiBody Markup Language.

ADA Anti-Drug Antibodies.

ADC Antibody Drug Conjugate.

AGL Assign Germ Line.

BCR B Cell Receptor.

BERT Bidirectional Encoder Representations from Transformer.

BiTE Bispecific T Cell Engager.

CDR Complementary Determining Region.

CV Cross-Validation.

Fab Antiven Binding Fragment.

FDA U.S. Food and Drug Administration.

G Score Germline Score.

H Score Humanness Score.

HELM Hierarchical Editing Language for Macromolecules.

HIC Hydrophobic Interaction Chromatography.

Ig Immunoglobulin.

LLM Large Language Model.

mAb Monoclonal Antibody.

MsAb Multispecific Antibody.

NB Naive Bayes.

OAS Observed Antibody Space.

PCA Principal Component Analysis.

pI Isoelectric Point.

PTM Post-Translational Modification.

sc*Fv* Single Chain Fragment Variable.

Sn Sensitivity.

Sp Specificity.

SVC Support Vector Machine Classifier.

Tagg Aggregation Temperature.

TAP Therapeutic Antibody Profiler.

TCR T Cell Receptor.

Tm Melting Temperature.

WHO-INN World Health Organisation International Non-proprietary Names Committee.

Chapter 1

Introduction

This chapter aims to give the reader a sufficient background on antibodies, their genetics and structure, explaining where antibodies come from and how they can be collected to assemble libraries. After this, the introduction covers therapeutic antibodies, developability and developability prediction. These sections are necessary to engage with the results chapters of the thesis.

1.1 Introduction to Antibodies

Antibodies, or immunoglobulins (Ig), are large 'Y'-shaped proteins that play a major role in the adaptive immune system's response to infection [1]. They carry this out by binding to antigens with high affinity and specificity to neutralise them or initialise the formation of a membrane attack complex [2]. Antibodies consist of four chains: two identical heavy chains and two identical light chains, where the variable domains of the heavy (V_H) and light (V_L) chains interact to construct an antigen binding region. Although antibody amino acid sequences are highly conserved, points of diversity give an antibody its specificity. Consequently, monoclonal antibodies

(mAbs), antibodies with the same binding affinity and specificity, are useful tools in experimental reagents, imaging, and therapeutics. mAbs have become an important class of biologic drugs capable of treating a variety of diseases including cancers, autoimmune diseases and recently, Covid-19 because they can target specific steps in a disease pathway [3, 4]. This thesis introduction covers antibody structure and function, the introduction of mAb therapeutics and developments to improve their discovery pipelines and functionality.

1.2 Antibody Genetics, Structure and Function

Antibodies are produced by B cells, each producing an antibody with a heavy and light chain unique to clones of that cell [4]. While all immune system cells originate from heamatopoetic stem cells in the bone marrow, B cells undergo a unique differentiation that is dependent on their ability to generate functional antibodies. At each stage of B cell maturation, if the antibody produced is not functional or the antibody binds to self-antigen expressed in the bone marrow, the differentiating B cell will commit apoptosis as part of the negative selection against non-functional antibodies [5]. This section will detail how those antibodies are produced and how diversity is introduced.

There is no single gene to code for antibodies, nor does each possible antibody have its own gene embedded in somatic cells. In fact, there are many genes that encode the regions of the antibody that are recombined during the process of B cell differentiation. At the IGH chain locus on human chromosome 14, a selection is made for one of each of 39 IGHV (variable), 27 DH (diversity) and 6 JH (junction)

genes, which are recombined to give a full V_H domain in the heavy chain. For the light chain, there are two loci offering a greater diversity for selection, one for IGK on chromosome 2 for κ light chains and IGL on chromosome 22, offering 40 V-kappa with 5 J-kappa, and 32 V-lambda and 4 J-lambda gene segments respectively [6]. For both V_H and V_L domains, it is standard to the V region gene to trace the 'germline family' which genes the antibody has come from.

Heavy chains have one variable domain (V_H) and a dependable number of constant domains (C_H1 , C_H2 , C_H3 , C_H4), usually with a hinge region between the C_H1 and C_H2 domains (except for antibodies of the IgM and IgE classes), whereas light chains have one variable domain (V_L) and one constant domain (C_L). These chains interact so the two heavy chains interact at the bottom prong of the 'Y' shape, which are then made up of the interaction of the hinge regions and constant regions. This gives the *Fc* fragment of the antibody. Furthermore, light chain and heavy chain pairs interact to form one fork of the 'Y' shape each, which makes up the *Fab* regions of the antibody, which is the interaction of the variable domains, and the C_H1 and C_L domains (Figure 1.1a). Lastly, the variable region (*Fv*) is where interaction of the V_H and V_L domains of the antibody takes place. This interaction gives the antibody its antigen-binding properties. The *Fv* region of the antibody demonstrates a high level of polymorphism in sequence and local structure, which is achieved through rearranging and selecting the genes that express the antibody inside the B cells which produce them [4, 7].

Within each V_H and V_L domain of the protein, there are three hypervariable loops known as 'complementarity determining regions' (CDRs) that have the most

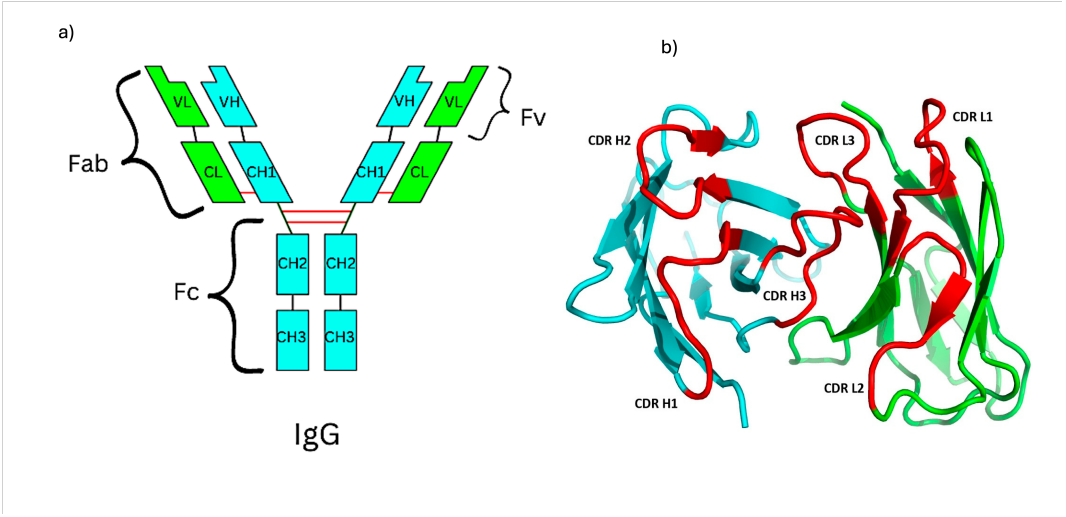

Figure 1.1: Antibody topology. a) IgG antibody showing heavy chains in blue and light chains in green. Domains have been labelled and fragments including *Fc*, *Fab* and *Fv* have been labelled. b) *Fv* fragment showing V_H and V_L domains as ribbons, highlighting the complimentarity determining loops in red at the interface of the domains [13].

interaction with the antigen and therefore are attributed the most importance in binding. CDR1 and CDR2 of the V_H and V_L regions of the protein are both located within the V gene segments and their structures follow canonical classes that have been relatively well characterised [8, 9]. However, the CDR3 sequence overlaps both of the V and J segments in the V_H and V_L domains, with the addition of the entire D gene segment in the V_H CDR3 loop (CDR-H3). In addition, junctional diversification occurs where insertions or deletions at these junctions contribute to even greater diversity and, in turn, high to binding affinity (Figure 1.1b)[10, 11, 12].

Before a B cell matures, it undergoes a series of checkpoints in the bone marrow, including pairing the heavy chain with a surrogate light chain, and an immune checkpoint to ensure that the antibody produced does not bind to self-antigens, a mechanism which fails in autoimmunity [14]. If these checkpoints are passed successfully, the pre-B cell may proliferate and then begin rearranging the light chain

genes in attempts to pair them, starting with a κ light chain. In cases where the light and heavy chain do not pair, the cell may either iterate its κ light chain recombination or switch to a λ light chain genes. If, after this recombination, the resulting combinations fail to pair, or the antibody is found to bind to the self-antigen, the cell undergoes apoptosis. Otherwise, the cell progresses to an immature B cell and is released into the blood and lymph tissue [12].

Initially, B cells produce antibodies of the IgM and IgD classes, which are generally low-affinity antibodies which then undergo somatic hypermutation. When signalled through interleukin T cell-dependent pathways, the antibody may switch classes to either IgG, IgE or IgA antibodies by switching the constant region domains that are expressed with the antibodies. The genes encoding these domains are found in a locus downstream of the variable domains and ordered: IgM, IgD, IgG, IgE and IgA. Appropriate constant domains are selected by removing the protein-coding sequences that are before it in the locus. For example, if an antibody switched to IgE: the genes encoding IgM, IgD and IgG classes and subclasses would all be removed but the gene for IgA would remain, but not be transcribed. The chosen class can depend on cytokines signalled to the B cell [15].

The mature B cells producing high affinity antibodies will go on to differentiate into memory cells. During differentiation into memory cells, somatic hypermutation may occur to generate higher affinity antibodies, usually belonging to the IgG class [4]. Once the antibody becomes highly specific to a given target, clones of the same antibody will continue to be made by the memory cells upon reactivation through encountering the original antigen that caused those antibodies to be raised.

To study this diversity and how these sequences of different lengths relate back to the original germlines and to compare residues in a position in relation to another sequence, numbering schemes were developed as opposed to traditional sequence alignments. As of 2024 there are five established schemes which have been published and adopted by researchers: Kabat [8]; Chothia [16]; Martin (also known as enhanced Chothia) [17]; AHo [18], and IMGT [19] which aim to apply antibody numbering in different scenarios by allowing different substitution and insertion mutations as well as changed definitions of CDR loops [7]. For instance, the IMGT numbering scheme is harmonized across antibody V_H and V_L domains and T Cell receptor variable domains, whereas the Chothia numbering scheme is more focused on structural alignment by incorporating the canonical CDR loops.

Taking this into account, there are numerous stages of B cell maturation where diversity may be introduced to the B cell receptor sequence. Estimates of the potential diversity of the antibody repertoire have varied between 10^{13} [20] to 10^{15} [21] through to 10^{18} [22]. Whatever the actual number may be, the importance of it is that the potential diversity is astronomically large and full of useful sequences for antibody applications. The next section will explore how collected data is starting to explore this space and how these datasets may be mined for those sequences.

1.3 Antibody Libraries

Although having an antibody with a known antigen is useful information, it does not give immediate clues as to how to pair other antibodies with antigens. It was not until phage display, a screening method in which a large number of antibodies stored

in physical libraries could be screened against an antigen of interest simultaneously and binding antibodies could be identified. These libraries are built by taking known sequences of antibodies obtained from Sanger sequencing. While these would have been limited in number, the library could be augmented by cloning and diversifying these sequences through multiple sets of amplification primers [23].

Phage display works by expressing different V_H and V_L domains on the surface protein coat of M13 bacteriophages as a single chain *Fv* (sc*Fv*) fragment. This is where the V_H and V_L domains are conjugated by an artificial peptide linker. The antigen of interest is be bound to a plate where the phage with a binding antibody variable domain will bind. Successive rounds of 'panning' where antibodies are allowed to bind to the target are carried out. Antibodies which do not bind are washed away with a buffer solution then antibodies which do bind are eluted from the solution with a low-pH buffer to stop the interaction between the antigen and bound phages. The phages that remain are then amplified to increase the number of binders. Antibodies that bind after successive washes of increasing concentration of buffer solution are expected to be rare strong binders within the library and can then be recovered and characterised [24, 25]. Display libraries have powerful applications to search for an antibody which binds to an antigen of interest . While phage display can offer large libraries (HuCAL offers up to 10^{11} [26, 27]), the chances of success of phage display are dependent on the quality and diversity of the library used [27]. Phage libraries require appropriate conditions including -80°C and protection from environmental factors for long-term physical storage.

Other research has been conducted on how the repertoire of B cell receptors

changes between age groups or in response to immunological challenge or immunisation. This work has been accelerated through improvements in single cell RNA sequencing allowing unpaired and paired antibody sequences to be recorded [28, 29, 30, 31, 32]. As a byproduct of this work, digitally stored libraries of hundreds or thousands of paired antibody DNA or amino acid sequences have been made available as a resource, mostly through the Observed Antibody Space (OAS) dataset [33].

Although protocols for single-cell B cell repertoire sequencing may differ between studies, they are mostly based on the same principles. Donor B cells may be collected from human blood samples, isolated, and then encapsulated in droplets [28] or loaded into flow chambers containing only one cell [34]. Paired libraries of V_H and V_L sequences are then generated after reverse transcription PCR and sequencing of amplicons [28, 35].

Furthermore, pairing heavy and light chains becomes possible through the use of DNA barcoding the transcripts of individual B cells, allowing them to be traced back to a single cell. Knowledge of pairing is important for understanding how the repertoire generates its diversity and effects [36]. This work has suggested that pairing between heavy and light chains may not be a random process and that some favoured pairings may be more stable [28], something that was previously demonstrated in smaller datasets [37]. However, it is acknowledged that what is captured in these datasets is only a fraction of the potential diversity that was discussed in the previous section as these datasets are usually technologically limited to around 10,000 paired sequences [35]. Therefore, efforts to explore this space are needed to

make use of this data.

There has also been much commercial interest in generating libraries using antibodies sequenced from immunised animals and human patients to screen them for useful sequences. Although several studies have made these data available, it is likely that there are many proprietary datasets that are not in the public domain for commercial or privacy reasons [38]. For example, Kymab generate around 10^6 sequences per week (personal communication). These proprietary datasets are generated either by sequencing particular groups of patients with potentially strong immune systems or by introducing novel mutations to existing sequences and augmenting the data to provide added value to these libraries [39]. Hypothetically, such libraries designed for specific applications have a higher chance of finding antibodies that have clinical potential, rather than the snapshots taken in generating more general antibody libraries.

1.4 mAb Therapeutics

The potential of antibody-based therapeutics was postulated long before their first regulatory approval. Between 1891-1896 Ehrlich described a hypothesis for the formation of antibodies (then antitoxins) as a ‘magic bullet’, meaning a drug that was perfectly suited for a specific target without side effects. These ideas were later applied by Behring and Kitasato in serum therapies for diphtheria and tetanus where the blood of animals exposed to weakened bacteria was inoculated into patients and offered immunity in future disease challenges [40].

While this offered hope that these therapies can be applied, there were short-

comings that Ehrlich would not have been aware of. Firstly, the antibody response produced in nature is polyclonal, which means that many different antibodies produced by different B cell lineages are made against the same infectious agent challenge, each with a different binding affinity that makes the process suboptimal and difficult to quality control. Secondly, due to the size and molecular complexity of antibodies, inoculating antibodies from one organism to another runs the risk of generating an immune response against those foreign antibodies, even if the organism is the same species. Thirdly, inoculating with serum could also risk infection with blood-borne diseases.

The first of these points was addressed by producing monoclonal antibodies (mAbs), which are clones of the same B cell with producing the same antibodies binding to the same epitope of an antigen with the same binding affinity [41]. This was especially directed toward the IgG class of antibodies as these are circulating antibodies that, after undergoing rounds of somatic hypermutation, bind to their antigen with a higher affinity than IgM. Kohler and Milstein became the first people to publish a method of producing mAbs using the hybridoma technique, where B cells that produce antibodies of the same lineage are hybridised with immortal myeloma cells. The resulting hybridoma would then produce and secrete many copies of the same antibody which is much easier to screen for quality and develop into a clinical therapeutic [41].

Despite the inception of mAbs, mass production of antibodies using hybridomas is not feasible, so production of mAb therapeutics is done as any other biologic drug is made. The antibody coding genes are excised into a bacterial plasmid

that may be transfected into, and expressed by a variety of eukaryotic cell lines on an industrial scale in a production culture. Most often it is the Chinese Hamster Ovary (CHO) cell line [42]. This made it possible for the first licensed mAb drug, Muromonab, to be approved by the US Food and Drug Administration (FDA) and released in 1986. This was a murine IgG antibody (OKT3) used to treat kidney transplant rejection by targeting the CD3 cell marker expressed by T cells and suppressing the immune response [43].

Although this was an incredible breakthrough in biologic drug design, there is still a significant risk that using a murine antibody for a drug in humans will lead to an immune response. However, in the case of kidney transplant rejection, it is likely that these patients will also be taking other immunosuppressive drugs, and so patients are less likely to mount a response in the first place. The phenomenon of immune response to mAb therapeutics causes great concern in their development pipeline. This immune response, known as an anti-antibody response (AAR), and the increase in anti-drug antibodies (ADAs) in the blood means that future administrations of that drug will be eliminated before the drug exerts its desired effect [44, 45, 1].

Before the first murine antibody therapeutic was approved for use, chimeric antibodies were developed using mouse variable domains conjugated to human constant domains [46]. The importance of this process was not only a reduced chance of immunity, but also the ability of mAb drugs to effect downstream effects via the human $Fc\gamma$ receptor with immune recruitment, complement and recruited basophil degranulation [47, 48]. Humanized antibodies take this principle a step further, and are

the result of grafting mouse CDRs onto human variable domain frameworks, again leaving the constant regions as human [49, 50]. 'Back mutations' are usually necessary to recreate some of the murine framework residues important for the structure and orientation of the CDR loops after humanization, which is usually done through computational analysis or artificial phage display libraries to find residues and positions which minimise immunogenicity and maximise binding. This offered a new method of reducing the immunogenicity of mAb therapeutics and allowing the drug to tackle a host of new drug targets that did not require immunosuppression.

While the majority of therapeutic mAbs have arisen from these immunisation programs, phage display became useful to screen many antibodies against the antigen of interest simultaneously [24, 25]. This technique is useful as a high-throughput screen to identify leads in drug screening programs, and to make antibodies a more feasible class of biologics. At least 17 approved mAb therapeutics have emerged from phage display including Adalimumab [51], Necitumab [52] and Avelumab [53, 27]. Furthermore, fully human antibodies can be produced from transgenic mice strains including the HuMab Mouse [54] and XenoMouse brought about by deleting or silencing the chromosomal loci which express mouse antibody regions and introducing human antibody genes using yeast artificial chromosomes [55, 56]. Successful drugs from this method of discovery include Ipilimumab [57], Brodalumab [58] and Cemiplimab [59].

More recently, Tixagevimab and Cilgavimab are antibodies isolated from recovered human Covid-19 patients that were given FDA approval to treat the same condition [60, 3]. This story has demonstrated the value for the human antibody

repertoire to be mined for potentially useful antibody sequences, particularly in recovered patients as a method of quickly responding to outbreaks of a new pathogen, or finding individuals with rare immunities to cancers or neurodegenerative diseases despite genetic predispositions.

Often, any antibody recovered from organisms, phage display libraries, or single B cells will undergo some sort of optimisation such as affinity maturation to maximise its binding ability or to minimise developability liabilities. General approaches to this include taking the lead antibodies and diversifying them by introducing mutations to the VDJ genes from which the antibody was transcribed through error-prone polymerase chain reaction [61]. Furthermore, these new mutants can be recombined with different chains by shuffling, as was the case in generating the HuCAL phage display library [62, 27]. These follow the assumption that these mutations may have a combined synergistic effect to increase the affinity of the antibody [63]. Phage display is then used again to identify the antibodies that bind the strongest to the antigen. These are then taken forward into pre-clinical and clinical trials [64].

The process of drug market approval is to recruit participants into clinical trials that measure the effectiveness of the drug against another treatment or a placebo. Usually, three phases, increasing in scope and participant number, are undertaken before approval is sought from regulatory bodies [65, 66]. By the time a drug reaches Phase 2 trials, it will usually be given a generic name by the World Health Organisation International Nonproprietary Names Committee (WHO-INN). These names are general terms used to describe drugs that are not linked to a particular

brand and are constructed from a set of syllables. Previously for mAb therapeutics, they end in the 'mab' suffix, with an indication of their source: 'omab' for mice; 'ximab' for chimeric; 'zumab' for huamnized and 'umab' for fully human antibodies. However this naming scheme was retired in 2017 due to market perceptions that fully human antibodies are better and the advent of more complex constructs which required a new naming scheme before the finite possibility of names is expended [67]. The new naming scheme now adopts the '-tug' and '-bart' suffixes for unmodified and artificial immunoglobulins respectively as well as '-ment' and '-mig' suffixes to denote antibody fragments and multispecific antibodies.

Using data from clinical trials, the sponsor, which is usually a pharmaceutical company, will apply to regulatory bodies such as the FDA for approval to use the drug in those jurisdictions for a specific indication. If the data demonstrate that the mAb gives a benefit that outweighs its potential risks, it will receive market approval; however in 75% of cases of all mAb therapeutics in clinical trials [68], the drug does not show efficacy in a clinical setting and will be discontinued or repurposed by the pharmaceutical company. The process of developing a biotherapeutic and taking it through clinical trials is estimated to cost around $2.6 billion (US Dollars), so drugs which fail at these later stages become hugely expensive for the sponsors[1].

Antibodies are now becoming a fast-growing sector of the biologics market and FDA-approved therapeutics are used to treat a wide range of ailments such as cancers, autoimmune diseases and infection, including Covid-19 [3, 65]. The ther-

[1] https://phrma.org/-/media/Project/PhRMA/PhRMA-Org/PhRMA-Org/PDF/P-R/proactive-policy-drug-discovery.pdf

apeutic mAb market received a valuation of $210 billion (US Dollars) in 2022[2]. At the time of writing, IMGT [19] reports 130 currently approved Whole mAb therapeutics. Additionally, three market-approved multi-specific antibodies (MsAbs): Blinatumomab (bispecific T-cell engager; [69]); Catumaxomab and Emicizumab (both bispecific IgGs); [70].

Engineering to express the antibody only as an sc*Fv*, or a chain of conjugated sc*Fv*s have many advantages including a reduced chance of immunogenicity on a number of fronts: sc*Fv*s are much smaller than a full immunoglobulin and therefore less likely to be identified by the immune system; the *Fc* fragment is not present to mediate downstream effects [71]. Additionally, these engineered molecules give the option of designing intracellular drug delivery systems through internalisation of sc*Fv*s [72]. However, if the *Fc* fragment is required to prolong half life of the drug, proprietary *Fc* silencing mutations that are introduced including LALA, LALAPG and STAR have shown greater ability of reducing *Fc*-mediated immune recruitment through inhibiting interactions of the *Fc* domain with the *Fc*γ receptor [73, 74]. These mutations have demonstrated better ability at reducing immune recruitment over using what was traditionally thought of as silent isotypes such as IgG4. Furthermore, the ability to design sequences *de novo* or generate diversity within a sequence using generative language models has now brought about a new possibility of where future therapeutics may come from [75, 33].

[2] https://www.grandviewresearch.com/industry-analysis/monoclonal-antibodies-market

1.5 Multi-Specific Antibodies

The success of mAbs in the clinic has inspired antibodies with multiple specificities to be put into clinical trials. These are non-natural, engineered molecules which can bind to two or more antigens [76]. The inception of these molecules started with the quadroma by fusing two different hybridoma cell lines used to generate mAbs. This fusion could generate a product where during heavy chain pairing, a heterodimer from the formation of the heavy and light chains of one antibody with another heavy and light chain of a different antibody would result in an antibody with two different antigen binding regions. However, this was a wasteful process as the correct assembly was only one of 10 possible products of the pairing of heavy and light chains this way [77, 78].

Consequently, efforts for more scalable synthesis have led to new techniques of multi-specific antibody (MsAb) generation [79]. DNA recombination has allowed greater flexibility in the design of MsAbs with IgG-like formats, which can be done by appending additional sc*Fv*s, or camelid single domain V_{HH} fragments (nanobodies) at the N-terminus or C-terminus of the heavy and light chains using engineered linkers [80, 76, 81]. All of these can give rise to symmetrical antibodies in which the correct pairings of light and heavy chains are not disfavoured, as seen in the quadroma.

Alternatively, asymmetric antibodies can be produced by introducing mutations that encourage heterodimerization of heavy chains or specific pairings of light and heavy chains with different specificities. Additional residue mutations for knobs-into-holes (KIH) formats [82] are typically used to form heavy chain

heterodimers by introducing mutations in the C_H3 domains, while introduction of positively and negatively charged residues in the C_H1 and C_L domains of one arm [83] helps in the correct pairing of light and heavy chains to make the desired asymmetric antibody format more favourable [79].

Protein engineering also allows the generation of smaller fragment-based MsAbs including 2-chained diabodies or Bispecific T-Cell Engagers (BiTE) [84]. These non-IgG-like molecules are advantageous because they are easier to produce (requiring no glycosylation), but they are limited by short half-lives, which can be extended through human serum albumin (HSA) fusion, PEGylation (addition of polyethylene glycol), or the addition of cysteine residues which form disulphide bonds [85, 86]. Antibody-drug conjugates (ADCs) have become popular for delivering small molecule drugs to an intended target [87]. Most recently chemical conjugation by thiol-thiol or amide-amide linkers for ligating antibody fragments in this way has been seen in the 'Dock and Lock' and by ligating two IgG molecules to give IgG-IgG molecules [79, 88]. Moreover, conditionally active mAbs which require cleavage of a protein or chemical linker to begin activity in response to a change in environment pH, temperature, or presence of an enzyme in a tumour microenvironment [89, 90].

In addition, molecules based around T cell receptors and fusions of these with sc*Fv*s (such as the ImmTAC format) are becoming popular and being able to describe and draw these is becoming more important [91]. While only a handful MsAbs have thus-far been given regulatory approval (all bispecifics), many more are in development and in clinical trials [66, 68]. Given the huge diversity of pos-

sible MsAb formats, a standardized format for description and annotation would be advantageous, for example when they are submitted to the WHO-INN or for regulatory approval. For small-molecule drugs, 'Simplified Molecular-Input Line-Entry System' (SMILES) strings [92] have been adopted as a standard for describing organic molecules. As yet, no such standard has been widely adopted for biologics.

The Hierarchical Editing Language for Macromolecules (HELM) [93] was introduced in 2012 as a general tool for describing biologics (including antibodies). It provides a visual editor and has the support of a number of large pharmaceutical companies including GlaxoSmithKline, Merck, Roche and Pfizer. Nonetheless, it has only gained limited traction in the annotation of antibodies and is not currently used by regulatory authorities, the WHO-INN, or the Chemical Abstracts Service (CAS) for description of antibody-based drugs. Current limitations which make HELM less suitable for MsAbs an inability to notate different specificities of given *Fv* on the structure; inability to notate modifications or mutations in a given domain; and additional complexity to support other macromolecule biologics. Furthermore, rather than allowing the user to draw a schematic for an MsAb using simple domain blocks, the HELM editor requires amino acid sequences in an attempt to draw a schematic automatically.

1.6 Developability Prediction

Despite success and developments in discovery and lead optimisation, still many mAbs fail to be suitable drug candidates. It is important that antibodies taken as leads are considered "developable" from their inception, where they can be pro-

duced on large scales, remain stable enough for storage and tolerated by the patient [94]. This is why the process comes with a high rate of attrition. Although antibody-based therapeutics make up a large number of current clinical trials, this does not guarantee that the drug will succeed in trials, as 75% are discontinued before market approval [68, 45]. Usually, the reasons why a drug fails clinical trials are not directly reported but are most likely due to: a lack of clinical efficacy; the immune response of the patient causing a loss of effect in the drug; adverse reactions to the drug or failure to produce an adequate titre of the drug at scale. However, commercial reasons, including funding withdrawal and poorly designed trials, can also cause a drug to be discontinued from trials [3]. As these commercial reasons are difficult to learn from as they are rarely reported, this has driven research into the physicochemical properties of mAbs which contribute to their chances of approval.

Several physicochemical features of antibodies have already been identified as possible contributors to in their approval as drugs. Possibly the most important of these is the propensity of the antibodies to aggregate. For example, the propensity for drug aggregation contributes to poor shelf life and immunogenicity, as it forms a large structure when aggregated in the blood [95]. The difficulty is then to untangle what exactly drives this aggregation propensity.

Although impurities in drug formulation may contribute to aggregation, more attention has been paid to the sequence and structure features of antibodies themselves [96]. Several of these features have already been identified. Thermostability (ΔG) is a measure of the likelihood that a protein will misfold at high temperatures. Proteins with poor thermostability and high feasibility of spontaneous un-

folding will have a high propensity to aggregate to maintain the distance between their hydrophobic residues normally present in their cores and the aqueous environment. Patches of charged or hydrophobic residues will also encourage aggregation in a similar manner, especially if they happen to be in the CDR loops, as these are exposed residues on the antibody surface. Post-translational modification ligation including glycosylation [97, 98]; deamidation [99] and oxidation [100] also present risks in protein stability that could result in instability or could incite an immune response in their own right [101].

Methods for measuring some of these developability characteristics *in vitro*, include differential scanning fluorimetry (DSF) assays to measure thermostability, and their melting temperature [27], hydrophobic interaction chromatography (HIC) to measure the hydrophobicity of proteins, and dynamic light scattering (DLS) to measure the propensity for aggregation [102, 103, 3]. However, the nature of these assays requires expensive protein expression at a suitable concentration and timely assays, meaning that it is impossible to run these tests in a high-throughput discovery pipeline, and so they are usually reserved for a small number of candidates that are found to bind to the target. As mentioned previously, another consideration for an antibody drug is its immunogenicity. Despite this phenomena being observed since the beginning of antibody therapeutics, there is a poor understanding of why a mAb would be immunogenic, and even scarcer data on the subject except for the proportion of patients that mount anti-drug antibodies to new therapeutics for a select few mAbs [104].

Although introducing mutations to these sequences may improve one of these

characteristics, it may have a negative impact on another, which has recently brought about understanding that these features are closely related and that antibodies with developability characteristics must occupy a space where all of these properties are in a tolerated range: a so-called "developability web". Despite a mAb performing well in an *in vitro* setting, this does not guarantee success in clinical trials and most will be discontinued due to poor efficacy, safety concerns such as anaphylactic shock, or resistance due to immunogenicity. This raises even more questions and blurs the line between if a drug is unsuccessful because of poor clinical efficacy or because of mounting an immune response.

The drive to overcome these approval barriers has led to many investigations into the properties described above and software has become available to predict the properties of naïve antibody sequences *in silico* using only their sequence. This allows for faster and cheaper high-throughput screening of candidates to evaluate which are more likely to be successful before they are taken to trials. The Developability Index [105, 106] was the first to offer scoring for aggregation propensity by identifying regions prone to aggregation using homology modelling to calculate the isoelectricity and solvent-accessibility-dependent spatial propensity with a scaled value. However, they also realised that they did not take into account the effect that post-translational modifications could have on the aggregation propensity of an antibody.

Attempts by Hebditch and Warwicker [107] to use computational means to predict physicochemical characteristics related to developability from an experimental data [102] set showed a weak correlation of predictive ability with experimental

data, but it has still been published for other researchers to use [107]. This has been confounded with previous efforts for the prediction of immunogenicity, which are mainly based on predicting the similarity of a given sequence to mouse sequences [108, 109, 104].

More recently, the Therapeutic Antibody Profiler (TAP) [110] was created as a development scorer, again generating a model of the antibody to assess surface-exposed hydrophobic or charged residues in the CDR loops to assess aggregation propensity. The Deane group used 242 clinical stage therapeutics and the database used to train TAP to understand the levels of hydrophobicity and aggregation propensity seen in clinical antibodies. Another recent method was produced by Negron *et al.* [111], the TA-DA pipeline which demonstrated an ability to separate clinical mAbs from a large library of naturally occurring antibodies by assessing *in silico* calculated properties correlated with clinical mAbs to calculate a developability score [111]. Although these are promising for candidate selection, the TAP software only assesses one antibody at a time and requires computationally expensive modelling, making it unsuitable for examining a large library, and the TA-DA algorithm has not been made available for use.

1.7 Hypothesis and Aims

With the landscape of antibody database expansion and pairing changing with the advent of B cell repertoire sequencing, tools are required for high-throughput drug screening that can be applied to paired V_H and V_L antibody sequences. We aim to develop a pipeline of software that can overcome current barriers to therapeutic

mAb approval by promoting early selection of leads with developability profiles that match previously approved mAbs to reduce the risk of failure later on in the clinical trials setting. Transcriptomic sequencing has been instrumental in building antibodies libraries that are used in the screening of new possible candidates. New methods of representing antibodies, including large language models (LLMs) trained on antibody sequences, are ways of learning subtle sequence patterns that may represent features that are not yet understood. It is the aim of this work to apply machine learning and deep learning to antibody sequences using these models to compare approved and discontinued datasets and to compare how these differ from the repertoire antibodies.

This project has been inspired by the need to develop an end-to-end bioinformatic pipeline that can make use of the large libraries of paired antibody sequences that are emerging from next-generation sequencing platforms. This pipeline should also have the capacity to analyse many sequences together for a high-throughput screening of candidates. This means that it will have to avoid modelling the antibodies to save computational time and work on sequence-based statistics only rather than modelling structures of antibodies. Consequently, it is proposed here to use encoding of antibody sequences by antibody language models and applying machine learning predictors to capture meaningful differences between library and clinical mAb drugs to better understand the reasons why some drugs become successful and others do not. In addition, the thesis attempts to address the problem of describing complex structures of (typically) MsAb-based drugs by further developing the Antibody Mark-Up Language developed in the Martin group and developing a drawing

display tool to work with AbML.

The first results chapter, Chapter 3, will cover separating clinical antibodies from repertoire sequences using both supervised and unsupervised learning using antibody language model encodings. Chapter 4 will examine using the antibody language models encodings to train linear models to predict physicochemical properties of clinical stage antibodies. Chapter 5 will then look at training binary classifiers on market approved mAb therapeutics and clinical stage mAbs which were discontinued at clinical trials. Chapter 6 then looks to combine the results from the previous chapters into a bioinformatics pipeline to detect antibodies with developability characteristics and are likely to pass clinical trials. An antibody repertoire dataset is used as an example input for this pipeline. Chapter 7 then outlines the developments in antibody annotation, AbML and abYdraw.

Chapter 2

Materials and Methods

2.1 Online Datasets and Resources

This chapter will describe the antibody sequence datasets and specialist antibody tools that were used throughout the project.

2.1.1 abYsis

abYsis [112] is a server that hosts 5896 paired human sequences taken from solved Protein Data Bank structures, and EMBL-ENA databank [113] and the Kabat antibody sequence database [114]. The server also hosts a number of useful antibody analysis tools including AbNum, abYmod and annotation for post-translational modification recognition sites.

2.1.1.1 AbNum

To number antibodies, AbNum is a tool supplied by the abYsis server which is capable of numbering V_H and V_L sequences according to the Kabat, Chothia, Martin, IMGT and AHo schemes. AbNum works by splitting the sequence into framework

and CDR loops using profiles derived from the Kabat database [114]. The profiles are slid across an input sequence to find the best match, to define anchor points in the sequence. AbNum then numbers the sequence from each pair of adjacent anchor points until a site where insertions are allowed is found. Any remaining residues are then given insertion codes [17].

2.1.1.2 abYmod

abYmod is another tool offered by abYsis where homology models are generated for the antibody framework regions and CDR1, CDR2 and CDR-L3 loops through templates of the Chothia canonical classes [16] selected as best sequence identity matches from the PDB. To find an appropriate loop for CDR-H3, the program searches a database of CDR-H3 loops from antibodies, but if no loops with the same length are found, it searches a database of loops from all proteins having appropriate takeoff geometry. It selects examples by correct length, ranks by sequence identity and inserts the top ranking loop into the template. Side-chains are modelled from the outside to the centre of the binding site using the minimum pertubation protocol [115]. The Gromacs energy minimization software [116] is used to optimise the model by removing side-chain clashes and errors in grafting loops to the framework. abYmod is available as a web server [1].

2.1.2 TheraSabDab

TheraSabDab is a publicly available dataset of therapeutic antibodies with assigned WHO-INN names [117]. Heavy chain and light chain sequences are reported, as

[1] abymod.abysis.org/

well as their current stage of clinical trials, approval or discontinuation, and the target of the antibody. This dataset collects sequences from the WHO-INN [67]. TheraSabDab was accessed in October 2021.

2.1.3 Observed Antibody Space

The Observed Antibody Space is an online repository of paired and unpaired V_H and V_L sequences compiled from studies that have resulted in large-scale repertoire sequencing [118, 33]. However, these studies have several methods for pairing antibody sequences and focus on a number of patient responses after exposure to different pathogens, including SARS-CoV2 [29], HIV [30] and Cytomegalovirus [32] as well as healthy patients [31], which can skew the observed result. OAS was accessed in June 2022.

2.1.4 Anti-Drug Antibody data

A dataset from Marks *et al.* [104] and Clavero-Alvarez *et al.* [119] assembled ADA incidence for therapeutics taken from clinical trial data found on DrugBank [120]. These data are simply reported as a percentage of patents in a trial that appeared to demonstrate an ADA titre above a given threshold used in that study. However, there can be disagreement between trials on the threshold to report ADAs, and in some cases multiple numbers for the same drug have been reported across trials. In the case of multiple incidences, the mean reported incidences was given. The result was a list of incidences for 217 therapeutics.

2.1.5 Pure2

Data was taken from Stewart *et al.* [121] where B cells were isolated from three healthy male blood donors were isolated and sorted into developmental stages with distinct phenotypes using fluorescence-activated cell sorting. The transcriptomes were sequenced to generate 10X Genomics v3.1.0 libraries through CellRanger [2]. Three additional donors had blood collected and data from their B cell repertoires representing an older cohort HB91 (male, 77 y.o.), HB86 (male, 69 y.o.) and HB7 (female, 68 y.o.) were added to the Data taken from Stewart *et al.* [121] as provided by Franca Fraternali and her group. Antibodies were paired by a unique cell barcode and in cases where multiple different chains were observed with the same barcode, the chain with the highest count was taken as the true antibody. In total, 10,492 paired antibodies were extracted from the six patients as an ideal example dataset for our pipeline to work on.

2.2 Statistical Tests

2.2.1 Mann-Whitney U test

The U test is a null hypothesis test carried out on two populations to determine whether they are statistical different, similar to a parametric t test. The U test is a nonparametric test, meaning that it does not assume a normal distribution of the two samples, which in the case of physicochemical property data may be the case. The values of both populations are combined into a single dataset and ranked from smallest to largest and the U statistic is calculated for each group and the final

[2] https://github.com/10XGenomics/cellranger

U statistic is then given as the minimum of the U statistics of both groups and compared to the critical value ($p < 0.05$) to determine significance (Equation 2.1). A lower value rejects the null hypothesis, meaning that the samples are different, where the opposite is true for a higher value [122].

$$U_x = mn + \frac{m(m+1)}{2} - R_x \tag{2.1}$$

$$U_y = mn + \frac{n(n+1)}{2} - R_y \tag{2.2}$$

$$U = min(U_x, U_y) \tag{2.3}$$

- x and y are two populations
- m and n are the number of data points taken from populations x and y respectively
- R_x and R_y is the sum of ranks from populations x and y respectively
- U_x and U_y are the respective U statistics of two populations whereas U is given as the minimum of these two

2.2.2 Independent (Unpaired) t test

An unpaired t test is used to determine the statistical significance between two populations where the data points within those populations are unrelated. Firstly the means, and variances for both groups are calculated and the t-test formula (Equation 2.4) is used to calculate the t-statistic. This statistic is then compared to a

critical value ($p < 0.05$) to either accept or reject a null hypothesis that the difference between groups is large enough to be statistically significant. Similarly to the U test, a lower value rejects the null hypothesis, meaning that the samples are statistically different [122].

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \tag{2.4}$$

- x and y are two populations
- $\bar{x}$ and $\bar{y}$ are the sample means of populations x and y.
- s_x^2 and s_y^2 are the sample variances of populations x and y.
- n_x and n_y are the sample sizes of populations x and y.

2.2.3 χ^2 Test

The χ^2 (Chi^2) test is is a measure of statistical differences between two categorical variables, whether the frequencies of an event in those categories differ from expected frequencies, assuming the variables are independent. The null hypothesis is that there is no association between the two variables. For each test, the squared difference between observed and expected is taken, and divided by the expected frequency giving a χ^2 value for each observation (Equation 2.5). The sum of all χ^2 values is taken and compared to a critical value ($p < 0.05$), with degrees of freedom corresponding to the number of independent comparisons, to either accept or reject

the null hypothesis.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{2.5}$$

- i is the instance of an observation
- E and O are the expected and observed frequencies respectively.

2.2.4 Multiple Testing Correction

In cases where many null hypothesis tests are conducted, the false discovery rate or the chance of finding a significant value due to chance is increased. Therefore, in some cases, an adjusted p-value (q-value) was calculated using the Benjamini-Hochberg (BH) method [123].

- p-values are ranked in ascending order
- A BH critical value for each p-value is calculated using the following formula where q is the critical value, p is the current p-value, n is the number of samples, and r is the rank of the current p-value (Equation 2.6).

$$q = \frac{pn}{r} \tag{2.6}$$

- For each p-value and critical value, the largest p-value which is larger than its critical value is significant, and all others smaller than it are significant.

This method of controlling false discovery rate was chosen because it is less conservative than the Bonferroni method, allowing additional sensitivity to discoveries

and because it is easy to implement in Python.

2.3 Physicochemical Feature Representation

2.3.1 Identifying CDR-H3 Loops

Antibody binding has largely been attributed to the complimentarity determining region loop three of the VH domain (CDR-H3) because it is the single most diverse region between sequences, residing on the overlap of the Variable, Diversity and Junction gene segments [124, 125]. CDR-H3 regions were identified by numbering VH chains with AbNum and identified using the Kabat, Chothia or Martin definitions. This thesis used the Chothia numbering scheme definitions of CDR-H3 to capture the start (H95) and end (H102) of the loop.

2.3.2 Thermostability

Thermostability is an important element of developability because proteins with low stability are more likely to unfold and expose hydrophobic residues, leading to aggregation [126]. Gibbs' Free Energy (ΔG) of unfolding (henceforth called ΔG) is a measure of the feasibility of a protein spontaneously unfolding with higher values less likely to unfold than lower values. Although there are many *in silico* tools for prediction that rely on structural or machine learning models [127], these are unsuitable for high-throughput screening because of the compute power required to build models. Therefore, this statistic was calculated for each antibody sequence using the Oobatake method, where experimental values for ΔH and ΔS for the each residue are summed across a protein sequence and the free energy of unfolding

is calculated using Gibbs' Free Energy equation (Equation 2.7). ΔH and ΔS are derived from empirical data to model the hydration energy of different residues during protein unfolding, based on the surface accessible area of the residue. Taken together, this means that hydrophilic residues have a negative ΔH and ΔS because of the favourable reaction to hydrogen with water molecules, while the opposite is true for hydrophobic residues because of the unfavourable interaction of water molecules. Values for ΔG are given in kJ mol^{-1}

$$\Delta G = \Delta H - (T + 273.15) \cdot \Delta S \tag{2.7}$$

- ΔH is enthalpy calculated through summing experimental values for each residue across a sequence.
- T is temperature in Centigrade taken as 25°C as performed in the original paper [128].
- ΔS is entropy calculated through summing experimental values for each residue across a sequence.

2.3.3 Isoelectric Point

The Isoelectric point (pI) is the pH required for a net-zero charge of a polypeptide. It depends on the dissociation constant (pKa) of the seven charged amino acids (positive: R, H, K; negative: D, C, E, Y) as well as the N and C termini. The method of calculating pI was taken from the IPC software [129] which uses experimentally obtained peptide pKa values from the EMBOSS database [130] substituted

into a rearranged Henderson-Hasselbach equation. The equations are iterated using different values of pH, starting at 6.5, and the results of the termini and each of charged residues are summed together. If the sum is 0±0.01, an isoelectric point is reached, otherwise, the iteration continues either increasing the pH if the summed net charge was positive or decreasing the pH if it was negative. While this does not take into account protein structure, the IPC method was chosen because it calculates pI quickly and the code was integratable into the pipeline.

The rearranged Henderson-Hasselbach equation for negatively charged amino acids (Equation 2.8):

$$\sum_{i=1}^{n} \frac{-n}{1+10^{pKa-pH}} \tag{2.8}$$

The rearranged Henderson-Hasselbach equation for positively charged amino acids (Equation 2.9):

$$\sum_{i=1}^{n} \frac{n}{1+10^{pH-pKa}} \tag{2.9}$$

- i refers to an ionisable group of a molecule. For instance, a given amino acid in a sequence
- n is the count of that particular amino acid in a sequence

2.3.4 PTM Sites

Sequence recognition sites for PTMs are potential risk factors for poor developability and homogeneity and may increase the immunogenicity of antibody drugs.

Table 2.1: Regular expressions used to identify post-translational modifications in amino acid sequences

PTM	Regular Expression
N-Linked Glycosylation	"(N)(?=.[STC]"
Amidation	"(.)(?=G[RK][RK])"
Tyrosine Kinase Phosphorylation site	"[RK].{2,3}[DE].{2,3}(Y)"
Deamidation	"(N)(?=G)"
Asparagine deamidation	"(D)(?=[PSNHD])"
Methionine Oxidation	"(M)"
Hydroxylation	"(?<=C.)([DN])(?=....[FY].C.C)"
Protein Kinase C phosphorylation Site	"([ST])(?=.[RK])"
Protein Kinase CK2 Phosphorylation Site	"([ST])(?=..[DE])"
ATPase Site Phosphorylation	"(D)(?=KTGT[LIVM][TI])"
Aspartate hydrolysis	"(D)(?=P)"

Regular expressions taken from abYsis [112] (Table 2.1) and adapted from Vatsa *et al.* [131] and Xu *et al.* [132] were used to search for recognition sites for a selection of post-translational modifications and report where they occur within the antibody sequences.

2.3.5 Key residues

Work by Laffy *et al.* [133] proposed that antibodies with a propensity to form Beta pleated sheets in the CDR-H3 region were more likely to be promiscuous and therefore may have off-target binding or poor binding ability. They identified a set of residues within the region that would increase the likelihood of this occurring: 100 L; 100B D; 100C H; 100E W (Chothia numbering definition). These residues were identified within sequences from the numbering obtained from AbNum.

2.3.6 Cluster Residues

Surface clusters may increase the likelihood of protein aggregation and immunogenicity. ClusterResidues[3] was written by Andrew Martin and works to cluster residues on a given metric. The sequence is mapped to the structure of a reference antibody HyHEL-5 (PDB: 1yqv) [134] to calculate the solvent accessibility for each residue and the distances between each residue. The distance between residues to be considered a cluster, the number of residues to be considered a cluster, and the thresholds for solvent accessibility can be tuned to change what is considered on the surface of the molecule. In this case, groups of three or more residues with a relative solvent accessibility of $>25\%$ within 4.5A of each other were considered a cluster if they are hydrophobic or unusual.

2.3.6.1 Hydrophobic Clusters

The first characteristic calculated with ClusterResidues was hydrophobicity using scores for each residue calculated by Eisenberg [135]. The threshold for hydrophobicity is default set at >0.05 on the Eisenberg scale, however this could be tuned.

2.3.6.2 Unusual Clusters

The second characteristic reported with ClusterResidues was the clusters of unusual residues, which were defined as clusters of residues which occur less than 5% of precalculated frequencies from sequences stored in abYsis at those positions [112].

[3] `github.com/ACRMGroup/clusterResidues`

2.3.7 Solvent Accessibility

Solvent Accessibility was calculated using *pdbsolv*[4], which uses the Lee and Richards method of calculating Accessible Surface Area (ASA) using the rolling ball model [136] (Equation 2.10). *pdbsolv* gives an accessibility value as a relative percentage for each residue in a pdb file compared with a G-X-G extended peptide, where X represents that given residue.

$$Accessibility = A/4\pi r^2 \cdot 100 \tag{2.10}$$

$$A = \sum (R/\sqrt{R^2 - Z_i^2}) \cdot D \cdot L \tag{2.11}$$

$$D = \Delta Z/2 + \Delta' Z \tag{2.12}$$

- L is the length of the arc drawn on a given section
- Z is the perpetual distance from the centre of the sphere to the section i
- ΔZ is the space between the sections
- R is the radius of the sphere given by the sum of the Van der Waals radius of the atom

2.3.8 Germline Identification

Assign Germline (AGL) was used in cases where germline genes needed to be assigned to antibody sequences. The software uses germline DNA data obtained from IMGT [113] to align sequences and assign the relevant region gene. This operation

[4] www.bioinf.org.uk/software/bioptools

can be done for V, D, J, and C genes, but for the purposes of this thesis, it was used only for V region identification. It selects a gene using the logic that if two or more germline genes with the same sequence identity score were found for a given sequence at the protein level, the germline family with the lowest family number would be selected based on lower family numbers that were likely discovered first and therefore likely to be more numerous. AGL was developed by Andrew Martin and is available for download on GitHub[5].

2.4 Methods of scoring Antibody Immunogenicity

2.4.1 HScore

The HScore (Humanness Score) [108] was implemented by aligning an input sequence with all human Heavy, Light κ or Light λ sequences in the KabatMan database [114], producing a mean alignment score. Then the mean of means of all human:human alignments was subtracted from that value and divided by the standard deviation, to give a Z score of the similarity of that sequence to the others, which gives the final humanness score. A high HScore means that a sequence is more representative of human sequences than the average and is expected to be less immunogenic than a sequence with a low HScore, which corresponds to an antibody that is less representative of human sequences and is more likely to generate an immune response. This software is available online[6]

[5] github.com/AndrewCRMartin/agl
[6] www.bioinf.org.uk/abs/shab/

2.4.2 GScore

Secondly, the GScore (Germline humanness) [109] was developed using the same principles as HScore, but instead grouped sequences derived from specific germlines were grouped and Z scores were calculated for each germline using alignments from BLAST [137]. Newly inputted sequences are aligned against each germline V gene, and a GScore is given for each of these germline families ranked by smallest first, which is presumed to be the correct alignment. In cases where GScore was used, the highest ranking score was always taken. GScore is available online [7].

2.4.3 Hu-mAb

Hu-mAb is used as a comparison immunogenicity score, which instead calculates a score between zero and one [104]. Hu-mAb gives a score for each V-gene germline and if that score is above the threshold set by Youden's J statistic [138], then the program classifies it as human. For each set of results, the highest score was selected and assigned to each chain. Hu-mAb is available online [8].

2.5 Methods of Encoding Protein Sequences

2.5.1 Residue Level Encodings for Protein Sequences

Details of residue-level encodings length are given in Table 2.2.

[7] www.bioinf.org.uk/abs/gscore/
[8] opig.stats.ox.ac.uk/webapps/sabdab-sabpred/sabpred/humab

2.5.1.1 One Hot Encoding

One Hot Encoding is a binary representation of amino acid presence in a $n \times 20$ matrix. For each residue, 20 encodings are given where the amino acid present is coded as 1, and the other possibilities as 0.

2.5.1.2 One Hot 6

One Hot 6 is a binary representation of amino acid presence in a six bit matrix [139].

2.5.1.3 Binary 5-bit

Binary encoding method using five bit [140].

2.5.1.4 AESNN3

This is a three-dimensional protein encoding method resulting from a pre-trained machine learning algorithm [141].

2.5.1.5 Atchley Factors

Atchley Factors are a numerical encoding for each residue of five physicochemical properties properties: Bipolar, secondary structure, molecular volume, relative amino acid composition, and electrostatic charge [142].

2.5.1.6 Meiler Parameters

Meiler Parameters are another method of encoding residues using seven physicochemical properties: steric, polarisability, volume, hydrophobicity, isoelectricity, helix probability and sheet probability [143].

2.5.1.7 ANN4D

Dimensionality reduction from 7 physicochemical properties described by [143] to 4 composite values.

2.5.1.8 Kidera Factors

Kidera Factors are a set of 188 physicochemical properties of a protein sequence reduced to 10 numerical features through reduction in dimensionality [144]. The encoded information includes: Helix preference; side-chain size; extended structure preference; hydrophobicity; double bend preference; partial specific volume; flat extended preference; occurrence in alpha region; pK-C and surrounding hydrophobicity.

2.5.1.9 ProtVec

Tripeptide encoding of amino acid compositions from a machine learning encoder [145].

2.5.1.10 PAM250

PAM250 matrix compares similarities in protein sequences using comparisons of homologous aligned sequences [146].

2.5.1.11 Hydrophobicity Matrix

This method encodes each amino acid with a hydrophobicity score as described by [147].

Table 2.2: Details of Residue-Level Encoding methods.

Encoding Method	Encodings per residue	Encoding size	Description	Reference
Binary 5	n x 5	1270	Residue class binary encoding	[140]
One Hot 6	n x 6	1524	Residue class binary encoding	[139]
One Hot	n x 20	5334	Residue binary encoding	[151]
AESNN3	n x 3	762	Protein mapping representations	[141]
ANN4D	n x 4	1016	Per residue representations of physicochemical properties	[143]
Atchley Factors	n x 5	1270	Per residue representations of physicochemical properties	[142]
Meiler Parameters	n x 7	1778	Per residue representations of physicochemical properties	[143]
Kidera Factors	n x 10	2540	Eigenvector values of physicochemical properties	[144]
BLOSUM62	n x 20	5080	Matrix scoring residues based on observed substitutions	[148]
Hydrophobicity Matrix	n x 20	5080	Matrix scoring residues based on observed hydrophobic interactions	[152]
Micheletti Potentials	n x 20	5080	Matrix scoring residues based on potential energy interactions	[150]
Miyazawa Energies	n x 20	5080	Matrix scoring residues based on observed energies interactions	[149]
PAM250	n x 20	5080	Matrix scoring residues based on known sequence alignments	[146]
ProtVec	n x 3 x 100	25000	Per tripeptide representations of physicochemical properties	[145]
All Encodings	n x 260	65894	All given encodings methods conjugated	

2.5.1.12 BLOSUM62

BLOSUM62 matrices represent an amino acid substitution matrix, indicating the likelihood of each residue being substituted by all other possibilities [148].

2.5.1.13 Miyazawa Energies

These numerical encodings represent the interaction energies between every residue and every other possible residue [149].

2.5.1.14 Micehletti Potentials

Micheletti potentials are based on potential energy interactions for every residue for every other possible residue [150].

2.5.2 Amino Acid Compositions

Amino acid compositions are feature vectors quantifying the proportion of each amino acid relative to the sequence length for each residue, dipeptides and tripeptides using scores defined by Spanig *et al.* [153]. Sequence statistics were calculated using ProPythia [154] where V_H and V_L sequences were encoded separately using a set number of Propythia descriptors given in Table 2.3.

Table 2.3: Details of Propythia descriptors.

Number	Number of descriptors	Description
1	n x 20	Residue binary encoding
4	1	Net sequence charge
7	4	Sum of bond composition for each type of bond
10	1	Aromacity
11	1	Isoelectric Point
13	3	Fraction of residues which tend to be in helix, turn or shee
14	2	Molar extinction coefficient
15	1	Flexibility according to Vihinen *et al.* [155]
18	1	Hydrophobic ratio of sequence
20	20	Amino acid composition
21	400	Dipeptide composition
22	8000	Tripeptide composition
23	8420	All descriptors form amino acid composition
31	720	Normalised Moreau-Broto autocorrelation, Moran autocorrelation, Geary autocorrelation
33	343	Conjoint triad

2.6 Ellipse Function

The ellipse function takes in the points of the two extremes on the major axis ($x1$, $y1$) and ($x2$, $y2$) as well as a value for h, the height of the minor axis. The major axis is taken as the principal component where clinical mAbs have the largest distribution, and the selected points are given as the points on the distribution closest to a given Z score in that distribution. The value of h is given as the distance between two points on minor axis. The method for producing the ellipse works as shown in Algorithm 1.

Algorithm 1: Method of Drawing Ellipse Given Coordinates of Major and Minor Axes.

- Calculate the major (a) and minor (b) radii of the ellipse. The major radius is calculated from the two given points (Equation 2.13) and the minor radius is calculated as half the value given for h. where Δx is the difference in x values and Δy is the difference in y values between two extreme points on the major axis.

$$a = \frac{\sqrt{\Delta x^2 + \Delta y^2}}{2}, b = \frac{h}{2} \tag{2.13}$$

- Use the parametric equation of an ellipse to generate the ellipse over 100 equally spaced points between 0 and 2π assuming it is centred at the origin (Equation 2.14).

- For a given point on the ellipse, the equation is:

$$x = a\cos(\theta), y = b\sin(\theta) \tag{2.14}$$

- Where a is the major axis radius, b is the minor axis radius and θ is a given angle between 0 and 2π

- Calculate the angle between given points to obtain angle of rotation using the Numpy arctan2 function for Δy and Δx [156].

