

Deep learning: a new technology in  
antibody design

**Michael Hall**

Login ID: mah51

School of Biosciences  
University of Kent

Final Year Project (Module BI600)

**2021**

Supervisor: Michelle Garrett  
Dissertation Project

Word Count: 8093

## Checklist for your project hand in

(Please tick completed items and return with your project report)

**NAME: Michael Hall**

Deadline for submission of project report 12 noon Wednesday Week E1 2021	
✓	Have you submitted an electronic version of your project report (Deadline 12 noon Wednesday Week 24 2019)?
✓	Margins throughout project report must be no smaller than 2.5 cm and text must be double spaced (see instructions on Moodle)
✓	Title page shows year as being 2021 and complies with instructions
✓	Have you signed the plagiarism form (below)?
✓	Have you completed the on-line evaluation form?

### DECLARATION

I confirm that the wording of this report is entirely my own. No part of the report has been copied from scientific journals, web sites or any other sources. For a detailed statement on plagiarism the student is referred to the guidelines for preparing project reports.

Signature of the student:



Date: 12/04/21

# Contents

<b>1</b>	<b>Abstract</b>	<b>1</b>
1.1	Abbreviations . . . . .	2
1.2	Acknowledgement . . . . .	3
<b>2</b>	<b>Introduction</b>	<b>4</b>
2.1	Background . . . . .	4
2.1.1	Monoclonal antibodies and their production . . . . .	5
2.1.2	Human Monoclonal Antibodies . . . . .	8
2.1.3	Current therapeutics and diagnostics . . . . .	8
2.1.4	The need for computational power . . . . .	10
2.1.5	Introduction to deep learning. . . . .	11
2.2	About this dissertation. . . . .	17
<b>3</b>	<b>Methods</b>	<b>18</b>
3.1	Searching Literature . . . . .	18
3.1.1	Structural prediction . . . . .	18
3.1.2	Antibody-Antigen binding predictions . . . . .	21
3.1.3	CASP13 comparison. . . . .	23
3.1.4	Testing for an overfit . . . . .	23
<b>4</b>	<b>Results</b>	<b>25</b>
4.1	Overview . . . . .	25
4.1.1	Residual Neural Networks . . . . .	25
4.1.2	Convolutional Neural Networks . . . . .	30
4.1.3	Comparison using CASP13 data . . . . .	33
4.1.4	Identifying performance factors . . . . .	36
4.1.5	AlphaFold2 . . . . .	37
<b>5</b>	<b>Discussion</b>	<b>39</b>
5.1	Difficulties faced . . . . .	40
5.1.1	Improvements . . . . .	41
5.2	Concluding remarks . . . . .	41
	<b>References</b>	<b>43</b>

## Abstract

Antibodies are highly diverse proteins, that specifically target a large range of molecules. This diversity can be utilised via engineering techniques to produce antibodies that bind specific targets. Currently the discovery process to identify antibodies that can bind to the desired target are drawn out and expensive to perform. However, a computational tool named deep learning is a rapidly developing technology that can be trained to recognise patterns within data. This dissertation discovered, that the lack of data available specific to antibody structures prevents effective teaching of more complex deep learning networks, resulting in most research being focused on general protein structure. However, programs focused on protein structure make fairly effective predictions, and are rapidly improving. Neural network architectures varied depending on their application, with convolutional neural networks being preferred when less training data was available and residual neural networks chosen to analyse complex relationships. One article utilised a long short-term memory network, which due to its data persistence, that acts like memory, utilised the large dataset it was provided with, to achieve a high accuracy of 92.43% on a test set of data. Unfortunately, due to poor reporting in a lot of the literature, it is hard to compare current programs quantitatively, however

a set of reporting standards and uniform test data sets could address this problem.

## 1.1. Abbreviations

(AUC) Area under the ROC

(CASP) Critical Assessment of Protein Structure Prediction [1]

(cDNA) Complementary DNA

(CDR) Complementarity-Determining Regions

(CNN) Convolutional Neural Network

(F<sub>ab</sub>) Fragment Antigen-Binding

(F<sub>c</sub>) Fragment Crystallisable

(FM) Free Modelling

(GDT) Global Distance Test

(GDT<sub>HA</sub>) Global Distance Test High Accuracy

(GDT<sub>TS</sub>) Global Distance Test Total Score

(GPU) Graphics Processing Unit

(HAMAs) Human Anti-Murine Antibodies

(HHblits) HMM-HMM–based lightning-fast iterative sequence search

(LSTM) Long Short-Term Memory

(mAbs) Monoclonal Antibodies

(MSA) Multiple Sequence Alignment

(NN) Neural Network

(PCR) Polymerase Chain Reaction

(PDB) Protein Data Bank [2]

(PointNets) Point Neural Networks

(PPI) Protein-Protein Interactions

(PRISMA) Preferred Reporting Items for Systematic Reviews and Meta-Analyses [3]

(rAbs) Recombinant Monoclonal Antibodies

(ResNet) Residual Neural Network

(RMSD) Root Mean Square Deviation

(RNN) Recurrent Neural Network

(ROC) Receiver Operating Characteristic Curve

(SAbDab) The Structural Antibody Database [4]

(scFvs) Single-Chain Fragment Variable

(TBM) Template Based Modelling

## 1.2. Acknowledgement

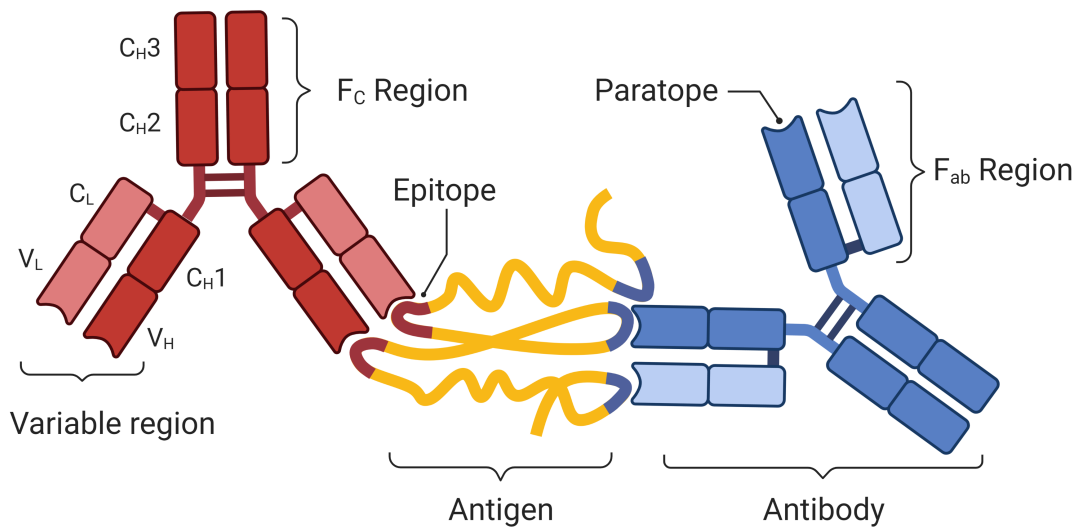
To preface this paper I would like to thank my supervisor Dr Michelle Garrett for all the assistance she provided during the creation of this dissertation. Including feedback, on the introduction, methods, and results as well as further support in investigating my chosen topic.

# 2

## Introduction

### 2.1. Background

We are constantly exposed to foreign antigens; they are present on the surface of microbes and viruses or as stand-alone molecules such as toxins. Our immune system uses a variety of mechanisms to protect us. Two important mechanisms are innate and adaptive immunity. Innate immunity acts as a first line of defence and includes physical and chemical barriers; as well as other immune action from complement and non-specific responses. Adaptive immunity utilises large y-shaped proteins known as antibodies to target specific molecules named antigens. Each arm of the antibody houses a binding region [5] named a paratope responsible for binding its complementary antigen on its epitope, a singular antigen can have multiple epitopes (**Figure 2.1**). The specificity of antibodies allows them to carry out several important functions to neutralise pathogens. For example, binding to virus surface proteins can block fusion events, preventing cell infection [6]. Furthermore, antigens can aid in innate immunity by binding to phagocytes via the  $F_C$  regions, inducing phagocytosis to neutralise pathogens faster [7]. The antibody paratopes are incredibly specific, due to their high



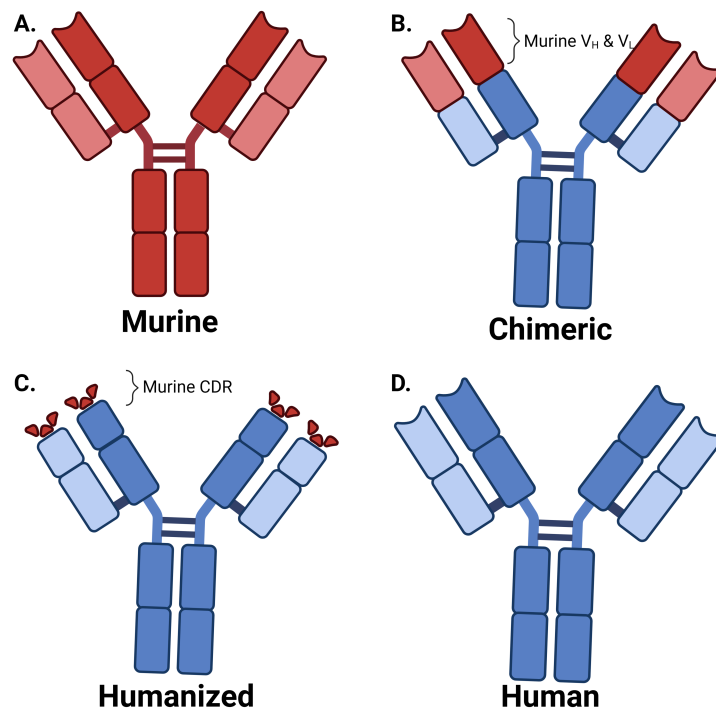
**Figure 2.1: Immunoglobulin G structure** Author adapted from “Antigen Recognition by Antibodies Description”, by BioRender.com. IgG is formed from four chains; Sections C<sub>H</sub>3, C<sub>H</sub>2, C<sub>H</sub>1, and V<sub>H</sub> make up the heavy chains of 50 kDa each and C<sub>L</sub> and V<sub>L</sub> make up the light chains of 25 kDa each [5].

binding affinity for their epitope counterparts and low binding affinity for other molecules, this specificity makes them useful for therapeutic and diagnostic applications.

### 2.1.1. Monoclonal antibodies and their production

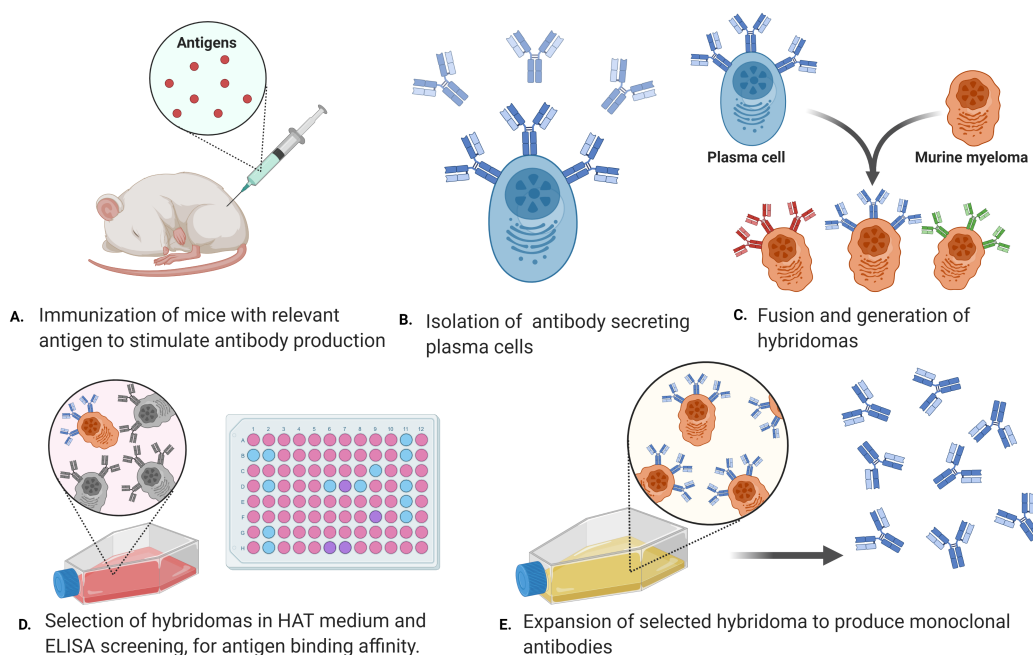
Monoclonal antibodies (mAbs) are antibodies generated from clones of a single B lymphocyte and therefore possess identical paratopes. Their use in therapeutics is becoming increasingly popular due to their ability to bind to a large range of targets, this is reflected in the doubling of their market value between 2012 and 2017 [8]. Four key types of mAbs are: murine, chimeric, humanised, and human (**Figure 2.2**). The first therapeutic mAb was of murine descent and was released in 1986 [9] under the name Orthoclone. It aimed to solve transplant rejection by targeting the CD<sub>3</sub> protein on T-cells. Murine mAbs are produced via cells known as hybridomas, a concept which was developed in 1975 by C. Milstein, G. J. F. Köhler, and N. K. Jerne for which they later received a Nobel Prize [10]. They injected their antigens (sheep red blood cells) into a mouse, allowing the B lymphocytes to mature and differentiate into antibody-producing plasma cells via CD4<sup>+</sup> T lymphocyte activation [11]. These B lymphocytes were then





