# Protein Structure, Antibody Specific B-Cell Epitope Complexes and Antibody-Antigen Interface Analysis

Machine Learning in Bioinformatics

**Bachelor Thesis**

Submitted in partial fulfillment of the requirements for the degree of

**Bachelor of Science in Engineering**

to the University of Applied Sciences FH Campus Wien

Bachelor Degree Program: Computer Science and Digital Communications


**Author:**

Mwape Chintu

**Student identification number:**

Number

**Supervisor:**

DI Dr. techn. Mugdim Bublin


**Date:**

17.01.2021

Declaration of authorship:

I declare that this Bachelor Thesis has been written by myself. I have not used any other than the listed sources, nor have I received any unauthorized help.

I hereby certify that I have not submitted this Bachelor Thesis in any form (to a reviewer for assessment) either in Austria or abroad.

Furthermore, I assure that the (printed and electronic) copies I have submitted are identical.

Date: 17.01.2021 Signature: Mwape Chintu

# Abstract

Antibody-Antigen Interface Analysis is the powerhouse of monoclonal therapeutics. Vaccines can be developed at an alarming rate through the integration of computational biology as molecular biologist have the possibility to examine binding probabilities of Antibodies-Antigens complexes and determine the antibody that would most effective against the specific B-Cell Epitope. This experiment will contribute to the development of custom Abs by evaluating Ab and Ag docking probabilities and addressing challenges that the state of the art computational protein modelling and Antibody-Antigen design currently faces. The effectiveness or binding affinities will be evaluated using existing tools, namely ZDOCK, pyDOCKWEB and HADDOCK for predictions on human Immunoglobulin E (IgE), an Ab that is responsible for an allergic reaction. There are in effect five main classes of antibodies with varying functions: immunoglobulin IgA IgD, IgE, IgG, and IgM. In this case, the focus is on IgE Abs PA12P1C07, and PA12P1C07 which coupled with the Arachis hypogaea (peanut) allergen Ags Ara h 1, Ara h 2, and Ara h 3.

# List of Abbreviations

| | |
|---|---|
| DNA | Deoxyribonucleic acid |
| A | Adenine |
| T | Guanine |
| G | Guanine |
| C | Cytosine |
| mRNA | Messenger Ribonucleic acid |
| Ab | Antibody |
| Ig | Immunoglobulin |
| IgA | Immunoglobulin A |
| IgD | Immunoglobulin D |
| IgE | Immunoglobulin E |
| IgG | Immunoglobulin G |
| IgM | Immunoglobulin M |
| Ag | Antigen |
| CNN | Convolutional Neural Network |
| RNN | Recurrent Neural Network |
| LSTM | Long Short Term Memory (RNN) |
| ANN | Artificial Neural Network |
| IMGT | International Immunogenetics Information System |
| PDB | Protein Databank |
| COOH | Carboxyl Group |
| NH3 | Amino Group |
| rRNA | Ribosomal Ribonucleic acid |
| $\alpha$ | Alpha Carbon Atoms |
| ER | Endoplasmic Reticulum |
| $\alpha$-Helix | Alpha Helix |
| $\beta$-Pleated Sheet | Beta Pleated Sheet |
| H | Hydrogen |
| C=O | Oxygen Atom |
| N-H | Amino H |

# Key Terms

DNA

mRMA

PDB

Ab

IgE

LSTM

# Contents


**1 Introduction** **1**

**2 Fundamentals of Proteins** **4**
2.1 Protein Functions and the Primary Structure . . . . . 4
2.2 Secondary Structure . . . . . 5
2.3 Tertiary Structure . . . . . 6
2.4 Quaternary Structure . . . . . 6

**3 Deep Learning for Homologous Sequences** **7**
3.1 The Protein Folding Problem: DeepMind AlphaFold . . . . . 7
3.2 CASP14 AlphaFold 2's Architecture and Transformers . . . . . 9
3.3 LSTMs for Monoclonal Antibody Design . . . . . 13

**4 Antibody-Antigen Interaction** **15**
4.1 Antibody Interface Analysis . . . . . 15
4.2 PyDOCKWEB . . . . . 18
4.3 HADDOCK . . . . . 18
4.4 ZDOCK . . . . . 20

**5 Results and Discussion** **21**

**6 Conclusion and Future Work** **22**

**Bibliography** **23**

**List of Figures** **24**

**List of Tables** **25**

# 1 Introduction

Proteins are the fundamental building blocks of life and have the most diverse range of functions across all macromolecules. They are the result of a chemical process that begins with the decryption of cell Deoxyribonucleic acid (DNA) which then forms into the structural and motor elements in a cell. DNA is a polynucleotide chain with sugar-phosphate mainchain, Adenine (A), thymine (T), guanine (G) and cytosine (C) attached to each sugar [1]. A common misconception is in respect to the size of the protein which is often mistaken for being miniscule, however this could not be any further from the truth. In order to comprehend the reasoning behind this, it is imperative to understand how proteins themselves are produced. They are produced in an analogous two-step process in virtually all organisms. DNA is first transcribed into Messenger Ribonucleic acid (mRNA), then mRNA is translated into the an amino acid sequence of a protein which then folds into a three dimensional (3D) structure. This process is illustrated in Figure 1.1

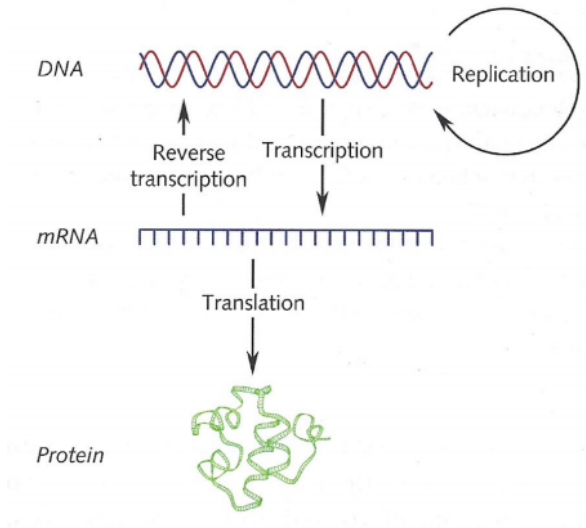



Figure 1.1: The process of DNA Replication. The information is transcripted to mRNA and is eventually translated into a Protein.

These are substantial and convoluted biological molecules that act as a powerhouse of a cell; carrying out and enforcing the function of the cell. In order to ensure that a Protein always has this "executive role", as author of the book "Introduction To Bioinformatics" Arthur M. Lesk, A Professor at the Pennsylvania State University always refers to this phenomenon, they have several functions with one of them being their ability to fold spontaneously into specific shapes. Protein folding serves as a very important function or tool as the shape that the protein folds into, determines its function and/or more complex structures can be generated. An Antibody (Ab) or immunoglobulin (Ig), is a Y-shaped protein designed by for combat and is used by the immune system to identify and neutralize foreign bodies named antigens (Ag).

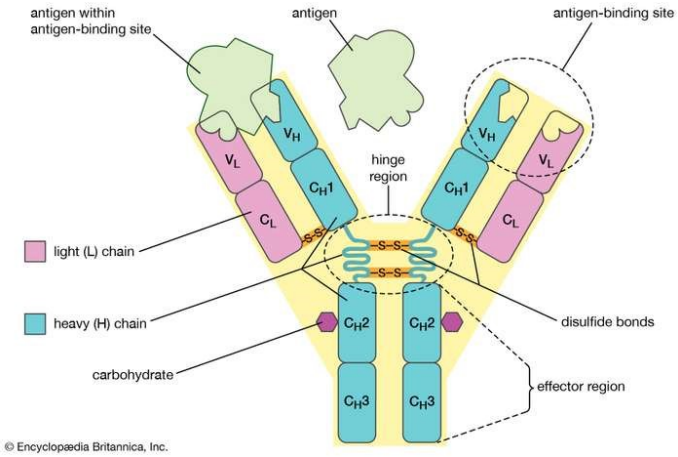



Figure 1.2: Antibody-Antigen binding illustration. At the binding site, the Paratope (analogous for a lock) binds with one and only one specific type of antigen on its Epitope (analogous for a key)

For an Ab to achieve this, it has an antigen-binding (antibody-antigen docking site) which is used for disabling the foreign body. The interaction between the two functions in the same way a key and a lock functions; each antibody binds to a specific antigen. The analogous "lock" is referred to a paratope, while the analogous "key" is referred to a epitope. The figure below, Figure 1.2, perfectly illustrates this mechanism.