- Calculate a rotation matrix (R) based on the angle of rotation (Equation 2.15)

$$R = [[cos(\theta), -sin(\theta)], [sin(\theta), cos(\theta)]] \tag{2.15}$$

- Where θ is the angle of rotation

- Apply the rotation matrix to the ellipse R

- Calculate the midpoint of the two given points (Equation 2.16)

$$x = \frac{x_1 + x_2}{2}, y = \frac{y_1 + y_2}{2} \tag{2.16}$$

- Translate the ellipse to the midpoint

- For each point, check if its x and y coordinates are inside the ellipse using the *Polygon* function

2.7 Introduction to Machine Learning

In an environment of high density data collection, it is useful to make meaningful interpretations of said data and to use it in making predictions by fitting or training predictive models on a given dataset. The advantage here is that not only are models more likely to learn patterns from datasets that are too subtle or too complex for human analysis, but that predictions can be made on unseen data that have yet to be generated, and data may be added to the training dataset to continuously improve future predictions [157, 158].

Generally, classification machine learning is split into supervised and unsupervised methods which are dependent on the data fed into the model. Supervised methods are usually used in cases where data is labelled (i.e. there are discrete categories to which the data points are assigned), and predictions are made to assign categories to new data points dependent on their features. Unsupervised learning is used mainly when there are no known categories in the data and the desired outcome is for the model to separate the data into categories. This could be done by looking at a feature or combination of features which differ between groups of data points or if these cannot be identified, reducing the dimensionality of the data by combining or eliminating features in the data.

The measurement of the difference between ideal and observed outputs is referred to as a loss functions. Different functions such as mean squared error and binary cross entropy can be applied depending on the question trying to be answered, and whether it is classification into discrete classes, or prediction of a continuous variable. The aim of training the model over successive iterations is to minimise

these loss functions so that performance is improved.

Models are all fitted with a set of parameters which can also be manipulated to improve predictions, however, there is a trade-off in how well a model may train on a given dataset. Models may overfit to the training data if their parameters are specifically tuned to the data, which then means they will make poor predictors on unseen data. If this is the case, using simpler, or more general models, with default parameters is probably a better approach. While training for good performance is strived for, it is unlikely that machine learning models will show perfect predictive ability. Cross-validation (CV) is a method of resampling the training dataset into a series of different training and testing datasets to verify the model's performance. An example of this is 'Jackknife' sampling, or 'leave one out CV' where for every data point, it is left out of the training dataset and predicted using on each other data point. This approach can be useful when using small training datasets. Otherwise, k-fold CV can be used for larger datasets.

By nature, biological data can be stochastic and difficult to interpret directly. With this in mind, it is a useful strategy, at least at first, to evaluate a number of models on the data and to then improve the selection seen on those which perform best by hyperparameterisation. In the following sections, a number of supervised and unsupervised machine learning models used in this thesis are outlined. The majority of these models were implemented through the *Scikit-learn* Python module [159]. To describe how modules were implemented, their location in the module has been provided.

2.8 Supervised Machine Learning

As explained above, supervised machine learning depends on labelled data from which to correlate data features. It is possible for binary or multinomial classification problems to be solved by supervised learning. In general, supervised models are trained by evaluating a model's predictive performance to categorise labelled data. At first, the model's guesses would be expected to be random and the loss to be high. In iterative training rounds, the weights are adjusted to give better performance, which should minimise the loss to an acceptable level over several interactions.

2.8.1 Supervised Machine Learning Classifiers

2.8.1.1 Logistic Regression

A linear model that can be applied to simple classification problems by calculating a predicted probability using a combination of features of an input data point and assigning an input data point to a class using a threshold value typically set at 0.5. Logistic regression was implemented through `sklearn.linear_model.LogisticRegression` with default parameters.

2.8.1.2 Decision Tree

Decision Trees are a method of simple classification which are able to learn a set of rules as to how a data point should be classified using its features. These rules are established by finding a split point that minimises the mean squared error for the new partitions [160]. This tree is arranged using these rules with a set of nodes along the tree, where a dichotomous decision is made at every node. Once trained, a

new data point starts at the root node and will move along the tree, where its path is decided by the already learnt set of rules, to a terminal node, a 'leaf' node, where its classification is assigned. Decision Trees can become overly complex and overfit to the training data, and so a number of pruning parameters are set where nodes with a small number of training samples are removed. A Decision Tree Classifier was implemented through `sklearn.tree.DecisionTreeClassifier` with default parameters.

2.8.1.3 Extra Trees Classifier

This method also includes decision trees, but also uses a number of additional randomised trees on subsamples of the data to improve accuracy and reduce overfitting changes. The Extra Trees Classifier was implemented through `sklearn.ensemble.ExtraTreesClassifier` with default parameters.

2.8.1.4 Random Forest Classifier

Forests are collections of decision trees trained on bootstrapping samples of a training dataset with replacement and additionally, a random sampling of the features of that dataset. The result of input data points is averaged across the collection to overcome the overfitting problem usually encountered in decision trees through a user-curated set of trees where the number of trees and the pruning of trees can be tuned [160] (Figure 2.1). A Random Forest Classifier was implemented through `sklearn.ensemble.RandomForestClassifier` with default parameters.

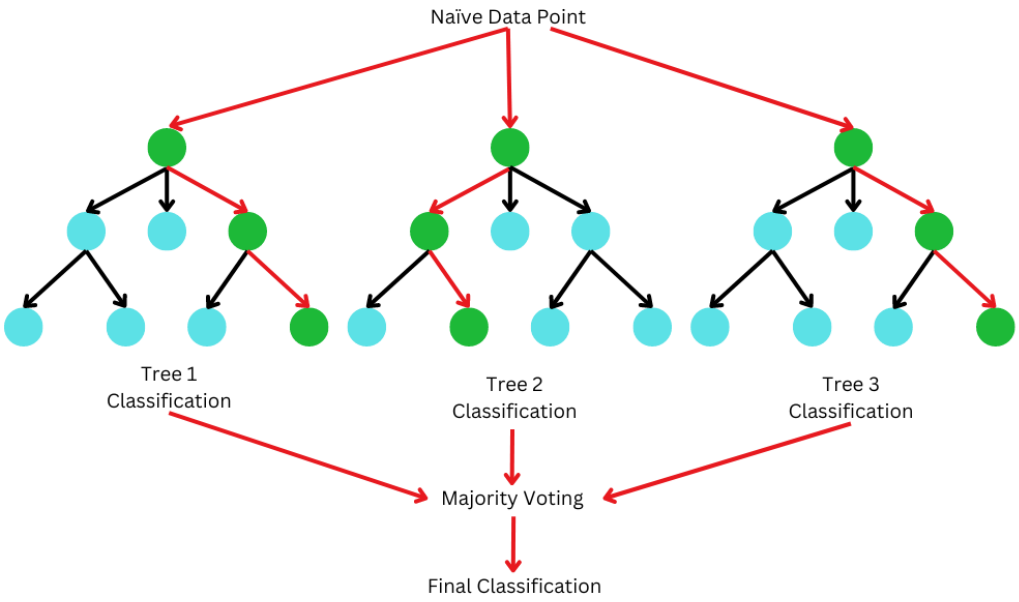

Figure 2.1: Schematic of Random Forest Classifier. Three decision trees, each with a maximum depth of two layers of nodes. A naïve data point is shown taking different paths through the tree, depicted by red arrows and nodes highlighted in green, and arriving at a leaf where a classification decision is made for each tree because each tree is trained on a different sample of the training data. All classifications are taken and a majority vote is taken giving the final classification.

2.8.1.5 Gradient Boosting Classifier

In cases where Random Forests do not effectively learn the patterns in the data, many weak learner decision trees with a single node, 'stumps', can be combined sequentially into an ensemble that collectively makes better predictions. This is known as boosting. It is performed by adding models sequentially to an ensemble of predictors so that the combined loss of all models is lower than the previous iterations of the model. Unlike Random Forests, this model uses gradient descent, an optimisation technique to minimise the loss function. A learning rate is chosen

as to define how much each tree should contribute to the answer. Usually a lower learning rate requires more trees, however, this could lead to overfitting the model. Gradient Boosting was implemented through `sklearn.ensemble.GradientBoostingClassifier` with default parameters.

2.8.1.6 AdaBoost Classifier

Adaptive Boosting (AdaBoost) works on similar principles to the Gradient Boost Classifier, but here, the model continues to correct previous iterations by placing more attention to underfitted results of the previous model. AdaBoost was implemented through `sklearn.ensemble.GradientBoostingClassifier` with default parameters.

2.8.1.7 Stochastic Gradient Descent Classifier

The Stochastic Gradient Descent (SGD) Classifier uses sampled data to improve gradient descent rather than using all the data. An SGD Classifier was implemented using `sklearn.linear_model.SGDClassifier` with default parameters.

2.8.1.8 Ridge Classifier

A Ridge Classifier is a linear classifier model with a penalty on large coefficients. Regularisation encourages a more stable and generalisable model. The model minimises an objective function that combines a loss term and regularisation, making it effective for handling correlated features. A Ridge Classifier was implemented using `sklearn.linear_model.RidgeClassifier` with default parameters.

2.8.1.9 Ridge Classifier CV

This is similar to the normal Ridge Classifier, but performs leave-one-out cross-validation. This CV training allows for a more dynamic hyperparameter tuning for learning rate and gradient. The Ridge Classifier CV was implemented using `sklearn.linear_model.RidgeClassifierCV` with default parameters.

2.8.1.10 Bagging Classifier

A Bagging Classifier also works similarly by combining weak learners where their voted, or mean score, forms a strong ensemble through random sampling with replacement from the dataset, known as 'bagging'. A Bagging Classifier was implemented through `sklearn.ensemble.BaggingClassifier` with default parameters.

2.8.1.11 Calibrated Classifier

A Calibrated Classifier aims to reflect the probability of an outcome in a class and to measure the confidence of that prediction. Unlike other classifiers that output a probability to assign classes, the probabilities are first calculated using a base estimator, another supervised learning model which is used to train on the data, and then calibrated according to the frequencies of each class in the dataset. A Calibrated Classifier was implemented through `sklearn.calibration.CalibratedClassifier` with default parameters where the base estimator was a Linear Support Vector Machine Classifier.

2.8.1.12 Gaussian Naïve Bayes Classifier

Gaussian Naïve Bayes follows an assumption that each feature has an independent ability to predict an output label, and seeks to combine all of these predictive powers into the final model. It also assumes that all features follow a normal distribution and that Bayes' Theorem of conditional probability can be applied to each feature by the way of a probability density function. Gaussian Naïve Bayes was implemented through `sklearn.naive_bayes.GaussianNB` with default parameters.

2.8.1.13 Bernoulli Naïve Bayes Classifier

Bernoulli models are designed to make predictions based on binary data values and are usually applied to text recognition problems. A Bernoulli Naïve Bayes model uses the probability of observing the set of features, given that class to assign a class depending whether the probability calculated reaches a desired threshold. This was implemented through `sklearn.naive_bayes.BernoulliNB` with default parameters.

2.8.1.14 Support Vector Machine Classifier

Support Vector Machines work by representing a dataset in a highly dimensional space using a kernel and work to separate the two classes of labelled data using a hyperplane. Using the data points of class 0 and class 1 that are physically closest to the hyperplane, the support vectors, the maximum marginal space between the two classes [161] using a special loss function called the Hinge Loss. Although these machines may follow linear or polynomial functions depending on what is best suited to the data, this current function was set as a radial basis function (Figure

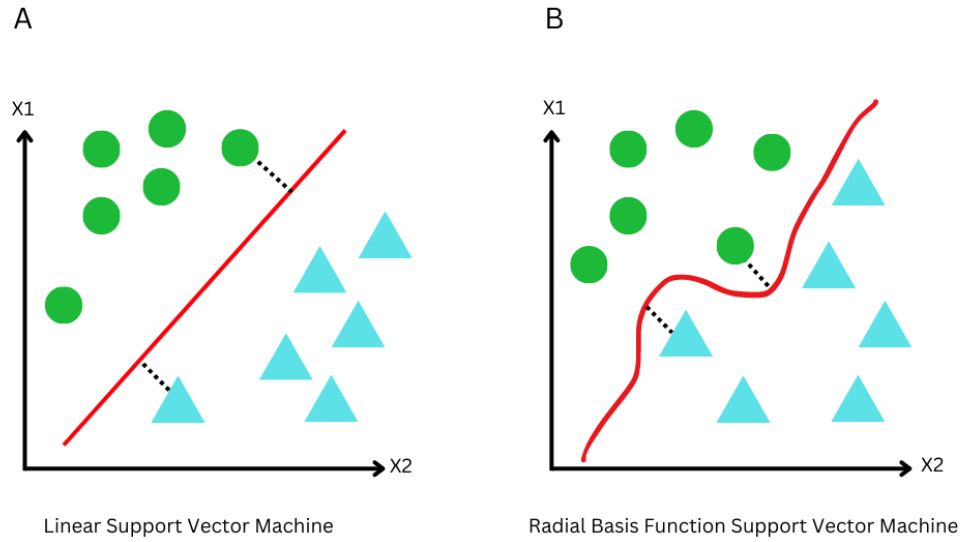

Figure 2.2: Schematic of Support Vector Machine Classifier. The separation of two data classes (green circles and blue triangles) by support vector machines can be achieved using a kernel (red line) with (A) a linear or (B) an exponential function such as a radial basis where the marginal space is maximised by the support vectors distance to the kernel (dashed lines).

2.2). The SVM classifier was implemented through `sklearn.svm.SVC` with default parameters.

2.8.1.15 Linear Support Vector Machine Classifier

LinearSVC is a similar algorithm to SVC but uses a linear kernel rather than a polynomial one. LinearSVC was implemented through `sklearn.svm.LinearSVC` with default parameters.

2.8.1.16 Linear Discriminant Analysis Classifier

Discriminant analyses work to separate data points by class and calculate descriptive statistics for each feature for each class. It is a more simplistic model where a new datapoint is classified based on conditional probability, calculated by Bayes' theorem, of it belonging to each class and selecting the class with the highest probability. It assumes features are independent of each other. Linear Discriminant Analysis works by firstly calculating a covariance matrix, and computing the discriminant score for a given observation using the covariance to classify new observations. This model was implemented through `sklearn.discriminant_analysis.LinearDiscriminantAnalysis` with default parameters.

2.8.1.17 Quadratic Discriminant Analysis Classifier

Like Linear Discriminant Analysis, Quadratic Discriminant Analysis aims to classify samples using conditional probabilities, however, using quadratic equations allows the use of a curved hyperplane to separate datapoints. This model was implemented through `sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis` with default parameters.

2.8.2 Linear Regression

When a calculated value is the desired outcome, rather than classification, linear regression is a regressive model that aims to fit a line of best fit to the data which minimises the difference, calculated as the sum of squares, between the actual values and the predicted values of the best fit. Linear regression was implemented through `sklearn.linear_model.LinearRegression`. Where possi-

ble, linear regression was trained in a Jackknife fashion. Other kinds of regressions models can be used, such as using polynomial or logistic functions, but these were not used for these problems in favour of a more simple regression method.

2.9 Unsupervised Machine Learning

Using unsupervised learning is a good strategy to overcome the lack of labelled data in a training dataset. Instead of relying on labels to train a model iteratively to these outcomes, unsupervised learning relies on clustering data using similarities in their features. This is done either through generating a pairwise matrix of correlation coefficients for each datapoint against all others in the case of dimensionality reduction or through a Euclidean distance between points in the case of K Nearest Neighbours. This section will outline the models used throughout this thesis.

2.9.1 Principal Component Analysis

When working with high dimensional data, it is probable that many dimensions will be irrelevant and misleading to classical supervised machine learning models. PCA works to combine features linearly into vectors, or principal components, that aim to explain the variance within the observed dataset by calculating the covariance between all of the features. The features are then reduced in an eigenvalue decomposition where if the covariance of two particular features is positive, then the features are combined together. The resulting principal components are sequentially ordered where PC1 is the vector that explains the most variance in the data, followed by PC2 and so on for the top k PCs that are desired (i.e. the eigenvector with the highest eigenvalue). This method is useful for visualising grouping in these data [162, 163]

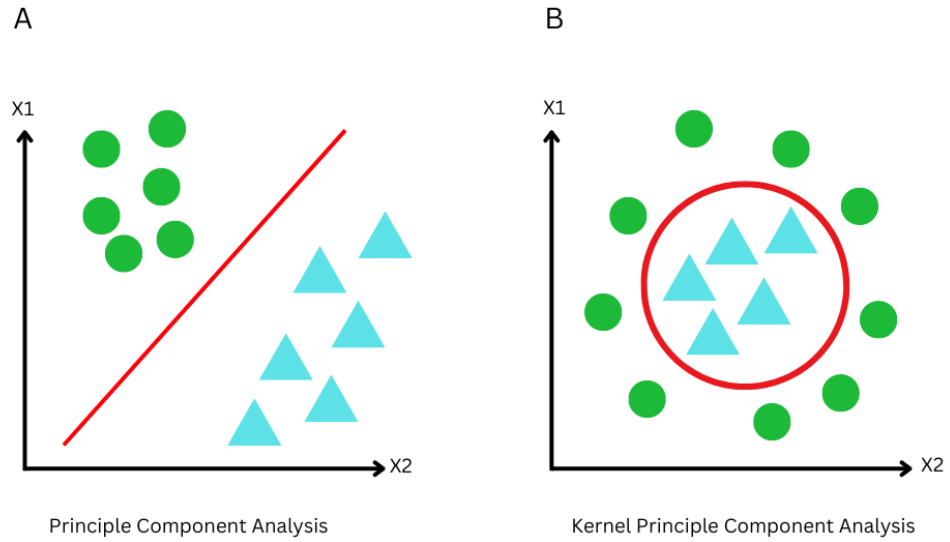

Figure 2.3: Schematic of Principal Component Analysis. The separation of two data classes (green circles and blue triangles) through principle component analysis can be achieved using a kernel (red line) with (A) a linear or (B) an exponential function such as a radial basis.

(Figure 2.3). PCA was implemented through `sklearn.decomposition.PCA`.

2.9.2 Kernel PCA

This is like PCA, however, this method uses non-linear dimensionality reduction to cluster data using either radial basis functions or sigmoid kernels. This is useful in cases where the given data are not linearly distributed [164]. A kernel matrix is calculated, which represents the pairwise similarity for the input data, and then a similar eigenvalue decomposition is performed on the data to lower its dimensionality. The eigenvectors are then sorted by their eigenvalues. KPCA was implemented through `sklearn.decomposition.KernelPCA`.

2.9.3 t-SNE

t-Distributed Stochastic Neighbour Embeddings is a non-linear data clustering algorithm that also reduces the dimensions of a dataset. It works by calculating pairwise similarities between datapoints, and minimising the difference between the similarities in the high-dimensional input and the low-dimensional output space iteratively [165]. t-SNE was implemented through `sklearn.manifold.TSNE`.

2.9.4 UMAP

Uniform Manifold Approximation and Projection (UMAP) instead uses a nearest neighbour graph based on the high-dimensional input, preserving the layout while minimising the cross-entropy between the pairwise similarities to give a lower-dimensional output. Unlike t-SNE, because UMAP places more emphasis on maintaining the layout of the nearest neighbours and the pairwise similarities, it can capture more complex grouping, so it is generally preferred for large datasets [166, 167]. UMAP was implemented through `sklearn.manifold.UMAP` with default parameters.

2.10 Methods of scoring and enhancing model performance

2.10.1 Evaluation Scoring

For supervised machine learning models, evaluations of accuracy can include a raw percentage of how many test data are categorised correctly as true positives (TP) and true negatives (TN), but this does not take into account false positives (FP) and false

negatives (FN), so for additional appreciation of how the model is learning, other statistics are used, especially in cases where false positives become important.

2.10.2 Sensitivity and Specificity

Sensitivity (Sn) and specificity (Sp) of machine learning classifiers are a measure of how many true positives are captured by the prediction (Equation 2.17), and how many true negatives are correctly captured by the prediction (Equation 2.18). A well-predictive model should have high sensitivity as well as high specificity.

$$Sn = \frac{TP}{TP+FN} \tag{2.17}$$

$$Sp = \frac{TN}{TN+FP} \tag{2.18}$$

2.10.2.1 False Omission Rate

False omission rate (FOR) quantifies how many false negative results are given by the model (Equation 2.19), normalised for the number of total negative predictions. A well-predictive model should have a low false omission rate.

$$FOR = \frac{FN}{FN+TN} \tag{2.19}$$

2.10.2.2 Negative Predictive Value

Negative predictive value (NPV) also quantifies how many true negative results are given by the model (Equation 2.20), normalised for the number of total negative

predictions. A well-predictive model should have a high negative predictive value.

$$NPV = \frac{TN}{FN + TN} \tag{2.20}$$

2.10.2.3 Pearson's Correlation Coefficient

The Pearson's correlation coefficient (r) is a measure of linear correlation between two sets of data (given as x and y) demonstrating the ability of a regression model to predict observed values (Equation 2.21) with a score between -1 and 1. A highly predictive model will have a correlation coefficient close to 1.

$$r = \frac{\sum(xy) - \sum x \sum y}{\sqrt{[\sum x^2 - (\sum x)^2][\sum y^2 - (\sum y)^2]}} \tag{2.21}$$

2.10.2.4 Spearman's Rank Correlation

Spearman's Rank correlation (ρ) demonstrates the correlation between the order of two datasets. This is useful for evaluating regression models where the precise predicted value is less important that the order of the sorted values (Equation 2.22). A highly predictive model will have a correlation coefficient close to 1.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{2.22}$$

- d_i refers to the difference between the two ranks for each observation

2.10.3 Mean Absolute Error

Mean absolute error (MAE) is a measure of the mean difference, whether positive or negative, between two sets of paired data, for instance measuring error in predicted and observed datasets (Equation 2.23).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.23}$$

- y_i refers to the predicted value
- $\hat{y}_i$ refers to the true value

2.10.4 Matthew's Correlation Coefficient

While accuracy is probably the most basic statistic that simply expresses the number of correct predictions of a test data as a score between 0 and 1, for large or unbalanced datasets, this statistic can be misleading. Take an example of a dataset with 90% data points belonging to one class and 10% to another class, if the model were simply to predict all data points as instances of the majority class, it would still have an accuracy of 0.9 [157], which is not reflective of the true performance.

Therefore, in this work, Matthew's Correlation Coefficient (MCC) is used to evaluate model performance [168] (Equation 2.24). This overcomes the previously mentioned problem with raw accuracy as it takes into account false positives and false negatives, so it is felt that MCC is better able to capture how a model understands the training data, and not just how much its predictions match what is seen in the training dataset [169]. MCC is implemented through `sklearn.metrics.matthews_corrcoef` and gives a score between -1

(inverse prediction) and 1 (perfect prediction) with 0 being random chance.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{2.24}$$

2.10.5 *k*-fold CV

Well known model benchmarking methods have been applied in order to give a more robust evaluation of their performance. *k*-fold cross-validation aims to train the model on samples of the training data by splitting it into *k* equal splits where, in turn, each split is held back and the model is trained on the remaining splits and tested on the held back split. The advantage of this strategy is to gain an understanding of how the model will perform on unseen data. If consistent performance is seen for all cross validations, it is indicative that the model will cope with new data and the reverse is true for inconsistent performance [170]. *k*-fold CV was implemented using `sklearn.model_selection.cross_val_score` using MCC as the performance metric.

2.10.6 Grid Search CV

Tuning the hyperparameters of a model is a method to increase its performance. Grid Search CV systematically retrains a model for each combination of a set of defined values for provided hyperparameters. The model found to be best performing on the given test dataset is kept with the tuned hyperparameters. Whilst this model can work for the given dataset, it is not reflective of the model performance on unseen data because this method can lead to overtraining, and so a held-back test dataset independent of the original training data is often useful to mark true

performance of hyperparameter tuning. Grid Search CV was implemented using `sklearn.model_selection.GridSearchCV` using MCC as the scoring function with 5-fold cross validation.

2.11 Feature Selection

2.11.1 F-Regression

Rather than combining features as seen in unsupervised learning models, it is also possible to increase model performance by selecting the most relevant values. F-regression calculates the cross-correlation of each data point and the label for all features (the raw encoding space), which is then converted to an F-score and then to a p-value. Features are then ranked by F-score, and correlation to the target. The top k features are then selected, where k is a number selected by the user [171]. F-regression was implemented through the module `sklearn.feature_selection.SelectKBest` using `sklearn.feature_selection.f_regression` as the score function and the variables for k were substituted.

2.11.2 Identifying the Position of Correlated Features

When using F-regression, it was necessary to understand which residues have features were selected across the antibody V_H or V_L sequences. This can be done when the number of features given per residue by an encoding method are consisted for all residues and the length of the V_H and V_L sequences are known. To identify the residues where a feature selected by F-regression was selected, the method is given in Algorithm 2:

Algorithm 2: Method of Identifying the Residue where a F-Regression Feature has been Selected.

- The index of each selected feature within the whole encoding space (i) was increased by one and divided by the number of encodings per residue (N) to give the normalised index (I') across the conjugated and spaced V_H and V_L sequence (Equation 2.25)

$$I'_i = \frac{i+1}{N} \tag{2.25}$$

- Indices were rounded down to nearest integer to give the index of the residue in the conjugated and spaced $V_H$and $V_L$sequence (Equation 2.26).

$$I_i = \lfloor I'_i \rfloor \tag{2.26}$$

- Indices less than or equal to 133 were assigned to the V_H chain whereas indices greater than 133 were assigned to the V_L chain. To obtain the light chain index, the index was subtracted by 132 (Equation 2.27). This gives the position of the residue in the numbered antibody V_H or V_L sequence (I_{VH} or I_{VL}) relating to the selected encoding (i).

$$if \;\; I \leq 133, \qquad I_{VH} = I \tag{2.27}$$
$$if \;\; I > 133, \quad I_{VL} = I - 132 \tag{2.28}$$

2.12 Deep Learning

Deep Learning is a type of Machine Learning that involves much more complex models, usually artificial neural networks (ANNs) [172]. These models consist of layers of nodes, which can also be called nodes, taking in multiple inputs through previous layers and outputting a new value to new layers. At each node, weights calculated during model training are applied to the inputs (Equation 2.29), and the output then becomes the input to the nodes of the next layer (Figure 2.4). Many of the hyperparameters of a model can be tuned including the number and size of hidden layers between the input and output, activation functions at each layer, which can tune a model between classification and regression tasks, the loss function the

model will use to measure training performance, and the maximum number of iterations the models can train. These can change the architecture of the model and therefore can become better suited to certain tasks such as more hidden layers and different connection patterns between layers.

$$f(\sum_{i=1}^{n} x_i w_i) \tag{2.29}$$

- f is the activation function
- x_i represents a single input with a total of n inputs
- w_i represents a weight applied to that input

Although ANNs can be applied to predictive and regressive tasks, they also have the ability to encode data, such as in the case of autoencoders. These are networks with an encoder and decoder module and work such that a dataset may be input and encoded into a new representation with a different dimensionality and then decoded such that the original dataset may be obtained from the encoded representation. Autoencoders demonstrate this ability for ANNs to effectively learn highly dense representations of input data [173] (Figure 2.5). Furthermore, these modules may be separated in that the embeddings made by the encoder, may be used in that of themselves to encode data for use to train other machine learning models.

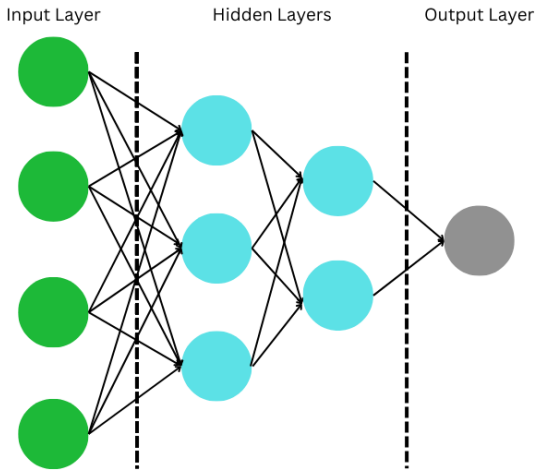

Figure 2.4: Schematic of Artificial Neural Network. A simple artificial neural network schematic with one input layer with 4 nodes, two input layers with 3 and 2 nodes respectively and one output layer of 1 node. Such models can be used for either classification or regression tasks depending on the optimiser used at the output layer.

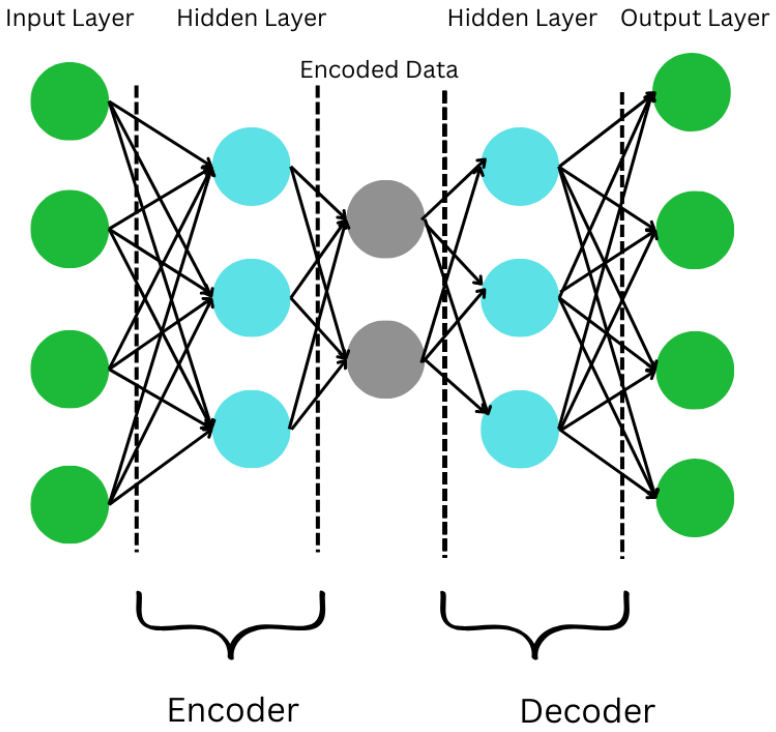

Figure 2.5: Schematic of autoencoder. The autoencoder shown takes in four input datapoints, and through one hidden layer, encodes the input at a lower dimensionality, shown as two data points. The decoder also transforms the encoded data back into the original datapoints. This encoder demonstrates a symmetrical encoder and decoder modules that may be used independently to encode the data for other uses.

2.13 Protein Language Models

Deep-learning Large Language Models (LLMs) are additional methods of encoding text-based information in a numerical format so that this data type can be applied to machine learning. LLMs have emerged from the field of natural language processing, specifically, how computers understand human speech by recognising patterns in text and generating sentences by predicting what should follow. These methods are based on autoencoders where complex input data can be encoded into numerical embeddings, and then using these embeddings, a response may be generated, or the embeddings may be used as a representation of the input data to be used for different tasks. Through encoding text into numerical formats, such as each word being represented by a number that is correlated with it (i.e. a token), these principles can be applied. While these techniques have become useful in the fields of text summarisation, translation, and sentiment analysis, LLMs can also be used to solve specific problems when trained on a curated dataset. Normally, they are trained on masses of text data from a number of sources. However, from the previous information it can be inferred that if an LLM is trained using protein sequences, it will become proficient at representing protein sequences, and this can also be restricted to particular classes of proteins, such as antibodies [174, 175, 176].

2.13.1 Training Antibody Large Language Models

The process for protein LLM training is much the same where amino acids are treated as words in a sentence. Before training a LLM using sequences using the V_H and V_L sequences, these sequences need to be numbered or padded with a given

padding character depending on the LLM, to make all input V_H and V_L sequences the same length as each other. This is because the shape of the output data is dependent on the shape of the input data and with antibody sequences being different lengths, it would be difficult to train a dataframe where datapoints had different numbers of features. The sequences of amino acids are tokenised, where a substitute digit can be used to replace the residue letter, to enter the data into the model.

Once the training data has been tokenised, a selection of architectures can be selected. The most reliable architecture chosen has been based on transformer architecture, which are modules able to apply attention that captures relationships between positions in the training data if it is seen consistently. This attention can be expressed as the relationship between a Query, Key and Value embeddings, representing the token, the context and the value of the tokens attended to. The similarity between query each query embedding and all key embeddings as a dot product, producing a matrix of scores informing how much attention the query token should pay attention to a particular key token.

In a given LLM, AntiBERTy [177], a self-attention mechanism is given, meaning that Query, Key and Value come from the same input sequence and relationships are captured from across the sequence. There is great importance in attention when training an antibody LLMs as the majority of antibody sequences share similar frameworks, and the main regions of diversity are seen at the CDRs and so the most attention is designated to these regions. The number of features that the model generates per token is also dictated by attention mechanisms and the architecture of the model (i.e. the number of nodes that output the encoded data), by splitting

attention into heads, which encode different information based on the Query, Key and Value embeddings. In the case of AntiBERTy where the size of the embeddings is 512, there are 8 heads, each head works to calculate 64 embeddings per token and the outputs are combined to give the full embedding space [177].

Language models are trained through a process called 'masking' in which a small number of tokens in input data are masked so that the model is made to predict what should be in its place. As the model retrains and its predictions are closer to the masked value (i.e. the loss function decreases over successive rounds of training), it is thought that the model gains a better understanding of how the words and sentences are structured to predict text. Furthermore, transformers are modules that may apply an enhanced focus on some aspects of the input data and diminish other aspects through dynamically generating a proportionate number of features to represent them depending on how important those features are in distinguishing data points. Because of the mass and complexity of the data these models are trained on, they may have millions or billions of parameters and require high compute power to implement, however, when LLMs are applied to specialists tasks like antibody or protein sequences, the demands for compute are not as high as for general purpose language models.

2.13.2 AntiBERTy

AntiBERTy [177] is based on the 'bidirectional encoder representation from transformers' (BERT) architecture which relies on each output feature representing each input feature and so learns representations of bidirectional data not just sequen-

tial, which means it can learn more meaningful representations of the input data through contextual information [178, 172]. It is trained on a dataset of 558 million non-redundant antibody sequences from the OAS representing: human, rat, mouse, camel, rabbit and rhesus monkey. A multiple instance framework was also used to identify antigen-binding residues and enhance attention onto the CDR loops of the antibody to increase the context-derived information available for the model to train on. For this reason, AntiBERTy is useful for antibody modelling and predicting binding. Implementation of AntiBERTy used for this project was 'model 1' that was included in the code of the Igfold software [179] to represent antibodies for structural prediction.

2.13.3 AbLang

AbLang is another LLM trained on 14,126,724 V_H sequences and 187,068 V_L sequences that were also taken from OAS [33]. By masking 70% of sequences during its training, it has been well trained to predict missing residues in antibody sequences resulting from sequencing errors. This has clear applications in database curating and antibody design. AbLang was used with the `ablang` Python package.

2.13.4 Sapiens

Another LLM is Sapiens, which is trained on 562,544,071 V_H and 55,826,963 V_L sequences also taken from the OAS, but with the use case of humanise antibodies, fill missing residues, and create feature embeddings [180]. This model can be accessed using the `sapiens` Python package.

2.13.5 ESM

ESM-2 is a transformer-based general protein language model with adjustable numbers of parameters ranging from 8 million to 15 billion depending on the complexity of the task at hand. The training data was an unspecified number of sequences from the UniRef database by predicting masked amino acids to learn sequence and structural data representations for protein modelling [176]. The iteration of the model used here included 8 million parameters, as found in the ESM tokeniser `facebook/esm2_t6_8M_UR50D` as this lightweight version of the model was faster to run on a CPU machine. This model can be accessed using the `esm` Python package.

2.13.6 Use of LLMs in this Thesis

To enter antibody sequences into the LLMs, throughout this thesis, sequences were made to a standardised format. This was done by numbering sequences using the Chothia numbering scheme, conjugating the V_H and V_L chain and padding missing residues with a padding character allowed by the LLM on an individual basis.

The LLMs used here included: AntiBERTy [179], AbLang [33] and Sapiens [180], which are all antibody-specific models, and ESM [176], which is a general protein language model. For each language model, V_H and V_L sequences were individually spaced according to the Chothia numbering scheme, where the padding character differed for each model according to its specifications. The encoding size observed for each language model using the dataset, and padding characters are given in Table 2.4.

Table 2.4: Details of language model encodings.

Language Model	Features ($V_H + V_L$)	Padding Character	Reference
AntiBERTy	130,048	"_"	[179]
AbLang	195,072	"*"	[33]
Sapiens	152,560	"*"	[180]
ESM (esm2_t6_8M_UR50D)	82,560	"X"	[176]

Chapter 3

Separating Clinical and Repertoire Antibodies to Identify Repertoire Antibodies with Clinical Potential

3.1 Introduction

In order to generate triaging criteria for separating clinical mAbs and human antibodies, several approaches were taken that are outlined in this chapter. This was partly inspired by the work in Negron *et al.* [111] that worked to solve the same problem using sequence statistics to train linear models to predict a score based on similarity to clinical stage mAbs.

This chapter concerns identifying clinical candidates from repertoire data through comparing antibodies from available human repertoire data with a dataset of human clinical mAbs. The first approach taken here included physicochemcial property-based filtering, however, better success was seen with machine learning.

This involved encoding these sequences with antibody language models and using machine learning models to classify groups in the data. Firstly, supervised machine learning using a voting function was used, but because the repertoire data is unlabelled, an unsupervised approach was seen as more appropriate. When this unsupervised approach was implemented, a number of methods were tested to extract antibodies which clustered closely with the clinical mAbs.

3.2 Datasets and Models

3.2.1 Human Repertoire Data

All human paired V_H and V_L sequences were downloaded from Observed Antibody Space [33] (accessed January 2022, Appendix A.1) and assimilated into one file consisting of 88,274 paired human sequences. Randomly selected paired sequences (n = 10,000) were collected into one file (Data File 1).

3.2.2 Human Clinical-Stage mAbs

TheraSabDab was accessed in October 2021 and filtered by Format 'Whole mAb'. Human antibodies were identified using the '-umab' suffix, and checked for source with a literature search. The remaining antibodies were then sorted using the 'Highest_Clin_Trial (Oct '21)' field for "Approved" (n = 31) (Data File 2); "Discontinued" (n = 77) (Data File 3) or in any phase of clinical trials ('Phase I', 'Phase-II', 'Phase-III') where Est. Status was 'Active' at the time of access (n = 35)(Data File 4), giving a total of 143 clinical stage antibodies. A further held back dataset of human-derived clinical mAbs was acquired (n=203), using the 2022 naming con-

Table 3.1: Means and standard deviation of sequence-calculated physicochemical properties for fully human mAb therapeutics (n=144) and library human antibodies (n=10,000).

Feature	Human Therapeutic mAbs	OAS Library Antibodies	p value
CDR-H3 Loop Length	12.08±6.65	15.01±10.54	0.00049
ΔG V_H kJ mol^{-1}	7614±3260	6583±3441	0.00014
ΔG V_L kJ mol^{-1}	1086±2381	796±2614	0.14
ΔG Concatenated V_HV_L kJ mol^{-1}	9248±3896	7944±4238	0.00015
Mean pI of V_HV_L	7.87±1.30	7.8±1.24	0.025

vention using '-tug' for unmodified whole immunoglobulins and '-bart' for whole immunoglobulins with engineered amino acid changes [181]. This was supplied by Andrew Martin in June 2024 (Data File 5).

3.2.3 Evaluation of Datasets Physicochemical Properties

It was thought that clinical antibodies with properties that separate them from library antibodies could be identified using physicochemical properties linked to developability characteristics as had been used in the TAP score [110] and TA-DA score [111]. However, it was found that the TAP score was not be suitable for high-throughput sequence analysis because of the need to model antibodies, which was time consuming on its online portal. For this reason, metrics that do not require modelling and could be calculated from sequence (ΔG, pI and CDR-H3 length) were selected.

3.3 Physicochemical Property Triaging

To establish a triaging pipeline from physicochemical properties, physiocochemical properties were evaluated and compared between the clinical and library datasets (Table 3.1). The maximum and minimum values of continuous variables were used to identify antibodies with developability characteristics which fall within the ob-

Table 3.2: Triaging effect of Z score filtering on all physicochemical properties using different values of Z.

	Z Score			
	None	2	1	0.5
Number of Antibodies	9653	7038	2600	351

served range of the market approved mAbs. Properties including CDR-H3 length, pI and ΔG were selected because methods were found to calculate them quickly using only sequences. Additionally, whilst they may not be very informative in terms of developability, these continuous variables could be used to make triaging-cutoffs using Z scores to apply to large databases and remove antibody sequences which show high divergence from the mean. This would allow a user to select a Z score based on how stringently they would like to triage the antibody library. It would be expected that using smaller Z scores would result in more antibodies being triaged out of the library as the range for acceptable features would become more narrow with respect to the mean seen in clinical mAbs.

It was observed that, decreasing the Z score had a dramatic triaging effect for all properties (Figure 3.1), and ΔG of unfolding had the largest filtering effect at the lowest Z score tested. When filtering using all properties, more antibodies were triaged out at each selected Z score than any of the individual filters (Table 3.2). This was a drastic filtering effect that, at the lowest Z score tested, removed over 95% of the entered antibodies. This result would be expected to give a subset of OAS antibodies with physicochemical properties similar to most clinical mAbs. However, it was decided to also use the machine learning models to evaluate this dataset and select candidates.

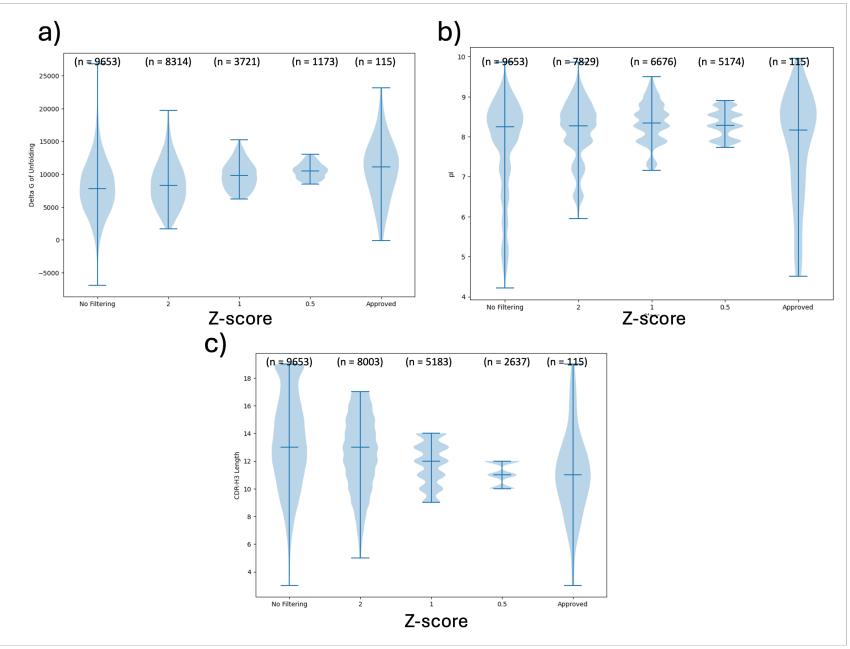

Figure 3.1: Triaging OAS antibodies using physicochemical properties. Violin plots of results of filtering OAS library antibodies based on based on Z score cutoffs for a) ΔG of unfolding (V_H+V_L), b) isolectric point and c) CDR-H3 Length.

3.4 Supervised Learning

3.4.1 Training models

As was introduced in the methods chapter, supervised machine learning may give a better solution approach to this problem, and the use of sequence encoding techniques may provide more insight into the mechanisms that determine these developability factors, more so than superficial physicochemical properties. This problem similarly has two groups which can be labelled library (class 0) and clinical (class 1) and so if a classifier could be trained which has a high performance of identifying clinical antibodies (i.e. high sensitivity), false positive examples (i.e. repertoire antibodies which the model thinks is a clinical antibody) are likely to have properties similar to clinical antibodies.

All of the human clinical dataset (class 1) and human repertoire data (class 0)

from the OAS were encoded with the AntiBERTy language model. A selection of 15 supervised machine learning models (see Section 2.8.1) were trained on this data using different numbers of antibodies randomly sampled from the library dataset (n=100, n=500, n=1000, n=10,000) to investigate whether this weighting balance of positive to negative examples made a difference to how the models learned. This was combined with F-regression to investigate if the number of features (k) used for the models to learn from had an effect. Different values for k were selected [1, 10, 50, 100, 500, 1000, 2500, 5000, 10000].

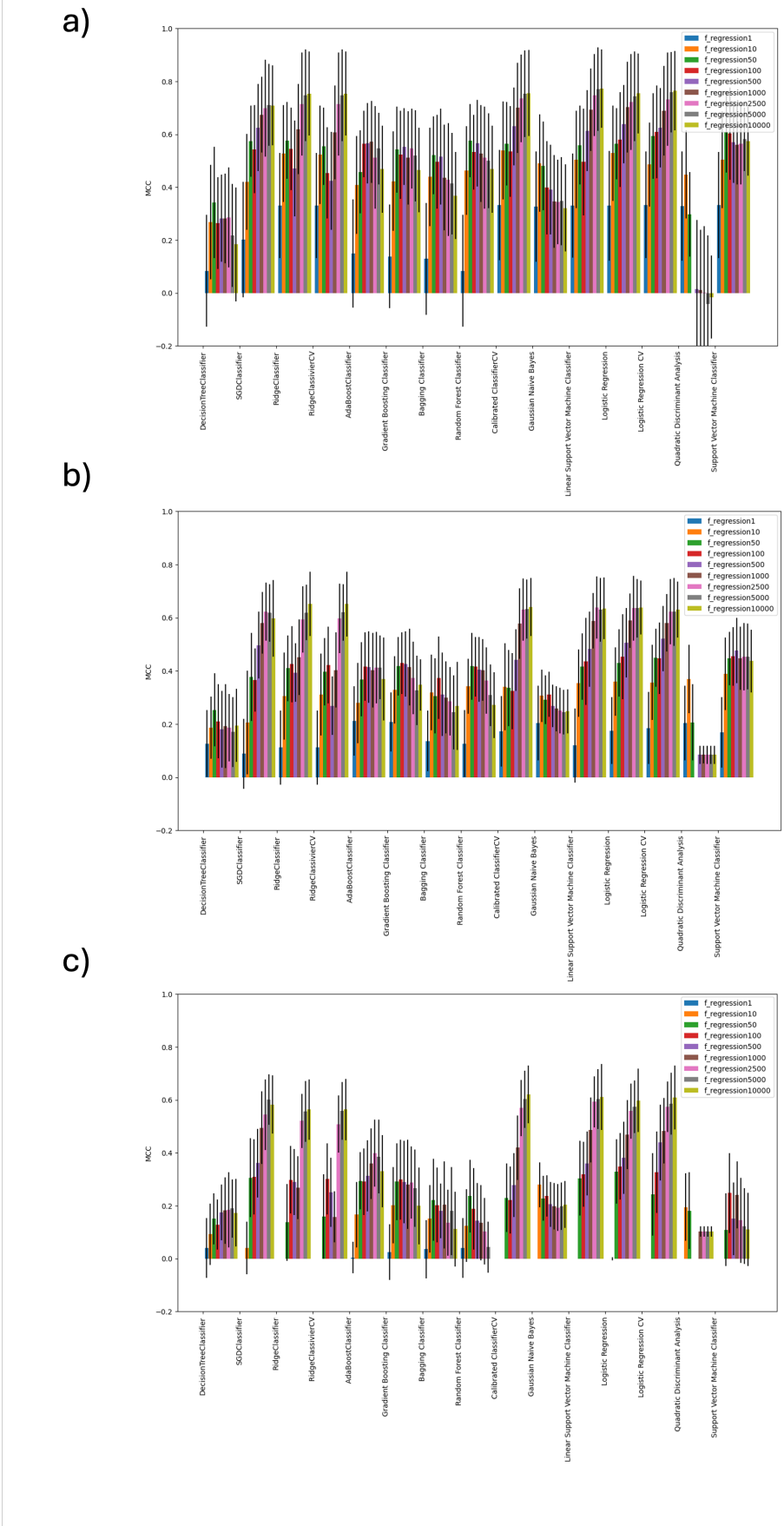

Figure 3.2: Classifiers trained on OAS repertoire antibodies and Clinical stage mAbs. MCC scores and standard deviation of 16 binary machine learning predictors classifying test split of repertoire antibodies a) (n = 100), b) (n = 500), c) (n = 1000) against clinical antibodies (n = 143) dataset encoded with the AntiBERTy language models. F-regression scores are colour coded.

Classifiers were trained on all splits of the data. It seems from inspection that performance for each split was similar across Support Vector Machine Classifier (SVC) models, Ridge classifiers, Logistic Regression and Calibrated Classifiers, and that there was an increase in performance was seen when more features were used in training, however, this increase in performance plateaued after k=1000.

The best overall performance was seen when training the clinical dataset against 100 library antibodies (Figure 3.2a), which was an aproximately balaced dataset,with k set to 10,000 features, and trained with the LinearSVC model was used (MCC=0.77±0.14). Performance decreased when training data increased to 500 library antibodies, but the best performance at this split was also observed using the same classifiers as before: the Ridge Classifier (MCC=0.65±0.12) where k=10,000 (Figure 3.2b). A similar decrease in performance was seen with 1000 library antibodies, where the best for the Calibrated Classifier (MCC=0.62±0.12) where k=10,000 (Figure 3.2c). It could be drawn from this experiment that by giving the library and clinical antibodies an equal weighting, the models are able to learn more to distinguish them, however, when using fewer library examples, it is more difficult to capture the possible diversity of the library leading to less generalisable models. So a method of manual cross-validation (CV) was devised that would allow for this balanced weighting and to allow the model to be trained on a greater diversity of the repertoire.

3.4.2 Manual CV Training

While it was seen that more balanced datasets of clinical and library antibodies to train classifiers was found to give the most reliable predictions, as previously mentioned, this may exclude a lot of the diversity seen in the repertoire, and it risks labelling potentially useful antibodies as class 0. In reality, the repertoire data are not labelled. It is uncertain whether a given library antibody would have the clinical features appropriate for a clinical antibody, and so labelling all of these library antibodies the same class would be misleading to the model. For this reason, a method of manually sampling the data was devised in order to train multiple models on the diversity of the repertoires and to use a method to take a vote of predictions from a collection of models as a method of not losing potentially useful antibodies.

It was decided to use sets of 500 library antibodies with a clinical data split as using a selection of features from the F-regression (k=2500) as training data collection of models from which a vote would be taken. This choice was made for a number of reasons. Firstly, performance was thought to be more consistent in these splits due to the reduced standard deviation of MCC scores between splits of the data than seen in the 100 library antibody split. Secondly, using the 100 library antibody method would also make 99 different models, per split to train on all antibodies from the OAS library, totalling 990 models, which seems to be unnecessarily large for this purpose. Thirdly, 500 library antibodies give the opportunity to introduce more diversity to the training dataset for individual models.

The 10,000 repertoire antibodies were randomly split into 20 splits of circa 500 library antibodies and joined with a given split of clinical data to give a total set

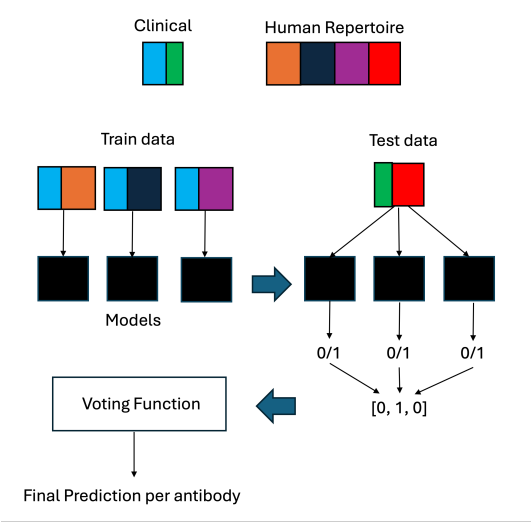

Figure 3.3: Schematic representation of the generation of weighted voted predictor. This shows how training clinical data has been concatenated to different splits of repertoire data to train individual classifiers and a held back split of both clinical and repertoire antibodies to test the overall performance of the model. This example is given imagining 4 splits of the human data and two splits of the clinical data.

of training data of 19 different training sets of 500 different library antibodies and 130 of the same clinical antibodies, and a test set of 500 library antibodies and 13 clinical antibodies. This would give 20 sets of predictions, each as a result of a vote taken from 19 models.

3.4.3 Voting Function

To then take a final vote for all predictions per antibody, two voting functions were tested to collect the final prediction of each antibody. This process has been schematically outlined in Figure 3.3. Because of this method, the antibodies to be taken forward would be seen as false positives to the model, so in addition to MCC other statistics have been used to evaluate the performance. False omission rate (FOR) to evaluate how often the models do not learn the positive examples, which would be expected to be low. Negative predictive value (NPV), to evaluate how well

the model can predict negative values, which would be expected proportional to the fraction of repertoire to clinical antibodies. Details of these statistics are given in Section 2.10.1.

3.4.3.1 Function one

The first voting method tried was to take the sum of positive and negative predicted probabilities and take the vote of whichever sum was higher (Equation 3.1).

$$S_{\text{out}} = \begin{cases} \frac{(\Sigma^{N^+} s_o^+)}{N^+}, \; if \; (V = C^+) \\ \frac{(\Sigma^{N^-} s_o^-)}{N^-}, \; if \; (V = C^-) \end{cases} \tag{3.1}$$

- s_o^+ is the confidence output of a predictor predicting C^+
- s_o^- is the confidence output of a predictor predicting C^-
- N^+ is the number of predictors predicting C^+
- N^- is the number of predictors predicting C^-

This gave 20 sets of predictions, which have been evaluated in Table 3.3. Some individual predictions using this method look impressive, especially predictions 1, 5, 12 and 20 where it seems there is a low FOR with high NPV as expected. Importantly, in each of these cases, more true positives are present than true negatives, indicating a trustworthy set of models that have learnt patterns that separate the clinical and library antibody.

Table 3.3: 20 sets of predictions made from voting classifier using Method 1.

Prediction	MCC	NPV	Sn	TP	FP	TN	FN
1	0.421	0.996	0.714	5	14	452	2
2	0.298	0.993	0.571	4	19	447	3
3	0.352	0.996	0.667	4	16	451	2
4	0.242	0.991	0.429	3	16	450	4
5	0.379	0.996	0.714	5	18	448	2
6	0.212	0.993	0.500	3	25	442	3
7	0.314	0.993	0.571	4	17	449	3
8	0.332	0.993	0.571	4	15	451	3
9	0.399	0.996	0.714	5	16	450	2
10	0.064	0.987	0.143	1	18	448	6
11	0.471	0.998	0.833	5	13	454	1
12	0.409	0.996	0.714	5	15	451	2
13	0.353	0.993	0.571	4	13	453	3
14	0.083	0.989	0.167	1	15	452	5
15	0.377	0.993	0.571	4	11	455	3
16	0.278	0.991	0.429	3	12	454	4
17	0.178	0.989	0.286	2	13	453	5
18	0.342	0.993	0.571	4	14	452	3
19	0.265	0.993	0.500	3	16	451	3
20	0.362	0.996	0.714	5	20	446	2
Mean±SD	0.307±0.108	0.993±0.003	0.548±0.185	3.700±1.261	15.800±3.189	450.450±3.120	3.050±1.234

3.4.3.2 Function two

The second function takes the difference in the sums of positive and negative predicted probabilities corrected by the number of predictions (Equation 3.2).

$$S_{\text{out}} = \left| \frac{(\sum_n^{N^+}(s_{o,n}^+ - s_{o,n}^-)) - (\sum_n^{N^-}(s_{o,n}^- - s_{o,n}^+))}{N^+ + N^-} \right| \tag{3.2}$$

- s_o^+ is the confidence output of a predictor predicting C^+
- s_o^- is the confidence output of a predictor predicting C^-
- N^+ is the number of predictors predicting C^+
- N^- is the number of predictors predicting C^-
- S_{out} is the final voted confidence output.

This method of voting produces different results than the previous method de-

Table 3.4: 20 sets of predictions made from voting classifier using Method 2.

Prediction	MCC	NPV	Sn	TP	FP	TN	FN
1	0.431	0.989	0.286	2	1	465	5
2	0.562	0.991	0.429	3	1	465	4
3	0.662	0.996	0.667	4	2	465	2
4	0.376	0.987	0.143	1	0	466	6
5	0.501	0.991	0.429	3	2	464	4
6	0.284	0.989	0.167	1	1	466	5
7	0.420	0.991	0.429	3	4	462	4
8	0.330	0.989	0.286	2	3	463	5
9	0.501	0.991	0.429	3	2	464	4
10	-0.010	0.985	0.000	0	3	463	7
11	0.406	0.989	0.167	1	0	467	5
12	0.562	0.991	0.429	3	1	465	4
13	0.371	0.989	0.286	2	2	464	5
14	0.000	0.987	0.000	0	0	467	6
15	0.672	0.994	0.571	4	1	465	3
16	0.180	0.987	0.143	1	3	463	6
17	0.431	0.989	0.286	2	1	465	5
18	0.501	0.991	0.429	3	2	464	4
19	-0.007	0.987	0.000	0	2	465	6
20	0.330	0.989	0.286	2	3	463	5
Mean±SD	0.375±0.203	0.990±0.002	0.293±0.187	2.000±1.257	1.700±1.129	464.550±1.356	4.750±1.164

spite using the same predictions as given above. In this case, the best looking sets of models are from prediction 3 (Table 3.4). No other set of models would be considered suitable for use as the sensitivity is lower, meaning that true positive examples have not been captured, and so the selected false positive results (repertoire antibodies expected to have clinical properties) become less trustworthy. Overall, this weighted method of voting would be seen as more conservative method than using the result of the summed probabilities.