**Figure 2.2: Four types of monoclonal antibody.** Created with biorender.com Murine domains are shown in red and human domains in blue. This shows the variation between a full murine mAb (a), a chimeric mAb (b) containing the variable regions from a murine antibody, a humanised mAb (c) that is mostly human domains apart from the hypervariable regions, a fully human mAb (d).

extracted from the spleen of the mouse and fused with murine myeloma cells lacking the hypoxanthine-guanine phosphoribosyltransferase (HGPRT) gene, to form hybridomas. To ensure the survival of HGPRT<sup>+</sup> fused hybridomas, they were grown on a hypoxanthine-aminopterin-thymidine (HAT) medium. Aminopterin inhibits dihydrofolate reductase [12], a vital enzyme in the *de novo* synthesis pathway. HGPRT catalyses the formation of purine nucleotides from guanine or hypoxanthine in a process known as the salvage pathway. Therefore, unfused myelomas in the HAT medium will not replicate as neither the salvage pathway nor *de novo* pathway can synthesise DNA. However, fused hybridomas have functional HGPRT translated from the B-cell genome can produce DNA via the salvage pathway and replicate. Once hybridomas have proliferated in the HAT medium their antibody products are tested using an enzyme-linked immunosorbent assay (ELISA) which confirms binding affinity for the antigen, the hybridoma with the highest affinity is selected and antibody proteins are extracted.



**Figure 2.3: Production of murine mAbs via hybridoma formation.** Author adapted from “**Monoclonal Antibodies Production**”, by BioRender.com A. Antigens are injected into the mouse stimulating plasma cell formation. B & C. Antibodies are isolated and fused with hybridomas. D. Hybridomas undergo selection pressure via the HAT medium to remove myeloma cells. E. Expansion to produce many monoclonal antibodies.

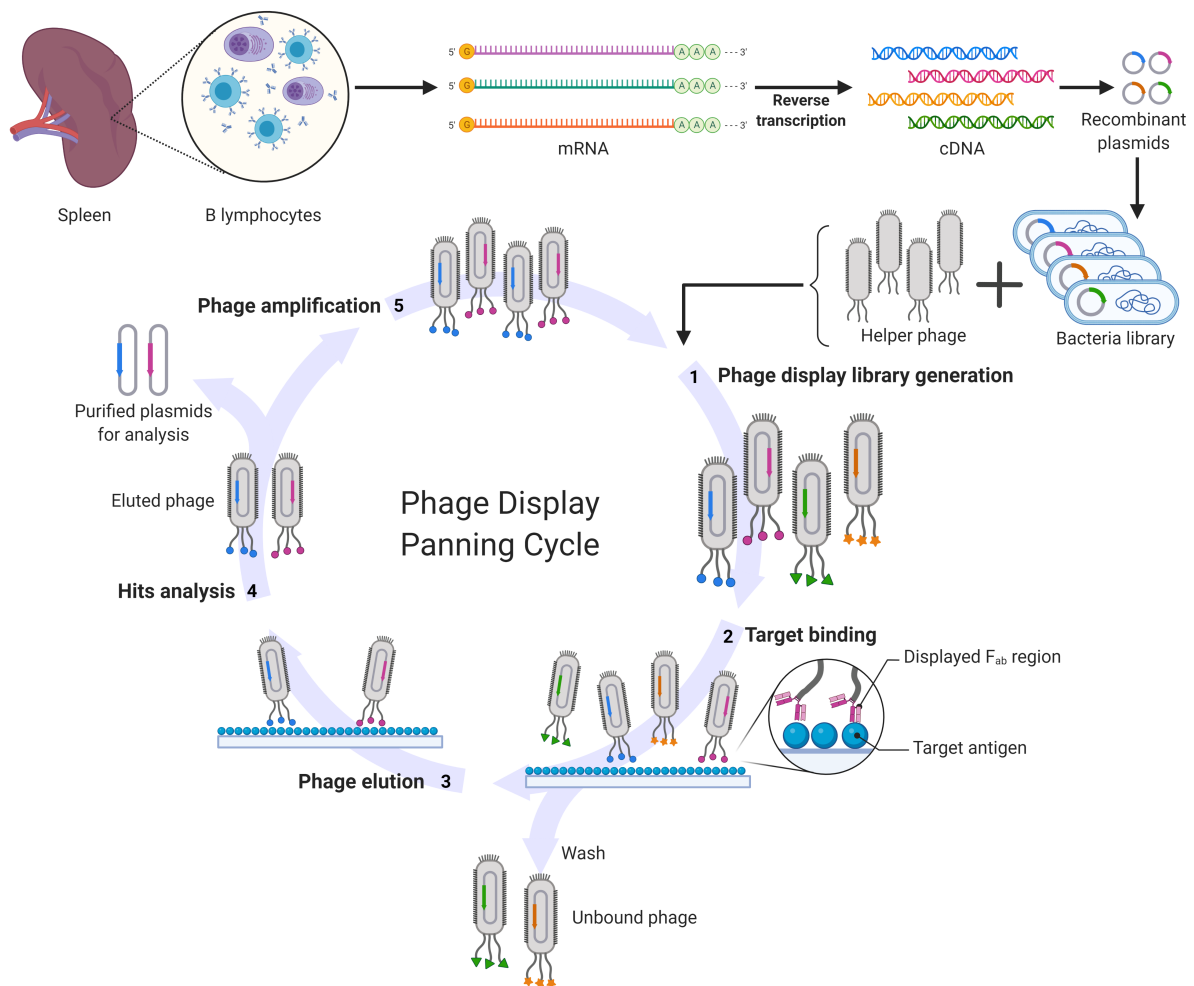
Unfortunately, murine mAbs do not act as effective therapeutics. On ingestion, the immune system targets these murine mAbs with human anti-murine antibodies (HAMAs), especially on repeated exposure. This leads to undesirable side effects such as an increased speed of clearance and possibly an allergic reaction [13, 14]. Chimeric mAbs use a human constant region ( $C_{H3}$ ,  $C_{H2}$ ,  $C_{H1}$ ,  $C_L$  (**Figure 2.1**)) with murine variable regions. This proved effective, resulting in a large immunogenicity reduction [15], suggesting HAMA antibodies mainly targeted one of the  $F_C$ ,  $C_L$ , or  $C_{H1}$  regions on the murine mAbs [13]. The next-generation humanised mAb (**Figure 2.2.C**) only included the murine hypervariable complementary determining regions (CDRs) [16]. CDRs are responsible for forming the paratope and are the loops that contact the antigen. This further reduces the HAMA response, but to minimise immune action a requirement for human antibodies arose.

### 2.1.2. Human Monoclonal Antibodies

Fully human monoclonal antibodies can be produced via hybridoma techniques with transgenic mice [17]. A faster but more expensive alternative to the hybridoma technique produces recombinant monoclonal antibodies (rAbs). The method was introduced by S.F.Parmley and G.P.Smith after Smith demonstrated a phage could present polypeptides on its surface [18, 19]. This technique, denominated phage display, requires a library of phage viruses. Initially, mRNA must be extracted from human B lymphocytes, which is used to create complementary DNA (cDNA) via a reverse transcription polymerase chain reaction (PCR) method. A set of primers designed to bind DNA responsible for the heavy and light chain regions of the antibody are used in PCR. Once amplified these cDNA copies are randomly joined and inserted into suitable phagemids via restriction enzymes and linker DNA, M13 phagemids are commonly used due to their mechanism of chronic infection, yielding a high output of reproduced phages from the bacteria. The recombinant phagemids are then inserted into *E.Coli* along with the rest of the phage genome from a helper phage. The *E.Coli* produce phages with single-chain fragment variable (scFvs) or  $F_{ab}$  regions (**Figure 2.1**) present on their cell surface. Once a library is established, an antigen can be presented, phages that bind with a high affinity to the antigen will remain in the library while weakly binding phages are washed away. This is repeated to find high-affinity mAbs, and is known as bio-panning (**Figure 2.4**) [20, 21]. Once phages with high binding affinity have been separated they can be sequenced and analysed to find common and conserved residues via computational methods. Whilst phage display is a faster method, it is more expensive and resulting antibodies often have less affinity for their antigen.

### 2.1.3. Current therapeutics and diagnostics

Some major examples of mAbs as therapeutics are adalimumab, rituximab, and bevacizumab. Adalimumab, a human rAb, treats various types of arthritis by targeting



**Figure 2.4: Phage panning as part of the phage display method.** Author adapted from “Phage Display Panning”, by BioRender.com mRNA extracted from B lymphocytes in the donor’s spleen undergoes reverse transcription to cDNA where PCR amplifies specific antibody associated DNA. DNA is then inserted into a plasmid vector and expressed in *E.Coli*. After formation of the phage library, target binding of antibody occurs and unbound phages are washed away. The phagemids are then extracted from eluted phage and sequenced.

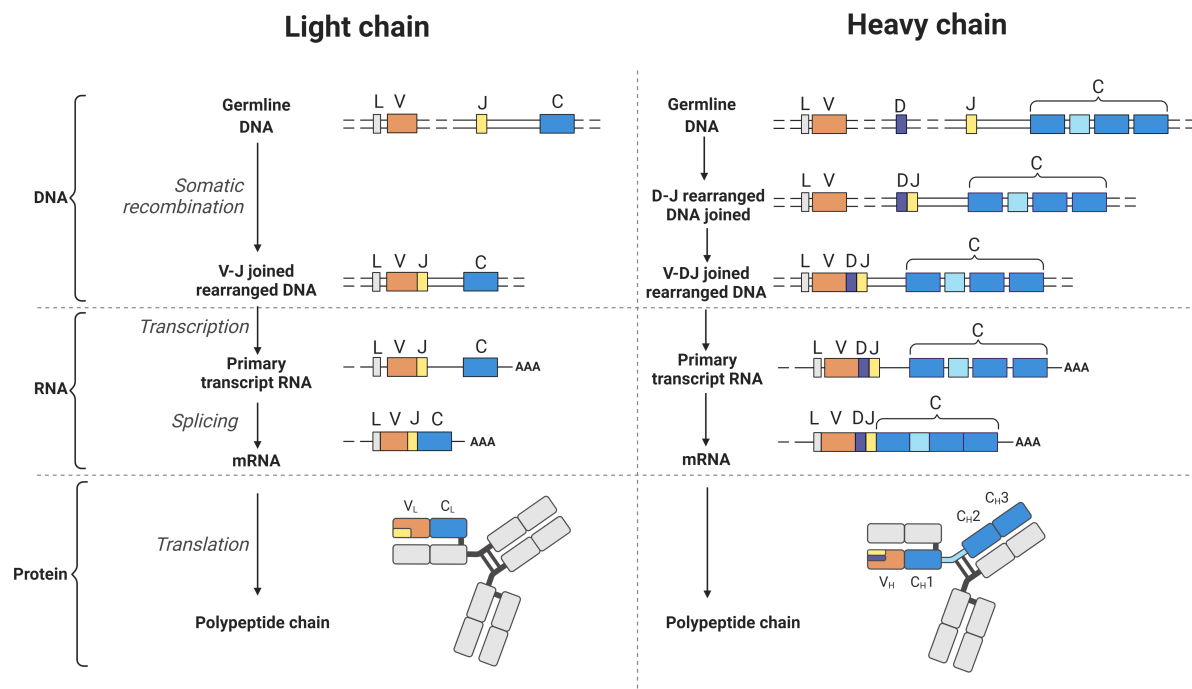
tumour necrosis factor, a substance associated with severe forms of arthritic diseases [22]. Bevacizumab, a humanised mAb cancer treatment, targets circulating VEGF which promotes rapid proliferation in cells, thus binding to this will prevent signaling via the VEGF receptor and cause a reduction in cell proliferation [23]. Rituximab, a chimeric antibody, targets the CD20 ligand on B lymphocytes in immune-related diseases such as autoimmunity, resulting in a depletion of these lymphocytes [24], trials of patients with Pemphigus vulgaris showed that B lymphocyte population dropped within 20 hours of treatment and remained low for 12 months [25]. Whilst there is a sizeable

market for these mAbs there are many downsides: they are incredibly expensive and time-consuming to produce, there are ethical considerations in producing hybridomas, and paratopes can be non-specific leading to undesired side effects [26].