Traditionally, B-Cells are the cells that make these Ab Specific to each pathogen. Depending on the pathogen, a specific type of T cell is responsible for aiding the B cells in this process. Evidently, viruses, as can be witnessed in the case of SARS-COV-2, the immune system response through the production of Abs may not be adequate to neutralize the threat. Videlicet, this is a major challenge that is the study of pathology faces at this current point of time. Viruses, unlike bacteria, are not classed as living organisms are tiny (e.g SARS-COV-2 is only 0.1 micrometers, 70 times smaller than a red blood cell) and as a result, difficult to target using traditional therapeutics and drugs. The solution then lies in the production of Antibodies. This is because of the binding property that they harness. It is by this virtue that they are at present the largest class of biotherapeutics. This paper aims contribute to the development of custom Abs by evaluating Ab and Ag docking probabilities and addressing challenges that the state of the art computational protein modelling and Antibody-Antigen design currently faces. The effectiveness or binding affinities will be evaluated using existing tools, namely ZDOCK [2], pyDOCKWEB [3] and HADDOCK [4] for prediction on human Immunoglobulin E (IgE), an Ab that is responsible for an allergic reaction. There are in effect five main classes of antibodies with varying functions: immunoglobulin IgA IgD, IgE, IgG, and IgM. In this case, the focus is on IgE Abs PA12P1C07, and PA12P1C07 which coupled with the Arachis hypogaea (peanut) allergen Ags Ara h 1 [5], Ara h 2, and Ara h 3. ZDOCK does not have a working 3D Visualizer. In such a case, a visualization tool called PyMOL [6] will be used to show the predicted structure of the binding of Ab-Ag (Receptor and Ligand). In addition, the primary input for the above mentioned tools is a PDB file. Swiss Model [7] will be used to convert the sequences which are in the FASTA format to a PDB file.

Introduction of computational protein modelling, custom Antibodies has been achieved through Convolutional Neural Network (CNN) based Deep Learning algorithms such as AlphaFold. Recurrent Neural Networks (RNN) or more specialized versions of RNNs such as

Long Short-Term Memory (LSTM) have been widely successful in long sequences predictions and for remote homology prediction such PSI-BLAST. The reason for this success stems from the major advantage that an LSTM has over regular Artificial Neural Networks (ANN) or RNNs, which is that they are capable of learning long-term dependencies. Antibody-specific computational protocols are now benefiting from the increasing amount of data available from next-generation sequencing and are being applied to relevant drug models based on traditional antibodies, such as nanobodies [8].

Machine Learning and Deep Learning Algorithms heavily rely on a large assortment of data for training process and testing the effectiveness of the mathematical model. Not only be the large quantity of such data be available, it must also be suitable, easily accessible, and usable. Fortunately, there already exist a wide variety of database repositories depending on the application. This article focuses on, the SAbDab-Therapeutic-Antibodies database is an already well-established databases that contain therapeutic antibody information. Sequence Databases are also important for Antibody design. As an example, the International Immunogenetics Information System (IMGT) is a sequence database that is used to derive gene assignments for recombined antibodies. If a Protein Data Bank (PDB) file containing the 3D structure of a Protein, Ab or Ag is present, it's structure can be analyzed and compared with several models from the PDB for the most suitable matches. Moreover, the PDB is not accessed directly but through tools such as Chimera X and Swiss Model.

# 2 Fundamentals of Proteins

## 2.1 Protein Functions and the Primary Structure

In order too grasp the concept of Ab Design and evaluate the Antibody-Antigen binding probabilities, a profound understanding of the fundamentals of how proteins functions and the composition of their structure is a prerequisite. The reason for this is that most protein analysis applications require some understanding of what to input, what the output represents, the advantages as well as the disadvantages. As already discussed in the introduction section, Proteins are the “powerhouse” of a cell and then they fold, larger complex structures such as antibodies can be formed. Protein folding is not the only function that is available have three main functions to help them with enforcing the function of the cell. The functions that permit this are:

- They have a vast assortment of functions and properties for the cell’s growth, structure, metabolism, and regulation/maintenance
- Proteins have a structural form: They no less than one long polypeptide, which is a single linear chain amino acid of more than fifty residues and short polypeptides containing less than thirty residues which are commonly referred to as peptides. This

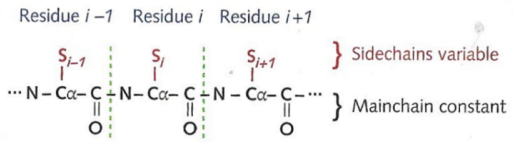



Figure 2.1: Graphic illustration of a Protein mainchain having different amino acid side chains (R Group) attached at regular intervals

linear Amino acid sequence within the protein is what is referred to as the Primary Structure of the protein and is the simplest level of its structure. This sequence is determined by DNA and as a result, a slightest change in the DNA sequence will likely lead to a change in the amino acid sequence. Any change in the amino acid sequence affects how the protein folds and its function. Each individual amino acid has an R group or more commonly referred to as a side chain, Carboxyl Group (COOH) and an Amino Group (NH3). The residue is bonded together using peptide bonds and neighboring amino acid residues. All of these components of a protein can be simplified into two parts; a polymer main chain, or more commonly referred to a “backbone” of a molecule, and a side chain sequence. This structural form is dictated by Ribosomal Ribonucleic acid (rRNA); a molecule within a cell that is part of the protein-synthesising organelle known as the "ribosome". That is then exported to the cytoplasm in order to aid with the translation process where the data in Messenger RNA (mRNA), is translated into proteins. The side chains of the amino acid residues are attached to the

alpha-carbon atoms ($\alpha$) of the amide mainchain at regular intervals as shown in Figure 1.2.

- Protein folding, the process of which a protein chain acquires its native 3D structure (folds into a specific shape), happens spontaneously and occurs inside of a compartment called the Endoplasmic Reticulum (ER). Protein folding is the most crucial function because it determines the function of the protein. It is from this folding that more complex structures and functions can be built. Using the example from the introduction and further expressing its point, As already mentioned, proteins can fold into a Y shape which functions as an Ab or Ig. These are large proteins that are used by the immune system to identify and neutralize foreign objects such as pathogenic bacteria and viruses. Naturally, due to the law of entropy, a key focus of the second law of thermodynamics, proteins can also fold incorrectly which has been associated with illnesses such as Alzheimer's disease, cystic fibrosis and many other degenerative and neurodegenerative disorders. The process is referred to as denaturing Not only is this folding happening spontaneously, there is another added complexity to the process. Apart from the already mentioned Primary Structure a protein, there are an additional three "levels" of protein folding. There remaining three levels of its structure are Secondary, Tertiary and Quaternary Structure.

## 2.2 Secondary Structure

At this level, the amino acids sequence locally folds inside a polypeptide. The two most common types of secondary structures are the Alpha Helix ($\alpha$-helix) and the Beta pleated sheet ($\beta$-pleated sheet). These secondary structures are held in shape by hydrogen (H) bonds, which form between the carbonyl O group of one amino acid and the amino H of the adjacent. A Carbonyl Group contains a carbon atom double-bonded to an oxygen atom (C=O). Which shape the protein folds into, depends on its arrange of the amino acid sequence. In an $\alpha$-helix structure, C=O of one amino acid is hydrogen bonded to the amino H (N-H) of an amino acid that is four down the chain (i-4). The result is that the polypeptide chain forms into a helix structure/repeated pattern.