3.5 Unsupervised Learning

To examine the groups of unlabelled data, unsupervised learning is the most suitable solution to allow natural clustering of data. This general approach has a number of advantages including that all antibodies in the sample can be included into a single model, and that these methods can be plotted in 2-dimensional space so that antibodies which are similar to each other would be positioned closer together than

antibodies that had different properties. Using this principle, it was hypothesised that if the clinical and repertoire antibodies underwent dimensionality reduction together, a cluster of clinical antibodies which would be assumed to have similar, developable properties would be identified together in one area and that antibodies with similar properties would cluster closely with the clinical mAbs. These would be the antibodies that would be selected to be taken as clinical candidates.

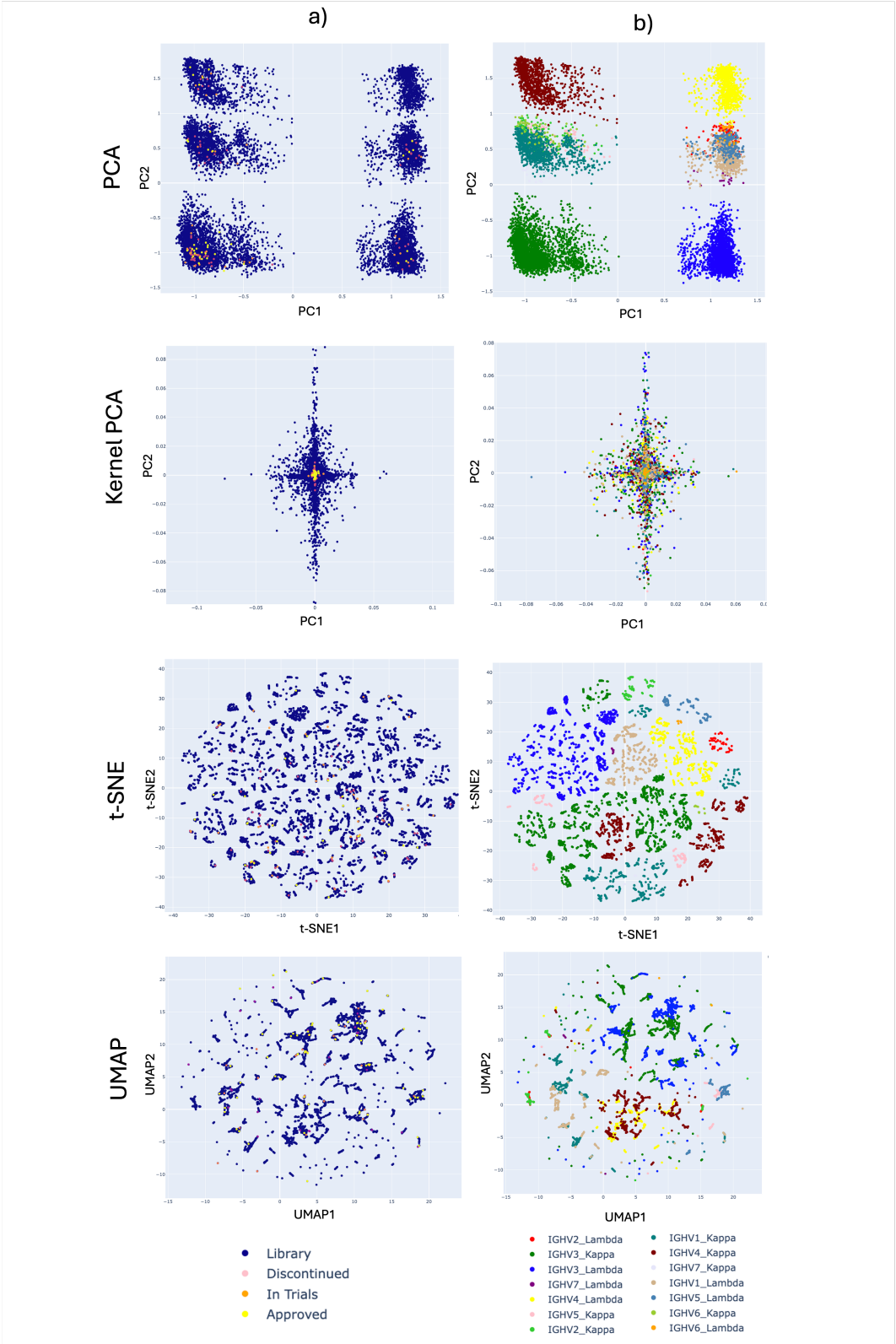

Figure 3.4: Scatter plots of unsupervised machine learning models trained on clinical (n=144) and library (n=10,000) paired antibody sequences encoded with the AntiBERTy language model. Plots are colour coded by clinical stage (a) or V_H chain V region germline gene and V_L chain type (b).

3.5.1 Principal Component Analysis

Linear principal component analysis was performed on the encoded data and the first two principal components were plotted. In total, 6 discrete groups were formed with three at one extreme of the principal component (PC) 1 (Figure 3.4a). Upon investigation, it was found that these clusters corresponded to different V_L chain types and heavy chain V_H germline families related to PC1 and PC2 respectively. Clinical antibodies were observed inside all these groups except for the IGHV4 with λ V_L chain, and were particularly over-represented in the groups of IGHV3 with κ V_L chain groupings. However, these groupings of clinical antibodies did not form tight clusters within their germline clusters as hypothesised, and so it was difficult to use these as a basis to identify clinical properties in library antibodies.

3.5.2 Kernel Principal Component Analysis

Non-linear or Kernel PCA (KPCA) using the radial basis function kernel showed the hypothesised outcome of clustered clinical antibodies amongst repertoire antibodies. Using increasing kernel coefficients (γ=[10, 100, 500, 1000]), the cluster of clinical antibodies persisted and this cluster always appeared at the origin of the dimensionality reduction plot. Furthermore, it appeared that increasing the values of γ caused the elimination in the importance of one of the principal components, which was most apparent when set at γ=1000 (Figure 3.5). Therefore, the plot at γ=500 was selected as the best way to demonstrate the clustering effect that was hypothesised. From inspection, it seems that the approved clinical antibodies have a tighter grouping effect than what is seen in the in trials and discontinued subgroup

along both axes, but no real effect was found (Table 3.5). Furthermore, when the plot was colour coded according to V_H germline family and V_L chain type pairing, no obvious clustering effect was found. This supports the idea that this method of dimensionality reduction is learning other characteristics than germlines (Figure 3.4b).

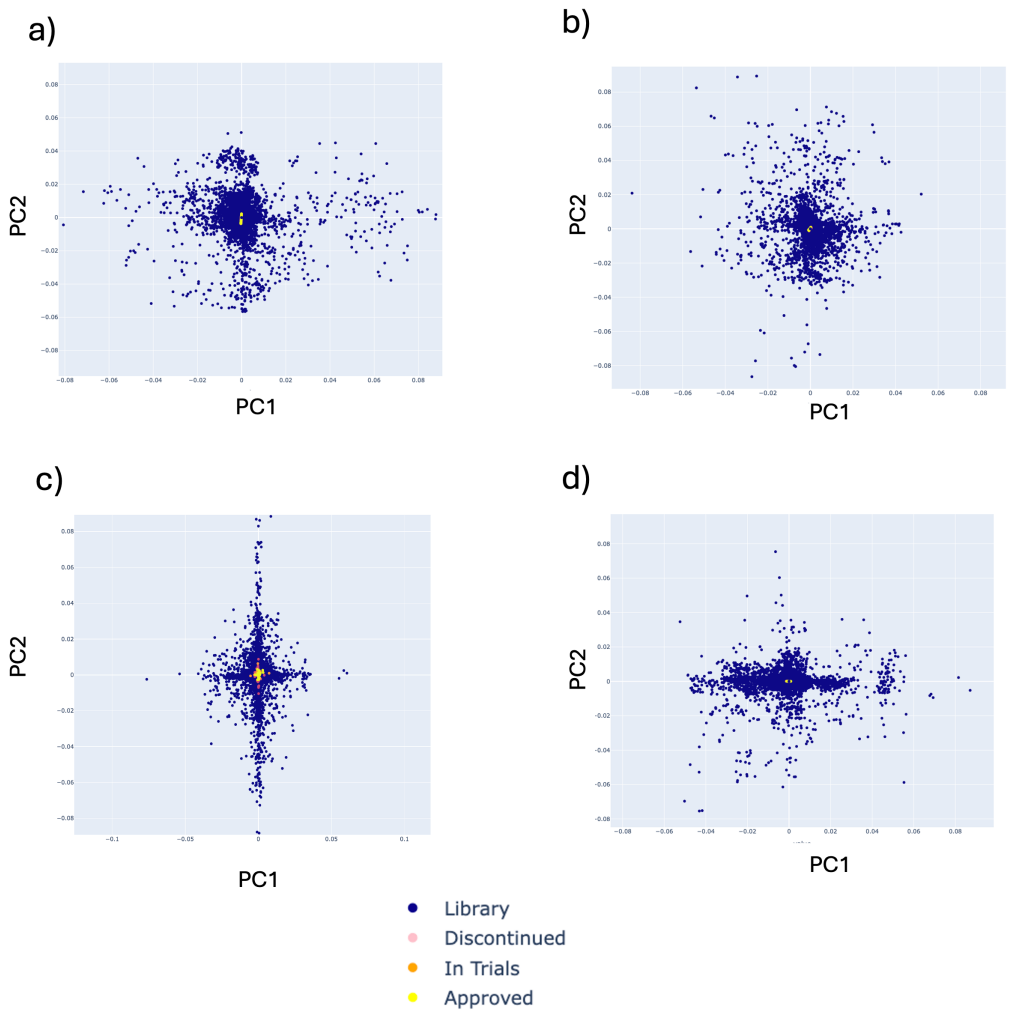

Figure 3.5: Scatter plots of kernel principal component analysis models trained on clinical (n=144) and library (n=10,000) paired antibody sequences encoded with the AntiBERTy language model. Plots are given for increasing kernel coefficients for a) γ=10, b) γ=100, c) γ=500 d) γ=1000. Antibodies are colour coded by clinical stage.

Table 3.5: Median values and standard deviation of PC1 and PC2 given for different groups of encoded antibody sequences of the KPCA (γ=500).

	PC1	PC2
OAS	-0.0007±0.01	0.0004±0.01
Approved	0.002±0.001	-0.0005±0.001
Trials	0.001±0.001	-0.0005±0.0008
Discontinued	0.001±0.001	-0.0002±0.001

3.5.3 t-SNE

More complex methods of dimensionality reduction include t-SNE, which has previously been shown to group the langauge model encodings in 2-dimensional space [33]. t-SNE was performed with perplexity set at 30, as this was the default value set in the Python module. This method of clustering formed much more discrete groupings, which these were highly linked to germline gene pairing but more granular than the linear PCA. Clinical mAbs were dispersed throughout the groupings, such that only one or two clinical antibodies would be in some clusters. Consequently so, it would be difficult to use this dimesionality reduction method is not suitable to identify groupings of clinical antibodies (Figure 3.4). It appears in this scenario that there are clusters of antibodies which have the same heavy chain V_H domain germline and light chain type (κ or λ), however, it would appear that IGVH3 paired with κ and IGHV3 paired with λ are widely spread amongst clusters representing more diversity within this germline pairing. Potentially, subgroups between common germline gene pairings would show this pairing effect at the V_L germline level. It was seen that increasing the perplexity could cause these clusters to converge in a similar manner to that seen in the linear PCA.

3.5.4 UMAP

UMAP is a popular method for dimensionality reduction that has previously been applied in demonstrating how antibody language models can separate repertoires of antibodies into groups [177]. The results presented here was set with the nearest neighbours set to 3, as this was the default value set by the program. As can be seen here, clusters have formed where clinical antibodies are dispersed among the clusters, rather than clustering together, as seen in the Kernel PCA. It can also be seen in Figure 3.4, as with previous methods of unsupervised learning, that these groups can be explained by V_H/V_L germline gene family pairing, however, there is overlap of λ (blue) and κ (green) pairing with IGHV3 and with IGHV4 and λ (red) and IGHV4 with κ (yellow) (Figure 3.4). So, this would demonstrate that there are similar patterns of clustering seen here as in t-SNE or linear PCA where V_H and V_L pairing account greatly for the observed clustering. Due to the dispersal of clinical antibodies through the clusters in these three dimensionality reduction and clustering methods, it was still considered unsuitable for pipeline purposes. Interestingly, it seems that V_L germline has less influence on clustering in UMPA than seen in other methods.

3.5.5 Selecting an Unsupervised Model

The Kernal PCA (KPCA) was selected as the most appropriate unsupervised machine learning model to separate clinical and library antibodies. To confirm its suitability, a test dataset made up of clinical mAbs which were named since 2022 using the new naming convention and therefore not included in the original train-

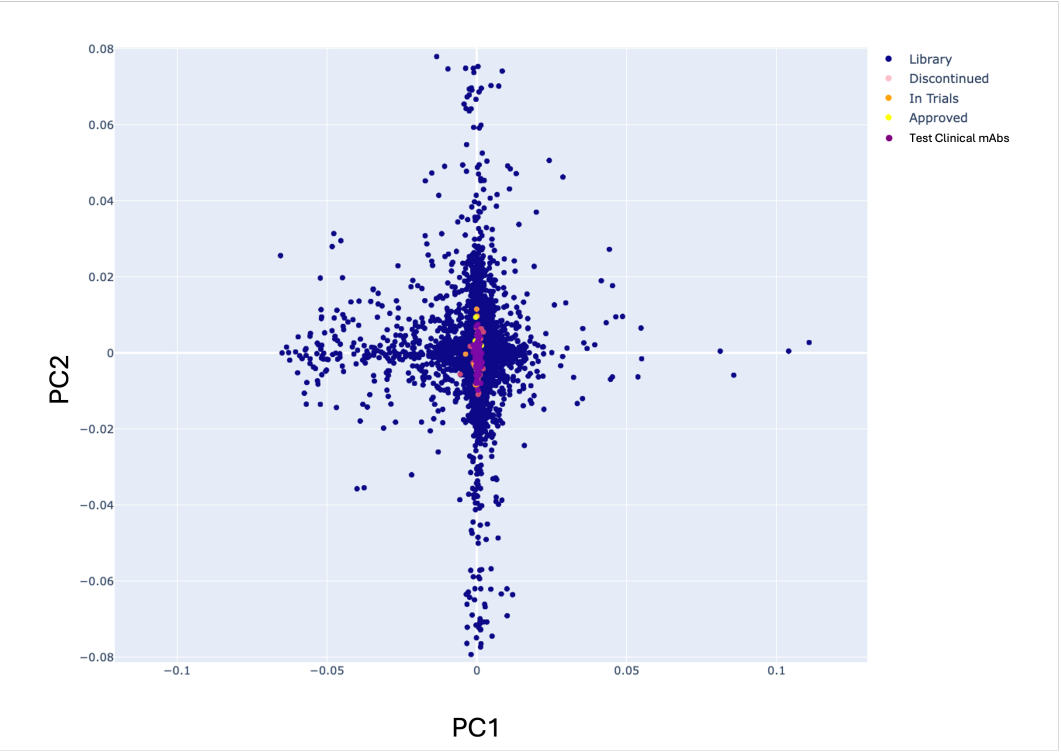

Figure 3.6: Scatter plots of Kernel PCA trained on clinical (n=144), library (n=10,000) and test clinical (n=203) paired antibody sequences encoded with the AntiBERTy language model.

ing data, was entered into the KPCA together with the OAS repertoire data and the original clinical dataset to examine if these antibodies were positioned closely to the original clinical dataset. It was seen that this was the case (Figure 3.6) and thus strengthened the argument that antibodies with suitable clinical features can be identified by positioning in the KPCA plot. The next consideration was how to identify antibodies positioned closely to the clinical dataset.

3.5.6 Selecting Clinical Candidates from KPCA

In order to use the result of unsupervised learning as a method of triaging naïve library antibodies, a method needed to be devised to identify candidates that closely align with clinical antibodies. Several approaches have been taken to do this.

3.5.6.1 Z Scores

The simplest imagined method was to look at the extremes of the clinical antibodies at each PC and use these as triaging criteria for antibodies outside these observed ranges. As shown with physiochemical filtering, Z scores may be used to give added stringency to the triaging process, so with lower Z scores used, antibodies closer to the origin are selected. The issue with this triaging method is that because of the radial shape of the KPCA, by using linear cutoffs, this effectively draws a box around the clinical antibodies and the similar library antibodies, which introduces unwanted library antibodies to the selection at the vertices of the box. For this reason, if using this triaging method, it would be better to use very low Z scores so the selection is entirely within the clinical antibodies, however this may introduce additional biases to the selection and would reject potentially useful antibodies.

3.5.6.2 Ellipse Function

A method for excluding unwanted antibodies at the extremes of a box method would be to use a circular function that captures only the desired library antibodies. Using similar principles in which antibodies on the extremes of the clinical cluster could be used as the major and minor axes of an ellipse, it can be drawn (see Section 2.6). Those antibodies which fall inside the drawn ellipse may be taken forward. Using the Z score cutoff method, the distance between the extremes on the major and minor axes can be altered and fewer antibodies are captured, and would be expected to be of higher quality as they cluster closer to the majority of clinical antibodies.

3.5.6.3 Pairwise Distances

Using the idea of capturing antibodies that are closer to the main group of clinical mAbs, pairwise distances between clinical mAbs can be used to identify antibodies which are as close to a clinical mAbs as they are to each other. By examining the largest distance between antibodies using the Euclidean distance formula (Equation 3.3), this distance can be applied to all of the library antibodies against all of the clinical antibodies, and those with a distance from a clinical antibody shorter than the maximum distance observed between clinical antibodies may be taken through. Stringency can be added to this model by shortening the maximum distance allowed or by requiring more than one antibody to be within the maximum distance.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{3.3}$$

3.5.6.4 Selecting an Approach

From the number of antibodies remaining in Table 3.6, it can be seen that the use of the ellipse function had the greatest effect of triaging out antibodies with each Z score cutoff compared with the other methods. It seemed surprising that the Z score box and pairwise distances method had low stringency, and even at the lowest selected Z score, 4627 and 6705 antibodies remained respectively, which would suggest minute scores would be needed to triage a manageable amount for the next stage of the pipeline. This is complimented by a much more conservative approach taken with the ellipse function, which only had 351 remaining library antibodies

Table 3.6: Counts of antibodies remaining from OAS library using different methods of selecting candidate antibodies from Kernal PCA result.

	Z score			
Method	2	1	0.5	0.1
Z score	7437	6371	5625	4627
Ellipse	4391	1521	813	351
Distances	8242	7792	7377	6705

with the lowest Z score threshold. This would make it the most appropriate approach to take forward candidates that cluster closely with clinical antibodies.

3.6 Discussion

The work carried out in this chapter describes approaches to identify library or repertoire antibodies that have properties similar to human clinical mAbs, which are assumed to have these developability properties. So, if entering a library of antibodies, the number of antibodies retained by the model would be unmanageable, so this additional approach was necessary to improve the result outputted by the pipeline.

It was seen that triaging on physicochemical properties could reduce the count of input antibodies to a manageable number. However, the properties used were considered to be superficial and did not give much insight as to how similar these antibodies are to clinical mAbs, even when the composite filter had a dramatic effect of triaging out input antibodies. Potentially, this is reflective of how few antibodies occupy the space where all of these conditions are satisfied. This approach was considered unrealistic for triaging antibodies because this would only retain antibodies as ones which sit very close to the mean of these properties seen in the clinical mAbs and not give opportunity to explore more diverse options. Previously, the

'developability web' analogy was mentioned, where one property may affect others; for this reason, it is probably unrealistic to expect many antibodies to occupy this narrow space, and so it would be more prudent to use this physicochemical filtering to triage out negative examples which sit outside the observed ranges, than to use this for selection. These physicochemical properties can therefore be used with low stringency to remove antibodies clearly unlikely to have suitable properties before using the machine learning approach, thus saving unecessary compute time.

Supervised learning appeared to generate voting models that were confident in identifying positive examples and triaging out negative examples. This was certainly true of the first method in cases where in most cases more true positives were found than true negatives, and so the false omission rate remained fairly low and consistent. For each test split of the data of around 500 antibodies, between 13-25 false positives were selected, which would be the library antibodies selected as clinical antibodies. This would mean that for a library of 10,000 around 260-500 antibodies could be selected. The second voting method overall scored worse, as *Sn* was much lower first voting method, potentially excluding useful antibodies if the models are not capable of learning the difference between clinical and library antibodies. This approach offers a clear cut method of identifying which antibodies to be taken forward, as the model outputs them.

Although the exercise in training supervised learning models was useful to demonstrate proof of concept, because these data are unlabelled in reality, methods of dimensionality reduction and unsupervised learning were used to cluster the

repertoire and clinical antibodies. It was hypothesised that candidates would cluster physically close to the clinical antibodies. The KPCA approach showed this relationship well and did not appear to be affected by biases in the germlines, which was the case for the other methods. However, some drawbacks to this approach would include that this method is best when taken with a large sample that is representative of the diversity of the repertoire, so future runs of the KPCA may have to include this OAS data as well as the library data to ensure that this diversity is taken into account by the model. Otherwise, if only a small number of antibodies are entered, they may all cluster close enough to the clinical antibodies that all may progress, which is not an informative outcome.

Because the clustering is more dense around the origin, a stringent triaging using Z scores was devised with the elliptical function, where more stringent Z scores led to fewer antibodies being selected through drawing a smaller ellipsis.

It was interesting that antibodies from the clinical dataset grouped around the origin of the radial basis function KPCA. This could indicate some notions about the properties of the clinical antibodies. Firstly, it supports the assumption of developability by demonstrating a shared property or set of properties that are also shared by some library antibodies, and the fact that these antibodies are typical examples of an antibody with no extremes shown. Secondly, biases in the germline will lead to antibodies of certain V_H and V_L germline family pairings being selected as clinical candidates from this pipeline. If these are proven to work, it does not appear to be a disadvantage, since the pipeline would be designed for libraries or repertoires of fully human antibodies. The ellipse function provided a robust solution for to

selecting out the antibodies which we want to take forward.

3.7 Conclusions

In conclusion, this chapter has shown that a sample of human repertoire or library antibodies can be triaged to select those with properties similar to clinical mAbs, assumed to fulfil the developability criteria needed to enter clinical trials. To continue this work, it seems necessary to investigate these physicochemical properties further and improve methods of prediction.

Chapter 4

Using LLMs to Predict Antibody Developability Features

4.1 Introduction

As shown throughout the thesis, the importance of producing antibodies with appropriate developability properties has become paramount in the pursuit of safe and viable therapeutics. These properties relate to the stability, hydrophobicity, charge and immunogenicity of the molecule. Although there are experimental methods to measure these metrics for proteins and antibodies, they are costly and time consuming, meaning they are not compatible with the high-throughput pipelines employed in antibody discovery campaigns until a manageably small number of leads are identified. Due to the high attrition rate of these pipelines, the focus has recently been on developability where *in silico* predictions of these metrics can be performed to avoid expensive late-stage failures [182, 103, 101].

This chapter is dedicated to using existing datasets to predict these properties.

It is split into two sections, the first relating to physicochemical properties including: thermostability, hydrophobicity; cross-reactivity and solubility using data from Jain *et al.* [102]. The second relates to immunogenicity prediction from Anti-drug antibody data from Marks *et al.* [104]. These were done using linear models, and machine learning approaches trained on experimental data trained using encodings from antibody LLMs. The use of these models would be important additions to the triage pipeline so that each antibody could receive additional predictions related to its properties, using the same encodings used by the pipeline thus far.

4.2 Predicting Physicochemical Properties of antibodies

4.2.1 Introduction

A landmark paper by Jain *et al.* [102] published experimental values measuring thermostability, hydrophobicity, self association and polyreactivity for 137 clinical stage antibodies. Measurements for thermostability included Melting temperature (Melting temperature); Accelerated Stability (AS) and Stand-up Monolayer Adsorption Chromatography (SMAC). Measurements for hydrophobicity included Hydrophobic Interaction Chromatography (HIC) and Salt-gradient Affinity-Capture Self-Interaction Nano-Particle Spectroscopy (SCAG) [183]. Measurements for self-association included: Cross-interaction chromatography (CIC); Clone self-interaction by bio-layer interferometry (CSI) and Affinity-capture self-interaction nanoparticle spectroscopy (AC-SINS) [184]. Measurements for polyreactivity

included Binding to baculovirus particles (BVP), Baculovirus particle ELISA (ELISA) [185].

Through these measurements, they identified key correlations between metrics of the same properties (AC-SINS and CIC; AC-SINS and CSI; ELISA and BVP) but also some unexpected correlations between unrelated properties (HIC and SMAC). Together, this demonstrates the interconnectivity of these developability features and how they affect each other. This work was continued in 2023 [103] where additional characteristics for clearance: *Fc*RN retention time (FcRN-RT) and binding to heparin (Hep.RT.3) [186], and polyspecificity were included: binding to 2,4-dinitrophenol (DNP), Induced polyspecificity in the presence of either iron or heme to Haemophilia Factor VIII (Fe.FVIII.2/Heme.FVIII.2), Complement (Fe.C3.2/Heme.C3.2), lysozyme (Fe.LysM.2/Heme.LysM.2), as well as folate binding (FA.2).

Previous attempts to train predictive models on these experimental results included the AbPred server, where machine learning models are available for *in silico* predictions of these metrics on an antibody-by-antibody basis [107]. This software used 20 amino acid propensities, previously established amino acid compositions to encode V_H and V_L sequences into a numerical input for training. Using a selection of linear models that were best suited to each metric, coefficient of determination (R^2) results for each predictor ranged from 0.08 (AS) to 0.39 (HIC). A similar approach to predict developability was taken by Negron *et al.* [111] using a logistic model fitted over the sequence characteristics of all clinical antibody sequences found in TheraSabDab [117]. Compared with the Jain data, rank-order correlations (ρ) were

Table 4.1: Descriptive statistics of the *in vitro* data taken from Jain *et al.* (2023).

Metric	Abbreviation	Median	Stdev	Desirability	n	Measurement
Melting temperature (Negated)	Tm(Negated)	-71	5.86	-	137	Stability
Accelerated Stability	AS	0.04	0.12	-	137	Aggregation
Salt-gradient affinity-capture self-interaction nano-particle spectroscopy	SGAC(Negated)	-700	319.23	+	137	Hydrophobicity
Hydrophobic interaction chromatography	HIC	9.89	1.23	-	137	Hydrophobicity
Stand-up monolayer adsorption chromatography	SMAC	8.88	3.05	-	137	Colloid stability
Poly-specificity Reagent	PSR	0	0.2	-	137	Non-specific binding
Affinity-capture self-interaction nanoparticle spectroscopy	AC-SINS	1.74	10.4	-	137	Self-interaction
Cross-interaction chromatography	CIC	8.87	0.85	-	137	Solubility
Clone self-interaction by bio-layer interferometry	CSI	-0.01	0.1	-	137	Self-interaction
FcRn relative retention time	FcRn-RT	0.445	0.56	+	132	Clearence
Baculovirus particle ELISA	ELISA	1.21	2.65	-	137	Non-specific binding
Binding to baculovirus particles	BVP	2.34	4.32	-	137	Non-specific binding
Binding to 2,4-dinitrophenol	DNP	0.245	0.79	-	112	Non-specific binding
Induced polyspecificity to FVIII (Fe^2+)	Fe.FVIII.2	3.27	1.69	-	114	Non-specific binding
Induced polyspecificity to C3 (Fe^2+)	Fe.C3.2	2.59	1.39	-	114	Non-specific binding
Induced polyspecificity to LysM (Fe^2+)	Fe.LysM.2	5.02	2.1	-	114	Non-specific binding
Induced polyspecificity to FVIII (Heme)	Heme.FVIII.2	2.62	2.19	-	113	Non-specific binding
Induced polyspecificity to C3 (Heme)	Heme.C3.2	3.1	2.73	-	113	Non-specific binding
Induced polyspecificity to LysM (Heme)	Heme.LysM.2	3.99	3.07	-	113	Non-specific binding
Heme binding	Heme.2	2.08	5.33	-	113	Non-specific binding
Folate binding	FA.2	1.25	2.73	-	113	Non-specific binding
Heparin chromatography retention time	Hep.RT.3	0.58	0.2	-	130	Non-specific binding

also between 0.2 and 0.3, which was notable as the models had not been trained using these data, showing that these characteristics can indeed be trained on.

Similarly this chapter will demonstrate that some of the large language model encodings are statistically correlated with experimental values provided by Jain *et al.* [102, 103] for clinical stage antibodies and these encodings may be used for *in silico* prediction of these metrics.

4.2.2 Datasets

4.2.2.1 Jain Developability Data

Experimental data for *in vitro* assays and amino acid sequences for 137 paired clinical stage antibodies were taken from the supplementary material of Jain *et al.* [103]. *In vitro* assays performed, and measurements, are given in Table 4.1. Paired V_H and V_L amino acid sequences for these therapeutics were found in the TheraSabDab [33]. These sequences were then encoded with a selection of antibody LLMs given in Section 2.4.

4.2.3 Encodings from Language Models are Statistically Correlated to Experimental Values

Similarly to Chapter 3, it was thought that feature selection could identify features encoding information relevant to these developability characteristics, but in this case, rather than using arbitrary numbers of features, features which were statistically correlated with the F-regression were selected. It was found that all of the LLMs had significantly correlated datapoints to each characteristic using the F-regression method. Significant features were identified from the p-value given by the F-regression for the chance of its significant correlation ($p < 0.05$) to the experimental measurement. Adjusted p-values (q-values) were given using the Benjamin-Hochberg method (see Section 2.2.4), however significant values were not always found for each metric. The counts of significantly associated features are given in Table 4.2 for p-value and q-value levels. Included in this chapter were versions of the ESM model with higher numbers of parameters: `esm_t35_650M_UR50D` (ESM_650M) and `esm2_t36_3B_UR50D` (ESM_3B) as larger LLMs to include.

4.2.4 Positions of Selected Features in the Antibody Sequences

This led to finding where such correlated features lay in the sequence that could be used to train machine learning classifiers (Figure 4.1). This figure demonstrated that most of the metrics had numbers of features selected skewed to one particular domain. For instance, HIC had more statistically associated features from the V_H domain than the V_L domain at the p-value level, whereas the opposite was true of FcRN-RT, SCAG, CIC which were more skewed to the V_L domain. Distributions

Table 4.2: Number of statistically correlated features ($p < 0.05$ and $q < 0.05$) found by F-regression for each metric and language model.

		Model					
Property	Significance	AntiBERTy	ESM_8M	ESM_650M	ESM_3B	AbLang	Sapiens
	All	130048	82560	330240	660480	195072	162560
Tm(Negated)	p	18754	12812	50671	103039	31922	28616
	q	248	82	969	1696	546	15
AS	p	9750	7378	27962	53923	18753	16067
	q	0	25	6	2	47	383
SGAC(Negated)	p	13052	9173	37545	75668	25836	21927
	q	1	0	0	0	0	0
HIC	p	11378	11985	42787	90730	22510	17823
	q	1	1158	586	2935	0	289
SMAC	p	11562	9023	33961	69858	21944	15063
	q	0	2	2	1	0	383
PSR	p	17317	13114	58020	118638	36754	33445
	q	1	0	1	0	1167	0
ACSINS	p	13354	7440	29737	63391	25395	18444
	q	0	0	0	0	1167	0
CIC	p	15548	10442	42327	80619	32345	22880
	q	38	90	812	1560	1875	1574
CSI	p	4608	2896	13560	24375	5924	4102
	q	0	0	0	0	0	200
FcRN-RT	p	17799	10494	43921	86816	38278	27800
	q	291	347	1234	2652	4248	6002
BVP	p	10120	6390	29570	55287	22168	13985
	q	0	0	0	0	2	3
DNP	p	11070	5939	24841	46712	25107	12951
	q	1	0	0	0	11	0
Fe.FVIII.2	p	31367	20234	81819	176026	66269	44311
	q	9427	6154	23284	56764	37201	17348
Fe.C3.2	p	20880	12057	45268	97088	45896	3465
	q	434	26	1	0	6862	0
Fe.LysM.2	p	31914	20790	85481	176069	69868	6286
	q	10576	7196	29855	66198	43862	0
ELISA	p	7438	4905	27922	49868	13409	11674
	q	0	0	0	0	0	60
Heme.FVIII.2	p	4631	1807	7315	13067	5501	3312
	q	0	0	0	0	0	0
Heme.C3.2	p	3850	1614	6930	11831	4679	3265
	q	0	0	0	0	0	0
Heme.LysM.2	p	6238	4107	16048	29370	9775	6286
	q	0	0	0	0	0	0
Heme.2	p	8433	6592	26117	53028	15216	12631
	q	0	0	0	0	0	0
FA.2	p	8613	6592	22442	41365	17446	8983
	q	0	5656	0	0	0	11
Hep.RT.3	p	8734	0	21477	37639	19923	7011
	q	0	5414	17	0	327	0

were roughly equal for Tm and Fe.FVIII.2. Noticeably, there are gaps in each of the plots at the CDR loops where no features from these residues have been used. This is especially true of residues in CDR1 and CDR3 of both the V_H and V_L domains. Thus, it can be concluded that the F-regression is identifying features from framework residues, rather than features from CDR loop residues to predict these properties, most likely because they are less variable and it is easier to learn patterns from framework encodings. This is very apparent from the high frequency of features selected from residues in Framework 1 of the V_L domain across all of the metrics examined, except for HIC. Interestingly, however, the only feature to be statistically correlated with HIC at the q-value level was found in CDR-H2.

4.2.5 Statistically Correlated Values May be Used for Prediction

For each metric, linear models were trained using all encodings and significantly correlated encodings ($p < 0.05$ and $q < 0.05$) if available. The predicted values for each antibody were calculated by Jackknife cross-validation (CV), and Spearman's rank correlation (ρ) was used to measure the predictive performance of the model.

Table 4.3 demonstrates that all models trained on all encodings were found to have poor performance for each experimental metric. Performance was then improved by selecting encodings that were statistically correlated with the experimental metrics ($p < 0.05$) however, in most cases a lower predictive performance was seen for encodings with higher statistical correlation ($q < 0.05$). This was seen in HIC, PSR, FcRN-RT and DNP metrics, but examples were found where better performance was found in the $q < 0.05$ significance level including Fe.FVIII2

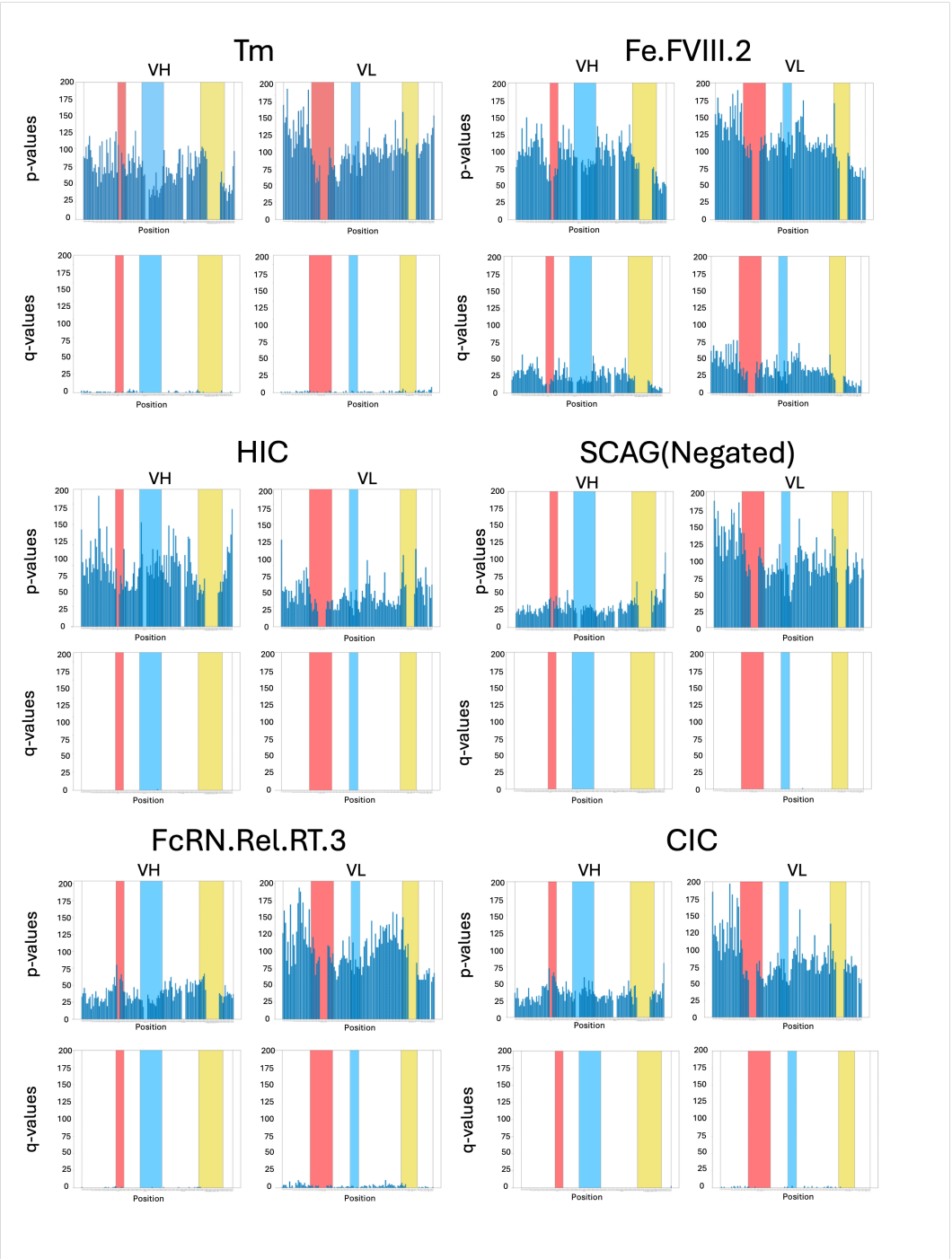

Figure 4.1: Correlated features to selected physicochemical properties across the antibody sequence. Bar plot of counts of significantly associated AntiBERTy encodings for each residue of $V_H$and $V_L$antibody domains along the Chothia numbering scheme for the selected metric dataset measured by [102]. CDR1 (red), CDR2 (blue) and CDR3 (yellow) are highlighted.

and Fe.LysM.2 (Table 4.3). It is easy to conclude that these relationships are due to more values being significantly correlated at the p-value level than the q-value level, however, for some features including Heme binding, models were not trained beyond random chance for any number of features, and so these are judged to be inadequate predictive power. The best examples of models trained were using p-value correlated encoding, included: Tm (Negated); SCAG (Negated); HIC; CIC and DNP which were all obtained using the encodings from the AntiBERTy LLM (Figure 4.2).

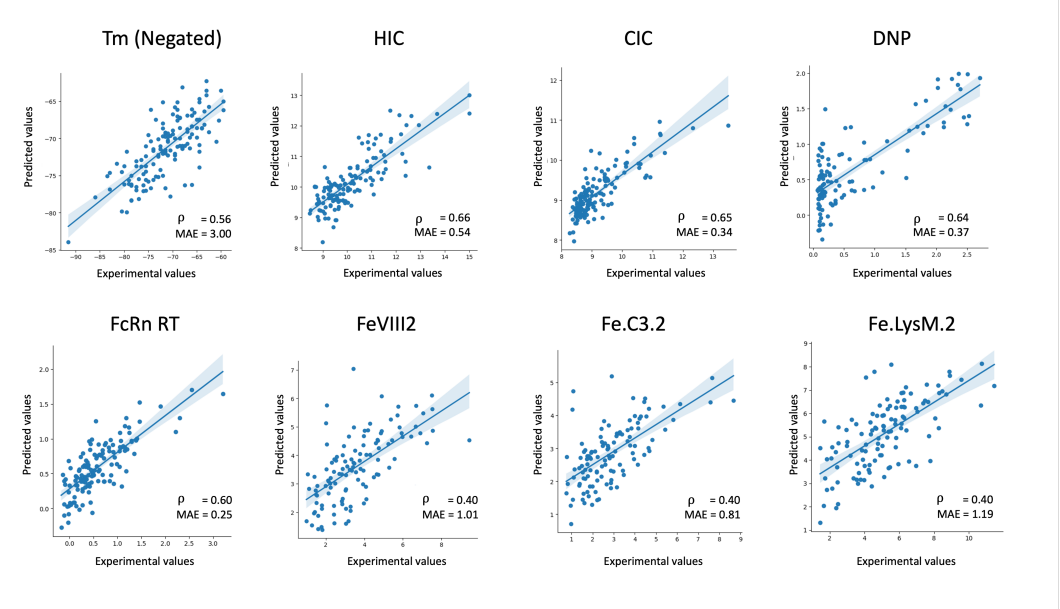

Figure 4.2: Scatter plots of linear models fitted to selected experimental metrics. Predictions were obtained by Jackknife sampling. Spearman's Rank Correlation (ρ) and mean absolute error (MAE) is given for each model. Blue hue indicates 95% confidence interval.

For both levels of significance, AntiBERTy had a greater overall ability to train linear models to predict these experimental values, and Sapiens had the poorest predictive power (where significant values were available). Individual examples of models better than those trained using the AntiBERTy encodings of the highest sta-

Table 4.3: Spearman's Rank Correlation (ρ) scores of linear models trained on experimental data each for language model and statistically significant ($p < 0.05$ and $q < 0.05$) features.

	Significance	AntiBERTy	ESM_8	ESM_650M	ESM_3B	AbLang	Sapiens
Tm(Negated)	All	0.13	-0.04	0.17	-0.06	-0.06	-0.13
	p	0.56	0.29	0.42	0.27	0.27	0.19
	q	0.43	-0.14	0.12	0.4	-0.45	0.24
AS	All		-0.36	-0.33	-0.29	-0.26	-0.41
	p		-0.04	0.01	0.03	-0.02	-0.07
	q		0.02	0.04	-0.09	-0.1	-1.04
SGAC(Negated)	All	0.05	-0.19	-0.12	-0.17	-0.08	-0.11
	p	0.58					
	q	0.16					
HIC	All	0.16	0.23	0.24	0.23	0.22	0.14
	p	0.66	0.45	0.47	0.46	0	0.43
	q	0.14	0.35	0.24	0.51	0	-1.35
SMAC	All	0.03	-0.14	-0.01	-0.09	-0.1	-0.15
	p		0.11	0.27	0.29	0	0.02
	q		0.11	0.15	0.15	0	-0.25
PSR	All	0.03	-0.31	-0.27	-0.26	-0.18	-0.16
	p	0.51	0	0.16	0	0.18	
	q	0.17	0	0.18	0	0.18	
ACSINS	All	-0.05	-0.35	-0.26	-0.37	-0.17	-0.28
	p					0.24	
	q					0.21	
CIC	All	0.26	0.15	0.17	0.16	0.23	0.11
	p	0.65	0.41	0.44	0.49	0.47	0.35
	q	0.44	-0.32	0.3	0.3	0.38	-0.22
CSI	All	-0.24	-0.58	-0.49	-0.6	-0.35	-0.47
	p						-0.21
	q						-2.38
FcRN-RT	All	0.17	-0.08	-0.01	-0.03	0.09	-0.04
	p	0.6	0.28	0.42	0.39	0.37	0.15
	q	0.26	0.19	0.07	-0.04	0.46	-0.36
BVP	All	-0.03	-0.33	-0.23	-0.29	-0.23	-0.33
	p					0.23	0.13
	q					0.3	0.25
DNP	All	0.01	-0.58	-0.62	-0.7	-0.22	-0.44
	p	0.65				0.24	
	q	0.2				0.42	
Fe.FVIII.2	All	0.15	0.04	-0.12	-0.02	0.01	-0.01
	p	0.4	0.2	0.04	0.14	0.19	0.26
	q	0.47	0.31	0.06	0.16	0.19	0.2
Fe.C3.2	All	-0.04	-0.2	-0.35	-0.25	-0.18	-0.19
	p	0.41	0.26	0.13		0.11	0.26
	q	0.46	0.45	0.24		0.25	-1.28
Fe.LysM.2	All	0.21	0.1	0.11	0.14	0.11	0.09
	p	0.45	0.38	0.3	0.32	0.22	0.29
	q	0.51	0.44	0.32	0.35	0.24	0.16
ELISA	All	-0.1	-0.43	-0.36	-0.37	-0.38	-0.36
Heme.FVIII.2	All	-0.13	-0.56	-0.44	-0.31	-0.39	-0.32
Heme.C3.2	All	-0.17	-0.41	-0.45	-0.35	-0.49	-0.36
Heme.LysM.2	All	-0.04	-0.23	-0.12	-0.05	-0.17	
Heme.2	All	-0.16	-0.5	-0.47	-0.56	-0.33	-0.5
FA.2	All	-0.09	-0.52	-0.44	-0.54	-0.52	-0.51
Hep.RT.3	All	0.1	0.12	0.08	0.13	0.16	-0.06

Missing values denote no features found at given significance level

tistical association ($q < 0.05$) were found for FcRN-RT using q-level associated encodings from AbLang (ρ=0.46, MAE=0.31) and HIC using q-value level encodings from ESM_3B (ρ=0.0.51, MAE=0.69) and (Figure 4.3). This demonstrates that the architecture of some models may be more advantageous than others in encoding particular experimental features.

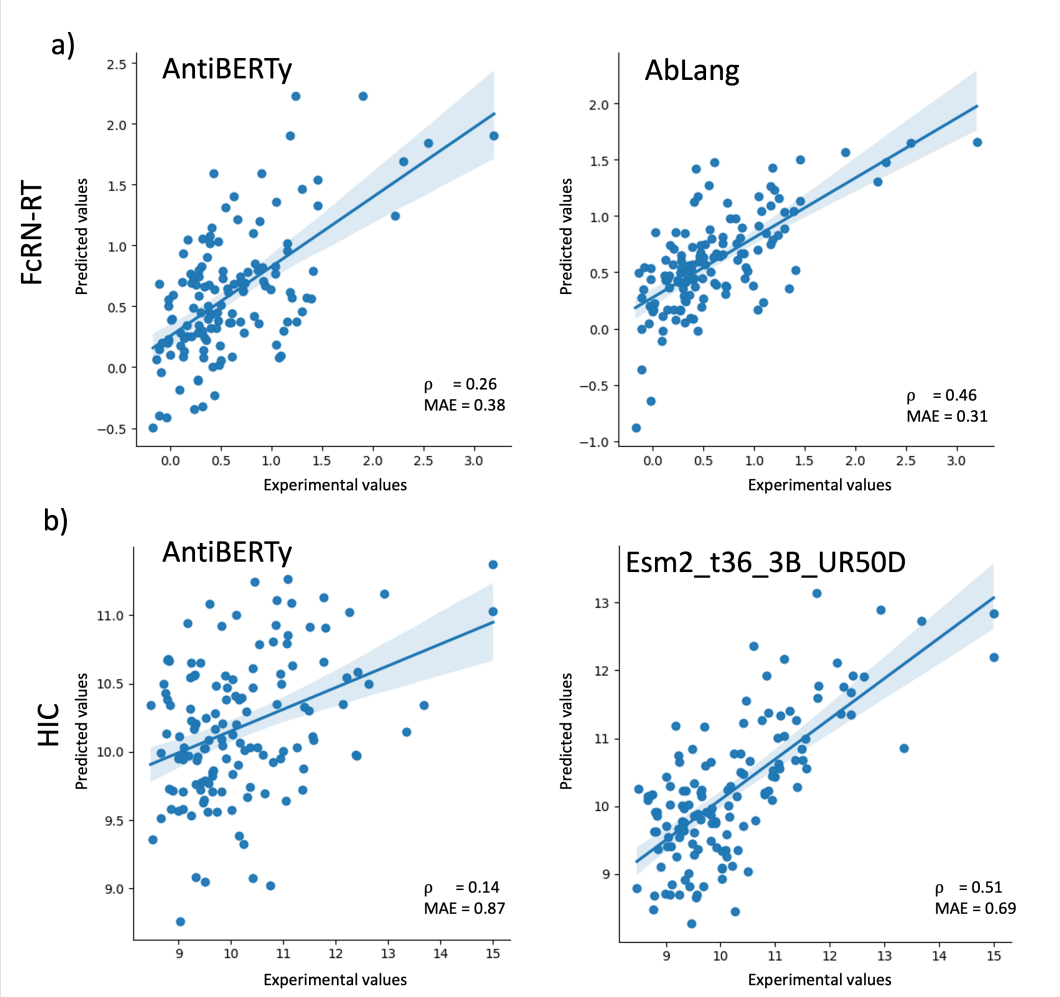

Figure 4.3: Scatter plots of comparing linear models fitted from different language models. a) FcRN-RT using AntiBERTy and the AbLang language models and b) HIC metrics from using AntiBERTy and the esm2_t36_3B_UR50D language models. Experimental training data was taken from [102, 103] for where predictions were obtained by Jackknife sampling. Spearman's Rank Correlation(ρ) and mean absolute error (MAE) is given for each model. Blue hue indicates 95% confidence interval.

4.3 Predicting Immunogenicity of Antibodies

4.3.1 Introduction

While antibodies are part of the human immune system, their exogenous origin makes them susceptible to patient immunogenicity, where anti-drug antibodies (ADAs) are generated as part of an anti-antibody response (AAR) [108, 1]. The most common pathways to this response are T cell-dependent mechanisms, including Th2 cell response, where fragments of antibodies are internalised by antigen-presenting cells and presented to T cells in the interaction of the MHC Class II molecule and the T cell receptor, releasing cytokines to raise an antibody response [44, 187]. Alternatively, B cell pathways have demonstrated a cross-linking of BCR with antibody drugs to generate the release of ADAs [188]. At best, the result of this pathway is the loss of efficacy of the drug as ADAs remove the therapeutic agent from the circulation and at worst, it can result in anaphylactic shock [45]. This can be a hugely costly obstacle to clinical trials, which this thesis aims to avoid. The most clinically relevant measure of immunogenicity is to examine the proportion of cohort patients who have raised ADAs to their treatment. However, this has not always been measured in trials, which in these cases, means that it is hard to separate failure from poor efficacy or failure for high immunogenicity [104, 188].

While, on average, antibodies of murine origin are more immunogenic to humans [104], examples exist of highly immunogenic fully human antibodies, adalimumab (ADA=28%) and golimumab (ADA=30%) [188]. On the other hand, there are chimeric antibodies where no ADAs have been observed: galiximab [189] and

futuximab [190]. Therefore, the assumption cannot be made that developing a fully human antibody would be non-immunogenic. For this reason, it has become important to develop tools with which to predict the immunogenicity of a new agent.

Although immune recruitment can drive immunogenicity [191, 192, 193], these risk factors can be mitigated by using an antibody isotype with perceived silent immunogenicity such as IgG4 [194], or by introducing silencing mutations in constant domains which interfere with the antibody binding to the $Fc\gamma$ receptor [195, 73]. These have shown an ability to reduce the immune response raised by these antibodies; however, it is suspected that these solutions are not always applied due to conflicting intellectual property and cost, so it is still necessary to investigate other drivers of immunogenicity that may be related to the V_H and V_L domains.

Therefore, it is important to investigate other features of the antibodies that are associated with immunogenicity. For example, the propensity for drug aggregation is believed to contribute to its immunogenicity as they form a large structure when aggregated in blood, which is more likely to interact with endogenous antibodies [95]. Although impurities in drug formulation may contribute to aggregation, more attention has been paid to the sequence and structure features of antibodies themselves [96]. This is likely to be driven by the presence of patches of hydrophobic residues the solvent-accessible surface of the protein [3]. Furthermore, sites for post-translational modification including glycosylation [97, 98]; deamidation [99] and oxidation [100] present risks in protein stability that could lead to aggregation or a higher chance of immune recognition through heterogeneity [101]. Furthermore, aggregation is a concern for the shelf life of a drug, so targeting this characteristic

would be an effective way to make antibody drugs more accessible [3].

This chapter aims to find predictors of antibody immunogenicity using the ADA dataset [104] available to find an acceptable cut-off for tolerated antibody immunogenicity.

4.3.2 Datasets

4.3.2.1 ADA Incidences

Immunogenicity data of therapeutics were taken from the supplementary material of Marks *et al.* [104] using data from Clavero-Alverez *et al.* [119], who report the mean ADA incidence of therapeutics. V_H and V_L sequences and clinical status of the specified antibodies were taken from the October 2021 release of TheraSabDab [110]. Some examples were excluded including bispecific antibody drugs, antibody drug conjugates and drugs which did not have sequences stored in TheraSabDab. These mAbs were then grouped by species origin which was established by a literature search. In total, the ADA data presented in this report represent 71 therapeutic human antibodies, 89 humanised antibodies, 19 chimeric antibodies and 8 murine, totalling 188 antibodies at varying stages of clinical approval.

4.3.3 Immunogenicity Scores of Clinical mAbs

The ADA data was examined to establish an appropriate cut-off point on which the highest immunogenicity (and therefore the highest ADA) rates are tolerated from approved mAbs. Hu-mAb is a software for predicting antibody immunogenicity, giving a score between 0 (mouse) and 1 (human), assuming that antibodies similar to human sequences would be less immunogenic [104] (see Section 2.4.3). Each of

the therapeutic antibodies collected here had their Hu-mAb score predicted using the software and plotted against their with ADA data collected from Marks *et al.*. The data was colour coded by stage of approval status and by species origin.

It is shown in Figure 4.4a that there is no clear separation in the data, contrary to what was previously hypothesised. Many approved mAbs show a high incidence of ADA, and while some highly immunogenic drugs are discontinued, it is also shown that some discontinued drugs have a no recorded incidences of ADA. In contrast, mAbs cluster much better by their source, as shown in Figure 4.4b. Human mAbs mostly but not always cluster with the highest Hu-mAb scores, regardless of whether they have a higher incidence of ADA. Humanised mAbs show more variability in scoring, whereas most chimeric mAbs score very poorly on the Hu-mAb scale, even if they have low immunogenicity. As expected, the five most immunogenic drugs are murine in origin and are correctly predicted as such being immunogenic. Despite this, a highly immunogenic humanised drug does not score poorly on the Hu-mAb scale, but an immunogenic chimeric does. No clear cut-off point for tolerated immunogenicity was found to separate approved and discontinued drugs.

4.3.4 Supervised Classification Methods for Predicting Immunogenicity

As no natural cut-off was observed for immunogenic and non-immunogenic mAb therapeutics, it was thought that binary classifiers of immunogenicity could be trained using arbitrary cut-offs for immunogenicity. Sequences of V_H and V_L se-

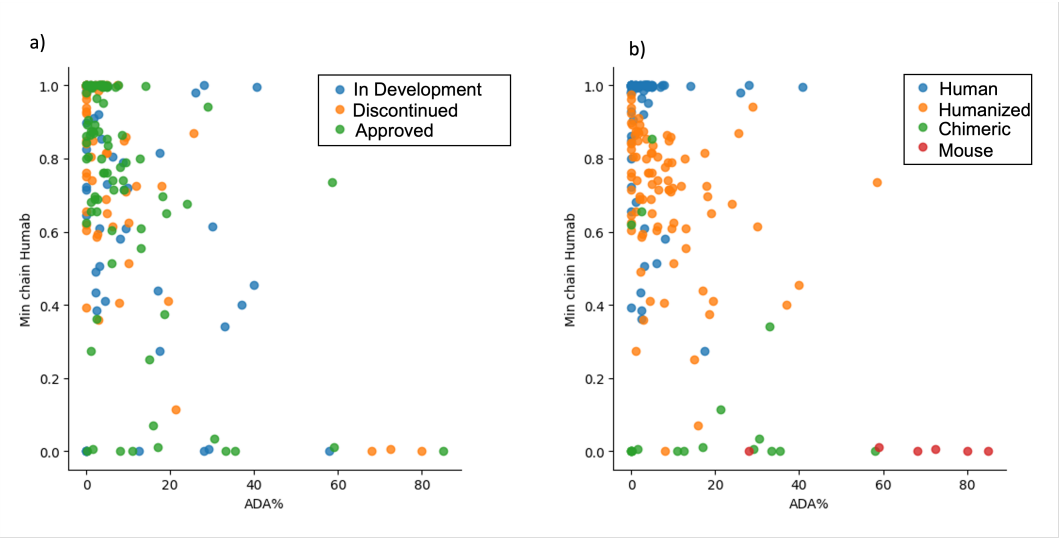

Figure 4.4: Scatter plots of Hu-mab [104] immunogenicity prediction scores against known Anti-Drug Antibody (ADA) incidence (%). Colour coded plots highlight a) Approval status as of October 2021 and b) Antibody origin.

Table 4.4: Group sizes according to ADA threshold split.