#### 2.1.4. The need for computational power

As explained, the processes to produce mAbs is currently very long winded, expensive and can result in antibodies with weak affinity for the target antigen. Antigens are any molecules that elicit an immune response and therefore can be virtually any substance; usually, they are proteins, peptides, or polysaccharides that form parts of the exterior of pathogens. Due to the specificity of mature antibodies, a hugely diverse range of binding regions are necessary to bind as many antigens as possible, this is achieved through a specialised splicing process named somatic recombination. The variable region of the light chain ( $V_L$ ) is encoded by two DNA segments: the variable and joining gene segments which contain many similar domains. Additionally, the heavy chain has an extra domain named the diversity gene segment; these separate exons are rearranged or spliced and joined via RNA splicing (**Figure 2.5**) [27]. Due to this random rearrangement and selection within each segment, it is thought that there could be as many as  $10^{11}$  unique antibody molecules in a single individual [28].

Computers perform billions of operations every second, and therefore the ability to model a 3D epitope structure, find an antibody with a complementary binding site and determine the sequence of said resulting structure would lead to faster and cheaper design of higher quality mAbs. This would be incredibly advantageous in designing antibodies. According to the Structural Antibody Database [4] only 3500 antibody structures are present in the Protein Data Bank (PDB) [2], the main source of protein 3D structures. Traditional techniques such as nuclear magnetic resonance (NMR) and X-ray crystallography are laborious and expensive tasks, therefore a computational based approach could save time and money ensuring lab experiments are targeted at struc-



**Figure 2.5: Somatic recombination in heavy and light variable regions** increases the variety of paratopes for antigen binding. Variable (V) and Joining (J) segments provide variety within the V<sub>L</sub> chain, whereas the V<sub>H</sub> chain has an additional diversity segment. These segments help form the variable paratopes of immunoglobulin proteins. Made with biorender.com

tures with potential, or even solve protein 3D structures from the genetic sequence to avoid hybridoma or phage display techniques altogether.

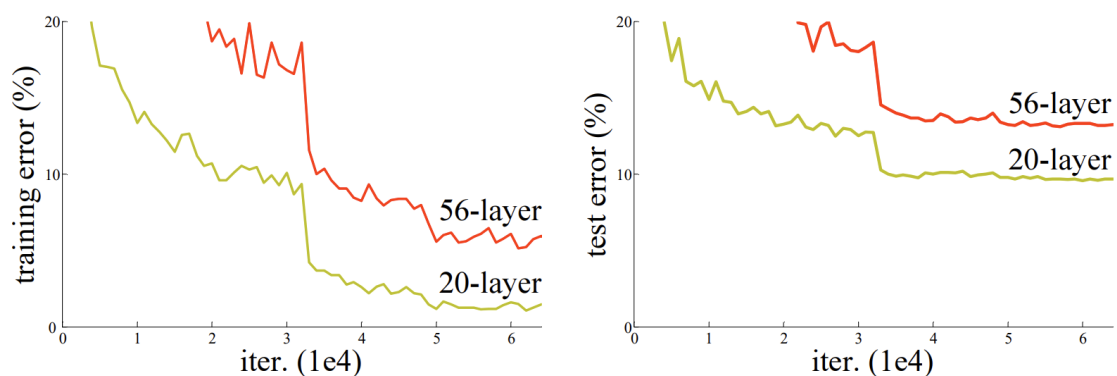
### 2.1.5. Introduction to deep learning

Machine learning is the study of computer algorithms that can improve themselves using data without human intervention. Deep learning is a subset of machine learning and applies many algorithms in layers to form a neural network (NN), like that of the human brain [29].

### Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are incredibly popular and have existed for quite some time, they are normally applied to image classification [30] and object recognition [31]. A CNN maintains the basic structure of a neural network with layers of in-

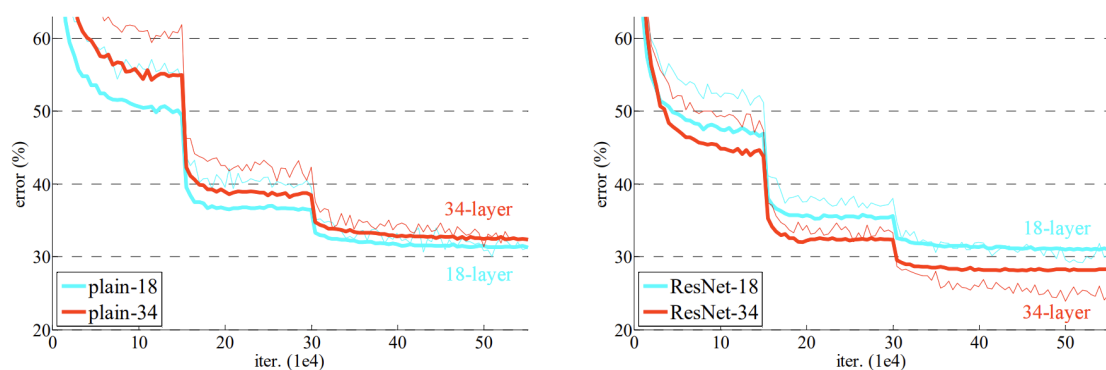
terconnected artificial neurons; but applies filters to the inputs of the NN. A filter is just a mathematical operation applied to every input in the input layer, the filter slides over each square to produce a feature map. When applied to images, early feature maps are often basic segments of objects such as edges or patches of colour. These feature maps are enriched with more convolutional layers, and more training iterations; which form more complex features describing the image, these are called high-level feature maps and can be used to classify objects in images. It is reasonable to assume that increasing the number of layers, leads to richer feature maps which lead to better object identification, which is true to an extent. More layers create higher-level features, but there is point at which the model becomes too deep and complex and proceeds to perform poorly (**Figure 2.6**)[32]. This poor performance is due to a problem known as the vanishing gradient problem [33], this relates to the mathematical function that neurons within the neural network apply to the inputs, if the inputs are too large, the gradient that updates the NN parameters can become diminishingly smaller until it has little to no effect. To solve this problem a new architecture was introduced, that forms residual neural networks (ResNets) [32].



**Figure 2.6: Training and test error of a 20 and 56 layer CNN.** training (left) and test (right) error vs iterations. Red line shows a deep CNN and the green a shallower CNN. Image sourced from (he et al. Deep Residual Learning for Image Recognition. 2016) [32]

## Residual Neural Networks

Residual neural networks (ResNets) are a sub-category of CNN which also utilise filters to automatically extract features from the provided data set during training. ResNets differ from CNNs in the structure of their networks; additional to the fully interconnected neural network there are skip connections. Skip connections connect the input of a neuron to its output, essentially allowing a neuron to be bypassed. This ensures two things: Firstly, each network layer will perform at least as well as it did on previous iterations as the data can skip this layer and remain unmodified, and this will occur naturally by chance on certain repetitions of training. If the unmodified data is the optimal output then it is selected thus preventing an irregular neuron or layer affecting output [32]. Furthermore, on each iteration of training, there is a backpropagation state, in which the parameters for each neuron are adjusted based on the difference between the output of the network data and the labelled training data. The option to travel through a skip connection on the backpropagation stage allows the gradient to jump back to previous neurons thus preventing the reduction of its value in every layer [32]. This allows ResNets to be trained to a much deeper level than CNNs (**Figure 2.7**), making them more accurate in labelling real world data with extracted features.



**Figure 2.7: Comparison of basic CNN (left) and ResNet (right) at different depths.** Plain CNN and ResNets had the same number of parameters to ensure fairness in comparison. Image sourced from (he et al. Deep Residual Learning for Image Recognition. 2016) [32]



### **Recurrent and Long Short-Term Memory Neural networks**

Recurrent neural networks (RNNs) [33] are iterative, thus a layers output loops back into the input as to consistently extract more features, they are commonly used in tasks such as speech recognition and translation. RNNs are more complex than conventional NNs, they possess an internal state, that can memorise information about previous inputs. Unfortunately, the promise of memory falls flat in practise as RNNs fail to learn over large numbers of iterations, as older inputs present in the state become vanishingly small. To tackle this problem, long short-term memory neural networks (LSTMs) were introduced [34]. LSTMs are a special kind of RNN that can learn and remember data over long periods of time. They achieve this by replacing the single repeating structure of a RNN with 4 channels, which each represent a different state. Layers range from no memory to only long term memory, and thus with numerous iterations, channels are selected by chance and different states of memory are expressed within the NN [35]. This recollection of data proves very useful in some scenarios.

### **Point Neural Networks**

Point Neural Networks (PointNets) [36] are optimised to interpret point cloud data. Point cloud refers to data that represents points in space, that contribute to a 3D structure. Unlike images which is a grid of pixels, 3D data is unstructured and therefore harder to convert into an input recognisable to a computer. PointNets utilise a spatial transformer network [37], which maintains the input's geometric features whilst applying filters similar to those in the CNN, this allows automatic feature extraction for 3D objects.

### **Difficulties**

A drawback to deep learning is the quantity of data the algorithms require to apply weights to their models and 'learn'. Although the boom in popularity of deep learning over the last few years is partly due to the recent developments in genome sequencing,

resulting in a large amount of accessible genomic data in the public domain. Another concern around more complex NNs such as CNNs, ResNets, and PointNets is overfitting. There can be multiple factors that play into overfitting but usually a complex model is trained on an insufficient volume of data, the result is a model that memorises the training data and specifically tailors its output to the training data set and thus will not perform well with real world data. There are a few different methods utilised in counteracting overfitting: Data augmentation, disturb label, weight decay, and dropout are a few of examples.

Data augmentation can diversify a training data set without the need for any additional data. The technique involves manipulating the existing data to produce slight variations of the same input, for example in image processing, rotating or cropping the images. This diversification essentially provides more training data for the NN to train from, reducing the likeliness of an overfit [38].

DisturbLabel [39] and dropout [40] are executed in a similar manner. DisturbLabel acts after each training iteration as the parameters are being updated in the neural network, by introducing random noise, and preventing the neural network from relying on specific bloated parameters. Dropout, affects neurons in the forward training route, and in the backwards readjustment phase, by randomly ignoring some neurons within each iteration, again preventing the reliance on small numbers of neurons for big changes in output. These both promote more robust features to remain in the feature maps, and reduces the chance of noise and other smaller connections being prioritised [39, 40].

Weight decay is a fairly simple method of reducing overfitting, and acts during the training phase on certain parameters that are being applied to the inputs of the hidden layers within a neural network. This technique penalises parameters that become too large, preventing a small group of parameters from dominating the output of the network [41].

All of these overfitting prevention techniques result in more iterations required during

the training stage, meaning more time spent training and higher server costs, but ensure only relevant features are being added to the feature map.

## Implementations

The research in this dissertation was partly inspired by a program named AlphaFold [42], it highlights the potential of deep learning techniques.

