Evaluating Protein Language Models as Antibody Structure Learners

Jake Pencharz

Technical University of Munich
Department of Electrical And Computer Engineering

In collaboration with

The Machine Learning Research Group
Bayer AG

Master's Thesis submission

Abstract

Antibodies are a critical component of the adaptive immune system. Their high target specificity combined with their modularity has served to popularize these molecules as a class of biotherapeutics. Characterizing antigen binding is contingent on having accurate knowledge of an antibody's structure. However, experimentally solving the structure of a protein is prohibitively expensive. Using efficient computational methods to extract structural features from a sequence is becoming feasible due to advances in deep learning. We investigated a class of self-supervised networks trained on large sets of proteins known as Protein Language Models (PLMs).

Here we present the extent to which these models grasp the structure of an antibody. A sparse, weighted combination of a PLM's internal embeddings has been shown to accurately predict inter-residues contacts in proteins. However, it is unclear whether these methods work for antibodies. Our experiments showed that pretraining the LM on antibodies provides little to no benefit for contact prediction. Rather, pre-training a PLM on general proteins and weighting attention maps based on their relevance to antibody structure yields the best results. Regardless, no models achieved high precision for rare contacts or those involved in hypervariable regions.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the Technical University of Munich or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the Technical University of Munich or any other University or similar institution except as declared in the Preface and specified in the text.

Jake Pencharz
March, 2023

Acknowledgements

This thesis was completed under the guidance of Santiago Villalba and Adrien Bitton within the Bayer Machine Learning Research group.

The work would not have been possible without the generosity and mentorship received from Prof. Ruben Portugues at the Technical University of Munich.

Contents

Glossary

APLM A PLM which is pretrained using only antibodies.

CDR Complementarity determining region within each FAB. These are highly variable loops which are largely responsible for an antibody's binding behaviour .

Fab Antigen binding fragment on the tip of each branch of a Y-shaped antibody .

GPLM A PLM which has been pre-trained using a set of "general" proteins. This is different to an APLM which is pretrained on antibodies.

LR Logistic Regression is a statistical model which, despite its name, is used for binary classification..

MI Mutual information.

NMR Nuclear Magnetic Resonance Spectroscopy is an experimental technique for protein structure determination.

PLM Protein language model: this usually refers to a transformer-based architecture which is trained on sequences of amino acids .

RF Random Forests are ensemble methods. They comprise a collection of decision trees, all trained independently on the same dataset..

VDJ The process by which variable (V), diversity (D) and joining (J) gene segments are combined during transcription. .

Chapter 1

Introduction

Antibodies, or immunoglobulins, are specialized proteins which form the foundation of all jawed vertebrates' adaptive immune response. Their role is to identify and bind to potentially pathogenic molecules known as antigens (Ag). Each antibody is finetuned to bind to a specific structural sequence motif (epitope) within the Ag. If bound to a functionally relevant area, the antibody will neutralise the threat directly. Otherwise, the antibody functions as a tag, recruiting other immune cells to eliminate the threat.

The ability to target a protein complex with high specificity has popularized monoclonal antibodies (mAbs) as a class of biotherapeutics [59]. In fact, mAbs have come to dominate the pharmaceutical market, comprising seven of the ten top-selling drugs of 2018 [25]. Oncology is arguably the branch of medicine in which mAbs have found the most success. For example, Trastuzumab, a popular drug for treating HER2-positive (human epidermal growth factor receptor 2) cases of breast cancer [25], is a rationally designed mAb. It acts by binding to the extracellular domain of the HER2 protein triggering the degradation of the cancer cell [8].

1.1 Antibody discovery and design

Antibodies have the potential to be an incredibly flexible tool in building therapeutics. Given that a person is ill, the ultimate goal is to be able to design an antibody which targets and neutralises the cause of illness. An instrumental goal to that end would be to develop an antibody that targets a known epitope. Discovering, or actively designing such a molecule, is riddled with difficulties.

In the vast landscape of possible antibody conformations, finding a molecule appropriate for medicinal use is challenging. Many characteristics need to be optimised simultaneously. The most important features of an antibody are its binding affinity, a measure of the strength with which an antibody binds to a target, and specificity, the ability to discriminate between the antigen and other proteins. However, other important attributes include immunogenicity, folding stability, effector functions, solubility and pharmacokinetics (these features are often summarised as the molecule's "developability" [106]). Depending on the requirements, the antibody might also need to facilitate the attachment of a cytotoxic drug forming an antibody-drug conjugate. This poses an additional challenge of optimising the link between the antibody and the drug molecule [113]. Unfortunately, these properties cannot be optimised for independently since improvements in one attribute generally lead to the degradation of another [34, 41]. Drawing inspiration from the immune system's natural optimisation pipeline, many approaches have been tested to solve this Pareto optimisation problem. Three broad categories of techniques can be used to find antibodies with desirable binding functionality and characteristics [106].

The first approach directly harnesses the power of the mammalian immune system, isolating antibodies produced by the body in response to infection. This *in vivo* discovery method has led to the production of mAbs that target new strains of influenza [121]; or neutralise severe acute respiratory syndrome (SARS) coronavirus [115]. Developing a feasible drug requires optimising around a wild-type antibody. These are often produced by murine or camelid immune systems and need to be humanized to prevent harmful side effects. Despite the huge number of naturally occurring, wild-type sequences, there is little control over how much of the conformational space is explored using this technique.

To avoid this limitation, the second approach does not rely on the design prowess of the immune system. *In vitro* discovery relies on laboratory screening of large libraries of diverse antibody sequences. In a common procedure known as phage display, immunoglobulins from these libraries are expressed on the surface of bacteriophages. They are subsequently exposed to a target peptide adhered to a solid surface. After washing, it is assumed that only the remaining phages express antibodies that bind to the target. This selection cycle can be repeated several times to maximise binding affinity [42]. One issue with this approach is that the selection process does not mimic the conditions of the human body. With a heavy focus on binding affinity, the resulting antibodies often possess poor biophysical properties and can even elicit an immune response [106]. Another challenge is targeting a specific epitope on the target antigen. Phage display selects for antibodies which bind to immunodominant epitopes since these locations facilitate the tightest binding, rather than a functionally relevant epitope with lower affinities [106]. Lastly, screening campaigns can also be costly and time-consuming [113].

Recently, to reduce costs and have more control, *in silico* approaches to library generation and screening have shown much promise. A computational design framework, referred to as rational design [106], allows parallel screening for several biophysical properties as well as selection based on the ability to target specific epitopes. Increased control over biophysical properties, such as solubility, is important since natural selection does not optimise for characteristics associated with developability (the ease with which a molecule can be administered as a drug). For example, a successful antibody candidate is required to be soluble at the concentrations required to avoid aggregation when administered as a drug. However, protein abundance *in vivo* has been shown to be highly correlated with solubility [110] which supports the hypothesis that natural proteins tend to be only as stable and soluble as is required for them to perform their function [109]. This example points to the need for techniques, other than evolutionary pressure, to discover developable antibodies. It is feasible that computational approaches could generate antibodies with improved biophysical properties by searching the space of possible proteins left unexplored by evolution. This space is incredibly vast and it is therefore important to be able to quickly and accurately determine whether or not a candidate molecule binds to a target.

Structure-based design is a sub-category of these emerging computational approaches. If the 3D structure of a candidate is known, computational tools such as SnugDock [120] can be used to dock the antibody to a target antigen. Rational design based on docking procedures is just one motivation for accurate structural modelling techniques. An *in silico* structure-based design pipeline was proposed by Hummer et al. [50]. First, the structure of both the antibody and antigen is predicted. Based on these predictions and using a separate model, the locations of the paratope and epitope are inferred. After docking the predicted antibody to the targeted region on the antigen, an affinity prediction method will quantify the binding affinity. Affinity data can be used to optimise the candidate antibody sequence. Much progress is required, but recent developments in structure prediction [55] indicate that a process such as this is on the horizon of possibilities.

1.2 Antibody structure

To understand why mAbs are so effective at targeting antigens, one must turn to their structure and, therefore, to the principles which govern protein folding.

1.2.1 Protein folding

As workhorses of the cell, proteins act as essential structural and motor elements serving as the catalysts for virtually every biochemical reaction neccesary to living. These macromolecules are comprised of one or many chains of amino acids, each of which is folded into a unique conformation. That intricate 3D structure allows the protein to perform its function via distinct surface characteristics that determine which molecules a protein interacts with and the nature of that interaction. To study a protein's function, it is therefore useful to be able to determine its shape. Due to their size, it is not possible to visualise a protein using a microscope. Techniques such as X-ray crystallography, Nuclear Magnetic Resonance Spectroscopy NMR, or the recently popular [16] Cryo-EM are used to uncover their complex atomic configurations.

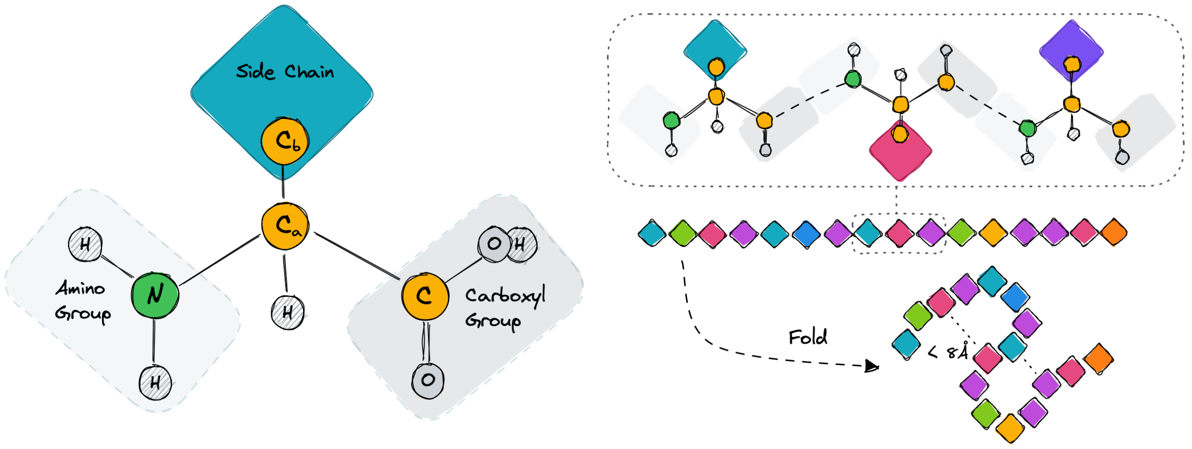

Figure 1.1: Cartoon illustration of how amino acids combine into chains. Proteins comprise amino acids, the backbones of which contain the same series of atoms: N, C_α, C. The first atom of the side chain is C_β for all amino acid types other than glycine. On the top right is an example of a short peptide chain formed by the joining of these amino acids. The bottom right illustrates how a chain of amino acids, the primary structure, can fold into a complex geometry. For the purposes of this document, two residues are considered to be in contact with one another if they are separated by at least six residues in the sequence and are at most 8Å apart in 3D space.

Protein structure is usually explained incrementally starting from the primary structure: a chain comprising amino acids joined sequentially by peptide bonds (Fig 1.1). Only twenty amino acids are directly encoded in the human genome. Each of these is uniquely distinguished by its side chains which determine the monomer's properties, such as whether it is acidic, basic, aliphatic or aromatic. Every amino acid shares a backbone of N, C_α, C atoms, and every side chain, other than glycine, attaches to this backbone via a C_β atom. Peptide bonds between amino acids are fixed and cannot rotate. However, all the other single bonds in the chain are free to rotate allowing the chain to twist and fold. Locally, chain topology tends to converge on a set of common structural motifs stabilised by hydrogen bonds. These folding patterns, referred to as the chain's secondary structure, include alpha helices and beta-pleated sheets. The ensemble of helices, sheets, and other shapes combine across the chain into a tertiary structure. Often, a chain can be conceptually split into separate domains which fold independently of one another. A protein can comprise multiple chains, intertwining to form the macromolecule's quaternary structure. The final shape of the folded chains, stabilised by thousands of non-covalent bonds, is the most energetically favourable.

The relationship between a protein's sequence and the shape it folds into was first formalised by Christian Anfinsen. Anfinsen's "thermodynamic hypothesis" states that a protein's native structure in a given environment is entirely determined by its amino acid sequence [5]. It is worth noting that there is some evidence against this hypothesis.

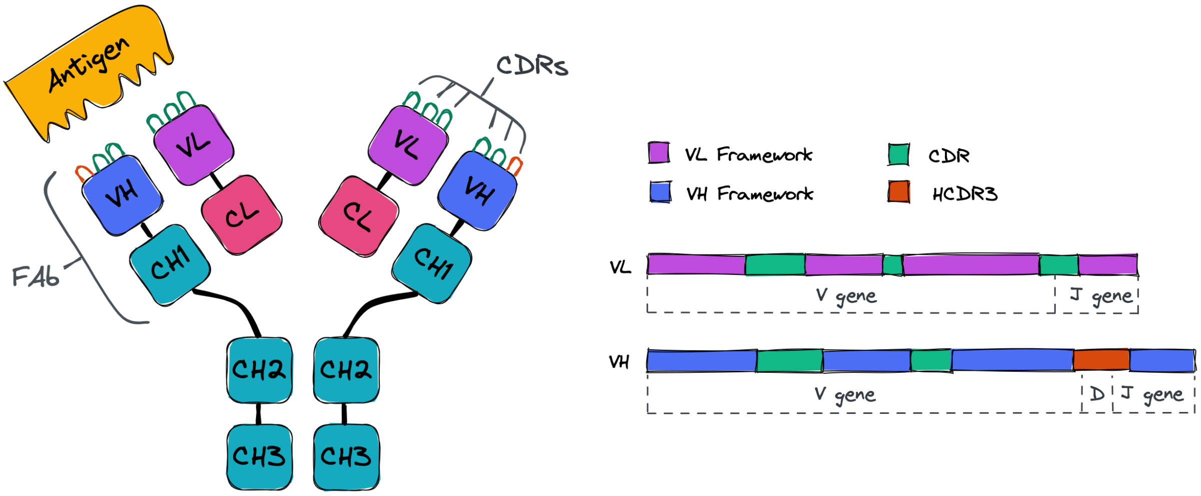

Figure 1.2: Cartoon illustration of an antibody's structure and variable domain sequences. The structure of an antibody (left) contains a heavy chain, shown in shades of blue, and a light chain shown in shades of pink. The variable fragment, F_v, is comprised of the variable regions from both chains (VH, VL). The crystallizable fragment of the antibody (F_c) contains the conserved regions of both chains (CH_1, CH_2, CH_3, CL). The two tips of the Y-shaped structure are known as antibody binding fragments (Fabs) and are responsible for binding to a matching antigen. The ability to bind to a target depends on the shape of this binding fragment and is highly determined by the complementarity-determining regions (CDRs) which lie in the variable regions of each chain and characterise the paratope. On the right is an illustration of the sequences of both chains' variable domains. Each domain has four framework regions and three CDRs. Approximate mappings between gene segments (VDJ in the heavy chain, VJ in the light chain) and regions of the variable domain are also shown.

For example, Rosenberg et al. [95] showed that synonymous codons, codons that code for identical amino acids, can affect how the protein folds. This implies that identical sequences can have diverging conformations. To add further complication to this relationship, proteins are not truly rigid bodies. Each molecule could have more than one energetically favourable state, and may even require flexibility to perform its function. An example of protein flexibility being beneficial is a phenomenon known as "induced fit". To promote an even tighter fit between itself and its target, a protein will change its shape upon binding. This behaviour is commonly observed among enzymes during substrate binding. Because of this, proteins are often described as being in one of two states: *apo* when unbound and *holo* when bound.

1.2.2 Standard model of antibody structure and diversity

An antibody typically consists of two heavy and two light chains linked by disulfide bonds (Fig 1.2). Each light chain comprises a variable (VL) and constant (CL) domain and each heavy chain includes a variable (VH) and three constant domains (CH1, CH2, CH3). The base of the antibody, known as the crystallizable fragment (Fc), only includes constant domains and is not involved in antigen binding. It contains conserved glycosylation sites that mediate interactions with other parts of the immune system. Instead, the tips of the symmetrical Y-shaped structure act as the interface between an antibody and its target. These are aptly named antigen-binding fragments (Fabs) and contain the variable domains of the heavy (VH) and light chains (VL). In an effort to identify previously unseen foreign invaders, the immune system spawns a massive variety of Fabs [29, 68]. Within each Fab, diversity is concentrated in six β-strand loops, three per chain, known as the complementarity determining regions (CDR). The 3D structure of these loops and their relative positions to one another and to the remainder of the chain are of critical importance: they are largely responsible for an antibody's binding properties.

The most diverse of these regions, and arguably the best predictor for epitope binding [68], is the third CDR loop on the heavy chain (HCDR3). Located at the centre of the binding site, it is largely responsible for the site's topology, makes the most contacts with the epitope [73], and strongly affects binding energetics [65]. H3 loops range in shape from long, finger-like projections to short loops that form cavities to interact with a protrusion on the epitope [73].

Although briefly defined in Sec 1, a more precise definition of the epitope is the set of antigen residues directly involved in antigen-antibody binding. When forming a complex, these residues come into contact with a corresponding set of residues on the surface of the antibody known as the paratope. Although CDR loops are a reasonable approximation of the paratope, they are not synonymous. A residue is only considered part of the paratope if it lies proximal to the antigen or significantly contributes to a decrease in Gibbs free energy upon binding [27].

There are approximately 10^{13} unique antibody sequences in the human antibody repertoire [44]. Although sequence diversity does not map directly to structural diversity, this number indicates the scale at which the immune system evolves new antibody conformations. Diversification of the antibody repertoire occurs in two stages. The first is during B-cell maturation in a process known as VDJ recombination [56]. During this process, B-cells randomly rearrange segments of the genes that code for the CDRs. The gene coding for VH contains multiple variable (V), diversity (D) and joining (J) segments and the gene that codes for VL contain multiple V and J segments. This means that LCDR3's sequence depends on the random recombination of a V and J segment whereas HCDR3 depends on the random recombination of V, D, and J segments (Fig 1.2). During transcription, the HCDR3 sequence is determined by a random selection of these segments being spliced together into one of 10^{15} conceivable combinations [24]. Dependence on more than one randomly selected gene makes both CDR3 regions hyper-variable.

The second source of variability occurs in a process known as somatic hypermutation.

This process is triggered when an immunoglobulin binds to a target. Binding triggers the B-cell to which the immunoglobulin is attached to replicate rapidly. During replication, mutations which increase binding affinity are selected for in an attempt to boost epitope specificity. As a result of gene segment recombination and stochastic mutations, there exists a huge variety of CDR sequences and structures capable of binding with high affinity and specificity to myriads of targets.

When comparing antibodies it is useful to represent the sequences such that the conserved regions overlap and the differences between regions (which are important for binding) can be examined. This is made complicated by occasional insertions, deletions and variability in the CDRs which cause variable domains to have different lengths. To solve this problem several "numbering schemes" have been proposed [123, 48, 3, 2, 69]. Based on patterns extracted through the analysis of thousands of antibodies, these schemes assign sequential numbers to each residue in the domain based on which region of the variable domain that residue is likely to belong. For example, in the IMGT numbering scheme, residues in HCDR3 will be numbered from 105-117 [69].

Modern numbering schemes are based on a combination of sequence identity and structural similarity. Early numbering schemes, such as that proposed by Wu and Kabat [123] who first identified the high entropy regions of the variable domain, only take sequence information into account. As the number of solved antibody structures grew, it became possible to integrate structural information into these descriptive frameworks [3, 69]. The location of CDRs is prone to shifting depending on data availability and analysis techniques. These dependencies mean that there is no consensus delineation between framework regions and CDRs and that schemes will differ in how they handle insertions and deletions.

By using structural analysis to develop a framework aimed at highlighting differences between antibodies, structural similarities between CDR loops were discovered. Other than HCDR3, all CDR loops adopt a limited number of canonical conformations [3]. These canonical classes are found by clustering the CDR loops based on their structural similarities [78, 19]. Clustering CDRs by their structural features rather than relying on sequence similarities reveals that a region's class is length-independent [79]. Clusters in the structural space are mapped to the sequence space by relying on key residues which dictate the shape of the loop [78, 3]. Therefore, the presence of these residues, despite small changes elsewhere in the sequence, is sufficient to identify the class of the CDR. Importantly, no canonical clusters exist for HCDR3 which makes identifying the loop's conformation using only sequence-level information an unsolved challenge (see Sec 1.4 for more detail).