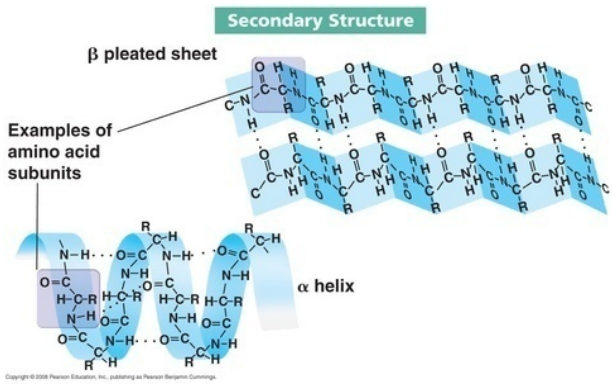



Figure 2.2: Graphic illustration of an $\alpha$-Helix and a $\beta$-Pleated Sheet Protein Structure

In the second secondary structure, the $\beta$-pleated sheet, at least two segments of a polypeptide chain align next to each other and form a sheet-like structure as shown in Fig 5. The

corner's in this structure are formed when hydrogen bonds between atoms on the backbone of the polypeptide chain. The $\alpha$-Helix and $\beta$-Pleated Sheet structures are found in many globular and fibrous proteins.

## 2.3 Tertiary Structure

The tertiary structure is the 3D structure of the protein consisting of the polypeptide chain. This occurs mostly due to interactions between the R groups of the amino acids that make up the protein. It contains a single chain of amino acids but with varying structure shapes within the chain. A single chain may have sections containing both an $\alpha$-Helix as well as a $\beta$-pleated sheet as shown the ribbon diagram in Figure 2.3. The $\alpha$-Helices are represented by curls while the $\beta$-pleated sheets are represented by arrows. " When protein folding takes place, the hydrophobic R groups of nonpolar amino acids lay in the interior of the protein, whereas the hydrophilic R groups lay on the outside. The former types of interactions are also known as hydrophobic interactions. Interaction between cysteine side chains forms disulfide linkages in the presence of oxygen, the only covalent bond forming during protein folding."

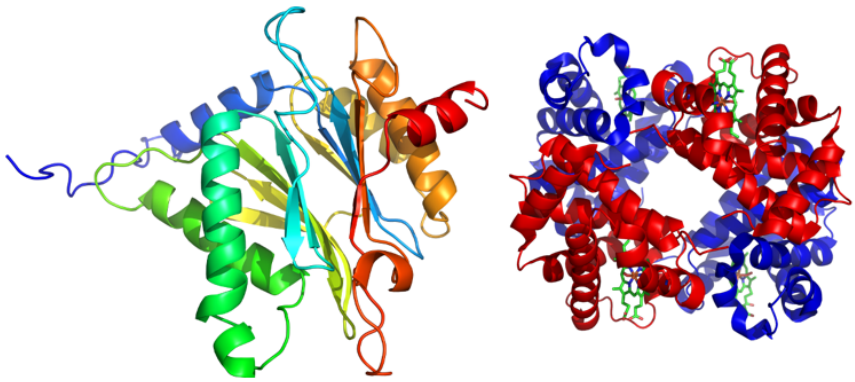


Figure 2.3: A Ribbon Diagram representation of the Tertiary Structure (Left) and the Quaternary Structure (Right) of a protein. Note how one figure consists of a single chain of amino acids while the other contains at least two.

## 2.4 Quaternary Structure

The main difference between the Tertiary Structure and the Quaternary Structure is the number of amino acid chains that the protein has. Whereas the former level only contains one chain of amino acids, the latter has at least more than two as shown in Figure 2.3. "In nature, some proteins are formed from several polypeptides, also known as subunits, and the interaction of these subunits forms the quaternary structure. Weak interactions between the subunits help to stabilize the overall structure"

# 3 Deep Learning for Homologous Sequences

## 3.1 The Protein Folding Problem: DeepMind AlphaFold

Traditionally, determining protein structures from an empirical point of view is difficult to achieve. As already discussed in the second chapter, Fundamentals of Proteins, a protein structure consists of four levels; Primary structure, Secondary Structure, Tertiary Structure and Quaternary Structure. This structure is determined by the linear amino acid sequence with each each individual amino acid having an R group, COOH group and an NH3 group. The residue is bonded together using peptide bonds and neighboring amino acid residues. However, there is one problem. It is still not clear how a proteins amino acid sequence leads to a fully three dimensional strucutre. This is the known as the Protein Folding Problem. It is ultimately a series of smaller questions that it centered around the overall question of how in the primary structure of a protein, how its amino acid sequence dictates this 3D structure. These smaller questions are:

- What is the folding code?
- What is the folding mechanism?
- Can we predict the native structure of a protein from its amino acid sequence?

The 1972 Nobel Peace Prize Winning American Biologist Christian Boehmer Anfinsen Jr, theorized that it should be possible to predict the 3D structure of any protein, in this case specifically Abs, based on their 1D amino acid sequence. According to the Levinthal Paradox, the number of ways in which a can protein fold by brute force before it settles in its intended shape is $10^{300}$, which would take longer than the age of the known universe and yet in nature, proteins fold spontaneously within miliseconds.

However, one can deduce these amino acid residues are bonded together by leveraging genetic data, calculating the local co-variance in the protein sequence alignments as given by:

$$MIp(i,j) = MI(i,j) - \frac{MI(i,\bar{x}) \cdot MI(j,\bar{x})}{\overline{MI}}$$

Figure 3.1: Formular for calculating covariance in a local protein sequence alignment [9]

Where MI(I, $\bar{X}$) is an mean score of the $i$-th column along with other columns in the protein sequence alignment and MI($j$, $\bar{x}$) represents the mean score of the $j$-th column with other columns in the sequence alignment. $\bar{MI}$ is the mean of all the MI scores from the $i$-th and the $j$-th columns of the sequence alignment. An analysis can then be performed on this locally calculated co-variance thus the structure of the protein can be predicted. A Deep Learning Artificial Neural Network can then be used to predict the distance between

all possible amino acid residue pairs of a three-dimensional protein structure using a binary 2D-Array [9]. This is referred to as a contact map as shown in the Figure 3.2:

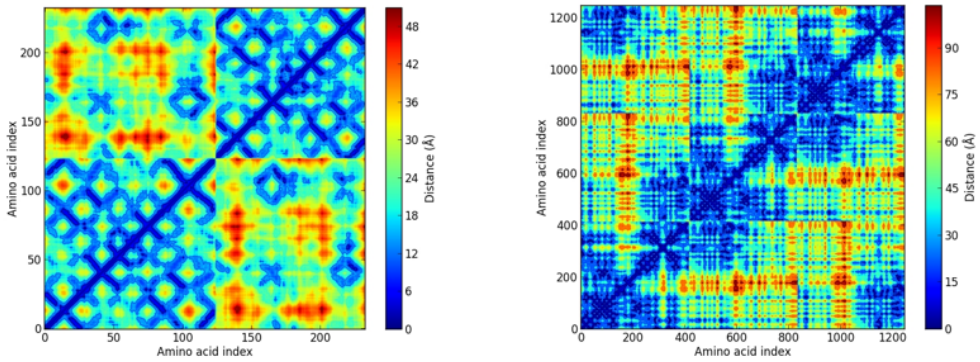



Figure 3.2: Distance Map computed for the Ab PA13P1H08 (Left) and the Ag Ara h 1 peanut allergen (Right) using the nanohub.org online protein contact map tool

For two residues $i$ and $j$, $ij$-th element of the matrix is 1 if the two residues are closer than a predetermined threshold, and 0 otherwise. The result can then be optimized using gradient descent or often referred to as backpropagation, common machine and deep learning approach used for updating the weights in a neural network. It is a mathematical optimization algorithm designed in finding the local extrema in order to obtain the highest degree of accuracy. In the case of Antibody-Antigen interaction, this same mechanism can be used to predict both structures of the Ab and Ag and from the results determine the binding probabilities. The system described above is what is AlphaFold, A deep learning algorithm that was trained using the sequences and structures of 100,000 estimated known proteins. The Critical Assessment of Protein Structure Prediction (CASP) is an experimental event held every two years for protein structure prediction and in its latest version at the time in the year 2018, CASP13, AlphaFold predicted the structures of the protein with a high accuracy. It achieved a template based modelling score (TM) of 0.7 and higher for 24 of the 43 domains while on the other hand, the second most best tool used achieved a TM score of 0.7 or and higher for 14 out of 43 domains as shown in Figure 3.3.