AlphaFold is a cutting-edge deep learning technology that predicts tertiary protein structure from its genetic sequence. The software was entered into the Critical Assessment of Protein Structure Prediction or CASP13 (2018) along with 97 other teams; Teams are ranked on many criteria across many protein predictions, but the main metric is known as the global distance test (GDT), which scores from 0-100, and counts the percentage of atoms within a threshold distance of the experimentally calculated protein structure. During the 2018 CASP13 assessment the AlphaFold team scored just under 60 in the GDT test average [42] and was recognised as the best software entered into the competition. Wanting to improve, the team resubmitted a completely re-engineered version: AlphaFold2 in CASP14 2020 and score a median of 92.4 across all targets, with their predictions having an average error (RMSD) of approximately 1.6 Angstroms [43].

Similar methods of deep learning could be applied in many ways to aid in the design of antibodies. Structural predictions can be made about the antibody and its paratope, using template models from the PDB or *ab initio* modelling [44] from just a peptide sequence. Structural predictions can also be made about the epitope on the surface of the antigen. From these structures, a docking simulation can be performed to calculate the binding affinity between antibody and antigen; then through a feedback loop with refinement this tool could accurately assign a highly specific protein structure for

a provided antigen [45].

## 2.2. About this dissertation

This dissertation aims to answer the question: "How effective is Deep Learning in therapeutic antibody development?". I will break the research into two distinct segments: 3D paratope and epitope structure prediction and paratope-antigen docking simulations. I will evaluate the varying methodologies throughout some selected literature to investigate which techniques and architectures are most effective at predicting antibody and antigen structure and their interactions. Additionally, I will highlight the limitations of certain techniques and explain how these papers can be applied in real world scenarios.

# 3

## Methods

### 3.1. Searching Literature

The PRISMA method [3], allows authors to report their findings from a literature search in a clear manner, and was utilised in this dissertation. The PubMed database (<https://pubmed.ncbi.nlm.nih.gov/>) houses over 32 million citations and abstracts of biomedical literature. I used PubMed to search for relevant literature relevant to my question: "How effective is deep learning in therapeutic antibody development.". Antibody development is a large area of research, in order to make the research more specific and easier to digest, I broke my research into two segments: Structural prediction of both the antigen and the antibody and then binding prediction between said antigen and antibody.

#### 3.1.1. Structural prediction

To identify literature applying deep learning to antigen and antibody structure I used the search terms shown in **Figure 3.1** which only produced one result. Whilst deep learning

**Antibody and antigen structure: Search terms**

(paratope structure [All Fields] OR ("CDR" [Title] AND "structure" [All Fields]) OR "antibody structure" [All Fields] OR "antibody paratope" [All Fields] OR "antigen-structure" [All Fields] OR "epitope structure" [Title]) AND "deep-learning" [All Fields]

**Figure 3.1:** The search terms applied to the PubMed database to identify relevant literature associated with antibody / paratope and antigen / epitope structure.

is not a novel technology with the first deep learning network originating in 1965 [46]. It has only recently gained popularity due to advancements in computing power and learning techniques, reinforcing it as a viable method of computational learning. This is reflected in the lack of literature specific to antibody and antigen structure prediction, a niche topic that is yet to be fully explored. In order to gain a broader set of literature for analysis, I performed a secondary search to find literature associated with protein structure prediction (**Figure 3.2**). Antibodies are a form of complex protein, and antigens are very commonly proteins, thus literature on protein structure prediction is applicable to both antibody and antigen structure prediction. I searched with the following search terms:

**Protein structure: Search terms**

('deep-learning'[Title] AND 'protein'[Title] AND 'structure'[Title])

**Figure 3.2:** The search terms applied to PubMed to expand literature associated with antibody and antigen structure prediction.

These terms produced 22 pieces of literature, totalling 23 results. I excluded any non-primary literature, which includes: reviews, systematic reviews, meta-analysis, and editorials. Then full-text articles were assessed for their eligibility in this dissertation.

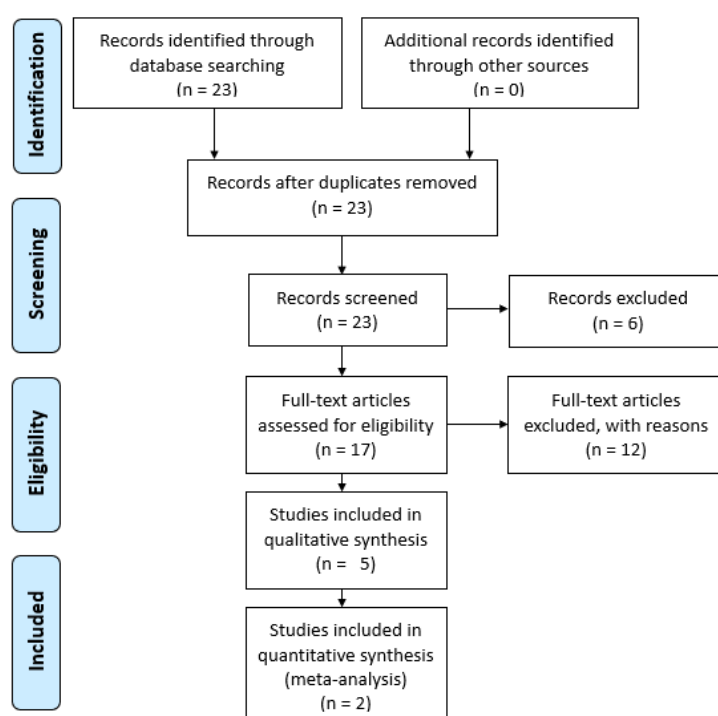
**Reasons for exclusion**

Firstly, any literature published prior to 2019 was excluded; rapid developments

of computer technologies cause newer deep learning models to rapidly outcompete older programs. This is in tandem with Moore's law [47] which states the number of transistors on integrated circuit boards doubles every year, resulting in a direct increase of processing power. Such advancements of computing power each year on top of novel deep learning techniques, renders older models much less effective and therefore were excluded from the review.

Additionally, only literature proposing deep learning programs that aimed to predict structure of the paratope or epitope from the proteins primary amino acid sequence were included. This allows a more accurate representation of how deep learnings could be implemented with other technologies to predict the proteins structure; whilst still modelling the protein structures *ab initio*.

After applying my exclusion criteria, there were five relevant papers remaining (**Figure 3.3**).



**Figure 3.3: PRISMA flow diagram** selection process for antibody + antigen structure literature. Searches used (**Figure 3.2**) and (**Figure 3.1**)

### 3.1.2. Antibody-Antigen binding predictions

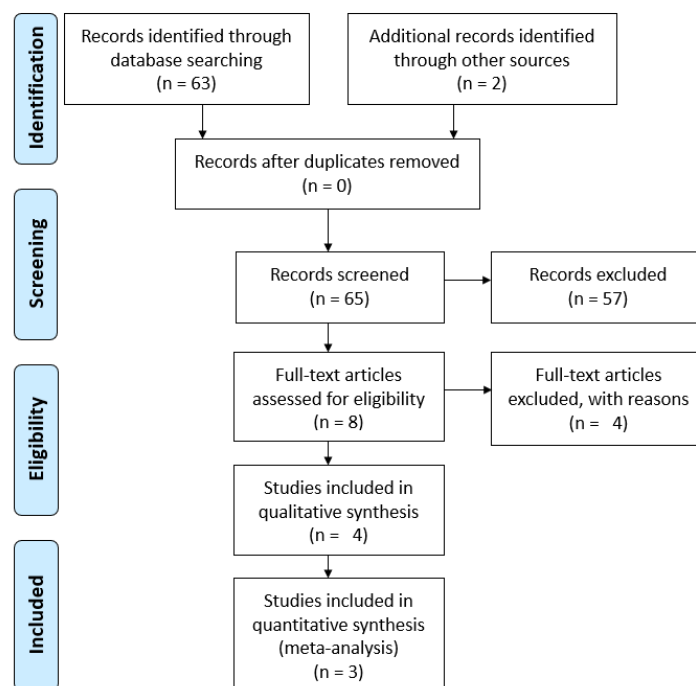
#### Protein-Protein interaction: Search terms

(antibody-antigen[Title] OR paratope-epitope[Title] or protein-protein-interactions or PPI) AND (deep learning)

**Figure 3.4:** The search terms applied to PubMed to expand literature associated with antibody and antigen binding prediction.

I began my search trying to identify models that predict antibody-antigen binding from their primary sequence, but I was unsuccessful. Again this is probably due to a lack of research specific to antibody-antigen interaction prediction. Fortunately, general protein-protein interactions have been researched a fair amount; and as antibody-antigen interactions often represent protein protein interactions, research into this area would be suitable to evaluate deep learnings ability in predicting antibody-antigen binding. My search terms produced 81 results **Figure 3.4**. Once I had excluded any non-primary literature as seen previously, there were 62 articles ready to be reviewed for eligibility (**Figure 3.5**).





**Figure 3.5: PRISMA flow diagram** selection process for antibody-antigen interaction literature.

I excluded any papers prior to 2018 and papers that did not predict binding interactions via contact prediction were excluded. Additionally, papers associated with MHC and TCR binding are not relevant to the binding interactions of antibodies, thus these papers were also excluded.

The literature selected for analysis is shown in (**Figure 3.1**).

**Table 3.1:** Deep learning programs from papers discovered in the literature search.

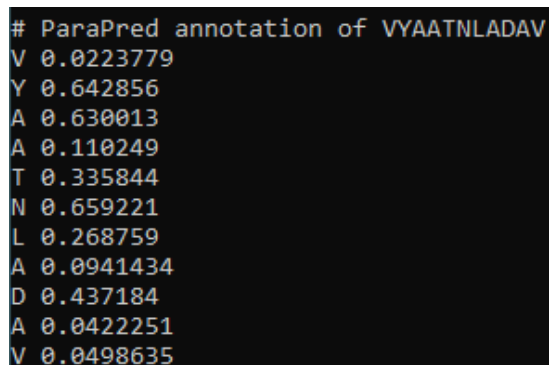
Structure Prediction	Interaction Prediction
AlphaFold [42]	DeepInterface [48]
DeepH3 [49]	DNN-PPI [50]
DeepAccNet [51]	DPPI [52]
MULTICOM [53]	ParaPred [54]
RaptorX [55]	PINet [56]

### 3.1.3. CASP13 comparison

I extracted global distance test total scores (GDT\_TS) and GDT high accuracy scores (GDT\_HA) from CASP13's official website [predictioncenter.org/casp13](http://predictioncenter.org/casp13), for groups 43 (AlphaFold), 89 (MULTICOM), and 324 (RaptorX). I separated the results with the following categories: template based modelling, free modelling, and contact prediction. I performed a one-way anova to test if any of the models performed significantly better than their counterparts for all three categories.

### 3.1.4. Testing for an overfit

To test for an overfit of the ParaPred model [54] I curated a test set of antibody-antigen structures, 3 from within the ParaPred training set, and 3 from the SAbDab [4] that were not in the ParaPred training set. The combined dataset totalled 476 residues from 36 CDRs. I entered the primary sequences of these antibodies to the ParaPred program, which extracted CDR sequences using the Cothia numbering scheme [57] and that produced a list of outputs between 1 and 0 for each residue representing bound or unbound from the antigen respectively (**Figure 3.6**).

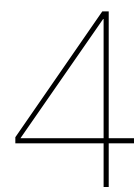


```
# ParaPred annotation of VYAATNLADAV
V 0.0223779
Y 0.642856
A 0.630013
A 0.110249
T 0.335844
N 0.659221
L 0.268759
A 0.0941434
D 0.437184
A 0.0422251
V 0.0498635
```

**Figure 3.6: Example output from the ParaPred program [54].** Each value represents a binding prediction between 1 and 0 for the associated residue to its left.

ParaPreds curators treated residues as bound if an atom belonging to a residue is within 4.5Å of any part of the antigen[54]. To confirm which residues I should treat

as bound for comparison to the ParaPred output, I used python and a library called BioPython [58], my code is available here: <https://github.com/mah51/determine-bound-residues>. I calculated the absolute error between the predicted value outputted from ParaPred and the actual bound value determined by my program, for each residue. I performed a Welch t-test using this absolute error to determine if the difference in means between the structures from within ParaPred's training set and those that were not, was significant.



