Exploring the Sequence vs. Binding Relationships for Monoclonal Antibodies and Other Proteins

by

Pritha Bisarad

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved June 2022 by the
Graduate Supervisory Committee:

Neal W. Woodbury, Chair
Alexander A. Green
Nicholas Stephanopoulos

ARIZONA STATE UNIVERSITY

August 2022

ABSTRACT

Molecular recognition forms the basis of all protein interactions, and therefore is crucial for maintaining biological functions and pathways. It can be governed by many factors, but in case of proteins and peptides, the amino acids sequences of the interacting entities play a huge role. It is molecular recognition that helps a protein identify the correct sequences residues necessary for an interaction, among the vast number of possibilities from the combinatorial sequence space. Therefore, it is fundamental to study how the interacting amino acid sequences define the molecular interactions of proteins. In this work, sparsely sampled peptide sequences from the combinatorial sequence space were used to study the molecular recognition observed in proteins, especially monoclonal antibodies. A machine learning based approach was used to study the molecular recognition characteristics of 11 monoclonal antibodies, where a neural network (NN) was trained on data from protein binding experiments performed on high-throughput random-sequence peptide microarrays. The use of random-sequence microarrays allowed for the peptides to be sparsely sampled from sequence space. Post-training, a sequence vs. binding relationship was deduced by the NN, for each antibody. This in silico relationship was then extended to larger libraries of random peptides, as well as to the biologically relevant sequences (target antigens, and proteomes). The NN models performed well in predicting the pertinent interactions for 6 out of the 11 monoclonal antibodies, in all aspects. The interactions of the other five monoclonal antibodies could not be predicted well by the models, due to their poor recognition of the residues that were omitted from the array. Furthermore, NN predicted sequence vs. binding relationships for 3 other proteins were experimentally probed using surface plasmon resonance (SPR). This was done to explore

the relationship between the observed and predicted binding to the arrays and the observed binding on different assay platforms. It was noted that there was a general motif dependent correlation between predicted and SPR-measured binding. This study also indicated that a combined reiterative approach using in silico and in vitro techniques is a powerful tool for optimizing the selectivity of the protein-binding peptides.

DEDICATION

To my father who always believed in me and let me choose my own path.

And to my mother who is my very first and most beloved teacher.

ACKNOWLEDGMENTS

As my journey as a PhD student is about to culminate, I cannot help but feel incredibly grateful to have received the support of so many people who have cheered me on through this journey. First and foremost, I would like to thank my advisor, Prof. Neal Woodbury, for his constant mentorship, support, and motivation. He has always inspired and encouraged me to be curious and learn and try out new things. He has also taught me to always be optimistic and enthusiastic about science, even when things aren't going so well. I am sincerely grateful to him for always being so understanding and considerate. I would also like to extend my sincerest thanks to my committee members, Prof. Nicholas Stephanopoulos, and Prof. Alexander Green, for their valuable guidance and input throughout all these years. They have always been very encouraging and supportive of me.

I am extremely grateful to Laimonas Kelbauskas, Bart Legutki, Chris Diehnelt, and Su Lin for their amazing and valuable mentorship. Whenever I was stuck on a challenging problem, they have always helped me by providing very thoughtful and crucial feedback on my research and pointing me in the right direction. I will always be very grateful to Su for being so caring and looking out for us. Also, I am very thankful to my collaborators from HealthTell (iCarbonX) as well as to Prof. Don Seo and Prof. Hao Yan, for allowing me to work on such diverse research projects. A big thank you to all my teachers who have instilled in me a life-long love and inquisitiveness for science.

I will always be thankful to my friends who have helped me navigate through a tumultuous six years in grad school. My dearest lab mates Akanksha Singh, Kirstie Swingle, and Robayet Chowdhury have always stayed with me through all of the craziness. I will always cherish the endless laughs, dances, debates, and drinks that we have shared.

Cheers to these amazing people for just being so awesome and making grad life so much more fun – Joydeep Banerjee, Anweysha Bhowmik, Stephanie Thibert, Tushar Modi, Soma Chaudhary, B. Sree Ganesh, Srivatsan Mohanarangan, Bharath Sampath, Skanda Vishnu, Shatabdi Roy Chowdhury, Nikita Kumari, Banashree Gogoi, Sohini Mukherjee, Gourango Charan, Saborni Chowdhury, Abesh Banerjee, and Souvik Poddar. I absolutely do not thank Mayukh Nandy for infecting me with Covid. Thanks to my squad from 42 Sikderbagan Street for giving me the courage to pursue this journey.

Heartfelt thanks to my apartment-mates in Tempe for always putting up with my antics. Raghu Pradeep Narayanan has always been a strict but extremely caring older brother to me. Sanchari Saha, with whom I started the journey of grad school, has always been the one to calm me down in case things went downhill. Leeza Abraham will always be the "best roommate" despite strong competition, for providing me with food and transportation, and listening to my rambles. They are my family away from family.

Special thanks to Annesha Lahiri and Shreya Bunk for sticking with me through all these years since school and college, despite the endless fights. I would also like to thank Snigdha Mondal for just being there beside me and keeping me grounded, no matter what.

Finally, I would like to express my deepest gratitude to my parents and grandmother for always believing in me and giving me the courage to believe in myself. I am deeply indebted for the sacrifices that my parents had to make to avail this opportunity to me. Even through difficult times, they have always supported me and never let me give up. And my grandmother will always be my biggest cheerleader. I thank them from the bottom of my heart for their unconditional love and support.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

Molecular recognition is a major driving force behind all the biological processes that are carried out in nature. As stated by Kricka, (1988), over many years nature has evolved to express a plethora of biomolecules that display a magnificent variation and diversity in recognizing other molecules. Many scientists are looking into uncovering the key interactions that drive molecular recognition, to understand biomolecular pathways better. For biomolecular polymers like proteins, molecular recognition is a function of their sequence information. Evolution has enabled the protein molecules to carry out their functions efficiently and specifically, through sparse sampling across the expanse of the sequence space and local optimization of sequences. However, there are regions of the sequence space that have not been explored through evolution, owing to the vastness of this combinatorial space. Though unexplored, these 'landscapes' are not devoid of information relevant molecular recognition. But given that most proteins have a few hundred amino acid residues, the size of this multidimensional landscapes become astronomically large. Thus, exploring this space in its entirety is a physically impossible task.

One thing that is known in case of biomolecular recognition, particularly in the case of proteins, is that only a few key residues play a crucial role in creating a direct interface during molecular recognition events. These residues are often part of one or more short, linear sequences on the protein. There are many tools available in modern biochemistry, including combinatorial methods like library screening, that allow one to identify these key regions in molecular recognition, that are important to the function of the proteins.

Combinatorial methods, such as peptide microarrays, and peptide and protein display methods, especially allow one to screen through a large library to find relevant interactions. However, such methods are generally directed towards searching in a particular region of sequence space (Mimmi et al., 2019, Ullman et al., 2011). But what about sampling the sequence space randomly, in a much broader sense? In 2020, Taguchi and others demonstrated that by training a neural network on binding information obtained from random-sequence peptide microarrays (as little as few thousand peptides), it was possible to derive a quantitative sequence-to-binding relationship between the array peptides and the assayed proteins, which could then be applied to the entire combinatorial sequence space for those peptides. In this previous work, the sequence vs. binding relationship was obtained for relatively weaker binding interactions between proteins and peptides. But can the same approach be used to derive a sequence-to-binding relationship in case of proteins with very high degree of specificity towards their target(s), for e.g., monoclonal antibodies. How much information about molecular recognition of monoclonal antibodies can be obtained by analyzing random peptide sequences? Would it be possible to predict stronger binding interactions by analyzing weaker ones? The aim of this study is to find answers to the questions posed above and therefore gain a better understanding of molecular recognition in proteins, especially antibodies.

1.1 MOLECULAR RECOGNITION AND SEQUENCE SPACE

1.1.1 What is Molecular Recognition?

The term “molecular recognition” can be used to describe a set of specific interactions between two or more molecules, primarily through non-covalent interactions. The non-covalent interactions may consist of hydrogen bonding, van der Waals forces,

coordination with ions, hydrophobic forces, π-π interactions etc. (Nano-Inspired Biosensors for Protein Assay with Clinical Applications, 2019). These forces along with the three-dimensional structural orientation of the participating molecules aid in forming a molecular complex. The assembling of these complexes is highly dependent on the complementarity between the interacting molecules. Some examples include host-guest interactions, molecular self-assemblies, and supramolecular interactions (Gellman, 1997). Molecular recognition plays an important role in chemistry and biology alike. In biology, molecular recognition is integral to carrying out almost all essential functions in a living organism (Kricka, 1988). In nature, a vast number of biomolecules have evolved to achieve molecular recognition with high-specificity. Receptor-ligand, self-assembly of nucleic acids, nucleic acid-protein/peptide, enzyme catalysis, protein-protein/peptide, small molecule-protein, and antibody-antigen interactions are all examples of biomolecular recognition.

1.1.2 The Concept of Sequence Space

One of the most striking features of any biomolecular recognition event is its dependance on the shape and chemical nature of the surface of the interacting molecules. In case of biomolecules like proteins or nucleic acids, the shape and chemical nature are a function of the sequence composition (Rebek, 2009). In proteins, the amino acid residues that compose the polypeptide chains are mainly responsible for determining the three-dimensional shape of the molecules. Therefore, the amino acid composition of proteins determines the shape, and therefore the function of the protein. This leads to the discussion about sequence space. The concept of sequence space was first introduced by John Maynard Smith (Smith, 1970) in which he compared the protein sequences to a word game.

A sequence of amino acids was considered to be a word made up of letters. The arrangement of the letters (amino acids) in that sequence would comprise a protein. Exchanging any letter with another would represent a change in the composition of amino acids, therefore indicating a mutation at that point. The combinatorial sequence space for a peptide or protein of certain length is represented by all possible combination of the 20 naturally available amino acids in that given length. Thus, for a polypeptide chain with N number of amino acids residues, the total number of unique sequences with the 20 canonical amino acids found in most proteomes is equal to 20^N.

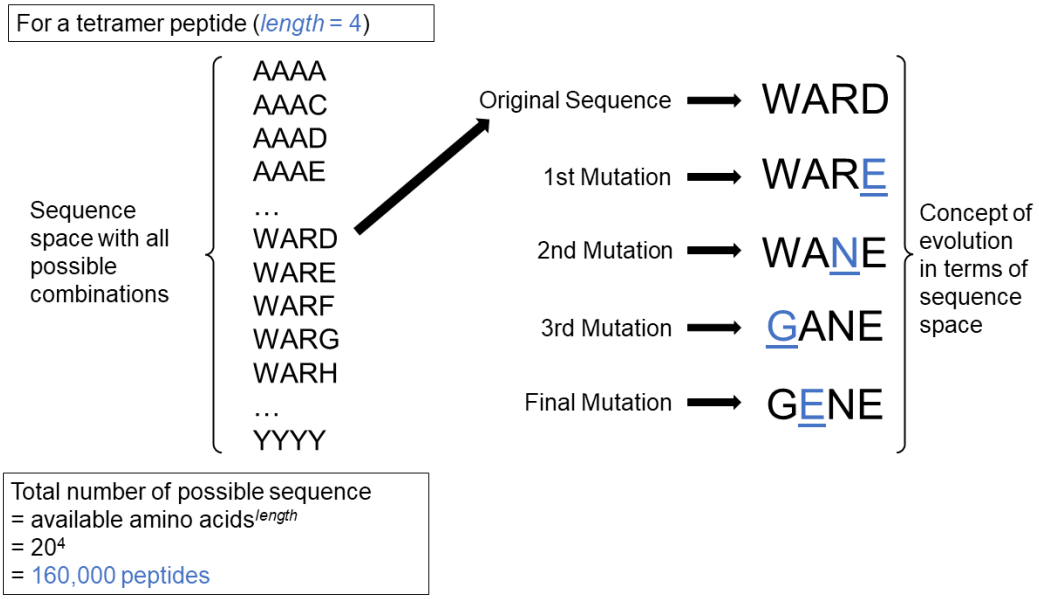


Figure 1.1. Schematic diagram showing the concept of combinatorial sequence space. Single letter representation of the 20 amino acids that are found in the genetic codes of humans have been used here. It also represents the concept of evolution in the context of sequence space. The blue underlined letters represent the mutated residues at each step.

Figure 1.1 explains the concept of combinatorial sequence space as well as the definition of evolution within a sequence space. It also shows the total number of possible peptides. The 'distance' between any two sequences in the combinatorial space is often measured as a Hamming distance. For example, in Figure 1.1, the sequence 'WARE' is 1

Hamming distance away from the original sequence, ‘WARD’. Similarly, the Hamming distances of the sequence ‘GANE’ is 3 from the sequence ‘WARD’ and 1 from the sequence ‘WANE’.

It can be calculated from this that the actual observable sequence space represented by the proteins is much smaller than the number of possible representations in the combinatorial sequence space. Not all of the possible sequence combinations are explored over the course of evolution. Rather, only changes (mutations) that maintain the functions of the protein or produce advantageous effects are retained (Clarke, 1970; Mirny et al., 1998; Gustafsson, 2001). Sequences which reduce the optimal functioning capacity of the proteins are eliminated. However, the combinatorial sequence space of a protein is a multi-dimensional landscape (the number of dimensions is directly proportional to the length of the peptide). Therefore, it can be expected that there are unexplored pockets in this combinatorial sequence landscape which will still conform to the structural and functional characteristics of the protein, resulting in molecular recognition. A thorough understanding of the sequence space landscape with respect to protein activity will lead to deriving accurate sequence-to-binding correlations. Understanding molecular recognition of proteins in context of sequence space would be essential if one were to study and apply the molecular interactions of proteins in various context. Through this work, it was explored if one can identify the contributing residues responsible for molecular recognition of monoclonal antibodies and other proteins, by studying sparsely sampled random peptide sequences (4 - 13 mers) from an extensive combinatorial peptide sequence space (10^{12} peptides).

1.2 PROTEINS AND ANTIBODIES

1.2.1 Introduction to Proteins

Proteins are highly complex polymeric biomolecules that mediate almost all essential biological functions in living organisms such as respiration, metabolic processes, immune responses, response to stimuli etc. (Lord et al.,1988; Getzoff et al., 1988; Terwilliger, 1998; Gosline et al., 2002; Sidhu et al., 2003; Chockalingam et al., 2007; Zhao et al., 2010). The term 'protein' was first used by a Dutch chemist, Gerard Johann Mulder, in 1838, following the suggestions of another chemist, Jöns Jacob Berzilius (Vickery, 1950). Proteins can be defined as polymers of different amino acids linked via α-peptide bonds (Watford and Wu, 2018). Amino acids are the basic building blocks that make up and constitute the proteins. They are organic molecules that contain a carboxyl (-COO) group and an amino ($-NH_2$) group. In α-amino acids, which are the primary monomeric units for most proteins found in nature, the carboxyl group and the amino group are attached to the same carbon atom (α-carbon). This carbon is also attached to another varying functional group that can be either polar or non-polar. There are 20 amino acids commonly used in natural proteins with different functionalities that constitute almost all of the proteins. A peptide can be defined as a chain of different amino acids linked together via an amide bond (also known as peptide bond, $-CONH_2$). This sequence of amino acids in a peptide chain that forms the primary structure of the protein. As the peptide chains get longer, they are folded together in three-dimensional space (secondary, and tertiary structures). Thus, the functionality and binding preferences of a protein are determined by the specific arrangement of different amino acids within a peptide chain.

As mentioned previously, the mediation of biological processes by proteins relies heavily on molecular recognition events. Thus, a detailed understanding of the molecular recognition between different proteins and their targets is essential to understand biological processes in-depth. Molecular recognition in proteins is generally dependent on their shape, which in turn is dependent on the sequence of amino acids. Depending on the function, proteins can bind to a variety of molecules like other proteins, smaller peptides, nucleic acids, and small organic molecules. The amino acid sequences determine the binding characteristics of a protein. In order to bind to a target, the binding site on the protein must spatially complement that of the target's. Such complementarity ensures that the interactions remain largely specific to the target molecule of the protein. However, often there are molecules in the combinatorial space that partially or entirely mimic the binder's physicochemical characteristics and therefore are recognized by the proteins. Comprehensive knowledge of such interactions would accelerate the development of therapeutics, diagnostics, and other protein-based applications. High-throughput combinatorial approaches, both in vitro and in silico, are a great way to study such molecular interactions thoroughly (Fout et al., 2017; Johansson-Åkhe et al., 2018; Xavier et al., 2016; Gabernet et al., 2019; Rose et al., 2003; Mimmi et al., 2019; Tiwari., 2016; Ulman et al., 2011). The efficacy of such an approach is always dependent on the extent and characteristics of combinatorial space that has been sampled. In 2020, Taguchi et al., described a neural-network based approach to predict molecular recognition behavior for nine different proteins using information captured from random peptides that were sparsely sampled from the combinatorial sequence space (elucidated later). To study how molecular recognition can be described by utilizing sparsely sampling from combinatorial sequence

space, one needs a model protein system which undergoes specific interactions with their targets and that can be well-characterized. Antibodies, which are discussed in the next section of this chapter, are an excellent example of such a system. Aside from studying the interactions of the antibodies, the binding interactions of three other proteins (diaphorase, ferredoxin, and ferredoxin-NADP reductase or FNR) were also probed and compared between different assaying platform as well as using predictive approaches.

1.2.2 Introduction to Antibodies

Antibodies are a special class of proteins that are part of the humoral immune system (Burnet, 1957). They are sometimes referred to as immunoglobulins (Igs) and are secreted by the B-cells. Antibodies were one of the first proteins from the immune system that were characterized. The target-binding region of the antibodies, also known as the variable (V) region, varies extensively. Thus, this enables the B-cells to produce a huge repertoire of antibodies that can recognize a broad variety of pathogens. This, alongside their specificity towards their targets, makes the antibodies an excellent candidate for investigating the principles of molecular recognition.

All antibodies are made from pairs of heavy (H) and light (L) polypeptide chains. These chains combine to form a constant region (C) which is the stem of the Y shaped molecule, and the variable (V) region which is the target binding region (Figure 1.2). These polypeptide chains are connected with disulfide bridges. Both the heavy chains and the light chains have constant (C) and variable (V) domains. The H chains consist of three C domains and one V domain (C_H1, C_H2, C_H3, V_H). The L chains consist of one C domain and one V domain (C_L, V_L). The variable regions from the heavy chains and the light chains (V_H and V_L) form the target-binding site. There are five different classes (or

isotypes) of immunoglobulins that are typically found in human blood – IgG, IgM, IgA, IgD, and IgE. Of these, the IgG is the most abundant isotype found in the sera samples. Therefore, from now on in this thesis, the word antibody refers to this isotype in particular. Each of the V_H and V_L domains contain three particularly variable loop segments known as hypervariable regions. When the V_H and the V_L domain are together, these total of six hypervariable regions at the tip of the antibody are commonly known as complementarity-determining regions (CDRs). CDRs from both light-chain and heavy-chain variable regions are responsible for identifying the target (molecular recognition).

The target of an antibody is also known as an antigen. Antibodies can recognize and bind to a variety of molecules including proteins, peptides, carbohydrates, nucleic acids, and small molecules. In the context of this thesis, binding interactions refer to either antibody-peptide interactions or antibody-protein interactions. The region on the antigen protein surface which is recognized by the antibody for binding is known as the epitope.

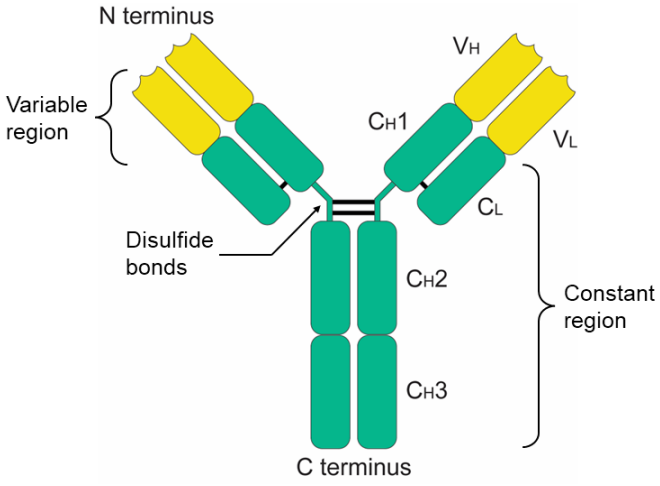


Figure 1.2. Schematic representation of an IgG. The subscripts H and L refer to the heavy and light chains respectively. C and V refer to constant and variable domains respectively.

Protein epitopes are sometimes described in terms of their cognate sequences, and here in this thesis they will be used interchangeably henceforth. Epitopes can be broadly classified into two different classes, linear, and conformational. Linear epitopes are composed of a single and continuous segment of a polypeptide chain. Hence, they are also known as continuous epitopes. On the other hand, conformational epitopes are small fragments on the polypeptide chains that are not continuous due to the three-dimensional folding of the chains. Therefore, they are known as discontinuous epitopes as well. Interactions between the antibody and the epitope region are governed by the electrostatic forces primarily. Sometimes, hydrogen bonding, Van der Waals forces, hydrophobic forces, and pi-interactions also play a role in the molecular recognition. Most antibodies recognize conformational epitopes, as they are raised against an intact antigen. However, many of them also recognize continuous peptide fragments, that may be the entire epitope itself or part of the conformational epitope.

1.2.3 Monoclonal Antibodies

Generally, the antibodies produced from B cells (polyclonal antibodies) as a result of a natural immune response are polyreactive and have heterogeneous specificities, also known as cross-reactivity (Frank, 2002). This heterogeneous binding behavior can be a potential limitation for using antibodies for various purposes (diagnostics, therapeutics etc.). Monoclonal antibodies are a special class of antibodies that have a very specific and known antigen specificity. In 1975, scientists Georges Köhler and César Milstein devised a novel method for the production of monoclonal antibodies with predefined specificity using hybridoma techniques (Köhler and Milstein, 1975). This was achieved by fusing antibody secreting mouse spleen cells (with high specificity towards a single antigen) with

mouse myeloma cells that produced no antibody by themselves and had the ability to grow indefinitely, to make a set of immortal hybridoma cell lines. Thus, the antibodies obtained from the resulting hybridomas were highly specific towards the intended target antigen. The high antigen specificity of the monoclonal antibodies has since been exploited by various research applications in the fields of immunology, biotechnology, biochemistry, therapeutics, and applied biology (Ansar and Ghosh, 2013). As therapeutic agents monoclonal antibodies have been used in the treatment of cancers, auto-immune disorders, and cardiovascular diseases (Brennan et al., 2010; Beck et al., 2010; Ansar and Ghosh, 2013; Kaplon et al., 2020). Other applications include biosensors, microarrays, purification, and imaging.

Although monoclonal antibodies have a higher specificity compared to polyclonal antibodies, they are not devoid of cross-reactivity (Flores-Moreno et al., 2014; Vojdani et al., 2021). Monoclonal antibodies may bind to peptide fragments on the antigen that are partially similar to the actual epitope sequence. Sometimes, they might bind to fragments that have completely different amino acid sequence than the epitope but mimic its physicochemical properties. Such peptide fragments are called mimotopes (Geysen et al., 1986). Whether to use an antibody for therapeutic purposes or for clinical and diagnostic applications, one must understand and characterize both on-target (cognate sequences) and off-target (near-cognate sequences, mimotopes) interactions thoroughly (Brennan et al., 2010; Uhlen et al., 2010; Ansar and Ghosh, 2013; Norman et al., 2020; Kaplon et al., 2020).

1.3 TOOLS TO STUDY MOLECULAR RECOGNITION

Using sequence information to determine the binding interactions of proteins is an essential part of modern molecular biology. There are many tools are available to

characterize different types of protein-protein interactions (PPIs). These tools can be broadly classified into experimental and computational techniques. Each technique has its own merits and challenges. Before delving into the techniques that were used for this thesis, some of the most commonly used techniques are discussed in the following paragraphs.

1.3.1 Experimental Tools

There are diverse categories of experimental tools that are available to characterize the interactions of the proteins. Methods like Förster resonance energy transfer (FRET), X-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electron microscopy (cryo EM), aim to understand the structural aspects of molecular recognition (Russell et al., 2004). Although these methods provide a comprehensive understanding of the protein interactions, they are not high-throughput in nature. Often at times, they are also time-consuming and cost-ineffective. There are other techniques that are not as detailed as the above methods, but nevertheless provide valuable information on the molecular interactions of the proteins (Nealon et al., 2017). A lot of them are high-throughput and are reliant on omics or combinatorial approaches. High-throughput methods have gained popularity over the years due to the ability to test large number of samples at once. There are several high-throughput techniques such as tandem affinity purification (TAP), surface plasmon resonance (SPR), mass spectrometry (MS), immunoassays like enzyme linked immunosorbent assays (ELISAs), high-throughput protein/peptide microarrays, and display systems like phage and bacterial displays (Sidhu et al., 2003; Schweitzer, 2003; Berggard et al., 2007; Nealon et al., 2017; Miura, 2018). These techniques, and many more, are widely used for studying the molecular recognition of many proteins, including antibodies.

1.3.2 Computational Tools

Although experimental methods are highly effective, sometimes they demand a lot of investment in terms of time and resources. Computational methods (also known as in silico methods) seek to address some of these challenges. Just like experimental techniques, there are several computational approaches as well to study molecular recognition. Computational methods are useful in narrowing down possibilities, from the enormous number of the entities in the combinatorial space (Colwell, 2018) which makes them a great tool to be used in extension with experimental methods. Machine learning algorithms have made these approaches even better over time, with the ability to process even larger sequence and binding datasets, and better prediction capabilities. According to Lim et al. (2022), machine learning is a method of data analysis that allows "machines" (computers) to learn and extract patterns from huge amounts of collected data and make predictions, accordingly. Therefore, machine learning is being widely used in computational analyses that intend to address molecular recognition challenges. With the emergence of deep learning algorithms, more people are trying to computationally solve the complex relationship between protein sequence and molecular recognition. Some of these approaches rely on structural modelling and prediction of structures of protein complexes (Russell et al., 2004; Nealon et al., 2017). Molecular docking is an example of such an approach where they try to predict and identify the residues present on the surfaces of interacting proteins. Docking strategies rely on the assumption that out of a number of possible orientations for an interaction between a protein and its target, the native orientation will be scored higher than the others. Although not entirely accurate, there are many docking algorithms that have achieved better prediction results over the years

(Russell et al., 2004; Nealon et al., 2017). Docking algorithms are however sometimes not feasible due to intensive use of computational resources. Another type of computational approach utilizes sequence alignments and binding motif discovery algorithms to identify molecular recognition interactions (Mohamed, 2016). Examples of such algorithms include SLiMDisc, STREME, and motif-x (Davey et al., 2006; Chou and Schwartz, 2011; Bailey, 2021). These motif discovery algorithms rely on a set of sequences that are related functionally. Given a set of sequences, these algorithms try to find *de novo* motifs that might not be defined by a formal criterion. Some of these algorithms use an alignment based method whereas others search for patterns which appear to be overrepresented in the provided set of sequences. Many of the other approaches rely on machine learning based algorithms (both supervised and unsupervised) to find answers related to molecular recognition of proteins by probing their evolutionary fitness landscapes (Freschlin et al, 2022). These approaches allow one to systematically search the sequence space represented by the fitness landscape, thus allowing them to derive sequence-binding relationships.

When it comes to predicting interactions of the antibodies specifically, much work has been done to use computational algorithms (including neural network) to predict their epitopes. Most approaches focus on predicting linear epitopes due to the structural complexities involved with conformational epitopes. Early attempts at predicting the linear epitopes of antibody relied on propensities and physicochemical properties of the amino acids (Hopp and Woods, 1981). Later, many more novel routines, for e.g., PREDITOP and BEPITOPE, followed similar approaches, increasing the variety and number of propensities used (Kolaskar and Tongaonkar, 1990; Pellequer and Westhof, 1993; Pellequer et al., 1993; Odrico and Pellequer, 2003). A study conducted by Blythe and

Flower (2005) showed that using propensity scales might not be the best approach in determining antigenic regions for antibodies. Instead, more sophisticated approaches are required. As machine learning algorithms started gaining popularity, Saha and Raghava (2006) developed the first neural network-based linear epitope prediction program, ABCPred (Potocnakova et al., 2016). Following suite, many other hidden Markov models (HMM), support vector machines (SVM), and neural networks based routines were developed that attempted to predict epitopes. Notable examples include BepiPred, BCPred, FBCPred, CBTOPE, and many other (Larson et al., 2006; EL-Manzalawy and Hanovar, 2010; Zhang and Niu, 2010; Ansari and Raghava, 2013). The training datasets of these predictors included a few thousand epitopes. Also, most of these models were trained on benchmark datasets that were obtained from IEDB (www.iedb.org, Vita et al., 2009) or similar databases. In 2017, a linear epitope predictor, DRPEP, which was based on deep neural networks was developed by Sher et al. It allowed for the prediction of linear epitopes of variable length and applied the predictor to entire protein sequences. There are also mimotope-based epitope predictors. These programs utilize mimotope discoveries from phage display experiments. In these approaches the mimotopes of the antibodies are mapped on to the overlapping patches on the antigen using sequence alignment and other statistical methods. Examples include MIMOX, MimoPro, and Pep-3D-Search (Potocnakova et al., 2016; Huang et al., 2008; Chen et al., 2012; Huang et al., 2011).

Some of the notable examples of in silico molecular recognition was covered here but this is not an exhaustive list. A plethora of many other such tools exist that seek to address the challenge of molecular recognition and provide a better understanding of it. However, the predictive capabilities of most of the computational approaches are limited

due to the unavailability of unbiased benchmark datasets. In the following section, a different approach that employs sparse sampling of peptide sequences from combinatorial space and uses that library for training a neural network to predict the binding interactions will be discussed. This approach was first used by Taguchi et al. (2020) to predict the molecular binding interactions of different proteins.

1.3.3 Molecular Recognition and Random Sequence Peptide Microarrays

Here in this section a new approach to identify the molecular recognition interactions of proteins will be discussed. This approach utilizes binding data acquired from random-sequence peptide microarrays to train a neural network that is later used to predict binding to peptides with amino acid sequences that are not present on the array. In order to comprehend the approach better, one needs to know what random-sequence peptide microarrays are. Microarrays are high-throughput tools that have been around since early 1990s and are frequently used to infer molecular recognition information from studying omics related features (Heiss et al., 2020). There are many different types of microarrays, for e.g., DNA, RNA, proteins, and peptides. Peptide microarrays consists of hundreds to thousands of peptides that are discretely arranged on a solid support (e.g., glass slides) (Meng et al., 2018). They are usually more stable chemically and easier to synthesize than full protein microarrays. Peptide microarrays have a variety of applications in the field of biochemistry and medicine, ranging from basic research o clinical diagnostics (Meng et al., 2018).

There are many different types of peptide microarrays available. Although these microarrays are useful for a variety of applications, they mostly do have an inherent bias towards specific target(s) (Richer et al., 2015), e.g., representing the tiled proteome of a

specific pathogen or a number of proteins. Random-sequence peptide microarrays as the ones used in this study are fully agnostic in their nature and have no bias towards any particular protein as the peptides on the array have been sampled nearly randomly from a combinatorial sequence space (~10^{12} peptides). These photolithographically synthesized arrays contain 126,050 unique peptide sequences that cover about 83% of the possible tetramers and more than 27% of all possible pentamers (Legutki et al., 2014; Richer et al., 2015). These arrays use an alphabet of 16 amino acid residues (A, D, E, F, G, H, K, L, N, P, Q, R, S, V, W, Y) to construct peptides whose length varies from 5 to 13 amino acid residues. 4 of the amino acids (M, I, T, and C) were excluded from the arrays due to limitations in the synthetic process. Thus, these peptide sequences are a very sparse and unbiased representation of the entire combinatorial sequence space for peptides with a median length of 9 amino acids. These arrays have been used for comprehensive health monitoring, diagnosis of different types of cancer and other diseases (immunosignatures), as well as for epitope identification of different antibodies (Legutki et al.,2010; Restrepo et al., 2011; Halperin et al., 2011; Hughes et al., 2012; Kukreja et al., 2012; Stafford et al., 2014; Legutki et al., 2014; Richer et al., 2015). However, the current sequence space covered by these arrays restricts them to linear sequences only, thus having limited utility for structural epitope binding.

In 2020, Taguchi and others used these random-sequence high-density peptide microarrays to predict sequence vs. binding relationship for diaphorase and other proteins with the help of a neural network. In the above stated work, nine different proteins (Diaphorase, Ferredoxin, FNR, PD1, PDL1, TNFα, TNFR, Transferrin, crystallizable fragment of an IgG) were fluorescently labeled and assayed on the random-sequence

microarrays. The sequence information along with the binding data from the assays were then used as an input to train a shallow feed-forward, backpropagated neural network. The sequence information of the peptides was encoded using 'one-hot encoding'. The neural network was then allowed to "learn" and recognize the binding patterns for these proteins. The results indicated that by training the model only on a few thousand sequences chosen randomly from the combinatorial sequence space one can derive a comprehensive predictive relationship between sequence and binding for the proteins that can be applied to the combinatorial sequence space in general. Not only was the neural network successfully able to identify the binding patterns of the proteins, but the predictions were specific to each protein. More importantly, this kind of predictive relationship is not dependent on structural knowledge available for the peptides or the proteins. This work demonstrated that it is possible to define the sequence vs. binding relationship for different proteins across sequence space by exploring only a small subset from it that has been sparsely and randomly sampled.

1.4 PROJECT OVERVIEW

The work described in this dissertation is an attempt to further explore the sequence vs. binding relationship of proteins, primarily monoclonal antibodies, using neural networks and high-density random-sequence microarrays. This study hopes to look into the utility of the approach in identifying binding targets of well-defined monoclonal antibodies and assessing the ability to predict and validate binding partners for yet uncharacterized proteins. The motivation for the study presented here is based on the work done by Taguchi et al. (2020) (Section 1.3.3). In order to characterize the contributions of combinatorial sequence space to protein molecular recognition interactions thoroughly, a

model protein system was required whose interactions were highly diverse, but specific, and easy to characterize. Monoclonal antibodies are very well suited to these criteria. Their interactions are known to be specific towards the target antigen(s) and are well-characterized by different experimental tools. Having such a model system allows for assessing the models performance with respect to the known interactions of the proteins. In this work, the sequence vs. binding relationship of eleven monoclonal antibodies (DM1A, p53Ab1, p53Ab8, 4C1, LNKB2, 9E10, rho-1D4, 3B5, AU1, Btag, and Htag) with known epitopes have been studied using a neural network model structure, similar to that of Taguchi et al. (2020). All of the eleven monoclonal antibodies were assayed on random-sequence microarrays with same sequence space representation (126,050 unique peptides), following the protocol laid out by Rowe et al. (2017).

In the first study using the five antibodies mentioned above, the neural network model was optimized to adapt to the binding interactions observed in case of monoclonal antibodies, as opposed to that of other proteins, through hyperparameter optimization. The predictive performance of the model was then evaluated based on the binding patterns recognized by the neural network models, and they were compared to the known cognate epitopes. It was also tested if the algorithm could predict strong binding of the epitope sequences among a library of a million random peptide sequences, sampled from the combinatorial sequence space. The specificity of the predictions with respect to each antibody was also probed. An in silico mutagenesis experiment on the epitopes of respective antibodies was carried out to show which amino acid residues were deemed important by the model in each case. Furthermore, the performance of the model was assessed with respect to different physicochemical propensities that were used as encoders

for the amino acid, and varying concentrations of the monoclonal antibodies which were used for the assays.

In the next study, the antibody-specific neural network models were projected on the antigen sequences of the five monoclonal antibodies that were studied, as well as on to the entire human proteome. This was done to observe how well the model performs when exposed to evolutionary sequence space as opposed to random sequences from the combinatorial sequence space.

In the following chapter, the molecular recognition of the monoclonal antibodies (9E10, rho-1D4, 3B5, AU1, Btag, and Htag) was studied. These six monoclonal antibodies were studied separately because their epitopes were not represented on the microarrays. The reported epitopes of four (out of six) of these antibodies also contains some of the residues that are omitted on the microarrays (M, I, T, C), which also resulted in distinctly different binding behavior in the assays, compared to the previous group of antibodies. Also, not all of them have human target antigens. The modeling and assessment approaches are similar to those carried out in the initial study.

Additionally, in the last study, the predictive performance of the model described by Taguchi et al., was experimentally tested by selecting three proteins from the study (diaphorase, ferredoxin, and FNR). Peptides predicted to be high, low, and mid-range binders, from a combinatorial library, for each of these three proteins were synthesized. The binding abilities of these peptides then to the respective proteins were estimated using surface plasmon resonance (SPR). The underlying aim behind this study was to determine how are protein-peptide interactions affected by changes in binding environment (microarray vs. SPR). As the neural network was trained on data from the microarray

experiments, what differences are observed when one attempts to translate the predicted interactions using another, orthogonal assaying platform?

Overall, the goal of these studies was to attempt and characterize the neural network based approach in terms of its ability to predict well-defined interactions of monoclonal antibodies and understand its potential limitations. The role of the neural network here is more akin to that of a pattern finding tool, which parses the sequence information available to it, to distinguish between specific and non-specific interactions in each case, through multiple iterations. What is more impressive is that one is able to predict stronger and highly specific binding interactions by analyzing weaker binders. While combinatorial studies are not a new field in science, it is interesting to observe how much information one can gather about molecular recognition just by looking at interactions with arbitrarily and sparsely sampled sequences from the sequence space. Hopefully, in the future, this work will serve as a useful stepping-stone for those who wish to combine combinatorial approaches with machine learning to investigate the enigmas of molecular recognition.

CHAPTER 2

USING NEURAL NETWORKS TO DERIVE SEQUENCE VS. BINDING RELATIONSHIP FOR MONOCLONAL ANTIBODIES WITH KNOWN BINDING TO MICROARRAY PEPTIDES

This work was initiated in collaboration with Akanksha Singh

2.1 INTRODUCTION

The term 'sequence space' often comes into play when one is describing interactions between different types of proteins, peptides and/or nucleic acids. The concept of sequence space was first introduced in 1970 (Smith, 1970) who called it 'protein space'. It can be described as the number of possible amino acid sequences, for a protein or peptide of a given length. For example, for a peptide with 10 amino acid residues, the sequence space consists of 20^{10} (or 10,240 billion) possible sequences. As the length of the peptide/protein gets bigger, the possible combinatorial sequence space becomes immensely huge. However, over the years of evolution, proteins have evolved to carry out biological functions efficiently while sampling only a small fraction of this space (Clarke, 1970; Mirny et al., 1998). It shows that finding the correct set of amino acid sequences from a large and vast landscape is of utmost importance. Thus, sequence space plays an integral part in any kind of molecular recognition event involving protein molecules. In order to study the role of sequence space and its contribution to molecular recognition events, a model protein system was required whose molecular interactions are of diverse nature and could be easily characterized.

Antibodies make a good example of such a system as they have high specificity towards their targets. Antibodies are a class of proteins that are produced by the B-cells of

the immune system. They play a very important role in the humoral immune response (Fagraeus, 1948; Burnet, 1957). The molecular and structural diversity of antibodies make them very versatile binders for a broad variety of targets (Peng et al., 2014). Because of the diversity and specificity of their binding interactions, antibodies have found broad applicability in the field of proteomics research, diagnostics, and therapeutics. They have consistently remained one of the most rapidly growing class of therapeutics (Uhlen et al., 2010; Stadler et al., 2013; Norman et al., 2020; Kaplon et al., 2020). Monoclonal antibodies are a special class of antibodies that are highly specific as they are raised against a single target (Nelson et al., 2000). Their interactions with their cognate sequences are very well characterized. This makes them highly desirable in the field of therapeutics and drug development, where small molecule-based approaches have not been sufficient (Graves et al., 2020). Many of the recent uses of monoclonal antibodies in the field of therapeutics have been related to the treatment of cancer and autoimmune disorders (Brennan et al., 2010; Beck et al., 2010; Kaplon et al., 2020). Due to the ever increasing demand of monoclonal antibodies in the field of research and therapeutics, significant effort has also been put in to optimizing and characterizing them (Clementi et al., 2013).

The most important step before utilizing a monoclonal antibody for any application is to characterize its interaction with the antigen/pathogen and discover the binding sites on them, which are called epitopes. It is essential to have the knowledge of antibody-antigen interactions, especially the interacting epitopes for the development of immunodiagnostic assays and tests, vaccines, and therapeutic antibody treatments (Uhlen et al., 2010; Shirai et al., 2014). Epitopes are a set of amino acid residues on the antigen/pathogen to which the antibody binds through molecular recognition. Epitopes can

be either linear or conformational. A linear epitope is a short and continuous peptide sequence on the antigen, whereas a conformational epitope is comprised of two or more such linear peptide sequences. The conformational epitopes are a result of the folding of the protein in 3-dimensional space (Barlow et al., 1986) and are most common in nature. There are several methods to characterize the epitope of an antibody. One of the most commonly used techniques is X-ray crystallography (Clementi et al., 2013) which helps determine the crystal structure of the antibody-antigen complex. Other than X-ray crystallography, one can also use NMR (Zuiderweg, 2002), cryo-EM (Fibriansah et al.,2015), mass spectrometry (Huang & Chen, 2014), ELISAs (Brennan et al., 2010), mutagenesis (Kowalsky et al., 2015), and display of different peptides on bacteria (Rockberg et al., 2008) and phage (Peterson et al., 1995). Although these methods are well established, they can be quite expensive and time consuming. Some of these methods are also low on accuracy and throughput. Also, these methods do not always provide a thorough understanding of the molecular recognition in the context of sequence space. To address some of these issues, the use of computational approaches has been on the rise in the past few years. High-throughput computational methods that employ machine learning, like motif discovery algorithms and epitope mapping suites, offer a fast, highly scalable and cost effective solution, due to their ability of handling vast amounts of sequence data. Hence much research has taken place in recent years to develop sophisticated computational methods that can assist or substitute existing experimental methods in understanding of the molecular recognition of monoclonal antibodies and mapping their epitopes (El-Manzalawy & Honavar, 2020; Manieri et al., 2020; Norman et al., 2020; Graves et al.,2020).

