Machine Learning for *in silico* Optimisation and Design of Therapeutic Antibodies

Alissa M. Hummer

Department of Statistics

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2024

Abstract

Antibodies are essential proteins in our immune systems, defending against foreign pathogens. Their unique ability to bind strongly and specifically to theoretically any target has also made them one of the most important classes of therapeutics. While target binding affinity lies at the heart of therapeutic antibody efficacy, a wide range of properties affecting safety and developability must be considered. Machine learning (ML) offers great promise to overcome the bottlenecks of laborious and trial-and-error experimental optimisation of these properties. In this thesis, I describe the development of ML models for *in silico* antibody optimisation.

I begin by detailing efforts to predict the effects of mutations on antibody-antigen binding affinity. Using experimental and synthetic data, and an equivariant graph neural network architecture, I demonstrate that there are currently orders of magnitude too little experimental data available for accurate, generalisable prediction. I also investigate the role of dataset diversity and suggest guidelines for robust ML model development and evaluation in this area.

In the next chapter, I explore the interpretability of graph neural network affinity predictions by examining the weighting of interface components. Overall, current methods were unable to provide meaningful insights into the factors most important for model predictions.

As antibody development requires solving a complex, multi-objective optimisation problem beyond affinity, I have also used ML to investigate additional properties. I outline our Random Forest-based approach, trained on millions of sequences, which can distinguish human from non-human antibodies with near-perfect accuracy. These models form the basis of Hu-mAb, our antibody humanization tool.

I then describe our fine-tuning strategy to produce an antibody inverse folding model. AntiFold can guide antibody optimisation by identifying mutations that are predicted to maintain the structure and, therefore, structure-related properties of an antibody.

In my DPhil, I have evaluated the applications and limitations of ML to accelerate multiple steps in the antibody design pipeline. These contributions set the foundation for simultaneous multi-objective optimisation, as well as biasing antibody design towards favourable properties.

Publications

This DPhil led to the following publications:

Gordon, G. L., Greenshields-Watson, A., Agarwal, P., Wong, A., Boyles, F., **Hummer, A.M.**, Lujan Hernandez, A.G. and Deane, C.M. PLAbDab-nano: a database of camelid and shark nanobodies from patents and literature. bioRxiv (2024).

Hoie, M.H.*, **Hummer, A.M.***, Olsen, T.H., Nielsen, M. and Deane, C.M. (2024). AntiFold: Improved antibody structure-based design using inverse folding. *arXiv*.

Glaser, P., Paul, S., **Hummer, A.H.**, Deane, C.M., Marks, D.S. and Amin, A.N. (2024). Kernel-Based Evaluation of Conditional Biological Sequence Models *International Conference on Machine Learning*

Chinery, L.*, **Hummer, A.M.***, Mehta, B.B.*, Akbar, R., Rawat, P., Slabodkin, A., Le Quy, K., Lund-Johansen, F., Greiff, V., Jeliazkov, J.R. and Deane, C.M. (2024). Baselining the Buzz: Trastuzumab-HER2 Affinity, and Beyond. *bioRxiv*.

Hummer, A.M. and Deane, C.M. Designing stable humanized antibodies. (2024) *Nature Biomedical Engineering News & Views* **8**:3-4.

Hummer, A.M.*, Hoie, M.H.*, Olsen, T.H., Nielsen, M. Deane, C.M. (2023). AntiFold: Improved antibody structure design using inverse folding. *NeurIPS Machine Learning in Structural Biology Workshop*.

Hummer, A.M., Schneider, C., Chinery, C. and Deane, C.M. (2023). Investigating the Volume and Diversity of Data Needed for Generalizable Antibody-Antigen $\Delta\Delta$G Prediction. *bioRxiv*.

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

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

* Denotes equal contribution

Abbreviations

AA	Amino acid
AAR	Amino acid recovery
Ab	Antibody
ADA	Anti-drug antibody
Ag	Antigen
AID	Activation-induced cytidine deaminase
AOR	Adjusted overlap ratio
AP	Average precision
BCR	B cell receptor
BSA	Buried surface area
CDR	Complementarity-determining region
CNN	Convolutional neural network
CV	Cross-validation
D	Diversity
EGC	Equivariant graph convolutional
EGNN	Equivariant graph neural network
Fab	Fragment antigen-binding
Fc	Fragment constant
FR	Framework
Fv	Fragment variable
GNN	Graph neural network
GVP-GNN	Graph Vector Perceptron Graph Neural Network

H-bond	Hydrogen bond
Ig	Immunoglobulin
IMGT	International ImMunoGeneTics information system
INN	International Nonproprietary Name
ITC	Isothermal titration calorimetry
J	Joining
K_D	Dissociation constant
KDE	Kernel density estimate
KinExA	Kinetic exclusion assay
lDDT	Local distance difference test
LSTM	Long short-term memory
MD	Molecular dynamics
MG	Multivariate Gaussian
MHC	Major histocompatibility complex
MI	Mutual information
ML	Machine learning
MLP	Multi-layer perceptron
MSA	Multiple sequence alignment
NN	Neural network
OAS	Observed Antibody Space
OR	Overlap ratio
PCR	Polymerase chain reaction
pLDDT	Predicted local distance difference test
PR AUC	Area under the precision recall curve
PSSM	Position-specific scoring matrix
pTM	Predicted template modelling
ReLU	Rectified linear unit
RF	Random Forest
RMSD	Root mean square deviation

ROC AUC	Area under the receiver operating characteristic curve
S_r	Spearman's rank correlation
SAbDab	Structural Antibody Database
scFv	Single-chain fragment variable
SHM	Somatic hypermutation
SiLU	Sigmoid linear unit
SPR	Surface plasmon resonance
TanH	Hyperbolic tangent
Thera-SAbDab	Therapeutic Structural Antibody Database
TM	Template modeling score
T_m	Melting temperature
V	Variable
VH	Variable heavy chain
VL	Variable light chain
WT	Wild-type
YJS	Youden's J Statistic

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Motivation

Therapeutic antibodies are best-in-class treatments for diseases ranging from cancers to viruses. Their development, however, is plagued by long timescales (averaging more than a decade (Beall et al., 2019)), high costs (hundreds of millions to billions of dollars (Schlander et al., 2021)) and high failure rates (nearly 80% (Kaplon and Reichert, 2019)). Antibody therapeutics that reach clinical development can fail for various reasons, including limited efficacy (e.g., arising from insufficient therapeutic effect in response to target binding) and safety (e.g., immunogenicity) (Sun and Benet, 2020). Machine learning (ML) holds great promise to overcome these challenges by accelerating therapeutic development and enabling the design of better therapeutics. In this thesis, I explore the successes and limitations of ML for antibody development, focusing primarily on creating tools for antibody optimisation after an initial candidate has been obtained.

This chapter will provide the background required to contextualise the research presented in this thesis. Following a description of the development and functions of antibodies in the human adaptive immune response, I will explain the sequence and structure of antibodies. These confer properties that underpin the importance of antibodies not only in the immune response but also as therapeutics. I will then discuss how therapeutic antibodies are developed experimentally. Finally, I will provide an overview of how computational, and particularly ML, approaches have been applied to advance steps in the therapeutic antibody development pipeline.

1.2 Contributions

This chapter contains text reproduced from:

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

1.3 The role of antibodies in the immune system

Antibodies are proteins that form an integral component of the adaptive immune response against foreign molecules and pathogens, which evolved in jawed vertebrates. The functions of antibodies are mediated by their ability to bind strongly and specifically to targets, known as antigens. These abilities arise from their sequence, structure and development.

1.3.1 B cell development

Antibodies are produced by B cells in the immune system. B cells undergo multiple stages of immunoglobulin gene arrangement, mutation and selection to produce a highly diverse repertoire of B cell receptors (BCRs, precursors to antibodies) and, ultimately, secreted antibodies. The early stages of B cell development, from early pro-B cell to immature B cell, are host to the V(D)J recombination process, in which an initial functioning BCR is generated (Lieber et al., 1987) (see Section 1.4.1). The immature B cells then undergo selection to eliminate those with autoreactive BCRs, i.e., which bind to self-antigens, both in the bone marrow and after migration to peripheral lymphoid organs (e.g., spleen) (Goodnow et al., 1989; Russell et al., 1991; Casellas et al., 2001). The surviving B cells are then exposed to circulating lymph. Here, they may bind to an antigen and then be activated in a T cell-dependent

manner (Noelle and Snow, 1991; Parker, 1993). After binding to a BCR, the antigen is internalised, degraded and presented on the B cell, enabling the activation of CD4+ T cells in a linked recognition mechanism (Bretscher and Cohn, 1970; Noelle and Snow, 1991; Parker, 1993) (Figure 1.1). The T cells then activate the antigen-recognising B cells and promote their proliferation. Some of these B cells form a germinal centre, together with the T cells, and undergo affinity maturation to increase the affinity for the foreign antigen (Allen et al., 2007; Kerfoot et al., 2011) (see Section 1.4.2). The resulting B cells are then selected on the basis of antigen affinity: higher-affinity BCRs bind, uptake and present more antigen, resulting in stronger survival signals from CD4+ T cells (Anderson et al., 2009; Gitlin et al., 2014).

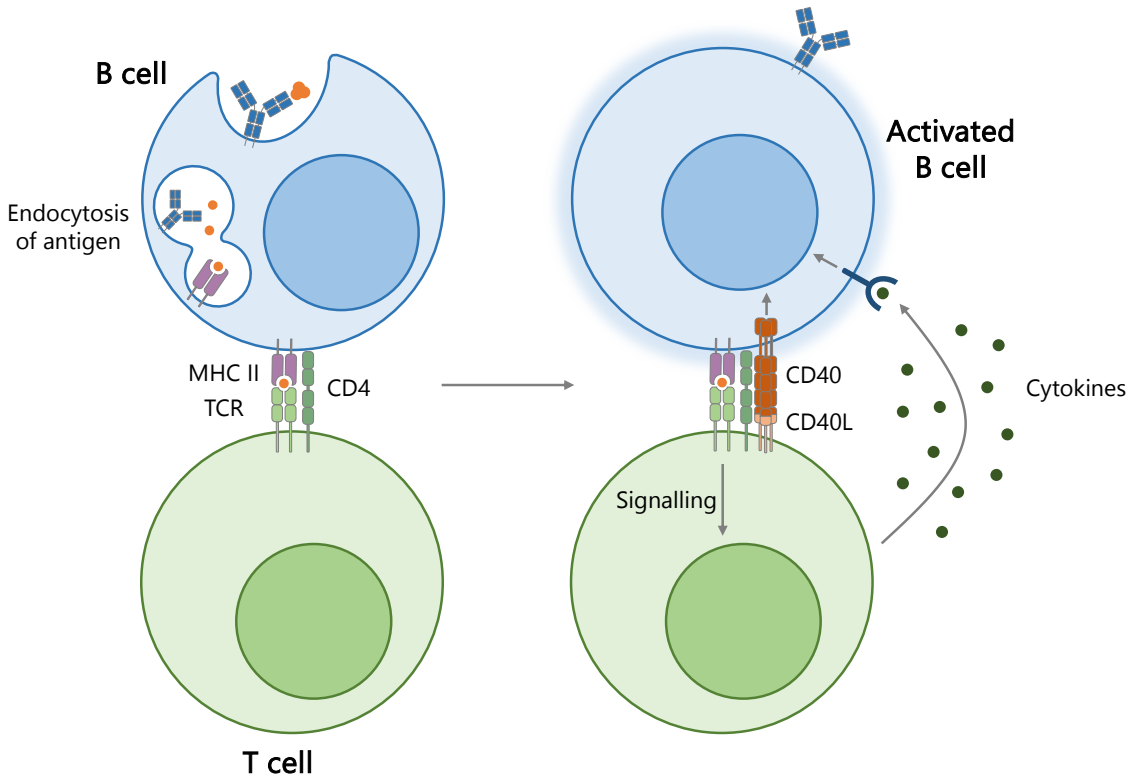

Figure 1.1: **B and T cell activation via linked recognition.** The antigen is bound by the B cell receptor (BCR) then internalised, degraded via endocytosis and presented on the surface of the cell. CD4+ T cells recognising this antigen become activated and in turn release signals, such as cytokines, to activate the B cell. This figure was adapted from Akiko Iwasaki.

There are five main antibody isotypes (IgM, IgD, IgG, IgA and IgE; Figure 1.2). While most of the text in this thesis centres on the variable regions of the antibody, which bind the antigen, the antibody isotypes differ in their constant regions, which can recruit immune effectors and therefore influence the antibody function. B cells originally express surface IgM and IgD. After activation following antigen binding, a B cell can undergo class switching, in which the antibody isotype it is expressing changes (Stavnezer, 1996). The immunoglobulin heavy chain gene contains exons for the different constant domains, each preceded by a 'switch region' (Dunnick et al., 1993). Double-strand breaks are introduced into two switch regions, initiated by the activation-induced cytidine deaminase (AID) enzyme, which also catalyses somatic hypermutation (Muramatsu et al., 1999, 2000; Revy et al., 2000) (see Section 1.4.2). The DNA repair process involves the removal of the DNA between these two breaks (Iwasato et al., 1990; Von Schwedler et al., 1990), thus placing a different constant region exon directly after the variable region DNA.

The vast majority of therapeutic antibodies are of the IgG isotype (Raybould et al., 2020), due to their prevalence in serum (Stoop et al., 1969; Zegers et al., 1975; Manz et al., 2005) and favourable properties, such as half-life (Tang et al., 2021). This thesis will therefore focus on IgG antibodies.

1.3.2 Antibody immune functions

IgG antibodies can act through a number of mechanisms, which are achieved through the antibody structure: the antibody variable domain binds the antigen and the constant domain can recruit immune effectors. Four main functions are neutralisation, opsonisation, complement activation and antibody-dependent cellular cytotoxicity. Neutralising antibodies bind to a toxin or the surface of a virus and prevent their uptake into cells (Bizebard et al., 1995; Forthal, 2015; Jiang et al., 2020). The latter

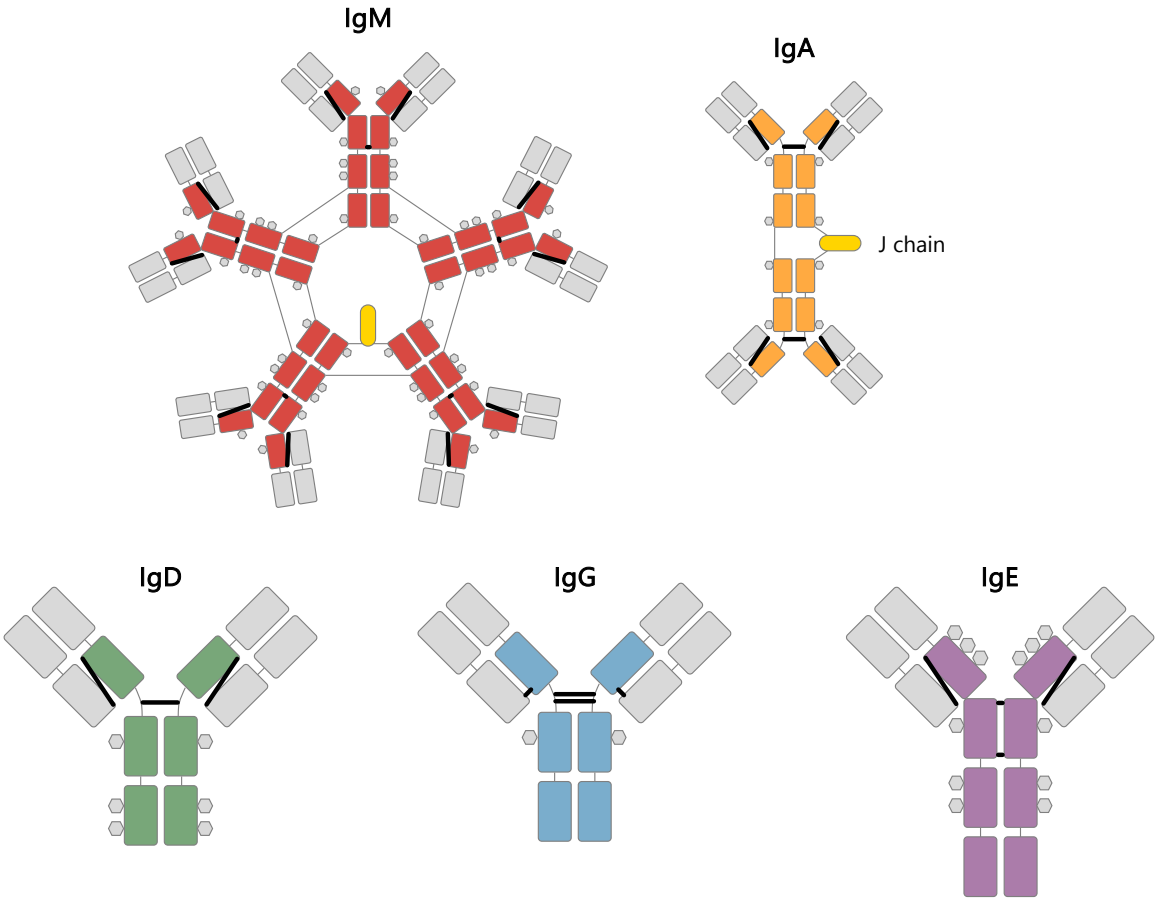

Figure 1.2: **Antibody isotypes in mammals.** The heavy chain constant domains, coloured, differ between antibody isotypes. IgM and IgA are multimeric (pentamer and dimer, respectively). Disulphide bonds are shown in thicker black lines; *N*-linked carbohydrate groups are shown as hexagons. This figure is adapted from Murphy et al. (2022).

three functions involve the recognition of antibody-bound pathogen surfaces/cells by phagocytes, complement proteins and natural killer cells, respectively, with each process resulting in the death of the bound cell (Forthal, 2015).

1.4 Antibody sequence and diversity

The B cell development process gives rise to an enormous level of antibody sequence diversity: theoretical estimates place the number of naïve (not yet exposed to antigen) antibodies up to $\sim 10^{15}$ (Briney et al., 2019; Schroeder, 2006), although this would

never be realised in a single human, as there are only on the order of 10^9 circulating B cells (Morbach et al., 2010; Rees, 2020). This diversity is achieved in the antibody variable domains (VH and VL, forming the fragment variable, Fv) through V(D)J recombination (Section 1.4.1), heavy-light chain pairing (Section 1.4.1) and somatic hypermutation during affinity maturation (Section 1.4.2). The high variability confers the ability of antibodies to bind nearly any molecule and underpins their role in the immune system. The constant domains, as the name suggests, are unchanged for a specific immunoglobulin isotype.

1.4.1 V(D)J recombination and heavy-light chain pairing: combinatorial and junctional diversity

The antibody Fv region is encoded by five separate genes: immunoglobulin heavy chain variable (V), diversity (D) and joining (J) genes, as well as immunoglobulin light chain V and J genes. There are two light chain isotypes: kappa (κ, K) and lambda (λ, L). The gene segments are combined for heavy and light chains, respectively, during the V(D)J recombination process in B cell development. The resulting heavy and light chains are then paired. This combinatorial diversity results in $2.9 * 10^6$ possible heavy-light chain variable domain gene pairs (gene counts from IMGT/GENE-DB (Giudicelli et al., 2005)), before allelic diversity and somatic hypermutation are even taken into account.

IGH: 57 IGHV genes × 23 IGHD genes × 6 IGHJ genes = 7866 combinations

IGK: 41 IGKV genes × 5 IGKJ genes = 205 combinations

IGL: 33 IGLV genes × 5 IGLJ genes = 165 combinations

Total: 7866 heavy chain × (205 + 165) light chain combinations

$= 2.9 * 10^6$ heavy-light chain variable domain combinations

In V(D)J recombination, the heavy chain is rearranged first: D to J_H followed by V_H to DJ_H (Alt et al., 1984). The latter step results in many unsuccessful rearrangements due to the need to achieve an in-frame junction, in which no stop codon is encountered during translation (Jung et al., 2006). Successfully rearranged heavy chains, termed μ heavy chains, are paired with 'surrogate' light chain-resembling proteins to form the pre-BCR (Sakaguchi and Melchers, 1986; Kudo and Melchers, 1987). Signalling through the pre-BCR results in the proliferation of the B cell (Loder et al., 1999; Levine et al., 2000) and subsequent VJ recombination of the light chain (V_L to J_L), which may take multiple attempts (Jung et al., 2006). The successfully rearranged light chain is paired with the μ heavy chain to result in BCRs on the surface of immature B cells.

The recombination steps are catalysed by the V(D)J recombinase, which includes the RAG-1 and RAG-2 proteins (Schatz et al., 1989; Oettinger et al., 1990; Van Gent et al., 1995; Mcblane et al., 1995). Recombination signal sequences at the ends of the immunoglobulin genes are aligned and excised out, along with the DNA in between (Rooney et al., 2004). The remaining double-strand DNA break between the immunoglobulin genes is repaired in an imprecise manner, in which nucleotides are

removed by DNA repair enzymes and introduced by the terminal deoxynucleotidyl transferase, resulting in junctional diversity (Desiderio et al., 1984; Komori et al., 1993; Gilfillan et al., 1993).

The resulting antibody Fv is classified into different regions: framework (FR) and complementarity-determining (CDR) (Figure 1.3). The V gene encodes the majority of the antibody variable domain, up to the beginning and middle of CDR3 in heavy and light chains, respectively. The J segment comprises the end of CDR3 and FR4, while the heavy chain D gene is part of the CDR3. The high level of variability in the CDR3 stems, in part, from the presence of gene junction(s), and therefore junctional diversity.

1.4.2 Affinity maturation: somatic hypermutation

Antibody diversity is further increased through somatic hypermutation (SHM), in which mutations are made to the antibody sequence, with those improving the affinity for the antigen being selected and propagated (Figure 1.6). In SHM, the AID enzyme converts cytosine to uracil bases at a rate of 10^{-3} per base pair per cell division (orders of magnitude greater than the rate of mutation for the rest of the DNA in the cell) (Maul and Gearhart, 2010). DNA repair pathways, including mismatch and base excision repair, are activated to address the foreign uracil (Di Noia and Neuberger, 2007). In mismatch repair, the uracil and adjacent nucleotides are removed and then filled in by error-prone DNA polymerases, typically affecting A-T base pairs (Zeng et al., 2001; Martomo et al., 2005; Delbos et al., 2007). The base excision pathway removes the uracil base to leave an abasic site, which, after two rounds of replication, will have been replaced by a random mutation; this path typically affects C-T base pairs (Seki et al., 2005; Prakash et al., 2005). Mutations accumulate in a step-wise manner, with those increasing antigen affinity carried forward. Affinity-

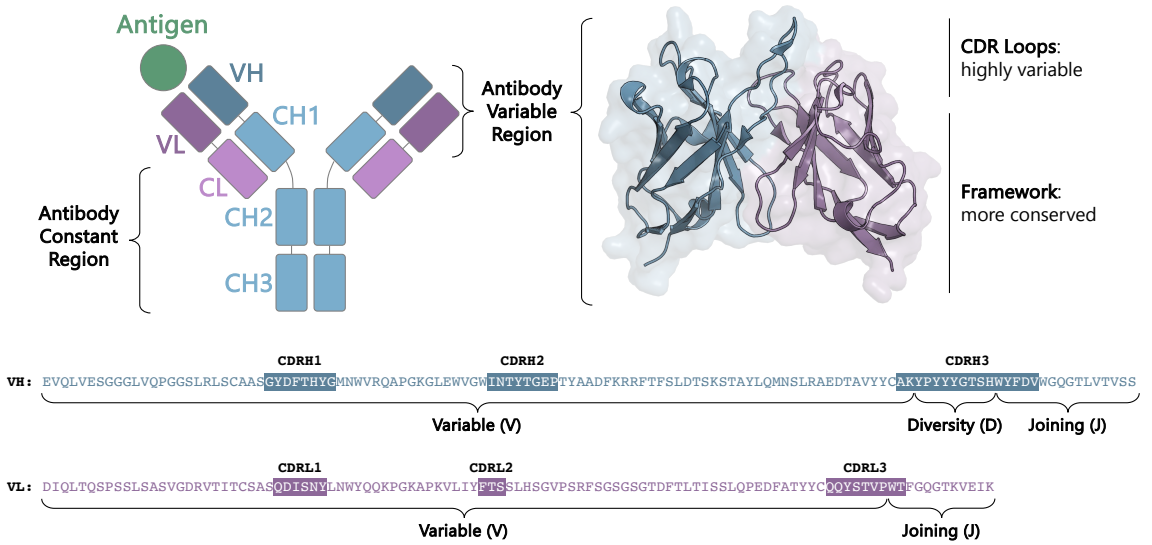

Figure 1.3: **Antibody structure and sequence.** Antibodies are formed from two heavy (blue) and two light (purple) chains. Binding to an antigen is mediated by the variable region, in particular the CDR loops. The antibody sequence is composed of heavy chain Variable (V), Diversity (D) and Joining (J) gene segments, as well as light chain V and J segments. Structure and sequence from PDB 1CZ8 (Chen et al., 1999).

improving mutations are expected to be much less frequent than those that reduce affinity, disrupt binding or prevent correct folding of the BCR. Most non-synonymous mutations that are positively selected for lie in the CDRs, consistent with their role in antigen-binding (Te Wu and Kabat, 1970; Jolly et al., 1996).

1.4.3 Antibody numbering

The antibody development process results in a pattern of more conserved FR and more variable CDR regions in the Fv (Figure 1.3). This pattern is captured in antibody numbering schemes, which are used to better understand the composition of and make comparisons between antibody sequences. Numbering schemes allow equivalent positions to be identified and thus different sequences to be aligned. Multiple antibody numbering schemes have been developed (e.g., Te Wu and Kabat, 1970; Chothia and Lesk, 1987; Honegger and Plückthun, 2001; Lefranc et al., 2003; Abhinandan and

Martin, 2008). While they differ in their basis for numbering (e.g., sequence and structural alignments of different depths and compositions) and insertion points, each method aims to distinguish between FRs and CDRs (although the boundaries are not always consistent).

The IMGT numbering scheme, used throughout this thesis, was introduced in 1997 and is applicable to both antibodies and T cell receptors (TCRs) (Lefranc, 1997; Lefranc et al., 2003). IMGT numbering is achieved using an alignment of germline sequences (initially limited to V genes and later extended). This scheme numbers heavy and light chains consistently, with the FR and CDR positions/boundaries matching between the two chains. Additionally, symmetrical insertions are incorporated in the CDRH3 at positions 111-112, which maintain the structural alignment of this loop for sequences of different lengths.

Various tools have been developed to number antibody sequences (e.g., Abhinandan and Martin, 2008; Ehrenmann et al., 2010; Adolf-Bryfogle et al., 2015; Dunbar and Deane, 2016). In this thesis, I use ANARCI, which aligns input sequences to hidden Markov models describing germline sequences for domain types and species (Dunbar and Deane, 2016).

1.4.4 Antibody sequence data

To better understand immune responses to disease or vaccination, explore antigen interactions and advance therapeutic development, antibody repertoires from individuals have been sequenced. These are hosted in databases including the Observed Antibody Space (OAS) database (Kovaltsuk et al., 2018; Olsen et al., 2022a), iReceptor (Corrie et al., 2018) and ImmuneDB (Rosenfeld et al., 2018). OAS, used throughout this thesis, stores antibody sequences in a standardised and easily accessible format. The sequences have been processed using the MiAIRR (Minimal

information about Adaptive Immune Receptor Repertoire) protocol and numbered using ANARCI (Dunbar and Deane, 2016). There are currently >2.4 billion unpaired sequences from 90 studies and >2 million paired sequences (in which the heavy-light chain pairings are known) from 12 studies in the database.

1.5 Antibody structure

The sequence patterns and functions of antibodies are reflected in their structure. Antibodies are formed from two heavy chains, consisting of three constant and one variable domain, and two light chains, consisting of one constant and one variable domain (Figure 1.3). Each domain adopts the immunoglobulin fold of 9 β-strands forming two β-sheets in a β-sandwich. In the variable domains, the CDR loops are at one end of the immunoglobulin fold, allowing them to form an interface which can interact with the antigen.

The structures of heavy chain CDR loops 1-2 (CDRH1-2) and light chain CDR1-3 (CDRL1-3) can be grouped into discrete 'canonical forms', respectively, which can be predicted from the sequence (North et al., 2011; Nowak et al., 2016; Wong et al., 2019; Kelow et al., 2022). Existing canonical forms have been identified from known structures and may increase as more structures become available. CDRH3 loop structures, however, owing to their especially high level of sequence variability, cannot be clustered (Weitzner et al., 2015; Regep et al., 2017). The FR regions can influence the antigen binding interface via the structure of the CDRs (Foote and Winter, 1992; Honegger and Plückthun, 2001), as well as the relative orientation of the VH and VL domains (Dunbar et al., 2013).

Antibodies can be separated into different fragments, due to the presence of multiple individually stable domains, via enzyme digestion (Andrew, 2003). Examples include the fragment crystallizable region (Fc: heavy chain CH2 and CH3 constant

domains), the fragment antigen-binding (Fab: heavy and light chain variable, CH1 and CL domains) and the fragment variable (Fv: variable domains only) (Figure 1.4).

1.5.1 Antibody-antigen interaction

Antibodies interact with antigens via non-covalent interactions, including hydrogen bond (H-bond), ionic and hydrophobic interactions. The interface residues on the antibody are referred to as the paratope and those on the antigen as the epitope. The antibody paratope is typically comprised mostly of CDR residues, with the CDRH3 loop often playing the most significant role due to its hypervariability (MacCallum et al., 1996; Sela-Culang et al., 2013). Tyr, Ser, Gly and Asn residues are prevalent in paratopes (Akbar et al., 2021). Epitopes exhibit a more even distribution of amino acid frequencies, although charged residues are among the most common (Akbar et al., 2021).

1.5.2 Antibody structure data

Solved structures of antibodies have formed the basis of our understanding of their function and our ability to develop therapeutics. The Structural Antibody Database (SAbDab) (Dunbar et al., 2014; Schneider et al., 2021) currently contains 8219 antibody structures. The vast majority (95%) were solved in complex with an antigen, providing insights into paratope-epitope interactions. The predominant methods used to solve these structures are X-ray diffraction – in which protein crystals are bombarded with X-rays to produce a diffraction pattern from which the spatial arrangement of atoms can be identified – and cryo-electron microscopy – in which frozen protein samples are exposed to electrons and the resulting electron densities are derived from image processing. These techniques are very time-consuming and expensive, limiting the number of solved structures that are available.

This is reflected in the antibody structure and sequence databases, which highlight the trade-offs for these types of data: structures are information-rich, providing atomistic-resolution insights into CDR loop conformations and antibody-antigen interfaces, but there are orders of magnitude fewer structures than sequences (both paired and unpaired) available. Computational methods are aiming to close this gap by generating high-accuracy predicted models of known sequences. These methods have made significant advances in recent years but limitations remain (see Section 1.7.1.1).

1.6 Antibodies as therapeutics

The properties that make antibodies such an important component of the adaptive immune response – the ability to bind strongly and specifically to a target antigen – have also made them one of the largest and most important classes of therapeutics. To date, there are 172 approved therapeutic antibodies (Raybould et al., 2020, 2024), used to treat diseases ranging from viruses and inflammatory diseases to autoimmune disorders and cancers. Antibodies are generally well-suited for diseases where a therapeutic effect can be achieved via binding to an extracellular target (e.g., through neutralisation or recruiting of immune effectors; see Section 1.3.2).

The majority of these therapeutics are full IgG antibodies, but there has been an increase in the adoption of alternate formats (Raybould et al., 2020; Carter and Rajpal, 2022). These include fragments of the full IgG structure, such as Fab and Fv fragments (see Section 1.5), as well as the single-chain Fv (scFv), comprised of heavy and light chain variable domains connected by a linker (Figure 1.4). The smaller size of antibody fragments yields favourable properties, such as greater tissue and tumour penetration (Yokota et al., 1992; Jain, 1990) and less expensive manufacture (Nelson, 2010). However, the lack of an Fc results in shorter half-lives (King et al., 1994) and

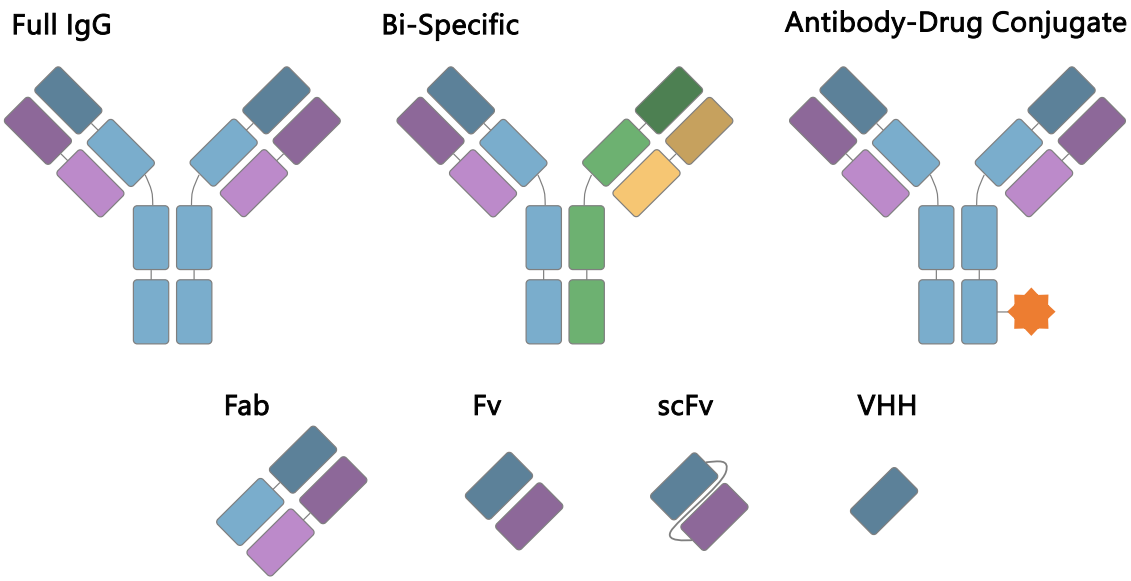

Figure 1.4: **Examples of therapeutic antibody formats.** Heavy chains are shown in blue and green; light chains are shown in purple and yellow. The drug component of the antibody-drug conjugate is shown in orange. IgG: immunoglobulin G; Fab: fragment antigen-binding; Fv: fragment variable; scFv: single-chain Fv; VHH: variable heavy domain of a heavy chain.

the loss of Fc-related immune effector functions (Section 1.3.2). In recent years, a diverse set of alternative antibody formats, beyond IgG fragments, has also emerged (Figure 1.4). The single-domain heavy chain based on camelid heavy-chain only antibodies, or VHH, is the smallest antibody therapeutic modality, consisting only of a VH domain with no partner light chain. Full IgG chains have also been adapted to form next-generation therapeutics. Examples include antibody-drug conjugates, in which drugs are linked to the antibody for specific delivery to a site in the body, and multi-specific antibodies, which have binding specificity for more than one antigen (e.g., through different Fv domains on each arm of the antibody).

The Therapeutic Structural Antibody Database (Thera-SAbDab) (Raybould et al., 2020) is a collection of all antibodies recognised by the World Health Organisation that have been approved for therapeutic use or are in clinical development (Phase I-III clinical trials).

1.6.1 Therapeutic antibody discovery

1.6.1.1 *In vivo* discovery

Therapeutic antibodies have traditionally been generated via animal immunisation, where the antigen of interest is injected into an animal such as a mouse (Greenfield, 2022a) (Figure 1.5). This approach leverages the natural immune response of the animal (see Section 1.3) to achieve antigen-specific antibodies. Typically the antigen is administered along with an adjuvant to stimulate a strong immune response (Greenfield, 2022a). Adjuvants can act via multiple pathways, including activating the innate immune system, which can strengthen an adaptive immune response, and promoting continuous antigen presentation (Coffman et al., 2010; Zhao et al., 2023). Repeated low doses of the target antigen can take advantage of *in vivo* affinity maturation (Section 1.4.2) to result in high-affinity antibodies (Greenfield, 2022a). To produce a homogenous set of monoclonal antibodies, immortal hybridomas are generated (Köhler and Milstein, 1975a). Activated B cells are obtained from the immunised animal's spleen and fused with myeloma cells. The resulting hybridomas are screened for antigen binding using methods such as flow cytometry, ELISA, immunoprecipitation-mass spectrometry and Western blot (Greenfield, 2022b). The hybridomas are immortal and can be continuously proliferated, as well as frozen and stored until needed. While this approach often results in antibodies with high affinity and specificity, it has limitations. Animal immunisation is time-consuming and expensive, does not allow for control over the specific epitope and results in non-human, potentially immunogenic antibodies. To overcome the immunogenicity risks, the binding portions of antibodies, the Fv or the CDR loops, can be grafted onto human antibodies to produce chimeric or humanised antibodies, respectively (see Section 4.3, Figure 4.1). Additionally, animals whose native immunoglobulin genes have been knocked out and replaced with human immunoglobulin genes have been

developed (Lee et al., 2014; Richardson et al., 2023).

Recent advances have also allowed antibodies to be identified from convalescent human patients who have been exposed to the target disease or a corresponding vaccine (Figure 1.5). Peripheral blood mononuclear cells can be collected from these patients and screened to identify antigen-specific memory B cells (Tiller et al., 2008; Smith et al., 2009). This approach has been adopted for infectious diseases including SARS-CoV-2 (e.g., Andreano et al., 2021; Rouet et al., 2023) and malaria (e.g., Alanine et al., 2019).

The properties of antibodies identified from *in vivo* techniques may need to be optimised. Beyond immunogenicity, improvements may be required for antibody stability, expression levels and solubility, for example (see Section 1.6.4).

1.6.1.2 *In vitro* discovery

In vitro methods based on display technologies provide an alternative to *in vivo* antibody discovery with more control over the selection process (Bradbury et al., 2011). Broadly, these approaches start with a large, diverse library of sequences (often antibody fragments such as scFvs or Fabs) and then undergo iterative steps of selection and propagation to enrich for antigen binders (Figure 1.5). Phage and yeast display are the most common display methods employed for the selection process. These exhibit a trade-off between the library size they can accommodate (up to 10^{11} for phage display versus at least one order of magnitude smaller in yeast) and antibody expression levels (typically higher in yeast due to the eukaryotic cell structure and protein-folding chaperones) (Zhao et al., 2012; Almagro et al., 2019). The initial antibody sequence libraries can be obtained from animals or human patients (including libraries of naïve antibodies and libraries from individuals exposed to a particular antigen or disease of interest), as well as made synthetically (Zhu and Dimitrov, 2009;

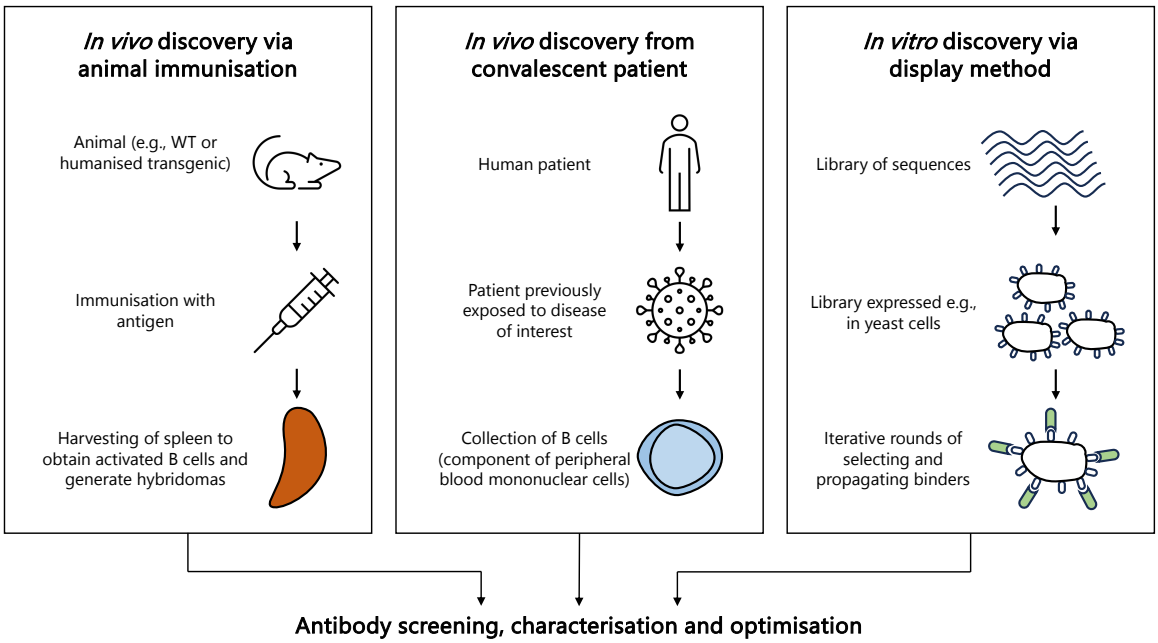

Figure 1.5: **Examples of antibody discovery methods.** Therapeutic antibodies can be obtained via *in vivo* (Section 1.6.1.1) and *in vitro* (Section 1.6.1.2) methods. *In vivo* methods include animal immunisation (of wild-type and transgenic animals), as well as discovery from convalescent human patients. *In vitro* methods are typically based on iterative rounds of binder enrichment from display methods (e.g., phage and yeast display). Libraries can be obtained from individuals (naïve repertoires, or after exposure to the target of interest), as well as produced synthetically. This figure is adapted from Laustsen et al. (2021).

Tiller et al., 2013; Weber et al., 2014; Kügler et al., 2018). Synthetic libraries are often derived from well-folded and -characterised framework scaffold(s), with diversity introduced in the CDR regions; some libraries only change the CDRH3 (Shim, 2015).

As for animal immunisation, display techniques do not enable control over the specific epitope targeted by an antibody. Correspondingly, while the variants are screened for binding, this does not always guarantee the desired functions, such as neutralising effects (see Section 1.6.3). The *in vitro* nature of this method does allow for greater control over other properties, however. Libraries derived from humans would theoretically produce human, low/non-immunogenic antibodies. Additionally, for example, the target affinity can be increased to levels exceeding that of antibodies identified from animal immunisation by making and screening more narrow variants

of known binders (Bradbury et al., 2011) (Section 1.6.2, Figure 1.6). The specificity profile of an antibody can also be tuned by altering the selectivity strategy, for example, to achieve cross/poly-reactive binders which may be beneficial for the treatment of certain diseases including viral infections (Bradbury et al., 2011). As for antibodies obtained from *in vivo* sources, further optimisation of properties is often required. To reduce this need, libraries have been developed around sequences that have 'drug-like' properties and which remove sequence liabilities (e.g., Teixeira et al., 2021).

1.6.2 Antibody-antigen binding affinity

The antigen binding affinity of antibody candidates obtained through *in vivo* or *in vitro* techniques often needs to be improved (Hudson and Souriau, 2003). Antigen binding affinity is an essential property of antibody therapeutics, conferring efficacy. High binding affinity can also reduce the therapeutic dose required, thereby decreasing the cost of treatment. Affinity maturation can be achieved *in vitro* through directed evolution and display techniques, similar to described above (Section 1.6.1.2). Conceptually, this approach mimics the diversification and selection stages of affinity maturation *in vivo*: sequence variants are generated, for example via error-prone polymerase chain reaction or targeted mutagenesis, and screened for binding (Figure 1.6). While these methods can improve affinity by enriching for strong binders in iterative selection cycles, they usually cannot provide an exact measurement of affinity.

Affinity is typically measured as the dissociation constant (K_D, units in Molar), from which ΔG can be calculated as: $\Delta G = RTln(K_D)$. There are multiple methods to measure binding affinity, with surface plasmon resonance (SPR) being one of the most widely used and reliable. In SPR, one binding partner is immobilised on a thin metal film while the other binding partner is passed over in solution. Binding

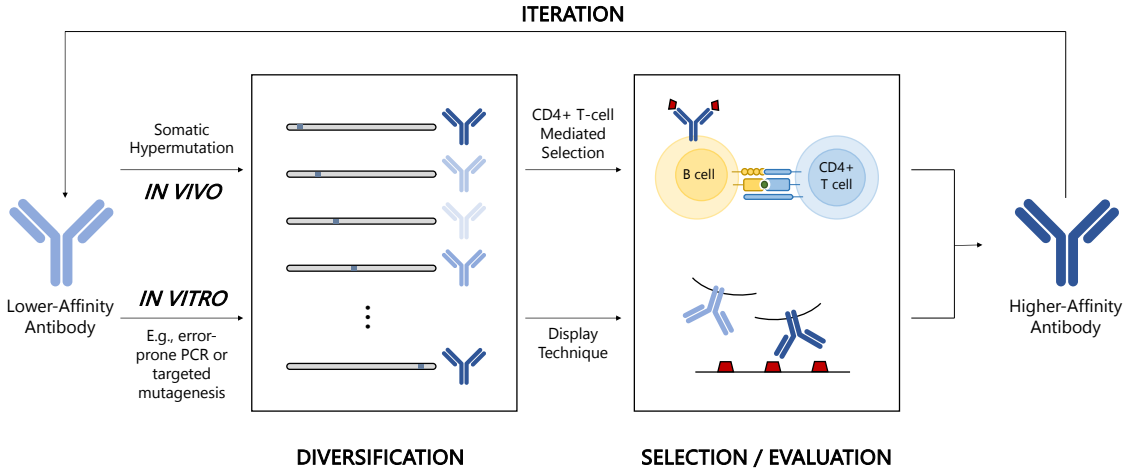

Figure 1.6: **Biological and experimental affinity maturation.** (Top) *In vivo* affinity maturation is achieved through somatic hypermutation followed by CD4+ T-cell mediated selection. Antibody precursor-expressing B cells compete such that cells with higher-affinity mutants proliferate. The selection panel (second from right) is adapted from Murphy et al. (2022). (Bottom) In *in vitro* affinity maturation, sequence diversification can be obtained through error-prone polymerase chain reaction (PCR) or targeted mutagenesis. The resulting sequences are then evaluated via a display technique (e.g., phage display) to select and propagate high-affinity binders.

events result in a change of mass and therefore change in the refractive index of the metal film, from which the K_D can be calculated (Douzi, 2017). Other affinity measurement methods include biolayer interferometry (Weeramange et al., 2020), isothermal titration calorimetry (ITC) (Lin and Wu, 2019) and kinetic exclusion assay (KinExA) (Darling and Brault, 2005).

1.6.3 Antibody function

Antibody discovery techniques typically identify, enrich and/or optimise for binders to a desired antigen. While binding may be a prerequisite, it does not necessarily guarantee the desired function (see Section 1.3.2 for descriptions of antibody functions). For example, the antibody may bind to a non-functional epitope and not be able to neutralise the target virus (i.e., prevent it from binding to the host cell membrane). Functional assays are therefore essential in the antibody discovery pipeline. These

tend to be more complex than binding assays, as a larger number of biological components and pathways are involved (Feavers and Walker, 2010). The specific assay will depend on the desired effect, but examples include neutralisation (e.g., Muruato et al., 2020), growth inhibition (e.g., Miura et al., 2023) and antibody-dependent cell cytotoxicity (e.g., Nelson et al., 1993) assays.

1.6.4 Developability properties

When developing a therapeutic antibody, a wide range of properties beyond affinity and efficacy must be considered (Figure 1.7). The requirements of a therapeutic antibody exceed those applied to antibodies *in vivo*. Therapeutics must be manufactured, stored and delivered to patients. Therapeutic antibody formulations can be many factors higher than the concentration at which antibodies are found in serum (typically less than 20 mg/mL) (Manz et al., 2005; Leeman et al., 2018; Ghosh et al., 2023). Additionally, most endogenous IgG antibodies have a half-life of approximately three weeks (Morell et al., 1970), while therapeutics may need to be stored for months after manufacture. The method from which the candidate antibodies are derived can introduce additional hurdles. For example, *in vitro* antibody discovery often does not filter out antibodies with off-target reactivity (Cunningham et al., 2021). Conversely, antibodies produced in non-human species may trigger harmful immune responses when administered to patients if recognised as foreign. Given the need to consider multiple different – and often incongruous (Rabia et al., 2018) – properties, antibody discovery processes commonly aim to produce multiple potential candidates which can then be selected for favourable properties and/or further optimised. Some of the key properties important for antibody safety and manufacturability are discussed below.

Polyreactivity describes off-target binding to unintended antigens, which can

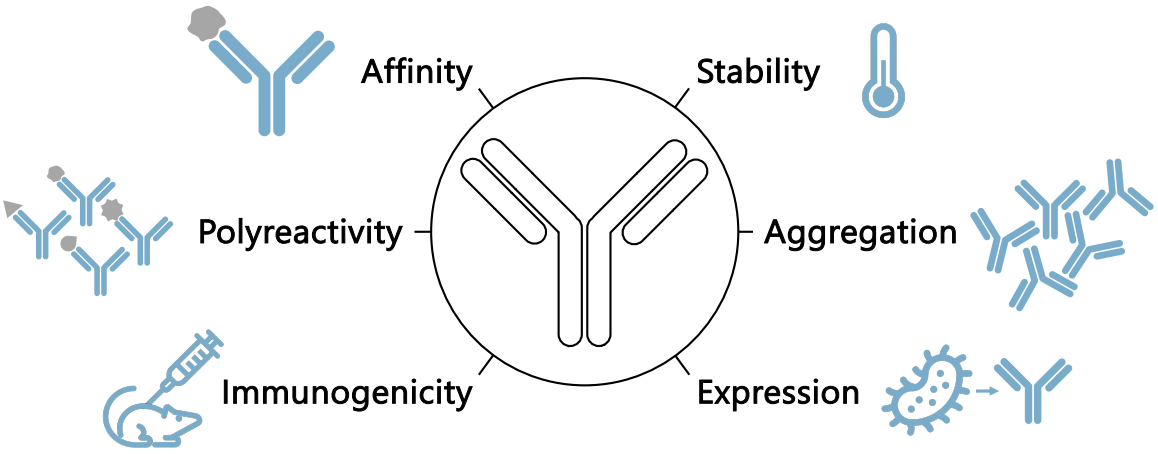

Figure 1.7: **Examples of antibody developability properties.**

result in harmful side-effects. Polyreactivity can be measured through experimental assays where the antibody is exposed to polyspecificity reagents, comprised of common or a heterogeneous mixture of potential off-target molecules (e.g., soluble and membrane protein fractions, double- and single-stranded DNA, lipopolysaccharide, etc.) (Wardemann et al., 2003; Hötzel et al., 2012; Xu et al., 2013; Makowski et al., 2021).

Immunogenicity, already briefly mentioned, is the propensity of a therapeutic to trigger an immune response in the patient it has been delivered to. This immune response can result in the formation of anti-drug antibodies (ADAs), which can reduce the efficacy of the therapeutic (by binding to and potentially neutralising it), and, in some cases, trigger harmful side-effects (including autoimmune reactions) (Hansel et al., 2010; Tovey and Lallemand, 2011). The risk of immunogenicity is higher in antibodies derived from non-human sources.

Aggregation arises from self-interactions between therapeutic antibody monomers to form aggregates. This property is inversely associated with **solubility**. Aggregation and solubility influence the 'shelf life' of a therapeutic antibody, as well as the concentration at which a therapeutic can be formulated (often required to be high

e.g., for subcutaneous injection (Davis et al., 2024)). Aggregation can also increase the risk of an immunogenic response (Bi et al., 2013; Ratanji et al., 2014; Kijanka et al., 2018; Lundahl et al., 2021; Swanson et al., 2022). Experimental techniques to measure antibody aggregation and solubility include dynamic/static light scattering and analytical ultracentrifugation (Geng et al., 2014).

Stability is the propensity of a therapeutic antibody to remain correctly folded. High stability contributes to easier manufacturing and storage of the antibody. The thermal shift assay (Huynh and Partch, 2015) is commonly used to measure stability as the melting temperature (T_m), at which half of the protein is unfolded.

Expression levels of an antibody from a recombinant expression system also have important implications for antibody manufacture (Frenzel et al., 2013). Depending on the format, antibodies may require eukaryotic/mammalian folding machinery for correct expression. Eukaryotic/mammalian expression systems are however more expensive than bacterial ones.

In sum, this leads to a complex, multi-parameter optimisation problem. The largest challenge is posed by the trade-offs between different properties: improving one can have a detrimental impact on another.

1.7 Computational antibody development

As described throughout this chapter, antibody development faces numerous difficulties and bottlenecks. Computational methods have been developed to accelerate steps in the development pipeline and to design better therapeutics. In recent years, there have been breakthroughs advancing the scope of ML methods for protein and antibody applications. Most notably, accurate structure prediction, led by AlphaFold2 (Jumper et al., 2021), has expanded the use cases of structure-based computational

methods to nearly any antibody, as structure inputs can be readily generated. Additionally, there have been great improvements in sequence-based language models, which can generate new sequences and provide information-rich embeddings. I will discuss some of the general principles and advances of computational methods applied to antibodies, as well as their limitations.

The computational techniques covered in this section and used throughout this thesis generally fall into two categories: 'physics-based' and ML. Physics-based methods are built on physical equations and empirical measurements. These methods, such as FoldX (Schymkowitz et al., 2005) and Rosetta (Leaver-Fay et al., 2011), calculate free energy using parameterised forcefields, which describe interactions between atoms and molecules. Molecular dynamics (MD) simulations also employ forcefields and solve Newton's equations of motion throughout simulations. ML methods, broadly, are derived from training ML architectures with relevant data (Figure 1.8). These methods can be trained in supervised or unsupervised manners, which differ in whether labels are provided for the input data or not. Supervised methods require labelled data (for example, experimental assay results for antibody properties outlined in Section 1.6.4), which is typically challenging and expensive to obtain for antibodies. With sufficient training data, supervised models are often able to accurately predict the property when applied to unseen sequences or structures. Unsupervised methods, in contrast, can be applied to large corpuses of unlabelled data, such as protein structures or sequences. This can be advantageous as there is far more unlabelled data available (e.g., millions to billions of sequences and thousands of structures, compared to only hundreds of binding affinity labels for antibodies). Unsupervised models can be used to create information-rich latent space embeddings of an input, as well as to generate new sequences or structures. However, the performance of unsupervised methods for predicting specific antibody properties, for example, is often limited as the model was not explicitly trained with this information (Chungyoun et al., 2023).

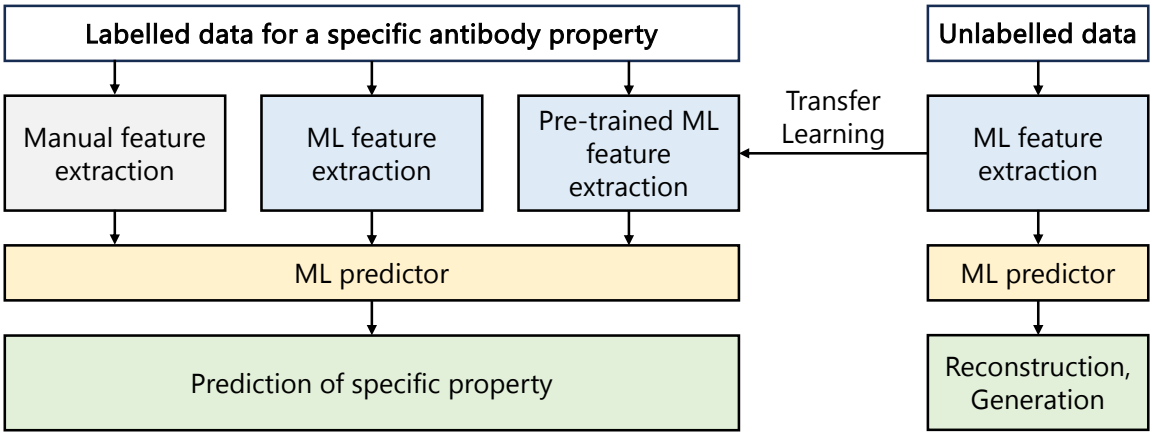

Figure 1.8: **ML strategies for biomolecular applications.** ML models can be trained in supervised and unsupervised manners. The former approach produces models that can be directly applied to predictive tasks. These are, however, limited by the availability of labelled data. Unsupervised models can be used for generative design (of e.g., structures and/or sequences), as well as to produce information-rich latent space embeddings and for transfer learning. This figure was adapted from Tobias H. Olsen.

The model components trained to extract feature information (and output embeddings) can form the basis for transfer learning (i.e., further training for a different task), though.