## Results

### 4.1. Overview

Neural Network architecture plays a significant role in the performance of a NN. The training period requires a balance between the incorporation of new information, learning and retention of old information, memory. Various architectures train themselves in a variety of ways and thus process training data very differently. Additionally, inputs fed into a NN as well as the quality of the training set can greatly impact its efficacy in making accurate predictions. **Table 4.1** shows the performance of each program from each article.

#### 4.1.1. Residual Neural Networks

Many of the selected literature provided programs that utilised a residual network (ResNet) architecture [42, 49, 51, 53–55] in some part of the program's structure. This is most likely due to ResNet's ability to apply a deeper neural network without reducing prediction accuracy, whilst still maintaining convolutional layers for automatic feature extrac-

	Performance
RaptorX	18 out of 32 domains (TM-score > 0.5) in CASP13 free modelling
AlphaFold	Achieved a summed z-score of 52.8 in the CASP13 free modelling category, with TM-score > 0.7 in 24 out of 43 domains
DeepH3	Pearsons coefficient of 0.87 and 0.79 for distance and phi values, and circular correlation coefficients of 0.52 and 0.88 for omega and theta dihedrals.
DeepAccNet	Average GDT_TS of 58.66 on CASP14 dataset free modelling
ParaPred	F-Score 0.690 on validation dataset
DeepInterface	75% accuracy and 61% precision on an unknown benchmark dataset
DNN-PPI	98.78% accuracy on held out test dataset
DPPI	94.55% accuracy on the PPI dataset
MULTICOM	17 out of 31 domains (TM-score > 0.5) in CASP13 free modelling category
Pinet	53.8% precision in DBD5 benchmark dataset

**Table 4.1: Performance of each NN within the selected literature.** Different metrics selected to indicate performance over multiple test sets makes analysis fairly difficult.

tion.

### RaptorX [55]

RaptorX uses a very deep ResNet to predict the Pythagorean distances between pairs  $C_\beta$  atoms from two distinct residues. The NN at its core aims to predict distances between  $C_\beta$  of pairs of residues. A multiple sequence alignment (MSA) is produced using homologous proteins, which provides co-evolutional data. Residues that mutate together in a similar time frame indicate close proximity in space. The creators of RaptorX employed two separate ResNets, one that accepts 2D matrix inputs, and another that accepts 3D matrices, utilising 7 and 60 convolutional layers respectively. Distances between  $C_\beta$  are placed into separate categories for classification. The sys-

tem was trained using 11,410 non-redundant proteins, with 900 forming the validation set.

### **AlphaFold [42]**

At its core AlphaFold has a deep two-dimensional residual network. The fundamentals that form this residual network originate from the RaptorX predecessor [59], thus there are many structural similarities.

Similar to RaptorX, AlphaFold organises spatial distances into discrete categories for classification and whilst the network architectures vary, they receive similar inputs such as co-evolutional MSA inputs and secondary structure. The heart of the inter-residue distance prediction is a deep two-dimensional dilated ResNet. The network is formed of 220 residual layers, with dilated convolutions; which produce dense predictions [60]. AlphaFold is trying to predict inter-residue distance between every residue within the structure, dissimilar from image recognition where an object is being recognised and annotated, thus dilated convolutions are used to provide better prediction capabilities for the given input. The NN's complexity is enabled by the ResNet structure and their vast training set of 31,247 protein domains sourced from the PDB [2]. Furthermore, AlphaFold employ a folding network that aims to fold the proteins by minimising the gradient descent, allowing them to minimise folding error.

### **DeepH3 [49]**

DeepH3 uses a deep residual network of 1D and 2D convolutions, to structurally model the H3 CDR loop. The NN predicts distances between  $C_{\beta}$  on separate residues, by taking samples of pairs within the loop. DeepH3 only utilises the protein sequence with no other data, MSAs etc. and still obtained valuable predictions. Unfortunately DeepH3 only reported the Pearson correlation and circular correlation coefficients (**Figure 4.1**), again making a comparison to other programs difficult.

### **DeepAccNet [51]**

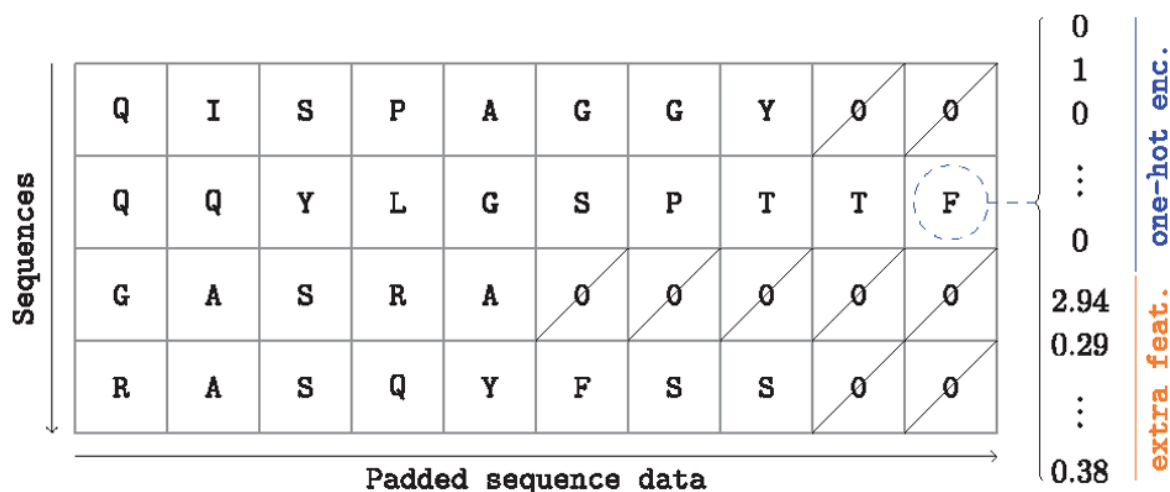
DeepAccNet takes inputs of atomic coordinates that are processed in convolution layers. Which is then flattened and passed through a ResNet with 24 layers. The model was trained on 8,718 protein chains, whilst this is a large dataset for a shallow network, the authors did not mention any overfitting prevention techniques or tests. However, a test data set was correctly applied, which mitigates this concern.

### **ParaPred [54]**

The ParaPred deep learning model utilises an antibody primary sequence, from which it extracts the CDR sequences using the Chothia numbering scheme [57] with an additional two residues on each side of the CDR as these are known to be involved with binding [61]. It then outputs a prediction for each residue between 1 and 0, depending on whether they are predicted to bind or not bind respectively. This model is composed of a CNN that feeds into a recurrent neural network (RNN), it also includes some ResNet properties, including shortcut connections within the RNN to allow a more complex model. The CNN processes amino acid sequences via a 3D input matrix with each sequence constituting a new row, and the max sequence length setting the number of columns (**Figure 4.1**), on the z-axis there are 21 layers that account for 20 amino acid types + 1 unknown type. A binary 1 or 0 in any of these slots will represent what amino acid is in the respective cell.

### **Potential Overfit**

Whilst reviewing the ParaPred paper, one of my initial concerns was the small dataset compiled from the SAbDab [4], a database of antibody-antigen crystal structures compiled from the protein data bank [2] a larger database of various protein structures. ParaPred's dataset only consisted of 277 antibody-antigen complexes. Additionally, the validation set was used to tune the parameters of the network, and then the NN was evaluated with this validation set instead of a separate test set. The retuning



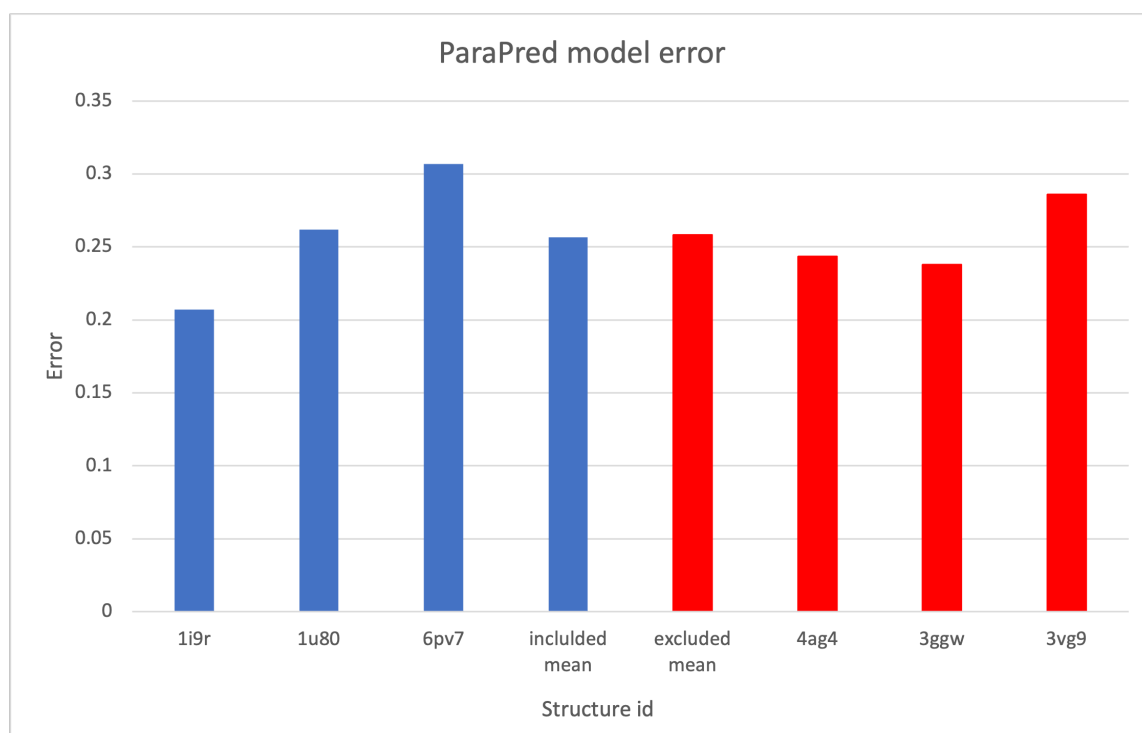
**Figure 4.1: ParaPred CNN input matrix.** Taken from ParaPred’s article [54]. Each row represents a new CDR sequence and the max number of columns is set by the longest sequence, with shorter sequences being padded by empty 0 values.

may have introduced a bias, meaning their results are potentially unreliable.

I tested ParaPred for an overfit and found no significant difference in prediction error of a test set of three antibody-antigen complexes that ParaPred was trained on and three that the model had never seen. (**Figure 4.2**) shows the model’s prediction error for each structure, and after performing a Welch t-test found that the difference in means was not significant.

ParaPred had not overfitted the data as I originally expected. Whilst this was surprising due to the small data set used to train a complex model, there are a few techniques employed to prevent overfitting. Firstly ParaPred is trained using 10-fold cross-validation; this method trains the model 10 times from scratch with 10 varying test sets and training sets, aiming to highlight if a model is overfit. Additionally, the model is regularized via weight decay and dropout [40]. On each training loop of the RNN, random inputs are removed from neurons in the network, to prevent the model relying on one particular set of inputs. Weight decay targets parameters that have grown too large during training and reduces their effect, essentially preventing aggressive output influence from small areas of the network. These factors combined, cleverly helped negotiate the issue of





**Figure 4.2: ParaPred overfitting test results.** Three blue bars represent structures 1i9r, 1u80, 6pv7 (PDB ids) that were not part of the training set. Three red bars represent structures 4ag4, 3ggw, 3vg (PDB ids) that were part of ParaPred’s training set. Mean of the errors shown as the two centre bars.

overfitting in creating a complex model that was trained using such a small dataset.

### 4.1.2. Convolutional Neural Networks

NNs that were created with a CNN architecture were limited in depth and complexity, but in many cases this was to prevent overfitting specifically in the interaction prediction literature [48], due to the lack of antigen-antibody complexes available for use as training data. ParaPred [54] utilised a CNN to analyse local interactions between

#### DeepInterface [48]

DeepInterface aims to classify proteins that are in a docked state and determine whether or not that is a natural bound conformation. Protein structures that are to be inputted to the model are presented as voxels, a voxel to a 3D structure is essentially what a pixel is to an image, it represents a point in 3D space. Protein complexes were collected from a variety of sources, and to ensure that there were some proteins

that were not bound with the correct residues contacting the authors used an algorithm named ZDOCK. This is a vital step, as without these 'negative' structures, the NN would not be able to differentiate between correctly bound and incorrectly bound structures. DeepInterface was trained on an enormous 271,830 separate interfaces, for such a shallow network this training set is incredibly large and hints at why the NN's performance exceeded expectations, achieving 75% accuracy on a test dataset. However, there was a possible overfit as the model did achieve a better accuracy of 92% and 88% in the training and validation sets respectively.