Though most of the antibody binding regions or epitopes are estimated to be conformational, majority of the computational approaches focus of finding linear sequences due to the structural complexities involved with the identification of conformational epitopes (Sanchez-Trincado et al.,2017). Most of the early approaches that sought to look into the molecular recognition of monoclonal antibodies and attempted to predict their epitopes were based on propensity indices that assigned a set of values to each amino acid residue on the basis of physicochemical properties like hydrophobicity, steric factors, and antigenicity among others (Hopp & Woods, 1981; Kolaskar & Tongaonkar, 1990; Pellequer & Westhof, 1993; Pellequer et al., 1993). These assigned set of values were then used to recognize a set of residues that would potentially be an epitope for an antibody. These approaches only looked at a couple of antigens for the identification of binding interactions, making the size of their training dataset very small. Therefore, the accuracy of their prediction was low (Blythe & Flower, 2009). Over time, the algorithms became more and more sophisticated, and many started employing machine learning approaches due to their largely automated feature extraction as opposed to manually electing each and every feature. Saha and Raghava developed the first machine learning model known as ABCpred in 2006 that utilized a combination of propensity indices and sequence complexity to recognize patterns in different types of linear epitopes, and therefore make predictions. After ABCpred, many other methods were developed in an attempt to predict linear epitopes correctly from sequence-based information. Many of them used used a support vector machine (SVM) based approach which utilized multiple propensities and inputs derived from sequences (Chen et al., 2007; El-Manzalawy et al., 2008; Wang et al., 2011; Gao et al., 2012). However, most of these methods were using

small set of peptides for training the algorithm (approx. 1500 peptides) which limited their performance (Sher et al., 2017). Later, another study (Singh et al., 2013) employed a SVM model along with K-nearest neighbor, where they were able to utilize a new dataset with much higher number of peptides (>30,000) which they obtained from IEDB (Vita et al., 2018).

The study that employed the first deep learning approach in this context was by Lian et al. (2015) who implemented and trained a deep maxout network (DMN) with dropouts. Their approach utilized the same dataset as used by Singh et al. (2013) and a slight increase in the performance of the classification was reported. As the predictive models are improving over time, the differences in the training datasets used by these models makes their general applicability a challenge. Also, these classification approaches use a known set of epitopes and non-epitopes for training the models which introduce an inherent bias to the algorithms during training. In this study, a different approach is presented to address the current limitations of predicting molecular recognition. This predictive approach is not dependent on the availability of benchmark dataset or structural knowledge of the target sequences. This makes the algorithm largely unbiased and versatile in predicting binding patterns for monoclonal antibodies.

For this work, a combined approach utilizing neural networks and high-throughput microarrays that enable sparse sampling from the combinatorial space has been laid out to characterize the molecular recognition of monoclonal antibodies. This combined approach was first explored in a study by our research group (Taguchi et al., 2020) where a neural network was implemented to characterize the binding behavior of different proteins on the microarrays. A simple feedforward neural network was developed to derive a quantitative

sequence vs. function relationship between the proteins and the array peptides. Once trained, this predictive model was then used to predict the relationship of the protein molecules over the combinatorial sequence space. It demonstrated that a sparse library of peptides (126,050 peptides) generated by randomly sampling over a combinatorial sequence space (10^{12} peptides), could be used to characterize the binding interactions of different proteins. As these peptides on the microarray are nearly randomly sampled from the available sequence space, there is little inherent bias to the library. Also, one does not need to have any structural knowledge about the target molecules beforehand. Although the interactions of the proteins were characterized by this method, monoclonal antibodies cannot be generalized in the same way due to the highly specific nature of their interactions. This study aims to shed light on the molecular recognition behavior of 5 different monoclonal antibodies whose epitopes are well-characterized. Given the highly specific binding nature of the monoclonal antibodies, will it be possible to characterize the binding interactions just from the information obtained from a sparse library of randomly sampled peptides? In other words, can the sequence and binding information available from weaker binders of the monoclonal antibodies be used to predict more specific interactions like that of the cognate sequences.

The 5 monoclonal antibodies that were chosen for this study are DM1A, p53Ab1, p53Ab8, 4C1, and LNKB2. These five monoclonal antibodies were selected because their binding to the sequences present on the peptide microarray were known and well-characterized. Although monoclonal antibodies are highly specific to their epitopes, they are prone to showing cross-reactivity with other non-cognate peptide sequences present on these high-throughput arrays (Stafford et al., 2012; Notkins, 2014; Horwacik et al., 2015).

The aim is to exploit these off-target interactions to derive a model that can correctly predict a set of target residues in a vast combinatorial space. Experiments carried out on the high-throughput microarrays with random sampling of peptides over the combinatorial space (Legutki et al., 2014; Rowe et al., 2017) are the source of data used for training the model. The advantage of using such a microarray based peptide library over methods like phage display is that they are faster, unbiased, less expensive, and allow direct measurement of peptide binding (Halperin et al., 2011). All of the peptides on these microarrays are between 4 to 13 residues in length. Experimental data captured from assaying fluorescently labeled antibodies on the microarray (Rowe et al., 2017) was used as the input for the neural network based on the work by Taguchi et al. (2020). The output in terms of predicted binding intensity was used to derive a comprehensive predictive relationship between the amount of monoclonal antibodies bound on the array and the peptides. It is to be noted that all data regarding the epitopes of the respective antibodies were omitted while training the model. The difference in the number of peptides on the array (126,050 unique peptides) and the number of total possible sequences (10^{12}) is huge, which is roughly about one peptide out of every 10 million. The predictive relationship obtained only from sampling such a small fraction of the combinatorial space can then be projected on larger libraries to check the accuracies of the prediction.

2.2 METHODS

2.2.1 Synthesis of High-throughput Peptide Microarrays

Peptide microarrays with 126,050 unique peptide sequences were synthesized using a photolithography-based approach (Legutki et al., 2014; Rowe et al., 2017) at HealthTell (http://www.healthtell.com/). A 200 mm silicon wafer coated with thermal-oxide was first

functionalized with aminosilane monolayer to create attachment sites for the peptides. This surface was then coated with *tert*-butylcarbonyl-glycine (BOC-glycine). A photoresist containing a photoacid generator was applied next to the wafer through spin coating. Further, exposing the wafer to a UV light source at 365 nm after applying a defined photomask, resulted in the deprotection of the BOC-protected amines at specified features on the array. A coupling solution consisting of BOC-protected amino acids was then spin-coated on the wafer. This process ensured that the coupling took place only at the deprotected features on the array and marked the completion of one cycle. The cyclic process was repeated several times to add amino acids at the N-terminus of the peptides. It created a combinatorial sequence space of peptides ranging in length from 3-13 amino acid residues (median length is 9). After completion of all the cycles, the wafer was cut into 13 rectangular pieces, each with the dimension of a microscope slide (25 mm X 75 mm). Each slide contained 24 identical arrays (8 rows and 3 columns) with 126,050 unique peptides on each array. The quality of the array manufacturing process was ensured by the characterization of arrays through MALDI-MS. The slides were stored in a dry nitrogen environment, post-manufacturing. Since these arrays were produced in a highly-automated fashion, the reproducibility is very high (Taguchi et al., 2020).

The peptide sequences synthesized on the arrays were pseudo-randomly generated using an algorithm designed to reduce the number of synthetic steps while covering a predetermined percentage of the desired sequence space. They use only 16 of the 20 naturally available amino acids (A, D, E, F, G, H, K, L, N, P, Q, R, S, V, W, Y). Cysteine, methionine, isoleucine, and threonine were excluded from array manufacturing process in order to simplify the synthetic procedures. Nevertheless, the sequence space represented

by the 16 amino acids used on the arrays were sufficient to cover the interactions of the monoclonal antibodies that were selected.

2.2.2 Labeling of the Monoclonal Antibodies and Assays

For this study, five different mAbs were used (DM1A, p53Ab1, p53Ab8, 4C1, and LNKB2). These antibodies were sourced from a mouse host. Table 2.1 contains detailed information about the sources and the targets of the mAbs. Binding assays on the array were performed following an ELISA-based format ((Rowe et al., 2017). The arrays were first soaked in distilled water for 1 hour to rehydrate them. They were then treated with PBS for 30 minutes, followed by 1 hour in primary incubation buffer (1% mannitol, PBST). They were again rinsed with distilled water to remove excess salt build-up. The rehydrated arrays were then loaded into custom cassettes to adapt them into a 96-well plate format. All the mAbs were serially diluted in the primary incubation buffer according to Table 2.2 and added to the arrays. After adding the mAbs, the arrays were incubated at 37°C for 1 hour with mixing, to facilitate antibody-peptide interactions. Array-bound mAbs were labeled using goat anti-mouse IgGs with Alexa Fluor 555 (Invitrogen, catalog #A21424) in a secondary incubation buffer consisting of 0.5% casein in PBST. The mixture was incubated for 1 hour at 37°C. Post-incubation, the arrays were washed 3 times in PBST, followed by distilled water using a BioTek microplate washer (BioTek Instruments, Inc., Winooski, VT). After removing them from the cassette, isopropanol was sprayed on them. Then they were centrifuged for drying. The arrays were then imaged using an ImageXpress Micro XLS (Molecular Devices, San Jose, CA) fluorescence imager for 375ms of exposure. The images were then processed to align the fluorescence measurements with corresponding peptides using Mapix software (Innopsys, Carbonne, France). Each assay

had two technical replicates. The correlations between two replicates are also mentioned in Table 2.2.

Table 2.1. Sources and Target Sequences for the Monoclonal Antibodies

Monoclonal Antibody	**Isotype**	**Supplier**	**Catalog Number**	**Target Protein (UniProt ID)**	**Epitope Sequence**
DM1A	IgG1	Millipore Sigma	05-829	Human α-tubulin (Q71U36)	ALEKDYE
p53Ab1 (clone PAb240)	IgG1	Millipore Sigma	CBL404	Human cellular tumor antigen p53 (P04637)	RHSVV
p53Ab8 (clone BP53-12)	IgG1	Invitrogen	MA1-19055	Human cellular tumor antigen p53 (P04637)	SDLWKL
4C1	IgG2a	GeneTex	GTX47974	Human thyrotropin receptor (P16473)	LQAFDSH
LNKB2	IgG1	Absolute Antibody	Ab00232-1.1	Human interleukin-2 (P60568)	PLEEVLN

Table 2.2. Serial Dilution of the Monoclonal Antibodies and Correlation of Replicates

	Correlation between technical replicates				
Concentration (pM)	**DM1A**	**p53Ab1**	**p53Ab8**	**4C1**	**LNKB2**
16000	N/A*	N/A*	0.991	N/A*	N/A*
8000	0.99	0.968	0.991	0.99	0.987
4000	0.987	0.983	0.842	0.987	0.997
2000	0.992	0.983	0.996	0.992	0.999
1000	0.994	0.959	0.997	0.994	0.004**

500	0.998	0.985	0.996	0.998	0.999
250	0.997	0.996	0.998	0.997	0.999
125	0.997	0.997	0.996	0.997	0.976
62.5	0.996	0.994	0.951	0.996	0.977
31.25	0.989	0.981	0.96	0.989	0.999
15.625	0.987	0.966	N/A*	0.987	0.998
7.81	0.974	0.862	N/A*	0.974	0.898

* N/A – Not applicable, concentration outside the range of dilution. **Technical replicates not considered during training the model due to poor correlation

2.2.3 Neural Network Model Architecture for Prediction

To explore the molecular recognition of monoclonal antibodies, a simple feed-forward neural network (NN) was used based on a previous work by Taguchi et al (2020). This NN model was used to quantitatively predict the sequence vs. activity relationship for a given set of peptides binding to a particular monoclonal antibody. The relative fluorescent intensities (0 to 65536 relative fluorescence units) associated with each peptide on the microarray were processed beforehand and transformed into a $\log_{10}$ scale for enabling them as inputs to the NN model. A value of 10 was added to intensities before log transformation to offset the fluctuations of lower binding intensities near the zero end in the log scale. The peptide sequences were (in silico) stripped of the GSG linkers at the C-terminal, before being used as inputs for training the model. Sequences shorter than five amino acid residues were also removed from the dataset. The remaining sequences were then represented using "one-hot encoding". Each peptide sequence was represented as a sparse binary matrix of dimension 13 × 16 (length of the longest peptide on the array × number of amino acid residues used on the array).

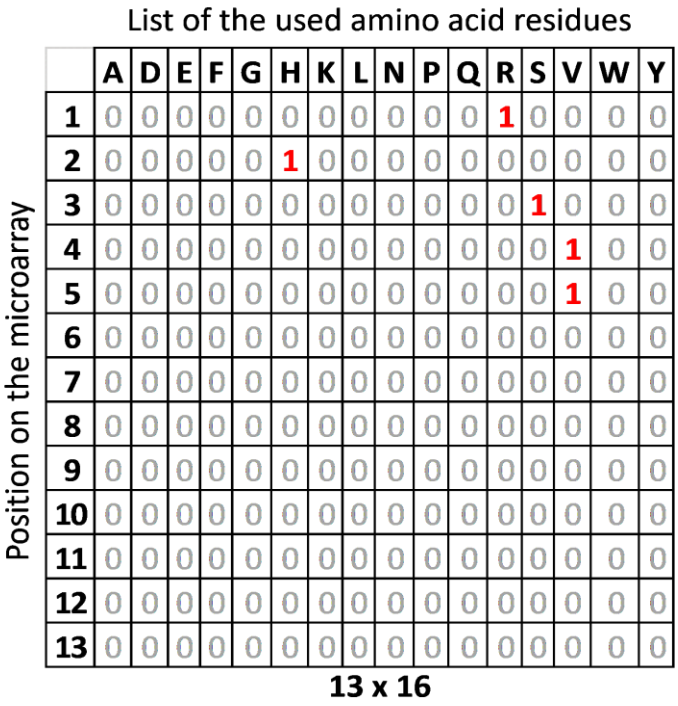


Figure 2.1. Representation of sparse binary matrix for one-hot encoding. The peptide represented here as an example is "RHSVV". The amino acid residue present in the corresponding position of the peptide is shown here in bold red. The rows represent the maximum peptide length on the microarray (13).

An example of the representation can be seen in Figure 2.1 where the peptide **RHSVV** is represented in its binary form. The activation function used for each layer except for the output layer was rectified linear unit (ReLU; Nair et al., 2010). The weights of the neural network were optimized using an Adam optimizer (Kingma et al., 2014). To account for the skewed distribution of the binding intensities, all the values and corresponding peptides have been divided into 100 equidistant bins and a peptide from each bin is randomly selected into batches of 100 peptides during training.

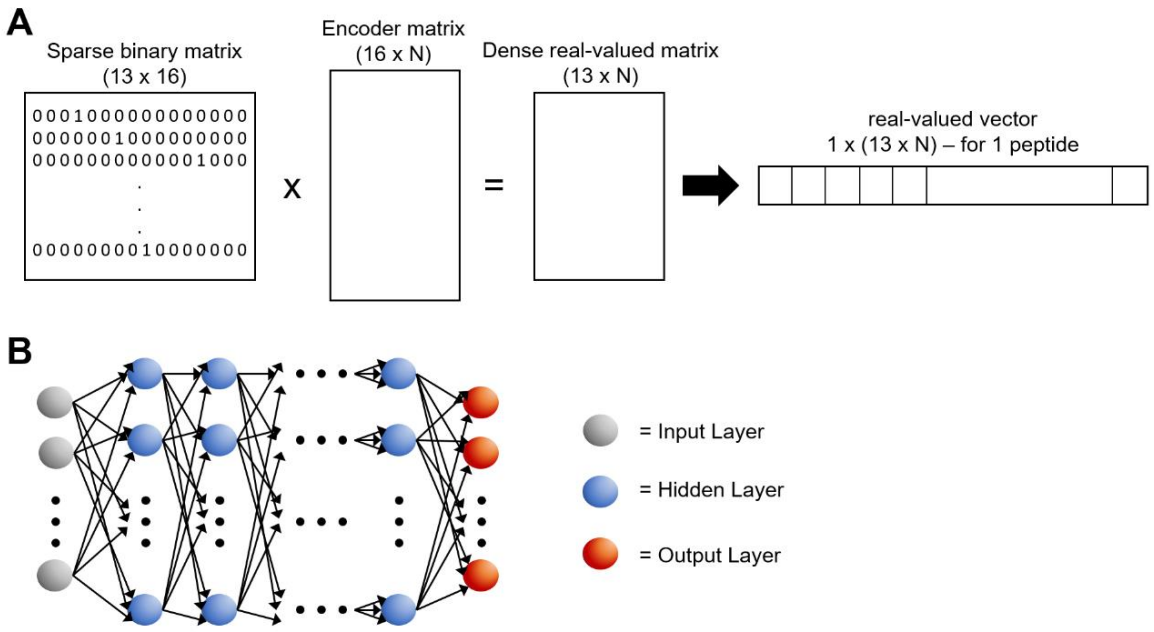


Figure 2.2. Schematics of the neural network architecture used for predicting epitopes of monoclonal antibodies. (A) Conversion of the sparse binary matrix into a dense real-valued vector for representing a peptide. (B) The real-vectors are then fed into the neural network as the input layer. All the hidden layers have rectified linear unit as activation.

An encoder matrix of 16 × N (where N is the number of descriptors for each amino acid; N<= 16), which is learned through training, was then multiplied with the sparse 13 × 16 matrix to linearly transform it into a 13 × N dense matrix with real-valued representation. This dimensionality reduction is driven by the first layer of the neural network, which also the encoder layer. It is during this process that the algorithm learns from the training process and generates optimized weights for the learned encoder matrix. Therefore, the value of the encoder matrix varies during each fitting by the model. By learning the weights through training, the encoder matrix is optimized to preserve the information about the sequences as closely to the observed binding as possible. Once transformed, each peptide's 13 × N matrix was flattened into a single vector representation. These flattened vectors were then used as input features to train the model (Figure 2.2).

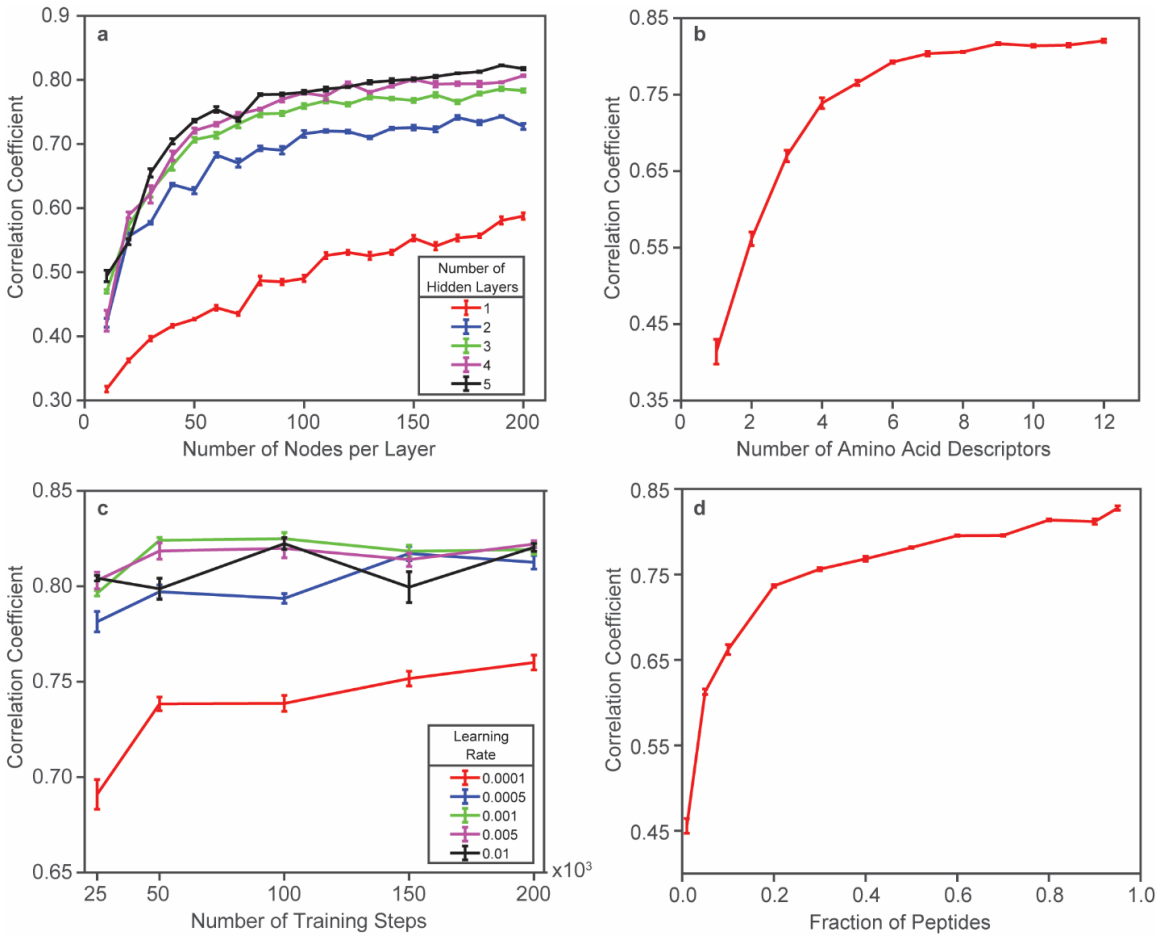


Figure 2.3. Neural network model performance optimization. Variation of the Pearson correlation coefficient (PCC) with respect to various hyperparameters for optimization of the model for DM1A monoclonal antibody binding data. Hyperparameters that were tested include the number of hidden layers and the number of nodes in each hidden layer (a), number of descriptors for each amino acid residue in a peptide (b), learning rate and number of training steps (c), and fraction of peptides from the dataset used for training the model (d). Points plotted are the mean of 5 independently trained models with randomly chosen peptides for training and testing of the model and the error bars represent the standard error of the mean (SEM). The correlations are based on the test set sequences only.

Next, the hyperparameters of the model were optimized. The hyperparameters that were optimized are the number of hidden layers, number of nodes in each hidden layer, number of descriptors for each amino acid, number of training steps, the learning rate, and the fraction of peptides from the dataset used for training. The parameters were optimized in a grid-like manner and Pearson correlation coefficient (PCC), calculated between the

predicted and the measured binding values, was used as the determining factor to evaluate the performance of the model. All the hyperparameter optimization was done using the dataset of the monoclonal DM1A, which is raised against α-tubulin. Note that the cognate sequences of each mAbs were removed from the training set while training the model and were placed in the test set to avoid biasing the model. Results from the hyperparameter optimization are shown in Figure 2.3. Table 2.3 summarizes the selected hyperparameters.

Table 2.3. List of Hyperparameters chosen from Optimization of the Model

Hyperparameters	**Optimized Value**
Number of Hidden Layers	5
Number of Hidden Nodes per Layer	200
Number of Amino Acid Descriptors	9
Learning Rate	0.001
Training Steps	50,000
Fraction of Peptides used for Training	0.95

After optimizing the hyperparameters, the model was trained and validated 100 times, randomizing the selection of training peptides each time, for each mAb. All the neural network models were developed using PyTorch 1.4.0 using Python 3.7 as an interpreter. Note that most of this computational analysis was done on a single workstation with 20 cores which took about an hour for 10 individual fits when a parallel batch approach was implemented.

2.2.4 Testing Model using Randomly Generated In silico Peptide Arrays

For validating the predictions further, the performance of the model was projected onto in silico libraries of 10^6 randomly generated unique peptides. All the peptides were nonamers (9 residues) and were composed of the 16 amino acids present on the microarray. The epitope of the mAb being analyzed was also included in these in silico libraries. It

must be noted here that the models were trained without the epitope sequences being present in the training set. The weights from all the trained models for a monoclonal antibody were projected on these libraries using a MATLAB code and corresponding binding values were obtained. The predicted binding values were then sorted in descending order of Z-score and the peptides were ranked accordingly. The ranking of the epitope in this sorted list was used as a measure for the model performance. Also, the top non-epitope sequences with Z-score value above 3 were selected from this list and were analyzed using STREME motif analysis tool (Bailey, 2021), and the most significant saliency logos with the lowest p-values were represented. Same thing was done with the top binders among the array peptides as well. The Z-score was calculated according to the equation below.

$$Z_i = \frac{X_i - \mu}{\sigma}$$

Here Z_i is the score of the i-th peptide, X_i is the measured binding intensity of the i-th peptide, μ is the mean binding intensity of the entire dataset, and σ is the standard deviation in binding data.

To test the specificity of the predictions, the difference between the binding values with respect to measured data as well as predicted data were calculated, for all of the common peptides for a pair of monoclonal antibodies. The difference was plotted as a scatter plot with the difference in measured values on the x-axis, and the difference in predicted values on the y-axis. All the differences shown here are in the $\log_{10}$ scale.

An in silico mutation experiment was also devised to study the variation in predicted binding due to mutation of the epitope sequences. Each residue of a cognate sequence was replaced with a different amino acid residue from the 16 residues used on

the array, one at a time and serially. The resulting in silico library of substituted peptides is always one Hamming distance away from the cognate epitope. The weights of the trained models were projected onto this library and predicted binding intensities were obtained. The binding intensity (in $\log_{10}$ scale) of the cognate sequence was subtracted from the binding intensities (in $\log_{10}$ scale) of the substituted peptides and the results were plotted as a heatmap.

2.2.5 Using Physicochemical Propensities as Amino Acid Encoders

In the next few steps of the study, the effects of using different propensity indices of amino acids as encoders were studied. A total of 17 different type of indices were used for the study. These propensities acted as the descriptors for each of the amino acids instead of self-learned parameters. A fixed matrix of 20x9 with values between 0 and 1 was used as a control among the supplied encoders. 10 independent training runs were carried out for each selected propensity. All the other hyperparameters used remained same as shown in Table 2.3. The trained models were projected on to in silico libraries similar to the method laid out in section 2.2.4 of this chapter.

2.2.6 Effects of Antibody Concentration on the Model Performance

Next, the change in the predictive model with respect to concentration of the monoclonal antibody was considered. All the available concentrations for each antibody were used to individually train the neural network (Table 2.2). 5 independent training runs were carried out for each available concentration. Post-training, these models were projected onto in silico libraries of randomly generated peptides (10^6) for evaluation of performance (similar to section 2.2.4). For fitting multiple concentrations simultaneously, the binding data from the concentrations were compiled together, so that there would be

multiple columns instead of one. The model treated these columns as different samples and fitted accordingly. The model was trained 5 times independently using each dataset with multiple columns corresponding to different concentrations.

2.3 RESULTS

2.3.1 Sequence vs. Binding Relationship for the Monoclonal Antibodies

In this study, the binding behavior of five monoclonal antibodies was probed utilizing a feedforward, back-propagated neural network, based on the work done by Taguchi et al (2020), to map their sequence-to-binding relationship. The immunoassays with the fluorescently-labeled monoclonal antibodies were carried out at HealthTell and fluorescence readings from the imager were processed using an array alignment software (see Methods; section 2.2.2). The processed dataset was then used for training the neural network.

Table 2.4. List of known cognate sequences present on the array for the antibodies

Monoclonal Antibody	Cognate Sequences on the array (total copies)
DM1A	AALEKDY, ALEKDYE, LEKDYEE (482)
p53Ab1	RHSVV, RHSVVV (311)
p53Ab8	SDLWKL, SDLWKLL (345)
4C1	LQAFDS, QAFDSH (200)
LNKB2	PLEEVLN (100)

The first step was optimizing the hyperparameters of the NN (Methods; Figure 2.3 and Table 2.3). The hyperparameters were optimized using the dataset of DM1A because the binding profile of this antibody indicates a broader coverage of the entire dynamic range of peptides as compared to other mAbs used in the study. Once the hyperparameters were finalized, the model was independently trained 100 times for each monoclonal antibody using randomly chosen training and test sets from the available dataset during

each training. 95% of the available peptides (~119750 peptides) were randomly chosen for training the NN model. The remaining 5% (~6300 peptides) were included as the test set for cross-validating the model. The known cognate sequences of the monoclonal antibodies present on the array (Table 2.4) were deliberately removed from the training set and included in the test set to avoid biasing the model's prediction. The training was carried out 100 times to maximize the probability of the occurrence of each array peptide in the test set. The results and the sequences from every iteration were accumulated. Approximately 700 peptides did not end up appearing in the combined test set from the 100 runs, for each monoclonal antibody. From the combined test set, all the unique peptides were identified and their predicted and measured binding data ($\log_{10}$ scale) were averaged. The mean predicted values were then plotted against the mean measured values as a scatter plot for each antibody, as shown in Figure 2.4. The cognate sequences are also plotted on the graph for each antibody. All the monoclonal antibodies shown here (Table 2.4) have very well characterized epitope sequences (Breitling & Little,1986; Stephen et al, 1995; Shepherd et al, 1999; Afonin et al, 2001).

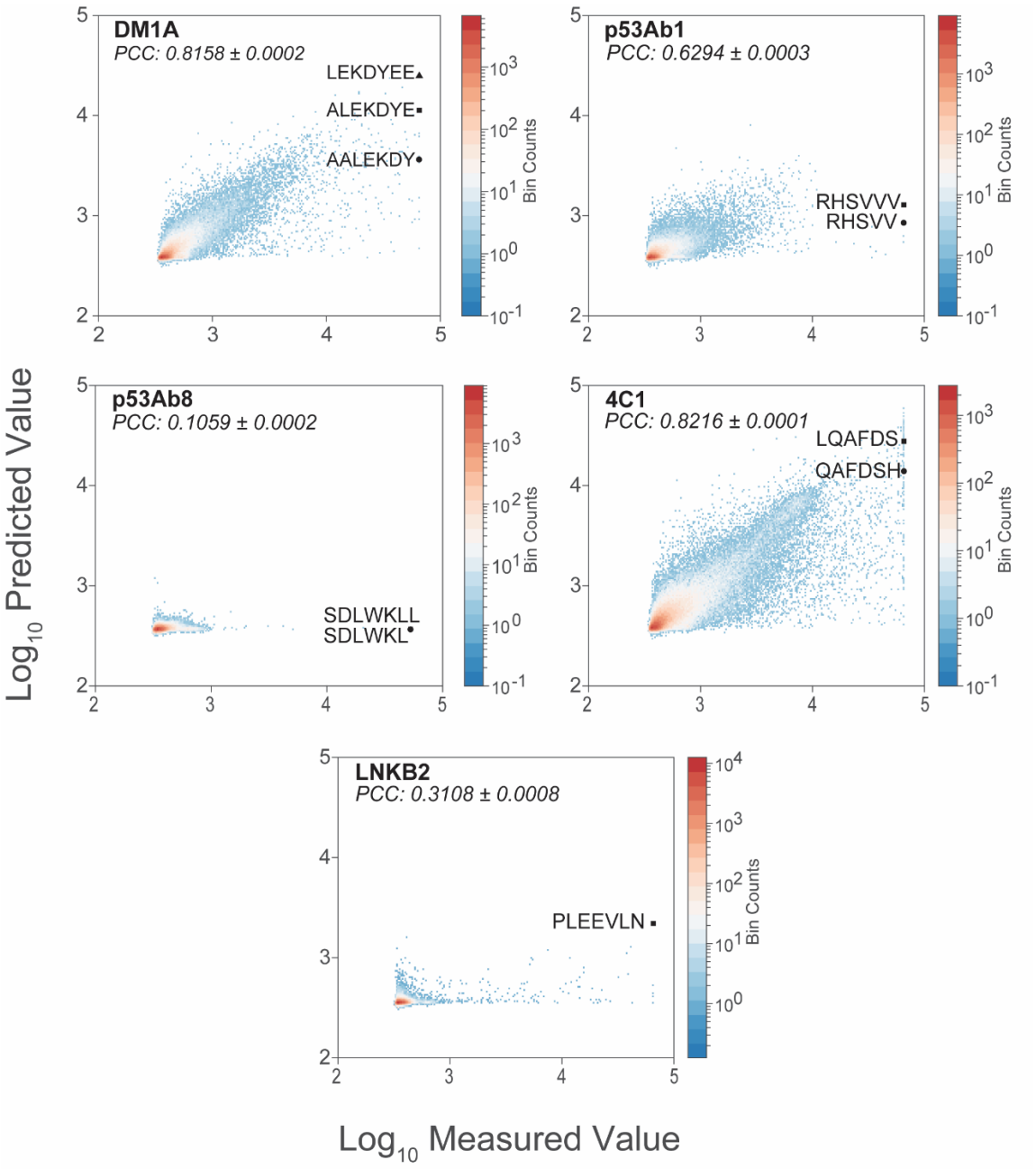


Figure 2.4. Scatter plots showing the correlation between $\log_{10}$ values of predicted binding measurements (y-axes) vs. actual binding measurements (x-axes) of peptides from the test datasets of monoclonal antibodies DM1A, p53Ab1, p53Ab8, 4C1, and LNKB2. The epitopes of each monoclonal antibody are also represented in the respective plot. The density of datapoints (peptides) is color-coded as the number of peptides per datapoint.

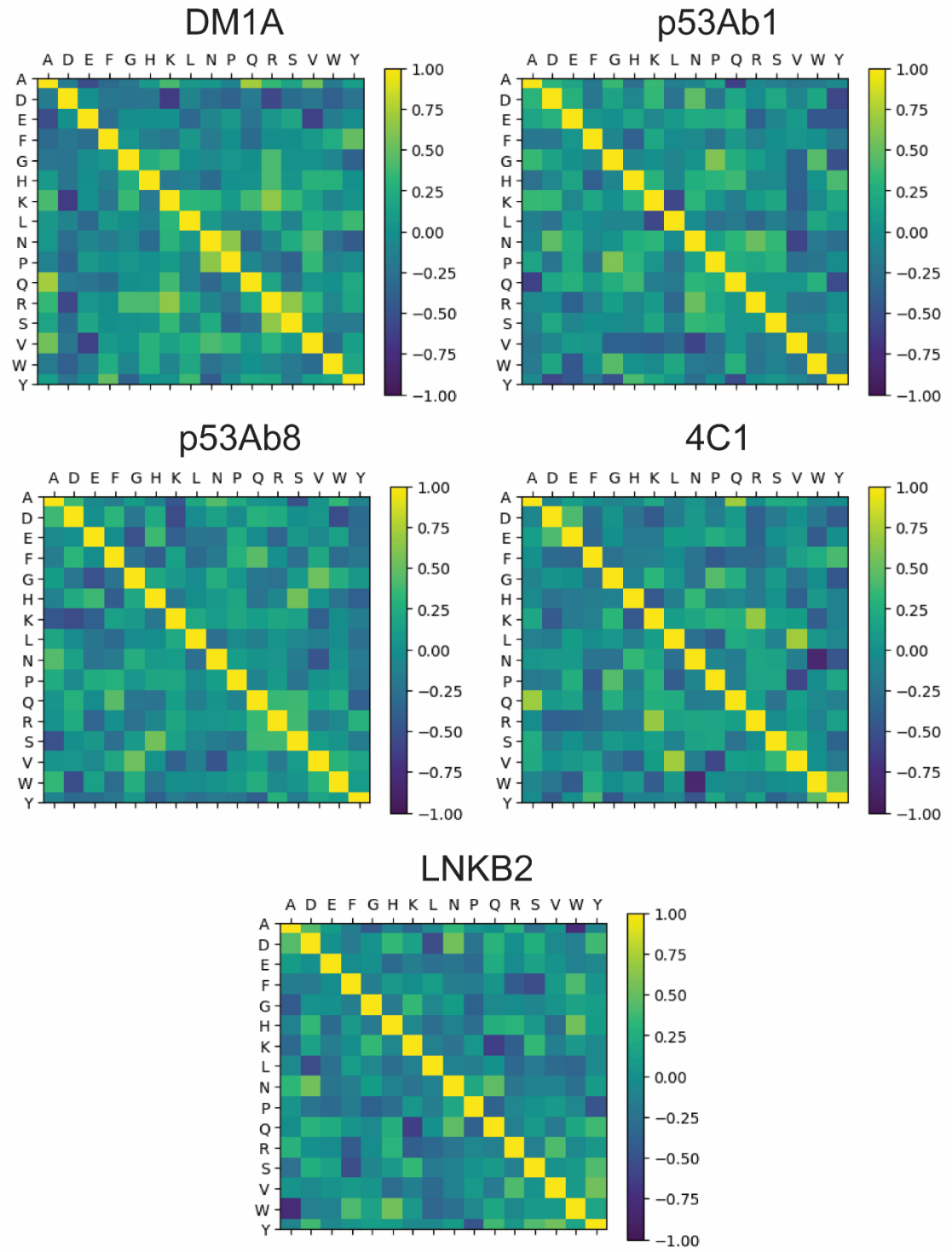


Figure 2.5. Heatmaps demonstrating the similarities between each pair of amino acid vectors, for each individual monoclonal antibody. These vectors were learned by the neural network model during training. The heatmaps were generated by calculating cosine of the angle between the vectors. The data shown here represents the average of 100 independent trainings. The number of descriptors used per amino acid were 9.

As can be seen in Figure 2.4, most of the peptides (>80%) are weak binders (<=3 on the x-axes) to the monoclonals and interact non-specifically with them. Thus, it is only a very small fraction of the peptides that interact with the antibodies above the noise cut-off that is responsible for defining the predicted sequence space. However, this small subset of peptides (<1000 peptides) must contain a diverse set of information in terms of peptide sequences and distribution of measured binding intensity, in order for the algorithm to predict the sequence space with a higher degree of accuracy.

However, the performance of the predictions also depends on the interacting nature of the monoclonal antibodies themselves. From the scatter plots it is seen that the epitopes of the antibodies which showed broader distribution range of binding intensities on the array (DM1A, 4C1) were predicted better. Antibodies that are highly specific and bind mostly to their cognate sequences only (p53Ab1, p53Ab8, and LNKB2) do not show such diversity in interacting with the peptides, with most datapoints in the noise range ($<10^3$) and hence the predicted binding values for the cognates of these antibodies were much lower than the observed values.

As the encoder values for all the 16 amino acids were learned during the training of the model, it varied from run to run, for any given antibody. The average values of these encoders were used to generate heatmaps that show the similarities between the amino acids, for every monoclonal antibody. The similarities were calculated as the cosine of the angle between the amino acid vectors (the dot product divided by the vector magnitude). In these heatmaps, the yellows represent the highest similarity between each amino acid pair, whereas blue represents the lowest similarity between the pair. These similarity

matrices show how similar or opposite the residues are to each other with respect to the interactions of the antibody.

Table 2.5. The mean rank of epitopes within libraries of 10^6 random 9-mer peptides

Monoclonal Antibodies	**Epitopes Included**	**Average Rank (20 iterations) ***	**Percentage rank in 1 million**	**Percentage rank of the epitope on the array ‡**
DM1A	ALEKDYE	23 ± 0.21	0.0023%	0.0156%
p53Ab1	RHSVV	1,727 ± 142	0.1727%	0.0467%
p53Ab8	SDLWKL	481,348 ± 9,382	48.1348%	0.0914%
4C1	LQAFDSH	359 ± 1	0.0036%	0.0078%
LNKB2	PLEEVLN	3 ± 1	0.0003%	0.3197%

* Error shown here is the standard error of the mean of the ranks
‡ Ranks calculated on the basis of the number of unique binding values as measured by the instrument, for each monoclonal antibody.

However, the predicted binding intensity is not the only factor to determine the accuracy of these predictions. Therefore, to further analyze the performance of the models, another method was employed that involved projecting the optimized weights onto in silico peptide libraries. Using MATLAB, a library of 10^6 random nonamer peptides was generated in each iteration and the epitope sequence of the monoclonal antibody being analyzed was included in it. Nonamers were used because the median length of peptides on the microarray is 9. It is to be noted that the sequence space for a nonamer peptide consists of nearly 68.7 billion peptides if one only uses the 16 amino acids used on the array. Therefore, the probability that any 2 randomly generated libraries with 1million peptides would be the same is extremely low ($<10^{-5}$). The optimized weights from all the 100 runs for the antibody were projected onto it. After the projection was complete, this

random peptide library was sorted in descending order of predicted binding values. The rank of the cognate sequence(s) was determined from the sorted peptide list. 20 iterations of this process were carried out and the mean rank of the cognate sequence from 20 different peptide lists was calculated. The results were tabulated and shown in Table 2.5. The mean values and the errors have been rounded off to the nearest integer wherever possible. As can be seen from Table 2.5, except for p53Ab8, the epitopes of all the other antibodies are predicted within top 0.2% of the in silico libraries.

Additionally, the sequences with Z-score above or equal to 3, for each monoclonal antibody were selected from representative libraries. These peptides were then analyzed using STREME motif analysis tool from the MEME suites (Bailey, 2021). The resulting motifs which were most significant in terms of p-value are shown in Figure 2.6. These logos are a good representation of the most prominent amino acid residues that are recognized and deemed important by the neural network model with respect to the binding behavior of a particular monoclonal antibody. Figure 2.6 also shows the most significant sequence logos that represents the binding motifs as observed on the arrays. These sequence logos were also created using the STREME motif analysis tool, from the array peptides with Z-score above or equal to 3.

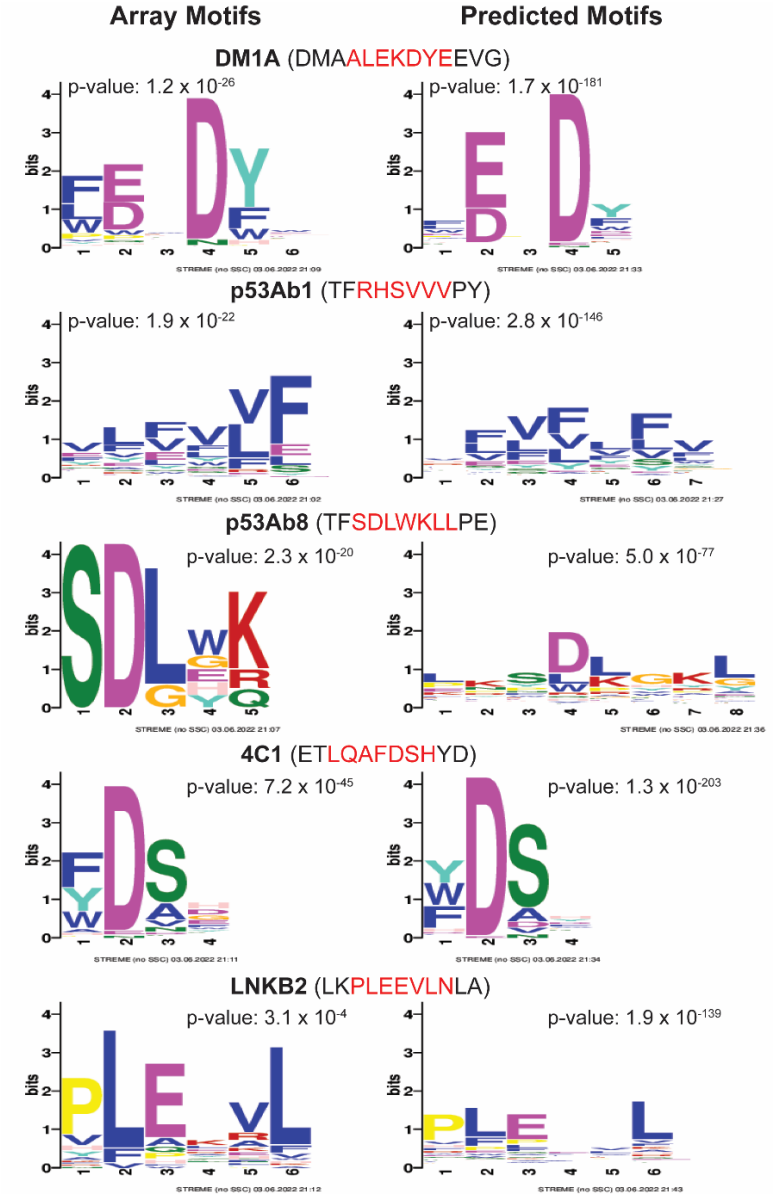


Figure 2.6. Sequence logos showing the residue patterns recognized by the monoclonal antibodies on the array, and as observed through the predictions. Array motifs represent the common motifs observed in the top array peptides (Z-score >=3). Predicted motifs represent the preference of amino acid residues at each position as predicted by the neural network model. The top predicted peptides (Z-score >=3) from random in silico peptide libraries were used to generate these motifs. Part of the proteome from the target proteins is shown on the top of each sequence logo, where the epitope is shown in red. All the sequence logos were generated using STREME (Bailey, 2021).