A wide range of ML architectures have been developed, with notable advances in the last decade particularly in neural network (NN) deep learning. There have also been improvements in the representation of proteins for input to ML. Protein sequences can be represented in multiple ways, with the simplest being one-hot encoding, which converts an amino acid sequence to a binary vector representation. Each position is replaced by a vector of dimension 20 (for 20 AAs), with one value in that vector set to 1 (corresponding to the index of the AA at that position within the 20 AAs) and the rest to 0. Sequences can also be represented as tokens, with one token per AA, for input to language models, as well as latent space embeddings (n-dimensional vectors) output by ML models. For structure-based inputs, graph representations of proteins have become increasingly popular, with graphs generated from 3-dimensional protein coordinates and options to featurise the nodes and edges

based on the physicochemical properties of the corresponding atoms and interactions.

1.7.1 Antibody structure modelling

1.7.1.1 Structure prediction

AlphaFold2 (Jumper et al., 2021) transformed protein structure prediction, cementing a shift toward deep learning for this task. In the CASP14 competition in 2020, AlphaFold2 produced structure models for unseen targets with median < 1 Å C_α root mean square deviation (RMSD) from true solved structures. The model architecture consists of two modules, the Evoformer and the Structure Module, and outputs a protein structure in an end-to-end (sequence-to-structure) manner. The model takes as input a MSA and, optionally with a template for a related structure, constructs an initial 'pair representation', capturing predicted residue interactions. These components are embedded, passed to and updated by the attention-based Evoformer. The Structure Module outputs 3-dimensional protein structure coordinates based on the single sequence embedding (extracted from the original sequence row of the MSA embedding) and pair representation embedding produced by the Evoformer. The amino acids are represented as triangles (of the backbone atoms) and their positions are updated in accordance with Invariant Point Attention and loss functions including Frame-Aligned Point Error. AlphaFold2 outputs confidence metrics, including predicted local-distance difference test (pLDDT), a local metric, and predicted template modelling score (pTM), a global metric, for predicted structures. Both correlate strongly with the respective true metrics (Pearson's correlation of 0.76 and 0.85). The model is trained to learn these metrics from true residue-level lDDT values (Mariani et al., 2013) and TM scores (Zhang and Skolnick, 2004).

The release of AlphaFold2 spurred the development of further methods, most of which built off of concepts used in AlphaFold2. Examples include RoseTTAFold2

(Baek et al., 2023), which explored which features of AlphaFold2 were essential for performance or could be removed, and ESMFold (Lin et al., 2023), which substantially sped up inference by replacing the need for an MSA input with a protein language model (ESM-2). AlphaFold2 was also extended to be applicable to protein complexes (AlphaFold-Multimer (Evans et al., 2022)).

Recently, AlphaFold3 was published (Abramson et al., 2024). The new model exhibits multiple changes compared to AlphaFold2. The MSA representation was removed and the Evoformer was replaced with a PairFormer (retaining only the pair and single representations). Additionally, the Structure Module was replaced with a Diffusion Module, which is applied directly to raw atom coordinates (see Section 1.7.3.3 for a brief explanation of diffusion models). A confidence head is trained to predict confidence metrics, including pLDDT. AlphaFold3 is no longer limited to proteins and can be used to predict the structures of diverse complexes (protein-protein, protein-ligand, protein-nucleic acids) and covalent modifications. It demonstrates substantial improvements over existing state-of-the-art models for most of these tasks, although results for its performance on single-chain and antibody (in the absence of an antigen) structure prediction are not yet available. At the time of writing this thesis, the code and weights for AlphaFold3 have not been publicly released.

Overall, these methods presented enormous advances for the field of protein structure prediction. However, they typically suffered from speed limitations and/or poor performance on antibody structures, particularly the CDRH3 loop. Antibody-specific structure prediction tools (e.g., Ruffolo et al., 2022, 2023; Abanades et al., 2023) aim to address this challenge. AlphaFold-Multimer (Evans et al., 2022) and ABodyBuilder2 (Abanades et al., 2023) are the state-of-the-art structure predictors for antibodies. ABodyBuilder2 is an antibody-specific adaptation of the AlphaFold-Multimer Structure Module. An ensemble of four models is used to generate predictions and the

structure closest to the average is selected to undergo refinement using OpenMM (Eastman et al., 2017). ABodyBuilder2 has an advantage in speed over AlphaFold-Multimer, being over two orders of magnitude faster, and is used throughout this thesis. The structure of the CDRH3 loop is still difficult to predict, though, with model RMSDs close to 3 Å as compared to < 1 Å RMSD for all other FR and CDR regions.

1.7.1.2 Docking

Docking is the process of predicting the structures of bound complexes. Accurate antibody-antigen docking remains a significant challenge. Factors contributing to the difficulty include the great structural and sequence diversity of the CDRH3 loop, the possibility that the binding partners adopt different conformations in the bound versus unbound states (Guest et al., 2021; Fernández-Quintero et al., 2019) and the lack of evolutionary information guiding antibody-antigen complexes.

Docking is computationally expensive, especially for large and flexible molecules such as antibodies and their protein targets (Dauzhenka et al., 2018). Paratope and epitope prediction can reduce the docking search space by providing predicted binding interfaces (Norman et al., 2020). Current methods for paratope prediction offer reasonable accuracy (e.g., Paragraph: area under the precision-recall curve (PR AUC) = 0.725, area under the receiver operating characteristic curve (ROC AUC) = 0.934 (Chinery et al., 2022)). Epitope prediction has proven to be more challenging, with a structure-based antibody-agnostic method like DiscoTope-3.0 (Høie et al., 2024) only achieving a PR AUC of 0.22 (ROC AUC = 0.81). Even perfect paratope and epitope prediction do not tell us the binding mode and relative orientations of the antibody and antigen in the bound complex, though.

Protein docking methods, which do output this information, can be classified into

two main categories, flexible and rigid body. There is a trade-off between the two approaches, with the latter being much faster but less accurate. Docking methods can often generate a near-native pose, but the scoring or ranking of the poses often has little correlation with true binding energies (Lensink et al., 2020, 2021). Deep learning has been applied to enhance antibody-antigen docking by re-scoring poses generated using physics-based docking tools. DLAB is one such method. It employs a convolutional neural network (CNN) to improve docking pose ranking and identify antibody-antigen pairs which are more likely to be docked accurately (Schneider et al., 2022). AlphaFold2 has been adapted for a similar purpose: docked poses were fed to the AlphaFold2 Structure Module as templates and the model returned an output structure along with confidence metrics (pLDDT and pTMscore) (Gaudreault et al., 2023). The poses were re-ranked based on a linear combination of the standardised confidence metrics. These approaches improve the ranking of poses, but it remains difficult to rank the correct pose in the top 1-5, particularly when starting from models or unbound structures of the binding partners.

Other approaches aim to go from sequences directly to complex structures using deep learning. While many of these, including AlphaFold-Multimer (Evans et al., 2022) and AlphaFold3 (Abramson et al., 2024), have achieved success in general protein-protein docking, antibody-antigen complexes have proven much more challenging to model accurately.

AlphaFold3 demonstrated large performance increases in antibody-antigen complex prediction over AlphaFold-Multimer. However, the output from one run of AlphaFold3 was still only correct in <40% of cases and highly accurate in ∼10% of cases. The accuracy of the top-ranked complex increases as more predictions are generated, but only to approximately 60% correct and 30% highly accurate, respectively, for 1000 predictions per target. AlphaFold3 has not yet been benchmarked against

models other than AlphaFold-Multimer for antibody-antigen complex modelling.

Additionally, RoseTTAFold2 was recently finetuned on antibody structures (Bennett et al., 2024). When 'hotspots' (i.e., interface residues) are provided, the model can discriminate between true and decoy complex structures. More importantly, the model confidence metric, pAE, corresponds somewhat with the RMSD to the true complex: when pAE is low, the predicted models tend to have low RMSD. However, the model performance is poor when no information is given about the binding site, limiting the use cases.

1.7.1.3 Modelling mutations

Although accurate at predicting protein structures, deep learning models, including AlphaFold2, are typically not sensitive to small changes in sequence and thus cannot be accurately used to predict the effects of (single-point) mutations on structures (Zhang et al., 2021; Buel and Walters, 2022; Pak et al., 2023). Physics-based methods are commonly used but suffer from a trade-off between accuracy and run-time.

FoldX (Schymkowitz et al., 2005), for example, only changes the protein side-chain, not the backbone, for a mutation but is very fast. FoldX models mutations using a rotamer library and side-chain energy minimisation, in which the lowest-energy rotamers for the mutated site and neighbouring residues are retained. The forcefield of FoldX, used to calculate the energy, includes terms for solvent, van der Waals, H-bond and electrostatic interactions, as well as a simplified entropy estimation and clash term. These terms are based on empirical measurements and the forcefield was calibrated with experimental protein stability $\Delta\Delta$G values (Schymkowitz et al., 2005). This energy function is also the basis for FoldX's calculations of predicted $\Delta\Delta$G (stability or binding).

Rosetta (Leaver-Fay et al., 2011; Barlow et al., 2018) more thoroughly models the

effects of mutations on the protein structure, including the backbone, but takes orders of magnitude more time to run. Conceptually similar to FoldX, Rosetta employs a rotamer library and energy function. The 'backrub' method can be used to sample conformational changes to the side chain and backbone near the mutated site (Smith et al., 2009; Barlow et al., 2018). The side chain conformations are then optimised via packing and the full structure undergoes energy minimisation. The Flex ddG protocol (Barlow et al., 2018) has been developed to predict binding $\Delta\Delta G$ with the Rosetta Talaris all-atom energy function (Shapovalov and Dunbrack, 2011; Song et al., 2011; O'Meara et al., 2015), following backrub-based modelling of mutations.

MD simulations fall at the far end of the time-accuracy spectrum, taking the longest but typically providing the strongest performance. Mutations can be modelled using MD by first changing the side chain (with, for example, a tool like FoldX) and then using the resulting structure as input to a simulation. Over the course of a simulation, Newton's equations of motion are solved in accordance with parameterised forcefields (e.g., Ponder and Case, 2003; Christen et al., 2005; Vanommeslaeghe et al., 2010; Maier et al., 2015). The positions of atoms are iteratively updated and the most energetically favourable rotamers or conformations adopted. Specific MD techniques, including free energy perturbation (Zwanzig, 1954), thermodynamic integration (Khavrutskii and Wallqvist, 2011) and potential of mean force with umbrella sampling (Kirkwood, 1935; Torrie and Valleau, 1977; Kästner, 2011), can be used to calculate the $\Delta\Delta G$ of these mutations (Klimovich et al., 2015; Aldeghi et al., 2018).

1.7.2 Machine learning to predict and optimise antibody properties

Beyond modelling structure, ML approaches have been applied to improve antibody properties (see Section 1.6.4). The aim is to increase the speed and efficiency of

antibody optimisation by computationally identifying favourable mutations and estimating the effect these will have on a property of interest. These methods are typically trained in a supervised manner with labelled data for the specific property. Given the scarcity of available labelled data and the complexity of many properties (proving challenging to predict from small data amounts), predicting properties in a generalisable manner (i.e., to be applicable to any antibody or antibody-antigen complex) is difficult. Some methods have been trained for a narrower scope – antigen-specific and/or variants of an antibody. In antigen-specific cases, simpler sequence-based models have been shown to achieve the same performance as or even outperform more complex models (in architecture and/or input) (Chinery et al., 2024). Simpler architectures offer benefits in interpretability, for example by training on a small number of features (Makowski et al., 2024) or identifying signal motifs (Harvey et al., 2022), but may not be suitable for broader generalisation for all tasks.

A wide range of properties have been tackled using ML approaches, including affinity (Mason et al., 2021; Hummer et al., 2023; Chinery et al., 2024), humanness (Marks et al., 2021; Prihoda et al., 2022; Ramon et al., 2024; Ucar et al., 2024), polyreactivity (Harvey et al., 2022), solubility (Sormanni et al., 2017) and viscosity (Makowski et al., 2024). Additionally, some methods aim to optimise for multiple properties, e.g., affinity and specificity (Makowski et al., 2022), solubility and stability (Rosace et al., 2023) and self-association and polyreactivity (Makowski et al., 2023). Currently, most of these approaches are limited to identifying the Pareto front for two properties, but they lay the foundation for future simultaneous consideration of a larger number of properties.

More detailed descriptions of the computational methods that have been developed for affinity and humanness prediction and optimisation, the focuses of Chapters 2 and 4, are included in the respective Introduction sections (Sections 2.3, 4.3).

1.7.3 Generative design

Recent advances, often achieved through adopting ML architectures which have been developed for language (Vaswani et al., 2017) and images (Ho et al., 2020), have enabled the generative design of proteins. Rather than predict specific properties or features, these models can be used to create novel protein sequences and structures. Additionally, a useful feature of generative models is the ability to produce information-rich embeddings of the inputs.

1.7.3.1 Language models

Language models are deep learning architectures, commonly based on Transformers (Vaswani et al., 2017), which can learn the rules that govern protein sequences. These models offer promise to surpass previous methods for capturing evolutionary information, conserved motifs and sequence variability, such as position-specific scoring matrices (PSSMs) and hidden Markov models (HMMs). PSSMs (e.g., Altschul et al., 1997) are generated based on amino acid frequency in a multiple sequence alignment (MSA). HMMs (e.g., Remmert et al., 2011), also derived from MSAs, are statistical models that enable greater flexibility to insertions/deletions and local sequence context. Language models leverage deep learning to capture even longer-range sequence dependencies and have the potential to generalize beyond specific sequence alignments.

Language models can be trained, for example, to predict masked tokens (representing amino acids), to predict the next token or to reconstruct sequences. These models can then output residue probabilities for each position in a sequence, which can be sampled from to generate new sequences. One of the key advantages Transformer architectures present, which has revolutionised not only protein design but also natural

language processing, is the attention mechanism, through which language models can learn weightings of and interdependencies between positions (Vaswani et al., 2017).

Language models have been developed for general proteins (e.g., Ferruz et al., 2022; Lin et al., 2023; Madani et al., 2023) as well as specific to antibodies (e.g., Ruffolo et al., 2021; Olsen et al., 2022b; Leem et al., 2022; Shuai et al., 2023; Madani et al., 2023; Olsen et al., 2024; Kenlay et al., 2024). In addition to sequence generation, language models have shown promise for antibody affinity maturation (Hie et al., 2023; Chinery et al., 2024) and downstream task prediction based on their embeddings. A challenge for language models is to produce diverse and novel outputs: antibody language models, for example, have been shown to revert to germline sequences (Olsen et al., 2024). Such results, which could be achieved without an ML approach, reflect naïve rather than affinity-matured antibodies. Non-germline residues are important for high-affinity binding and therefore therapeutic effect.

1.7.3.2 Inverse folding

Inverse folding, the inverse task of structure prediction, takes a protein structure as input and predicts sequence(s) that would adopt that structure (Ingraham et al., 2019). While perhaps counterintuitive (as structure is usually thought to follow from sequence, not vice versa), this strategy can be used to re-design sequences for a given structure (for example, improving features including expression and stability (Sumida et al., 2024)) and to generate latent space embeddings of an input structure. Inverse folding models have been shown to learn meaningful information about protein properties they were not explicitly trained on, which can be leveraged for zero-shot prediction (Hsu et al., 2022) and transfer learning (Dieckhaus et al., 2024). Inverse folding models have been developed for general proteins (Ingraham et al., 2019; Strokach et al., 2020; Anand et al., 2022; Jing et al., 2021; Hsu et al., 2022; Dauparas et al., 2022), as well as specifically for antibodies (Høie et al., 2024; Dreyer

et al., 2023; Shanehsazzadeh et al., 2023; Shanker et al., 2023). For more detail, see Chapter 5.

1.7.3.3 Diffusion models

Diffusion models represent a great step forward in structure-based *de novo* protein design, with the capability to generate new protein structures. Denoising diffusion probabilistic models are built by iteratively adding noise and training to predict the pre-noised state (Ho et al., 2020). These models thereby learn the distribution of the input space. A key advantage of diffusion models is the ability to condition the inference process, for example, to generate specific protein architectures, binders for targets or re-design a portion of an existing sequence.

Notable successes in diffusion models for protein generation included RFdiffusion (Watson et al., 2023) and Chroma (Ingraham et al., 2023), which have both been experimentally validated. RFdiffusion, which is built on the RoseTTAFold structure prediction network, achieved double-digit percentage success rates for most tasks, including binder design (Watson et al., 2023). The diffusion models of RFdiffusion and Chroma only generate protein backbones and must be combined with an inverse folding model to generate a corresponding sequence. All-atom diffusion models, which can model side chains directly, including RFdiffusion All-Atom (Krishna et al., 2023) and Protpardelle (Chu et al., 2023), have been released more recently. Additionally, there are antibody-specific diffusion models, including RFantibody (RFdiffusion fine-tuned on antibodies) (Bennett et al., 2024), AbDiffuser (Martinkus et al., 2023), DiffAb (Luo et al., 2022) and IgDiff (Cutting et al., 2024). The only model which has been experimentally validated for the *de novo* design of antibody binders to date is RFantibody, although this model starts from an existing framework (albeit not one previously known to bind the target and without prespecifying the binding site and orientation). RFantibody produced binders for a range of therapeutically

relevant targets but was still limited by success rates (around 1% (Callaway, 2024)) and affinities (ranging from high nM to μM).

These models form an important basis for reaching entirely *in silico* therapeutic antibody design. However, they also highlight the challenges antibodies pose, necessitating antibody-specific models and resulting in substantially lower success rates than general protein binder design. Future directions will require a two-pronged approach of advancing ML architectures alongside expanding the data available to train them.

1.8 Thesis outline

This thesis describes my contributions to the development of ML tools to predict and optimise antibody properties.

In Chapter 2, I present an equivariant graph neural network (EGNN) approach to predict antibody-antigen binding affinity in a generalisable manner. Using experimental and synthetic data, I demonstrate that orders of magnitude more experimental data will be required to realise accurate, generalisable prediction. I also assess the importance of dataset diversity.

In Chapter 3, I explore the interpretability of affinity predictions made by the EGNN architecture, aiming to understand how different components of the graph (atoms/residues and interactions) are weighted for scoring interfaces.

Chapter 4 delves into another essential antibody property, humanness and immunogenicity. I describe Random Forest models which achieved near-perfect accuracy in discriminating human from non-human sequences. I discuss how these models are used for humanness optimisation, as well as the interpretability and context-sensitivity of these models.

In Chapter 5, I present an antibody-specific inverse folding model, fine-tuned from

a general protein model. In addition to sampling structure-consistent sequences, I show the applicability of this model for the downstream task of affinity prediction.

This thesis finishes with Chapter 6, describing the key conclusions from this work and future directions.

Chapter 2

Investigating the Volume and Diversity of Data Needed for Generalisable Antibody-Antigen ΔΔG Prediction

Contents

2.1 Motivation

Antibody-antigen binding affinity is the driving consideration in therapeutic antibody development. It underpins therapeutic efficacy and thus control of affinity must not be lost during the optimisation of other properties. Predicting the effects of mutations on affinity *in silico* would address this need, but remains a challenge.

In this chapter, I present an EGNN architecture, Graphinity, to predict change in binding affinity for antibody-antigen complexes. I investigate the challenges posed by the limited available experimental data, as well as the amount of data which will be required to accurately and generalisably predict antibody-antigen $\Delta\Delta$G.

2.2 Contributions

This chapter contains material reproduced from:

Hummer, A.M., Schneider, C., Chinery, C. and Deane, C.M. (2023). Investigating the Volume and Diversity of Data Needed for Generalizable Antibody-Antigen $\Delta\Delta$G Prediction. *bioRxiv*.

I carried out all of the model development, dataset preparation and analysis presented in this chapter unless otherwise stated. Constantin Schneider contributed to writing the EGNN code. Lewis Chinery prepared the Trastuzumab dataset, originally published by Mason et al. (2021).

2.3 Introduction

As described in Chapter 1, antibodies mediate their functions, both physiologically and therapeutically, by binding specifically to a target antigen. Many other properties, in addition to binding affinity, often referred to collectively as developability, also

play important roles. There have been substantial advances in recent years in using ML to predict such properties (see Section 1.7.2). However, changes to the antibody sequence to improve these properties must not come at the cost of binding affinity.

Experimental techniques for affinity quantification are typically slow and laborious (Jarmoskaite et al., 2020) (see Section 1.6.2). A fast and accurate computational predictor of change in affinity would fill a need in the antibody design pipeline. Furthermore, computational approaches can, in principle, incorporate information from different predictors to simultaneously optimise multiple properties, while still controlling binding affinity. *In silico* prediction of antibody-antigen affinity remains a challenge. As discussed in Section 1.7.1.3, traditional affinity prediction tools, such as FoldX (Schymkowitz et al., 2005) and Rosetta Flex ddG (Barlow et al., 2018), are based on physical equations and empirical measurements. FoldX was calibrated with an experimental protein stability dataset and both tools have been benchmarked on antibody-antigen $\Delta\Delta$G datasets, such as AB-Bind (Sirin et al., 2016). These physics-based methods have proven effective for certain applications (Leman et al., 2020) but can be limited in speed (taking on the order of seconds to hours per mutation prediction) and accuracy (with Pearson's correlations of 0.34 and 0.61 on the AB-Bind dataset, respectively) (Barlow et al., 2018; Pires and Ascher, 2016). In recent years, there has been a shift towards ML approaches, which can be divided into two main categories: sequence- and structure-based. Sequence-based methods have been successfully applied to predict affinity for a specific antigen in cases where a large amount of data is available (Mason et al., 2021; Bachas et al., 2022). These methods are not broadly generalisable: the information they are trained on is antigen-specific and the models cannot be readily applied to another antigen without further training. Structure-based methods promise greater generalisability by aiming to capture the interaction patterns across many different antibody-antigen complexes. Current methods are trained on features derived from antibody-antigen complex structures,

such as binding surface area, interatomic interactions and energy-based terms (Pires and Ascher, 2016; Wang et al., 2020; Myung et al., 2020). However, these methods appear to not predict well outside their training data (Wang et al., 2020; Geng et al., 2019). Given the costs associated with solving protein structures, there is substantially less structural than sequence data available (see Sections 1.4.4, 1.5.2). Additionally, they require the extraction of features, which can be slow and is subject to human bias.

Here I present Graphinity, an EGNN architecture for predicting change in antibody-antigen binding affinity. The deep learning models are built directly from protein complex structures, potentially enabling scalability and generalisability.

Graphinity achieved state-of-the-art performance for $\Delta\Delta$G prediction on single-point mutations from the experimental AB-Bind dataset (Sirin et al., 2016), achieving test Pearson's correlations of up to 0.80. However, further investigation indicated that this high performance stemmed from model overtraining and was not robust to train-test cutoffs, an observation which has been hinted at by results from previous methods (Wang et al., 2020; Geng et al., 2019; Liu et al., 2021; Behbahani et al., 2022).

To examine if the Graphinity architecture could be used to robustly predict change in binding affinity, I generated a large synthetic dataset of nearly 1 million $\Delta\Delta$G values using FoldX (Schymkowitz et al., 2005). Pearson's correlations close to 0.9, which were robust to train-test sequence identity cutoffs and noise, were achieved on this dataset.

This far larger dataset enabled the investigation of the volume and type of data needed for a potentially generalisable antibody-antigen $\Delta\Delta$G predictor. Investigating model performance with varying amounts of synthetic data demonstrated that there is currently insufficient experimental data to accurately predict $\Delta\Delta$G, with orders of magnitude more likely to be needed. The results also highlighted the importance of

dataset diversity for model predictiveness.

I validated that Graphinity can learn not only the FoldX forcefield but also experimental binding affinity by adapting and successfully applying it to a dataset of >36,000 HER2-binding and non-binding Trastuzumab variants (Mason et al., 2021).

2.4 Methods

2.4.1 Dataset preparation

2.4.1.1 Experimental ΔΔG data preparation

The AB-Bind dataset consists of 645 single-point mutations and ΔΔG measurements from 29 complexes. I downloaded this dataset, which was originally compiled by Sirin et al. (2016), from Wang et al. (2020). The sign on the ΔΔG labels was reversed to reflect $\Delta\Delta G = \Delta G_{WT} - \Delta G_{Mutant}$, as is done by Myung et al. (2020) and in the synthetic datasets. The structures were 'repaired' using FoldX (version 5) RepairPDB and the mutations modelled using FoldX BuildModel (Schymkowitz et al., 2005). In this step, the side chain conformations are modelled using a rotamer library and subsequently undergo energy minimisation. This dataset is referred to as Experimental_ΔΔG_645 (Appendix Table A.1, Appendix Figure A.1a).

As in mCSM-AB2 (Myung et al., 2020), reverse mutations were generated by mutating the forward mutant model back to the wild-type (WT) using FoldX BuildModel and setting the ΔΔG label to the negative value of the forward mutation (Experimental_ΔΔG_645 + Reverse Mutations). For some model development regimes, the training and validation datasets were augmented with reverse mutations. Reverse mutations were never included in the test dataset.

The Experimental_ΔΔG_645 dataset has multiple limitations: 5 of the complexes

do not contain an antibody (PDBs: 1AK4, 1FFW, 1JTG, 1KTZ, 3K2M)[1], 27 of the mutations are non-binders whose change in binding affinity has arbitrarily been set to -8 kcal/mol and there are 3 duplicated mutations with different ΔΔG values. This dataset was however used here to compare against the performance of previous methods, which were applied to it (Pires and Ascher, 2016; Wang et al., 2020; Myung et al., 2020).

These limitations prompted me to propose a new experimental antibody-antigen ΔΔG benchmarking dataset, Experimental_ΔΔG_608 (Appendix Table A.1, Appendix Figure A.1b), consisting of 608 single-point mutations, obtained by rigorous filtering of the SKEMPI 2.0 database (Jankauskaite et al., 2019). The total database consists of 7085 entries, 1150 of which are for antibody-antigen interactions. Non-antibody-antigen complexes, multi-point mutations, non-binder mutations, mutations in which the affinity could not be measured exactly and duplicates of mutations were removed. When filtering duplicate mutations, I preferentially retained those with kinetic data, based on the measurement method (SPR > KinExA > ELISA; IASP > SP)[2] and based on the temperature (298 > 296 > 303 > 310 > 283 > 298(assumed)). The final filtered dataset contained 608 single-point mutations from 44 complexes.

ΔΔG was calculated from the SKEMPI 2.0 database as:

$$\Delta G = RT * ln(K_d) \tag{2.1}$$

$$\Delta\Delta G = \Delta G_{WT} - \Delta G_{Mutant} \tag{2.2}$$

[1]The protein complex components in these 5 PDBs are listed. 1AK4: cylophilin A–HIV-1 capsid; 1FFW: chemotaxis proteins CheY-CheA; 1JTG: β-lactamase TEM–β-lactamase inhibitory protein; 1KTZ: TGF-β3–TGF-β Type II Receptor; 3K2M: Proto-oncogene tyrosine-protein kinase ABL1–Monobody HA4.

[2]SPR: Surface Plasmon Resonance; KinExA: Kinetic Exclusion Assay; ELISA: Enzyme-Linked Immunosorbent Assay; IASP: Inhibition Assay Spectroscopy; SP: Spectroscopy. The filtering order reflects approximate reliability (Geng et al., 2016).

A limitation common to both experimental datasets is that the entries were collected from multiple sources, and therefore include affinity measurements from different labs and experimental set-ups, which is likely to introduce noise.

See Figure 2.1a for an example of ΔΔG data.

2.4.1.2 Synthetic ΔΔG data preparation

To investigate affinity prediction without the constraint of dataset size, I generated a synthetic dataset orders of magnitude larger than the experimental datasets (Figure 2.1b).

Structurally resolved antibody-protein antigen complexes were downloaded from SAbDab (Dunbar et al., 2014; Schneider et al., 2021), resulting in 6077 non-redundant entries from 3065 PDB files (SAbDab accession date: 19 May 2022). Twenty-seven PDBs with only C_α residues resolved were removed from the dataset. The PDB files were renumbered using a custom script, to prevent issues with insertion numbering in subsequent steps with FoldX, and repaired using FoldX RepairPDB (Schymkowitz et al., 2005). Twenty-five PDBs for which the repair did not run to completion were removed from the dataset. The antibody-antigen complexes were then clustered based on a 90% length-matched CDR sequence identity threshold (see below), resulting in 1475 clusters. One complex per cluster was carried forward for exhaustive interface mutagenesis: all interface residues, defined as being within 4 Å of the binding partner, were mutated to every other amino acid using FoldX BuildModel (Schymkowitz et al., 2005). The Interaction Energy for each WT and mutant complex was estimated with FoldX AnalyseComplex (Schymkowitz et al., 2005), and the FoldX ΔΔG determined as:

$$\Delta\Delta G = InteractionEnergy_{WT} - InteractionEnergy_{Mut} \tag{2.3}$$

such that a negative ΔΔG represents a destabilising mutation.

Mutations where the WT amino acid was 'X', the chain identifier was a number, the antibody and antigen were >4 Å apart and the FoldX Interaction Energy calculation failed were excluded. The final dataset (Synthetic_ΔΔG_942723, Appendix Table A.1, Appendix Figure A.1c) consisted of 942,723 mutations from 1471 antibody-antigen complexes from 1409 PDBs. The dataset is publicly available at `https://github.com/oxpig/Graphinity`.

2.4.1.3 Train-validation-test cutoffs

All antibody sequences were numbered with ANARCI (Dunbar and Deane, 2016) using the IMGT numbering scheme (Lefranc et al., 2003). The CDRs were extracted, concatenated and binned based on length. CD-HIT (Fu et al., 2012), with varying sequence identity cutoffs, was applied to cluster the length-matched CDRs. Seventy percent was the lowest threshold rounded to 10 for which CD-HIT ran (for CDRs and antigen sequences).

The AB-Bind Experimental_ΔΔG_645 dataset contained non-antibody-antigen complexes, which could not be clustered by CDR sequence identity. The sequences of each of the chains in these complexes had less than 90% sequence identity with each other and each complex was considered as its own cluster.

A synthetic dataset split was generated with an antigen sequence identity cutoff in addition to the antibody CDR sequence identity cutoff. In this case, antigen sequences were extracted from the PDB structures using the Bio.PDB.PDBParser module and clustered using CD-HIT with a 70% sequence identity cutoff. Clusters from the antibody CDR- and antigen-based sequence identity cutoffs were merged such that no cluster had a complex with >70% length-matched CDR sequence identity to an

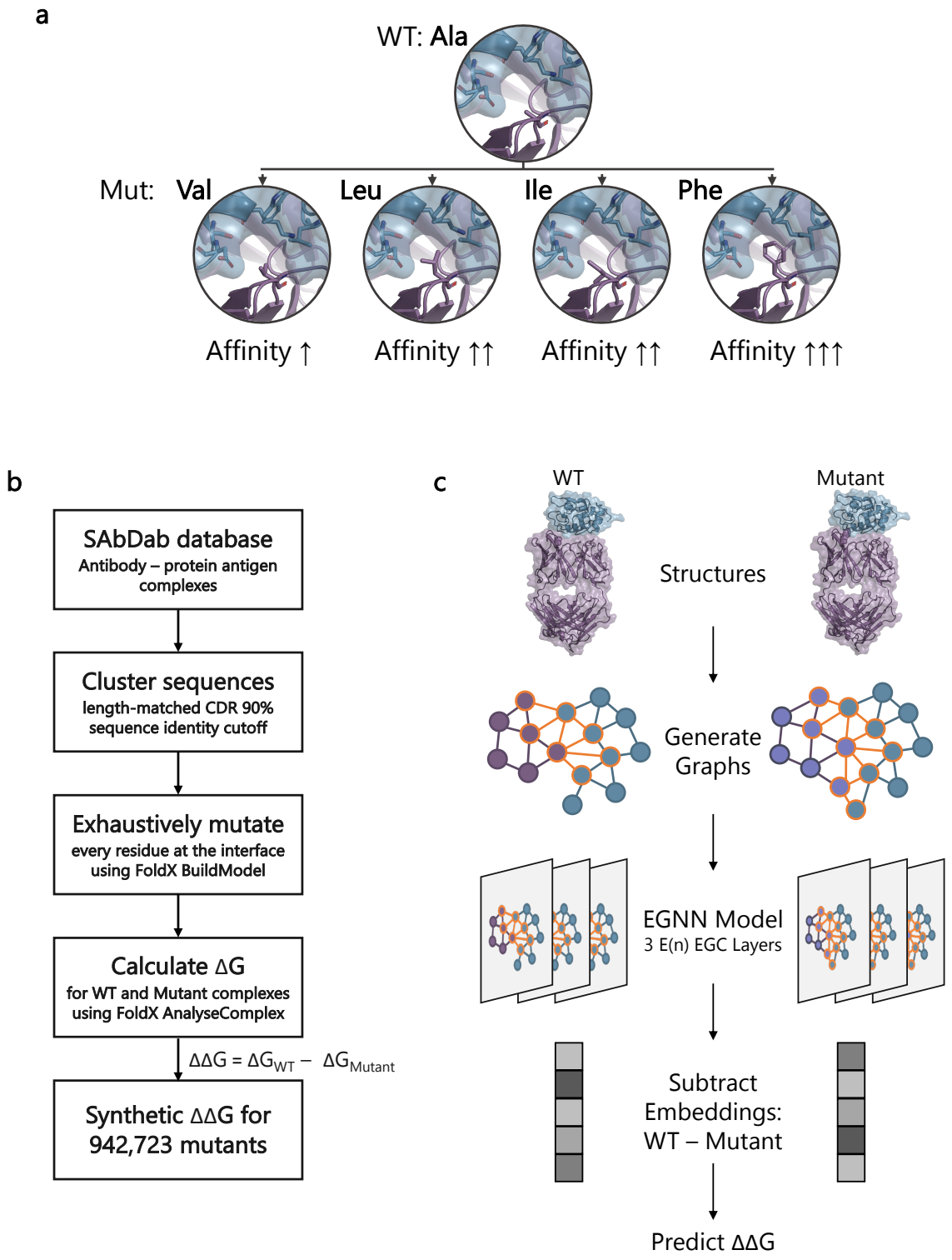

Figure 2.1: **Graphinity architecture and synthetic dataset preparation.** (a) Example of ΔΔG data for a complex. PDB: 1XGP (Li et al., 2005); affinity values from SKEMPI 2.0 (Jankauskaite et al., 2019). (b) Outline of the method for generating the synthetic ΔΔG dataset using FoldX. (c) Schematic of the EGNN deep learning model architecture.

antibody in another cluster nor >70% sequence identity to an antigen in another cluster.

Train-validation-test datasets were generated with an 80%-10%-10% split, with respect to the full dataset size. The datasets were sampled such that no cluster had members in more than one dataset, with the exception of datasets split with no cutoff. For 10-fold cross-validation, 10 dataset folds were created using the CD-HIT clusters, such that no cluster had members in more than one fold.

Unless otherwise specified, models were built from a single-fold 80%-10%-10% split with a 90% length-matched CDR sequence identity cutoff.

2.4.1.4 Varying synthetic dataset amounts

To investigate the role of dataset size on model performance, I trained models on subsets of the full, large-scale synthetic dataset (Synthetic_ΔΔG_942723). These subsets were randomly sampled from the respective train and validation datasets (Synthetic_ΔΔG_{580-450000}, Appendix Table A.1). All models were evaluated on the same test set, consisting of 94,126 mutations (one fold, held-out test set). A 90% length-matched CDR sequence identity cutoff was applied between respective train, validation and test sets.

2.4.1.5 Varying synthetic dataset diversity

The importance of dataset diversity for model performance was explored via the following three metrics:

- The number of antibody clusters, following clustering with a 90% length-matched CDR sequence identity cutoff
- The number of amino acid substitution types (e.g., Arg to Lys; Arg to Ala)

- The distribution of amino acid substitutions in the complex: mutation locations were classified based on the binding partner (antibody/antigen) and proximity to the interface centre; for the latter, the interface was divided into two areas (inner and outer shell) defined by concentric circles where, assuming that the interface is approximately flat, the outer shell circle was defined with a radius $\sqrt{2}$ times the radius of the inner shell circle, to produce two equal areas)

Training and validation datasets minimising and maximising the different metrics of diversity (Synthetic_ΔΔG_100000_{sequence/substitution_type/substitution_distribution}_{min/max}, Appendix Table A.1) were created. The test data was kept the same in each case. The respective training, validation and test datasets consisted of 100,000 mutations combined. A 90% length-matched CDR sequence identity cutoff was applied between each.

2.4.1.6 Investigating model robustness to noise

I assessed the robustness of Graphinity to noise by (1) shuffling and (2) applying random noise from a Gaussian distribution to the training and validation dataset affinity labels. In each of these cases, the test data remained unmodified.

- **Shuffling:** Varying percentages of the training and validation ΔΔG dataset labels were shuffled. The effective shuffling percentage was not necessarily equal to the percentage of the dataset which was shuffled, as some labels are the same and others were shuffled back into the same place.
- **Gaussian noise:** Gaussian noise was applied by adding random values generated from a normal distribution, using numpy.normal, with a set scale (0.5, 1, 2, 5 or 10) to the training and validation datasets.

2.4.1.7 Evolutionarily grounded mutations

A recent study demonstrated that the likelihood of FoldX incorrectly predicting a mutation to be stabilising (in this case, independent of an antigen) could be decreased by up to 11% by limiting FoldX predictions to mutations that are observed naturally (Rosace et al., 2023). I investigated the effect of limiting the test dataset to such 'evolutionarily grounded' mutations on model performance.

PSSMs generated from subsets (Prihoda et al., 2022) of the OAS database (Olsen et al., 2022a; Kovaltsuk et al., 2018) and corresponding custom code for calculating log-likelihoods were obtained from the authors of Rosace et al. (2023). As in Rosace et al. (2023), the 'evolutionarily grounded' mutations were defined as those with a positive log-likelihood and which have a log-likelihood greater than is seen for the wild-type residue.

The log-likelihood scores were mapped to the dataset mutations via the Aho numbering scheme (Honegger and Plückthun, 2001), as this was used for the PSSMs, with sequences numbered using ANARCI (Dunbar and Deane, 2016). There were 10 PDBs where ANARCI failed to number a chain with the Aho numbering scheme (3U2S, 4DQO, 4Y5Y, 6BPE, 6E1K, 6OPA, 6U0N, 7EY0, 7LF8, 7LY9). This approach was applied to mutations from antibody chains from humans or mice, as identified in SAbDab (Dunbar et al., 2014; Schneider et al., 2021), as the PSSMs were restricted to these species.

Complexes with human or mouse antibodies made up 75% percent (710,562 mutations) of the full synthetic dataset. Just over half of these (366,862) were for mutations to an antibody chain. The final 'evolutionarily grounded' dataset consisted of 47,983 mutations. This set encompassed nearly every possible amino acid mutation (374 out of 380 total; the 6 that were not included involved mutations to or from Cys). Of

the 10 most common 'evolutionarily grounded' mutations, half could be achieved via a single-base pair change to the codon (e.g., Ser to Asn, codon AGU to AAU).

2.4.2 Graphinity: equivariant graph neural network architecture

We developed a deep learning EGNN architecture to predict change in antibody-antigen binding affinity (Figure 2.1c). GNNs are well-suited for protein structures, as they can operate directly on atomic coordinates. Although the architecture has limitations (for example, with respect to scalability with increasing graph size (Hu et al., 2020), over-smoothing (Rusch et al., 2023) and noise robustness (Dai et al., 2021; Wang et al., 2024)), GNNs have proven effective for diverse molecular modelling and generation tasks (e.g., Abanades et al., 2022; Dauparas et al., 2022; Trippe et al., 2022; Kong et al., 2022; Soleymani et al., 2024).

Our model, Graphinity, is composed of three E(n) Equivariant Graph Convolutional (EGC) layers (Satorras et al., 2021) with a hidden dimension of 128. The model takes the 3D coordinates of a protein complex structure (PDB file) as input and generates an atomic-resolution graph with nodes representing non-hydrogen atoms and edges representing interactions between nodes <4 Å apart. The node features are a one-hot encoded vector describing the LibMolGrid atom type (Sunseri and Koes, 2020) and the edge features a one-hot encoded vector describing whether the edge is intra-binding partner, i.e. between atoms on the same binding partner, or inter-binding partner, i.e. between atoms on different binding partners. The graphs represent the mutation site neighbourhood (for $\Delta\Delta$G prediction: atoms on the same chain as the mutated residue within 4 Å of the mutated residue (local neighbourhood) and atoms on the binding partner chain within 4 Å of these local neighbourhood atoms). The models were trained with Mean Squared Error loss on the $\Delta\Delta$G value predictions. The architecture was implemented using PyTorch and PyTorch Geometric.

Models were trained using PyTorch Lightning for 500 epochs, with the exception of the synthetic $\Delta\Delta$G dataset models, which, due to the high computational costs, were trained for 10 epochs.

For $\Delta\Delta$G prediction, graphs of the WT and mutant structures were aggregated. Both graphs were fed through the three E(n) EGC layers in a Siamese manner and the resulting embeddings were subtracted from one another (WT − Mutant) prior to the last linear layer.

The model parameters were set as:

Optimizer: Adam

Learning rate: 0.001

Batch size: 32

Dropout: 0.2

Weight decay: 1e-16

Graph readout: global_max_pool over nodes

TanH activation at the output of the coordinate function: True

Update coords: True

In the models generated with datasets limited to a specific amino acid substitution and transfer learning, model weights were initialised with those from the model trained on the full dataset. In these cases, the learning rate was set to 0.0001.

The model training times are given below for training with 1 GPU (NVIDIA RTX 6000) and 4 CPUs on a single data fold (80/10/10 train-validation-test data split).

Experimental_$\Delta\Delta$G_645 (500 epochs): ca. 1 hour

Experimental_$\Delta\Delta$G_608 (500 epochs): ca. 1 hour

Synthetic_ΔΔG_942723 (10 epochs): ca. 19.5 hours

Trastuzumab Variants (500 epochs): ca. 35 hours

2.4.3 Tree-based model trained on featurised structures

To investigate the role of model architecture, I generated a tree-based model trained on featurised structures. Features were derived from the antibody-antigen structures as in mCSM-AB2 (Myung et al., 2020):

- FoldX AnalyseComplex energetic terms: The FoldX AnalyseComplex function (Schymkowitz et al., 2005) was used to calculate interaction energetic terms (e.g., Van der Waals clashes, Van der Waals contributions, hydrogen bond contributions, electrostatic interactions and polar and hydrophobic solvation) for the WT and mutant complexes.
- Arpeggio interactions: The inter-protein interface interactions (e.g., H-bonds and ionic interactions) of the complexes were calculated using Arpeggio (Jubb et al., 2017).
- Pharmacophore vectors: To represent the change in amino acid upon mutation, a change in the pharmacophore counts, adapted from Pires et al. (2014), was calculated. Pharmacophores (e.g., hydrophobic, H-bond acceptor or H-bond donor) were assigned to each atom in each amino acid and summed across the amino acid (Appendix Table A.2). To note, an atom can have more than one pharmacophore.
- Buried surface area: The buried surface area (BSA) for each binding partner (antibody and antigen) in each complex was calculated using the PSA program (Lee and Richards, 1971): $\mathrm{BSA} = \mathrm{SA}_{\mathrm{free}} - \mathrm{SA}_{\mathrm{bound}}$. An average change in BSA across the two binding partners was calculated.

- PSSM evolutionary term: A measure of residue conservation at a position was captured in PSSMs. The PSSM scores were calculated using PSI-BLAST (Altschul et al., 1997) (parameters: evolutionary scoring matrix = PAM30, num_iterations = 3, evalue = 1E-10, seg = Yes, comp_based_stats = 1, and db = swissprot) as in mCSM-AB2 (Myung et al., 2020).

An Extra Trees model with 300 estimators was generated as in mCSM-AB2. This is not a direct comparison to the mCSM-AB2 model, as the graph-based features of the CSM-based models were not incorporated.

This featurisation and subsequent Extra Trees model was applied to the Experimental_ΔΔG_608 dataset. Given the time required for featurisation (on the order of minutes per mutation), it was computationally infeasible to apply this approach to the large synthetic dataset.

2.4.4 Trastuzumab variants

We obtained the dataset of Trastuzumab CDRH3 variants and corresponding binary binding labels from Mason et al. (2021). The sequences were mutated at 10 amino acid positions in the CDRH3. The variants which had been labelled as both binding and non-binding were assigned the binding label, as in Mason et al. (2021). This resulted in 36,391 variants, 11,277 of which were labelled as binding. The dataset was split (1) randomly using sklearn.model_selection.train_test_split and (2) with a clonotype plus sequence-identity split. For (2), variants were clustered based on the V- and J-gene assignments, as labelled by ANARCI (Dunbar and Deane, 2016), and sequence identity of the CDRH3 (limited to the 10 mutated positions). Sequence identity in this case describes the maximum allowed edit distance from a representative sequence (cluster centre). For example, a minimum identity of 70% allows edit distances of up to three residues from the cluster centre. We used the clonotype and sequence

identity approach as CD-HIT did not run with the 10-position variant sequences due to their short length. The clonotype plus sequence-identity split data was prepared by Lewis Chinery.

The Trastuzumab datasets were prepared with a 70%-15%-15% train-validation-test split to allow comparison with Mason et al. (2021).

Structures of the Trastuzumab variants in complex with HER2 were modelled using the FoldX BuildModel function (Schymkowitz et al., 2005) starting from a FoldX-'repaired' structure of PDB 1N8Z (Cho et al., 2003). Although this approach is unlikely to capture the true structural effect of the mutations, as FoldX does not model changes to the backbone (Van Durme et al., 2011), it is fast and allows docking to be avoided by starting from a structure of a bound complex.

The Graphinity architecture was adapted for this task. The input was changed to be one graph only (and subsequently there was no subtraction of embeddings before the final layer) and the graphs were formed from the 10 mutated CDRH3 residues and surrounding neighbourhood (antibody atoms within 4 Å of CDRH3 atoms (antibody neighbourhood), antigen atoms within 4 Å of the antibody neighbourhood and antigen atoms within 4 Å of these antigen atoms). I also updated the model for classification by changing the loss function (to Binary Cross-Entropy with Logits) and accuracy metrics (to ROC AUC and average precision, AP).

2.5 Results

2.5.1 Graphinity performance for predicting experimental ΔΔG

I applied Graphinity to the experimental ΔΔG dataset from AB-Bind (Sirin et al., 2016) (Experimental_ΔΔG_645, Appendix Table A.1, Appendix Figure A.1a), includ-

ing and excluding hypothetical reverse mutations in the training and validation data (Experimental_ΔΔG_645 ± Reverse Mutations) as well as non-binder mutations with ΔΔG values arbitrarily set to −8 kcal/mol (Experimental_ΔΔG_645 ± Non-Binders).

Graphinity achieved Pearson's correlations of up to 0.80 on 10-fold cross-validation (Figure 2.2a), similar in performance to existing methods which report correlations of up to 0.76 (Wang et al., 2020; Myung et al., 2020). However, delving into the robustness of the model – by imposing sequence identity cutoffs between folds – indicated that these high correlations were the result of overtraining as opposed to true learning (Figure 2.2b). When a 100% length-matched CDR sequence identity cutoff was imposed, ensuring that mutations from the same complex cannot be in both the training and test dataset, the Pearson's correlations decreased by an average of 63%. The results were also highly sensitive to the inclusion of non-binders (Figure 2.2b) and, across all train-test cutoffs, there was substantial variation in the Pearson's correlation across different cross-validation folds (Appendix Figure A.2). Poor model robustness on experimental ΔΔG prediction has been found for previous approaches (Wang et al., 2020; Geng et al., 2019; Liu et al., 2021; Behbahani et al., 2022).

Tests of existing methods for antibody-antigen ΔΔG prediction have not imposed cutoffs between cross-validation folds, with the exception of leave-one-complex-out cross-validation, in which mutations from the same PDB cannot be in both the training and test set (although mutations from identical or closely related complexes in separate PDBs could be) (Wang et al., 2020; Myung et al., 2020). For one method, TopNetTree, this leave-one-complex-out test caused a drop in the average Pearson's correlation to 0.17 (Wang et al., 2020). For another method, mCSM-AB2, only a minor drop in performance was reported but they appear to include hypothetical reverse mutations in their test data (Myung et al., 2020).

Although widely used, the AB-Bind dataset suffers from several limitations, in-

cluding that five of the included complexes do not contain an antibody, despite the dataset being described as an "antibody binding mutational database" (Sirin et al., 2016) (more details in Section 2.4.1.1). I therefore propose a new experimental antibody-antigen single-point mutation ΔΔG dataset (Experimental_ΔΔG_608, Appendix Table A.1, Appendix Figure A.1b), consisting of 608 mutations filtered from the SKEMPI 2.0 database (Schymkowitz et al., 2005), for model benchmarking. Although this dataset has fewer single-point mutations, these mutations come from a slightly larger number of complexes (44). The performance of Graphinity on this dataset is similar to that for the Experimental_ΔΔG_645 dataset (Appendix Figure A.3a): model correlation is high when the data is split randomly and with reverse mutations, but once again is not robust to train-test CDR sequence identity cutoffs.

On this more rigorous dataset, I also investigated the role of model architecture. I applied a tree-based model built from features derived from the WT and mutant complex structures (more details in Section 2.4.3), similar to the method employed by the mCSM-based models (Pires and Ascher, 2016; Myung et al., 2020). This different model architecture gave similar correlations but also suffered from overtraining (Appendix Figure A.3b), suggesting that the problem lies in the data.

2.5.2 Using a synthetic dataset of ∼1 million mutations

The poor robustness of model performance on the limited experimental data led me to investigate how well ΔΔG could be predicted if more data was available. I generated a synthetic dataset of nearly 1 million ΔΔG data points (Synthetic_ΔΔG_942723, Appendix Table A.1, Appendix Figure A.1c) by exhaustively mutating the interfaces of structurally-resolved complexes from SAbDab (Dunbar et al., 2014; Schneider et al., 2021) using FoldX (Schymkowitz et al., 2005) (Figure 2.1b). This synthetic dataset will not completely mimic the complexity of true ΔΔG values. The Pearson's cor-

relation between FoldX predictions and experimental values is 0.34 for the AB-Bind dataset (Sirin et al., 2016). The accuracy is higher for mutations with a larger effect on binding affinity though. The ROC AUC for predicting whether a mutation is stabilising or not is 0.87 for mutations with an absolute value greater than 1 kcal/mol (Sirin et al., 2016), suggesting that this data does contain some of the characteristics of experimental values.

On this synthetic dataset, Graphinity achieved a test Pearson's correlation of 0.87 with 10-fold cross-validation and a 90% length-matched CDR sequence identity cutoff imposed between folds (Figure 2.2c,d). Training the model for longer (100 epochs, as opposed to 10) improved the correlation slightly, to 0.91 on a single fold, but was not explored further due to computational cost. Graphinity substantially outperformed a simple baseline for predicting $\Delta\Delta$G: the correlation between the change in number of contacts between the WT and mutant structure (4 Å interaction distance cutoff) and the synthetic $\Delta\Delta$G is 0.42, less than half the correlation the EGNN model achieves.

The performance of Graphinity was robust to train-validation-test sequence identity cutoffs (Figure 2.2b, Figure 2.4a). The most stringent split, a length-matched CDR sequence identity cutoff of 70% plus an antigen sequence identity cutoff of 70%, maintained a Pearson's correlation above 0.87. For reference, 88% of antibody therapeutics share at least 70% heavy and light chain CDR sequence identity with a natural antibody sequence (Olsen et al., 2023).

Another way to assess model performance is with the Spearman's rank correlation. This value (ca. 0.6) was lower than the Pearson's correlation. This appears to be due, in large part, to the very high density of $\Delta\Delta$G values close to 0, which the EGNN did not always rank in the correct order. The Spearman's rank correlation rose to ca. 0.7 when values between -1 and $+1$ kcal/mol were excluded. FoldX is known to be less accurate at predicting the $\Delta\Delta$G values for mutations with only a small effect on

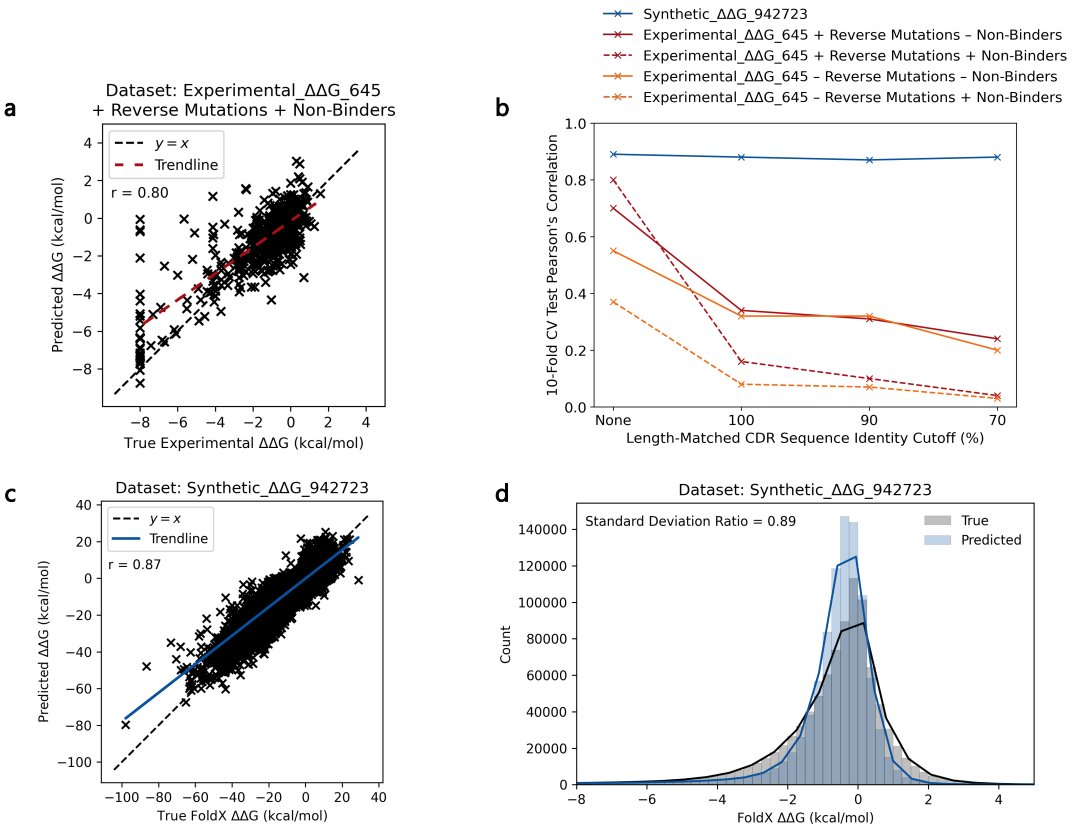

Figure 2.2: **Graphinity model performance for ΔΔG prediction.** (a) Correlation between Graphinity predictions and true values for the Experimental_ΔΔG_645 + Reverse Mutations + Non-Binders dataset. Graphinity was trained with a random train-validation-test split. Reverse mutations were used for training/validation only and were not included in the test dataset. An ensemble of 10 models was trained for 500 epochs with 10-fold cross-validation (CV) on the datasets. The trendline, shown in red, is a least squares polynomial fit. (b) The effect of introducing length-matched CDR sequence-identity cutoffs when splitting the train, validation and test data. This figure is included with error bars representing the standard deviation across the 10 folds in Appendix Figure A.2. (c) Correlation between Graphinity predictions and true values for the Synthetic_ΔΔG_942723 dataset. Graphinity was trained with a 90% length-matched CDR sequence identity cutoff applied for the train-validation-test split. An ensemble of 10 models was trained for 10 epochs with 10-fold cross-validation. The trendline, shown in blue, is a least squares polynomial fit. (d) Histograms of the true and predicted FoldX ΔΔG values (x-axis limited to -8 to +5 kcal/mol for clarity) shown in (c). The solid lines are kernel density estimates (KDEs).

binding affinity (Sirin et al., 2016) and therefore there may be less signal in the data in this region.

A recent study found that FoldX accuracy was higher for mutations that are observed naturally, with an 11% decrease in incorrectly predicting a mutation to be stabilising (Rosace et al., 2023). I explored the performance of Graphinity on a test dataset limited to such 'evolutionarily grounded' mutations, as defined by Rosace et al. (2023), from human and mouse sequences and found that the Pearson's correlation was stable at 0.89. Conversely, Graphinity also performed well (Pearson's correlation = 0.85) on non-evolutionarily grounded mutations from human and mouse sequences.

I also investigated model performance with different graph inputs – of the full interface rather than just the mutation site neighbourhood, reflecting the input for potential multi-point mutation data – and found that performance was maintained (Pearson's correlation = 0.85 on a single fold held-out test dataset, 90% length-matched CDR sequence identity cutoff).

These results serve as a proof of concept that ΔΔG can be accurately predicted when sufficient data is available.