### **DNN-PPI [50]**

DNN-PPI utilised primary protein sequence without any additional information, the sequence was passed into a CNN and then to a long short-term memory NN. Over 70,000 structures were extracted for training with 8000 being used as a validation set (N.B. in the article the authors mislabel this as a testing set, but the testing set is the mus musculus structures). There are an additional 22,870 structures used as a test set. Additionally, DNN-PPI implement good techniques to prevent overfitting such as 5-cross fold validation. On the test set DNN-PPI scored a prediction accuracy of 92.43% which is very high, and likely due to their significant dataset. Additionally, the selection of a LSTM network appears to be an important feature. To maximise the value of the dataset, some features of previously seen inputs must be retained so they can be combined with the current sequence of the data that the model is processing, which the LSTM network is excellent at, likely contributing to the high accuracy of this NN.

### **DPPI [52]**

DPPI, like DeepH3 utilises minimal input data including the primary protein sequence and a PSI-BLAST search. It aims to predict antigen The core architecture is formed of a Siamese-like CNN, which are primed to learn relationships between two entities. Interestingly DPPI is trained using an unsupervised data set, that is unlabelled;

this allows the model to train itself without having to source labelled data, but can lead to lower quality predictions. Unlike AlphaFold and RaptorX who use classification techniques using categories for inter-residue distances, DPPI uses a regression model, which can predict a continuous output for residues between the proteins. 10-fold cross validation is employed like that with ParaPred. Even though the NN was tested on 'benchmark' datasets, the model was trained using these datasets with validation sets being used as a make shift test set, with the lack of a true test set; results are likely biased, therefore their high accuracy of 94.55 percent could be unreliable.

### **MULTICOM [53]**

MULTICOM [62] like other neural networks utilised co-evolution methods, by generating MSAs with homologous proteins and predict secondary structure of the protein sequence which is also fed into the NN. MULTICOM employs a remarkably shallow NN to process all of this information, called DNCON2 described as a separate NN [63]. This CNN utilises 7 convolutional layers for feature extraction in comparison to AlphaFold's 220, and is trained over 1600 epochs in comparison to AlphaFold's 500 epochs, one epoch refers to a single training loop where all training data has passed through the NN. The curators attempted to test up to 9 convolutional layers, but experienced problems caused by overloading the graphics processing unit (GPU) being used to run the NN. No overfitting preventative measures were taken, but overfitting was unlikely to be prevalent on such a shallow network [64]. Additionally the DNCON2 model was only trained with 1230 proteins from previous CASP assessments.

### **Point Neural Networks**

Plnet was the only NN to use a PointNet architecture, it utilises pairs of point clouds of two protein structures to predict interfacing regions. After the local features were extracted using the PointNet, global features were extracted by using pooling together local features, this quickly and with comparatively little computation provides a summary

of the proteins' surfaces. The authors implemented a data augmentation technique to prevent overfitting while training on small datasets, this was achieved by randomly rotating certain structures in the training data, producing slightly altered duplicates as further training data. Again it was unclear whether the test dataset remained excluded until parameters were tweaked, therefore the results may not be reliable. PInet did achieve a fairly high AUC-ROC score of 0.867, but a low precision of 51.8%. Whilst this NN was designed for general protein-protein interactions it was trained with a dataset consisting of antibody-antigen complexes, it scored slightly worse in this dataset potentially suggesting antibody-antigen interactions are more complicated. However, this dataset was significantly smaller than the other training datasets, which could have affected the performance instead.

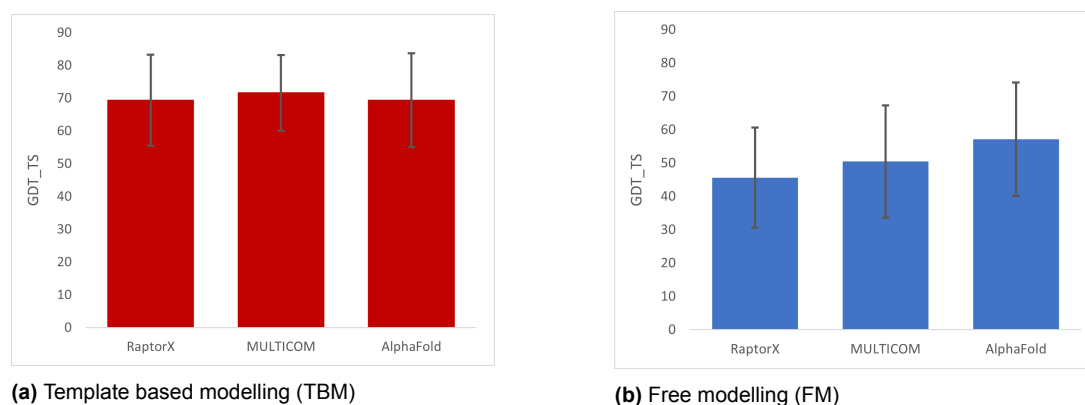
#### **4.1.3. Comparison using CASP13 data**

Three out of the five papers relating to protein structure prediction described models that were entered into the 13th Critical Assessment of Protein Structure Prediction (CASP13): AlphaFold, MULTICOM, and RaptorX [53, 55, 65]. CASP13 is an independent accessible competition to identify computational programs that produce the best protein structure predictions. The competitors are presented with the amino acid sequences of target proteins, from which they have to determine the proteins tertiary and quaternary structure. Participants submit five of models towards each of the 84 protein structures.

CASP13 provides two main categories for assessment: free modelling (FM) and template based modelling (TBM). Structures fall into the TBM category if part of their structure is available on the public domain via a sequence search, meanwhile FM proteins have no identifiable segments available; thus the computational program has to predict the structure without assistance.

Neural networks are ranked by a global distance test total score (GD\_TS). The GDT\_TS provides a more accurate measurement of similarity between two protein structures, in

this context, the modelled structure vs experimentally determined structure; allowing the NN accuracies to be determined. It represents a percentage of residue's alpha carbons falling within a certain cut-off distance of the experimentally determined structure, thus a score must be between 0 and 100. CASP13 produces a GDT\_TS average of 4 GDT with different cut-off values of 1,2,4, and 8 Å, forming the final GDT\_TS.

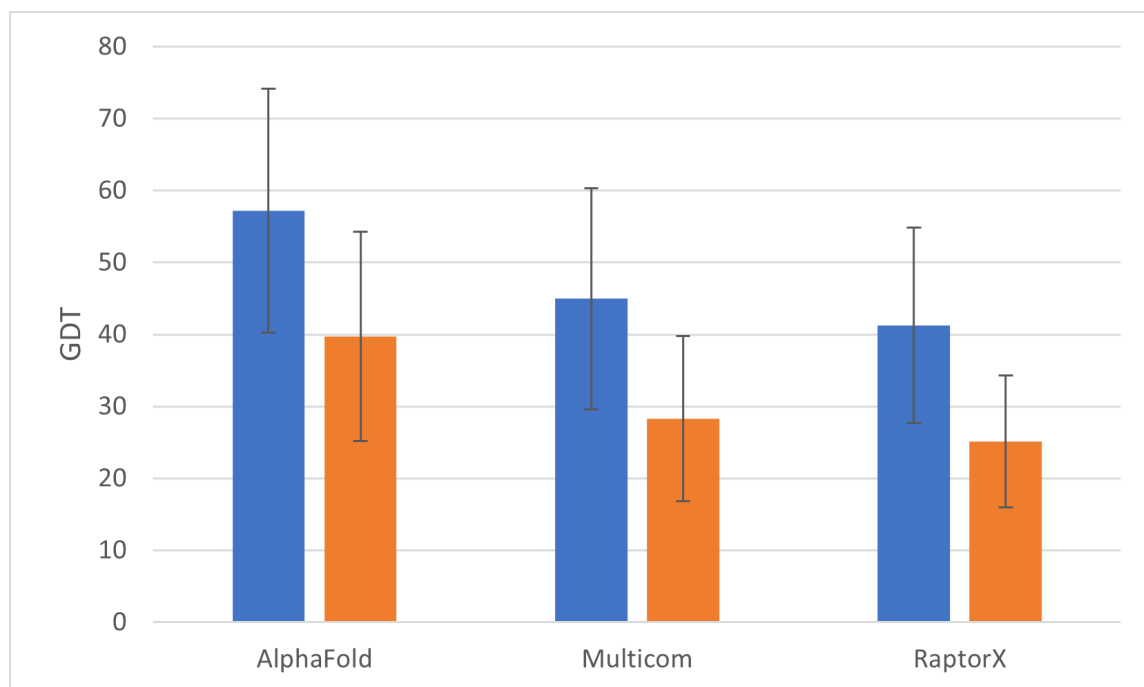


**Figure 4.3: Performance of three deep learning neural networks in CASP13.** GDT total score shown on y-axis for a. Template based models, and b. free models. The error bars represent the standard deviation.

**Figure 4.3** shows the data available from the CASP13 website ([predictioncenter.org/casp13](http://predictioncenter.org/casp13)). In the template based modelling dataset (**Figure 4.3a**) MULTICOM performs slightly better than both Raptor and AlphaFold, but this does not represent a significant difference. In the free modelling category (**Figure 4.3b**) AlphaFold performs significantly better than RaptorX, and slightly better than MULTICOM, but not a significant amount. It is important to note, that whilst RaptorX does not perform as well as MULTICOM and AlphaFold in the results shown here; the model performed significantly better in a separate contact prediction category (**Figure 4.5**).

Models perform remarkably well in the TBM category and fairly well in the FM category. FM represents a harder but more realistic challenge, as many protein structures are still unsolved. Whilst these seem promising, it is thought that a different metric known as GDT high accuracy (GDT\_HA) performs better in the assessment of structural accuracy producing "more stringent" results, that promote structures useful in drug discovery [66]. GDT\_HA is calculated in the same manner as GDT\_TS, but different

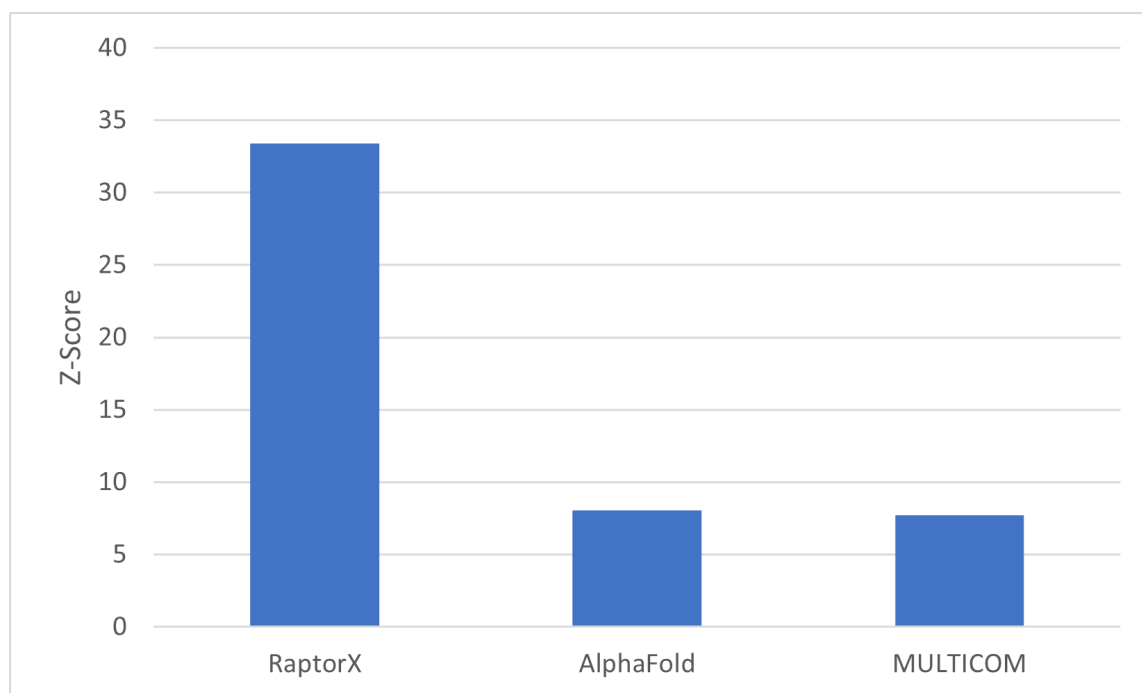
cut-off values are incorporated of 0.5, 1, 2, 4 Å, half the distance of the cut-off points used in GDT\_TS.