Furthermore, HCDR3 loops vary significantly in length. Typically, these loops comprise 3 to 20 residues [73]. However, some bovine antibodies have HCDR3 loops with more than

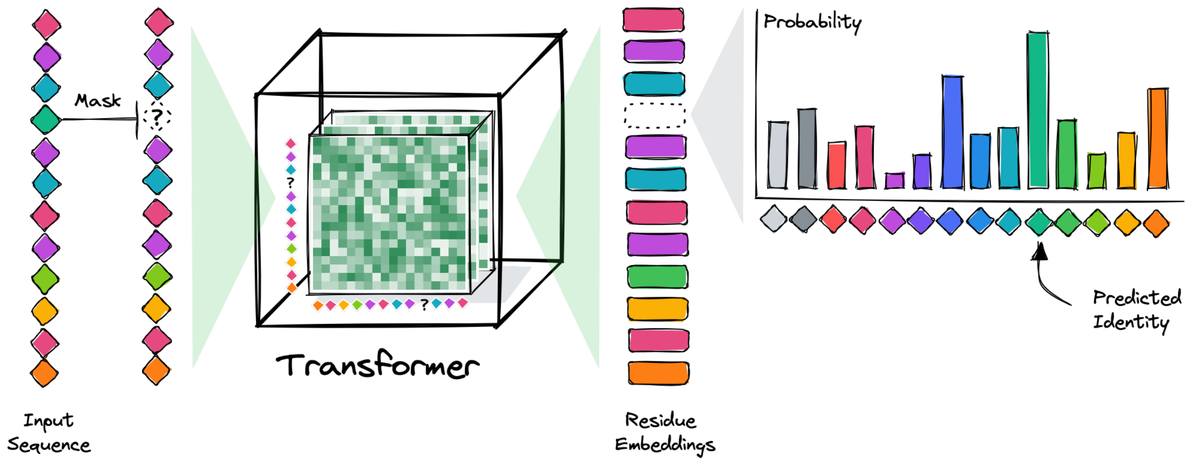

Figure 1.3: Training a Transformer using Masked Language Modelling. A mask is applied to one or many input tokens. A well-trained Transformer should output a distribution over its vocabulary at a masked position which displays a peak at the masked input token. Correctly guessing the most likely token at a masked position requires understanding the surrounding sequence. The relationships between the input tokens are captured in the model's attention weights.

60 residues. In contrast, other CDR loops do not vary drastically in length. Each loop has, at most, 8 unique lengths and is far shorter than HCDR3 [73].

Although the orthodox model of Ab-Ag binding claims that CDRs are solely responsible for the antibody's binding behaviour, this may not be entirely true. CDRs do not correspond directly to the residues directly participating in binding (paratope). It is estimated that only 20-33 % of CDR residues are involved in Ag binding [84]. Furthermore, there is evidence that the crystallisable fragment of the antibody also plays a role in determining binding affinity [104]. These factors are important to keep in mind during structure-informed design or analysis of antibodies.

1.3 Protein language models

Self-supervision is a training paradigm in machine learning which does not require labelled data. Since large labelled datasets are expensive to produce this class of methods has recently gained a lot of traction. In the field of natural language processing (NLP), self-supervised models are trained on proxy tasks such as predicting the next word in a sentence given all previous words [86, 14] or filling in masked words using the context of the entire sentence [26]. These tasks are designed to guide the language models (LM) in learning the statistical properties of natural languages. LM's abilities increase with larger training datasets [58] allowing them to take advantage of huge corpora of unlabelled text.

Next-generation sequencing has made prodigious amounts of sequence data from ostensibly functional proteins available. Few of these sequences are well-characterised or have solved structures. To take advantage of the copious unlabelled sequence data, techniques from NLP have been applied to proteins. Protein Language Models (PLMs) are self-supervised models which learn useful representations of proteins from their sequence data. UniRep [4] is a recurrent neural network (RNN) that learns to reconstruct sequences sequentially, one amino acid at a time. To accurately predict the successive amino acid, the model learns a rich, fixed-length representation of the preceding sequence. These representations were used in predicting structural clusters as well as protein stability and functional effects. Folding of protein domains often causes residues which are far apart in sequence space to be proximal in 3D space. Although UniRep uses a Long Short Term Memory (LSTM) network [64] it is limited to modelling dependencies between earlier residues in the chain. This quirk introduced during training may result in less structurally informative representations.

Transformers [116] are a class of self-supervised models which were introduced for natural language processing (NLP) and have been used with great success in a broad range of applications [55, 87]. Their main innovation is a self-attention mechanism which enables the network to learn pairwise relationships between its input tokens. Since attention is of central importance in my investigation it is worth describing in some detail.

Given an input set of vectors $\mathcal{X} = \{x_1, x_2, ..., x_n\}$, self-attention aims to find a set of weights which model the dependencies within the set. In the version introduced by Vaswani et al, an input vector x_i is mapped, using a learnable function such as a neural network, to a query, key and value vector where $q_i, k_i, v_i \in \mathbb{R}^d$. This terminology is borrowed from a dictionary data structure where the query is used to retrieve a value based on which key it matches. Similarity scores between the query and key value are computed using inner products:

$$z_{ij} = q_i \cdot k_j \quad \forall j = [1, ...n] \tag{1.1}$$

By stacking query and key vectors as rows of two matrices $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{n \times d}$, the inner products between every query and key vector can be computed using matrix multiplication $\mathbf{Z} = \mathbf{Q}\mathbf{K}^T \ \in \mathbb{R}^{n \times n}$. The result is then normalised by the size of the vectors and softmax [13] is applied such that the similarity scores, now referred to as attention weights, resemble probability distributions:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right) \quad \in \mathbb{R}^{n \times n} \tag{1.2}$$

Each weight in the attention matrix, a_{ij}, captures the dependency of x_i on x_j. These dependencies are then used to generate a weighted sum of value vectors $\mathbf{O} = \mathbf{A}\mathbf{V} \ \in \mathbb{R}^{n \times d}$ where a row o_i will be most heavily influenced by the value vectors relevant to token i.

Many layers M of self-attention are usually applied to an input set, continuously transforming the input tokens by incorporating information from its context. Furthermore, each self-attention layer is "multi-headed" where each of the H attention heads implements the same operation but uses independent sets of weights. A transformer therefore generates $M \times H$ square attention matrices and a set of output embeddings $\mathcal{E} = \{e_1, e_2, ...e_n\}$ where $o_i \in \mathbb{R}^d$.

In natural languages, there are a fixed number of possible input tokens, a vocabulary $\mathcal{V}$. Another transformation is learned which maps the output embedding at each position to a probability distribution over that vocabulary. The distribution computed from embedding e_i should display a peak associated with the word used to generate input token x_i. As mentioned above, one method used to train transformers entails masking a percentage of the words in the input sentence and tasking the model to predict the identity of those missing tokens (see Fig 1.3). The idea is for the model to use the context provided by the rest of the sequence to generate output embeddings that will be mapped to distributions assigning high probability to the masked words. The likelihood of predicting the correct token is, therefore, $p_\theta(x_i|\mathcal{X} \setminus \{x_i\})$ where the set difference operator $\setminus$ is being abused to indicate that a token x_i is masked and θ represents the model's trainable parameters. Through this proxy task, known as Masked Language Modelling (MLM), Transformers implicitly learn the grammar of natural languages and are highly adept at understanding long-range dependencies between words in a sentence [86].

Translating this technique from sentences to proteins is straightforward. Instead of words, use amino acids, and instead of the English or Japanese lexicon use the 20 amino acids that the human body encodes as a vocabulary. Now, rather than modelling the statistical relationships between words, the Transformer strives to understand the relationship between each amino acid and its chain thereby learning what some [32] have referred to as the "language of life".

Several Transformer-based PLMs have recently been released after being trained on vast amounts of sequence data. Many of them, such as ESM [91], ProtTrans [32] and ProteinBERT [10] use a BERT-based architecture [26] and are trained using various denoising proxy tasks, of which MLM is an example, to reconstruct their input sequences. The goal of these models is to generate rich sequence embeddings that can be used for downstream tasks such as predicting structure, or biophysical properties [88, 10]. Generative models, such as ProtGPT2 [35], are another class PLMs focused on *de novo* protein design. Motivated by the success of the GPT-x [86] family of LMs, the authors used an autoregressive training protocol, predicting the next amino acid in a sequence given the preceding residues. Although these models show promise, the focus of the investigation presented here is on the prior class of sequence-embedding PLMs and their relevance to structure prediction.

During masked language modelling, a Transformer-based PLM is required to predict the identity of a subset of masked amino acids. Consider a structural biologist given the same task. If they had access to the structure of the protein they would utilise their profound understanding of the biochemical properties of each amino acid to solve the puzzle. First, they might examine the masked positions in their 3D context to find other residues that are close by. Given the charge, shape and other biochemical properties of these surrounding residues, the expert would carefully pick out an amino acid to fill the masked position that would "fit" into its surroundings maintaining stability.

Of course, a language model does not possess this expertise and it only has access to sequence data. However, following Anfinsen's hypothesis [5], one can assume that all of this information is encoded in the sequence. Therefore, as it strives to improve on its language modelling objective, through complex pattern matching, the model might learn to extract this information. Self-attention intuitively captures the interaction of every residue in the sequence to a query residue. In trying to predict the identity of a masked residue an accurate map of residue-residue interactions is invaluable. Since proximity in 3D space implies that two residues interact, it should, therefore, not be entirely surprising if some attention maps model the distances between residues. Rao et al. [89] used a simple method to demonstrate that this is in fact the case (see Sec 2.1). As we shall see later, this emergent phenomenon makes PLMs appealing components in modern structure prediction pipelines.

As well as their attention matrices, PLMs derive several useful representations of a sequence. At the final self-attention layer, a fixed-length embedding $e_{res}^{(i)} \in \mathbb{R}^d$ is generated for each residue in the sequence at the output of the network. After being trained on a large set of proteins, residue embeddings have been shown to capture biochemical exchangeability [91] between amino acids. Averaging along the length dimension returns a fixed-length sequence embedding $e_{seq} = \frac{1}{L}\sum_{i=1}^{L} e_{res}^{(i)}$. Analysis of sequence embeddings have been shown to encode remote homology and protein-family alignment [91].

1.4 Review of antibody structure prediction

Rational design of antibodies requires the ability to accurately gauge binding affinity with an antigen. However, sequence-level information is not sufficient for characterising this behaviour. A consistent mapping between a sequence and its interaction with a given epitope has proved challenging to define. Most CDRs adopt canonical conformations (see Sec 1.2.2), but predicting the activity of the antibody is made complicated by the variability of HCDR3. Despite differing in length and containing different amino acids, HCDR3 sequences have been shown to adopt similar structures and target the same epitope [63, 24, 92]. Furthermore, identical HCDR3 sequences embedded in different chains can have different binding properties [24]. For all their differences, these sequences represent functionally equivalent proteins. Therefore an HCDR3 sequence is "necessary but

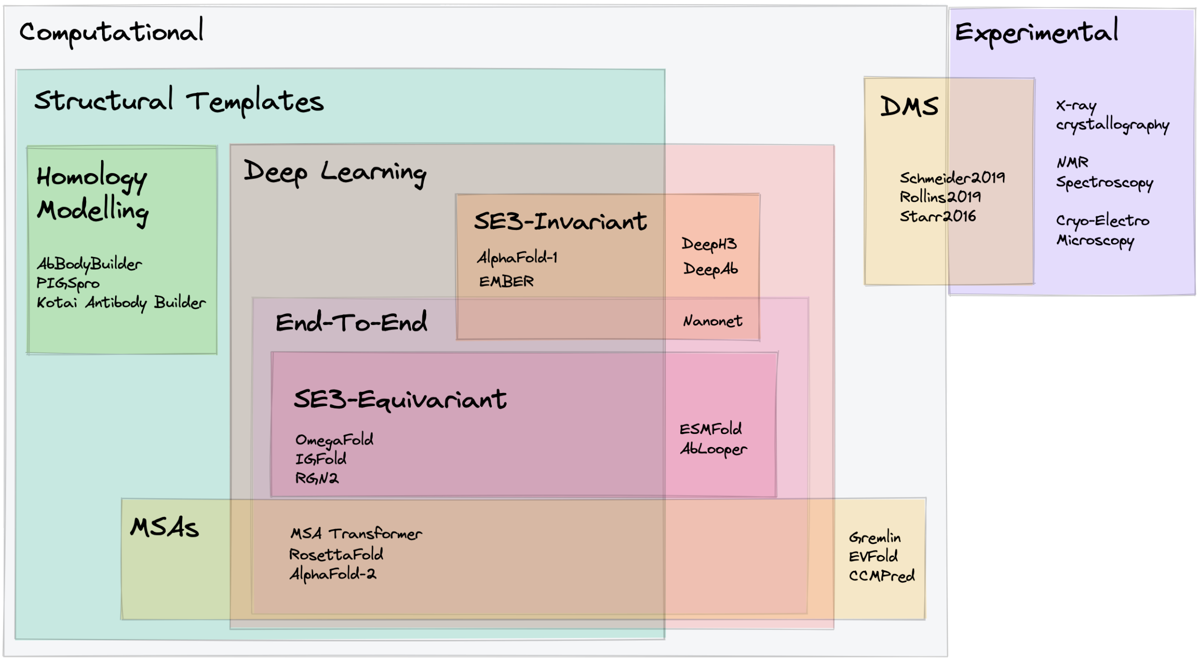

Figure 1.4: Simplified taxonomy of antibody structure determination techniques. Some examples of techniques to solve or predict the structure of antibodies. There is a range of experimental and computational approaches which have been used in a variety of combinations. One category of techniques which rely on epistatic effects of mutations makes use of deep mutational scanning and is therefore neither a purely computational nor experimental approach. Many of these techniques are protein structure prediction methods and were not developed for antibodies specifically. Also note that there is some overlap between categories (for example, Cryo-EM has a large computational component) but the placement in the taxonomy is dependent on the core component of the structure determination technique.

insufficient" [24] to characterise binding. For this reason, when designing a biotherapeutic a model of the structure is required to forecast antigen binding [18].

This is not unique to antibodies. The structure of all proteins is of fundamental importance in determining their function. Unfortunately, structural information is not readily available. Experimentally determining the structure of a protein is expensive and time-consuming. This means that the total number of solved protein structures, antibodies in particular, is low. Furthermore, there are even fewer experimentally solved protein complexes. Solved complexes are desirable because, as mentioned earlier, the conformation of a protein can change upon binding. Despite being less readily available, the bound or *holo* conformation of the protein is of more interest than the protein in its *apo* state.

In contrast to this, the amount of protein sequence data is growing exponentially thanks to advances in next-generation sequencing [39, 44]. With access to this wealth of sequences, a data-driven system which can routinely predict the structure of a protein from its sequence is highly desirable - this is often referred to as the "protein folding problem". The ability to quickly predict how a chain of amino acids folds will allow efficient exploration of sequence space [20, 49] facilitating the rational engineering of sequences with desirable functions.

If we assume Anfinsen's hypothesis is correct, predicting a protein's structure from its sequence seems entirely feasible. Anfinsen postulated that the molecule's conformation corresponds to a stable minima in free energy which is entirely determined by interactions between the residues in the chain. Unfortunately, modelling this energy landscape is combinatorially intractable. As the length of the chain increases, the number of possible conformations that it can assume grows exponentially. Calculating the free energy of each possible fold is simply not practical.

Several methods have been proposed to overcome these hurdles. Most computational methods for predicting the structure of antibodies are distributed into one of two broad categories: template-based and template-free (or *ab-initio*) methods (see Fig 1.4). Template-based methods utilize a database of solved structural templates, taking advantage of antibodies' highly conserved structures. Template-free methods build structures *ab initio* from the amino acid sequence alone. A few approaches from each of these categories are described below. However, to contextualise the antibody-specific approaches, a brief review of important ideas in protein structure prediction is presented.

1.4.1 Protein structure from co-evolution

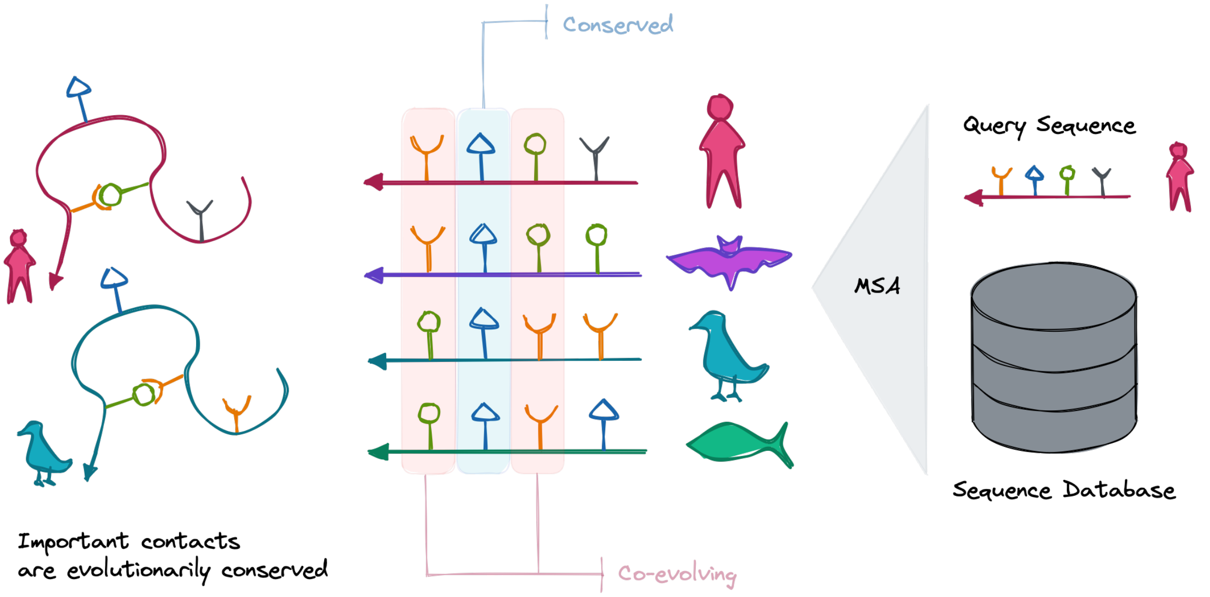

Figure 1.5: Multiple sequence alignments identify contacts. For a mutation to become persistent it cannot interfere with the function of a protein. By aligning a query sequence with a set of similar sequences from different species, one can observe co-evolutionary patterns. This means that residues at certain positions in the chain mutate in step with one another implying that they might be in contact.

Mutations are an essential component of evolution. Most mutations have a negative effect on protein fitness and, therefore, only a small set is positively reinforced. Some mutations can have non-independent effects on protein fitness [114, 107]. Correlated mutations of residues at distal positions in a protein have been utilised to infer the structure of proteins.

In an aligned protein family residues in different positions frequently mutate together [43]. Evolutionary constraints that maintain protein function can be used to explain this behaviour. If two positions in a sequence must co-evolve for the protein to remain functional, it is probable that these residues interact with one another (see Fig 1.5). This interaction implies that they are proximal in 3D space. Thus, the 3D structure of a protein leaves echoes in the evolutionary record. Several methods [43, 30, 31, 103, 57, 74, 75, 83] have exploited this fact, using correlated mutations extracted from multiple sequence alignments (MSAs) to predict which residues in a chain are close together. Rather than predicting distance as a continuous value, many methods aim to predict residue-residue contacts, a binary metric generated by thresholding distances, or distograms that bin distances into a fixed number of ranges. Knowing which residues are proximal in 3D space greatly constrains the number of possible conformations the molecule could adopt.

One of the main problems encountered when using correlations to predict contacts is indirect couplings. For example, if x is coupled to y and y is coupled to z, a naive correlation-based method would assume that x is coupled to z which might not be the case. The network of residues which interact directly can be conceptualised as hidden beneath the observable correlations. To overcome this, one can model the dependencies between positions in a chain using a discrete Markov Random Field (MRF) known as a Potts Model. Using this model, one can distinguish between direct and indirect positional couplings. Methods such as Direct Coupling Analysis (DCA) [77] and Protein Sparse Inverse COVariance (PSICOV) [54] fit observations from the alignment to an approximation of the statistical model by inverting a sparse residue-residue covariance matrix. Achieving a higher accuracy, GREMLIN [57], plmDCA [31] and the more computationally efficient CCMPred [103], instead maximise the pseudolikelihood of the MRF. In both cases, the direct couplings established in the statistical model are used as a proxy for residue-residue contacts. GREMLIN's authors found that their approach was likely to produce accurate contact predictions when the depth of the MSA (with sequence redundancy at a maximum of 90%) is greater than five times the length of the protein [57].

More recently, MSAs have been used in conjunction with deep learning to predict contacts from co-evolutionary patterns. Rather than using expensive feature extraction techniques such as DCA, Jones et al[53] use construct a pairwise frequency tensor as an input feature to a convolutional neural network. To do this, they count the number of every possible amino acid pair between every column in an alignment. For an alignment with m columns and a set of 21 amino acids (20 true amino acids and a gap character), the resulting feature has dimensions $441 \times m \times m$. In 2018 DeepMind beat the other competitors in the 13th Critical Assessment of Techniques for Protein Structure Prediction (CASP13) by a significant margin. Their method, AlphaFold1 [105], used the parameters from a regularised pseudo-likelihood trained Potts model (similar to CCMPred) as well as features that explicitly represented gaps and deletions in the MSA as an input feature set to their network. Rather than predicting contacts, AlphaFold1 applies a dilated [125] convolutional residual network to predict a distogram (64 bins deep). Binned, pairwise distances are then smoothed, parameterised according to the residues' torsion angles and used to optimise for those angles using gradient descent.

Two years later DeepMind repeated its success at CASP14 with AlphaFold2. The new method, considered a breakthrough in protein structure prediction [76], trounced its competition, achieving near experimental accuracy in the majority of cases [55] and inspiring a new generation of methods such as RoseTTAFold [7], and AlphaFold-Multimer which can predict the structure of protein complexes [33].