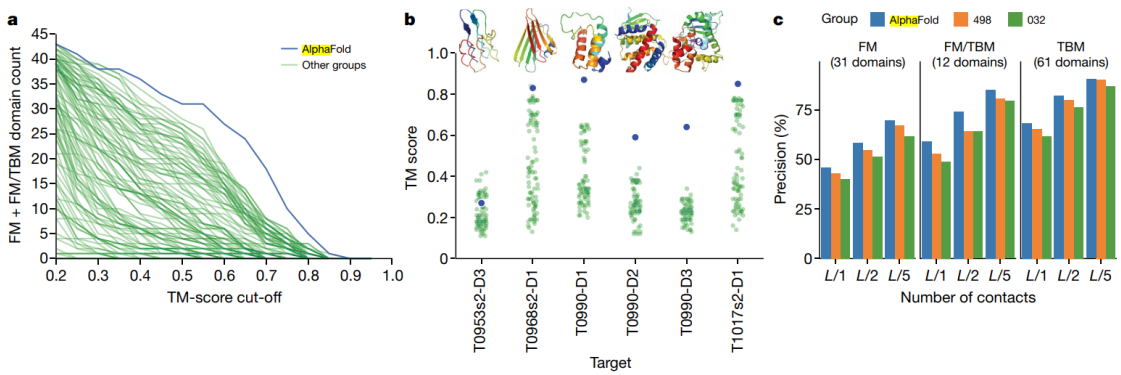



Figure 3.3: Graph Depiction of the TM Scores achieved by AlphaFold compared to its counterparts in the CASP13 Assessment

AlphaFold uses a convolutional neural network that is trained in PDB structures to predict the distances d$ij$ between the atoms of pairs, $ij$, of residues in a protein C$\beta$. Based on the representation of the sequence of amino acids, S, a protein and the characteristics derived

from the MSA(S)-sequence, the net, which has a similar structure to those used for imaging recognition tasks, 29 predicts a discrete probability distribution P(dij|S, MSA(S)) for each pair *ij* in any region 64 × 64 of the distance matrix L × L. The complete set of distance distribution predictions constructed by the combination of these predictions covering the entire distance map is called the histogram, short for distance histogram.

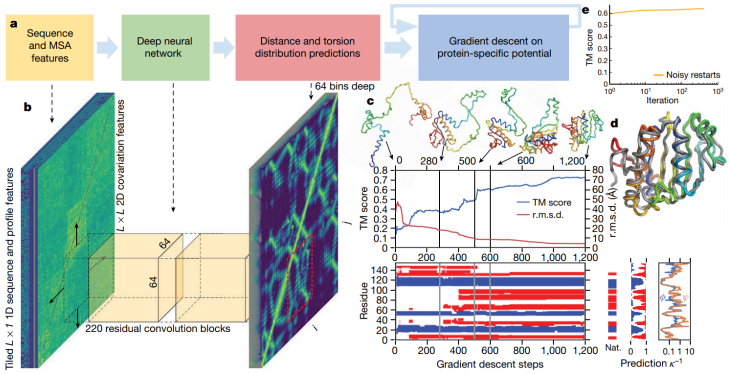



Figure 3.4: Graph Depiction of the TM Scores achieved by AlphaFold compared to its counterparts in the CASP13 Assessment

## 3.2 CASP14 AlphaFold 2's Architecture and Transformers

The latest iteration of Google's Deep Mind algorithm showcased at CASP14, AlphaFold 2 overhauled it's first iteration in attempt to achieve a much more accurate structure prediction among all protein types. The recent primary method used by CASP to assess accuracy of the prediction made by a tool is what is commonly refered to as the Global Distance Test (GDT). This is a metric that has a range of 0-100 where the higher the number, the more accurate the prediction is deemed.In effect, this can be reffered to as a percentage of amino acid residues within a limiting distance of the correct position. According to the CASP14 evaluation, the latter AlphaFold system achieves an average score of 92.4 GDT on all goals. Simply put, these results indicate that the predictions have an average error (RMSD) of about 1.6 Angstroms, which is proportional to the width of an atom. In comparison, the first iteration of AlphaFold had a GDT score of 58.

The exact architecture of the ANN of DeepMind's AlphaFold 2 is not fully known, However it can be speculated, based on the team's presentation [10] that it consists of either a transformer or LSTM networks as both types of ANN have proved to successful in capturing possible long range dependencies between binding sequence and structure motifs identified by the CNNs.

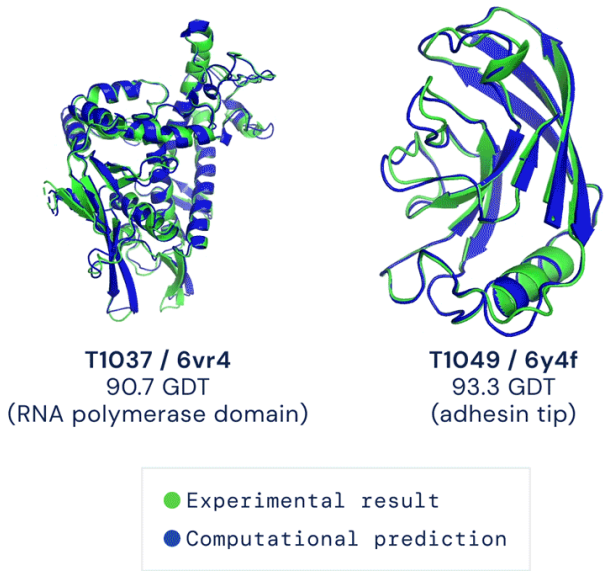



Figure 3.5: Demonstration of the accuracy achieved by the AlphaFold 2 System represented by the blue structure

Referring to the protein folding problem discussed in section 3.1, A protein structure should be predicted just from its 1D amino acid sequence. As a means of illustration, Given the following sequence of Arah h 1 Arachis Hypogaea:

```
sp|P43238|ALL12_ARAHY  RREEGGRWGPAGPRERERREEDWRQPREDWR    54

sp|P43238|ALL12_ARAHY  RPSHQQPRKIRPEGREGEQEWGTPGSHVRE    84

sp|P43238|ALL12_ARAHY  ETSRNNPFYFPSRRFSTRYGNQNGRIRVLQ   114

sp|P43238|ALL12_ARAHY  RFDQRSRQFQNLQNHRIVQIEAKPNTLVLP   144

sp|P43238|ALL12_ARAHY  KHADADNILVIQQGQATVTVANGNNR   170

X   88.1 (G); Standard sidechain area (Å²)
X  118.2 (A); 129.8 (S)
X  146.1 (C); 146.8 (P)
X  152.5 (T); 158.7 (D); 164.5 (V); 165.5 (N)
X  181.0 (I); 186.2 (E)
X  193.1 (L); 193.2 (Q); 202.5 (H); 203.3 (M)
X  222.8 (F); 225.8 (K)
X  238.8 (Y)
X  256.0 (R); 266.2 (W)
```

The X, Y, and Z coordinates of the protein backbone atoms (structure) should be returned including the symmetry of roto-translation from that sequence. The geometry of a protein can then be mathematically represented as follows:

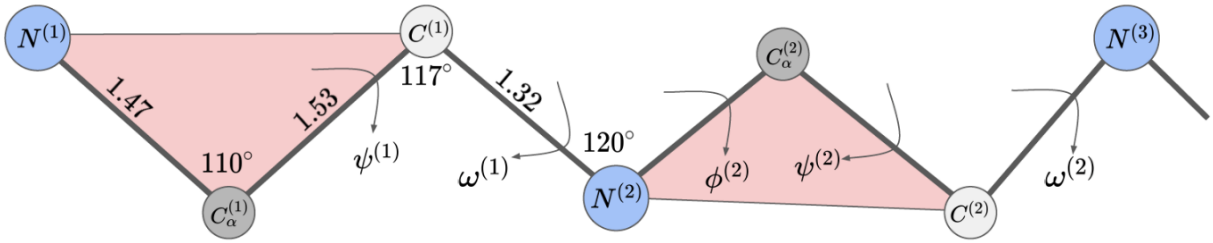



Figure 3.6: Mathematical representation of the geometry of a protein [11]

In the first iteration of AlphaFold, The system used distance distribution predictions between the atoms and then a gradient descent algorithm was applied to predict the coordinates for the protein. In DeepMind's second iteration, AlphaFold 2, on the other hand, the coordinates are predicated directly using end-to-end regression techniques. This has the undesired effect of increased inaccurate symmetry.