**Figure 4.4: Three analysed deep learning models: AlphaFold, MULTICOM, RaptorX [42, 53, 55] performance in CASP13.** Global distance test total score (GDT\_TS) shown in blue (left columns) and GDT high accuracy (GDT\_HA) shown in orange columns (right columns). Error bars represent the standard deviation.

**Figure 4.4** shows the free modelling values for GDT\_TS (blue) and GDT\_HA (orange), as expected GDT\_HA values are significantly smaller. Values here show a more accurate representation of the computer model's ability in predicting protein structure.

Another category, known as contact prediction involves predicting which residues in a protein are contacting one another; this data is normally incorporated into free modelling programs to provide more data for better identification of possible homologues. In the CASP13 assessment, residues were determined as bound if two  $C_\beta$  (or  $C_\alpha$  in the case of glycine) atoms were within 8Å of each other. Participants were judged on their Z-score which incorporates the expected and obtained values. **Figure 4.5** shows RaptorX's performance compared to MULTICOM and AlphaFold.



**Figure 4.5: RaptorX, MULTICOM, and AlphaFold performance in CASP13, in the contact prediction category.** Z-score provided for each of the three compared programs.

#### 4.1.4. Identifying performance factors

My analysis will be focused around the performance in the FM category as this is thought to be the 'harder' of the protein categories to predict, as the computer models are having to predict the full protein structure *ab initio*.

AlphaFold [42] performed the best out of the three models, outperforming the nearest competitor by 40% in the more stringent GDT\_HA assessment. The amino acid sequence is applied using co-evolution methods such as multiple sequence alignments (MSA) with homologous proteins, and residues that mutate together over time are identified. They are likely a result of one residue mutating and the other following to maintain protein structure, thus suggesting a close proximity to one another. It also accepts, protein secondary structure predicted from the SST web server [67]. Results are outputted for every pair of residues in a sequence: distance and psi and phi torsion angles are produced. AlphaFold utilise a very deep 220 block ResNet which is trained over 500

epochs. To prevent overfitting within such a complex network the NN is trained using 31,247 non-redundant protein domains, extracted from the protein data bank (PDB) [2]. Additionally, the creators employed both data augmentation and dropout techniques to reduce the prevalence of overfitting, increasing the viability of such a deep network. Additionally, to a complex network, dilated convolutions [60] are used, which can span a larger range than traditional convolution levels, allowing interactions across the whole protein sequence to be analysed.

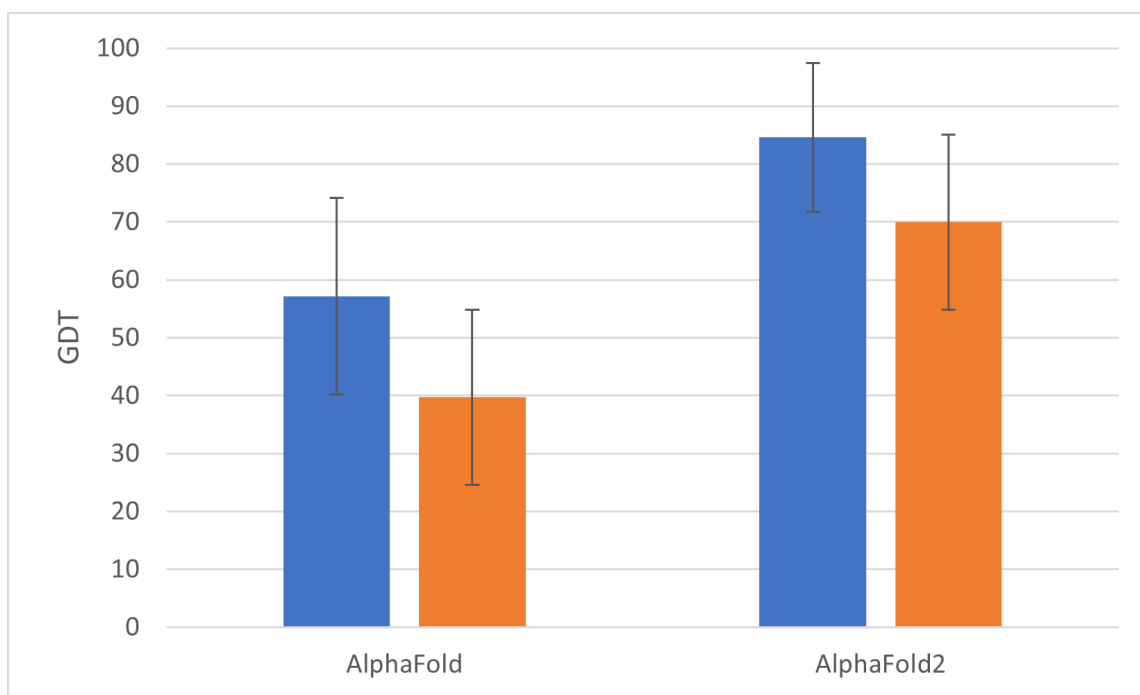
MULTICOM has come under criticism from other teams [55], due to their unclear methodology and surprising performance considering the shallow network utilised. Whilst it is speculative, suggestions were made that MULTICOM "heavily relied on consensus analysis" [55] rather than producing a high performance neural network. Potentially providing an explanation for the surprising performance of such a shallow network.

RaptorX [55] initially process input features with a 1D ResNet, with a depth of 7 convolutional layers. They produce a MSA using HHBlits like that utilised by AlphaFold and MULTICOM [68], additional inputs include sequence profile, secondary structure, and solvent accessibility. The overall network consists of the 1D ResNet of 7 layers and a 2D ResNet of 60 convolutional layers trained with 10,000 proteins from the PDB. RaptorX massively outcompetes both AlphaFold and MULTICOM in the contact prediction category (**Figure 4.5**), but performed significantly worse than AlphaFold in the FM category; potentially indicating a deficiency in their protein folding methodology when compared to AlphaFold.

#### 4.1.5. AlphaFold2

CASP14 took place at the beginning of 2020, and the articles have not yet been released for many of the models or the results; but a notable mention is the successor to AlphaFold, AlphaFold2 [43]. AlphaFold performed significantly better in the free modelling category with both GDT\_TS and GDT\_HA scores. Whilst the article is not yet





**Figure 4.6: Comparison between AlphaFold [42] and AlphaFold 2 entered into CASP14.** Blue columns represent GDT\_TS and orange columns denote GDT\_HA scores. Standard deviation is represented by error bars.

available for review, there is a blog post available providing some detail on how the team achieved such a result (<https://deepmind.com/blog>). This progress is inspiring, and shows how rapid the advancements in deep learning are.

# 5

## Discussion

The literature analysed in this dissertation displays that deep learning has useful applications in both structural prediction and binding prediction of antigens and antibodies. Whilst applications are not yet accurate enough for real world use, rapid advancements are being made each year.

Of those papers that could be analysed thanks to the public CASP13 data, AlphaFold performed the best in the free modelling category, even though RaptorX massively outperformed both MULTICOM and AlphaFold in contact prediction. This is suggestive that AlphaFold has a much superior folding architecture than RaptorX. There could be a few factors playing into AlphaFold's success: The dataset used to train AlphaFold is much larger than that of RaptorX and MULTICOM. Additionally, the depth of the AlphaFold ResNet, allows for much more complex feature maps to be generated; thus enabling better classifications.

Features implemented such as MSA construction in all three papers suggests the resulting spatial data is of importance in producing high quality predictions. Convolutions are used in every paper, which enable the automatic feature extraction in the deep

learning programs. ResNets were the most popular architecture within my papers, which makes sense due to their use in mapping complex relationships; although in cases where training data was scarce, more simple CNNs were chosen. DNN-PPI showed an excellent use of the LSTM neural network, which was made viable by their large training datasets. The LSTMs retention of data most likely aided in their impressive accuracy on the test dataset of 92.43%.

## 5.1. Difficulties faced

There were numerous difficulties while trying to analyse the selected literature, this highlighted some shortcomings of a few articles.

Cross analysis of the literature proved difficult. Firstly, authors reported on their programs with various metrics (**Figure 4.1**); standardised testing units to confirm the accuracy of protein structure prediction and protein protein interactions would enable a direct comparison to other NNs. Additionally, test datasets are not standardised and sometimes not utilised at all. Even with the performance being reported in a comprehensive manner, programs would not be able to be reliably compared if the models are not tested on identical datasets. This is due to various structures being easier or harder to model depending on their size, existing protein domains and other factors. Furthermore, the unclear dialect throughout some articles caused difficulties in understanding testing and training protocols; a recurring example is the use of 'validation' interchangeably with 'test' in reference to datasets. A validation dataset is used to manually tune the parameters of a NN during its development, allowing the creator to determine optimal values for dropout, learning rate, weight decay etc. A test set is used to evaluate the performance of a NN, and this should be used to provide valuable metrics such as accuracy and precision. Once the model has been evaluated with the test set it should not be tuned, this is to prevent bias towards the test data set. This often caused confusion whilst reviewing the literature, as it was hard to determine if the test set had been

used as a validation set, and if an unbiased assessment of the network was being presented. Adopting a standardised nomenclature consistent throughout literature would ensure clarity.

As shown by AlphaFold, DeepInterface, and RaptorX, large datasets for training produce significantly better results. A well known disadvantage of deep learning is the volume of data required to train and test the resulting networks. Whilst there is a large amount of structural data available for proteins, antibodies only make up 1.75% of the 176,773 protein structures available on the PDB [2], only 8,000 constitute full or segments of antibodies. Therefore, specialising networks for antibody structure prediction remains a challenge.

### 5.1.1. Improvements

The code for most of the programs analysed is publicly available, and ideally with more time I would have liked to curate two large test datasets which could be used for the protein structure prediction and protein-protein interaction prediction networks. This would allow me to compare all of the NNs with the same dataset, which would help highlight the most useful programs based on their accuracy, consistency, and how easy their output is to interpret, as well as how much information the NN actually produces about the target antibody or antigen.

## 5.2. Concluding remarks

Whilst deep learning is not accurate enough in its current state to replace traditional antibody selection techniques, an incorporation between the two could be useful to narrow down search targets or produce more specialised phage display libraries for a higher chance of obtaining an antibody with a higher affinity for the target antigen. Furthermore, binding affinity could be analysed initially with a deep learning program

to give a rough estimate of the antigen interface.

This dissertation has shown deep learning is already fairly successful in predicting protein structure, and is rapidly improving. Whilst antibody and antigen specific structure remains untapped, protein structure prediction provides a good indication of how well deep learning will tackle these tasks. However, more complex antigens not of protein descent will require further research to investigate respective structures. Thankfully mAbs are normally utilised in targeting various proteins for example membrane proteins that can affect signaling within the cell and therefore influence biological processes. Additionally, there still remains a potential for predicting specific interactions between paratope and epitope structures, as PPI networks prove to be fairly accurate. Hopefully more research and developments like those made by the AlphaFold team, will aid the production of therapeutic antibodies helping to optimise binding activity and stability. Therefore, I believe deep learning is already a useful technology in therapeutic antibody design, with plenty more potential.