ADA Threshold	Non-immunogenic	Immunogenic
1%	61	136
2%	77	120
5%	128	69
10%	155	42

quences were encoded using residue level encodings, amino acid compositions and language model encodings as used in previous chapters and then divided by arbitrary immunogenicity thresholds 1%, 2%, 5% and 10%. 10% was considered the maximum, as only a minority of the data reports incidences of ADA above this value (Table 4.4).

4.3.4.1 Residue Level Encodings

Each split of the ADA dataset was encoded using the 14 methods of residue level encodings (see Section 2.5.1), which were concatenated together to give 65894 features per encoded paired sequence. These encoded sequences were used to train 15

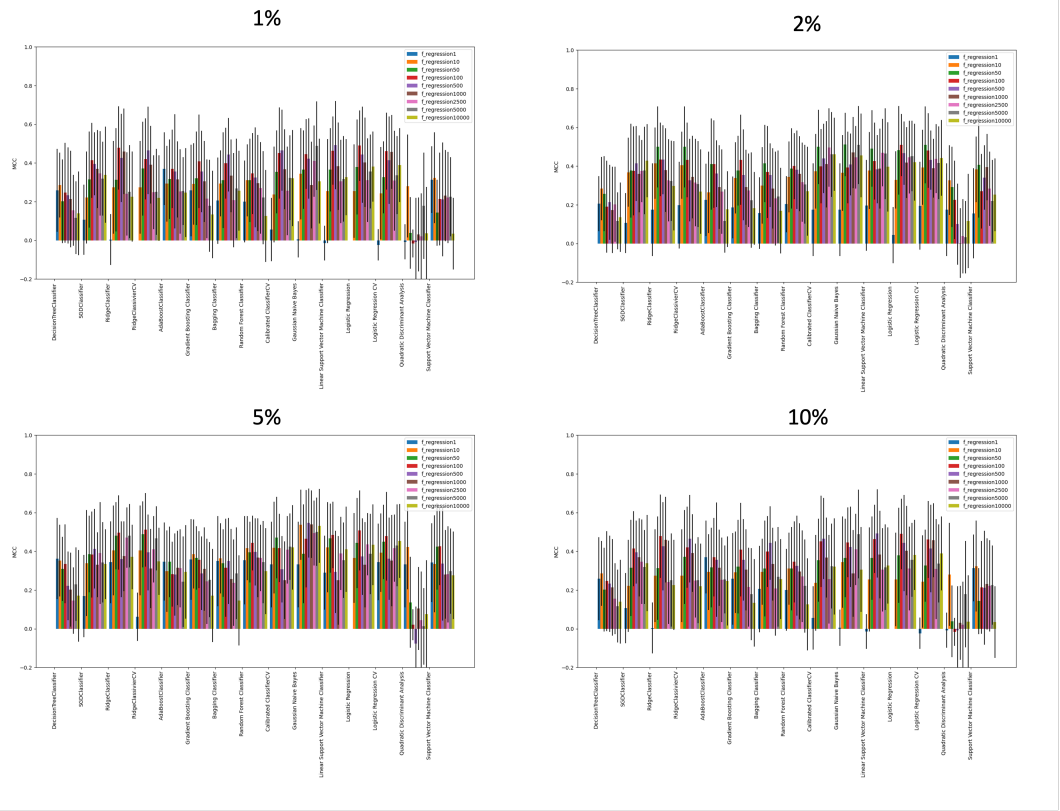

Figure 4.5: Classifiers trained against ADA scores cut offs for 188 therapeutics encoded with residue level encodings. MCC scores and standard deviation of 15 binary machine learning predictors with 10-fold cross validation classifying test split of immunogenic and non immunogenic clinical antibodies dataset encoded with 14 different residue-level encoding methods. Cutoffs for considering immunogenic and non-immunogenic are set at 1%, 2%, 5% and 10% ADA incidence.

different supervised machine learning classifiers with 10-fold CV at each threshold for the previously used values of features selected by the F-regression approach (k=[1, 10, 50, 100, 500, 1000, 2500, 5000]). Generally, performance was moderate and similar across all classifiers for all values of k (Figure **??**). Details of machine learning classifiers are given in Section 2.8.1.

4.3.4.2 Amino Acid Encodings

Amino acid compositions were additionally used in place of residue level encodings to see if these predictions could be improved. 19330 features were encoded using

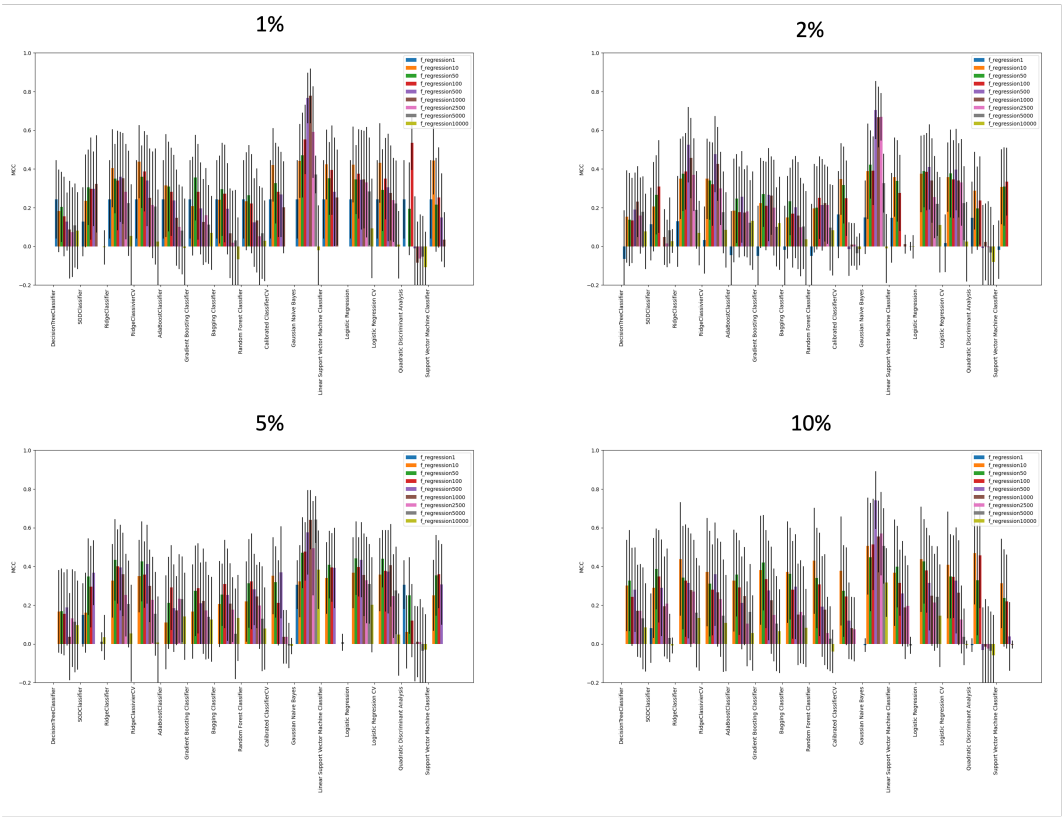

Figure 4.6: Classifiers trained against ADA scores cut offs for 188 therapeutics encoded with amino acid compositions. MCC scores and standard deviation of 15 binary machine learning predictors with 10-fold cross validation classifying test split of immunogenic and non immunogenic clinical antibodies dataset encoded with amino acid composition encoding methods. Cutoffs for considering immunogenic and non-immunogenic are set at 1%, 2%, 5% and 10% ADA incidence.

ProPythia software [154] (see Section 2.5.2). These encodings were used to train the same set of 15 machine learning classifiers with 10-fold CV. The best performance with these encoding methods was seen for the Gaussian Naive Bayes with all cut-offs, with different values of k for the feature selection. Arguably the best performance was seen at the 1% threshold (MCC = 0.77±0.14) where k=1000 and 10% (MCC=0.74±0.18) where k=500) (Figure 4.6).

4.3.4.3 Language Model Encodings

After the amino acid composition statistics, the same sequences were encoded with the AntiBERTy language model [179] to then train the 15 machine learning predictors with 10-fold CV. Similar performance was observed for SGDClassifier, Ridge Classifier, Ridge Classifier CV, Calbrated Classifier CV, LinearSVC, Logistic Regression and Logistic RegressionCV for 1% and 2% cut-offs, but general predictive performance decreased for the 5% and 10% cut-offs where all classifiers performed similarly. In the case of the 1% threshold, the best performance observed was around the same for the Ridge Classifier (MCC=0.73±0.17) where k=1000 than as the Logistic Regression CV at the 2% threshold (MCC=0.72±0.15) where k=500. It was seen that for higher thresholds, the MCC performance was seen to be poorer with larger standard deviations, as demonstrated by the best predictors from these thresholds. The best restuls were seen when the partition between immunogenic and non-immunogenic was set at 5% (MCC=0.63±0.18) with Logistic Regression and at 10% (MCC=0.58±0.25) with LinearSVC, both where k=500 (Figure 4.7).

To examine this further, the probability thresholds required to accept a positive prediction were adjusted from 0.5, 0.6, 0.7, 0.8 and 0.9 to see if a better MCC score could be achieved. However, it was not observed that this significantly improved the MCC scores (Figure 4.8).

Like before with the previous experimental metrics, the positions of the top k=1000 features from the encodings related to immunogenicity were searched for at each threshold. As expected the distribution of the 1% and 2% threshold look very similar to each other, but as the threshold for what is considered immunogenic

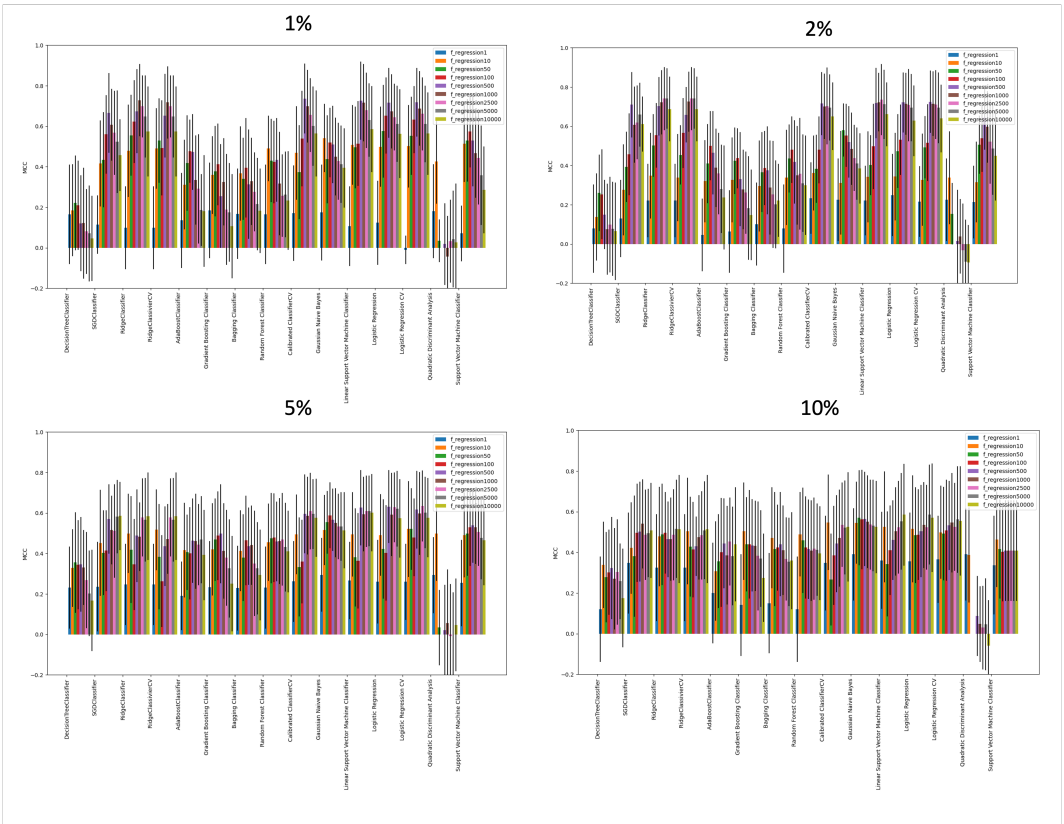

Figure 4.7: Classifiers trained against ADA scores cut-offs for 188 therapeutics encoded with the AntiBERTy LLM. MCC scores and standard deviation of 15 binary machine learning predictors with 10-fold cross validation classifying test split of immunogenic and non immunogenic clinical antibodies dataset encoded AntiBERTy Language Model. Cutoffs for considering immunogenic and non-immunogenic are set at 1%, 2%, 5% and 10% ADA incidence.

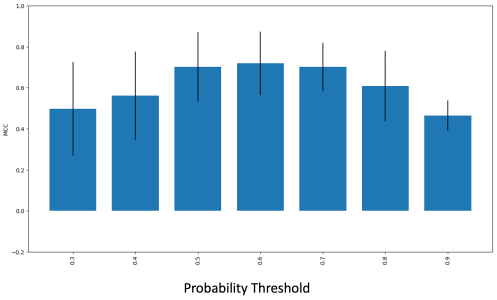

Figure 4.8: Using probability thresholds to improve MCC predictions. MCC scores and standard deviation of 15 binary machine learning predictors classifying test split of immunogenic and non immunogenic clinical antibodies dataset encoded AntiBERTy Language Model. Cutoffs for considering immunogenic and non-immunogenic are set at 1% over a set of thresholds to consider a positive value.

increases, there are fewer features taken from the V_L domain and an increase in features taken from the Framework regions of the V_H domain (Figure 4.9). This is especially true for Framework 1, 3 and 4 where there are a number of residues with high frequencies of features correlated with immunogenicity. Despite this, however, these could be species differences between low immunogenic human mAbs and highly immunogenic murine mAbs.

In general, the classifiers trained in this section demonstrate a strong discrimination of immunogenic and non-immunogenic sequences, particularly when set at the 1% ADA cut-off point.

4.3.5 G2Score

It was then thought a newer statistical approach could then overcome the binary classifiers by capitalising on the wealth of antibody sequence data available in online repositories. The G2score was devised as a similar statistic to GScore [109] but with a more extensive database of sequences corresponding to antibody germlines, enabling us to better capture the variation seen within them. We demonstrate its application in indicating the immunogenicity of human antibodies and predicting the incidence of ADA generation in a cohort of patients.

4.3.5.1 Development of the G2score

The dataset used to generate the original GScore [109] (Section 2.4.2) was limited by the relatively small dataset available in KabatMan [114], in some cases, very small numbers of proteins sequences were available from a given germline. With so many more sequences available from NGS, it was decided to reproduce this

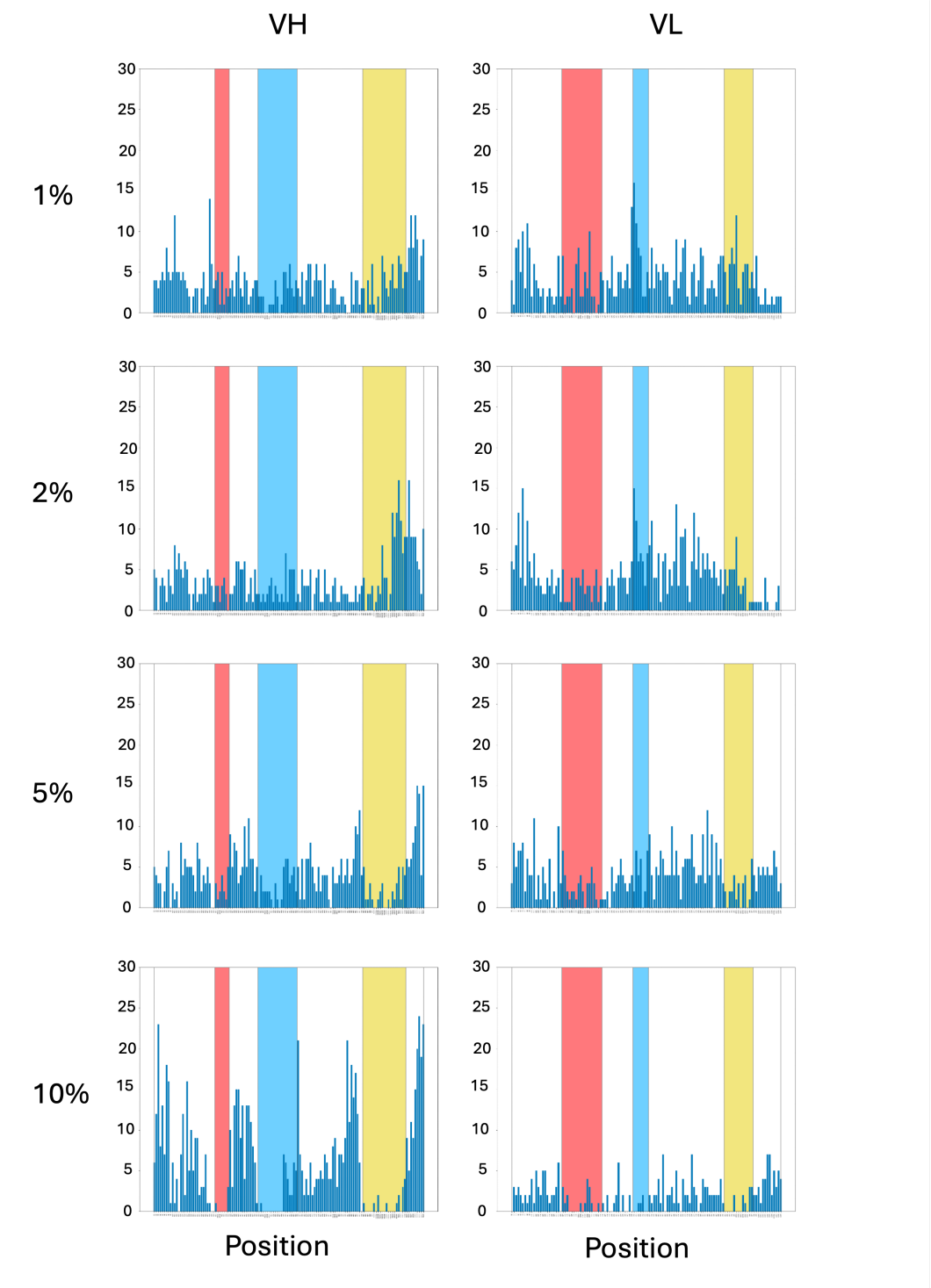

Figure 4.9: Frequency counts for features selected by F-regression (k=1000) at different ADA thresholds. Bar plot of counts of significantly associated AntiBERTy encodings for each residue of V_H and V_L antibody domains along the Chothia numbering scheme for the selected metric dataset measured by Jain *et al.* [102]. CDR1 (red), CDR2 (blue) and CDR3 (yellow) are highlighted.

statistic with human sequences from the OAS [118]. Firstly, datasets of unpaired heavy and light sequences were downloaded from the OAS and joined into a total of 34,768,256 V_L and 16,871,584 V_H human sequences representing 87 IGHV, 41 IGLV and 56 IGKV germline genes. Accession numbers for V_H domain sequences can be found in Data files 6 for V_H sequences and Data file 7 for V_L sequences. The corresponding germline genes for each sequence were found using AGL (see Section 2.3.8). The germline gene data sets were sorted by family, where families with a large number of representatives were reduced to a maximum of 10,000 by random sampling.

The G2Score was then made in the same way as the HScore and the GScore, but the alignments were carried out using BLAST [137], which gives a identity score for the target sequence against all other sequences for representatives of that germline family. For each germline, there is a mean identity score which is then used to calculate a Z score for newly input sequences.

4.3.5.2 G2Score Pipeline

For an input sequence, the germline family is found using AGL and it is aligned to every other sequence in the dataset for that germline family using BLAST, giving a similarity score. The mean similarity score is then calculated using the Z score equation (Equation 4.1).

$$Z = \frac{x - \mu}{\sigma} \tag{4.1}$$

- Z is the final G2Score
- x is the mean similarity score of the naive sequence to all other sequences in the germline family
- μ is the mean of means similarity score of each sequence in the germline compared to every other sequence
- σ is the standard deviation of the mean of means similarity score

4.3.5.3 Benchmarking the G2Score

Immunogenicity prediction scores for all antibodies of the ADA dataset were calculated and presented for Human, Humanised, Chimeric and Mouse mAbs, respectively, and shown as scatter plots. The minimum score out of their V_H and V_L domains with the assumption that if it were an immunogenic antibody, the minimum score would be more representative of this than taking the higher score or the mean of the two scores.

The G2Score shows an improved correlation coefficient score (r) compared with the original GScore for human mAbs (r=-0.23, r=-0.16, respectively) and HScore (r=0). Figure 4.10 demonstrates poor performance of Hu-mAb with human mAbs. Hu-mAb had the lowest prediction ability amongst human antibodies, but showed the best correlated predictive ability for humanised antibodies (r=-0.27). HScore (r=-0.03), GScore (r=-0.09) and G2score (r=-0.08) did not perform well when predicting the immunogenicity of humanised mAbs as no correlation of scores to immunogenicity was found with any of these metrics.

Despite G2Score being the most predictive of human antibody immunogenicity, the most immunogenic human antibody (Utomilumab: G2Score=0.58, ADA=40.7) received a higher G2 score, compared with less immunogenic antibodies. However, the ClusterResidues software for detecting patches of unusual residues found that this had three patches, which was the most number of patches found on an input antibody. The second most immunogenic antibody, (Nirsevimab: G2Score=0.00, ADA=28) had two patches, but it cannot be ignored that there are many low-immunogenic antibodies with two patches.

4.3.6 Predicting ADA Incidence using Regression Models

Another hypothesis was that if the incidence of ADA could be predicted using these encodings of the mAb therapeutics, it would be a clinically relevant prediction of antibody immunogenicity and more informative about the care needed when an antibody is entered into trials. This was first attempted using linear models, which are lightweight and fast to train. This was done by encoding the sequences from the ADA dataset with both the amino acid compositions, or AntiBERTy LLM encodings using features from those encodings which were statistically correlated to the ADA score to train these linear models.

4.3.6.1 Linear Models

Linear models were trained on the encoded sequences using the `sklearn.linear_model.LinearRegression` Python module. Performance of these models was evaluated in a jackknife fashion as described in the methods section of this thesis, where a prediction is given for each data point based on the model

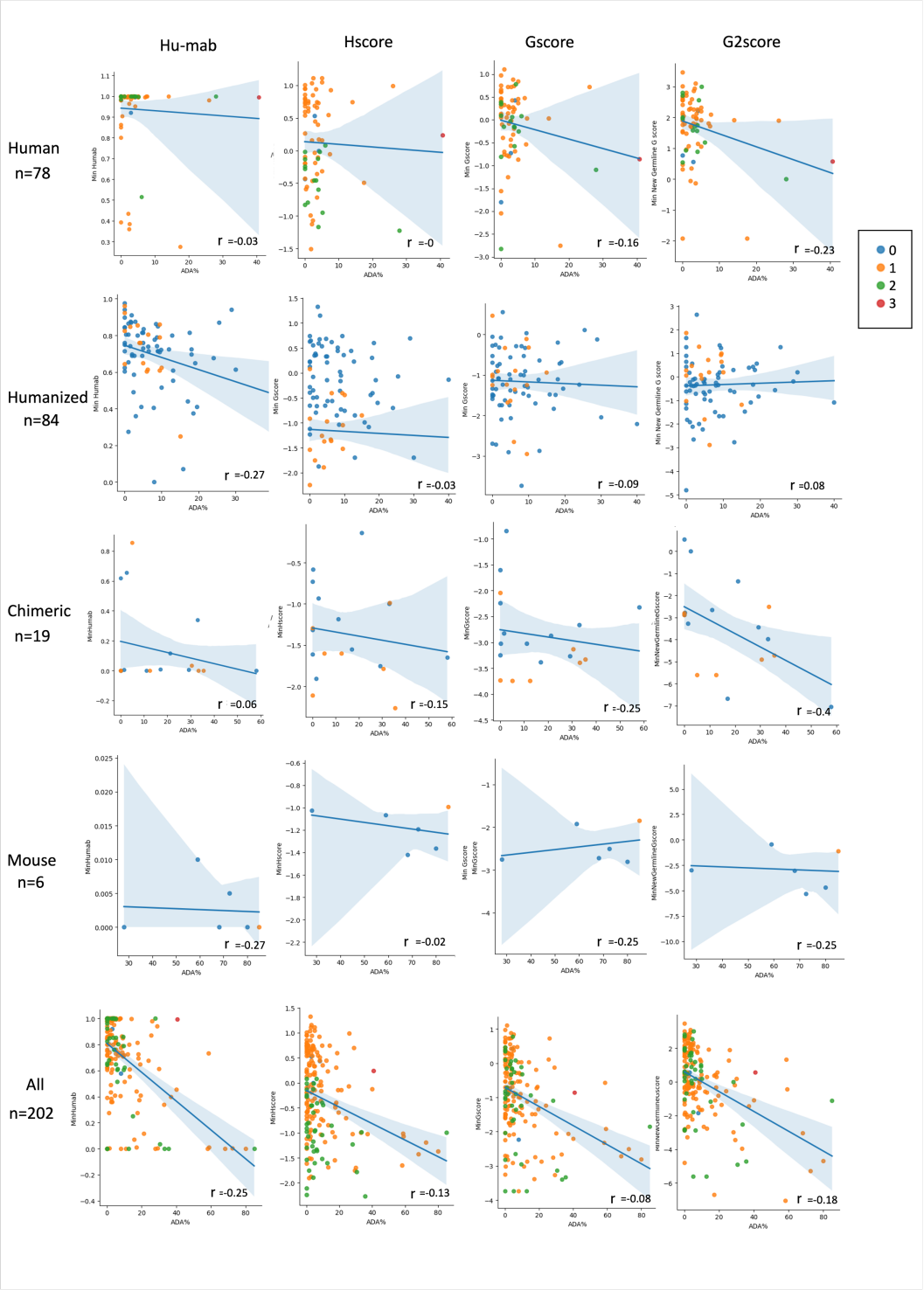

Figure 4.10: Comparison of immunogenicity prediction software for human, humanized, chimeric, mouse and all clinical antibodies. Scatter plots are given for each combination of predictor and antibody type with regression line plotted alongside showing correlation coefficient (r). Points show number of unusual patches found in each antibody (blue=0, yellow=1, green=2, red=>2). Blue hue represents 95% confidence interval across ADA% scores.

being trained on each other data point. Overall performance was measured with Spearman's rank correlation (ρ). When all features were used to train the linear model, poor performance was observed with amino acid encodings (Figure 4.11a), but improvement was observed when using AntiBERTy encodings (Figure 4.11b), especially when using statistically correlated features (ρ=0.34, $p < 0.05$; ρ=0.37, $q < 0.05$). This increase in performance was also seen to have the ability to predict the higher incidence of ADA in mouse antibodies but there was only a small performance increase between the p-value significant and q-value significant features.

It was also checked to see if removing mouse and chimeric antibodies from this dataset would improve these predictions, however, in the case of amino acid compositions, no observable improvement was seen. An improvement was seen in p-value significant AntiBERTy encodings (ρ=0.44; Figure 4.12), however this training dataset excludes many of the high incidence ADA mAbs that are important to learn from, and so including mouse and chimeric antibodies may result in a more generalisable model.

4.3.7 Deep Learning Models

It was checked if these predictions could be improved by applying a deep learning regression model to the problem. A basic model was established in pytorch to take an input layer of k nodes, where k was the number of features per data point there were 4096 nodes in the first hidden layer. Following that 6 additional hidden layers using the ReLU activation function, all halving in node number, were used until at the seventh layer, a single value is outputted from 32 input nodes. Other architec-

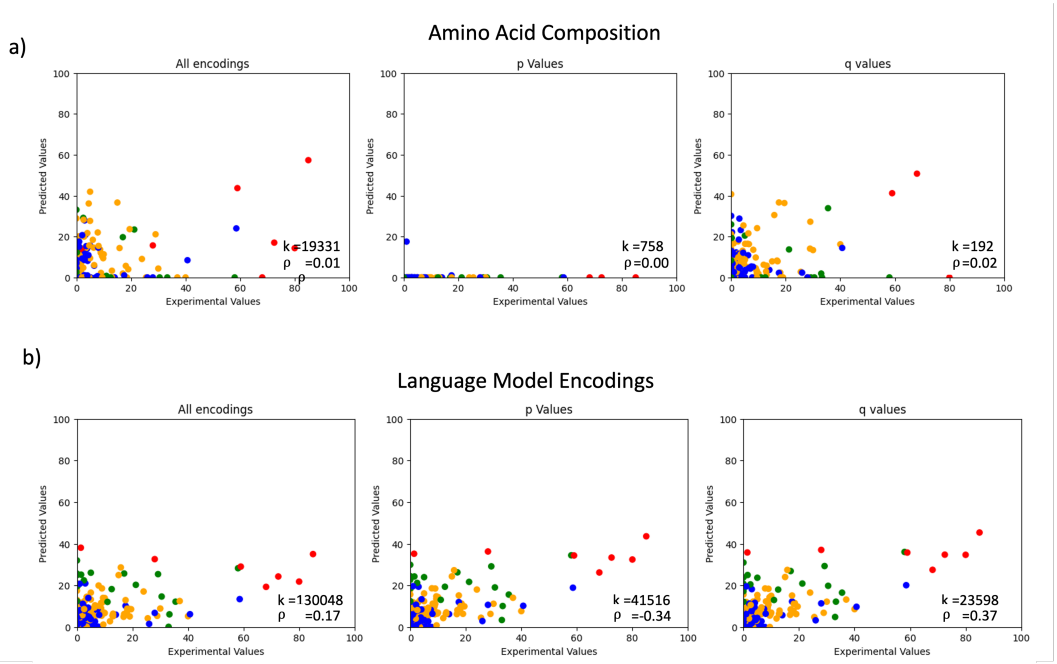

Figure 4.11: Linear models for ADA prediction. Scatter plots of jackknife predictions of linear models trained against all ADA incidence and a) amino acid encodings or b) language model encodings given for all data points in the encoding, all data points correlated with ADA incidence ($p < 0.05$) and all data points correlated with ADA incidence ($q < 0.05$).

tures including 5 hidden layers with more numbers of nodes were tried, however, the model with 7 hidden layers was found to be the best achieved at predicting.

This model was also used to predict the incidence of ADA using amino acid compositions (Figure 4.13a) and AntiBERTy encodings (Figure 4.13b). An ADA prediction was given for each data point using each model using cross-validation with 40 splits rather than Jackknifing, to be more time efficient. It was seen that performance given for language model encodings was on the whole better than that for amino acid compositions, however, in terms of correlation, it was not an improvement over the more simple linear models given. This would indicate that there is merit to this predictive model.

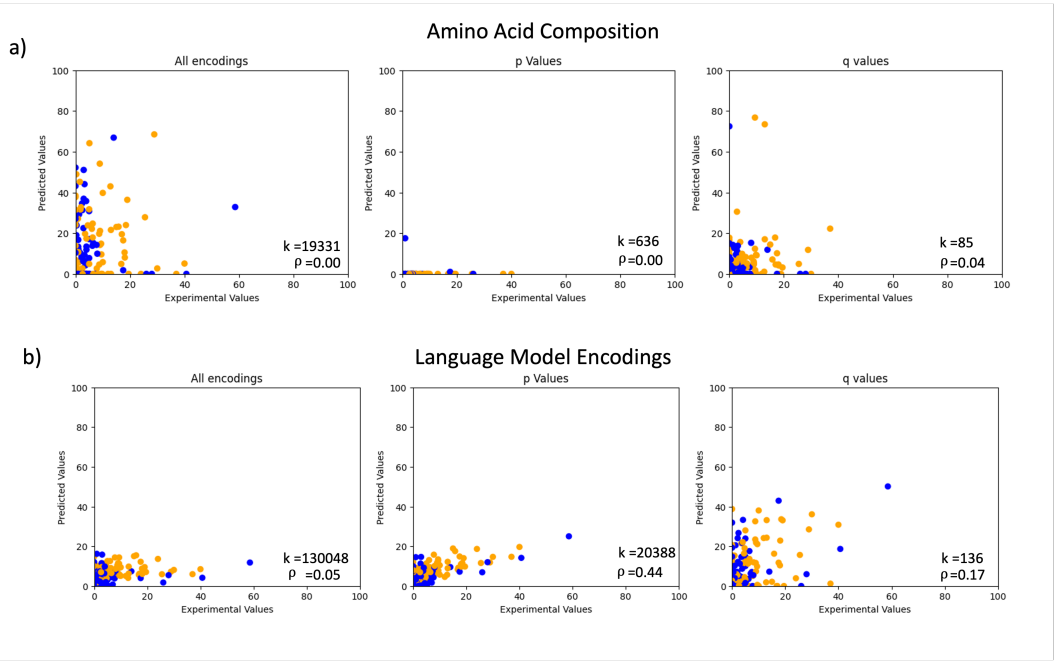

Figure 4.12: Linear models for human and humanised antibody ADA prediction. Scatter plots of jackknife predictions of linear models trained against all human and humanised incidence and a) amino acid encodings or b) language model encodings given for all data points in the encoding, all data points correlated with ADA incidence ($p<0.05$) and all data points correlated with ADA incidence ($q<0.05$).

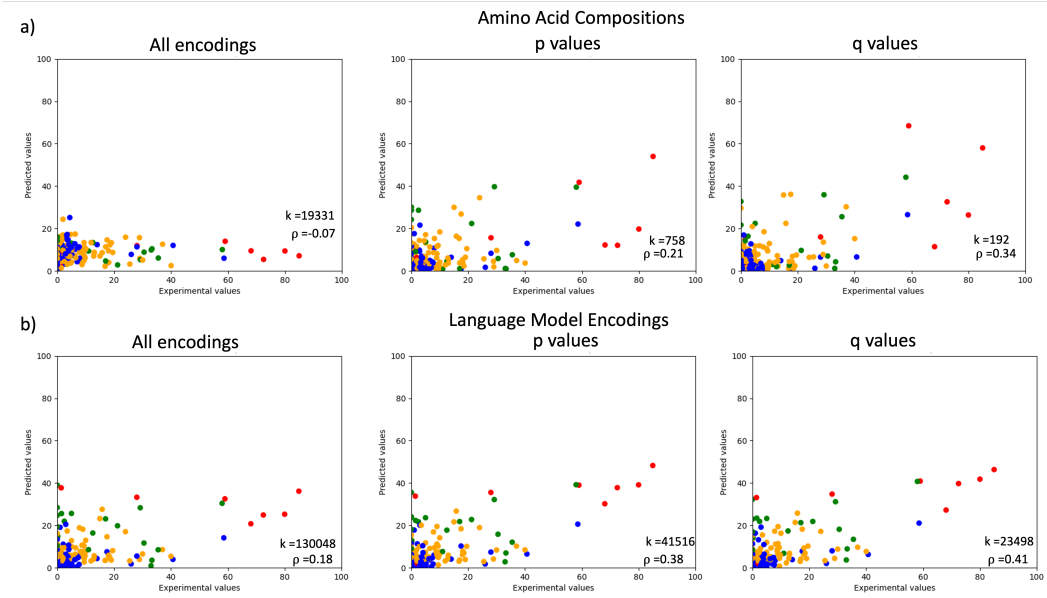

Figure 4.13: Deep learning models for ADA prediction. Scatter plots of predictions of Deep learning model trained against all ADA incidence data using sequences encoded by a) amino acid encodings or b) language model encodings given for all data points in the encoding, all data points correlated with ADA incidence ($p<0.05$) and all data points correlated with ADA incidence ($q<0.05$).

4.4 Discussion and Conclusions

This chapter has demonstrated that antibody LLMs can be used to predict physicochemical properties and immunogenicity of given antibodies which are relevant characteristics to developability. This was done by training predictive models on LLM encodings of mAbs where experimental data could be found, and have shown that they may outperform other training data using sequence-based statistics [107] or classifiers based on Random Forests [104]. From here these models can be made into useful tools for high-throughput screening pipelines used in drug candidate selection. While antibody LLMs require a larger compute power than sequence-based statistics, this can be increased when using GPUs and this computational cost can be outweighed by increased performance.

One of the greatest hurdles in this prediction with machine learning approaches is the lack of available physicochemical and ADA data for mAbs of the clinical stage and library antibodies [104, 102]. The data used in this chapter reflects only a small portion of WHO-INN-described mAbs [67], and may not be transferable to library antibodies, as it has been shown earlier in this thesis that clinical stage mAbs are selected dataset because they are expected to already satisfy the developability properties required. This prompts the need for standardising these metrics and gathering more data on more unusual and poorly developable antibodies to improve these kinds of prediction in the future. However, it is good that from the limited available data there is a breadth of ADA incidence rates among them.

4.4.1 Physicochemical Properties Predictions

The approach taken agrees with the sentiments of the AbPred authors that linear models are simple, with the advantage of interpretability, which gives additional support that there is a real relationship between these encodings and the experimental features [107]. However, the known bias seen in language models has meant that they have a tendency to revert to mean, and recognise features they have seen in training [196]. This would lead to the hypothesis that point mutations necessary for removing developability liabilities and lead optimisation may not affect the encodings enough to predict significant changes in these developability characteristics. This leads to the conclusion that these models work well to reduce the need for but not entirely replace, experimentation to discover antibodies suitable for therapeutic use.

Because these models were only trained on clinical stage mAbs which mostly occupied a narrow range in a given metric, it was seen that there was a tendency to overfit. This notion is supported by the observation that the models underestimated extreme values because few of them have been included in the dataset. This was also observed by Hebditch and Warwicker [107]. Although an obvious solution to this would be to add experimental values for antibodies calculated by other papers, these datasets are not directly comparable as stated in Licari *et al.* [197], different experimental conditions, including pH and salinity, can yield different results for given properties due to conformational changes that expose or connect small hydrophobic surface patches. Together, this incentivises the need not only to publish more varied experimental data which include negative examples, but also to stan-

dardise protocols so more generalisable models can be trained with less error.

Although models with high predictive correlation scores could be trained using p-value level encodings, the results warn of cases where q-value level encodings could not be found for that metric. Because multiple tests must be used to find these relationships and language model encodings have hundreds of thousands of data points, q-value significant encodings indicate a real relationship between those encodings and the experimental value. Using encodings where there is no indication of a real statistical relationship could result in an overfitted model based on encodings that happen to be correlated by chance. For this reason, allowances were made for predictors which are trained on all p-value level encodings when these include encodings which pass $q < 0.05$ statistical threshold.

4.4.2 Discussing Immunogenicity Prediction

In this chapter, it was attempted to establish an estimated ADA incidence 'cut-off' for where immunogenic drugs are no longer tolerated, however this was not observed. Different methods of identifying immunogenic antibodies were then devised through binary classifiers, similarity statistics and predictive models to predict the ADA for therapeutic antibodies.

As expected, the highest ADA incidence were seen to come from mice [104]. Moxetumomab [198] and Racotumomab [199] are murine mAb drugs for the treatment of tumours where patients will likely be placed on other immunosuppressive drugs and treated acutely and may be less likely to mount AAR (Moxetumomab was discontinued in 2023 for lack of sales, not efficacy or safety reasons [1]). This is in

[1] https://www.onclive.com/view/astrazeneca-to-discontinue-moxetum

contrast to Adalimumab, a human antibody used for chronic treatment of rheumatoid arthritis where there is likely to be an over-activation of the immune system at the site it is needed [200]. From this it was concluded that the clinical context of a drug has a large bearing on how much the immunogenicity will affect its action. This includes patients who are immunosuppressed or have overactive immune systems, which could affect how readily they would mount an AAR. Additionally, species differences can influence these models and so more examples of highly immunogenic human antibodies would greatly improve the reliability of these predictors.

This led to the selection of arbitrary cutoffs to establish immunogenicity with binary classifiers, where the best performance was observed in classifying antibodies where the immunogenicity AAR threshold was set at 1% or 2%. However in all cases, large standard deviations across CV folds indicates that the performance of the classifier may have a heavy reliance on the training dataset.

In the case of the statistical approach, the newly developed G2Score showed an improvement in the ability to score immunogenic human antibodies with a lower score than the previously established GScore statistic and Hu-mAb [104]. Despite this, since the G2Score is a relative score, choosing a cut-off where there is no tolerated cut-off is a difficult task.

The third approach taken was to use antibody encoding methods to train linear models to predict ADA incidences in antibodies. Moderate correlation between predicted and experimental values was found when the p-value associated data points

omab-pasudotox-in-us-for-hairy-cell-leukemia

were used to train a linear model, however, because of the dataset being heavily skewed to lower immunogenicity antibodies, these models were poorly predictive of highly immunogenic antibodies, and so the binary classifier approach was preferred.

The use of binary classifiers was the most successful approach to predict antibodies that would have and incidence of ADA above a given threshold. In this case, the most consistent performance was observed for the 2% threshold. This leads to the hypothesis that the characteristics linked to immunogenicity are encoded with the AntiBERTy language model. For the purposes of the pipeline constructed throughout this thesis, the results suggest that this kind of binary classifier is more informative than using a score because we demonstrate that it has similar performance for antibodies of human, chimeric, and murine origin unlike Hu-Mab [104].

4.5 Conclusion

To conclude this chapter, the ambition of using experimental datasets to predict physiochemcial properties of antibodies can only be achieved by collecting datasets which demonstrate a range of values for given metrics so, which would give more generalisable and less overfitted models. Having said that, this chapter has demonstrated that the emphasis of how much these metrics are taken into account for a given mAbs is dependent on the clinical context where it is used. For this reason, binary immunogenicity classifiers for predicting immunogenicity and linear models predicting features linked to developability can be added as an optional step to the

pipeline constructed in the thesis to identify antibodies with clinical stage properties.

Chapter 5

Separating Approved and Discontinued Clinical mAbs

5.1 Introduction

Once a lead clinical candidate with suitable developability properties has been identified, there is no guarantee it will be successful in clinical trials. This has been shown throughout the thesis in that mAbs used from different clinical stages share a narrow range of developability properties, and yet some have failed at clinical trials and others were successful. Therefore, an additional set of experiments was conducted in order to investigate whether predicting success at clinical trials could be done. Using the previously established method for encoding these sequences, it was hypothesised that there could be a difference learnt between market-approved mAbs and mAbs which were discontinued during clinical trials.

This chapter sees binary classifiers trained to classify approved and discontinued clinical mAbs, and an investigation as to how these classifiers were able to do

so. What follows is a detailed characterisation of the physicochemical properties of the approved and discontinued groups and identifying what features the models are learning, because this has application in learning what antibodies will pass clinical trials in the future.

5.2 Datasets

5.2.1 Approved and Discontinued Antibodies

V_H and V_L sequences of 115 approved mAb drugs and 154 discontinued mAbs labelled 'Whole mAb' were collected from the October 2021 release of the TheraSabDab database [117]. Unlike previous chapters in which only human-derived mAbs were selected, in this case mAbs of all sources were selected to maximise the training dataset. A literature search was carried out to establish the reasons for discontinuation for all of the discontinued mAbs. Eight drugs were discontinued for marketing or financial reasons, and removed from the discontinued dataset because these could potentially be mAbs that otherwise would have passed clinical trials. Edrecolomab was also removed from the approved dataset and added to the discontinued dataset because this was withdrawn for efficacy reasons [201]. The result of this is a dataset of 115 approved antibodies (Data File 2) and 147 discontinued antibodies (Data File 9). The excluded sequences and sources for their basis of exclusion are found in Table 5.1.

Table 5.1: Discontinued clinical stage with non-clinical reasons for discontinuation.

Therapeutic	Source
Opicinumab	Fierce Biotech
Sifalimumab	Guide to Pharmacology
Vorsetuzumab	Creative Biolabs
Birtamimab	Biopharma Dive
Epratuzumab	GenEng News
Fulranumab	JNJ
Zanolimumab	Fierce Biotech

5.2.2 Held Back Dataset

From the October 2023 release of TheraSabDab, newly approved and discontinued "Whole mAb" human therapeutic mAbs that were not found in the October 2021 release were used as a held back dataset for validation. This included 10 approved mAbs (Data File 10) and 11 discontinued mAbs (Data File 11).

5.3 Encoding Amino Acid Sequences for Machine Learning

A supervised machine learning approach to this problem was taken which would train a model to separate market-approved (class 1) and discontinued (class 0) mAbs. It was expected that this approach would NOT be successful because clinical mAbs clustered closely in the unsupervised learning method, and because there are so many reasons an antibody may fail in clinical trials, many of which are influenced by external factors. However, it seemed necessary to predict this success for the pipeline because including these discontinued clinical mAbs at previous stages could potentially introduce liabilities to the pipeline that should be removed at later stages. V_H and V_L sequences were spaced according to the Chothia numbering

scheme and then encoded using residue level encodings, amino acid compositions and language models for training with machine learning models.

5.3.1 Residue Level Encodings for machine learning

Residue-level encodings were performed using the SequenceEncoding software obtained from Jing *et al.* [202] (see Section 2.5.1). The program works through parsing through a sequence, and for each residue, the software looks up a stored array of values corresponding to that residue, or that residue and those around it for the specified encoding method. The encoding methods used by this software are described in Table 2.2. The advantage of using multiple methods of encoding is that a diverse set features may be taken into account, including features that have already been investigated including pI (Meiler Parameters), thermostability (Kidera Factors) [144] as well as hydrophobicity (hydrophobicity matrix) [147] and binary encodings (One Hot Encoding) [202].

15 supervised machine learning models (see Section 2.8.1) were trained on each encoded dataset to classify approved and discontinued mAbs. After 10-fold CV, good performance was not seen with any of the encodings used (Figure 5.1). All predictions have standard deviation error bars that pass through MCC=0, which can be interpreted as the predictions being no better than random chance. This method was attempted again using all the encoding methods joined together. This gave 65894 numerical features for each antibody. Because the number of features was much greater compared with any of the other methods used, feature selection through F-regression was used to select the top k = [1, 10, 50, 100, 500, 1000, 2500,

5000, 10000]. The result was an increased performance, with a top mean performance (MCC=0.55±0.4) for the Gaussian Naive Bayes model with the selection of 1000 F-regression features (Figure **??**). However, while this is the mean highest performance, the standard deviation is 0.4, which is a larger spread, indicating that this performance is highly dependent on the test-train split and not a trustworthy result. From the experiment given here, it is likely that these methods of encodings using preselected values for each amino acid do not allow sufficient data encoding necessary information for training. Next, more dynamic methods of generating encodings were tried.

5.3.2 Amino Acid Compositions

Another method of encoding protein sequences for machine learning was by amino acid composition statistics. Amino acid compositions for V_H and V_L sequences were individually encoded using the Propythia software [154] (see Section 2.5.2) and concatenated to give a total of 19330 encodings per paired antibody sequence. The best performing models were given by the GaussianNB classifier, particularly with F-regression k=1000, (MCC=0.88±0.09) (Figure 5.3). This model appeared to have much greater predictive performance than any of the other algorithms used in this training, and a large improvement in predictive performance from the residue level encodings method (MCC=0.55±0.4) despite the amino acid compositions method having fewer features to use. Overall, there is an improvement in prediction performance with other classifiers compared to the residue level encodings.

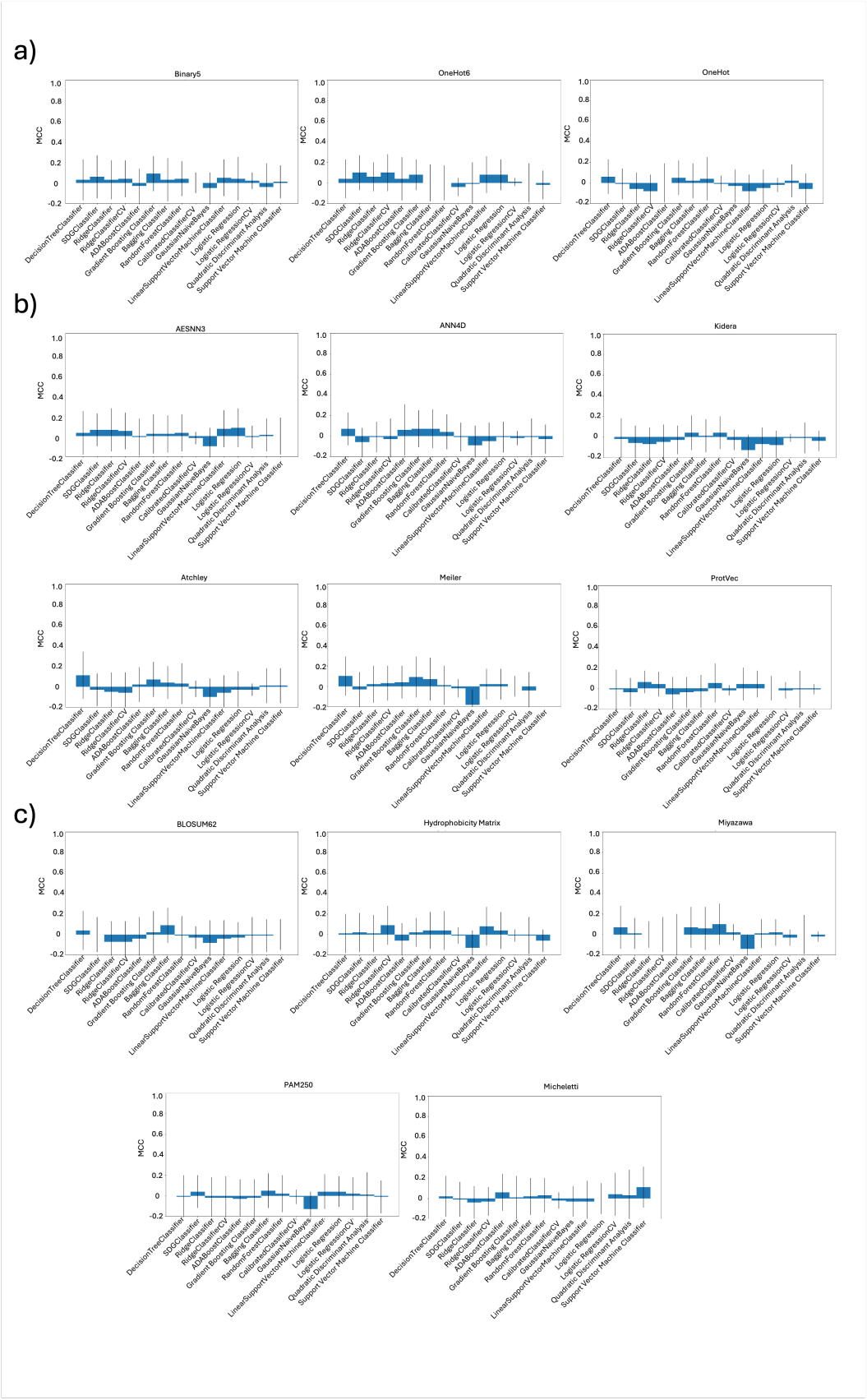

Figure 5.1: Classifiers trained on market-approved and discontinued mAbs encoded with 14 different residue level encodings. MCC scores and standard deviation of 15 binary machine learning predictors classifying test split of approved and discontinued therapeutic antibodies dataset. Charts have been split between a) binary encodings b) physicochemical property encodings and c) interaction matrices.

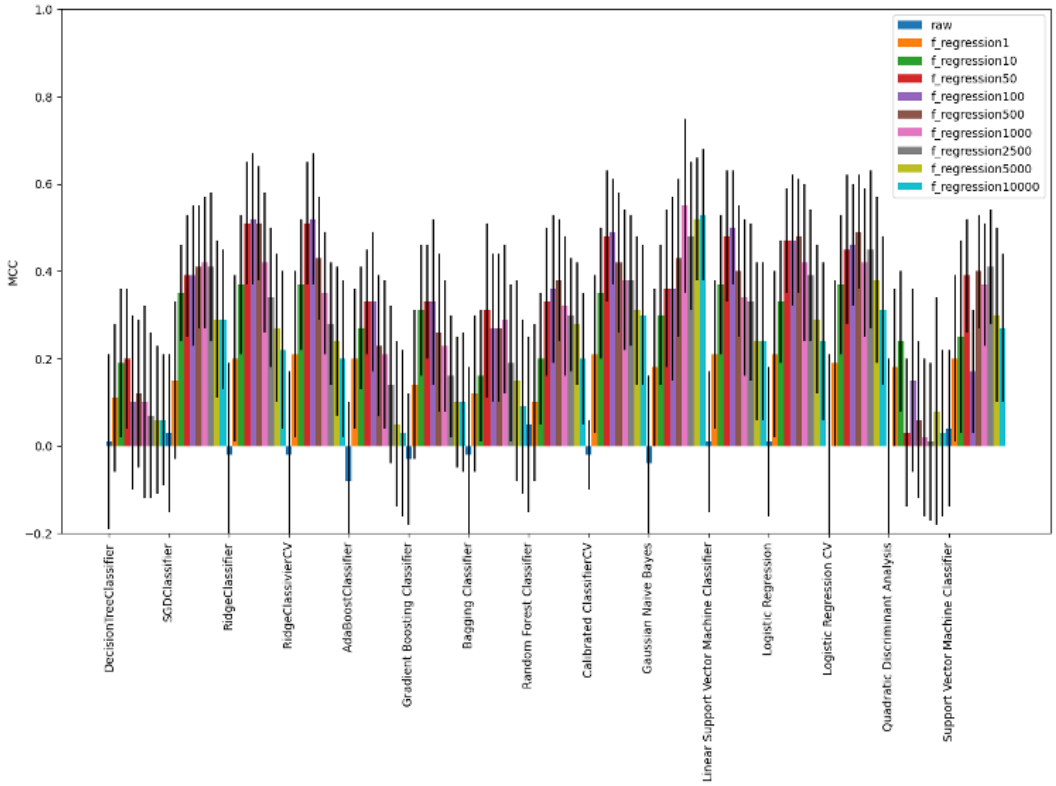

Figure 5.2: Classifiers trained on market-approved and discontinued mAbs encoded with 14 concatenated residue level encodings. MCC scores and standard deviation of 15 binary machine learning predictors classifying test split of approved and discontinued therapeutic antibodies. F-regression thresholds are colour coded.

5.4 Language Model Encodings for Supervised Machine Learning

Despite good performance with amino acid composition encodings, it was decided to use protein and antibody LLMs. Protein language models are an additional method of encoding sequences into a numerical representation; however, the feature space is much denser and more complex than any of the previously used methods of encodings. It was hypothesised that this feature space could encode features that would distinguish between mAbs that have been successful or unsuccessful at clinical trials.

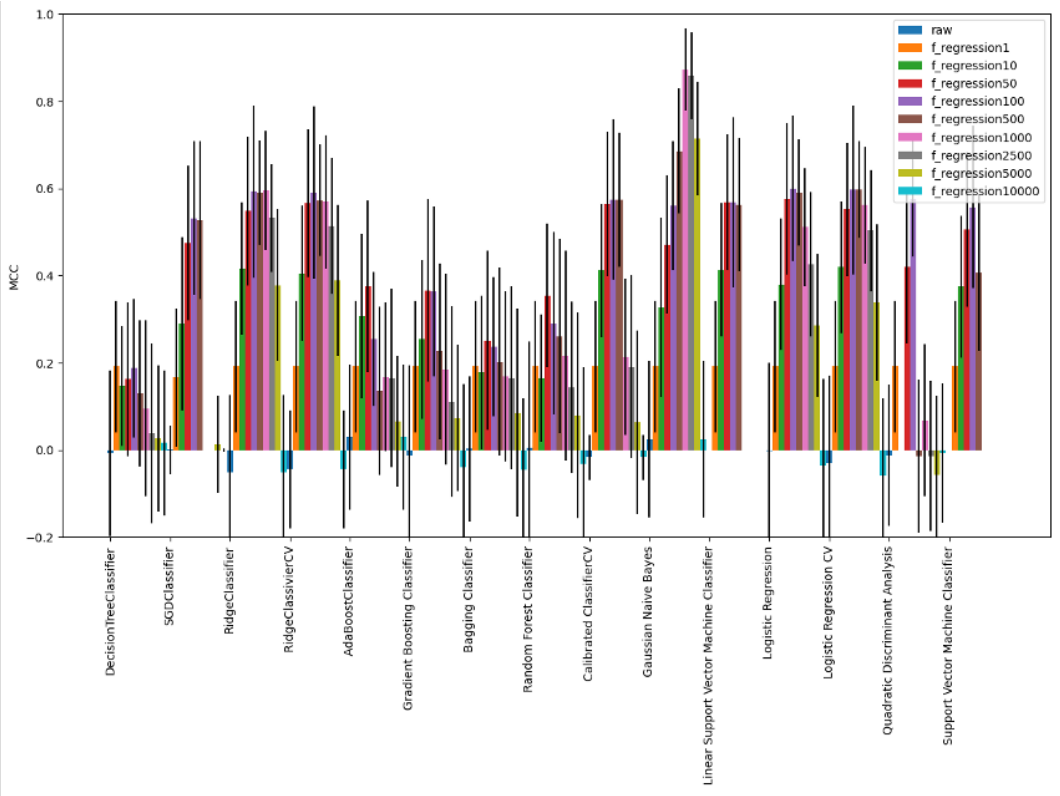

Figure 5.3: Classifiers trained on market-approved and discontinued mAbs encoded with amino acid compositions. MCC scores and standard deviation of 15 binary machine learning predictors classifying test split of approved and discontinued therapeutic antibodies dataset encoded with amino acid compositions. F-regression thresholds are colour coded.