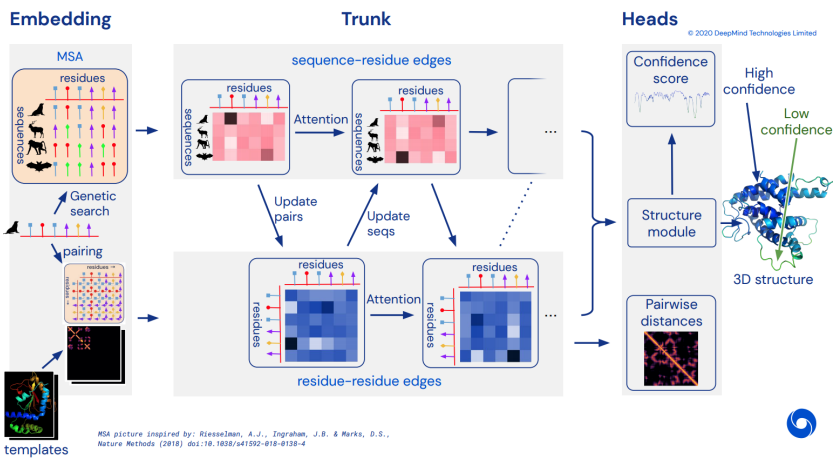



Figure 3.7: AlphaFold 2 Protein Sequencing Architecture showcased during the DeepMind AlphaFold 2 presentation [10]

According to Figure 3.7 which was taken from the AlphaFold 2 presentation [10], the architecture consists of three main sections:

- Embedding
- Trunk
- Heads

### Embedding

In the embedding section, The target protein sequence is encoded along with Multiple Sequence Alignment (MSA) as well as protein structures that share similar characteristics called templates in a technique referred to as template based modelling. Multiple sequence alignment (MSA) can refer to the process or result of aligning the sequences of three or more biological sequences. These sequences are either proteins, DNA or RNA.

### Trunk

The Trunk learns the residue to residue graph edges and the sequence to residue graph edges as shown in Figure 3.6. The residue to edge sequence of a residue graph is the pairwise information between all residues that are the same relative distances in three-dimensional space and angles. The sequence to residue edges can contain sequence evolutionary information that can be used to predict pairwise distances and additionally, the output is passed to the structure module of the Heads section in order to generate the X, Y, and Z coordinates of the protein or antibody structure.

### Heads

Finally, in the structure module of the Heads section, a 3D equivariant transformer architecture is used to refine backbone coordinates and to predict side chains. The neural network should then predict new positions and orientations of so called “red triangles” and produce a confidence score.

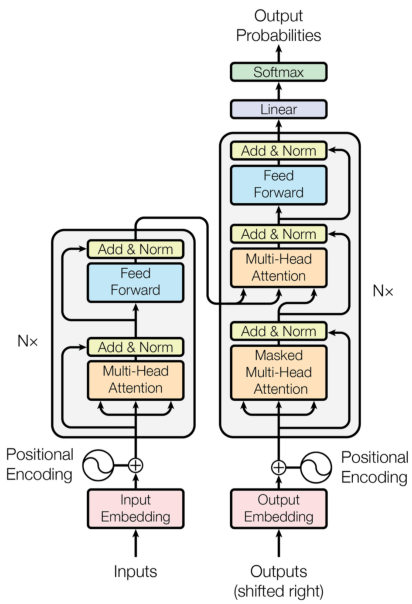



Figure 3.8: Depiction of a Transformer [12]

Simply put, the Transformer is a Attention Mechanism neural network has an encoder on the left and a decoder on the right, examining the input sequence and calculating at each

step which information is important and which is not as shown in Figure 3.8 . It is an architecture that converts one sequence to another with the help of an encoder and decoder, but is different from other sequence-to-sequence networks such as LSTMs because it does not imply any Recurrent network.

## 3.3 LSTMs for Monoclonal Antibody Design

Antibodies are large proteins with large sequences containing a heavy and a light chain. Information is carried within this sequence and the context matters in determining which information is important and should be kept. This means that a neural network that is tasked with making a prediction based on a sequence must be able to provide long-dependencies. Standard RNNs do not provide long-term dependencies due to the vanishing gradient problem. This is because the weights are updated by back-propagation through time and therefore unfold as a multi-layer feed-forward network. As the gradients are back-propagated through many time steps, the gradients tend to grow or disappear, as is the case in deep feed-forward networks [13].

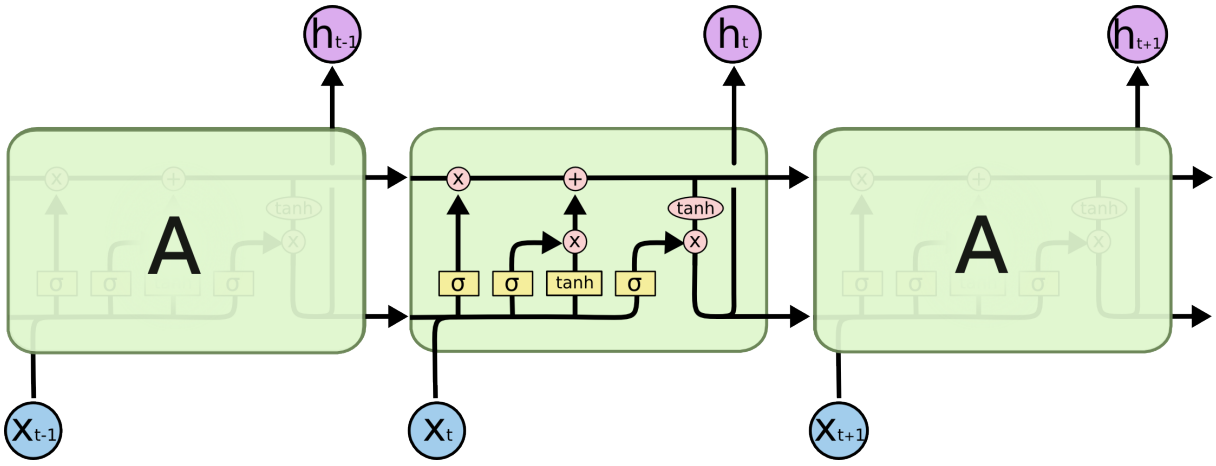



Figure 3.9: Depiction of an LSTM neural network [13]

LSTM networks have Several components:

- Forget Gate that is comprised of a Sigmoid function ($\sigma$-function) and an "x" multiplier. Here, the previous hidden state (ht-1) is combined with an input Xi and the result is computed. The closer the computed value or information is to 1, the more important this information. Likewise, if the output is closer to 0, it is considered insignificant. Ultimately, this is kept in a cell state.

- The Cell State keeps track of the current state of the neural network and is represented by the top straight horizontal line shown in Figure 3.9.