2.5.3 Considerations for generating experimental ΔΔG datasets

Having demonstrated the potential of the EGNN architecture for predicting ΔΔG when input data is abundant, I next attempted to quantify the amount of data that will be required for the accurate prediction of experimental values. I built models with varying training plus validation dataset sizes (datasets Synthetic_ΔΔG_{580-450000}, Appendix Table A.1) and applied them to a test set of 94,126 mutations (90% length-matched CDR sequence identity cutoff). The test Pearson's correlations

only began to plateau, reaching 0.85, for models trained with at least 90,000 mutations (Figure 2.3a).

Comparing the distributions of the predicted and true values revealed that models built from smaller datasets often regressed towards the mean and achieved high correlations despite predictions not covering the full range of true values. To quantify this effect, I calculated the standard deviation ratio, the relative ratios of the standard deviations of the true and predicted values. The standard deviation ratio does not plateau at any stage and only exceeds 0.8 with a dataset size of 450,000 mutations (Figure 2.3a).

Diversity is a known important characteristic of any dataset used for model training. The role of dataset diversity was evaluated using three metrics: the diversity of antibody sequences, amino acid substitution types and structural distribution of mutations in the interface. I constructed training and validation datasets to minimise and maximise each respective metric (Synthetic_ΔΔG_100000_{sequence/substitution_type /substitution_distribution}_ {min/max}, Appendix Table A.1). For example, the Synthetic_ΔΔG_100000_sequence_min training dataset contained mutations from 75 antibody-antigen complexes, while the corresponding maximum-diversity dataset contained mutations from 1177 complexes. All models built from these datasets were evaluated on the same test data, consisting of 10,000 mutations (Appendix Table A.1). A 90% length-matched CDR sequence identity cutoff was imposed between all the training, validation and test datasets.

The distribution of mutations in the interface had only a marginal effect, which may be explained by the input graphs, which represent only the neighbourhood of the mutated site. However, sequence and substitution type diversity impacted model performance, particularly the test standard deviation ratio (Figure 2.3b). The minimum sequence and substitution type diversity datasets achieved 31% and 59% lower

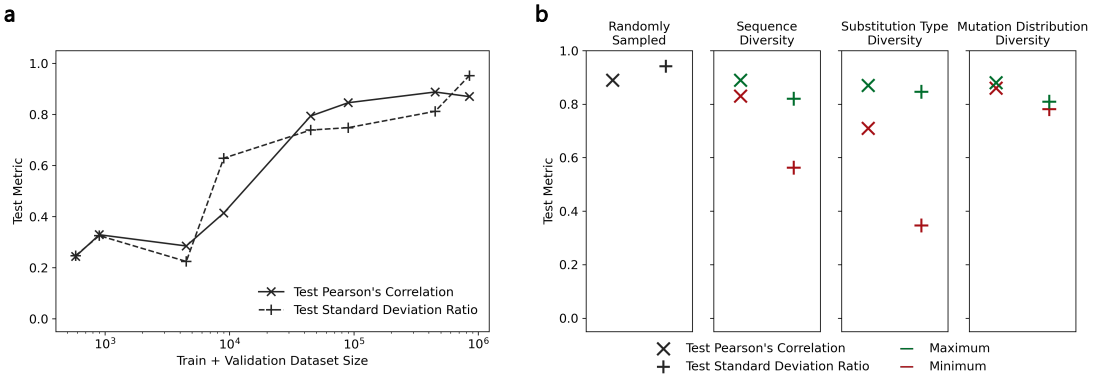

Figure 2.3: **Considerations for experimental ΔΔG dataset generation, with respect to ML predictiveness.** (a) Graphinity performance when trained with training and validation datasets of varying sizes. Datasets used: Synthetic_ΔΔG_{580-450000} (Appendix Table A.1). (b) The effect of dataset diversity (antibody CDR sequence identity, amino acid substitution type frequency and the distribution of mutated positions in the complex) on model performance. Datasets used: Synthetic_ΔΔG_100000_randomly_sampled, Synthetic_ΔΔG_100000_{sequence/substitution_type/substitution_distribution}_{min/max} (Appendix Table A.1).

standard deviation ratios than the corresponding maximum diversity datasets, respectively.

2.5.4 Graphinity is robust to noise on large synthetic ΔΔG dataset

Experimental ΔΔG data is noisy, particularly if acquired from different experimental setups and/or labs (Jankauskaite et al., 2019; Landrum and Riniker, 2024). I therefore explored the robustness of Graphinity to noise by perturbing the training and validation datasets of Synthetic_ΔΔG_942723 in two ways: (1) shuffling the affinity labels corresponding with mutations (Synthetic_ΔΔG_942723_shuffled) and (2) adding Gaussian-distributed random noise to the labels (Synthetic_ΔΔG_942723_gaussian_noise).

The Pearson's correlations on held-out test sets remained remarkably constant, at approximately 0.85 for datasets with 0-60% shuffled labels (Figure 2.4b). However,

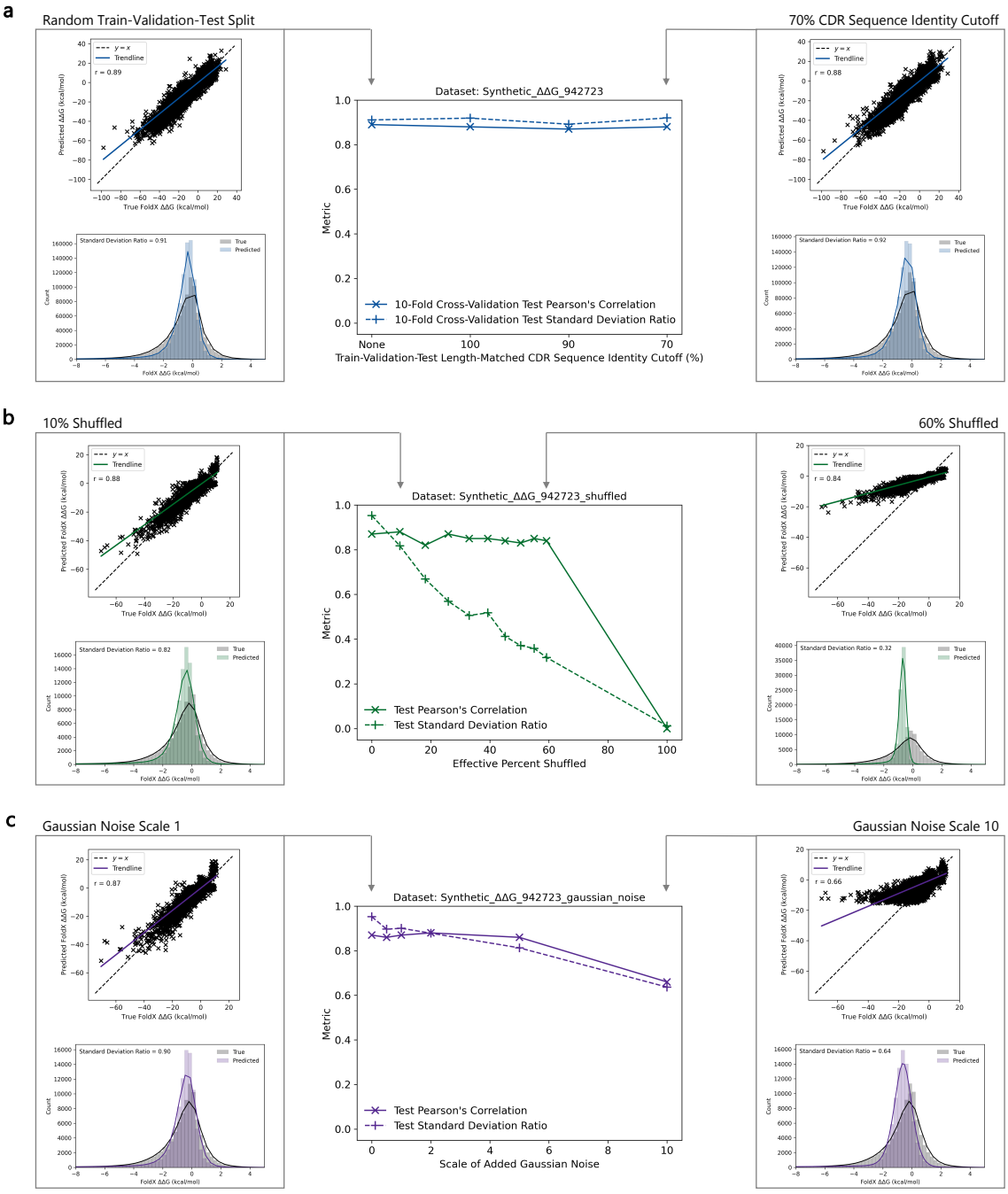

Figure 2.4: **Graphinity robustness to train-validation-test cutoffs and noise on synthetic data.** (a) Train-validation-test cutoffs: Length-matched CDR sequence identity cutoffs were applied when splitting the train, validation and test datasets. The synthetic dataset was already filtered such that no complex had more than 90% CDR sequence identity with any other complex and as such, the 100% and 90% cutoffs are functionally identical (although these were sampled from the full dataset separately). (b) Shuffling (Synthetic_ΔΔG_942723_shuffled): Noise was added by shuffling varying percentages of the labels, such that a subset of the labels

Figure 2.4: (cont.) were incorrect. (c) Gaussian noise (Synthetic_ΔΔG_942723 _gaussian_noise): Random noise sampled from a Gaussian distribution was added to the training and validation datasets. Results are shown for 10-fold cross-validation in (a) and for a single fold, held-out test set in (b-c). For all histograms, the x-axes were limited to -8 to +5 kcal/mol for clarity and the solid lines are KDEs.

the relative distributions of the predicted and true FoldX ΔΔG values revealed that the model lost predictiveness with increased shuffling: the predicted values began to fall in increasingly narrow distributions as compared to the true spread of ΔΔG (Figure 2.4b). These results underscore the importance of looking beyond the traditional evaluation metric of Pearson's correlation and also assessing the standard deviation ratio. Model performance was 0 when 100% of the labels were shuffled, supporting that, while the FoldX-generated values are not as accurate as experimental data, there is true signal that can be learned from the input complex structures.

There are 82 duplicated antibody-antigen single-point mutations with ΔΔG values in SKEMPI 2.0 (Schymkowitz et al., 2005). Across these, the average ΔΔG standard deviation between duplicates is 0.19 kcal/mol and the maximum 0.90 kcal/mol. Graphinity maintained Pearson's correlations and standard deviation ratios above 0.8 with added noise in this range, and indeed up to a Gaussian noise scale of 5 (Figure 2.4c).

2.5.5 Performance by amino acid substitution

I further investigated how Graphinity performs for specific amino acid substitutions (e.g., Arg to Lys). The FoldX ΔΔG values varied widely for a specific substitution, with standard deviations ranging from 0.5 to 10.6 kcal/mol (Figure 2.5c). The pattern of the mean ΔΔG values and corresponding standard deviations closely matched between FoldX and predicted values (Figure 2.5), suggesting the model is learning the structural context of the mutations rather than just the average value for the

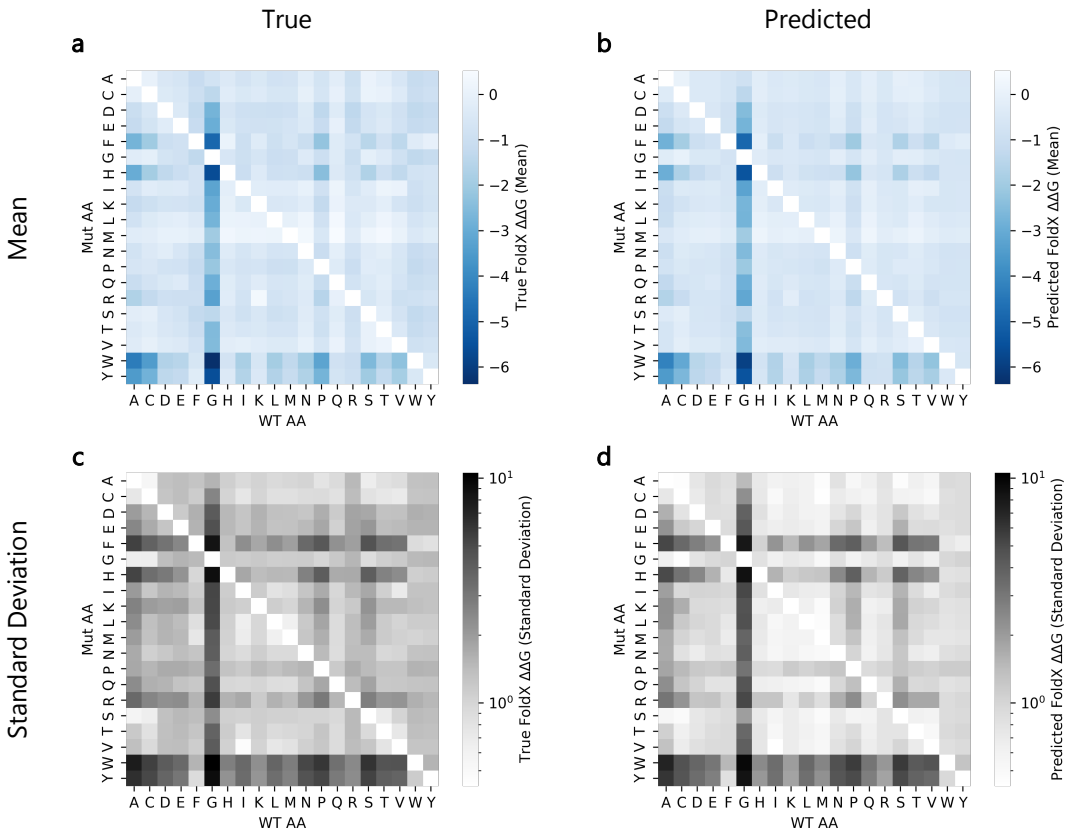

Figure 2.5: **Average of and variation in ΔΔG values.** (a,c) ΔΔG values from the Synthetic_ΔΔG_942723 dataset, separated by amino acid substitution. (b,d) ΔΔG values predicted by Graphinity applied to the Synthetic_ΔΔG_942723 dataset (10-fold cross-validation, 90% length-matched CDR sequence identity cutoff), separated by amino acid substitution. Top row: mean, bottom row: standard deviation. WT: wild-type, Mut: mutant; AA: amino acid.

mutation. If the predicted values were set as the average ΔΔG value for the specific substitution, the Pearson's correlation would be just 0.35 as compared with the trained model's performance of 0.87.

Models were also trained on datasets limited to each substitution type to explore whether Graphinity could learn the effect of a substitution better when trained only on data for this substitution. However, model performance for a specific substitution decreased as compared to the model trained on the full dataset (Appendix Figure A.4). Performance could be rescued, reaching or exceeding that of the model trained on the full dataset, by initialising with weights from the full model (Appendix Figure

A.4).

2.5.6 Validation on experimental binding dataset

To test whether Graphinity can learn the distribution of experimental data, not just FoldX predictions, the architecture was adapted and applied to a dataset of 36,391 CDRH3 variants of Trastuzumab (Mason et al., 2021). The variants are classified as binders or non-binders for the antigen, HER2. While a single-antigen task is not necessarily the intended aim of the Graphinity architecture, this dataset was sufficiently large that prediction would be expected to be successful.

Graphinity learned to separate the binding and non-binding variants, achieving a ROC AUC of 0.88 and AP of 0.77 (Figure 2.6). This performance is close to that of the sequence-based CNN reported by the authors of the study (ROC AUC = 0.91, AP = 0.83) (Mason et al., 2021). Furthermore, the performance was robust to train-validation-test cutoffs with ROC AUC values maintained above 0.83 when V- and J-gene clonotype plus CDRH3 sequence identity cutoffs (90%, 70%) were applied (Figure 2.6).

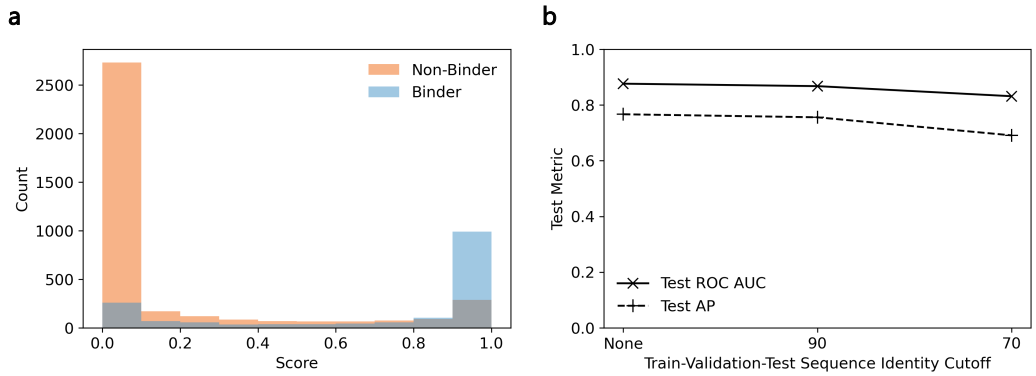

Figure 2.6: **Application of Graphinity to 36,391 Trastuzumab CDRH3 variants.** (a) Graphinity scores of binding and non-binding Trastuzumab CDRH3 variants (Mason et al., 2021) (randomly split data). (b) Model performance with clonotype and sequence identity cutoffs imposed between the train, validation and test datasets. In cases where a sequence identity cutoff (value shown on the x-axis) was applied, the data was also separated by clonotype (V- and J-gene assignments).

2.6 Discussion

Antigen binding affinity, essential to the function and efficacy of an antibody, is complex and challenging to predict computationally. Graphinity is built directly from the coordinates of antibody-antigen structures and does not rely on featurisation, which is slow and may miss information that could be important for predicting affinity. This architecture was applied to $\Delta\Delta G$ prediction on both the limited available experimental data and a large constructed synthetic dataset. Graphinity achieved state-of-the-art performance on the experimental data from the AB-Bind database (Sirin et al., 2016). However, the high correlations were the result of overtraining, as has been found across existing ML methods for $\Delta\Delta G$ prediction (Wang et al., 2020; Geng et al., 2019; Liu et al., 2021; Behbahani et al., 2022). Overtraining is also gaining increased recognition in related fields, such as protein-protein interaction prediction (Bernett et al., 2023) and molecular discovery (Crusius et al., 2024). Effective cutoffs between train and test datasets are highly dependent on the task, intended application and available data, making it challenging to set widely applicable guidelines and standards. Addressing this difficulty is however essential to achieve confidence in model performance and will require continuous discussion throughout the field, particularly between model developers and intended users.

To test whether affinity could be accurately and robustly predicted, I applied Graphinity to a synthetic dataset of nearly 1 million mutants (Schymkowitz et al., 2005). Test Pearson's correlations on this dataset neared 0.9 and the model generalised well beyond its training data, with performance being maintained with stringent sequence identity cutoffs for both antibody and antigen between the train, validation and test datasets.

Performance was also robust to levels of noise that have been observed in experimental data. Applying Graphinity to noisy data emphasised the importance of going

beyond the test Pearson's correlation when evaluating a model. A high correlation can be achieved when the model regresses towards the mean, predicting values in only a small range and with a trendline much flatter than $y = x$. The standard deviation ratio, a metric comparing the relative distributions of the true and predicted values, exposes poor predictiveness by identifying when predicted values cover only a fraction of the true $\Delta\Delta$G distribution.

The results on the synthetic data must be considered in light of the source of the data. The synthetic data points were all produced by the same software and are thus expected to be more self-consistent and less noisy than experimental data. The synthetic values may also follow a different distribution to the true values. However, FoldX can accurately predict whether mutations with a substantial effect on binding affinity will be stabilising or destabilising, suggesting there is signal in this dataset (Sirin et al., 2016).

To test if the Graphinity architecture can also learn the distribution of experimental data, not just the FoldX forcefield, it was evaluated on an experimental dataset of 36,391 Trastuzumab variants. Graphinity separated binding from non-binding variants with a ROC AUC similar to that achieved by a CNN trained on the variant sequences. The EGNN architecture offers further benefits over the CNN, most notably the potential for generalisability to different antibody-antigen complexes. This application also highlights the modularity of the architecture: Graphinity can be applied for regression and classification, single- and multi-point mutations, as well as affinity and change in affinity prediction.

The success of the EGNN model on large datasets lends support to the idea that the major challenge with experimental $\Delta\Delta$G prediction lies in the availability of experimental data rather than the model architecture. I explored the amount of data that would be required for accurate and generalisable prediction of experimental $\Delta\Delta$G

using the synthetic dataset. The results suggest that there is currently vastly insufficient data available and orders of magnitude more, tens to hundreds of thousands of data points, will likely be needed. These predictions may even be underestimates, as FoldX, and computational methods more generally, are unable to accurately account for entropic effects and the impacts of mutations on the protein backbone structure and folding process. Model development may be able to achieve greater success on smaller datasets by focusing on the search space where entropic effects and folding perturbations are minimal. Additionally, there is potential for limitations in data to be compensated for, to some extent, by machine learning know-how such as by identifying model architectures that require less data, using stratified sampling or transfer learning from a related data-rich task or from synthetic data. Future model design could also be augmented by considering physiological features that are typically ignored in current methods, such as water molecules and protein conformational flexibility.

In addition to dataset size, the results underscore the importance of dataset diversity, particularly with respect to antibody sequence identity and amino acid substitution type. Both of these diversity metrics are currently very limited in experimental data. For example, the antibody-antigen single-point mutations in SKEMPI 2.0 (Jankauskaite et al., 2019) derive from less than 50 complexes and are highly skewed in substitution type, with mutations to alanine making up over half of the dataset.

These results highlight the need to move towards 'machine learning-grade data', where model development is considered in the data generation process.

Large datasets generated for ML, and accurate ML models trained on these, will allow us to shift our focus from initial prediction to better understanding the factors that contribute to affinity. In the next chapter, I explore the interpretability of the Graphinity architecture and how protein interface components are weighted for

prediction.

Chapter 3

Assessing the Interpretability of Deep Learning for Antibody-Antigen Binding Affinity Prediction

Contents

3.1 Motivation

Neural network deep learning models can be 'black boxes', with limited to no understanding of the factors that contribute to a particular outcome or prediction. In this

chapter, I aim to shine light into this black box by exploring the weighting of nodes and edges in the Graphinity architecture (described in Chapter 2) when applied to a large experimental affinity dataset.

I trained Graphinity on >500,000 Trastuzumab CDRH3 variants with associated binding affinity labels for the target HER-2 (experimental data from the lab of Victor Greiff, University of Oslo). The model achieved near-perfect performance in separating high- from low- and medium-affinity variants.

To investigate the contributions of model components (edges and nodes, representing interactions and atoms, respectively), I implemented GNNExplainer (Ying et al., 2019), as well as attention multi-layer perceptrons (MLPs) within the EGNN architecture.

3.2 Contributions

This chapter contains material reproduced from:

Chinery, L.*, **Hummer, A.M.***, Mehta, B.B.*, Akbar, R., Rawat, P., Slabodkin, A., Le Quy, K., Lund-Johansen, F., Greiff, V., Jeliazkov, J.R. and Deane, C.M. (2024) Baselining the Buzz: Trastuzumab-HER2 Affinity, and Beyond. *bioRxiv*.

I carried out all of the model development and analysis presented in this chapter unless otherwise stated. The experimental Trastuzumab dataset and description of its generation were prepared by the lab of Victor Greiff, University of Oslo (led by Brij Bhushan Mehta). Lewis Chinery processed and prepared the resulting dataset for ML. The EGNN edge and node attention approach was adapted from Satorras et al. (2021) and PointVS (Scantlebury et al., 2023).

3.3 Introduction

ML models tend to suffer from a trade-off between accuracy and interpretability. Simpler models, such as decision trees, can more easily be probed to identify the importance of features for prediction, whilst more complex NNs often achieve higher accuracy, but are harder to explain. Interpretability can give greater confidence in model predictions, as well as provide insights into what models are learning – and therefore, for example, reveal model biases or information that can be used to better understand the problem.

A number of techniques have been generated for NN interpretability. These can be broadly split into two categories. One set of approaches approximates the prediction space or extracts rules from the NN using a separate, 'surrogate' (often decision tree-based) model (Schmitz et al., 1999; Gethsiyal Augasta and Kathirvalavakumar, 2012; Ribeiro et al., 2016; Lakkaraju et al., 2017). Other methods investigate the relevant features of the NN model, for example by backpropagating neuron contributions to input features (Shrikumar et al., 2017), based on Shapley values (Lundberg and Lee, 2017), proposing counterfactual explanations (Kang et al., 2019) or maximising mutual information (Chen et al., 2018).

Strategies designed specifically for graph neural networks (GNNs) have also been developed in recent years (e.g., Ying et al., 2019; Huang et al., 2020; Luo et al., 2020; Vu and Thai, 2020; Yuan et al., 2020; Schlichtkrull et al., 2020; Duval and Malliaros, 2021; Yuan et al., 2021; Lucic et al., 2021; Mastropietro et al., 2022; Zhang et al., 2022). These approaches vary in their applicability to prediction tasks: many are focused on node rather than graph classification, some require specific GNN architectures and others can be applied post-hoc to nearly any GNN. Attention is one commonly used approach which has been implemented directly within GNN architectures, including graph attention networks (Veličković et al., 2017) and learned

attention weightings of edges (Satorras et al., 2021; Scantlebury et al., 2023). GNNExplainer, used in this chapter, is widely applicable and derives importance via computing the change in the probability of the prediction for subgraphs and maximising the mutual information (Ying et al., 2019).

GNN interpretability approaches have been applied to molecular tasks, although the field has not yet reached a consensus on the best methods. For example, attention-based attribution has been used to interpret small molecule-protein affinity prediction (Scantlebury et al., 2023; Hadfield et al., 2023), a Myerson value approach has been developed for small molecule graph prediction (Homberg et al., 2023), gradient-weighted class activation mapping has been implemented for kinase functional state prediction (Ravichandran et al., 2024) and specific interpretable model architectures have been designed for protein binding site prediction (Tubiana et al., 2022) and protein interactions (Jha et al., 2022).

Here, I explore the interpretability of EGNN binding predictions of the Trastuzumab-HER2 complex. Trastuzumab (brand name Herceptin) is a therapeutic antibody, which targets HER2 (human epidermal growth factor receptor 2) and is used to treat breast and stomach cancers. Every CDR loop in both the VH and VL contributes to the interface (Figure 3.1). This antibody has been used in multiple experimental studies exploring the binding of CDRH3 variants (Mason et al., 2021; Shanehsazzadeh et al., 2024; Chinery et al., 2024). The Victor Greiff lab generated a new dataset of >500,000 Trastuzumab CDRH3 variants with high/medium/low-affinity labels. This large dataset provided an opportunity to investigate the interpretability of the Graphinity architecture. I trained Graphinity on this dataset, achieving strong performance, and applied the GNNexplainer and attention MLP approaches to quantify the contributions of graph components to the predictions.

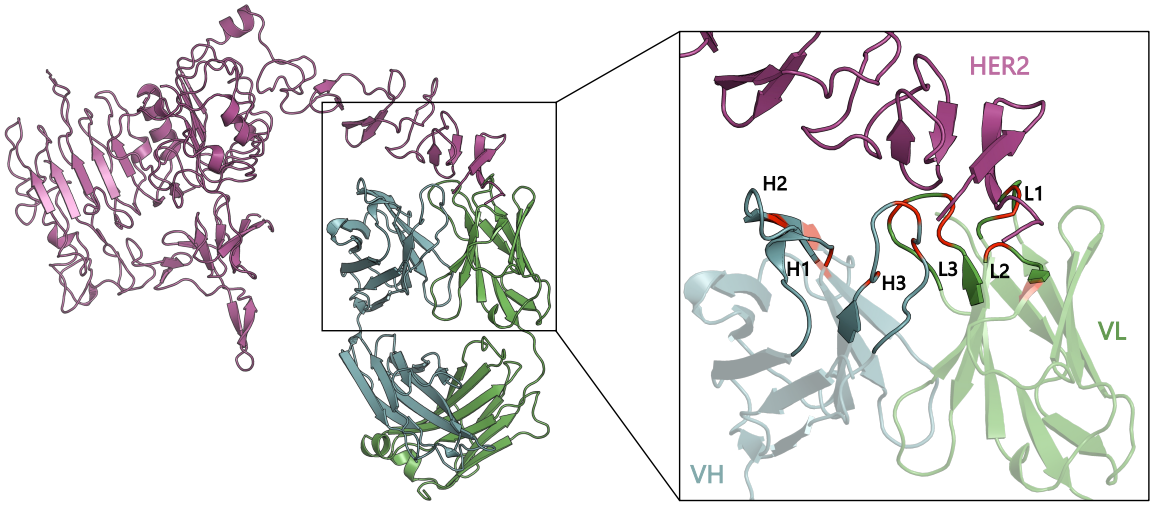

Figure 3.1: **Structure of the Trastuzumab-HER2 interface.** The CDR loops are shown with no transparency. Antibody interface residues (have atom(s) within 4 Å of the antigen) are shown in red. The Trastuzumab heavy and light chains are shown in blue and green, respectively; HER2 is shown in purple. PDB 1N8Z (Cho et al., 2003). The CDR loops (H1-3, L1-3) are labelled.

3.4 Methods

3.4.1 Trastuzumab dataset

3.4.1.1 Experimental data generation

The Trastuzumab scFv CDRH3 dataset used in this chapter was guided by the site-specific Deep Mutational Scanning results generated previously by Mason et al. (2021). The new dataset, referred to as HER2-aff-large, was generated by the Victor Greiff lab (led by Brij Bhushan Mehta) and they also provided the methods description given below.

Briefly, a Trastuzumab scFv antibody library was cloned in a pSYD yeast display vector, a variant of the pDNL6 yeast display vector (pSYD uses N-terminal fusion for scFv-aga2 display, while pDNL6 uses a C-terminal fusion of aga2-scFv). The Trastuzumab scFv antibody library cloned in pSYD vector was transformed in

EBY100 yeast cells (ATCC #MYA-4941DQ) selected on SD + CAA plates (2% dextrose, 0.67% yeast nitrogen base, and 0.5% casamino acids yeast selection media) at 30°C for 48-72 hours. Yeast display analysis of the Trastuzumab scFv library was performed as described previously by Ferrara et al. (2012) and Chao et al. (2006).

The next day, the cell pellet was resuspended in SG + CAA (containing 2% galactose and 0.1% dextrose) at 0.5 OD/ml and incubated at 20°C with shaking for one to two doublings, as determined by OD. The cells were washed with the wash buffer and processed for staining to check HER2 binding. Around 1-10 $\times$ 10^7 cells were labelled with 100μg/ml anti-V5 tag antibody followed by the addition of 100nM HER2 and incubated for 30 minutes on ice. The cells were then washed twice more with wash buffer and labelled with a 1:200 dilution of secondary reagents (goat anti-mouse - Alexa 488 and streptavidin-PE).

Finally, the cells were incubated for 30 minutes on ice, washed twice with a wash buffer, and resuspended in 1mL of sorting buffer. To determine their affinity, the cells were sorted for the brightest V5 FITC positive (scFv expression) antigen binding population (PE positive) and labelled as high-affinity binders (Appendix Figure B.1). The cells were further sorted for the brightest V5 FITC positive medium and low-affinity antigen binding populations. The populations were sorted into tubes containing YPD media and grown in SD + CAA liquid media at 30°C with shaking overnight as described previously (Ferrara et al., 2012).

Plasmid DNA was isolated using a yeast plasmid isolation kit (Zymoprep Yeast Plasmid Miniprep I #D2100) following user protocol. The VH gene containing the CDRH3 sequence for each population was PCR amplified using in-house NGS-specific primers. The amplicons were PCR-cleaned and prepared for NGS. The DNA libraries were sequenced on Illumina using NovaSeq 6000 S2 Reagent Kit v1.5 (300 cycles) and the raw data has been deposited on Zenodo - `doi.org/10.5281/zenodo.10549115`.

The primers used for generating variable heavy amplicons were:

NGSVH Fwd: 5´CACCCGTTATGCCGACAG3´

NGSVH Rev: 5´GGGATTGGTTTGCCGCTAG3´

The raw paired NGS reads were merged using PEAR (v0.9.6). The subsequent dataset consisted of 618,585, 799,368, and 663,397 high, medium and low-affinity unique CDRH3 sequences, respectively. Singleton (count=1) sequences were removed from the dataset to improve the quality of the data. The final Trastuzumab variant dataset comprised 178,160, 196,392, and 171,732 sequences in 'high', 'medium', and 'low' affinity binder classes, respectively. The heavy and light chain sequences (from PDB 1N8Z (Cho et al., 2003)) were numbered according to the IMGT scheme.

Heavy chain insertion start and stop positions are 107 and 116, respectively.

Heavy chain sequence:

EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVKGR

FTISADTSKNTAYLQMNSLRAEDTAVYYCSRWGGDGFYAMDYWGQGTLVTVSSA

Light chain sequence:

DIQMTQSPSSLSASVGDRVTITCRASQDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGSRS

GTDFTLTISSLQPEDFATYYCQQHYTTPPTFGQGTKVEIKR

3.4.1.2 Dataset preparation for ML

The 'medium' and 'low' binding affinity classes of the HER2-aff-large dataset clustered with the negative binders from the Mason et al. (2021) Trastuzumab variant dataset using tSNE visualisations (Appendix Figure B.2). Consistent with this observation, and the motivation to achieve high-affinity binders in antibody optimisation,

'high' affinity sequences were assigned positive labels and 'medium' and 'low' affinity sequences were grouped together as negative binders.

There was a small amount of overlap in the sequences between the different classes. Removing any overlapping sequence, with more than one label, resulted in a dataset of 524,346 sequences and a class imbalance of 32.8% (more non-binders than binders).

A train-validation-test dataset size ratio of 70/15/15 was used as in Mason et al. (2021). The data was split data by clonotype, with sequences clustered according to their V and J genes (as annotated by ANARCI (Dunbar and Deane, 2016)), and by 70% sequence identity across the CDRH3. All HER2-aff-large sequences share the same V-gene (IGHV3-66) and one of two J-genes (IGHJ4 or IGHJ1). All members of a clonotype were added to the same train, validation or test set. Train, validation and test sets have the same class imbalances (i.e., the ratio of binders to non-binders). The data processing and splitting were conducted by Lewis Chinery.

Due to the computational resources required, I applied the interpretability analyses to a subset of the test dataset achieved via CDRH3 sequence identity- and clonotype-based clustering (70% sequence identity cutoff for initial analyses, n=706; and 90% for a more in-depth subset, n=22945).

3.4.2 Graphinity: equivariant graph neural network architecture

The Graphinity EGNN architecture (see Section 2.4.2) was adapted for application to the Trastuzumab dataset. The model input was changed to be residue-level graphs of the Trastuzumab-HER2 complex. These graphs include the C_α atoms of the 10 mutated CDRH3 residues and surrounding neighborhood (antibody C_α atoms within 10 Å of CDRH3 C_α atoms (antibody neighborhood), antigen C_α atoms within 10 Å of the antibody neighborhood and antigen C_α atoms within 10 Å of these antigen

atoms). The node features were a one-hot encoded vector describing the residue type and chain type (antibody or antigen). The edge features were a one-hot encoded vector describing whether the edge is intra-binding partner, i.e., between atoms on the same binding partner, or inter-binding partner, i.e., between atoms on different binding partners.

The graphs were fed through a network composed of three E(n) EGC layers (Satorras et al., 2021) with a hidden dimension of 128. The models were trained with Binary Cross-Entropy with Logits loss. The architecture was implemented using PyTorch and PyTorch Geometric. The models were trained for 10 epochs, with a training time of ca. 5.5 hrs.

As in Section 2.4.4, FoldX BuildModel (Schymkowitz et al., 2005) was used to generate the structural inputs for the EGNN. Mutations were introduced to the Trastuzumab CDRH3, starting from a FoldX-'repaired' structure in complex with HER2 (PDB 1N8Z (Cho et al., 2003)). As FoldX does not model changes to the backbone (Van Durme et al., 2011), the true structural effects of the mutations are unlikely to be represented. However, this approach has the advantages of speed (and therefore compatibility with high-throughput datasets) and avoiding the need for docking by starting from the structure of a bound complex.

3.4.3 Edge and node weighting

3.4.3.1 GNNExplainer

The PyTorch Geometric (version 2.5.2) implementation of GNNExplainer was used. The edge_mask_type and node_mask_type parameters were set to 'object', such that each edge and feature are masked, respectively. The 'explanation_type' parameter was set to model, 'task_level' to graph, 'model_mode' to binary classification and 'result_type' to raw.

3.4.3.2 Edge and node weighting with attention multi-layer pereceptron

I implemented an MLP within the EGNN architecture to learn the weighting of inputs to the model – edges and nodes, representing interactions and atoms, respectively. These edge and node weights are multiplied by the respective feature embeddings during model training and thus reflect the contribution of the embeddings toward the output score. For example, a weight of zero for an edge would mean that the edge is effectively ignored or removed from the graph, while a weight of one would mean the embedding for the edge is unchanged.

I assessed a range of MLP architectures differing in the numbers of linear layers (1-3) and non-linear activation functions (ReLU, Sigmoid, SiLU, Scatter Softmax, TanH; Appendix Figure B.3), implemented using PyTorch and PyTorch Geometric. The term 'attention MLP' is used for consistency with the Satorras et al. (2021) EGNN code and also refers to 'MLPs' with fewer than three layers. The parameters of the EGNN architecture were kept the same as in Sections 2.4.2 and 3.4.2, with the exception of dropout, which was set to 0.

The edges in each graph are bidirectional and thus duplicated (e.g., for source = A, destination = B, edges A-B and B-A will be present). Within a graph, I took the maximum value for each edge (i.e. maximum of A-B and B-A). I collated the edge values from multiple graphs by averaging over matching edges.

Each Trastuzumab interface graph input is constructed from 90 nodes and 676 edges, centred around the CDRH3 (see Section 3.4.2). The residue-level graph structure is the same for all inputs and allows for direct comparison between and averaging across graphs of edge/node weights.

3.5 Results

3.5.1 Graphinity accurately separates high- from medium/low-affinity binders

Graphinity achieved near-perfect accuracy (ROC AUC = 0.98 and PR AUC = 0.97) in separating high- from medium/low-affinity binders (Figure 3.2), consistent with the results of Section 2.5.6 and the larger training dataset used here.

3.5.2 Interpretability of Trastuzumab graphs

I implemented two different approaches to explore the factors contributing to the EGNN model's predictions. GNNExplainer quantifies the importance of model components by masking them and examining the effect on the prediction (Ying et al., 2019); it is applied post-hoc to a trained model. The attention MLP is an additional trainable element within the EGNN architecture, with the model learning weightings to apply to edges and nodes, respectively. I will refer to the outputs of both approaches as edge/node 'weights', which are metrics of the importance of these components for the model's predictions.

3.5.2.1 GNNExplainer

GNNExplainer output edge and node weights in a relatively narrow range (0.4-0.7), with not much differentiation between the most and least important graph components. Furthermore, there was little consistency in the ranking of the edges and nodes. Every edge and every node was observed in the top 10% of respective scores in at least one graph of the 706 in the clustered test dataset (see Section 3.4.1.2, 70% sequence identity cutoff). The edges and nodes most frequently observed in the top 10% are found in this category in less than 15% of the graphs, indicating that there are no

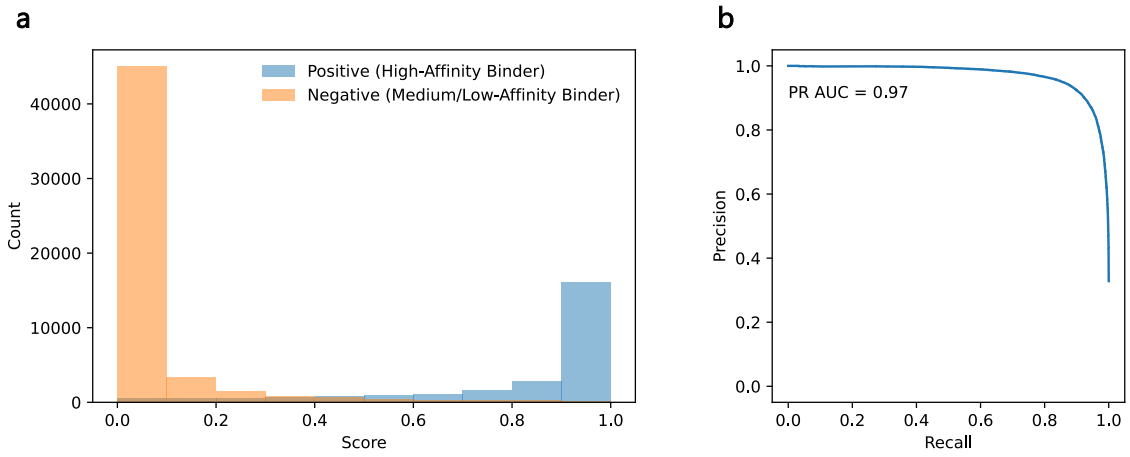

Figure 3.2: **Application of Graphinity to 524,346 Trastuzumab CDRH3 variants.** (a) Graphinity scores of positive (high-affinity) and negative (medium/low-affinity) Trastuzumab CDRH3 variants (HER2-aff-large dataset, train-validation-test split with a 70% CDRH3 sequence identity cutoff and by clonotype). (b) Precision recall curve for the corresponding predictions.

consistently important or dominant components. Of these edges, the vast majority (>90%) are intra-protein, counter to what one might expect for key interactions in a protein-protein complex.

3.5.2.2 Attention MLP

I tested multiple attention MLP architectures with a different number of layers and activation functions. These approaches all gave different weightings to the edges and nodes.

Edge weights

Most of the MLP architectures did not result in easily identifiable outputs for the edge weights. Many produced weights that provided only limited differentiation between the edges (e.g., Sigmoid 1-3 layers; SiLU 1-3 layers; TanH 1,3 layers). The ReLU 3-layer MLP assigned weights of 0 to every edge and concordantly resulted in a drop in accuracy (PR AUC down to 0.78). In MLP cases where the edge weights differed more substantially, the ranking was unintuitive, with intra-protein edges scored most

highly.

The only architecture which achieved more evidently explainable edge weighting was the Scatter Softmax activation function applied after 1 linear layer. This was also the approach employed in PointVS for small molecule-protein binding affinity prediction attribution (Scantlebury et al., 2023). The Scatter Softmax activation function applies a softmax to all edges for each node, respectively. As such, the edge weights sum to 1 for every node (for example, if one node were to have two edges, possible edge weightings could be 0 and 1 or 0.5 and 0.5; the sum across all edge weightings will equal the number of nodes, 90). The edge weightings from the Scatter Softmax MLP were unstable, with weights typically equal to 0, 0.5 or 1, rather than a more continuous distribution. This indicates that only one or two edges from each node, typically, will be considered in the graph. However, the ranking was more consistent with expectations, with inter-protein edges weighted highest. Subsequent analysis was continued for edge weights derived using the Scatter Softmax activation function applied to a larger subset of the test data clustered with a 90% CDRH3 sequence identity cutoff (see Section 3.4.1.2).

All edges with an average weighting > 0.5 (n = 8) were inter-protein (Figure 3.3, solid lines). Of these, 5 edges are between HER2 and the Trastuzumab VL, and 3 with the VH. Although the CDRH3 is the only variable component in this dataset, the Trastuzumab light chain forms a substantial part of the interface with HER2 (Figure 3.1). In the subsection of the WT interface complex included in the graph input, there is one inter-protein H-bond (between HER2 and the CDRH3; Figure 3.3, dashed line). While this edge appears to be ignored by the model (average weight of 0), there is a highly weighted edge directly adjacent, which could play a role in structuring the interface to maintain the H-bond.

I also tested whether the edge weights were simply a proxy for the distance of the

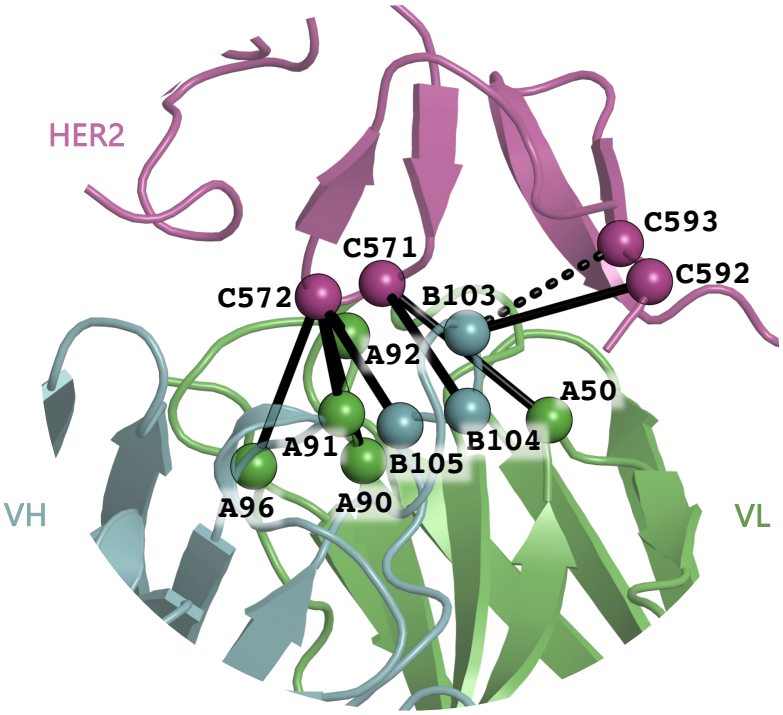

Figure 3.3: **EGNN edge weighting of Trastuzumab-HER2 interface.** The highly weighted edges (average weight > 0.5, n=8), obtained using an attention MLP of one linear layer followed by the Scatter Softmax activation function, are shown as solid black lines. The inter-protein H-bond is indicated with a dashed black line. The Trastuzumab heavy and light chains are shown in blue and green, respectively; HER2 is shown in purple. PDB 1N8Z (Cho et al., 2003).

edges but found this not to be the case (Pearson's correlation between edge weight and distance = 0.02). However, the weightings appear to represent what the model has learned about the interface structure, rather than differentiate between factors which could be used for design. I grouped the Trastuzumab variants into categories based on the model scores (high- vs. low-scoring binders, at a range of thresholds; and by classification category: true positive, false positive, true negative, false negative) and compared the edge weights using a Wilcoxon signed-rank test, with matching edges paired between different groupings. With the exception of true negatives vs. false positives ($p = 0.047$), there is no statistically significant difference ($p > 0.05$) between the edge weightings for groupings (Appendix Tables B.1, B.2). Additionally, for example, in the 10 highest- and 10 lowest-scored graphs, all but one of the top-

scored edges (average score of 1) are identical.

Node weights

As seen for edge attention, many architectures resulted in node weightings with limited differentiation (e.g., ReLU 2-3 layers, Sigmoid 2-3 layers, TanH 1-3 layers). For the remaining architectures (ReLU 1 layer, Sigmoid 1 layer, SiLU 1-3 layers), it is unclear which is most suitable for further analysis. There is little consensus in the highest scoring nodes between different MLPs and little overlap ($\leq 1/3$ by chain) between the highest scoring nodes and the residues that are part of or within 4 Å of the CDRH3.

3.6 Discussion

In this chapter, I implemented strategies to investigate the importance of graph components for affinity predictions. The aim was to generate outputs that provide interpretable information on the interactions and atoms in the binding site that contribute most strongly towards the model's predictions. This information could then, theoretically, be used to inform future antibody design, as well as to glean insights into what the NN model is learning and its relationship with physics/chemistry.

However, my results demonstrate that it is still challenging to identify what GNNs applied to antibody-antigen binding affinity prediction are weighting strongly. Each method I implemented (GNNExplainer, as well as attention MLPs with different architectures and activation functions) resulted in different interpretations of model weighting. As such, there is no clear consensus on the approach to trust. This analysis is further complicated by our own expectations (that the model will learn the most chemically and physically relevant components, such as inter-protein interactions or H-bonds) and we may miss out on the actual features the model is extracting, which

could potentially provide new information about protein-protein interactions or reveal biases intrinsic to the model.

In the case of the Graphinity prediction of Trastuzumab variant binding affinity for HER2, it is unclear if the model is not learning meaningful features or if the interpretability strategies are not appropriate for this task. The Trastuzumab-HER2 complex was used here because it is the only antibody-antigen complex for which hundreds of thousands (or even tens of thousands) of binding data points are available. However, there are aspects of this task which may complicate interpretability investigations. The graphs included in this study were limited to the neighbourhood around the CDRH3, the only part of the antibody that varies in the dataset, and thus perhaps is already constrained to the most important components. Additionally, the FoldX-modelled inputs are unlikely to capture the true structural differences between variants, as only side chain changes are modelled. HER2-binding for CDRH3 variants of Trastuzumab also introduces limitations, as the majority of interactions between Trastuzumab and HER2 are mediated by non-CDRH3 regions of Trastuzumab.

Future work could explore different graph structures (perhaps extending to the full interface), graph inputs (such as atomistic inputs, although this will make comparison between different graphs more challenging) and, when available, binding datasets for other antibody-antigen complexes. There are also non-GNN interpretability methods, which could be implemented for comparison. While sequence-based CNN saliency analyses (completed by Lewis Chinery) also did not yield clear insights for predictions applied to this Trastuzumab dataset, other strategies that could be investigated include a hierarchical Bayesian model-based approach (Tonner et al., 2022) and SHAP values to calculate feature importance (Lundberg and Lee, 2017). Additionally, the interpretability of predictions for other antibody properties, beyond affinity, should be assessed to obtain a broader understanding of what ML models are learning.

Chapter 4

Humanization of Antibodies Using a Machine Learning Approach on Large-Scale Repertoire Data

Contents

4.1 Motivation

Antibody-antigen binding affinity is only one consideration in the complex, multi-parameter antibody development task. There are numerous additional properties affecting manufacturability and safety (outlined in Section 1.6.4). If an antibody therapeutic is recognised as foreign, the patient's immune system can mount an immune response against the therapeutic. Humanization may be required to reduce immunogenicity and remains a bottleneck in antibody development.

We developed ML classifiers that can discriminate human from non-human sequences with near-perfect accuracy. Building on these models, we created a computational humanization method to improve the humanness, and thereby reduce the immunogenicity, of an antibody in a more rational and efficient manner.

4.2 Contributions

This chapter contains material reproduced from:

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

The main method development for Hu-mAb was done by Dr Claire Marks and Mark Chin before I joined the Oxford Protein Informatics Group. I conducted analyses on model performance and updated the results for the revisions and final publication.

I continued this work to explore the interpretability of humanness classifications

and, in collaboration with Ashley Wong, to extend the applicability of Hu-mAb to VHH antibody formats.

4.3 Introduction

As discussed in Chapter 1, therapeutic antibodies can be derived from convalescent human patients (e.g., Bullen et al., 2021; Raybould et al., 2021a,b; Wu et al., 2020)), animals with humanized immune systems (e.g., Mendez et al., 1991), human sequence display libraries (e.g., de Bruin et al., 1999) and non-humanized animals (e.g., Köhler and Milstein, 1975b). Despite the former three methods yielding human antibodies, approximately half of the antibodies currently in development are still obtained from animals (Raybould et al., 2020). Patients can mount an immune response against therapeutic antibodies if peptides derived from the therapeutic presented on the major histocompatibility complex (MHC) are recognised as foreign (Roche and Furuta, 2015; Sekiguchi et al., 2018). Such an immune response can have detrimental impacts on both the safety and efficacy of a therapeutic, for example by leading to the development of neutralising ADAs (Gunn et al., 2016).

Various approaches have been developed to reduce the risk of immunogenicity, including by making an antibody more 'human'. For example, chimeric antibodies can be created by combining a non-human Fv, which mediates binding to the desired antigen, with human constant domains (Morrison et al., 1984) (Figure 4.1). In another technique, antibodies are humanized by grafting the non-human CDR loops onto a human antibody (Jones et al., 1986) (Figure 4.1). However, chimerisation retains a substantial portion of the non-human antibody, and humanization risks impacting the antibody structure and function, including binding properties. For example, framework residues have been known to affect the structure of CDR loops (Foote and Winter, 1992; Kettleborough et al., 1991). To address these unintended effects, back-

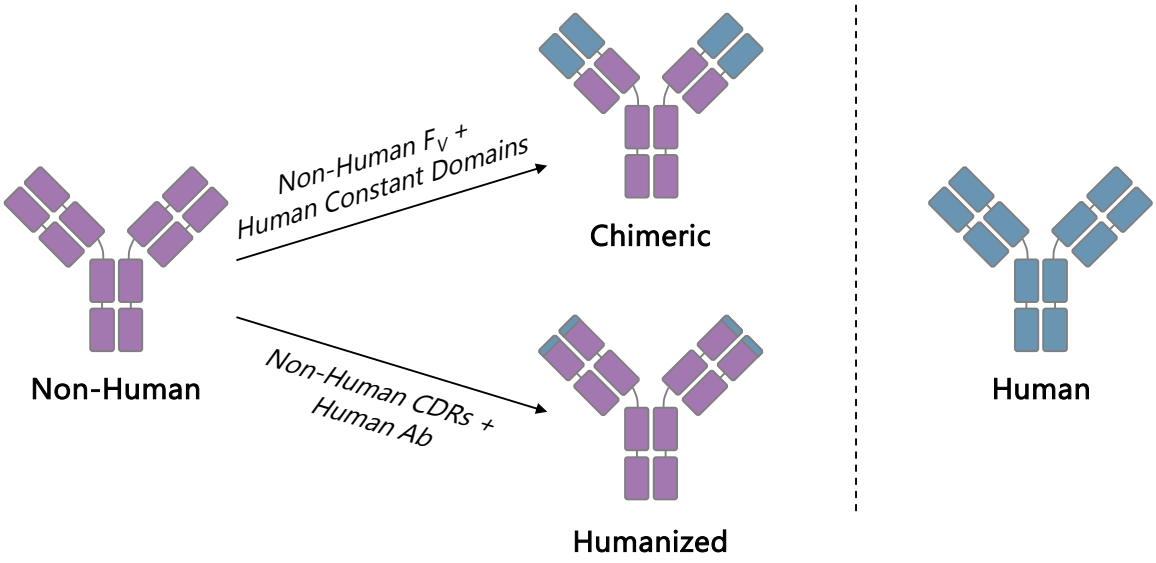

Figure 4.1: **Experimental humanization techniques.** Experimental methods for antibody humanization include chimerisation (top), in which non-human Fv regions are combined with human constant domains, and humanization (bottom), in which non-human CDRs are grafted onto a human antibody (Ab). The non-human antibody portions are shown in purple and the human antibody portions in blue.

mutations to the original, non-human residues may be necessary. These mutations are often made in an arbitrary and trial-and-error manner.

Computational approaches to evaluate and increase antibody humanness have been developed with the aim of facilitating systematic humanization. These approaches have evolved from scoring antibody humanness based only on sequence similarity with existing human sequence(s) (Abhinandan and Martin, 2007; Gao et al., 2013; Pelat et al., 2008; Thullier et al., 2010) to considering more complex relationships between positions (Choi et al., 2015; Clavero-Álvarez et al., 2018; Seeliger, 2013; Wollacott et al., 2019). To be effective for therapeutic applications, computational models must be able to distinguish human from non-human sequences, as well as predict immunogenicity. A Multivariate Gaussian (MG) model was able to achieve good accuracy in classifying human and mouse sequences but the output score correlated only weakly with experimentally determined immunogenicity (Clavero-Álvarez et al.,

2018). Long short-term memory (LSTM) network models outperformed previous methods, including the MG model, in identifying the species of sequences, but had only marginally higher performance in predicting immunogenicity (Wollacott et al., 2019).

Since the development of Hu-mAb, presented in this chapter, multiple other methods, adopting a range of strategies, have been released. Prihoda et al. took a simpler approach for scoring humanness, inspired by the biology of immunogenicity: BioPhi OASis scores parts of antibody sequences (9-residue, or 9mer, peptides) based on the frequency of the 9mer in human antibody repertoire sequences (Prihoda et al., 2022). The authors created a separate humanization pipeline using a Transformer-based ML model. AbNativ (Ramon et al., 2024), a vector-quantized variational auto-encoder, is trained on a large corpus of human antibody sequences from OAS (Kovaltsuk et al., 2018; Olsen et al., 2022a). The model decoder returns a reconstructed, humanized, version of an input sequence and quantifies the humanness score based on the difference between these two sequences. The SelfPAD approach, in contrast, does not train on large sequence datasets derived from OAS but instead applies a self-supervised contrastive learning strategy to a smaller number of antibody sequences (<300,000) with noisy labels extracted from patents (Ucar et al., 2024). Another method, CUMAb, is inspired by traditional humanization and is built on structural modelling and energy-based ranking. CUMAb computationally grafts CDR loops onto human acceptor frameworks and ranks the resulting constructs based on Rosetta (Leaver-Fay et al., 2011) energy (Tennenhouse et al., 2023). CUMAb cannot, however, be used for scoring antibody humanness or immunogenicity. Experimental validation demonstrated that many AbNativ and CUMAb constructs retained favourable properties including binding, stability and expression.