These sequence logos help recognize the important residues for binding interactions on the microarray. The actual cognate sequences along with the flanking residues from

their target proteins are also shown in the figure for reference. The positions at the bottom of each sequence logo are a frame of reference and do not represent the actual positions of these residues in the target protein. The height of each amino acid residue within the stack represents the relative frequency with which that particular residue appears at that particular position after the alignment of the sequences.

It must be noted here that the sequence logos from the predicted peptides have a similar motif as that of the sequence logos of the top peptides from the array. This was especially observed in case of DM1A, 4C1, and LNKB2, where the more important residues highlighted in both the cases are the same. In case of p53Ab1 and p53Ab8 as well, the residues that were found to be of common occurrence on the array were also favored by the neural network predictions.

The sequence motif observed in case of p53Ab8 (Figure 2.6) indicated that the algorithm was able to identify the relevant residues correctly. However, the results of projection from Table 2.5 indicated otherwise. It was observed in the motifs that the cognate residues (SDLWKL) were preceded by two other residues (L and K). Taking this into account, the p53Ab8 model was projected again on the in silico peptide libraries, but this time the sequence LKSDLWKL was used instead of SDLWKL. The results significantly improved this time with the sequence LKSDLWKL being ranked at the top of the list (Table 2.6). Following this observation, the cognate motif from the target protein p53 (TFSDLWKL) was introduced among the randomly generated peptides and the model was projected again on this library. This time, the cognate interactions ranked around 43 (Table 2.6), which was a significant improvement of performance from the one observed in Table 2.5. Thus, it was found that the model was actually taking positional information

into account in case of p53Ab8. Therefore, the model was able to successfully predict the relevant residues in case of all the five monoclonal antibodies within top 0.2% of the randomly in silico peptide libraries with 10^6 sequences. In accordance with this observation, the epitope of p53Ab8 will be represented as TFSDLWKL henceforth.

Table 2.6. The mean rank of p53Ab8 sequences among 10^6 random 9-mer peptides

p53Ab8 sequences	**Average Rank (20 iterations) ***	**Percentage rank in 1 million**
LKSDLWKL	1 ± 0	0.0001%
TFSDLWKL	43 ± 6	0.0043%

* Error shown here is the standard error of the mean of the ranks

2.3.2 Specificity of the Binding Predictions

In order to get a better understanding of the sequence-to-function relationships of these antibodies and how the binding information obtained from the sequence space represented by the arrays dictates the binding behavior, it was important to study the specificity of these predictions. The binding experiments were all done using identical peptide arrays. Therefore, the sequences represented on the array are common to all the monoclonal antibodies. So, the binding interactions between the sequences and the antibodies become a very important deciding factor in distinguishing the binding patterns from one another, both in vitro and in silico. Not all sequences present on the array contribute equally to this distinction of binding behavior. To determine, how specific the predictions are and if the sequences that interact highly with one monoclonal antibody were also interacting (or predicted to interact) with another antibody, specificity plots (Figure 2.7) were drawn. Previous works have established the peptide microarrays as an effective tool for mapping linear epitopes. Each monoclonal antibody also binds to a unique set of peptides from the peptide library that are weakly bound by other monoclonal antibodies.

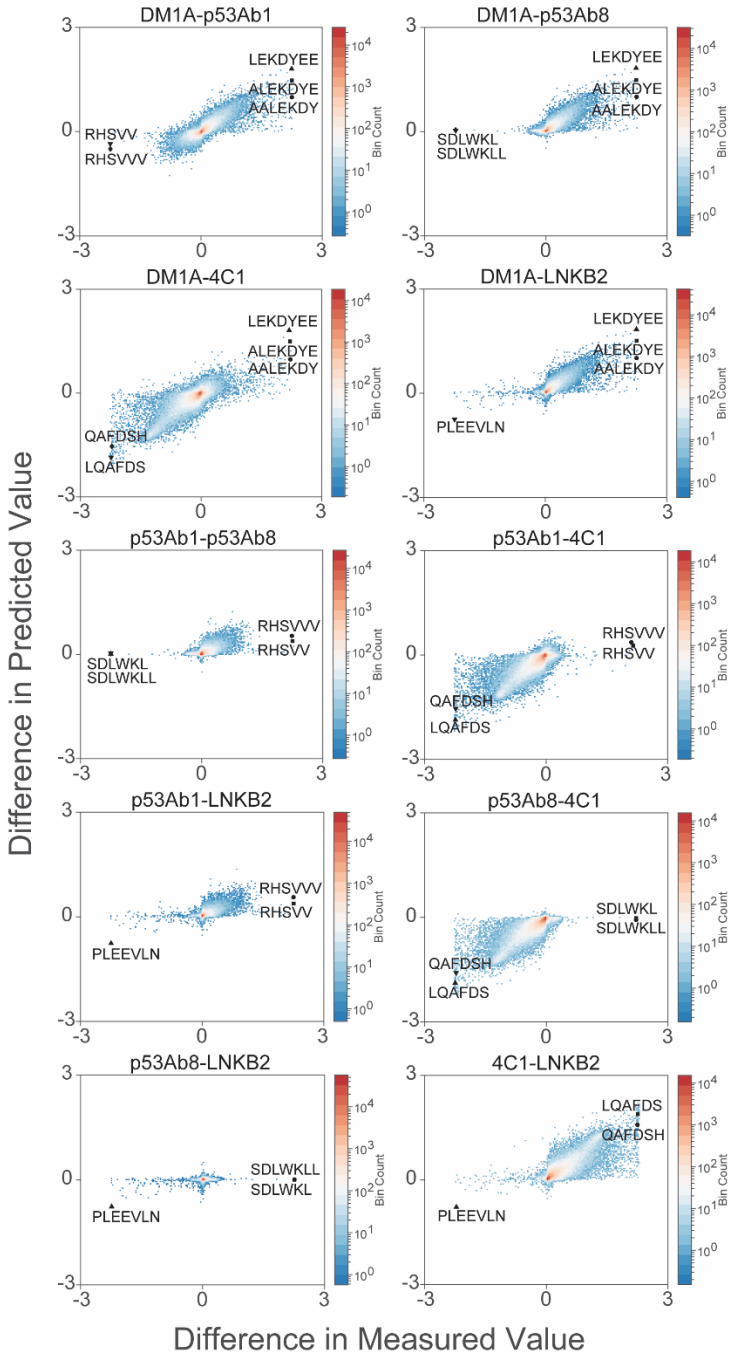


Figure 2.7. Scatter plots showing the specificity of the predictions. The specificity was calculated as the differences in the measured and predicted binding intensities between a pair of monoclonal. Along the x-axis, the difference between the measured binding values is plotted. The y-axis represents the difference in the predicted binding values. In each case, the positions of the cognate sequences for the corresponding monoclonal antibodies are indicated. The colorbar indicates the density of distribution for the datapoints. All calculations are in the $\log_{10}$ scale.

It would be important for the algorithm to recognize and predict the specificity as observed on the microarrays. Figure 2.7. represents the specificity between all the pairs of monoclonal antibodies used for the study. The x-axes show the difference in $\log_{10}$ measured binding values and the y-axes show the difference in $\log_{10}$ predicted binding values. All the cognate sequences for each pair of antibodies are represented on the graph as well.

The cognate sequences of each pair of antibodies are observed to have very high specificity with a high degree of separation with respect to measured data (x-axes) on the array, as expected. As can be seen from the figure, the majority of the array peptides (>80%) have the same binding affinity towards either of the monoclonals, both in terms of measured and predicted data. Therefore, they are centered around the zero-value in both axes. These peptides are mostly weak binders to the monoclonals and interact non-specifically with them. Thus, it is only a very small fraction of the peptides that interact with the antibodies above the noise cut-off that is responsible for defining the predicted sequence space. The predicted specificity for DM1A and 4C1 epitopes is on par with respect to the measured specificity. The lowest predicted specificity is observed in the case of p53Ab8 and p53Ab1. This can be attributed to the broad absence of peptide binding in the mid-range values (typically on the order of $\sim 10^3$ intensity units) in the actual array experiments of these antibodies (see Figure 2.4). These monoclonals show high binding specificity to their specific epitope sequences and do not exhibit marked binding to the majority of other peptide sequences on the array. This however was found not to affect the predictive performance of the model, which still identified the cognate interactions well in both the cases.

2.3.3. In silico Substitution of the Cognate Sequences

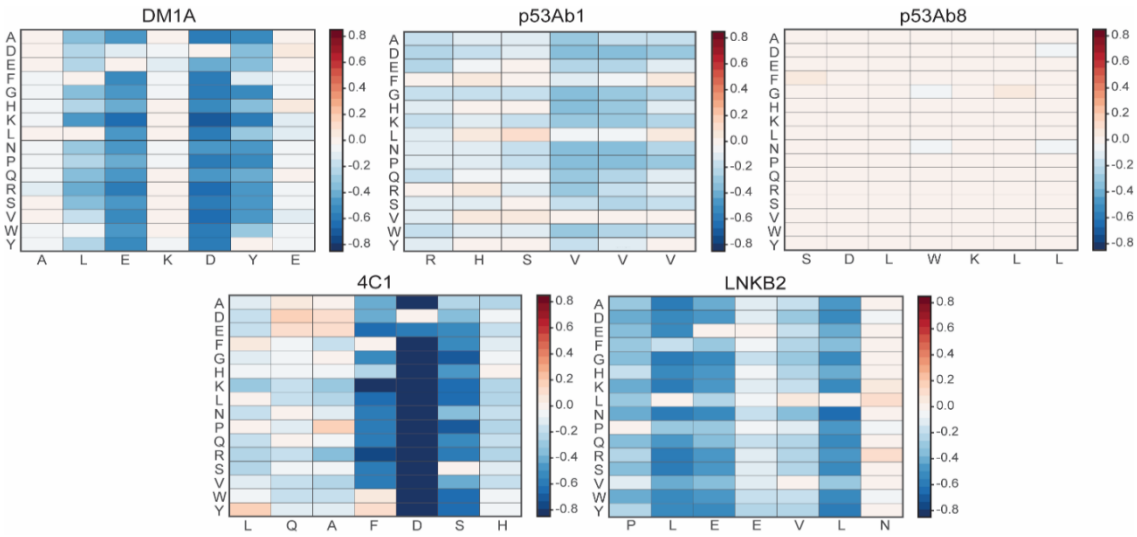


Figure 2.8. Heatmaps representing the results of in silico single-point substitution of the epitopes for each monoclonal antibody used in this project. Each residue in an epitope sequence (x-axes), was replaced with the 16 amino acid residues used on the microarray (y-axes), one at a time to create an in silico library. The NN model was used to predict binding values for the substituted peptide in the library. The predicted value of the actual cognate sequence was then subtracted from those of the other peptides. The colorbar represents the difference of $\log_{10}$ binding intensities between the substituted and the cognate peptide.

The next prerogative was to investigate whether the NN model was capable of distinguishing between key residues of the cognate sequences. In order to do so, each cognate sequence was taken, and using a MATLAB script, an in silico library of substituted peptides was created from the cognate. The mutations were carried out using only the 16 amino acids used in the array and in a position-specific manner such that the substituted peptide sequences are always one Hamming distance away from the actual cognate sequence. The antibody-specific NN model was then used to predict the binding intensities of the substituted peptides. The relative change in $\log_{10}$ binding of the substituted peptide was then calculated according to the following formula:

$$C_{ij-E} = PB_{ij} - PB_E$$

where PB_E is the predicted $\log_{10}$ binding value of the actual cognate sequence, PB_{ij} is the predicted binding of the substituted sequence with i-th residue from the 16 amino acids at the j-th position of the cognate sequence. C_{ij-E} is the relative predicted binding. The predicted binding intensity of the actual cognate epitope was then subtracted from these predicted intensities to calculate differential binding. The results are represented as heatmaps as shown in Figure 2.7. In the figure, the shades of blue indicate a predicted lower binding preference for that residue compared with the original residue whereas shades of red indicate the opposite.

The heatmaps can also be considered as an indicator of the conserved regions of the cognate sequence as recognized by the predictive model. The residues with more blue shading are the ones that are more sensitive to replacement with other residues. It was observed that in the case of DM1A, the predictive NN model predicts high sensitivity for substitutions of the residues L(2), E(3), D(5), and Y(6) at positions 2,3,5, and 6, respectively. Correspondingly, in the case of p53Ab1, the residues R(1) and V(4,5) were shown to be moderately conserved sequences. In the case of p53Ab8, no such preference was observed. This may be attributed to the lack of mid-range values while training the model. For 4C1, residues F(4), D(5), and S(6) were determined to be the most conserved regions by the predictive model. All the other residues can be considered to be moderately conserved. For LNKB2, the highly conserved residues were P(1), L(2), E(3), and L(6).

2.3.4 Using Propensity Scales as Encoders for the Amino Acids

In the next step of the study, the effects of using propensities of the amino acids as an encoder for the neural network were studied. The propensities of an amino acid represent

the physicochemical characteristics of that residue like isoelectric point, molecular weight etc. So far in the study, the algorithm has been using an encoder (first layer) that was learned during the training of the model. Here, various physicochemical propensities of the amino acids will be supplied to the algorithm to encode the different residues, instead of using a learned encoder. Again, the dataset for the DM1A monoclonal antibody was chosen for this as the distribution of the data covered a much broader range of values compared to the other antibodies. Different sets of propensities were taken into consideration for this. Propensities like different hydropathy indices (Kyte & Doolittle, 1982; Eisenberg, 1984; Engelman et al., 1986; Cornette et al., 1987; Rose et al, 2003), molecular weight, and isoelectric point of the amino acids (Gasteiger et al., 2003) were used. Also, other propensities that highlighted the structure activity relationship of the amino acids were used as well. Kidera factors (Kidera et al., 1985) are a scale of 10 components that were derived from multivariate analysis of 188 physicochemical properties of amino acids. The Z scale (Sandberg et al., 1998) consists of 5 components and was derived from analyzing 26 different physicochemical properties including lipophilicity, bulk, and charge. MSWHIM is a set of 3 descriptors derived from considering 36 different electrostatic potential properties (Zaliani & Gancia, 1999). Cruciani properties are a set of 3 principal properties (PP 1-3) that characterize the amino acids based on polarity, hydrophobicity, and H-bonding capabilities (Cruciani et al., 2004). T-scales are a set of 5 topological descriptors for the amino acids (Tian et al., 2007). Factor analysis scales of generalized amino acids information (FASGAI) are a set of 6 components that were derived from 335 physicochemical properties of the 20 naturally available amino acid residues (Liang et al., 2007). BLOSUM is a matrix derived amino acid descriptor set that employed

physicochemical properties that have been subjected to varimax analyses and BLOSUM62 alignment of the 20 amino acids (Georgiev, 2009). ST scales are again topological descriptors (8) which also includes 3D structural information (Yang et al., 2010). The 8 vectors of hydrophobic, steric, and electronic (VHSE) properties were derived from the respective properties by Mei et al., 2005. Lastly, protein fingerprint or ProtFP descriptors are a set of 8 descriptors compiled by selecting different amino acid indices and eliminating the most co-varying indices (van Westen et al., 2013).

Table 2.7. Performance of the model with respect to different propensities used (mean correlation coefficient and mean ranking of epitope)

Propensities Used	**Number of Amino Acid Descriptors**	**Mean Correlation Coefficient ***	**Mean Rank of Epitope[#]**
None (Learned encoder)	9	0.8243 ± 0.0071	35 ± 2
Random, normally distributed number between [0,1]	9	0.7635 ± 0.0095	51,127 ± 54
Hydropathy Indices, Molecular Weight, Isoelectric Point	7	0.7890 ± 0.0050	818,680 ± 75
Representative properties from each different propensity scale (PP, KF, Z, F, T, VHSE, ProtFP, ST, BLOSUM, MSWHIM)	10	0.8294 ± 0.0060	290,176 ± 200
VHSE	8	0.8396 ± 0.0055	63,292 ± 54
ST Scales	8	0.7968 ± 0.0066	508,516 ± 3,112
ProtFP	8	0.8239 ± 0.0064	86,694 ± 84
Kidera Factors	10	0.8537 ± 0.0104	726,006 ± 106
BLOSUM	10	0.8511 ± 0.0042	586,987 ± 133

* Mean correlation coefficients have been calculated over 10 independent trainings of the model for each antibody. [#]Mean ranks are out of 1 million and have been calculated from projecting the respective models on to 10 individual libraries of random 9-mer peptide sequences. None of the libraries had any common peptides except for the epitope sequence of the respective antibody. All the errors represent the standard errors of the means. All the experiments were done using the binding data available for DM1A.

Before using propensity scales as encoders, a 20 x 9 matrix containing a set of fixed random values between the range of 0 and 1 was tested as an encoder matrix to evaluate the performance of the model on the basis of random numbers. This matrix represented the 20 naturally available amino acids with 9 descriptors for each. These random numbers were normally distributed. The average correlation coefficient for the trained model was 0.7635 (10 independent runs, Table 2.7). The mean rank of the epitope sequence in random in silico peptide libraries was ~50,000. Therefore, the model definitely did perform worse than when using a learned encoder for the training. For training the model next, only physicochemical properties such as, molecular weight and isoelectric point of the amino acids were used along with different 5 different hydropathy indices. The total number of descriptors for each amino acid was equal to 7. All the values were normalized beforehand between 0 and 1. The average correlation coefficient of 5 independent training runs was 0.789, which was lower compared to the average correlation of 0.8156 from the runs where no external encoder was supplied to the algorithm. The average ranking of the epitope in libraries of random million peptides was found out to be ~800,000 which meant that the model did not perform well. Due to the poor performance of the model, another set of propensity parameters were chosen as encoders. This time, a representative scale from each different type of propensity was selected. However, physicochemical properties were not included in this case, as the other propensities already account for them. The total number of descriptors for each amino acid was equal to 10. It was observed that the average correlation between the predicted and the measured values in this case was 0.8294 which was slightly higher than the average correlation of the trainings where no external encoders were used. However, when the predicted ranking of the epitopes among random peptides

were checked again, the average rank was ~290,000 out of 1 million. Finally, individual scales were considered for training the model. As according to the data shown in this study, the model performs best if the number of descriptors per amino acid is around 9, therefore scales with 8 or higher components/factors were chosen only (VHSE, ST scales, ProtFP, Kidera factors, BLOSUM descriptors). The different encoders chosen and the variation in performance of the model has been summarized in Table 2.7. Although, VHSE and ProtFP scales did produce better results than rest of the encoders, comparable to that of the random numbers, it is interesting to note that using propensity scales in general produced worse results than using a learned encoder, in terms of predicting the epitope sequences. The general speculation from this observation is that using a pre-assigned encoder for training the model might be biasing the algorithm towards a favored set of residues as determined by the propensities used. This in turn might be prohibiting the neural network from correctly evaluating and interpreting the actual amino acid residues responsible for binding to the monoclonal antibody.

2.3.5 Evaluation of Model Performance with respect to Concentration

All of the above studies were performed on data available at the highest concentration for each monoclonal antibody (Table 2.2). However, if one were to make changes in the concentration of the monoclonal antibody during the measurement of the binding on the arrays, what changes would be observed in the performance of the model? To answer this question, a series of different binding experiments were performed where the concentrations of the monoclonal antibodies were serially diluted (Table 2.2). These different datasets were then used to train the neural network model, using the same hyperparameters that were optimized beforehand. Figure 2.9 shows the Pearson correlation

coefficients between the predicted and measured binding intensities at each different concentration for all the antibodies. As seen in the figure, at very low concentrations (<125pM), the predicted binding values did not correlate very well with the experimental data. As the concentration kept on increasing, the correlation between the predicted and the measured values went up, as there was more information available to the neural network.

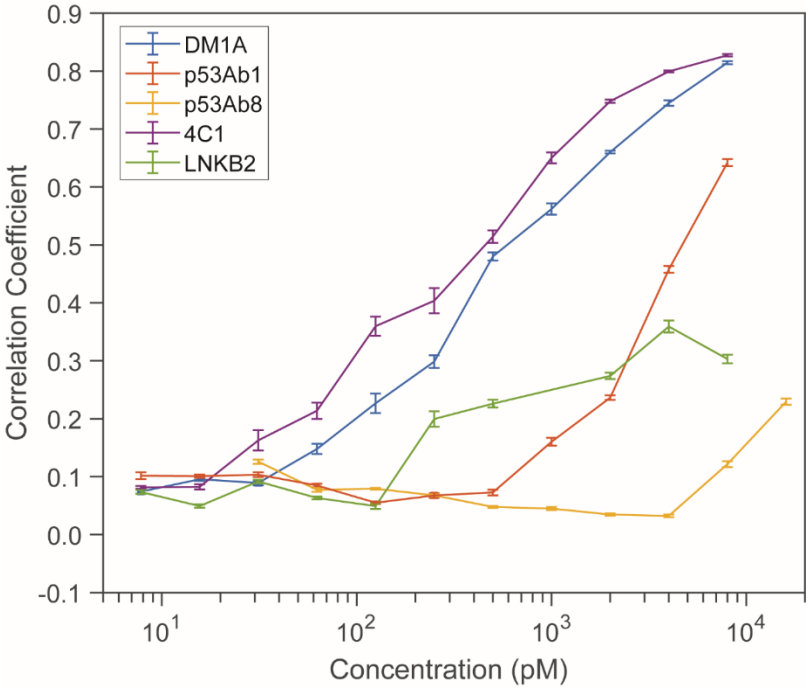


Figure 2.9. Variation of Pearson correlation coefficient (y-axis) with gradually increasing concentration (x-axis) for each monoclonal antibody. Points plotted are means of 5 independently trained replicates of the model and the error bars represent the standard error of the mean.

It might be due to the fact that at lower concentrations the measured binding values were in the order of $10^{2.5}$ fluorescence units (Figure 2.10) in all the cases. As these values were not above the noise cut-off of the intensity signals, the neural network was not able to distinguish between the interactions observed. As the concentrations of the monoclonal antibodies were increased the signal intensities of different interactions became considerably higher than the observed noise, therefore allowing the model to learn the

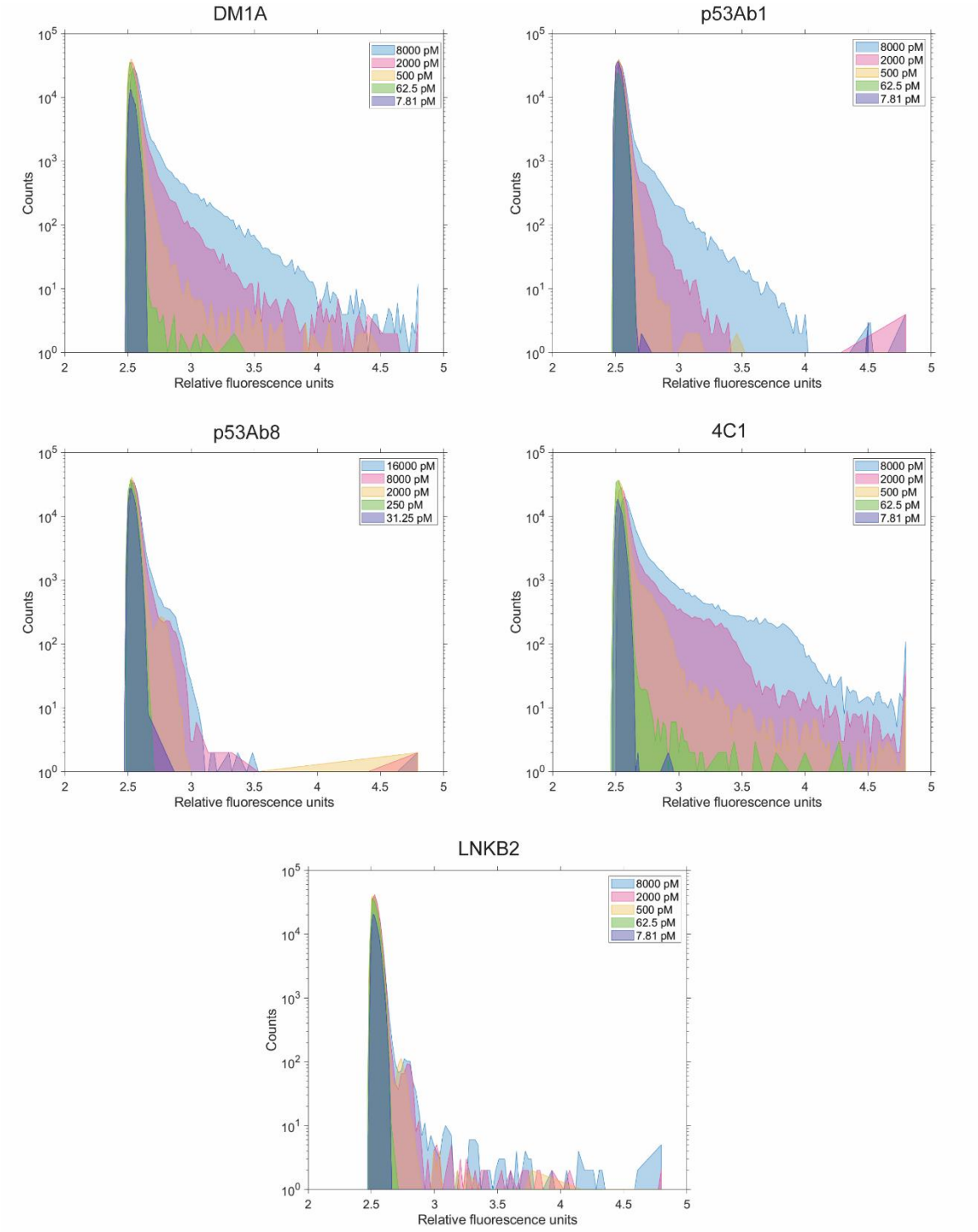


Figure 2.10. Distribution of fluorescence intensities (in log10 scale) across five different concentrations for each of the 5 monoclonal antibodies. The x-axis represents the range of measurement in terms of fluorescence intensities. The y-axis is the count of data points available against the relative fluorescence intensity.

observed binding patterns on the array better. Significantly better results were obtained when the intensities were greater than $10^{3.5}$ (Figure 2.10). This was because at higher concentrations, fewer of the interactions are needed to establish binding above the noise. It was observed that if the distribution of the intensities covered a broader section of the dynamic range of the instrument, the predictive models were correlated better, and the predictions improved accordingly. Figure 2.10 shows the distribution of datapoints across various concentrations.

Table 2.8. The mean rank of epitopes with respect to different concentrations of the monoclonal antibodies

Conc (pM)	**DM1A (ALEKDYE)**	**p53Ab1 (RHSVV)**	**4C1 (LQAFDSH)**	**LNKB2 (PLEEVLN)**		**Conc (pM)**	**p53Ab8 (TFSDLWKL)**
7.81	664,944 ± 207,562	409,086 ± 188,045	115,689 ± 20,455	516,512 ± 120,709		31.25	716,285 ± 219
62.5	1,844 ± 215	219,671 ± 140,501	31,470 ± 4,242	151,857 ± 20,873		250	870,210 ± 357
500	75 ± 2	10,808 ± 23,925	5,918 ± 570	121 ± 24		2000	21,183 ± 33
2000	31 ± 2	4,012 ± 2,749	1,405 ± 135	9 ± 5		8000	3,122 ± 20
8000	35 ± 2	2,686 ± 1,035	722 ± 8	5 ± 1		16000	51 ± 3

The ranks of epitopes were determined using random libraries of 1 million peptides (5 independent runs). Error shown is the standard error of mean in ranks. Both mean and error were rounded off to the nearest integer.

Table 2.8 shows the mean ranks of the epitopes as predicted in random libraries of 1 million peptides across the concentrations mentioned in Figure 2.10. Considering these results, the best set of concentrations were chosen for each antibody. These binding data from different concentrations were then compiled and were simultaneously fitted using the algorithm.

None of the hyperparameters were changed for this experiment. The results are as tabulated below (Table 2.9).

As can be seen in Table 2.9, when the data from these concentrations were compiled and fitted simultaneously, the observed correlation coefficients were slightly lower than when data from a singular higher concentration was used for training. However, the predictive performance of the models was nearly identical if one compares the results in terms of epitope ranking. This goes on to show data available from multiple experiments with varying concentrations of the antibodies can also be used to train the algorithm instead of using a singular binding data.

Table 2.9. The correlations and mean rank of epitopes of the monoclonal antibodies when multiple concentrations are fitted simultaneously

Monoclonal Antibodies	Concentrations used (pM)	Mean Pearson Correlation Coefficient*	Mean Rank in 1 million[#]
DM1A	500; 1000; 2000; 4000; 8000	0.7513 ± 0.0062	222 ± 6
Ab1	2000; 4000; 8000	0.5224 ± 0.0084	1,051 ± 3,752
Ab8	8000; 16000	0.1745 ± 0.0060	13 ± 1
4C1	2000; 4000; 8000	0.8081 ± 0.0036	191 ± 4
LNKB2	500; 1000; 2000; 4000; 8000	0.1352 ± 0.0178	2 ± 1

* Mean correlation coefficients have been calculated over 5 independent trainings of the model for each antibody. [#]Mean ranks have been calculated from projecting the respective models on to 10 individual libraries of random 9-mer peptide sequences (10^6 peptides). None of the libraries had any common peptides except for the epitope sequence of the respective antibody. All the errors represent the standard errors of the means.

2.4 DISCUSSION

Many attempts have been made before to characterize the binding behavior of monoclonal antibodies using various types of computational approaches. These computational approaches aim to provide helpful aid to the already existing experimental

methods that can be time-consuming and expensive often (Norman et al., 2020; Manieri et al., 2020; Graves et al., 2020). Many of these approaches, however, have often been restricted due to limited availabilities of related structural data in existing databases. Often, characterizing the interactions of antibodies prove to be a challenging task because of the lack of specific structural information. Through this work, an alternative approach has been adapted that considers the sequence vs. function relationship between the microarray peptides that covers a certain combinatorial space and the monoclonal antibodies. In this approach one does not need to have any information about the structure of the antibody beforehand. Rather, information obtained from interactions of the antibodies with a given set of peptides that sparsely sample the combinatorial chemical space, is used to predict its molecular interactions, with the help of a neural network. A similar approach has already been established to quantify protein-peptide relationships through a previous work by our group (Taguchi et al., 2020), where one does not need to obtain a resolved structure of the protein beforehand.

The work described here utilizes the methodology laid out by Taguchi et al, to characterize and quantify the binding interactions of different monoclonal antibodies, whose epitopes are represented among the combinatorial space covered by the microarray peptides. This combinatorial space consists of 126,050 unique peptides whose length varies from 4 to 13 amino acid residues. The distribution of peptides of various length across the array is as shown below in Figure 2.11. The median length of the peptides is 9.

These peptides were randomly sampled from a combinatorial space of nearly 10^{12} molecules considering that only 16 out of the 20 amino acids were used. All of the peptides are linear, therefore only allowing the characterization of linear target epitopes of the

antibodies. However, this removes the structural complexities associated with conformational epitopes and simplifies the model. All the interactions of the monoclonals with the peptides on the array are considered for training the models and deriving predictive models, whether on-target or off-target. These provided a more in-depth insight, especially about interactions that were seemingly random.

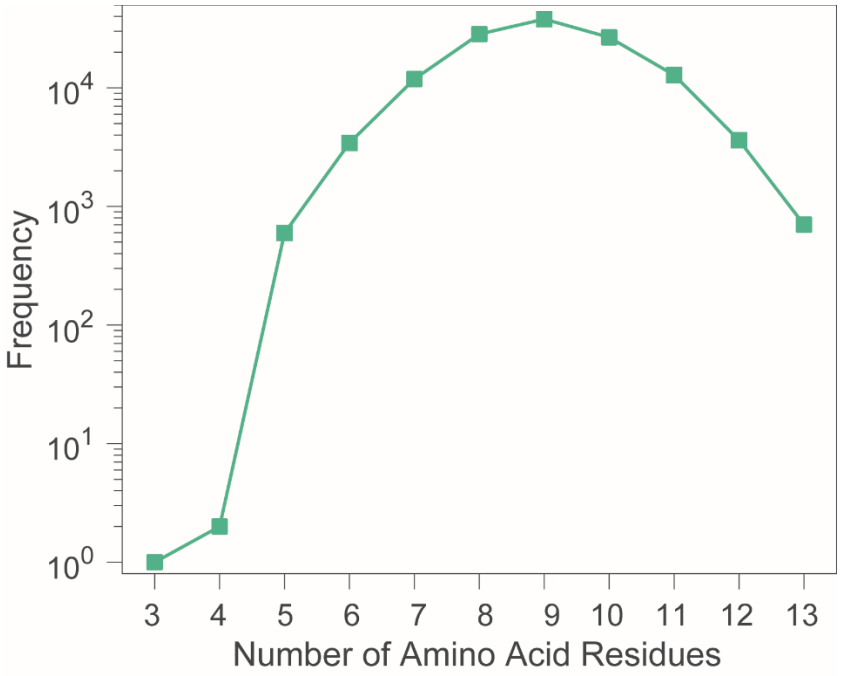


Figure 2.11. Length distribution of HT-V13 arrays obtained from HealthTell. The mean length of peptides on the array is 9 with a standard deviation of 1.37.

Choosing the right hyperparameters before training any neural network is of utmost importance. In this work, the hyperparameters were tested in a grid-like fashion, using DM1A as a model system. A set of different hyperparameters were tried out and the ones that resulted in the best performance of the model were chosen for further training with the rest of the monoclonal antibodies. The performance of the model was estimated on the basis of Pearson's correlation coefficient and the rank of the epitope within in silico libraries of 10^6 randomly generated peptides. However, correlation coefficients alone

would not prove to a be a sufficient measure for the performance of the model as it varies greatly depending on the interacting antibodies. In Figure 2.4, the average correlation coefficients of the models are shown alongside the distribution of the predicted and the measured values. In that figure, it can be seen that the correlation coefficients of DM1A and 4C1 are significantly higher than the rest of the antibodies. This is largely dependent on the information available to the neural network in terms of the distribution of the measured data. For the antibodies where the binding intensities are fairly well-distributed across the dynamic range of measurement, the correlations obtained were significantly higher (DM1A, and 4C1). In contrast, antibodies like p53Ab1, p53Ab8, and LNKB2 have a highly skewed distribution of binding intensities where most of the datapoints (>99%) are very close to the noise. The rest of the available peptides above the noise range are the ones which are very strong binders to the antibody, thus giving a very high signal. However, there is not much signal available in the middle of the dynamic measurement range, resulting in poor correlation. Hence another method was employed to test out the performance of the model. It must be noted, however, that even when the correlation is lower, the model still manages to successfully identify the epitopes of the antibodies in 4 out of 5 cases. As all the epitope sequences were excluded during training, this points towards the fact that a small minority of non-cognate peptide sequences (<500) that interact with the antibodies hold enough information in terms of sequence space to let the models differentiate between specific and non-specific interactions (Figure 2.7).

Correlation is a direct comparison between the measured and predicted values of the peptides. However, the performance of the model also depends on the weightage given by the model to each amino acid residue. When the models were tested against random

peptide libraries of 10^6 peptides, all of the five antibodies (DM1A, p53Ab1, p53Ab8, 4C1, and LNKB2) had their epitopes correctly predicted within the top 0.2% of the libraries (Table 2.5 and 2.6) which is statistically very high. This indicates that although the correlation between the predicted and measured values for a given antibody might be lower, it does not prevent the model from distinguishing between specific and non-specific interactions. In Figure 2.6, the sequence logo patterns demonstrate the amino acid residues that were deemed important and necessary by the neural network to interact with the respective antibodies (B). These sequence logos were generated by using ranked (based on predicted binding) peptides from randomly generated libraries and do not contain any cognate sequences. The residues highlighted were favored by the neural network simply on the basis of the weights learned during training. If these predicted motifs (Figure 2.6) were compared to the actual observed motifs from the array, one would observe striking similarities between the two. In case of DM1A, residues L, F, E, D, and Y appear to be more prominent, and the motif pattern observed from the non-cognate sequences is in fact very similar to the actual cognate sequence of the antibody. Similar results are observed in the rest of the cases as well. In case of p53Ab1, the residue V is identified correctly along with F and L, which also appear in the array motif. In fact, F being a flanking residue to the cognate also assists in the interaction. However, it appears from the array and the predicted motifs that the p53Ab1 antibody interacts with a lot of hydrophobic sequences apart from its cognate, which is why the residues R, H, and S were not given more importance by the algorithm. In the motif for p53Ab8, the residues S, D, L, and K have been identified in correct order of occurrence. Interestingly enough, the sequence motif of p53Ab8 indicates a position bias of the particular residues which was not observed in case

of other antibodies. This position bias was learned by the algorithm from the interactions observed on the arrays. This was an interesting observation that needs more thorough exploration. However, due to time constraints, a more in-depth explanation could not be provided in this thesis, with regards to the same. In case of 4C1, the residues F, D, and S have been highlighted correctly and comparable to the pattern observed on the array. In fact, it was shown experimentally (Shepherd et al., 1999) that the motif FDSH is the most important contributor in binding to the antibody 4C1. Residues Y, and W also show up in both the motifs, largely due to their structural similarities with F (aromatic), which is an important contributor to the interaction. For LNKB2, the residues P, L, E, and V again appear in both the motifs, very similar to what is observed in the cognate sequence. Residue N does not appear very likely because of the fact that there is lesser representation of N on the array compared to the other residues.

Next the specificities of these predictions were considered. It has been already mentioned previously that the algorithm is capable of distinguishing between specific and non-specific interactions, for each antibody. However, one also needs to consider how specific these predictions are when compared with one another. Figure 2.7 shows the specificities between the pairs of antibodies as differences between predicted and measured binding values. From these plots, one can see that a large majority of the peptides interact with either of the antibodies in a non-specific manner and therefore they are found towards the center of the plots. This is especially prominent in the cases of antibodies which are known to show highly specific behavior on the array and bind only to the motif represented by the cognate sequences (p53Ab1, p53Ab8, and LNKB2). However, in all of the cases, a small fraction of the peptides can be identified which are highly specific to a particular

antibody (top right quadrant or bottom left quadrant). The specificities of these peptides have been predicted mostly accurately in all of the cases.

The amino acid residues present in the cognate sequences play the most important role in facilitating molecular recognition between the antigens and the antibodies. Figure 2.8 demonstrates the performance of the predictive models, with respect to recognition of the amino acid residues present in the epitope, with the help of an in silico mutation experiment. As can be seen in Figure 2.6, not all the residues of the epitopes are given equal weighting by the models as the epitopes themselves were excluded from training examples. As the learning of the neural network is dependent on the distribution of the sequence space of the arrays, some amino acid residues are preferred over the others. This in silico mutation experiment was designed to observe the changes that occurred in predicted binding when known residues of the epitopes were replaced with any of the 16 amino acids used on the array. In Figure 2.8, one can see which mutations are unfavored by the model, leading to a negative change in binding, in the case of each antibody. In case of DM1A, Residues L, D, and Y are favored over all other residues in their respective positions, whereas residues A, K are shown to be more favorable towards substitution. Residue E is shown to be mostly conserved at the 2^{nd} position but much more favorable to substitution at the 7^{th} position. In case of p53Ab1, none of the residues show a high degree of conservation. However, the residues present in the actual cognate sequence are more preferred over other substitutions. Residues F, D and S appear to be most conserved in 4C1, while L, Q, A, and H are moderately preferred over other substituting residues. For LNKB2, residues P, L, E, and V are favored over other mutations in the respective positions, whereas residue N is not particularly resistant to mutational changes. In case of

p53Ab8, none of the residues are shown to be strongly preferred, however the predictive performance of the model was not affected by it. This analysis demonstrates how strongly the residues present in the epitopes are favored over other residues in their respective positions, as learned by the predictive models. It also gives an idea as to which residues are considered as favorable mutations at a particular position as determined by the algorithm.

So far, the algorithm has only been using learned encoders during training, and all the characterization of the performance was solely based on that. No physicochemical or structural information was taken into account for those analyses. However, amino acids are chemical entities and have many physicochemical properties that play a dominant role in determining the structure and therefore functions of the peptides and proteins. Considering this factor, a set of different propensity indices were tested as encoders for the algorithm and eventually the performances of the models were characterized. These propensity scales encoded various physicochemical, structural, and topological aspects of the 20 naturally available amino acids. Table 2.7 summarizes the characterization results. One can see that the general predictive capabilities of the models were reduced in all of the cases where propensity indices were used as encoders (in terms of epitope ranking), when compared to the results from the learned encoders. Although, propensity scales like the VHSE, and the ProtFP fare better than the rest, they are still nowhere close to the predictive performance observed from the learned encoders. Rather their performances are somewhat similar to the performance observed in case of randomly generated numbers which were used as encoders. The likelihood is that while using a preassigned encoder, one is unwittingly introducing a bias to the algorithm that prevents the model from accurately determining the orthogonality between each amino acid residue as observed through the interactions on the

arrays. Even with a combination of different properties, it was not possible to remove this inherent bias. These biases which are learnt from the propensity indices, or a pre-supplied encoder in general, tend to favor a particular set of residues, that might not be relevant to the interactions observed on the arrays. Hence, the predictive performance of the algorithm was not accurate. On the other hand, a learned encoder obtains the values during the training of the model where only the interactions observed on the arrays are considered. Therefore, the predictive capabilities are higher in such case.