The number of innovations required to achieve such remarkable results is too numerous to summarise in this brief review. However, the introduction of attention-based mechanisms is important to draw attention to. The body of the AlphaFold2 network, the EvoFormer, iteratively refines raw MSAs using axial attention while incorporating information from the sequence's pairwise representation (initialised using structural homologs). The modified query sequence at the top of the refined MSA is then translated into a 3D structure using a novel SE(3) equivariant [1] transformer. By avoiding a physics-based folding engine, the network became end-to-end differentiable. It is important to note that raw MSAs, rather than the complex features derived from MSAs used in AlphaFold1, were used as a primary input to the AlphaFold2 network. This choice was validated in an ablation study (see Supplementary material, Sec 1.13 for [55]). According to their own analysis, the accuracy of their model decreased substantially with an MSA depth lower than 30 sequences [55]. Therefore, in step with GREMLIN, AlphaFold2 and its related methods benefit from high-quality, deep MSAs.

This shortcoming may indicate a prerequisite for any method which relies on MSAs to extract co-evolutionary information and is of particular importance for antibody structure prediction. Orphan proteins, by definition, lack known homologs. Antibodies are often grouped together with orphan proteins because they evolve quickly and independently according to the suite of pathogens an organism is exposed to. Sequence databases do not

[1] SE(3) is the "special Euclidean" group of distance-preserving transformations, such as rotations and translations, in 3D space. If a model is not SE(3) equivariant, two identical molecules in different poses would induce entirely unrelated internal representations. The model essentially treats them as different molecules. These symmetries are essential to take into account if training an end-to-end differentiable network.

contain homologs for the most variable regions of these sequences. MSAs, therefore, do not capture the protein's evolutionary progression and may not provide useful co-evolutionary information.

Recently it was suggested that AlphaFold2, RosettaFold, and other structure prediction tools which rely on deep learning do not understand the underlying physics of protein folding [82]. The jump in performance is therefore attributed to successfully leveraging the statistical properties observed in the training dataset. In scenarios where the statistics differ from the training proteins, such as antibody structure prediction, it is possible that these methods will be less successful.

Although PLMs (see Sec 1.3) traditionally are trained on single sequences to extract patterns across protein families, AlphaFold2 is not the only Transformer-based method that uses sets of aligned sequences as input. The aptly named MSA Transformer also abandons this convention. Axial attention [47] is alternatively applied along the row and column dimensions of the alignment. Supporting their hypothesis that the homologous sequences should share structural features, the authors found that sharing an attention map across all rows (for all sequences), what they termed tied-row attention, boosted performance. The MSA Transformer does not predict 3D structure but can be used for contact prediction according to the procedure introduced by Rao et al. [89] and elaborated on in Sec 2.1. Even for sequences with few (< 100) available homologs the MSA Transformer outperforms state-of-the-art single sequence PLMs, as well as a best-in-class Potts model at contact prediction.

1.4.2 Homology modelling techniques

Because antibody conformations are highly conserved, there are many methods which rely on grafting together previously solved fragments based on sequence similarities. This is an example of template-based structure prediction, known as homology modelling. Relevant templates are chosen based on their similarity to the target sequence. Some online services using this technique include PIGSpro [70], Kotai Antibody Builder [124], and ABodyBuilder [67].

Homology modelling is capable of generating structures which are accurate (RMSD $<$ 1Å) in highly conserved regions [99]. The difficulty arises when modelling hypervariable regions of the antibody (VH and VL) which lack high-quality templates.

Fragment assembly methods generally follow a four-stage workflow [67, 63]. First, a template structure is chosen based on its sequence's similarity to the target sequence. This template can include both VH and VL or can be generated using separate templates for VH and VL [67]. For these hybrid templates, the orientation between VH and VL is determined in the next stage. Due to their high variability, template structures are usually incapable of accurately modelling CDRs. In an effort to overcome this shortcoming,

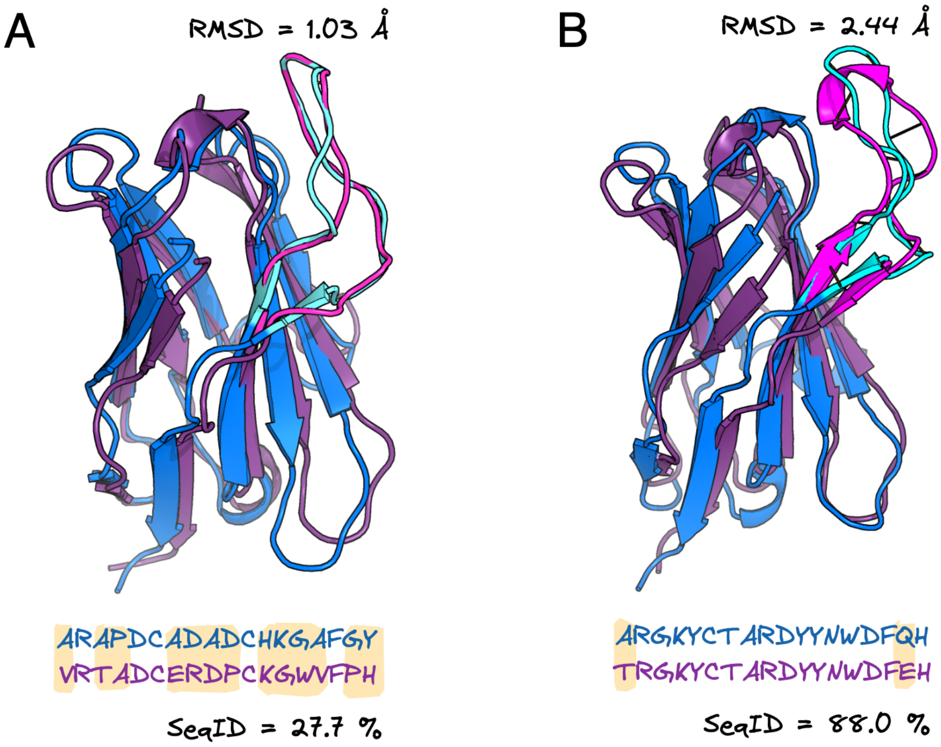

Figure 1.6: Sequence-level similarity does not represent structural similarity. By visually comparing pairs of aligned VH chains from SAbDab [29], one can appreciate that high sequence identity (SeqID) in complementarity-determining region 3 (HCDR3) does not imply similar structure. These highly illustrative pairs of chains were identified by Kovaltsuk et al. [63]. Structural disparities are reported as a root mean squared difference (RMSD). Mismatched amino acids in the sequences are highlighted in yellow. **A** The variable domains of 4s1s (blues) and 4nzu (pinks) are overlaid and the sequences of the HCDR3s are included below. Despite a blatant mismatch in sequence space, the HCDR3 loops are highly aligned (RMSD $\approx$ 1Å). **B** Same as A, but 3U7W (pinks) is compared to 4JDV (blues). In this case, the sequences are almost identical but the shapes of the loops are distinct (RMSD $>$ 2Å).

homology modelling techniques search through several CDR-specific databases, replacing these highly variable regions in the template with the most similar structures found. Finally, the side chains rotamers of residues which differ between the target and template are predicted.

If a template structure is available, homology modelling can generate structures fast. However, their weakness lies in their dependence on the availability of similar template structures. This is especially problematic for longer HCDR3 loops [73]. Structural accuracy is usually evaluated using the Root Mean Squared Deviation (RMSD) between the backbone atoms of the solved and predicted structures. Sub-Angstrom RMSD is considered perfect, where short sequences with an RMSD $>$ 2Å are structurally distinct. Framework regions and canonical CDRs can usually be predicted to within 1.5Å where HCDR3's predictions are often more than 3Å away from the native structure using these methods [63]. Even in cases where homologs are present, sequence similarity does not always correspond to structural similarity (see Fig 1.6).

1.4.3 Predicting antibody structure in the absence of homologs

For proteins which lack known homologs, such as orphan and de-novo designed proteins, it seems that models like AlphaFold are inherently handicapped by their reliance on MSAs. Similar to orphan proteins, antibodies evolve independently and are somewhat isolated from the evolutionary record. Furthermore, because HCDR3 loops are hypervariable, MSAs tend to be noisy in those regions [122]. Unlike the methods presented in Sec 1.4.1, some structure prediction techniques, including antibody-specific approaches, eschew MSAs, often in an effort to take these edge cases into account.

Mutations occur naturally in the evolutionary record but, with the advent of genetic engineering, they can also be artificially introduced. By introducing mutations at every position in a protein sequence and observing mutations that have non-independent effects on the protein's fitness, it is also possible to elucidate residue-residue contacts. Deep Mutational Scanning (DMS) [37] is an experimental technique in which large libraries of mutants are functionally assessed. A library of double mutants (two positions are mutated in each sequence) can then be used to investigate the energetic couplings between positions. Using the same principle as co-evolutionary analysis, mutations which exhibit strong epistatic effects [107] are likely to indicate structural proximity. A number of recent techniques [102, 93] illustrated that epistatic interactions found using DMS are sufficient to construct accurate ($<$ 1.9Å RMSD) models of protein structures without requiring many homologous sequences. Such experimental methods are resource intensive requiring DMS of each protein or domain. Furthermore, they rely on *in vitro* or *in vivo* selection assays which may not be available if a protein's function is not known. Regardless, for antibodies, which have well-established selection assays, DMS is an interesting alternative to purely computational approaches.

To overcome the shortcomings of homology modelling (see Sec 1.4.2) a host of methods have attempted to use deep learning to predict the structure of antibodies. Because there is rarely relevant evolutionary history available for HCDR3 loops, MSAs are ostensibly ineffective at predicting their structure. DeepH3 [96], a method that does not use any coevolutionary information, employs 1D convolutions to predict the distances and orientations (a set of dihedral angles) between residues in HCDR3 using the concatenated, one-hot-encoded heavy and light chain sequences as input. These predictions were converted to geometric potentials which are used to constrain an energy minimisation protocol that generates a 3D conformation for the HCDR3. The method was also used to differentiate between native and decoy HCDR3 sequences. DeepAb [99] builds on the groundwork of DeepH3 as well as the success of UniRep [4], a recurrent neural network

(RNN) that learned useful representations of protein sequences. The authors pretrained an RNN-based autoencoder on a large ($> 100,000$) set of sequences from the Observed Antibody Space (OAS) [80]. Sequence embeddings are then generated by concatenating the encoder hidden states for each residue. To train the structure prediction module they combine the representation generated by the 1D convolutions, similar to DeepH3, with the representation generated by the pretrained encoder. A residual, 2D convolutional network combined with six separate attention heads utilise the combined representation to predict six geometric features. Once again using a Rosetta-based energy minimisation method to generate structures from the predicted geometric potentials, DeepAb outperformed the state-of-the-art in antibody structure prediction.

Neither DeepH3 nor DeepAb can be trained end-to-end. This means that they cannot be trained directly on the atomic coordinates of known proteins. In contrast, AbLooper [1] is an end-to-end CDR structure prediction tool. To achieve this, they used an equivariant graph neural network [100] making their method SE(3) equivariant. After generating predictions for the CDR loops, ABodyBuilder [67] is used to construct a complete model of the protein, incorporating the loop predictions. AbLooper matches the performance of AlphaFold2 and DeepAb while being substantially faster.

Another end-to-end method focussed on speed is Nanonet - a nanobody [2] structure predictor [23]. Taking advantage of the highly conserved framework regions, the 2,000 training structures were aligned to an arbitrarily chosen reference frame thereby achieving translational invariance. Rather than using a complex equivariant graph-based method, this “trick” allowed the authors to use a relatively simple 1D convolutional neural network (CNN) to predict C_α coordinates directly (side chains are not modelled). When tested on heavy chains of antibodies, NanoNet achieves similar performance to DeepAb.

PLMs (see Section 1.3), which learn informative representations from single sequences, are a promising avenue to exploring poorly characterised regions of protein space. These models learn from common patterns in their input sequences. Understanding which amino acid goes where is similar to learning substitution probabilities. This can be thought of as the model learning evolutionary constraints, previously extracted from MSAs, based on the data that it has observed. At inference time, the inferred co-evolutionary context for a sequence is only available through the model’s learned parameters. PLMs have been shown to capture structural information in their attention matrices (see Sec 1.5.1). Encouraged by these findings, and the time saved by avoiding searching through sequence databases to build an alignment, PLMs have become popular components of 3D structure prediction pipelines. ESM [91] gave rise to ESMFold [71], ProtTrans [32] enabled EMBER [119], OmegAPLM opened the door for OmegaFold [122] and AminoBERT facilitated RGN2 [20]. Even though these models do not use MSAs, their performance tends to decrease on

[2] Nanobodies, common in camelid immune systems, are heavy-chain-only antibodies.

sequences which have few homologs [90, 71]. Because they are trained on a general set of proteins, usually from UniRef50 [108], I refer to them as general-PLMs (GPLM).

In an effort to better model the biophysical properties of antibodies, another set of recent PLMs, such as AntiBERTy [97], AntiBERTa [68] and AbLang [81] were trained only on antibodies from OAS. I refer to these models as antibody-PLMS (APLM). These models were created for antibody-specific tasks. For example, AntiBERTy was initially used to predict evolutionary trajectories of antibodies and later incorporated into a fast and accurate antibody-specific structure prediction network IgFold [98]. AbLang was shown to restore missing residues in sequences from OAS more accurately than ESM-1b [81]. AntiBERTa was better able to distinguish between naive and memory B-cell receptors (BCR) [3] than a ProtBERT, a GPLM, and showed a propensity to focus its attention on putative paratope residues [68].

Despite not directly using MSAs, PLMs seem to indirectly capture co-evolutionary relationships between positions by observing patterns in sequence data. If this is the case, the prediction accuracy for HCDR3 sequences remains limited by the number of neighbouring sequences in the training dataset.

1.5 Previous work analysing PLM features

1.5.1 PLMs as self-supervised structure learners

A structural evaluation of PLMs is not straightforward since these models output a set of residue embeddings rather than a fully realised 3D dimensional structure. Structures can, however, be represented in two dimensions as well. A distogram, $\mathbf{D} \in \mathbb{R}^{R \times R}$, is a square, symmetrical matrix comprising the distances between the R residues in a sequence. Although this representation does not capture dihedral angles between the backbone residues, distograms have been used as input to energy minimisation algorithms to faithfully recreate 3D conformations [99, 105]. By setting a threshold in separation in 3D space (t_{euc}) and a minimum separation in sequence space (t_{seq}), distograms can be used to determine which residues are in contact with one another. Rather than expressing the distance between residues as $d(\mathrm{C}^i_\beta, \mathrm{C}^j_\beta)$, to take glycines into account, it is convenient to simply write $d(i, j)$. Contacts are therefore defined according to:

$$c_{ij} = \begin{cases} 1, & \text{if } d(i,j) < t_{euc}\text{Å and } i - j > t_{seq} \\ 0, & \text{otherwise} \end{cases} \tag{1.3}$$

Applying this rule to a distogram, one derives a contact matrix $\mathbf{C}$. Knowing which residues are in contact with one another greatly constrains the space of possible conforma-

[3] B-cell receptors are immature antibodies bound to the surface of B-cells.

tions. Precise contact maps have proved helpful in, among other applications, accurately modelling *de novo* structures [60, 62, 74, 119, 83]. For this reason, contact prediction has been enshrined as CASP benchmark and a standard metric for comparing structure prediction methods [88].

Given a transformer with L layers and H heads, a forward pass over a sequence of R residues will generate a stack of $N = L \times H$ attention maps. Each attention matrix $A_n \in \mathbb{R}^{R \times R}$ where $n = \{1...N\}$ is square but not symmetrical. Since the Euclidean distance between two residues is symmetrical, $d(i,j) = d(j,i)$, the attention matrices must be symmetrised. This is trivially accomplished by adding the transpose of each matrix to itself: $S_n = A_n + {A_n}^T$.

Dunn et al. introduced an efficient method to estimate and remove background mutual information (MI) between pairs of residues in a protein family [30]. MI is relevant to contact prediction since it explicitly captures the dependence of one position on another. Given that two residues are in contact, knowledge about one greatly reduces the uncertainty in identifying its partner. High MI between two residues is, therefore, an indication that those positions might co-evolve and be in contact with one another. However, this signal can be corrupted by shared ancestry, other structural and functional constraints, and random noise [6]. This is especially relevant when dealing with a CDR since positions with higher entropy tend to have higher levels of random MI [36]. Given an aligned set of sequences from a protein family, found using multiple sequence alignment, Dunn et al. estimate the background MI between positions i and j in the sequence using a term they call average product correction (APC).

$$APC(i,j) = \frac{MI(i,\bar{x})MI(j,\bar{x})}{\overline{MI}} \tag{1.4}$$

Where $MI(i,\bar{x})$ is the mean mutual information between column i and all other columns and $\overline{MI}$ is the overall mean MI.

One can build a coupling matrix $\mathbf{M} \in \mathbb{R}^{R \times R}$ from MI values, where $\mathbf{M}(i,j) = \mathbf{M}(j,i)$ represents the MI between two positions. Since values in this matrix represent how informative position i is to position j Rao et al. argued that it is highly similar to an attention matrix [88]. Therefore, Equation 1.4 can be applied to every pair of positions in each symmetrised attention matrix to find background correlations between positions. The background can then be removed yielding a “corrected” attention matrix, F_n. This can be expressed using vector notation:

$$F_n = S_n - \frac{(S_n\mathbb{1})(S_n\mathbb{1})^T}{\mathbb{1}^T S_n \mathbb{1}} \tag{1.5}$$

Finally, the corrected contact matrices can be combined into a single coupling matrix which can be interpreted as a contact map (Fig 2.1). However, Vig et al. [118] found that attention heads in particular layers tend to display a disproportionate correlation to distances between residues. Therefore, some attention maps need to be weighted more highly than others. To determine which maps are most important to predicting the contacts between residues, Rao et al. [89] proposed fitting an L1-regularised logistic regression model to a small set of known structures. If $\mathcal{D}$ is a set of solved proteins, regressing a logistic regression model to this set of structures will determine an optimal set of weights $\beta \in \mathbb{R}^{N+1}$ to modulate the importance of a PLM's attention matrices. Given a model with $N = L \times H$ attention maps, after featurising the sequence of $d \in \mathcal{D}$, the model estimates the probability of residues i and j being in contact as:

$$p\left(c_d^{ij} | \beta\right) = \left(1 + \exp\left(-\beta_0 - \sum_{n=1}^{N} \beta_n {F_n}^{ij}\right)\right)^{-1} \tag{1.6}$$

Since residues which are close together in sequence space are trivially proximal in Euclidean space, a sequence separation threshold k is introduced. The likelihood of observing the (non-trivial) contacts in the training set is, therefore:

$$\mathcal{L}(\mathcal{D}|\beta) = \prod_{d \in \mathcal{D}} \prod_{i=1}^{L_d - k} \prod_{j=i+k}^{L_d} p\left(c_d^{ij} \mid \beta\right) \tag{1.7}$$

I then use `scikit-learn` [85] to solve the following optimisation problem:

$$\hat{\beta} = \max_{\beta} \ \mathcal{L}(\mathcal{D}|\beta) - \frac{1}{\lambda} \sum_{n=1}^{N} |\beta_n| \tag{1.8}$$

Where λ is a regularisation hyperparameter that encourages the combined magnitude of weights to be minimal. This results in a large percentage of the weights being close to zero, thereby identifying a sparse set of attention heads which have learned to reason about the distances between residues.

1.6 Motivation

It would seem that the trajectory of the antibody design field is gravitating towards the widespread use of PLMs. These self-supervised models, trained on large databases of protein sequences, are capable of generating informative features from single sequences [68, 32, 118] and their self-attention mechanism is surprisingly proficient at capturing structurally relevant relationships between residues [118, 98, 71, 122]. However, as is the case with all large language models, clear explanations for their internal representations are not immediately forthcoming. It can be difficult to interpret the how and why behind a model's outputs. Therefore, if the aim is to featurise a library of sequences, it remains challenging to choose a featuriser which offers the desired feature properties.

Standardised testing exists to compare structure prediction methods. Since 1994 scientists have participated in the bi-annual Critical Assessment of protein Structure Prediction (CASP) competition. Intended to incentivise progress in computational methods to predict protein structure from sequence, contestants compete to predict the structures of a set of previously unseen proteins. Predictions can be submitted in three formats: as atomic coordinates, pairs of residues in contact, or an assessment of their model's accuracy on the task.

To my knowledge, no extensive or standardised analysis exists to assess the capabilities (or potential lack thereof) of PLMs as featurisers of antibody sequence data. Since structural information is valuable to the rational design of antibodies, it seems fitting to base this analysis on a PLM's ability to reason about the structure of an antibody despite only having access to sequence-level data.

Attention matrices of transformers naturally represent relationships between residues. Some have postulated that this relationship is structurally informed [118, 91]. Furthermore, Rao et al. [89] demonstrated that contacts between residues can be readily extracted from attention weights (see Sec 1.5.1). However, this analysis was not targeted at antibodies. Therefore, using contact prediction as a proxy for structural understanding, this investigation aims to answer fundamental questions such as: Are PLMs trained on antibody sequences preferable to PLMs trained on a general set of proteins for antibody structure prediction? Is it possible to finetune a PLM trained on a general set of proteins to improve its performance when dealing with antibodies? Can PLMs be used to reason about CDR regions, and thus the binding modes of a sequence? And, does the perplexity of the antibody sequence correlate with how successfully the model captures structural information?

1.6.1 Contact maps for antibody design

Recently, it has been demonstrated that distograms, rather than 3D structures, can be used as a tool in protein design. Verkuil et al. [117] tested PLM's ability to generalise to unexplored regions of protein space using two experiments: fixed backbone design which aims to find a sequence which folds into a predetermined structure; and free design where the sequence and structure are both allowed to vary.