These methods represent important advances in the field but suffer from limi-

tations. Most notably, all ML approaches released before Hu-mAb were trained on limited numbers of sequences (on the order of thousands to tens of thousands at most). Additionally, none of these were trained on non-human sequences; instead, most models were only trained on human sequences and, in some cases, separate models were generated for different species. Recent methods, including BioPhi, AbNativ and SelfPAD do not consider V gene type in humanization. This risks capturing a blurred sequence representation and humanizing a sequence towards an unphysiological mixture of different V gene types.

Leveraging the extensive OAS sequence data, we developed Random Forest (RF) classifiers that accurately distinguish between each human V gene and non-human variable domain sequences. The humanness scores produced by our RF classifiers negatively correlated with observed immunogenicity levels. We used these models to build Hu-mAb, a computational tool that can systematically humanize VH and VL sequences of interest by suggesting mutations that increase humanness. Hu-mAb humanizes the sequence in an optimal manner, minimizing the number of mutations made to the sequence to limit the impact on efficacy. The mutations made by our humanizer were very similar to those made in experimental therapeutic humanization studies that produced sequences with low immunogenicity. Hu-mAb offers a powerful alternative to time-consuming, trial-and-error-based approaches to reducing immunogenicity.

4.4 Methods

4.4.1 Development of Hu-mAb Random Forest models and humanization protocol

4.4.1.1 Data collection and preparation

The IgG VH and VL antibody sequences in the OAS database (Kovaltsuk et al., 2018; Olsen et al., 2022a) were downloaded by Mark Chin (August 2020). Redundant sequences, as well as sequences missing the conserved Cys residues (at positions 23 and 104) and framework 1 residues, were filtered out. The final dataset was comprised of >65 million sequences, which were separated into human (positive, split by V gene type, Table 4.1) and non-human (negative, Table 4.2) sequences. Different classifiers were constructed for each V gene as principal component analysis demonstrated clear clustering of sequences by their respective V gene type (Appendix Figure C.1). In humans, there are 7 VH, 6 VL kappa and 10 VL lambda gene families (Vargas-Madrazo et al., 1997; Williams et al., 1996). The non-human (negative) sequences were derived from three species: mouse, rat and rhesus (Appendix Figure C.2). All sequences were aligned using ANARCI (Dunbar and Deane, 2016) with the IMGT numbering scheme (Lefranc et al., 2003).

Table 4.1: **Numbers of human sequences downloaded from the OAS database after filtering.**

	VH	**VL (kappa)**	**VL (lambda)**
V1	1,189,145	8,445,547	7,343,760
V2	52,673	2,873,511	9,005,751
V3	2,680,192	8,678,865	4,788,775
V4	1,075,999	3,245,968	747,946
V5	87,227	32,593	256,729
V6	29,894	131,586	429,196
V7	17,989		637,720
V8			409,564
V10			32,503
Total	5,133,119	23,408,070	23,209,877

Table 4.2: **Numbers of non-human sequences downloaded from the OAS database after filtering.**

	VH	**VL (kappa)**	**VL (lambda)**
Total	12,284,297	950,335	655,826

4.4.1.2 Model training and evaluation

Binary RF classifiers were developed using the scikit-learn Python module, with default parameters unless stated otherwise. Separate models were created for each human V gene type, with the aim of generating V gene type-specific representations of an antibody sequence, rather than an unphysiological representation of a mixture of V gene types.

An 80/10/10 split of the data was employed for training, validation and testing. Each model was trained using 80% of all human sequences of the respective V gene type and 80% of all negative sequences. RF models were built using 200 estimators (performance was found to plateau beyond 200 estimators).

The classifiers output a humanness score on a scale from 0 (least human) to 1 (most human). This score is generated using the scikit-learn predict_proba function, which takes an average of the predicted class probabilities across trees in the forest. The validation dataset was used to set a classification threshold, above which sequences would be classified as human. The threshold was set as the value that maximised Youden's J Statistic (YJS = sensitivity + specificity − 1). Model performance was evaluated using the test dataset and the ROC AUC metric.

4.4.1.3 Kappa and lambda classifier

An RF model to classify whether a light chain sequence is of type kappa or lambda was trained on 25% of the total human VL dataset (12 million sequences). The model demonstrated perfect accuracy, correctly classifying every sequence as kappa

or lambda within the entire VL dataset (both human and negative).

4.4.1.4 Humanness of therapeutic antibodies

To examine how Hu-mAb ranks the humanness of existing therapeutic antibodies, the models were applied to the sequences of 481 approved and Phase 1-3 antibodies, for which both the VH and VL sequences were available, obtained from Thera-SAbDab (April 2020) (Raybould et al., 2020). The therapeutics were categorised by therapeutic origin using the International Nonproprietary Name (INN) infix (Parren et al., 2017) (Appendix Table C.1). We predicted the humanness scores of each VH and VL sequence for these therapeutics and classified therapeutics for which both sequences exceeded the YJS threshold as human.

4.4.1.5 Immunogenicity of therapeutic antibodies

ADA response data, reflecting therapeutic immunogenicity, was collected for 217 antibody therapeutics from clinical papers as described by Clavero-Álvarez et al. (2018) by Claire Marks and Mark Chin. As above, the VH and VL chains were classified separately and the overall therapeutic was only classified as human if both the chains were classified as human.

4.4.1.6 Automated humanization protocol

An automated humanization protocol, which suggests mutations that are predicted to increase the humanness score of an antibody, was created (Figure 4.2). This is an iterative process in which (1) the input antibody sequence is scored by the RF models, (2) exhaustive mutations are made to the framework region of the input sequence (one mutation at a time), (3) the single-point mutated sequences are scored by the RF models and (4) the highest-scoring single-point mutated sequence is selected for

further mutagenesis or output. This process is repeated until a desired humanness threshold is achieved. The V gene type-specific model used for the humanization procedure can be set by the user or, alternatively, is selected as the model which scores the input sequence the highest.

We evaluated the humanization protocol on a dataset of 25 antibody therapeutics, originally derived from non-human sources and subsequently experimentally humanized, for which non-human precursor sequences were available (collected by Claire Marks and Mark Chin). The precursor sequences were used as inputs for the humanization protocol and the threshold was set as the humanness score of the experimentally humanized sequence. The V gene type model was selected as the model which scored the experimentally humanized sequence the highest. Following computational humanization, we compared the amino acid composition and total number of mutations suggested by Hu-mAb to those made experimentally.

We investigated the importance of developing V gene type-specific models by humanizing these 25 therapeutics precursor sequences as described above but using the RF model which output the lowest score for the experimentally humanized sequence.

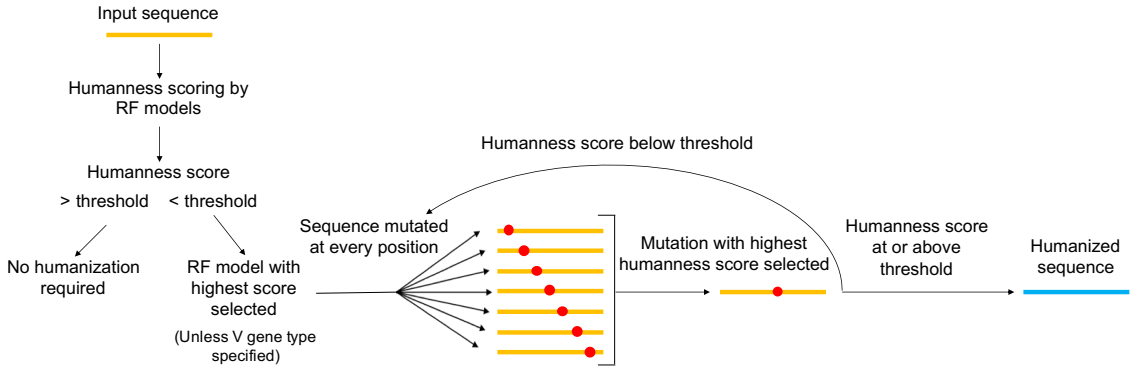

Figure 4.2: **The automated Hu-mAb humanization protocol.** An input sequence is iteratively mutated until a desired humanness threshold is achieved.

4.4.2 Interpretability of humanness predictions

4.4.2.1 Random Forest feature importance

The importance of sequence positions for scoring by the RF models was calculated using the scikit-learn feature_importances_ property.

4.4.2.2 Mutual information analysis of antibody sequences

We analysed the amino acid composition of positions in an MSA of antibody sequences using a mutual information (MI) calculation adapted from Mirny and Gelfand (2002) (Equation 4.1). This calculation was originally devised for an MSA of ortholog (genes with the same function in different species) and paralog (genes with homologous functions arising from gene duplication) sequences (Mirny and Gelfand, 2002). For our purpose, we adapted the input MSA and calculation such that the paralog groups were replaced with 'Human' and 'Non-Human' classification groups. The MI score of a position, i, in the MSA was calculated as follows:

$$\text{MI Score}_i = \sum_{\substack{x=1,\ldots,20 \\ y=Human,Non-Human}} f_i(x,y) \log\left(\frac{f_i(x,y)}{f_i(x)f(y)}\right) \tag{4.1}$$

where $f_i(x)$ represents the frequency of a residue type at position i; $f(y)$, the fraction of sequences belonging to the respective classification group; and $f_i(x,y)$, the frequency of a residue type at a position within the sequences of a classification group (Mirny and Gelfand, 2002).

High MI scores represent positions which are highly conserved in the sequences of both classification groups, but which have a different amino acid at that position between the groups.

The MI calculation was applied to a subset of the Hu-mAb data (Section 4.4.1.1),

due to the computational resources required. A total of 1,000,000 sequences were randomly sampled from the respective datasets: 400,000 human VH3 training sequences, 400,000 negative VH training sequences, 100,000 human VH3 test sequences and 100,000 negative VH test sequences. This will be referred to as the Hu-mAb-1e6-Subset dataset from hereon.

4.4.3 Extension of Hu-mAb to camelid VHH antibody formats

4.4.3.1 Data collection: camel VH and VHH sequences

Camel VH and VHH sequences were obtained from Li et al. (2016) as given in the OAS database (Kovaltsuk et al., 2018; Olsen et al., 2022a). These sequences originated from 3 camels and totalled to 814,456 VH and 750,787 VHH sequences. The sequences were numbered using ANARCI (Dunbar and Deane, 2016) with the IMGT numbering scheme (Lefranc et al., 2003). Redundant sequences, as well as sequences missing the conserved Cys residues (at positions 23 and 104) and FR1 residues, were filtered out. The final datasets consisted of 433,798 VH and 504,894 VHH sequences.

4.4.3.2 Retraining Hu-mAb models with camel sequences

The Hu-mAb RF models were retrained to be applicable to camel sequences. The positive datasets consisted of human sequences from the VH3 V gene type, which has the highest sequence similarity with camel VH and VHH sequences (Appendix Figure C.7 and (Vu et al., 1997; Klarenbeek et al., 2015; Asaadi et al., 2021)). The negative datasets used to train Hu-mAb (Section 4.4.1.1) were augmented with the camel VH and VHH sequences.

4.4.3.3 VHH therapeutics with ADA data

A dataset of 26 therapeutic VHH sequences from ten (multimeric) drugs for which ADA data was available (Appendix Table C.8) was compiled by Ashley Wong. Building on a set previously collected by Rossotti et al. (2022), further sequences were identified from clinical papers and patent literature. All sequences were numbered using ANARCI (Dunbar and Deane, 2016) with the IMGT numbering scheme (Lefranc et al., 2003). For multimeric therapeutics, the highest ADA value across monomers was selected.

4.5 Results

4.5.1 Hu-mAb Random Forest models for evaluating and improving antibody humanness

4.5.1.1 Classifier performance

The binary RF classifiers achieved perfect or near-perfect accuracy in separating human from non-human sequences (ROC AUC scores >0.9999 for every model, Appendix Table C.2). Every VH model perfectly discriminated between human and negative sequences in both the validation and test datasets. Performance on the light chain was also extremely high, but not perfect. This may stem from the smaller number of non-human VL (∼950,000 kappa, ∼650,000 lambda) than VH (>12 million) sequences. We also assessed model performance on a subset of our test dataset limited to sequences with <97% sequence identity with any training/validation sequence, identified using CD-HIT (Fu et al., 2012), and found no drop off in performance (Appendix Table C.2).

These RF models outscored the previous best-in-class approach, an LSTM model (Wollacott et al., 2019). This may result from the larger datasets, as well as the

inclusion of human and non-human sequences in training. Hu-mAb achieves similar or better levels of performance as humanness scoring methods that have been released since, including BioPhi OASis (Prihoda et al., 2022), AbNativ (Ramon et al., 2024) and SelfPAD (Ucar et al., 2024).

4.5.1.2 Humanness of therapeutic antibodies

The RF models were applied to a set of 481 antibody therapeutics (Phase I to approved) obtained from Thera-SAbDab (Raybould et al., 2020). Each VH and VL sequence was scored by the respective set of RF classifiers (VH, VL kappa or VL lambda) and was classified as human if a single model scored it as human (exceeding the YJS threshold). For the VL sequences, an additional RF model was trained to first identify a sequence as kappa or lambda.

The RF models classified more therapeutics as human as the human content of the antibody sequences increased (Figure 4.3): all but 1 of the 176 human antibodies were classified as human and all 14 mouse antibodies were classified as non-human. For the one human antibody incorrectly classified (VH+VL), the light chain humanness score (0.850) fell slightly short of the respective humanness threshold (0.856).

Chimeric antibodies are expected to have a completely non-human variable domain as only the constant domains are replaced with human sequences. However, two VH sequences and one VL sequence (out of 43) were labelled as human by our classifiers. This is likely to be because these sequences were of *Macaca irus* origin, a species that was not present in the OAS training dataset. Two-thirds of the humanized therapeutics had both VH and VL sequences classified as human. Humanized sequences often have arbitrary back mutations in the framework regions to improve efficacy, which might explain why not all humanized sequences were classified as human. Moreover, the INN definition was changed in 2014 such that sequences with

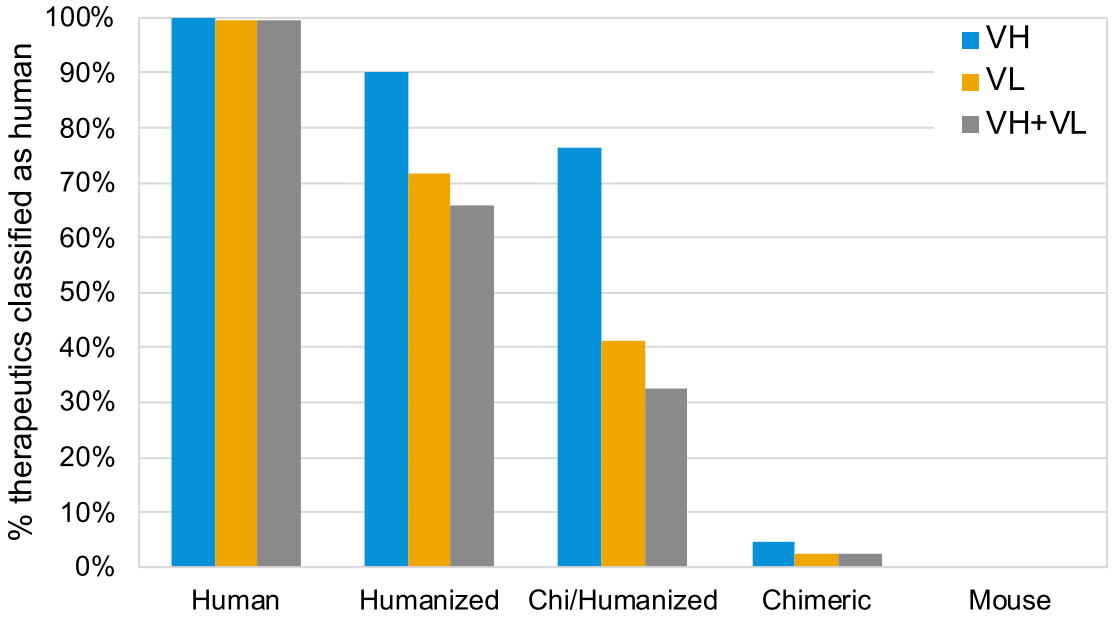

Figure 4.3: **Humanness scores of therapeutics.** The percentage of antibody therapeutics classified as human by our RF models, split by their origin: Human (176 sequences), Humanized (214 sequences), Chi/Humanized (34 sequences), Chimeric (43 sequences) and Mouse (14 sequences). Chi/Humanized are sequences which are part humanized and part chimeric. Therapeutics were classified based on their VH and VL sequences separately, as well as combined; to be classified as human, both VH and VL scores had to be above the respective YJS threshold.

a chimeric origin could be given an INN that implied a humanized sequence (Jones et al., 2016). A lower proportion of VL than VH sequences were classified as human. This could be potentially attributed to the lower number of mutations made in VL sequences during humanization (on average 75% of the number of mutations made in VH sequences, Table 4.3).

4.5.1.3 Relationship between Hu-mAb humanness scores and therapeutic antibody immunogenicity

The aim of humanization is to reduce the risk of eliciting an immunogenic response. A strong predictive score for humanness classification is not sufficient for humanization as it does not explicitly account for immunogenicity. The relationship of the

model scores with observed immunogenic responses, as measured by the appearance of ADAs, was therefore investigated. The fraction of patients with observed immunogenic responses was obtained from FDA labels of approved antibody therapeutics and clinical studies of therapeutics still in clinical trials. There are limitations to this data: for example, there are differences in patient demographics (age, physical conditions, illness), therapeutic dosage levels, length of dosage and whether the treatment was administered in combination with other drugs. In addition, the murine therapeutics within the dataset are likely to be inherently biased toward lower levels of immunogenicity as they are approved therapeutics.

We assessed the correlation between the percentage of patients who developed ADAs and the minimum humanness score of a therapeutic's VH and VL chains, as the least human chain is expected to dictate the level of immunogenicity, across 217 therapeutics. Higher minimum model scores tended to relate to lower immunogenicity, although the correlation was weak with an R^2 of 0.31 (Figure 4.4). This correlation is substantially higher than the R^2 of 0.18 observed in previous work (Clavero-Álvarez et al., 2018). Additionally, our RF models achieved the highest accuracy in paired-sequence immunogenicity prediction when compared with more recently released approaches (Ucar et al., 2024).

High RF humanness scores are linked with low immunogenicity, as visualized when grouping the set of 217 therapeutics by their scores (Figure 4.5). For example, 90% of therapeutics that had both their VH and VL sequence above the YJS threshold exhibited low observed immunogenicity and only one sequence (0.7%) had high immunogenicity. In contrast, >50% of the therapeutics with scores below the YJS threshold had medium or high immunogenicity.

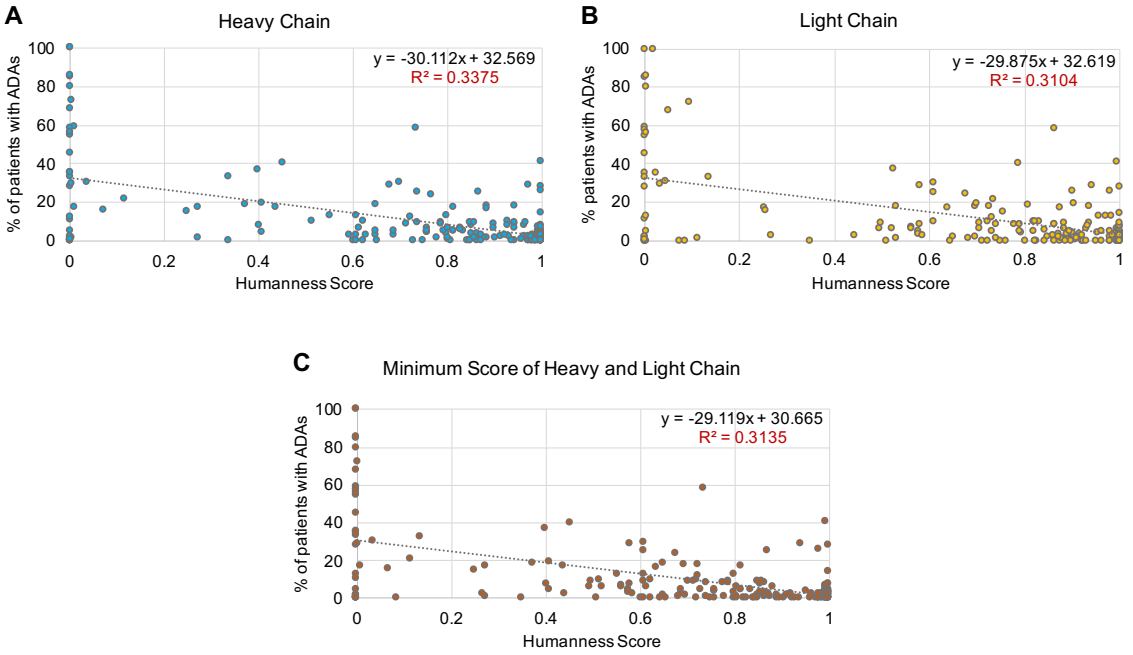

Figure 4.4: **Comparison between RF humanness scores and experimental immunogenicity (fraction of patients that develop ADAs) of therapeutic antibodies.** A) VH sequences (Pearson's correlation: $r = -0.58, p = 5.74^{-21}$). B) VL sequences (Pearson's correlation: $r = -0.42, p = 9.30^{-11}$). C) VH/VL – the minimum humanness score of the respective VH and VL sequence (Pearson's correlation: $r = -0.56, p = 2.95^{-19}$). The Pearson's correlation coefficient was calculated using the scipy.stats.pearsonr Python module.

4.5.1.4 Humanization protocol: comparison to experimental humanization

As high model scores were associated with lower levels of immunogenicity, the RF models were used as a basis for a computational humanization tool, Hu-mAb. Hu-mAb suggests optimal mutations that would increase the model score of the input sequence, therefore lowering immunogenicity. Residues in the CDRs are not mutated to maintain antigen-binding properties. The humanizer should ideally produce as few mutations as possible to reduce the loss of efficacy of the therapeutic. To investigate the similarity between mutations suggested by Hu-mAb and experimentally derived mutations, 25 experimentally humanized sequences that demonstrated low immunogenicity and for which the precursor sequence was available were collected (Appendix

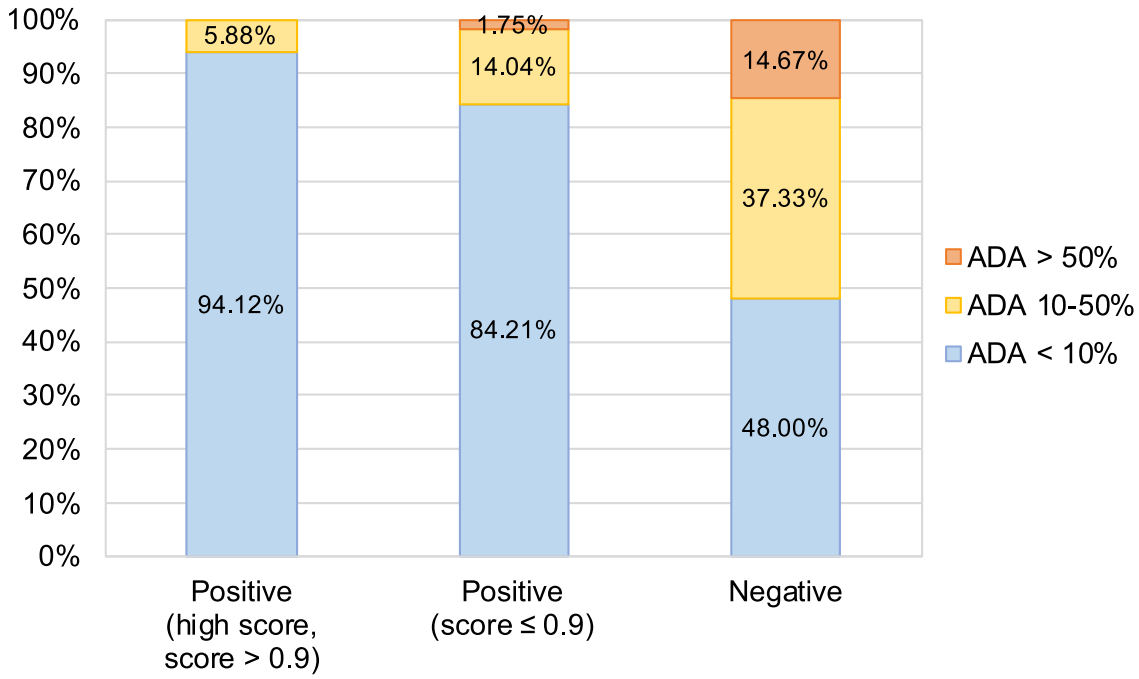

Figure 4.5: **Relationship between RF humanness scores and experimentally determined immunogenicity.** Therapeutics were split into three categories according to the minimum humanness score of the VH and VL chains: positive with a score above 0.9 ['Positive (high score, score > 0.9)'] (85 sequences), above the YJS threshold for the relevant RF model but with a score ≤ 0.9 ['Positive (score ≤ 0.9)'] (57 sequences) and below the YJS threshold ('Negative') (75 sequences). Both the VH and VL sequences had to be above the threshold to be classified as 'Positive'. The immunogenicity of a therapeutic is also represented by three levels: over 50% of patients develop ADAs (orange), 10–50% of patients develop ADAs (yellow) and under 10% of patients develop ADAs (blue).

Table C.3). The VH and VL sequence of each therapeutic was scored by each RF model, and the V gene identified by selecting the model that produced the highest score. The precursor sequence was used as the input sequence into the humanizer, along with its target humanness score (the score achieved by the experimentally humanized sequence) and V gene type.

All precursor sequences were of murine, rat or rabbit origin and most had model scores close to 0. Two therapeutics had precursor sequences which scored above the respective YJS threshold, which is likely due to sequences of their species origin not

being present in the training dataset (Clazakizumab: rabbit VH/VL, Campath: rat VL).

Hu-mAb consistently suggested fewer mutations than the number introduced experimentally (59% and 58% for VH and VL sequences, respectively). Of the mutations suggested by Hu-mAb, an average of 68% and 77% (for VH and VL sequences, respectively) were also made experimentally (overlap ratio, OR). Including mutations to similar residue types (see Appendix Table C.4 for groupings) resulted in an average adjusted OR (AOR) of 77% and 85% for VH and VL, respectively. In contrast, a randomly humanized sequence would be expected to produce an average OR and AOR of ∼2% and ∼6%, respectively (Appendix Table C.5). Hu-mAb is exploiting the information found in the antibody repertoires to more efficiently humanize therapeutic sequences.

We investigated the significance of considering the V gene type in humanization by humanizing these therapeutics using an RF classifier of a different V gene type (e.g., humanization of a sequence that is of the VH1 gene type with the VH2 classifier). Humanization success was substantially lower as compared to above. Of the 25 therapeutics, the humanization of 19 heavy and 8 light chains was unable to reach the humanness threshold of the experimentally humanized sequence. Where the threshold was reached, an average of 12 and 14 more mutations, for heavy and light chains, respectively, were required to achieve the target humanness score. Furthermore, the OR and AOR, calculated for all mutations suggested even if the threshold was not reached, with the experimentally humanized mutations were on average only 10% and 35% (heavy chain) and 12% and 43% (light chain), respectively.

Table 4.3: **Comparison between experimental humanization and our computational tool, Hu-mAb.** The mutation ratio is the average number of mutations Hu-mAb suggested relative to the number of mutations made experimentally in the framework regions; Hu-mAb never suggests mutations to the CDRs. The overlap ratio is the number of mutations that were both suggested by Hu-mAb and made experimentally, relative to the number of mutations suggested by Hu-mAb. For the 'unadjusted' overlap ratio, only mutations to identical amino acid types were considered; the 'adjusted' version considers mutations to similar amino acid types (Appendix Table C.4) to be a match.

	VH						VL					
Therapeutic	**Gene**	**Unadjusted Overlap Ratio**	**Adjusted Overlap Ratio**	**# Hu-mAb Mutations**	**# Experimental Mutations**	**Mutation Ratio**	**Gene**	**Unadjusted Overlap Ratio**	**Adjusted Overlap Ratio**	**# Hu-mAb Mutations**	**# Experimental Mutations**	**Mutation Ratio**
AntiCD28	V3	63%	79%	19	33	58%	KV4	64%	73%	11	19	58%
Campath	V4	75%	88%	16	39	41%	KV1	67%	67%	3	14	21%
Bevacizumab	V3	50%	57%	14	25	56%	KV1	89%	100%	9	16	56%
Herceptin	V3	59%	78%	27	32	84%	KV1	88%	88%	8	22	36%
Omalizumab	V3	62%	76%	21	34	62%	KV1	89%	95%	19	25	76%
Eculizumab	V1	73%	73%	15	23	65%	KV1	83%	83%	12	20	60%
Tocilizumab	V4	64%	86%	14	23	61%	KV1	78%	89%	9	19	47%
Pembrolizumab	V1	73%	73%	11	23	48%	KV3	75%	75%	12	20	60%
Pertuzumab	V3	68%	79%	19	32	59%	KV1	80%	90%	10	20	50%
Ixekizumab	V1	75%	75%	12	29	41%	KV2	78%	100%	9	12	75%
Palivizumab	V2	75%	83%	12	18	67%	KV1	77%	92%	13	26	50%
Certolizumab	V3	61%	78%	18	31	58%	KV1	80%	90%	10	20	50%
Idarucizumab	V4	80%	80%	15	24	63%	KV2	67%	67%	6	8	75%
Reslizumab	V3	50%	80%	10	21	48%	KV1	83%	100%	6	20	30%
Solanezumab	V3	50%	70%	10	16	63%	KV2	88%	100%	8	10	80%
Lorvotuzumab	V3	90%	90%	10	13	77%	KV2	82%	82%	11	13	85%
Pinatuzumab	V3	61%	78%	23	33	70%	KV1	74%	79%	19	23	83%
Etaracizumab	V3	58%	83%	12	16	75%	KV3	62%	69%	13	25	52%
Talacotuzumab	V5	78%	83%	18	33	55%	KV4	73%	73%	11	16	69%
Rovalpituzumab	V1	67%	67%	21	30	70%	KV3	64%	79%	14	26	54%
Clazakizumab	V3	86%	86%	7	27	26%	KV1	75%	75%	4	22	18%
Ligelizumab	V1	64%	64%	11	21	52%	KV3	64%	91%	11	21	52%
Crizanlizumab	V1	64%	64%	11	23	48%	KV1	85%	95%	20	23	87%
Mogamulizumab	V3	67%	67%	6	15	40%	KV2	67%	67%	6	12	50%
Refanezumab	V7	87%	87%	15	17	88%	KV4	92%	100%	12	17	71%
Average		**68%**	**77%**			**59%**		**77%**	**85%**			**58%**
Median		**67%**	**78%**			**59%**		**78%**	**88%**			**56%**

The geometry of the antibody binding site is dependent on the orientation of the VH and VL, which is in turn affected by the residues present at the interface between the two domains. The proportion of mutations suggested by Hu-mAb to key VH–VL interface residues is slightly lower than the proportion made by experimental procedures (Appendix Table C.6) and the OR calculated for these residues is also higher than the average (74%/96% for VH/ VL compared to an average across all mutations of 68%/77%). Since Hu-mAb suggests fewer mutations, the average number of interface mutations per sequence is around half that of experimental procedures (0.8 vs 1.6 for heavy chains, 0.8 vs 1.8 for light chains). A similar pattern was observed for the Vernier zone – Hu-mAb proposed fewer mutations to these residues, which are known to affect CDR conformations (Foote and Winter, 1992) (Appendix Table C.6). The binding properties of the antibody are therefore more likely to be preserved by using Hu-mAb.

4.5.1.5 Hu-mAb is responsive to sequence contexts

Analysis of the Hu-mAb protocol showed that identical mutations (i.e., mutations of position X to residue type Y) in different sequences do not result in an identical increase in humanness score; the effect depends on the rest of the sequence. Moreover, Hu-mAb occasionally made more than one mutation to the same position in the sequence over the course of the humanization procedure (e.g., as shown for the heavy chain of Pembrolizumab in Figure 4.6, where position 13 is mutated in step 1 as well as step 5). These observations suggest that our RF models do not consider positions in the sequence independently, but rather they incorporate interactions between residues to more realistically evaluate humanness.

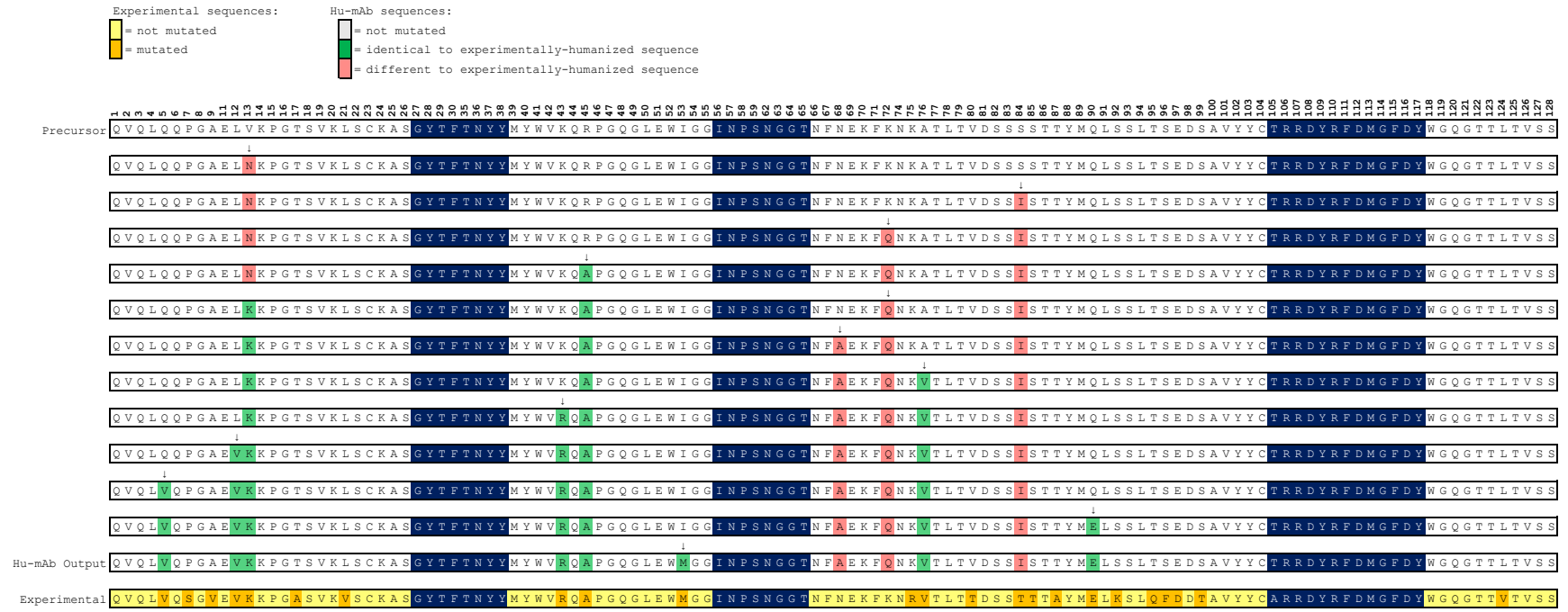

Figure 4.6: **The Hu-mAb humanization procedure demonstrated using the heavy chain sequence of the therapeutic Pembrolizumab.** The IMGT position numbers are shown across the top. The humanized sequence produced experimentally is shown at the bottom of the figure (conserved residues in yellow, mutated residues in orange). Starting with the unhumanized precursor sequence (top), Hu-mAb makes every possible mutation to the framework residues (white) and selects the one that produces the largest increase in humanness score. CDR residues (dark blue) are not mutated to preserve binding. This procedure is performed iteratively until the humanness score reaches a given threshold. Mutations suggested by Hu-mAb are coloured depending on whether they are the same (green) or different (red) to mutations made experimentally. In this case, Hu-mAb suggested 11 mutations (compared to 23 from the experiment), 8 of which were the same as those made experimentally. Hu-mAb made two mutations to position 13: first V $\rightarrow$ N and then later N $\rightarrow$ K, the latter of which is the mutation at this position in the experimentally mutated sequence.

4.5.2 Interpretability of humanness predictions

4.5.2.1 RF model feature importance

To better understand the RF models and what makes a sequence human, we explored model feature importance using the scikit learn feature_importances_ attribute. The results discussed here are for the VH3 model; feature importances for other V gene models are included in Appendix Figures C.4-C.6. For the VH3 model, IMGT position 20 is the most important feature (Figure 4.7a). Position 20 is highly conserved in human and non-human sequences, albeit with a different amino acid in the two classes of sequences: in human sequences, position 20 is primarily Arg, while in non-human sequences, it is primarily Lys (Figure 4.7c). However, a small number of human sequences have the canonical non-human amino acid, Lys, at position 20. To determine the weight the RF models place on a single, but highly discriminating position such as this, model performance was evaluated on 1680 human sequences with Lys at position 20 obtained from the Hu-mAb-1e6-Subset test dataset. The RF models correctly classified each sequence as human and only assigned them marginally lower humanness scores (Figure 4.7d), further supporting that the models are sensitive to sequence context.

Every therapeutic antibody of the VH3 germline (n=195) in the therapeutic antibody dataset (see Section 4.4.1.4) has an Arg at position 20. However, there is insufficient data to determine whether a Lys at position 20 would result in an immunogenic response, particularly given the physicochemical similarity between Arg and Lys. Position 20 is solvent-facing (Figure 4.7b) and would therefore be unlikely to affect the antibody structure or heavy-light chain interface. The amino acid identity may however influence the immune response via altering the binding of the peptide containing this position to the major histocompatibility complex and/or TCR, both of which are sensitive to even conservative single amino acid changes (Sloan-Lancaster

et al., 1993; Hemmer et al., 2000). To note, Lys is found at position 20 in therapeutic antibodies from other germlines (i.e., other sequence contexts; VH1, VH5, VH7) and sequences classified as non-human by Hu-mAb.

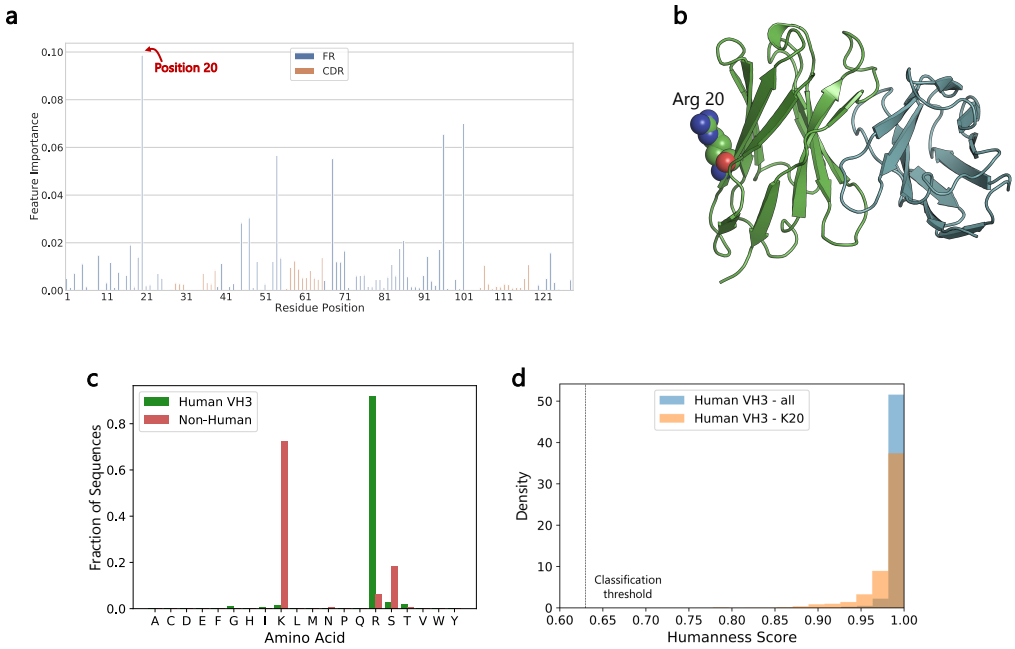

Figure 4.7: **IMGT position 20, the most important feature in the VH3 RF model.** (a) Feature importance of the VH3 RF model. The x-axis shows the residue positions in a sequential manner (IMGT numbering scheme). Position 20 has the highest feature importance. FR regions of the antibody are shown in blue, while the CDRs are shown in orange. (b) Structure of a VH3 antibody (human therapeutic Adalimumab, PDB: 4NYL) with the Arg at position 20 shown as spheres. (c) Amino acid composition at position 20 in human VH3 and non-human training sequences (400,000 sequences each) from the Hu-mAb-1e6-Subset dataset. (d) Distribution of humanness scores of all human VH3 sequences (100,000 sequences) and of human VH3 sequences with a Lys at position 20 (K20, 1680 sequences) in the Hu-mAb-1e6-Subset test dataset. The classification threshold of the VH3 model (0.630) is shown in a dotted line.

4.5.2.2 Mutual information to identify species-discriminating positions in antibody sequences

We next identified further VH3 positions with a similar conservation pattern using a MI calculation adapted from Mirny and Gelfand (2002) (see Section 4.4.2.2; Figure

4.8a). In addition to position 20, positions 54, 68 and 96 had high MI scores, indicating they are conserved in human and non-human sequences, but with a different amino acid between the classes (Figure 4.8b).

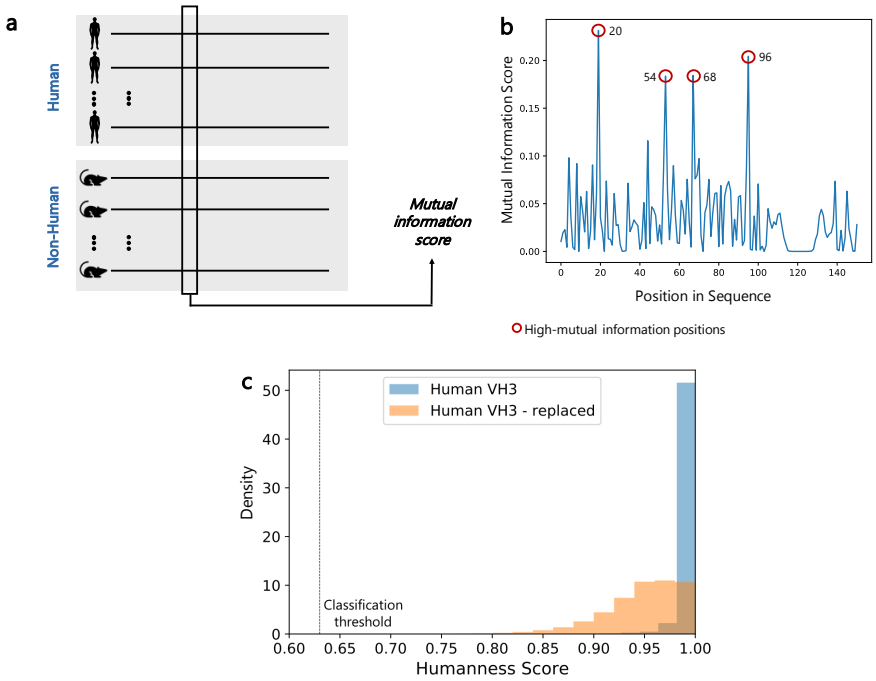

Figure 4.8: **High-mutual information (MI) positions in human VH3 antibody sequences**. (a) Schematic of the MI calculation (adapted from Kleist et al., unpublished). The input consists of sequences belonging to two classification groups: human or non-human. A MI score is calculated for each position in the sequence alignment based on the amino acid frequencies. (b) MI score results for the Hu-mAb-1e6-Subset test sequences (100,000 human VH3 and 100,000 non-human sequences). Four high-MI positions were identified: 20, 54, 68 and 96. (c) Performance of the VH3 RF model on human VH3 sequences from the Hu-mAb-1e6-Subset test dataset in which each of the four high-MI positions was replaced with the respective least-frequent amino acid, compared to original test human VH3 sequences from the Hu-mAb-1e6-Subset dataset (100,000 sequences each). The classification threshold of the VH3 model (0.630) is shown in a dotted line.

To assess the importance of these discriminating positions for the RF model, we simulated a loss of information from these positions in human sequences. The least frequent amino acid at each of these four positions (found in <0.01% of sequences, Appendix Table C.7) was identified and substituted into the respective positions of the Hu-mAb-1e6-Subset human test dataset. As these amino acids appeared so

infrequently in the training data, the model will not have a strong association with a particular label (human or non-human). The model correctly classified >99.9% (all but 12 of 100,000) of the position-replaced sequences as human, albeit with lower humanness scores (Figure 4.8c).

4.5.3 Extension of Hu-mAb to camelid VHH antibody formats

The Hu-mAb RF models demonstrated extremely high accuracy in discriminating human from non-human sequences. However, the models are limited by the species represented in the training data: mouse (VH, VL), rat (VH only) and rhesus (VL only). When applied to therapeutics sequences derived from other species, the models incorrectly classified these as human (Sections 4.5.1.2, 4.5.1.4).

While most therapeutics of non-human origin are derived from murine sources, alternative antibody formats from diverse species, in particular camelid VHHs (Jovčevska and Muyldermans, 2020), are becoming increasingly prevalent. To address this challenge, we expanded the applicability of the RF models to camel VH and VHH sequences.

Camel VH and VHH sequences exhibit the highest sequence similarity with human VH3 sequences (Appendix Figure C.7 and (Vu et al., 1997; Klarenbeek et al., 2015; Asaadi et al., 2021)). The VH3 RF model is therefore used for all downstream analysis and development.

4.5.3.1 Scoring of camel sequences

The original RF models, whose training data did not include any camel sequences, poorly discriminated between human and camel VH/VHH sequences (Figure 4.9a).

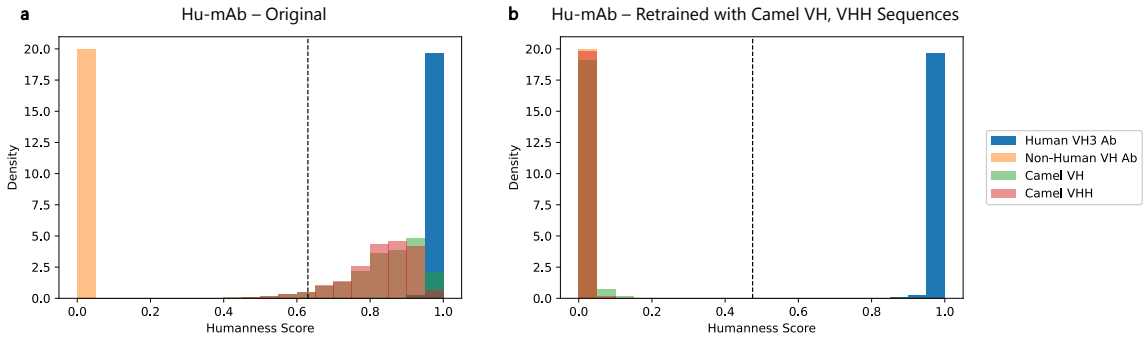

Figure 4.9: **Humanness scores of camel VH and VHH sequences.** The scores of human, non-human (non-camel), camel VH and camel VHH sequences predicted by (a) the original RF VH3 model and (b) the RF VH3 model retrained with camel sequences in the negative dataset. The respective humanness thresholds (0.630 and 0.475) are shown in dotted lines.

Although the latter were assigned lower scores, most fell above the classification threshold.

We retrained the RF models to include camel VH and VHH sequences in the negative data (∼3% and ∼4% of the total negative dataset, respectively). The resulting models achieved a ROC AUC >0.9999 and correctly classified the camel sequences as non-human (Figure 4.9b).

Similar separation is achieved even when only the VH or VHH sequences were included for training, indicating that the model is learning camelid features generalisable beyond the VH/VHH format (Appendix Figure C.8).

4.5.3.2 Humanness of VHH therapeutics

We applied the retrained RF models to a dataset of 10 therapeutics containing single-domain heavy-chain antibodies for which ADA data was available (Section 4.4.3.3, Appendix Table C.8). Eight of the 10 therapeutics exceeded the human classification threshold (0.475), but no therapeutic achieved a minimum score greater than 0.9 (Figure 4.10). Of these 8 therapeutics, 5 had low ADA values, 2 medium and 1 high.

The humanness score (0.520) of the therapeutic with the high ADA value that was classified as human was slightly above the classification threshold. The original RF models, for comparison, scored each therapeutic one category higher (Positive (score ≤ 0.9) $\rightarrow$ Positive (high score > 0.9); Negative $\rightarrow$ Positive (high score ≥ 0.9)).

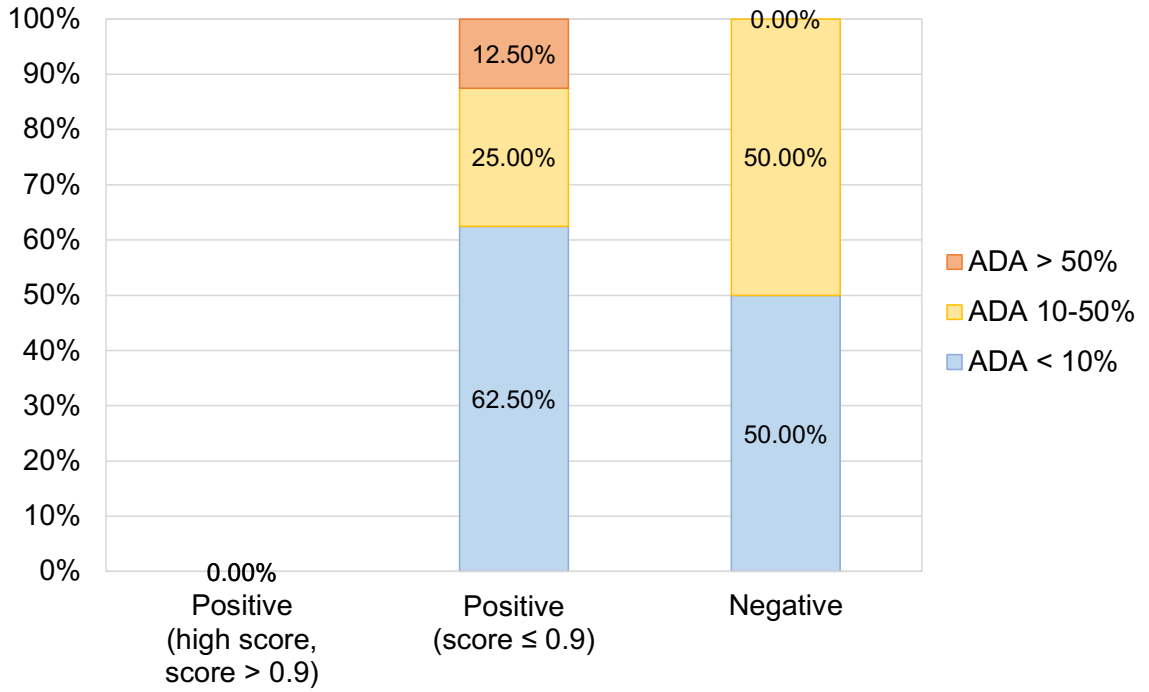

Figure 4.10: **Relationship between the humanness scores produced by our retrained RF models and experimentally determined immunogenicity of VHH antibodies.** Therapeutics were split into three categories according to the minimum humanness score of single-domain heavy chain monomers: positive with a score above 0.9 ['Positive (high score, score > 0.9)'] (0 sequences), above the YJS threshold for the retrained RF model but with a score ≤ 0.9 ['Positive (score ≤ 0.9)'] (8 sequences) and below the YJS threshold ('Negative') (2 sequences). The immunogenicity of a therapeutic is also represented by three levels: over 50% of patients develop ADAs (orange), 10–50% of patients develop ADAs (yellow) and under 10% of patients develop ADAs (blue).

4.6 Discussion

We have developed a novel humanization tool, Hu-mAb, that can humanize potential antibody therapeutics. The tool is based on RF models trained on large-scale

repertoire sequence data, which demonstrate very high levels of accuracy in the classification of human versus non-human antibodies. The humanness scores of the models exhibited a negative relationship with observed experimental immunogenicity. Therefore, sequences that have higher humanness scores are likely to have lower levels of immunogenicity.

Experimental approaches to humanization are largely trial-and-error processes involving the grafting of CDRs onto a human scaffold. If efficacy is lost, arbitrary back mutations are made to attempt to restore it (Safdari et al., 2013). Hu-mAb was constructed as a greedy algorithm and is optimised to select the mutations that provide the highest increase in humanness score, thus suggesting as few mutations as possible to reduce the likelihood of impacting the efficacy of the therapeutic. By utilising RF classifiers that have only been trained on a particular V gene type, the humanizer should produce a realistic sequence with a single V gene origin.

Hu-mAb is efficient and only proposes mutations to the key residues in the framework region responsible for humanness, it incrementally suggests additional mutations to reduce immunogenicity if necessary and back mutations can be suggested in a sequential and non-arbitrary manner (the mutation with the lowest impact on the humanness score). Compared to experimentally humanized therapeutics, Hu-mAb suggested ∼60% of the number of mutations, with high similarity to those made experimentally (average AOR of 77/85%). Hu-mAb offers a promising alternative to experimental humanization approaches, allowing mutations to be made in a more systematic and efficient manner, and achieving similar results in a fraction of the time.

The RF models are able to capture complex sequence information with sensitivity to position interdependency. The predicted humanness scores of mutations differed for different sequence contexts. The models also maintained accuracy even when the

canonical non-human residue was found at the most discriminating position, as well as when a loss of information was simulated for high-MI positions.

Hu-mAb is however limited by its training data: the model often misclassifies sequences from species it has not been exposed to. The original RF models are primarily well-suited for use on murine precursor sequences. As most therapeutics of non-human origin are derived from murine sources, the RF models and Hu-mAb humanizer should already be applicable in many cases.

To extend the applicability of Hu-mAb to alternative antibody formats derived from camelid VHH domains, we updated the training data to include camel sequences. Unlike the original model, the retrained model correctly identified camel sequences as non-human. When applied to 10 single-domain heavy chain therapeutics, the retrained model classified most of the sequences as human, but with scores below 0.9. Analysis on a larger number of single-domain heavy chain therapeutics, when available, will be needed to more definitively assess the relationship between the model's predictions and immunogenicity.

There are additional remaining challenges in antibody humanization, most notably the potential impact on other properties, such as binding affinity and stability, which the RF models do not account for. Future humanization method development could limit suggested mutations to ones predicted to retain antibody properties essential for efficacy and developability.

Chapter 5

Antibody Inverse Folding for Improved Structure-Based Sequence Design

Contents

5.1 Motivation

As discussed throughout this thesis, a wide range of properties must be considered and optimised during therapeutic antibody development (see Section 1.6.4). There is therefore great promise for tools that can capture features relating to multiple properties, and which can be used to bias design towards favourable properties. Inverse folding involves training a model to predict sequence from structure (i.e., to predict the sequence that will fold into a given structure). Models of this type can be used to identify mutations that will be structurally tolerated.

We fine-tuned an existing general protein inverse folding model, ESM-IF1 (Hsu et al., 2022), on antibody structures to produce an antibody inverse folding model, AntiFold. This approach takes advantage of the large general protein datasets ESM-IF1 was trained on, which may allow the model to learn some of the underlying physical properties of protein structures while improving performance on antibody-specific tasks. This model could be used to guide antibody optimisation by limiting mutations to ones predicted to retain the structure, and therefore structure-related properties, of an antibody.

5.2 Contributions

This chapter contains material reproduced from:

Hoie, M.H.*, **Hummer, A.M.***, Olsen, T.H., Nielsen and Deane, C.M. (2024). AntiFold: Improved antibody structure-based design using inverse folding. *arXiv*.

I collaborated with Magnus Haraldson Høie on this project. We have each been involved in all aspects and have contributed equally. I prepared antibody structure data, wrote code for fine-tuning the model, evaluated fine-tuning parameters and

conducted model evaluation and data analysis.

5.3 Introduction

Computational, and in particular ML, tools can be used to reduce antibody liabilities such as immunogenicity and aggregation or to rationally optimise for desirable properties such as binding affinity and developability (e.g., Marks et al., 2021; Prihoda et al., 2022; Tennenhouse et al., 2023; Makowski et al., 2022, 2023; Harvey et al., 2022) (see Section 1.7.2). However, any changes to the antibody sequence may detrimentally impact other features and most current approaches only focus on one or a very small number of properties.