The results presented so far were all on the basis of a singular concentration of the monoclonal antibody. Changes in concentration of the monoclonal antibody during assays would definitely change the recorded binding intensities, therefore also altering the output of the neural network. In section 2.3.5, it was shown that with increase in concentration of the antibodies, the correlation between the predicted and measured values also got better (Figure 2.9). The reason for this observation was explained with the help of Figure 2.10 which shows the distribution of measured binding intensities across various concentrations. As one can see in this figure, as the concentrations of the monoclonal antibodies are increased, the relative range of recorded binding intensities also broadens. As more and more measurements are recorded above the noise cutoff ($>10^3$), more information is available to the algorithm during training. This results in more accurate learning of the parameters, hence the predictive capabilities of the models also become better with increasing concentration (Table 2.8) in all the 5 cases.

In Table 2.8, it can be seen that one does not need to have a very high concentration to get a good predictive model. Even at lower concentrations, the algorithm achieves a decent prediction, provided that the recorded binding intensities capture enough

information in terms of interacting peptides. However, it does depend on the nature of the molecular interactions of the monoclonal antibody that is being assayed. As has been previously discussed, these peptides are not large in number, but they contain enough information in terms of represented sequence space for the model to accurately distinguish between specific and non-specific interactions. Using the observations from Table 2.8, it was considered to train the algorithm on data from multiple concentrations simultaneously. The concentrations which were used were selected on the basis of their predictive performance in terms of ranking of the epitopes. The results are as shown in Table 2.9. Here it can be seen that although the overall correlation coefficients of the models are lower than their single-concentration counterparts, but the ranked prediction of epitopes are comparable. This goes on to show that multiple concentrations can also be used simultaneously in training the neural network. Also, assaying at multiple concentrations captures an even broader range of interactions which might be absent in case of a singular concentration. Therefore, it might aid the model in training more accurately.

It is impressive that by training a simple neural network on a set of random sequences that were sparsely sampled from a huge combinatorial space (~10^5 out of a combinatorial space of ~10^{12}), one can predict the binding behavior of monoclonal antibodies in all the 5 cases presented in this study. These results were obtained solely from analyzing binding motifs in the represented sequence space, without any structural data. In the next step, learning how this predictive model holds against peptides from the actual proteomes would be interesting. Combining the versatility of the machine learning algorithms with the advantage of large combinatorial libraries, one can create powerful

tools that have ample opportunities for the discovery, development, and optimization of various other biomolecular interactions.

CHAPTER 3

USING NEURAL NETWORKS TO MAP PREDICTED BINDING MOTIFS OF MONOCLONAL ANTIBODIES ON THE TARGET PROTEINS AND THE HUMAN PROTEOME

This work was initiated in collaboration with Akanksha Singh

3.1 INTRODUCTION

Over the recent years, monoclonal antibodies have gained prominence in the field of therapeutics and diagnostics (Uhlen et al., 2010; Stadler et al., 2013; Norman et al., 2020; Kaplon et al., 2020). They have been used in the treatment of cancer, infectious diseases, and autoimmune disorders (Beck et al., 2010; Brennan et al., 2010; Kaplon et al., 2020; Nelson et al., 2010). Aside from being used in therapeutics, they are also used in several immunodiagnostic assays as reagents (Shi et al., 1995; Jansen et al., 2015). Although monoclonal antibodies are raised against a specific antigen, they do have some cross-reactivity (Haspel et al., 1983; Dighiero et al., 1983; Ghosh et al.,1986; Notkins, 2014; Bradbury et al., 2018). Identifying these on-target and off-target interactions of a monoclonal antibody is necessary and crucial for its utilization in any kind of application, be it therapeutics or diagnostics. Identifying and characterizing these interactions between the antibody and their target(s) will help the development of antibody-based therapeutics and vaccines immensely (Sette & Fikes, 2003; Roggen, 2008).

There are many tools available for such identification and characterization of cognate binding region on antibody targets, both experimental and computational (Manieri et al., 2020; Hua et al., 2017; Fibriansah et al., 2015; Clementi et al., 2013; Singh et al., 2013; Krawczyk et al., 2017; Sher et al, 2017; Vita et al, 2018). A more detailed

introduction to some of these approaches have been provided in the introduction to Chapter 2 (Section 2.1).

High-throughput peptide microarrays consisting of randomly sampled peptides from a combinatorial space have served as one such effective tool to characterize the different interactions of monoclonal antibodies (Stafford et al., 2012; Sykes et al., 2013; Legutki et al., 2014). These microarrays feature 126,050 unique peptide sequences that are 4-13 residues long in length (median length 9) and have been randomly chosen from a possible sequence space of ~10^{12} peptides. In 2020, Taguchi and others showed that one can exploit the binding information obtained from the sparse sequence space represented by these peptides, to derive a quantitative relationship about molecular recognition, using a neural network. They used nine different proteins for their study. Following Taguchi's work, in chapter 2, it was demonstrated that by training a simple feed-forward, back-propagated neural network with the sequence and binding information obtained from these high-throughput peptide microarrays, one could predict the binding behavior of five different monoclonal antibodies (DM1A, p53Ab1, p53Ab8, 4C1, and LNKB2), across the combinatorial space. It must be mentioned that the epitopes of these five monoclonal antibodies were present on the array, although they were deliberately excluded from the training dataset of the neural network to avoid potential bias to the algorithm. This work showed that the sequence space represented by a sparse library of nearly 126,000 randomly sampled peptides provided sufficient information to the neural network to differentiate between specific and non-specific binders across combinatorial chemical space, for the five monoclonal antibodies that were studied. More importantly, the cognate binding information for these five monoclonal antibodies were predicted accurately, by analyzing

non-cognate binding sequences (Figure 2.4). The cognate sequences were also correctly and consistently identified among the highest binders when tested against randomly generated in silico libraries of peptides (Table 2.5). These in silico libraries were generated by randomly sampling a million peptides at a time from the combinatorial sequence space available for all 9-mer peptides (16^9 peptides), and therefore had no biological significance.

However, the amino acid sequences that represent proteins are not in random order as they define the structure and function of these molecular entities. The model-predicted motifs and performances observed so far (Figure 2.6, B) were from random sequences that had no biological frame of reference. All the five monoclonal antibodies that were studied in chapter 2 are known to bind to targets from the human proteome. Therefore, to completely assess the predictive capabilities of the algorithm, in context of biology and proteomics, the relationship was projected on to the target proteins of the five monoclonal antibodies, as well as the human proteome, in this study. Projecting the neural network on the target protein sequences of the monoclonal antibodies allowed one to discern between the predicted and the known interactions. Visualizing the results of the projection on the structures of the target proteins also helped to estimate neighboring regions on the protein, besides the cognate sequence, that might play a role in the protein-antibody interaction. To further evaluate the predictive ability of the models, they were projected on the human proteome, which consisted of 20,361 unique peptide sequences. The idea behind this experiment was to see how well the model can distinguish between the target protein and other proteins from the proteome.

3.2 METHODS

3.2.1 Binding Experiments of Monoclonal Antibodies on the Microarray

Binding experiments with the monoclonal antibodies were carried out on high-throughput peptide microarrays. These microarrays were synthesized using a photolithographic approach that was previously described (Legutki et al., 2014; Rowe et al., 2017). A more detailed overview on the manufacturing process is provided in section 2.2.1 (Synthesis of High-throughput Peptide Microarrays) of Chapter 2. The binding assay experiments were carried out using five different monoclonal antibodies (DM1A, p53Ab1, p53Ab8, 4C1, and LNKB2). Section 2.2.2 (Labeling of the Monoclonal Antibodies and Assays) of the previous chapter provides a detailed procedure on labelling the antibodies and performing assays with them. The supplier information and sources of these antibodies are included in Table 2.1, along with their known epitopes.

3.2.2. Training the Neural Network Model

The binding data obtained from the assays were used to train a simple feedforward, backpropagated neural network, similar to that of Taguchi et al (2020). The architecture of the neural network has been laid out in detail in section 2.2.3 (Neural Network Model Architecture for Prediction) of the previous chapter. Table 2.3 shows the hyperparameters that were selected for training the neural network. The neural network was trained 100 independent times for each monoclonal antibody. Each time a random set of peptides was selected from the peptides present on the arrays, for training the algorithm. Cognate sequences related to each antibody were removed from all training sets and included in the test set. Once these models were tested on random *in-silico* peptide libraries (see section

2.3.1; Table 2.5 and Figure 2.6), their performance on their target proteins and the human proteome was tested.

3.2.3 Projection of the Models on Target Proteins and Human Proteome

Table 3.1. Target proteins of the monoclonal antibodies and the linear cognate sequences of the respective monoclonal antibodies on these proteins

mAbs[≠]	**Target Protein**	**UniProt ID of protein sequence**	**Known Length of the Protein Sequence**	**Cognate Sequence**	**Location of the Cognate Region on the Antigen**	**PDB ID of Available Protein Structures**[ǂ]
DM1A	Human α-tubulin	P68363	451	ALEKDYE	427-433	5IJ0
p53Ab1	Human cellular tumor antigen p53	P04637	390	RHSVV	212-216	4MZI
p53Ab8	Human cellular tumor antigen p53	P04637	390	SDLWKL	18-23	N/A*****
4C1	Human thyrotropin receptor	P16473	764	LQAFDSH	378-384	N/A*
LNKB2	Human interleukin-2	P60568	150	PLEEVLN	85-91	1M47

[≠]mAbs = monoclonal antibodies. [ǂ] www.rcsb.org * Crystal structure of the protein available on PDB does not cover the cognate region that the monoclonal antibody binds to.

Once trained, the models for the monoclonal antibodies were then projected on to their respective target antigen. The UniProtKB database (The UniProt Consortium, 2021) was used to retrieve the sequence information for these target antigen proteins (Table 3.1). As these models were trained to accommodate peptides only, the protein sequences had to be broken down into peptides to be able to project the model on to them. Thus, the sequence

of each individual protein was broken into peptide sequences of 9 amino acid residue each, starting from the first residue of the protein and gradually shifting up by one residue at a time. This generated an in silico array of overlapping peptides, that represented the entire protein. Figure 3.1 shows the manner in which the protein sequence was broken down into individual peptides in silico. The amino acids C, I, M, and T are not present on the array and therefore are beyond the scope of recognition by the trained models. Hence, these residues were substituted with A, V, L, and S respectively in their positions because of their physicochemical similarities with the respective omitted amino acids. However, the binding interactions of the monoclonal antibodies studied here could be represented with the 16 amino acids present on the arrays. The model of the respective monoclonal antibody was then projected on to this peptide library that represented the entire target protein.

After the projection was complete, these peptides were sorted out in descending order of predicted binding values. Before sorting out the peptides, the predicted binding values were normalized between [0,1] and these normalized predicted values were plotted against each peptide obtained from the target protein. After sorting the peptides, the top 10 sequences were considered, and the results were tabulated. The positions of the first residues of these sequences within the target protein was also indicated, to highlight the region of most importance according to the neural network. The top 5 peptides according to the predicted results were then visually represented on the crystal structure of the target protein. Out of a total of 5 different target proteins (one for each monoclonal antibody), only 3 have resolved crystal structures in the RCSB-PDB (www.rcsb.org; Berman et al., 2000) database (human α-tubulin, human cellular tumor antigen p53, human interleukin-2) that also covers the cognate region that binds to the respective monoclonal antibodies

(Table 3.1). For the other two, AlphaFold web version (Jumper et al., 2021; Varadi et al., 2021) was used to obtain predicted crystal structure and the relevant peptides were represented on those.

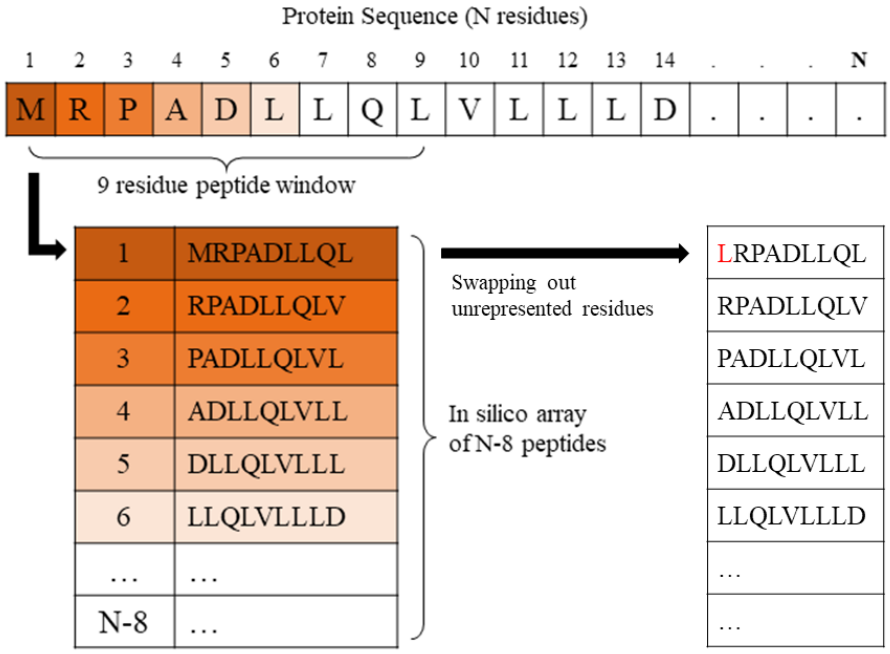


Figure 3.1. Schematic diagram showing the in silico breakdown of target protein sequence with N residues into a tiled peptide library of overlapping 9-mer peptides. After the generation of the in silico library, the amino acids residues that are not present on the array (C, I, M, and T) are swapped out with residues that have similar physicochemical properties to that of the omitted ones (A, V, L, and S respectively).

Next these models were projected on to human proteome (UniProt KB Proteome ID: UP000005640). The human proteome sequence dataset consists of 20,361 reviewed proteins along with their sequence information. These retrieved protein sequences were then broken down into an in silico library of 9-mer peptides, as shown in the Figure 3.1. However, before replacing the residues C, I, M, and T, all the sequences with more than 3 repeating residues were removed. This was done because the sequence space represented on the arrays does not have any peptides that have more than 3 of the same amino acids in

a single sequence. After removal of those peptide sequences, the total number of peptides came to 10,386,533. The models were then projected on to this library and the peptides were sorted in descending order of predicted binding values. The proteins associated with all the peptides were also located and sorted accordingly. The rank of the target protein and sequence was determined from this sorted list and was highlighted alongside the top 10 non-cognate peptides and proteins that were predicted to be the highest binders. The sequence motif from the predictive models were also shown to estimate the binding pattern recognized by the neural network among these sequences. The sequence logos were created using Weblogo (https://weblogo.berkeley.edu/) from the top 10 sequences predicted to bind the highest after aligning them using Clustal Omega (www.ebi.ac.uk/Tools/msa/clustalo/).

3.3 RESULTS

3.3.1 Projection Results on Target Antigens of Monoclonal Antibodies

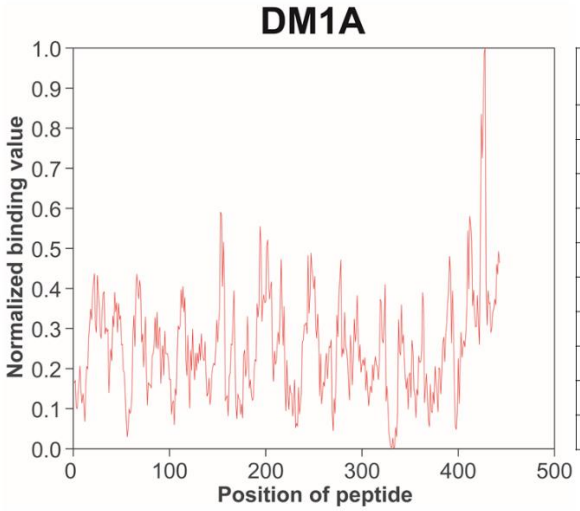
In chapter 2, the binding behavior of five different monoclonal antibodies were probed with the help of neural network models. These models were trained to recognize and predict binding sequence motifs for the antibodies through analyzing 126,050 peptides that were randomly sampled from a combinatorial chemical sequence space. The cognate sequences of the monoclonal antibodies were not included during training of these models. One must keep in mind that the biological sequence space is represented by all the 20 canonical amino acids, which is not the case for the array peptides. Therefore, substitutions had to be made to accommodate those omitted residues from the array in order to project the neural network on the biological sequence space, as the model otherwise would not recognize these residues. The omitted residues were substituted with other residues from

the array based on the closest physicochemical similarities between the residues. However, these substitutions are not accurate representations of the actual residues themselves. In future, using peptide arrays with all 20 amino acids would help represent the combinatorial space better, without the need for substitution.

The performance of the models solely relies on information from off-target interactions (mimotopes and near-cognates mostly). Figures 2.4 and 2.6, along with Table 2.5 gives a summary of the predictive performance of the model. Table 2.5 and 2.6 summarizes the results from projecting the antibody-specific models on to in silico libraries of 10^6 random peptides, which were also sparsely sampled from the combinatorial sequence space, sparing the inclusion of the cognate sequences. From this table, it can be seen that the models work quite well for the five monoclonal antibodies, correctly identifying the epitopes among a large number of randomly sampled peptides. Figure 2.6 also depicts the motifs identified by the antibody-specific models from the non-cognate peptides predicted to be the top binders from the random peptides' library. However, these random peptides are not representative of the available biological sequence space (proteomes and individual target proteins). Therefore, it was a prerogative to probe how do the models' predictive capabilities fare against actual biologically available sequence space in the form of target proteins. In order to investigate this, the target protein sequence of each antibody was deconstructed into a tiled library of overlapping 9-mer peptides. For a protein with N number of amino acid residues, the total number of tiled peptides obtained would be N-8, which covers the entirety of the protein sequence. The model for each respective antibody, consisting of 100 independently trained runs, with randomly selected

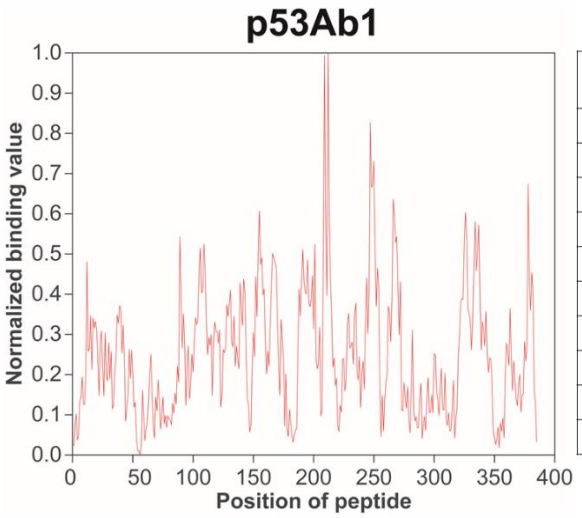
training set for each run, were then projected onto projected onto this in silico library of 9-mer peptides that represent the entire antigenic protein sequence.

Figures 3.2, 3.3, 3.4, 3.5 and 3.6 represents the performance of the antibody-specific models on their respective target antigens. The plots shown in these figures represent normalized predicted binding intensities for all the peptides that constitute the respective target proteins. The tables show the highest binding peptides within the target proteins as predicted by the neural network models. In all the cases, the cognate motifs of the respective monoclonal antibodies were predicted to be the highest binders by the models, corresponding to the highest peaks in the binding intensities vs. peptide position plots. For three out of five antibodies (DM1A, 4C1, and LNKB2), five or more peptides out of the predicted top 10 peptides represent the cognate interactions either entirely or partially (Figures 3.2, 3.5, and 3.6).



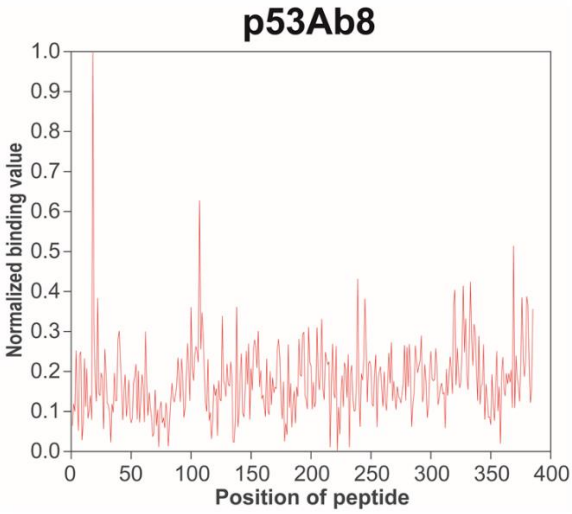
Peptides	Residue Positions
LEKDYEEVG	428 - 436
ALEKDYEEV	427 - 435
DLAALEKDY	424 - 432
AALEKDYEE	426 - 434
LAALEKDYE	425 - 433
LLERLSVDY	153 - 161
LERLSVDYG	154 - 162
GLEEGEFSE	412 - 420
LEEGEFSEA	413 - 421
SLEHSDAAF	194 - 202

Figure 3.2. Projection of the neural network model for DM1A on to the peptide sequences obtained from its target antigen, human α-tubulin (451 residues). The plot on the left shows the variation in predicted binding intensities for each peptide normalized between [0,1]. The table on the right represents the top 10 peptide sequences from the target protein, as predicted by the model, in descending order of binding intensity. The residues highlighted in red represents the cognate region and parts of it.



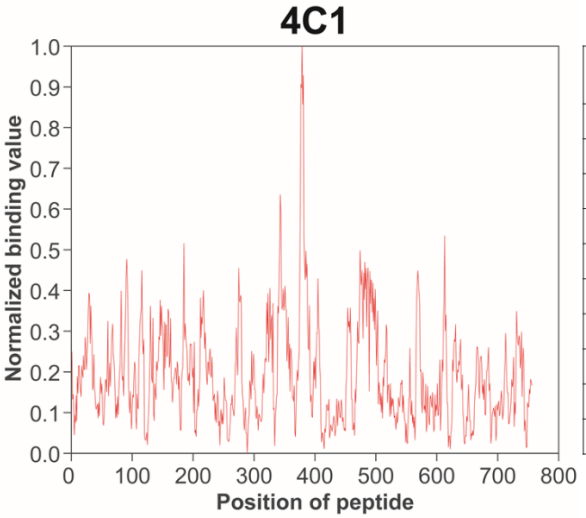
Peptides	Residue Positions
FRHSVVVPY	212 - 220
RNSFRHSVV	209 - 217
NRRPVLSVV	247 - 255
PVLSVVSLE	250 - 258
SRHKKLLFK	378 - 386
RPVLSVVSL	249 - 257
RRPVLSVVS	248 - 256
GRNSFEVRV	266 - 274
SRVRALAVY	155 - 163
EYFSLQVRG	326 - 334

Figure 3.3. Projection of the neural network model for p53Ab1 on to the peptide sequences obtained from its target antigen, human p53 (390 residues). The plot on the left shows the variation in predicted binding intensities for each peptide normalized between [0,1]. The table on the right represents the top 10 peptide sequences from the target protein, as predicted by the model, in descending order of binding intensity. The rightmost column indicates the location of the peptides within the protein. The residues highlighted in red represents the cognate region and parts of it.



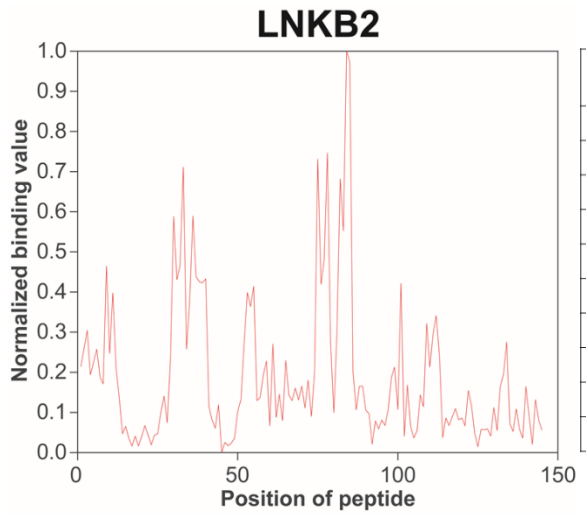
Peptides	Residue Positions
SFSDLWKLL	18 - 26
YGFRLGFLH	107 - 115
LKSKKGQSS	369 - 377
NSSALGGLN	239 - 247
RGRERFELF	333 - 341
YFSLQVRGR	327 - 335
KKPLDGEYF	320 - 328
HKKLLFKSE	380 - 388
SSSRHKKLL	376 - 384
LWKLLPENN	22 - 30

Figure 3.4. Projection of the neural network model for p53Ab8 on to the peptide sequences obtained from its target antigen, human p53 (390 residues). The plot on the left shows the variation in predicted binding intensities for each peptide normalized between [0,1]. The table on the right represents the top 10 peptide sequences from the target protein, as predicted by the model, in descending order of binding intensity. The residues highlighted in red represents the cognate region and parts of it.



Peptides	Residue Positions
QAFDSHYDY	379 - 387
FDSHYDYSV	381 - 389
SLQAFDSHY	377 - 385
LQAFDSHYD	378 - 386
AFDSHYDYS	380 - 388
FQDSHNNAH	343 - 351
ESLQAFDSH	376 - 384
QDSHNNAHY	344 - 352
YNPGDKDSK	613 - 621
DSHYDYSVA	382 - 390

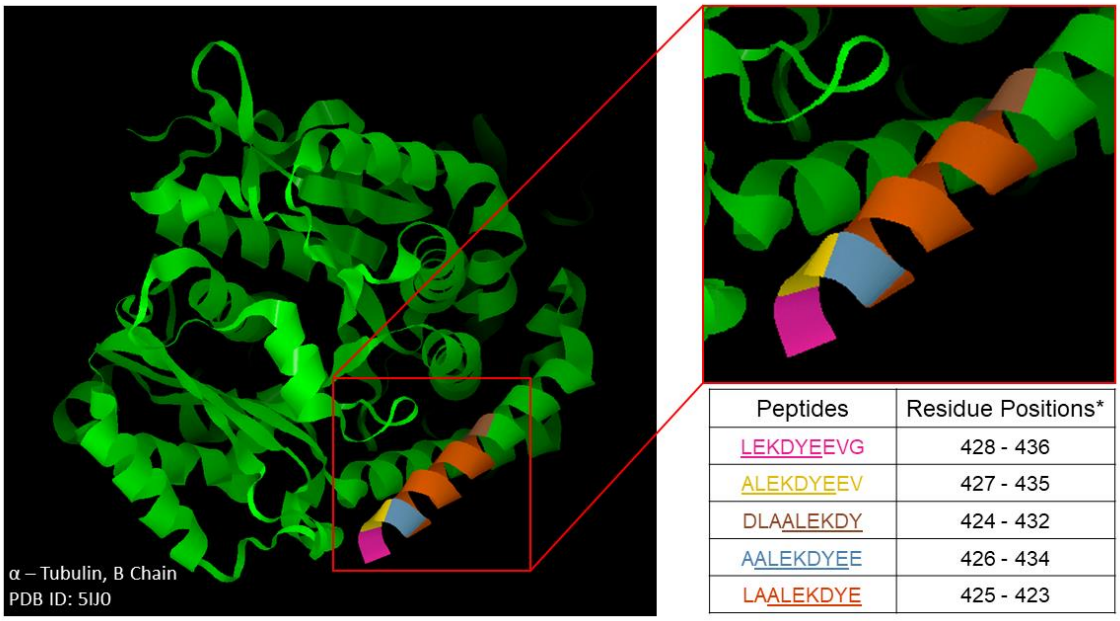
Figure 3.5. Projection of the neural network model for 4C1 antibody on to the peptide sequences obtained from its target antigen, human thyrotropin receptor (764 residues). The plot on the left shows the variation in predicted binding intensities for each peptide normalized between [0,1]. The table on the right represents the top 10 peptide sequences from the target protein, as predicted by the model, in descending order of binding intensity. The residues highlighted in red represents the cognate region and parts of it.



Peptides	Residue Positions
KPLEEVLNL	84 - 92
PLEEVLNLA	85 - 93
ALEEELKPL	78 - 86
HLQALEEEL	75 - 83
QLEHLLLDL	33 - 41
ELKPLEEVL	82 - 90
HLLLDLQLV	36 - 44
SQLQLEHLL	30 - 38
LKPLEEVLN	83 - 91
QALEEELKP	77 - 85

Figure 3.6. Projection of the neural network model for LNKB2 antibody on to the peptide sequences obtained from its target antigen, human interleukin-2 (150 residues). The plot on the left shows the variation in predicted binding intensities for each peptide normalized between [0,1]. The table on the right represents the top 10 peptide sequences from the target protein, as predicted by the model, in descending order of binding intensity. The residues highlighted in red represents the cognate region and parts of it.

For antibodies p53Ab1 and p53Ab8 (Figures 3.3 and 3.4), the cognate rose to the top ten in the context of two sequences and the other nearby sequences in the structure made up a number of the other sequences. Some of the non-cognate sequences that were predicted as top binders were found to be in close proximity of the cognate binding region.



Peptides	Residue Positions*
LEKDYEEVG	428 - 436
ALEKDYEEV	427 - 435
DLAALEKDY	424 - 432
AALEKDYEE	426 - 434
LAALEKDYE	425 - 423

Figure 3.7. Top five peptides predicted as the strongest binders by the neural network model for DM1A represented on the crystal structure of its target protein, human α-tubulin, B chain (PDB ID: 5IJ0, residues 1-437). The peptides are part of the actual protein sequence and ranked in descending order of predicted binding value. The position of each peptide on the crystal structure have been color-coded as shown in the table included in the figure. *Residue positions correspond to the actual residues on the protein. Regions of cognate residues in the peptides are highlighted through underlining.

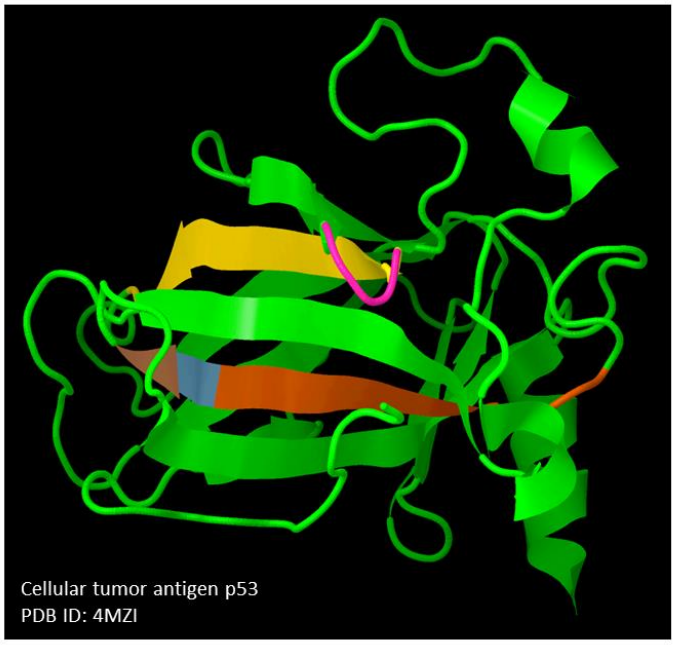
To visualize this better in a structural context, the top peptides were represented on the 3-dimensional (3D) structure of the target proteins. For monoclonal antibodies DM1A, p53Ab1, and LNKB2, the resolved 3D crystal structures of their respective target proteins that also covered their cognate regions were available on the RCSB-PDB (www.rcsb.org; Berman et al., 2000) database (Table 3.1). For 4C1 and p53Ab8, the resolved crystal structures of the target proteins did not contain the relevant cognate binding regions.

Therefore, AlphaFold web program was to obtain the predicted 3D structure of their target proteins, and the top binding peptides were represented on them. Figures 3.7, 3.8, 3.9, 3.10, and 3.11 show the locations of the top 5 predicted peptides on the crystal structures of the target proteins.

In Figure 3.7, the target protein of DM1A, α-tubulin is represented (PDB ID: 5IJ0; Ti et al., 2016). The strongest binding regions predicted according to the trained neural network model are shown on the crystal structure, in a color coded manner. As one can see in the figure, all of the top 5 predicted peptides (residues 424 – 436) form the part of a continuous α-helix and they contain some or all residues of the cognate sequence for DM1A (ALEKDYE). Especially, the motif LEKDY is common to all of the five top predicted regions, indicating that the model was able to correctly identify the residues that play the most crucial role in the interaction of α-tubulin with DM1A. All of the top peptides that were predicted to be the strongest binders, overlap with the cognate sequence to an extent, primarily the motif LEKDY. It also hints towards the observation that the neighboring non-cognate residues that are not part of the reported epitope might also contribute to the binding interactions of the protein with the monoclonal antibody.

In case of human cellular tumor antigen p53 (Figures 3.8 and 3.9) which is the target protein for both p53Ab1 and p53Ab8, the resolved crystal structure for the entire protein is not available on RCSB-PDB database due to the flexible nature of the molecule which makes its structure difficult to determine. Therefore, only domains that are stable enough have been resolved structurally. The domain of p53 with a β-barrel (residues 96-288; PDB ID: 4MZI; Emamzadah et al., 2014), is the region where the monoclonal antibody p53Ab1 binds. In Figure 3.8, the crystal structure of this domain is shown along

with the peptides that are predicted to be the highest binders of p53Ab1 by the neural network algorithm (residues 209-220 and 248-258). The predicted binding region for the antibody, along with the known cognate sequence (RHSVV, residues 213-217) are part of the strands of β-barrel. One might observe that the 3 peptides that represent the non-cognate residues on the protein (residues 248-258) are in close proximity to the cognate sequence. These sequences bind outside the cognate region and therefore suggest these noncognate sequences may be important.

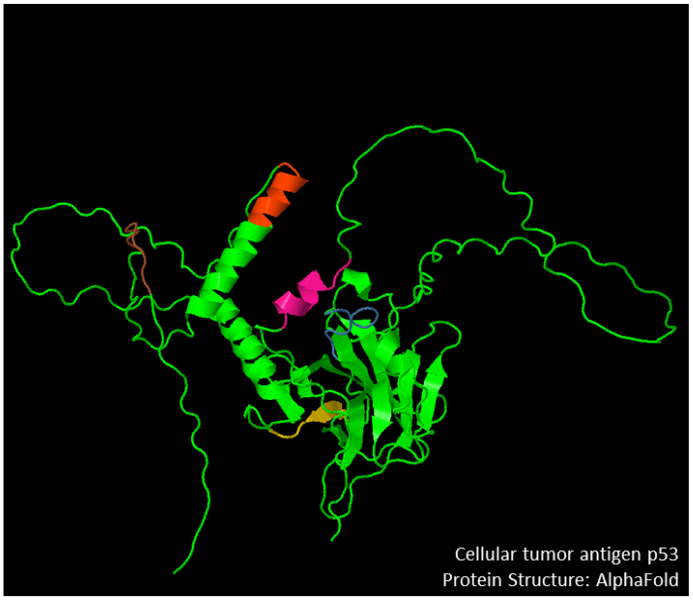


Peptides	Residue Positions*
PNSFRHSVV	209 - 217
FRHSVVPY	212 - 220
PVLVVVSLE	250 - 258
RPVLVVVSL	249 - 257
RRPVLVVVS	248 - 256

Figure 3.8. Top five peptides predicted as the strongest binders by the neural network model for p53Ab1 represented on the crystal structure of its target protein, human cellular tumor antigen p53 (PDB ID: 4MZI, residues 96-288). The peptides are part of the actual protein sequence and ranked in descending order of predicted binding value. The position of each peptide on the crystal structure have been color-coded as shown in the table included in the figure. *Residue positions correspond to the actual residues on the protein. Regions of cognate residues in the peptides are highlighted through underlining.

The crystal structure for the domain of p53 where p53Ab8 binds was not found in literature. Hence AlphaFold was used to predict an estimate structure of the protein p53.

The top 5 predicted binders of p53Ab8, were represented on this predicted structure of the protein (Figure 3.9). As can be seen from the figure, the cognate sequence SDLWKL (also predicted to be the highest binder), is a small helical region that looks like a part of a very flexible domain of the protein, according to the predicted structure. The positions of the other predicted regions, which are non-cognate residues, cannot be determined with respect to the location of the cognate region from the predicted structure alone because of the flexibility of the p53 protein molecule. Therefore, it is difficult to estimate if any of the non-cognate regions listed here also bind to p53Ab8 or not.

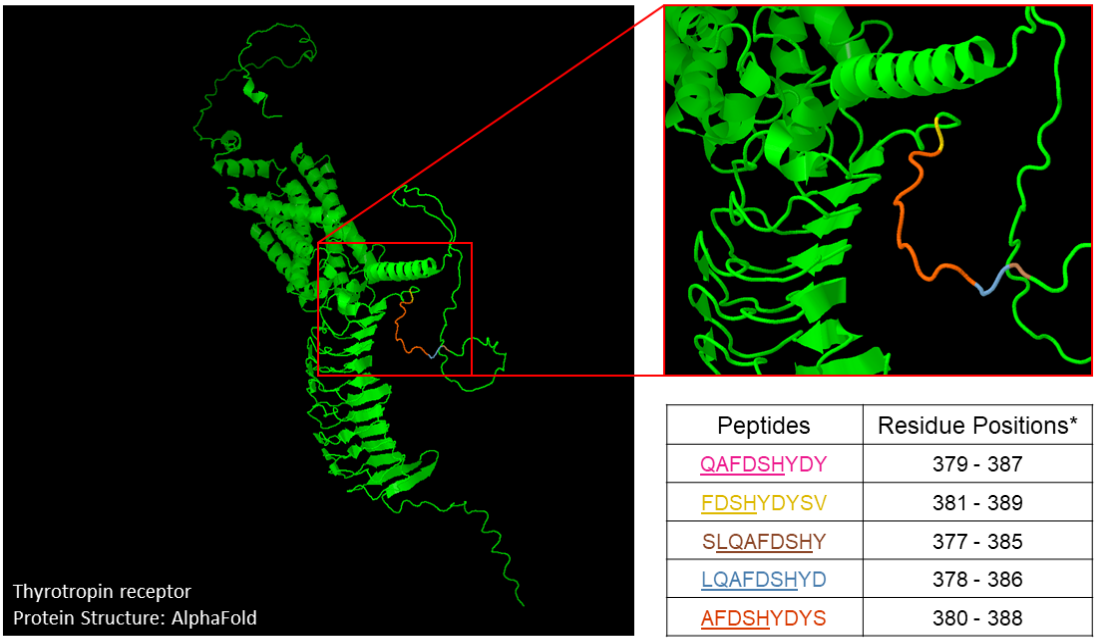


Peptides	Residue Positions*
SFSDLWKLL	18 - 26
YGFRLGFLH	107 - 115
LKSKKGQSS	369 - 377
NSSALGGLN	239 - 247
RGRERFELF	333 - 341

Figure 3.9. Top five peptides predicted as the strongest binders by the neural network model for p53Ab8 represented on the predicted crystal structure of its target protein, human cellular tumor antigen p53 (https://alphafold.ebi.ac.uk/). The peptides are part of the actual protein sequence and ranked in descending order of predicted binding value. The position of each peptide on the crystal structure have been color-coded as shown in the table included in the figure. *Residue positions correspond to the actual residues on the protein. Regions of cognate residues in the peptides are highlighted through underlining.

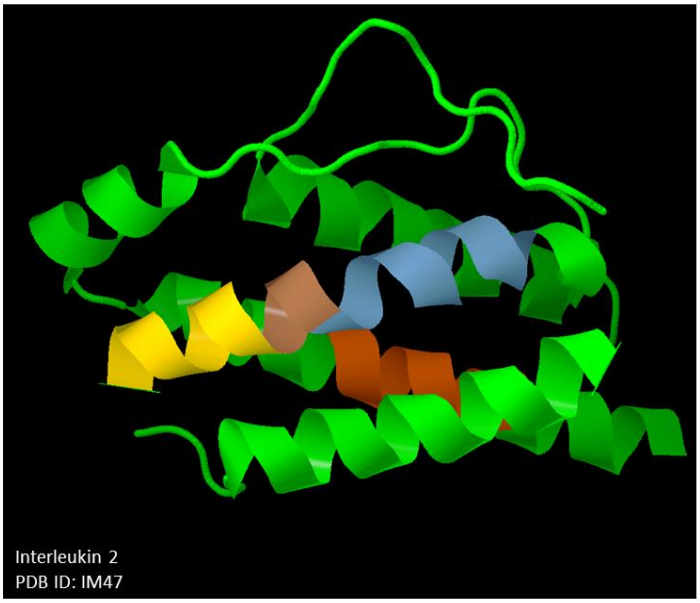
For 4C1, the target protein is human thyrotropin receptor. The entire crystal structure of this protein is also not available in literature. Therefore, the structure predicted by AlphaFold was used to demonstrate the results of the projection. As can be seen from

Figure 3.10, the highest binding peptides predicted by the model cover the group of residues between 377 – 389, which also covers the cognate binding region for 4C1. All of the five predicted high binders include parts of the cognate sequence, particularly the motif FDSH (381-384), which is considered to be an essential set of residues that bind to 4C1 (Shepherd et al., 1998). The other residues that are not part of the known linear epitope for 4C1 still might play an important role in the binding interactions because of their close proximity to the epitope (flanking residues). However, according to the predicted structure, the cognate region and its flanking residues fall in a non-rigid region. Therefore, it is difficult to estimate the type of interactions that they undergo.



Peptides	Residue Positions*
QAFDSHYDY	379 - 387
FDSHYDYSV	381 - 389
SLQAFDSHY	377 - 385
LQAFDSHYD	378 - 386
AFDSHYDYS	380 - 388

Figure 3.10. Top five peptides predicted as the strongest binders by the neural network model for 4C1 represented on the predicted crystal structure of its target protein, human thyrotropin receptor (https://alphafold.ebi.ac.uk/). The peptides are part of the actual protein sequence and ranked in descending order of predicted binding value. The position of each peptide on the crystal structure have been color-coded as shown in the table included in the figure. *Residue positions correspond to the actual residues on the protein. Regions of cognate residues in the peptides are highlighted through underlining.



Peptides	Residue Positions*
KPLEEVLNL	64 - 72
PLEEVLNLA	65 - 73
ALEEELKPL	58 - 66
HLAQALEEEL	55 - 63
QLEHLLLDL	13 - 21

Figure 3.11. Top five peptides predicted as the strongest binders by the neural network model for LNKB2 represented on the crystal structure of its target protein, interleukin 2 (PDB ID: 1M47, residues 6-133). The peptides are part of the actual protein sequence and ranked in descending order of predicted binding value. The position of each peptide on the crystal structure have been color-coded as shown in the table included in the figure. *Residue positions correspond to the actual residues on the protein. Regions of cognate residues in the peptides are highlighted through underlining.