These experiments did not focus on antibody design. A PLM which can generate precise distograms or contact matrices for antibodies might, therefore, be invaluable in designing antibodies which bind to a known target using fixed-backbone design. It is therefore directly valuable to the design of therapeutics to determine which PLMs are most successful at this task.

Chapter 2

Evaluating PLMs as antibody structure predictors

Pretrained PLMs are evaluated based on their innate ability to represent the structural features of antibodies. These featurisers vary in their architecture, pretraining, and, in cases where MSAs are utilised, the method by which the input sequences are found. Three categories of PLMs are considered: PLMs pretrained on single antibody sequences (APLM); PLMs trained on single sequences of general proteins (GPLM); and a PLM trained using MSAs of general proteins (MSA-PLM).

In all cases, using the procedure outlined in Sec 1.5.1, contact predictions are extracted from the PLM's attention maps. Each model in the evaluation comprises a featurising PLM and a contact prediction head (CPH). These composite models are uniquely identified using the naming convention `<featuriser>` → `<contact-prediction-head>`. As described in Sec 1.5.1 analyses presented in this chapter limit the CPH to regularised logistic regression. By using a low-complexity model to map attention matrices to contacts, the precision of the prediction will depend on information already present in the PLM's internal representations.

Thresholds defining a contact between residues (see Equation 1.3) are not standardised. For the purpose of this investigation, two residues are considered to be in contact if they are separated by at least six residues in sequence space ($t_{seq} < 6$), and the Euclidean distance between their C_β (C_α for glycine) atoms is less than 8Å (Fig 1.1).

To understand how PLM performance translates from the context of general proteins to that of antibodies, a set of evaluation metrics, presented in order of difficulty and are increasingly more antibody-focussed, are used to compare the featurisers.

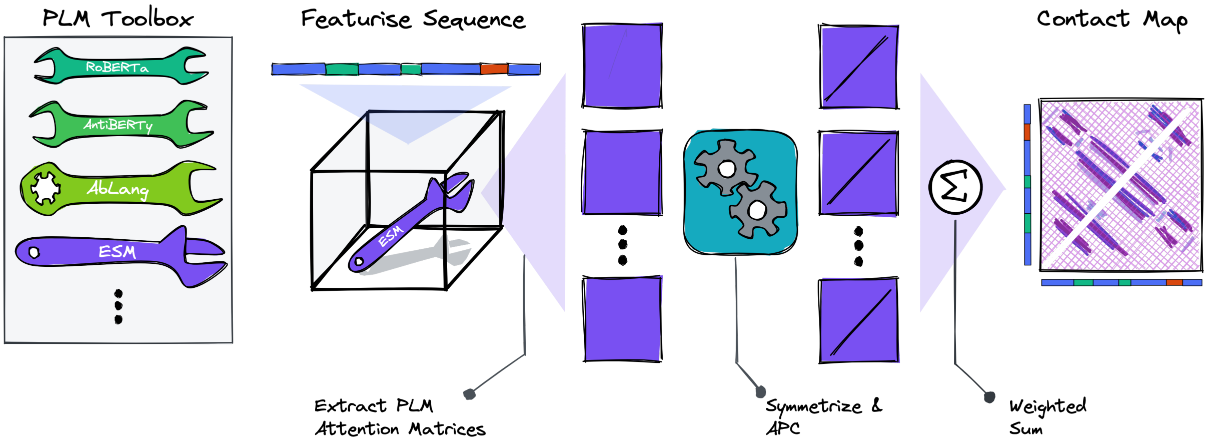

Figure 2.1: Contacts from attention. Contact prediction can be used to compare sequence-featurising tools known as Protein Language Models (PLMs) to determine which tool understands the structure of a given sequence best. To accomplish this, after featurising a sequence, the self-attention maps are extracted from the PLM. This stack of matrices is minimally processed. Each is symmetrised and has the average product correction (APC) value removed [30]. Contacts between the residues in the sequence are then predicted using a weighted sum of these maps.

2.1 Data

Measuring the accuracy of contact predictions requires solved antibody structures. SAbDab [29] is a data curation tool which searches through proteins in the Protein Data Bank (PDB) [9] and extracts antibodies. All of the files, therefore, follow the PDB's conventions. Ideally, each file contains the coordinates of all atoms in the antibody's heavy and light chains and, occasionally, a matching antigen as well. Conveniently, SAbDab provides sequences pre-numbered using ANARCI [28] according to the IMGT standard (see Sec 1.2.2 for an explainer of antibody numbering). This enables relatively simple alignment of antibodies of different lengths.

Many of the antibodies are crystallised without their antigen counterpart (these are known as apo structures). Because proteins are not always rigid, the apo structure can differ substantially from the holo-structure (the conformation of the bound antibody) [21]. Ideally, the dataset would only comprise holo-structures since these represent the true binding conformation of the protein.

All structures made available via the SAbDab online portal as of 08/11/2022 were downloaded. To ensure minimal redundancy in the dataset, a subset of these structures was extracted where each sequence has a maximum of 90 % sequence similarity (sequence identity) with other structures in the set. Furthermore, filtering out structures with a resolution lower than 3 Å was required to curate a high-quality evaluation dataset.

Most modern structure prediction tools can rival experimental methods when predicting framework regions and CDRs H1, H2, L1, L2, and L3. Since HCDR3 has no canonical conformations and is highly variable, accurate predictions in this region are not trivial. To

PLM	Training dataset			Model Details		
	Ab only	No. Seq	Source	Arch	Params	Layers×Heads
DeepAb [99]	✓	118 Th	OAS	LSTM	6 M	-
AntiBERTy [97]	✓	558 M	OAS	BERT	26 M	8×8
AbLang (h) [81]	✓	14 M	OAS	RoBERTa	86 M	12×12
AntiBERTa [68]	✓	58 M	OAS	RoBERTa	86 M	12×12
UniRep [4]	×	24 M	UniRef50	LSTM	18 M	-
ProtT5-XL [32]	×	45 M	UniRef50	T5	3 B	24×32
ProtBERT [32]	×	216 M	UniRef100	BERT	420 M	30×16
ProtAlbert [32]	×	216 M	UniRef100	ALBERT	224 M	12×64
ESM-1b [91]	×	27 M	UniRef50	BERT	650 M	33×20

Table 2.1: Comparison of a few self-supervised PLMs. Most of the transformer-based models are based on an encoder-only architecture known as BERT [26] (Bidirectional Encoder Representations from Transformers). RoBERTa [72] (a Robustly Optimised adaptation) tweaks the training procedure of BERT while AlBERT [66] (a lite BERT) uses parameter sharing and lower dimension representations to dramatically reduce the number parameters required to achieve SOTA performance. All of these methods are trained using masked language modelling (MLM). In contrast, autoregressive models such as LSTMs [40] and the T5 transformer have an encoder-decoder architecture. The number of proteins used for pretraining models can vary even when sourcing data from identical sources. This could be the result of sampling from the dataset to set aside an evaluation dataset, or simply a function of data being collected from the online repository at different dates.

use HCDR3 as a benchmark for contact prediction requires all PDBs to contain complete heavy chains. Files that exclude the heavy chain are therefore removed from the dataset. Since they are rare, chains with HCDR3s longer than 13 residues are also excluded. Additionally, any single chain variable fragment (SCFv) [1] entries are ignored since they require special handling and are exceptionally rare after the other filters have been applied.

Imperfections in the crystallisation of proteins make the precise location of some atoms impossible to determine. Due to these experimental complications, PDB files are often incomplete. Errors such as these have been shown to cause inaccuracies in structural modelling tasks [111]. In cases where a backbone atom or entire residue is missing, it is not possible to describe the chain's complete structure accurately. Since this information is critical to maintaining a dependable ground truth, every chain must be meticulously inspected in order to exclude erroneous files.

Unless otherwise stated, the remaining 1942 antibodies are used throughout the analysis.

[1] A single chain variable fragment (SCFv) is an engineered chain comprising the variable regions of the heavy and light chain fused together. Encapsulating the binding region in a single chain makes manufacturing the antibody more convenient.

2.2 Evaluation using standard metrics

The CASP competition served the invaluable function of setting benchmarks for protein structure prediction. One of the metrics used to evaluate the quality of predictions is the precision of the predicted contact map. This section interrogates a number of PLM-based contact predicters using these standard metrics.

2.2.1 Methodology

ESM-1b is a BERT-based [26] language model trained on 27 million representative sequences from UniProt [112], the vast majority of which are not antibodies. To examine how well a PLM trained on a general set of proteins (GPLM) was able to predict contacts between residues in an antibody, sequences from SAbDab were used to train the contact prediction network and evaluate its performance. Following the procedure in Sec 1.5.1, regularised logistic regression is used to predict contacts from processed attention matrices. This network is referred to as a contact prediction head (CPH). A random sample of antibodies is removed from the set of 1942 solved structures (see Sec 2.1) and is used to fit the logistic regression predictor. To distinguish it from the pretraining dataset, this dataset is referred to as a transfer learning training dataset (TLTD). The remaining antibodies are used as a test dataset.

The usefulness of predicted contacts depends on their precision [53]. A high false positive rate could place additional constraints on the protein's conformation that result in an egregious final fold. Furthermore, false negatives may not be of the utmost importance since it has been demonstrated that a relatively small fraction of contacts can reliably produce accurate topology-level structures [60]. It is therefore common [88, 101, 60, 31] to evaluate contact prediction on the precision of a reduced set of highly confident predictions. Precision on the L, $L/2$, $L/5$ most probably contacts is calculated where L represents the number of residues in a chain. Adopting the standard notation for these metrics, they are subsequently reported as $P@L,\ P@L/2,\ P@L/5$ respectively.

$$P@L = \frac{TP_{\text{top-L}}}{TP_{\text{top-L}} + FP_{\text{top-L}}} \tag{2.1}$$

It is sometimes desirable to distinguish between the contact precision between residues at constant separations in sequence space. Based on the separation between two residues $i,\ j$ contacts can be classified into three non-overlapping ranges which are most relevant to the tertiary structure of proteins [101]: short-range where $6 \leq |i - j| < 12$; medium range where $12 \leq |i - j| < 24$; and long-range where $24 \leq |i - j|$.

Choosing the size and character of the training set which is used to fit the logistic regression contact predictor offers a range of hyperparameters that can be tuned. To determine the effects of regularisation, and the size of the training dataset, some exploratory experiments were run using ESM-1b [91], a state-of-the-art single-sequence PLM at the time of writing, as a featuriser (see Sec 2.2.2).

Another important contributing factor is the composition of the pre-training dataset. Sec 2.2.3 uses the standard measures of contact precision to compare featurisers trained on different sets of proteins. Of particular interest is the contribution made by a pretraining dataset exclusively comprising antibodies.

2.2.2 Regularised regression on a set of antibodies improves performance

TLTD size has a minor effect Varying the size of the random sample used for training to CPH (for each size the experiment is repeated with five random samples) demonstrates that precision reaches a plateau above a sample size of 20 (Fig 2.2A). This echoed the findings of Rao et al. [89] when predicting contacts on a general set of proteins. Therefore, unless otherwise stated, subsequent experiments can be assumed to use 20 randomly selected Abs from the curated set of 1942.

Regularisation strength has a minor effect To test the effect of regularisation strength on precision and sparsity [2] , logistic regression models were fit to a random set of 20 antibodies (resampled five times for each value of λ) while varying these parameters. The effect of regularisation strength on precision is unpronounced with precision slowly dropping as the strength increased (Fig 2.2B). As expected, increasing the regularisation strength coerces a larger percentage of the learned weights to tend towards zero (Fig 2.2C). The self-attention heads which are picked out by the model as being most important to predicting contacts are concentrated in the final layers of the transformer. This seems to indicate that a sparse set of attention maps, collected from deeper layers of the network, is sufficient to precisely predict contacts. Having observed a similar result, Weissenow et al. [119] discarded most of the attention matrices from ProtT5 [32] and used only 50 attention matrices as features for downstream structure prediction.

Check for sampling bias If the training dataset is highly similar to the test dataset, results are bound to be biased. After filtering the combined test and train dataset to only comprise representative sequences with a maximum of 90 % sequence identity, high sequence similarity between test and train sets is not likely. However, it is important to ensure that the precision of the contacts predicted for a sequence is independent of the similarity between the sequences in the test and train sets. An alignment score is

[2] Sparsity in this context is used to refer to the ratio of zero to non-zero learned parameters.

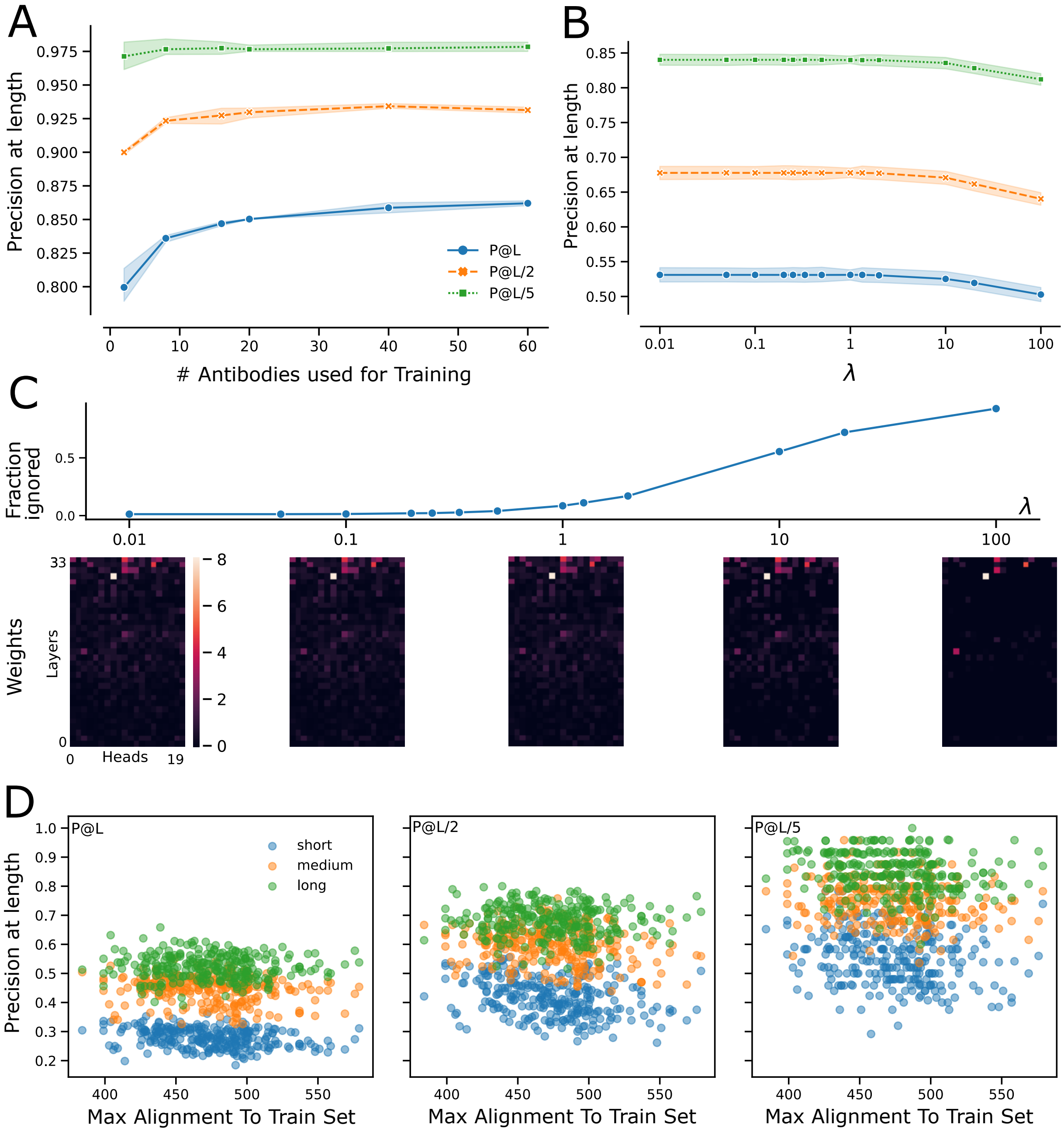

Figure 2.2: Training a logistic regression model to predict contacts from raw attention maps extracted from ESM1-b. **A** Effect of increasing the number of antibodies used to train a logistic regression model on the precision of its contact predictions. **B** Effect of changing the regularisation strength (L1-regularisation) on precision. Increased regularisation encourages sparsity. Eventually, as fewer attention maps are considered in an increasingly sparse network, the precision decreases. **C** The learned weights at varying regularisation values are visualised as a sequence of grids accompanied by the trend in the fraction of ignored attention maps (learned weight of 0). The sparsity of the important attention heads increases with regularisation with a greater focus on deeper layers in the network. **D** Testing the dependence of contact precision to sequence similarity among training sequences. The precision of predicted contacts on test sequences is plotted against the maximum alignment score to the training dataset for each sequence in the test dataset.

generated between each of the 1922 sequences in the test dataset and the 20 sequences in the train set. Sequences are globally aligned using `Biopython.pairwise2` [22] and the BLOSUM62 substitution matrix [46]. For each of the test sequences, the maximum alignment score is visualised against its contact prediction precision in Fig 2.2D. This is repeated for the top-L, top-L/2, and top-L/5 predicted contacts at short, medium, and long ranges. Of these nine combinations all are negatively correlated to the alignment score with Pearson correlation coefficients ranging from -0.32 to -0.02. There is, therefore, no positive correlation between the precision of the predictions and the similarity to the training dataset sequences.

A TLTD comprising only antibodies boosts precision ESM was not specifically developed for use in antibody design. Rather, its designers trained the network on a massive set of general proteins in the hopes that diversity would enable the model to generalise into unseen regions of protein space. In a single forward pass on an antibody heavy chain, GPLMs, such as ESM, might extract many features of which only a subset are useful to antibody structure prediction. It is possible to identify which attention maps are most relevant using L1-regularised logistic regression (LR). Rao et al. [89] released weights for their LR model which was trained to predict contacts from the features of ESM-1b by regressing against a general set of proteins. ESM-1b→LR* refers to this composite model where the $\star$ indicates that the LR head is "pre-trained" and was not tuned to the structure of antibodies. The precision of this model's predictions is a useful benchmark against which one can compare a CPH trained on antibody structures.

By regressing against the TLTD, comprising exclusively of antibodies, rather than a general set of protein structures a CPH is better able to pick out the most important attention maps. The resulting long-range contact predictions are 16% more precise (see the comparison of ESM-1b→LR and ESM-1b→LR* in Fig 2.3). Clearly, transfer learning on antibodies rather than general proteins is more suitable to pick out the transformer's most useful attention maps.

2.2.3 Pretraining on general proteins rather than antibodies

Language models are flexible tools which, after being trained on a large dataset, can be applied to myriad downstream tasks. Since, in this case, the downstream task is assumed to be structure prediction, it is important to identify a pretraining tactic which is likely to enable accurate results. Possibly the most important component of a pretraining strategy is the data. I, therefore, compared several transformer-based language models pretrained on different datasets.