A guiding consideration in optimisation is to select mutations that maintain the structure, and thus biophysical characteristics, such as stability and antigen binding mode, of the antibody. There is therefore a need for models that can suggest mutations which will be structurally tolerated at particular positions. Inverse folding models are trained to predict sequence given structure (Ingraham et al., 2019) and can be used to generate novel sequences without altering the antibody backbone structure. In recent years, there have been many advances in the development of inverse folding models for general proteins (Ingraham et al., 2019; Strokach et al., 2020; Anand et al., 2022; Jing et al., 2021; Hsu et al., 2022; Dauparas et al., 2022).

Antibodies, however, have distinct structure and sequence properties (Stanfield and Wilson, 2014; Regep et al., 2017). For example, over two-thirds of CDRH3 loops adopt structures not found in other general protein structures (Regep et al., 2017). The CDR loops are especially challenging for modelling tasks but are of great interest as they form most of the antigen binding site (MacCallum et al., 1996; Sela-Culang et al., 2013). Training inverse folding models specifically on antibody structures could

therefore improve our understanding of the immunoglobulin fold sequence-structure relationship.

Antibody inverse folding models, AbMPNN (Dreyer et al., 2023) and IgMPNN (Shanehsazzadeh et al., 2023), based on ProteinMPNN (Dauparas et al., 2022), demonstrated the performance gains that can be realised from fine-tuning. However, their sequence recovery on the CDR loops was limited. Additionally, this architecture has several features, including the occasional reordering of antibody heavy and light chains, reversal of residues in the CDRH3 112 positions and insertion of residues into gaps in IMGT-numbered antibodies, incompatible with antibody structures.

Here we present AntiFold, an antibody inverse folding model fine-tuned from ESM-IF1 (Hsu et al., 2022) on solved and predicted antibody structures. AntiFold achieved state-of-the-art performance on antibody sequence recovery across FR and CDR regions. Structural models of AntiFold-predicted sequences showed high similarity with input experimentally solved structures. Furthermore, the AntiFold residue probability outputs correlated with experimental antibody-antigen binding affinity and enabled loss-of-binding variants to be de-selected. The use of AntiFold in tandem with other property prediction tools, to guide mutations, could therefore improve the success rates of *in silico* antibody optimisation.

5.4 Methods

5.4.1 Data

We fine-tuned ESM-IF1 on solved and predicted antibody structures. To enable a direct comparison with AbMPNN (Dreyer et al., 2023), our model was trained, validated and tested on the same data, which was split with a 90% length-matched CDR sequence identity cutoff.

5.4.1.1 Experimental antibody structures from SAbDab

The AbMPNN dataset contains 2,074 structures of antibodies in complex with a protein antigen, after filtering for redundancy and experimental resolution <5 Å (Dreyer et al., 2023). Structures of the corresponding Fv domains, numbered with the IMGT antibody numbering scheme (Lefranc et al., 2003), were obtained from SAbDab (Dunbar et al., 2014; Schneider et al., 2021). Structures of the validation and test set were predicted using ABodyBuilder2 (Abanades et al., 2023) to evaluate AntiFold performance on modelled inputs. One and three structures were removed from the validation and test datasets, respectively, because these could not be modelled with ABodyBuilder2 due to the VL sequences being deemed to be too short (Abanades et al., 2023).

5.4.1.2 Predicted antibody structures from ABodyBuilder2

The structures of 148,832 paired antibody sequences from OAS (Kovaltsuk et al., 2018; Olsen et al., 2022a) modelled using ABodyBuilder2 were released as part of ImmuneBuilder (Abanades et al., 2023). Filtering out structures with identical concatenated CDRs, as in AbMPNN (Dreyer et al., 2023), resulted in a dataset of 147,458 structures.

5.4.2 Fine-tuning strategy

We trained AntiFold by fine-tuning the ESM-IF1 inverse folding architecture (Hsu et al., 2022) on antibody structures. The inverse folding problem can be formalised as learning the conditional probability distribution, $p(Y|X)$, of the protein sequence, Y, consisting of amino acids $(y_1, \ldots, y_i, \ldots, y_n)$, given the structure, X, with spatial coordinates of the backbone atoms (N, C_α and C) $(x_1, \ldots, x_i, \ldots, x_{3n})$ (Hsu et al., 2022):

$$p(Y|X) = \prod_{i=1}^{n} p(y_i|y_{i-1}, \ldots, y_1; X) \tag{5.1}$$

The ESM-IF1 architecture consists of 4 Geometric Vector Perceptron Graph Neural Network (GVP-GNN) layers (Jing et al., 2021), 8 generic Transformer (Vaswani et al., 2017) encoder layers and 8 decoder layers (Hsu et al., 2022) (Figure 5.1). The architecture is invariant to the rotation and translation of the input coordinates.

The ESM-IF1 model was trained only on single-chain structures. In order to represent complexes of antibody heavy and light chains, the backbone coordinates were concatenated with a 10-position padding of "gap" tokens, represented as missing coordinates in the input structure.

5.4.2.1 Fine-tuning parameter evaluation

We evaluated the effect of the parameters described below on model performance, as applied to the validation dataset. To assess the masking and layer-wise learning rate decay parameters, the model was trained on the dataset of solved structures for 10 epochs. To assess the effects of Gaussian noise applied to the predicted structures, and for our final model, we trained for one epoch on the predicted structures followed by up to 100 epochs with early stopping (see details below) on the solved structures.

Masking

We masked portions of the input antibody structure for model training and calculated loss over model predictions for the masked positions. The coordinates of masked positions were hidden for input to the model.

Three different masking schemes were evaluated:

- Shotgun masking: individual positions were randomly selected for masking

- Span masking: consecutive stretches of positions were masked by randomly selecting starting positions and sampling the span length from a geometric distribution where p = 0.05, with a maximum span length of 30 positions, as in Hsu et al. (2022)
- Shotgun plus span masking: 7.5% of the structure was first masked using span masking and a further 7.5% was subsequently masked using the shotgun approach

As our model loss was calculated over masked positions and the FR regions are more conserved than CDRs, we explored whether performance could be improved by biasing the selection of masked positions towards CDR residues. There are more than 2.5 times as many FR as CDR positions in the sequence. For shotgun masking, a 3:1 weighting was implemented for the selection of CDR versus FR positions. For span masking, selection was biased to be low (weight = 1) for most FR positions, high (weight = 3) for most CDR positions and medium (weight = 2) for FR positions immediately preceding CDRs as well as CDR positions immediately preceding FRs.

Layer-wise learning rate decay

The learning rate was decayed for each previous layer in the ESM-IF1 architecture by an alpha factor:

$$LR_i = LR \times \alpha^i \tag{5.2}$$

where i ranges from zero to the number of layers in the model (20) and alpha is set to 0.85.

Gaussian noise

In the case of predicted structures, noise sampled from a Gaussian distribution with a scale of 0.1 Å was added to the backbone (N, C_α and C) 3-dimensional coordinates, following the approach taken in ESM-IF1 (Hsu et al., 2022).

5.4.2.2 Early stopping

Model training was stopped when the validation loss did not decrease after 10 epochs. The model with the lowest validation loss was carried forward.

5.4.3 Model performance evaluation

Amino acid recovery (AAR) was calculated as the percent of positions for which the highest-probability amino acid as predicted by AntiFold was the true amino acid.

Model output probabilities were given by:

$$logits = raw\ model\ outputs \tag{5.3}$$

$$probabilities(i) = \frac{e^{\text{logits}(i)}}{\sum_{j=1}^{20} e^{\text{logits}(j)}} \tag{5.4}$$

Perplexity for each position was calculated as:

$$perplexities = 2^{-\sum_{i=1}^{20} probabilities(i) \times log_2(probabilities(i))} \tag{5.5}$$

5.4.3.1 Sampling and refolding sequences

During sequence sampling, residues were sampled for each position in the CDRs proportional to their probability, using a temperature of 0.20. We used the same method as ProteinMPNN (Dauparas et al., 2022) of applying temperature directly to the logits before converting to probabilities:

$$scaled\ logits = \frac{logits}{t} \tag{5.6}$$

ProteinMPNN (Dauparas et al., 2022) and AbMPNN (Dreyer et al., 2023) were run with default settings and the flags –conditional_probs_only, –sampling_temp 0.20, –num_seq_per_target 20 and –seed 37. Structures of sampled sequences were then predicted with ABodyBuilder2 (Abanades et al., 2023) at default settings. We corrected for ProteinMPNN reordered chains, reversal of insertions in IMGT positions 112 and invalid gaps.

RMSD between the solved and predicted backbone (N, C_α and C atoms) for each region was calculated using Pymol's rms_cur method (Schrödinger, LLC, 2015) after aligning on the framework.

5.4.3.2 Bootstrapping

For bootstrapping, we resampled with replacement 1000 times, with the bootstrapped values used to calculate means and confidence intervals.

5.4.4 Binding affinity prediction

Inverse folding and ESM-2 (650M) log-likelihoods were predicted for antibody variants in the Warszawski et al. (2019) deep mutational scan (PDB 1MLC (Braden et al., 1994) and extracted sequence inputs, respectively; heavy and light variable domains (IMGT positions 1-128)). Experimental scores were mapped to a log_2 fold-change and correlated with log-likelihood scores using scipy.stats.spearmanr.

Structures of antibody variants in the Hie et al. (2023) study were identified by searching the PDB for the extracted antibody sequence and selecting the highest

sequence identity match. In cases where multiple matches of the same sequence identity were available, the X-ray structure with the highest resolution was selected.

5.4.4.1 Rank normalisation

When assessing the model rankings of improved amino acid variants, we first rank-normalised all single amino variant scores (N = L × 20) for each antibody separately. Next, we selected the 124 experimentally measured variants (N = 124) and calculated their ranks using the same formula.

Rank normalisation of scores was calculated as

$$\text{Normalised Rank} = \frac{\text{Rank} - 1}{N - 1} \tag{5.7}$$

where Rank is the variant's score rank and N is the total number of variants.

5.4.5 Statistical tests

All reported p-values were calculated using the Mann-Whitney one-tailed U test unless otherwise stated.

5.4.6 Model speed

AntiFold samples ∼300 antibody structures per minute on a Nvidia GTX 1080 Ti GPU.

5.4.7 Model availability

AntiFold is available at `https://github.com/oxpig/AntiFold`.

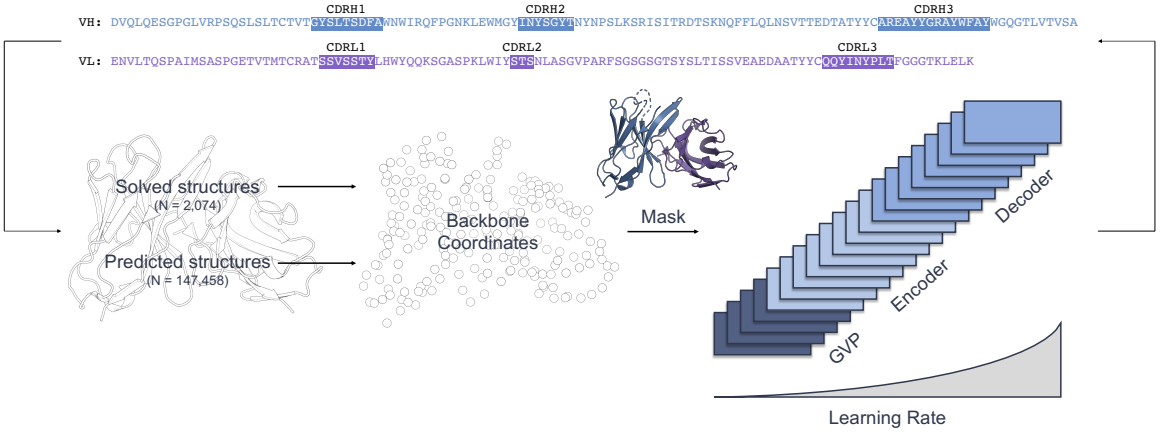

Figure 5.1: **AntiFold model architecture and training.** AntiFold was initialised with weights from ESM-IF1 (Hsu et al., 2022), then fine-tuned on antibody variable domain structures: solved antibody structures from SAbDab (Dunbar et al., 2014; Schneider et al., 2021) and structures of antibody sequences from OAS (Kovaltsuk et al., 2018; Olsen et al., 2022a) modelled with ABodyBuilder2 (Abanades et al., 2023). A subset of positions were masked and layer-wise learning rate decay was applied during training. Sequence and structure from PDB 3W2D (Xia et al., 2014).

5.5 Results

5.5.1 Fine-tuning strategy

Fine-tuning from a general protein inverse folding model enabled us to benefit from existing knowledge learned by ESM-IF1, which was trained on millions of protein structures. We explored the effect of multiple parameters in our fine-tuning strategy.

When fine-tuning on a new task or domain, there is a risk of "catastrophically forgetting" previously learned knowledge. We therefore applied a strategy of layer-wise learning rate decay, successfully used to fine-tune BERT models (Sun et al., 2019). The learning rate was exponentially decayed from the last to the first layer, preserving the weights of earlier parts of the model during training (Figure 5.1). Layer-wise learning rate decay did not further improve sequence recovery (Appendix Table D.1-D.3), however, it was retained for subsequent training to reduce the risk of overfitting and maintain generalisation towards untested properties.

We also investigated different masking schemes in training. Shotgun masking hides the coordinates of randomly selected single positions, while span masking is applied to a consecutive stretch of positions. As FR and CDR regions in the antibody structure have different levels of variability, the selection of masked positions was biased towards the more variable CDR residues (3x weight, IMGT-weighted masking). In total, 15% of the backbone residues were masked during training. As previously reported (Hsu et al., 2022), stronger performance was found for shotgun than span masking on test structures with no masking. However, span masking improved CDR sequence recovery for test cases with masked CDR loops, a realistic design use case (Figure 5.2, Appendix Table D.1-D.3). IMGT-weighted masking further improved performance on CDR loops, while only slightly reducing sequence recovery on FR regions (Figure 5.2, Appendix Table D.1-D.2).

We included a large dataset of 147,458 predicted structures from OAS (Kovaltsuk et al., 2018; Olsen et al., 2022a) in our fine-tuning strategy, in an aim to boost performance by training on more diverse antibodies. The effects of adding Gaussian noise at a scale of 0.1 Å to the modelled protein backbone, previously found to improve performance (Hsu et al., 2022; Dauparas et al., 2022), was evaluated. No substantial effect was found, but Gaussian noise was included in our final model for robustness towards minor variations in input structures (Appendix Table D.3).

Based on these results, the final AntiFold model was trained with IMGT-weighted shotgun and span masking, layer-wise learning rate decay and added Gaussian noise on predicted structures. These augmentations, along with the use of the larger pre-trained ESM-IF1 architecture (142M parameters) instead of ProteinMPNN (1.7M parameters), comprise the main differences with AbMPNN and IgMPNN. The training of AntiFold was split into two phases. First, ESM-IF1 was fine-tuned on one pass of the training dataset of predicted structures from OAS. Next, the model was fine-

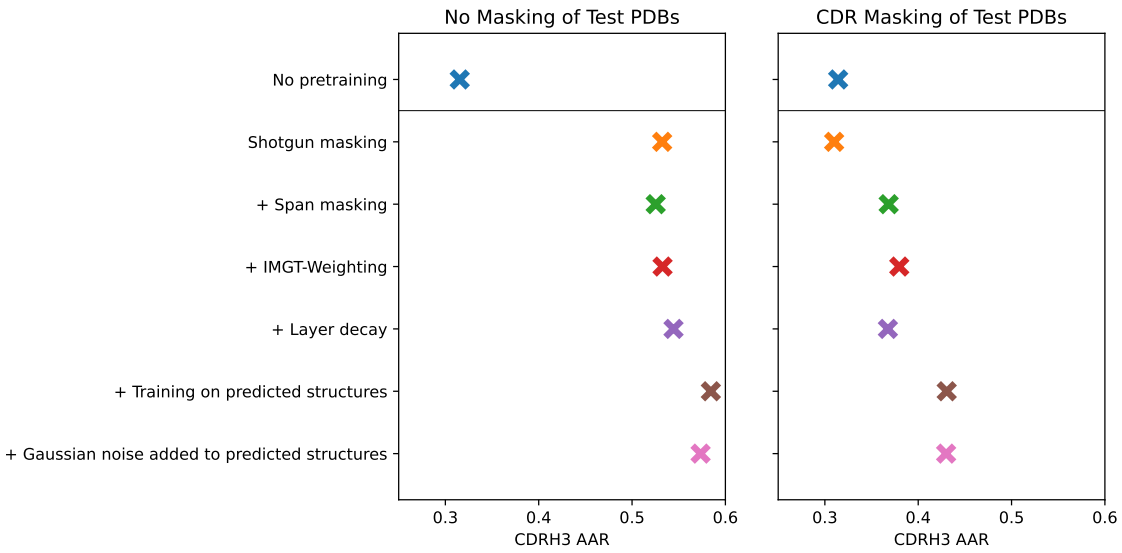

Figure 5.2: **The effect of AntiFold fine-tuning parameters on CDRH3 amino acid sequence recovery (AAR).** The models trained only on solved structures (second to fifth reported values) were trained for 10 epochs. The models trained with predicted structures as well (top one and bottom two reported values) were trained for 1 epoch on the predicted structures and up to 100 epochs with early stopping on the solved structures. The model with no pretraining (top value) was generated with otherwise final model parameters (IMGT-weighted shotgun plus span masking, layer decay, Gaussian noise added to predicted structures). More detailed results are reported in Appendix Table D.1-D.4.

tuned on the solved training dataset for a maximum of 100 epochs, stopping training when there was no further improvement in validation loss for 10 epochs. This model, termed AntiFold, was used for all subsequent analyses.

We quantified the value of pretraining by comparing the final model against a model trained only on antibody structures without initialising from the ESM-IF1 model weights. The recovery of true amino acids, AAR (see Section 5.4.3), on the validation dataset is on average 24% higher across FR regions and 29% higher across CDR regions for the pretrained model (Figure 5.2, Appendix Table D.4).

5.5.2 Fine-tuning improves amino acid recovery on antibody sequences

AntiFold demonstrated a substantial improvement in AAR on the test set of solved structures as compared to the original ESM-IF1 model (43 to 60% for CDRH3; $p < 0.05$, Mann-Whitney one-tailed U test; Figure 5.3a). AntiFold also outperformed AbMPNN across all CDR regions (Antifold 60-84%, AbMPNN 56-76%, Figure 5.3a) and most framework regions (Antifold 87-94%, AbMPNN 85-89%, Figure 5.3b). Performance was lowest across all models for CDRH3 (AntiFold 60% AAR), corresponding with the challenge of predictive tasks for this loop (see Section 1.5). Additionally, AntiFold's performance was lower for antibodies with longer CDRH3 loops, with a median AAR of 71% for shorter loops (6-9 residues) and 48% for longer loops (16-28 residues) (Appendix Figure D.1). We did not directly compare against IgMPNN (Shanehsazzadeh et al., 2023), as the model weights were not made publicly available.

We confirmed that AntiFold can be accurately applied to modelled input structures by testing on ABodyBuilder2 predictions of structures in the test set. AntiFold achieved a similar AAR for solved and predicted structures, in contrast to AbMPNN, which performed slightly worse on solved structures (Figure 5.3c).

Additionally, we calculated the perplexity, representing the average number of amino acid suggestions per position, across positions in the solved structures (see Section 5.4.3). A random model (assigning equal probability to all 20 possible amino acids) would have a perplexity of 20, while an oracle model, assigning 100% confidence to a single amino acid, would have a perplexity of 1. AntiFold suggests on average ∼2-8 amino acids in the CDRH3 which are likely to preserve the fold of the loop, as compared to ∼3-10 amino acids for AbMNN (Figure 5.4). Both models achieved lower perplexity than the observed frequency of different amino acids in centres of

the CDRH3 loops of the test set (Figure 5.4, grey).

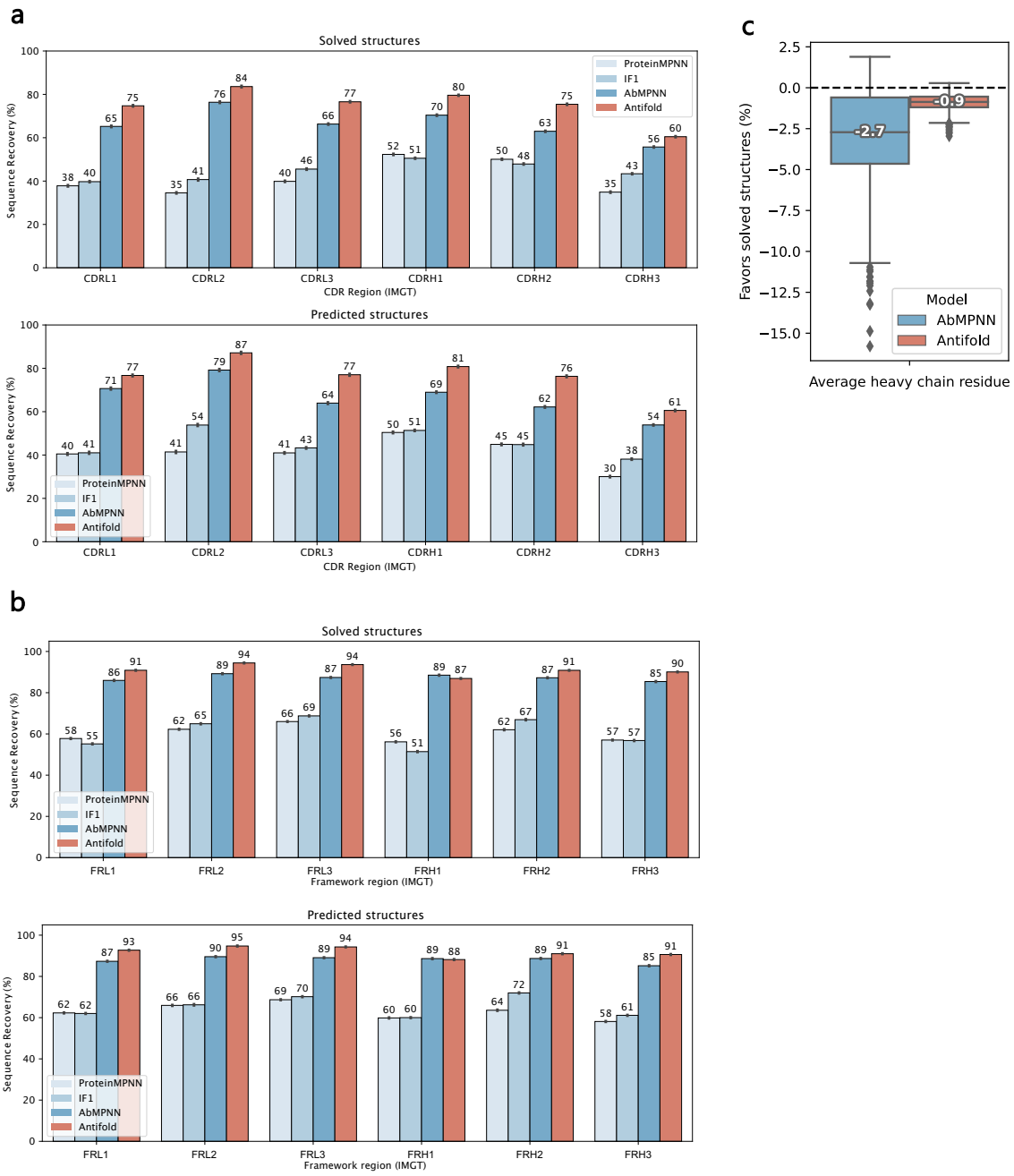

Figure 5.3: **AntiFold sequence recovery.** Amino acid sequence recovery (AAR) for solved (top) and predicted (bottom) structures in the test set: (a) CDR and (b) FR regions. (c) Percent change in AAR when applied to predicted versus solved heavy chain structures of the test set.

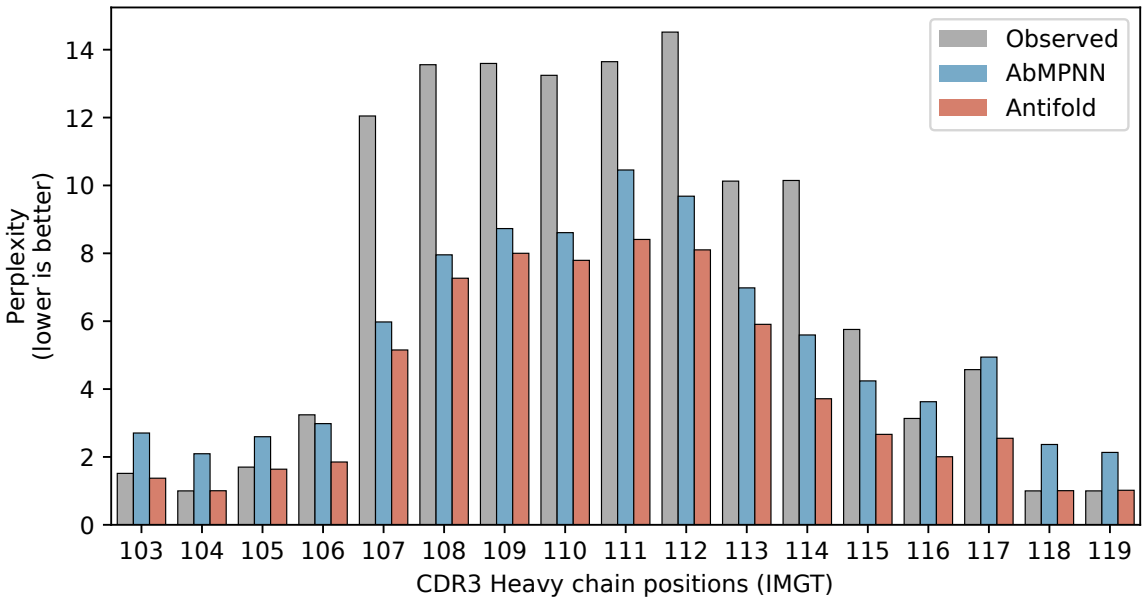

Figure 5.4: **Perplexity across the CDRH3 loop.** The 'observed' perplexity (grey) reflects the frequency of amino acids observed in the true CDRH3 sequences of the test dataset. A value of 1 would indicate that only one amino acid is ever observed, while 20 would represent all twenty amino acids being observed at equal frequency.

5.5.3 Predicted sequences have good structural agreement with experimental structures

To assess whether model predictions preserve the fold of the CDRs, we compared true structures with predicted structures of AntiFold-sampled sequences. This approach was applied to 56 high-quality antibody structures in the test set, which were solved using X-ray crystallography with a resolution below 2.5 Å.

For each antibody, 20 sequences were sampled using AntiFold, AbMPNN, ESM-IF1 and ProteinMPNN with a sampling temperature of 0.20. We modelled the outputs using ABodyBuilder2, aligned the predicted structures with the framework backbone of their experimentally solved counterpart, then calculated the RMSD over the CDR backbone. The RMSD calculation does not take the side chains into account. As a baseline, the true sequences were modelled with ABodyBuilder2 (native).

AntiFold generated sequences with high structural similarity to the original CDR

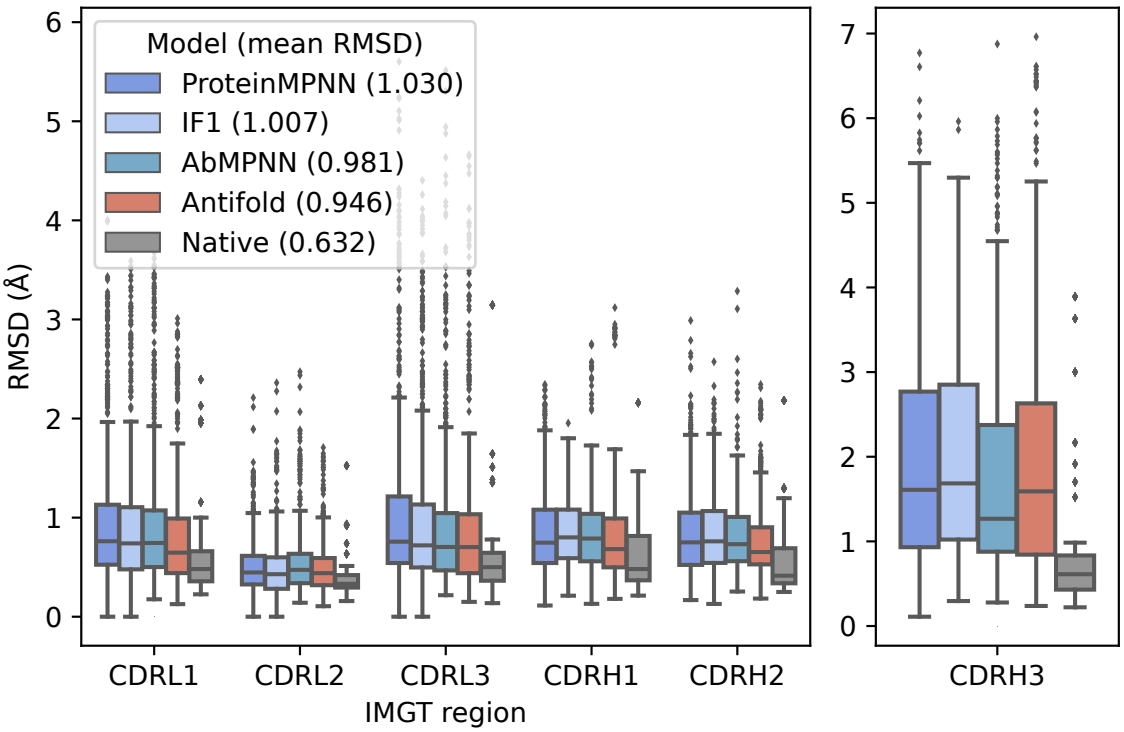

Figure 5.5: **Refolding of inverse folding-sampled sequences.** Sequences were sampled with AntiFold, AbMPNN, ESM-IF1 and ProteinMPNN (sampling temperature 0.20) and the structures predicted with ABodyBuilder2. The CDR backbone RMSD between the experimentally solved structures and predicted structures of the sampled sequences are shown. A comparison to the ABodyBuilder2-predicted structure of the true sequence is included as a baseline (native). Mean CDR region RMSD values are shown in parentheses in the legend.

backbones, with a mean CDR region RMSD of 0.946 Å (versus native RMSD 0.632 Å, AbMPNN 0.981 Å, ESM-IF1 1.007 Å, ProteinMPNN 1.030 Å; Figure 5.5).

5.5.4 Inverse folding probabilities correlate with antibody-antigen binding affinity

We next investigated whether AntiFold had captured information relevant to a property it had not directly been trained on, antibody-antigen binding affinity. AntiFold and other inverse folding models were used to calculate the log-likelihoods of 2209 variable domain variants of an anti-lysozyme antibody (D44.1) generated in a deep muta-

tional scanning dataset (Warszawski et al., 2019) (PDB 1MLC (Braden et al., 1994)). As a sequence-only comparison, ESM-2 (650M parameters) (Lin et al., 2023) was included. AntiFold significantly outperformed the other models with a Spearman's rank correlation (S_r) of 0.418 (Figure 5.6a,c; $p < 0.05$, Mann-Whitney one-tailed U test). All inverse folding models achieved higher correlations than the sequence-based ESM-2 ($S_r = 0.264$) ($p < 0.05$).

Next, the importance of including the antigen chain(s) when calculating inverse folding probabilities was quantified. With this additional context, AntiFold's performance improved in the case of all CDR regions, in particular the CDR2 (S_r of 0.427 to 0.473, $p < 0.05$) and CDR3 (S_r 0.298 to 0.320, $p < 0.05$), but not for framework residues (Figure 5.6b, Appendix Figure D.2). Contrary to this, ProteinMPNN and AbMPNN lost performance for most CDR regions when the antigen chain was included, while performance on framework residues was largely unchanged (Figure 5.6b, Appendix Figure D.2).

To investigate which mutations AntiFold performed best on, the variants were separated into those increasing (log_2 fold-change > 0) and decreasing (log_2 fold-change ≤ 0) binding affinity. At a variant log-likelihood threshold of -11, 40% of disruptive mutations were de-selected while >95% of improved variants were maintained (Figure 5.6d).

We further explored the ability of inverse folding models to de-select affinity-reducing variants by applying them to a dataset of 124 mutations made across 7 antibodies in protein language model-guided affinity maturation experiments (Hie et al., 2023). For each antibody, the PDB structure with the highest sequence identity and, if multiple structures were available, the highest X-ray crystal structure resolution was used.

Each model's scores were rank-normalised across all single-amino acid variants.

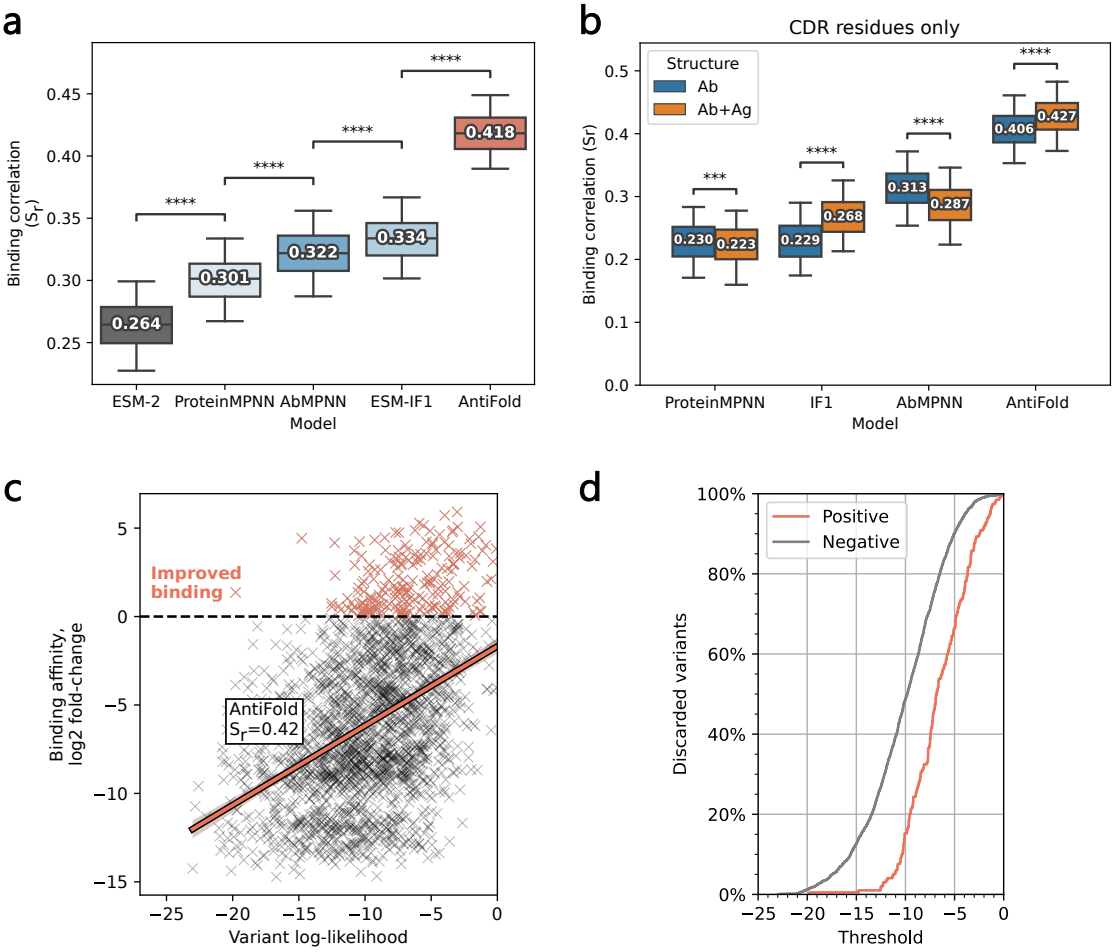

Figure 5.6: **Zero-shot prediction of D44.1 antibody-antigen affinity by inverse folding models.** (a) Spearman's rank correlation between log-likelihood scores of the 2209 variants of the D44.1 anti-lysozyme antibody and the log_2 fold-change in binding affinity (Warszawski et al., 2019). (b) Spearman's rank correlation (CDR residues only), excluding (Ab) and including (Ab+Ag) the antigen chain. (a-b) Error bars indicate the 5-95th percentile range for the Spearman's rank correlation after bootstrapping 1000 times. Statistical significance from Mann-Whitney one-tailed U tests is shown (**** = p < 0.00005). (c) Scatterplot of AntiFold variant log-likelihood scores versus experimental binding affinity values. Spearman's rank correlation (S_r) and fitted ordinary least squares model are shown. Variants with improved binding affinity (log_2 fold-change > 0) are shown in orange (positives). (d) Percent of discarded variants (i.e., labelled as decreasing affinity) at varying log-likelihood score thresholds.

Next, the 124 experimentally measured variants were selected and the rank-normalisation updated. Using the experimental binding affinity values, variants were separated into lower (fold-change < 0.75), maintained (0.75-1.25) and higher (> 1.25) binding affin-

ity groups.

AntiFold achieved significantly improved separation of these groups, scoring the improved variants with a median rank score of 80% versus 73% for ProteinMPNN, 57% for ESM-IF1 and 55% for AbMPNN (Figure 5.7; $p < 0.05$, Mann-Whitney one-tailed U test).

5.6 Discussion

Therapeutic antibody development requires solving a complex, multi-parameter problem where optimising one property can detrimentally impact another. Inverse folding models, trained to predict sequences which fold into a desired structure, could be used to identify structurally tolerated mutations and improve the efficiency of optimisation. For example, these models could be used in conjunction with predictors of other properties (e.g., immunogenicity) to limit the search space to mutations with a lower likelihood of disrupting the antibody structure.

We fine-tuned a general protein inverse folding model, ESM-IF1 (Hsu et al., 2022), on antibody structures. This approach allowed us to take advantage of what ESM-IF1 learned from the millions of structures in its training dataset while improving performance further on antibody tasks. The benefits of this strategy are highlighted by the increased antibody sequence recovery (1) when initialising with weights from ESM-IF1, as opposed to training only on antibody structures, and (2) for AntiFold as compared to ESM-IF1. The former is likely the result of the orders-of-magnitude difference in the number of structures for general proteins versus antibodies, allowing ESM-IF1 to better learn fundamental properties about protein sequence-structure relationships. With regard to the latter and our fine-tuning strategy, we evaluated a range of parameters and found that the model was relatively robust to their values. The largest improvements came from the masking scheme – with IMGT-weighted

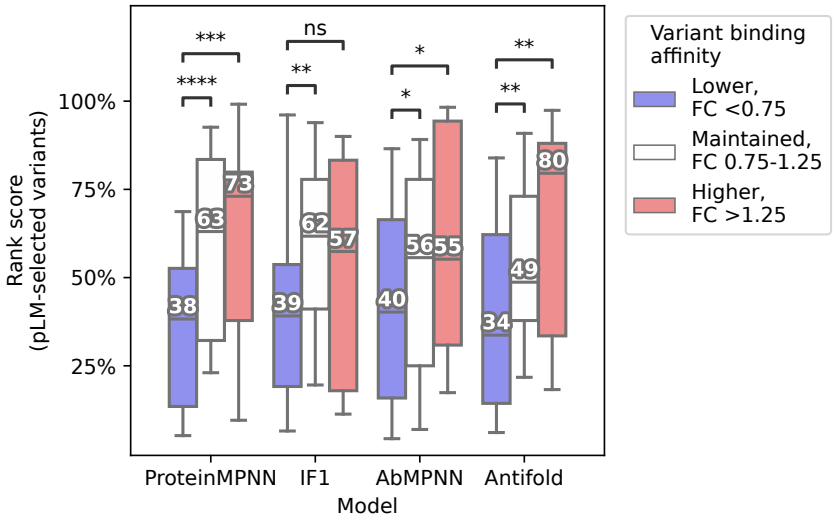

Figure 5.7: **Inverse folding model ranking of mutations identified for affinity maturation by a protein language model**. Variants of 7 antibodies (Hie et al., 2023) were separated into mutations which resulted in lower (fold-change, FC, <0.75), maintained (fold-change 0.75-1.25) and higher (fold-change >1.25) binding affinity. The 5-95th percentile of variant scores (rank-normalised across the 124 variants), median values and Mann-Whitney one-tailed (less) U test significance are shown (**** = $p < 0.00005$).

shotgun plus span masking optimising performance and flexibility to different design tasks – and training on predicted in addition to solved antibody structures.

The resulting model, AntiFold, not only outperformed ESM-IF1 but also another general protein and a fine-tuned antibody-specific inverse folding model, ProteinMPNN and AbMPNN, achieving state-of-the-art antibody sequence recovery. AntiFold exhibited greater confidence (lower perplexity) than AbMPNN, which would allow downstream experimental validation to be prioritized to a smaller and more efficient search space. Additionally, although all tested inverse folding models performed well on the task of refolding sampled sequences (mean CDR backbone RMSD $\sim$ 1 Å), the improved AAR indicates that AntiFold predictions would recapitulate antibody sequences with higher fidelity.

Fine-tuning also improved performance on the zero-shot task of affinity prediction, for which the model was not explicitly trained. AntiFold residue log-likelihoods

achieved the highest correlation with antigen affinity values for an anti-lysozyme antibody. AntiFold achieved this performance primarily by identifying and predicting a lower log-likelihood for mutations with a destabilising effect on binding affinity. All inverse folding models outperformed the sequence-based ESM-2, underscoring the value of incorporating structural information. Applying the inverse folding models to a dataset from ESM-2-guided affinity maturation experiments indicates that the structure- and sequence-based models learn complementary information, as inverse folding models, and particularly AntiFold, were able to further separate mutations.

AntiFold predictions exhibit high similarity to true antibody sequence and structure and capture information relevant to antibody-antigen affinity. These results demonstrate the promise of AntiFold for guiding antibody optimisation by identifying mutations that maintain the antibody structure and structure-related properties.

Chapter 6

Conclusions and Future Directions

Contents

6.1 Conclusions

This thesis presents several novel ML-based approaches for therapeutic antibody optimisation. I identified challenges, including data availability and interpretability, as

well as successful applications, including humanization and inverse folding, of ML. Overall, I believe this research lays the groundwork for future multi-objective *in silico* antibody optimisation and, eventually, the one-shot design of optimal therapeutic antibodies.

6.1.1 Antibody-antigen binding affinity prediction

In Chapter 2, I detailed efforts to predict the effects of mutations on antibody-antigen binding affinity, a problem at the core of therapeutic antibody development. I found that our – and all existing – models were overtraining on the few hundred experimental data points available. My results indicate that, given current methods, we will need orders of magnitude more data, and better consideration of dataset diversity, to build a generalisable predictor of affinity. These findings establish guidelines for ML model development and experimental data generation, which will not only bring the field closer to achieving generalisable antibody-antigen affinity prediction but are also broadly applicable across biomolecular ML.

6.1.2 Interpretability of affinity predictions

In Chapter 3, I applied the EGNN architecture developed in Chapter 2 to a new, large experimental dataset of Trastuzumab variants. I aimed to identify the antibody-antigen interface graph components that contribute most strongly to model predictions. This proved to be a challenging task, with inconclusive and not easily understandable outputs. Neural network interpretability is an area that requires future work, including applications to different graph inputs and protein-protein complexes.

6.1.3 Antibody humanness and immunogenicity

In addition to antibody-antigen binding affinity, multiple properties affecting safety and manufacturability must be considered. In Chapter 4, I outlined our ML approach to accurately identify non-human antibodies and systematically improve humanness, thereby reducing immunogenicity. Our Random Forest classifiers achieved near-perfect, best-in-class accuracy in discriminating human from non-human antibodies. Moreover, the model scores assigned to existing therapeutics were predictive of immunogenicity. Building on our classifiers, we developed a humanization tool to iteratively increase antibody humanness. Hu-mAb offers a promising alternative to trial-and-error experimental humanization, enabling more systematic and efficient optimisation.

6.1.4 Antibody inverse folding

Due to the often antagonistic nature of different properties, it will be necessary to move beyond optimising properties in isolation or even in series. Chapter 5 described our development of an antibody-specific inverse folding model through fine-tuning. AntiFold achieved state-of-the-art antibody sequence recovery and displayed high confidence in its predictions, which would prioritise downstream experimental validation to a narrow search space. Additionally, predicted structures of sequences sampled with AntiFold exhibited high agreement with input structures. These results highlight the ability of our inverse folding model to capture antibody sequence-structure relationships and the value of fine-tuning to increase task-specific performance. AntiFold has promise for guiding antibody optimisation by identifying mutations that preserve the antibody structure and structure-related properties.

6.2 Future directions: next steps

The findings of this thesis open the doors to a number of pathways for future exploration, which I outline below.

6.2.1 Machine learning strategies to overcome limited data availability

The research presented in Chapter 2, investigating the data that will be required for the generalisable prediction of antibody-antigen $\Delta\Delta G$, could be extended via application to other synthetic datasets, such as one created using Rosetta Flex ddG (Barlow et al., 2018). This could provide further evidence for the disparity between existing and needed data, as well as more insights into the amount/range of data required. Additionally, it would be valuable to better understand the consistency (or lack thereof) between different physics-based techniques.

In a longer timeframe, as more data becomes available, ML engineering and/or data augmentation strategies can be implemented to improve the accuracy of antibody-antigen $\Delta\Delta G$ prediction. Recent advances in experimental methods, for example, MAGMA-seq, which can measure binding affinity for multiple antibodies and multiple antigens simultaneously (Petersen et al., 2024), could allow tens to hundreds of thousands of antibody-antigen $\Delta\Delta G$ values to be generated. Even if the resulting data is not sufficient for immediate accurate prediction, it could contain enough signal to be extracted by various approaches. For example, models could be pre-trained on synthetic datasets (such as the one generated in Chapter 2) or general protein-protein complexes and subsequently fine-tuned on experimental $\Delta\Delta G$ values. (This strategy has been unsuccessful thus far due to the limitations on the experimental $\Delta\Delta G$ datasets for fine-tuning.) Additionally, semi-supervised learning, in which a model trained on a moderate amount of labelled data is used to produce labels for

unlabelled data, could be used to augment the training data. Another approach, meta-learning has already shown promising initial results for antibody-antigen affinity prediction, although limited to a single-antigen task (Minot and Reddy, 2024). Meta-learning can be used to reweight labels, placing a stronger weight on those the meta-model believes are accurate, as well as to relabel data points, correcting labels the model believes are inaccurate.

6.2.2 Machine learning interpretability

The main research direction that follows on from Chapter 3 will require identifying a suitable interpretability method that truly reflects what the EGNN model is learning. Subsequently, it will be important to explore whether the model is strongly weighting components that a human would, i.e., whether the weights align with our understanding of chemistry and physics. If not, this could uncover biases in the model or training data that enable high performance by weighting the 'wrong' features (as, for example, was seen for a model trained to differentiate wolves from huskies which instead learned to detect whether there was snow in the background (Besse et al., 2018)). Alternatively, if the model weightings are indeed meaningful, they could expand our understanding of antibody-antigen interactions and be used to guide antibody design.

6.2.3 Humanness of antibodies from diverse species and alternative formats

As shown in Chapter 4, predicting whether a sequence is human or not is a relatively straightforward task when the origin species is included in the training data: simple Random Forest models trained directly on one-hot encoded sequences achieved near-perfect accuracy. Challenges arise when the models are applied to sequences derived from species they have not seen before and potentially, although as yet not thoroughly explored, computational design.

The former difficulty can be countered by regular retraining of the Random Forest models, as more species are included in OAS (Kovaltsuk et al., 2018; Olsen et al., 2022a). However, they may still struggle with computationally designed antibodies, which are becoming more common with advances in language models (e.g., Madani et al., 2023; Olsen et al., 2024; Kenlay et al., 2024) and diffusion-based *de novo* design (Luo et al., 2022; Martinkus et al., 2023; Bennett et al., 2024). A recent ML model released after Hu-mAb achieved strong predictive performance when trained only on positive (human) sequences (Ramon et al., 2024). This strategy is likely to be less biased towards or against particular non-human species. However, this approach does not stratify by V gene type and may produce non-physiological sequences of mixed V genes. Subsequent models could train only on human sequences, but ensure the resulting outputs are V gene type-specific (either through model design or extensive testing).

There is also more work required to extend the applicability of humanness predictors to the wide variety of antibody formats that are being developed. While VHH formats are typically closely linked to species (e.g., camelid), to which the retrained Random Forest models presented in Chapter 4 are applicable, further consideration is needed for the residues that form the VH-VL interface in standard antibody formats but are solvent-exposed in VHHs (Gordon et al., 2024). Other formats, such as multi-specifics, may pose unknown challenges, such as from the non-antibody protein linkers used to connect VHH modules in some newer therapeutics.

6.2.4 Antibody language models for property prediction

Language models have been shown to capture underlying properties of proteins, even when information about these is not directly included in training. The latent space embeddings or sequence scores (e.g., log-likelihood scores) generated from these mod-

els can be used for prediction on downstream tasks with (few- to many-shot prediction) or without (zero-shot prediction) further training data. For example, in Chapter 5, we found that AntiFold has some predictive power for a task, affinity prediction, it was not explicitly trained for. Furthermore, these predictions appeared to contain information not captured by the sequence-based protein language model, ESM-2.

There are no clear guidelines for which tasks structure- versus sequence-based and general protein versus antibody-specific models are most suitable for. Future work could benchmark models on a diverse set of tasks to identify whether there are evident use cases for each type of model. Initial results for downstream antibody applications of language models indicate that there is no consistently successful model or class of models for zero-shot prediction, even for different datasets for the same task (Chungyoun et al., 2023). Few-shot evaluation is likely to be more successful, necessitating guidelines on the amount of labelled data required for different tasks.

Additionally, it would be useful to explore in more depth whether structure- and sequence-based language models learn orthogonal information, as is hinted at by the affinity-based ranking results from AntiFold. If true, a subsequent avenue could be to investigate architectures that jointly embed sequence and structure.

6.3 Future directions: longer-term perspectives

The overarching goal this thesis and the broader field are working towards is the ability to design a therapeutic antibody entirely computationally, such that experimental validation can be saved for the final stages before clinical development. Tremendous steps have been made to bring us closer to this goal in recent years, but challenges remain.

6.3.1 Multi-property optimisation

Antibody optimisation must move beyond predicting and improving properties individually to a multi-objective strategy. The current approach of improving only one or two properties at a time risks detrimental impacts to other properties, introducing time-consuming inefficiencies to the development pipeline. A range of mathematical approaches exist for general multi-objective optimisation (including scalarisation methods like weighted sums, surrogate models, multi-objective evolutionary algorithms and multi-objective reinforcement learning) but it remains to be seen which are most efficient and appropriate. Additionally, predictive models will be required to produce the inputs (the predicted effects of mutations on various properties) for multi-objective optimisation.

6.3.2 Machine learning-grade data

Predictive models remain limited by the availability of labelled training data. As such, a two-pronged approach of model development and dataset generation will be essential. We will need to create, and make publicly available, 'machine learning-grade data'. At its core, this data is designed specifically for ML. The measurements should be made using standard processes in a high-throughput manner and include uncertainty quantifications. Additionally, the datasets must be designed from the outset to consider diversity and bias.

Data generation and ML should also be integrated into a seamless cycle driven by active learning, in which further data collection is guided by the model to prioritise areas of the search space that have high uncertainty or are undersampled.

6.3.3 One-shot antibody design

The ultimate aim is to design therapeutic antibodies in one shot, fully *in silico*. To achieve this, predictive and generative models could be incorporated to directly produce optimised antibody outputs. For example, predictive models could be used to guide adversarial training, design loss and/or reward functions and condition or constrain generative model inference. These strategies have the potential to not only accelerate the time to obtain a lead candidate but also to increase the success rates of clinical development.

Bibliography

Abanades, B., Georges, G., Bujotzek, A., and Deane, C. M. ABlooper: Fast accurate antibody CDR loop structure prediction with accuracy estimation. *Bioinformatics*, btac016, 2022.

Abanades, B., Wong, W. K., Boyles, F., Georges, G., Bujotzek, A., and Deane, C. M. ImmuneBuilder: Deep-Learning models for predicting the structures of immune profteins. *Communications Biology*, 6(1):1–8, 2023. doi: 10.1038/s42003-023-04927-7.

Abhinandan, K. R. and Martin, A. C. Analyzing the "Degree of Humanness" of Antibody Sequences. *Journal of Molecular Biology*, 369(3):852–862, 2007. doi: 10.1016/j.jmb.2007.02.100.

Abhinandan, K. R. and Martin, A. C. Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Molecular Immunology*, 45(14):3832–3839, 2008. doi: 10.1016/J.MOLIMM.2008.05.022.

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S.,

Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D., and Jumper, J. M. Accurate structure prediction of biomolecular interactions with AlphaFold3. *Nature 2024*, 1–3, 2024. doi: 10.1038/s41586-024-07487-w.

Adolf-Bryfogle, J., Xu, Q., North, B., Lehmann, A., and Dunbrack, R. L. PyIgClassify: a database of antibody CDR structural classifications. *Nucleic Acids Research*, 43(D1):D432–D438, 2015. doi: 10.1093/NAR/GKU1106.

Akbar, R., Robert, P. A., Pavlović, M., Jeliazkov, J. R., Snapkov, I., Slabodkin, A., Weber, C. R., Scheffer, L., Miho, E., Haff, I. H., Haug, D. T. T., Lund-Johansen, F., Safonova, Y., Sandve, G. K., and Greiff, V. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Reports*, 34(11):108856, 2021. doi: 10.1016/J.CELREP.2021.108856.

Alanine, D. G. W., Quinkert, D., Kumarasingha, R., Gilson, P. R., Higgins, M. K., and Draper, S. J. Human Antibodies that Slow Erythrocyte Invasion Potentiate Malaria-Neutralizing Antibodies. *Cell*, 178:216–228, 2019. doi: 10.1016/j.cell.2019.05.025.

Aldeghi, M., Bluck, J. P., and Biggin, P. C. Absolute Alchemical Free Energy Calculations for Ligand Binding: A Beginner's Guide. *Methods in Molecular Biology*, 1762:199–232, 2018. doi: 10.1007/978-1-4939-7756-7_11.

Allen, C. D., Okada, T., and Cyster, J. G. Germinal-Center Organization and Cellular Dynamics. *Immunity*, 27(2):190–202, 2007. doi: 10.1016/J.IMMUNI.2007.07.009.

Almagro, J. C., Pedraza-Escalona, M., Arrieta, H. I., and Pérez-Tapia, S. M. Phage Display Libraries for Antibody Therapeutic Discovery and Development. *Antibodies (Basel)*, 8(3):44, 2019. doi: 10.3390/ANTIB8030044.

Alt, F. W., Yancopoulos, G. D., Blackwell, T. K., Wood, C., Thomas, E., Boss, M., Coffman, R., Rosenberg, N., Tonegawa, S., and Baltimore, D. Ordered rearrangement of immunoglobulin heavy chain variable region segments. *The EMBO Journal*, 3(6):1209–1219, 1984. doi: 10.1002/J.1460-2075.1984.TB01955.X.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997. doi: 10.1016/B978-1-4832-3211-9.50009-7.

Anand, N., Eguchi, R., Mathews, I. I., Perez, C. P., Derry, A., Altman, R. B., and Huang, P. S. Protein sequence design with a learned potential. *Nature Communications*, 13(1):1–11, 2022. doi: 10.1038/s41467-022-28313-9.

Anderson, S. M., Khalil, A., Uduman, M., Hershberg, U., Louzoun, Y., Haberman, A. M., Kleinstein, S. H., and Shlomchik, M. J. Taking Advantage: High-Affinity B Cells in the Germinal Center Have Lower Death Rates, but Similar Rates of Division, Compared to Low-Affinity Cells. *The Journal of Immunology*, 183(11): 7314–7325, 2009. doi: 10.4049/JIMMUNOL.0902452.

Andreano, E., Nicastri, E., Paciello, I., Pileri, P., Manganaro, N., Piccini, G., Manenti, A., Pantano, E., Kabanova, A., Troisi, M., Vacca, F., Cardamone, D., Santi, C. D., Torres, J. L., Ozorowski, G., Benincasa, L., Jang, H., Genova, C. D., Depau, L., Brunetti, J., Agrati, C., Capobianchi, M. R., Castilletti, C., Emiliozzi, A., Fabbiani, M., Montagnani, F., Bracci, L., Sautto, G., Ross, T. M., Montomoli, E., Temperton, N., Ward, A. B., Sala, C., Ippolito, G., and Rappuoli, R. Extremely potent human monoclonal antibodies from COVID-19 convalescent patients. *Cell*, 184:1821–1835.e16, 2021. doi: 10.1016/j.cell.2021.02.035.

Andrew, S. M. Enzymatic Digestion of Monoclonal Antibodies. *Protein Protocols Handbook, The*, 1047–1052, 2003. doi: 10.1385/1-59259-169-8:1047.

Asaadi, Y., Jouneghani, F. F., Janani, S., and Rahbarizadeh, F. A comprehensive comparison between camelid nanobodies and single chain variable fragments. *Biomarker Research 2021 9:1*, 9(1):1–20, 2021. doi: 10.1186/S40364-021-00332-6.