In case of LNKB2, the crystal structure of its target protein, interleukin-2 was documented in the RCSB-PDB database (PDB ID: 1M47; Arkin et al., 2003). Figure 3.11 shows the protein along with the top peptide binders that were predicted by the model. The known linear epitope for 4C1 lies between residues 65 – 71 (PLEEVLN). Four out of the top five predicted binders lie between residues 55 – 73, that form a helical arm in the protein. Similar to the previous cases (DM1A, p53Ab1, and 4C1), the non-cognate residues flanking the known epitope residues might play a role in the interaction between the antibody and the protein, given that the residues are in such close proximity. These results show that the algorithm is able to identify the relevant binding motifs on the target protein

precisely. It is able to identify non-cognate residues that might play a part in the binding interactions of the antibody.

3.3.2 Projection Results on the Human Proteome

In the previous section, the performance of the algorithm was tested on target proteins of the respective monoclonal antibodies. The predictions were very specific in all the cases with the model correctly identifying the cognate binding motifs. It also successfully predicted regions on the antigen that might be contributory to the binding phenomena. But what happens when one wants to test the predictive capabilities of the neural network to see if it can identify the target protein/motif correctly from an entire proteome consisting of thousands of proteins? Will the algorithm also be able to identify similar binding motifs present in other proteins of the proteome?

In order to investigate these questions, the antibody-specific trained neural network models were projected on to the entire human proteome. The human proteome database from UniProtKB (Proteome ID: UP000005640) consists of 20,361 unique protein sequences. These protein sequences were parsed into an in silico library of 9-mer peptides. After parsing, sequences with more than 3 repeats of the same amino acid residue were removed. The total number of peptides in the resulting library was equal to 10,386,533. The size of this library was an order higher than the largest libraries (10^6 peptides) on which the models had been projected so far. Thus, it provided a larger sequence space to test the performance of the models. The models that were used for projecting on the human proteome were the same as the ones used for projection on to the target proteins (section 3.2.2). After the projection was complete, the peptides were ranked in descending order of

binding intensities (log10 scale). The ranks of the proteins were determined from their highest predicted peptide, which represents a binding motif on that protein itself. The results were tabulated as shown below (Table 3.2, 3.3, 3.4, 3.5, and 3.6).

Table 3.2. Top predicted peptides from the human proteome for DM1A and their corresponding proteins. Rank shown here is the rank of the protein target. Rank 1 corresponds to highest predicted binding interaction. The rank and protein ID of the actual target protein is represented in bold letters. The residues from the epitope of DM1A in the target protein are highlighted in red. Regions of the non-cognate peptides that are similar to the cognate residues are highlighted in bold letters.

Rank	Protein ID	Top Predicted Peptide
1	sp|Q6AI14|SL9A4_HUMAN Sodium/hydrogen exchanger	VF**E**L**DY**DYV
2	sp|Q9NQY0|BIN3_HUMAN Bridging integrator 3	V**E**R**D**F**E**REY
3	sp|Q9HC78|ZBT20_HUMAN Zinc finger and BTB domain-containing protein 20	E**ME**D**DY**DYY
4	sp|Q15393|SF3B3_HUMAN Splicing factor 3B subunit 3	**LE**M**DYE**EAD
5	sp|Q5TBA9|FRY_HUMAN Protein furry homolog	**ME**S**D**F**E**FEY
6	sp|Q6EMK4|VASN_HUMAN Vasorin	L**LE**L**DY**ADF
7	sp|Q9P281|BAHC1_HUMAN BAH and coiled-coil domain-containing protein 1	V**EE**D**FE**FDD
8	sp|O94915|FRYL_HUMAN Protein furry homolog-like	**LE**S**DYE**YEY
9	sp|Q8NFA0|UBP32_HUMAN Ubiquitin carboxyl-terminal hydrolase 32	**M**DE**D**F**E**SD
10	sp|Q9HB65|ELL3_HUMAN RNA polymerase II elongation factor ELL3	Y**EQD**F**E**TDY
…	…	…
39	**sp|P68363|TBA1B_HUMAN Tubulin alpha-1B chain**	LEKDYEEVG

In Table 3.2, the performance of the DM1A-specific model on the human proteome is shown. It is worth noting here and in the following cases that the 4 amino acids that were

not present on the array are being substituted with other amino acids. This was done to incorporate the information of those amino acids in a way that is recognizable to the model. However, these substitutions do not necessarily accurately represent the amino acids that were omitted during the array manufacturing process. From the table, it can be seen that the target protein is ranked 39th among the list of top predicted targets for DM1A, according to the model. In percentage, the algorithm places the target protein, α-tubulin, within the top 0.2% out of the 20,361 proteins. The top predicted regions in the other proteins (top 10) have a similar motif (LEXDY, where X can be any residue) to that of the cognate sequence of DM1A (ALEKDYE, Figure 3.12). As all occurrences of M were substituted with L during the projection, methionine here was treated similar to leucine.

Table 3.3. Top predicted peptides from the human proteome for p53Ab1 and their corresponding proteins. Rank shown here is the rank of the protein target. Rank 1 corresponds to highest predicted binding interaction. The rank and protein ID of the actual target protein is represented in bold letters. The residues from the epitope of p53Ab1 in the target protein are highlighted in red. Regions of the non-cognate peptides that are similar to the cognate residues are highlighted in bold letters.

Rank	Protein ID	Top Predicted Peptide
1	sp|P55286|CADH8_HUMAN Cadherin-8	IILLL**VIV**V
2	'sp|Q96DX8|RTP4_HUMAN Receptor-transporting protein 4	FILLLVF**IV**
3	sp|Q14627|I13R2_HUMAN Interleukin-13 receptor subunit alpha-2	FILIL**VI**FV
4	sp|P43355|MAGA1_HUMAN Melanoma-associated antigen 1	**F**LII**V**LVMI
5	sp|Q7Z333|SETX_HUMAN Probable helicase senataxin	**F**LLILVSVI
6	sp|O60637|TSN3_HUMAN Tetraspanin-3	**VII**LLLVFV
7	sp|Q14643|ITPR1_HUMAN Inositol 1,4,5-trisphosphate receptor type 1	**F**FFM**VII**IV
8	sp|Q8WXG9|AGRV1_HUMAN Adhesion G-protein coupled receptor V1	F**VVI**LLIVI

9	sp|Q96R09|OR5B2_HUMAN Olfactory receptor 5B2	F**F**VLL**VI**FI
10	sp|Q86UG4|SO6A1_HUMAN Solute carrier organic anion transporter family member 6A1	ILL**V**FIIFV
…	…	…
1918	**'sp|P04637|P53_HUMAN Cellular tumor antigen p53**	FRHSVVVPY

Table 3.4. Top predicted peptides from the human proteome for p53Ab8 and their corresponding proteins. Rank shown here is the rank of the protein target. Rank 1 corresponds to highest predicted binding interaction. The rank and protein ID of the actual target protein is represented in bold letters. The residues from the epitope of p53Ab8 in the target protein are highlighted in red. Regions of the non-cognate peptides that are similar to the cognate residues are highlighted in bold letters.

Rank	**Protein ID**	**Top Predicted Peptide**
1	sp|Q6UVJ0|SAS6_HUMAN Spindle assembly abnormal protein 6 homolog	LK**T**L**M**G**KL**K
2	sp|O95498|VNN2_HUMAN Vascular non-inflammatory molecule 2	MK**TEL**G**KL**L
3	sp|P53778|MK12_HUMAN Mitogen-activated protein kinase 12	MG**TDL**G**KL**M
4	sp|Q9BXL8|CDCA4_HUMAN Cell division cycle-associated protein 4	CK**SDL**GE**L**D
5	sp|Q4AE62|GTDC1_HUMAN Glycosyltransferase-like domain-containing protein 1	LRP**DL**G**KL**K
6	sp|Q7Z7B0|FLIP1_HUMAN Filamin-A-interacting protein 1	LKD**DL**T**KL**K
7	sp|Q4L180|FIL1L_HUMAN Filamin A-interacting protein 1-like	LKE**DL**T**KL**K
8	sp|P0C645|OR4E1_HUMAN Olfactory receptor 4E1	MK**S**A**L**N**KL**V
9	sp|Q8NGL6|O4A15_HUMAN Olfactory receptor 4A15	MK**S**A**M**R**KL**W
10	sp|P51170|SCNNG_HUMAN Amiloride-sensitive sodium channel subunit gamma	NK**TDL**A**KL**L
…	…	…
452	**'sp|P04637|P53_HUMAN Cellular tumor antigen p53**	TFSDLWKLL

In case of p53Ab1 (Table 3.3), the predicted rank of the target protein, p53, is at the 1918th position in the sorted list, which is around top 10% of the total number of proteins. That is because the arginine (R), histidine (H), and serine (S) residues that are present in the cognate sequence are not very well recognized by the model (Figure 2.6). Instead, aliphatic residues like valine (V), leucine (L), isoleucine (I), and phenylalanine (F) are preferred more. While V and F are essential residues in interacting with the antibody, L and I are not part of the cognate region. However, it has been indicated in experimental literature that the V(4) and V(5) of the known cognate sequence can be replaced with I and/or L and the mutation is favorable in binding towards p53Ab1 (Stephen et al., 1995). Also, all occurrences of I in the proteome were replaced with V, as I was not represented on the microarrays.

Table 3.5. Top predicted peptides from the human proteome for 4C1 and their corresponding proteins. Rank shown here is the rank of the protein target. Rank 1 corresponds to highest predicted binding interaction. The rank and protein ID of the actual target protein is represented in bold letters. The residues from the epitope of 4C1 in the target protein are highlighted in red. Regions of the non-cognate peptides that are similar to the cognate residues are highlighted in bold letters.

Rank	Protein ID	Top Predicted Peptide
1	sp|Q8N4C6|NIN_HUMAN Ninein	**FDS**F**D**TTGT
2	sp|Q8WU20|FRS2_HUMAN Fibroblast growth factor receptor substrate 2	W**DT**G**YD**SDE
3	sp|O43548|TGM5_HUMAN Protein-glutamine gamma-glutamyltransferase 5	**FDS**G**H**DTDG
4	sp|Q9UPX8|SHAN2_HUMAN SH3 and multiple ankyrin repeat domains protein 2	Y**DSFDT**SSD
5	sp|Q9ULX6|AKP8L_HUMAN A-kinase anchor protein 8-like	Y**DS**YESCDS
6	sp|Q14517|FAT1_HUMAN Protocadherin Fat 1	Y**DSH**F**D**VDK

7	sp|O60264|SMCA5_HUMAN SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A member 5	**FDS**WF**D**TNN
8	sp|Q9BVM4|GGACT_HUMAN Gamma-glutamylaminecyclotransferase	HH**DS**YDSEG
9	sp|Q9H0J4|QRIC2_HUMAN Glutamine-rich protein 2	**FDSH**DSMYP
10	sp|Q6P2Q9|PRP8_HUMAN Pre-mRNA-processing-splicing factor 8	Y**DSH**DIERY
…	…	…
18	**sp|P16473|TSHR_HUMAN Thyrotropin receptor**	QAFDSHYDY

Cellular tumor antigen p53 is also the target protein for p53Ab8. The cognate binding sequence (SDLWKL) is different from that of p53Ab1. In this case the target protein was ranked at 452nd position with respect to the correct cognate region on the protein, therefore coming up in the top 2.2%. It was observed before that model actually prefers the motif LKSDLXKL (X can be any amino acid) instead of SDLXKL, where the serine (S) residue of the cognate peptide sequence comes at the third position instead of the first position. However, mutations at the first (L) and second (K) positions are well tolerated by the algorithm, as long as the S residue is at the third position. In case of the top peptides observed here however, threonine (T) is more prominent at the third position, because all the occurrences of T were replaced with S prior to projecting the model on the proteome, as T was one of the omitted residues that were not represented on the microarray.

For 4C1, the target protein, thyrotropin receptor, appears at the 18th rank in the list (top 0.09%). The motif that appears to be more common in the other peptides that XDS where X could be either F or Y. The residues L, Q, and A from the cognate sequence are not represented well by the top predicted binders, indicating that they are not considered to essential by the algorithm to represent binding motif pattern of 4C1 (Figure 3.12).

The target protein of LNKB2, interleukin-2 was predicted to be at the 105th position which was within the top 0.52% of all the proteins present in the human proteome. According to the model, XPLE (X = R, P) appears to be the most prominent motif in the top predicted peptides from the proteome, that are not part of the target protein itself (Figure 3.12).

Table 3.6. Top predicted peptides from the human proteome for LNKB2 and their corresponding proteins. Rank shown here is the rank of the protein target. Rank 1 corresponds to highest predicted binding interaction. The rank and protein ID of the actual target protein is represented in bold letters. The residues from the epitope of LNKB2 in the target protein are highlighted in red. Regions of the non-cognate peptides that are similar to the cognate residues are highlighted in bold letters.

Rank	**Protein ID**	**Top Predicted Peptide**
1	sp|A6NGG8|PCARE_HUMAN Photoreceptor cilium actin regulator	P**PME**VL**M**DK
2	sp|Q9NX74|DUS2L_HUMAN tRNA-dihydrouridine (20) synthase [NAD(P)+]-like	R**PLEEV**MQK
3	sp|O95072|REC8_HUMAN Meiotic recombination protein REC8 homolog	**PME**MP**L**VLP
4	sp|P52895|AK1C2_HUMAN Aldo-keto reductase family 1 member C2	RL**LE**M**IL**NK
5	sp|A8K0R7|ZN839_HUMAN Zinc finger protein 839	P**PLE**KI**L**SV
6	sp|Q9BZC7|ABCA2_HUMAN ATP-binding cassette sub-family A member 2	R**MEE**L**L**LAP
7	sp|P42285|MTREX_HUMAN Exosome RNA helicase MTR4	R**LEE**LLRQM
8	sp|O43566|RGS14_HUMAN Regulator of G-protein signaling 14	**PLE**V**VL**HRP
9	sp|P51530|DNA2_HUMAN DNA replication ATP-dependent helicase/nuclease DNA2	P**PLE**KL**L**NH
10	sp|Q9HCU4|CELR2_HUMAN Cadherin EGF LAG seven-pass G-type receptor 2	R**PLE**AIMSV
…	…	…
105	**sp|P60568|IL2_HUMAN Interleukin-2**	KPLEEVLNL

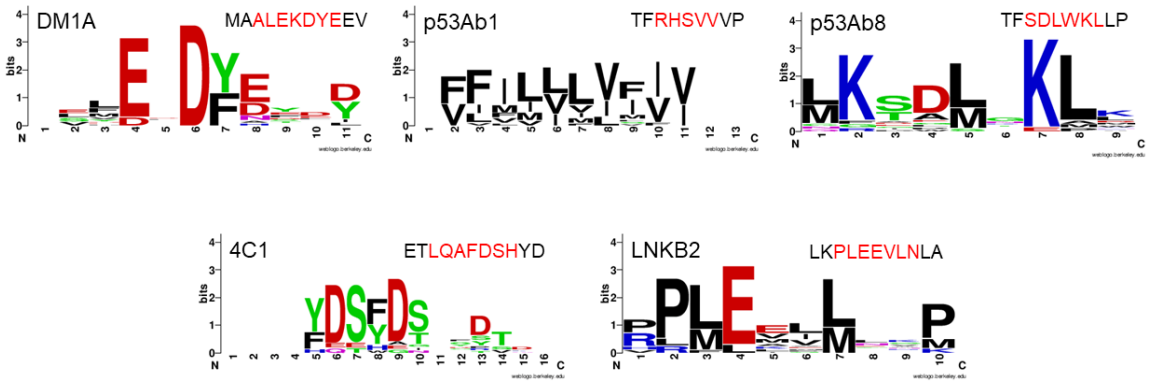


Figure 3.12. Sequence logos showing the residue motifs recognized by the neural network algorithm for each monoclonal antibody from the top 10 peptides predicted to be their best binders from the human proteome (Tables 3.2 – 3.6). Part of the target protein sequence for the respective antibody is shown on the top right corner of each sequence logo, where the epitope region is shown in red. All the sequence logos are created with the help of WebLogo (https://weblogo.berkeley.edu/). Negatively charged residues (D, E) are colored in red, positively charged residues (H, K, R) are blue and residues with a hydroxyl group (S, Y) are in green. All other residues are represented in black.

These results show that the antibody-specific models are able to recognize the correct target motifs even among a large library of non-random peptides (~10^7) that represent the biological sequence space as covered by the human proteome. All the target proteins for the respective monoclonal antibodies were identified among the top 10% out of the total number of available proteins. In four out of five cases studied here (DM1A, p53Ab8, 4C1 and LNKB2), the target protein was predicted well within top 2.2%. Thus, by using a neural network to analyze the interactions of these five monoclonal antibodies with combinatorial peptides that were randomly and sparsely sampled from sequence space, it was possible to greatly narrow the range of potential protein targets and that in the off target proteins, sequences similar to the cognate were recognized.

3.4. DISCUSSION

In this chapter the results of projecting the neural network models on the target protein of each antibody and the human proteome was shown. Compared to the previous

chapter, where the performance of the algorithm was tested on randomly generated combinatorial peptide sequences, actual biological sequences were used here. This was done because, random peptides that were generated in silico are not comparable to the actual biological sequence space represented by various proteins. Out of a total combinatorial possibility of 16^9 unique sequences (because only 16 amino acids were used on the array) for a 9-mer peptide, comparatively a much smaller fraction (10^7 9-mer peptides) is representative of the actual biological sequence space covered by the human proteome. But the proteins that constitute the human proteome have evolved over many years to carry out specific biological functions. The aim was to be able to use a neural network trained on random sequences for predicting interactions between the target and the monoclonal antibody from among such specific sequence space.

At first, the algorithm was tested only on the target proteins of the monoclonal antibodies that were studied. The antibody-specific models were projected onto the target proteins, which were parsed into tiled arrays of 9-mer peptides that represented the entire protein sequences. The reason for such parsing was that the model had been trained on array peptides only that ranged from 4 to 13-mers in length. Since the median length of those peptides were 9, therefore the protein sequence was broken down to 9-mers. The models were trained using ‘one-hot encoding’ only (see section 2.2.3, Figure 2.1). Each antibody-specific model consisted of 100 independently trained runs that did not include the cognate sequence(s) of that antibody in the training set. No physicochemical propensities of the amino acids were used as external encoders for training the models in this case. The results from these projections are shown in Figures 3.2 to 3.6.

In the case of all the five monoclonal antibodies studied here, projecting the models on their target proteins resulted in the cognate region being predicted as the highest binder, corresponding to the highest peaks in Figures 3.2 to 3.6. The tables included in the figures also list the other peptides that were predicted to be the highest binders (top 10) of their respective monoclonal antibody. For DM1A, each of the top 5 predicted peptides in its target protein, α-tubulin, constitute of the known cognate binding region (ALEKDYE, residues 427 - 433) either entirely or partially. This has been depicted when these peptides were mapped on to the crystal structure of the protein (Figure 3.7). From the overlapping regions of the top predicted peptides, it can be estimated that residues that flank the main cognate motif (LEKDY) and are not part of the reported cognate region might also have a role to play in the interaction with DM1A. Indeed, an experimental study (Breitling and Little, 1986) did report residues 426 to 451 of α-tubulin as the region responsible for binding to DM1A, but only residues 427 to 431 (ALEKD) were considered essential for binding according to the article. Therefore, the other residues might play an assistive role in the binding interaction. The model was able to correctly identify those interactions in this case.

p53Ab1 and p53Ab8 interact with the same target protein, cellular tumor antigen p53, but at different regions. p53Ab1's known epitope (Stephen et al., 1995) consists of residues 213-217 (RHSVV) whereas p53Ab8's epitope is from residue 20 to 25 (SDLWKL). While the crystal structure of the domain of p53 where p53Ab1 interacts have been resolved (Emamzadah et al., 2014), the structure of the region where p53Ab8 binds to have not been experimentally resolved yet. In case of p53Ab1, the top two peptides predicted to be the highest binders (Figure 3.3, residues 212 - 220, and 209 – 217) contained

the actual cognate residues. When the predicted top binders were mapped on to the crystal structure of p53 (peptides which are not covered by the available crystal structure were avoided for mapping), it was observed that the cognate region as well as the other peptides were a part of the β-barrel in that domain (Figure 3.8). The other predicted high binders which are not a part of the cognate region, lie very close to it (one adjacent strand away). Although the reported epitope of p53Ab1 is linear, usually epitope regions have been found out to be conformational in nature consisting of multiple linear regions. Therefore, it is a possibility that these sequences might also contribute to the interaction with p53Ab1. However, this remains a theoretical speculation given the lack of experimental evidence consisting of resolved structure of the antibody-protein complex.

For p53Ab8, when the model was projected on to p53, only two of the top 10 highest predicted binders contained parts of its cognate region (Figure 3.4). While the actual cognate region was at the very top, the only other sequence that represented the cognate partially was at the bottom of the list. Due to the unavailability of an experimentally resolved crystal structure of p53 that housed its cognate region, a predicted structure obtained from AlphaFold was used to map the top five peptides (Figure 3.9). Although the other sequences bore semblances to the cognate sequence, they were not located adjacent to the cognate binding region. Also, most of the regions where the peptides are mapped are largely unstructured, save for the cognate itself which is predicted to be a part of a small helix. Therefore, it is difficult to gauge the orientation of the predicted sequences with respect to the cognate, just from the predicted structural information. Whether or not, these other peptides contribute to binding to p53Ab8 can only be concluded if an experimentally resolved crystal structure can be made available in the future.

In case of 4C1, majority of the top 10 predicted highest binders (7 peptides) on its target protein (thyrotropin receptor) contained the entire known cognate region or parts of it (Figure 3.5). Among the top five sequences, the common residues observed were FDSH (residues 381 – 384) indicating that the those are essential for binding to 4C1. An experimentally resolved crystal structure was not available for thyrotropin receptor that covered the cognate region for 4C1. Therefore, AlphaFold was used here as well to predict the structure (Figure 3.10). The peptides mapped on to the predicted structure are overlapping each other, indicating that the cognate binding region might extend beyond the reported residues (LQAFDSH, 378 - 384). As this is a predicted structure, the orientation of the cognate region in space remains unclear, but as the non-cognate residues are flanking the cognate sequence, it is highly likely that they might be playing an assisting role in the binding interactions between the cognate region of the protein and the antibody.

In case of LNKB2, the highest predicted peptides from its target protein, interleukin-2 mostly comprise of the cognate residues or parts of them (Figure 3.6). According to the model, the predicted cognate binding region extends from residue 77 to 93. When the top five peptides were mapped on to the crystal structure of interleukin-2, it was found that four out of the five peptides were overlapping and constituted the same helical arm (Figure 3.11). As these overlapping peptides are adjacent to each other, it is highly likely that the non-cognate residues that constitute these peptides also interact with LNKB2, similar to observations made in case of DM1A, p53Ab1, and 4C1. This goes on to show that the neural network models for these antibodies are not only be able to correctly identify the cognate motifs, but they are also able to predict some of the non-cognate

residues that are also highly likely to take part in the binding activity between the monoclonal antibodies and their targets.

After demonstrating that the neural network was able to identify antibody-specific binding motifs on the target antigens, the performance of the models on the entire human proteome was tested. This was done to assess the performance of the neural network in recognizing the relevant cognate information present on the target protein when thousands of other protein sequences are also presented to it. The total number of proteins present in the human proteome dataset is equal to 20,361. When parsed into a library of 9-mer peptides, the total size of this peptide library was equal to 10,386,533 peptides. The neural network model for each antibody was projected on this library. All the proteins were ranked based on their highest binding peptide regions. The rank of the target protein was also determined from the predicted binding intensity of the cognate motif. It must be kept in mind that the residues C, I, M, and T were replaced with A, V, L, and S respectively as the former residues were not present in the arrays. Due to such substitution, the neural network treats the omitted residues the same as the residues they are replaced with. As some of the latter residues are also part of the cognate motifs for the mAbs, therefore peptides where a substitution had taken place have made it to the highest binders list. The 3rd peptide (EMEDDYDYY) in Table 3.2 can be used as an example. In this case, there is a methionine residue at the 2nd position of the peptide, that is replaced with leucine before the projection. Therefore, the model essentially treats the methionine as same as a leucine residue. The same holds true for the other omitted residues as well. Tables 3.2 to 3.6 depict the results from these projections. Except for p53Ab1, the target proteins for all of the other four antibodies were found within the top 2.2% from the list of available proteins. The target

protein of DM1A (α-tubulin) was in top 0.2% (Table 3.2), whereas the target proteins of 4C1 (thyrotropin receptor) and LNKB2 (interleukin-2) were found to be within top 0.09% and 0.52% respectively (Tables 3.5 and 3.6). For p53Ab8, the target protein cellular tumor antigen p53, was ranked within top 2.2% (Table 3.4). In case of p53Ab1, the target protein (p53) was ranked at 1918th position (Table 3.3), which was a much lower rank compared to that of the other cases. That is likely because the dipeptide VV (VI or IV in some cases) is considered more important by the model than the residues R, H, and S when defining the interactions of p53Ab1. As the dipeptide is a common motif that occurs across biological sequences, it made it harder for the model to discern the target sequences. The motif observed in the top 10 non-cognate sequences from the proteome for each antibody is shown in the Figure 3.12.

Despite the varying performances in case of every antibody, the neural network was successfully able to significantly narrow down the possibilities for finding the target motifs from a pool of biological sequence space. All of the known cognate interactions were predicted within top 10% of the list of proteins. This was a fascinating study because the algorithm which was trained on randomly sampled peptides that were only represented by 16 out of the 20 canonical amino acids, and devoid of any biological context, was able to distinguish clearly between cognate and non-cognate interactions from a set of biological sequences. Therefore, for the five monoclonal antibodies shown here, this study, in conjunction with the study elucidated in chapter 2, successfully demonstrates that by analyzing a sparsely sampled combinatorial sequence space, one can obtain very specific information about binding characteristics. In future, such machine learning algorithms could be used as tools to identify immunogenic regions on proteins as well as identify

residues that play an important role in the binding interaction between an antibody and a protein. Using such models alongside currently available tools could make the process of understanding the interactions of monoclonal antibodies much easier.

CHAPTER 4

USING NEURAL NETWORKS TO DERIVE SEQUENCE VS. BINDING RELATIONSHIPS FOR MONOCLONAL ANTIBODIES WITH UNKNOWN BINDING TO MICROARRAY PEPTIDES

4.1 INTRODUCTION

In the previous chapters (chapters 2 and 3), a neural network based approach was laid out to predict the sequence vs. binding relationships of five monoclonal antibodies (DM1A, p53Ab1, p53Ab8, 4C1, and LNKB2) whose binding behavior on the high-density random sequence peptide microarrays were well characterized (Legutki et al., 2014, Richer et al., 2015). Data from the binding experiments of these monoclonal antibodies on the random-sequence microarrays were used to train a feed-forward, backpropagated deep neural network to derive a comprehensive and quantitative sequence vs. binding relationship for the monoclonal antibodies. The performance of these derived relationships was evaluated on the basis of various criteria like cognate recognition, motif analysis etc. The results from the study presented in chapter 2 show that the trained neural network models were able to accurately identify the necessary motifs relevant to the binding interactions of the five monoclonal antibodies. Not only that, but the models were also able to predict highly specific cognate interactions by analyzing weaker binding interactions. Aside from identifying the relevant binding motif from the combinatorial sequence space (chapter 2), the trained models were also able correctly identify the cognate interactions from the biological sequence space as well (chapter 3).

As previously mentioned, these predictive models were trained on binding data of monoclonal antibodies whose binding characteristics on the random-sequence microarrays

is thoroughly understood. These antibodies are known to bind to the sequences present on the microarray including their cognates, though the cognate sequences are deliberately removed from training dataset during of the neural network to avoid potential bias. But how would the neural network perform if it were trained on data from binding interactions of monoclonal antibodies whose binding interactions on the microarrays were not as well described as the five antibodies previously mentioned. Also, the cognate sequences of these antibodies use some amino acids that are not present on the array and must be approximated by the most similar amino acids. How would the neural network fare with monoclonal antibodies whose binding patterns with the sequence space represented on the random-sequence microarrays are not characterized well? Also, how does the absence of the amino acid residues found in some of their molecular interactions affect the performance of the model?

In order to evaluate the performance of the neural network under such conditions, six monoclonal antibodies (3B5, 1D4, 9E10, AU1, Btag, and Htag) were chosen to be assayed on the microarrays. The binding data from the assays were then used to train the neural network to obtain a sequence vs. binding relationship for each of the antibodies. The hyperparameters of the neural network were the same as was used in the case of the previous five monoclonal antibodies. The derived relationship was then characterized using in silico projections on random combinatorial peptide libraries and motif analyses. In silico substitutions on the cognate sequences of these antibodies were also carried out to identify which residues from the cognate sequences were favored by the algorithm. Additionally, different propensity indices (physicochemical properties of the amino acids) were used as encoders for the amino acids to compare the performance of learned vs.

supplied encoders in case of these monoclonal antibodies. Lastly, the models of these antibodies were projected on to their respective target antigens and proteomes to evaluate their prediction ability in the biological sequence space.

4.2 METHODS

4.2.1 Array Synthesis and Binding Assays with Monoclonal Antibodies

Six monoclonal antibodies were used in this study (3B5, 1D4, 9E10, AU1, Btag, and Htag). All the binding assays with the monoclonal antibodies were carried out at HealthTell (http://www.healthtell.com/). The V13 high-throughput peptide microarrays were synthesized at HealthTell as laid out in section 2.2.1 in chapter 2 of this dissertation. Labeling and binding assays were carried out according to the protocol laid out in section 2.2.2 of the same chapter. The monoclonal antibodies that were used and their target proteins are listed in Table 4.1. This table also shows the correlation between the technical assay replicates. There were 3 technical replicates for each monoclonal antibody that were assayed.

4.2.2 Predicting Sequence vs. Binding Relationship

The sequence and binding data from the array experiments were used to train a feed-forward, backpropagated neural network, using the hyperparameters laid out in Table 2.3 in chapter 2. The encoding principle is the same as shown in Figure 2.1 (one-hot encoding). For each antibody, the neural network was trained 10 times independently, each time using a randomly selected training set from among the array peptides. The evaluation of the models was carried out using the methodologies laid out in section 2.2.4. The antibody-specific models were projected on to in silico libraries of 10^6 9-mer peptides that were randomly generated. These in silico libraries also included the target epitopes of the

respective monoclonal antibodies. There are four antibodies (1D4, 9E10, Au1, and Btag) used in this study whose interactions were not represented by the 16 amino acids used on the array. The model was unable to recognize those omitted residues. Therefore, in order to make accommodations for those sequences during projection of the models, the cognate sequences of those antibodies were replaced with other residues in silico. The substitutions were made with new residues that were among the array amino acids, and closely resembled the physicochemical properties of the original residues which were not present on the arrays. These substitutions enabled those sequences to be incorporated into the in silico libraries, for assessment of the performance of the model. The performance was evaluated by determining the rank of the epitope among randomly generated 10^6 peptides in an in silico library, which was sorted in descending order of Z-score values, after projecting the model on to it. Comparisons of sequence motifs were also performed, for both array peptides and predicted peptides, with the help of STREME motif analysis platform from the MEME suite (Bailey, 2021). Peptides with Z-score values greater than or equal to 5 were chosen for motif analysis, in case of both array and predictions. The most significant motifs with the lowest p-values, in both the cases, were chosen for comparison. Effects of using physicochemical propensities of amino acids as encoders for the model were also tested on 3B5 dataset using the same 17 propensities as mentioned in section 2.2.5.

Table 4.1. Information about the monoclonal antibodies used in this study. The omitted residues from the array are highlighted in red.

Monoclonal Antibodies	**Isotype**	**Target Protein (UniProt ID)**	**Epitope**	**Assay concentration (pM)**	**Corr≠ ± s.e.m**
3B5	IgG1	Human Receptor tyrosine-protein kinase erbB-2 (P04626)	EYLGLD	1000	0.8397 ± 0.0050
1D4	IgG1	Human rhodopsin (P08100)	TETSQVAPA	8000	0.7196 ± 0.0199
9E10	IgG1	Human myc-proto-oncogene protein (P01106)	QKLISEEDL	1000	0.8179 ± 0.0356
AU1	IgG2a	Bovine papillomavirus major capsid protein L1 (P03103)	DTYRYI	8000	0.7550 ± 0.1093
Btag	N/A*	Blue Tongue Virus Core Protein VP7 (Q5U8S6)	QYPALT	8000	0.8995 ± 0.0227
Htag	IgG1	N/A‡	HNHNHN	8000	0.7532 ± 0.0308

* N/A = not applicable; Btag is a mixture of two different isotypes. ‡ N/A = not applicable; the epitope of Htag is a synthetic peptide. ≠Corr = correlation between the technical replicates of the assays. The errors shown are standard errors of the means (s.e.m).

4.2.3 Projection on Antigen Proteins and Proteomes

The trained antibody-specific models for 3B5, 1D4, and 9E10. were projected in silico on to the crystal structures of their respective target proteins, using the same approach as mentioned in section 3.2.3 of chapter 3. The crystal structures of the target proteins were obtained from AlphaFold web version (Jumper et al., 2021; Varadi et al., 2021). The top five predicted peptides on the antigen were highlighted on the crystal structure of the antigen along the relevant cognate sequences. The trained algorithms were also projected

on the respective target proteomes (except for Htag which is raised against a synthetic peptide), once again following the approach mentioned in section 3.2.3.

4.3 RESULTS

4.3.1 Predicting the Sequence vs. Binding Relationship of the Antibodies

In this study, the molecular recognition of six monoclonal antibodies was probed using neural network algorithms. All the six monoclonal antibodies have very well characterized epitopes. For antibodies 3B5, 1D4, and 9E10, their respective target proteins are part of the human proteome (Vijver et al., 1988; Hodges et al., 1988; Schüchner et al., 2020). The target protein of AU1 is from bovine papillomavirus (Jenson et al., 1997) and the target protein of Btag is from bluetongue virus (Wang et al., 1996). Htag does not have a target protein in particular but is raised against a synthetic peptide instead (Hochuli et al., 1988). What separates these six monoclonal antibodies from the ones that were studied in chapter 2, is that their epitopes are not present among the sequences represented on the peptide arrays. Additionally, for four of these antibodies (1D4, 9E10, AU1, and Btag) the epitope sequences contain residues that were not represented on the microarrays (C, I, M, and T). Table 4.1 shows the target antigen and epitope sequence for each antibody used in this study.

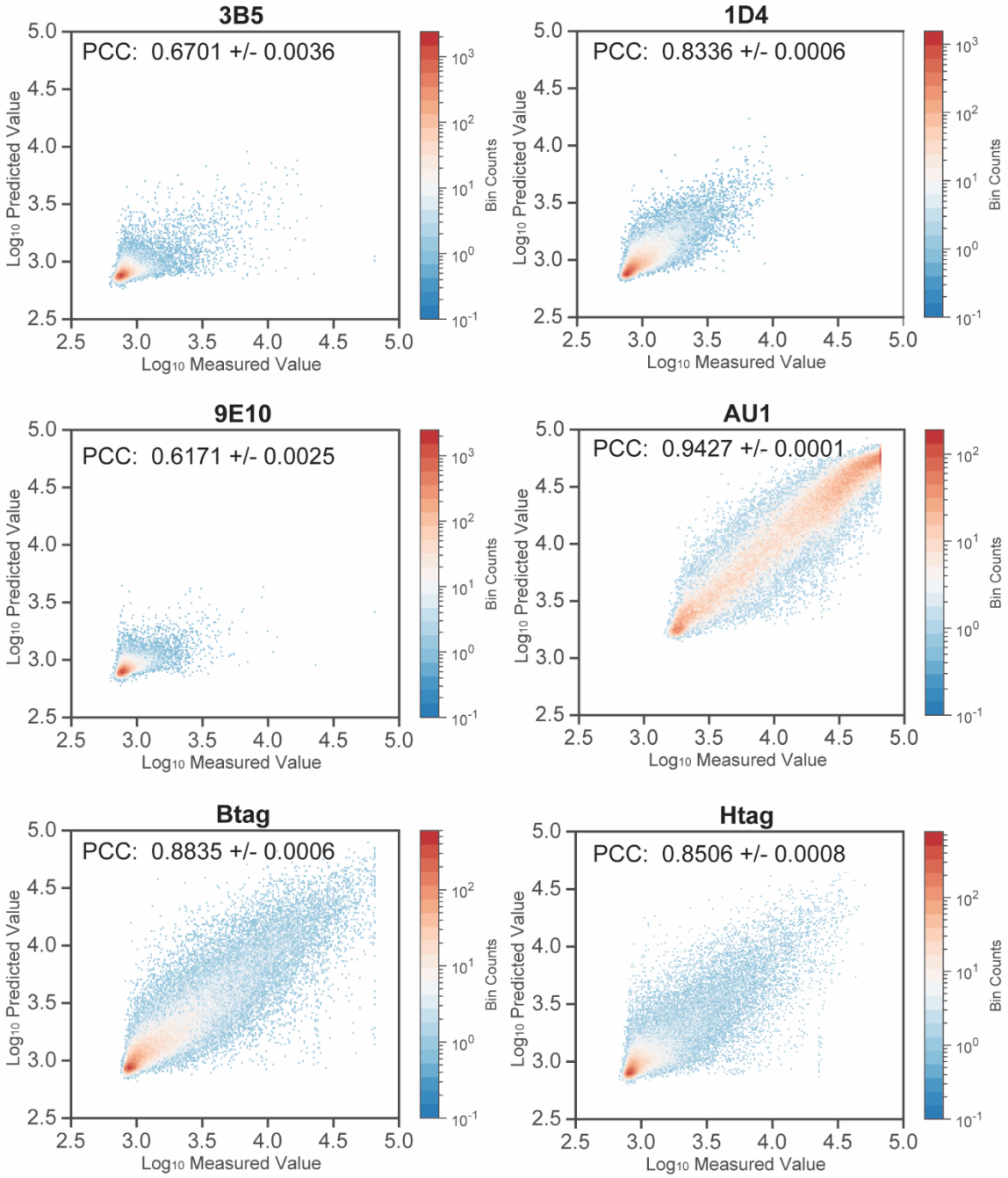


Figure 4.1. Scatter plots showing the correlation between $\log_{10}$ values of predicted binding measurements (y-axes) vs. actual binding measurements (x-axes) of peptides from the test datasets of monoclonal antibodies 3B5, 1D4, 9E10, AU1, Btag, and Htag. The density of datapoints (peptides) is color-coded as the number of peptides per datapoint. The Pearson correlation coefficient (PCC) between predicted and actual binding data is mentioned in each plot.

Similar to the approach taken in chapter 2, the antibodies were first assayed on peptide microarrays and then the processed data from the binding experiments were used to train the neural network. The hyperparameters used in this case were the same as shown in Table 2.3, Of the total number of peptides present on the array, 95% of the peptides were randomly chosen at a time for training the algorithm during each independent run. The rest of the 5% peptides were used as test set for cross-validating the performance of the model. 10 independent runs were carried out for each monoclonal antibody, each time with a random selection of peptides for training the model. It must be pointed out here, that unlike the study performed in chapter 2, no peptide sequences were deliberately excluded from the training set, as no known target of these monoclonal antibodies were present on the arrays. The results from each training were accumulated and the data from the test peptides of each run was used to compare the predicted and the measured binding values. Figure 4.1 show the scatter plots for each monoclonal antibody with comparison between predicted (x-axes) and measured (y-axes) data, as compiled from all the 10 independent trainings. The Pearson correlation coefficient (PCC) between the predicted and the measured data is also mentioned in each case.

The trained models were also projected on in silico libraries of peptides randomly sampled from combinatorial sequence space, to evaluate their performance. The methodology is similar to chapter 1 (section 2.2.4). However, in case of four out of the six antibodies (1D4, 9E10, AU1, and Btag), the epitope sequences contain amino acid residues that are not represented on the arrays (threonine and isoleucine). As the neural network algorithm does not identify the omitted residues, the ones present in the epitopes of the previously mentioned antibodies were substituted in silico with other appropriate residues

before being evaluated by the model. Table 4.2 show the original cognate sequences for antibodies 1D4, 9E10, AU1, and Btag, and the in silico modified sequences with substitutions. These substituted sequences will be used henceforth in this chapter to represent the respective cognate epitopes, beside the original sequences. Post-substitution, the epitope sequences were included in the randomly-generated peptide libraries, and the model was projected on to them. The results of the projections are shown in Table 4.3.

Table 4.2. Original and modified epitope sequences for 1D4, 9E10, AU1, and Btag

Monoclonal Antibody	**Original epitope sequences***	**Substituted epitope sequences≠**
1D4	TETSQVAPA	**S**E**S**SQVAPA
9E10	QKLISEEDL	QKL**V**SEEDL
AU1	DTYRYI	D**S**YRY**V**
Btag	QYPALT	QYPAL**S**

*The residues that are not present on the arrays but in the epitopes are represented in red.
≠The residues in red were substituted with residues represented in bold underlined letters.

Table 4.3. Rank of the epitopes within library of 10^6 random 9-mer peptides

Monoclonal Antibodies	**Epitope Sequences**	**Mean rank among 1 million peptides ≠**
3B5	EYLGLD	208 ± 2
1D4	SESSQVAPA*	801,584 ± 22
9E10	QKLVSEEDL*	679,996 ± 18
AU1	DSYRYV*	543,869 ± 1,818
Btag	QYPALS*	214,298 ± 2,835
Htag	HNHNHN	744,616 ± 3,335

* The substituted variants of the actual epitopes that were used for the projections. ≠ Rank values are averaged over 20 iterations of projections for each monoclonal antibody. Error shown here is the standard error of the mean of the ranks.