Since antibodies are distinct from general proteins in many ways, it seems intuitive that a PLM pre-trained on only antibodies (APLM) would be better equipped to handle

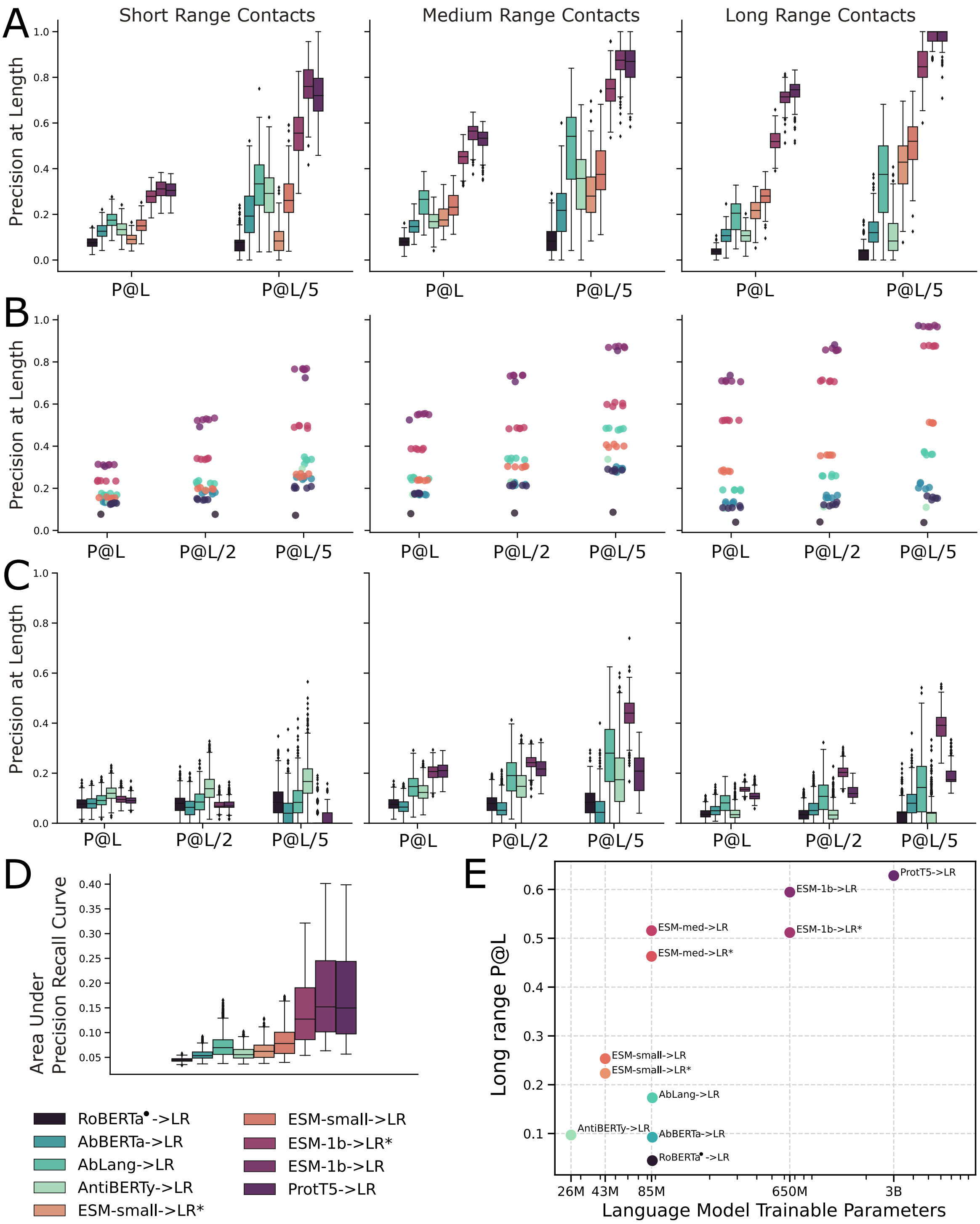

Figure 2.3: Comparison of different protein language models as contact predictors. Blue-green hues represent language models trained on antibodies, while orange-purple hues represent those trained on a generic set of proteins. The first, middle and last columns represent the precisions for short-range, medium-range and long-range contacts. **A** Precisions (P@L, P@l/5) where predictions were made by fitting a regularized logistic regression head. P@L/2 is not shown to rather focus on a wider variety of models. **B** Same as B, but repeated using several random test train splits. In this case, P@L/2 is included. **C** To detangle the effects of the logistic regression model, the top-performing models' attention matrices were simply summed across the layers × heads dimension to predict contacts. **D** To avoid only observing the precision on the models' most confident predictions, the area under the precision-recall curve is also reported. **E** Scaling laws for protein language models (ESM was retrained at several scales).

domain-specific quirks. There is evidence that, for certain downstream tasks, this might be the case. Sequence-embeddings from APLMs seem to cluster more strongly around the antibody source (human, murine, humanized, chimeric) than those generated by GPLMs [68]. Using a similar analysis, Olsen et al. [81] found that both AbLang and ESM1-b are able to identify the heavy chain's V-gene family but that clustering according to B-cell maturity is stronger for the APLM. Furthermore, they found that AbLang was better equipped to accurately restore sequences with missing residues.

Although useful, these abilities do not imply that APLMs extract more structurally relevant features than their general counterparts. To directly compare their ability to reason about immunoglobulin structure, the models' attention matrices are used to predict residue-residue contacts (as per the method described in Sec 1.5.1). A low-complexity model (L1-regularised logistic regression) is intentionally used to predict contacts since the goal is to ascertain the amount of structural information that is captured after self-supervised training. Using a more advanced pattern-matching algorithm as a CPH would make the benefit of the PLM difficult to isolate. For each featuriser, a logistic regression (LR) head is trained to predict contacts from its attention matrices using the same set of 20 solved antibodies. The remaining 1922 antibodies are used as an evaluation dataset to quantify the precision at short, medium and long ranges. The composite models compared in Fig 2.3A are chosen to disentangle the effects of several potentially confounding factors.

Check for sampling bias To check for sampling bias in the TLTD, five randomly sampled sets of 20 antibodies are used to retrain the CPHs. Once again, the remaining set of 1922 proteins is used for evaluation. Results from the five rounds of re-sampling and evaluation are presented in Fig 2.3B. The final precision score is stable despite the different training datasets.

Setting a Benchmark Randomly initialised networks can generate useful representations for downstream tasks [51]. To ensure that pre-training the transformer does impart a boost in precision an untrained language model, RoBERTa$^\bullet$ (where the $\bullet$ indicates random initialisation), is used as a baseline.

GPLM > APLM When using an LR head trained on Abs, GPLM outperforms APLM. ESM-1b and ProtT5 (state-of-the-art GPLMs) surpass AbLang and AntiBERTy (state-of-the-art APLMs) as well as AbBERTa (an APLM trained by the Machine Learning Research team at Bayer). Using these metrics, GPLMs consistently deliver, at minimum, a 2x performance boost.

Against the benchmark of ESM-1b$\rightarrow$LR* APLMs underperform. This is surprising since the structure of immunoglobulin is highly conserved and an LR model tuned to predict those common contacts is at a clear advantage. This may imply that the features extracted by GPLM are richer with structural information. Indeed, when using these features combined with a contact prediction head trained on Abs, the precision significantly improves over the benchmark. Since, for a given sequence, the attention maps being used by ESM-1b→LR* and ESM-1b→LR are identical, one can assume that the LR model simply identifies which of these attention maps are most relevant to the structure of antibodies rather than proteins in general.

Considering capacity Language model performance follows a set of empirically determined scaling laws [58]. Performance scales with training data, model capacity (see Fig 2.3E), and train compute. ProtT5→LR is the largest model under evaluation. ESM-1b→LR performs similarly but is substantially larger than the APLMs. Rives et al. [91] released several versions of ESM by varying the number of trainable parameters. Therefore, a smaller version of ESM-1b, containing around 43M parameters is also evaluated. By approximately matching the model sizes, APLMs and GPLMs can be more fairly compared. ESM-1b-small→LR* appears to have difficulty in predicting short-range contacts, barely surpassing the performance of the randomly initialised benchmark. However, at longer ranges, its performance exceeds the most proficient APLM. Another version of ESM containing 85 M trainable parameters, ESM-1b-med→LR*, substantially outperforms almost identically sized APLMs (see Fig 2.3E) even without its CPH being tuned to the structure of antibodies. As expected, retraining the contact prediction heads of ESM-1b-small→LR and ESM-1b-med→LR on Abs boosts their performance significantly. Once tuned to the structure of antibodies both of these models exceed the capabilities of the APLMs. These trends indicate that given more data and endowed with more trainable parameters, APLMs would improve but still would not match the precision of equivalently sized GPLMs.

Large GPLMs, with more attention maps, also provide their LR head with extra degrees of freedom providing an unfair advantage over the smaller APLMs. To remove this effect, the absolute values of every element in the processed attention maps are computed abs : $F_n \to F_n^{|\cdot|}$ and they are summed along the $N = L \times H$ dimension: $\sum_{n=1}^{N} {F_n}^{|\cdot|}$. Note that there is no need to normalise this aggregate value since the top-L values are interpreted as contacts. In the absence of a contact prediction head with learnable parameters, the pairwise features extracted by the GPLM are not obviously more expressive than those extracted by APLMs. However, for longer-range contacts, it appears that the magnitudes of attention weights in ESM-1b are highly correlated to contacts (see Fig 2.3).

2.2.4 Perplexity is not a predictor of precision

Perplexity (PPL) is a metric commonly used to evaluate how well a language model "understands" an input sequence. This score is based on the model's ability to correctly predict masked tokens using the surrounding context. More formally, given an ordered

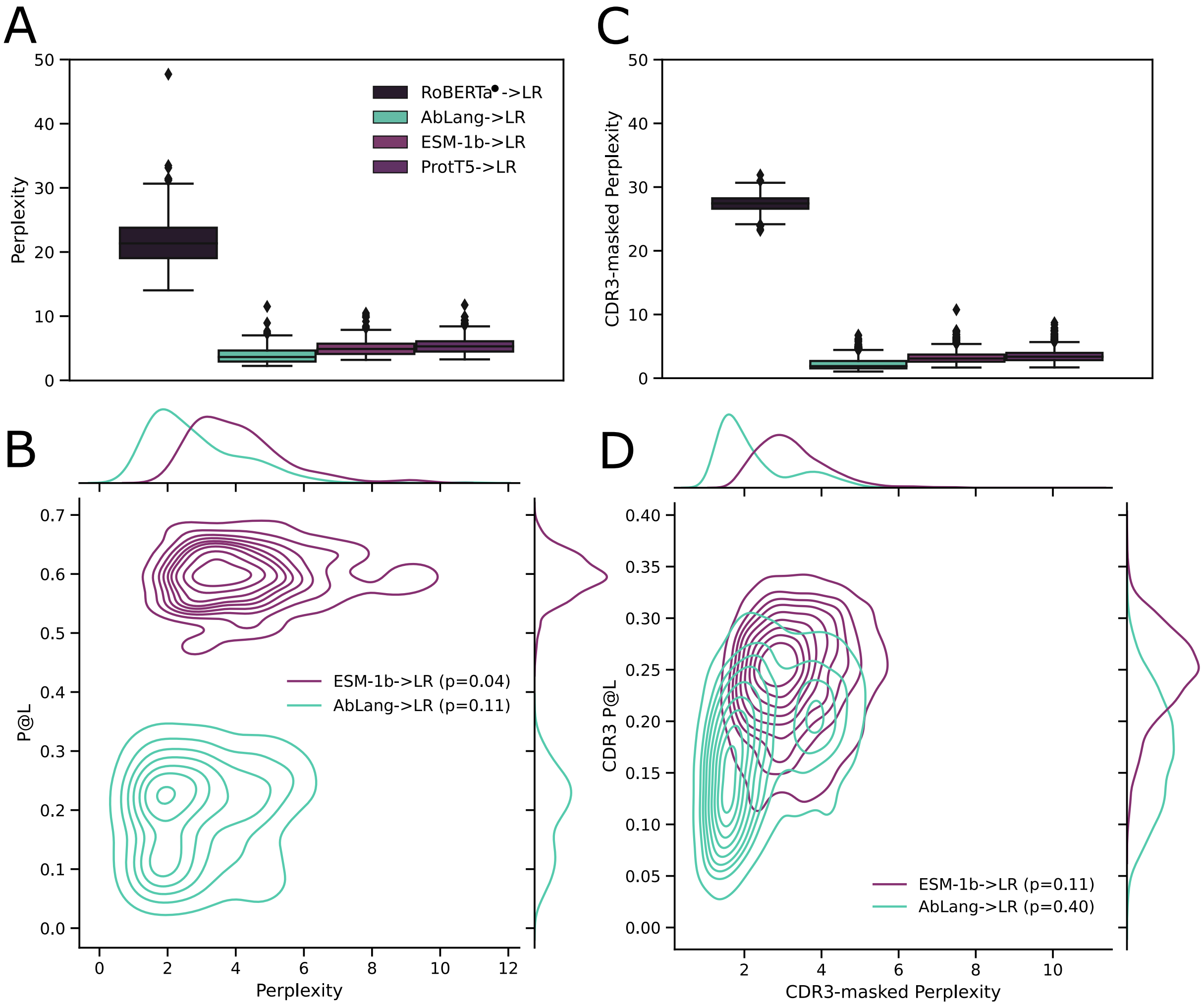

Figure 2.4: Evaluating sequence perplexity over the test dataset. A Perplexity (exponential of the negative, average log-likelihood) for several protein language models calculated across the entire test set. Lower is better. **B** Relationship between perplexity and precision for a PLM trained on a general set of proteins, and one trained on antibodies only. Note that the logistic regression head is not relevant to the calculation of perplexity. **C** Perplexity of the HCDR3 sequences from the test set. **D** Relationship between the perplexity of the HCDR3 sequences and the precision on those regions of the antibody. The APLM shows some dubious correlation (see text for details) while the GPLM shows no appreciable correlation between the two measures.

set of amino acids $\mathbf{x} = \{x_1, x_2, ..., x_L\}$ where position i is masked in the input sequence, the LM will use the sequence to generate a probability distribution for the missing token x_i over its vocabulary, $\mathcal{V}$. The likelihood of predicting the correct token is, therefore, $p_\theta(x_i|\mathbf{x} \setminus \{x_i\})$. By sequentially masking each token in the set, one can calculate the exponential of the negative, average log-likelihood of the tokens:

$$PPL(\mathbf{x}) = \exp\left(-\frac{1}{L}\sum_{i=1}^{L}\log p_\theta(x_i|\mathbf{x} \setminus \{x_i\})\right) \quad \in (0, |\mathcal{V}|) \tag{2.2}$$

A PLM which makes predictions at random will receive a perplexity score of $|\mathcal{V}|$ while a perfect model would score a 1. The score can be interpreted as the average number of amino acids that the model is uncertain between. A model that is able to reliably predict the identity of a masked token must have the ability to extract relevant information from the surrounding context. Arguably, this ability can be construed as understanding.

A well-calibrated confidence measure would be invaluable when deciding how trustworthy a model's predictions are. One would expect that low perplexity sequences would correspond to more precise predictions allowing the perplexity score to act as a confidence measure. This was in fact demonstrated for ESM by Rives et al. [91, 71] where low perplexity sequences reliably generate better quality structures.

Unsurprisingly, pretraining on antibodies allows APLMs to better predict missing residues, demonstrated by lower perplexity scores, than their generalist counterparts (see Fig 2.4A). However, in the case of antibodies, there seems to be no significant correlation between a sequence's perplexity and the precision of the predictions (Fig 2.4B).

To interrogate this relationship further, one can isolate particular regions of the antibody. Strangely, for AbLang (an APLM), if the perplexity is only calculated by masking residues in HCDR3 (see Fig 2.4C) there seems to be a positive correlation (Pearson correlation coefficient of ≈ 0.4) between the masked-perplexity score and precision in the same region (see Fig 2.4D). This correlation does not exist for GPLM such as ESM-1b. On face value, this result is discouraging: AbLang should generate higher precisions at lower perplexities ideally resulting in a negative correlation. However, the reliability of this correlation does not hold up under scrutiny. The Pearson correlation coefficients between the perplexity of HCDR3 sequences, and the model's precision across all framework regions and HCDR1 are 0.57 and 0.41 respectively. The model's ability to predict missing residues in HCDR3 should not be correlated to its ability to predict contacts between framework residues. This indicates that the observed correlation in Fig 2.4D is likely to be spurious. In either case, perplexity should not be used as a proxy for the model's confidence.

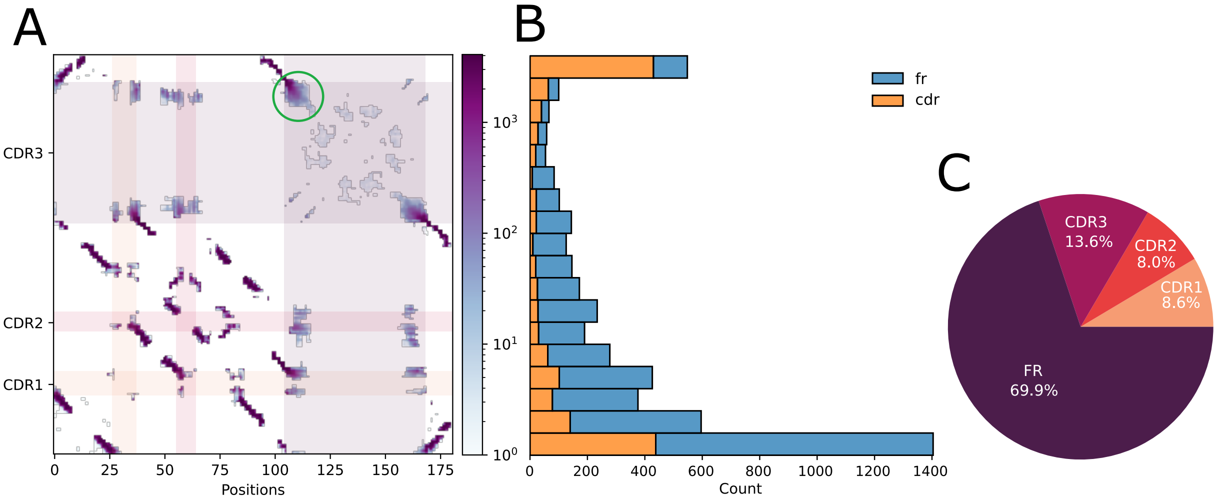

Figure 2.5: Cumulative contacts across aligned antibodies. **A** Cumulative contact matrix found by summing aligned (according to IMGT positioning) contact maps from 4092 antibodies. Conserved contacts rarely lie in CDRs. The canonical "stem" of the HCDR3 loop is circled in green. **B** A histogram over the values in A (only taking into account positions in the matrix which are non-zero, of which there are 5104 out of 32,761 possible contact positions in the aligned matric) reveals a bimodal distribution. The peaks are generated by contacts which are either extremely rare ($< 2/4092$), or conserved ($> 3500/4092$). **C** Relative contribution of contacts in the cumulative matrix conditioned on the region.

2.3 Evaluation using antibody-specific metrics

Evaluation using standard metrics seems to indicate that sufficiently large GPLMs capture more salient structural features than their antibody-specific counterparts. However, using metrics which are designed to evaluate structure prediction on general proteins may introduce a bias. To truly separate the two classes of models, more antibody-specific metrics are required.

Predicting the conformation of an antibody, on average, is straightforward since the majority of the structure is conserved. Contacts which are most commonly conserved, and therefore most trivial to accurately predict, usually exclude residues within the CDRs (Fig 2.5A). Furthermore, these regions of the antibody do not directly determine binding. In the context of antibodies, a structure prediction technique is truly valuable if it can facilitate the study of the molecule's binding dynamics. Adjusting the evaluation criteria based on these unique criteria of antibody structure prediction can be accomplished by focusing on contacts involving residues within CDRs.

2.3.1 Focusing on CDRs

Only considering the model's most confident predictions could imply that only contacts in the framework regions are being considered. Since these contacts are conserved it is trivial for a model with sufficient capacity to store this information in its parameters.

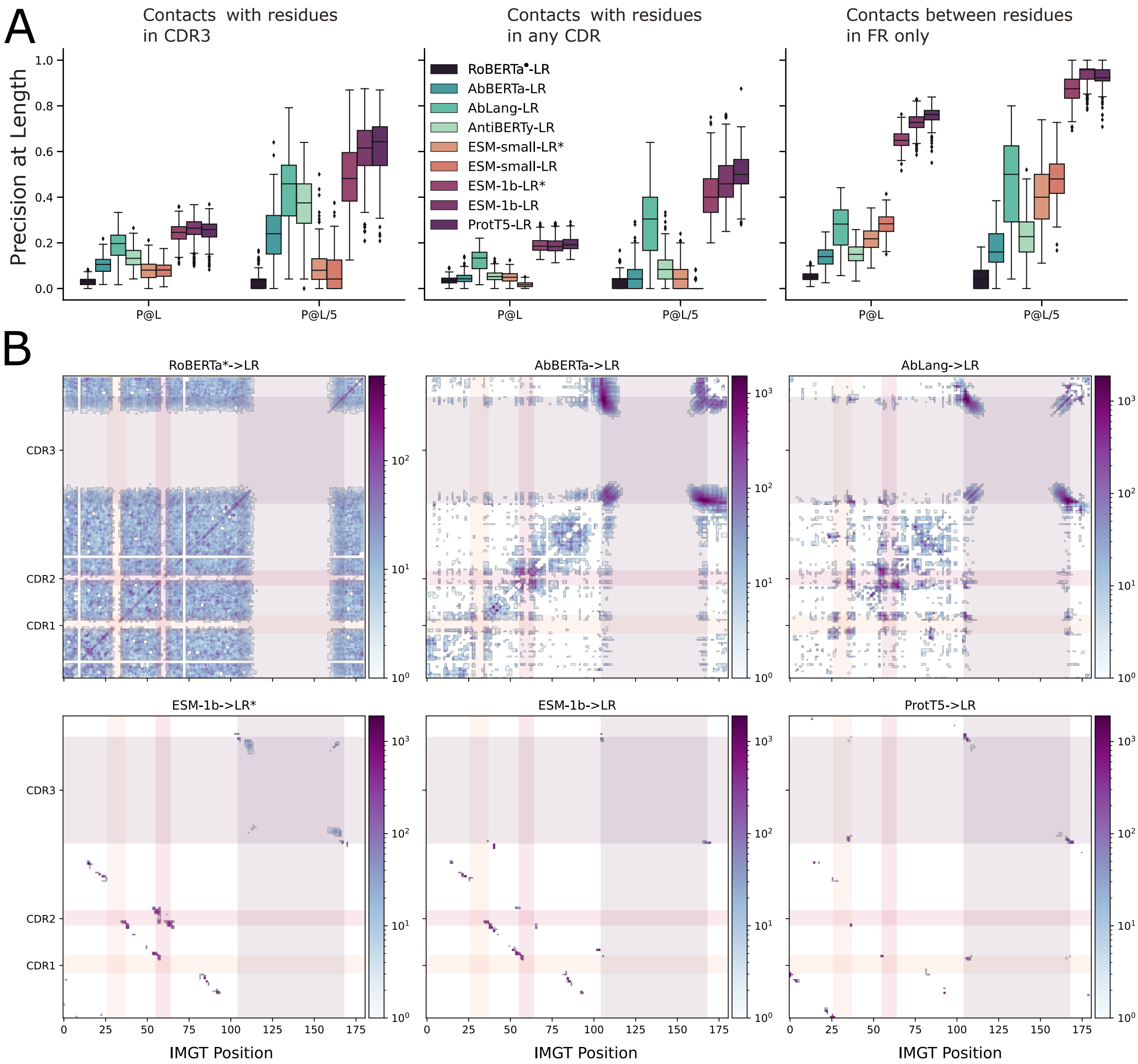

Figure 2.6: Evaluating precision within CDRs. A Precisions of contacts where at least one residue is in HCDR3(left) or any HCDR (middle) are compared to the precisions of predicted contacts where neither residue is in a CDR (right). Blue-green hues represent language models trained on antibodies, while orange-purple hues represent those trained on a generic set of proteins. Precisions (P@L, P@l/5) are reported where predictions were made by fitting a regularized logistic regression head to the attention maps of the PLMs. **B** The aligned, top-L most confident predictions are summed across the test dataset to generate a frequency map. A randomly initialised language model coupled with a logistic regression head ($\text{RoBERTa}^{\bullet}$ →LR) demonstrates no pattern in its predictions. On the other end of the spectrum, large GPLMs make confident predictions which are tightly packed within FRs. Frequencies are coloured using a logarithmic scale to make less common contacts more visible and emphasise highly consistent predictions.