Bachas, S., Rakocevic, G., Spencer, D., Sastry, A. V., Haile, R., Sutton, J. M., Kasun, G., Stachyra, A., Gutierrez, J. M., Yassine, E., Medjo, B., Blay, V., Kohnert, C., Stanton, J. T., Brown, A., Tijanic, N., McCloskey, C., Viazzo, R., Consbruck, R., Carter, H., Levine, S., Abdulhaqq, S., Shaul, J., Ventura, A. B., Olson, R. S., Yapici, E., Meier, J., McClain, S., Weinstock, M., Hannum, G., Schwartz, A., Gander, M., and Spreafico, R. Antibody optimization enabled by artificial intelligence predictions of binding affinity and naturalness. *bioRxiv*, 2022.08.16.504181, 2022.

Baek, M., Anishchenko, I., Humphreys, I. R., Cong, Q., Baker, D., and DiMaio, F. Efficient and accurate prediction of protein structure using RoseTTAFold2. *bioRxiv*, 2023.05.24.542179, 2023. doi: 10.1101/2023.05.24.542179.

Barlow, K. A., Ó Conchúir, S., Thompson, S., Suresh, P., Lucas, J. E., Heinonen, M., and Kortemme, T. Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation. *Journal of Physical Chemistry B*, 122(21):5389–5399, 2018. doi: 10.1021/acs.jpcb.7b11367.

Beall, R. F., Hwang, T. J., and Kesselheim, A. S. Pre-market development times for biologic versus small-molecule drugs. *Nature Biotechnology*, 37(7):708–711, 2019. doi: 10.1038/s41587-019-0175-2.

Behbahani, Y. M., Laine, E., and Carbone, A. Deep local analysis estimates effects of mutations on protein-protein interactions. *bioRxiv*, 2022. doi: 10.1101/2022.10.09.511484.

Bennett, N. R., Watson, J. L., Ragotte, R. J., Borst, A. J., See, D. L., Weidle, C., Biswas, R., Shrock, E. L., Leung, P. J. Y., Huang, B., Goreshnik, I., Ault, R., Carr, K. D., Singer, B., Criswell, C., Vafeados, D., Sanchez, M. G., Kim, H. M., Torres, S. V., Chan, S., and Baker, D. Atomically accurate de novo design of single-domain antibodies. *bioRxiv*, 2024.03.14.585103, 2024. doi: 10.1101/2024.03.14.585103.

Bernett, J., Blumenthal, D. B., and List, M. Cracking the black box of deep sequence-based protein-protein interaction prediction. *bioRxiv*, 2023. doi: 10.1101/2023.01.18.524543.

Besse, P., Castets-Renard, C., Garivier, A., and Loubes, J.-M. Can everyday AI be ethical. Fairness of Machine Learning Algorithms. *arXiv*, 2018. doi: 10.48550/arXiv.1810.01729.

Bi, V., Jawa, V., Joubert, M. K., Kaliyaperumal, A., Eakin, C., Richmond, K., Pan, O., Sun, J., Hokom, M., Goletz, T. J., Wypych, J., Zhou, L., Kerwin, B. A., Narhi, L. O., and Arora, T. Development of a human antibody tolerant mouse model to assess the immunogenicity risk due to aggregated biotherapeutics. *Journal of Pharmaceutical Sciences*, 102(10):3545–3555, 2013. doi: 10.1002/jps.23663.

Bizebard, T., Gigant, B., Rigolet, P., Rasmussen, B., Diat, O., Bösecke, P., Wharton, S. A., Skehel, J. J., and Knossow, M. Structure of influenza virus haemagglutinin complexed with a neutralizing antibody. *Nature 1995*, 376(6535):92–94, 1995. doi: 10.1038/376092a0.

Bradbury, A. R., Sidhu, S., Dübel, S., and McCafferty, J. Beyond natural antibodies: the power of in vitro display technologies. *Nature Biotechnology*, 29(3):245–254, 2011. doi: 10.1038/nbt.1791.

Braden, B. C., Souchon, H., Eiselé, J.-L., Bentley, G. A., Bhat, T., Navaza, J., and Poljak, R. J. Three-dimensional structures of the free and the antigen-complexed

fab from monoclonal anti-lysozyme antibody d44.1. *Journal of Molecular Biology*, 243(4):767–781, 1994. doi: 10.1016/0022-2836(94)90046-9.

Bretscher, P. and Cohn, M. A Theory of Self-Nonself Discrimination. *Science*, 169 (3950):1042–1049, 1970. doi: 10.1126/SCIENCE.169.3950.1042.

Briney, B., Inderbitzin, A., Joyce, C., and Burton, D. R. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*, 566(7744): 393–397, 2019. doi: 10.1038/s41586-019-0879-y.

Buel, G. R. and Walters, K. J. Can AlphaFold2 predict the impact of missense mutations on structure? *Nature Structural & Molecular Biology*, 29(1):1–2, 2022. doi: 10.1038/s41594-021-00714-2.

Bullen, G., Galson, J. D., Hall, G., Villar, P., Moreels, L., Ledsgaard, L., Mattiuzzo, G., Bentley, E. M., Masters, E. W., Tang, D., Millett, S., Tongue, D., Brown, R., Diamantopoulos, I., Parthiban, K., Tebbutt, C., Leah, R., Chaitanya, K., Ergueta-Carballo, S., Pazeraitis, D., Surade, S. B., Ashiru, O., Crippa, L., Cowan, R., Bowler, M. W., Campbell, J. I., Lee, W. Y. J., Carr, M. D., Matthews, D., Pfeffer, P., Hufton, S. E., Sawmynaden, K., Osbourn, J., McCafferty, J., and Karatt-Vellatt, A. Cross-Reactive SARS-CoV-2 Neutralizing Antibodies From Deep Mining of Early Patient Responses. *Frontiers in Immunology*, 12(June):678570, 2021. doi: 10.3389/fimmu.2021.678570.

Callaway, E. 'A landmark moment': scientists use AI to design antibodies from scratch. *Nature*, 2024. doi: 10.1038/D41586-024-00846-7.

Carter, P. J. and Rajpal, A. Designing antibodies as therapeutics. *Cell*, 185(15): 2789–2805, 2022. doi: 10.1016/J.CELL.2022.05.029.

Casellas, R., Tien-An Yang Shih, Kleinewietfeld, M., Rakonjac, J., Nemazee, D., Rajewsky, K., and Nussenzweig, M. C. Contribution of receptor editing to the antibody repertoire. *Science*, 291(5508):1541–1544, 2001.

Chao, G., Lau, W. L., Hackel, B. J., Sazinsky, S. L., Lippow, S. M., and Wittrup, K. D. Isolating and engineering human antibodies using yeast surface display. *Nature Protocols*, 1(2):755–768, 2006. doi: 10.1038/nprot.2006.94.

Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. *35th International Conference on Machine Learning*, 2:1386–1418, 2018.

Chen, Y., Wiesmann, C., Fuh, G., Li, B., Christinger, H. W., McKay, P., de Vos, A. M., and Lowman, H. B. Selection and analysis of an optimized anti-vegf antibody: crystal structure of an affinity-matured fab in complex with antigen11edited by i. a. wilson. *Journal of Molecular Biology*, 293(4):865–881, 1999. doi: 10.1006/jmbi.1999.3192.

Chinery, L., Wahome, N., Moal, I., and Deane, C. M. Paragraph—antibody paratope prediction using graph neural networks with minimal feature vectors. *Bioinformatics*, 39(1):btac732, 2022. doi: 10.1093/bioinformatics/btac732.

Chinery, L., Hummer, A. M., Mehta, B. B., Akbar, R., Rawat, P., Slabodkin, A., Quy, K. L., Lund-Johansen, F., Greiff, V., Jeliazkov, J. R., and Deane, C. M. Baselining the Buzz Trastuzumab-HER2 Affinity, and Beyond. *bioRxiv*, 2024. doi: 10.1101/2024.03.26.586756.

Cho, H.-S., Mason, K., Ramyar, K. X., Stanley, A. M., Gabelli, S. B., Denney, D. W., and Leahy, D. J. Structure of the extracellular region of her2 alone and in complex with the herceptin fab. *Nature*, 421(6924):756–760, 2003. doi: 10.1038/nature01392.

Choi, Y., Hua, C., Sentman, C. L., Ackerman, M. E., and Bailey-Kellogg, C. Antibody humanization by structure-based computational protein design. *mAbs*, 7(6):1045–1057, 2015. doi: 10.1080/19420862.2015.1076600.

Chothia, C. and Lesk, A. M. Canonical structures for the hypervariable regions of immunoglobulins. *Journal of Molecular Biology*, 196(4):901–917, 1987. doi: 10.1016/0022-2836(87)90412-8.

Christen, M., Hünenberger, P. H., Bakowies, D., Baron, R., Bürgi, R., Geerke, D. P., Heinz, T. N., Kastenholz, M. A., Kräutler, V., Oostenbrink, C., Peter, C., Trzesniak, D., and Van Gunsteren, W. F. The GROMOS software for biomolecular simulation: GROMOS05. *Journal of Computational Chemistry*, 26(16):1719–1751, 2005. doi: 10.1002/JCC.20303.

Chu, A. E., Cheng, L., Nesr, G. E., Xu, M., and Huang, P.-S. An all-atom protein generative model. *bioRxiv*, 2023. doi: 10.1101/2023.05.24.542194.

Chungyoun, M., Ruffolo, J., and Gray, J. FLAb: Benchmarking deep learning methods for antibody fitness prediction. *NeurIPS Machine Learning for Structural Biology Workshop*, 2023.

Clavero-Álvarez, A., Di Mambro, T., Perez-Gaviro, S., Magnani, M., and Bruscolini, P. Humanization of Antibodies using a Statistical Inference Approach. *Scientific Reports*, 8(1):1–11, 2018. doi: 10.1038/s41598-018-32986-y.

Coffman, R. L., Sher, A., and Seder, R. A. Vaccine adjuvants: Putting innate immunity to work. *Immunity*, 33(4):492–503, 2010. doi: 10.1016/J.IMMUNI.2010.10.002/ASSET/29566C4A-2698-48A8-8CA8-1983B456EF09/MAIN.ASSETS/GR1.JPG.

Corrie, B. D., Marthandan, N., Zimonja, B., Jaglale, J., Zhou, Y., Barr, E., Knoetze, N., Breden, F. M., Christley, S., Scott, J. K., Cowell, L. G., and Breden, F. iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunological Reviews*, 284(1):24–41, 2018. doi: 10.1111/IMR.12666.

Crusius, D., Cipcigan, F., and Biggin, P. C. Are we fitting data or noise? Analysing the predictive power of commonly used datasets in drug-, materials-, and molecular-discovery. *ChemRxiv*, 2024. doi: 10.26434/CHEMRXIV-2024-Z0PZ7.

Cunningham, O., Scott, M., Zhou, Z. S., and Finlay, W. J. Polyreactivity and polyspecificity in therapeutic antibody development: risk factors for failure in preclinical and clinical development campaigns. *mAbs*, 13(1), 2021. doi: 10.1080/19420862.2021.1999195.

Cutting, D., Dreyer, F. A., Errington, D., Schneider, C., and Deane, C. M. De novo antibody design with se(3) diffusion. *arXiv*, 2024. doi: 10.48550/arXiv.2405.07622.

Dai, E., Aggarwal, C., and Wang, S. NRGNN: Learning a Label Noise-Resistant Graph Neural Network on Sparsely and Noisily Labeled Graphs. *27th Conference on Knowledge Discovery and Data Mining*, 10, 2021. doi: 10.1145/3447548.3467364.

Darling, R. J. and Brault, P. A. Kinetic Exclusion Assay Technology: Characterization of Molecular Interactions. *Assay and Drug Development Technologies*, 2(6): 647–657, 2005. doi: 10.1089/ADT.2004.2.647.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P., and Baker, D. Robust deep learning–based

protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187.

Dauzhenka, T., Kundrotas, P. J., and Vakser, I. A. Computational Feasibility of an Exhaustive Search of Side-Chain Conformations in Protein-Protein Docking. *Journal of Computational Chemistry*, 39(24):2012–2021, 2018. doi: 10.1002/JCC.25381.

Davis, J. D., Bravo Padros, M., Conrado, D. J., Ganguly, S., Guan, X., Hassan, H. E., Hazra, A., Irvin, S. C., Jayachandran, P., Kosloski, M. P., Lin, K. J., Mukherjee, K., Paccaly, A., Papachristos, A., Partridge, M. A., Prabhu, S., Visich, J., Welf, E. S., Xu, X., Zhao, A., and Zhu, M. Subcutaneous Administration of Monoclonal Antibodies: Pharmacology, Delivery, Immunogenicity, and Learnings From Applications to Clinical Development. *Clinical Pharmacology & Therapeutics*, 115(3):422–439, 2024. doi: 10.1002/CPT.3150.

de Bruin, R., Spelt, K., Mol, J., Koes, R., and Quattrocchio, F. Selection of high-affinity phage antibodies from phage display libraries. *Nature Biotechnology*, 17 (C):397–399, 1999.

Delbos, F., Aoufouchi, S., Faili, A., Weill, J. C., and Reynaud, C. A. DNA polymerase η is the sole contributor of A/T modifications during immunoglobulin gene hypermutation in the mouse. *Journal of Experimental Medicine*, 204(1):17–23, 2007. doi: 10.1084/JEM.20062131.

Desiderio, S. V., Yancopoulos, G. D., Paskind, M., Thomas, E., Boss, M. A., Landau, N., Alt, F. W., and Baltimore, D. Insertion of N regions into heavy-chain genes is correlated with expression of terminal deoxytransferase in B cells. *Nature 1984*, 311(5988):752–755, 1984. doi: 10.1038/311752a0.

Di Noia, J. M. and Neuberger, M. S. Molecular mechanisms of antibody somatic hypermutation. *Annual Review of Biochemistry*, 76:1–22, 2007. doi: 10.1146/annurev.biochem.76.061705.090740.

Dieckhaus, H., Brocidiacono, M., Randolph, N. Z., and Kuhlman, B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proceedings of the National Academy of Sciences*, 121(6):e2314853121, 2024.

Douzi, B. Protein–Protein Interactions: Surface Plasmon Resonance. *Methods in Molecular Biology*, 1615:257–275, 2017. doi: 10.1007/978-1-4939-7033-9_21.

Dreyer, F. A., Cutting, D., Schneider, C., Kenlay, H., and Deane, C. M. Inverse folding for antibody sequence design using deep learning. *ICML Workshop on Computational Biology*, 2023.

Dunbar, J., Fuchs, A., Shi, J., and Deane, C. M. ABangle: characterising the VH–VL orientation in antibodies. *Protein Engineering, Design and Selection*, 26(10):611–620, 2013. doi: 10.1093/PROTEIN/GZT020.

Dunbar, J. and Deane, C. M. ANARCI: Antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016. doi: 10.1093/bioinformatics/btv552.

Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J., and Deane, C. M. SAbDab: The structural antibody database. *Nucleic Acids Research*, 42:1140–1146, 2014. doi: 10.1093/nar/gkt1043.

Dunnick, W., Hertz, G. Z., Scappino, L., and Gritzmacher, C. DNA sequences at immunoglobulin switch region recombination sites. *Nucleic Acids Research*, 21(3):365–372, 1993. doi: 10.1093/NAR/21.3.365.

Duval, A. and Malliaros, F. D. GraphSVX: Shapley Value Explanations for Graph Neural Networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12976 LNAI: 302–318, 2021. doi: 10.1007/978-3-030-86520-7_19.

Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L. P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., and Pande, V. S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, 13 (7):e1005659, 2017. doi: 10.1371/JOURNAL.PCBI.1005659.

Ehrenmann, F., Kaas, Q., and Lefranc, M. P. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Research*, 38:D301–D307, 2010. doi: 10.1093/NAR/GKP946.

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J., and Hassabis, D. Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021.10.04.463034, 2022. doi: 10.1101/2021.10.04.463034.

Feavers, I. and Walker, B. Functional Antibody Assays. *Methods in Molecular Biology*, 626:199–211, 2010. doi: 10.1007/978-1-60761-585-9_14.

Fernández-Quintero, M. L., Kraml, J., Georges, G., and Liedl, K. R. CDR-H3 loop ensemble in solution – conformational selection upon antibody binding. *mAbs*, 11 (6):1077–1088, 2019. doi: 10.1080/19420862.2019.1618676.

Ferrara, F., Naranjo, L. A., Kumar, S., Gaiotto, T., Mukundan, H., Swanson, B., and Bradbury, A. R. Using Phage and Yeast Display to Select Hundreds of Monoclonal Antibodies: Application to Antigen 85, a Tuberculosis Biomarker. *PLOS ONE*, 7 (11):e49535, 2012. doi: 10.1371/JOURNAL.PONE.0049535.

Ferruz, N., Schmidt, S., and Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1):1–10, 2022. doi: 10.1038/s41467-022-32007-7.

Foote, J. and Winter, G. Antibody framework residues affecting the conformation of the hypervariable loops. *Journal of Molecular Biology*, 224(2):487–499, 1992. doi: 10.1016/0022-2836(92)91010-M.

Forthal, D. N. Functions of Antibodies. *Microbiology spectrum*, 2(4):1, 2015. doi: 10.1128/9781555817411.ch2.

Frenzel, A., Hust, M., and Schirrmann, T. Expression of recombinant antibodies. *Frontiers in Immunology*, 4:51304, 2013. doi: 10.3389/FIMMU.2013.00217.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012. doi: 10.1093/bioinformatics/bts565.

Gao, S. H., Huang, K., Tu, H., and Adler, A. S. Monoclonal antibody humanness score and its applications. *BMC Biotechnology*, 13, 2013. doi: 10.1186/1472-6750-13-55.

Gaudreault, F., Corbeil, C. R., and Sulea, T. Enhanced antibody-antigen structure prediction from molecular docking using AlphaFold2. *Scientific Reports*, 13(1): 1–14, 2023. doi: 10.1038/s41598-023-42090-5.

Geng, C., Vangone, A., and Bonvin, A. M. Exploring the interplay between experimental methods and the performance of predictors of binding affinity change upon mutations in protein complexes. *Protein Engineering, Design and Selection*, 29(8): 291–299, 2016. doi: 10.1093/PROTEIN/GZW020.

Geng, C., Xue, L. C., Roel-Touris, J., and Bonvin, A. M. J. J. Finding the g spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? *WIREs Computational Molecular Science*, 9(5):e1410, 2019. doi: 10.1002/wcms.1410.

Geng, S. B., Cheung, J. K., Narasimhan, C., Shameem, M., and Tessier, P. M. Improving monoclonal antibody selection and engineering using measurements of colloidal protein interactions. *Journal of Pharmaceutical Sciences*, 103(11):3356–3363, 2014. doi: 10.1002/jps.24130.

Gethsiyal Augasta, M. and Kathirvalavakumar, T. Reverse engineering the neural networks for rule extraction in classification problems. *Neural Processing Letters*, 35(2):131–150, 2012. doi: 10.1007/S11063-011-9207-8/METRICS.

Ghosh, I., Gutka, H., Krause, M. E., Clemens, R., and Kashi, R. S. A systematic review of commercial high concentration antibody drug products approved in the US: formulation composition, dosage form design and primary packaging considerations. *mAbs*, 15(1), 2023. doi: 10.1080/19420862.2023.2205540.

Gilfillan, S., Dierich, A., Lemeur, M., Benoist, C., and Mathis, D. Mice Lacking TdT: Mature Animals with an Immature Lymphocyte Repertoire. *Science*, 261(5125): 1175–1178, 1993. doi: 10.1126/SCIENCE.8356452.

Gitlin, A. D., Shulman, Z., and Nussenzweig, M. C. Clonal selection in the germinal centre by regulated proliferation and hypermutation. *Nature*, 509(7502):637–640, 2014. doi: 10.1038/nature13300.

Giudicelli, V., Chaume, D., and Lefranc, M. P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Research*, 33:D256–D261, 2005. doi: 10.1093/NAR/GKI010.

Goodnow, C. C., Crosbie, J., Jorgensen, H., Brink, R. A., and Basten, A. Induction of self-tolerance in mature peripheral B lymphocytes. *Nature 1989*, 342(6248): 385–391, 1989. doi: 10.1038/342385a0.

Gordon, G. L., Raybould, M. I. J., Wong, A., and Deane, C. M. Prospects for the computational humanization of antibodies and nanobodies. *Frontiers in Immunology*, 15, 2024. doi: 10.3389/fimmu.2024.1399438.

Greenfield, E. A. Immunizing Animals. *Cold Spring Harbor Protocols*, 2022(7): pdb.top100180, 2022a. doi: 10.1101/PDB.TOP100180.

Greenfield, E. A. Generating Monoclonal Antibodies. *Cold Spring Harbor Protocols*, 2022(8):pdb.top103036, 2022b. doi: 10.1101/PDB.TOP103036.

Guest, J. D., Vreven, T., Zhou, J., Moal, I., Jeliazkov, J. R., Gray, J. J., Weng, Z., and Pierce, B. G. An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants. *Structure*, 29 (6):606–621.e5, 2021. doi: 10.1016/J.STR.2021.01.005.

Gunn, G. R., Sealey, D. C., Jamali, F., Meibohm, B., Ghosh, S., and Shankar, G. From the bench to clinical practice: understanding the challenges and uncertainties in immunogenicity testing for biopharmaceuticals. *Clinical and Experimental Immunology*, 184(2):137–146, 2016. doi: 10.1111/cei.12742.

Hadfield, T. E., Scantlebury, J., and Deane, C. M. Exploring the ability of machine learning-based virtual screening models to identify the functional groups responsible for binding. *Journal of Cheminformatics*, 15(1):1–15, 2023. doi: 10.1186/S13321-023-00755-3.

Hansel, T. T., Kropshofer, H., Singer, T., Mitchell, J. A., and George, A. J. The safety and side effects of monoclonal antibodies. *Nature Reviews Drug Discovery*, 9(4):325–338, 2010. doi: 10.1038/nrd3003.

Harvey, E. P., Shin, J.-E., Skiba, M. A., Nemeth, G. R., Hurley, J. D., Wellner, A., Shaw, A. Y., Miranda, V. G., Min, J. K., Liu, C. C., Marks, D. S., and Kruse, A. C. An in silico method to assess antibody fragment polyreactivity. *Nature Communications*, 13:7554, 2022. doi: 10.1038/s41467-022-35276-4.

Hemmer, B., Pinilla, C., Gran, B., Vergelli, M., Ling, N., Conlon, P., McFarland, H. F., Houghten, R., and Martin, R. Contribution of Individual Amino Acids Within MHC Molecule or Antigenic Peptide to TCR Ligand Potency. *The Journal of Immunology*, 164(2):861–871, 2000. doi: 10.4049/JIMMUNOL.164.2.861.

Hie, B. L., Shanker, V. R., Xu, D., Bruun, T. U., Weidenbacher, P. A., Tang, S., Wu, W., Pak, J. E., and Kim, P. S. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 42(2):275–283, 2023. doi: 10.1038/s41587-023-01763-2.

Ho, J., Jain, A., and Abbeel, P. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 2020. doi: 10.48550/arXiv.2006.11239.

Høie, M. H., Gade, F. S., Johansen, J., Würtzen, C., Winther, O., Nielsen, M., and Marcatili, P. DiscoTope-3.0: improved B-cell epitope prediction using inverse folding latent representations. *Frontiers in Immunology*, 15:1322712, 2024. doi: 10.3389/FIMMU.2024.1322712.

Homberg, S. K. R., Menke, J., Morris, G. M., and Koch, O. Interpreting Graph Neural Networks with Myerson Values for Cheminformatics Approaches. *ChemRxiv*, 2023. doi: 10.26434/CHEMRXIV-2023-1HXXC.

Honegger, A. and Plückthun, A. Yet Another Numbering Scheme for Immunoglobulin Variable Domains: An Automatic Modeling and Analysis Tool. *Journal of Molecular Biology*, 309(3):657–670, 2001. doi: 10.1006/JMBI.2001.4662.

Honegger, A. and Plückthun, A. Yet another numbering scheme for immunoglobulin variable domains: An automatic modeling and analysis tool. *Journal of Molecular Biology*, 309(3):657–670, 2001. doi: 10.1006/jmbi.2001.4662.

Hötzel, I., Theil, F. P., Bernstein, L. J., Prabhu, S., Deng, R., Quintana, L., Lutman, J., Sibia, R., Chan, P., Bumbaca, D., Fielder, P., Carter, P. J., and Kelley, R. F. A strategy for risk mitigation of antibodies with fast clearance. *mAbs*, 4(6):753–760, 2012. doi: 10.4161/MABS.22189.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. *bioRxiv*, 2022. doi: 10.1101/2022.04.10.487779.

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., Leskovec, J., Barzilay, R., Battaglia, P., Bengio, Y., Bronstein, M., Günnemann, S., Hamilton, W., Jaakkola, T., Jegelka, S., Nickel, M., Re, C., Song, L., Tang, J., Welling, M., and Zemel, R. Open Graph Benchmark: Datasets for Machine Learning on Graphs Steering Committee. *34th Conference on Neural Information Processing Systems*, 2020.

Huang, Q., Yamada, M., Tian, Y., Singh, D., and Chang, Y. GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):6968–6972, 2020. doi: 10.1109/TKDE.2022.3187455.

Hudson, P. J. and Souriau, C. Engineered antibodies. *Nature Medicine*, 9(1):129–134, 2003. doi: 10.1038/nm0103-129.

Hummer, A. M., Schneider, C., Chinery, L., and Deane, C. M. Investigating the volume and diversity of data needed for generalizable antibody-antigen g prediction. *bioRxiv*, 2023. doi: 10.1101/2023.05.17.541222.

Huynh, K. and Partch, C. L. Analysis of Protein Stability and Ligand Interactions by Thermal Shift Assay. *Current Protocols in Protein Science*, 79(1):28.9.1–28.9.14, 2015. doi: 10.1002/0471140864.PS2809S79.

Høie, M. H., Hummer, A. M., Olsen, T. H., Aguilar-Sanjuan, B., Nielsen, M., and Deane, C. M. Antifold: Improved antibody structure-based design using inverse folding, 2024.

Ingraham, J., Garg, V., Barzilay, R., and Jaakkola, T. Generative models for graph-based protein design. *Advances in Neural Information Processing Systems*, 32, 2019.

Ingraham, J. B., Baranov, M., Costello, Z., Barber, K. W., Wang, W., Ismail, A., Frappier, V., Lord, D. M., Ng-Thow-Hing, C., Van Vlack, E. R., Tie, S., Xue, V., Cowles, S. C., Leung, A., Rodrigues, J. V., Morales-Perez, C. L., Ayoub, A. M., Green, R., Puentes, K., Oplinger, F., Panwar, N. V., Obermeyer, F., Root, A. R., Beam, A. L., Poelwijk, F. J., and Grigoryan, G. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023. doi: 10.1038/s41586-023-06728-8.

Iwasato, T., Shimizu, A., Honjo, T., and Yamagishi, H. Circular DNA is excised by immunoglobulin class switch recombination. *Cell*, 62(1):143–149, 1990. doi: 10.1016/0092-8674(90)90248-D.

Jain, R. K. Physiological Barriers to Delivery of Monoclonal Antibodies and Other Macromolecules in Tumors. *Cancer Research*, 1990.

Jankauskaite, J., Jiménez-García, B., Dapkunas, J., Fernández-Recio, J., and Moal, I. H. SKEMPI 2.0: An updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019. doi: 10.1093/bioinformatics/bty635.

Jarmoskaite, I., AlSadhan, I., Vaidyanathan, P. P., and Herschlag, D. How to measure and evaluate binding affinities. *eLife*, 9:e57264, 2020. doi: 10.7554/eLife.57264.

Jha, K., Saha, S., and Singh, H. Prediction of protein–protein interaction using graph neural networks. *Scientific Reports*, 12(1):1–12, 2022. doi: 10.1038/s41598-022-12201-9.

Jiang, S., Hillyer, C., and Du, L. Neutralizing Antibodies against SARS-CoV-2 and Other Human Coronaviruses. *Trends in Immunology*, 41(5):355–359, 2020. doi: 10.1016/j.it.2020.03.007.

Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L., and Dror, R. Learning from protein structure with geometric vector perceptrons. *International Conference on Learning Representations*, 2021. doi: 10.48550/arXiv.2009.01411.

Jolly, C. J., Wagner, S. D., Rada, C., Klix, N., Milstein, C., and Neuberger, M. S. The targeting of somatic hypermutation. *Seminars in Immunology*, 8(3):159–168, 1996. doi: 10.1006/SMIM.1996.0020.

Jones, P. T., Dear, P. H., Foote, J., Neuberger, M. S., and Winter, G. Replacing the complementarity- determining regions in a human antibody with those from a mouse. *Nature*, 321:522–525, 1986.

Jones, T. D., Carter, P. J., Plückthun, A., Vásquez, M., Holgate, R. G., Hötzel, I., Popplewell, A. G., Parren, P. W., Enzelberger, M., Rademaker, H. J., Clark, M. R., Lowe, D. C., Dahiyat, B. I., Smith, V., Lambert, J. M., Wu, H., Reilly, M., Haurum,

J. S., Dübel, S., Huston, J. S., Schirrmann, T., Janssen, R. A., Steegmaier, M., Gross, J. A., Bradbury, A. R., Burton, D. R., Dimitrov, D. S., Chester, K. A., Glennie, M. J., Davies, J., Walker, A., Martin, S., McCafferty, J., and Baker, M. P. The inns and outs of antibody nonproprietary names. *mAbs*, 8(1):1–9, 2016. doi: 10.1080/19420862.2015.1114320.

Jovčevska, I. and Muyldermans, S. The Therapeutic Potential of Nanobodies. *BioDrugs*, 34(1):11–26, 2020. doi: 10.1007/s40259-019-00392-z.

Jubb, H. C., Higueruelo, A. P., Ochoa-Montaño, B., Pitt, W. R., Ascher, D. B., and Blundell, T. L. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of Molecular Biology*, 429(3):365–371, 2017. doi: 10.1016/j.jmb.2016.12.004.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Jung, D., Giallourakis, C., Mostoslavsky, R., and Alt, F. W. Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annual Review of Immunology*, 24(Volume 24, 2006):541–570, 2006. doi: 10.1146/ANNUREV.IMMUNOL.23.021704.115830.

Kang, B., Lijffijt, J., and De Bie, T. ExplaiNE: An Approach for Explaining Network Embedding-based Link Predictions. *arXiv*, 2019. doi: 10.48550/arXiv.1904.12694.

Kaplon, H. and Reichert, J. M. Antibodies to watch in 2019. *mAbs*, 11(2):219–238, 2019. doi: 10.1080/19420862.2018.1556465.

Kästner, J. Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(6):932–942, 2011. doi: 10.1002/WCMS.66.

Kelow, S., Faezov, B., Xu, Q., Parker, M., Adolf-Bryfogle, J., and Dunbrack, R. L. A penultimate classification of canonical antibody CDR conformations. *bioRxiv*, 2022.10.12.511988, 2022. doi: 10.1101/2022.10.12.511988.

Kenlay, H., Dreyer, F. A., Kovaltsuk, A., Miketa, D., Pires, D., and Deane, C. M. Large scale paired antibody language models. *arXiv*, 2024. doi: 10.48550/arXiv.2403.17889.

Kerfoot, S. M., Yaari, G., Patel, J. R., Johnson, K. L., Gonzalez, D. G., Kleinstein, S. H., and Haberman, A. M. Germinal Center B Cell and T Follicular Helper Cell Development Initiates in the Interfollicular Zone. *Immunity*, 34(6):947–960, 2011. doi: 10.1016/j.immuni.2011.03.024.

Kettleborough, C. A., Saldanha, J., Heath, V. J., Morrison, C. J., and Bendig, M. M. Humanization of a mouse monoclonal antibody by CDR-grafting: The importance of framework residues on loop conformation. *Protein Engineering, Design and Selection*, 4(7):773–783, 1991. doi: 10.1093/protein/4.7.773.

Khavrutskii, I. V. and Wallqvist, A. Improved Binding Free Energy Predictions from Single-Reference Thermodynamic Integration Augmented with Hamiltonian Replica Exchange. *Journal of Chemical Theory and Computation*, 7(9):3001–3011, 2011. doi: 10.1021/CT2003786.

Kijanka, G., Bee, J. S., Korman, S. A., Wu, Y., Roskos, L. K., Schenerman, M. A., Slütter, B., and Jiskoot, W. Submicron Size Particles of a Murine Monoclonal

Antibody Are More Immunogenic Than Soluble Oligomers or Micron Size Particles Upon Subcutaneous Administration in Mice. *Journal of Pharmaceutical Sciences*, 107(11):2847–2859, 2018. doi: 10.1016/j.xphs.2018.06.029.

King, D. J., Turner, A., Farnsworth, A. P. H., Adair, J. R., Owens, R. J., Pedley, R. B., Baldock, D., Proudfoot, K. A., Lawson, A. D. G., Beeley, N. R. A., Millar, K., Millican, T. A., Boyce, B. A., Antoniw, P., Mountain, A., Begent, R. H. J., Shochat, D., and Yarranton, G. T. Improved tumor targeting with chemically cross-linked recombinant antibody fragments. *Cancer Research*, 54(23):6176–6185, 1994.

Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *The Journal of Chemical Physics*, 3(5):300–313, 1935. doi: 10.1063/1.1749657.

Klarenbeek, A., El Mazouari, K., Desmyter, A., Blanchetot, C., Hultberg, A., de Jonge, N., Roovers, R. C., Cambillau, C., Spinelli, S., Del-Favero, J., Verrips, T., de Haard, H. J., and Achour, I. Camelid Ig V genes reveal significant human homology not seen in therapeutic target genes, providing for a powerful therapeutic antibody platform. *mAbs*, 7(4):693–706, 2015. doi: 10.1080/19420862.2015.1046648.

Klimovich, P. V., Shirts, M. R., and Mobley, D. L. Guidelines for the analysis of free energy calculations. *Journal of Computer-Aided Molecular Design*, 29(5):397–411, 2015. doi: 10.1007/S10822-015-9840-9/METRICS.

Köhler, G. and Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature*, 256(5517):495–497, 1975a. doi: 10.1038/256495a0.

Köhler, G. and Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature*, 256:495–497, 1975b.

Komori, T., Okada, A., Stewart, V., and Alt, F. W. Lack of N Regions in Antigen Receptor Variable Region Genes of TdT-Deficient Lymphocytes. *Science*, 261 (5125):1171–1175, 1993. doi: 10.1126/SCIENCE.8356451.

Kong, X., Huang, W., and Liu, Y. Conditional Antibody Design as 3D Equivariant Graph Translation. *11th International Conference on Learning Representations*, 2022.

Kovaltsuk, A., Leem, J., Kelm, S., Snowden, J., Deane, C. M., and Krawczyk, K. Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. *The Journal of Immunology*, 201(8):2502–2509, 2018. doi: 10.4049/jimmunol.1800708.

Krishna, R., Wang, J., Ahern, W., Sturmfels, P., Venkatesh, P., Kalvet, I., Lee, G. R., Morey-Burrows, F. S., Anishchenko, I., Humphreys, I. R., McHugh, R., Vafeados, D., Li, X., Sutherland, G. A., Hitchcock, A., Hunter, C. N., Baek, M., DiMaio, F., and Baker, D. Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom. *bioRxiv*, 2023.10.09.561603, 2023. doi: 10.1101/2023.10.09.561603.

Kudo, A. and Melchers, F. A second gene, VpreB in the lambda 5 locus of the mouse, which appears to be selectively expressed in pre-B lymphocytes. *The EMBO Journal*, 6(8):2267–2272, 1987. doi: 10.1002/J.1460-2075.1987.TB02500.X.

Kügler, J., Tomszak, F., Frenzel, A., and Hust, M. Construction of Human Immune and Naive scFv Libraries. *Methods in Molecular Biology*, 1701:3–24, 2018. doi: 10.1007/978-1-4939-7447-4_1.

Lakkaraju, H., Research, M., Caruana, R., and Leskovec, J. Interpretable Explorable Approximations of Black Box Models. *KDD'17, Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

Landrum, G. A. and Riniker, S. Combining IC50 or Ki Values from Different Sources Is a Source of Significant Noise. *Journal of Chemical Information and Modeling*, 64(5):1560–1567, 2024. doi: 10.1021/ACS.JCIM.4C00049.

Laustsen, A. H., Greiff, V., Karatt-Vellatt, A., Muyldermans, S., and Jenkins, T. P. Animal Immunization, in Vitro Display Technologies, and Machine Learning for Antibody Discovery. *Trends in Biotechnology*, 39(12):1263–1273, 2021. doi: 10.1016/j.tibtech.2021.03.003.

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K. W., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y.-E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., and Bradley, P. Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology*, 487:545–574, 2011. doi: 10.1016/B978-0-12-381270-4.00019-6.

Lee, B. and Richards, F. M. The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology*, 55(3), 1971. doi: 10.1016/0022-2836(71)90324-X.

Lee, E. C., Liang, Q., Ali, H., Bayliss, L., Beasley, A., Bloomfield-Gerdes, T., Bonoli, L., Brown, R., Campbell, J., Carpenter, A., Chalk, S., Davis, A., England, N., Fane-Dremucheva, A., Franz, B., Germaschewski, V., Holmes, H., Holmes, S., Kirby, I., Kosmac, M., Legent, A., Lui, H., Manin, A., O'Leary, S., Paterson, J., Sciarrillo, R., Speak, A., Spensberger, D., Tuffery, L., Waddell, N., Wang, W., Wells, S., Wong, V., Wood, A., Owen, M. J., Friedrich, G. A., and Bradley, A. Complete humanization of the mouse immunoglobulin loci enables efficient therapeutic antibody discovery. *Nature Biotechnology*, 32(4):356–363, 2014. doi: 10.1038/nbt.2825.

Leem, J., Mitchell, L. S., Farmery, J. H., Barton, J., and Galson, J. D. Deciphering the language of antibodies using self-supervised learning. *Patterns*, 3(7):100513, 2022. doi: 10.1016/J.PATTER.2022.100513.

Leeman, M., Choi, J., Hansson, S., Storm, M. U., and Nilsson, L. Proteins and antibodies in serum, plasma, and whole blood—size characterization using asymmetrical flow field-flow fractionation (AF4). *Analytical and Bioanalytical Chemistry*, 410(20):4867–4873, 2018. doi: 10.1007/S00216-018-1127-2.

Lefranc, M. P. Unique database numbering system for immunogenetic analysis; Current literature. *Immunology Today*, 18(11):509, 1997. doi: 10.1016/S0167-5699(97)01163-8.

Lefranc, M.-P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V., and Lefranc, G. Imgt unique numbering for immunoglobulin and t cell receptor variable domains and ig superfamily v-like domains. *Developmental Comparative Immunology*, 27(1):55–77, 2003. doi: 10.1016/S0145-305X(02)00039-3.

Leman, J. K., Weitzner, B. D., Lewis, S. M., Adolf-Bryfogle, J., Alam, N., Alford, R. F., Aprahamian, M., Baker, D., Barlow, K. A., Barth, P., Basanta, B., Bender, B. J., Blacklock, K., Bonet, J., Boyken, S. E., Bradley, P., Bystroff, C., Conway, P., Cooper, S., Correia, B. E., Coventry, B., Das, R., De Jong, R. M., DiMaio, F., Dsilva, L., Dunbrack, R., Ford, A. S., Frenz, B., Fu, D. Y., Geniesse, C., Goldschmidt, L., Gowthaman, R., Gray, J. J., Gront, D., Guffy, S., Horowitz, S., Huang, P.-S., Huber, T., Jacobs, T. M., Jeliazkov, J. R., Johnson, D. K., Kappel, K., Karanicolas, J., Khakzad, H., Khar, K. R., Khare, S. D., Khatib, F., Khramushin, A., King, I. C., Kleffner, R., Koepnick, B., Kortemme, T., Kuenze, G., Kuhlman, B., Kuroda, D., Labonte, J. W., Lai, J. K., Lapidoth, G., Leaver-Fay, A., Lindert, S., Linsky, T., London, N., Lubin, J. H., Lyskov, S., Maguire, J., Malmström, L.,

Marcos, E., Marcu, O., Marze, N. A., Meiler, J., Moretti, R., Mulligan, V. K., Nerli, S., Norn, C., Ó'Conchúir, S., Ollikainen, N., Ovchinnikov, S., Pacella, M. S., Pan, X., Park, H., Pavlovicz, R. E., Pethe, M., Pierce, B. G., Pilla, K. B., Raveh, B., Renfrew, P. D., Burman, S. S. R., Rubenstein, A., Sauer, M. F., Scheck, A., Schief, W., Schueler-Furman, O., Sedan, Y., Sevy, A. M., Sgourakis, N. G., Shi, L., Siegel, J. B., Silva, D.-A., Smith, S., Song, Y., Stein, A., Szegedy, M., Teets, F. D., Thyme, S. B., Wang, R. Y.-R., Watkins, A., Zimmerman, L., and Bonneau, R. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature Methods*, 17(7):665–680, 2020. doi: 10.1038/s41592-020-0848-2.

Lensink, M. F., Nadzirin, N., Velankar, S., and Wodak, S. J. Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition. *Proteins: Structure, Function, and Bioinformatics*, 88(8):916–938, 2020. doi: 10.1002/PROT.25870.

Lensink, M. F., Brysbaert, G., Mauri, T., Nadzirin, N., Velankar, S., Chaleil, R. A., Clarence, T., Bates, P. A., Kong, R., Liu, B., Yang, G., Liu, M., Shi, H., Lu, X., Chang, S., Roy, R. S., Quadir, F., Liu, J., Cheng, J., Antoniak, A., Czaplewski, C., Giełdoń, A., Kogut, M., Lipska, A. G., Liwo, A., Lubecka, E. A., Maszota-Zieleniak, M., Sieradzan, A. K., Ślusarz, R., Wesołowski, P. A., Zięba, K., Del Carpio Muñoz, C. A., Ichiishi, E., Harmalkar, A., Gray, J. J., Bonvin, A. M., Ambrosetti, F., Vargas Honorato, R., Jandova, Z., Jiménez-García, B., Koukos, P. I., Van Keulen, S., Van Noort, C. W., Réau, M., Roel-Touris, J., Kotelnikov, S., Padhorny, D., Porter, K. A., Alekseenko, A., Ignatov, M., Desta, I., Ashizawa, R., Sun, Z., Ghani, U., Hashemi, N., Vajda, S., Kozakov, D., Rosell, M., Rodríguez-Lumbreras, L. A., Fernandez-Recio, J., Karczynska, A., Grudinin, S., Yan, Y., Li, H., Lin, P., Huang, S. Y., Christoffer, C., Terashi, G., Verburgt, J., Sarkar, D., Aderinwale, T., Wang, X., Kihara, D., Nakamura, T., Hanazono, Y., Gowthaman,

R., Guest, J. D., Yin, R., Taherzadeh, G., Pierce, B. G., Barradas-Bautista, D., Cao, Z., Cavallo, L., Oliva, R., Sun, Y., Zhu, S., Shen, Y., Park, T., Woo, H., Yang, J., Kwon, S., Won, J., Seok, C., Kiyota, Y., Kobayashi, S., Harada, Y., Takeda-Shitaka, M., Kundrotas, P. J., Singh, A., Vakser, I. A., Dapkūnas, J., Olechnovič, K., Venclovas, Č., Duan, R., Qiu, L., Xu, X., Zhang, S., Zou, X., and Wodak, S. J. Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1800–1823, 2021. doi: 10.1002/PROT.26222.

Levine, M. H., Haberman, A. M., Sant'Angelo, D. B., Hannum, L. G., Cancro, M. P., Janeway, C. A., and Shlomchik, M. J. A B-cell receptor-specific selection step governs immature to mature B cell differentiation. *Proceedings of the National Academy of Sciences*, 97(6):2743–2748, 2000. doi: 10.1073/PNAS.050552997.

Li, X., Duan, X., Yang, K., Zhang, W., Zhang, C., Fu, L., Ren, Z., Wang, C., Wu, J., Lu, R., Ye, Y., He, M., Nie, C., Yang, N., Wang, J., Yang, H., Liu, X., and Tan, W. Comparative analysis of immune repertoires between bactrian Camel's conventional and heavy-chain antibodies. *PLOS ONE*, 11(9):1–15, 2016. doi: 10.1371/journal.pone.0161801.

Li, Y., Huang, Y., Swaminathan, C. P., Smith-Gill, S. J., and Mariuzza, R. A. Magnitude of the hydrophobic effect at central versus peripheral sites in protein-protein interfaces. *Structure*, 13(2):297–307, 2005. doi: 10.1016/j.str.2004.12.012.

Lieber, M. R., Hesse, J. E., Mizuuchi, K., and Gellert, M. Developmental stage specificity of the lymphoid V(D)J recombination activity. *Genes Development*, 1 (8):751–761, 1987. doi: 10.1101/GAD.1.8.751.

Lin, K. and Wu, G. Isothermal Titration Calorimetry Assays to Measure Bind-

ing Affinities In Vitro. *Methods in Molecular Biology*, 1893:257–272, 2019. doi: 10.1007/978-1-4939-8910-2_19.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Liu, X., Luo, Y., Li, P., Song, S., and Peng, J. Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. *PLOS Computational Biology*, 17(8):1–28, 2021. doi: 10.1371/journal.pcbi.1009284.

Loder, F., Mutschler, B., Ray, R. J., Paige, C. J., Sideras, P., Torres, R., Lamers, M. C., and Carsetti, R. B Cell Development in the Spleen Takes Place in Discrete Steps and Is Determined by the Quality of B Cell Receptor–Derived Signals. *Journal of Experimental Medicine*, 190(1):75–90, 1999. doi: 10.1084/JEM.190.1.75.

Lucic, A., ter Hoeve, M., Tolomei, G., de Rijke, M., and Silvestri, F. CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks. *Proceedings of Machine Learning Research*, 151:4499–4511, 2021. doi: 10.48550/arXiv.2102.03322.

Lundahl, M. L., Fogli, S., Colavita, P. E., and Scanlan, E. M. Aggregation of protein therapeutics enhances their immunogenicity: causes and mitigation strategies. *RSC Chemical Biology*, 2(4):1004–1020, 2021. doi: 10.1039/D1CB00067E.

Lundberg, S. M. and Lee, S. I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 2017-December:4766–4775, 2017. doi: 10.48550/arXiv.1705.07874.

Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H., and Zhang, X. Parameterized Explainer for Graph Neural Network. *34th Conference on Neural Information Processing Systems*, 2020. doi: 10.48550/arXiv.2011.04573.

Luo, S., Su, Y., Peng, X., Wang, S., Peng, J., Ma, J., Research, H., and Allen, P. G. Antigen-Specific Antibody Design and Optimization with Diffusion-Based Generative Models for Protein Structures. *bioRxiv*, 2022. doi: 10.1101/2022.07.10.499510.

MacCallum, R. M., Martin, A. C., and Thornton, J. M. Antibody-antigen Interactions: Contact Analysis and Binding Site Topography. *Journal of Molecular Biology*, 262(5):732–745, 1996. doi: 10.1006/JMBI.1996.0548.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., Fraser, J. S., and Naik, N. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023. doi: 10.1038/s41587-022-01618-2.

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8): 3696–3713, 2015. doi: 10.1021/ACS.JCTC.5B00255.

Makowski, E. K., Wu, L., Desai, A. A., and Tessier, P. M. Highly sensitive detection of antibody nonspecific interactions using flow cytometry. *mAbs*, 13(1), 2021. doi: 10.1080/19420862.2021.1951426.

Makowski, E. K., Kinnunen, P. C., Huang, J., Wu, L., Smith, M. D., Wang, T., Desai, A. A., Streu, C. N., Zhang, Y., Zupancic, J. M., Schardt, J. S., Linderman, J. J., and Tessier, P. M. Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nature Communications*, 13(1), 2022. doi: 10.1038/s41467-022-31457-3.

Makowski, E. K., Wang, T., Zupancic, J. M., Huang, J., Wu, L., Schardt, J. S., De Groot, A. S., Elkins, S. L., Martin, W. D., and Tessier, P. M. Optimization of therapeutic antibodies for reduced self-association and non-specific binding via interpretable machine learning. *Nature Biomedical Engineering*, 8:45–56, 2023. doi: 10.1038/s41551-023-01074-6.

Makowski, E. K., Chen, H. T., Wang, T., Wu, L., Huang, J., Mock, M., Underhill, P., Pelegri-O'Day, E., Maglalang, E., Winters, D., and Tessier, P. M. Reduction of monoclonal antibody viscosity using interpretable machine learning. *mAbs*, 16(1), 2024. doi: 10.1080/19420862.2024.2303781.

Manz, R. A., Hauser, A. E., Hiepe, F., and Radbruch, A. Maintenance of serum antibody levels. *Annual Review of Immunology*, 23:367–386, 2005. doi: 10.1146/ANNUREV.IMMUNOL.23.021704.115723.

Mariani, V., Biasini, M., Barbato, A., and Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013. doi: 10.1093/BIOINFORMATICS/BTT473.

Marks, C., Hummer, A. M., Chin, M., and Deane, C. M. Humanization of antibodies using a machine learning approach on large-scale repertoire data. *Bioinformatics*, 37(22):4041–4047, 2021. doi: 10.1093/bioinformatics/btab434.

Martinkus, K., Ludwiczak, J., Cho, K., Liang, W.-C., Lafrance-Vanasse, J., Hotzel, I., Rajpal, A., Wu, Y., Bonneau, R., Gligorijevic, V., and Loukas, A. AbDiffuser: Full-Atom Generation of in vitro Functioning Antibodies. *Advances Neural Information Processing Systems*, 2023.

Martomo, S. A., Yang, W. W., Wersto, R. P., Ohkumo, T., Kondo, Y., Yokoi, M., Masutani, C., Hanaoka, F., and Gearhart, P. J. Different mutation signatures in

DNA polymerase η- and MSH6-deficient mice suggest separate roles in antibody diversification. *Proceedings of the National Academy of Sciences*, 102(24):8656–8661, 2005. doi: 10.1073/PNAS.0501852102.

Mason, D. M., Friedensohn, S., Weber, C. R., Jordi, C., Wagner, B., Meng, S. M., Ehling, R. A., Bonati, L., Dahinden, J., Gainza, P., Correia, B. E., and Reddy, S. T. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nature Biomedical Engineering*, 5(6):600–612, 2021. doi: 10.1038/s41551-021-00699-9.

Mastropietro, A., Pasculli, G., Feldmann, C., Rodríguez-Pérez, R., and Bajorath, J. EdgeSHAPer: Bond-centric Shapley value-based explanation method for graph neural networks. *iScience*, 25(10):105043, 2022. doi: 10.1016/J.ISCI.2022.105043.

Maul, R. W. and Gearhart, P. J. AID AND SOMATIC HYPERMUTATION. *Advances in immunology*, 105(C):159, 2010. doi: 10.1016/S0065-2776(10)05006-6.

Mcblane, J. F., Van Gent, D. C., Ramsden, D. A., Romeo, C., Cuomo, C. A., Gellert, M., and Oettinger, M. A. Cleavage at a V(D)J Recombination Signal Requires Only RAG1 and RAG2 Proteins and Occurs in Two Steps. *Cell*, 83:387–395, 1995.

Mendez, M. J., Green, L. L., Corvalan, J. R. F., Jia, X.-c., Maynard-currie, C. E., Yang, X.-d., Gallo, M. L., Louie, D. M., Lee, D. V., Erickson, K. L., Luna, J., Roy, C. M., Abderrahim, H., Kirschenbauni, F., Noguchi, M., Smith, D. H., Hales, J. F., Finer, M. H., Davis, C. G., Zsebo, K. M., and Jakobovits, A. Functional transplant of megabase human immunoglobulin loci recapitulates human antibody response in mice. *Nature Genetics*, 15:146–156, 1991.

Minot, M. and Reddy, S. T. Meta learning addresses noisy and under-labeled data in machine learning-guided antibody engineering. *Cell Systems*, 15(1):4–18.e4, 2024. doi: 10.1016/j.cels.2023.12.003.

Mirny, L. A. and Gelfand, M. S. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *Journal of Molecular Biology*, 321(1):7–20, 2002. doi: 10.1016/S0022-2836(02)00587-9.

Miura, K., Diouf, A., Fay, M. P., Barrett, J. R., Payne, R. O., Olotu, A. I., Minassian, A. M., Silk, S. E., Draper, S. J., and Long, C. A. Assessment of precision in growth inhibition assay (GIA) using human anti-PfRH5 antibodies. *Malaria Journal*, 22 (1):1–14, 2023. doi: 10.1186/S12936-023-04591-6.

Morbach, H., Eichhorn, E. M., Liese, J. G., and Girschick, H. J. Reference values for B cell subpopulations from infancy to adulthood. *Clinical and Experimental Immunology*, 162(2):271–279, 2010. doi: 10.1111/j.1365-2249.2010.04206.x.

Morell, A., Terry, W. D., and Waldmann, T. A. Metabolic properties of IgG subclasses in man. *The Journal of Clinical Investigation*, 49(4):673–680, 1970. doi: 10.1172/JCI106279.

Morrison, S. L., Johnson, M. J., Herzenberg, L. A., and Oi, V. T. Chimeric human antibody molecules: Mouse antigen-binding domains with human constant region domains. *Proceedings of the National Academy of Sciences*, 81(21 I):6851–6855, 1984. doi: 10.1073/pnas.81.21.6851.

Muramatsu, M., Sankaranand, V. S., Anant, S., Sugai, M., Kinoshita, K., Davidson, N. O., and Honjo, T. Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells. *Journal of Biological Chemistry*, 274(26):18470–18476, 1999. doi: 10.1074/jbc.274.26.18470.

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., and Honjo, T. Class switch recombination and hypermutation require activation-induced cytidine

deaminase (AID), a potential RNA editing enzyme. *Cell*, 102(5):553–563, 2000. doi: 10.1016/S0092-8674(00)00078-7.

Murphy, K., Weaver, C., Berg, L., and Barton, G. *Janeway's Immunobiology*. W. W. Norton Company, 10th edition, 2022.

Muruato, A. E., Fontes-Garfias, C. R., Ren, P., Garcia-Blanco, M. A., Menachery, V. D., Xie, X., and Shi, P. Y. A high-throughput neutralizing antibody assay for COVID-19 diagnosis and vaccine evaluation. *Nature Communications*, 11(1):1–6, 2020. doi: 10.1038/s41467-020-17892-0.

Myung, Y., Rodrigues, C. H., Ascher, D. B., and Pires, D. E. mCSM-AB2: Guiding rational antibody design using graph-based signatures. *Bioinformatics*, 36(5):1453–1459, 2020. doi: 10.1093/bioinformatics/btz779.

Nelson, A. L. Antibody fragments. *mAbs*, 2(1):77–83, 2010. doi: 10.4161/MABS.2.1.10786.

Nelson, D. L., Kurman, C. C., and Serbousek, D. E. 51Cr Release Assay of Antibody-Dependent Cell-Mediated Cytotoxicity (ADCC). *Current Protocols in Immunology*, 8(1):7.27.1–7.27.8, 1993. doi: 10.1002/0471142735.IM0727S08.

Noelle, R. J. and Snow, E. C. T helper cell-dependent B cell activation. *The FASEB Journal*, 5(13):2770–2776, 1991. doi: 10.1096/FASEBJ.5.13.1833257.

Norman, R. A., Ambrosetti, F., Bonvin, A. M., Colwell, L. J., Kelm, S., Kumar, S., and Krawczyk, K. Computational approaches to therapeutic antibody design: established methods and emerging trends. *Briefings in Bioinformatics*, 21(5):1549–1567, 2020. doi: 10.1093/BIB/BBZ095.

North, B., Lehmann, A., and Dunbrack, R. L. A New Clustering of Antibody CDR Loop Conformations. *Journal of Molecular Biology*, 406(2):228–256, 2011. doi: 10.1016/J.JMB.2010.10.030.

Nowak, J., Baker, T., Georges, G., Kelm, S., Klostermann, S., Shi, J., Sridharan, S., and Deane, C. M. Length-independent structural similarities enrich the antibody CDR canonical class model. *mAbs*, 8(4):751–760, 2016. doi: 10.1080/19420862.2016.1158370.

Oettinger, M. A., Schatz, D. G., Gorka, C., and Baltimore, D. RAG-1 and RAG-2, Adjacent Genes That Synergistically Activate V(D)J Recombination. *Science*, 248 (4962):1517–1523, 1990. doi: 10.1126/SCIENCE.2360047.

Olsen, T. H., Boyles, F., and Deane, C. M. Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1):141–146, 2022a. doi: 10.1002/pro.4205.

Olsen, T. H., Moal, I. H., and Deane, C. M. AbLang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2(1):vbac046, 2022b. doi: 10.1093/bioadv/vbac046.