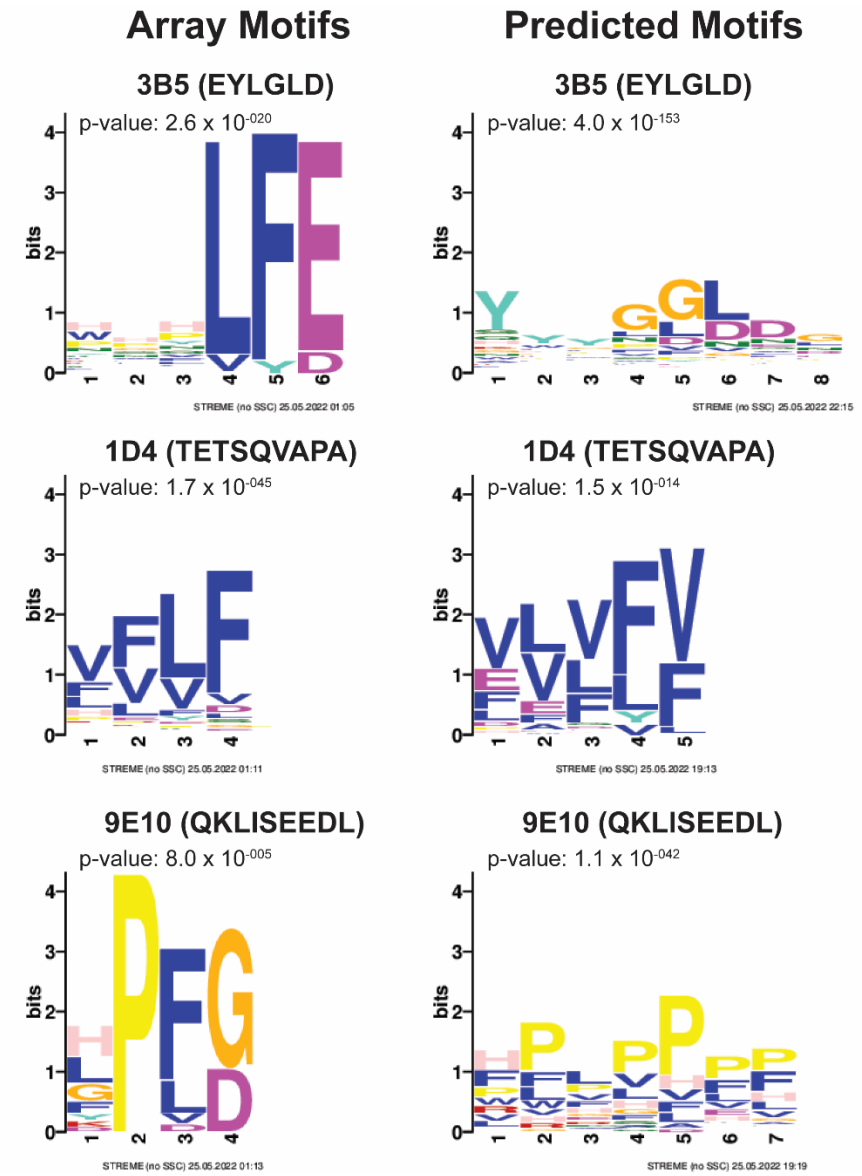


Figure 4.2. Sequence logos showing the motif analysis results for antibodies 3B5, 1D4, and 9E10. All analyses were carried out using STREME. Peptides from both arrays as well as in silico libraries (z-score >= 5) were selected. The p-value represents the statistical significance of the motif represented.

It can be seen from Table 4.3, that with the exception of 3B5, the model did not do a good job to identify the cognate sequences. Furthermore, the Z-scores of the peptides were calculated, both from the arrays as well as from the in silico libraries. Peptides with Z-score values higher than 5 were selected from both the datasets. The selected peptides

were then analyzed with the help of STREME motif analysis tool (Bailey, 2021) to find commonly occurring residues among the peptides with the highest Z-scores. Figures 4.2 and 4.3 demonstrate the results of the STREME analysis of the array and in silico peptides.

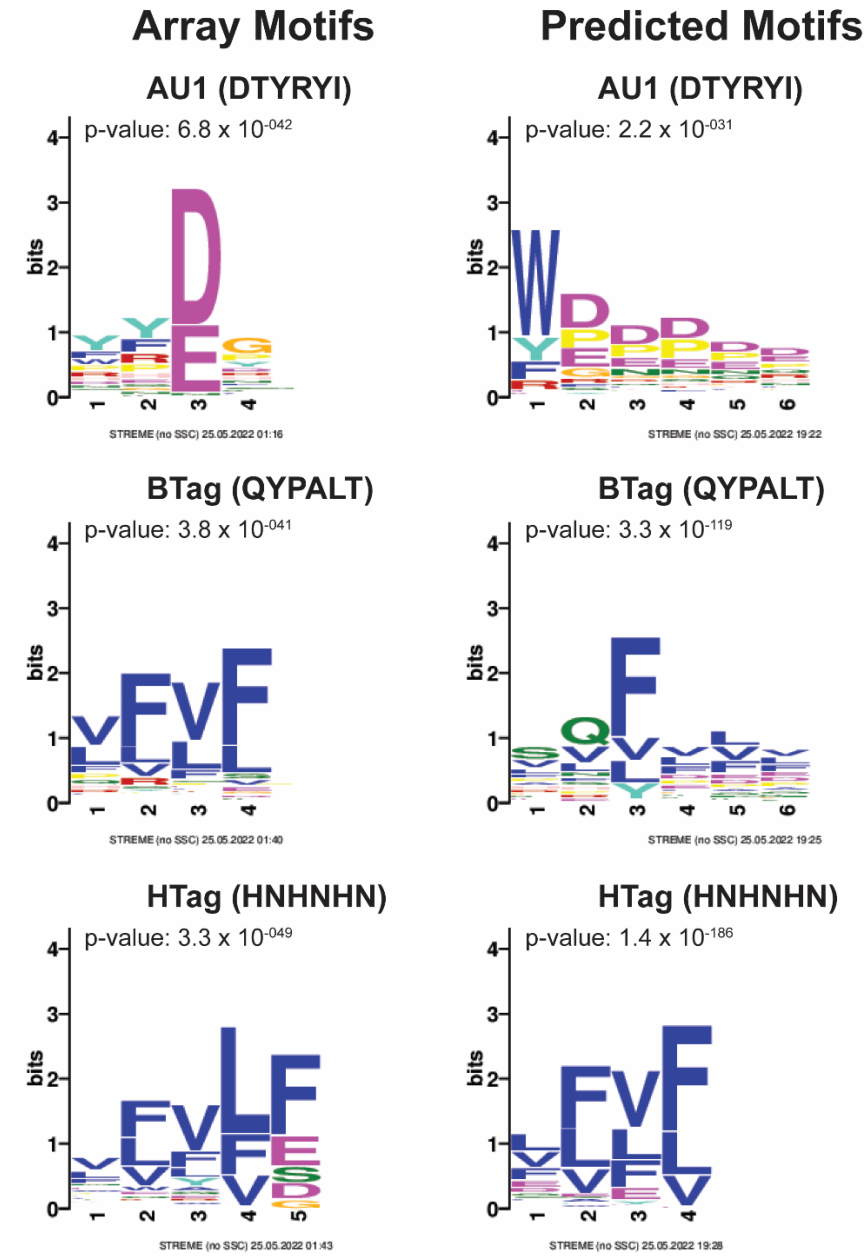


Figure 4.3. Sequence logos showing the motif analysis results for antibodies AU1, Btag, and Htag. All analyses were carried out using STREME. Peptides from both arrays as well as in silico libraries (z-score >= 5) were selected. The p-value represents the statistical significance of the motif represented.

It can be seen from Figures 4.2 and 4.3, that none of the peptide motifs observed on the arrays are representative of the actual epitopes of the monoclonal antibodies. This indicates that the monoclonal antibodies are reacting with peptide sequences on the microarray that contain residues that are inherently different from the residues present in the cognate sequences. In case of 1D4 (Figure 4.2), Btag, and Htag (Figure 4.3), the sequence motifs from the arrays show a lot of hydrophobic residues like valine , leucine, and phenylalanine. Correspondingly, the predictions also show an abundance of these residues in the highest-predicted binders. For 3B5 (Figure 4.2), the most statistically significant (p-value = 2.6×10^{-20}) motif from the array primarily contains lysine, phenylalanine, and glutamic acid residues. The motif from the top predicted peptides, however, more closely resembles the epitope of 3B5, with tyrosine, glycine, leucine, and aspartic acid being the more prominent residues. It should be noted that the phenylalanine and glutamic acid observed in the 3B5 array motifs bear physicochemical similarities with tyrosine and aspartic acid residues respectively. For 9E10 (Figure 4.2), the most abundant residue in the both the array and the predicted motif is shown to be proline, which is not a part of the actual cognate sequence. Other residues that are evident from the motif, but not part of the actual cognate sequence are phenylalanine, histidine, and glycine. Overall, the motif observed from the array and the in silico sequences bear no semblance to the epitope sequence. In case of AU1 (Figure 4.3), aspartic acid, which also present in the cognate sequence, is one of the most prominent residues in the array motif as well as the predicted motifs. Aside from aspartic acid, tyrosine is also present in the motif, but not as major residue. Other residues that are not part of the AU1 cognate sequence include glutamic acid (array motif), tryptophan, and proline (predicted motifs). It can be seen from the motif

analysis that the neural network has not been successful in identifying the correct residues in most cases, except for 3B5 and to some extent AU1.

4.3.2 In silico Substitutions of Cognate Sequence Residues

Following the evaluation of the model, an experiment was carried out where the cognate sequence of each antibody was substituted in silico with other amino acid residues from the array, generating new peptides. The antibody-specific models were then projected on to these libraries of substituted peptides, to obtain predicted binding values for each peptide. The predicted binding intensities of the parent peptide (cognate sequence) was then subtracted from that of the substituted peptides. The resultant values were represented as heatmaps as shown in Figures 4.4 and 4.5.

In Figure 4.4, the results of amino acid substation in the epitopes of 3B5, 1D4, and 9E10 are shown. In case of 3B5, it can be seen that for most of the residues of the cognate sequence (EYLGLD), substitutions are not favored well. Y(2), G(4), L(5), and D(6) are the residues most resistant towards substitution, not favoring any other amino acid in their respective position. L(3) is the least resistant to substitution, according to this in silico study. This shows that the model strongly recognizes and favors the cognate residues over the other amino acids. In case of 1D4, the in silico study shows that the model does not strongly favor the cognate sequence residues over the substitutions. It can be seen that the residue phenylalanine is favored strongly at all positions despite it not being present in the epitope sequence. Other favored residues include valine, tryptophan, and tyrosine. In 9E10, residues L(3), V(4), S(5), E(6), E(7), and L(9) are fairly conserved. Interestingly, the serine in the 5^{th} position highly favors a substitution with a proline residue. Residues Q(1), K(2), and D(3) are shown to favor substitution with other amino acid residues.

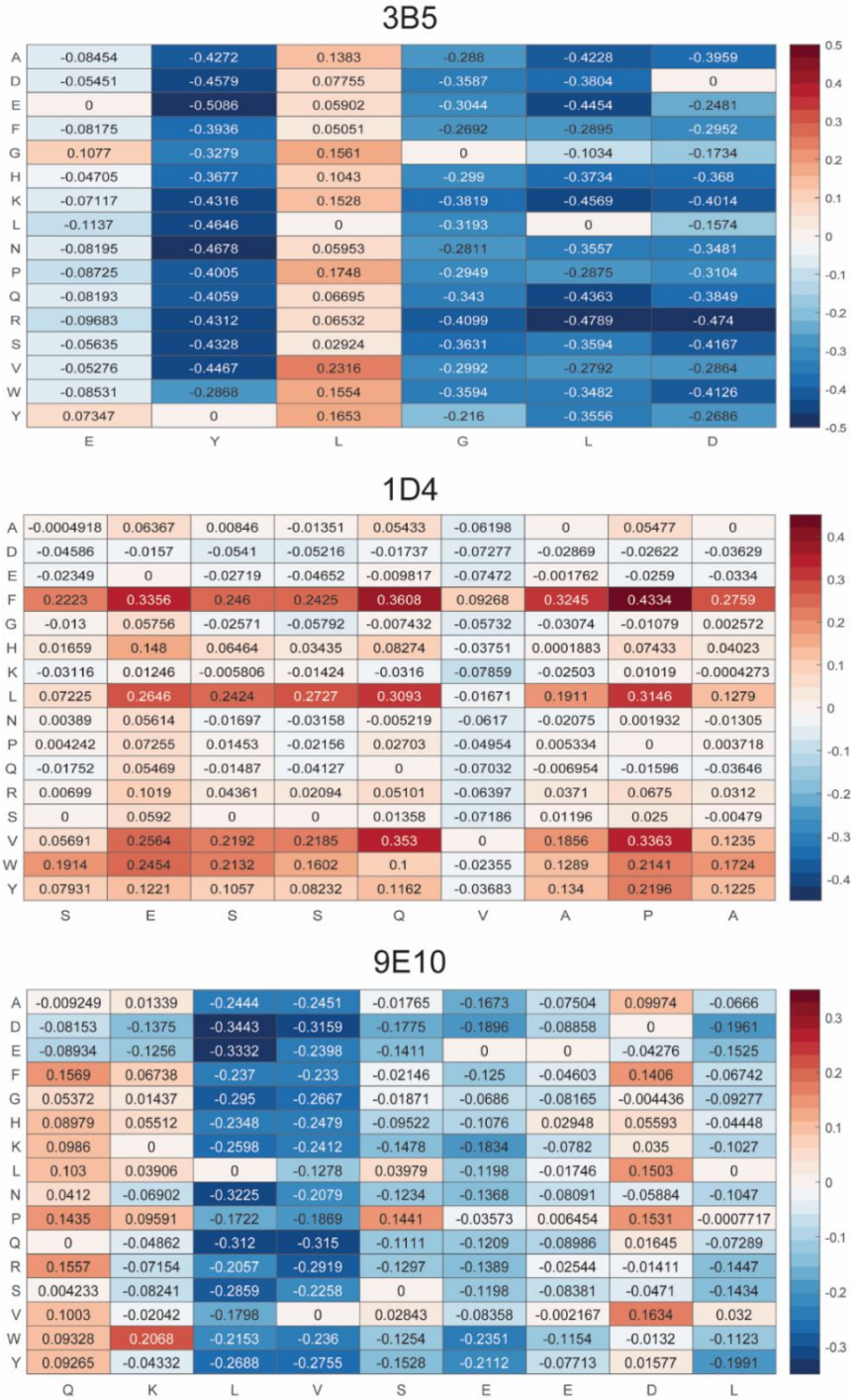


Figure 4.4. Heatmaps representing the results of in silico single-point substitution of the epitopes for 3B5, 1D4, and 9E10. Each residue in an epitope sequence (x-axes), was replaced with the 16 amino acid residues used on the microarray (y-axes), one at a time to create an in silico library. The colorbar represents the deviation of the peptides from the predicted binding of the actual epitope and is scaled the same for all the monoclonal antibodies.

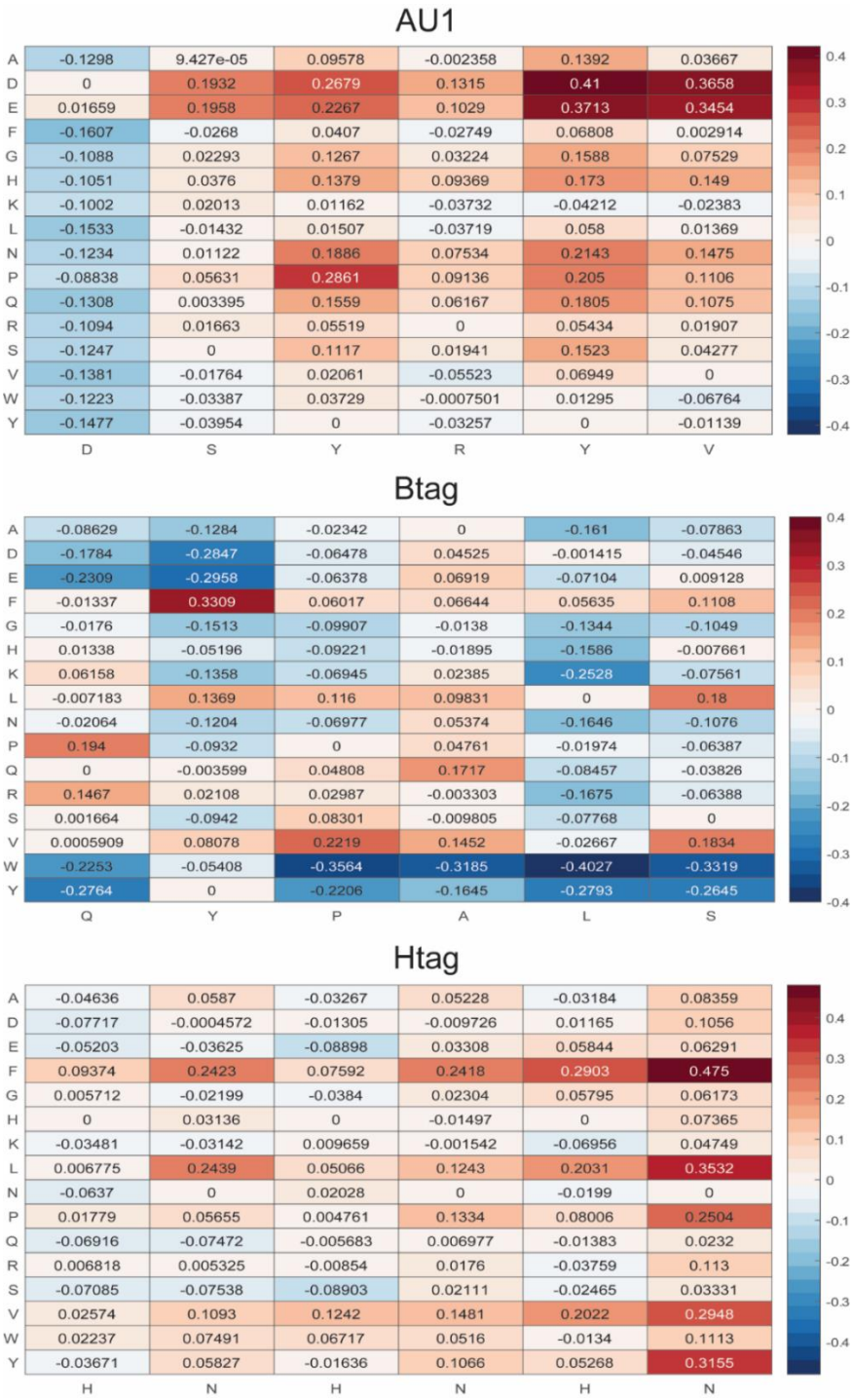


Figure 4.5. Heatmaps representing the results of in silico single-point substitution of the epitopes for AU1, Btag, and Htag. Each residue in an epitope sequence (x-axes), was replaced with the 16 amino acid residues used on the microarray (y-axes), one at a time to create an in silico library. The colorbar represents the deviation of the peptides from the predicted binding of the actual epitope and is scaled the same for all the monoclonal antibodies.

In Figure 4.5, it can be seen that only the aspartic acid in the 1st position of the cognate sequence of AU1 (DSYRYV) does not favor substitution. The rest of the residues show favorable predicted binding when substituted. High preference is given to aspartic acid and glutamic acid residues at all positions. In case of Btag, some residues are strongly favored while other are strongly disfavored at different positions. For example. tryptophan and tyrosine residues are strongly disfavored at all positions except the tyrosine at the 2nd position. No residue appears to be extensively conserved across all substitutions except for the leucine at the 5th position. For Htag, none of the cognate residues (histidine and asparagine) are favored, according to the model. Rather, substitutions with hydrophobic residues like phenylalanine, valine, and proline are preferred. It can be concluded from here that the cognate residues were not well-recognized by the model in this case.

4.3.3 Using Amino Acid Propensities as Encoders for the Model

So far, the model had been using a learned encoder for encoding the amino acids during training the neural network. It would also be interesting to see how the neural network models performs when physicochemical propensities of amino acids are used as encoders in this case. To study the effects of using physicochemical propensities as encoders, 3B5 was chosen as a model system among the antibodies studied here, because of the better performance of its model compared to the others. The 17 propensity scales that were used as encoders are listed in section 2.3.4 of chapter 2. The results of using different propensities as encoders are listed in Table 4.4, as shown below. The results are shown in comparison to using a learned encoder. Also, an encoder with random numbers between 0 and 1 was used as a negative control.

Table 4.4. Performance of the model with respect to different propensities used (mean correlation coefficient and mean ranking of epitope)

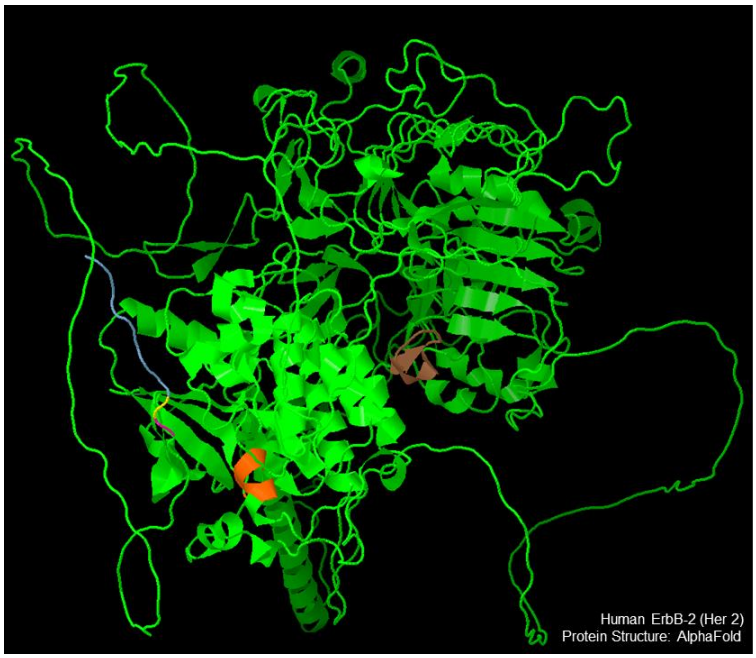
Propensities Used	**Number of Amino Acid Descriptors**	**Mean Correlation Coefficient ***	**Mean Rank of Epitope[#]**
None (Learned encoder)	9	0.6701 ± 0.0036	208 ± 2
Random, normally distributed number between [0,1]	9	0.6313 ± 0.0140	876,184 ± 56
Hydropathy Indices, Molecular Weight, Isoelectric Point	7	0.6507 ± 0.0076	821,712 ± 28,794
Representative properties from each different propensity scale (PP, KF, Z, F, T, VHSE, ProtFP, ST, BLOSUM, MSWHIM)	10	0.7017 ± 0.0110	413,717 ± 167
VHSE	8	0.7071 ± 0.0101	878,120 ± 27,876
ST Scales	8	0.6881 ± 0.0174	770,043 ± 20,154
ProtFP	8	0.7217 ± 0.0145	563,368 ± 166
Kidera Factors	10	0.7146 ± 0.0067	731,013 ± 26,045
BLOSUM	10	0.7545 ± 0.0076	226,585 ± 12,922

* Mean correlation coefficients have been calculated over 5 independent trainings of the model for each antibody. [#]Mean ranks are out of 1 million and have been calculated from projecting the respective models on to 10 individual libraries of random 9-mer peptide sequences. None of the libraries had any common peptides except for the epitope sequence of the respective antibody. All the errors represent the standard errors of the means. All the experiments were done using the binding data available for 3B5.

It can be seen from Table 4.4 that although the correlation coefficients are higher in most cases compared to the learned encoder, the model fails to identify the cognate sequence correctly. The learned encoder based model performs extensively better than the other encoders in terms of predicting the cognate epitope. Among the propensity encoders, the BLOSUM descriptors perform better than the rest, in terms of correlation between predicted and measures values as well as in identifying the epitope. Interestingly, compared

to the correlation observed in case of BLOSUM descriptors, the learned encoder had a lower correlation value but performed better in terms of predicting the cognate residues.

4.3.4 Projection on the Target Antigens and Proteomes

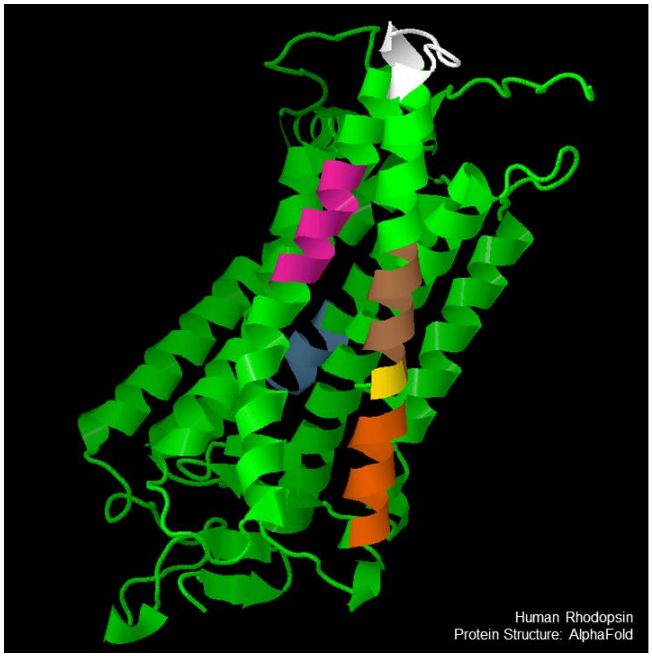


Peptide	Position
NPEYLGLDV	1245 – 1253
PEYLGLDVP	1246 – 1254
AYGLGLEHL	342 – 350
EYLGLDVPV	1247 – 1255
PYGALLDHV	802 – 810

Figure 4.6. Top five peptides predicted as the strongest binders by the neural network model for 3B5 represented on the predicted crystal structure of its target protein, human receptor tyrosine-protein kinase erbB-2 (https://alphafold.ebi.ac.uk/). The peptides are part of the actual protein sequence and ranked in descending order of predicted binding value. The position of each peptide on the crystal structure have been color-coded as shown in the table included in the figure. Regions of cognate residues in the peptides are highlighted through underlining.

Next, the trained models for 3B5, 1D4, and 9E10 were projected on to their target proteins from the human proteome. Predicted structures were obtained from the web version of AlphaFold (www.alphafold.ebi.ac.uk) to visualize the highest predicted binders on the proteins. The target proteins of AU1 and Btag were from viruses and their predicted structures were not available from AlphaFold. The ones available from RCSB-PDB (www.rcsb.org) did not represent the whole proteins and were therefore not used here. Htag

is raised against a synthetic peptide and therefore any target protein was not available in its case.



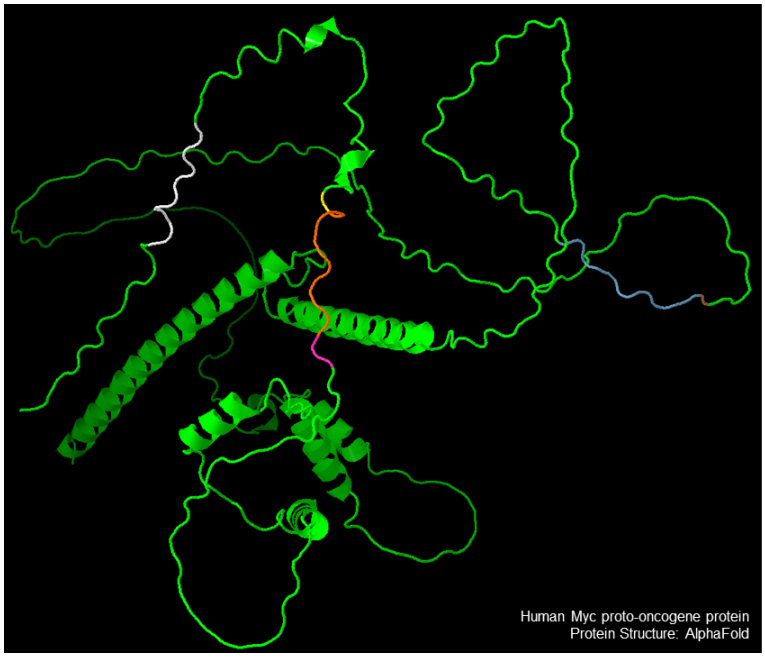
Peptide	Position
RLVVVLVV	251 – 259
SVPLVVVFF	213 – 221
VPLVVVFFA	214 – 222
DLFLVLGGF	83 – 91
ESFVVYLFV	201 – 209

TETSQVAPA	300 – 308

Figure 4.7. Top five peptides predicted as the strongest binders by the neural network model for 1D4 represented on the predicted crystal structure of its target protein, human rhodopsin (https://alphafold.ebi.ac.uk/). The peptides are part of the actual protein sequence and ranked in descending order of predicted binding value. The position of each peptide on the crystal structure have been color-coded as shown in the table included in the figure. The epitope has been shown in white.

In case of 3B5, the cognate region (residues 1245 – 1255) on its target protein, human receptor tyrosine-protein kinase erbB-2 (Her-2), was correctly identified as shown in Figure 4.6. The other peptides that were predicted among the top five highest binders were not found to be structurally close to the cognate region. For 1D4 (Figure 4.7) and 9E10 (Figure 4.8), there were no parts of the cognate region among the top five predicted peptides. In case of 1D4 (Figure 4.7), the epitope along with the top five predicted binders are highlighted as shown. The cognate sequence in the target protein, rhodopsin, did not show up among the top 20 binders. All the five highest predicted binders from rhodopsin

have large number of hydrophobic residues like valine and phenylalanine and are likely not involved in interacting with the monoclonal antibody.



Peptide	Position
LLPSPPLSP	55 – 63
KFELLPSPP	52 – 60
KPPHSPLVL	289 – 297
PPHSPLVLK	290 – 298
FELLPSPPL	53 – 61

QKLISEEDL	411 – 419

Figure 4.8. Top five peptides predicted as the strongest binders by the neural network model for 9E10 represented on the predicted crystal structure of its target protein, human myc proto-oncogene protein (https://alphafold.ebi.ac.uk/). The peptides are part of the actual protein sequence and ranked in descending order of predicted binding value. The position of each peptide on the crystal structure have been color-coded as shown in the table included in the figure. The epitope has been shown in white.

In case of 9E10 (Figure 4.8), the top five predicted binders are on non-rigid regions in the structure of myc proto-oncogene protein (c-Myc), as predicted by AlphaFold. The cognate region also lies on one of the non-rigid arms. The five highest predicted binders are not close to the cognate region (410 – 419). The peptide KLISEEDLL, which contains part of the cognate region was predicted at the 11th position, but most of the highest binders predicted on the target protein were part of regions with lot of proline residues, as can be seen from Figure 4.8.

Table 4.5. Top predicted peptides from the human proteome (UP000005640) for 3B5 and their corresponding proteins. Epitope is highlighted in red.

Rank	Protein ID	Top Predicted Peptide
1	sp|P09958|FURIN_HUMAN Furin	SYGYGLLDA
2	sp|P32243|OTX2_HUMAN Homeobox protein OTX2	SYFGGMDCG
3	sp|Q8IVF4|DYH10_HUMAN Dynein axonemal heavy chain 10	GYEYMGLNG
4	sp|Q5T5U3|RHG21_HUMAN Rho GTPase-activating protein 21	SYDEGLDDY
5	sp|Q00444|HXC5_HUMAN Homeobox protein Hox-C5	RYCYGGLDL
6	sp|P40818|UBP8_HUMAN Ubiquitin carboxyl-terminal hydrolase 8	SNHYGGLDG
7	sp|Q8WYR1|PI3R5_HUMAN Phosphoinositide 3-kinase regulatory subunit 5	SHYLGMLDP
8	sp|P20719|HXA5_HUMAN Homeobox protein Hox-A5	GYGYNGMDL
9	sp|Q09013|DMPK_HUMAN Myotonin-protein kinase	EYYVGGDLL
10	sp|O75427|LRCH4_HUMAN Leucine-rich repeat and calponin homology domain-containing protein 4	RYDGGLDSG
…	…	…
294	**sp|P04626|ERBB2_HUMAN Receptor tyrosine-protein kinase erbB-2**	NPEYLGLDV

Next, the models were projected on the target proteomes of the antibodies. For the human proteome (UP000005640) a total of 20,361 proteins were present, which were parsed into 10,386,533 9-mer peptides. 3B5, 1D4, and 9E10- specific models were then projected on these peptides. After projection, the peptides were ranked in descending order of binding values. The results are shown in Tables 4.5, 4.6, and 4.7. The target protein of 3B5, Her-2 was ranked at the 294th position (Table 4.5) out of the ~10 million peptides (top 0.0028%) indicating that the model was able to correctly identify the cognate residues in this case. The residues G, L, and D are frequently seen among the top 10 predicted

sequences from the proteome, with GLD being a common motif among the peptides. For 1D4, the cognate sequence from target protein rhodopsin was ranked over 8 million (Table 4.6), The model was clearly not able to identify the cognate region or the target protein in this case. Also, the highest predicted peptides contain a lot of hydrophobic residues and bear no resemblance to the cognate motif.

Table 4.6. Top predicted peptides from the human proteome (UP000005640) for 1D4 and their corresponding proteins. Epitope is highlighted in red.

Rank	**Protein ID**	**Top Predicted Peptide**
1	sp|Q8NCS7|CTL5_HUMAN Choline transporter-like protein 5	EVIVILMLI
2	sp|Q9UI40|NCKX2_HUMAN Sodium/potassium/calcium exchanger 2	DLIMLIIFF
3	sp|Q8IZ96|CKLF1_HUMAN CKLF-like MARVEL transmembrane domain-containing protein 1	EICIVVFFI
4	sp|O15360|FANCA_HUMAN Fanconi anemia group A protein	EELLVFLFF
5	sp|O60721|NCKX1_HUMAN Sodium/potassium/calcium exchanger 1	DLIMLILFF
6	sp|A6NJZ3|O6C65_HUMAN Olfactory receptor 6C65	ELQVVIFFF
7	sp|O60741|HCN1_HUMAN Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 1	DLIMLIMMV
8	sp|P54840|GYS2_HUMAN Glycogen [starch] synthase, liver	DITVMVFFI
9	sp|P29275|AA2BR_HUMAN Adenosine receptor A2b	PPLLIMLVI
10	sp|Q9UBR5|CKLF_HUMAN Chemokine-like factor	EVTVILFFI
…	…	…
8,783,703	**sp|P08100|OPSD_HUMAN Rhodopsin**	TETSQVAPA

In case of 9E10 (Table 4.7), the cognate sequence was ranked as the 426,947th peptide out of 10 million which would be around top 4.11%. Among the top predicted non-

cognate peptides, the motif LXPE appears frequently, where X is either I or V. It has been mentioned previously in this chapter that the serine residue present in the cognate prefers a substitution with proline and that can be seen here as well.

Table 4.7. Top predicted peptides from the human proteome (UP000005640) for 9E10 and their corresponding proteins. Epitope is highlighted in red.

Rank	**Protein ID**	**Top Predicted Peptide**
1	sp|Q9NRS6|SNX15_HUMAN Sorting nexin-15	LHILPPPLI
2	sp|Q9P2J8|ZN624_HUMAN Zinc finger protein 624	ILIPEPGIA
3	sp|Q9Y5F0|PCDBD_HUMAN Protocadherin beta-13	CLVPEGPLP
4	sp|P52569|CTR2_HUMAN Cationic amino acid transporter	PFLPFLPAF
5	sp|O75445|USH2A_HUMAN Usherin	ILIPEIPVE
6	sp|Q6RI45|BRWD3_HUMAN Bromodomain and WD repeat-containing protein 3	HLMPPPFLV
7	sp|Q8TCX1|DC2L1_HUMAN Cytoplasmic dynein 2 light intermediate chain 1	PFPVPLVII
8	sp|Q9UK22|FBX2_HUMAN F-box only protein 2	GLVPEGGVE
9	sp|P10912|GHR_HUMAN Growth hormone receptor	MLILPPVPV
10	sp|O94911|ABCA8_HUMAN ABC-type organic anion transporter ABCA8	PFLVFLIPF
…	…	…
426,947	**sp|P01106|MYC_HUMAN Myc proto-oncogene protein**	QKLISEEDL

The bovine papillomavirus has 9 proteins in its proteome which were parsed into 2395 individual 9-mer peptides. The target protein for AU1 is the major capsid protein L1. The cognate region on the target protein was predicted at the 97th position (Table 4.8) out of the total number of peptides (top 4.05%). The most common motif observed among the top 10 non-cognate sequences in this case is DRP. These residues are part of the cognate sequence also, though they are spaced apart from each other with other residues.

Table 4.8. Top predicted peptides from the bovine papillomavirus proteome (UP000006567) for AU1 and their corresponding proteins. Epitope is highlighted in red.

Rank	**Protein ID**	**Top Predicted Peptide**
1	sp|P03122|VE2_BPV1 Regulatory protein E2	SDFRDRPDG
2	sp|P03122|VE2_BPV1 Regulatory protein E2	DFRDRPDGV
3	sp|P03109|VL2_BPV1 Minor capsid protein L2	LDDFSETHR
4	sp|P03122|VE2_BPV1 Regulatory protein E2	SSDFRDRPD
5	sp|P03122|VE2_BPV1 Regulatory protein E2	SRFGDEAAR
6	sp|P0DOD6|VE8E2_BPV1 Protein E8^E2C	RPSRDRPDG
7	sp|P03122|VE2_BPV1 Regulatory protein E2	YSRFGDEAA
8	sp|P0DOD6|VE8E2_BPV1 Protein E8^E2C	LRPSRDRPD
9	sp|P03109|VL2_BPV1 Minor capsid protein	ELQPLDRPT
10	sp|P03109|VL2_BPV1 Minor capsid protein L2	DDFSETHRL
…	…	…
97	**sp|P03103|VL1_BPV1 Major capsid protein L1**	DTYRYIESP

In case of the antibody Btag, the target protein, core protein VP7, is part of the bluetongue virus proteome. This proteome also has 9 proteins which were parsed into 5857 peptides. The cognate ranked at the 118th position which was among the top 2% of the peptides from the proteome. The more common residues in the non-cognate sequences are mostly hydrophobic like valine or leucine, glutamine which is a part of the cognate residues was also found on some of the non-cognate sequences. Overall, it can be seen that the cognate sequences which had residues that were not present on the array were not predicted very well by the model.

Table 4.9. Top predicted peptides from the bluetongue virus proteome (UP000112999) for Btag and their corresponding proteins. Epitope is highlighted in red.

Rank	Protein ID	Top Predicted Peptide
1	tr|C5IWV8|C5IWV8_9REOV Outer capsid protein VP2	SQVIVLVFD
2	tr|C5IWV8|C5IWV8_9REOV Outer capsid protein VP2	ISQVIVLVF
3	tr|C5IWV7|C5IWV7_9REOV RNA-directed RNA polymerase	PQLIVTLPL
4	tr|C5IWV8|C5IWV8_9REOV Outer capsid protein VP2	QVIVLVFDL
5	tr|C5IWW4|C5IWW4_9REOV Non-structural protein NS2	DMSLIILPV
6	tr|C5IWV8|C5IWV8_9REOV Outer capsid protein VP2	VIVLVFDLI
7	tr|C5IWV7|C5IWV7_9REOV RNA-directed RNA polymerase	QLIVTLPLN
8	tr|C5IWV8|C5IWV8_9REOV Outer capsid protein VP2	VLVFDLIFE
9	tr|C5IWV7|C5IWV7_9REOV RNA-directed RNA polymerase	DLVTVFTLM
10	tr|C5IWV8|C5IWV8_9REOV Outer capsid protein VP2	DISQVIVLV
…	…	…
118	**tr|Q71TX7|Q71TX7_9REOV Core protein VP7**	TLNQYPALT

4.4 DISCUSSION

In this study, the sequence vs. binding relationship of six antibodies were probed whose binding interactions with the peptides on the microarray were not well-known. This was an extension of the approach showed in chapter 2 to investigate the predicted binding of antibodies whose interactions with the sequences present on the array were not well-characterized. The neural network architecture that was used is the same as the one described in chapters 2 and 3. The algorithm was trained on data from the binding experiments carried out on the microarrays. Of the six monoclonal antibodies that were

studied, four of the antibodies have cognate sequences that contain at least one of the amino acid residues that were omitted from the array due to manufacturing constraints (Table 4.2). It can be seen from Figure 4.1, that some of these antibodies, like 1D4 and 9E10, did not bind to the array peptides very well, whereas antibodies like AU1 and Btag bound strongly to a lot of sequences.

When the trained neural networks were projected on to the in silico libraries (Table 4.3), the models were not able to identify the respective cognate sequences among the top 1% binders in five out of the six cases. This was in stark contrast to the performance of the models observed in chapter 2 (Tables 2.5 and 2.6), where antibodies whose interactions were well-characterized on the arrays were used. The model, however, successfully predicted the cognate interaction of the antibody 3B5 among the top 0.02%. It should be kept in mind that the cognate residues of 3B5 do not contain any of the amino acids omitted from the microarrays. Further analyses of the motifs (Figures 4.2 and 4.3) from the highest binders, both in case of the array peptides as well as the in silico libraries, show that except for 3B5, the cognate residues of the other antibodies were not very well recognized by the models. In most of the cases (1D4, 9E10, Btag, and Htag), a lot of hydrophobic residues like leucine, valine, and proline were identified as most commonly occurring residues among the highest peptide binders from the array, as well as the predicted peptides. For the antibodies whose epitopes contain one of the omitted residues, one can speculate that in absence of the necessary residues, these antibodies bound to a lot of hydrophobic sequences present on the array, that were not representative of the cognate interactions or residues. Previous studies (James et al., 2003, James and Tawfik, 2003, Sykes et al., 2013) with other monoclonals have suggested that this kind of observation is not necessarily a phenomenon

of cross-reactivity related to ‘hydrophobic stickiness’, but that of ‘multispecificity’ which involves identification of different residues than that of the cognate sequence. Though such interactions have not been reported for the antibodies presented in this study, observation of such interactions on the microarray can be considered as a proof of such multispecificity. This in turn led the algorithm to prioritize those interactions, therefore leading to a prediction of these residues. These binding interactions were not necessarily strong as seen in the case of 1D4 and 9E10. In case of AU1, the reported target residues are considered crucial for interacting are T(2), Y(3), R(4), and Y(5) according to a study done by Jenson et al. (1997). The adjacent aspartic acid residue, D(1), which flanks the tetramer was not considered as necessary. However, from the interactions observed on the array, it can be seen that the antibody reacted with sequences that had the aspartic acid residue in them, alongside the tyrosine residues, which was also ultimately reflected in the predictions (Figure 4.3). For Htag, although the target sequence does not contain any omitted residues, the antibody was seen to prefer hydrophobic interactions more on these arrays, possibly due to the same phenomenon as explained above. But due to the absence of conclusive literature exploring such interactions, this can be only restricted to speculations. Hence, resulting predictions were also biased towards those residues.