- An Input Gate is used for inputting and updating the neural network. According to Figure 3.9, this is the second Sigmoid function depicted in the image. The process works as follows. The previous hidden state is combined with the current state and the Sigmoid function computes a value between 1 and 0. Simultaneously, the $\tan x$ function computes an output of values between 1 and -1, then both outputs are combined using

an "x" multiplier and finally, a bit-wise operation is performed on this output and the output from the Forget Gate. The result is then kept in the Cell State

- An Output Gate contains the last Sigmoid function and as well as the Forget Gate and the Input Gate, the result is computed from the previous hidden state and the current state. The cell state then computes the $\tan x$ of the value currently stored and this result is combined with the result from the Sigmoid function using a "x" multiplier. This information of the new hidden state and output is then carried out to the next timestamp. This ensures that information is never lost through the network, and thus in the case of antibody design

# 4 Antibody-Antigen Interaction

## 4.1 Antibody Interface Analysis

As briefly discussed in the Introduction of the paper, An antibody is a large Y- shaped protein that functions by binding itself to an antigen or a virus or a pathogen. During this binding process, it triggers the immune system to disable or eliminate the foreign body. Antibodies can be highly specific while having low immunogenicity, which is achieved by their unique structure. The fragmented crystallisable region also commonly referred to as the Fc region as shown in Figure 4.1, activates the immune response and is species-specific. For instance, the human Fc region should not trigger an immune response in humans. The binding region of the antigen fragments, or in short the Fab region, must be highly variable to bind to antigens of different types.

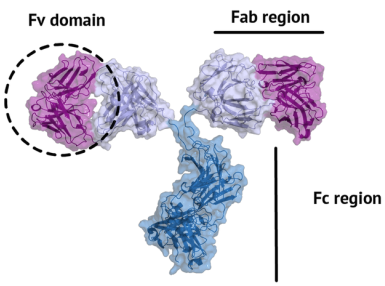



Figure 4.1:

A tiny section of the Fab region that binds to the antigen is called the paratope and is analogous to a lock. On the flip side, the tiny section of the antigen that binds to an antibody is called the epitope and is analogous to a key. The paratope consists of six highly flexible loops called complementary regions (CDRs) or hypervariable loops whose sequence and conformation are altered to bind to different antigens. The CDRs are shown in red in Figure 4.2:

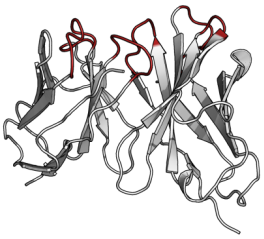


Figure 4.2: CDR loops of of an Antibody [14]

Computational antibody methods vary in great detail with each method having its own unique set of steps. State of the art methods include:

- Antibody Modelling
- Antibody Interface Analysis
- Antibody Design
- Therapeutic-specific Applications
- Antibody Repertoire Characterization

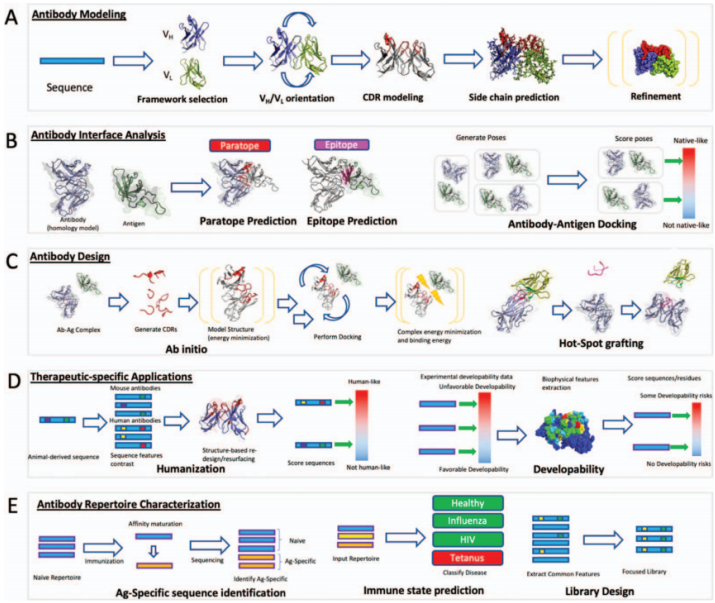



Figure 4.3: Depiction of State of the art Computational antibody methods [8]

The exact steps for each are shown in Figure 4.3. In terms of Antibody Interface Analysis, Antibody interface prediction identifies the residues on the paratope of the antibody that are in contact with the epitope of the antigen. This is a special case of molecular docking, where the purpose of antibody-antigen docking is to reproduce the complex between the antibody and the antigen. After the paratope and epitope prediction, Poses are generated and through Antibody-Antigen Docking, Score poses are are predicted. The result is their Native-Like or Not native-like [8]. The Antibody-Antigen binding probability is measured by the interaction

energy between the two binding sites. The higher the energy, the more energy, the higher the binding probability. In this case, this analysis is performed on the following tools:

- PyDOCKWEB
- ZDOCK
- HADDOCK

### Input Data

The input data is the protein sequence of the three antigens Ara h 1, Ara h 2, Ara 3.

Ara h 1:

```
sp|P43238|ALL12_ARAHY  NTLVLPKHADADNILVIQQGQATVTVANGN   168

sp|P43238|ALL12_ARAHY  NR   170

                       X   88.1 (G); Standard sidechain area (Å²)
                       X  118.2 (A); 129.8 (S)
                       X  146.1 (C); 146.8 (P)
                       X  152.5 (T); 158.7 (D); 164.5 (V); 165.5 (N)
                       X  181.0 (I); 186.2 (E)
                       X  193.1 (L); 193.2 (Q); 202.5 (H); 203.3 (M)
                       X  222.8 (F); 225.8 (K)
                       X  238.8 (Y)
                       X  256.0 (R); 266.2 (W)
```

Ara h 2:

```
sp|Q6PSU2|CONG7_ARAHY  FKRELRNLPQQCGLRAPQRCDLEVESGGRD   149

sp|Q6PSU2|CONG7_ARAHY  RY   151

                       X   88.1 (G); Standard sidechain area (Å²)
                       X  118.2 (A); 129.8 (S)
                       X  146.1 (C); 146.8 (P)
                       X  152.5 (T); 158.7 (D); 164.5 (V); 165.5 (N)
                       X  181.0 (I); 186.2 (E)
                       X  193.1 (L); 193.2 (Q); 202.5 (H); 203.3 (M)
                       X  222.8 (F); 225.8 (K)
                       X  238.8 (Y)
                       X  256.0 (R); 266.2 (W)
```

Ara h 3:

```
tr|Q8LKN1|Q8LKN1_ARAHY  PTGVAFWMYNDHDTDVVAVSLTDTNNNDNQ   168

tr|Q8LKN1|Q8LKN1_ARAHY  LD   170
```

```
X  88.1 (G); Standard sidechain area (Å²)
X 118.2 (A); 129.8 (S)
X 146.1 (C); 146.8 (P)
X 152.5 (T); 158.7 (D); 164.5 (V); 165.5 (N)
X 181.0 (I); 186.2 (E)
X 193.1 (L); 193.2 (Q); 202.5 (H); 203.3 (M)
X 222.8 (F); 225.8 (K)
X 238.8 (Y)
X 256.0 (R); 266.2 (W)
```

The binding probabilities are evaluated against the Antibodies PA13P1H08 and PA12P3F10

## 4.2 PyDOCKWEB

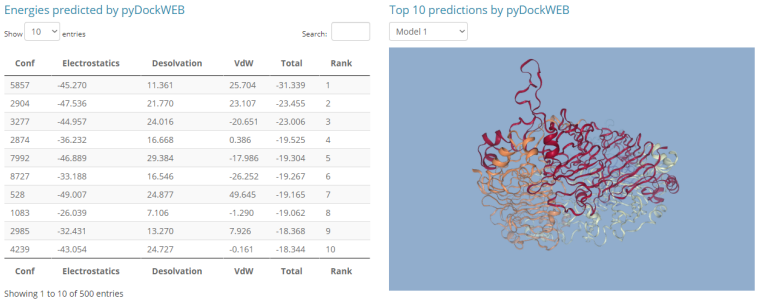
pyDockWEB is a web server for rigid-body docking prediction of protein-protein complex structures using a new version of the pyDock scoring algorithm. Starting from the 3D coordinates of two interacting proteins, pyDockWEB provides the best docking orientations, mainly evaluated by electrostatics and desolvation energy.