Olsen, T. H., Abanades, B., Moal, I. H., and Deane, C. M. KA-Search, a method for rapid and exhaustive sequence identity search of known antibodies. *Scientific Reports*, 13(1):1–11, 2023. doi: 10.1038/s41598-023-38108-7.

Olsen, T. H., Moal, I. H., and Deane, C. M. Addressing the antibody germline bias and its effect on language models for improved antibody design. *bioRxiv*, 2024. doi: 10.1101/2024.02.02.578678.

O'Meara, M. J., Leaver-Fay, A., Tyka, M. D., Stein, A., Houlihan, K., Dimaio, F., Bradley, P., Kortemme, T., Baker, D., Snoeyink, J., and Kuhlman, B. Com-

bined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *Journal of Chemical Theory and Computation*, 11(2):609–622, 2015. doi: 10.1021/CT500864R.

Pak, M. A., Markhieva, K. A., Novikova, M. S., Petrov, D. S., Vorobyev, I. S., Maksimova, E. S., Kondrashov, F. A., and Ivankov, D. N. Using AlphaFold to predict the impact of single mutations on protein stability and function. *PLOS ONE*, 18(3):e0282689, 2023. doi: 10.1371/JOURNAL.PONE.0282689.

Parker, D. C. T cell-dependent B cell activation. *Annual Review of Immunology*, 11 (Volume 11, 1993):331–360, 1993. doi: 10.1146/ANNUREV.IY.11.040193.001555.

Parren, P. W., Carter, P. J., and Plückthun, A. Changes to International Nonproprietary Names for antibody therapeutics 2017 and beyond: of mice, men and more. *mAbs*, 9(6):898–906, 2017. doi: 10.1080/19420862.2017.1341029.

Pelat, T., Bedouelle, H., Rees, A. R., Crennell, S. J., Lefranc, M. P., and Thullier, P. Germline Humanization of a Non-human Primate Antibody that Neutralizes the Anthrax Toxin, by in Vitro and in Silico Engineering. *Journal of Molecular Biology*, 384(5):1400–1407, 2008. doi: 10.1016/j.jmb.2008.10.033.

Petersen, B. M., Kirby, M. B., Chrispens, K. M., Irvin, O. M., Strawn, I. K., Haas, C. M., Walker, A. M., Baumer, Z. T., Ulmer, S. A., Ayala, E., Rhodes, E. R., Guthmiller, J. J., Steiner, P. J., and Whitehead, T. A. An integrated technology for quantitative wide mutational scanning of human antibody Fab libraries. *Nature Communications*, 15(1):1–15, 2024. doi: 10.1038/s41467-024-48072-z.

Pires, D. E. and Ascher, D. B. mCSM-AB: a web server for predicting antibody–antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Research*, 44(W1):W469–W473, 2016. doi: 10.1093/nar/gkw458.

Pires, D. E., Ascher, D. B., and Blundell, T. L. mCSM: Predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30(3):335–342, 2014. doi: 10.1093/bioinformatics/btt691.

Ponder, J. W. and Case, D. A. Force Fields for Protein Simulations. *Advances in Protein Chemistry*, 66:27–85, 2003. doi: 10.1016/S0065-3233(03)66002-X.

Prakash, S., Johnson, R. E., and Prakash, L. Eukaryotic translesion synthesis DNA polymerases: Specificity of structure and function. *Annual Review of Biochemistry*, 74(Volume 74, 2005):317–353, 2005. doi: 10.1146/ANNUREV.BIOCHEM.74.082803.133250.

Prihoda, D., Maamary, J., Waight, A., Juan, V., Fayadat-Dilman, L., Svozil, D., and Bitton, D. A. Biophi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *mAbs*, 14(1):2020203, 2022. doi: 10.1080/19420862.2021.2020203.

Rabia, L. A., Desai, A. A., Jhajj, H. S., and Tessier, P. M. Understanding and overcoming trade-offs between antibody affinity, specificity, stability and solubility. *Biochemical Engineering Journal*, 137:365–374, 2018. doi: 10.1016/j.bej.2018.06.003.

Ramon, A., Ali, M., Atkinson, M., Saturnino, A., Didi, K., Visentin, C., Ricagno, S., Xu, X., Greenig, M., and Sormanni, P. Assessing antibody and nanobody nativeness for hit selection and humanization with AbNatiV. *Nature Machine Intelligence*, 6(1):74–91, 2024. doi: 10.1038/s42256-023-00778-3.

Ratanji, K. D., Derrick, J. P., Dearman, R. J., and Kimber, I. Immunogenicity of therapeutic proteins: Influence of aggregation. *Journal of Immunotoxicology*, 11 (2):99–109, 2014. doi: 10.3109/1547691X.2013.821564.

Ravichandran, A., Araque, J. C., and Lawson, J. W. Predicting the functional state of protein kinases using interpretable graph neural networks from sequence and

structural data. *Proteins: Structure, Function, and Bioinformatics*, 92(5):623–636, 2024. doi: 10.1002/PROT.26641.

Raybould, M. I. J., Marks, C., Lewis, A. P., Shi, J., Taddese, B., Deane, C. M., and Bujotzek, A. Thera-SAbDab : the Therapeutic Structural Antibody Database. *Nucleic Acids Research*, 48:383–388, 2020. doi: 10.1093/nar/gkz827.

Raybould, M. I. J., Rees, A. R., and Deane, C. M. Current strategies for detecting functional convergence across B-cell receptor repertoires. *mAbs*, 13(1):1996732, 2021a. doi: 10.1080/19420862.2021.1996732.

Raybould, M. I., Marks, C., Kovaltsuk, A., Lewis, A. P., Shi, J., and Deane, C. M. Public Baseline and shared response structures support the theory of antibody repertoire functional commonality. *PLOS Computational Biology*, 17(3):1–23, 2021b. doi: 10.1371/journal.pcbi.1008781.

Raybould, M. I., Turnbull, O. M., Suter, A., Guloglu, B., and Deane, C. M. Contextualising the developability risk of antibodies with lambda light chains using enhanced therapeutic antibody profiling. *Communications Biology 2024 7:1*, 7(1): 1–13, 2024. doi: 10.1038/s42003-023-05744-8.

Rees, A. R. Understanding the human antibody repertoire. *mAbs*, 12(1):1–16, 2020. doi: 10.1080/19420862.2020.1729683.

Regep, C., Georges, G., Shi, J., Popovic, B., and Deane, C. M. The H3 loop of antibodies shows unique structural characteristics. *Proteins: Structure, Function, and Bioinformatics*, 85(7):1311–1318, 2017. doi: 10.1002/PROT.25291.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2):173–175, 2011. doi: 10.1038/nmeth.1818.

Revy, P., Muto, T., Levy, Y., Geissmann, F., Plebani, A., Sanal, O., Catalan, N., Forveille, M., Dufourcq-Lagelouse, R., Gennery, A., Tezcan, I., Ersoy, F., Kayserili, H., Ugazio, A. G., Brousse, N., Muramatsu, M., Notarangelo, L. D., Kinoshita, K., Honjo, T., Fischer, A., and Durandy, A. Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the hyper-IgM syndrome (HIGM2). *Cell*, 102(5):565–575, 2000. doi: 10.1016/S0092-8674(00)00079-9.

Ribeiro, M. T., Singh, S., and Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, 97–101, 2016. doi: 10.18653/v1/n16-3020.

Richardson, E., Binter, Kosmac, M., Ghraichy, M., von Niederhausern, V., Kovaltsuk, A., Galson, J., Trück, J., Kelly, D. F., Deane, C. M., Kellam, P., and Watson, S. J. Characterisation of the immune repertoire of a humanised transgenic mouse through immunophenotyping and high-throughput sequencing. *eLife*, 12, 2023. doi: 10.7554/ELIFE.81629.

Roche, P. A. and Furuta, K. The ins and outs of MHC class II-mediated antigen processing and presentation. *Nature Reviews Immunology*, 15(4):203–216, 2015. doi: 10.1038/nri3818.

Rooney, S., Chaudhuri, J., and Alt, F. W. The role of the non-homologous end-joining pathway in lymphocyte development. *Immunological Reviews*, 200(1):115–131, 2004. doi: 10.1111/J.0105-2896.2004.00165.X.

Rosace, A., Bennett, A., Oeller, M., Mortensen, M. M., Sakhnini, L., Lorenzen, N., Poulsen, C., and Sormanni, P. Automated optimisation of solubility and confor-

mational stability of antibodies and proteins. *Nature Communications*, 14(1):1–15, 2023. doi: 10.1038/s41467-023-37668-6.

Rosenfeld, A. M., Meng, W., Luning Prak, E. T., and Hershberg, U. ImmuneDB, a novel tool for the analysis, storage, and dissemination of immune repertoire sequencing data. *Frontiers in Immunology*, 9(SEP):413882, 2018. doi: 10.3389/FIMMU.2018.02107.

Rossotti, M. A., Bélanger, K., Henry, K. A., and Tanha, J. Immunogenicity and humanization of single-domain antibodies. *The FEBS Journal*, 289(14):4304–4327, 2022. doi: 10.1111/FEBS.15809.

Rouet, R., Henry, J. Y., Johansen, M. D., Sobti, M., Balachandran, H., Langley, D. B., Walker, G. J., Lenthall, H., Jackson, J., Ubiparipovic, S., Mazigi, O., Schofield, P., Burnett, D. L., Brown, S. H., Martinello, M., Hudson, B., Gilroy, N., Post, J. J., Kelleher, A., Jäck, H. M., Goodnow, C. C., Turville, S. G., Rawlinson, W. D., Bull, R. A., Stewart, A. G., Hansbro, P. M., and Christ, D. Broadly neutralizing SARS-CoV-2 antibodies through epitope-based selection from convalescent patients. *Nature Communications*, 14(1):1–13, 2023. doi: 10.1038/s41467-023-36295-5.

Ruffolo, J. A., Gray, J. J., and Sulam, J. Deciphering antibody affinity maturation with language models and weakly supervised learning. *NeurIPS Machine Learning for Structural Biology Workshop*, 2021.

Ruffolo, J. A., Sulam, J., and Gray, J. J. Antibody structure prediction using interpretable deep learning. *Patterns*, 3(2):100406, 2022. doi: 10.1016/J.PATTER.2021.100406.

Ruffolo, J. A., Chu, L. S., Mahajan, S. P., and Gray, J. J. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature Communications*, 14(1):1–13, 2023. doi: 10.1038/s41467-023-38063-x.

Rusch, T. K., Zurich, E., Bronstein, M. M., and Mishra, S. A Survey on Oversmoothing in Graph Neural Networks. *arXiv*, 2023.

Russell, D. M., Dembić, Z., Morahan, G., Miller, J. F., Bürki, K., and Nemazee, D. Peripheral deletion of self-reactive B cells. *Nature 1991*, 354(6351):308–311, 1991. doi: 10.1038/354308a0.

Safdari, Y., Farajnia, S., Asgharzadeh, M., and Khalili, M. Antibody humanization methods – a review and update. *Biotechnology and Genetic Engineering Reviews*, 29(2):175–186, 2013. doi: 10.1080/02648725.2013.801235.

Sakaguchi, N. and Melchers, F. $\lambda 5$, a new light-chain-related locus selectively expressed in pre-B lymphocytes. *Nature 1986*, 324(6097):579–582, 1986. doi: 10.1038/324579a0.

Satorras, V. G., Hoogeboom, E., and Welling, M. E(n) equivariant graph neural networks. *arXiv*, 2021. doi: 10.48550/ARXIV.2102.09844.

Scantlebury, J., Vost, L., Carbery, A., Hadfield, T. E., Turnbull, O. M., Brown, N., Chenthamarakshan, V., Das, P., Grosjean, H., Von Delft, F., and Deane, C. M. A Small Step Toward Generalizability: Training a Machine Learning Scoring Function for Structure-Based Virtual Screening. *Journal of Chemical Information and Modeling*, 63(10):2960–2974, 2023. doi: 10.1021/ACS.JCIM.3C00322.

Schatz, D. G., Oettinger, M. A., and Baltimore, D. The V(D)J Recombination Activating Gene, RAG-1. *Cell*, 59:1035–1048, 1989.

Schlander, M., Hernandez-Villafuerte, K., Cheng, C. Y., Mestre-Ferrandiz, J., and Baumann, M. How Much Does It Cost to Research and Develop a New Drug? A Systematic Review and Assessment. *PharmacoEconomics*, 39(11):1243–1269, 2021. doi: 10.1007/S40273-021-01065-Y.

Schlichtkrull, M. S., De Cao, N., and Titov, I. Interpreting Graph Neural Networks for NLP With Differentiable Edge Masking. *9th International Conference on Learning Representations*, 2020. doi: 10.48550/arXiv.2010.00577.

Schmitz, G. P., Aldrich, C., and Gouws, F. S. ANN-DT: An algorithm for extraction of decision trees from artificial neural networks. *IEEE Transactions on Neural Networks*, 10(6):1392–1401, 1999. doi: 10.1109/72.809084.

Schneider, C., Raybould, M. I. J., and Deane, C. M. SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. *Nucleic Acids Research*, 50(D1):D1368–D1372, 2021. doi: 10.1093/nar/gkab1050.

Schneider, C., Buchanan, A., Taddese, B., and Deane, C. M. DLAB: deep learning methods for structure-based virtual screening of antibodies. *Bioinformatics*, 38(2): 377–383, 2022. doi: 10.1093/BIOINFORMATICS/BTAB660.

Schrödinger, LLC. The PyMOL molecular graphics system, version. 2015.

Schroeder, H. W. Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Developmental and Comparative Immunology*, 30(1-2):119–135, 2006. doi: 10.1016/j.dci.2005.06.006.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. The FoldX web server: An online force field. *Nucleic Acids Research*, 33:382–388, 2005. doi: 10.1093/nar/gki387.

Seeliger, D. Development of Scoring Functions for Antibody Sequence Assessment and Optimization. *PLOS ONE*, 8(10), 2013. doi: 10.1371/journal.pone.0076909.

Seki, M., Gearhart, P. J., and Wood, R. D. DNA polymerases and somatic hypermutation of immunoglobulin genes. *EMBO Reports*, 6(12):1143–1148, 2005. doi: 10.1038/SJ.EMBOR.7400582.

Sekiguchi, N., Kubo, C., Takahashi, A., Muraoka, K., Takeiri, A., Ito, S., Yano, M., Mimoto, F., Maeda, A., Iwayanagi, Y., Wakabayashi, T., Takata, S., Murao, N., Chiba, S., and Ishigai, M. MHC-associated peptide proteomics enabling highly sensitive detection of immunogenic sequences for the development of therapeutic antibodies with low immunogenicity. *mAbs*, 10(8):1168–1181, 2018. doi: 10.1080/19420862.2018.1518888.

Sela-Culang, I., Kunik, V., and Ofran, Y. The structural basis of antibody-antigen recognition. *Frontiers in Immunology*, 4:302, 2013. doi: 10.3389/fimmu.2013.00302.

Shanehsazzadeh, A., Alverio, J., Kasun, G., Levine, S., Khan, J. A., Chung, C., Diaz, N., Luton, B. K., Tarter, Y., McCloskey, C., Bateman, K. B., Carter, H., Chapman, D., Consbruck, R., Jaeger, A., Kohnert, C., Kopec-Belliveau, G., Sutton, J. M., Guo, Z., Canales, G., Ejan, K., Marsh, E., Ruelos, A., Ripley, R., Stoddard, B., Caguiat, R., Chapman, K., Saunders, M., Sharp, J., da Silva, D. G., Feltner, A., Ripley, J., Bryant, M. E., Castillo, D., Meier, J., Stegmann, C. M., Moran, K., Lemke, C., Abdulhaqq, S., Klug, L. R., Bachas, S., and Corporation, A. In vitro validated antibody design against multiple therapeutic antigens using generative inverse folding. *bioRxiv*, 2023. doi: 10.1101/2023.12.08.570889.

Shanehsazzadeh, A., McPartlon, M., Kasun, G., Steiger, A. K., Sutton, J. M., Yassine, E., McCloskey, C., Haile, R., Shuai, R., Alverio, J., Rakocevic, G., Levine, S., Cejovic, J., Gutierrez, J. M., Morehead, A., Dubrovskyi, O., Chung, C., Luton, B. K., Diaz, N., Kohnert, C., Consbruck, R., Carter, H., LaCombe, C., Bist, I., Vilaychack, P., Anderson, Z., Xiu, L., Bringas, P., Alarcon, K., Knight, B., Radach, M., Bateman, K., Kopec-Belliveau, G., Chapman, D., Bennett, J., Ventura, A. B., Canales, G. M., Gowda, M., Jackson, K. A., Caguiat, R., Brown, A., da Silva, D. G., Guo, Z., Abdulhaqq, S., Klug, L. R., Gander, M., Yapici, E., Meier, J., and

Bachas, S. Unlocking de novo antibody design with generative artificial intelligence. *bioRxiv*, 2024. doi: 10.1101/2023.01.08.523187.

Shanker, V. R., Bruun, T. U., Hie, B. L., and Kim, P. S. Inverse folding of protein complexes with a structure-informed language model enables unsupervised antibody evolution. *bioRxiv*, 2023.12.19.572475, 2023. doi: 10.1101/2023.12.19.572475.

Shapovalov, M. V. and Dunbrack, R. L. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19(6):844–858, 2011. doi: 10.1016/j.str.2011.03.019.

Shim, H. Synthetic approach to the generation of antibody diversity. *BMB Reports*, 48(9):489, 2015. doi: 10.5483/BMBREP.2015.48.9.120.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning Important Features Through Propagating Activation Differences, 2017. ISSN 2640-3498.

Shuai, R. W., Ruffolo, J. A., and Gray, J. J. IgLM: Infilling language modeling for antibody sequence design. *Cell Systems*, 14(11):979–989.e4, 2023. doi: 10.1016/j.cels.2023.10.001.

Sirin, S., Apgar, J. R., Bennett, E. M., and Keating, A. E. AB-Bind: Antibody binding mutational database for computational affinity predictions. *Protein Science*, 25 (2):393–409, 2016. doi: 10.1002/pro.2829.

Sloan-Lancaster, J., Evavold, B. D., and Allen, P. M. Induction of T-cell anergy by altered T-cell-receptor ligand on live antigen-presenting cells. *Nature*, 363(6425): 156–159, 1993. doi: 10.1038/363156a0.

Smith, K., Garman, L., Wrammert, J., Zheng, N. Y., Capra, J. D., Ahmed, R., and Wilson, P. C. Rapid generation of fully human monoclonal antibod-

ies specific to a vaccinating antigen. *Nature Protocols*, 4(3):372–384, 2009. doi: 10.1038/nprot.2009.3.

Soleymani, F., Paquet, E., Viktor, H. L., and Michalowski, W. Structure-based protein and small molecule generation using EGNN and diffusion models: A comprehensive review. *Computational and Structural Biotechnology Journal*, 23:2779–2797, 2024. doi: 10.1016/J.CSBJ.2024.06.021.

Song, Y., Tyka, M., Leaver-Fay, A., Thompson, J., and Baker, D. Structure-guided forcefield optimization. *Proteins: Structure, Function, and Bioinformatics*, 79(6): 1898–1909, 2011. doi: 10.1002/PROT.23013.

Sormanni, P., Amery, L., Ekizoglou, S., Vendruscolo, M., and Popovic, B. Rapid and accurate in silico solubility screening of a monoclonal antibody library. *Scientific Reports*, 7(1):1–9, 2017. doi: 10.1038/s41598-017-07800-w.

Stanfield, R. L. and Wilson, I. A. Antibody structure. *Microbiology Spectrum*, 2(2): 10.1128/microbiolspec.aid–0012–2013, 2014. doi: 10.1128/microbiolspec.aid-0012-2013.

Stavnezer, J. Immunoglobulin class switching. *Current Opinion in Immunology*, 8 (2):199–205, 1996. doi: 10.1016/S0952-7915(96)80058-6.

Stoop, J. W., Zegers, B. J. M., Sander, P. C., and Ballieux, R. E. Serum immunoglobulin levels in healthy children and adults. *Clinical and Experimental Immunology*, 4(1):101, 1969.

Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A., and Kim, P. M. Fast and flexible protein design using deep graph neural networks. *Cell Systems*, 11(4): 402–411.e4, 2020. doi: 10.1016/j.cels.2020.08.016.

Sumida, K. H., Núñez-Franco, R., Kalvet, I., Pellock, S. J., Wicky, B. I., Milles, L. F., Dauparas, J., Wang, J., Kipnis, Y., Jameson, N., Kang, A., De La Cruz, J., Sankaran, B., Bera, A. K., Jiménez-Osés, G., and Baker, D. Improving Protein Expression, Stability, and Function with ProteinMPNN. *Journal of the American Chemical Society*, 146(3):2054–2061, 2024. doi: 10.1021/JACS.3C10941.

Sun, A. and Benet, L. Z. Late-Stage Failures of Monoclonal Antibody Drugs: A Retrospective Case Study Analysis. *Pharmacology*, 105(3-4):145–163, 2020. doi: 10.1159/000505379.

Sun, C., Qiu, X., Xu, Y., and Huang, X. How to fine-tune BERT for text classification? *arXiv*, abs/1905.05583, 2019. doi: 10.48550/arXiv.1905.05583.

Sunseri, J. and Koes, D. R. libmolgrid: Graphics processing unit accelerated molecular gridding for deep learning applications. *Journal of Chemical Information and Modeling*, 60(3):1079–1084, 2020. doi: 10.1021/acs.jcim.9b01145.

Swanson, M. D., Rios, S., Mittal, S., Soder, G., and Jawa, V. Immunogenicity Risk Assessment of Spontaneously Occurring Therapeutic Monoclonal Antibody Aggregates. *Frontiers in Immunology*, 13:915412, 2022. doi: 10.3389/FIMMU.2022.915412.

Tang, Y., Cain, P., Anguiano, V., Shih, J. J., Chai, Q., and Feng, Y. Impact of IgG subclass on molecular properties of monoclonal antibodies. *mAbs*, 13(1), 2021. doi: 10.1080/19420862.2021.1993768.

Te Wu, T. and Kabat, E. A. An Analysis Of The Sequences Of The Variable Regions Of Bence Jones Proteins And Myeloma Light Chains And Their Implications For Antibody Complementarity. *Journal of Experimental Medicine*, 132(2):211–250, 1970. doi: 10.1084/JEM.132.2.211.

Teixeira, A. A. R., Erasmus, M. F., D'Angelo, S., Naranjo, L., Ferrara, F., Leal-Lopes, C., Durrant, O., Galmiche, C., Morelli, A., Scott-Tucker, A., and Bradbury, A. R. M. Drug-like antibodies with high affinity, diversity and developability directly from next-generation antibody libraries. *mAbs*, 13(1), 2021. doi: 10.1080/19420862.2021.1980942.

Tennenhouse, A., Khmelnitsky, L., Khalaila, R., Yeshaya, N., Noronha, A., Lindzen, M., Makowski, E. K., Zaretsky, I., Sirkis, Y. F., Galon-Wolfenson, Y., Tessier, P. M., Abramson, J., Yarden, Y., Fass, D., and Fleishman, S. J. Computational optimization of antibody humanness and stability by systematic energy-based ranking. *Nature Biomedical Engineering*, 8:30–44, 2023. doi: 10.1038/s41551-023-01079-1.

Thullier, P., Huish, O., Pelat, T., and Martin, A. C. The Humanness of Macaque Antibody Sequences. *Journal of Molecular Biology*, 396(5):1439–1450, 2010. doi: 10.1016/j.jmb.2009.12.041.

Tiller, T., Meffre, E., Yurasov, S., Tsuiji, M., Nussenzweig, M. C., and Wardemann, H. Efficient generation of monoclonal antibodies from single human B cells by single cell RT-PCR and expression vector cloning. *Journal of Immunological Methods*, 329 (1-2):112–124, 2008. doi: 10.1016/J.JIM.2007.09.017.

Tiller, T., Schuster, I., Deppe, D., Siegers, K., Strohner, R., Herrmann, T., Berenguer, M., Poujol, D., Stehle, J., Stark, Y., Heßling, M., Daubert, D., Felderer, K., Kaden, S., Kölln, J., Enzelberger, M., and Urlinger, S. A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. *mAbs*, 5(3):445–470, 2013. doi: 10.4161/MABS.24218.

Tonner, P. D., Pressman, A., and Ross, D. Interpretable modeling of genotype–phenotype landscapes with state-of-the-art predictive power. *Proceed-*

ings of the National Academy of Sciences, 119(26):e2114021119, 2022. doi: 10.1073/pnas.2114021119.

Torrie, G. M. and Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23 (2):187–199, 1977. doi: 10.1016/0021-9991(77)90121-8.

Tovey, M. G. and Lallemand, C. Immunogenicity and other problems associated with the use of biopharmaceuticals. *Therapeutic Advances in Drug Safety*, 2(3):113–128, 2011. doi: 10.1177/2042098611406318.

Trippe, B. L., Yim, J., Tischer, D., Baker, D., Broderick, T., Barzilay, R., and Jaakkola, T. Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. *11th International Conference on Learning Representations,*, 2022.

Tubiana, J., Schneidman-Duhovny, D., and Wolfson, H. J. ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nature Methods*, 19(6):730–739, 2022. doi: 10.1038/s41592-022-01490-7.

Ucar, T., Ramon, A., Oglic, D., Croasdale-Wood, R., Diethe, T., and Sormanni, P. Improving antibody humanness prediction using patent data. *arXiv*, 2024. doi: 10.48550/arXiv.2401.14442.

Van Durme, J., Delgado, J., Stricher, F., Serrano, L., Schymkowitz, J., and Rousseau, F. A graphical interface for the FoldX forcefield. *Bioinformatics*, 27(12):1711–1712, 2011. doi: 10.1093/bioinformatics/btr254.

Van Gent, D. C., Mcblane, J. F., Ramsden, D. A., Sadofsky, M. J., Hesse, J. E., and Gellert, M. Initiation of V(D)J Recombination in a Cell-Free System. *Cell*, 81: 925–934, 1995.

Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., and Mackerell, A. D. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry*, 31 (4):671–690, 2010. doi: 10.1002/JCC.21367.

Vargas-Madrazo, E., Lara-Ochoa, F., Ramirez-Benites, M. C., and Almagro, J. C. Evolution of the structural repertoire of the human V(H) and V(κ) germline genes. *International Immunology*, 9(12):1801–1815, 1997. doi: 10.1093/intimm/9.12.1801.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. doi: 10.48550/arXiv.1706.03762.

Veličković, P., Casanova, A., Liò, P., Cucurull, G., Romero, A., and Bengio, Y. Graph Attention Networks. *6th International Conference on Learning Representations*, 2017. doi: 10.1007/978-3-031-01587-8_7.

Von Schwedler, U., Jäck, H. M., and Wabl, M. Circular DNA is a product of the immunoglobulin class switch rearrangement. *Nature 1990*, 345(6274):452–456, 1990. doi: 10.1038/345452a0.

Vu, K. B., Ghahroudi, M. A., Wyns, L., and Muyldermans, S. Comparison of llama VH sequences from conventional and heavy chain antibodies. *Molecular Immunology*, 34(16-17):1121–1131, 1997. doi: 10.1016/S0161-5890(97)00146-6.

Vu, M. N. and Thai, M. T. PGM-Explainer: Probabilistic Graphical Model Explanations for Graph Neural Networks. *Advances in Neural Information Processing Systems*, 2020.

Wang, M., Cang, Z., and Wei, G.-w. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nature Machine Intelligence*, 2:116–123, 2020.

Wang, Z., Sun, D., Zhou, S., Wang, H., Fan, J., Huang, L., and Bu, J. NoisyGL: A Comprehensive Benchmark for Graph Neural Networks under Label Noise. *arXiv*, 2024.

Wardemann, H., Yurasov, S., Schaefer, A., Young, J. W., Meffre, E., and Nussenzweig, M. C. Predominant autoantibody production by early human B cell precursors. *Science*, 301(5638):1374–1377, 2003. doi: 10.1126/SCIENCE.1086907.

Warszawski, S., Borenstein Katz, A., Lipsh, R., Khmelnitsky, L., Ben Nissan, G., Javitt, G., Dym, O., Unger, T., Knop, O., Albeck, S., Diskin, R., Fass, D., Sharon, M., and Fleishman, S. J. Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces. *PLOS Computational Biology*, 15(8):1–24, 2019. doi: 10.1371/journal.pcbi.1007207.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., De Bortoli, V., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M., and Baker, D. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

Weber, M., Bujak, E., Putelli, A., Villa, A., Matasci, M., Gualandi, L., Hemmerle, T., Wulhfard, S., and Neri, D. A Highly Functional Synthetic Phage Display Library Containing over 40 Billion Human Antibody Clones. *PLOS ONE*, 9(6):e100000, 2014. doi: 10.1371/JOURNAL.PONE.0100000.

Weeramange, C. J., Fairlamb, M. S., Singh, D., Fenton, A. W., and Swint-Kruse, L. The strengths and limitations of using biolayer interferometry to monitor equilibrium titrations of biomolecules. *Protein Science*, 29(4):1004–1020, 2020. doi: 10.1002/PRO.3827.

Weitzner, B. D., Dunbrack, R. L., and Gray, J. J. The origin of CDR H3 structural diversity. *Structure*, 23(2):302–311, 2015. doi: 10.1016/j.str.2014.11.010.

Williams, S. C., Frippiat, J. P., Tomlinson, I. M., Ignatovich, O., Lefranc, M. P., and Winter, G. Sequence and evolution of the human germline Vλ repertoire. *Journal of Molecular Biology*, 264(2):220–232, 1996. doi: 10.1006/jmbi.1996.0636.

Wollacott, A. M., Xue, C., Qin, Q., Hua, J., Bohnuud, T., Viswanathan, K., and Kolachalama, V. B. Quantifying the nativeness of antibody sequences using long short-term memory networks. *Protein Engineering, Design and Selection*, 32(7): 347–354, 2019. doi: 10.1093/protein/gzz031.

Wong, W. K., Georges, G., Ros, F., Kelm, S., Lewis, A. P., Taddese, B., Leem, J., and Deane, C. M. SCALOP: sequence-based antibody canonical loop structure annotation. *Bioinformatics*, 35(10):1774–1776, 2019. doi: 10.1093/BIOINFORMATICS/BTY877.

Wu, Y., Wang, F., Shen, C., Peng, W., Li, D., Zhao, C., Li, Z., Li, S., Bi, Y., Yang, Y., Gong, Y., Xiao, H., Fan, Z., Tan, S., Wu, G., Tan, W., Lu, X., Fan, C., Wang, Q., Liu, Y., Zhang, C., Qi, J., Gao, G. F., Gao, F., and Liu, L. A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. *Science*, 368(6496):1274–1278, 2020. doi: 10.1126/science.abc2241.

Xia, T., Liang, S., Wang, H., Hu, S., Sun, Y., Yu, X., Han, J., Li, J., Guo, S., Dai, J., Lou, Z., and Guo, Y. Structural basis for the neutralization and specificity

of staphylococcal enterotoxin b against its mhc class ii binding site. *mAbs*, 6(1): 119–129, 2014. doi: 10.4161/mabs.27106.

Xu, Y., Roach, W., Sun, T., Jain, T., Prinz, B., Yu, T. Y., Torrey, J., Thomas, J., Bobrowicz, P., Vásquez, M., Wittrup, K. D., and Krauland, E. Addressing polyspecificity of antibodies selected from an in vitro yeast presentation system: a FACS-based, high-throughput selection and analytical tool. *Protein Engineering, Design and Selection*, 26(10):663–670, 2013. doi: 10.1093/PROTEIN/GZT047.

Ying, R., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. GNNExplainer: Generating Explanations for Graph Neural Networks. *Advances in Neural Information Processing Systems*, 32, 2019. doi: 10.48550/arXiv.1903.03894.

Yokota, T., Milenic, D. E., Whitlow, M., and Schlom1, J. Rapid Tumor Penetration of a Single-Chain Fv and Comparison with Other Immunoglobulin Forms. *Cancer Research*, 52:3402–3408, 1992.

Yuan, H., Tang, J., Hu, X., and Ji, S. XGNN: Towards Model-Level Explanations of Graph Neural Networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 430–438, 2020. doi: 10.1145/3394486.3403085.

Yuan, H., Yu, H., Wang, J., Li, K., and Ji, S. On Explainability of Graph Neural Networks via Subgraph Explorations, 2021. ISSN 2640-3498.

Zegers, B. J., Stoop, J. W., Reerink-Brongers, E. E., Sander, P. C., Aalberse, R. C., and Ballieux, R. E. Serum immunoglobulins in healthy children and adults levels of the five classes, expressed in international units per millilitre. *Clinica Chimica Acta*, 65(3):319–329, 1975. doi: 10.1016/0009-8981(75)90257-0.

Zeng, X., Winter, D. B., Kasmer, C., Kraemer, K. H., Lehmann, A. R., and Gearhart, P. J. DNA polymerase η is an A-T mutator in somatic hypermutation of immunoglobulin variable genes. *Nature Immunology*, 2(6):537–541, 2001. doi: 10.1038/88740.

Zhang, S., Liu, Y., Shah, N., and Sun, Y. GStarX: Explaining Graph Neural Networks with Structure-Aware Cooperative Games. *Advances in Neural Information Processing Systems*, 35, 2022. doi: 10.48550/arXiv.2201.12380.

Zhang, Y. and Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57 (4):702–710, 2004. doi: 10.1002/PROT.20264.

Zhang, Y., Li, P., Pan, F., Liu, H., Hong, P., Liu, X., and Zhang, J. Applications of AlphaFold beyond Protein Structure Prediction. *bioRxiv*, 2021. doi: 10.1101/2021.11.03.467194.

Zhao, Q., Zhu, Z., and Dimitrov, D. S. Yeast display of engineered antibody domains. *Methods in Molecular Biology*, 899:73–84, 2012. doi: 10.1007/978-1-61779-921-1_5.

Zhao, T., Cai, Y., Jiang, Y., He, X., Wei, Y., Yu, Y., and Tian, X. Vaccine adjuvants: mechanisms and platforms. *Signal Transduction and Targeted Therapy*, 8(1):1–24, 2023. doi: 10.1038/s41392-023-01557-7.

Zhu, Z. and Dimitrov, D. S. Construction of a Large Naïve Human Phage-Displayed Fab Library Through One-Step Cloning. *Methods in molecular biology (Clifton, N.J.)*, 525:129, 2009. doi: 10.1007/978-1-59745-554-1_6.

Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics*, 22(8):1420–1426, 1954. doi: 10.1063/1.1740409.

Appendix A

Supplementary materials accompanying Chapter 2: Investigating the Volume and Diversity of Data Needed for Generalisable Antibody-Antigen $\Delta\Delta$G Prediction

Table A.1: **Descriptions of the experimental and synthetic ΔΔG datasets to which Graphinity was applied.** Ab: antibody, Ag: antigen, AA: amino acid. For definitions of inner and outer shell see Section 2.4.1.5.

Dataset Name	Experimental/ Synthetic	Description		Train-Val-Test Split (CDR Sequence Identity Cutoff)	Number of Mutations	Number of Complexes	Number of AA Substitution Types	Mutation Distribution (# of muts. in Ab Inner, Ab Outer, Ag Inner, Ag Outer)
Experimental_ΔΔG_645 – Reverse Mutations + Non-Binders	Experimental	Dataset of single-point mutations from AB-Bind (Sirin et al., 2016)	– reverse mutations, + non-binders	None (Random), 100%, 90%, 70%	645	24 (plus 5 homology models)	141	185, 148, 172, 140
Experimental_ΔΔG_645 – Reverse Mutations – Non-Binders			– reverse mutations, – non-binders		618		136	170, 136, 172, 140
Experimental_ΔΔG_645 + Reverse Mutations + Non-Binders			+ reverse mutations, +non-binders		1,290		224	370, 296, 344, 280
Experimental_ΔΔG_645 + Reverse Mutations – Non-Binders			+ reverse mutations, – non-binders		1,236		216	340, 272, 344, 280
Experimental_ΔΔG_608		Dataset of single-point mutations filtered from SKEMPI 2.0 (Jankauskaite et al., 2019)	– reverse mutations		608	33	163	232, 138, 151, 87
Experimental_ΔΔG_608 + Reverse Mutations			+ reverse mutations		1,216		232	464, 276, 302, 174
Synthetic_ΔΔG_ 942723	Synthetic	Synthetic single-point mutation ΔΔG data generated using FoldX		None (Random), 100%, 90%, 70%, 70% + 70% Ag seq. identity cutoff	942,723	1,471	380	326,990, 155,439, 328,130, 132,164
Synthetic_ΔΔG_ 942723_shuffled		Synthetic_ΔΔG_942723 with a percentage of ΔΔG labels shuffled (i.e. incorrect)		90%				
Synthetic_ΔΔG_ 942723_gaussian_ noise		Synthetic_ΔΔG_942723 with random noise sampled from Gaussian distributions with varying scales added						
Synthetic_ΔΔG_580		Train + validation datasets of varying sizes randomly sampled from the respective Synthetic_ΔΔG_942723 datasets			580	462	269	200, 87, 212, 81
Synthetic_ΔΔG_900					900	645	313	296, 137, 331, 136
Synthetic_ΔΔG_4500					4,500	1,264	377	1,519, 711, 1,602, 668
Synthetic_ΔΔG_9000					9,000	1,316	380	3,052, 1,468, 3,180, 1,300
Synthetic_ΔΔG_45000					45,000	1,324	380	15,522, 7,443, 15,843, 6,192
Synthetic_ΔΔG_90000					90,000	1,324	380	31,337, 14,800, 31,387, 12,476
Synthetic_ΔΔG_450000					450,000	1,324	380	156,449, 74,016, 156,945, 62,590
Synthetic_ΔΔG_848597					848,597	1,324	380	294,614, 139,802, 296,305, 117,876
Synthetic_ΔΔG_94126		Test dataset to which models trained on the train + validation datasets of varying sizes were applied			94,126	147	380	32,376, 15,637, 31,825, 14,288
Synthetic_ΔΔG_ 100000_sequence_min		Datasets of 90,000 mutations sampled from Synthetic_ΔΔG_ 942723 to: [NB train and validation datasets only, which are 80,000 and 10,000 mutations respectively]	minimize antibody CDR sequence diversity		Train: 80,000; Val: 10,000	86 (Train + Val)	380	34,143, 10,678, 32,785, 12,394
Synthetic_ΔΔG_ 100000_sequence_max			maximize antibody CDR sequence diversity			1,324 (Train + Val)	380	30,744, 15,420, 31,172, 12,664
Synthetic_ΔΔG_ 100000_substitution_ type_min			minimize antibody substitution type diversity			1,293 (Train + Val)	16	48,747, 20,314, 14,940, 5,999
Synthetic_ΔΔG_ 100000_substitution_ type_max			maximize antibody substitution type diversity			1,324 (Train + Val)	380	35,549, 12,464, 27,987, 14,000
Synthetic_ΔΔG_ 100000_mutation_ distribution_min			minimize antibody mutation distribution diversity			1,321 (Train + Val)	380	0, 0, 90,000, 0
Synthetic_ΔΔG_ 100000_mutation_ distribution_max			maximize antibody mutation distribution diversity			1,324 (Train + Val)	380	22,505, 22,498, 22,499, 22,498
Synthetic_ΔΔG_ 100000_randomly_ sampled		Train and validation datasets randomly sampled from Synthetic_ΔΔG_ 942723, for which no complex overlaps with any in Synthetic_ΔΔG_100000_diversity_test_set				1,332 (Train + Val)	380	31,283, 14,800, 31,345, 12,572
Synthetic_ΔΔG_ 100000_diversity_ test_set		Test dataset for all train/val diversity datasets – consists of 10,000 mutations, for which no complex overlaps with any in the train and validation sets			Test: 10,000	139 (Test)	380	3,468, 3,377, 1,763, 1,392

Table A.2: **Pharmacophore counts for each amino acid, as used in the tree-based model featurisation.**

AA	Neutral	H-bond Donor	H-bond Acceptor	Hydro-phobic	Aromatic	Positive	Negative	Sulphur
A	2	1	1	1	0	0	0	0
C	2	1	1	1	0	0	0	1
D	3	1	3	1	0	0	2	0
E	3	1	3	2	0	0	2	0
F	2	1	1	7	6	0	0	0
G	2	1	1	0	0	0	0	0
H	2	3	3	6	5	2	0	0
I	2	1	1	4	0	0	0	0
K	2	2	1	4	0	1	0	0
L	2	1	1	4	0	0	0	0
M	2	1	1	4	0	0	0	0
N	3	2	2	1	0	0	0	0
P	1	0	1	5	0	0	0	0
Q	3	2	2	2	0	0	0	0
R	3	4	1	3	0	3	0	0
S	2	2	2	1	0	0	0	0
T	2	2	2	2	0	0	0	0
V	2	1	1	3	0	0	0	0
W	2	2	1	10	9	0	0	0
Y	2	2	2	7	6	0	0	0

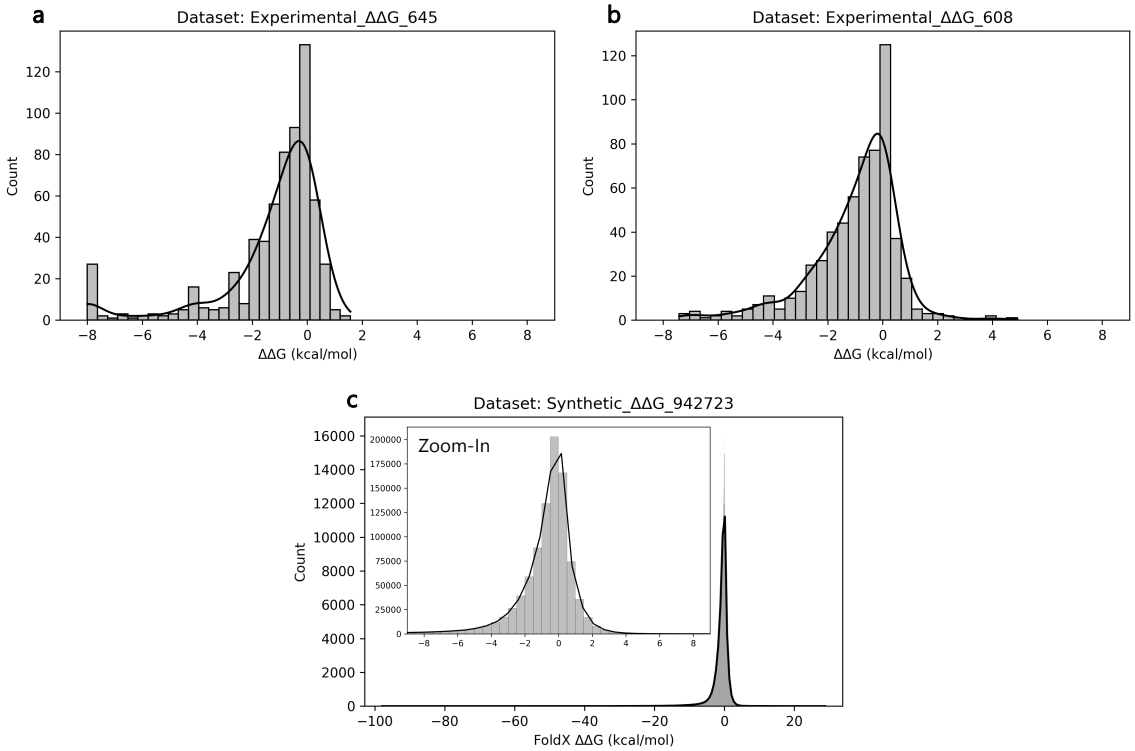

Figure A.1: **The distributions of the ΔΔG values of the base datasets to which Graphinity was applied**: (a) Experimental_ΔΔG_645, (b) Experimental_ΔΔG_608, (c) Synthetic_ΔΔG_942723. The solid lines are kernel density estimates.

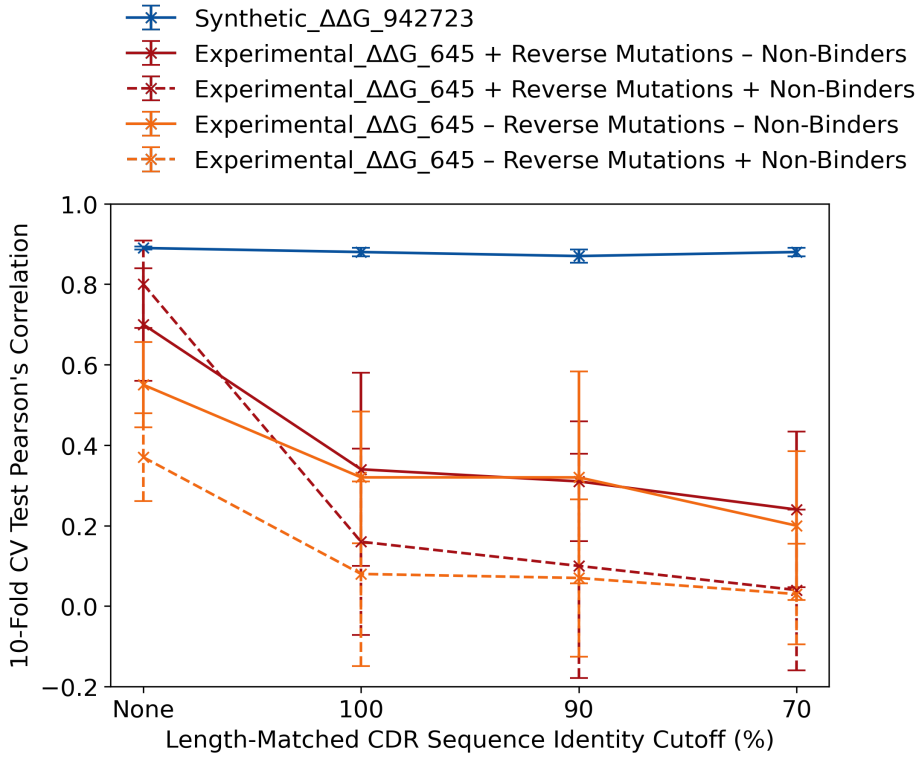

Figure A.2: **The Pearson's correlations of Graphinity on different train-validation-test cutoffs applied to the Experimental_ΔΔG_645 dataset (red, orange) and Synthetic_ΔΔG_942723 dataset (blue).** This is Figure 2.2 including error bars, which represent the standard deviation in Pearson's correlation across the 10 folds of 10-fold cross-validation (CV).

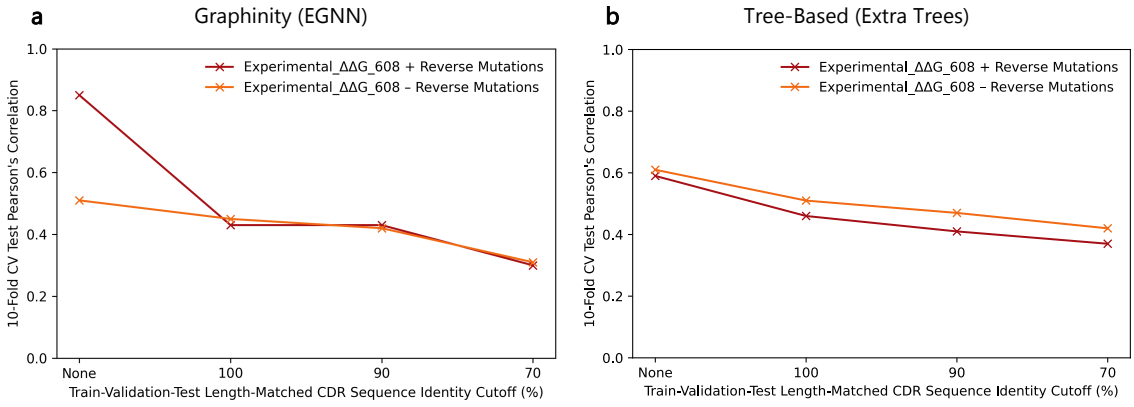

Figure A.3: **Performance of (a) Graphinity (EGNN architecture) and (b) a tree-based (Extra Trees) model on the Experimental_ΔΔG_608 dataset**, with and without reverse mutations, at different length-matched CDR sequence identity cutoffs.

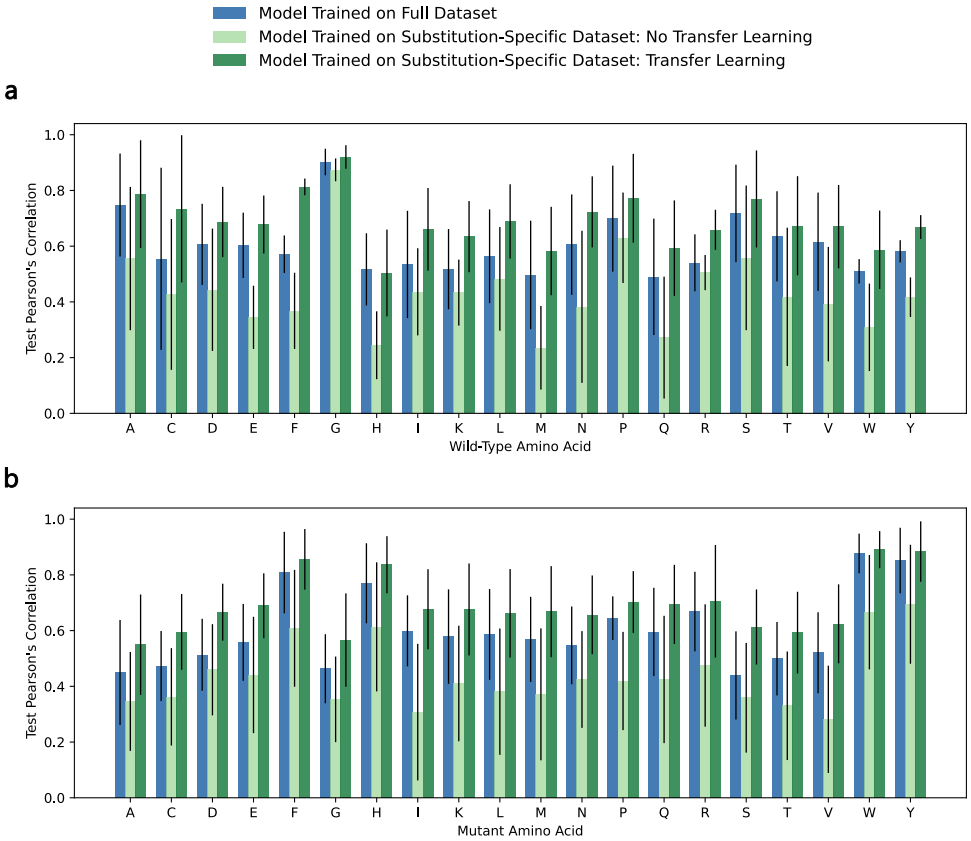

Figure A.4: **Graphinity performance on amino acid substitutions.** Pearson's correlations are shown for models trained on the full dataset (blue), substitution-specific datasets (light green) and substitution-specific datasets with weights initialised from the model trained on the full dataset (dark green). The results were grouped and averaged by (a) wild-type amino acid and (b) mutant amino acid. The error bars represent the standard deviation.

Appendix B

Supplementary materials accompanying Chapter 3: Assessing the Interpretability of Deep Learning for Antibody-Antigen Binding Affinity Prediction

Table B.1: **Comparison of edge weighting in high- versus low-scored Trastuzumab variant graphs.** The edge weights were derived from the Scatter Softmax 1-layer attention MLP. The p-values were calculated using a Wilcoxon signed-rank test.

High Score Cutoff	Low Score Cutoff	n High	n Low	p-Value
0.9	0.1	3055	14609	0.42
0.1	0.1	8336	14609	0.26
0.2	0.2	6990	15955	0.40
0.3	0.3	6291	16654	0.55
0.4	0.4	5753	17192	0.49
0.5	0.5	5296	17649	0.35
0.6	0.6	4837	18108	0.36
0.7	0.7	4379	18566	0.30
0.8	0.8	3853	19092	0.37
0.9	0.9	3055	19890	0.61
Top 10	Bottom 10	10	10	0.89
Top 100	Bottom 100	100	100	0.86

Table B.2: **Comparison of edge weighting in true positive, true negative, false positive and false negative ('Pred Labels') classified Trastuzumab variant graphs.** The edge weights were derived from the Scatter Softmax 1-layer attention MLP. The p-values were calculated using a Wilcoxon signed-rank test.

Pred Label 1	**Pred Label 2**	**n Pred Label 1**	**n Pred Label 2**	**p-Value**
TP	TN	5024	16333	0.24
TP	FP	5024	726	0.40
TP	FN	5024	862	0.79
TN	FP	16333	726	0.05
TN	FN	16333	862	0.79
FP	FN	726	862	0.29

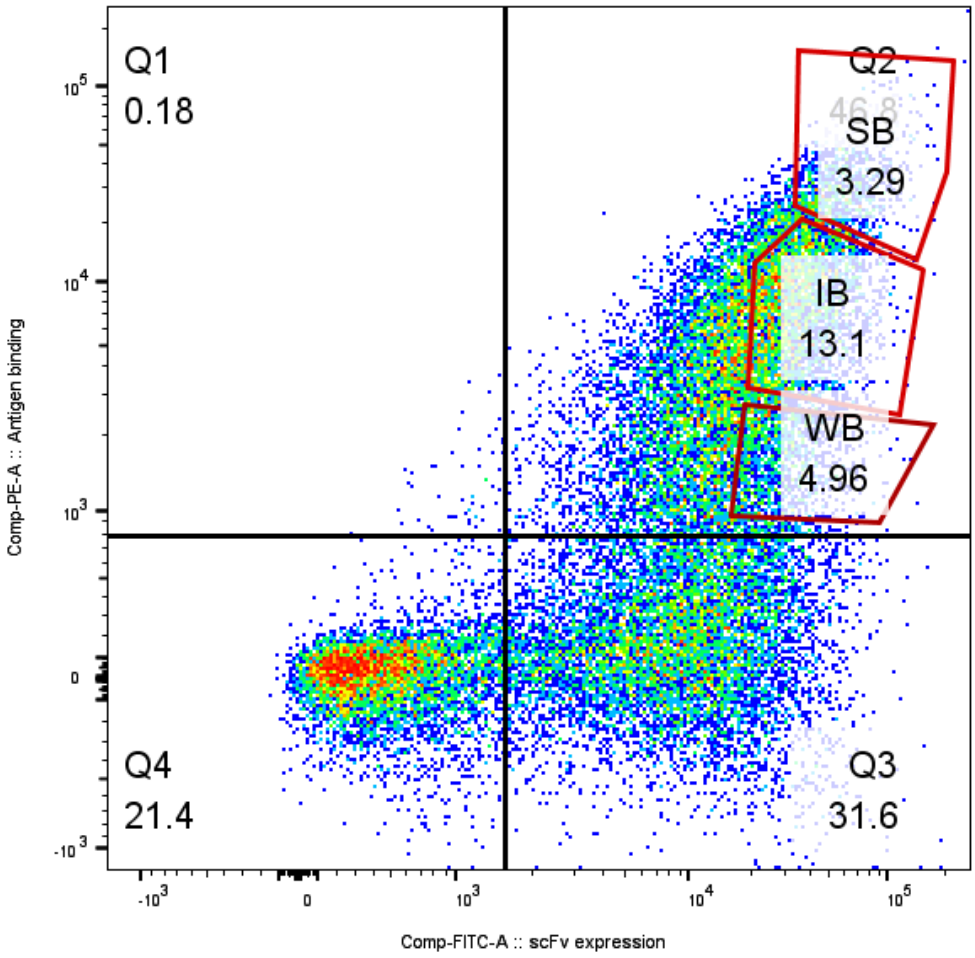

Figure B.1: **Bivariate flow-cytometric analysis of Trastuzumab-variant library highlights different antigen-binding populations.** Cells are double-labelled with biotinylated antigen/streptavidin–phycoerythrin (y-axis), and anti-V5/anti-mouse FITC labels (x-axis). The percentages of events measured in each quadrant and covered by each gate are displayed as numbers. The sub-population with the brightest antigen labelling at a given scFv expression is termed 'strong binder' (SB), referred to in the main text as 'high-affinity'. Sub-populations showing intermediate (medium-affinity) and weak (low-affinity) labelling for HER2 at a given scFv expression are termed intermediate (IB) and weak binders (WB). This experiment was completed and figure prepared by the lab of Victor Greiff.