Fig 2.6B demonstrates a tendency for the GPLM to make confident predictions in framework regions while the predictions of the APLM are much more varied. This is interesting for two reasons. Firstly, by comparing these frequency maps to the one in Fig 2.5A, it seems that the APLMs, despite only ever being exposed to antibodies, do not capture the canonical contacts within the framework regions. In comparison, GPLMs focus on these positions without having to train their accompanying contact prediction head on antibodies. Because the FR contacts are more common than those involving CDRs, it is likely that this explains the discrepancy in precision between APLM and GPLMs seen in Sec 2.2 and Fig 2.2A.

Secondly, these frequency maps indicate that the evaluation metrics being used are insufficient in the context of antibody structure prediction. If only examining the most confident predictions, a model which predicts the structure of framework regions perfectly but is uncertain about contacts in the more interesting variable regions will score higher.

To account for this, each region in the chain (defined by IMGT) is evaluated independently. In practice, this entails predicting contacts across the entire chain and then separating contacts into non-overlapping sets. For example, if at least one of the residues in a pair lies in HCDR1, that pair will be included in the analysis of HCDR1. However, to be included in the analysis of the framework region, both residues must reside in one of the four framework segments of the chain.

As suspected, the large GPLMs have close to perfect precision in the framework regions. More interestingly, despite not being specialised for antibody structure prediction the larger GPLMs outperform the APLMs in CDRs - including CDR3 - as well (Fig 2.6A). Even ESM-1b→LR* manages to achieve a higher precision than the best-in-class AbLang→LR. This seems to indicate that pretraining on antibodies also does not endow a PLM with the ability to make accurate predictions in hypervariable regions. To that end, it seems that exposing a language model to variable protein conformations during training is more advantageous.

2.3.2 Predicting rare contacts

Evaluating contacts within CDRs is more revealing than considering the entire chain at once. However, the results of this experiment remain susceptible to bias introduced by highly conserved positions. One should expect the contacts comprising residues within HCDR3 to score lower than other CDRs. However, this is not the case (see Fig 2.6A). One possible reason for this unexpected result is that the contacts in HCDR3 are more common than other complementarity-determining regions in this dataset (see Fig 2.5C). This would encourage the contact prediction head to focus on HCDR3 during training, placing a higher priority on those positions to achieve an optimal fit to the dataset.

Another possible explanation is the presence of a few conserved contacts in HCDR3.

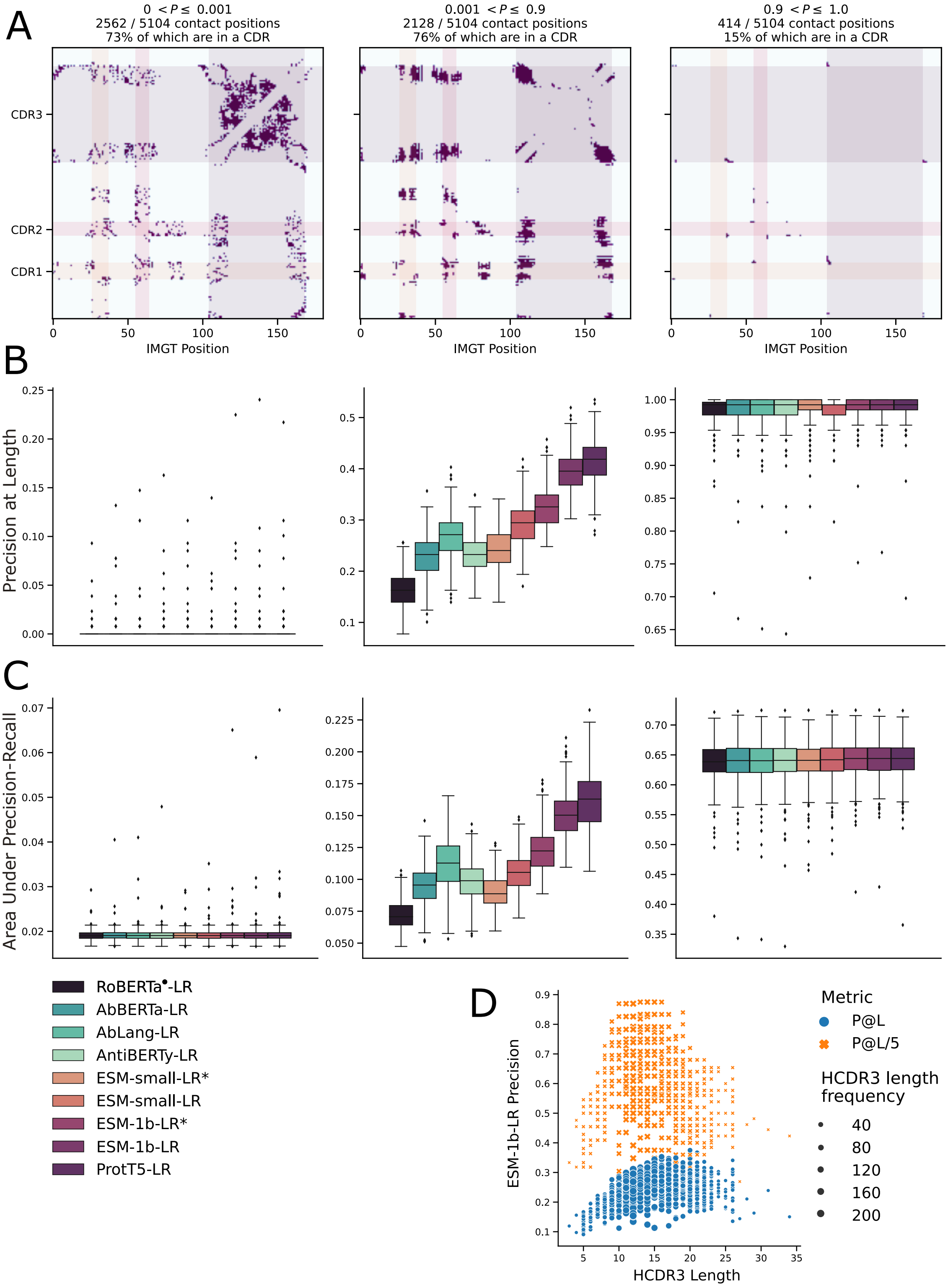

Figure 2.7: Evaluating Language models on rare contacts. **A** Observed contacts between IMGT positions split by frequency of occurrence in SAbDab. **B** P@L of contact predictions masked by the frequency maps in A. The leftmost panel represents the precision on the rarest contacts, whereas the rightmost represents precision on common contact positions. **C** Area under the precision-recall curve after predictions are masked by the frequency maps in A. **D** Relationship between prediction precisions and HCDR3 length where the frequency of each length is calculated from a filtered version of SAbDab.

Metric	Best APLM		Best GPLM		Difference
	Model	Score	Model	Score	
P@L, long range	AbLang	0.4	ProtT5	1.0	0.6
P@L, CDR	AbLang	0.35	ProtT5	0.45	0.1
P@L, common	AbLang	0.28	ProtT5	0.42	0.14

Table 2.2: Summary of evaluation metrics

Immunoglobulin domains fold into a canonical "beta-sandwich" comprising antiparallel pairs of beta strands that form two adjacent sheets. Since the antiparallel strands on either side of the sandwich are close together, they result in canonical contact patterns perpendicular to the diagonal of the contact matrix (note the antiparallel patterns of contacts in Fig 2.5C). This sandwich forms the framework onto which the CDR loops attach. Some numbering schemas, such as IMGT, include a small portion of this framework in HCDR3. These "stems" act as anchor points, connecting the haphazard looping structure to the conserved beta strands. Because they serve this essential function, the stems' representative antiparallel lines are found consistently in most antibody contact maps between the start and end of the HCDR3 loop (see the green circle in Fig 2.5). These stems could skew the analysis described in the previous section since a small number of highly conserved contacts is likely to boost the P@L observed for HCDR3. Because it has been established that GPLMs are proficient in predicting conserved contacts, the stems could explain their surprisingly high precision observed in Fig 2.7A.

Standard numbering schemes are constructed for convenient analysis and comparison. Like all simplifying models, they are imperfect. To remove any bias introduced by the overlap of numbering schemes with conserved regions, one can directly evaluate the model's performance on rare contacts. The frequencies with which particular contacts appear to follow a bimodal distribution: peaks are formed around exceptionally rare, and consistently conserved contacts respectively (Fig 2.5B). By using the frequencies with which contacts appear in SAbDab, and the shape of the estimated distribution as a guide, one can create masks that partition the possible contact positions in an aligned contact map into rare ($p < 0.001$), common ($0.001 < p < 0.9$), and highly conserved ($p > 0.9$) contacts (Fig2.7A). The rare contacts are found in CDRs 73 %. This set often comprises contacts between rare insertions in the HCDR3 loop. There are also many interactions between residues close to the borders of CDRs which may be an artefact of the artificial boundaries enforced by the chosen numbering scheme. On the opposite side of the spectrum, highly conserved contacts are only found within CDRs 15 % of the time. Interestingly, this subset of contacts is dominated by sections of residues in HCDR1 and HCDR3 which are consistently close together in 3D space. In between these two extremes are "common" contacts most of which (76 %) lie in CDRs. These include the stems of HCDR3.

After inference and alignment, these masks are independently applied to contact map predictions and ground truth. The P@L and average precision (area under the precision-recall curve) of the masked predictions are reported in Fig 2.7B and Fig 2.7C respectively.

On rare and conserved contacts there is no significant performance difference between the featurisers. In both of these cases, pretraining the language model seems to offer no benefit (note the comparable performance of RoBERTa$^\bullet$). This implies that rare contacts, notably contacts between HCDR3 insertion positions, are not captured by the model's attention matrices. In fact, 2.7D demonstrates that ESM, one of the most successful PLMs according to all metrics presented, have poor precision for HCDR3 loops with uncommon lengths. This hints at an inability to generalise to uncommon conformations.

Prediction precision on contacts which are not rare or highly conserved is more varied between featurisers. Once again, the performance of GPLM exceeds that of the APLMs despite most of these positions falling into CDRs. This result is comparable (see Table 2.2) to evaluating contacts within CDRs, where the drop in performance can be explained by the exclusion of highly conserved contacts, and the drop in variance is likely caused by the removal of rare contacts.

Chapter 3

Contact Prediction Heads

Using an LR-based CPH to predict contacts is useful to identify and extract structural information which has already been generated by the PLM. However, it is feasible that more complex models could take advantage of more subtle patterns in the attention maps to more precisely predict contacts. A precise contact predictor could prove useful in a number of downstream tasks including 3D structure prediction [119] and antibody design [117].

3.1 Beyond logistic regression

By stacking attention maps one can extract $L \times L$ fixed-length representations - one for each pair of positions. Even when considering these fixed-length representations in isolation from one another, there are a number of models which can be used to predict whether or not the residues in this pair are in contact.

One such model is a binary decision tree classifier. Each node in a decision tree splits the data into two parts using a threshold on one of the feature dimensions. By tuning a set of sequential thresholds (decisions) and iteratively splitting up the feature space into regions, a decision tree attempts to perfectly non-linearly separable classes. Note that LR can only determine a linear decision boundary between classes. Given a new datapoint, the tree's learned thresholds are applied to determine the class label of the region in which the datapoint lies. A Random Forest (RF) [11] is an ensemble learning method whereby several decision trees are fit to random samples of the training data. At inference time, each tree independently classifies new data points. A consensus is usually formed by selecting the most commonly predicted class.

Using the same set of 20 antibodies to train the LR models in Sec 2.2.3, RF classifiers were trained for the top-performing featurisers. The RF models consistently outperform LR models in predicting contacts in all regions using GPLM attention maps (see Fig 3.1F). For AbLang, the RF was outperformed by the baseline LR-based CPH. This is surprising because an ensemble of decision trees is a more powerful discriminator than LR. However, overly complex trees are prone to overfitting. With hyperparameter tuning, it is likely that the RF will outperform the simpler model for any PLM.

3.1.1 Moving towards an optimal architecture

To further explore a range of CPH architectures, ProtT5 was chosen as a featuriser. Although this was the largest model evaluated in Sec 2, it achieved the highest precision on rare contacts as well as contacts involving residues in HCDR3. LR is used as a baseline model to compare different estimator's abilities to predict rare (Fig 3.1A) and regional (Fig 3.1B) contacts. In all cases, the hyperparameters of each model are tuned to achieve maximal precision. To achieve this, once again 20 antibodies (identical to those used in Sec 2.2.3) were randomly sampled from the cleaned selection of SAbDab. Because proximal residues are rare in comparison with residue pairs which are far apart, the dataset was rebalanced to avoid introducing a bias towards the "no-contact" class. Class balancing was achieved by selecting a random set of non-contacting pairs of equal cardinality with the set of pairs in contact. After rebalancing, the dataset comprised just under 70,000 data points each of which is in $\mathbb{R}^{768}$. This training dataset, mapping fixed-length vectors extracted from attention maps to a binary output, was kept constant for all CPHs. To compare the performance of the CPHs a test set comprising 100 antibodies was set aside from the SAbDab dataset, taking care to exclude any of those used during training.

To maximise performance, for each model architecture, a combination of random search[1] and cross-validation (5 splits) is used to tune a selection of hyperparameters. In the interest of time, hyperparameter tuning was not exhaustive. Some configurable parameters were set. For example, at each node, the RF is restricted to considering the square root of the number of available features and splits are evaluated using Gini impurity [12]. However, the number of trees and the maximum allowable depth of those trees were tuned. The most successful RF used 152 estimators with a maximum depth of 15.

Alongside an RF, a variant tree-based ensemble method implemented using AdaBoost [45] is tested. As well as fitting several trees on random subsets of the training data, this training protocol "boosts" the model's ability to correctly classify data points that are initially misclassified. It does this by iteratively training new classifiers on samples from the same dataset while adjusting the weights of incorrectly classified items to encourage the next iteration to learn from the previous estimators' mistakes. For this architecture, the number of trees needs to be balanced by the importance of each tree (referred to as

[1] Random search is one technique for searching through the space of possible hyperparameter combinations. In each iteration, each parameter is randomly selected from a user-specified distribution over a set range of permissible values.

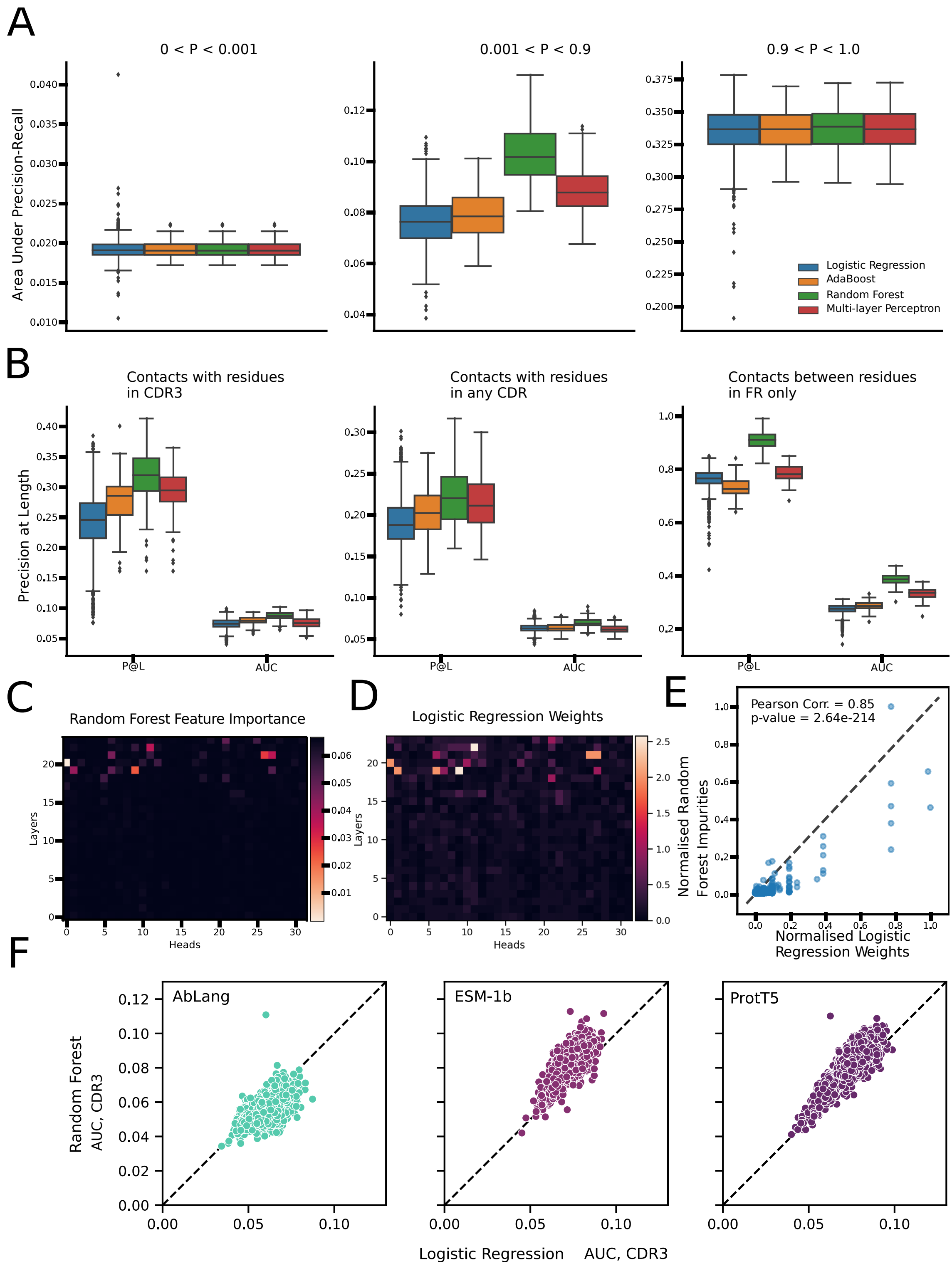

Figure 3.1: Comparison of different contact prediction heads. **A** Comparing different CPH architectures using the average precision achieved on rare contacts. **B** Comparing different CPH architectures using the precision achieved in different regions of the antibody. **C** Normalised Gini Importance from a Random Forest contact classifier trained on ProtT5 attention maps. Gini Importance is a measure which indicates the amount by which a feature reduces the impurity of a split at a node in a decision tree. The features which are most effective at splitting the data occur at deeper layers in the network. **D** Weights of an LR contact classifier where the most important features occur at deeper layers in the network. **E** Correlation between different measures of feature importance from C and D. **F** Comparing the precision of logistic regression and random forest classifiers in predicting HCDR3 contacts across the full 1942 test antibodies.

the “learning rate” in Scikit Learn). A random search discovered 87 estimators with a learning rate of 1.0 to be optimal in the range considered.

Additionally, a multi-layer perceptron (MLP) is assessed. Inspired by a simplified model of a neuron, each node in an MLP computes a weighted sum of its inputs and passes the result through a non-linear function (in this case a Rectified Linear Unit, or ReLU [38]). The network parameters are optimised to minimise the binary cross entropy loss using the Adam optimser [61], a variation of stochastic gradient descent, with a batch size of 128. The model tested had three hidden layers with 64, 32, and 16 nodes respectively. This design decision was arbitrary, following a common heuristic of halving the number of nodes with each layer. With more time, the depth and dimensionality of the network could have been tuned for better performance. Hyperparameter tuning identified a learning of 0.01 and a regularisation strength (referred to as “alpha” by Scikit learn) of 0.001.

Evaluating these architectures over the same test set and comparing them to the LR-based CPH as a baseline indicates that an RF is the most successful at predicting both rare (see Fig 3.1A) contacts and contacts in CDR contacts (see Fig 3.1B).

3.1.2 Important features are shared between architectures

Gini impurity is one way to measure the quality of a split at a single node in a decision tree. At every node, the splitting function utilises a feature and corresponding threshold which minimised the impurity of the two child sets. This can be loosely interpreted as a measure of the misclassification rate in each of the splits. Features which can be used to create cleaner splits in the data contain more information about the class of the data. One can therefore rank features using their Gini importance: the normalised reduction in the Gini impurity brought about by that feature.

Interestingly, there is a high correlation between an RF’s feature importances and the weights of an LR model (see Fig 3.1E). As was shown for ESM in Fig 2.2 LR models place high weights on the deeper layers of the transformer. Examining the most important features identified by an RF reveals the same pattern (see Fig3.1C).