In silico substitutions were carried out on the cognate sequences to further analyze what residues are favored towards substitution as deemed by the model (Figures 4.4 and 4.5). From Figure 4.4, it can be seen that the model was able to identify the residues that are considered crucial for the interactions of the antibody 3B5, and it also shows that no substitutions are tolerated at most positions. In case of 1D4, no such preference is observed towards any residue from the cognate sequence. The cognate sequence of 1D4

(TETSQVAPA) has two threonine residues that were not represented on the arrays. In absence of these residues, the model was provided with a substituted peptide as an alternative and as can be seen from the figure, it identifies the substituted cognate sequence very poorly. Similar observations were made in case of AU1 and Btag (Figure 4.5). In the case of 9E10 though, some preference was given to the residues from the substituted cognate sequence as can be seen from Figure 4.4. The motif LVSE appears to be particularly favored over other residues, with the exception of proline substitution in place of serine. Therefore, some of the cognate interactions were identified in this case, though the flanking residues were again not identified well by the model. In case of Htag, the substitution results again show no preference towards the cognate residues (Figure 4.5). The residues at all positions favor substitutions with other residues, which indicates that the model was not able to identify the key residues in this case as well. This is because Htag interacted with mostly hydrophobic residues on the array, which made the model biased towards those residues during training. In future, a detailed experimental exploration of the molecular interactions of these antibodies would hopefully shed some light on these observations.

Different physicochemical parameters were also used to encode the amino acids, to determine the changes in the model's performance compared to the learned encoders. The 3B5 antibody was used as a model system for this study owing to its relatively better performance compared to the others. The results are as shown in Table 4.4. It can be seen that the predictive performance of the model did not get better when different types of propensity indices were used as encoders, although an increase in correlation values were observed in some cases. It has been observed previously in this chapter that higher

correlations do not necessary relate better predictions, which actually depends on the eigenvalues assigned to the amino acids to identify the relevant residues. A learned encoder ensures that appropriate orthogonality is maintained between the amino acid residues, whereas a preassigned encoder is more likely to introduce a bias in the algorithm. This might be the reason why a learned encoder performs better than the supplied propensity indices in this case.

Now all the previous studies were carried out on in silico libraries of peptides that were sampled from the combinatorial sequence space. To test the performance of the models on biological sequence space, the models of 3B5, 1D4, and 9E10 were projected on their respective target proteins and the results were visualized in Figure 4.6, 4.7, and 4.8 respectively. Only in case of 3B5, the cognate region appeared among the top five predicted binders from its target protein, Her-2 (Figure 4.6). In case of antibodies 1D4 (Figure 4.7) and 9E10 (Figure 4.8), the cognate region was not predicted among the top five binders from their respective antigens. None of the top five predicted interactions are close to the cognate regions on the target proteins, in both the cases, as can be seen from these figures. Rather they represent hydrophobic regions on the surface of the target antigens. So, the model failed to identify the cognate residues in both these cases, even among a small number of peptides.

Additionally, each of the antibody-specific models were projected on their target proteomes. Among the antibodies whose target proteins lie within the human proteome, the model performed quite well in predicting the cognate sequence for 3B5, predicting the epitope among top 0.002% of the total number of peptides (Table 4.5). GLD was found to be a common motif among the top binders in this case. For 9E10 (Table 4.7), the cognate

sequence was predicted among the top 4.11% with the LXPE motif being dominant among the top predicted binders. In both these cases, the models were able to recognize the relevant binding interactions to varying extent. However, the algorithm absolutely did not perform well in case of 1D4 (Table 4.6), where the cognate sequence was predicted around 8.7 million out of the 10 million peptides. In case of AU1 (Table 4.8), the cognate sequence from the target protein was ranked 97th out of a total 2395 (top 4.05%) that represented the entire proteome of the virus. The residues D, R, and P was found to be commonly occurring across the highest predicted binders. The cognate sequence of Btag was ranked 118th out of the 5857 peptides from the proteome of the bluetongue virus (top 2%). Qualitatively speaking, in both these cases the models performed slightly better with respect to biological sequences, as opposed to what was observed in case of combinatorial sequences. Such an observation can be most likely be attributed to the type of residues and sequences that were represented by these proteomes, as these biological sequences are not fully representative of the combinatorial sequence space, but a small specific fraction of it.

In conclusion, it was found that the binding interactions of only one of the monoclonal antibodies (3B5) could be successfully predicted, out of the six that were studied. For the rest of these antibodies, the predictive sequence vs. relationships which were derived were not found to identify the cognate interactions in most of the cases. The success of prediction in case of 3B5 indicates that it is possible to identify the relevant binding interactions even in case of antibodies which do not have known binding to the sequences on the array. However, there are some limitations to the current approach as was observed from the predictions in case of the rest of the antibodies. One of the major reasons would be their less-defined binding to the peptides on the array. Also, four of these

antibodies bound to cognate sequences that had isoleucine and threonine residues, which are residues omitted from the array. Therefore, the neural network does not recognize these residues well. Thus, in order to accommodate those cognate sequences for the algorithm, those residues (I and T) were replaced with other amino acids that were physicochemically similar to them (V and S respectively). This was of course not an accurate representation of the actual cognate interaction as these residues were also present in other peptides with whom the antibodies may or may not have interacted. Therefore, such substitution did not work well to help the model define the sequence vs. binding relationship. This could be avoided in future by carrying out binding experiments on arrays with all the 20 amino acid residues that constitutes most of the proteomes. Then no substitution would be required when representing these residues in the algorithm. Previous literature (Sykes et al., 2013) has stated that enabling more randomly sampled peptides on the microarray could allow better capture of information with respect to the assays. That would help one look into a higher diversity of interactions. Therefore, designing bigger arrays with an even higher number of peptides might also be worth looking into.

CHAPTER 5

PROBING PROTEIN-PEPTIDE BINDING INTERACTIONS USING IN SILICO AND IN VITRO APRROACHES

This study was conducted jointly with Kirstie Swingle

5.1 INTRODUCTION

In the previous chapters (chapters 2, 3, and 4), the ability of the neural network to predict the sequence vs. binding relationship in case of eleven monoclonal antibodies (DM1A, p53Ab1, p53Ab8, 4C1, LNKB2, 9E10, AU1, 3B5, 1D4, Btag, and Htag) was studied. The algorithm used for these predictive studies are based on the work done by Taguchi et al. (2020). In that particular study, the authors investigated the sequence vs. binding relationships for nine different proteins using a similar neural network based approach as shown here. It was shown that using sparsely sampled sequences from a combinatorial library of peptides as the input to a neural network algorithm, one could derive a predictive relationship that can be used to predict the binding between the protein(s) and any peptide from the combinatorial sequence space. The nearly random sparsely sampled peptides were represented on a high-density microarray (~125,000 unique peptides), that uses only 16 α-amino acids. Because the sampling procedure was nearly random, the algorithm was exposed to a number of sequences that might not have any biological relevance but represented a wide range of sequence patterns from the possible combinatorial space. Thus, by optimizing and training the algorithm on these sequences, it was possible to derive a comprehensive sequence vs. binding relationship between the proteins and the represented peptides. This relationship was later projected on random sequences from the combinatorial space, that were not represented on the

microarrays. As many of these non-array random peptide sequences, predicted to bind to the protein(s), represented different pockets or 'regions' from the combinatorial 'landscape', it was interesting to probe how much of the predicted binding relationship was actually corroborating with experimental evidence. The molecular interactions between the proteins and the peptide sequences are often context dependent, be it the experimental assays or predictive models. Therefore, exploring these molecular interactions with respect to changes in the environment would shed more light on how different conditions can cause changes in observed binding. The binding trends observed from the microarrays and the predictions were compared to the observations from surface plasmon resonance (or SPR) assays in this study. The microarray-based assays on which the neural network models were trained, represented a different interaction environment than that of an SPR assay, owing to the high density of peptides. As the predictive relationship between the protein(s) and the peptides were derived from the microarray experiments, it was intriguing to probe how the binding relationship changed when tested on a different platform like SPR. A thorough understanding of this predictive binding vs. experimental binding will help improve the applicability of the algorithm, thus opening up new avenues to help understand molecular recognition better.

Characterizing protein interactions using peptides have several experimental advantages over using full proteins (Benyamini and Friedler, 2010). First off, the synthesis of peptides is largely automated compared to expression and purification of proteins. Secondly, using peptides allows one to focus on identifying exact residues contributing to the interactions. Some of the assay tools to characterize protein-peptide interactions are listed in chapter 1 (section 1.3.1). For initial measurement of the binding interactions of

the protein, random-sequence peptide microarrays were used. Peptide microarrays provide an efficient tool for the simultaneous detection of interactions between a protein and several peptides. Such a platform is convenient not only for the detection of known binders, but also for identifying new sequences that are not known for interacting with the proteins (Benyamini and Friedler, 2010). In the machine learning approach laid out by Taguchi et al. (2020), the algorithm exploits the distinctive binding profile observed on the microarray in case of each protein to derive an individual sequence vs. binding relationship for them. But when a large number of different peptides are put so close to each other, the binding interactions observed can vary greatly with respect to other assays (Benyamini and Friedler, 2010). Therefore, in order to derive a thorough binding relationship between the protein and peptides, often another assay is recommended. Needless to say, peptides that were predicted to be highest binders from the combinatorial chemical space should also be experimentally probed to develop a more thorough and quantitative sequence vs. binding relationship.

Understanding how protein-peptide interactions vary across different assaying platforms is necessary and crucial for properly characterizing these interactions, as each different assay comes with its own specific set of characteristics. In microarrays the peptide features are immobilized on the glass surface. Compared to that, in SPR, the peptides are a part of the analyte solution, floating freely. Also in the microarrays, the protein analytes are fluorescently labeled whereas SPR is label-free method. SPR also allows for the measurement of binding kinetics between the proteins and peptides in real time. A study done by Greving et al. (2010), which aimed at finding peptide targets for the protein TNFα, showed a comparison between binding kinetics measured on the arrays vs. SPR. They

observed that the peptides determined to have a measurable off-rate on the microarrays, generally also had measurable off-rates in SPR. The peptides corresponding to the slowest off-rates on the array were also the slowest to come off when measured using SPR. However, the opposite was not necessarily true. Differences in binding kinetics also arises due to the fact that in SPR, the peptides are in solution, as opposed to the microarray where they are immobilized on the surface. Other studies also validated peptide array results using SPR among other methods (Katz et al., 2008, Rotem et al., 2008). In these studies, as well, the researchers found that all the peptides that demonstrated binding on the array, also had measurable dissociation constants when validated using SPR. One key point to be noted however in this case is that the peptide library they used had sequences from the target proteins that were of interest.

If one takes these studies into consideration, it is shown that there are similarities observed between interactions observed on the peptides microarrays and SPR. Whether or not these similarities also apply to the sequence vs. binding relationship determined by Taguchi et al. (2020) will be probed further in this chapter. If the relationship predicted by the neural network based on array experiments, is also validated for protein-peptide interactions in solution phase, it will expand the applicability of the predictive algorithm.

For this study, three of the proteins studied by Taguchi et al (2020) were considered - Diaphorase, Ferredoxin, and FNR. In that study, after carrying out binding experiments with these proteins on high-density, random-sequence peptide microarrays, the binding data was collected with respect to each sequence and a neural network was trained to predict specific sequence-to-binding relationship for each protein. In this work, that trained algorithm was taken and projected onto in silico libraries of random peptides that were

sampled from the combinatorial sequence space. The peptides were then sorted in descending order of predicted binding intensity. From the list, a set of peptides with varying ranges of propensity for binding to diaphorase were chosen and synthesized. Alongside these peptides, a set of sequences from the measured array that bound these proteins were also chosen. The binding of these peptides was then compared to the binding observed using SPR. The observed results from SPR were then compared to the predictive analyses and assay results from the arrays. The overall study showed that the results from the predictive analyses and the SPR experiments were related but the proportionality was found to be dependent on the assay conditions. This indicated that the assay environment played a big role in the interaction of proteins and peptides and that the predictive sequence vs. binding relationship is definitely influenced by these changes.

5.2 METHODS

5.2.1 Protein Binding Experiments on Peptide Microarrays

The data for this section was taken from the work done by Taguchi et al. (2020). Protein binding experiments were carried out on high-density peptide microarrays using fluorescently labeled proteins. The experiments were carried out on V13 microarrays from HealthTell (http://www.healthtell.com/). The array synthesis procedure is laid out in detail in chapter 2 (section 2.2.1). The V13 arrays have 126,050 peptides that are composed of 16 amino acids. For the assays, the proteins (diaphorase, ferredoxin, and FNR) were labeled using AlexaFluor 555 NHS ester (ThermoFisher cat. # A37571) (Taguchi et al., 2020). The list of the proteins is provided in Table 5.1 along with their source and molecular weight. The final concentration of the labeled proteins used for binding measurements on the arrays was 10 nM.

For the assays, the slides containing the microarrays were removed from vacuum sealed storage and loaded into custom microarray cassettes, that arranged them in 96 well plate format. The arrays were then hydrated with PBST (81 µL/well) for 30 minutes at 37 °C. Following that, each array was incubated with 9 µL of the labeled protein (in PBST) at 37 °C for 1 hour, using a shaker for mixing. After the completion of incubation time, the arrays were washed three times with PBST, followed by washing with distilled water. The slides were then removed from the custom cassettes and sprayed with isopropyl alcohol. They were then centrifuged to ensure dryness. After drying, the arrays were imaged using a fluorescence imager at 532 nm excitation wavelength and 750 ms exposure time. The images were quantified to align the fluorescence measurements with the peptides using Mapix software (Innopsys, Carbonne, France). There were three technical replicates in terms of assays, for each protein. The correlation between all three replicates in each case was 0.99. It must be noted here, that for these experiments the diaphorase was obtained from Sigma, but later experiments in this study the diaphorase was sourced in-house.

Table 5.1. Proteins used and their sources and molecular weight

Protein	**Source**	**Catalog Number**	**Molecular Weight (kDa)**
Diaphorase	Sigma	D1315	30.1
Ferredoxin	Sigma	F3013	11.1
FNR	Prof. Kevin Redding (ASU)	N/A	35.3

5.2.2 Training the Neural Networks

The sequence and binding data from the array experiments was then used to train a feed-forward, backpropagating neural network. It must be noted here that the highest binding ~2% of the measured data was excluded from the training set to avoid fitting the

model with saturated data. The excluded sequences were included in the validation set (~12,350 peptides) for the model. The encoding principle of the neural network is the same as shown in chapter 2 (Section 2.2.3). However, different hyperparameters were used. Table 5.2 lists the hyperparameters that were used for training the model. A total of 100 independent training runs were carried out for each antibody. The average correlation between predicted and measured data was greater than >0.98 in all the three cases. More information on this could be found out in the supplementary information of the work published by Taguchi et al. (2020).

Table 5.2. Hyperparameters used for the neural network

Hyperparameters	**Values**
Number of Hidden layers	2
Number of Hidden Nodes per Layer	100
Fraction of Peptides used for Training	0.9
Number of Amino Acid Descriptors	10
Training Steps	50,000
Learning Rate	0.001

After completion of training, the diaphorase-specific model was projected on 1 million randomly generated 9-mer peptide sequences, that were broken down in silico into sub-groups of 100,000 sequences. These subgroups were sorted in alphabetical order with respect to the first 4 residues from the N-terminus of the peptide. This was done to ensure that there was an even distribution and representation of different peptides in all the subgroups. After projecting the model on to each subgroup, the sequences were sorted in descending order of predicted binding value. Nearly the top two-thirds (~65,500) of the sequences from the sorted list were then chosen to be added to the final compiled list of sequences. As there were 10 subgroups, this was repeated 10 times and the number of

peptides in the final resulting dataset was 655,361. The ferredoxin and FNR-specific models were also projected on these sequences.

5.2.3 Selection and Synthesis of Peptides

After the projection was completed, a number of peptides were chosen to be synthesized from the compiled final list of peptides (655,361 sequences). It must be noted here that these sequences were randomly generated so there was no experimental record of any interaction of these peptides with the proteins, beforehand. The peptides were chosen according to their predicted binding specificity towards diaphorase and isoelectric point. The predicted binding specificity was calculated by taking the ratio between the predicted binding intensity of the peptide for diaphorase and predicted binding intensity for ferredoxin or FNR. Using this information, 43 peptides were chosen from the list to represent as much sequence variability and an even distribution of isoelectric point as possible. Of these chosen peptides, 5 sequences had a predicted binding intensity of 50,000 – 80,000 with respect to diaphorase, 10 between 20,000 – 30,000, 24 between 5,000 – 15,000, and 5 between 2000 – 3000. Sequences grouped by binding values had a somewhat similar sequence space representation. Aside from this, 10 peptides from the arrays were also chosen to be synthesized. All the peptides were synthesized by Sigma-Aldrich and the complete list can be found in Appendix C. The peptides were received in 2 – 10 mg powder form and came with mass spectrometric reports (MALDI-TOF) for quality control (QC). The mass of the peptides was also ensured in-lab using MALDI-TOF to verify the QC reports. A representative QC MALDI-TOF report is shown in Figure 5.1. Before any experiments were conducted, all the peptides were solubilized in 20% DMSO/distilled water mixture and stored at 4 °C.

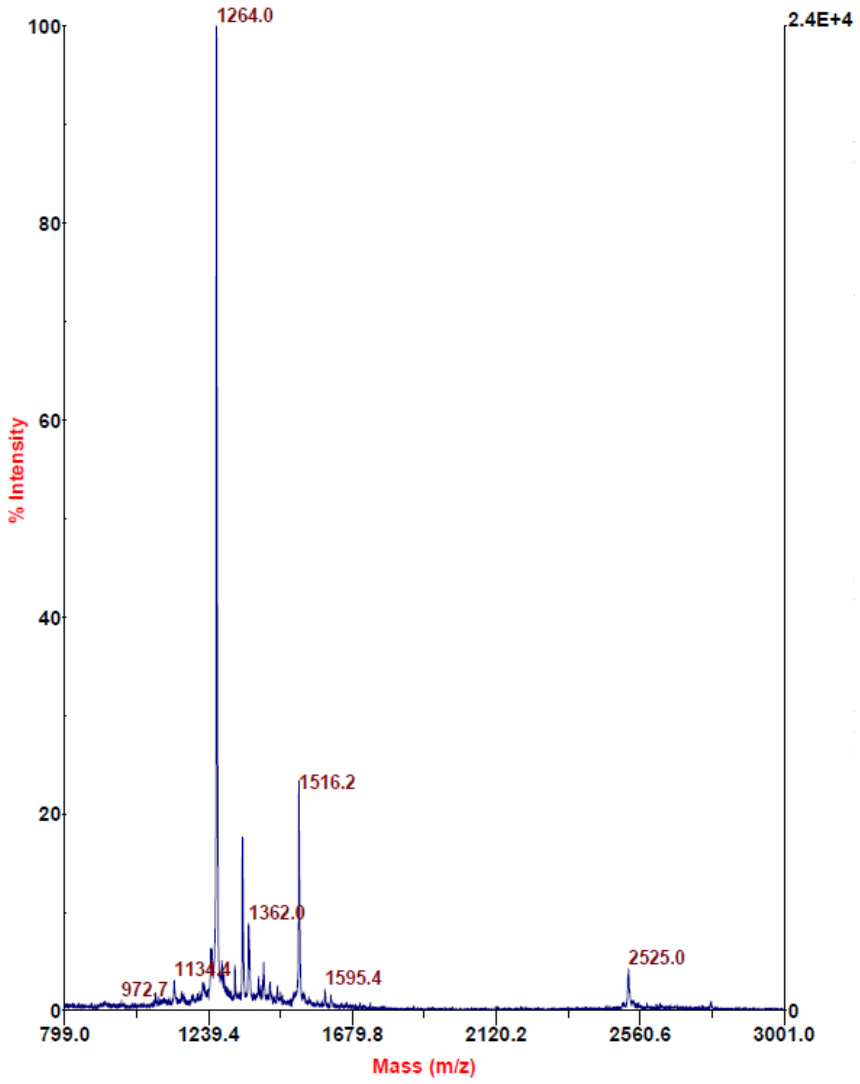


Figure 5.1. Representative MALDI-TOF quality control report for the peptide GERWVYYEY as provided by Sigma-Aldrich. The theoretical molecular weight of the peptide is 1263.397.

5.2.4 Binding Experiments using Surface Plasmon Resonance

The solubilized peptides were then used to test their binding against three proteins - diaphorase, ferredoxin, and FNR using SPR. All of the SPR assays were carried out using a Biacore T200 system from Cytiva (www.cytivalifesciences.com/en/us). In these assay experiment, the proteins were immobilized over a carboxymethyl dextran matrix

covalently attached to the CM5 gold chip, also from Cytiva (Catalog # BR100530). the dextran matrix coating of the chip was important for ligand binding to unmodified surfaces. The peptides remained in solution and were flowed over the chip. The buffer used for the assays was HBS-N buffer (1X, pH = 7.4) with P20 (0.1%), and carboxymethyl dextran (1mg/ml), unless noted otherwise. All the assay protocols were maintained using the T200 software.

Before immobilizing the proteins, a pH scouting study was conducted on the chip to gauge the pH at which the immobilization will be carried out. This was done to ensure the pH conditions under which the dextran surface was coated with the maximum amount of protein ligand, resulting in the highest signal intensity. Usually, a pH values which is slightly under the isoelectric point of the protein is an optimum condition. Nevertheless, scouting different pH values is helpful in finding the ideal condition. In case of all the 3 proteins, he solubilized protein was introduced to the chip using 10 mM sodium acetate buffers at 4 different pH levels (4.0, 4.5, 5.0, and 5.5). After each scouting, the surface was washed with 50 mM NaOH. The conditions of pH scouting for the protein, diaphorase, is given below in Table 5.3. After scouting for pH, the optimum pH condition that was found for diaphorase was pH 5.0. The protein was then immobilized on the 1 flow cell of the CN5 chip using 10mM sodium acetate buffer at pH 5.0. The condition for immobilization of diaphorase is given below in Table 5.4. Similar conditions were used for other proteins.

Table 5.3. pH Scouting Conditions for Diaphorase

Protein	**Concentration**	**Flow rate**	**Contact time**	**Isoelectric Point**	**Selected Buffer pH**
Diaphorase	50 µg/ml	5 µl/minute	180 s	4.2	4.0,4.5, 5.0, 5.5

Table 5.4. Immobilization of Diaphorase on CM5 Chip

Protein	**Concentration**	**Flow rate**	**Contact time**	**Buffer**	**pH**	**Immobilization response unit (RU)**
Diaphorase	50 µg/ml	10 µl/minute	360 s	Sodium Acetate	5.0	8191

The immobilization of diaphorase on the dextran matrix was carried out covalently using N-hydroxysuccinimide/1-Ethyl-3-(3-dimethylaminopropyl) carbodiimide (or NHS/EDC) coupling. The dextran matrix of the chip was activated using 400 mM EDC and 100 mM NHS which changes the carboxymethyl groups present on the surface to NHS esters. The NHS esters then reacted with the amine group present in the proteins to immobilize it on the surface. Lastly, the immobilized surface was deactivated using 1 M ethanolamine that also washed away unreacted protein. After this, the quality of the immobilization was verified by testing the binding results of 14 nM anti-diaphorase antibody solution. After each binding event, regeneration of the surface was done using 10 mM glycine solution at pH 3.0.

Following the completion of the immobilization procedure, the chips were stored in 1X HBS-EP buffer (Cytiva) at 4 °C, until utilization (max. 1 week). For the next steps, the solubilized peptides were screened for binding to the protein. The screening assays were carried out at 25 °C. The peptides were made into stock solutions of 5 µM using 1X HBS-N buffer. Before screening, the surface-modified CM5 chip was equilibrated to the assay temperature by running 2 cycles of assay buffer. 150 µl of each peptide was pipetted into the wells of standard 96-well titer plate, that was inserted into the Biacore T200 system. Each peptide had 60 seconds of contact time at a flow rate of 30 µl/s. The

dissociation time allotted for each peptide was 60 seconds. After each association and dissociation event, the response unit was recorded, and the protein surface was regenerated using 10 mM glycine (pH 3.0). All response unit values were subtracted from a reference cell (no protein) to remove background noise. To ensure that the protein activity was not lost over multiple cycles of regeneration, anti-diaphorase antibodies were tested on the chip surface periodically during peptide screening. Similar steps were repeated for ferredoxin and FNR as well. The excess stocks of peptide solutions were stored in -20 °C for future use.

Post-screening, the peptides that gave rise to a response greater than or equal to 4 units were chosen to evaluate their dissociation constants for binding to the three proteins. For evaluating dissociation constant, each peptide stock solution was serially diluted to obtain 7-10 different concentrations. 150 µl of each peptide solution at different concentration was then added to a standard 96-well titer plate. During the assay, each sample was allotted 90 seconds of association and 120 seconds dissociation time, followed by regeneration. The response units were adjusted with respect to the reference cell. For each peptide, the set of final response units obtained from the assay was used to calculate the dissociation constant, using a graphing software. The equation used for calculating dissociation constant is as follows:

$$RU = \frac{RU_{max} * conc}{K_D + conc}$$

In this equation, RU is response unit, RU_{max} is the maximum recorded response unit, conc stands for the concentration of the peptide, and K_D is the dissociation constant.

5.3 RESULTS

5.3.1 Prediction of Protein Interactions using Machine Learning

The binding behavior of three protein were probed in this study (diaphorase, ferredoxin, and FNR). All of these three proteins are a part of the light-dependent electron transport pathway related to Photosystem I, that is responsible for the reduction of nicotinamide adenine dinucleotide phosphate ($NADP^+$) to NADPH. In the in silico study, a neural network was first trained on sequence and binding data available from assaying these proteins on high-density random sequence peptide microarrays (126,050 unique peptides). The random sequence peptide microarrays allowed sparsely sampling sequences from the combinatorial space. The predicted binding interactions of the protein were based on this sparse sampling of sequence space. Pearson correlation coefficient was used as a measure to assess the performance of the models. The correlations between the predicted and the measured binding intensities in all the three cases were observed to be greater than 0.983 (average of 10 runs). The results of the training are shown in Figure 5.2. Upon training the neural networks, a comprehensive and quantitative sequence vs. binding relationship was obtained for each protein. This quantitative relationship could be used to predict the binding relationship between the protein and any peptide from the combinatorial sequence space ($\sim 10^{12}$ peptides). The trained models, specific to each individual protein, were then projected on an in silico library of randomly generated peptide sequences (10^6 peptides) from the combinatorial sequence space (section 5.2.2).

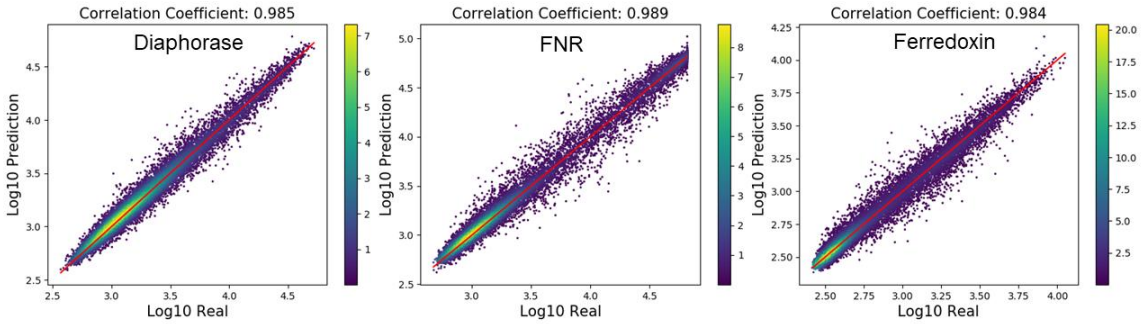


Figure 5.2. Scatter plots showing Pearson correlation coefficients between predicted and measured binding of the array peptides (testing set peptides only) for diaphorase, FNR, and ferredoxin. Colorbar indicates density of plot points.

5.3.2 Selection and Synthesis of Predicted and Array Peptides

After projecting the diaphorase-specific models on the in silico peptide library (10^6 sequences), 655,361 peptides were chosen that represent various ranges of predicted binding to diaphorase (section 5.2.2). Following the selection, the ferredoxin and FNR-specific models were also projected on to this library to calculate the predicted binding intensities of the sequences with respect to these two proteins. The isoelectric points of these sequences were also calculated. Aside from these sequences, the isoelectric points of the peptides represented on the array were calculated as well. Figure 5.3 shows the variation in number of sequences and isoelectric point against a range of binding values, both for predicted peptides that were selected among a million, as well as array sequences. The binding intensities shown in this figure are for diaphorase related measurements only. The selection of the predicted peptides was done on the basis of distribution of binding intensities and isoelectric points.

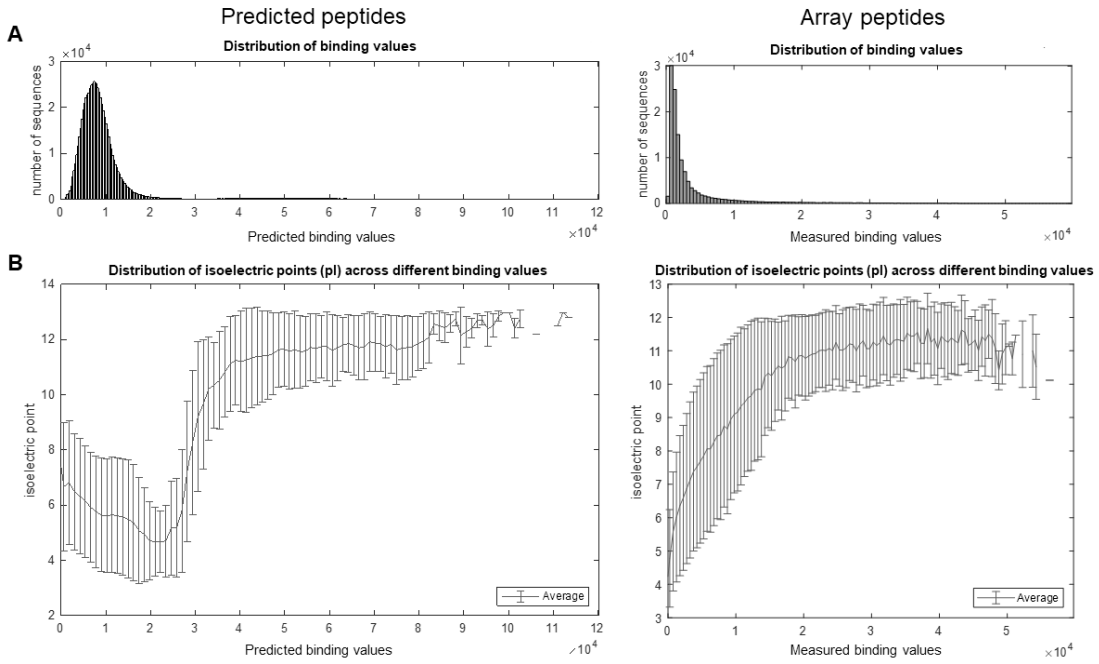


Figure 5.3. Distribution of number of sequences vs. binding values (A), and distribution isoelectric points against binding values (B), for predicted peptides and array peptides. These values are for diaphorase specific models only. All binding values are in relative fluorescence units (RFUs). The error bars shown here are standard deviations from the mean values.

In Figure 5.3, it can be seen that in case of both predicted and array peptides, increase in binding value corresponds with increase in isoelectric point to a certain extent. In both the cases, binding intensities greater than 40,000 correspond with mean pI values between 10 – 11. However, if lower binding intensities (<30,000) are considered where majority of the sequences lie in both the cases, then the observed trends are slightly different in the two cases. In case of predicted sequences, the mean pI value dips (pI ~5) until it starts going up again, whereas in the case of array sequences the mean pI value starts around 5 even in the case of lowest binders and increases gradually. Overall, it can be concluded that the highest binding sequences (both measured and predicted) have a higher pI than the rest of the sequences.

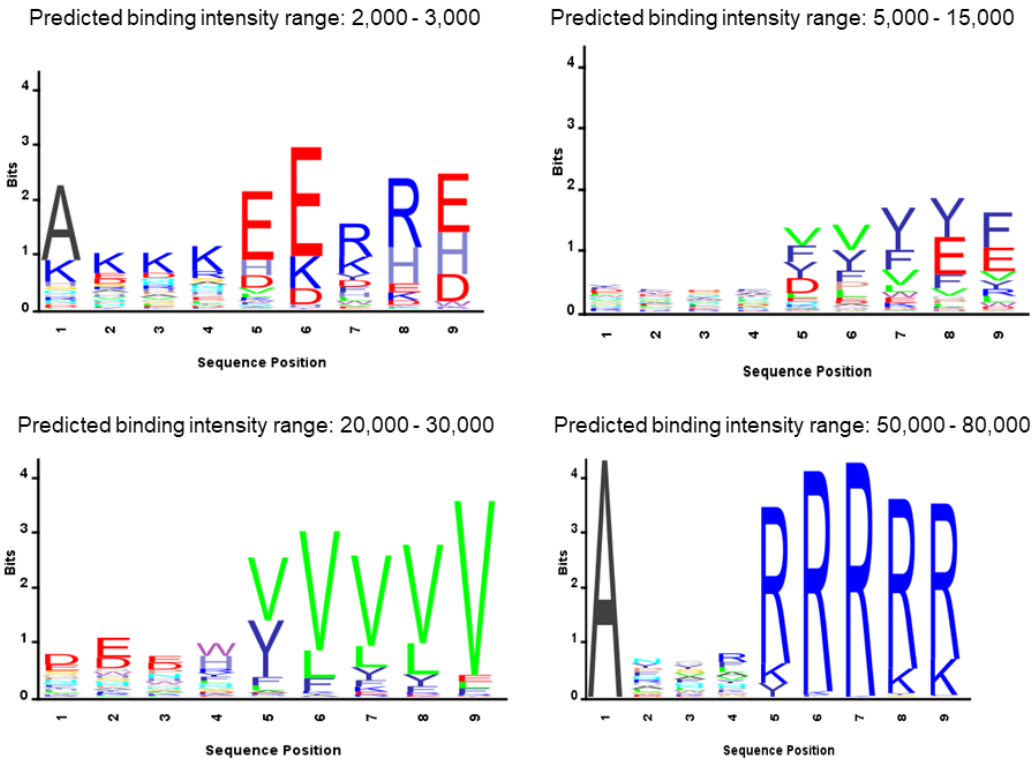


Figure 5.4. Sequence logos of random in silico peptides from predicted binding intensity ranges (2000 – 3000, 5000 – 15,000, 20,000 – 30,000, and 50,000 – 80,000 RFUs) for diaphorase. The N-terminus is on the left-hand side. Red indicates acidic residues, blue indicates basic residues, green and grey indicate non-polar residues, and navy blue indicates aromatic residues.

Out of the total number of random peptides that were chosen in silico, 43 peptides were chosen to be synthesized, and validate their binding to the three proteins experimentally. In order to choose such a small number of peptides from a relatively huge list, the whole range of predicted binding intensities were broken down into smaller subranges (2000 – 3000, 5000 – 10,000, 20,000 – 30,000, and 50,000 – 80,000). The peptides in each of these subranges were aligned using Clustal Omega multiple sequence alignment and the sequence logos of the aligned peptides were plotted using MATLAB. Figure 5.4 show the sequence logos for each of these binding value subranges.

Table 5.5. Number of peptides selected from each source and binding range

Peptide Source	Number of sequences
High-density peptide microarray	10
Predicted from projection (2000 – 3000)	5
Predicted from projection (5000 – 15,000)	24
Predicted from projection (20,000 – 30,000)	10
Predicted from projection (50,000 – 80,000)	5

This is done to analyze if there were any common motifs observed among the sequences that have similar binding intensities. From Figure 5.4, it can be seen that at lower predicted binding range (2000 – 3000) a combination of basic (K, R, H) and acidic (D, E) are prevalent. However, acidic residues are slightly favored over the basic residues in this case. As the predicted binding intensities go higher up (5000 – 15,000 and 20,000 – 30,000), charged residues become less prevalent and hydrophobic residues are more preferred, especially near the C-terminus (right side). These binding ranges also had the greatest number of sequences (Figure 5.3 (A)). It must be noted that the first four positions from the N-terminus (left) show almost no preference towards any amino acid residue. This is because the randomly generated library of 1 million peptides was divided into smaller subsets, which were arranged alphabetically such that the N-terminus and the next few consecutive positions (positions 1-4) had more or less equally distributed propensity for all the amino acids. Such was done to emulate the representation of the array peptides which have a similar property, with more variable representation of residues near N-terminus, due to the nature of the array manufacturing process. In case of the highest predicted binding ranges (50,000 – 80,000), an overwhelming majority of basic residues, like arginine, can be observed near the C-terminus. This is very likely because the basic residues, like arginine and lysine, are highly charged causing them to be ‘sticky’ in nature. Owing to

their extremely charged nature, they can interact not just with diaphorase, but with a variety of other proteins as well.

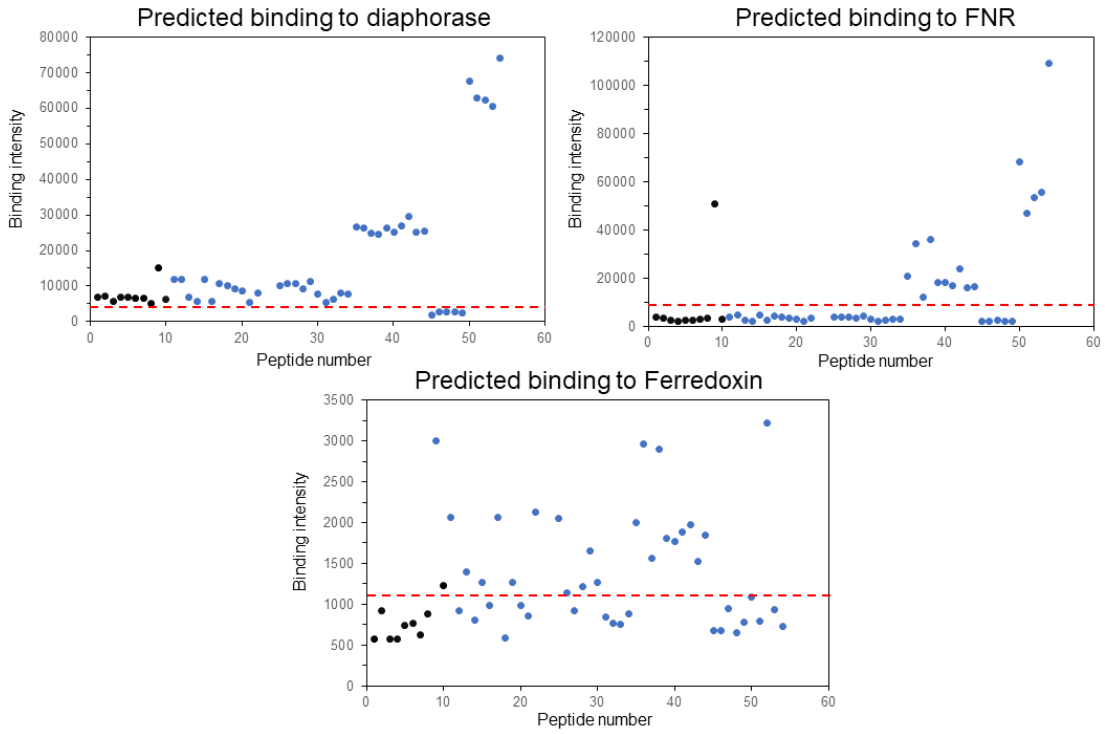


Figure 5.5. Predicted binding distribution of the synthesized peptides for diaphorase, FNR, and ferredoxin. The black circles are array peptides and blue circles are predicted peptides. The red dashed line represents median of array binding experiments for each protein.

Based on the above observations, a set of 43 representative peptides were chosen from each predicted binding intensity range. 10 peptides were chosen from the array sequences as well, bringing the total number of peptides up to 54. The number of peptides selected from each category of binding range is given below in Table 5.5. The full list of selected peptides can be found in Appendix C. The predicted binding intensities of these synthesized peptides for diaphorase, ferredoxin, FNR is shown in Figure 5.5.

It must be noted that in case of diaphorase, the predicted intensities for almost all the peptides are above the median measured binding value from the array, but not for FNR and ferredoxin. In case of FNR, most of the peptides are below the array mean cut-off, and in case of ferredoxin they are more or less evenly distributed across the mean cut-off. Based on these observations, and calculations from in silico projections, it was expected that these peptides will have higher specificity towards diaphorase than towards the other two proteins. Thus, by considering the variation of charge, binding intensities, and observed motifs across the peptides, a set of sequences were selected that were likely representative of a wide variety of binding interactions with the proteins.

5.3.3 Initial Screening of Peptides using Surface Plasmon Resonance

For experimentally verifying the binding interactions of the selected peptides to the three proteins, SPR assay was used. SPR assay is a highly-sensitive label free detection method to characterize the protein interactions. The phenomenon of surface plasmon resonance occurs when photon from the incident light hits the surface of the thin gold chip at a particular angle (Nguyen et al., 2015). The incident light stimulates the surface electrons and a portion of the photons couples with the electrons. These electrons then start moving parallel to the surface (plasmons). The oscillations from the plasmons generate an electric field with a range of ~300nm from the boundary between the gold surface and the analyte solution. In the SPR instrument, there's a light source that constantly interacts with the gold surface and incident light is passed through a glass prism with high-refractive index. In the absence of any surface interactions, when plasmon oscillations are taking place due to SPR, the photons from the incident light are reflected at a certain constant angle due to total internal reflection through the prism. When molecular interactions are

taking place on the surface of the gold chip, it causes perturbances to the surface plasmon movement. These perturbances consequently change the refractive index, therefore changing the angle of the reflected light which is then measured by the detector. For example, immobilization of a protein on the surface, or ligands binding to the immobilized protein will cause a change in mass of the gold surface, causing the refractive index to change. The difference in the angle of the reflected light before and after an interaction is what is recorded by the instrument. A schematic diagram of SPR is shown in Figure 5.6. Aside from the high-sensitivity and label-free detection of this technique, it has also been noted that there is a good correlation between the results obtained from SPR and from microarray-based assays (Katz et al., 2008, Rotem et al., 2008, Greving et al., 2010). Therefore, SPR was chosen as a suitable method to verify the protein-peptide interactions in this case as well.