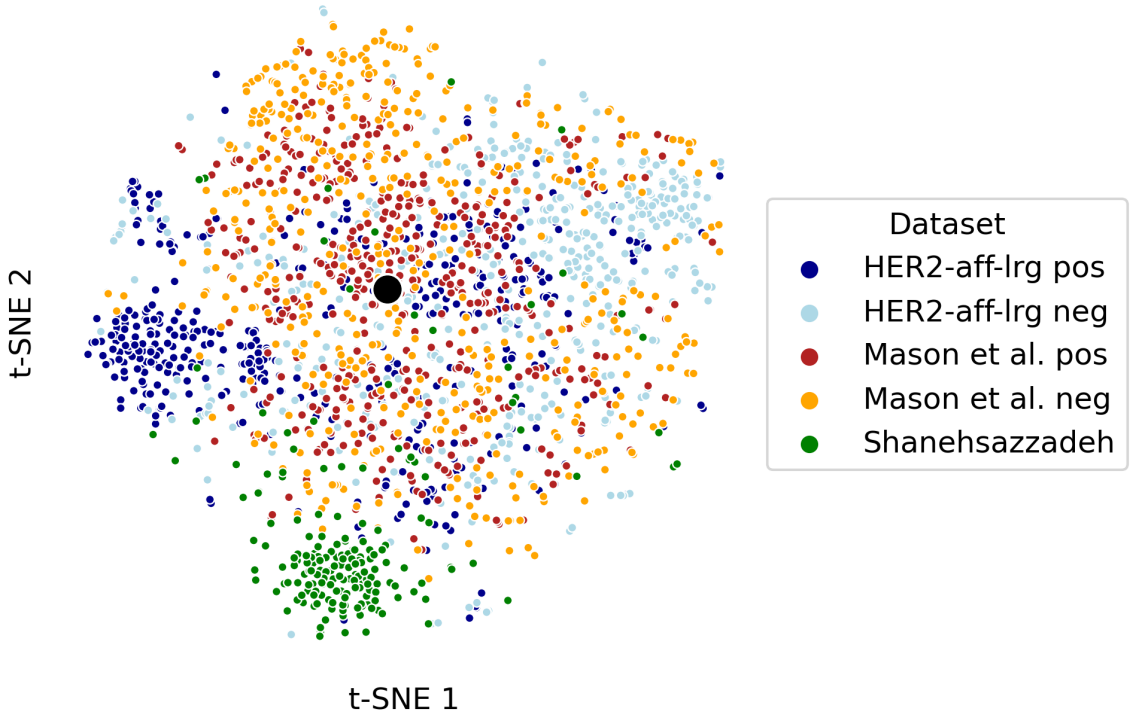

Figure B.2: **Clustering of HER2-binding Trastuzumab variant datasets.** A tSNE visualisation of the 198 Trastuzumab-length-matched designs (all binding HER2) from Shanehsazzadeh et al. (2024) along with 500 sequences randomly sampled from the positive and negative members of the dataset from Mason et al. (2021) and HER2-aff-large. Trastuzumab is shown as a large black circle in the centre of the plot. This analysis was completed and figure prepared by Lewis Chinery.

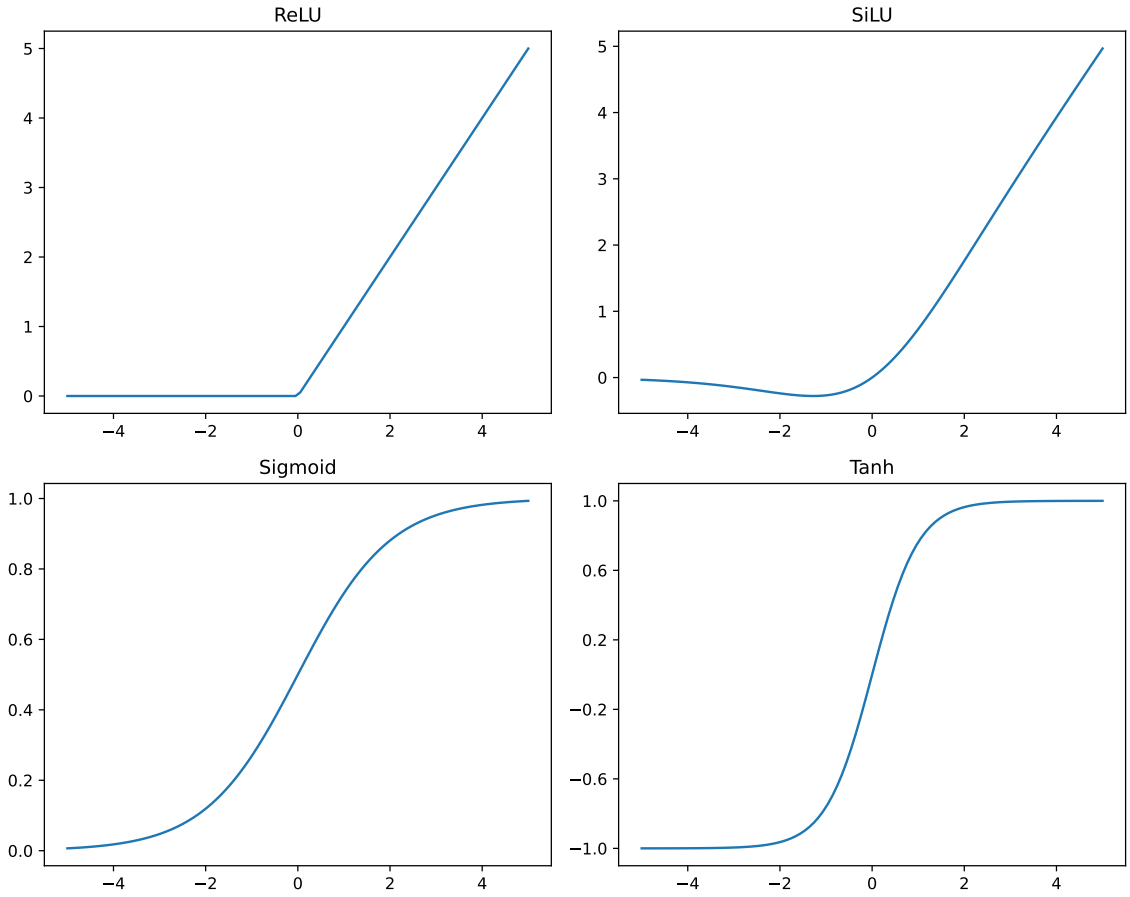

Figure B.3: **Activation functions ReLU, Sigmoid, SiLU and TanH.**

Appendix C

Supplementary materials accompanying Chapter 4: Humanization of Antibodies Using a Machine Learning Approach on Large-Scale Repertoire Data

Table C.1: **'Infixes' used for therapeutic classification.**

Source Infix	Origin
-u-	Human
-zu-	Humanized
-xizu-	Chimeric/Humanized
-xi-	Chimeric
-o-	Mouse

Table C.2: **Testing performance of RF models** on the full test dataset (left) and on a subset of the test dataset limited to sequences with <97% sequence identity with any training/validation sequence (right).

	100% Sequence Identity Cutoff			97% Sequence Identity Cutoff		
V Gene	ROCAUC	YJS	MCC	ROCAUC	YJS	MCC
HV1	1.00000000000	1.000000	1.000000	1.00000000000	1.000000	1.000000
HV2	1.00000000000	1.000000	1.000000	1.00000000000	1.000000	1.000000
HV3	1.00000000000	1.000000	1.000000	1.00000000000	1.000000	1.000000
HV4	1.00000000000	1.000000	1.000000	1.00000000000	1.000000	1.000000
HV5	1.00000000000	1.000000	1.000000	1.00000000000	1.000000	1.000000
HV6	1.00000000000	1.000000	1.000000	1.00000000000	1.000000	1.000000
HV7	1.00000000000	1.000000	1.000000	1.00000000000	1.000000	1.000000
KV1	0.99999981855	0.999552	0.999540	0.99999910846	0.999526	0.999344
KV2	0.99999999844	0.999958	0.999970	0.99999998828	0.999918	0.999937
KV3	0.99999999770	0.999526	0.999740	0.99999999130	0.999894	0.999760
KV4	0.99999999998	0.999997	0.999990	1.00000000000	0.999990	0.999985
KV5	1.00000000000	1.000000	1.000000	1.00000000000	1.000000	1.000000
KV6	1.00000000000	1.000000	1.000000	1.00000000000	1.000000	1.000000
LV1	0.99999999994	0.999996	0.999980	0.99999999897	0.999978	0.999820
LV2	0.99999999998	0.999997	0.999980	0.99999999982	0.999997	0.999970
LV3	0.99999998860	0.999950	0.999960	0.99999996100	0.999818	0.999844
LV4	1.00000000000	0.999987	0.999990	1.00000000000	0.999965	0.999954
LV5	0.99999999941	0.999995	0.999920	0.99999999418	0.999831	0.999761
LV6	1.00000000000	1.000000	1.000000	1.00000000000	1.000000	1.000000
LV7	1.00000000000	1.000000	1.000000	1.00000000000	1.000000	1.000000
LV8	1.00000000000	1.000000	1.000000	1.00000000000	1.000000	1.000000
LV10	1.00000000000	0.999692	0.999840	1.00000000000	0.999217	0.999580

Table C.3: **References and reported immunogenicity for precursor and experimentally humanized sequences of 25 therapeutics.**

Therapeutic	Immunogenicity (% of patients with ADA)	Reference
AntiCD28	NA	https://www.jimmunol.org/content/169/2/1119
Campath	5.1	https://journals.lww.com/transplantjournal/Fulltext/1999/11150/ANTI_GLOBULIN_RESPONSES_TO_RAT_AND_HUMANIZED.32.aspx
Bevacizumab	0.32	http://www.imgt.org/IMGTrepertoire/GenesClinical/humanized/bevacizumab/bevacizumab_ProteinDisplay.html#igh
Herceptin	8.1	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC49066/pdf/pnas01084-0075.pdf
Omalizumab	0	https://www.jimmunol.org/content/151/5/2623
Eculizumab	2	https://patentimages.storage.googleapis.com/f2/c4/09/171125042450cd/EP2298808A1.pdf
Tocilizumab	2	https://cancerres.aacrjournals.org/content/canres/53/4/851.full.pdf
Pembrolizumab	1.7	https://cancerres.aacrjournals.org/content/74/19_Supplement/5024
Pertuzumab	2.8	https://pubmed.ncbi.nlm.nih.gov/16151804/
Ixekizumab	8.5	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4846058/#SD1-jir-9-039
Palivizumab	1.1	https://academic.oup.com/jid/article/176/5/1215/831423
Certolizumab	8	https://patentimages.storage.googleapis.com/54/71/03/400fe464c8bb2d/US20050042219A1.pdf
Idarucizumab	4	https://patentimages.storage.googleapis.com/51/ff/74/48026e9919861a/EP2525812B1.pdf
Reslizumab	5	https://patentimages.storage.googleapis.com/b1/be/fc/ee66a606c6ed4f/CA2192543C.pdf
Solanezumab	3.5	https://patentimages.storage.googleapis.com/8a/ed/d6/645d49a2b2fae4/WO2004071408A2.pdf
Lorvotuzumab	0	https://patentimages.storage.googleapis.com/6e/13/b2/f740eceb58298a/US5639641.pdf
Pinatuzumab	1.4	https://patentimages.storage.googleapis.com/e1/b4/6f/b94b77b6b5806f/ES2543475T3.pdf
Etaracizumab	0	https://www.pnas.org/content/pnas/95/15/8910.full.pdf
Talacotuzumab	17.4	https://patentimages.storage.googleapis.com/74/d1/05/9d3a61813b2985/US8492119.pdf
Rovalpituzumab	0	https://patentimages.storage.googleapis.com/4a/00/25/e01f76b1cb6ec6/US9089616.pdf
Clazakizumab	1.82	https://patentimages.storage.googleapis.com/52/b8/4f/0146181ade3705/US20090104187A1.pdf
Ligelizumab	6.32	https://patentimages.storage.googleapis.com/9d/5c/b5/f8789f9a5c7722/US7531169.pdf
Crizanlizumab	0.94	https://patentimages.storage.googleapis.com/b9/22/2d/5e02e51d7e935a/US8377440.pdf
Mogamulizumab	4.2	https://patentimages.storage.googleapis.com/20/fc/8e/b206ab42434698/US8491902.pdf
Refanezumab	9.38	https://patentimages.storage.googleapis.com/1b/f7/0c/0e94c7d7cf18ad/US8974782.pdf

Table C.4: **Amino acid groupings based on physicochemical characteristics.** Amino acids are denoted in single-letter code.

Type	Amino Acids
Positive	K, R, H
Negative	D, E
Hydrophobic	V, M, I, L, A
Hydrophilic	Q, N, S, T
Aromatic	W, F, Y
Others	C, G, P

Table C.5: **Random humanization of Certolizumab, Omalizumab, Eculizumab.** A random humanization model was constructed to generate mutations randomly up to the same number of mutations as Hu-mAb. 100 million randomly humanized VH sequences were generated and the average Overlap Ratios and Adjusted Overlap Ratios were calculated. This analysis was completed by Claire Marks.

Therapeutic	Overlap Ratio	Adjusted Overlap Ratio
Certolizumab	1.8%	6.0%
Omalizumab	1.9%	6.8%
Eculizumab	1.3%	5.2%

Table C.6: **Comparison of mutation locations for humanization performed experimentally and by Hu-mAb.** This analysis was completed by Claire Marks.

Residues		VH					VL				
		Proportion of Muts.		Muts. Per Sequence		OR	Proportion of Muts.		Muts. Per Sequence		OR
		Hu-mAb	Exp.	Hu-mAb	Exp.		Hu-mAb	Exp.	Hu-mAb	Exp.	
Interface	Mean	6.2%	7.3%	0.8	1.6	73.8%	8.2%	10.1%	0.8	1.8	96.4%
	Median	4.3%	5.7%	1.0	2.0	100.0%	7.7%	10.0%	1.0	2.0	100.0%
Vernier Zone	Mean	14.1%	10.4%	2.2	3.0	51.7%	4.8%	5.0%	0.4	1.0	70.0%
	Median	14.3%	12.5%	2.0	3.0	58.3%	0.0%	4.6%	0.0	1.0	100.0%
Surface	Mean	67.1%	67.6%	9.6	17.1	72.5%	63.3%	64.9%	6.7	12.4	78.1%
	Median	66.7%	66.7%	9.0	16.0	70.0%	63.6%	66.7%	6.0	13.0	83.3%
Buried	Mean	32.8%	29.8%	5.1	8.1	51.5%	36.7%	32.6%	3.9	6.3	74.3%
	Median	33.3%	33.3%	5.0	9.0	60.0%	36.4%	31.3%	4.0	7.0	75.0%

Table C.7: **The most and least frequent amino acids found at high-mutual information positions.** Analysis from 400,000 human VH3 and 400,000 non-human sequences (Hu-mAb-1e6-Subset dataset), as well as 814,455 camel VH and 729,784 camel VHH sequences. Sequences were numbered using ANARCI with the IMGT format.

High-MI Position	Most Frequent AA (%)								Least Frequent AA (%)	
	Human		Non-Human, Non-Camel		Camel VH		Camel VHH		Human	
20	R	(92)	K	(73)	R	(83)	R	(75)	F	(0.0003)
54	S	(58)	G	(61)	S	(51)	A	(85)	M	(0.0015)
68	A	(76)	N	(44)	A	(73)	A	(67)	C	(0.0070)
96	A	(60)	S	(69)	P	(52)	P	(80)	W	(0.0003)

Table C.8: **Single-domain VH therapeutics with available sequences and anti-drug antibody (ADA) data.** This dataset was collected by Ashley Wong.

Name	# Monomers	Phase	# Participants	% ADA	Reference
2Rs15d	1	-	20	0	Ackaert et al., 2021; Keyaerts et al., 2016
		-	20	5	
ALX-0061	2	I/II	37	0	Ablynx, a Sanofi company, 2019a, 2019b; Holz et al., 2013; Van Roy et al., 2015
		II	250	41	
		IIb	187	31	
ALX-0081	2	I/Ib	64	0	Bartunek et al., 2013; Peyvandi et al., 2017, 2016; Scully et al., 2019
		II	36	9	
		III	145	3	
ALX-0141	3	I	42	0	Schoen et al., 2013
ALX-0171	3 3	I	60	0	Detalle et al., 2015
		IIb	135	34	
ALX-0761	3	I	33	30.3	Merck KGaA, Darmstadt, Germany, 2016
ATN-103	3	I/II	266	3	Ablynx, a Sanofi company, 2016, 2013a, 2013b
M6495: Construct 579	3	-	50	6	Buyse et al., 2018
M6495: Construct 581	2	-	50	0	Buyse et al., 2018
TAS266	4	I	4	75	Papadopoulos et al., 2015

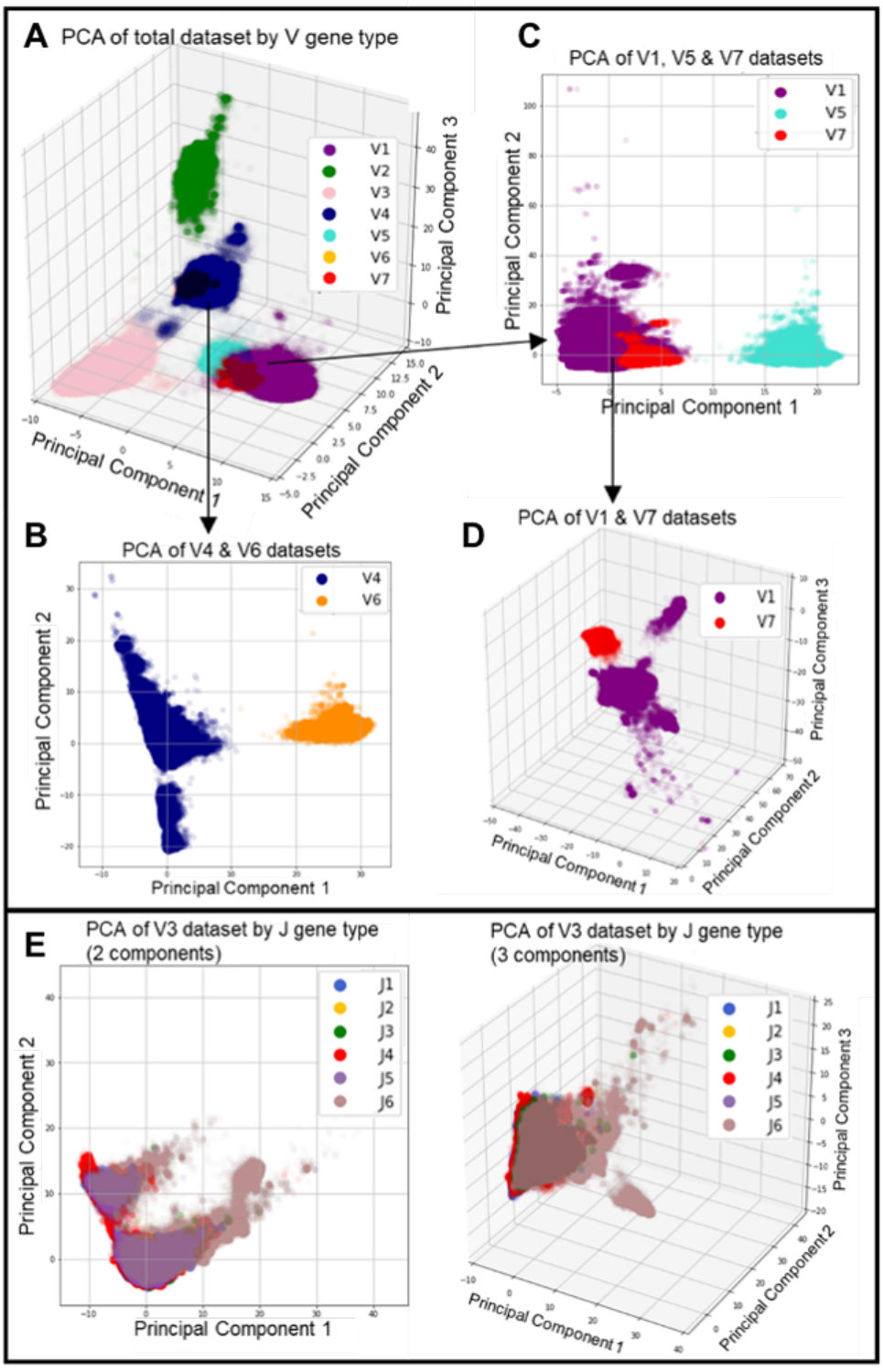

Figure C.1: **Principal component analysis of VH sequences by V gene type and J gene type.** Sequences were labelled by their V gene type (A,B,C,D) or by their J gene type (E). This analysis was completed and figure prepared by Mark Chin.

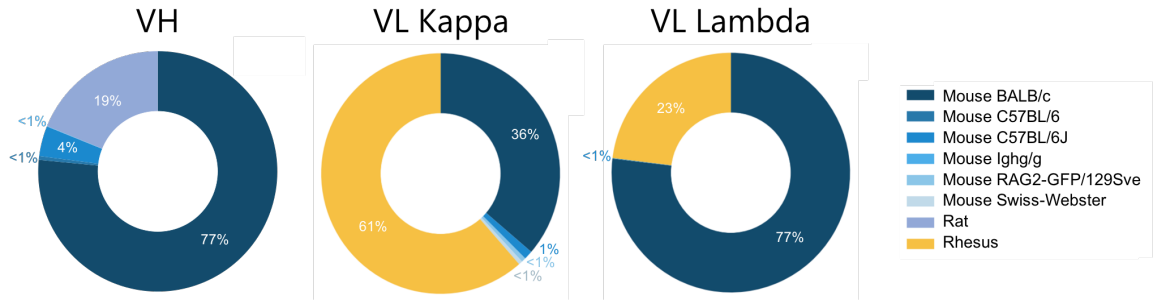

Figure C.2: **Breakdown by species of negative sequences downloaded from the Observed Antibody Space database after filtering.**

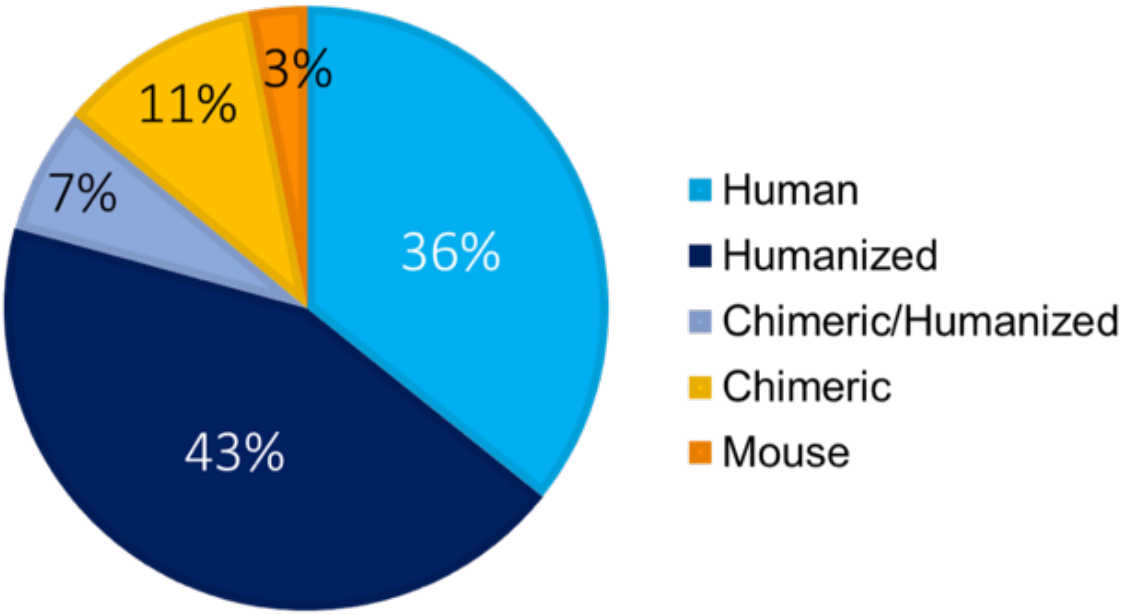

Figure C.3: **Therapeutic antibodies split by origin.** The therapeutics, approved or in phase 1-3 trials, were gathered from the Therapeutic Structural Antibody Database (Raybould et al., 2020), which had a total of 481 antibody therapeutics intended for human use. This figure was prepared by Claire Marks.

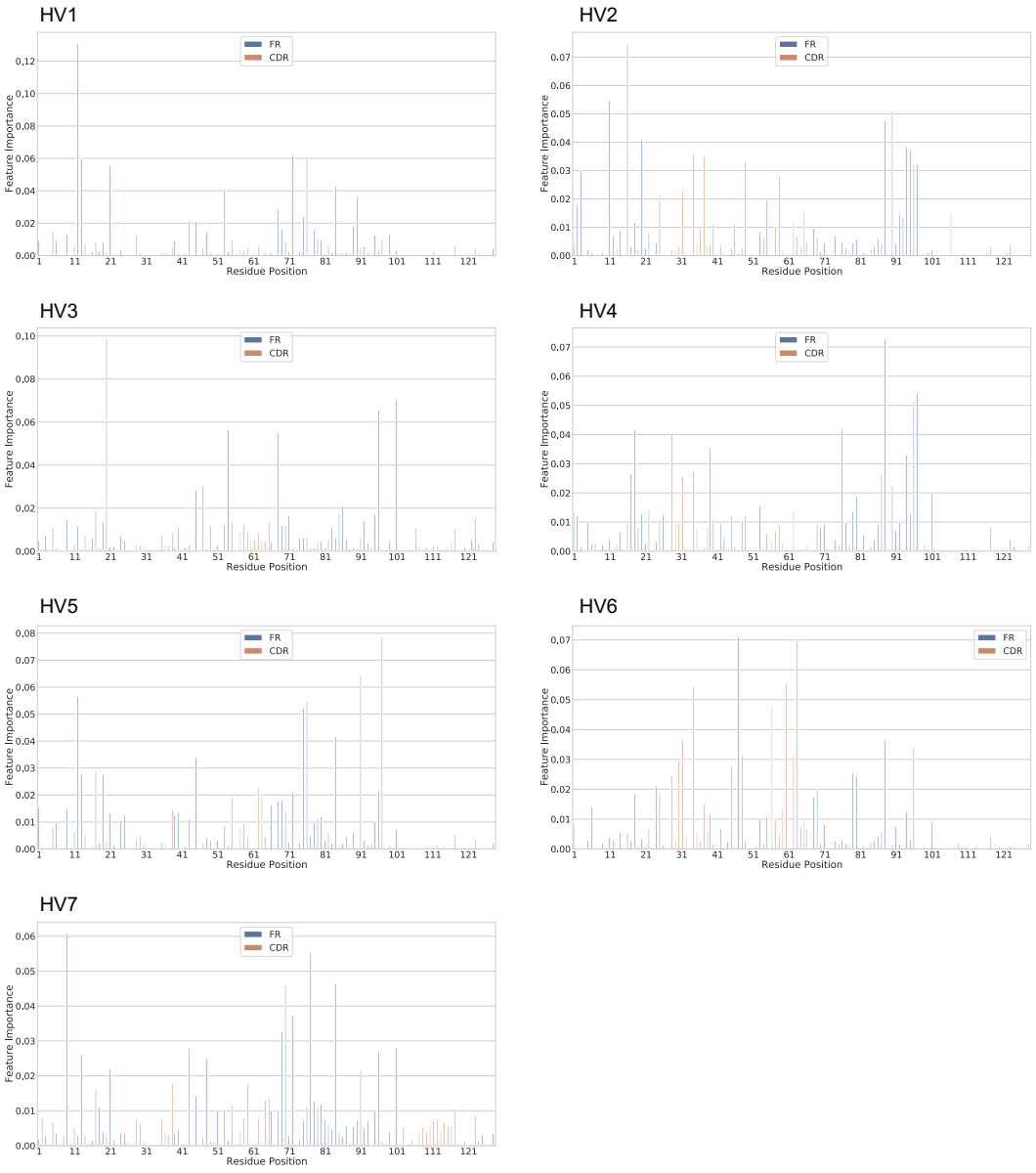

Figure C.4: **Feature importance of heavy chain RF models.** The x-axis consists of the residue positions in a sequential manner (left to right, IMGT numbering scheme).

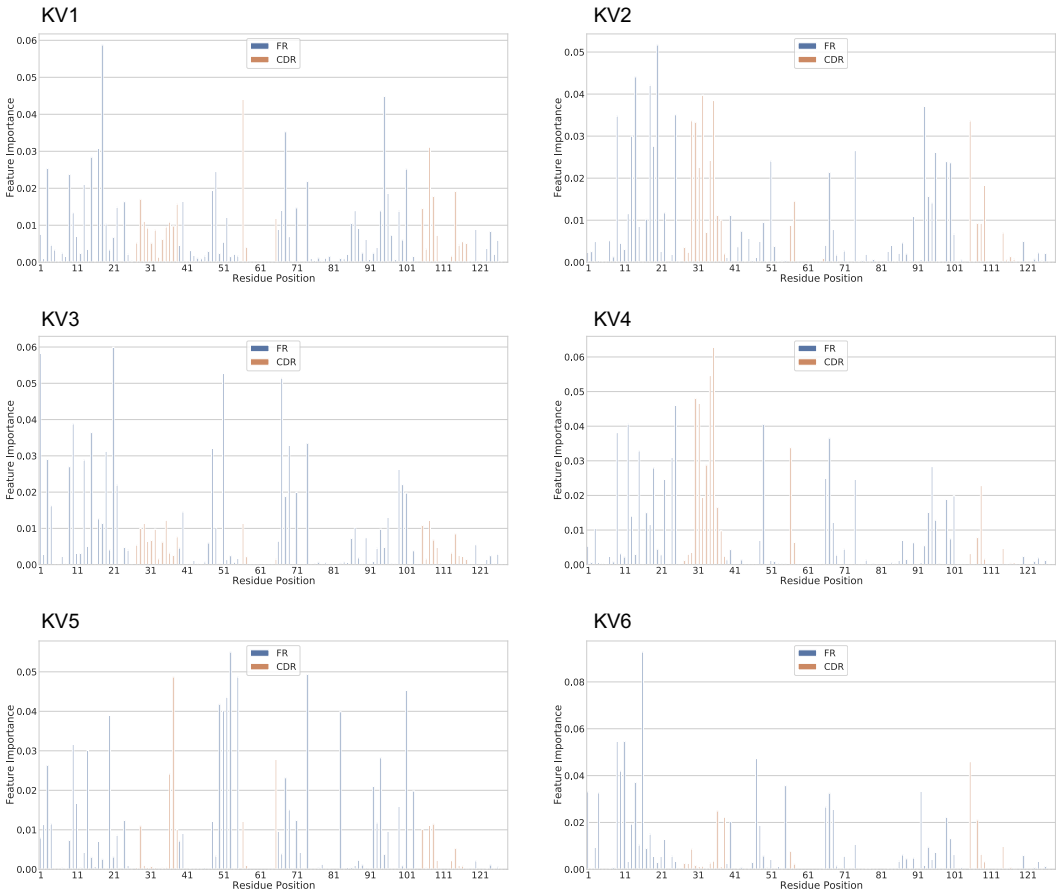

Figure C.5: **Feature importance of kappa light chain RF models.** The x-axis consists of the residue positions in a sequential manner (left to right, IMGT numbering scheme).

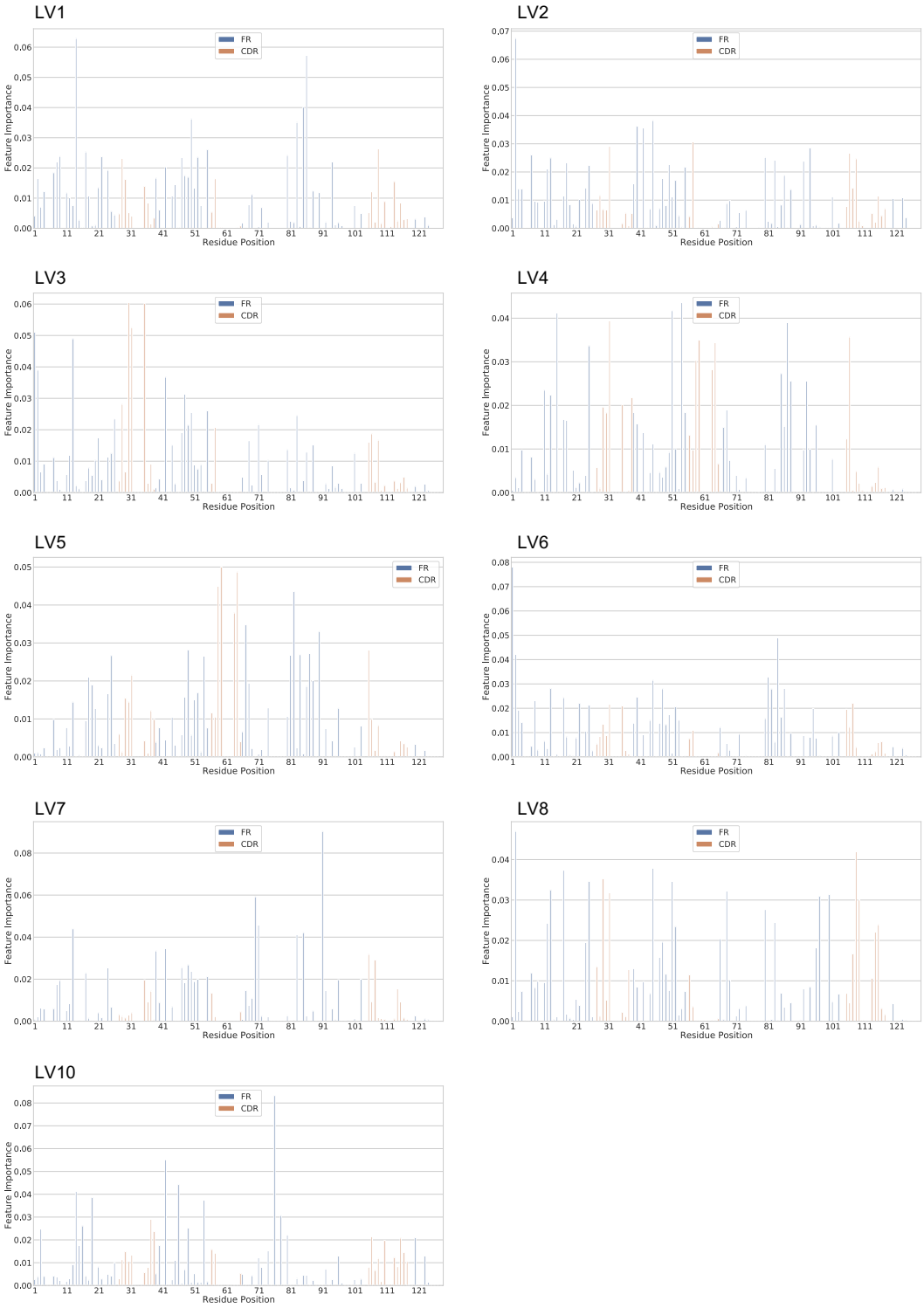

Figure C.6: **Feature importance of lambda light chain RF models.** The x-axis consists of the residue positions in a sequential manner (left to right, IMGT numbering scheme).

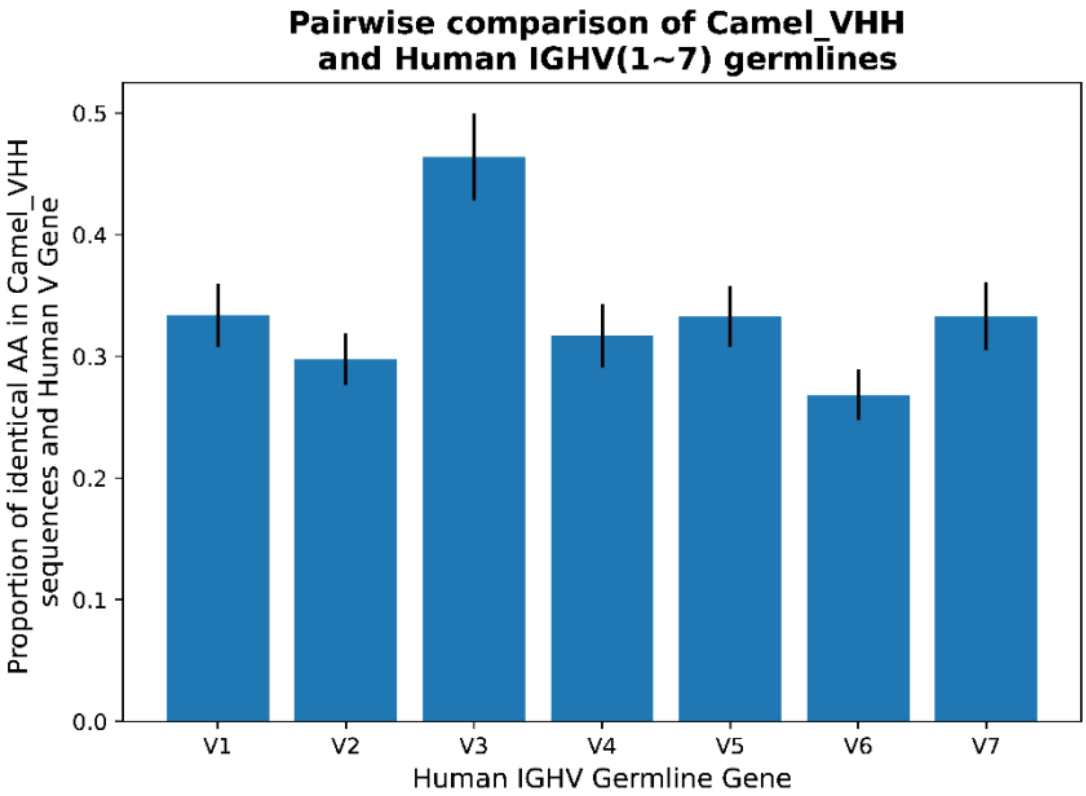

Figure C.7: **Sequence identity between camel VHH sequences and Human IGHV 1-7 germlines.** This analysis was completed and figure prepared by Ashley Wong.

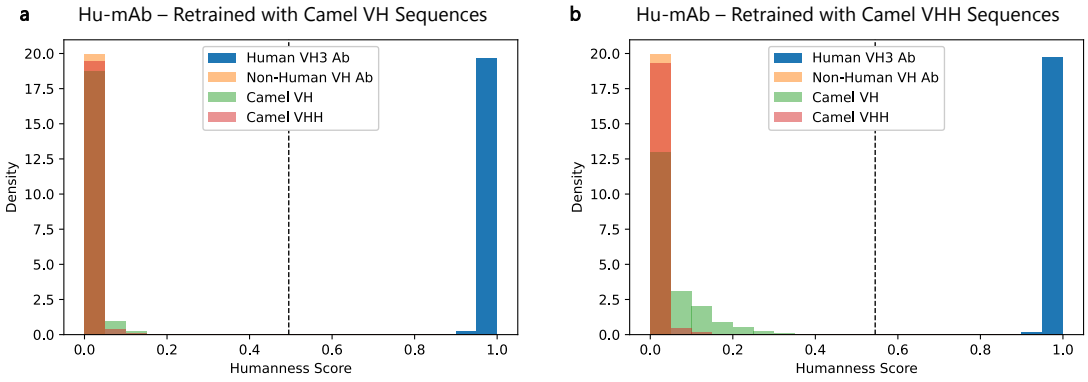

Figure C.8: **Humanness scores of camel VH and VHH sequences.** The scores of human, non-human (non-camel), camel VH and camel VHH sequences predicted by the RF VH3 model retrained with (a) camel VH sequences or (b) camel VHH sequences in the negative dataset. The respective humanness thresholds (0.495 and 0.545) are shown in dotted lines.

Appendix D

Supplementary materials accompanying Chapter 5: Antibody Inverse Folding for Improved Structure-Based Sequence Design

Table D.1: **AntiFold fine-tuning parameter evaluation, applied to validation dataset (experimental, "Exp", structures).** The training (layer-wise learning rate decay, train masking) and testing (test masking) parameters are indicated. The values in the right side of the table represent amino acid recovery for a particular IMGT region (FR: framework, CDR: complementarity-determining region). The highest value is shown in bold, the second-highest in italics.

Exp/Pred	Layer Decay	Train Masking	Test Masking	FR Avg.	CDRH1	CDRH2	CDRH3	CDRL1	CDRL2	CDRL3
Exp	–	Shotgun	None	**0.845**	0.695	0.606	0.532	0.597	0.584	*0.609*
Exp	–	Span	None	0.814	0.635	0.506	0.364	0.521	0.516	0.505
Exp	–	Shotgun + Span	None	*0.842*	0.675	0.601	0.525	0.570	0.559	0.582
Exp	–	Shotgun – IMGT-Weighted	None	0.835	**0.708**	**0.640**	*0.543*	**0.613**	**0.628**	**0.626**
Exp	–	Span – IMGT-Weighted	None	0.807	0.636	0.511	0.365	0.535	0.521	0.516
Exp	–	Shotgun + Span – IMGT-Weighted	None	0.837	0.688	0.631	0.533	0.591	0.611	0.601
Exp	✓	Shotgun	None	*0.842*	**0.708**	0.620	*0.543*	0.601	0.567	*0.609*
Exp	✓	Span	None	0.803	0.621	0.500	0.364	0.513	0.492	0.487
Exp	✓	Shotgun + Span	None	0.838	0.684	0.609	0.538	0.587	0.577	0.596
Exp	✓	Shotgun – IMGT-Weighted	None	0.832	**0.708**	*0.636*	0.541	*0.611*	*0.614*	**0.626**
Exp	✓	Span – IMGT-Weighted	None	0.798	0.614	0.498	0.354	0.507	0.502	0.494
Exp	✓	Shotgun + Span – IMGT-Weighted	None	0.833	*0.699*	0.629	**0.544**	0.600	0.598	0.606
Exp	–	Shotgun	CDRs	**0.832**	0.520	0.388	0.310	0.439	0.438	0.437
Exp	–	Span	CDRs	0.811	*0.622*	0.507	0.348	0.521	*0.521*	0.485
Exp	–	Shotgun + Span	CDRs	**0.832**	0.587	0.477	*0.368*	0.506	0.484	0.485
Exp	–	Shotgun – IMGT-Weighted	CDRs	0.827	0.608	0.496	0.343	0.520	**0.545**	*0.499*
Exp	–	Span – IMGT-Weighted	CDRs	0.807	**0.623**	*0.512*	0.354	*0.532*	0.511	**0.509**
Exp	–	Shotgun + Span – IMGT-Weighted	CDRs	*0.828*	0.604	**0.532**	**0.380**	**0.541**	0.511	0.493
Exp	✓	Shotgun	CDRs	*0.828*	0.524	0.386	0.307	0.428	0.446	0.434
Exp	✓	Span	CDRs	0.800	0.599	0.483	0.330	0.494	0.467	0.470
Exp	✓	Shotgun + Span	CDRs	0.825	0.582	0.483	0.348	0.476	0.466	0.465
Exp	✓	Shotgun – IMGT-Weighted	CDRs	0.824	0.580	0.476	0.350	0.478	0.498	0.466
Exp	✓	Span – IMGT-Weighted	CDRs	0.795	0.606	0.508	0.343	0.490	0.479	0.485
Exp	✓	Shotgun + Span – IMGT-Weighted	CDRs	0.822	0.609	0.497	*0.368*	0.509	0.485	0.498

Table D.2: **AntiFold fine-tuning parameter evaluation, applied to validation dataset (predicted, "Pred", structures).** The training (layer decay, train masking) and testing (test masking) parameters are indicated. The values in the right side of the table represent amino acid recovery for a particular IMGT region (FR: framework, CDR: complementarity-determining region). The highest value is shown in bold, the second-highest in italics.

Exp/Pred	Layer Decay	Train Masking	Test Masking	FR Avg.	CDRH1	CDRH2	CDRH3	CDRL1	CDRL2	CDRL3
Pred	–	Shotgun	None	**0.856**	0.703	0.617	0.519	0.600	0.611	0.604
Pred	–	Span	None	0.816	0.639	0.505	0.373	0.531	0.506	0.499
Pred	–	Shotgun + Span	None	0.851	0.697	0.602	0.510	0.580	0.563	0.596
Pred	–	Shotgun – IMGT-Weighted	None	0.850	*0.708*	*0.640*	*0.520*	**0.636**	*0.625*	**0.635**
Pred	–	Span – IMGT-Weighted	None	0.810	0.643	0.506	0.377	0.545	0.519	0.516
Pred	–	Shotgun + Span – IMGT-Weighted	None	0.844	0.701	0.628	0.513	0.589	0.604	0.602
Pred	✓	Shotgun	None	*0.853*	**0.710**	0.626	*0.520*	0.585	0.597	0.603
Pred	✓	Span	None	0.808	0.618	0.487	0.361	0.503	0.464	0.481
Pred	✓	Shotgun + Span	None	0.848	0.693	0.615	0.507	0.587	0.585	0.593
Pred	✓	Shotgun – IMGT-Weighted	None	0.847	0.704	**0.645**	**0.526**	*0.620*	**0.632**	*0.624*
Pred	✓	Span – IMGT-Weighted	None	0.803	0.615	0.509	0.359	0.512	0.499	0.493
Pred	✓	Shotgun + Span – IMGT-Weighted	None	0.842	0.706	0.634	0.518	0.596	0.612	0.605
Pred	–	Shotgun	CDRs	**0.844**	0.535	0.395	0.327	0.438	0.444	0.444
Pred	–	Span	CDRs	0.814	0.618	0.501	0.351	0.517	0.508	0.486
Pred	–	Shotgun + Span	CDRs	0.840	0.603	0.481	0.374	0.517	0.473	0.478
Pred	–	Shotgun – IMGT-Weighted	CDRs	*0.841*	0.622	0.504	0.356	0.522	**0.534**	0.492
Pred	–	Span – IMGT-Weighted	CDRs	0.810	**0.630**	*0.512*	0.356	*0.536*	*0.529*	*0.499*
Pred	–	Shotgun + Span – IMGT-Weighted	CDRs	0.836	*0.627*	**0.536**	**0.394**	**0.537**	0.509	**0.502**
Pred	✓	Shotgun	CDRs	0.840	0.540	0.388	0.319	0.435	0.445	0.426
Pred	✓	Span	CDRs	0.805	0.600	0.488	0.341	0.498	0.481	0.464
Pred	✓	Shotgun + Span	CDRs	0.836	0.600	0.464	0.351	0.494	0.468	0.463
Pred	✓	Shotgun – IMGT-Weighted	CDRs	0.837	0.586	0.476	0.361	0.482	0.498	0.475
Pred	✓	Span – IMGT-Weighted	CDRs	0.800	0.610	0.503	0.344	0.493	0.496	0.487
Pred	✓	Shotgun + Span – IMGT-Weighted	CDRs	0.835	0.612	0.487	*0.377*	0.523	0.471	0.494

Table D.3: **AntiFold final model parameter evaluation, applied to validation dataset (experimental, "Exp", and predicted, "Pred", structures)**. Each model was trained with IMGT-weighted shotgun plus span masking for 1 epoch on the large predicted OAS structure dataset, followed by training on the experimental SAbDab dataset. The other training parameters (layer-wise learning rate decay and application of Gaussian noise to the predicted OAS structures) are indicated. The values in the right side of the table represent amino acid recovery for a particular IMGT region (FR: framework, CDR: complementarity-determining region). The highest value is shown in bold, the second-highest in italics.

Exp/Pred	Layer Decay	OAS Gaussian Noise	Test Masking	FR Avg.	CDRH1	CDRH2	CDRH3	CDRL1	CDRL2	CDRL3
Exp	-	-	None	**0.898**	0.731	**0.712**	0.569	**0.723**	*0.736*	0.718
Exp	-	✓	None	**0.898**	*0.735*	0.698	0.566	0.716	0.702	0.713
Exp	✓	-	None	*0.895*	**0.741**	0.700	**0.584**	0.716	**0.741**	*0.725*
Exp	✓	✓	None	0.894	0.727	*0.702*	*0.573*	*0.720*	0.728	**0.727**
Exp	-	-	CDRs	**0.894**	0.680	0.637	*0.432*	*0.677*	*0.689*	**0.661**
Exp	-	✓	CDRs	**0.894**	**0.696**	0.651	**0.434**	**0.692**	0.680	*0.659*
Exp	✓	-	CDRs	0.890	0.675	**0.657**	0.431	0.666	*0.689*	0.658
Exp	✓	✓	CDRs	*0.891*	*0.681*	*0.653*	0.430	0.666	**0.698**	0.655
Pred	-	-	None	**0.909**	**0.753**	*0.716*	*0.561*	0.738	0.731	*0.722*
Pred	-	✓	None	0.905	0.749	0.704	0.558	0.729	0.725	*0.722*
Pred	✓	-	None	*0.907*	*0.750*	**0.730**	**0.572**	**0.746**	**0.737**	**0.730**
Pred	✓	✓	None	0.903	0.744	0.713	0.554	*0.744*	*0.733*	0.718
Pred	-	-	CDRs	**0.904**	*0.706*	0.650	**0.445**	*0.691*	*0.687*	**0.665**
Pred	-	✓	CDRs	0.901	**0.709**	**0.657**	*0.435*	**0.701**	**0.690**	*0.658*
Pred	✓	-	CDRs	*0.903*	0.695	*0.654*	*0.435*	0.675	0.675	0.654
Pred	✓	✓	CDRs	0.898	0.699	0.647	0.433	0.682	0.682	*0.658*

Table D.4: **AntiFold performance without ESM-IF1 pretraining** (i.e., with weights not initialised from ESM-IF1), with final parameters (layer-wise learning rate decay, IMGT-weighted shotgun plus span masking and application of Gaussian noise to the predicted OAS structures). The values in the right side of the table represent amino acid recovery for a particular IMGT region (FR: framework, CDR: complementarity-determining region). The highest value is shown in bold, the second-highest in italics.

Exp/Pred	**Test Masking**	**FR Avg.**	**CDRH1**	**CDRH2**	**CDRH3**	**CDRL1**	**CDRL2**	**CDRL3**
Exp	None	0.653	0.548	0.362	0.315	0.338	0.335	0.343
Exp	CDRs	0.651	0.547	0.364	0.314	0.339	0.338	0.345
Pred	None	0.662	0.583	0.387	0.329	0.345	0.315	0.351
Pred	CDRs	0.662	0.583	0.387	0.328	0.347	0.323	0.352

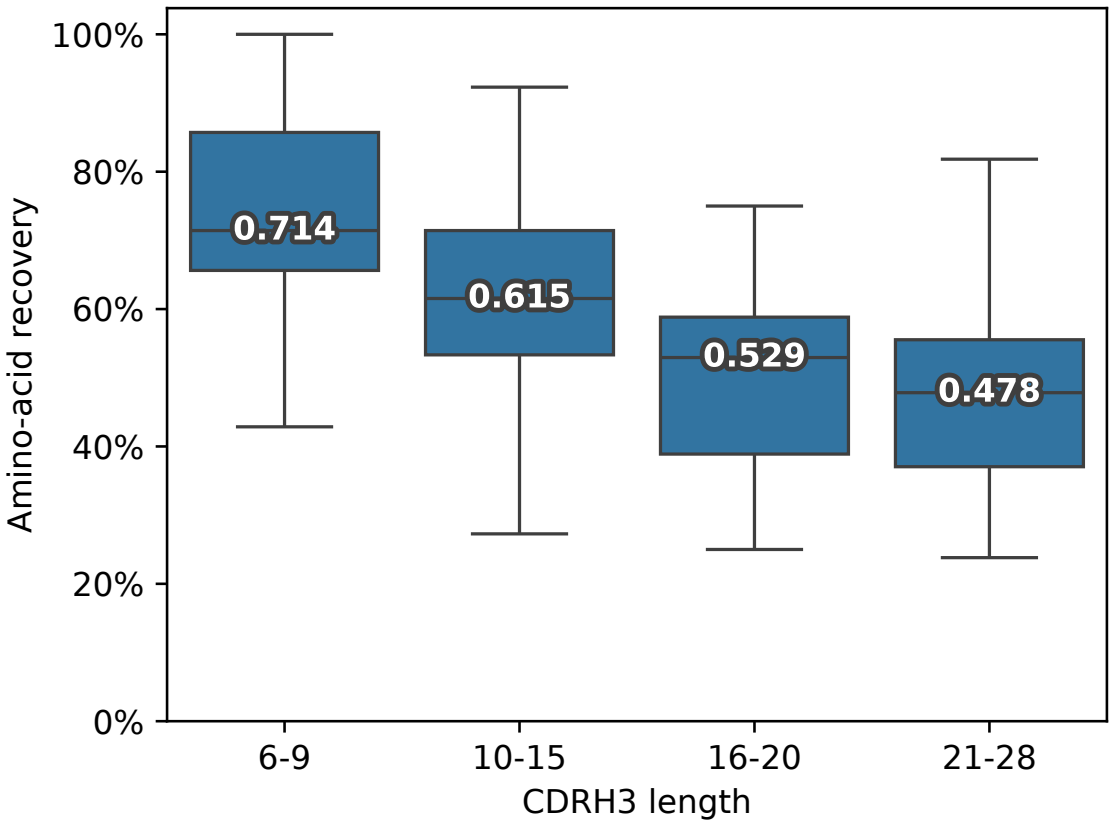

Figure D.1: **AntiFold amino acid recovery is higher for shorter CDRH3 loops.** Test-set amino acid recovery stratified by CDRH3 length.

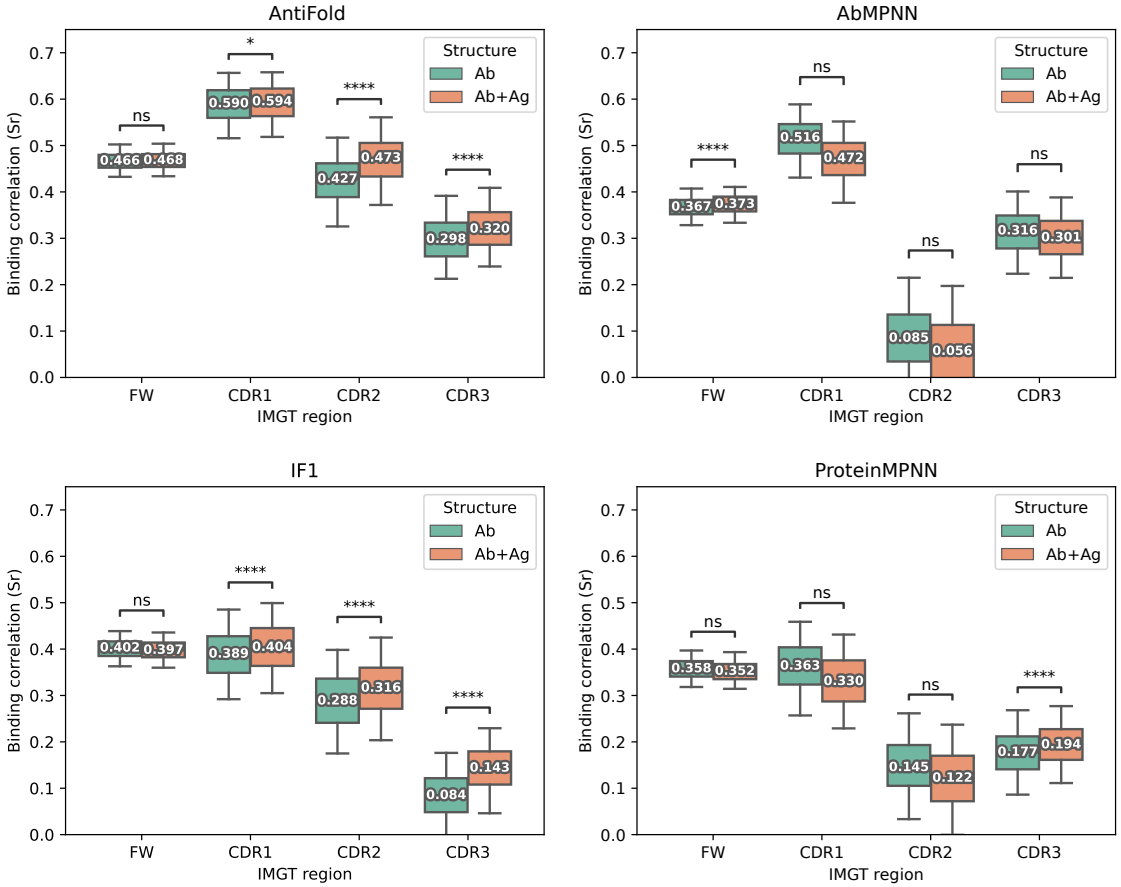

Figure D.2: **Effect of including antigen context on inverse folding model antibody-antigen binding affinity prediction.** Change in Spearman's rank correlation between inverse folding model scores and experimental affinity values in the Warszawski et al. (2019) deep mutational scan, excluding (Ab) and including (Ab+Ag) the antigen chain and split by regions. Values were bootstrapped 1000 times. Mann-Whitney one-tailed (less) U test results are shown (* = $p < 0.05$).