3.2 Strategies to further improve CPH performance

Based on the investigation presented in the previous section, RF-based CPHs demonstrate a strong ability to predict inter-protein contacts from attention maps. Even more complex architectures are likely to further improve performance. For example, other successful structure predictors share information between positions in the attention maps using standard [119] or graph-based [98] convolutions. However, the intention is to achieve maximally precise contact predictions using only the structural features extracted via

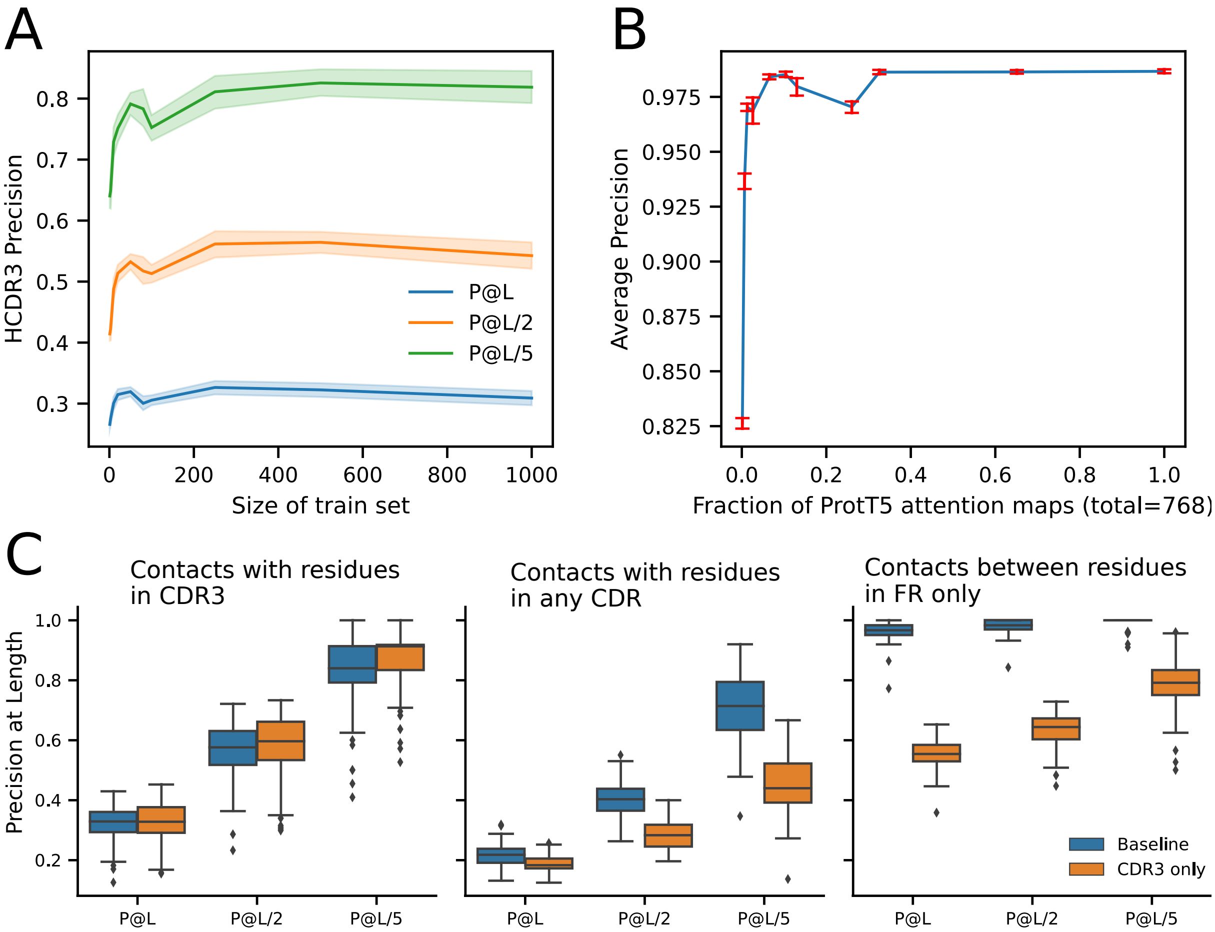

Figure 3.2: Identifying factors that affect the precision of an RF-based contact predictor. **A** Effect of the number of antibodies used during training on the precision of contacts involving residues in HCDR3. **B** Effect of changing the number of attention maps extracted from each training datapoint on the average precision. The attention maps were chosen based on a secondary logistic regression model (see Sec 1.5.1) which uses L1 regularisation to identify a sparse set of the most important attention maps. In each run, the N most important attention maps are chosen according to the learned weights of the logistic regression model where $N \in \mathbb{N}^{[1,768]} = \{x \in \mathbb{N} | 1 \leq x \leq 768\}$. **C** Comparing models with identical architectures, trained on the same set of antibodies. However, the orange model is only trained on data points which involve HCDR3, whether the residue pair is in contact or not. Performance on HCRD3 increases slightly while precision in framework regions drops dramatically.

self-supervised learning. A model which has access to features at multiple positions in the chain might learn to extract structural information itself thereby diluting this analysis.

Furthermore, a highly complex model would overfit the relatively small dataset available. It is therefore interesting to see if better results can be achieved using only traditional machine learning. This restricts the structure predictor to relying on the contextual information encoded in a single slice of the stack of attention matrices rather than allowing it to continue building new relationships between positions in a protein Therefore, this section explores a few approaches to increase precision using an RF-based CPH.

Increasing the size of the training dataset Up until this point, all CPH models have been trained on the same selection of 20 antibodies. Reserving the 100 antibodies from Sec 3.1.1 for evaluation, the size of the training dataset was varied from 1 to 1000. Each dataset was rebalanced such that the number of residue pairs in contact matched the number of non-proximal residue pairs. Furthermore, in every iteration, the hyperparameters of the model are tuned using cross-validation and a random search to find the optimal number of trees, and maximum tree depth associated with each training dataset.

Using more than 200 antibodies does not seem to provide any benefit to predicting contacts, especially those associated with HCDR3 (see Fig 3.2).

Fitting CPH on a subset of attention matrices ProtT5 is a large model, with 768 attention maps generated for every input sequence. As illustrated in Fig 3.1 C and D, only a small set of those attention maps are significantly more important than the others. It should then be possible to train the contact predictor on a subset of these highly informative features rather than the entire stack of attention maps. This reduced feature set may reduce overfitting by removing noisy features in the data. Another advantage is a decrease in computational complexity. Although this speedup is insignificant for a single protein, it could make a substantial difference if predicting the contacts of a large set of sequences.

To test the effect of this feature selection tactic a set of models were trained on 100 randomly sampled training antibodies and evaluated on the same test dataset as above. In each training run, a set of features is selected (ranging from 1 to 768 in number), a new model is trained, and its hyperparameters are tuned.

As illustrated in Fig 3.2C, with anything more than 50 features it is possible to match the performance of a model trained with access to all attention matrices. A similar observation was made by Weissenow et al. [119] which led them to use only 50 attention maps for their CNN-based structure predictor.

Region specific training One can think about the role of a contact predictor as a mechanism to select the attention maps which contain structural information and combine them. With this framing, one can imagine a contact predictor trained to extract the attention maps which are most relevant to structural information in a particular region of the antibody. By simply masking out all regions other than HCDR3, it is possible to train a contact predictor that outperforms an unmasked estimator in that region (see Fig 3.2). However, this new model is unsurprisingly poor at predicting contacts in framework regions.

Since the regions of an antibody can be isolated using a tool like ANARCI [28] it is entirely feasible to use an ensemble of region-specific models to maximise precision. For example, one could predict HCDR3 contacts with one specialised model, and use a more general model to make predictions for the remainder of the chain. Ultimately a highly specialised HCDR3 contact predictor is immensely more valuable than a model that can generalise across the chain.

Chapter 4

Discussion and Conclusion

The application of deep learning in protein structure prediction has revolutionized the field of protein structure prediction, providing access to a vast number of previously unknown protein structures. However, a consistently accurate model for folding a sequence into its lowest energy conformation without external information has not yet been developed. Although AlphaFold2 has shown some promise in approximating a biophysical energy function [94], it still requires a multiple sequence alignment to locate a starting point on the energy landscape. Moreover, some argue that AlphaFold2 fails to capture the dynamics of folding [82], thereby limiting its capacity to provide insight into the underlying physics of the system. Clearly, the problem of protein folding remains unsolved. Thus, it is crucial to carefully evaluate the ability of recent deep-learning-based methods to handle proteins with untraceable lineages before translating their successes to the design of *de novo* antibody therapies.

This paper presents a thorough comparison of pre-trained language models' (PLMs) intrinsic capacity to capture antibody structure, evaluating featurisers using both standard and antibody-specific metrics. Although they are not perfect, these models aim to generalise to unknown regions of the protein landscape and may, therefore, be suitable for capturing useful structural information in sequences with few homologs. By carefully evaluating the proficiency with which different PLMs reason about the structure of functionally important regions of antibody sequences, it was found that PLMs trained on general sets of proteins extract more information than those trained on antibodies alone. All featurisers considered were transformer-based, an architecture known to improve with parameter count and training time. It is therefore possible that the observed improvements could be attributed to the size of the GPLMs as well as the resources well-funded research projects have during training. An important future line of research would be to train much larger APLMs and compare their abilities to an equivalent-sized GPLM.

It was also demonstrated that these transformers capture structurally relevant information in a *subset* of their internal representations. After symmetrisation and APC, a sparse combination of these attention maps can closely approximate distances between residues which are close or far apart in sequence space. Unsurprisingly, the most important attention maps for predicting the structures of general proteins are not the same set of matrices relevant to antibody structure. When restricted to using an LR-based contact predictor, maximal precision was achieved using by training an L1-regularised CPH on a small set of antibodies. This allowed the model to identify which attention maps were important and ignore redundant or useless information. Extending this idea, a regularised LR model can be used to extract only information useful to predicting contacts with residues in HCDR3 (or any other desirable region). Therefore, it seems possible to improve prediction accuracy by training a model to select the most appropriate representations for the class of proteins being modelled. This selection process can be taken even further by selecting attention maps for a region of interest, HCRD3 for example. If speed is a priority, these findings indicate that downstream processes can efficiently utilise a small subset of representations without reducing performance.

Swapping out the LR-based model with more powerful estimators significantly improves performance on rare contacts as well as those inside HCDR3. This performance can be maintained using a subset of attention maps chosen by the LR model. The models explored were limited to making predictions based only on pairwise information extracted through self-attention. It is likely that even more complex models which combine these pairwise features, using convolutions for example, would further increase performance.

However, in all cases, regardless of the pretraining strategy used, none of the featurisers were capable of precisely predicting exceptionally rare contacts inside and outside the hypervariable regions of antibodies. If the transformer is incapable of exposing these relationships, it seems unreasonable to expect a downstream model to make accurate predictions. It remains unclear exactly what causes this failure but it does gesture towards an inability to generalise to unexplored regions of protein space. If these models truly understood the physics of folding, they would display no difficulty predicting uncommon contacts between positions in a chain. Based on this, it is possible that the abilities of PLMs remain bound by the data that they are exposed to during training. If this is true, more effort is required to build rich and varied antibody datasets before the potential of PLMs can be fully utilised for rational design. For a clearer picture of their ability to generalise, further research is required to determine whether structural motifs in PLMs' training datasets can also be observed in their test sets.

Furthermore, this evaluation is not exhaustive, and further research is necessary to assess the sensitivity of PLMs. For example, during lead optimization, a PLM must distinguish between structurally distinct proteins with highly similar sequences. Studies have shown that PLMs often fail to predict misfolding caused by single-point mutations (SPMs) [15]. This limitation can be measured by demonstrating that a transformer generates highly similar contact maps for a wild-type and a slightly mutated defunct sequence. The extent to which these adversarial [17] single-point mutations affect contact predictions in an antibody's binding region remains unexplored. Conversely, there is evidence that biologically irrelevant perturbations to the sequence can induce radical conformational changes to the PLM's predictions [52]. Therefore, before these models can be confidently used for the rational design of complementarity-determining region (CDR) loops, these limitations must be more clearly understood.

Despite this, it is still possible to generate value using PLMs in their current form. The findings presented in this report could inform the application of these new tools.

For example, precise contact map predictions can be used directly for 3D structure prediction. Contact maps and distograms are useful but for many applications, the 3D structure of a protein is more desirable. It is possible, using a tool like Rosetta, to initialise a physics-based folding engine using a contact map. Therefore, a potentially fast and efficient antibody structure predictor could be built using a pretrained GPLM, and a CPH which utilises its most important attention heads and is trained on antibody structures. The CPH's predictions will then be used to generate 3D conformations. Although this is not an end-to-end method, training a CPH is much faster than training an end-to-end model from scratch. Furthermore, inference speed is expected to improve as a result of efficient feature selection allowing for fast library characterisation.

It may also be possible to improve antibody-specific end-to-end methods. For example, IgFold utilises an APLM to replace the MSA in a network which bears strong similarities to AlphaFold2. This is because, empirically, AlphaFold2's Invariant Point Attention appears to be a particularly proficient network design to translate residue-level features into a 3D structure. Based on the findings presented above, it is likely that IgFold would be more successful if it simply replaced its APLM with a GPLM.

In addition to its utility for protein structure prediction, precise contact maps have the potential to aid in the exploration of protein space by identifying structurally similar proteins based on the distances between their contact maps. However, to achieve successful mapping of protein clusters to antibody functions, it is necessary to use powerful PLMs that can capture intricate structural relationships between residues. It is therefore feasible to employ a GPLM to create contact maps which can be used to cluster antibodies based on similarities in CDRs. This could provide valuable insights into which sequences are likely to bind to the same antigen. As models become larger and more expressive CPHs are used, one would expect the utility of these clusters to improve.

Furthermore, given an antigen structure, it is now possible to design an antibody-binding fragment which binds to that target. Contacts generated from a PLM's attention maps can be used to iteratively formulate a sequence which is likely to fold into a fixed structure designed to bind to the antigen. Based on their performance in predicting contacts in regions relevant to antibody binding, it seems advisable to use a GPLM rather than an APLM for this task although both approaches should be explored.

Overall, until larger APLMs are available, this report suggests that a base heuristic should be to leverage large GPLMs for design and prediction tasks which are structure-informed.

Bibliography

[1] Brennan Abanades, Guy Georges, Alexander Bujotzek, and Charlotte M Deane. ABlooper: Fast Accurate Antibody CDR Loop Structure Prediction with Accuracy Estimation. *Bioinformatics*, 38(7):1877–1880, April 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac016.

[2] K. R. Abhinandan and Andrew C. R. Martin. Analysis and Improvements to Kabat and Structurally Correct Numbering of Antibody Variable Domains. *Molecular Immunology*, 45(14):3832–3839, August 2008. ISSN 0161-5890. doi: 10.1016/j.molimm.2008.05.022.

[3] B. Al-Lazikani, A. M. Lesk, and C. Chothia. Standard Conformations for the Canonical Structures of Immunoglobulins. *Journal of Molecular Biology*, 273(4): 927–948, November 1997. ISSN 0022-2836. doi: 10.1006/jmbi.1997.1354.

[4] Ethan C. Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church. Unified Rational Protein Engineering with Sequence-Based Deep Representation Learning. *Nature Methods*, 16(12):1315–1322, December 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0598-1.

[5] Christian B Anfinsen. Principles That Govern the Folding of Protein Chains. *Science*, 181(4096):223–230, 1973.

[6] William R. Atchley, Kurt R. Wollenberg, Walter M. Fitch, Werner Terhalle, and Andreas W. Dress. Correlations Among Amino Acid Sites in bHLH Protein Domains: An Information Theoretic Analysis. *Molecular Biology and Evolution*, 17(1):164–178, January 2000. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a026229.

[7] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke,

K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science*, 373(6557):871–876, August 2021. doi: 10.1126/science.abj8754.

[8] Rupert Bartsch, Catharina Wenzel, and Guenther G Steger. Trastuzumab in the Management of Early and Advanced Stage Breast Cancer. *Biologics : Targets & Therapy*, 1(1):19–31, March 2007. ISSN 1177-5475.

[9] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The Protein Data Bank. *Nucleic acids research*, 28(1):235–242, 2000.

[10] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. ProteinBERT: A Universal Deep-Learning Model of Protein Sequence and Function. *Bioinformatics*, 38(8):2102–2110, April 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac020.

[11] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324.

[12] Leo Breiman. *Classification and Regression Trees*. Routledge, New York, October 2017. ISBN 978-1-315-13947-0. doi: 10.1201/9781315139470.

[13] John Bridle. Training Stochastic Model Recognition Algorithms as Networks Can Lead to Maximum Mutual Information Estimation of Parameters. In *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989.

[14] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models Are Few-Shot Learners, 2020.

[15] Gwen R. Buel and Kylie J. Walters. Can AlphaFold2 Predict the Impact of Missense Mutations on Structure? *Nature Structural & Molecular Biology*, 29(1):1–2, January 2022. ISSN 1545-9985. doi: 10.1038/s41594-021-00714-2.

[16] Ewen Callaway. Revolutionary Cryo-EM Is Taking over Structural Biology. *Nature*, 578(7794):201–201, February 2020. doi: 10.1038/d41586-020-00341-9.

[17] Ginevra Carbone, Francesca Cuturello, Luca Bortolussi, and Alberto Cazzaniga. Adversarial Attacks on Protein Language Models, October 2022.

[18] Mark L. Chiu, Dennis R. Goulet, Alexey Teplyakov, and Gary L. Gilliland. Antibody Structure and Function: The Basis for Engineering Therapeutics. *Antibodies*, 8(4): 55, December 2019. ISSN 2073-4468. doi: 10.3390/antib8040055.

[19] Cyrus Chothia and Arthur M. Lesk. Canonical Structures for the Hypervariable Regions of Immunoglobulins. *Journal of Molecular Biology*, 196(4):901–917, August 1987. ISSN 0022-2836. doi: 10.1016/0022-2836(87)90412-8.

[20] Ratul Chowdhury, Nazim Bouatta, Surojit Biswas, Christina Floristean, Anant Kharkar, Koushik Roy, Charlotte Rochereau, Gustaf Ahdritz, Joanna Zhang, George M. Church, Peter K. Sorger, and Mohammed AlQuraishi. Single-Sequence Protein Structure Prediction Using a Language Model and Deep Learning. *Nature Biotechnology*, pages 1–7, October 2022. ISSN 1546-1696. doi: 10.1038/s41587-022-01432-w.

[21] Jordan J. Clark, Mark L. Benson, Richard D. Smith, and Heather A. Carlson. Inherent versus Induced Protein Flexibility: Comparisons within and between Apo and Holo Structures. *PLoS Computational Biology*, 15(1), January 2019. doi: 10.1371/journal.pcbi.1006705.

[22] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics*, 25(11): 1422–1423, June 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp163.

[23] Tomer Cohen, Matan Halfon, and Dina Schneidman-Duhovny. NanoNet: Rapid End-to-End Nanobody Modeling by Deep Learning at Sub Angstrom Resolution, August 2021.

[24] Sara D'Angelo, Fortunato Ferrara, Leslie Naranjo, M. Frank Erasmus, Peter Hraber, and Andrew R. M. Bradbury. Many Routes to an Antibody Heavy-Chain CDR3: Necessary, Yet Insufficient, for Specific Binding. *Frontiers in Immunology*, 9, 2018. ISSN 1664-3224.

[25] Biopharma Dealmakers. Moving up with the Monoclonals. *Biopharma Dealmakers*, September 2019. doi: 10.1038/d43747-020-00765-2.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019.

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

[28] James Dunbar and Charlotte M. Deane. ANARCI: Antigen Receptor Numbering and Receptor Classification. *Bioinformatics (Oxford, England)*, 32(2):298–300, January 2016. ISSN 1367-4811. doi: 10.1093/bioinformatics/btv552.

[29] James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane. SAbDab: The Structural Antibody Database. *Nucleic Acids Research*, 42(D1):D1140–D1146, January 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1043.

[30] S.D. Dunn, L.M. Wahl, and G.B. Gloor. Mutual Information without the Influence of Phylogeny or Entropy Dramatically Improves Residue Contact Prediction. *Bioinformatics*, 24(3):333–340, February 2008. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm604.

[31] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved Contact Prediction in Proteins: Using Pseudolikelihoods to Infer Potts Models. *Physical Review E*, 87(1):012707, January 2013. doi: 10.1103/PhysRevE.87.012707.

[32] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing, May 2021.

[33] Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein Complex Prediction with AlphaFold-Multimer. *bioRxiv : the preprint server for biology*, pages 2021–10, 2022.

[34] Diego U. Ferreiro, Elizabeth A. Komives, and Peter G. Wolynes. Frustration in Biomolecules. *Quarterly Reviews of Biophysics*, 47(4):285–363, November 2014. ISSN 1469-8994. doi: 10.1017/S0033583514000092.

[35] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. ProtGPT2 Is a Deep Unsupervised Language Model for Protein Design. *Nature Communications*, 13(1):4348, July 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-32007-7.

[36] Anthony A. Fodor and Richard W. Aldrich. Influence of Conservation on Calculations of Amino Acid Covariance in Multiple Sequence Alignments. *Proteins: Structure, Function, and Bioinformatics*, 56(2):211–221, 2004. ISSN 1097-0134. doi: 10.1002/prot.20098.