# References

- [1] Kryshchak A, Schwede T, Topf M, Fidelis K, Mout R. Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics*. 2019;87(12):1011–1020.
- [2] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Research*. 2000 Jan;28(1):235–242.
- [3] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *Journal of Clinical Epidemiology*. 2021 Mar;0(0).
- [4] Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, et al. SAbDab: The Structural Antibody Database. *Nucleic Acids Research*. 2014 Jan;42(D1):D1140–D1146.
- [5] Edelman GM. Antibody Structure and Molecular Immunology. *Science*. 1973;180(4088):830–840.
- [6] Chen X, Li R, Pan Z, Qian C, Yang Y, You R, et al. Human Monoclonal Antibodies Block the Binding of SARS-CoV-2 Spike Protein to Angiotensin Converting Enzyme 2 Receptor. *Cellular & Molecular Immunology*. 2020 Jun;17(6):647–649.
- [7] Forthal DN. Functions of Antibodies. *Microbiology Spectrum*. 2014 Aug;2(4).
- [8] Grilo AL, Mantalaris A. The Increasingly Human and Profitable Monoclonal Antibody Market. *Trends in Biotechnology*. 2019 Jan;37(1):9–16.
- [9] Ecker DM, Jones SD, Levine HL. The Therapeutic Monoclonal Antibody Market. *mAbs*. 2014 Dec;7(1):9–14.
- [10] Köhler G, Milstein C. Continuous Cultures of Fused Cells Secreting Antibody of Predefined Specificity. *Nature*. 1975 Aug;256(5517):495–497.
- [11] Breitfeld D, Ohl L, Kremmer E, Ellwart J, Sallusto F, Lipp M, et al. Follicular B Helper T Cells Express Cxc Chemokine Receptor 5, Localize to B Cell Follicles, and Support Immunoglobulin Production. *The Journal of Experimental Medicine*. 2000 Dec;192(11):1545–1552.
- [12] Stone SR, Montgomery JA, Morrison JF. Inhibition of Dihydrofolate Reductase from Bacterial and Vertebrate Sources by Folate, Aminopterin, Methotrexate and Their 5-Deaza Analogues. *Biochemical Pharmacology*. 1984 Jan;33(2):175–179.
- [13] Hosono M, Endo K, Sakahara H, Watanabe Y, Saga T, Nakai T, et al. Human/Mouse Chimeric Antibodies Show Low Reactivity with Human Anti-Murine Antibodies (HAMA). *British Journal of Cancer*. 1992 Feb;65(2):197–200.

- [14] Legouffe E, Liautard J, Gaillard JP, Rossi JF, Wijdenes J, Bataille R, et al. Human Anti-Mouse Antibody Response to the Injection of Murine Monoclonal Antibodies against IL-6. *Clinical & Experimental Immunology*. 1994;98(2):323–329.
- [15] Hwang WYK, Foote J. Immunogenicity of Engineered Antibodies. *Methods*. 2005 May;36(1):3–10.
- [16] Harding FA, Stickler MM, Razo J, DuBridge RB. The Immunogenicity of Humanized and Fully Human Antibodies. *mAbs*. 2010;2(3):256–265.
- [17] Lonberg N, Taylor LD, Harding FA, Trounstine M, Higgins KM, Schramm SR, et al. Antigen-Specific Human Antibodies from Mice Comprising Four Distinct Genetic Modifications. *Nature*. 1994 Apr;368(6474):856–859.
- [18] Parmley SF, Smith GP. Antibody-Selectable Filamentous Fd Phage Vectors: Affinity Purification of Target Genes. *Gene*. 1988 Dec;73(2):305–318.
- [19] Smith GP. Filamentous Fusion Phage: Novel Expression Vectors That Display Cloned Antigens on the Virion Surface. *Science*. 1985 Jun;228(4705):1315–1317.
- [20] Azzazy HME, Highsmith WE. Phage Display Technology: Clinical Applications and Recent Innovations. *Clinical Biochemistry*. 2002 Sep;35(6):425–445.
- [21] Marks JD, Hoogenboom HR, Bonnert TP, McCafferty J, Griffiths AD, Winter G. Bypassing Immunization. *Journal of Molecular Biology*. 1991 Dec;222(3):581–597.
- [22] Mease PJ. Adalimumab in the Treatment of Arthritis. *Therapeutics and Clinical Risk Management*. 2007 Mar;3(1):133–148.
- [23] Kazazi-Hyseni F, Beijnen JH, Schellens JHM. Bevacizumab. *The Oncologist*. 2010 Aug;15(8):819–825.
- [24] Thurlings RM, Vos K, Wijbrandts CA, Zwinderman AH, Gerlag DM, Tak PP. Synovial Tissue Response to Rituximab: Mechanism of Action and Identification of Biomarkers of Response. *Annals of the Rheumatic Diseases*. 2008 Jul;67(7):917–925.
- [25] Eming R, Nagel A, Wolff-Franke S, Podstawa E, Debus D, Hertl M. Rituximab Exerts a Dual Effect in Pemphigus Vulgaris. *Journal of Investigative Dermatology*. 2008 Dec;128(12):2850–2858.
- [26] Saylor C, Dadachova E, Casadevall A. Monoclonal Antibody-Based Therapies for Microbial Diseases. *Vaccine*. 2009 Dec;27:G38–G46.
- [27] Roth DB. V(D)J Recombination: Mechanism, Errors, and Fidelity. *Microbiology spectrum*. 2014 Dec;2(6).
- [28] Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, et al. Precise Determination of the Diversity of a Combinatorial Antibody Library Gives Insight into the Human Immunoglobulin Repertoire. *Proceedings of the National Academy of Sciences of the United States of America*. 2009 Dec;106(48):20216–20221.

- [29] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A Guide to Deep Learning in Healthcare. *Nature Medicine*. 2019 Jan;25(1):24–29.
- [30] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*. 2017 May;60(6):84–90.
- [31] Zhang Y, Sohn K, Villegas R, Pan G, Lee H. Improving Object Detection With Deep Convolutional Networks via Bayesian Optimization and Structured Prediction. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2015. p. 249–258.
- [32] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016. p. 770–778.
- [33] Bengio Y, Simard P, Frasconi P. Learning Long-Term Dependencies with Gradient Descent Is Difficult. *IEEE Transactions on Neural Networks*. 1994 Mar;5(2):157–166.
- [34] Graves A, Schmidhuber J. Framewise Phoneme Classification with Bidirectional LSTM Networks. In: *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.. vol. 4; 2005. p. 2047–2052 vol. 4.*
- [35] Hochreiter S, Schmidhuber J. Long Short-Term Memory. *Neural Computation*. 1997 Nov;9(8):1735–1780.
- [36] Qi CR, Su H, Mo K, Guibas LJ. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. *arXiv:161200593 [cs]*. 2017 Apr. Comment: CVPR 2017.
- [37] Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial Transformer Networks. *arXiv:150602025 [cs]*. 2016 Feb.
- [38] Mikołajczyk A, Grochowski M. Data Augmentation for Improving Deep Learning in Image Classification Problem. In: *2018 International Interdisciplinary PhD Workshop (IIPhDW)*; 2018. p. 117–122.
- [39] Xie L, Wang J, Wei Z, Wang M, Tian Q. DisturbLabel: Regularizing CNN on the Loss Layer. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016. p. 4753–4762.
- [40] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research*. 2014 Jan;15(1):1929–1958.
- [41] Krogh A, Hertz JA. A Simple Weight Decay Can Improve Generalization. In: *Proceedings of the 4th International Conference on Neural Information Processing Systems. NIPS'91. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1991. p. 950–957.*



- [42] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature*. 2020 Jan;577(7792):706–710.
- [43] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Tunyasuvunakool K, et al.. High Accuracy Protein Structure Prediction Using Deep Learning; 2021.
- [44] Choong YS, Lee YV, Soong JX, Law CT, Lim YY. Computer-Aided Antibody Design: An Overview. In: Lim TS, editor. *Recombinant Antibodies for Infectious Diseases*. Advances in Experimental Medicine and Biology. Cham: Springer International Publishing; 2017. p. 221–243.
- [45] Kringelum JV, Lundegaard C, Lund O, Nielsen M. Reliable B Cell Epitope Predictions: Impacts of Method Development and Improved Benchmarking. *PLOS Computational Biology*. 2012 Dec;8(12):e1002829.
- [46] Ivakhnenko AG, Lapa VG. CYBERNETIC PREDICTING DEVICES,. PURDUE UNIV LAFAYETTE IND SCHOOL OF ELECTRICAL ENGINEERING; 1966.
- [47] Schaller RR. Moore's Law: Past, Present and Future. *IEEE Spectrum*. 1997 Jun;34(6):52–59.
- [48] Balci AT, Gumeli C, Hakouz A, Yuret D, Keskin O, Gursoy A. DeepInterface: Protein-Protein Interface Validation Using 3D Convolutional Neural Networks. *bioRxiv*. 2019 Apr:617506.
- [49] Ruffolo JA, Guerra C, Mahajan SP, Sulam J, Gray JJ. Geometric Potentials from Deep Learning Improve Prediction of CDR H3 Loop Structures. *Bioinformatics (Oxford, England)*. 2020 Jul;36(Suppl\_1):i268–i275.
- [50] Li H, Gong XJ, Yu H, Zhou C. Deep Neural Network Based Predictions of Protein Interactions Using Primary Sequences. *Molecules (Basel, Switzerland)*. 2018 Aug;23(8).
- [51] Hiranuma N, Park H, Baek M, Anishchenko I, Dauparas J, Baker D. Improved Protein Structure Refinement Guided by Deep Learning Based Accuracy Estimation. *Nature Communications*. 2021 Feb;12(1):1340.
- [52] Hashemifar S, Neyshabur B, Khan AA, Xu J. Predicting Protein–Protein Interactions through Sequence-Based Deep Learning. *Bioinformatics*. 2018 Sep;34(17):i802–i810.
- [53] Hou J, Wu T, Guo Z, Quadir F, Cheng J. The MULTICOM Protein Structure Prediction Server Empowered by Deep Learning and Contact Distance Prediction. *Methods in Molecular Biology (Clifton, NJ)*. 2020;2165:13–26.
- [54] Liberis E, Veličković P, Sormanni P, Vendruscolo M, Liò P. Parapred: Antibody Paratope Prediction Using Convolutional and Recurrent Neural Networks. *Bioinformatics*. 2018 Sep;34(17):2944–2950.

- [55] Xu J, Wang S. Analysis of Distance-Based Protein Structure Prediction by Deep Learning in CASP13. *Proteins*. 2019 Dec;87(12):1069–1081.
- [56] Dai B, Bailey-Kellogg C. Protein Interaction Interface Region Prediction by Geometric Deep Learning. *Bioinformatics (Oxford, England)*. 2021 Mar.
- [57] Al-Lazikani B, Lesk AM, Chothia C. Standard Conformations for the Canonical Structures of immunoglobulins<sup>11</sup> Edited by I. A. Wilson. *Journal of Molecular Biology*. 1997 Nov;273(4):927–948.
- [58] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics*. 2009 Jun;25(11):1422–1423.
- [59] Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Computational Biology*. 2017 Jan;13(1):e1005324.
- [60] Yu F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions. arXiv:151107122 [cs]. 2016 Apr. Comment: Published as a conference paper at ICLR 2016.
- [61] Krawczyk K, Baker T, Shi J, Deane CM. Antibody I-Patch Prediction of the Antibody Binding Site Improves Rigid Local Antibody-Antigen Docking. *Protein engineering, design & selection: PEDS*. 2013 Oct;26(10):621–629.
- [62] Hou J, Wu T, Cao R, Cheng J. Protein Tertiary Structure Modeling Driven by Deep Learning and Contact Distance Prediction in CASP13. *Proteins*. 2019 Dec;87(12):1165–1178.
- [63] Adhikari B, Hou J, Cheng J. DNCON2: Improved Protein Contact Prediction Using Two-Level Deep Convolutional Neural Networks. *Bioinformatics (Oxford, England)*. 2018 May;34(9):1466–1472.
- [64] Xu Q, Zhang M, Gu Z, Pan G. Overfitting Remedy by Sparsifying Regularization on Fully-Connected Layers of CNNs. *Neurocomputing*. 2019 Feb;328:69–74.
- [65] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Protein Structure Prediction Using Multiple Deep Neural Networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins: Structure, Function, and Bioinformatics*. 2019;87(12):1141–1148.
- [66] Read RJ, Chavali G. Assessment of CASP7 Predictions in the High Accuracy Template-Based Modeling Category. *Proteins: Structure, Function, and Bioinformatics*. 2007;69(S8):27–37.
- [67] Konagurthu AS, Lesk AM, Allison L. Minimum Message Length Inference of Secondary Structure from Protein Coordinate Data. *Bioinformatics*. 2012 Jun;28(12):i97–i105.

- 
- [68] Remmert M, Biegert A, Hauser A, Söding J. HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment. *Nature Methods*. 2012 Feb;9(2):173–175.