The same selection of supervised machine learning classifiers was used to classify approved and discontinued mAbs for each language model encoding method with the same set of values of k for the F-regression. Generally, performance across all classifiers was good, and it was seen that best overall performance was obtained for the AntiBERTy encodings. Of these, the best performing model was LinearSVC (MCC=0.80±0.1), Ridge Classifier CV (MCC=0.78±0.12) and Logistic regression (MCC=0.80±0.1) when F-regression was set to k=2500 (Figure 5.4).

The LinearSVC model was selected as the best model with a mean sensitivity and specificity of Sn=0.92 and Sp=0.93 respectively across each CV split at the de-

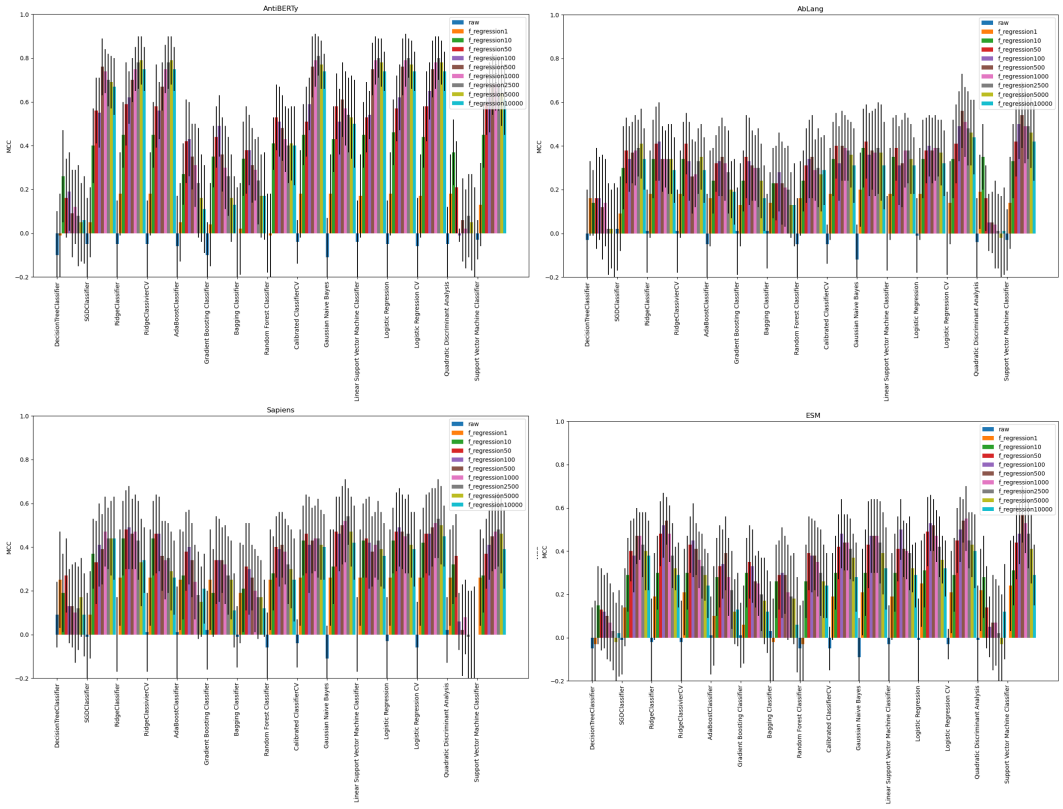

Figure 5.4: Classifiers trained on market-approved and discontinued mAbs encoded with protein LLMs. MCC scores and standard deviation of 15 binary machine learning predictors classifying test split of approved and discontinued therapeutic antibodies dataset encoded with four protein language models. F-regression thresholds are colour coded.

fault probability threshold (0.5). As an experiment, the model was assessed in a case where a higher probability threshold (0.8) was used to accept a positive result. This did result in a loss in mean sensitivity and a small increase in specificity mean specificity, given as Sn=0.57 and Sp=0.98, resulting in a decreased MCC=0.61±0.16. Confusion matrices for the raw outputs of this model at both probabilities can be seen in Figure 5.5.

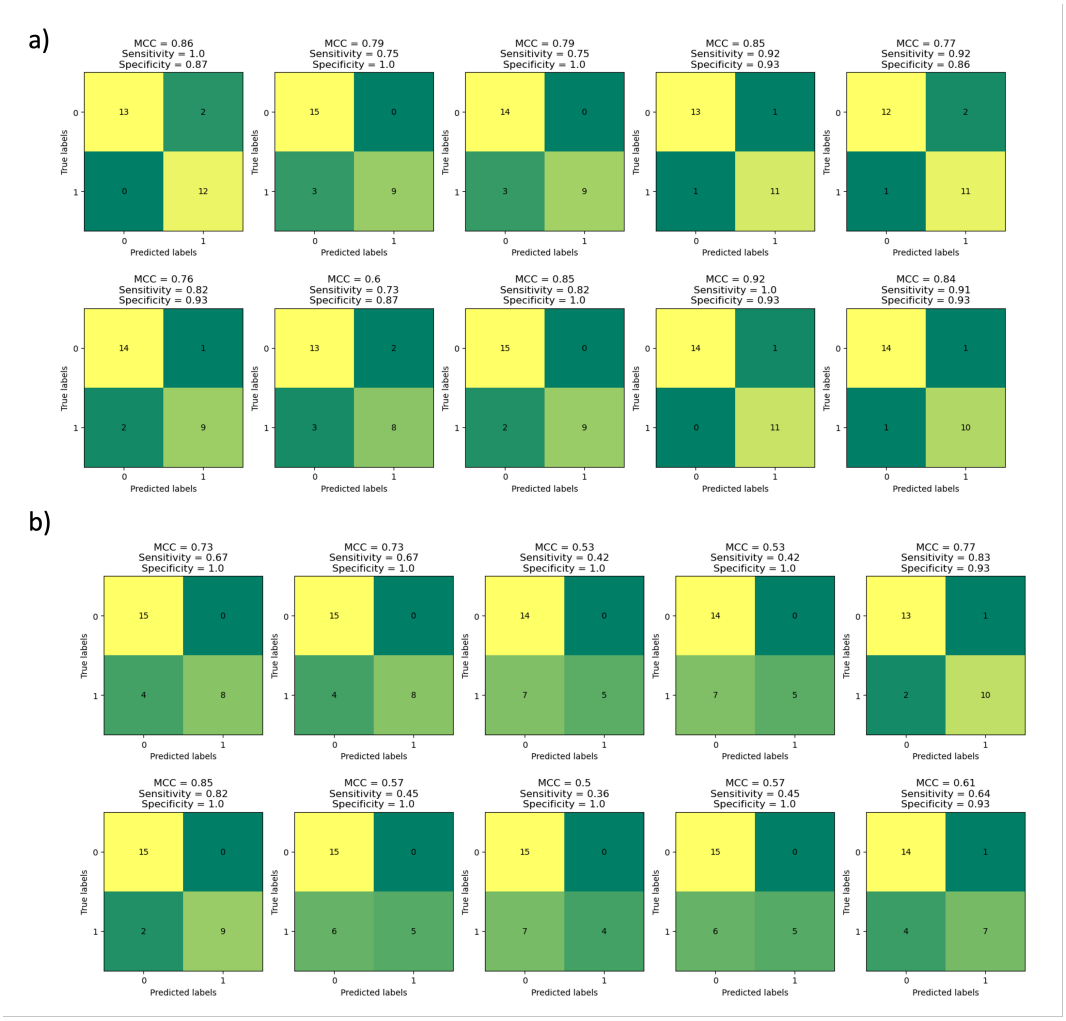

Figure 5.5: Confusion matrices for each of 10 split of a dataset of approved (class 1) and discontinued (class 0) antibodies. mAbs were encoded with the AntiBERTy language model and trained using LinearSVC using a threshold was of a) 0.5 or b) 0.8.

5.5 Locating Features Across V_H and V_L Domains

It was then investigated whether there are concentrations of features selected by the F-regression in particular areas of the sequences. Because CDR-H3 is mostly associated with binding affinity, it was hypothesised that this would be where the majority of characteristics were selected. The location of the features was found using the method outlined in Section 2.11.2 (Figure 5.6). For k=1000 and k=2500 it was seen that CDR-H3 demonstrated some favouring of selected features on the

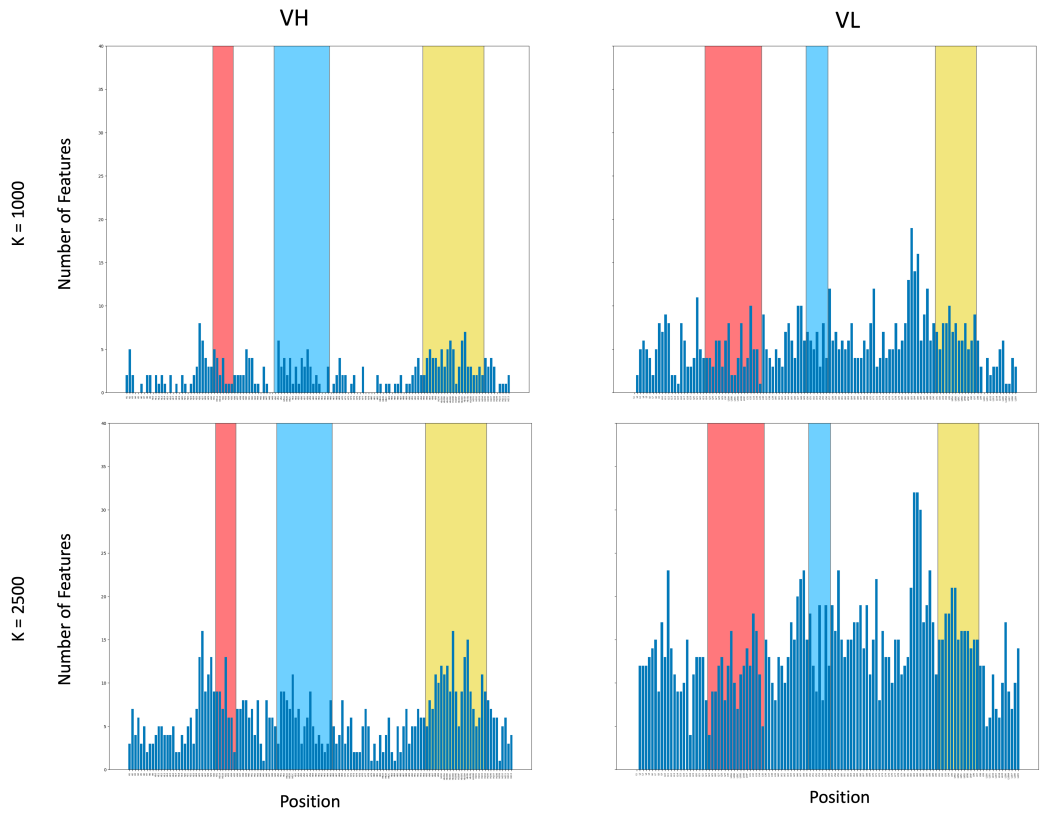

Figure 5.6: Locations of selected AntiBERTy features across V_H and V_L domains of approved and discontinued mAbs. These are given for values of k=1000 and k=2500. CDR loops are highlighted in red (CDR1), Blue (CDR2) and Yellow (CDR3).

V_H domain, particularly for *k*=2500. However, in both cases, more than twice the number of features selected from the V_H domain were selected from the V_L domain, indicating that the predictor uses more information from the V_L domain to make these predictions. This was particularly true for positions in the framework sequence before the CDR-L3 loop.

5.6 Improving the best classifiers

At this point, it has been decided that the best encoding methods to separate approved and discontinued antibodies are the amino acid compostitions with the

Table 5.2: GaussianNB parameters for GridSearchCV.

Parameter	Values
priors	None, [0.1, 0.9], [0.2, 0.8], [0.3, 0.7], [0.4, 0.6], [0.5, 0.5], [0.6, 0.4], [0.7, 0.3], [0.8, 0.2], [0.9, 0.1]
var_smoothing	10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1

Table 5.3: LinearSVC parameters for GridSearchCV.

Parameter	Values
C	0.0001,0.001, 0.01, 0.1, 1, 10, 100
penalty	l1, l2
tol	100,10,1,1e-1,1e-2,1e-3,1e-4,1e-5
class_weight	None, balanced
max_iter	500, 1000, 2000
multi_class	ovr, crammer_singer
fit_intercept	True, False

Gaussian Naive Bayes model (MCC=0.87±0.1, k=1000) and the encodings of the AntiBERTy language model with the linearSVC and Calibrated Classifier models (MCC=0.8±0.1, k=2500). It was checked if these performances could be further improved using including GridsearchCV.

5.6.1 GridSearchCV

GridsearchCV is a systematic hyperparameterisation technique in which every combination of a selection of values for parameters for a given model is tested for the model the GaussianNB model (Table 5.2) and Linear SVC model (Table 5.3). The result was a small change in performance for both models (Table 5.4). Consequently, it was decided that this operation was not better than the default parameters.

5.6.2 Increasing Probability Threshold

Rather than using binary results to classify predictions, probability scores for both positive and negative predictions can be retrieved. This allows for additional param-

Table 5.4: Best parameters selected from GridSearchCV with default and parameter MCC scores.

Model	Encodings	k	Default Parameters (MCC)	Best Parameters (MCC)	Best Parameters
LinearSVC	AntiBERTy	2500	0.86±0.09	0.80±0.09	C: 10 class_weight: balanced fit_intercept: True max_iter: 2000, multi_class: ovr, penalty: l2, tol: 1
GaussianNB	Amino Acid Compositions	1000	0.85±0.09	0.85±0.09	priors: None var_smoothing: 1e-09

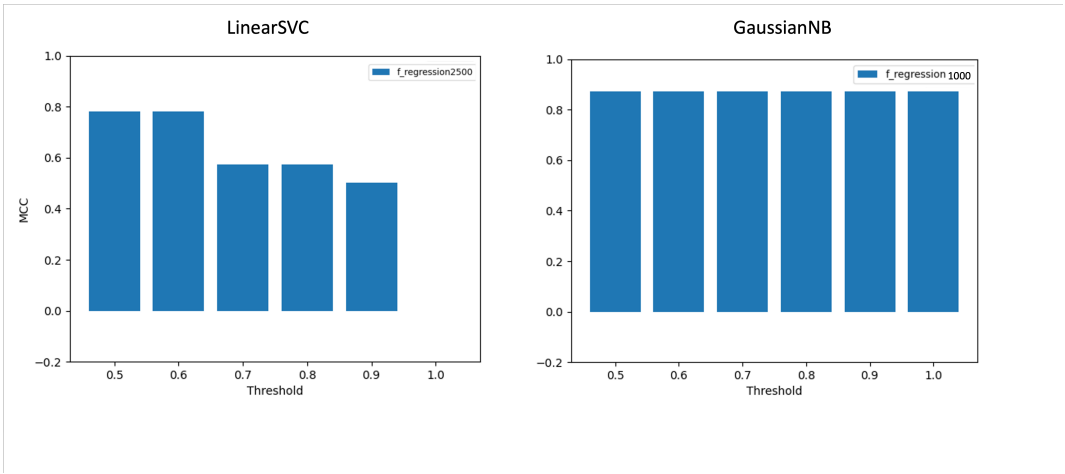

Figure 5.7: MCC scores of predictions of test split data set at different positive prediction probability thresholds. Predictions were made using Linear SVC model (k=2500) and GaussianNB model (k=1000).

eterisation by adjusting the threshold required to give a positive prediction above the default (0.5). It was seen that MCC decreased for LinearSVC, but no change was seen for the GaussianNB model when increasing the probability threshold (Figure 5.7).

5.7 Selecting a Model to Take Forward

As it was seen that optimising the models did not give an obvious increase in predictive performance, a held back dataset was used to finally comment on which model should be considered the best for use in the pipeline.

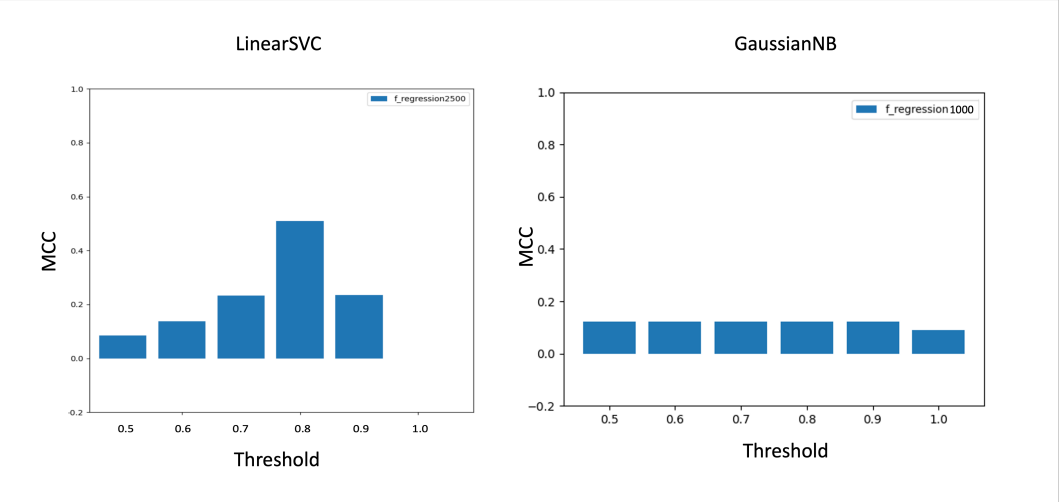

Figure 5.8: Using probability thresholds to improve MCC prediction of approved and discontinued mAbs. MCC scores for the predictions on a held-back dataset at different positive prediction probability thresholds. Predictions were made using the LinearSVC model (k=2500) and GaussianNB model (k=1000).

5.7.1 Testing Classifiers with a Held-Back Dataset

One of the dangers of a small training dataset is overfitting models to that dataset, which is characterised by good predictive performance with examples from the training dataset, but poor performance for examples not included. For this reason, a held-back dataset of newly approved and discontinued antibodies that were not part of the original training dataset was collected to check for overfitting. Adjusting the probability threshold, the best MCC achieved for this group was MCC=0.5 when the LinearSVC model was with the probability threshold was set to 0.8 (Figure 5.8). In contrast, the GaussianNB classifier prediction was MCC=0.12 for all predictive thresholds, indicating the model was overtrained to the original dataset.

To support the notion that the signal resulting from the F-regression is genuine, pairwise distances between the clinical antibodies used in the training data and the held-back dataset were calculated using `sklearn.metrics.pairwise_d`

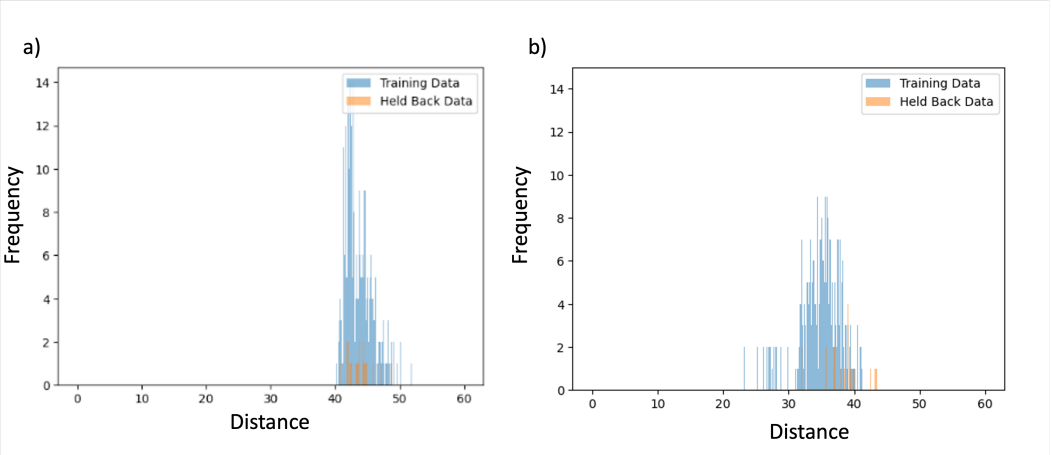

Figure 5.9: Average Pairwise distances between held back antibodies and antibodies used in training dataset using a) all AntiBERTy encodings and b) top encodings (k=2500) selected using F-regression.

`istances` function in Python. It returns a Euclidean distance matrix for each sequence in the training data against each sequence in the held-back data. It was found that there were no significant differences between the two groups for the average distance between clinical and held back using an unpaired t-test (Figure 5.9a; p=0.77) or the minimum distance and held back (Figure 5.9b; p=0.66). This means that the two datasets are closely enough related to support the notion that a signal found in the training dataset should also be found in the held back dataset. Furthermore, held back antibodies were found to cluster closely with clinical antibodies used in the training dataset when used in a kernel principal component analysis (kernel= radial basis function, γ=500) (Figure 5.10).

5.7.2 Speed of Encodings

The LinearSVC model was deemed the best for identifying approved antibodies within clinical antibodies. Amino acid encodings have been slow to compute, taking 1.4 seconds per antibody, and may not be suitable for high-throughput encoding. In

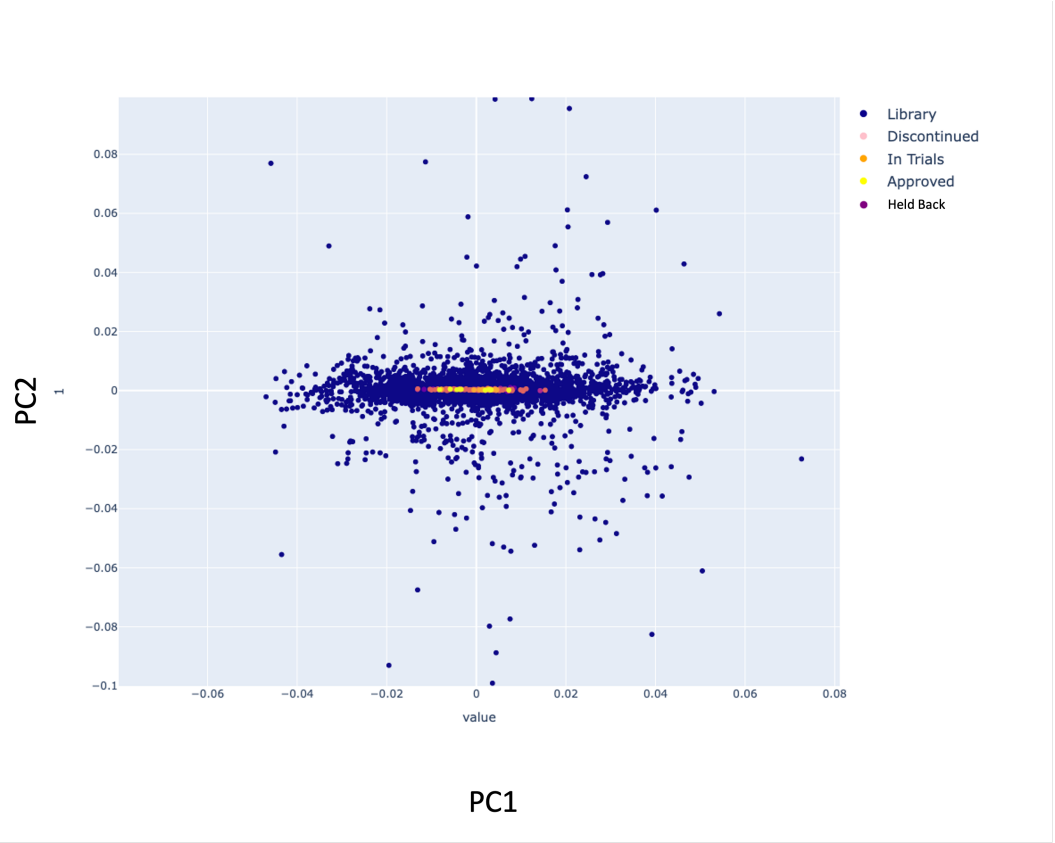

Figure 5.10: Kernel PCA demonstrating clustering of clinical antibodies used in training with held back therapeutic mAbs.

contrast, because of the ability to utilise GPU hardware, denser protein language model encoding only requires 0.02 seconds per antibody.

5.7.3 Approved *vs.* Discontinued Classifier on Repertoire Dataset

The performance of this model using unlabelled repertoire sequences was tested. These sequences have much more diversity than the clinical sequences this model was trained on. An experiment was set up where the encoded OAS antibody dataset (n=10,000) (see Section 3.2.1) was used as a test dataset for the LinearSVC model trained in this chapter. This was not expected to perform well. The encoded data

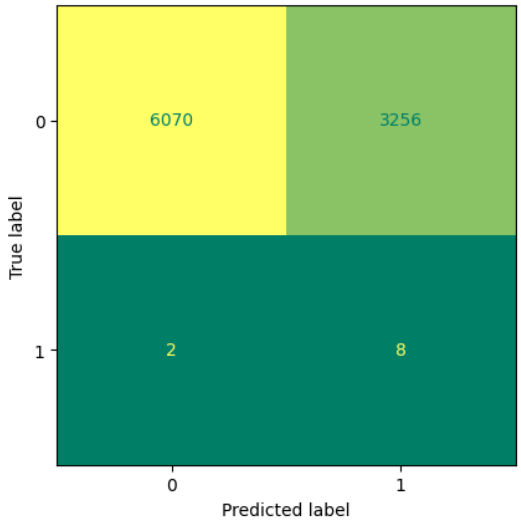

Figure 5.11: Confusion matrix of the predictions of approved (class 1) and discontinued (class 0) for 10,000 human repertoire antibodies.

had the same features used by the LinearSVC model (k=2500), and the model was tasked with classifying these antibodies. The approved held back antibodies (n=10) were used as positive controls for this experiment.

As can be seen in Figure 5.11, the classifier predicted a roughly 2:1 ratio of discontinued and approved antibodies (MCC=0.03, Sn=0.8, Sp=0.65). Potentially because the previous approved *vs.* discontinued mAb dataset roughly followed this ratio, the model has inherited this assumption, which could potentially be an overestimation of the number of clinical antibodies in the library. From this, it was concluded that because this classifier had been trained on clinical sequences, it was not suited to classifying repertoire sequences, demonstrating the need for the separate model classifying clinical from repertoire.

5.7.4 Selection

The LinearSVC model trained on the AntiBERTy encodings was decidedly taken forward to be the approved *vs.* discontinued predictor because it had better predic-

tive performance on the held back dataset, and encodings take less time, making it suitable for high-throughput analysis.

5.8 Physicochemical Properties of Approved and Discontinued mAbs

Due to the persistence of this learning effect, it was hypothesised that a statistical difference in the two groups might be observed in one or more of the physicochemical properties that could be likely to jeopardise a given mAb's success in trials. A larger number of physicochemical properties were investigated than before in order to find any such differences and to explain why these models could be trained.

5.8.1 CDR-H3 Loop

The length of the CDR-H3 loop in antibodies could potentially be a liability if it is too long [9]. Each approved and discontinued antibody was numbered with AbNum and the CDR-H3 regions were identified using Chothia definitions. Both the approved and discontinued groups had the same maximum and minimum CDR-H3 length, 19 and 3, respectively. The mean length of the CDR-H3 regions between approved and discontinued antibodies was 11 and 10, respectively, and there was no significant difference between their distributions through unpaired t-test (p=0.23) (Figure 5.12).

5.8.2 Thermostability

Thermostability (ΔG) was predicted with the Oobatake Method [128] for the conjugated V_H and V_L sequences, the lone V_H sequence and lone V_L sequence for the

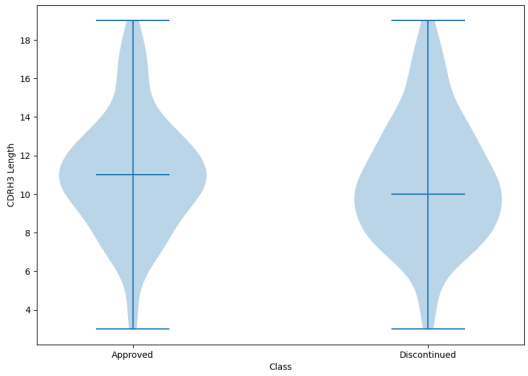

Figure 5.12: Approved and discontinued mAb CDR-H3 length. Violin plots of the distributions describing the length of CDR-H3 loops in approved and discontinued antibodies as defined by the Chothia numbering scheme.

dataset of approved and discontinued antibodies. It was seen that no statistical differences were observed between the approved and discontinued dataset in any of these groups by unpaired t-test: combined (p=0.88), V_H (p=0.78), V_L (p=0.55) and mean of V_H and V_L sequence Δ G per antibody (p=0.88) (Figure 5.13).

5.8.3 Isoelectric point

Isoelectric point (pI) was calculated with the IPC Method [129] for the conjugated V_H and V_L sequences (p=0.99), the lone V_H sequence (p=0.22) and lone V_L sequence (p=0.17) for the dataset of approved and discontinued antibodies. It was seen that no statistical difference was observed between any of these groups between the approved and discontinued dataset. This included the mean pI of the lone V_H and V_L sequences (p=0.17) (Figure 5.14).

5.8.4 Key Residues

The idea of key residues is to identify residues in the CDR-H3 region linked to the propensity to form beta sheets [133]. These are associated with antibody promiscuity and could potentially lead to off-target effects in clinical scenarios. 9 approved

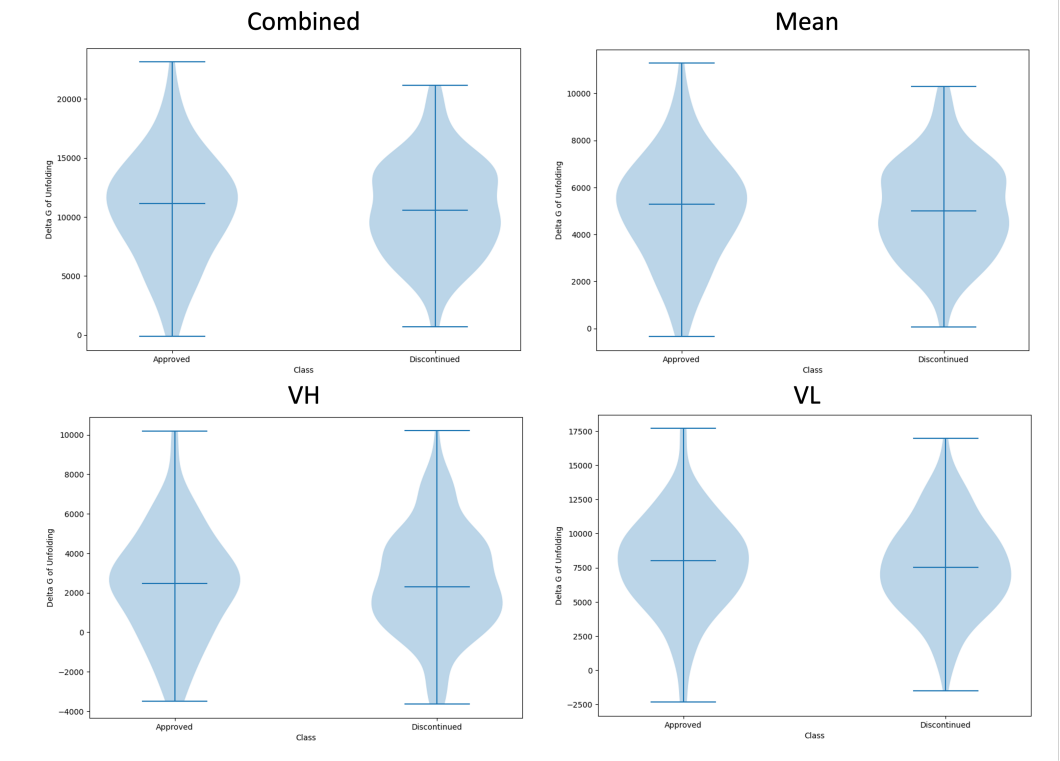

Figure 5.13: Approved and discontinued mAb ΔG. Violin plots of the distributions describing ΔG values for market-approved and discontinued antibodies calculated using the Oobatake method for the combined V_H andV_L sequence, mean value of the V_H and V_L sequence, lone V_H sequence and loneV_L sequence.

Table 5.5: Approved and discontinued antibodies with key residues in CDR-H3 Loop.

	100 L	100C H	100EW	100 L, 100C H
Approved	Fremanezumab Mepolizumab Odesivimab Polatuzumab Secukinumab Ciutumumab	Bevacizumab Omalizumab	Emapalumab	Belimumab
Discontinued	Dectrekumab Etokimab Fresolimumab Iladatuzumab Tabalumab	Icrucumab Tavolimab	Enoticumab Firivumab Lesofavumab Gedivumab	

and 12 discontinued antibodies were found to have these key residues in their CDR-H3 loops (Figure 5.5). Only one antibody was found to have more than one key residue, and this was from the approved dataset (Belimumab).

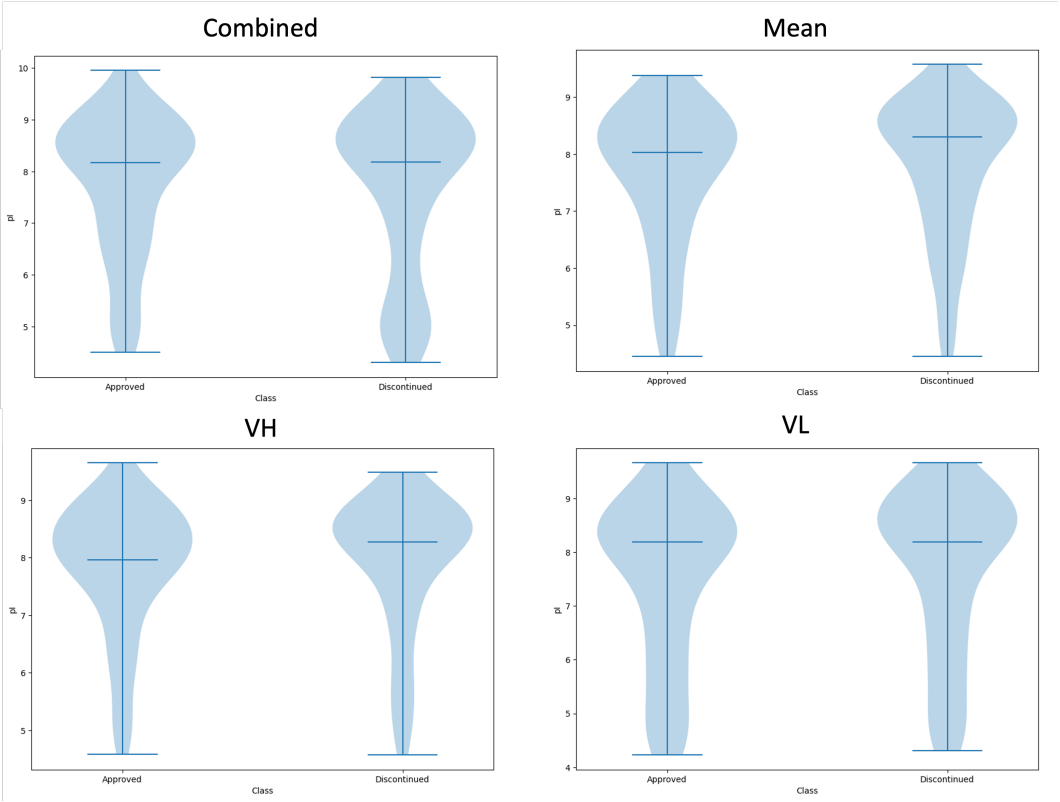

Figure 5.14: Approved and discontinued mAb isoelectric point. Violin plots of the distributions describing isoelectric point (pI) values of market-approved and discontinued antibodies calculated using the IPC method for the combined V_H and V_L sequence, mean value of the V_H andV_L sequence, lone V_H sequence and loneV_L sequence.

5.8.5 V-region Germline Gene Pairing

The germline genes of all clinical stage antibodies were evaluated with AGL. For both approved and discontinued antibodies the most popular V_H and V_L V gene families were IGHV3 and IGKV1 respectively. These germlines were also the most popular pairing of gene families in each group (Figure 5.15). In both approved and discontinued, the second most popular V_H and V_L V region gene families were IGHV1 and IGKV3, however, in the approved group, the pairing of IGHV1/IGKV1 was proportionally greater than the pairing of IGHV3/IGKV3 in each group, but this is the opposite case in the discontinued. It was seen using χ^2 test that there was

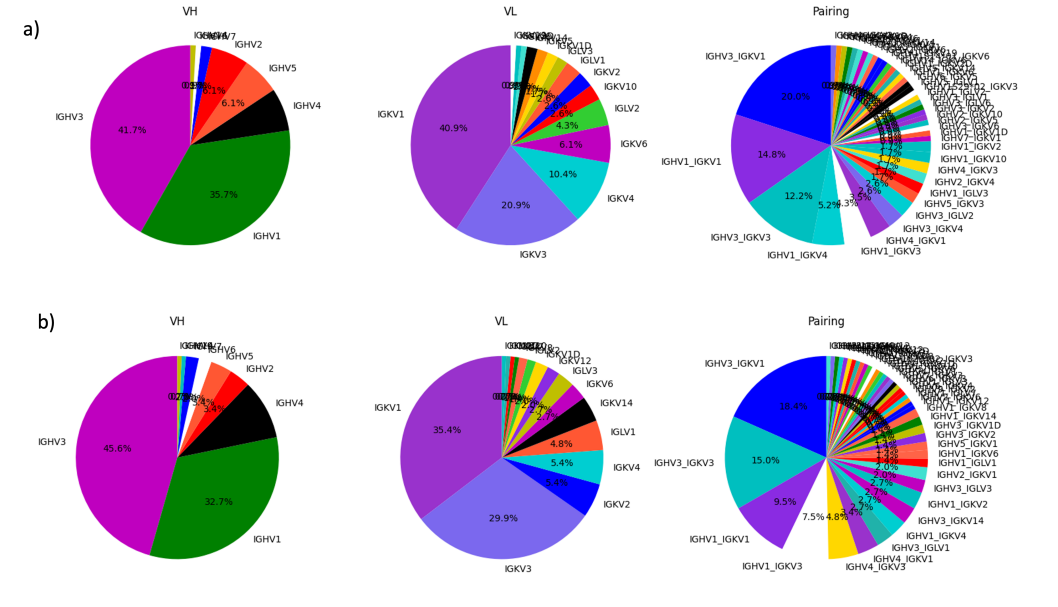

Figure 5.15: Approved and discontinued germline pairing proportions. Proportions of germline gene families in approved (a) and discontinued (b) antibodies. Colour coding for V_H , V_L and Pairing categories are consistent for approved and discontinued figures.

no significant difference between approved and discontinued germline frequency in V_H domains (p=0.09, 30 degrees of freedom) or V_L domain (p=0.08, 72 degrees of freedom). There was a significant difference seen in the V_H/V_L pairings (p=0.00, 90 degrees of freedom) but none of the individual pairings were found to be significant using the Bonferroni-Hochberg q value adjustment.

5.8.6 Post-Translational Modifications

Post translational modification sites were detected by scanning along the V_H and V_L sequences of each antibody and recording at which positions a regular expression site corresponding to a recognition site was matched (see Section 2.3.4). The results show little difference between the approved and discontinued data set for the V_H domains (Figure 5.16) and V_L domains (Figure 5.17). None of the PTMs at any positions showed a significant p-value when checked with χ^2 test with the

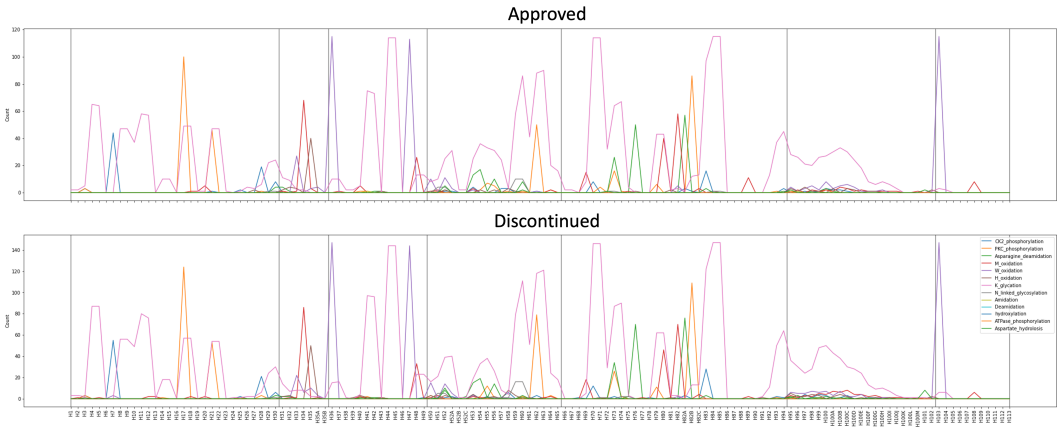

Figure 5.16: Approved and discontinued mAb V_H PTM recognition sites by sequence position. Frequency of post-translational modification recognition sites in market-approved and discontinued antibodies for each position in the V_H sequence according to the Chothia numbering scheme.

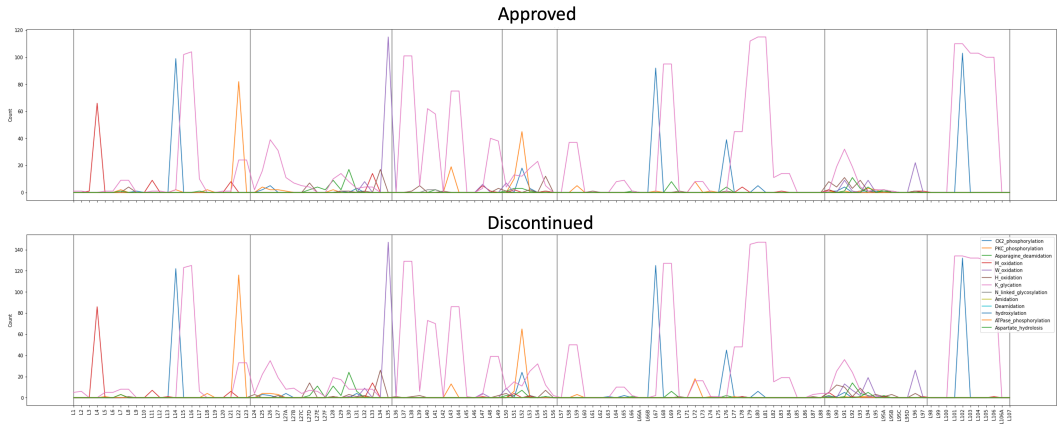

Figure 5.17: Approved and discontinued mAb V_L PTM recognition sites by sequence position. Frequency of post-translational modification recognition sites in market-approved and discontinued antibodies for each position in the V_L sequence according to the Chothia numbering scheme.

Bonferroni-Hochberg q value adjustment.

5.8.7 Hydrophobicity

Clusters were calculated using the 'ClusterResidues' programme (see Section 2.3.6) using Chothia-numbered antibody sequences from AbNum and structural models made by abYmod for all antibodies. For all mAbs, the number of different profiles

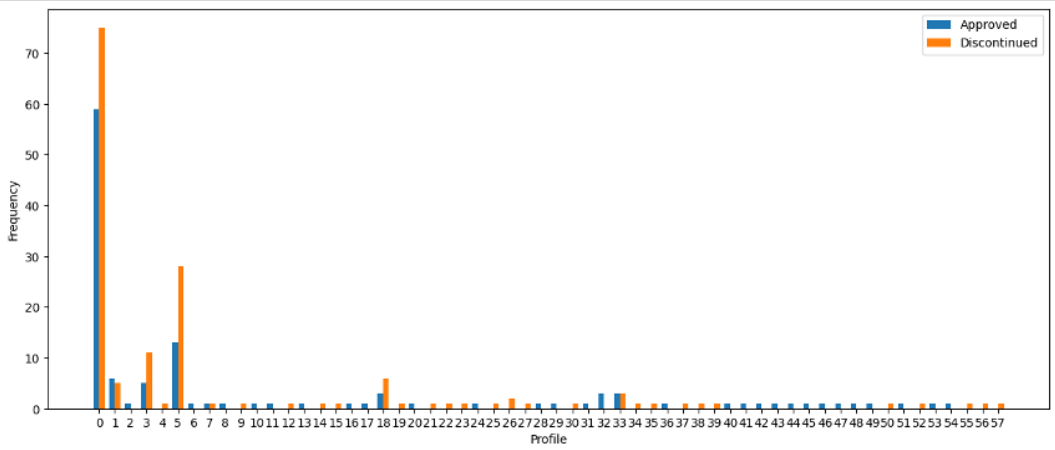

Figure 5.18: Counts of different hydrophobic patch profiles observed between the approved and discontinued antibody datasets.

Table 5.6: Most popular hydrophobic cluster profiles between market-approved and discontinued antibodies.

Rank	Profile Number	Approved Count (%)	Discontinued Count (%)	Clusters
1	0	59 (51.7%)	75 (51.7%)	None
2	5	13 (11.4%)	28 (19.3%)	Cluster 1: L106, L15, L83
3	3	5 (4.4%)	11 (7.6%)	Cluster 1: H108, H89, H9
4	1	6 (5.3%)	5 (3.4%)	Cluster 1: H108, H89, H9 Cluster 2: L106, L15, L83
5	18	3 (2.6%)	6 (4.1%)	Cluster 1: H108, H89, H9 Cluster 2: L106, L15, L80, L83}

for hydrophobic patches was tallied, and counts of the frequency of each profile (combinations of different clusters) were taken for the approved and discontinued groups. In both groups, the most common profile was that there were no hydrophobic patches. The five most popular profiles are summarised in Table 5.6 showing that the proportions of these profiles in the approved and discontinued groups are similar. None of the differences in frequencies of hydrophobic patch profiles between approved and discontinued mAbs had a significant p-value when checked with χ^2 test.

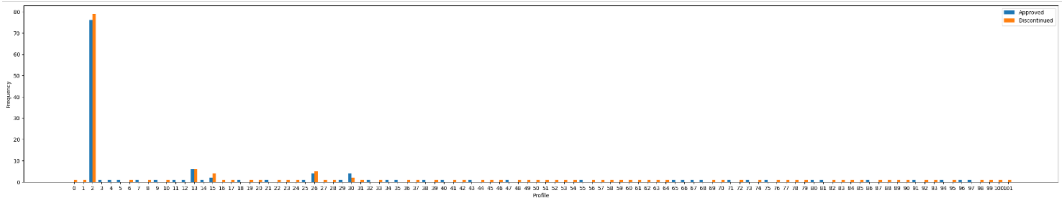

Figure 5.19: Counts of different unusual patch profiles observed between the approved and discontinued antibody datasets.

Table 5.7: Most popular unusual residue cluster profiles between approved and discontinued antibodies.

Rank	Profile Number	Approved Count (%)	Discontinued Count (%)	Clusters
1	2	76 (66.6%)	79 (54.5%)	None
2	13	6 (5.3%)	6 (4.1%)	Cluster 1: H54, H56, H57
3	26	4 (3.5%)	5 (3.4%)	Cluster 1: L30A, L30B, L92, L93
4	15	2 (1.75%)	4 (2.8%)	Cluster 1: H61, L94, L95 Cluster 2: H98, L50, L53
5	30	4 (3.5%)	2 (1.4%)	Cluster 1: H100C, H95, H97 Cluster 2: L28, L30, L32, L92

5.8.8 Unusual Clusters

Clusters were calculated using the clusterresidues programme using Chothia-numbered antibody sequences from AbNum and structural models made by abYmod for all antibodies. For all mAbs, the number of different profiles for unusual residue patches was tallied, and counts of the frequency of each profile were taken for the approved and discontinued groups. The five most popular profiles are summarised in Table 5.7. None of the differences in frequencies of unusual patch profiles between approved and discontinued mAbs had a significant p-value when checked with χ^2 test.

5.8.9 Solvent Accessibility

abYmod was used to generate a structure of each market-approved mAb and discontinued mAb. Each structure was entered into *pdbsolv*, and a solvent accessibility value was obtained for each residue. The residues were then numbered according

to the Chothia scheme, and for each residue, the distributions of solvent accessibility for approved and discontinued antibodies were compared. Figure 5.20 demonstrates the mean and standard deviation for all solvent accessibility values for each residue. The nonparametric Mann-Whiteney test did not find significant differences between the means of approved and discontinued reletive solvent accessibility values for most of the residues in the V_H or V_L sequences. Some individual positions were shown to have significant ($p < 0.05$) differences between the approved and discontinued groups for solvent accessibility: H58; H63; H85; L47; L62 and L65 (Figure 5.20). All of these positions are found in the Framework 3 region of the V_L sequence, except for L47, which is in CDR-L2. On inspection of the figure, it seems that these may be artefacts of multiple testing, because in all cases, error bars overlap. When this was corrected for multiple tests using the BH method, none of these positions remained significantly different between groups.

5.9 Discussion

In order to predict which antibodies selected by the pipeline would be more likely to pass clinical trials, the work in this chapter sought to train a predictor between market-approved and discontinued mAbs taken from TheraSabDab [33] in order to analyse characteristics that the model considered important for market approval.

A major limitation of the dataset is that the reasoning for discontinuation is rarely disclosed by pharmaceutical companies. Although this can involve safety aspects in immunogenicity, lack of efficacy, or other adverse effects, marketing reasons can also play in this, as the stakes are so great when entering a drug into

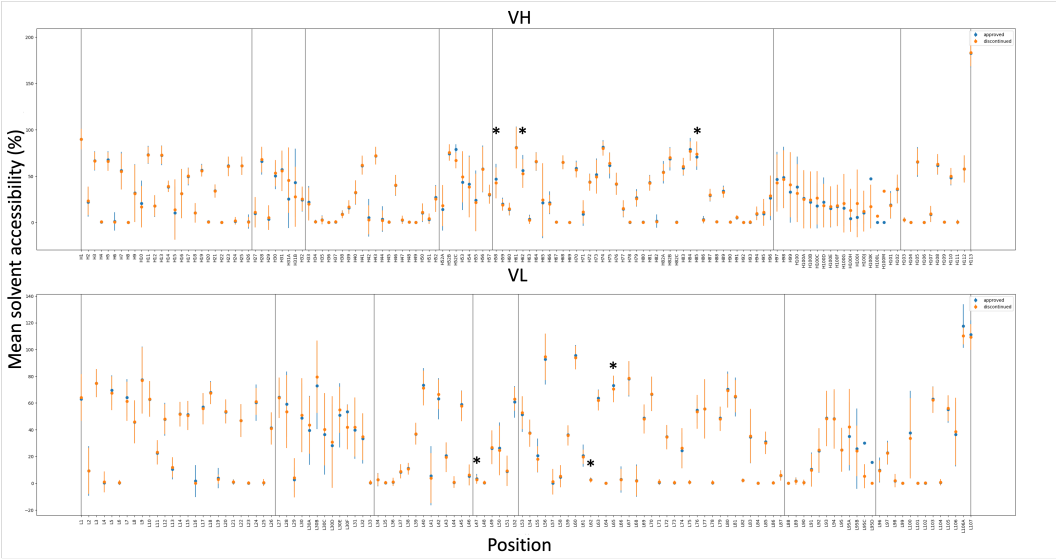

Figure 5.20: Approved and discontinued mAb solvent accessibility values. Values are given for each residue of the Chothia V_H and V_L sequence numbering scheme calculated by pdbsolv for market-approved and discontinued antibodies. * denotes significant difference between approved and discontinued groups ($p < 0.05$).

clinical trial. When the literature search to find reasons for discontinuation was carried out, rather than finding scientific literature, the reporting was usually from anecdotal press releases, citing efficacy reasons. The problem with this is that efficacy is difficult to disentangle from poor bioavailablility, poor binding, potentially a poorly understood target/pathway or poorly understood clinical endpoint [45]. Furthermore, these reports cannot be wholly relied upon as pharmaceutical companies may want to protect their reputations by claiming that a discontinuation was due to marketing reasons, rather than admit to any safety or efficacy issues with a drug in trials, so even here, this study assumes that the press releases are truthful. Despite this, it was hoped that the collected samples would at least be enough to identify criteria from which clinical antibodies could be identified.

No statistical differences were found in these properties of this grouping of

approved and discontinued clinical mAbs, so it was interesting that the models were able to achieve such good predictive performance. Generally, it was seen that these groups had similar physicochemical property profiles, including thermostability and isoelectric point, and therefore it was not possible to generate distinct classification criteria based on these properties. This was extended to within-group frequencies of post-translational modifications, hydrophobic patches and germline pairings, which seemed the most surprising result, because it was hypothesised that discontinued mAbs would show a higher frequency of these developability 'red flags' that would have led to them being discontinued. However, this result is expected to be linked to approved and discontinued datasets that have similar proportions of their most popular V_H and V_L chain germline pairings. The result of this adds an additional layer of bias to the dataset as the majority of these drugs come from the IGHV3 or IGHV1 germline families.

Whilst intuitive, it has not been clearly stated before in the literature, but the clinical dataset is already a selected dataset with antibodies that can be assumed to meet the developability profiles necessary to be developed at scale, but do not all have the characteristics to make it to market. Therefore, the developability characteristics that have been previously identified in research are necessary to get to the clinical stage but not sufficient to get a drug from clinical trials to market. Work in this chapter was useful since these were the only examples available as true positives and true negatives, so was the best place to start to identify clinical antibodies that could in fact pass trials.

What was interesting, however, was that using classifiers trained on amino acid

encodings and protein LLM encodings could distinguish these two groups with high performance with roughly equal prediction capability, which was markedly better than predictions made with the residue level encodings, even when using all of the methods concatenated together. This would indicate that amino acid compositions and language models capture features that are linked to the drug becoming accepted onto the market.

The feature selection via F-regression was found to have a vast improvement in predictive performance in all cases where it was used, compared to the raw encodings, which usually had poor performance most likely to overfitting to less informative features. To explain this, a number of reasons can be interrogated: firstly, the language models have been trained on a selection of millions of sequences and so have learnt important features of the antibody sequences that can in fact distinguish these groups, unlike the residue level encodings, and secondly, the fact that these models generate more dense encodings than residue level encodings, there is greater probability of finding more features which are correlated with the two groups due to chance to be learned from when feature selection is employed.

Potentially, there is a mixture of both of these reasons at play. From initial studies, it has been shown that these language models are heavily biased towards distinct germlines [177] due to the biases in their training data, however, it was shown in this chapter that approved and discontinued antibodies have similar frequencies of the same germline gene families in their groups, most likely due to biases in candidate selection processes from biases in the repertoires of the antibody source. This would indicate that this bias towards certain germlines is not what is causing the

selection and that other characteristics relating to antibody effectiveness may make a meaningful difference. These features include: poor binding affinity *in vivo*, off-target effects or poor bioavailability due to low antibody titre from administration route [45, 201], but it is unlikely that the language models themselves have an understanding of these characteristics. However, since this study failed to establish the reasons for the discontinuation of all the discontinued antibodies, it is difficult to draw meaningful conclusions from these encodings to understand these reasons for the failure.

This chapter ends with a selected machine learning model that shows good predictive performance with approved vs. discontinued antibodies, which is maintained with a held back dataset. It is not so surprising that the performance with the held back dataset was not as strong as the training dataset as it can be expected that these antibodies may be more diverse, or have different properties from the antibodies in the training dataset. Furthermore, it is suspected the original model may be somewhat overfitted as the best performance was seen when the positive example probability threshold was increased to 0.8, however in our case, it is preferred that the model would be more stringent and only select antibodies which it is confident are positive examples from the learnt dataset.

5.10 Conclusions

This chapter concludes with the notion that clinical antibodies are part of a selected dataset with similar developability properties that make them unsuitable to discover triaging material. Dense encoding methods, like protein LLMs, can be used to

train machine learning models in order to separate these groups with high degrees of performance, however, the features that led to this performance could not be identified. This would indicate the relationship here could be due to coincidence, but a good prediction was also achieved with a held back dataset, indicating that these models are still useful. To improve the interpretability of this model, it would be important to know why certain mAbs were discontinued to try and relate these features more clearly.

Chapter 6

Assembling the pipeline

6.1 Introduction

Throughout the preceding chapters of this thesis, an approach has been built to generate a triaging pipeline that will take a sample of paired human antibody sequences and output those which satisfy developability features seen in the clinical dataset, and are more likely to pass clinical trials.

This chapter will assemble the pipeline using the results of the previous chapters. Firstly, by ordering physicochemical property triaging based on Z score filtering of previously calculated properties; secondly the previously used kernel PCA for unsupervised categorisation of repertoire data and clinical mAbs, and finally the supervised labelling of antibodies more or less likely to pass clinical trials. A test library of antibodies from Stewart *et al.* [121] was used to test the pipeline at different Z score parameters and measure how sensitive the pipeline is to mutations relevant to mAb developability. Performance was compared with other developability screening software [110] and a selection of antibodies triaged out at different

stages of the pipeline were expressed and experimentally measured for developability properties.