[37] Douglas M. Fowler and Stanley Fields. Deep Mutational Scanning: A New Style of Protein Science. *Nature Methods*, 11(8):801–807, August 2014. ISSN 1548-7105. doi: 10.1038/nmeth.3027.

[38] Kunihiko Fukushima. Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics*, 20(3-4):121–136, September 1975. ISSN 0340-1200. doi: 10.1007/BF00342633.

[39] George Georgiou, Gregory C. Ippolito, John Beausang, Christian E. Busse, Hedda Wardemann, and Stephen R. Quake. The Promise and Challenge of High-Throughput Sequencing of the Antibody Repertoire. *Nature Biotechnology*, 32(2):158–168, February 2014. ISSN 1546-1696. doi: 10.1038/nbt.2782.

[40] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, 12(10):2451–2471, October 2000. ISSN 0899-7667. doi: 10.1162/089976600300015015.

[41] Stefano Gianni, Carlo Camilloni, Rajanish Giri, Angelo Toto, Daniela Bonetti, Angela Morrone, Pietro Sormanni, Maurizio Brunori, and Michele Vendruscolo. Understanding the Frustration Arising from the Competition between Function, Misfolding, and Aggregation in a Globular Protein. *Proceedings of the National Academy of Sciences*, 111(39):14141–14146, September 2014. doi: 10.1073/pnas.1405233111.

[42] Ricardo J. Giordano, Marina Cardó-Vila, Johanna Lahdenranta, Renata Pasqualini, and Wadih Arap. Biopanning and Rapid Analysis of Selective Interactive Ligands. *Nature Medicine*, 7(11):1249–1253, November 2001. ISSN 1546-170X. doi: 10.1038/nm1101-1249.

[43] U. Göbel, C. Sander, R. Schneider, and A. Valencia. Correlated Mutations and Residue Contacts in Proteins. *Proteins*, 18(4):309–317, April 1994. ISSN 0887-3585. doi: 10.1002/prot.340180402.

[44] Victor Greiff, Enkelejda Miho, Ulrike Menzel, and Sai T. Reddy. Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires. *Trends in Immunology*, 36 (11):738–749, November 2015. ISSN 1471-4906. doi: 10.1016/j.it.2015.09.006.

[45] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class AdaBoost. *Statistics and Its Interface*, 2(3):349–360, 2009. ISSN 19387989, 19387997. doi: 10.4310/SII.2009.v2.n3.a8.

[46] S Henikoff and J G Henikoff. Amino Acid Substitution Matrices from Protein Blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89 (22):10915–10919, November 1992. ISSN 0027-8424.

[47] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial Attention in Multidimensional Transformers, December 2019.

[48] A. Honegger and A. Plückthun. Yet Another Numbering Scheme for Immunoglobulin Variable Domains: An Automatic Modeling and Analysis Tool. *Journal of Molecular Biology*, 309(3):657–670, June 2001. ISSN 0022-2836. doi: 10.1006/jmbi.2001.4662.

[49] Xiaoqiang Huang, Robin Pearce, and Yang Zhang. EvoEF2: Accurate and Fast Energy Function for Computational Protein Design. *Bioinformatics*, 36(4):1135–1142, February 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz740.

[50] Alissa M. Hummer, Brennan Abanades, and Charlotte M. Deane. Advances in Computational Structure-Based Antibody Design. *Current Opinion in Structural Biology*, 74:102379, June 2022. ISSN 0959-440X. doi: 10.1016/j.sbi.2022.102379.

[51] Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. What Is the Best Multi-Stage Architecture for Object Recognition? In *2009 IEEE 12th International Conference on Computer Vision*, pages 2146–2153, September 2009. doi: 10.1109/ICCV.2009.5459469.

[52] Sumit Kumar Jha, Arvind Ramanathan, Rickard Ewetz, Alvaro Velasquez, and Susmit Jha. Protein Folding Neural Networks Are Not Robust. *arXiv preprint arXiv:2109.04460*, 2021.

[53] David T Jones and Shaun M Kandathil. High Precision in Protein Contact Prediction Using Fully Convolutional Neural Networks and Minimal Sequence Features. *Bioinformatics*, 34(19):3308–3315, October 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty341.

[54] David T. Jones, Daniel W. A. Buchan, Domenico Cozzetto, and Massimiliano Pontil. PSICOV: Precise Structural Contact Prediction Using Sparse Inverse Covariance

Estimation on Large Multiple Sequence Alignments. *Bioinformatics*, 28(2):184–190, January 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr638.

[55] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2.

[56] David Jung, Cosmas Giallourakis, Raul Mostoslavsky, and Frederick Alt. Mechanism and Control of V(D)J Recombination at the Immunoglobulin Heavy Chain Locus. *Annual review of immunology*, 24:541–70, February 2006. doi: 10.1146/annurev.immunol.23.021704.115830.

[57] Hetunandan Kamisetty, Sergey Ovchinnikov, and David Baker. Assessing the Utility of Coevolution-Based Residue– Residue Contact Predictions in a Sequence- and Structure-Rich Era. *Proceedings of the National Academy of Sciences*, 110(39):15674–15679, September 2013. doi: 10.1073/pnas.1314045110.

[58] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*, 2020.

[59] Hélène Kaplon, Alicia Chenoweth, Silvia Crescioli, and Janice M Reichert. Antibodies to Watch in 2022. In *MAbs*, volume 14, page 2014296. Taylor & Francis, 2022.

[60] David E. Kim, Frank DiMaio, Ray Yu-Ruei Wang, Yifan Song, and David Baker. One Contact for Every Twelve Residues Allows Robust and Accurate Topology-Level Protein Structure Modeling. *Proteins: Structure, Function, and Bioinformatics*, 82(S2):208–218, 2014. ISSN 1097-0134. doi: 10.1002/prot.24374.

[61] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[62] Tomasz Kosciolek and David T. Jones. De Novo Structure Prediction of Globular Proteins Aided by Sequence Variation-Derived Contacts. *PLOS ONE*, 9(3):e92197, March 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0092197.

[63] Aleksandr Kovaltsuk, Konrad Krawczyk, Jacob D. Galson, Dominic F. Kelly, Charlotte M. Deane, and Johannes Trück. How B-Cell Receptor Repertoire Sequencing Can Be Enriched with Structural Antibody Data. *Frontiers in Immunology*, 8, 2017. ISSN 1664-3224.

[64] Ben Krause, Liang Lu, Iain Murray, and Steve Renals. Multiplicative LSTM for Sequence Modelling, 2016.

[65] Vered Kunik and Yanay Ofran. The Indistinguishability of Epitopes from Protein Surface Is Explained by the Distinct Binding Preferences of Each of the Six Antigen-Binding Loops. *Protein Engineering, Design and Selection*, 26(10):599–609, October 2013. ISSN 1741-0126. doi: 10.1093/protein/gzt027.

[66] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A Lite Bert for Self-Supervised Learning of Language Representations. *arXiv preprint arXiv:1909.11942*, 2019.

[67] Jinwoo Leem, James Dunbar, Guy Georges, Jiye Shi, and Charlotte M Deane. ABodyBuilder: Automated Antibody Structure Prediction with Data– Driven Accuracy Estimation. In *MAbs*, volume 8, pages 1259–1268. Taylor & Francis, 2016.

[68] Jinwoo Leem, Laura S. Mitchell, James H. R. Farmery, Justin Barton, and Jacob D. Galson. Deciphering the Language of Antibodies Using Self-Supervised Learning. *Patterns*, 3(7):100513, July 2022. ISSN 2666-3899. doi: 10.1016/j.patter.2022.100513.

[69] Marie-Paule Lefranc, Christelle Pommié, Manuel Ruiz, Véronique Giudicelli, Elodie Foulquier, Lisa Truong, Valérie Thouvenin-Contet, and Gérard Lefranc. IMGT Unique Numbering for Immunoglobulin and T Cell Receptor Variable Domains and Ig Superfamily V-like Domains. *Developmental and Comparative Immunology*, 27(1): 55–77, January 2003. ISSN 0145-305X. doi: 10.1016/s0145-305x(02)00039-3.

[70] Rosalba Lepore, Pier P. Olimpieri, Mario A. Messih, and Anna Tramontano. PIGSPro: Prediction of immunoGlobulin Structures V2. *Nucleic Acids Research*, 45(W1):W17–W23, July 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx334.

[71] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, and Alexander Rives. Language Models of Protein Sequences at the Scale of Evolution Enable Accurate Structure Prediction, July 2022.

[72] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A Robustly Optimized Bert Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

[73] C. Marks and C. M. Deane. Antibody H3 Structure Prediction. *Computational and Structural Biotechnology Journal*, 15:222–231, January 2017. ISSN 2001-0370. doi: 10.1016/j.csbj.2017.01.010.

[74] Debora S. Marks, Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLOS ONE*, 6(12):e28766, December 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0028766.

[75] Debora S. Marks, Thomas A. Hopf, and Chris Sander. Protein Structure Prediction from Sequence Variation. *Nature Biotechnology*, 30(11):1072–1080, November 2012. ISSN 1546-1696. doi: 10.1038/nbt.2419.

[76] Vivien Marx. Method of the Year: Protein Structure Prediction. *Nature Methods*, 19(1):5–10, January 2022. ISSN 1548-7105. doi: 10.1038/s41592-021-01359-1.

[77] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. Direct-Coupling Analysis of Residue Coevolution Captures Native Contacts across Many Protein Families. *Proceedings of the National Academy of Sciences*, 108(49): E1293–E1301, December 2011. doi: 10.1073/pnas.1111471108.

[78] Benjamin North, Andreas Lehmann, and Roland L. Dunbrack. A New Clustering of Antibody CDR Loop Conformations. *Journal of Molecular Biology*, 406(2):228–256, February 2011. ISSN 0022-2836. doi: 10.1016/j.jmb.2010.10.030.

[79] Jaroslaw Nowak, Terry Baker, Guy Georges, Sebastian Kelm, Stefan Klostermann, Jiye Shi, Sudharsan Sridharan, and Charlotte M. Deane. Length-Independent Structural Similarities Enrich the Antibody CDR Canonical Class Model. *mAbs*, 8 (4):751–760, June 2016. ISSN 1942-0870. doi: 10.1080/19420862.2016.1158370.

[80] Tobias H. Olsen, Fergus Boyles, and Charlotte M. Deane. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1):141–146, 2022. ISSN 1469-896X. doi: 10.1002/pro.4205.

[81] Tobias H Olsen, Iain H Moal, and Charlotte M Deane. AbLang: An Antibody Language Model for Completing Antibody Sequences. *Bioinformatics Advances*, 2 (1):vbac046, January 2022. ISSN 2635-0041. doi: 10.1093/bioadv/vbac046.

[82] Carlos Outeiral, Daniel A Nissley, and Charlotte M Deane. Current Structure Predictors Are Not Learning the Physics of Protein Folding. *Bioinformatics*, 38(7): 1881–1887, April 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab881.

[83] Sergey Ovchinnikov, David E. Kim, Ray Yu-Ruei Wang, Yuan Liu, Frank DiMaio, and David Baker. Improved de Novo Structure Prediction in CASP11 by Incorporating Coevolution Information into Rosetta. *Proteins: Structure, Function, and Bioinformatics*, 84(S1):67–75, 2016. ISSN 1097-0134. doi: 10.1002/prot.24974.

[84] Eduardo A. Padlan. Anatomy of the Antibody Molecule. *Molecular Immunology*, 31 (3):169–217, February 1994. ISSN 0161-5890. doi: 10.1016/0161-5890(94)90001-9.

[85] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-Learn: Machine Learning in Python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[86] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models Are Unsupervised Multitask Learners. 2019.

[87] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, July 2021.

[88] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating Protein Transfer Learning with TAPE, June 2019.

[89] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer Protein Language Models Are Unsupervised Structure Learners. *bioRxiv*, page 2020.12.15.422761, January 2020. doi: 10.1101/2020.12.15.422761.

[90] Roshan Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F. Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA Transformer, February 2021.

[91] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proceedings of the National Academy of Sciences*, 118(15): e2016239118, April 2021. doi: 10.1073/pnas.2016239118.

[92] Sarah A. Robinson, Matthew I. J. Raybould, Constantin Schneider, Wing Ki Wong, Claire Marks, and Charlotte M. Deane. Epitope Profiling Using Computational Structural Modelling Demonstrated on Coronavirus-Binding Antibodies. *PLOS*

Computational Biology, 17(12):e1009675, December 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1009675.

[93] Nathan J. Rollins, Kelly P. Brock, Frank J. Poelwijk, Michael A. Stiffler, Nicholas P. Gauthier, Chris Sander, and Debora S. Marks. Inferring Protein 3D Structure from Deep Mutation Scans. *Nature Genetics*, 51(7):1170–1176, July 2019. ISSN 1546-1718. doi: 10.1038/s41588-019-0432-9.

[94] James P. Roney and Sergey Ovchinnikov. State-of-the-Art Estimation of Protein Model Accuracy Using AlphaFold. *Physical Review Letters*, 129(23):238101, November 2022. doi: 10.1103/PhysRevLett.129.238101.

[95] Aviv A. Rosenberg, Ailie Marx, and Alex M. Bronstein. Codon-Specific Ramachandran Plots Show Amino Acid Backbone Conformation Depends on Identity of the Translated Codon. *Nature Communications*, 13(1):2815, May 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-30390-9.

[96] Jeffrey A Ruffolo, Carlos Guerra, Sai Pooja Mahajan, Jeremias Sulam, and Jeffrey J Gray. Geometric Potentials from Deep Learning Improve Prediction of CDR H3 Loop Structures. *Bioinformatics*, 36(Supplement_1):i268–i275, July 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa457.

[97] Jeffrey A. Ruffolo, Jeffrey J. Gray, and Jeremias Sulam. Deciphering Antibody Affinity Maturation with Language Models and Weakly Supervised Learning, December 2021.

[98] Jeffrey A. Ruffolo, Lee-Shin Chu, Sai Pooja Mahajan, and Jeffrey J. Gray. Fast, Accurate Antibody Structure Prediction from Deep Learning on Massive Set of Natural Antibodies, April 2022.

[99] Jeffrey A. Ruffolo, Jeremias Sulam, and Jeffrey J. Gray. Antibody Structure Prediction Using Interpretable Deep Learning. *Patterns*, 3(2):100406, February 2022. ISSN 2666-3899. doi: 10.1016/j.patter.2021.100406.

[100] Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) Equivariant Graph Neural Networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9323–9332. PMLR, July 2021.

[101] Joerg Schaarschmidt, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Alexandre M.J.J. Bonvin. Assessment of Contact Predictions in CASP12: Co-evolution and Deep Learning Coming of Age. *Proteins: Structure, Function, and Bioinformatics*, 86(S1):51–66, 2018. ISSN 1097-0134. doi: 10.1002/prot.25407.

[102] Jörn M. Schmiedel and Ben Lehner. Determining Protein Structures Using Deep Mutagenesis. *Nature Genetics*, 51(7):1177–1186, July 2019. ISSN 1546-1718. doi: 10.1038/s41588-019-0431-x.

[103] Stefan Seemayer, Markus Gruber, and Johannes Söding. CCMpred— Fast and Precise Prediction of Protein Residue– Residue Contacts from Correlated Mutations. *Bioinformatics*, 30(21):3128–3130, November 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu500.

[104] Inbal Sela-Culang, Vered Kunik, and Yanay Ofran. The Structural Basis of Antibody-Antigen Recognition. *Frontiers in Immunology*, 4, 2013. ISSN 1664-3224.

[105] Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature*, 577 (7792):706–710, January 2020. ISSN 1476-4687. doi: 10.1038/s41586-019-1923-7.

[106] Pietro Sormanni, Francesco A. Aprile, and Michele Vendruscolo. Third Generation Antibody Discovery Methods: In Silico Rational Design. *Chemical Society Reviews*, 47(24):9137–9157, December 2018. ISSN 1460-4744. doi: 10.1039/C8CS00523K.

[107] Tyler N. Starr and Joseph W. Thornton. Epistasis in Protein Evolution. *Protein Science*, 25(7):1204–1218, 2016. ISSN 1469-896X. doi: 10.1002/pro.2897.

[108] Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, Cathy H. Wu, and the UniProt Consortium. UniRef Clusters: A Comprehensive and Scalable Alternative for Improving Sequence Similarity Searches. *Bioinformatics*, 31(6): 926–932, March 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu739.

[109] Gian Gaetano Tartaglia, Sebastian Pechmann, Christopher M. Dobson, and Michele Vendruscolo. Life on the Edge: A Link between Gene Expression Levels and Aggregation Rates of Human Proteins. *Trends in Biochemical Sciences*, 32(5): 204–206, May 2007. ISSN 0968-0004. doi: 10.1016/j.tibs.2007.03.005.

[110] Gian Gaetano Tartaglia, Sebastian Pechmann, Christopher M. Dobson, and Michele Vendruscolo. A Relationship between mRNA Expression Levels and Protein Solubility in E. Coli. *Journal of Molecular Biology*, 388(2):381–389, May 2009. ISSN 1089-8638. doi: 10.1016/j.jmb.2009.03.002.

[111] Alexey Teplyakov, Jinquan Luo, Galina Obmolova, Thomas J. Malia, Raymond Sweet, Robyn L. Stanfield, Sreekumar Kodangattil, Juan Carlos Almagro, and Gary L.

Gilliland. Antibody Modeling Assessment II. Structures and Models. *Proteins: Structure, Function, and Bioinformatics*, 82(8):1563–1582, 2014. ISSN 1097-0134. doi: 10.1002/prot.24554.

[112] The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 36(suppl_1):D190–D195, January 2008. ISSN 0305-1048. doi: 10.1093/nar/gkm895.

[113] Kathryn E. Tiller and Peter M. Tessier. Advances in Antibody Design. *Annual review of biomedical engineering*, 17:191–216, 2015. ISSN 1523-9829. doi: 10.1146/annurev-bioeng-071114-040733.

[114] Nobuhiko Tokuriki and Dan S Tawfik. Stability Effects of Mutations and Protein Evolvability. *Current Opinion in Structural Biology*, 19(5):596–604, October 2009. ISSN 0959-440X. doi: 10.1016/j.sbi.2009.08.003.

[115] Elisabetta Traggiai, Stephan Becker, Kanta Subbarao, Larissa Kolesnikova, Yasushi Uematsu, Maria Rita Gismondo, Brian R. Murphy, Rino Rappuoli, and Antonio Lanzavecchia. An Efficient Method to Make Human Monoclonal Antibodies from Memory B Cells: Potent Neutralization of SARS Coronavirus. *Nature Medicine*, 10 (8):871–875, August 2004. ISSN 1546-170X. doi: 10.1038/nm1080.

[116] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, December 2017.

[117] Robert Verkuil, Ori Kabeli, Yilun Du, Basile IM Wicky, Lukas F Milles, Justas Dauparas, David Baker, Sergey Ovchinnikov, Tom Sercu, and Alexander Rives. Language models generalize beyond natural proteins. *bioRxiv*, pages 2022–12, 2022.

[118] Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. BERTology Meets Biology: Interpreting Attention in Protein Language Models, March 2021.

[119] Konstantin Weissenow, Michael Heinzinger, and Burkhard Rost. Protein Language-Model Embeddings for Fast, Accurate, and Alignment-Free Protein Structure Prediction. *Structure*, 30(8):1169–1177.e4, August 2022. ISSN 0969-2126. doi: 10.1016/j.str.2022.05.001.

[120] Brian D. Weitzner, Jeliazko R. Jeliazkov, Sergey Lyskov, Nicholas Marze, Daisuke Kuroda, Rahel Frick, Jared Adolf-Bryfogle, Naireeta Biswas, Roland L. Dunbrack, and Jeffrey J. Gray. Modeling and Docking of Antibody Structures with Rosetta.

Nature Protocols, 12(2):401–416, February 2017. ISSN 1750-2799. doi: 10.1038/nprot.2016.180.

[121] Jens Wrammert, Kenneth Smith, Joe Miller, William A. Langley, Kenneth Kokko, Christian Larsen, Nai-Ying Zheng, Israel Mays, Lori Garman, Christina Helms, Judith James, Gillian M. Air, J. Donald Capra, Rafi Ahmed, and Patrick C. Wilson. Rapid Cloning of High-Affinity Human Monoclonal Antibodies against Influenza Virus. *Nature*, 453(7195):667–671, May 2008. ISSN 1476-4687. doi: 10.1038/nature06890.

[122] Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, and Jian Peng. High-Resolution de Novo Structure Prediction from Primary Sequence. *bioRxiv*, page 2022.07.21.500999, January 2022. doi: 10.1101/2022.07.21.500999.

[123] T. T. Wu and E. A. Kabat. An Analysis of the Sequences of the Variable Regions of Bence Jones Proteins and Myeloma Light Chains and Their Implications for Antibody Complementarity. *The Journal of Experimental Medicine*, 132(2):211–250, August 1970. ISSN 0022-1007. doi: 10.1084/jem.132.2.211.

[124] Kazuo Yamashita, Kazuyoshi Ikeda, Karlou Amada, Shide Liang, Yuko Tsuchiya, Haruki Nakamura, Hiroki Shirai, and Daron M. Standley. Kotai Antibody Builder: Automated High-Resolution Structural Modeling of Antibodies. *Bioinformatics*, 30(22):3279–3280, November 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu510.

[125] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv preprint arXiv:1511.07122*, 2015.