Table 4.1: pyDOCKWEB PA13P1H08 and PA12P3F10 Antigen Energy Scores

| | **Configuration** | **Electrostatics** | **Desolvation** | **VdW** | **Total** |
|---|---|---|---|---|---|
| PA13P1H08, Ara h 1 | 5857 | -45.270 | 11.361 | 25.704 | -31.339 |
| PA13P1H08, Ara h 2 | 4862 | -24.517 | 0.126 | 19.578 | -22.433 |
| PA13P1H08, Ara h 3 | - | - | - | - | - |
| PA12P3F10, Ara h 1 | 8910 | -30.717 | 1.349 | 16.722 | -27.696 |
| PA12P3F10, Ara h 2 | 2237 | -14.250 | -12.017 | 34.149 | -22.852 |
| PA12P3F10, Ara h 3 | - | - | - | - | - |

## 4.3 HADDOCK

HADDOCK is a antibody-antigen docking, not antibody specific platform for the modelling of biomolecular complexes. On a fundamental level, it is a collection of python scripts derived from ARIA that use a Crystallography and NMR System (CNS) to calculate the complex structure of molecules. One major advantage that the HADDOCK platform has over other tools with a similar purpose is its ability, which is inherited from the CNS, to incorporate



Energies predicted by pyDockWEB

Show 10 entries Search:

| Conf | Electrostatics | Desolvation | VdW | Total | Rank |
|---|---|---|---|---|---|
| 5857 | -45.270 | 11.361 | 25.704 | -31.339 | 1 |
| 2904 | -47.536 | 21.770 | 23.107 | -23.455 | 2 |
| 3277 | -44.957 | 24.016 | -20.651 | -23.006 | 3 |
| 2874 | -36.232 | 16.668 | 0.386 | -19.525 | 4 |
| 7992 | -46.889 | 29.384 | -17.986 | -19.304 | 5 |
| 8727 | -33.188 | 16.546 | -26.252 | -19.267 | 6 |
| 528 | -49.007 | 24.877 | 49.645 | -19.165 | 7 |
| 1083 | -26.039 | 7.106 | -1.290 | -19.062 | 8 |
| 2985 | -32.431 | 13.270 | 7.926 | -18.368 | 9 |
| 4239 | -43.054 | 24.727 | -0.161 | -18.344 | 10 |
| Conf | Electrostatics | Desolvation | VdW | Total | Rank |

Showing 1 to 10 of 500 entries

Top 10 predictions by pyDockWEB

Model 1

Figure 4.4: The top ten models output by pyDOCKWEB

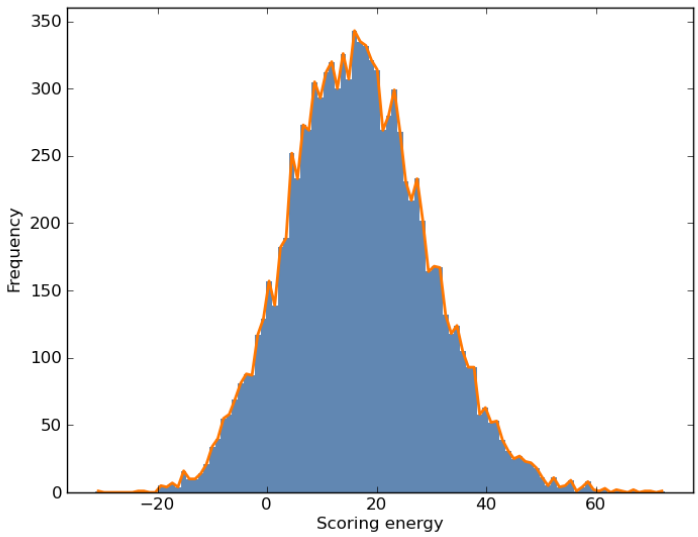



Figure 4.5: Energy Score of the binding of PA13P1H08 and Ara h 1

experimental data as constraints and use it to guide the docking process alongside traditional energy complementarity and form. In addition, the close coupling with the CNS gives HADDOCK the ability to effectively produce models of sufficient quality to be archived in the protein database.

A central aspect of the HADDOCK is the definition of Ambiguous Interaction Restrictions (AIR). These allow raw data, such as NMR chemical shift perturbation or mutagenesis experiments to be translated into distance restrictions that are incorporated into the energy function used in the calculations. RIAs are defined by a list of wastes which are divided into two categories:

- Active Site
- Passive Site

In general, active wastes are those that are of central importance to the interaction, such as those whose outlets eliminate the interaction or those where the chemical displacement perturbation is greater. Throughout the simulation, these active residues are limited to being

part of the interface, if possible, otherwise they are penalized.

Table 4.2: HADDOCK PA13P1H08 and PA12P3F10 Antigen Energy Scores

| | **HADDOCK Score** | **Cluster Size** | **Electrostatics** | **Desolvation** | **VdW** |
|---|---|---|---|---|---|
| PA13P1H08, Ara h 1 | -105.0 | 6 | -42.270 | 11.361 | 25.704 |
| PA13P1H08, Ara h 2 | -85.3 | 3 | -23.520 | 0.122 | 17.578 |
| PA13P1H08, Ara h 3 | -92.2 | 6 | -45.021 | 12.342 | 24.322 |
| PA12P3F10, Ara h 1 | -101.2 | 6 | -30.717 | 1.349 | 16.702 |
| PA12P3F10, Ara h 2 | -93.5 | 3 | -12.250 | -12.015 | 34.153 |
| PA12P3F10, Ara h 3 | -131.1 | 6 | -32.250 | 12.361 | 25.704 |

## 4.4 ZDOCK

ZDOCK and M-ZDOCK are fast Fourier transform-based protein docking programs written and maintained by the laboratory of Zhiping Weng (ZLAB) at the University of Massachusetts Medical School. ZDOCK functions by searching for all possible binding modes in the translation and rotation space between two proteins and evaluates each pose using an energy-based scoring function [2].

Unforunately, ZDOCK produces data that does not provide information on that can be easily transcribed and thus the binding data cannot be reviewed. However, ZDOCK combined with PyMOL returns a three dimensional strucuture of the binding as shown in the figure below:

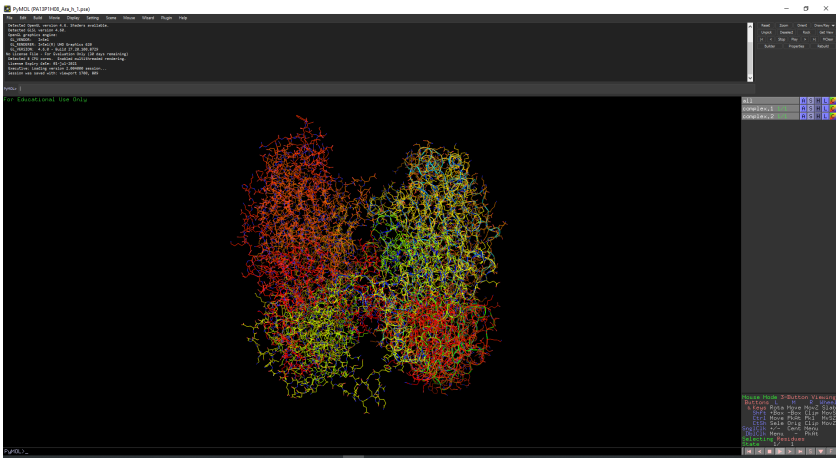


Figure 4.6: PyMOL visualisation from ZDOCK Ab-Ag output

# 5 Results and Discussion

From the experiment conducted, is it clear that both pyDOCKWEB and HADDOCK seemingly produce very similar results in terms of the Electrostatics, Dosolvation, Van der Waals Energy and Desolvation. The other categories cannot be compared as HADDOCK does not have a total or a configuation number included in the output. In terms of pyDOCKWEB The binding between PA13P1H08 and Ara h 1, the configuration given was 5857 with an electrostatics of -45.270 and a Van der Waals energy constant of 24.704. While with the HADDOCK interface, a HADDOCK score of -105.0, a electrostatics of -42.270 and a VdW of 25.704 was recorded. Both these results can be interpreted as both Ab-Ag having a reasonably high binding probability compared to the other combinations shown in both table's. Interestingly, the combination with the least binding probability was PA13P1H08 and Ara h 2. This result suggest that this type of Antibody is not specific to the antigen and would be the least effective in the case of an allergen. Predicting the the Antibody-Antigen binding probabilities is unfortunately a very complex process which was anticipated. The main reason behind this is that most protein structure tools produce a vast array of different outputs. For instance, when comparing pyDOCKWEB to HADDOCK this becomes very clear. While pyDOCKWEB returns several binding information such as the configuration number, electrostatics, desolvation and vdW (Van der Waals energy), HADDOCK also returns RMSD value from the overall lowest energy structure, a restraints violation energy, Buried Surface Area, Z-Score, Cluster size and a HADDOCK score. It also includes a three dimensional view of both Ab and Ag proteins. In addition, the operator of the tool must know which residues from the sequence are active and which ones are passive. This makes it more difficult to compare the different tools with each other even though they are widely recognized as Ab specific Epitope prediction tools. However, both tools offer the possibility to instigate one or more chains from the Ab and Ab PDB file which allows for a more diverse analysis. ZDOCK offered the least amount of features and parameter options as the tool itself is outdated. No information about the binding probabilities could be drawn as only a three dimensional view was available. However, in it's defence, it allowed for constraints to be applied to both the Receptor and Ligand. pyDOCKWEB had a major advantage over its counter parts because of it's ease to use however it's biggest downfall was the lack of other features such as an optional selection of the active and passive sites. In addition, pyDOCKWEB had a limitation on the size of the molecule, namely Ara h 3, and for this reason it was unable to evaluate the binding probability. HADDOCK and ZDOCK do not have such a limitation.

# 6 Conclusion and Future Work

In Conclusion, The Antibody PA13P1H08 produced the highest binding probability when paired with the peanut allergen Ara h 1. This is hardly surprising, as described in the Introduction Section 1, an Antibody binds with a specific antigen due the the shape of the binding site. The interaction between the two functions in the same way a key and a lock functions; each antibody binds to a specific antigen. The analogous "lock" is referred to a Paratope, while the analogous "key" is referred to an Epitope. However, what the results do not show, is if the Antibody PA13P1H08 does indeed have the highest binding probability amongst all Human Immunoglobulin IgM. In order to achieve this, new and innovative methods need to be addressed. A suggestion would be a Deep Learning algorithm that has access to the Protein Data Bank and computes the binding probabilities of as many Antibodies as possible. The user would then input a PDB strucutre of an antigen and have the system predict which Antibody out of the ones that are currently known would provide the highest binding probability. Computational Antibody-Antibody predictions do have their limits however tools such as HADDOCK and pyDOCKWEB can offer molecular biologists platforms for modelling and designing monoclonal antibodies. The world currently faces the SARS-COV-2 and the development of antibodies as therapeutics has never been as important as in modern times. As these computational methods produce more accurate results, their integration into the therapeutic development process is inevitable and will ensure the faster delivery of new drugs and effective drugs to patients.

# Bibliography


[1] A. M. Lesk, *Introduction to bioinformatics.* Oxford university press, 2019. 1

[2] "Zdock server: An automatic protein docking server." [Online]. Available: http://zdock.umassmed.edu/ 2, 20

[3] P. I. Group and Docking. [Online]. Available: https://life.bsc.es/pid/pydockweb/ 2

[4] "Haddock 2.4." [Online]. Available: https://wenmr.science.uu.nl/haddock2.4 2

[5] H. Shinmoto, Y. Matsuo, Y. Naganawa, S. Tomita, and Y. Takano-Ishikawa, "Epitope analysis of peanut allergen ara h1 with human monoclonal igm antibody 92-2," Aug 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2978308/ 2

[6] "Pymol is a user-sponsored molecular visualization system on an open-source foundation, maintained and distributed by schrödinger. we are happy to introducepymol 2.4!" [Online]. Available: https://pymol.org/2/ 2

[7] "Model interactive workspace." [Online]. Available: https://swissmodel.expasy.org/interactive 2

[8] R. A. Norman, F. Ambrosetti, A. M. J. J. Bonvin, L. J. Colwell, S. Kelm, S. Kumar, and K. Krawczyk, "Computational approaches to therapeutic antibody design: established methods and emerging trends," *Briefings in Bioinformatics*, vol. 21, no. 5, pp. 1549–1567, 10 2019. [Online]. Available: https://doi.org/10.1093/bib/bbz095 3, 16, 24

[9] K. Kreth and A. Fodor. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1401/1401.1141.pdf 7, 8, 24

[10] G. DeepMind, "Alphafold 2," 2020. [Online]. Available: https://predictioncenter.org/casp14/doc/presentations/2020_12_01_TS_predictor_AlphaFold2.pdf 9, 11, 24

[11] F. Fuchs, "Fabian fuchs." [Online]. Available: https://fabianfuchsml.github.io/alphafold2/ 11, 24

[12] [Online]. Available: https://medium.com/inside-machine-learning/what-is-a-transformer-d07dd1fbec04 12, 24

[13] "Understanding lstm networks." [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/ 13, 24

[14] "Haddock2.4 antibody - antigen tutorial using pdb-tools web-server." [Online]. Available: https://www.bonvinlab.org/education/HADDOCK24/HADDOCK24-antibody-antigen/ 16, 24

# List of Figures


1.1 The process of DNA Replication. The information is transcripted to mRNA and is eventually translated into a Protein. . . . . 1
1.2 Antibody-Antigen binding illustration. At the binding site, the Paratope (analogous for a lock) binds with one and only one specific type of antigen on its Epitope (analogous for a key) . . . . . 2

2.1 Graphic illustration of a Protein mainchain having different amino acid side chains (R Group) attached at regular intervals . . . . . 4
2.2 Graphic illustration of an $\alpha$-Helix and a $\beta$-Pleated Sheet Protein Structure . 5
2.3 A Ribbon Diagram representation of the Tertiary Structure (Left) and the Quaternary Structure (Right) of a protein. Note how one figure consists of a single chain of amino acids while the other contains at least two. . . . . 6

3.1 Formular for calculating covariance in a local protein sequence alignment [9] . 7
3.2 Distance Map computed for the Ab PA13P1H08 (Left) and the Ag Ara h 1 peanut allergen (Right) using the nanohub.org online protein contact map tool 8
3.3 Graph Depiction of the TM Scores achieved by AlphaFold compared to its counterparts in the CASP13 Assessment . . . . . 8
3.4 Graph Depiction of the TM Scores achieved by AlphaFold compared to its counterparts in the CASP13 Assessment . . . . . 9
3.5 Demonstration of the accuracy achieved by the AlphaFold 2 System represented by the blue structure . . . . . 10
3.6 Mathematical representation of the geometry of a protein [11] . . . . . 11
3.7 AlphaFold 2 Protein Sequencing Architecture showcased during the DeepMind AlphaFold 2 presentation [10] . . . . . 11
3.8 Depiction of a Transformer [12] . . . . . 12
3.9 Depiction of an LSTM neural network [13] . . . . . 13

4.1 . . . . . 15
4.2 CDR loops of of an Antibody [14] . . . . . 16
4.3 Depiction of State of the art Computational antibody methods [8] . . . . . 16
4.4 The top ten models output by pyDOCKWEB . . . . . 19
4.5 Energy Score of the binding of PA13P1H08 and Ara h 1 . . . . . 19
4.6 PyMOL visualisation from ZDOCK Ab-Ag output . . . . . 20

# List of Tables

4.1 pyDOCKWEB PA13P1H08 and PA12P3F10 Antigen Energy Scores . . . . . 18
4.2 HADDOCK PA13P1H08 and PA12P3F10 Antigen Energy Scores . . . . . 20