6.2 Pipeline Outline

The pipeline takes a combination of approaches to triaging antibodies. Input antibodies are triaged using physicochemical properties to remove antibodies with properties that clearly fall outside of the ranges observed in the clinical dataset. This reduces both the computation time for numbering and encoding, to select higher-quality antibodies later in the pipeline. Antibodies are numbered according to the Chothia Numbering Scheme [16], and missing residues are spaced out using AbNum [112]. The spaced sequences are encoded with the AntiBERTy langauge model and then the sequences are entered into the machine learning part of the pipeline. Firstly a filter using an unsupervised learning KPCA model 'Layer 1' to triage out antibodies which do not have similar properties to those observed in clinical antibodies and then a Linear SVC binary classifier 'Layer 2' to select antibodies which a supervised model predicts are more likely to be approved at clinical trials. Using Z scores and probability thresholds, additional stringency can be added at different steps.

Optional models trained on anti-drug antibody (ADA) incidence data to predict immunogenicity above a threshold of 1% can be used if the clinical setting of the antibody being developed demands this. Linear models predicting relevant developability properties trained on experimental data can also optionally be employed.

It is expected that there would be some triaging out of antibodies at each step of

the pipeline, including spacing and encoding due to incompatible sequences which the AntiBERTy model cannot encode. A pipeline schematic is shown in Figure 6.1.

6.3 Testing the Pipeline with a test dataset

The assembled pipeline was tested for its triaging effect in using a library of antibodies from healthy human donors.

6.3.1 Pure2 Dataset

6.3.1.1 Library Preparation

The Pure2 B cell receptor (BCR) sequence resource was provided by the Franca Fraternali Group at UCL. These datasets are an expansion of the library of three healthy young blood donors published in Stewart *et al.* [121] with three additional older blood donors. In all cases, blood donors had their B cells isolated and FACS-sorted by developmental stage. The transcripts from each individual cell were barcoded and therefore V_H and V_L pairing is possible. Antibody V_H and V_L pairs were taken as B cells that shared the same barcode where an IGH and IGL or IGK chain was present. In cases where both IGL and IGK chains were present, the chain with the highest count number was taken as the V_L chain pair. No filtering based on the type, or stage of development, of the BCR was performed for the purposes of assembling this dataset. Individual sequences for the frameworks and CDR loops were concatenated to give the full antibody *Fv* domain sequence. In total, 10,492 paired antibodies were extracted from the library in nucleotide format (Data File 12) which was then translated into amino acid format (Data File 13).

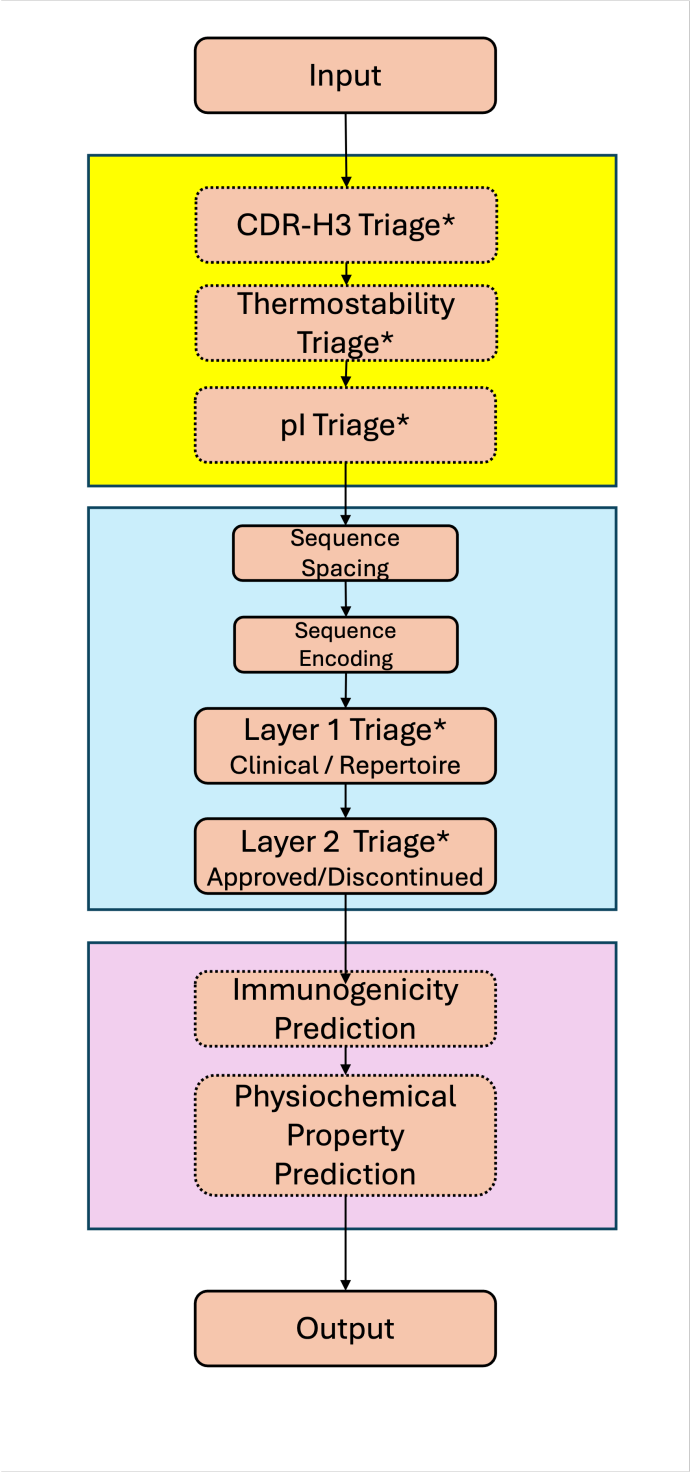

Figure 6.1: Schematic of the antibody triaging pipeline from input to output. The yellow box indicates optional physicochemical feature triaging steps calculating CDR-H3 length using AbNum [112]. Thermostability (ΔG of unfolding) using the Oobatake Method [128] and pI using the IPC method [129]. The blue box indicates machine learning elements including spacing and encoding, as well as 'Layer 1' triage which is based on the Kernel PCA model for separating antibodies with similar properties to clinical mAbs from the repertoire. The selection of antibodies to take forward is made using the ellipse function. 'Layer 2' is the F-regression and supervised LinearSVC model trained to distinguish approved and discontinued clinical mAbs. The purple box indicates optional physicochemical property prediction as done by linear models. * indicates stages where stringency can be adjusted using Z score thresholds.

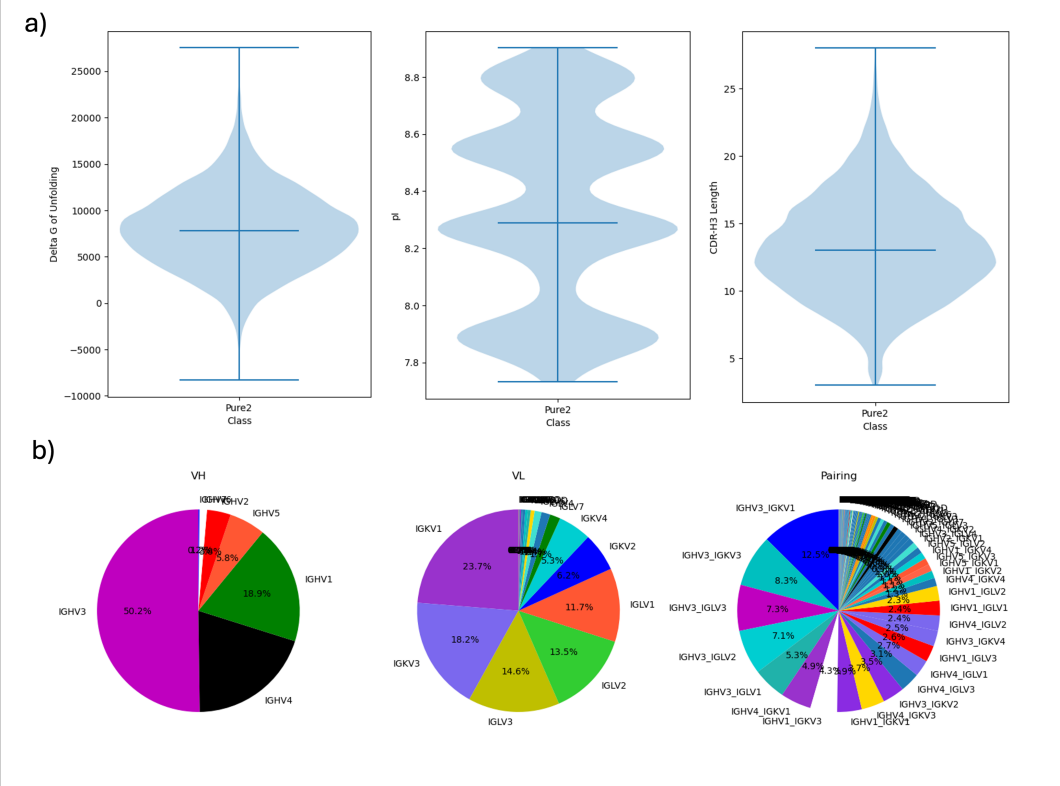

Figure 6.2: Physicochemical properties of Pure2 dataset. a) Violin plots of physicochemical properties of the Pure2 dataset and b) proportions of V_H and V_L germline gene families and pairings.

6.3.1.2 Library Statistics

The Physicochemical properties of this library were examined (Figure 6.2). The mean ΔG of these antibodies was 7944 and the majority of these antibodies had non-negative values. The mean pI was 8.3 and the mean length of CDR-H3 was 13. Approximately half of the heavy chains came from the IGHV3 V_H domain germline family, which was a higher proportion than that seen in the approved or discontinued mAb dataset. Despite this, there was more variability in the V_L domain germlines and therefore more variability in the frequency pairing combinations. The most frequent pairings all involved IGHV3.

As a comparison developability score the TAP score [117] was calculated for

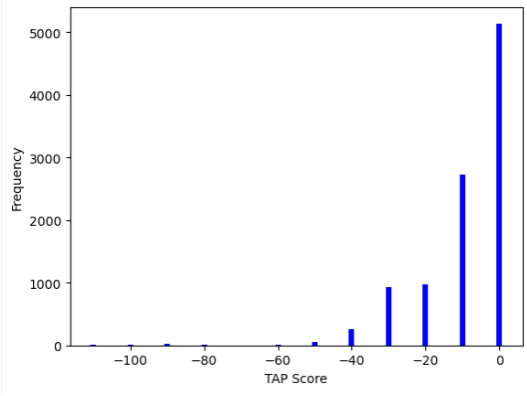

Figure 6.3: TAP scores assigned to Pure2 dataset. Histogram of TAP scores seen for 10,132 paired antibodies from the Pure2 Dataset.

each antibody from the Pure2 library. The TAP score is a developability score where an antibody with values for the selected physicochemical properties that are seen within the clinical mAb dataset is given a perfect score of 0. Antibodies with increasing numbers of 'red flags' where the values are outside the range observed in clinical mAbs are given negative scores. This is only a predictive indicator of developability and so the use of this score was only as a comparison, not a benchmark. TAP scores were calculated for 10,132 of the paired V_H and V_L nucleotide sequences from the Pure2 dataset in batches of 500 using the IGX platform [1] in August 2023 using the set of default penalties. Details of the statistics measured and penalties assigned can be found in Raybould *et al.* [117].

The median TAP score for all the Pure2 antibodies was 0, suggesting that half of these antibodies would have no developability red flags according to the TAP score. Each successive negative score was seen less frequently until the most negative score, -110, which was observed for seven of these antibodies, indicating numerous developability issues (Figure 6.3). For the antibodies from the Pure2 dataset

[1] `https://igx.bio/`

that could not be assigned a score, probably they could not be numbered, and this suggests that they are unusual, and that there would also be developability problems with these antibodies.

6.3.2 Training a model on the TAP score output

6.3.2.1 Binary Classifiers

As an exploratory experiment, it was investigated whether, using the AntiBERTy [177] encodings, a binary classifier could be trained to predict that it would have developability red flags (n=4971) or not (n=5123). These two groups were used to train a series of 15 machine learning models with 10-fold CV, however, it appeared that none of these models could effectively train on the input data with the best performance (MCC= 0.15 ± 0.03) (Figure 6.4). Performance was not improved significantly by using F-regression feature selection, so it is possible that these groupings are not different enough to allow models to distinguish groupings.

6.3.2.2 LLM Fine Tuning

Fine tuning using the LoRA method [203] was used to train a binary classifier for the same sets of antibodies to test the same hypothesis. This would fine-tune the AntiBERTy language model by adding more layers to the end of the model according to the training data, to learn from these sequences with developability red flags. It was also seen here that the model was not able to learn from the training data as the MCC remained around 0, and raw accuracy remained around 0.5, meaning that the predictions are not better than random chance. The loss functions observed in training did not decrease much through successive epochs demonstrating that the

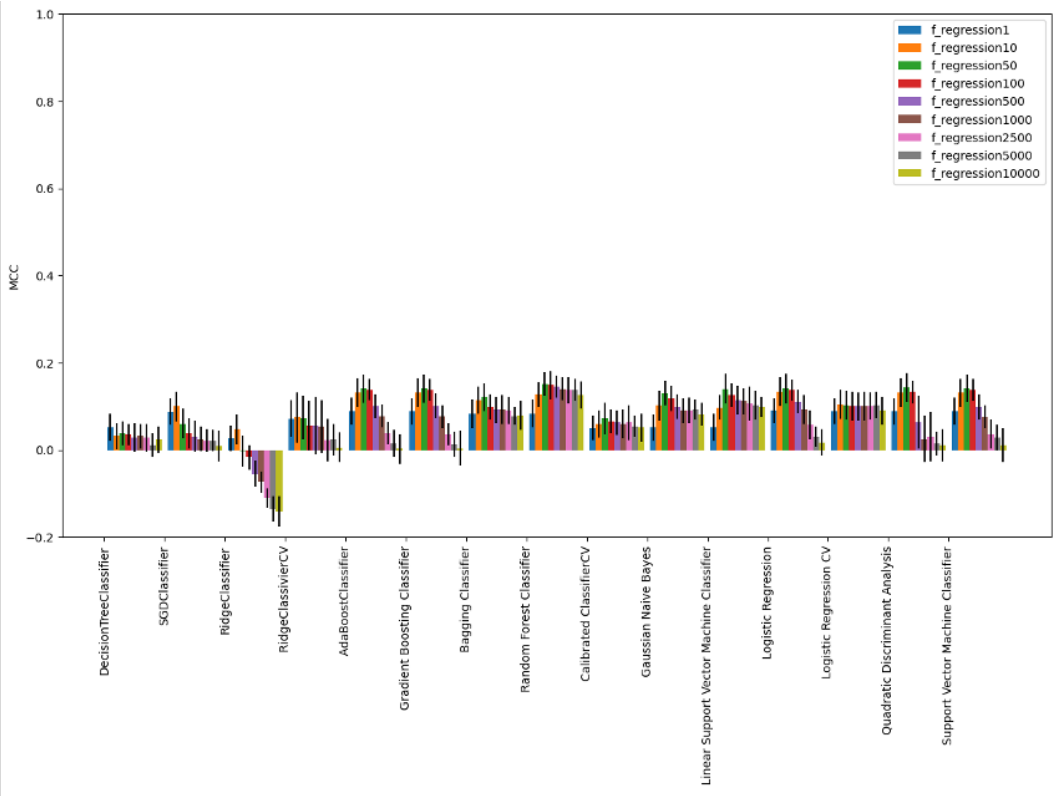

Figure 6.4: Classifiers trained on Pure2 dataset TAP scores. MCC scores and standard deviation of 15 binary machine learning predictors with 10-fold CV classifying test split of human library antibodies with a negative and non-negative TAP score. Error bars represent standard deviation.

models were not learning. What can be concluded from this is that the encodings of the two groups were two similar so the features that the TAP score is searching for are not represented in the encodings, such that antibodies with red flags were not identifiable.

6.3.3 Evaluating the Pipeline

6.3.3.1 Physicochemical Property Triage

Using the mean and standard deviation of the properties taken from the market approved antibodies, triaging criteria based on Z scores can be applied to a test dataset to remove antibodies.

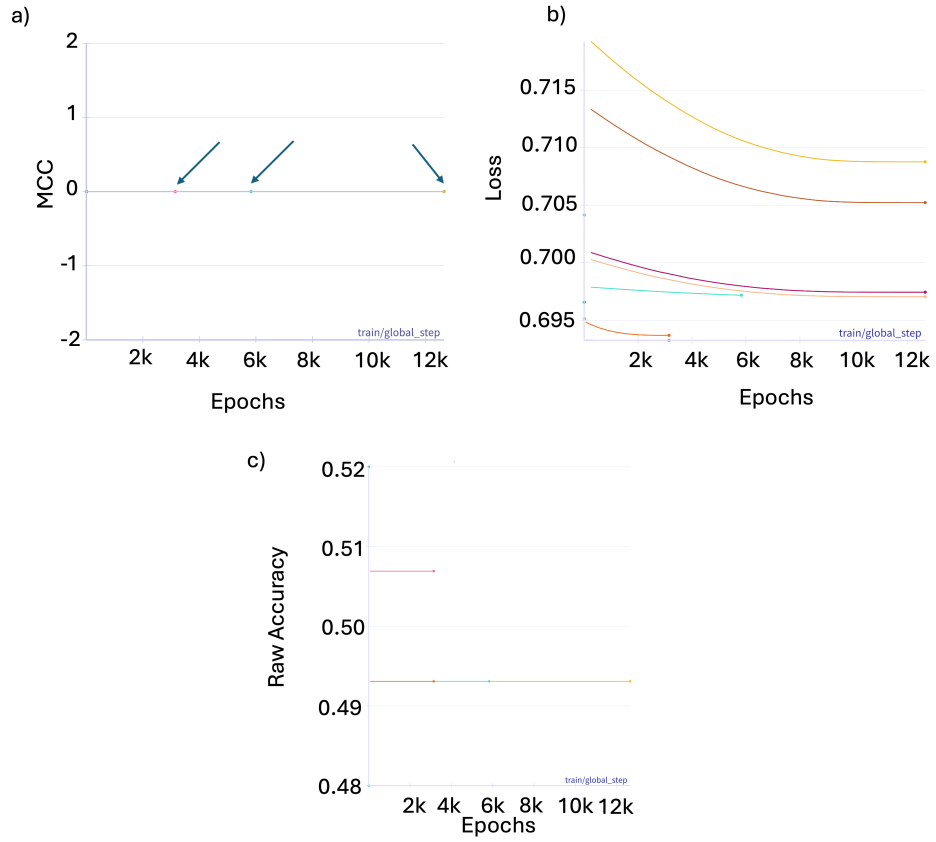

Figure 6.5: LoRA language model fine-tuning demonstrating the a) MCC b) loss and c) raw accuracy of different runs of the training (shown in different colours). In cases where there is no movement across epochs, arrows have been added for visibility.

Figure 6.6 demonstrates both the variance in thermostability, pI, and length of CDR-H3 in the Pure2 dataset and the triaging effect these filters have on the Pure2 dataset of human antibodies. It is easily identified that the values of ΔG have the largest spread with respect to the approved dataset when compared to the other properties (Figure 6.6a). The triaging based on ΔG had the greatest effect in removing antibodies from the sample when based on the combined ΔG (V_HΔG + V_LΔG), than the ΔG of either the V_H or V_L domains individually (Figure 6.6b). It was also found that pI-based classification tended to remove antibodies with low pI

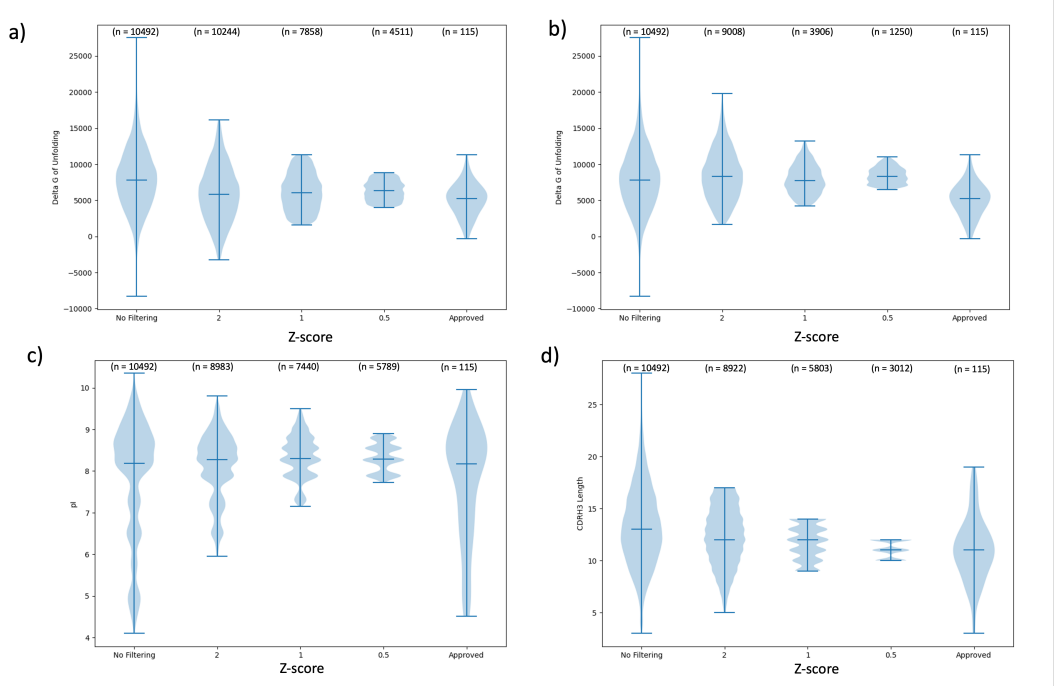

Figure 6.6: Triaging the Pure2 dataset using physicochemical properties of clinical stage mAbs. Violin plots of physicochemical properties: a) ΔG of unfolding (V_H+V_L), b)ΔG of unfolding (V_H , V_L), c) isoelectric point and d) CDR-H3 length of Pure2 dataset with various Z score filters applied, compared to the market approved dataset.

values from the dataset, however, because pI does not have a normal distribution, the Z score cutoff loses more antibodies with low pI because it assumes a normal distribution (Figure 6.6c). It was also surprising how strong an effect of the CDR-H3 length triaging had on the resulting dataset size.

Because the triaging filter was set to triage based on the individual values of V_H+V_LΔG, V_HΔG and V_LΔG separately, it had the largest effect on the triaging of the antibodies at each stage. Like before, it seems that using many individual filters in conjunction, a larger effect will be observed. This was shown to be the case when measuring all of the predicted properties (V_H+V_LΔG, V_HΔG and V_LΔG separately, pI and CDR-H3 length) were used to make one filter, and it was shown that the number of antibodies remaining after any of the Z score filters was fewer than any of the

Table 6.1: Triaging effect of the filtering of the Z score filtering using ΔG, pI and CDR-H3 length together with different values of Z.

	Z Score			
	None	2	1	0.5
Number of Antibodies	10492	8045	2740	386

individual filters (Table 6.1). By setting filtering criteria for all antibodies within the range of $0.5 < Z < 0.5$ for each metric, a large reduction in dataset size was observed where only 386 antibodies remained. This effect was most likely due to the small range observed for ΔG (Figure 6.6b) demonstrating that this approach was not necessarily the best. For the sake of simplicity for this thesis, this approach was used to demonstrate the pipeline, however, the full software would allow individual Z scores triaging for each of these properties.

What has been shown in the above results is that there are some clear differences in the observed ranges of physicochemical properties of clinical and library antibodies that allow triaging of antibodies that fall outside the range found in clinical antibodies. What follows is the use of the unsupervised learning models trained in previous chapters can be used in the pipeline (Layer 1).

6.3.3.2 Clinical *vs.* Library Triage

Three methods were tested to integrate the inputted Pure2 data set with the KPCA model from Chapter 3.

6.3.3.3 Method 1

The first method was to use the trained PCA model with the OAS and clinical mAb data and to transform the Pure2 data set using the model. The advantage of this

method would be that any biases in the Pure2 dataset would not affect the clustering as it is being done on a pre-trained model. This approach was unsuccessful and produced a plot in which the different groups of data were tightly condensed together and did not overlap as seen in the other methods (Figure 6.7a).

6.3.3.4 Method 2

The second method was to fit a model to the OAS and clinical data, then train a new model on the Pure2, and then to concatenate the results into one plot. This was an improvement in the sense that the Pure2 antibodies conformed to the same grouping pattern as seen when this model was trained (Figure 6.7b).

6.3.3.5 Method 3

The final method was to fit a model to the OAS, clinical data and the inputted data together. The advantages of this approach is that the OAS library would provide a snapshot of the repertoire that the inputted antibodies would be mapped onto and reduce bias in future clustering if the inputs were biased to a particular set of germlines. However, this result would likely be the most computationally intensive and may produce stochastic results with every run (Figure 6.7c).

Method 3 was chosen as the most robust, despite the increased compute time, so that the dimensionality reduction would be applied in the same way for all antibodies entered into the pipeline. Interestingly, however, when the plot produced by Method 3 was coloured by the TAP score, there was no obvious pattern in the clustering (Figure 6.8) . It was not seen, as expected, that the antibodies with the most negative tap scores would be in the extremes of the principal components. In-

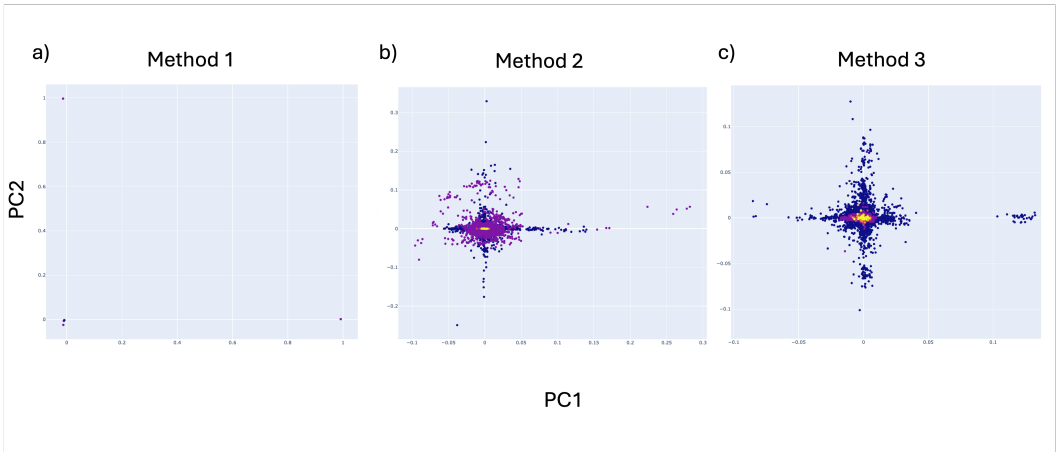

Figure 6.7: Methods of adding new data existing PCA model. with Scatter plot of Kernel PCA (KPCA) results (γ=500) of three different methods of joining together pretrained KPCA model of OAS (blue) and clinical (yellow, orange, pink) data and input Pure2 dataset (purple).

stead, what was found was that by using the physicochemical property filtering with more stringent Z scores cut-offs, the proportion of antibodies with highly negative TAP scores was removed. This would be an indicator that these physicochemical property triaging is highly effective in removing antibodies with poor developability. However, as stated, the TAP score is only a prediction of developability, and not ground truth. In conclusion, it was observed that altering the Z score of the ellipse function in Layer 1 affects the number of antibodies that are carried forward to Layer 2 (Figure 6.9).

Using decreasing values for Z scores for the physicochemical feature triaging the input sequences, drawing the ellipse function in Layer 1, reduces the number of sequences retained (Table 6.2). Increasing the probability threshold required to accept a positive prediction also decreases the number of antibodies that are output.

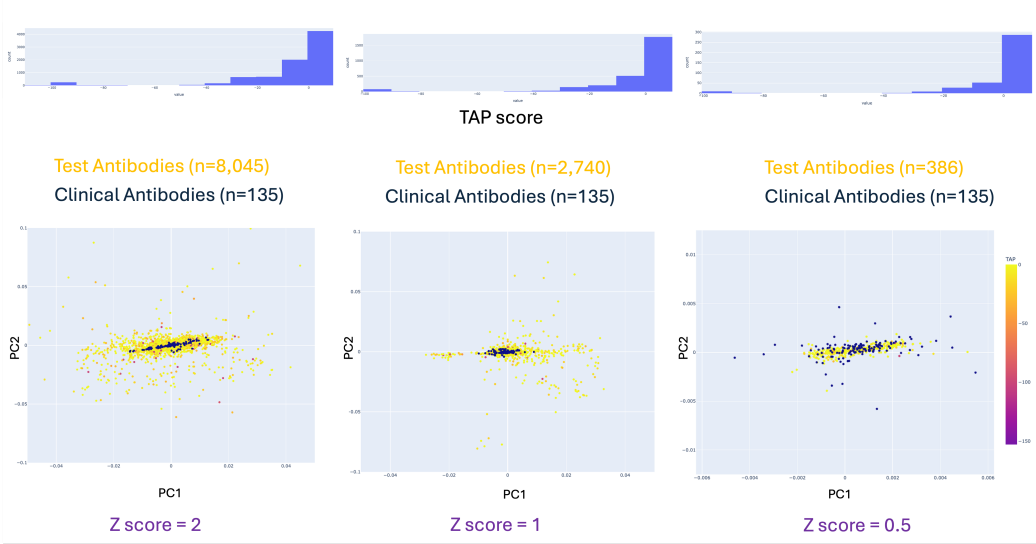

Figure 6.8: Visualising kernel PCA by TAP score. Scatter plot of Kernel PCA results (gamma=500) of three different methods of joining together pretrained PCA model of OAS (blue) and clinical (yellow, orange, pink) data and input Pure2 dataset (purple).

6.3.3.6 Approved *vs.* Discontinued Triage

In cases where no physicochemical filtering was applied, the median TAP score for each antibody output was -10, however, it was 0 in all cases where filtering was applied. From the data, it appears that the fewest negative TAP scores are obtained when the physicochemical filtering Z score is set to its most stringent setting (Z=0.5). Generally, the minimum TAP score is not improved by changing the Layer 1 or Layer 2 stringency. This result is to be expected as the properties being filtered on are similar to those used by the TAP score, and potentially, Layer 2 is trained on an antibody's chances to pass clinical trials, not directly on developability properties.

Table 6.2: Number of antibodies from the Pure2 library output from the triaging pipeline given different parameters of filtering.

Physicochemical Filtering Z Score		No Z Score				2				1				0.5			
Layer 1 Z Score	Layer 2 Threshold	PC Filtering	Layer1	Layer 2	Min TAP Score	PC Filtering	Layer 1	Layer 2	Min TAP Score	PC Filtering	Layer1	Layer 2	Min TAP Score	PC Filtering	Layer1	Layer 2	Min TAP Score
No Z Score	0.5	10492	9875	3587	-110	8045	7508	2571	-110	2740	2359	808	-40	386	308	113	-40
	0.6			2742	-110			1924	-110			580	-40			86	-40
	0.7			1894	-110			1294	-110			363	-40			56	-40
	0.8			1155	-110			772	-110			195	-40			29	-20
	0.9			449	-90			291	-110			67	-30			9	-20
2	0.5	10492	8165	2981	-110	8045	7333	2514	-110	2740	2329	797	-40	386	231	80	-40
	0.6			2263	-110			1880	-110			571	-40			60	-40
	0.7			1159	-110			1268	-110			358	-40			40	-40
	0.8			959	-110			758	-110			193	-40			21	-20
	0.9			376	-90			283	-110			66	-30			7	-20
1	0.5	10492	8186	2978	-110	8045	5855	1981	-110	2740	2056	705	-40	386	157	57	-30
	0.6			2268	-110			1482	-110			505	-40			47	-30
	0.7			1564	-110			997	-110			319	-40			28	-30
	0.8			965	-110			592	-110			171	-40			16	-20
	0.9			371	-90			210	-90			60	-30			5	-20
0.5	0.5	10492	6107	2232	-110	8045	3753	1272	-90	2740	1086	361	-40	386	39	14*	-20
	0.6			1699	-110			958	-90			260	-40			13	-10
	0.7			1187	-110			643	-90			168	-30			8	0
	0.8			729	-110			379	-90			83	-30			4	0
	0.9			281	-90			131	-90			26	-30			1	0

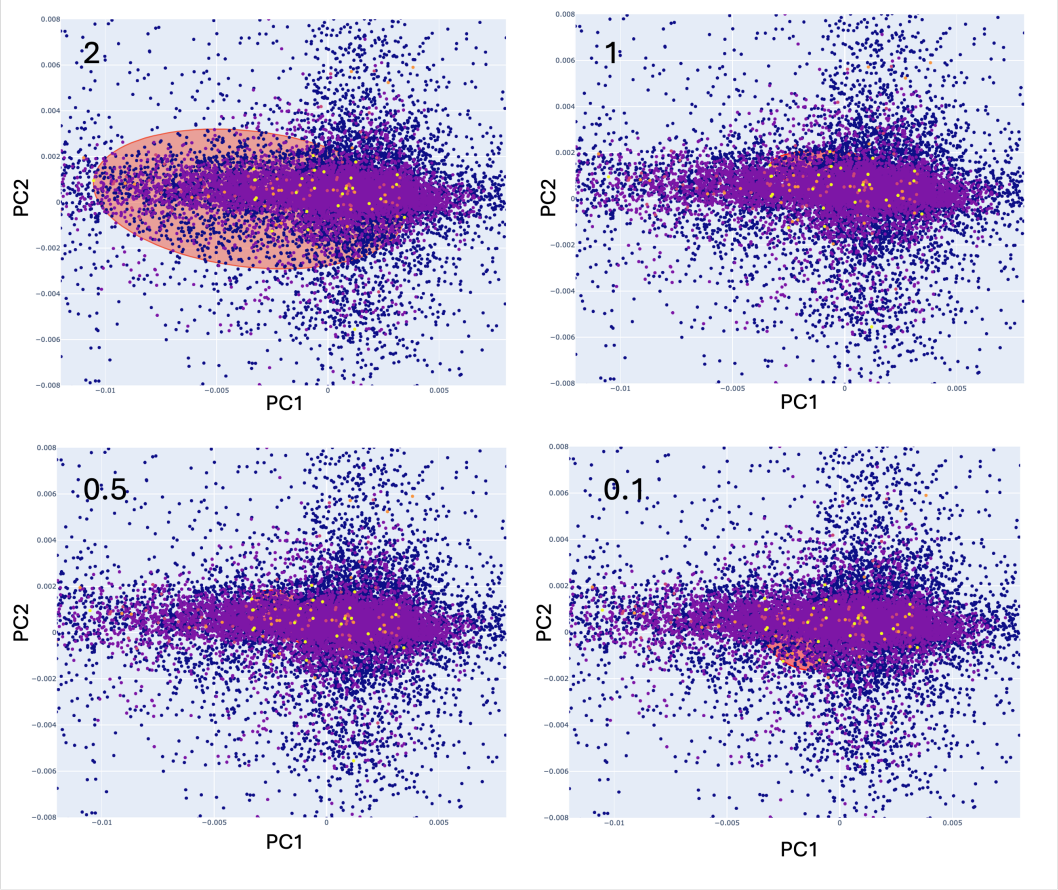

Figure 6.9: Ellipse function to select closely clustered clinical antibodies. Scatter plot of Kernel PCA results (γ=500) (blue), Pure2 (purple) and clinical (yellow, orange, pink) data and input Pure2 dataset (purple) with ellipse function drawn in red for a given number of Z scores.

6.3.4 Physicochemical Property Prediction

6.3.4.1 Immunogenicity

It was expected that the vast majority of these antibodies would not be predicted to be immunogenic because they are all human-derived. In this case, it was surprising that the binary immunogenicity predictor model predicted that 8105 antibodies would have ADA incidence >1% and 2340 antibodies below. This would imply that the model is overly sensitive, may reflect the balance of the training data, or that it has been trained on clinical mAbs which look different from the repertoire.

Table 6.3: Median values of developability properties predicted by linear models trained on experimental data.

Property	Tm (Negated)	HIC-RT	CIC	DNP	Fe.FVIII.2	Fe.C3.2	Fe.LysM.2
Pure2 Predicted Values	-69.3 ± 4.32	10.2 ± 0.62	9.4± 0.5	0.6±0.3	2.8 ± 0.7	2.5 ± 0.7	4.5 ± 0.8
Top 14 (*) Predicted Values	-71.6 ± 3.3	10.0 ± 0.5	9.5 ± 0.5	0.5 ±0.2	3.2 ± 0.6	2.6 ± 0.7	4.6 ± 0.8
Clinical Experimental Values	-71.0 ± 5.82	9.9 ± 1.23	8.9 ±0.8	0.25 ±0.8	3.3 ± 1.7	2.6 ± 1.4	5.0 ± 2.1

A similar case, where the majority of antibodies were predicted as being immunogenic, was seen for the 14 antibodies which were output by the model with its most stringent parameters (marked * in Table 6.2) where 11 were marked as immunogenic and 3 were predicted as non-immunogenic, demonstrating a similar ratio of immunogenic to non-immunogenic antibodies.

6.3.5 Predicting Developability Properties of a Test Dataset

It was seen that there was low variance in the linear model predictions of the developability properties trained on the data by Jain *et al.* [102, 103]. However, when comparing this data with the predicted data of the input antibodies from the Pure2 dataset, it was found that the Pure2 data overall had lower median stability and higher median hydrophobicity, but not polyreactivity when binding to iron with Complement (C3). In contrast, the top 14 antibodies showed overall higher stability than the rest of the Pure2 dataset, but without large changes in hydrophobicity or polyreactivity. This is not surprising since these features are not identified or triaged on by the pipeline. Overall, this result demonstrates that antibodies outputted from the pipeline ΔG was more aligned to clinical mAbs, which supports the idea that the pipeline identifies antibodies with properties similar to clinical mAbs.

6.4 Evaluating the Sensitivity of the Pipeline to Mutations

Another test for this pipeline would be to evaluate its sensitivity to point mutations in the sequence that are relevant to the pharmaceutical pipeline, where a given sequence may have point mutations introduced to optimise binding or remove liabilities. A series of experiments was carried out. In each experiment, a procedure was performed to remove sequence liabilities in which mutated sequences were substituted into the dataset, so that for each run of the pipeline 10,492 antibodies were input, but some with edited sequences. For each procedure, it was hypothesised that a greater number of antibodies from the mutated dataset would pass through the pipeline than from the original Pure2 dataset if the pipeline was left with default parameters.

6.4.0.1 Mutating Deamidation Sites

Asparagine deamidation sites can affect protein structural properties linked to a myriad of antibody immunogenicity problems [99], and are therefore undesirable in therapeutic antibodies. The antibodies in the Pure2 dataset were aligned with the Chothia numbering scheme and where an asparagine residue (N) was present, if the surface accessibility (%) determined by *pdbsolv* was over 80%, N residues were substituted for a glutamine (Q) as a neutral substitution. 10206 antibodies were edited.

6.4.0.2 Mutating Surface Methionines

Methionine oxidation sites can also affect antibody-antigen binding through changed structural properties [100]. Antibodies in the Pure2 dataset were aligned to the Chothia numbering scheme and where a methionine (M) residue was present, if surface accessibility (%) as determined by *pdbsolv* was over 20%, M residues were substituted for whichever residue was given at that position as the most frequent through the consensus data from abYsis [112]. As a result, 3125 antibodies were edited.

6.4.0.3 Mutating N-linked Glycosylation Sites

N-linked glycosylation sites can increase the risk of antibody immunogenicity or heterogeneity [97]. This motif was searched for in antibodies of the Pure2 dataset using regular expressions to locate instances of "NX[S/T]X" (where X is not P). In cases where it was found, N was mutated to Q. 1410 antibodies were edited and entered into the pipeline as before.

6.4.0.4 Results

Results show that there was an increase in antibodies outputted from the final pipeline when the deamidation and glycosylation sites were removed, but not for oxidation (Table 6.4). This effect is also seen in the unsupervised learning component ('Layer 1'), suggesting more clustering of antibodies in the centre of the KPCA when these changes were made. Overall, this was interesting because the removal of the deamidation sites changed nearly all antibodies in the Pure2 dataset, and removing glycosylation sites only changed around 10% of the dataset, yet more antibodies

were retained in this. It could be argued that, in the case of deamidation sites being removed, the number of output antibodies does not seem to reflect the number of antibodies changed. However, removing oxidation sites appeared to affect 30% of the dataset, yet the output is decreased by 747 than in the Pure2 dataset. It was thought that this could be due to the removal of conserved methionine residues or that these antibodies have additional liabilities, as seen in the chapter comparing approved and discontinued antibodies, which may result in encodings that look very different from the unedited antibodies.

6.5 Evaluating the Sensitivity of the Pipeline to Mutations in CDR-H3 Regions

This procedure was repeated in which only instances of post-translational modification sites within the CDR-H3 sequence were edited (Table 6.5). The rationale was to overcome the effect of removing conserved residues and only target the areas that are thought to affect binding the most. In the case of oxidation sites being removed, it was seen that the number of antibodies output had increased from the previous experiment, but for glycosylations and deamidations there were fewer antibodies in the outputs than in the previous experiment, despite fewer edits being made in this experiment. This was a surprising result, demonstrating that these point mutations can influence the output of the pipeline, but to investigate how, the proportions of the output that were edited were measured.

With the exception of the deamidation sites being removed, the edited sequences made up small proportions of the output of the pipeline, potentially demon-

Table 6.4: Counts of antibodies (n=10,492) output from the pipeline following removal of post-translational modification sites.

Stage	Pure2	Oxidation	Glycosylation	Deamidation
Edits	0	3125	1410	10206
PCfiltering	8085	7671	8073	7995
Level1	6689	4543	7153	7095
Level2	2284	1537	2595	2708
Edits Out		521	302	2625

Table 6.5: Counts of antibodies (n=10,492) output from the pipeline following removal of post-translational modification sites in only CDR-H3 loops.

	Pure2	Oxidation	Glycosylation	Deamidation
Edits	0	357	106	755
PC Filtering	8085	7759	7770	7771
Level 1	6689	6060	5788	5003
Level 2	2284	2110	2011	1708
Edits Out		70	24	119

strating that the models are not trained on these sequences with point mutations and that these are different enough to be reflected in the encodings. Overall, this illustrated that removing post-translational modification sites may make a given sequence more liable to be triaged out of the pipeline because the models expect the sequences to have them, and so editing these sites should be done after the pipeline has outputted a selection.

6.6 Expressing Representative Examples from the Pipeline

To truly assess how real these predictions are, it was decided to express a small number of representative antibodies from the Pure2 dataset. The antibodies to be expressed were selected from each part of the pipeline and tested in a laboratory setting for thermostability (Tm), hydrophobicity (HIC) and aggregation (Tagg).

6.6.0.1 Sample Selection

Firstly, it was decided that an antibody that would have been removed in the prediction of physicochemical properties should be tested. A representative was chosen which also happened to cluster closely with the centre of the KPCA, challenging the idea that antibodies positioned at the origin of the KPCA have good developable properties. If the antibody was shown to have poor properties under experimental conditions, this would defend the need for this early triaging step. Secondly, it was necessary to test the hypothesis that the KPCA plots were a suitable measure of developability, so antibodies from the extremes of PC1 and PC2 were selected to be expressed. It would be hypothesised that these would have poor developability features because they are located away from clinical antibodies and would have been triaged out at this stage of selection.

Antibodies were selected from their position in the KPCA. Firstly, if antibodies from the extremes of the PC turn out to be developable (A, B), it will challenge the assumption that being closer to the clinical antibodies means better developability characteristics, and will potentially mean similar germlines used. Similarly, using antibodies which the model predicts to be good (D) and bad (E) antibodies which are towards the origin of the cross plot to additionally challenge this assumption. Furthermore, this experimental validation aims to test the linear models by selecting the antibody with the highest predicted Tm (C). The positions of the selected antibodies from the KPCA are shown in Figure 6.10 and details of their predicted properties are given in Table 6.6.

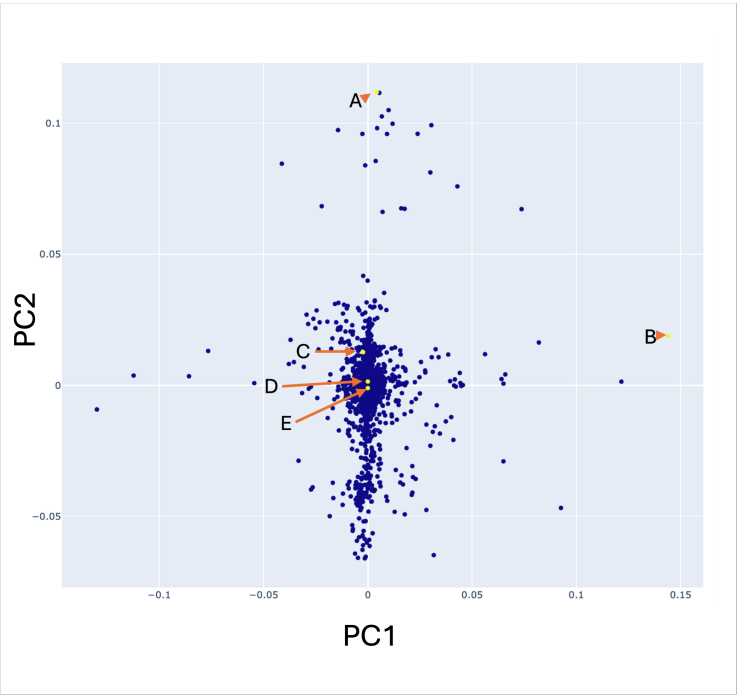

Figure 6.10: Selecting representative examples for expression. Scatter plot of kernel PCA Result (γ=500, kernel=rbf) of Pure2 dataset with highlighted antibodies considered for expression.

Table 6.6: Details of antibodies chosen for expression and predicted physicochemical properties.

Label	Identifier	Tm(Negated)	HIC-HPCL RT	CIC	DNP	Fe.FVIII.2	Fe.C3.2	Fe.LysM.2	BVP	ELISA	ACSINS	Immunogenicity
A	ACGAGGATCGCATGAT	-73.75	10.5	9.8	0.7	2.7	1.4	4.7	5.0	3.9	7.4	1
B	ACGAGCCCAAGCGCTC	-68.1	10.4	9.2	1.0	3.8	4.3	5.0	3.7	2.8	8.6	1
C	CCTACCAGTATGAAAC	-82.1	9.9	8.7	-0.2	4.0	2.71	5.8	1.4	0.7	-5.0	1
D	CTAGTGAGTAGAGTGC	-71.4	11.3	10.0	0.5	3.1	3.0	4.5	4.8	0.9	8.8	1
E	TGCGCAGGTTTGTTTC	-59.3	10.6	10.0	1.3	3.4	2.9	4.2	5.0	1.0	7.2	1

6.6.0.2 Results of Experimental Validation

Antibody expression and developability tests were performed by GenScriptBio. The method of expression and experimental protocols are described in Appendix C. The performance of the test for the five antibodies is given in Table 6.7. The degree

Table 6.7: Developability assay performance of selected antibodies.

Label	Reasoning	HIC-HPLC RT(min)	Tm (°C)	Tagg (°C)
A	Most extreme (PC2)	28.1	70.8	72.1
B	Most extreme (PC1)	27.7	70.5	63.1
C	Highest predicted Tm	26.6	68.1	59.3
D	Finalist Antibody Proximal to Origin	32.3	68.7	67.2
E	Triaged Antibody Proximal to Origin	28.3	74.8	65.1

of correlation between experimental and predicted values for Tm and HIC as given in Table 6.6 using Spearman's rank correlation (ρ). Rank correlation was used over correlation coefficient to remove the effect of the difference in absolute values between experimental set-ups, and because of the small dataset. A near-perfect rank correlation was found for the predicted HPLC-HIC RT (ρ = 0.97) but a poor correlation was found for the Tm (ρ = -0.88) compared to the experimental results here. Both predictions for HIC and Tm were made using linear models trained on experimental data [102]. Jain *et al.* [102] did not take the Tagg measurement, so there is no direct prediction to compare against. Despite this, a strong correlation was found between the experimental values of Tagg and the predicted values for Cross-Interaction Chromatography using models trained from the Jain data, which another measure of aggregation or self-interaction (ρ = 0.80).

The results from the experimental procedures demonstrated that the finalist antibody (D) showed a high Tagg but also the highest HPLC-HIC RT of all of the antibodies selected. In contrast, the antibody triaged in the early pipeline (E) showed the highest Tm and the shortest HPLC-HIC RT, which seems counterintuitive. Furthermore, antibody which had the highest predicted Tm from the Pure2 dataset (C), was found to have the lowest experimental Tm and the lowest Tagg. The antibodies at the extremes of the PCs (A & B) were found to be in the middle for most of the experimental values, except for antibody A which has the highest Tagg. Both antibodies have higher experimental Tm antibody D.

6.7 Discussion

The construction of the pipeline was decided on the basis of the results discovered throughout this thesis for separating clinical antibodies from library using encodings from the AntiBERTy language model.

The Pure2 library was made available to the author by the Franca Fraternali's group as a representative dataset to be used on the pipeline. The dataset was made up of antibodies from healthy young and old individuals, keeping this test agnostic of any known target and potentially eliminating biases of over-representing antibodies for a specific antigen. However, the intended application of this pipeline would be used in libraries where a large proportion of antibody would be expected to bind to a given antigen, such as immunisation campaigns. The differences between Pure2 and clinical antibodies especially apparent when different proportions of different heavy and light chain pairings were observed in the Pure2 dataset than in the clinical approved and discontinued mAb dataset as used in Chapter 3, but this does make sense as it has come to be shown that clinical mAbs are a highly selected group. With this in mind, it was surprising that more than half of the Pure2 antibodies received no red flags for developability by the TAP score software, suggesting that most of these antibodies sit within the ranges of the selected physicochemical properties observed in the clinical dataset. However, calculating these scores did require 60 hours of computation time to complete the dataset, and it begs the question how to select the best antibodies if there is no other means of differentiating the antibodies.

In terms of physicochemical filtering, the observation that setting filter triaging

based on ΔG had more filtering effect is due to the repertoire used having a greater variance of ΔG values compared to the marketed antibodies than for any of the other properties that were checked, however, this method also removes antibodies with larger ΔG values than seen in the approved antibodies, which would in principal have greater stability and unlikely to be discarded in other selection pipelines. However, in defence of this approach, other developabilty detection software also focuses on identifying antibodies with traits similar to the currently approved antibodies will also triage out antibodies with perceived positive traits where the value is above what is observed in the clinical antibody dataset [110, 107]. Perhaps, it would have potentially been better to have treated the two chains as a "weakest link" problem, and filtered based on the chain with the furthest deviation from the mean, however, the approach taken in here aimed to be hollistic, and to take into account the relationship between both V_H and V_L chain.

What was more surprising was the large synergistic effect of all these filters combining to reduce the sample size from 10,492 to 386 when the Z score was set to 0.5, a 96% reduction in the size of the data set. Arguably this kind of stringency makes the result of large libraries more manageable to work with to do more computationally intensive tests, such as software which requires modelling, or even making use of the TAP software. As has been shown given the overlap between approved and discontinued antibodies, it is inevitable that this sample will still have antibodies that are likely to fail clinical trials, even if they have developable traits, and so additional triaging would be needed. Furthermore, it was surprising that only one antibody remained out of 10,432 when all pipeline parameters were set to their

highest stringency, which is why the 14 top antibodies of this setting were chosen to select from for experimental testing, although it would have been interesting to have tested this one too. Although the TAP score was used only as a comparison for the pipeline developed here, it was reassuring to see that there was agreement when more stringent parameters were used to triage antibodies out of the Pure2 dataset. More stringent parameters demonstrated that the minimum TAP score observed for the antibodies in that output was less negative. This was mostly observed when smaller Z scores for physicochemical property filtering were selected, rather than when the ellipse function or the clinical trials model probability function was selected. This was to be expected, as the phyicochemical property triage would be working to remove antibodies with developability red flags that were more similar than those of the other models.

The pipeline was evaluated for its sensitivity to mutations usually employed to remove deleterious post-translational modification sites. It was surprising that removing deamidation sites edited so many antibodies and only gave a small increase in the number of antibodies output from pipeline, whereas removing methionine oxidation sites caused a reduction output antibodies. Furthermore, removing the sites from only CDR-H3 loops only caused a marginal difference in the number of antibodies out, in each case, it was a small proportion of the antibodies which were edited. This has suggested that while the pipeline is sensitive to mutations, it may not be in the way that would be expected. When they are mutated out, the encodings must change significantly enough that this changes how the fate of an antibody would be predicted at clinical trials, or whether it is a repertoire or clinical

sequence.

The predicted immunogenicity of the antibodies was surprising as it was expected that the majority would be considered non-immunogenic, however it appears the model trained was probably overly sensitive and predicted the majority would be immunogenic. This suggests that the ADA dataset used to train this model was not suitable for one or of for the following reasons: It was made up of clinical mAbs which have been shown to be a selected dataset which may not translate to library antibodies; clinical mAbs used in the dataset have a number of different sources and possibly the model is biased to the training dataset and giving rough proportions of outputs as it received for inputs. For this reason, the immunogenicity model was considered untrustworthy.

Although the budget for this experiment only allowed for 5 antibodies to be expressed, it was attempted to select antibodies that would test multiple hypotheses, rather than selecting 5 antibodies from the Pure2 dataset that were passed through the model. For this reason, it was thought that testing antibodies in the extremities and the origins of the KPCA plot of Pure2 antibodies, the OAS dataset and clinical antibodies to explore if developability features were being selected. This led to the selection of antibodies based on their position within the KPCA, regardless of their predicted physicochemical properties, as it was unclear at this point if these predictions were reliable.

It was reassuring that while the absolute values demonstrated differences that could be accounted for by the experimental setup, there was a strong rank correlation between the predicted and experimental values for HIC in the selected antibod-

ies, which strengthens the reliability of this predictor as a relative indicator of HIC. This was not the case for the Tm predictor, suggesting an element of overfitting in the predictive model. An explanation for this effect could be that impurities or different compositions in the samples from the Jain *et al.* data affected the recorded melting temperature in this dataset but not hydrophobicity [197]. Thus, making the results from these experiments less reproducibile using models trained on their data and applied to the antibodies used in this thesis.

Another explanation, though more speculative, is that in experimental procedures for Tm and Tagg are measured, the antibodies were seen to melt, or to aggregate at temperatures which would not be physiologically relevant (e.g 68°C-74.8°C). This is counter to a hydrophobic antibody, which would would have an effect at physiologically relevant temperatures. Therefore, there would be no evolutionary selection against proteins which unfold at this temperature range, and are tolerated as long as they are stable within physiologically relevant temperatures (i.e. body temperature). This would mean that it is impossible for the LLMs to learn to predict Tm or Tagg because it is out of context from the sequences that it has been presented, whereas analogous information has been encoded by the LLMs that can be used to predict HIC.

Although it appears that the antibody chosen to represent the output of the pipeline at its most stringent parameters had poorer developability experimental values of all of the other selected antibodies, this was not seen as a problem, as this antibody also had the highest predicted HIC of all of the other outputted antibodies at these model parameters. From this it could be assumed that an output antibody

with a low predicted HIC would reflect the truth, and that had HIC filtering also been used, this antibody would have been rejected. From this, it is reasonable to suggest that when using this pipeline, optional cut-off points for predicted developability values, particularly HIC, would be an additional useful step to triage out antibodies with poor developability.

This raises questions about why an antibody output by the model would still have poor developability features. An obvious hypothesis is that these 14 output antibodies with the pipeline's highest stringency parameters conform to the germline biases seen in the approved and discontinued dataset. This was investigated but it was seen that while most of these came from IGHV3/IGKV1 (n=5) and IGHV3/IGKV3 (n=4) pairings, antibodies from IGHV3/IGKV4 (n= 2), IGHV1/IGLV2 (n=2) and IGHV4/IGKV6 (n=1) pairings were also represented, suggesting that a feature other than germline gene pairings has been selected. However, it would be advised not to use the pipeline at its highest parameter stringency for fear of removing useful sequences, and potentially over-parameterising can lead to selecting antibodies which conform to another bias which has not been identified in this thesis.

What can be concluded from laboratory experimentation is that antibodies with developable properties can be found throughout the KPCA space and are not necessarily proximal to the origin, which raises the question of why this seems to be the case for the clinical antibodies used in this study. This could be because of biases in the language model that have caused them to cluster this way, and a germline bias still persists.

6.8 Conclusion

To conclude, it has been shown that the pipeline assembled from the models trained in previous chapters of this thesis can be used to select antibodies from a test library of repertoire BCR sequences that are predicted to have characteristics similar to current clinical mAbs. By setting more stringent values for the parameters of the models used, it is argued that a better quality output can be obtained, which is supported by the comparison with the TAP score. Furthermore while the experimental work has supported the relative predictions made by models trained on experimental libraries, it has also demonstrated that *in silico* predictions are not yet replacements for the experimental work necessary to demonstrate an antibody's success in the clinic.

Chapter 7

A New Annotation Language and Interactive software for Multispecific Antibodies

The work presented in this chapter has been published in Sweet-Jones, et al. (2022) Antibody markup language (AbML) - a notation language for antibody-based drug formats and software for creating and rendering AbML (abYdraw). mAbs 14:e2101183.

7.1 Introduction

Multi-specific antibodies (MsAbs) are an up-and-coming class of biologic drugs aimed at building on the success of mAbs in the clinic. These therapeutics require molecular engineering techniques to graft multiple *Fv* fragments onto one structure that can bring together two cells, a signaling molecule to a cell, or enhance a step in a drug pathway [78, 79]. This can be done using a number of genetic and chemical

methods, which have generated a host of unique MsAb formats that have become recognised by the WHO-INN. The MsAb formats which have reached market approval by governing bodies at the time of writing: Emicizumab; Blinatumomab and Catumaxomab, are all bispecific (binding to two different epitopes). However, trispecific and tetraspecific antibodies have also been developed, which are of great interest in cancer immunomodulatory therapies.

The growing diversity of these molecules and their increasing entry into clinical trials have demonstrated a need to encourage users of antibody discovery pipelines to adopt a machine-readable standardised description of these formats in this thesis to engage in MsAb engineering. The development of an annotation language that could address the shortcomings of previously used annotation methods, including HELM [93], was thought necessary. As previously described, HELM has a number of shortfalls when annotating MsAbs, including the inability to notate specificities of different *Fv*s, the requirement to provide sequence data for the therapetic and an overly complex editor.

In this chapter of the thesis, the author presents a new antibody annotation language, Antibody Markup Language (AbML), is presented. This is designed to encourage users of the antibody discovery pipeline to explore bispecific antibody engineering by addressing the needs of the antibody community to describe these formats.

7.2 Development of Antibody Markup Language (AbML)

Work outlined in this section 'Development of Antibody Markup Language (AbML)' was originally done by Maham Ahmed with Andrew Martin as part of the MSc Bioinformatics Titled 'Describing the Format of Antibody Based Drugs' (submitted in 2020). The contributions made by James Sweet-Jones are listed below:

- *Naming AbML was done by Andrew Martin and James Sweet-Jones.*
- *Additions of TCR domains to AbML after request from reviewers.*
- *Sequentially numbering domains. This was done to make rendering images more convenient and logical.*
- *Chemically bonded linker domains (-C-) after reading the work of Szijj* et al. *[88].*
- *Removing whitespace in AbML expressions to avoid confusion in readability.*
- *Positioning heavy chains first, and light chains and antibody fragments to the end of AbML expressions for convenience in rendering images.*
- *Allow ASEQ and DSEQ expressions if a specified sequence is provided after request from reviewers.*
- *Specifying which domains may be compatibly paired for the convenience of drawing schematics.*

- *CLASS comment to specify which immunoglobulin class a domain originates after reading the work from Heinke* et al. *[204].*

- *Development of the abYdraw software to draw and render AbML descriptors*

AbML was designed to have a structure similar to HELM [93], but adapted for MsAbs using Ig-like and TCR-like domains. For example, HELM would require one to specify a constant heavy (`'CH'`) domain and add a comment to specify which C_H type it is (C_H1, C_H2, etc.). To simplify this, AbML adopts separate domain types (`'CH1'`, `'CH2'`, etc.). With these ideas in mind, the requirements for AbML were devised as follows:

- The language needed to be simple to encourage its use.

- It needed to be sufficiently flexible to describe all current MsAb formats and all those that could be envisioned in future.

- In addition to standard antibody domains, it needed to be able to describe modified domains (e.g. knobs-into-holes), non-antibody domains, and chemical conjugation.

- Interactions between domains and (multiple) disulfide bonds linking domains needed to be described.

- The specificity of different V_H/V_L domains needed to be indicated.

- Three types of connection between domains needed to be allowed: normal peptide connections between domains, natural (or engineered) hinge regions and artificial (engineered) peptide linkers.

- AbML needed to support additional optional comments including general notes, types of additional domains, modifications and region lengths.

AbML is based on the description of a chain of antibody domains, separated by connectors, from the N-terminus to the C-terminus in a plain text string. The aim is to provide as simple a format as possible while conveying all necessary information and allowing users to specify additional information such as sequences and modifications.

Each domain is numbered sequentially in order of its appearance in the expression. In this respect, hinges and artificial linkers can be considered more like domains, as they are numbered and separated from neighbouring domains with a '–' character, which represents naturally occurring linkers. Whitespace, including line breaks, is ignored in AbML except for comments given in square brackets.

Antibody chains are separated by '|' characters. Chains that are part of the antibody molecule can be presented in any order, but any additional chains that interact with antibody chains via a disulphide or a domain pairing with a domain conjugated to the antibody) are placed last. In a multi-chain structure, every chain must have at least one domain that interacts with a domain on a different chain, and pairings may only be between compatible domains. These are specified V_H/V_L, C_H1/C_L, C_H2/C_H2, C_H3/C_H3, C_H4/C_H4, C_H5/C_H5, L/L, X/X, VA/VB, CA/CB, VG/VD, CG/CD. The properties of the language for Ig-like domains are shown in Figure 7.1 with full specifications provided with the publication [205].

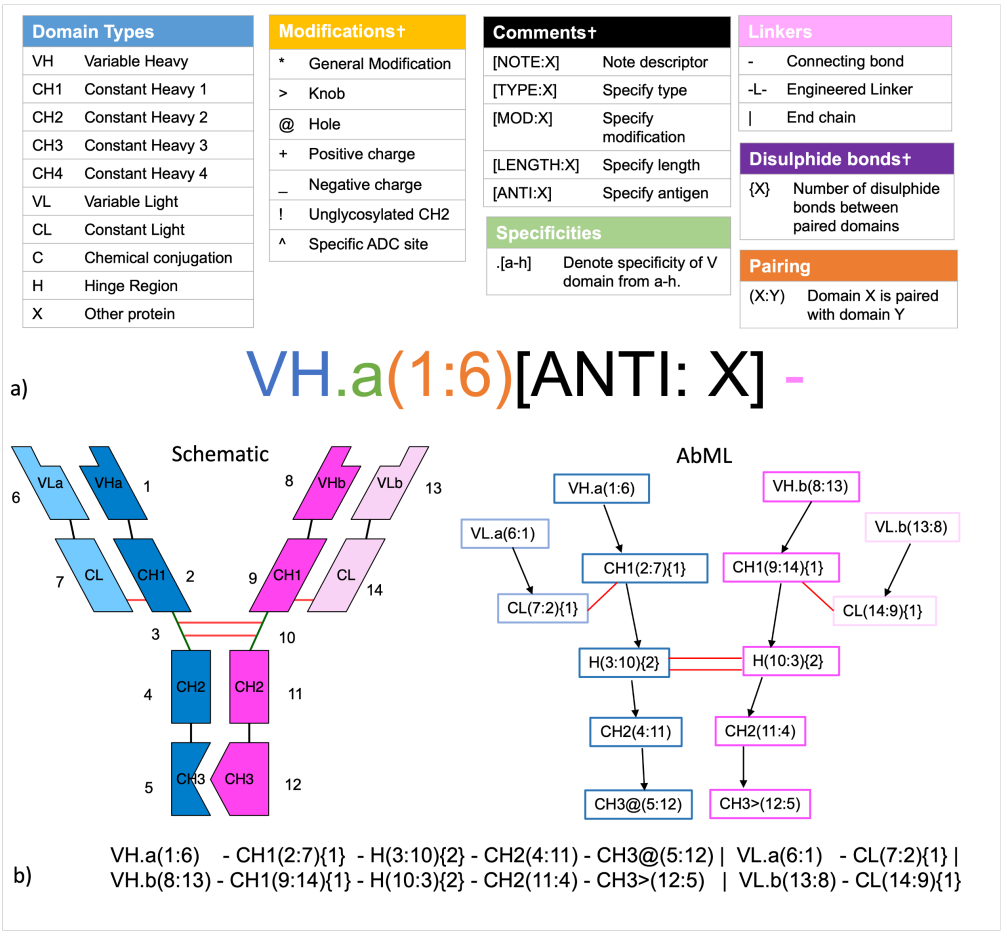

Figure 7.1: Building AbML expressions for an antibody structure from domains to chains. a) Each domain is an individual unit containing information about the domain type, modifications and interactions with other domains. Order of information is shown through the largeprint domain text colour-coded to above tables of all possible entries. b) Chains are made up of a group of domains conjugated by a selection of linkers. The numbered schematic and AbML schematic of the bispecific knobs in holes (KIH) MsAb demonstrates how its corresponding expression below is constructed for each chain from the N-terminus to C-terminus. Blue and Pink colouring represent different antibody specificities.

7.3 Development of abYdraw

It was suggested by Ahmed in the discussion of her thesis that a graphical editor for the antibody annotation language would be a good tool to promote it, so this work was also carried out as part of this thesis. It was to be called abYdraw to follow the naming convention of other software released from Andrew Martin's group, and while it was initially developed only to render AbML strings as images, it was extended so that point-and-click actions and assembling schematics of MsAbs could then produce the AbML string. abYdraw was implemented in Python3 using TKinter to design the interface, which is presented in Figure 7.2. However, a command-line interface allowing abYdraw to be used for automatically rendering AbML strings was written to preempt a time where it would be useful to generate images on the fly in web pages.

7.3.1 Drawing MsAb formats from AbML expressions

Users may enter a valid AbML expression in the text box highlighted in the schematic to obtain a schematic of their designed antibody by clicking the `'Get Structure'` button. Originally, it was planned to include the AbML checker developed by Ahmed as part of this pipeline, but this was not found to work when the specifications of the language were changed (Algorithm 3).

The 'Domainmaker' function was written to calculate the coordinates of a given domain given the centre of the apical edge of the domain. Each domain was drawn as a rectangle with the short edges being 40 pixels across and the long edges 80 pixels tall. If the domain is drawn before the hinge, these were drawn at a slant

Algorithm 3: abYdraw method drawing MsAb formats from AbML expressions.

- Identify the number of chains and check for AbML errors.
  - An unrecognised domain type that is not allowed is entered in the AbML expression
  - A domain is missing its numbering, shares its number with another domain or is paired with a domain that does not exist
  - A domain with incompatible modifications (both positive and negative charges, both knob and hole mutations) has been entered

if *Error in AbML* **then**
  Raise Error
else

- Loop through each chain to identify which domains interact with a domain on another chain. This performs pairing. Using the information about where these interactions are, the software decides where the C-terminus domain of each chain should be drawn. This is given as a set of coordinates in 2D space.
- Starting with the lowest number domain on the C-terminus of the first chain, the software loops through each domain and connecting bonds of the AbML and decides: what the domain should look like; what colour it should be; what the label should be using the information given in the string. Furthermore, the location of where the domain will start is calculated, and this most likely depends on what domains come before or after it. For instance, if the current domain is paired with the previous domain, it will be drawn adjacent to it, if not, it will be drawn below it. Bonds are drawn from the centre of the basal edge of one domain to the center of the apical edge of the next.
- The step of calculating the coordinates for each domain is done using the DomainMaker function, which takes into account all of this information and realises the coordinates of each vertex of that domain. These coordinates, alongside a suitable domain label containing any modifications, interaction information and bonding information are saved in a domain object.
- Once all of the domains in each chain have had their coordinates saved in an object, these coordinates are first checked for any clashing domains, and if this is the case, they are spaced accordingly so they do not overlap. Then, they are rendered and filled with an appropriate colour based on which antigen binding specificities are on that chain. Any domain modification notes are written alongside the MsAb figure.

end

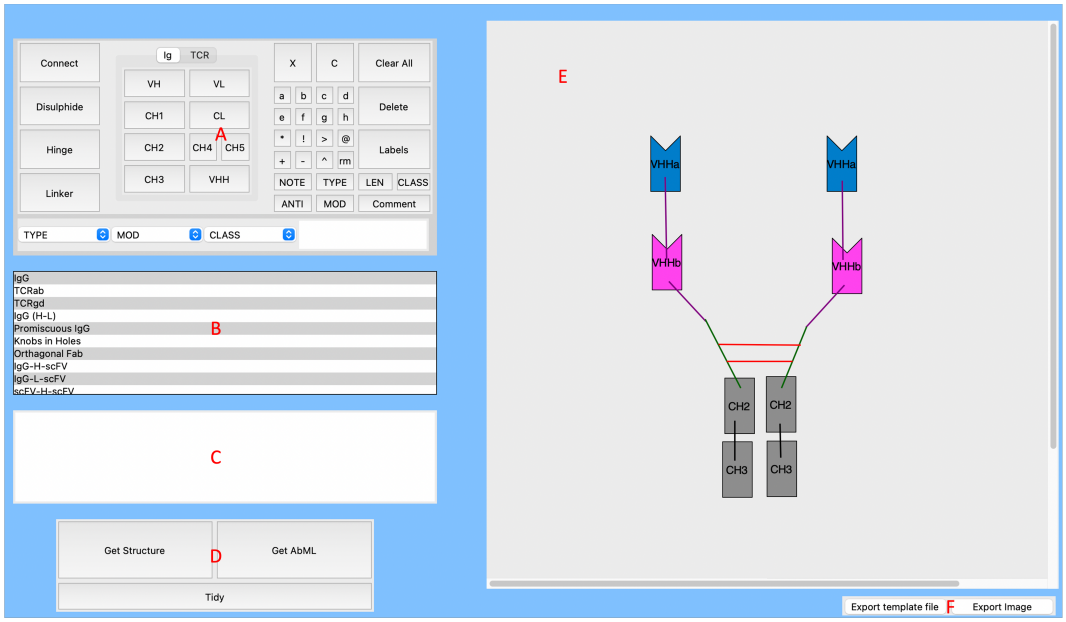

Figure 7.2: abYdraw interface. a) Domain palette which has buttons necessary for drawing antibody domains. b) Library of commonly used antibody AbML expressions. c) Textbox for inputting AbML expressions and receiving output AbML expressions.. d) Button pad that will render antibody schematics or output AbML to the textbox. e) Canvas for drawing and rendering antibody schematics. f) underneath there are two buttons which are involved in exporting the schematic.

where the coordinates were calculated using trigonometry.

7.3.2 Generating AbML Expressions from Drawings

The user is also able to draw antibodies by using the palate to select a domain, any specificities and modifications, and clicking the canvas to place the domain. The domain object is then written with those coordinates and modifications. These domains can be moved using a click-and-drag method or edited by selecting a modification and clicking the desired domain, and the data stored in that object are edited accordingly. Bonds can also be drawn between domains using a click-and-drag method, ensuring that the bond starts and ends inside two different domains in a N-terminus to C-terminus direction. Bonds are also saved as objects, but can be

natural connectors, hinges, or engineered linkers. Domains can be paired by placing compatibly paired domains adjacent to each other, which for V_H and V_L domains, may require using the right click to reverse the orientation of the domains so that they interact to form a complete antigen binding site. Furthermore, a library of popular structures is included which can be easily loaded onto the canvas and can then be edited.

Users may draw antibody-based drugs from scratch or begin with a template design of common formats (including MsAbs) that may be manipulated by the user. Normal connections between each domain are given by black lines that are drawn from the bottom of one domain to the top of the next domain.By default, artificial linkers are shown as purple lines, disulphide bonds are shown as red lines and hinges are shown in dark green. These default colours for all domain and bond types may be changed in the settings menu. Variable domains appear with a cutout at the top of the domain referring to their antigen-combining site, which pairs with another to give a complete *Fv* fragment. Nanobody domains (i.e. a V_H domain that does not interact with anything else and indicated as `VHH`) have a unique domain shape reflecting their single-domain binding site.

KIH adaptations are displayed in constant domains with either a cut-out or an extension to their side, which slots together to demonstrate how these domains are paired. By default, domains are coloured according to their specificity descriptor. Consequently, it is possible that chains will have blocks of different colours when domains of different specificities are given in the same chain. Disulphide bonds can be drawn starting from either of the interacting domains (including linkers and

hinges). To insert a comment (e.g. NOTE), the appropriate comment type button is clicked and, in the case of TYPE and MOD which have restricted allowed values, the required value is selected from a drop-down list. If the desired comment is not available, comment text is typed into the text entry box and the required domain is clicked to associate the comment with that domain. AbML also allows sequence information to be associated with each domain or chain using ASEQ and DSEQ keywords for amino acid and nucleotide sequences, respectively. Once the drawing is arranged, by using the 'Get AbML' button, the appropriate AbML expression can be retrieved (Algorithm 4).

Once the AbML is obtained for the drawing, using 'Get Structure' will re-render the schematic automatically. Both functions can be run in sequence using the 'Tidy' button. In the case of the negative charge modification, the '_' is replaced with a minus sign in the rendered image. For KIH modifications, the characters '@' and '>' are omitted, as these modifications are used to affect the shape of the rendered domain. abYdraw can be used to export these schematics as figures for publication and to generate a standardised expression that may be used in MsAb annotations. The interface draws domains as blocks labelled with their domain type and any specified modifications. A selection of structures included in the library of the software is given in Figure 7.3.

7.3.3 Software Availability

Compiled apps for Linux, Mac OS and Windows are freely downloadable [1] while an introduction to AbML and the latest AbML Format Description (i.e. any updates

[1] http://www.bioinf.org.uk/software/abydraw/

Algorithm 4: abYdraw method of constructing AbML from drawn MsAb domains.

while *Domains Remain on Canvas* **do**

- Loop through all domain objects and locate the object with the highest Y coordinate. This will be the domain at the start of the first chain.
- Loop through bonds to find a connection between the current domain and another (allowing some margin of error in the drawing). If a comment has been added to a domain, add it to a domain object. Continue looping through the chain until a domain is reached that does not have a connection. If no connection is found, check if a hinge or engineered linker has a connection to the current domain. If so, continue from there, otherwise, take this as the end of the chain.
- As domains are found, remove these domains from the pool of domains on the canvas.This will give plaintext representations of AbML strings, but without numbering

end

while *Domains remain unnumbered* **do**

- Loop through string and assign numbering to each domain, hinge, and linker then pair domains based on closeness:
  - For each domain, loop through all other domains and identify if any other compatible domains are within a given distance threshold to consider them paired. If multiple domains are observed, the closest is taken.
  - Write pairing data for that domain to the domain object and the interacting domain and number of disulphide bonds drawn between domains, then remove both domains from the loop. If domains are V_H and V_L , depending on the specificity given when drawing, that information will also be saved. If no information was given, all antigen binding sites are assumed to have the same specificity.
- Rewrite AbML expression with numbering, pairing, and modification information and print domain to text box.

end

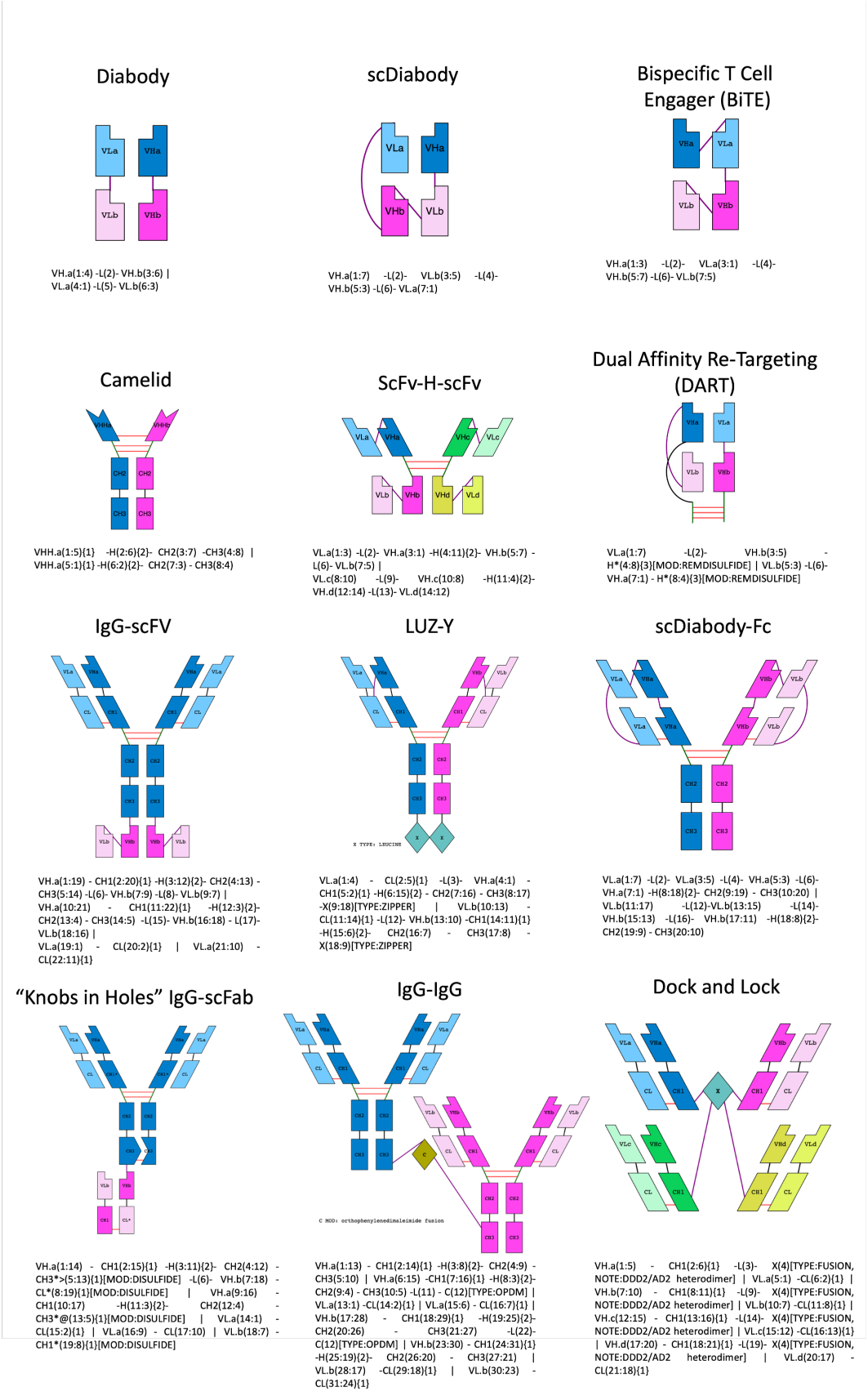

Figure 7.3: Popular MsAb formats and their AbML expressions. Given formats are included in the library of the abYdraw program and rendered through abYdraw. Domains are color-coded to demonstrate different specificities where lighter shades show light chains and darker shades show heavy chains. Peptide bonds are shown in black, hinge regions in green, artificial linkers in purple and disulfide bonds in red.

to Supplementary File 2) are available [2]. The source code for abYdraw, released under GPL3, is also available [3]

7.3.4 Use Cases of AbML

In summer 2022, all mAbs and MsAb submissions to WHO-INN received AbML annotations by generating a script that transformed the WHO-INN annotations into AbML, which included 567 therapeutics, both mAb and MsAb, but also ADCs and conditionally active therapeutics. It is also used for future WHO-INN submissions as part of their detailed annotation format. This has led to some reworking of the code to allow new formats of MsAbs antibodies to render. This has included: structures from IgM type; structures where a heavy chain may start with a H region; allowing structures with two Hinge regions; X proteins on different chains forming homodimers and allowing structures with X domains rather than Hinges. This has led to the demonstration of how lively the MsAb field can be and the fact that abYdraw is required to be flexible.

7.4 Discussion

By addressing the pitfalls of currently available annotation languages, we have developed AbML which is loosely based on the established HELM notation for macromolecule biologics but has been simplified and adapted specifically to describe antibody formats in a straightforward manner. AbML has been carefully designed to allow annotation of future possible formats, and we have demonstrated that it can be applied to all existing MsAbs described by Spiess *et al.* [79] as well

[2] http://www.bioinf.org.uk/abs/abml/
[3] https://github.com/JamesSweetJones/abYdraw

as newer antibodies listed by the WHO-INN.

The simplicity of AbML over HELM allows greater accessibility as well as allowing the potential to extend the language in the future by inserting additional modification symbols and domain types that will futureproof the language to cope with the inevitably expanding formats of recombinant and chemically conjugated antibody-based drugs. In general, the `'X'` and `'C'` domains can be used to describe a multitude of possible fusion proteins, drug conjugates and chemical bonds using the comments system, and consequently we do not expect the language to require constant updating.

Limitations of abYdraw are anticipated and may need to be addressed in the future. It only supports eight specificities (i.e. letters a–h), but this should be enough for all currently conceivable constructs. abYdraw also limits domain pairings to those normally seen, however, there is nothing to prevent non-standard pairings being present in the AbML language. In addition, interactions may be specified between 'extra' (non-antibody) domains and chemical conjugation moieties. This could be improved by allowing better rendering when linking two non-identical domains e.g. X/L pairings which are possible to specify in AbML, however, this was not considered when writing the software and so would require some rewriting of the software to allow this. Although it supports ADC drug conjugation (random or site-specific) in AbML, these are not currently rendered or supported in abYdraw and we foresee the need to support associated features, including spacers and specific payloads [206].

The work completed in this chapter was the first section of work to be com-

pleted for the thesis while the author was learning the programming techniques and machine learning skills required for the other chapters of the thesis. Had this programme been written with more experience, it would be clear from the start that object-orientated programming would be required to make the most efficient programme, instead of using Python Dictionaries, as done here. Furthermore, it would have been more suitable to have written abYdraw in JavaScript, which would allow the graphical interface to be used via a web page without the need to install software locally. This would be something to return to if the popularity of abYdraw was maintained. Some initial attempts to do this together with a more dynamic 'energy'-based layout algorithm have been made.

Since this work was published, the work by Biswas *et al.* [207] saw the inception of VERITAS, another markup language that criticised AbML for appearing overly complex and difficult to read when comparing different strings. Although it could be agreed that it might be difficult to miss these details, that was the whole point of abYdraw, so that these strings could be rendered and compared visually. By comparison, VERITAS aims to condense these domains and chains into functional units such as Sc*Fv*s, Sc*Fab*s and proteins. However, they decided to omit information about binding specificity or allow added comments in the written language, making it unsuitable for MsAbs and unsuitable for describing antibodies with modifications. Condensing the information to make it more readable has led to critical information being lost, and the auuthor would still advocate for using AbML to obtain full descriptions of MsAbs.

7.5 Conclusions

To conclude, the annotation language AbML is a new descriptor language for MsAb formats, and its ability to annotate all existing MsAb formats has been demonstrated. We expect this language and its corresponding tool abYdraw to become useful in the development of future MsAb drugs, allowing for standardisation of MsAb description as part of ushering in a new era of MsAb drug development. Improved descriptions of their formats and graphical interface for design is anticipated to allow accessibility for MsAb engineering for sequences discovered by the pipeline in question in the thesis. It is planed to reimplement abYdraw in a more flexible and dynamic manner to develop software to compare ASML strings since the format allows the same structure to be described in different ways.

Chapter 8

Conclusions and Future Directions

The purpose of this thesis was to develop *in silico* methods to identify antibodies with developability characteristics similar to those of clinical stage mAbs from sequence libraries. To do this, methods of encoding library sequences into numerical data to be input into machine learning algorithms were devised in order to identify library antibodies that cluster closely with clinical mAbs in a 2-dimensional projection of a high dimensional space, and for prediction of physicochemical properties relevant to developability. Furthermore, work here has developed an end-to-end pipeline with parameters capable of being toggled to suit researchers' needs or criteria, and a graphical user interface program for an annotation language developed for describing MsAb formats. This chapter concludes the investigation that was carried out in this thesis.

8.1 Construction of the Pipeline

Although it was originally set out only to predict these features on the basis of sequence statistics, it became clear that this would be difficult to perform in a timely

manner due to the time taken to calculate the amino acid encodings. It is understood that LLMs would encode some structural information, which is likely necessary to distinguish between clinical and repertoire, or approved and discontinued, and could be encoded quickly using a graphics processing unit. The drawback of using these LLMs is that it becomes difficult to identify what information is being highlighted for this purpose. When this was investigated for models trained on approved and discontinued mAbs, it was found that more features were used from the V_L chain than the V_H chain, which was unexpected but could suggest more unknown functions or features of the light chain in therapeutic mAbs relevant to their success at clinic [208].

As stated above and concluded in Chapter 5, the clinical stage mAb dataset is a selected and biased dataset, where each therapeutic would have had to be chosen from a pool of possibilities for its binding efficacy, and developability and these characteristics may have been optimised *in vitro*. At this stage, it is necessary to overcome preclinical toxicology tests and must be produced on a moderate scale, so it is not surprising that this dataset was seen to cluster closely in the KPCA. However, achieving a developability profile similar to current clinical mAbs does not guarantee success and, for a variety of reasons, most therapeutics fail in clinical trials. Although there were no statistical differences in the physicochemical properties between discontinued and approved antibodies, it was found that machine learning classifiers could be trained on amino acid compositions and LLM encodings, which could distinguish them with a high degree of performance using feature selection. Although this association could be due to chance because of the sheer number of

data points used per sequence, it was seen that these trained models also showed modest success with a set of held back data, indicating the relationship found was real. This information, and the models from it, have potential to be used to predict features early on which may go on to compromise a given antibody in clinical trials. Despite this, it is nearly impossible to disentangle what these features would mean and how they are influence developability.

8.2 Applications for the Pipeline in Therapeutic Antibody Discovery

The pipeline does not contain any information on binding to a given target, and throughout the thesis it has remained agnostic to the target. This was intentional to ensure that a given target would not influence the machine learning models and thus could be used to search for any target. It would be expected that the library entered into the pipeline would be a library generated as the result of immunisation campaigns in humanised mice toward a given target. Following immunisation where the mouse generates a high proportion of antibodies toward the target, these proteins can be sequenced to generate a library of about 10,000 paired nucleotide sequences in the same way other libraries have been generated. However, not all of these sequences will bind to the target and not all of these antibodies will bind with high affinity, or to a useful epitope but the aim of the pipeline is to identify antibodies in this sample with a developability profile similar to that of clinical stage mAbs and those that have a high probability of passing clinical trials thus reducing the number that need to be analyzed experimentally. Once these antibodies are

identified, the binding affinity can be tested using the phage display, with the advantage of knowing that any hits are predicted to be good quality antibodies before optimisation.

The experimental validation did contradict the hypothesis that developable antibodies would only be found close to the clinical mAbs in the KPCA. It was demonstrated that for the properties tested (Tm, Tagg and HIC) that antibodies with good properties were found throughout the KPCA, not just clustered at the origin with the clinical mAb dataset and that the output antibody had a poorly validated HIC-RT. It remains to be seen why this was the case, as germline biases were checked and not observed. However, it was seen that the predictor trained on HIC-RT data could have flagged this as showing poor developability, as the antibody selected from the pipeline with the highest stringency also had the highest predicted and experimental HIC retention time, so this could be used as a reliable additional triage at this stage. Furthermore, what is reassuring is that this particular antibody was not output when the probability threshold of the model predicting clinical trials outcome was set at 0.7, and so this has demonstrated that perhaps the threshold of this model should be set at this stringency for future runs. Antibodies output by the pipeline above these thresholds had lower predicted HIC retention time values and probably would have performed better in experimental validation. Had more funding been available, more antibodies output by the pipeline could have been tested to show its success, but it was felt more important to use this opportunity to test multiple hypotheses throughout the thesis.

Taken together, it is recommended that the pipeline should be used with strin-

gent settings to ensure that all of the antibodies output are good quality antibodies which would have a lower chance of failing during clinical trials.

8.3 Comparing the Pipeline to Other Available Software

One advantage of the pipeline is that by using LLMs, antibodies can be encoded in 0.05 seconds using a graphics processing unit, which means that a library of 10,000 sequences can be encoded in around 8 minutes. The bottleneck of the pipeline is the AbNum encoding and the KPCA. To make these more efficient, a version of AbNum was installed locally on the server where the pipeline runs, rather than using the AbNum API. This reduced the times for encoding all of the sequences from 8 hours to 20 minutes. Despite this, no faster method to run the KPCA, using the GPU was found. Despite this, this pipeline is still faster than running the TAP score which, when using the IGX platform, took around 3 hours for each batch of 500 antibodies. Furthermore, it was found that TAP itself is not a predictor of clinical success, since clincally approved mAbs were seen with score of as low as -20, so it is only suitable to triage out very poorly developable antibodies with highly negative scores. The developability prediction (TA-DA score) given by Negron *et al.* [111], while demonstrating a combination of descriptors that can clearly separate clinical mAbs from library antibodies (high TA-DA score near to 1, denotes clinical antibodies) has not been made available for use despite its publication in 2022. Additionally, like the TAP score, this requires homology modelling the given sequences, to calculate the descriptors, which is an intense process when a library

of thousands of sequences are to be evaluated. Despite this, because their descriptors have been assigned by themselves, it is easier to identify the properties that are most important to identify clinical mAbs from a repertoire. However, the TA-DA score did not correlate well with any of the individual physicochemical properties measured by Jain *et al.* [102]. This is potentially because antibody developability is not dependent on one property but on a set of properties that are in balance with each other, where this relationship can be difficult to account for.

Furthermore, the AbPred development prediction software [107] requires more time to encode the given antibody sequences, and it was shown in their paper that their correlation is only moderate between their predictions and the experimental data given in Jain *et al.* [102]. For this reason, it is difficult to use as a predictor of developability and probably should be reserved for when a manageable number of candidates remain before expensive *in vitro* studies commence.

The pipeline presented in this thesis tries to address the shortcomings of other software by using an end-to-end triaging pipeline where an entire library can be input. The architecture of the pipeline ensures that poor antibodies can be removed at each stage, making each successive step in the pipeline more efficient. This is especially clear when using the sequence-derived physicochemical properties before the bottlenecks of the pipeline which are numbering, encoding, and the KPCA steps. It is expected that at each of these steps, some sequences will be lost as they cannot be numbered or encoded, but this is not seen as an issue because if this is the case, it seems reasonable to assume that they were likely to be very unusual antibodies having poor predicted developability anyway.

8.4 Future Work

8.4.1 The Role of *Fc* Domains in Developability

Throughout this thesis, the role of the constant regions of the antibodies has been sidelined, despite their known role in downstream effector functions that can affect immunogenicity and clearance, as well as thermostability and isoelectric point. The simple reason for this is that *Fc* regions lack sequence diversity beyond immunoglobulin classes and subtypes [209], while *Fv* domains are involved in binding, which is the area of research with the highest focus. However, *Fc* domains are often engineered to introduce silencing mutations, or to select a subtype with desirable downstream-mediated effects (usually IgG1 or IgG4 [73, 210]). While such differences could have effects, for the purposes of this study where the input would usually be paired V_H and V_L sequences, it was considered irrelevant. Having said this, it was interesting to see from Jain *et al.* [103] that antibodies with different *Fv* sequences, but grafted onto the same Ig subtype demonstrated differences in HIC-RT, which also lends to the idea of the ”developability web” where features not immediately associated lead to changes in a given property.

8.4.2 The Role of Deep Learning

Something to be noted throughout this thesis is that it has avoided using neural networks where possible. True, the LLMs are the result of deep learning training, but in cases where simple machine learning models have been sufficient to train classifiers, it seems overindulgent then to use deep learning models, which must be constructed to a suitable architecture with appropriate optimisers and loss functions

to do so. Using simple models, which perform well, is better than using complex models which may overfit and be non-generalisable. This is not to say that deep learning does not have its place in this field of research in terms of antibody modelling and predicting antibody-antigen interactions, but these are potentially much more difficult problems. Furthermore, antibody LLMs are known to overfit and when predicting missing residues, will often revert back to residues found at that position in germline sequences, or in cases with multiple missing residues predict strings of the same residue [211]. Even in the cases where neural networks were used, it was not any more successful than using simple machine learning classifiers, which was shown when training ADA predictors in Chapter 4 and training classifiers based on TAP scores in Chapter 6. Perhaps this would be an avenue for future exploration, but as stated previously, it is difficult to compare the results between studies that include different criteria to consider whether a patient has raised ADAs, different populations, or whether the expression of ADAs is relevant to the clinical treatment of interest, so efforts have been made to harmonise these findings to make them more comparable [212].

8.4.3 The Need for More Complete Datasets

The obvious criticism of this work is that LLMs are highly biased towards their training data sets (i.e. BCR repertoire data), and in the context of antibody LLMs, certain germline gene pairings would be preferred over others, and this would prevent the discovery of the therapeutic potential of novel pairings. However, if these clinical mAbs from available online repositories are shown to work, it is not bad

that similar antibodies have been selected as these are likely to be stable and less immunogenic.

While these biases in the clinical dataset may be perpetuated by using this pipeline, it is also important to appreciate that this bias exists only because it is a selected dataset. Having that more sequence data from antibody libraries, more data transparency on why some clinical mAbs have been discontinued, more experimental data on non-clinical antibodies were made available so could allow biases to be overcome. It is understood that pharmaceutical companies have been collating decades of proprietary data, as well as generating new experimental data to train their own predictive models [213], leaving academic research unable to access these valuable data and relying on what has been published online. A more collaborative environment where data can be pooled from different sources using federated learning, (which has been shown to use sensitive patient data for machine learning while maintaining anonymity [214]), could offer a faster solution to training the generalisable models required.

8.5 Conclusions

To conclude, this thesis has explored methods of therapeutic antibody developability prediction using antibody-trained LLMs to encode sequences and apply them to machine learning tasks. Using this, it has been possible to train predictors that can separate clinical mAbs, and library antibodies, as well as approved and discontinued mAbs. It has been investigated how these encodings may be used to predict experimental properties relevant to developability and to facilitate engineering of

MsAbs. These components have been assembled into an end-to-end bioinformatics pipeline which can be used to triage a library of antibody sequences to be left with a selection that are predicted to have properties similar to those of clinical mAbs. This pipeline hopes to accelerate the discovery of good quality candidates to decrease the risk of failure in trials.

Appendix A

Supplementary Tables

Table A.1: Files accessed from Observed Antibody Space.

DS Name	Sequences	Organism	Isotype	Chain	Disease	Individual
King_2020_2	3358	human	All	Paired	Tonsillitis/Obstructive-Sleep-Apnea	Subject-BCP6
King_2020_2	1075	human	All	Paired	Tonsillitis/Obstructive-Sleep-Apnea	Subject-BCP8
King_2020_2	3390	human	All	Paired	Tonsillitis	Subject-BCP9
King_2020_2	1120	human	All	Paired	Obstructive-Sleep-Apnea	Subject-BCP3
King_2020_2	425	human	All	Paired	Tonsillitis	Subject-BCP4
King_2020_2	2935	human	All	Paired	Tonsillitis	Subject-BCP5
King_2020_2	2812	human	All	Paired	Tonsillitis/Obstructive-Sleep-Apnea	Subject-BCP6
King_2020_2	2888	human	All	Paired	Tonsillitis/Obstructive-Sleep-Apnea	Subject-BCP8
King_2020_2	2978	human	All	Paired	Tonsillitis	Subject-BCP9
Mor_2021	4393	human	All	Paired	SARS-COV-2	Patient-10
Mor_2021	4025	human	All	Paired	SARS-COV-2	Patient-9
Mor_2021	2946	human	All	Paired	SARS-COV-2	Patient-8
Mor_2021	1584	human	All	Paired	SARS-COV-2	Patient-7
Mor_2021	1605	human	All	Paired	SARS-COV-2	Patient-6
Mor_2021	3574	human	All	Paired	SARS-COV-2	Patient-5
Mor_2021	2032	human	All	Paired	SARS-COV-2	Patient-4
Mor_2021	1812	human	All	Paired	SARS-COV-2	Patient-3
Mor_2021	3105	human	All	Paired	SARS-COV-2	Patient-16
Mor_2021	3849	human	All	Paired	SARS-COV-2	Patient-15
Mor_2021	4232	human	All	Paired	SARS-COV-2	Patient-14
Mor_2021	3482	human	All	Paired	SARS-COV-2	Patient-13
Mor_2021	3314	human	All	Paired	SARS-COV-2	Patient-12
Mor_2021	4162	human	All	Paired	SARS-COV-2	Patient-2
Mor_2021	3340	human	All	Paired	SARS-COV-2	Patient-1
Setliff_2019	4103	human	All	Paired	HIV	Donor-45
Setliff_2019	1444	human	All	Paired	HIV	Donor-N90
Woodruff_2020	1896	human	All	Paired	SARS-COV-2	Patient-1
Woodruff_2020	1534	human	All	Paired	SARS-COV-2	Patient-1
Eccles_2020	100	human	All	Paired	None	Healthy-1
Eccles_2020	47	human	All	Paired	None	Healthy-1
Eccles_2020	624	human	All	Paired	None	Healthy-1
King_2020_2	2207	human	All	Paired	Obstructive-Sleep-Apnea	Subject-BCP3
King_2020_2	2090	human	All	Paired	Tonsillitis	Subject-BCP4
King_2020_2	5793	human	All	Paired	Tonsillitis	Subject-BCP5

Appendix B

Data Files

Presented here are large data files which were used in the project presented in

FASTA format

$>[identifier]\_VH|[identifier]$

$[HeavyChain]$

$>[identifier]\_VL|[identifier]$

$[LightChain]$

As an example

```
>Adalimumab_VH|Adalimumab
EVQLVESGGGLVQPGRSLRLSCAASGFTFDDYAMHWVRQAPGKGLEWVSAITWN
SGHIDYADSVEGRFTISRDNAKNSLYLQMNSLRAEDTAVYYCAKVSYLSTASSL
DYWGQGTLVTVSS
>Adalimumab_VL|Adalimumab
DIQMTQSPSSLSASVGDRVTITCRASQGIRNYLAWYQQKPGKAPKLLIYAASTL
QSGVPSRFSGSGSGTDFTLTISSLQPEDVATYYCQRYNRAPYTFGQGTKVEIK
```

B.1 Data File URLs

Data File 1: OAS Dataset

https://mega.nz/file/lsxCFJba#LTOngpbRGzZNzV-3bUThkmRM
mZTi2HmFWjTmQkMm8GU

Data File 2: Approved Human mAbs from the TheraSabSab as of October 2021

https://mega.nz/file/0toQBKiK#NfcKFmESFADK_rKrlkSHeN0c
lXGY_sZhgH_x_2z1I8Q

Data File 3: Discontinued Human mAbs from the TheraSabSab as of October 2021

https://mega.nz/file/woYXjZLR#0bEQnawF2-__3XFezUWuAL-0
_D11UEBE1y1odQDQK1k

Data File 4: Clinical Trial Human mAbs from the TheraSabSab as of October 2021

https://mega.nz/file/1xZinaAY#OTkLFsg_MPnH5DEKxbgv9aFv
GqUQju46_8zDnEwGDuQ

Data File 5: Human-Derived mAbs named after 2022

https://mega.nz/file/F9YzyRTI#-Mzcm2KdDl6rdLbi3vjwa4Zd
1DNn08PZgBl_e3ATQMU

Data File 6: Accession Numbers for Unpaired V_H chains

https://mega.nz/file/YhwWSAjT#M8AkWSNHCkzP02vRHUUWYXCL

47mqh_4-yXYidGDFhkc

Data File 7: Accession Numbers for Unpaired V_L chains

https://mega.nz/file/1lhVGAiJ#9dHQDLZG0mPLP2KAlQY2_R-P5mEwHwQLhSUdQ2SNHt4

Data File 8: Approved Therapeutics mAb from the TheraSabDab as of October 2021

https://mega.nz/file/p8QyVD6b#g22YuonW5FoHyxi2Yui8kgwtSPdjxjIFNvG9CHqU220

Data File 9: Discontinued mAb Therapeutics from the TheraSabDab as of October 2021

https://mega.nz/file/00Y1yBSJ#ojgDNWTgTgk0oSuZ_OMoOns_KDbwKaACXKcY6IMTeQ8

Data File 10: Approved mAb Therapeutics from the TheraSabDab from October 2021

https://mega.nz/file/JoxEBapD#GhauimPl53qQfzOlT4NYhiQJIMMd5italzxtO7-zoLE

Data File 11: Discontinued mAb Therapeutics from the TheraSabDab from October 2021

https://mega.nz/file/Ahx0zB7a#1gHpR8sljdJkSL3rqtASTXobA0oWJUpZta4QrwW_ZfA

Data File 12: Pure2 Dataset (NT)

https://mega.nz/file/BopQgJKR#X4Nl-1GX84-hjmjiL0UdWCZihKIS63vt2nsaR6B20tQ

Data File 13: Pure2 Dataset (AA)

`https://mega.nz/file/l9IzmYpa#5sUkVpoKepdd_oNV3rn9GgzLZ-QpEPSTleF0LBNXpMU`

Appendix C

Experimental Procedures

Experimental procedures were carried out by GenScript as no in-house lab facilities were available at the time required. Details of protocols were kept confidential and so described here is an overview of the assays from details provided by GenScript.

C.1 Expression

Original amino acid sequences from Pure2 were provided to GenScript which were then conjugated to a Human IgG1 backbone (Table C.1) and expressed with a proprietary Chinese Hamster Ovary cell line (TurboCHO-HT 2.0) to a volume of 30ml. Proteins were purified using Protein A for antibody targets and quantified using SDS-PAGE electrophoresis.

C.2 Melting Temperature (Tm) Assay

Melting temperature was tested from for each antibody using differential scanning fluorometry by increasing the temperature of the sample 1°C/min from start temperature 20°C to end temperature 90°C. A fluorescent dye is added to the solution

Table C.1: Antibody heavy and light chain expression sequences as expressed by GenScript.

Label	Identifier	Heacy Chain Expression Sequence	Light Chain Expression Sequence
A	ACGAGGATCGCATGAT	MGWSCIILFLVATATGVHSEVQLVQSGGGLVQPGGSLKLSCVASGFTVSASAMHWVRQAPGKGLE WIGRIRKKVDDNVTQYAASVRGRFAISRDDSKNMAHLQMDSLKTEDTAVYFCARRVGGYYGLDV WGQGTTVTVSSASTKGPSVFPLAPSSKSTSGGKSCDKTHTCPPCTAALGCLVKDYFPEPVTVSWNS GALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEPPAPELLGGPSV FLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREEQYNSTYRVVSV LTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQVYTLPPSRDELTKNQVSLTCLVKG FYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTVDKSRWQQGNVFSCSVMHEALHNHY TQKSLSLSPGK*	MGWSCIILFLVATATGVHSDIQMTQSPSSLSASVGDRVTITCRAGEQISTYLNWYQQKPGKAPKLLI YVASTLQSGVPSRFSGSASGTHFTLSISNLQPDDSATYYCQQTYQTPYTFGQGTSLEIRRTVAAPSV FIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTL SKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC*
B	ACGAGCCCAAGCGCTC	MGWSCIILFLVATATGVHSQVQLVQSGAEVKKPGASVKVSCKASGYTFTNYYMHWVRQAPGQGL EWMGIINPSSGTTTYAQKFQGRVTMTRDTSTSTLYMEVSSLRSEDTAIYYCVRERGGSHNAFDIWG QGTRVTVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQ SSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEPKSCDKTHTCPPCPAPELLGGPSVFL FPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREEQYNSTYRVVSVLT VLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQVYTLPPSRDELTKNQVSLTCLVKGFY PSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTVDKSRWQQGNVFSCSVMHEALHNHYTQ KSLSLSPGK*	MGWSCIILFLVATATGVHSDIVLTQSPDSLAVSLGERATINCKSSQSVLFRSNNKDYLAWYQQKPGQ PPKLLIYWASTRESGVPDRFSGSGSGTDFTLTISSLQAEDVAVYYCQQYYSGPLYTFGQGTKLEIKRT VAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSL SSTLTLSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC*
C	CCTACCAGTATGAAAC	MGWSCIILFLVATATGVHSQITLKESGPTLVKPTQTLTLTCTFSGFSLTTNGVGVGWIRQPPGKALEW LALIYWNDDKRYSPSLNKRFTITKDTSKNQVVLTMTNMDPVDTATYYCAREVVTATTSFDYWGQ GIMVTVASASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQ SSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEPKSCDKTHTCPPCPAPELLGGPSVFL FPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREEQYNSTYRVVSVL TVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQVYTLPPSRDELTKNQVSLTCLVKG FYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTVDKSRWQQGNVFSCSVMHEALHNH YTQKSLSLSPGK*	MGWSCIILFLVATATGVHSDIQMTQSPSTLSASVGDRVIITCRASQTINRWLAWYQQKPGKAPKLLIFK ASSLEDGVPSRFSGSGSGTEFTLTISSLQPDDFATYYCQQYNSYWTFGQGTKVEIKRTVAAPSVFIFPPS DEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEK HKVYACEVTHQGLSSPVTKSFNRGEC*
D	CTAGTGAGTAGAGTGC	MGWSCIILFLVATATGVHSEVQLVQSGAEVKKPGESLKISCKGSGYSFTSYWIGWVRQMPGKGLE WMGIIYPGDSDTRYSPSFQGQVTISADKSISTAYLQWSSLKASDTAMYYCARHMEVRGVDKRR DYSQQYYYYGMDVWGQGTTVTVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVS WNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEPKSCDK THTCPPCPAPELLGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHN AKTKPREEQYNSTYRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQVYT LPPSRDELTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTVDKSR WQQGNVFSCSVMHEALHNHYTQKSLSLSPGK*	MGWSCIILFLVATATGVHSQSVLTQPPSASGTPGQRVTISCSGSSSNIGSNYVYWYQQLPGTAPKLLIYRN NQRPSGVPDRFSGSKSGTSASLAISGLRSEDEADYYCAAWDDSLSGVVFGGGTKLTVLGQPKAAPSVT LFPPSSEELQANKATLVCLISDFYPGAVTVAWKADSSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWK SHRSYSCQVTHEGSTVEKTVAPTECS*
E	TGCGCAGGTTTGTTTC	MGWSCIILFLVATATGVHSQVQLVQSGAEVKKPGSSVKVSCKASGGTFSTYAFYWVRQAPGQGL EWMGRIIPLLGMTIYAQNFQGRVTMTADKSTSTAYMELSSLRSGDTAVYYCARVDDYGDLRFDD WGQGTLVTVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFP AVLQSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEPKSCDKTHTCPPCPAPELLGG PSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREEQYNSTYR VVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQVYTLPPSRDELTKNQVSLT CLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTVDKSRWQQGNVFSCSVMHE ALHNHYTQKSLSLSPGK*	MGWSCIILFLVATATGVHSQSALTQPRSVSGSPGQSVTISCTGTSSDVGSYNYVSWYQQHPGKAPKLMIYD VNKRPSGVPARFSGSKSGNTASLTISGLQSEDEADYYCCSYVGSHTFLFGAGTKVTVLGQPKAAPSVTLFP PSSEELQANKATLVCLISDFYPGAVTVAWKADSSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHR SYSCQVTHEGSTVEKTVAPTECS*

containing the protein of interest. As the protein unfolds from the increase in temperature, the dye associates with the hydrophobic regions of the protein which are now free to interact. This interaction causes a conformational change in the dye that causes an enhancement in fluorescence. The assay was carried out using a Wyatt DynaPro Plate Reader III which can detect the fluorescence as it increases.

C.3 HIC-HPLC Assay

The principles of Hydrophobic Interaction Chromatography-High Performance Liquid Chromatography (HIC-HPLC) is to separate a sample based on their hydrophobicity, which leads to a longer elution time if a given molecule is more hydrophobic. This assay is usually performed in a column with a hydrophobic stationary phase and a concentrated salt mobile phase which binds to the water molecules at equilibrium with the stationary phase. The sample is added to the system which binds to the stationary phase. As the salt concentration in the mobile phase is decreased over time, the sample is eluted as the water molecules become more free to

interact. Molecules in the sample are detected as they are eluted and the time taken for elution is recorded. Results of HIC-HPLC assays can differ in experimental setups depending on the kind of salts used and its concentration as well as the temperature and pH. Details of this and apparatus for detection were kept confidential.

C.4 Aggregation Temperature (Tagg) Assay

Melting temperature was tested from for each antibody using dynamic light scattering (DLS) by increasing the temperature of the sample 1°C /min from start temperature 25°C to end temperature 85°C. DLS measures the size of the particles in the sample, which increase with temperature as proteins unfold and aggregate in solution. Proteins with higher Tagg are considered less likely to aggregate because higher temperature is required for them to do so. The assay was carried out using a Wyatt DynaPro Plate Reader III.

Bibliography

[1] Katharine Bray-French, Katharina Hartman, Guido Steiner, Céline Marban-Doran, Juliana Bessa, Neil Campbell, Meret Martin-Facklam, Kay-Gunnar Stubenrauch, Corinne Solier, Thomas Singer, and Axel Ducret. Managing the Impact of Immunogenicity in an Era of Immunotherapy: From Bench to Bedside. *Journal of Pharmaceutical Sciences*, 110(7):2575–2584, 2021.

[2] Jason R. Dunkelberger and Wen-Chao Song. Complement and Its Role in Innate and Adaptive Immune Responses. *Cell Research*, 20(1):34–50, 2010.

[3] Rahul Khetan, Robin Curtis, Charlotte M. Deane, Johannes Thorling Hadsund, Uddipan Kar, Konrad Krawczyk, Daisuke Kuroda, Sarah A. Robinson, Pietro Sormanni, Kouhei Tsumoto, Jim Warwicker, and Andrew C. R Martin. Current Advances in Biopharmaceutical Informatics: Guidelines, Impact and Challenges in the Computational Developability Assessment of Antibody Therapeutics. *mAbs*, 14(1):2020082, 2022.

[4] Charles A. Janeway Jr., Paul Travers, Mark Walport, and Mark J. Shlomchik. *Immunobiology: the Immune System in Health and Disease. the Structure of a Typical Antibody Molecule*. New York, 5th edition, 2001.

[5] Christian Vettermann and Mark S. Schlissel. Allelic Exclusion of Immunoglobulin Genes: Models and Mechanisms. *Immunological Reviews*, 237(1):22–42, 2010.

[6] Harry W. Schroeder Jr. and Lisa Cavacini. Structure and Function of Immunoglobulins. *Journal of Allergy and Clinical Immunology*, 125(2):S41–S52, 2010.

[7] Mathieu Dondelinger, Patrice Filée, Eric Sauvage, Birgit Quinting, Serge Muyldermans, Moreno Galleni, and Marylène S. Vandevenne. Understanding the Significance and Implications of Antibody Numbering and Antigen-Binding Surface/Residue Definition. *Frontiers in Immunology*, 9:2278, 2018.

[8] Tai Te Wu and Elvin A. Kabat. An Analysis of the Sequences of the Variable Regions of Bence Jones Proteins and Myeloma Light Chains and Their Implications for Antibody Complementarity. *The Journal of Experimental Medicine*, 132(2):211–250, 1970.

[9] Monica L. Fernández-Quintero, Johannes R. Loeffler, Johannes Kraml, Ursula Kahler, Anna S. Kamenik, and Klaus R. Liedl. Characterizing the Diversity of the CDR-H3 Loop Conformational Ensembles in Relationship to Antibody Binding Properties. *Frontiers in Immunology*, 9:3065, 2019.

[10] Tai Te Wu, George Johnson, and Elvin A. Kabat. Length Distribution of CDRH3 in Antibodies. *Proteins: Structure, Function, and Bioinformatics*, 16(1):1–7, 1993.

[11] Jeliazko R. Jeliazkov, Rahel Frick, Jing Zhou, and Jeffrey J. Gray. Robustification of RosettaAntibody and Rosetta SnugDock. *PLOS ONE*, 16(3):e0234282, 2021.

[12] Wing Ki Wong, Jinwoo Leem, and Charlotte M. Deane. Comparative Analysis of the CDR Loops of Antigen Receptors. *Frontiers in Immunology*, 10:2454, 2019.

[13] Mark L. Chiu, Dennis R. Goulet, Alexey Teplyakov, and Gary L. Gilliland. Antibody Structure and Function: The Basis for Engineering Therapeutics. *Antibodies*, 8(4):55, 2019.

[14] Klaus Rajewsky. Clonal Selection and Learning in the Antibody System. *Nature*, 381(6585):751–758, 1996.

[15] Janet Stavnezer and Carol E. Schrader. IgH Chain Class Switch Recombination: Mechanism and Regulation. *The Journal of Immunology*, 193(11):5370–5378, 2014.

[16] Cyrus Chothia and Arthur M. Lesk. Canonical Structures for the Hypervariable Regions of Immunoglobulins. *Journal of Molecular Biology*, 196(4):901–917, 1987.

[17] K. R. Abhinandan and Andrew C. R. Martin. Analysis and Improvements to Kabat and Structurally Correct Numbering of Antibody Variable Domains. *Molecular Immunology*, 45(14):3832–3839, 2008.

[18] Annemarie Honegger and Andreas Plückthun. Yet Another Numbering Scheme for Immunoglobulin Variable Domains: An Automatic Modeling and Analysis Tool. *Journal of Molecular Biology*, 309(3):657–670, 2001.

[19] Marie-Paule Lefranc. Unique Database Numberings System for Immunogenetic Analysis. *Immunology Today*, 18(11):509, 1997.

[20] Victor Greiff, Enkelejda Miho, Ulrike Menzel, and Sai T. Reddy. Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires. *Trends in Immunology*, 36(11):738–749, 2015.

[21] Michael Zemlin, Martin Klinger, Jason Link, Cosima Zemlin, Karl Bauer, Jeffrey A. Engler, Harry W. Schroeder, and Perry M. Kirkham. Expressed Murine and Human CDR-H3 Intervals of Equal Length Exhibit Distinct Repertoires That Differ in Their Amino Acid Composition and Predicted Range of Structures. *Journal of Molecular Biology*, 334(4):733–749, 2003.

[22] Bryan Briney, Anne Inderbitzin, Collin Joyce, and Dennis R. Burton. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*, 566(7744):393–397, 2019.

[23] Changshou Gao, Shenlan Mao, Gunnar Kaufmann, Peter Wirsching, Richard A. Lerner, and Kim D. Janda. A Method for the Generation of Combinatorial Antibody Libraries Using pIX Phage Display. *Proceedings of the National Academy of Sciences*, 99(20):12612–12616, 2002.

[24] Mohamed A. Alfaleh, Hashem O. Alsaab, Ahmad Bakur Mahmoud, Almo-

hanad A. Alkayyal, Martina L. Jones, Stephen M. Mahler, and Anwar M. Hashem. Phage Display Derived Monoclonal Antibodies: From Bench to Bedside. *Frontiers in Immunology*, 11:1986, 2020.

[25] Rongzhi Wang, Shuangshuang Xiang, Youjun Feng, Swaminath Srinivas, Yonghui Zhang, Mingshen Lin, and Shihua Wang. Engineering Production of Functional scFv Antibody in E. Coli by Co-Expressing the Molecule Chaperone SKP. *Frontiers in Cellular and Infection Microbiology*, 3:72, 2013.

[26] Juan C. Almagro, Martha Pedraza-Escalona, Hugo Iván Arrieta, and Sonia Mayra Pérez-Tapia. Phage Display Libraries for Antibody Therapeutic Discovery and Development. *Antibodies*, 8(3):44, 2019.

[27] Yang Zhang. Evolution of phage display libraries for therapeutic antibody discovery. *mAbs*, 15(1):2213793, 2023.

[28] Saravanan Rajan, Michael R. Kierny, Andrew Mercer, Jincheng Wu, Andrey Tovchigrechko, Herren Wu, William F. DallAcqua, Xiaodong Xiao, and Partha S. Chowdhury. Recombinant Human B Cell Repertoires Enable Screening for Rare, Specific, and Natively Paired Antibodies. *Communications Biology*, 1(1):5, 2018.

[29] Matthew C. Woodruff, Richard P. Ramonell, Doan C. Nguyen, Kevin S. Cashman, Ankur Singh Saini, Natalie S. Haddad, Ariel M. Ley, Shuya Kyu, J. Christina Howell, Tugba Ozturk, Saeyun Lee, Naveenchandra Suryadevara, James Brett Case, Regina Bugrovsky, Weirong Chen, Ja-

cob Estrada, Andrea Morrison-Porter, Andrew Derrico, Fabliha A. Anam, Monika Sharma, Henry M. Wu, Sang N. Le, Scott A. Jenks, Christopher M. Tipton, Bashar Staitieh, John L. Daiss, Eliver Ghosn, Michael S. Diamond, Robert H. Carnahan, James E. Crowe, William T. Hu, F. Eun-Hyung Lee, and Ignacio Sanz. Extrafollicular B Cell Responses Correlate With Neutralizing Antibodies and Morbidity in COVID-19. *Nature Immunology*, 21(12):1506–1516, 2020.

[30] Ian Setliff, Andrea R. Shiakolas, Kelsey A. Pilewski, Amyn A. Murji, Rutendo E. Mapengo, Katarzyna Janowska, Simone Richardson, Charissa Oosthuysen, Nagarajan Raju, Larance Ronsard, Masaru Kanekiyo, Juliana S. Qin, Kevin J. Kramer, Allison R. Greenplate, Wyatt J. McDonnell, Barney S. Graham, Mark Connors, Daniel Lingwood, Priyamvada Acharya, Lynn Morris, and Ivelin S. Georgiev. High-Throughput Mapping of B Cell Receptor Sequences to Antigen Specificity. *Cell*, 179(7):1636–1646.e15, 2019.

[31] Jacob D. Eccles, Ronald B. Turner, Nicole A. Kirk, Lyndsey M. Muehling, Larry Borish, John W. Steinke, Spencer C. Payne, Paul W. Wright, Deborah Thacker, Sampo J. Lahtinen, Markus J. Lehtinen, Peter W. Heymann, and Judith A. Woodfolk. T-Bet+ Memory B Cells Link to Local Cross-Reactive IgG Upon Human Rhinovirus Infection. *Cell Reports*, 30(2):351–366.e7, 2020.

[32] David B. Jaffe, Payam Shahi, Bruce A. Adams, Ashley M. Chrisman, Peter M. Finnegan, Nandhini Raman, Ariel E. Royall, FuNien Tsai, Thomas

Vollbrecht, Daniel S. Reyes, N. Lance Hepler, and Wyatt J. McDonnell. Functional Antibodies Exhibit Light Chain Coherence. *Nature*, 611(7935):352–357, 2022.

[33] Tobias H. Olsen, Fergus Boyles, and Charlotte M. Deane. Observed Antibody Space: a Diverse Database of Cleaned, Annotated, and Translated Unpaired and Paired Antibody Sequences. *Protein Science*, 31(1):141–146, 2022.

[34] Leonard D. Goldstein, Ying-Jiun J. Chen, Jia Wu, Subhra Chaudhuri, Yi-Chun Hsiao, Kellen Schneider, Kam Hon Hoi, Zhonghua Lin, Steve Guerrero, Bijay S. Jaiswal, Jeremy Stinson, Aju Antony, Kanika Bajaj Pahuja, Dhaya Seshasayee, Zora Modrusan, Isidro Hötzel, and Somasekar Seshagiri. Massively Parallel Single-Cell B-Cell Receptor Sequencing Enables Rapid Discovery of Diverse Antigen-Reactive Antibodies. *Communications Biology*, 2(1):304, 2019.

[35] Sergio E. Irac, Megan Sioe Fei Soon, Nicholas Borcherding, and Zewen Kelvin Tuong. Single-Cell Immune Repertoire Analysis. *Nature Methods*, 21(5):777–792, 2024.

[36] William H. Robinson. Sequencing the Functional Antibody Repertoire—diagnostic and Therapeutic Discovery. *Nature Reviews Rheumatology*, 11(3):171–182, 2015.

[37] Narayan Jayaram, Pallab Bhowmick, and Andrew C. R Martin. Germline

VH/VL Pairing in Antibodies. *Protein Engineering, Design and Selection*, 25(10):523–530, 2012.

[38] Roger Dodd, Darren J. Schofield, Trevor Wilkinson, and Zachary T. Britton. Generating Therapeutic Monoclonal Antibodies to Complex Multi-Spanning Membrane Targets: Overcoming the Antigen Challenge and Enabling Discovery Strategies. *Methods*, 180:111–126, 2020.

[39] Jorge Dias. Antibodies from Resilient Individuals: Identifying a Potential Novel Treatment for Huntington's Disease Modification. Antibody Engineering & Therapeutics, Amsterdam, Netherlands, June 2023.

[40] Peter Valent, Bernd Groner, Udo Schumacher, Giulio Superti-Furga, Meinrad Busslinger, Robert Kralovics, Christoph Zielinski, Josef M. Penninger, Dontscho Kerjaschki, Georg Stingl, Josef S. Smolen, Rudolf Valenta, Hans Lassmann, Heinrich Kovar, Ulrich Jäger, Gabriela Kornek, Markus Müller, and Fritz Sörgel. Paul Ehrlich (1854-1915) and His Contributions to the Foundation and Birth of Translational Medicine. *Journal of Innate Immunity*, 8(2):111–120, 2016.

[41] Georges. Köhler and César. Milstein. Continuous Cultures of Fused Cells Secreting Antibody of Predefined Specificity. *Nature*, 256(5517):495–497, 1975.

[42] Frank W. F Lee, Cynthia B. Elias, Paul Todd, and Dhinakar S. Kompala. Engineering Chinese Hamster Ovary (CHO) Cells to Achieve an Inverse Growth

– Associated Production of a Foreign Protein, -Galactosidase. *Cytotechnology*, 28(1):73–80, 1998.

[43] Peter A. Todd and Rex N. Brogden. Muromonab CD3. *Drugs*, 37(6):871–899, 1989.

[44] Vibha Jawa, Frances Terry, Jochem Gokemeijer, Shibani Mitra-Kaushik, Brian J. Roberts, Sophie Tourdot, and Anne S. De Groot. T-Cell Dependent Immunogenicity of Protein Therapeutics Pre-Clinical Assessment and Mitigation–Updated Consensus and Review 2020. *Frontiers in Immunology*, 11:1301, 2020.

[45] Amy Sun and Leslie Z. Benet. Late-Stage Failures of Monoclonal Antibody Drugs: a Retrospective Case Study Analysis. *Pharmacology*, 105(3-4):145–163, 2020.

[46] Sherrie L. Morrison, M. Jacqueline Johnson, Leonard A. Herzenberg, and Vernon T. Oi. Chimeric Human Antibody Molecules: Mouse Antigen-Binding Domains With Human Constant Region Domains. *Proceedings of the National Academy of Sciences*, 81(21):6851–6855, 1984.

[47] Man Sung Co, Marguerite Deschamps, Richard J. Whitley, and Cary Queen. Humanized Antibodies for Antiviral Therapy. *Proceedings of the National Academy of Sciences*, 88(7):2869–2873, 1991.

[48] Michael S. Neuberger, Gareth T. Williams, E. Bruce Mitchell, S. S. Jouhal, John. G. Flanagan, and Terrance H. Rabbitts. A Hapten-Specific Chi-

maeric IgE Antibody With Human Physiological Effector Function. *Nature*, 314(6008):268–270, 1985.

[49] Peter T. Jones, Paul H. Dear, Jefferson Foote, Michael S. Neuberger, and Greg Winter. Replacing the Complementarity-Determining Regions in a Human Antibody With Those From a Mouse. *Nature*, 321(6069):522–525, 1986.

[50] Fiona A. Harding, Marcia M. Stickler, Jennifer Razo, and Robert B. DuBridge. The Immunogenicity of Humanized and Fully Human Antibodies: Residual Immunogenicity Resides in the CDR Regions. *mAbs*, 2(3):256–265, 2010.

[51] Laurent S. Jespers, Andy Roberts, Stephen M. Mahler, Greg Winter, and Hennie R. Hoogenboom. Guiding the Selection of Human Antibodies From Phage Display Repertoires to a Single Epitope of an Antigen. *Bio/Technology*, 12(9):899–903, 1994.

[52] Karly P. Garnock-Jones. Necitumumab: First Global Approval. *Drugs*, 76(2):283–289, 2016.

[53] Esther S. Kim. Avelumab: First Global Approval. *Drugs*, 77(8):929–937, 2017.

[54] Nils Lonberg, Lisa D. Taylor, Fiona A. Harding, Mary Trounstine, Kay M. Higgins, Stephen R. Schramm, Chiung-Chi Kuo, Roshanak Mashayekh, Kathryn Wymore, James G. McCabe, Donna Munoz-O'Regan, Susan L.

O'Donnell, Elizabeth S. G Lapachet, Tasha Bengoechea, Dianne M. Fishwild, Condie E. Carmack, Robert M. Kay, and Dennis Huszar. Antigen-Specific Human Antibodies From Mice Comprising Four Distinct Genetic Modifications. *Nature*, 368(6474):856–859, 1994.

[55] Arun K. Kashyap, John Steel, Ahmet F. Oner, Michael A. Dillon, Ryann E. Swale, Katherine M. Wall, Kimberly J. Perry, Aleksandr Faynboym, Mahmut Ilhan, Michael Horowitz, Lawrence Horowitz, Peter Palese, Ramesh R. Bhatt, and Richard A. Lerner. Combinatorial Antibody Libraries From Survivors of the Turkish H5N1 Avian Influenza Outbreak Reveal Virus Neutralization Strategies. *Proceedings of the National Academy of Sciences*, 105(16):5986–5991, 2008.

[56] Marianne Brüggemann, Michael J. Osborn, Biao Ma, Jasvinder Hayre, Suzanne Avis, Brian Lundstrom, and Roland Buelow. Human Antibody Production in Transgenic Animals. *Archivum Immunologiae et Therapiae Experimentalis*, 63(2):101–108, 2015.

[57] Michael A. Morse. Technology Evaluation: Ipilimumab, Medarex/Bristol-Myers Squibb. *Current Opinion in Molecular Therapeutics*, 7(6):588–597, 2005.

[58] Kim A. Papp, Craig Leonardi, Alan Menter, Jean-Paul Ortonne, James G. Krueger, Gregory Kricorian, Girish Aras, Juan Li, Chris B. Russell, Elizabeth H. Z Thompson, and Scott Baumgartner. Brodalumab, an Anti–Interleukin-

17–Receptor Antibody for Psoriasis. *New England Journal of Medicine*, 366(13):1181–1189, 2012.

[59] Michael R. Migden, Danny Rischin, Chrysalyne D. Schmults, Alexander Guminski, Axel Hauschild, Karl D. Lewis, Christine H. Chung, Leonel Hernandez-Aya, Annette M. Lim, Anne Lynn S. Chang, Guilherme Rabinowits, Alesha A. Thai, Lara A. Dunn, Brett G. M Hughes, Nikhil I. Khushalani, Badri Modi, Dirk Schadendorf, Bo Gao, Frank Seebach, Siyu Li, Jingjin Li, Melissa Mathias, Jocelyn Booth, Kosalai Mohan, Elizabeth Stankevich, Hani M. Babiker, Irene Brana, Marta Gil-Martin, Jade Homsi, Melissa L. Johnson, Victor Moreno, Jiaxin Niu, Taofeek K. Owonikoko, Kyriakos P. Papadopoulos, George D. Yancopoulos, Israel Lowy, and Matthew G. Fury. PD-1 Blockade With Cemiplimab in Advanced Cutaneous Squamous-Cell Carcinoma. *New England Journal of Medicine*, 379(4):341–351, 2018.

[60] Susan J. Keam. Tixagevimab + Cilgavimab: First Approval. *Drugs*, 82(9):1001–1010, 2022.

[61] Nigel M. Low, Philipp Holliger, and Greg Winter. Mimicking Somatic Hypermutation: Affinity Maturation of Antibodies Displayed on Bacteriophage Using a Bacterial Mutator Strain. *Journal of Molecular Biology*, 260(3):359–368, 1996.

[62] James D. Marks. Antibody Affinity Maturation by Chain Shuffling. *Methods in Molecular Biology*, pages 327–343, 2004.

[63] Mitchell Ho, Robert J. Kreitman, Masanori Onda, and Ira Pastan. *in vitro* Antibody Evolution Targeting Germline Hot Spots to Increase Activity of an Anti-Cd22 Immunotoxin. *Journal of Biological Chemistry*, 280(1):607–617, 2005.

[64] Ruei-Min Lu, Yu-Chyi Hwang, I-Ju Liu, Chi-Chiu Lee, Han-Zen Tsai, Hsin-Jung Li, and Han-Chung Wu. Development of Therapeutic Antibodies for the Treatment of Diseases. *Journal of Biomedical Science*, 27(1):1, 2020.

[65] Hélène Kaplon and Janice M. Reichert. Antibodies to Watch in 2021. *mAbs*, 13(1):1860476, 2021.

[66] Hélène Kaplon, Alicia Chenoweth, Silvia Crescioli, and Janice M. Reichert. Antibodies to Watch in 2022. *mAbs*, 14(1):2014296, 2022.

[67] Sofia S. Guimaraes Koch, Robin Thorpe, Nana Kawasaki, Marie-Paule Lefranc, Sarel Malan, Andrew C. R. Martin, Gilles Mignot, Andreas Plückthun, Menico Rizzi, Stephanie Shubat, Karin Weisser, and Raffaella Balocco. International Nonproprietary Names for Monoclonal Antibodies: An Evolving Nomenclature System. *mAbs*, 14(1):2075078, 2022.

[68] Silvia Crescioli, Hélène Kaplon, Alicia Chenoweth, Lin Wang, Jyothsna Visweswaraiah, and Janice M. Reichert. Antibodies to Watch in 2024. *mAbs*, 16(1):2297450, 2024.

[69] Aran F. Labrijn, Maarten L. Janmaat, Janice M. Reichert, and Paul W. H I.

Parren. Bispecific Antibodies: a Mechanistic Review of the Pipeline. *Nature Reviews Drug Discovery*, 18(8):585–608, 2019.

[70] Christophe Schmitt, Joanne I. Adamkewicz, Jin Xu, Claire Petry, Olivier Catalani, Guy Young, Claude Negrier, Michael U. Callaghan, and Gallia G. Levy. Pharmacokinetics and Pharmacodynamics of Emicizumab in Persons With Hemophilia a With Factor VIII Inhibitors: HAVEN 1 Study. *Thrombosis and Haemostasis*, 121(3):351–360, 2021.

[71] Yu Zhou, Lequn Zhao, and James D. Marks. Selection and Characterization of Cell Binding and Internalizing Phage Antibodies. *Archives of Biochemistry and Biophysics*, 526(2):107–113, 2012.

[72] Eunhee G. Kim, Jieun Jeong, Junghyeon Lee, Hyeryeon Jung, Minho Kim, Yi Zhao, Eugene C. Yi, and Kristine M. Kim. Rapid Evaluation of Antibody Fragment Endocytosis for Antibody Fragment–Drug Conjugates. *Biomolecules*, 10(6):955, 2020.

[73] Ian Wilkinson, Stephen Anderson, Jeremy Fry, Louis Alex Julien, David Neville, Omar Qureshi, Gary Watts, and Geoff Hale. Fc-Engineered Antibodies With Immune Effector Functions Completely Abolished. *PLOS ONE*, 16(12):e0260954, 2021.

[74] Alastair Douglas Davy Koen Sandra Geoff Hale, Jelle De Vos and Ian Wilkinson. Systematic analysis of fc mutations designed to reduce binding to fc-gamma receptors. *mAbs*, 16(1):2402701, 2024.

[75] Richard W. Shuai, Jeffrey A. Ruffolo, and Jeffrey J. Gray. IgLM: Infilling language modeling for antibody sequence design. *Cell Systems*, 14(11):979–989, 2023.

[76] Ulrich Brinkmann and Roland E. Kontermann. The Making of Bispecific Antibodies. *mAbs*, 9(2):182–212, 2017.

[77] César. Milstein and A. Claudio Cuello. Hybrid Hybridomas and Their Use in Immunohistochemistry. *Nature*, 305(5934):537–540, 1983.

[78] Roland E. Kontermann and Ulrich Brinkmann. Bispecific Antibodies. *Drug Discovery Today*, 20(7):838–847, 2015.

[79] Christoph Spiess, Qianting Zhai, and Paul J. Carter. Alternative Molecular Formats and Therapeutic Applications for Bispecific Antibodies. *Molecular Immunology*, 67(2, Part A):95–106, 2015.

[80] Fabrice Le Gall, Sergey M. Kipriyanov, Gerhard Moldenhauer, and Melvyn Little. Di-, Tri- and Tetrameric Single Chain Fv Antibody Fragments Against Human CD19: Effect of Valency on Cell Binding. *Federation of European Biochemical Societies Letters*, 453(1-2):164–168, 1999.

[81] Caroline Rozan, Amélie Cornillon, Corinne Pétiard, Martine Chartier, Ghislaine Behar, Charlotte Boix, Brigitte Kerfelec, Bruno Robert, André Pèlegrin, Patrick Chames, Jean-Luc Teillaud, and Daniel Baty. Single-Domain Antibody–Based and Linker-Free Bispecific Antibodies Targeting FcRIII Induce

Potent Antitumor Activity Without Recruiting Regulatory T. Cells. *Molecular Cancer Therapeutics*, 12(8):1481–1491, 2013.

[82] John B. B Ridgway, Leonard G. Presta, and Paul Carter. 'Knobs-Into-Holes' Engineering of Antibody CH3 Domains for Heavy Chain Heterodimerization. *Protein Engineering, Design and Selection*, 9(7):617–621, 1996.

[83] Kannan Gunasekaran, Martin Pentony, Min Shen, Logan Garrett, Carla Forte, Anne Woodward, Soo Bin Ng, Teresa Born, Marc Retter, Kathy Manchulenko, Heather Sweet, Ian N. Foltz, Michael Wittekind, and Wei Yan. Enhancing Antibody Fc Heterodimer Formation Through Electrostatic Steering Effects: Applications to Bispecific Molecules and Monovalent IgG. *Journal of Biological Chemistry*, 285(25):19637–19646, 2010.

[84] Klaus Brischwein, Larissa Parr, Stefan Pflanz, Jörg Volkland, John Lumsden, Matthias Klinger, Mathias Locher, Scott A. Hammond, Peter Kiener, Peter Kufer, Bernd Schlereth, and Patrick A. Baeuerle. Strictly Target Cell-Dependent Activation of T. Cells by Bispecific Single-Chain Antibody Constructs of the BiTE Class. *Journal of Immunotherapy*, 30(8):798–807, 2007.

[85] Roland E. Kontermann. Strategies for Extended Serum Half-Life of Protein Therapeutics. *Current Opinion in Biotechnology*, 22(6):868–876, 2011.

[86] Jiabing Ma, Yicheng Mo, Menglin Tang, Junjie Shen, Yanan Qi, Wenxu Zhao, Yi Huang, Yanmin Xu, and Cheng Qian. Bispecific Antibodies: From

Research to Clinical Application. *Frontiers in Immunology*, 12:626616, 2021.

[87] Samaresh Sau, Hashem O. Alsaab, Sushil Kumar Kashaw, Katyayani Tatiparti, and Arun K. Iyer. Advances in Antibody-Drug Conjugates: A New Era of Targeted Cancer Therapy. *Drug discovery today*, 22(10):1547–1556, 2017.

[88] Peter Szijj and Vijay Chudasama. The Renaissance of Chemically Generated Bispecific Antibodies. *Nature Reviews Chemistry*, 5(2):78–92, 2021.

[89] Roberta Lucchi, Jordi Bentanachs, and Benjamí Oller-Salvia. The Masking Game: Design of Activatable Antibodies and Mimetics for Selective Therapeutics and Cell Control. *ACS Central Science*, 7(5):724–738, 2021.

[90] F. Donelson Smith, Robert H. Pierce, Thomas Thisted, and Edward H. van der Horst. Conditionally Active, pH-Sensitive Immunoregulatory Antibodies Targeting VISTA and CTLA-4 Lead an Emerging Class of Cancer Therapeutics. *Antibodies*, 12(3):55, 2023.

[91] Ian B. Robertson, Rachel Mulvaney, Nele Dieckmann, Alessio Vantellini, Martina Canestraro, Francesca Amicarella, Ronan O'Dwyer, David K Cole, Stephen Harper, Omer Dushek, and Peter Kirk. Tuning the potency and selectivity of ImmTAC molecules by affinity modulation. *Clinical and Experimental Immunology*, 215(2):105–119, 2023.

[92] David Weininger. SMILES, a Chemical Language and Information System.

1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

[93] Tianhong Zhang, Hongli Li, Hualin Xi, Robert V. Stanton, and Sergio H. Rotstein. HELM: a Hierarchical Notation Language for Complex Biomolecule Structure Representation. *Journal of Chemical Information and Modeling*, 52(10):2796–2806, 2012.

[94] Adriana-Michelle Wolf Pérez, Nikolai Lorenzen, Michele Vendruscolo, and Pietro Sormanni. Assessment of Therapeutic Antibody Therapeutic Antibodies Developability by Combinations of *in vitro* and *in silico* Methods. *Therapeutic Antibodies: Methods and Protocols*, pages 57–113, 2022.

[95] Ehab M. Moussa, Jainik P. Panchal, Balakrishnan S. Moorthy, Janice S. Blum, Marisa K. Joubert, Linda O. Narhi, and Elizabeth M. Topp. Immunogenicity of Therapeutic Protein Aggregates. *Journal of Pharmaceutical Sciences*, 105(2):417–430, 2016.

[96] Ziyan Wang, Jianwei Zhu, and Huili Lu. Antibody Glycosylation: Impact on Antibody Drug Characteristics and Quality Control. *Applied Microbiology and Biotechnology*, 104(5):1905–1914, 2020.

[97] Anshu Kuriakose, Narendra Chirmule, and Pradip Nair. Immunogenicity of Biotherapeutics: Causes and Association With Posttranslational Modifications. *Journal of Immunology Research*, 2016:1298473, 2016.

[98] Miranda M. C van Beers and Muriel Bardor. Minimizing Immunogenicity

of Biopharmaceuticals by Controlling Critical Quality Attributes of Proteins. *Biotechnology Journal*, 7(12):1473–1484, 2012.

[99] Shuji Noguchi. Structural Changes Induced by the Deamidation and Isomerization of Asparagine Revealed by the Crystal Structure of Ustilago Sphaerogena Ribonuclease U2B. *Biopolymers*, 93(11):1003–1010, 2010.

[100] Riccardo Torosantucci, Victor S. Sharov, Miranda van Beers, Vera Brinks, Christian Schöneich, and Wim Jiskoot. Identification of Oxidation Sites and Covalent Cross-Links in Metal Catalyzed Oxidized Interferon Beta-1a: Potential Implications for Protein Aggregation and Immunogenicity. *Molecular Pharmaceutics*, 10(6):2311–2322, 2013.

[101] Marc Bailly, Carl Mieczkowski, Veronica Juan, Essam Metwally, Daniela Tomazela, Jeanne Baker, Makiko Uchida, Ester Kofman, Fahimeh Raoufi, Soha Motlagh, Yao Yu, Jihea Park, Smita Raghava, John Welsh, Michael Rauscher, Gopalan Raghunathan, Mark Hsieh, Yi-Ling Chen, Hang Thu Nguyen, Nhung Nguyen, Dan Cipriano, and Laurence Fayadat-Dilman. Predicting Antibody Developability Profiles Through Early Stage Discovery Screening. *mAbs*, 12(1):1743053, 2020.

[102] Tushar Jain, Tingwan Sun, Stéphanie Durand, Amy Hall, Nga Rewa Houston, Juergen H. Nett, Beth Sharkey, Beata Bobrowicz, Isabelle Caffry, Yao Yu, Yuan Cao, Heather Lynaugh, Michael Brown, Hemanta Baruah, Laura T. Gray, Eric M. Krauland, Yingda Xu, Maximiliano Vásquez, and K. Dane

Wittrup. Biophysical Properties of the Clinical-Stage Antibody Landscape. *Proceedings of the National Academy of Sciences*, 114(5):944–949, 2017.

[103] Tushar Jain, Todd Boland, and Maximiliano Vásquez. Identifying Developability Risks for Clinical Progression of Antibodies Using High-Throughput *in vitro* and *in silico* Approaches. *mAbs*, 15(1):2200540, 2023.

[104] Claire Marks, Alissa M. Hummer, Mark Chin, and Charlotte M. Deane. Humanization of Antibodies Using a Machine Learning Approach on Large-Scale Repertoire Data. *Bioinformatics*, 37(22):4041–4047, 2021.

[105] Timothy M. Lauer, Neeraj J. Agrawal, Naresh Chennamsetty, Kamal Egodage, Bernhard Helk, and Bernhardt L. Trout. Developability Index: a Rapid in Silico Tool for the Screening of Antibody Aggregation Propensity. *Journal of Pharmaceutical Sciences*, 101(1):102–115, 2012.

[106] Daniel Seeliger. Development of Scoring Functions for Antibody Sequence Assessment and Optimization. *PLOS ONE*, 8(10):e76909, 2013.

[107] Max Hebditch and Jim Warwicker. Charge and Hydrophobicity Are Key Features in Sequence-Trained Machine Learning Models for Predicting the Biophysical Properties of Clinical-Stage Antibodies. *PeerJ*, 7(1):e8199, 2019.

[108] K. R. Abhinandan and Andrew C. R. Martin. Analyzing the "Degree of Humanness" of Antibody Sequences. *Journal of Molecular Biology*, 369(3):852–862, 2007.

[109] Philippe Thullier, Oliver Huish, Thibaut Pelat, and Andrew C. R Martin. The Humanness of Macaque Antibody Sequences. *Journal of Molecular Biology*, 396(5):1439–1450, 2010.

[110] Matthew I. J. Raybould, Claire Marks, Konrad Krawczyk, Bruck Taddese, Jaroslaw Nowak, Alan P. Lewis, Alexander Bujotzek, Jiye Shi, and Charlotte M. Deane. Five Computational Developability Guidelines for Therapeutic Antibody Profiling. *Proceedings of the National Academy of Sciences*, 116(10):4025–4030, 2019.

[111] Christopher Negron, Joyce Fang, Michael J. McPherson, W. Blaine Stine, and Andrew J. McCluskey. Separating Clinical Antibodies From Repertoire Antibodies, a Path to *in silico* Developability Assessment. *mAbs*, 14(1):2080628, 2022.

[112] Mark B. Swindells, Craig T. Porter, Matthew Couch, Jacob Hurst, K. R. Abhinandan, Jens H. Nielsen, Gary Macindoe, James Hetherington, and Andrew C. R. Martin. abYsis: Integrated Antibody Sequence and Structure-Management, Analysis, and Prediction. *Journal of Molecular Biology*, 429(3):356–364, 2017.

[113] Marie-Paule Lefranc. IMGT, the International ImMunoGeneTics Database. *Nucleic Acids Research*, 31(1):307–310, 2003.

[114] Andrew C. R Martin. Accessing the Kabat Antibody Sequence Database by

Computer. *Proteins: Structure, Function, and Bioinformatics*, 25(1):130–133, 1996.

[115] H. H.-L. Shih, John Brady, and Martin Karplus. Structure of Proteins with Single-Site Mutations: a Minimum Perturbation Approach. *Proceedings of the National Academy of Sciences*, 82(6):1697–1700, 1985.

[116] Mark Abraham, Andrey Alekseenko, Vladimir Basov, Cathrine Bergh, Eliane Briand, Ania Brown, Mahesh Doijade, Giacomo Fiorin, Stefan Fleischmann, Sergey Gorelov, Gilles Gouaillardet, Alan Grey, M. Eric Irrgang, Farzaneh Jalalypour, Joe Jordan, Carsten Kutzner, Justin A. Lemkul, Magnus Lundborg, Pascal Merz, Vedran Miletic, Dmitry Morozov, Julien Nabet, Szilard Pall, Andrea Pasquadibisceglie, Michele Pellegrino, Hubert Santuz, Roland Schulz, Tatiana Shugaeva, Alexey Shvetsov, Alessandra Villa, Sebastian Wingbermuehle, Berk Hess, and Erik Lindahl. Gromacs 2024.3 manual, 2024.

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

[118] Aleksandr Kovaltsuk, Jinwoo Leem, Sebastian Kelm, James Snowden, Charlotte M. Deane, and Konrad Krawczyk. Observed Antibody Space: a Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. *The Journal of Immunology*, 201(8):2502–2509, 2018.

[119] Alejandro Clavero-Álvarez, Tomas Di Mambro, Sergio Perez-Gaviro, Mauro Magnani, and Pierpaolo Bruscolini. Humanization of Antibodies Using a Statistical Inference Approach. *Scientific Reports*, 8(1):14820, 2018.

[120] David S. Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36(1):D901–D906, 2008.

[121] Alexander Stewart, Emma Sinclair, Joseph Chi-Fung Ng, Joselli Silva O'Hare, Audrey Page, Ilaria Serangeli, Christian Margreitter, Federica Orsenigo, Katherine Longman, Cecile Frampas, Catia Costa, Holly-May Lewis, Nora Kasar, Bryan Wu, David Kipling, Peter J. M Openshaw, Christopher Chiu, J. Kenneth Baillie, Janet T. Scott, Malcolm G. Semple, Melanie J. Bailey, Franca Fraternali, and Deborah K. Dunn-Walters. Pandemic, Epidemic, Endemic: B Cell Repertoire Analysis Reveals Unique Anti-Viral Responses to SARS-CoV-2, Ebola and Respiratory Syncytial Virus. *Frontiers in Immunology*, 13, 2022.

[122] Michael P. Fay and Michael A. Proschan. Wilcoxon-Mann-Whitney or T-Test? On Assumptions for Hypothesis Tests and Multiple Interpretations of Decision Rules. *Statistics Surveys*, 4:1–39, 2010.

[123] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A. Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[124] George Johnson and Tai Te Wu. Preferred CDRH3 Lengths for Antibodies With Defined Specificities. *International Immunology*, 10(12):1801–1805, 1998.

[125] Abigail V. J. Collis, Adam P. Brouwer, and Andrew C. R. Martin. Analysis of the antigen combining site: Correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *Journal of Molecular Biology*, 325(2):337–354, 2003.

[126] Audrey D. McConnell, Xue Zhang, John L. Macomber, Betty Chau, Joseph C. Sheffer, Sorena Rahmanian, Eric Hare, Vladimir Spasojevic, Robert A. Horlick, David J. King, and Peter M. Bowers. A General Approach to Antibody Thermostabilization. *mAbs*, 6(5):1274–1282, 2014.

[127] Lei Jia, Ramya Yarlagadda, and Charles C. Reed. Structure Based Thermostability Prediction Models for Protein Single Point Mutations With Machine Learning Tools. *PLOS ONE*, 10(9):e0138022, 2015.

[128] Motohisa Oobatake and Tatsuo Ooi. Hydration and Heat Stability Effects on Protein Unfolding. *Progress in Biophysics and Molecular Biology*, 59(3):237–284, 1993.

[129] Lukasz P. Kozlowski. IPC – Isoelectric Point Calculator. *Biology Direct*, 11(1):55, 2016.

[130] Peter Rice, Ian Longden, and Alan Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(6):276–277, 2000.

[131] Shabdita Vatsa. *in silico* Prediction of Post-Translational Modifications in Therapeutic Antibodies. *mAbs*, 14(1):2023938, 2022.

[132] Xiaobin Xu, Yu Huang, Hao Pan, Rosalynn Molden, Haibo Qiu, Thomas J. Daly, and Ning Li. Quantitation and Modeling of Post-Translational Modifications in a Therapeutic Monoclonal Antibody From Single- and Multiple-Dose Monkey Pharmacokinetic Studies Using Mass Spectrometry. *PLOS ONE*, 14(10):e0223899–e0223899, 2019.

[133] Julie M. J Laffy, Tihomir Dodev, Jamie A. Macpherson, Catherine Townsend, Hui Chun Lu, Deborah Dunn-Walters, and Franca Fraternali. Promiscuous Antibodies Characterised by Their Physico-Chemical Properties: From Sequence to Structure and Back. *Progress in Biophysics and Molecular Biology*, 128:47–56, 2017.

[134] Gerson H. Cohen, Enid W. Silverton, Eduardo A. Padlan, Fred Dyda, Jamie A. Wibbenmeyer, Richard C. Willson, and David R. Davies. Water Molecules in the Antibody–Antigen Interface of the Structure of the Fab HyHEL-5–lysozyme Complex at 1.7Åresolution: Comparison With Results From Isothermal Titration Calorimetry. *Acta Crystallographica Section D*, 61(5):628–633, 2005.

[135] David Eisenberg, Robert M. Weiss, Thomas C. Terwilliger, and William Wilcox. Hydrophobic Moments and Protein Structure. *Faraday Symposia of the Chemical Society*, 17(0):109–120, 1982.

[136] B. Lee and F. M. Richards. The Interpretation of Protein Structures: Estimation of Static Accessibility. *Journal of Molecular Biology*, 55(3):379, 1971.

[137] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. BLAST+: Architecture and Applications. *BMC Bioinformatics*, 10(1):421, 2009.

[138] Marcus D. Ruopp, Neil J. Perkins, Brian W. Whitcomb, and Enrique F. Schisterman. Youden Index and Optimal Cut-Point Estimated From Observations Affected by a Lower Limit of Detection. *Biometrical Journal*, 50(3):419–430, 2008.

[139] J. T. L. Wang, Q. Ma, D. Shasha, and C. H. Wu. New Techniques for Extracting Features From Protein Sequences. *IBM Systems Journal*, 40(2):426–441, 2001.

[140] William White, Giblert and Steffans. Using a Neural Network to Backtranslate Amino Acid Sequences. *Electronic Journal of Biotechnology*, 1(3):17–18, 1998.

[141] Kuang Lin, Alex C. W. May, and William R. Taylor. Amino Acid Encoding Schemes From Protein Structure Alignments: Multi-Dimensional Vectors to Describe Residue Types. *Journal of Theoretical Biology*, 216(3):361–365, 2002.

[142] William R. Atchley, Jieping Zhao, Andrew D. Fernandes, and Tanja Drüke.

Solving the Protein Sequence Metric Problem. *Proceedings of the National Academy of Sciences*, 102(18):6395–6400, 2005.

[143] Jens Meiler, Michael Müller, Anita Zeidler, and Felix Schmäschke. Generation and Evaluation of Dimension-Reduced Amino Acid Parameter Representations by Artificial Neural Networks. *Molecular modeling annual*, 7(9):360–369, 2001.

[144] Akinori Kidera, Yasuo Konishi, Masahito Oka, Tatsuo Ooi, and Harold A. Scheraga. Statistical Analysis of the Physical Properties of the 20 Naturally Occurring Amino Acids. *Journal of Protein Chemistry*, 4(1):23–55, 1985.

[145] Ehsaneddin Asgari and Mohammad R. K. Mofrad. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLOS ONE*, 10(11):e0141287, 2015.

[146] Schwartz R.M. Orcutt B.C. Dayhoff, M.O. A Model of Evolutionary Change in Proteins. *Atlas of Protein Sequence and Structure*, 5:345–352, 1978.

[147] Jack Kyte and Russell F. Doolittle. A Simple Method for Displaying the Hydropathic Character of a Protein. *Journal of Molecular Biology*, 157(1):105–132, 1982.

[148] Patrice Koehl, Henri Orland, and Marc Delarue. Numerical Encodings of Amino Acids in Multivariate Gaussian Modeling of Protein Multiple Sequence Alignments. *Molecules*, 24(1), 2019.

[149] Sanzo Miyazawa and Robert L. Jernigan. Estimation of Effective Interresidue Contact Energies From Protein Crystal Structures: Quasi-Chemical Approximation. *Macromolecules*, 18(3):534–552, 1985.

[150] Cristian Micheletti, Flavio Seno, Jayanth R. Banavar, and Amos Maritan. Learning Effective Amino Acid Interactions Through Iterative Stochastic Techniques. *Proteins: Structure, Function, and Bioinformatics*, 42(3):422–431, 2001.

[151] Kevin K. Yang, Zachary Wu, Claire N. Bedbrook, and Frances H. Arnold. Learned Protein Embeddings for Machine Learning. *Bioinformatics*, 34(15):2642–2648, 2018.

[152] Ugo Bastolla, Markus Porto, H. Eduardo Roman, and Michele Vendruscolo. Principal Eigenvector of Contact Matrices and Hydrophobicity Profiles in Proteins. *Proteins: Structure, Function, and Bioinformatics*, 58(1):22–30, 2005.

[153] Sebastian Spänig and Dominik Heider. Encodings and Models for Antimicrobial Peptide Classification for Multi-Resistant Pathogens. *BioData Mining*, 12(1):7, 2019.

[154] Ana Marta Sequeira, Diana Lousa, and Miguel Rocha. ProPythia: a Python Package for Protein Classification Based on Machine and Deep Learning. *Neurocomputing*, 484:172–182, 2022.

[155] Mauno Vihinen, Esa Torkkila, and Pentti Riikonen. Accuracy of Protein

Flexibility Predictions. *Proteins: Structure, Function, and Bioinformatics*, 19(2):141–149, 1994.

[156] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array Programming with NumPy. *Nature*, 585(7825):357–362, 2020.

[157] Joe G. Greener, Shaun M. Kandathil, Lewis Moffat, and David T. Jones. A Guide to Machine Learning for Biologists. *Nature Reviews Molecular Cell Biology*, 23(1):40–55, 2022.

[158] Iqbal H. Sarker. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3):160, 2021.

[159] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[160] Danielle Denisko and Michael M. Hoffman. Classification and Interaction in Random Forests. *Proceedings of the National Academy of Sciences*, 115(8):1690–1692, 2018.

[161] Asa Ben-Hur, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, and Gunnar Rätsch. Support Vector Machines and Kernels for Computational Biology. *PLOS Computational Biology*, 4(10):e1000173, 2008.

[162] Jake Lever, Martin Krzywinski, and Naomi Altman. Principal Component Analysis. *Nature Methods*, 14(7):641–642, 2017.

[163] Ian T. Jolliffe and Jorge Cadima. Principal Component Analysis: a Review and Recent Developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[164] Quan Wang. Kernel Principal Component Analysis and Its Applications in Face Recognition and Active Shape Models. *arXiv*, page 1207.3538, 2014.

[165] Matthew C. Cieslak, Ann M. Castelfranco, Vittoria Roncalli, Petra H. Lenz, and Daniel K. Hartline. T-Distributed Stochastic Neighbor Embedding (T-Sne): A. Tool for Eco-Physiological Transcriptomic Analysis. *Marine Genomics*, 51:100723, 2020.

[166] Van Hoan Do and Stefan Canzar. A Generalization of T-Sne and UMAP to Single-Cell Multimodal Omics. *Genome Biology*, 22(1):130, 2021.

[167] Yang Yang, Hongjian Sun, Yu Zhang, Tiefu Zhang, Jialei Gong, Yunbo Wei, Yong-Gang Duan, Minglei Shu, Yuchen Yang, Di Wu, and Di Yu. Dimensionality Reduction by UMAP Reinforces Sample Heterogeneity Analysis in Bulk Transcriptomic Data. *Cell Reports*, 36(4):109442, 2021.

[168] Davide Chicco and Giuseppe Jurman. The Advantages of the Matthews Correlation Coefficient (MCC) Over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics*, 21(1):6, 2020.

[169] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. Optimal Classifier for Imbalanced Data Using Matthews Correlation Coefficient Metric. *PLOS ONE*, 12(6):e0177678, 2017.

[170] Yoonsuh Jung and Jianhua Hu. A K-Fold Averaging Cross-Validation Procedure. *Journal of Nonparametric Statistics*, 27(2):167–179, 2015.

[171] Xiaowei Yang, Qing Shen, Hongquan Xu, and Steven Shoptaw. Functional Regression Analysis Using an F Test for Longitudinal Data With Large Numbers of Repeated Measures. *Statistics in Medicine*, 26(7):1552–1566, 2007.

[172] Nicolae Sapoval, Amirali Aghazadeh, Michael G. Nute, Dinler A. Antunes, Advait Balaji, Richard Baraniuk, C. J. Barberan, Ruth Dannenfelser, Chen Dun, Mohammadamin Edrisi, R. A. Leo Elworth, Bryce Kille, Anastasios Kyrillidis, Luay Nakhleh, Cameron R. Wolfe, Zhi Yan, Vicky Yao, and Todd J. Treangen. Current Progress and Open Challenges for Applying Deep

Learning Across the Biosciences. *Nature Communications*, 13(1):1728, 2022.

[173] Shuangshuang Chen and Wei Guo. Auto-Encoders in Deep Learning;A Review With New Perspectives. *Mathematics*, 11(8):1777, 2023.

[174] Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large Language Models Generate Functional Protein Sequences Across Diverse Families. *Nature Biotechnology*, 41(8):1099–1106, 2023.

[175] Céline Marquet, Michael Heinzinger, Tobias Olenyi, Christian Dallago, Kyra Erckert, Michael Bernhofer, Dmitrii Nechaev, and Burkhard Rost. Embeddings From Protein Language Models Predict Conservation and Variant Effects. *Human Genetics*, 141(10):1629–1647, 2022.

[176] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[177] Jeffrey A. Ruffolo, Jeffrey J. Gray, and Jeremias Sulam. Deciphering Antibody Affinity Maturation With Language Models and Weakly Supervised Learning. *arXiv*, page 2112.07782, 2021.

[178] Alexander Turchin, Stanislav Masharsky, and Marinka Zitnik. Comparison of BERT Implementations for Natural Language Processing of Narrative Medical Documents. *Informatics in Medicine Unlocked*, 36:101139, 2023.

[179] Jeffrey A. Ruffolo, Lee-Shin Chu, Sai Pooja Mahajan, and Jeffrey J. Gray. Fast, Accurate Antibody Structure Prediction From Deep Learning on Massive Set of Natural Antibodies. *Nature Communications*, 14(1):2389, 2023.

[180] David Prihoda, Jad Maamary, Andrew Waight, Veronica Juan, Laurence Fayadat-Dilman, Daniel Svozil, and Danny A. Bitton. BioPhi: a Platform for Antibody Design, Humanization, and Humanness Evaluation Based on Natural Antibody Repertoires and Deep Learning. *mAbs*, 14(1):2020203, 2022.

[181] Sofia S. Guimaraes Koch, Robin Thorpe, Nana Kawasaki, Marie-Paule Lefranc, Sarel Malan, Andrew C. R Martin, Gilles Mignot, Andreas Plückthun, Menico Rizzi, Stephanie Shubat, Karin Weisser, and Raffaella Balocco. International Nonproprietary Names for Monoclonal Antibodies: An Evolving Nomenclature System. *mAbs*, 14(1):2075078, 2022.

[182] Monica L. Fernández-Quintero, Anne Ljungars, Franz Waibl, Victor Greiff, Jan Terje Andersen, Torleif T. Gjølberg, Timothy P. Jenkins, Bjørn Gunnar Voldborg, Lise Marie Grav, Sandeep Kumar, Guy Georges, Hubert Kettenberger, Klaus R. Liedl, Peter M. Tessier, John McCafferty, and Andreas H. Laustsen. Assessing Developability Early in the Discovery Process for Novel Biologics. *mAbs*, 15(1):2171248, 2023.

[183] Patricia Estep, Isabelle Caffry, Yao Yu, Tingwan Sun, Yuan Cao, Heather Lynaugh, Tushar Jain, Maximiliano Vásquez, Peter M. Tessier, and Yingda Xu. An Alternative Assay to Hydrophobic Interaction Chromatography for High-Throughput Characterization of Monoclonal Antibodies. *mAbs*, 7(3):553–561, 2015.

[184] Wayne G. Lilyestrom, Sandeep Yadav, Steven J. Shire, and Thomas M. Scherer. Monoclonal Antibody Self-Association, Cluster Formation, and Rheology at High Concentrations. *The Journal of Physical Chemistry B*, 117(21):6373–6384, 2013.

[185] Amita Datta-Mannan, Jirong Lu, Derrick R. Witcher, Donmienne Leung, Ying Tang, and Victor J. Wroblewski. The Interplay of Non-Specific Binding, Target-Mediated Clearance and FcRn Interactions on the Pharmacokinetics of Humanized Antibodies. *mAbs*, 7(6):1084–1093, 2015.

[186] Thomas E. Kraft, Wolfgang F. Richter, Thomas Emrich, Alexander Knaupp, Michaela Schuster, Andreas Wolfert, and Hubert Kettenberger. Heparin Chromatography as an in Vitro Predictor for Antibody Clearance Rate Through Pinocytosis. *mAbs*, 12(1):1683432, 2020.

[187] Anna-Lisa Schaap-Johansen, Milena Vujović, Annie Borch, Sine Reker Hadrup, and Paolo Marcatili. T Cell Epitope Prediction and Its Application to Immunotherapy. *Frontiers in Immunology*, 12:712488, 2021.

[188] Anna Vaisman-Mentesh, Matias Gutierrez-Gonzalez, Brandon J. DeKosky,

and Yariv Wine. The Molecular Mechanisms That Underlie the Immune Biology of Anti-Drug Antibody Formation Following Treatment With Monoclonal Antibodies. *Frontiers in Immunology*, 11:1951, 2020.

[189] J P. Leonard, J. W Friedberg, A. Younes, D. Fisher, L. I Gordon, J. Moore, M. Czuczman, T. Miller, P. Stiff, B. D. Cheson, A. Forero-Torres, N. Chieffo, B. McKinney, D. Finucane, and A. Molina. A Phase I/II Study of Galiximab (An Anti-Cd80 Monoclonal Antibody) in Combination With Rituximab for Relapsed or Refractory, Follicular Lymphoma. *Annals of Oncology*, 18(7):1216–1223, 2007.

[190] Takashi Kojima, Kentaro Yamazaki, Ken Kato, Kei Muro, Hiroki Hara, Keisho Chin, Thomas Goddemeier, Stefan Kuffel, Morihiro Watanabe, and Toshihiko Doi. Phase I Dose-Escalation Trial of Sym004, an Anti-Egfr Antibody Mixture, in Japanese Patients With Advanced Solid Tumors. *Cancer Science*, 109(10):3253–3262, 2018.

[191] Wentao Chen, Leopold Kong, Stephen Connelly, Julia M. Dendle, Yu Liu, Ian A. Wilson, Evan T. Powers, and Jeffery W. Kelly. Stabilizing the CH2 Domain of an Antibody by Engineering in an Enhanced Aromatic Sequon. *ACS Chemical Biology*, 11(7):1852–1861, 2016.

[192] Leander Meuris, Francis Santens, Greg Elson, Nele Festjens, Morgane Boone, Anaëlle Dos Santos, Simon Devos, François Rousseau, Evelyn Plets, Erica Houthuys, Pauline Malinge, Giovanni Magistrelli, Laura Cons, Laurence Chatel, Bart Devreese, and Nico Callewaert. GlycoDelete Engineering

of Mammalian Cells Simplifies N-Glycosylation of Recombinant Proteins. *Nature Biotechnology*, 32(5):485–489, 2014.

[193] Ngoc Phuong Lan Le, Thomas A. Bowden, Weston B. Struwe, and Max Crispin. Immune Recruitment or Suppression by Glycan Engineering of Endogenous and Therapeutic Antibodies. *Biochimica et Biophysica Acta*, 1860(8):1655–1668, 2016.

[194] Marianne Brüggemann, Gareth T. Williams, Carol I. Bindon, Michael R. Clark, Matthew R. Walker, Roy Jefferis, Herman Waldmann, and Michael S. Neuberger. Comparison of the Effector Functions of Human Immunoglobulins Using a Matched Set of Chimeric Antibodies. *Journal of Experimental Medicine*, 166(5):1351–1361, 1987.

[195] John Lund, Greg Winter, Peter T. Jones, John D. Pound, Toshiyuki Tanaka, Matthew R. Walker, Peter J. Artymiuk, Yogi Arata, Dennis R. Burton, Royston Jefferis, and Jennifer M. Woof. Human Fc Gamma RI and Fc Gamma RII Interact With Distinct but Overlapping Sites on Human IgG. *The Journal of Immunology*, 147(8):2657–2662, 1991.

[196] Sarah M. Burbach and Bryan Briney. Improving Antibody Language Models With Native Pairing. *arXiv*, page 2308.14300, 2023.

[197] Giuseppe Licari, Kyle P. Martin, Maureen Crames, Joseph Mozdzierz, Michael S. Marlow, Anne R. Karow-Zwick, Sandeep Kumar, and Joschka Bauer. Embedding Dynamics in Intrinsic Physicochemical Profiles of

Market-Stage Antibody-Based Biotherapeutics. *Molecular Pharmaceutics*, 20(2):1096–1111, 2023.

[198] Sohita Dhillon. Moxetumomab Pasudotox: First Global Approval. *Drugs*, 78(16):1763–1767, 2018.

[199] Ana María Vázquez, Ana María Hernández, Amparo Macías, Enrique Montero, Daniel Gómez, Daniel Alonso, Mariano Gabri, and Roberto Gómez. Racotumomab: An Anti-Idiotype Vaccine Related to N-Glycolyl-Containing Gangliosides – Preclinical and Clinical Data. *Frontiers in Oncology*, 2:150, 2012.

[200] Jocelyn Quistrebert, Signe Hässler, Delphine Bachelet, Cyprien Mbogning, Anne Musters, Paul Peter Tak, Carla Ann Wijbrandts, Marieke Herenius, Sytske Anne Bergstra, Gülşah Akdemir, Martina Johannesson, Bernard Combe, Bruno Fautrel, Sylvie Chollet-Martin, Aude Gleizes, Naoimh Donnellan, Florian Deisenhammer, Julie Davidson, Agnès Hincelin-Mery, Pierre Dönnes, Anna Fogdell-Hahn, Niek De Vries, Tom Huizinga, Imad Abugessaisa, Saedis Saevarsdottir, Salima Hacein-Bey-Abina, Marc Pallardy, Philippe Broët, and Xavier Mariette. Incidence and Risk Factors for Adalimumab and Infliximab Anti-Drug Antibodies in Rheumatoid Arthritis: a European Retrospective Multicohort Analysis. *Seminars in Arthritis and Rheumatism*, 48(6):967–975, 2019.

[201] Richard M. Goldberg. Lessons Learned From the Edrecolomab Story: How a Checkered Past Became a Checkered Flag for Monoclonal Antibodies in

Colorectal Cancer Therapy. *Oncology Research and Treatment*, 28(6):311–312, 2005.

[202] Xiaoyan Jing, Qiwen Dong, Daocheng Hong, and Ruqian Lu. Amino Acid Encoding Methods for Protein Sequences: a Comprehensive Review and Assessment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(6):1918–1931, 2020.

[203] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *Computing Research Repository*, page 2106.09685, 2021.

[204] Florian Heinkel, Meghan M. Verstraete, Siran Cao, Janessa Li, Patrick Farber, Elizabeth Stangle, Begonia Silva-Moreno, Fangni Peng, Surjit Dixit, Martin J. Boulanger, Thomas Spreter Von Kreudenstein, and Eric Escobar-Cabrera. Engineering a Pure and Stable Heterodimeric IgA for the Development of Multispecific Therapeutics. *mAbs*, 14(1):2141637, 2022.

[205] James Sweet-Jones, Maham Ahmad, and Andrew C. R. Martin. Antibody Markup Language (AbML) — a Notation Language for Antibody-Based Drug Formats and Software for Creating and Rendering AbML (abYdraw). *mAbs*, 14(1):2101183, 2022.

[206] Alberto Dal Corso, Luca Pignataro, Laura Belvisi, and Cesare Gennari. Innovative Linker Strategies for Tumor-Targeted Drug Conjugates. *Chemistry – A European Journal*, 25(65):14740–14757, 2019.

[207] Riti Biswas, Ed Belouski, Kevin Graham, Michelle Hortter, Marissa Mock, Christine E. Tinberg, and Alan J. Russell. VERITAS: Harnessing the Power of Nomenclature in Biologic Discovery. *mAbs*, 15(1):2207232, 2023.

[208] Matthew I. J. Raybould, Oliver M. Turnbull, Annabel Suter, Bora Guloglu, and Charlotte M. Deane. Contextualising the Developability Risk of Antibodies With Lambda Light Chains Using Enhanced Therapeutic Antibody Profiling. *Communications Biology*, 7(1):62, 2024.

[209] Gestur Vidarsson, Gillian Dekkers, and Theo Rispens. IgG Subclasses and Allotypes: From Structure to Effector Functions. *Frontiers in Immunology*, 5:520, 2014.

[210] Rena Liu, Robert J. Oldham, Emma Teal, Stephen A. Beers, and Mark S. Cragg. Fc-Engineering for Modulated Effector Functions—Improving Antibodies for Cancer Treatment. *Antibodies*, 9(4):64, 2020.

[211] Roberto Navigli, Simone Conia, and Björn Ross. Biases in Large Language Models: Origins, Inventory, and Discussion. *Journal of Data and Information Quality*, 15(2):1–21, 2023.

[212] Heather Myler, João Pedras-Vasconcelos, Kelli Phillips, Charles Scott Hottenstein, Paul Chamberlain, Viswanath Devanaryan, Carol Gleason, Joanne Goodman, Marta Starcevic Manning, Shobha Purushothama, Susan Richards, Honglue Shen, Jad Zoghbi, Lakshmi Amaravadi, Troy Barger, Steven Bowen, Ronald R. Bowsher, Adrienne Clements-Egan, Dong Geng,

Theresa J. Goletz, George R. Gunn, William Hallett, Michael E. Hodsdon, Brian M. Janelsins, Vibha Jawa, Szilard Kamondi, Susan Kirshner, Daniel Kramer, Meina Liang, Kathryn Lindley, Susana Liu, ZhenZhen Liu, Jim McNally, Alvydas Mikulskis, Robert Nelson, Mohsen Rajabi Ahbari, Qiang Qu, Jane Ruppel, Veerle Snoeck, An Song, Haoheng Yan, and Mark Ware. Anti-Drug Antibody Validation Testing and Reporting Harmonization. *The American Association of Pharmaceutical Scientists Journal*, 24(1):4, 2021.

[213] Rebecca Crossdale-Wood. Benchmarking the Impact of AI Biologics Discovery & Optimisation for Pharma. Protein Engineering Summit Europe, Lisbon, Portugal, Nov 2023.

[214] Wonsuk Oh and Girish N. Nadkarni. Federated Learning in Health Care Using Structured Medical Data. *Advances in Kidney Disease and Health*, 30(1):4–16, 2023.