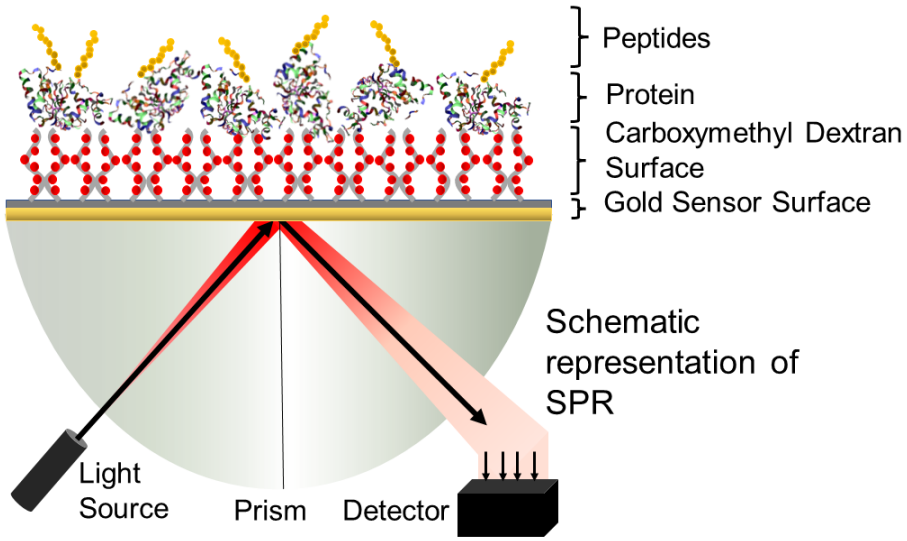


Figure 5.6. Schematic representation of a surface plasmon resonance assay to record protein interactions.

In order to characterize the protein-peptide interactions, the protein has to be first immobilized on the surface of the gold chip. To ensure that the proteins can be bound well

to the surface, the gold chip was coated with carboxymethyl dextran, which facilitates covalent bonding with the protein. Before immobilizing the protein on the chip surface, one needs to determine the pH at which maximum immobilization takes place. Usually, the immobilization is carried out at a pH which is lower than the over pI of the protein. To find the optimal pH for immobilization, the protein binding on the surface is tested at different pH conditions. This is known as pH scouting. The pH of the solution was maintained using 10 mM sodium acetate of the required pH value. Figure 5.8 shows the result of diaphorase immobilization at 4 different pH values (4.0, 4.5, 5.0, and 5.5). It can be seen from the figure that at pH 4.0 and 4.5, the protein associated itself to the surface and achieved saturation but almost came of completely during the dissociation phase. Therefore, they were not suitable for immobilization. At pH 5.5, the protein kept gradually attaching to the surface (steeper curve) without reaching saturation. The sudden drop on response in the dissociation phase could be attributed to washing away of aggregated protein by the buffer solution. At pH 5.0, a gradual increase in protein binding was observed which slowed down towards the end of the association phase, indicating that it was near saturation. Therefore, pH 5.0 was determined to be the best for carrying out diaphorase immobilization. For immobilization, the dextran surface on the chip was activated using NHS/EDC coupling, which rendered the surface with NHS-ester functionalization. This functional group then covalently reacted with the free amines on the protein molecules to immobilize them completely on the surface.

After immobilization, whether the protein bound correctly or not was checked by using anti-diaphorase antibodies. Figure 5.7 also shows the responses recorded from the diaphorase immobilized chip in the absence and presence of the antibody. It can be seen

that there was an increase in the response signal (maximum response 8493) when 14nM antibody was introduced to the protein coated surface, but no significant response (response value 8438) was recorded in the absence of the antibody, when compared with the base response value of 8435. This indicated that the immobilization of diaphorase on the surface of the gold chip was successful. Similar procedures were followed for FNR and ferredoxin as well to immobilize them on the surface of the gold chip.

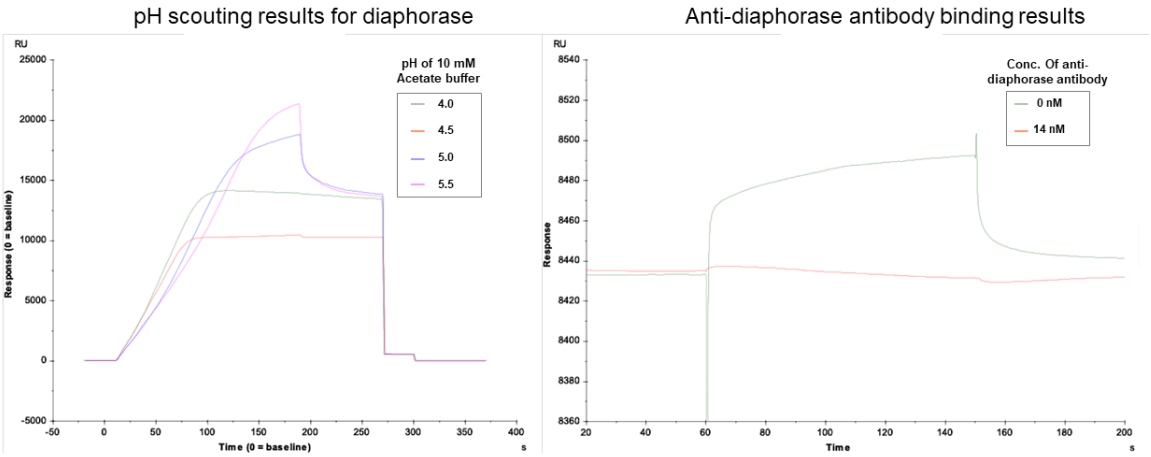


Figure 5.7. SPR Sensorgrams showing results of pH scouting for diaphorase immobilization (left) and post-immobilization quality control using anti-diaphorase antibodies (right).

When the protein was captured via NHS/EDC coupling of the surface, the orientation of the protein on the surface could not be controlled precisely, as any reactive amine group on the protein would form a bond with NHS ester. This might lead to the blockage of some potentially reactive sites on the protein. Therefore, another method of protein immobilization was also explored in this study, which involved capturing the proteins on the gold surface with the help of streptavidin-biotin interactions. This method helped in immobilizing the proteins with a controlled orientation on the surface. Additionally, the use of the streptavidin coated chip was expected to reduce the interactions

of the peptides with the surface of the reference cell (no protein). A streptavidin coated gold chip was used for this purpose, instead of a carboxymethyl dextran coated one. The protein of interest (diaphorase, FNR, or ferredoxin) was conjugated with biotin, and the capture on the chip surface took place through streptavidin-biotin interactions. Although it is a non-covalent interaction, it is still a strong interaction because of the high binding affinity between streptavidin and biotin.

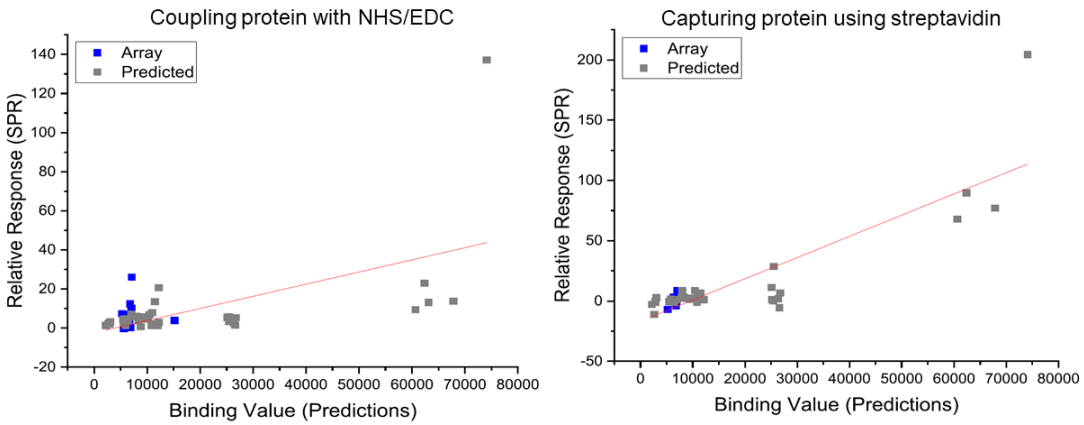


Figure 5.8. Results of peptide screening (54 peptides) for diaphorase binding using NHS/EDC coupling (left) and streptavidin-biotin capture (right). Blue markers are array peptides and grey markers are predicted peptides. red line indicates the best fit in both the cases. Screening was performed using SPR.

Following the immobilization procedures, screening of the synthesized peptides was carried out using SPR, to select the binders for diaphorase among the 54 peptides. During the screening process, all the peptides first tested for binding to diaphorase at 1 µM concentration each (section 5.2.4). But it was observed that 1 µM was not a sufficiently high concentration as SPR response of all the peptides to the protein were low at this concentration. The peptide concentration was then increased to 5 µM and better binding responses were observed in this case. The comparison of peptide screening for diaphorase binding on both the surfaces are shown in Figure 5.8.

From Figure 5.8, it can be seen how the binding results vary when two different immobilization methods are used. The response values shown here are the relative responses, where the reference baseline have already been subtracted from the final recorded response. It was observed that the overall SPR responses of the peptides were higher in case of streptavidin-biotin capture compared to that of the NHS-EDC capture. It is also shown here that the SPR responses of the peptides belonging to different binding ranges have a much more pronounced correlation with the observed trend in predicted binding. In indicated that the protein orientation on the surface of the gold chip mattered when considering interactions with these peptides. Orienting the protein in a consistently across the chip surface minimized the non-specific interactions of the analyte, therefore resulting in higher recorded response. These results show the importance of physicochemical context when considering the molecular interactions of proteins and peptides. However, further studies with the streptavidin-biotin capture method could not be conducted due to instrument malfunction and constraints with funding. It must be noted though that the peptides were in solution in both these cases, as opposed to being immobilized on the surface like the microarrays.

Hence all of the following studies were conducted on SPR chips which had the proteins immobilized with the help of NHS-EDC coupling. Figure 5.9 A show the peptide screening results carried out for all three proteins, using the same capture method. From the results, it was found that peptides PWELYFWRD (peptide 2), QERWFYYEF (peptide 11), and ARYRRYRRK (peptide 54) showed high binding activity towards all three proteins. Peptide 54 showed very high binding compared to the other peptides, for all the three proteins, due to its highly positively charged nature, owing to the five arginine and

one lysine residue. The peptides that were more selective towards diaphorase compared to the other proteins were FAWPAWVAWFE (peptide 5), GEKWFFYEF (peptide 29), and AWVDYRRRF (peptide 52). In general, it can be seen from the screening process for diaphorase (Figure 5.9 A) that the recorded response units (RUs) from SPR matches the trend of the predicted binding value ranges. The peptides from the lower predicted binding ranges have lower RUs, whereas peptides from the higher predicted binding ranges have higher RU values.

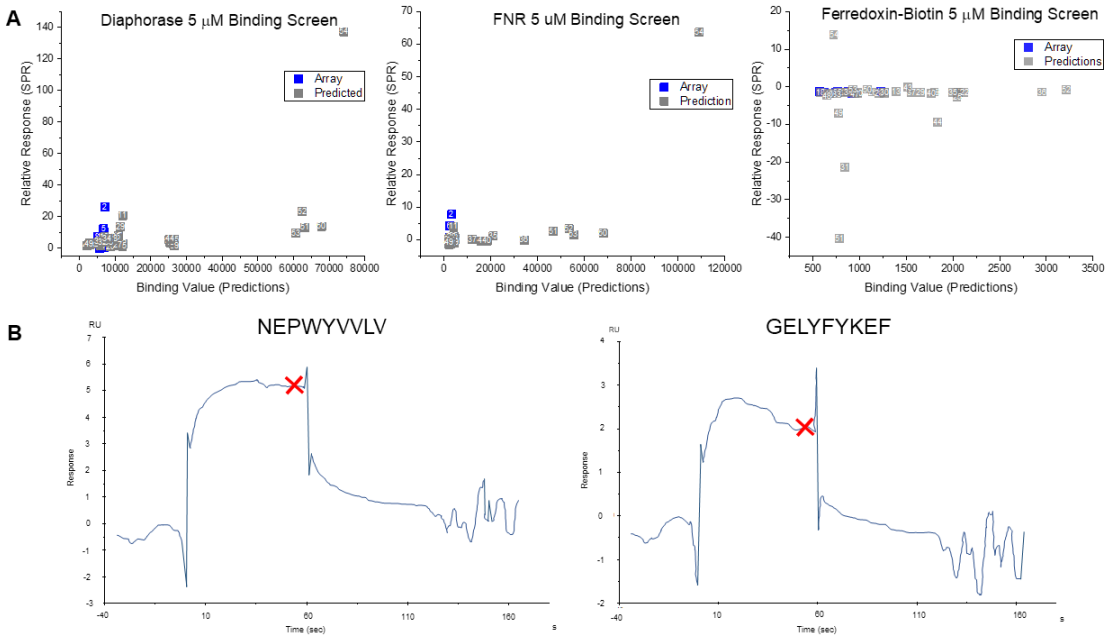


Figure 5.9. (A) Results of peptide binding screening for diaphorase, FNR, and ferredoxin using SPR. Array peptides are denoted in blue and predicted peptides are denoted in grey. (B) Examples of sensorgrams showing association and dissociation curves from peptides with response values greater than 4 (left) and less than 4 (right).

After the screening process was complete, peptides that recorded RU greater than or equal to 4 for binding to diaphorase, were short-listed for further assessment and evaluation of dissociation constant. The RU threshold for the peptide selection was determined based on the shape of the sensorgram observed in each case. Peptides with RU

values above 4 were found to have clearly defined association and dissociation curves in their sensorgrams, whereas peptides with RUs below 4 had signal-to-noise ratios that were insufficient for an accurate determination. An example of each case is shown in Figure 5.9 B. Based on the RU threshold, 4 out of 10 array peptides, and 18 out of 44 predicted peptides made the cut-off. Therefore, the total percentage of successful candidates were found to 40% in both the cases.

It is worth noting that the experimentally recorded response in all the cases were much lower than the theoretical maximum response. The theoretical maximum response (Response $_{max}$) is calculated using the formula given below:

$$Response_{\ max} = \frac{Bound\ Protein\ Reponse \times Molecular\ Weight\ of\ Analyte}{Molecular\ Weight\ of\ Bound\ Protein}$$

where bound protein response is equivalent is equivalent to the response recorded during protein immobilization. In this case the analytes are the peptides. It was found using this equation that the approximate maximum response is around 160. However, most of the responses recorded experimentally were under 20% of the theoretical maximum. This observation suggested that the assay conditions were probably not optimal at the time which will be discussed later in the chapter.

5.3.4 Measuring Dissociation Constants of the Selected Peptides

The next step was to measure the dissociation constant (K_D) of the short-listed peptides by studying their binding affinities at different concentrations. 4 out of the 10 array peptides and 18 out of the 44 predicted peptides were chosen for this purpose. In order to conduct these studies, the peptides were prepared at different concentrations (minimum 4 concentrations) with the help of serial dilution. The binding response of the

individual peptides to the protein, at different concentrations, was measured using SPR. Dissociation constant was then calculated from the response signals recorded at various concentrations using the equation given in section 5.2.4. The sensorgram responses from the SPR and plots showing the fitting curves for calculation of dissociation constant for two of the selected peptides are presented in Figure 5.10.

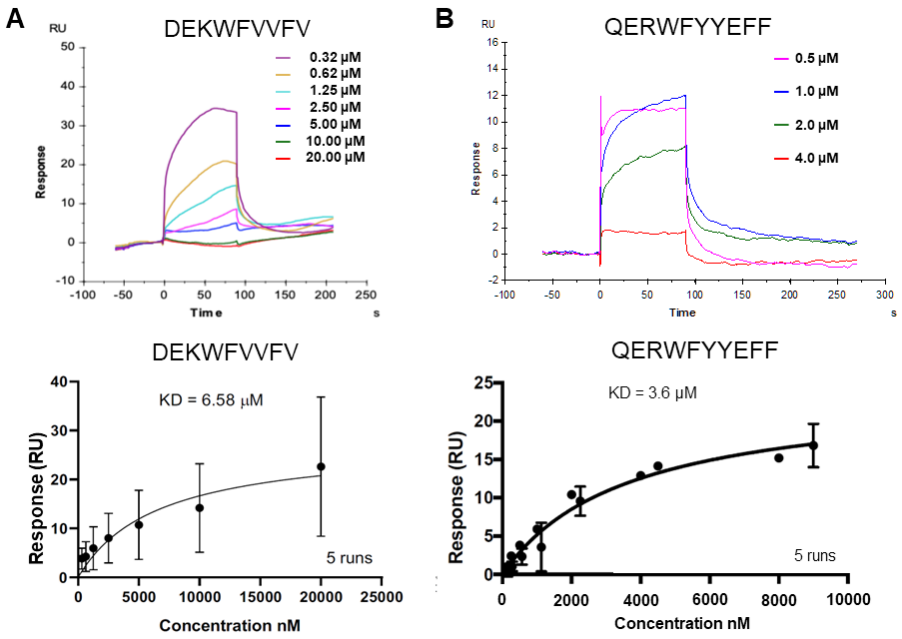


Figure 5.10. Sensorgrams and plots showing recorded SPR response at different concentrations and curve fitting for calculation of dissociation constant respectively, for peptides DEKWFVVFV (A) and QERWFYYEFF (B). Each experiment was repeated five times and the average response was recorded.

The relationship between measured K_D and predicted binding values for the predicted peptides have been shown in Figure 5.11. All the measurements presented here are with respect to diaphorase binding. No peptides that were predicted to have a binding value lower than mean predicted binding intensity (9271.19) were found to have K_D values less than 10 µM. Interactions with K_D value higher than 10 µM are not considered physically significant. Among the peptides whose predicted binding value were higher than

mean predicted binding, a few peptides were found to have dissociation constants lower than 10 µM. In total, six peptides out of the 18 were found to have a K_D value that were below the acceptable cut-off of 10 µM. Out of the six peptides, 1 was from predicted binding range of 5000 -15,000, 3 were from range 20,000 – 30,000, and 2 were from range 50,000 – 80,000. The two peptides that belonged to the last binding range group were highly positively charged, which caused issues with peptide aggregation and unwanted binding to reference cell. Although dissociation constants might not be the best measure for such reasons to determine binding interactions effectively in this case, it was interesting to note that there was some correlation between the predicted binding values and the calculated K_D values. The list of the peptides whose K_D values were less than 10 µM are shown in Table 5.6.

Table 5.6. List of peptides with dissociation constants less than 10 µM as measured by SPR during diaphorase binding experiments.

Peptide number	**Sequence (N-terminus to C-terminus)**	**Calculated K_D value (µM)**
23	QERWFYYEFF	3.60
35	DEKWFVVVFV	6.58
38	EERWYVVLV	5.10
39	PENWVLVVV	4.70
53	ADGFRRRKR	6.08
54	ARYRRYRRK	1.99

Following the analyses of the shortlisted predicted peptides, the four array peptides were also tested for binding to diaphorase and evaluation of dissociation constant. The binding affinities observed on the array differ from the binding affinities observed in SPR even for the same protein-peptide pair, due to the changes in the physical and chemical environment. The four peptides previously selected were shortlisted from the 10 peptides

that were selected for binding to diaphorase on the array. It shows that not all of the binding interactions observed on the array are translated to SPR, due to change of assay environments. Of the four peptides that were chosen for evaluation of K_D, only one peptide consistently succeeded in giving reproducible dissociation constant values that were in the sub-10 µM range. This further goes on to show that the variability of environments in different assays impact the interactions between the proteins and the peptides greatly. The responses from the SPR analyses of the four peptides at different concentrations are shown in Figure 5.12, as well as the fitting curve of the peptide with reproducible values of dissociation constants.

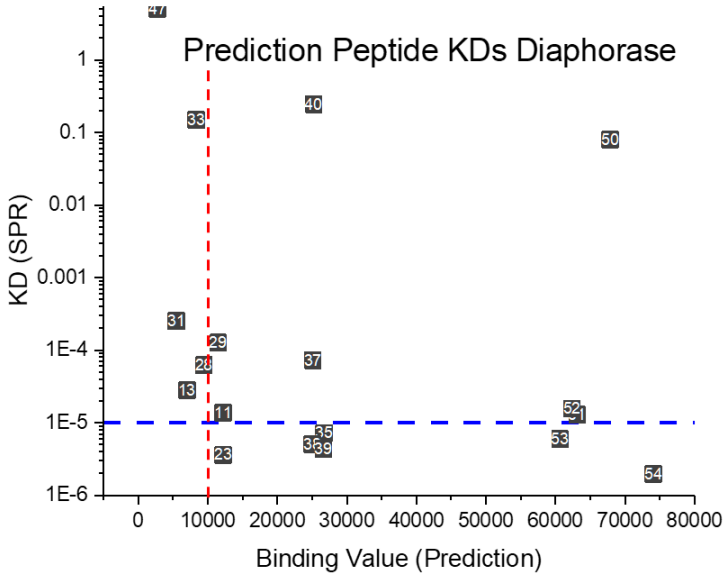


Figure 5.11. Distribution of dissociation constant against predicted binding intensities for the predicted peptides. The blue line represents 10 µM cut-off for dissociation constant. The red line represents the mean binding intensity of the array peptides.

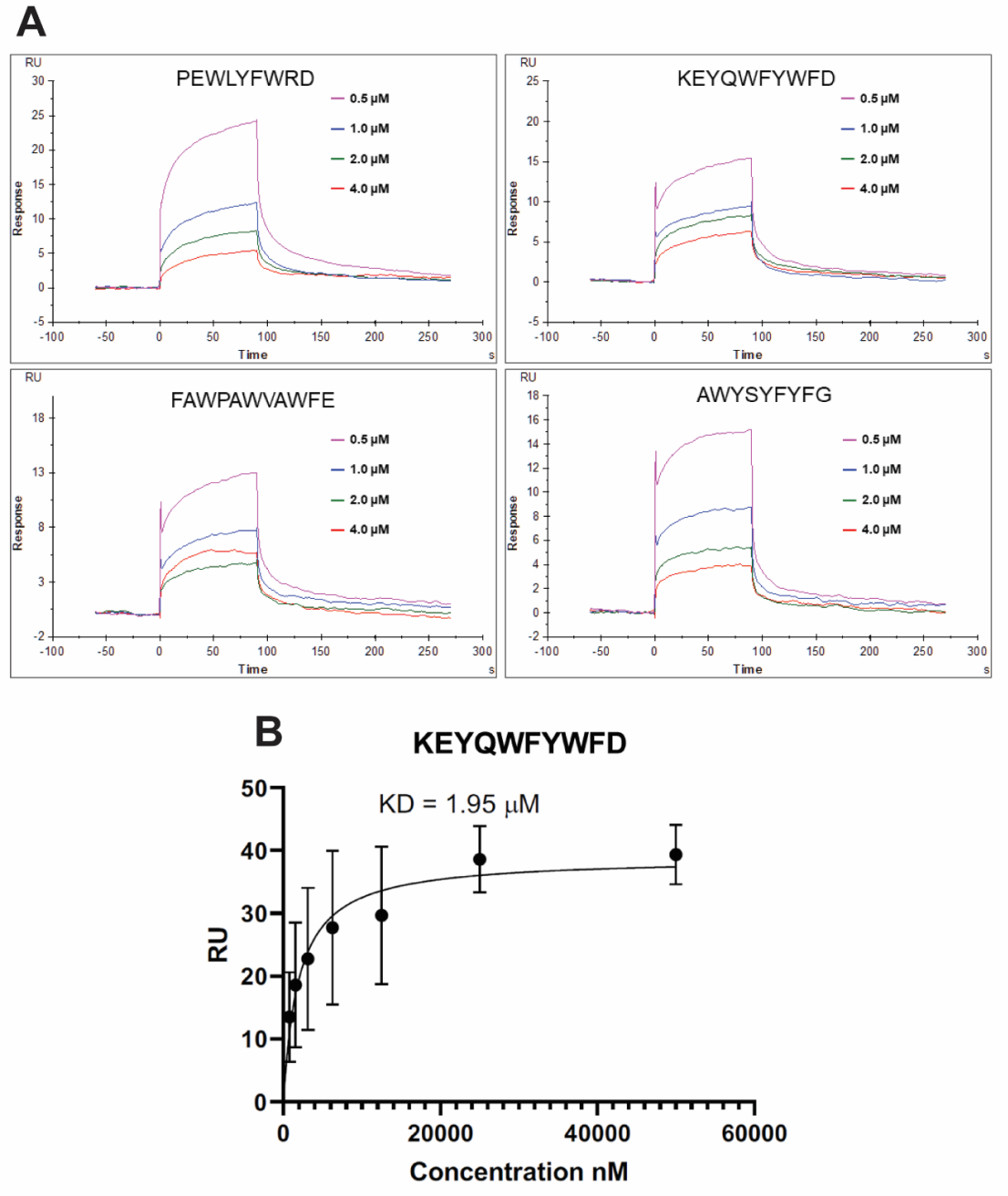


Figure 5.12. (A) Sensorgrams showing responses measured at different concentrations for the shortlisted array peptides binding to diaphorase. (B) Fitting curve for the calculation of dissociation constant for peptide KEYQWFYWFD, from the SPR responses at different concentrations. Experiment carried out in triplicate. RU is relative response unit.

It should be kept in mind here, that not only are the assay conditions different but also the immobilized entities are switched in case of array and SPR. In the arrays, the

peptides were the one that were immobilized whereas in case of SPR, the protein is immobilized on the surface. If the immobilized entities were kept the same in both cases, one might have observed more similar results between SPR and assay experiments. Unfortunately, due to limited supply of the proteins and SPR chips, this option was not experimentally feasible to further exploration.

5.3.5 Effects of Substitutions on Peptide Binding

One of the most compelling benefits of combining computational and experimental approaches, is that they can be used together in an iterative manner to further enhance optimization of binding interactions. This beneficial feature was explored in this study as well, where substituted sequences of two peptides with known binding to diaphorase (seen during screening), were studied computationally as well as experimentally to determine the effects of position specific mutations. The two parent peptide sequences which were studied for this purpose were AYELVLD (peptide 1, also referred to as B6) and QERWFYYEF (peptide 11, also referred to as F12). Firstly, in silico studies were carried out to see how well mutations were tolerated at each position, and what effect they would have on the binding and specificity. The in silico predicted binding for peptide F-12 and its mutagen is shown in Figure 5.13.

Following the results of the in silico studies, arrays were synthesized at HealthTell and assayed for binding to diaphorase. Considering the results from these assays, a handful of peptides were synthesized to be studied using SPR. Not only substituted sequences with enhanced binding activities were selected but sequences with decreased binding activities were selected as well. The list of the synthesized peptides is given Table 5.7. One major experimental change that was attempted while assaying the substituted sequences using

SPR, was immobilizing the peptides on the chip surface, similar to the array experiments. The parent peptides and the substituted sequences were synthesized with polyethylene glycol (PEG) – Lysine – biotin linker and streptavidin capture chips were used from Cytiva (Catalog # BR100531), which contained a dextran matrix pre-functionalized with streptavidin. However, it was realized that diaphorase was interacting too strongly with the surface thus resulting in lower relative responses. Attempts were made to thwart these interactions by adding surfactants (1% bovine serum albumin), but they were of little avail. Therefore, the traditional method where the protein was immobilized on the chip surface was used for further assays.

	Q	E	R	W	F	Y	Y	E	F
A	13563.4	14473.6	11025.75	9580.04	6883.23	7070.7	8115.7	16073.7	7932.74
D	10616.7	11150.5	7433.889	5745.33	4897.4	4634.31	5906.55	8477.75	6726.42
E	11164.2	**12186.6**	7952.059	5995.22	4119.37	4396.03	6696.78	**12186.6**	6445.72
F	16740.3	13386.4	17677.85	15664.1	**12186.6**	14199.5	14485.9	26209.8	**12186.6**
G	13233.4	14042.8	9020.904	7933.27	5721.27	5553.68	6201.63	13886.4	6506.69
H	11479.6	12520.5	11026.1	7236.94	5471.53	4170.21	4967.42	8015.06	7734.03
K	9682.87	12918.7	8256.398	7038.81	5969.29	5393.74	5750.88	21500.9	6248.87
L	14167.8	13485.9	16191.95	12415.2	10406.4	12746.1	13753.3	24057.7	11184.9
N	11978.6	12655.2	8172.631	6367.5	4932.31	4392.01	4968.41	10599.3	5398.79
P	13137.4	12927.7	8429.682	6080.89	4835.89	4522.35	4616.78	12049.1	5499.77
Q	**12186.6**	13248.4	9153.872	7050.26	5176.97	4626.14	5527.27	11370.9	5683.86
R	11115.9	29408.9	**12186.6**	10268.2	11106.8	9504.73	10730.7	38281.3	10002.2
S	12788.4	13845.5	9198.816	8416.91	6168.86	6090.49	6914.49	15147.7	7337.38
V	15263	14472.3	17742.73	13542.9	10624.3	12122	14045.9	23425	12221.2
W	14884.4	15563.2	14313.02	**12186.6**	9125.38	8548.24	8868.3	17351.6	9230.03
Y	16463.2	15116.6	15221.96	14729.9	11143.6	**12186.6**	**12186.6**	23018.9	11251.7

Figure 5.13. Heatmap showing the predicted binding values for the in silico substituted sequences of F12_3 (QERWFYYEF) with position specific mutations. The X-axis is the representation of the original residues. The Y-axis represents the residue introduced through in silico mutation. The predicted binding intensities of the parent peptide are underlined.

Similar to the screening approach described in previous sections, the substituted sequences from Table 5.7 were also screened for binding to diaphorase, however the

peptide concentration was increased to 10 µM. The results of the initial binding screen are shown in Figure 5.14 (A). It can be seen from this figure that the mutagen B6_5 showed a higher binding affinity to diaphorase than the parent peptide AYELVLD (B6_1). B6_2 showed a decreased affinity towards diaphorase. In the case of F12, the substituted sequences showed better binding interactions in the screening than the parent peptide (F12_3). The negative control, F12_4, showed the least affinity for binding as expected.

Table 5.7. Synthesized substituted sequences of peptides B6 and F12

Name	Peptide Sequence	Observed Feature*
B6_1	AYELVLD-PEG_4-Lys(Biotin)	Parent peptide
B6_2	AYEHVLD-PEG_4-Lys(Biotin)	L4H -decreased diaphorase binding
B6_3	QAYELVLDRS-PEG_4-Lys(Biotin)	Extended sequence – increased diaphorase binding
B6_4	ASYKFLLDY-PEG_4-Lys(Biotin)	Extended sequence – increased diaphorase binding
B6_5	YKFLPLRY-PEG_4-Lys(Biotin)	Specific high-affinity binder
B6_6	SYFPLEY-PEG_4-Lys(Biotin)	Non-specific binder
F12_1	QERWFYYEFF-PEG_4-Lys(Biotin)	Extra F at 10 – improves selectivity
F12_2	QERWFHYEFF-PEG_4-Lys(Biotin)	Y6H – improves selectivity towards diaphorase
F12_3	QRWFYYEF-PEG_4-Lys(Biotin)	Parent peptide
F12_4	QERWDHYEFF-PEG_4-Lys(Biotin)	Putative negative control
F12_5	FEYFRFQEWY-PEG_4-Lys(Biotin)	Scrambled sequence of F12_1

* Binding features of each sequence observed during array experiments

After the initial screening, some of the substituted sequences were chosen for evaluation of dissociation constant to determine the strength of their binding to diaphorase. The results shown here are for substituted sequences of F12 (Figure 5.14 B). Substituted variants of the B6 peptides were also tested using SPR for determining their K_D values but reliable K_D measurements were not achieved in their case, very likely due to the higher concentrations that were required for them to attain saturation (~200 µM). At

concentrations this high, peptides are prone to aggregation and are therefore the concentration of the free peptide is difficult to control.

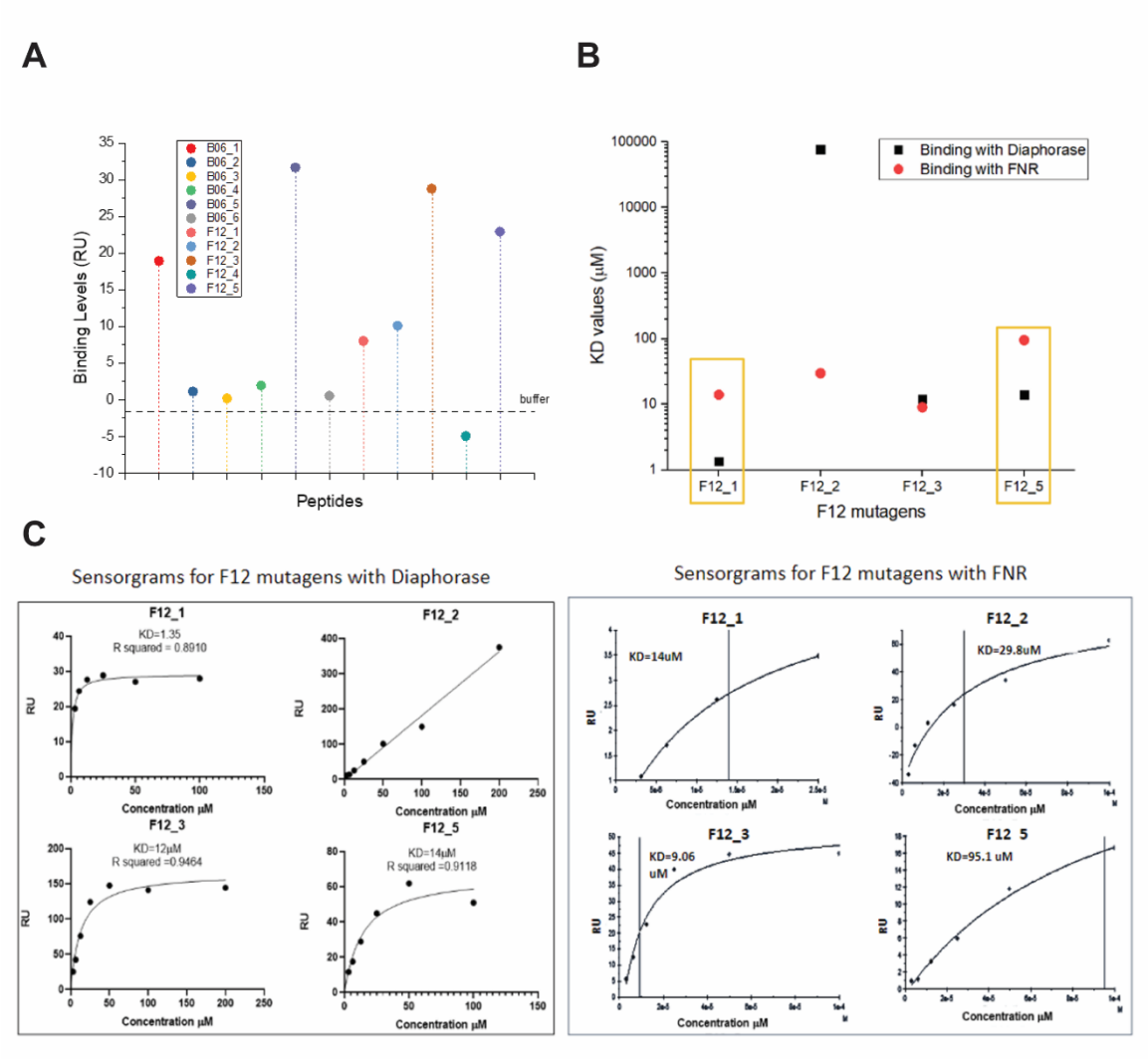


Figure 5.14. (A) SPR Screening results of substituted sequences binding to diaphorase. (B) Calculated K_D values of F12 mutagen peptides for Diaphorase (black) and FNR (red). F12_1 and F12_5 represent more selectivity towards diaphorase (yellow boxes). (C) Plots showing curve fitting for the calculation of dissociation constant for F12 peptides. Plots in the left box represents K_D calculated for binding to diaphorase. Plots on the right represents binding to FNR.

Different concentrations of the peptides (F12_1, F12_2, F12_3, and F12_5) were prepared using serial dilution. K_D was evaluated in a similar manner as laid out before in

section 5.3.4. K_D values were evaluated not only for diaphorase, but for FNR as well to determine the specificity of binding. It was observed that the K_D values, for binding to diaphorase, were lower than 15 μM in case of peptide F12_1, F12_3, and F12_5, while K_D could not be calculated for peptide F12_2 due to poor signal-to-noise ratio (Figure 5.14 B and C). The peptides F12_1 and F12_5 had more binding affinity towards diaphorase than FNR, compared to the other two peptides. F12_2 had higher binding affinity towards FNR and F12_3 had comparable binding affinity towards both the proteins (Figure 5.14 C). These results show that 2 of the substituted sequences of F12 demonstrate an overall enhancement in selectivity towards diaphorase. To summarize the findings of these experiments, it can be concluded that that the results of the initial binding screening do concur with the trends observed from prediction. However, the observed dissociation constants of the peptides were not following the trends observed in the predictions. This shows that the molecular interactions between proteins and peptides is largely dependent on the context of the experiment and findings from any one experiment cannot be used to generalize these interactions.

5.4 DISCUSSION

In this the chapter how the binding affinities of the same protein-peptide interactions differed, based on the context of the interactions and their physical and chemical environment, was probed. To evaluate these interactions the measured (arrays) and predicted sequence vs. binding relationship for diaphorase, FNR, and ferredoxin, as explored by Taguchi et al. (2020), was experimentally probed SPR based binding assays. This was done to explore if there was any plausible relationship between predicted binding trends and observed binding data for protein-peptide interactions. For the evaluation, a set

of 54 peptides were studied. Out of the 54 peptides, 10 sequences were selected from the peptides present on the array, with experimentally recognized binding interactions to diaphorase and the other two proteins. The rest of the sequences (9-mer) were selected from an in silico peptide library which was sampled randomly from the combinatorial sequence space. The predictive sequence vs. binding relationships obtained for the three proteins was projected on to this library and predicted peptide binders with a wide range of binding values (655,361 peptides) were selected by analyzing the projected outcomes. The selected group of peptides were further analyzed for isoelectric point distribution and observed sequence motifs across different binding ranges (Figures 5.3 and 5.4). Upon studying the trends observed in peptides belonging to different binding ranges, 44 peptides were chosen out of the 655,361 sequences to be synthesized for experimental verification.

The results from the following SPR experiments suggests that there is some correlation between observed binding and predicted trends. Sequences with more or less equal number of negatively charged and positively charged residues were predicted to lower binders for diaphorase (Figure 5.4). Consequently ,in the SPR assays also they were found out to be very weak binders to diaphorase (Figure 5.8 and 5.11). Sequences that had a smaller number of polar residues and higher number of hydrophobic or non-polar residues were found to have higher predicted binding intensities. The peptides were screened using SPR with two different protein immobilization method (streptavidin capture and NHS-EDC coupling). It was found that the results of SPR screening were largely dependent on the assay environment and methods of protein immobilization (Figure 5.8). In the SPR experiments, some of the sequences were relatively stronger binders to diaphorase with consistent K_D values below 10 μM. It was found that sequences which had multiple

positively charged residues like R and/or K, were predicted as very high binders. In the SPR experiments as well, their response signals were relatively higher than the rest of the peptides (Figures 5.8 and 5.11). However, not all peptides conform equally to this observed relationship. There were several outliers from all the different binding ranges that showed little correlation between predicted and measured values. This suggests that while the predicted sequence vs. binding relationships broadly matches the observed binding trends for different binding ranges, an absolute linear correlation might be hard to define for individual peptides. Nevertheless, a generalized relationship such as the one observed here, is helpful for picking out suitable binding candidates from a desired range of binding. The binding interactions of the selected candidates can then be optimized using a reiterative approach, employing both computational and experimental tools. The combined reiterative approach mentioned previously, was used in this study, to optimize a peptide with higher diaphorase selectivity than the parent peptide (Figures 5.13 and 5.14).

It also should be pointed out here that the peptides studied here were synthesized from the combinatorial space, based on the results of neural network predictions, with no known biological significance in literature. If the predicted relationships were to be projected on known biologically relevant sequences instead, the observed correlation could be expected to have a more clearly defined relationship.

The binders that were picked from the arrays also behaved differently on the SPR platform. There were four peptides that made through the cut-off of the initial binding (Figure 5.10) screening, reproducible K_D values were obtained only for one of the peptides (KEYQWFYWFD). It is essential to note that in case of SPR, the protein was the entity immobilized on the surface, but on the arrays the peptides were the ones that were fixed.

This causes a reversal of the analytes which are in solution phase. This may have had a role to play for the binding differences observed between array and SPR. The issue could be that in case of high-density peptide microarrays, the peptides are present very close to each other. Thus, these interactions are of different nature than the ones observed SPR. Interestingly, according to previously reported literatures (Katz et al., 2008; Katz et al., 2010, Greving et al., 2008), good correlation was found between array results and SPR studies. As mentioned in the results section (section 5.3.4), attempts were made to immobilize the peptides on the SPR chip, to somewhat emulate the binding environment of the arrays. But this approach could not be pursued further due to constraints in the supply of the proteins.

It was observed that the peptides that interacted well with diaphorase mostly had a lot of aromatic residues (W,F, and Y) in their sequences, indicating that pi-pi interactions might play a crucial role in the binding of these peptides to the surface of diaphorase. The predicted candidate peptide was thoroughly probed in this study (QERWFYYEF) has five aromatic residues (Figure 5.14). Even among the 4 peptides from the array whose K_D values were evaluated, the one that produced the best reproducible results was KEYQWFYWFD which has six aromatic residues (Figure 4.12).

There were some technical issues that were observed while conducting the assays using SPR, such as reference cell binding of the peptides and uneven orientation of the protein on the chip surface, that have contributed to measurement errors. Another challenge that was a potential cause of error in measurement was the solubility of highly hydrophobic peptides. Many of the peptides synthesized for this study had very hydrophobic nature, causing them to aggregate in assay buffer solutions. Such aggregations resulted in

reproducibility errors during the assay. Although this issue was partially resolved by altering the polarity of the buffer solution, it was not completely eliminated.

In conclusion, this chapter looked into the experimental verification of predicted sequence-to-binding relationships and also explored the variations in measurements observed between two different assaying platforms. The results suggest that although there are some parallels between measured and predicted binding, it varies greatly depending on the context. It also explores the possibility of using a combined reiterative methodology to optimize the binding and selectivity of computationally predicted peptide. Additionally, it sheds light on the quantitative differences in measurements between a microarray and an SPR set-up, and challenges that arises with respect to them. For future studies, a larger library, comprising of hundreds of peptides, could be used to derive a more comprehensive predicted vs. measured relationship, when comparing in silico results using SPR. Care should also be taken to develop an effective immobilization strategy with minimal reference binding.

CHAPTER 6

DISCUSSION

Molecular recognition is the main driving force behind all the biomolecular functions in living organisms. In case of biopolymers like proteins, the key to molecular recognition lies in their sequence information. Even though there are many different possible permutation and combinations of sequences for a protein or peptide of a given length, molecular recognition helps a protein to identify the correct set of amino acid residues that are necessary for a biomolecular interaction. Thus, understanding molecular recognition and how it can be defined in terms of sequence space is of fundamental significance in biological sciences. The objective of this dissertation has been to understand the molecular recognition of proteins, especially monoclonal antibodies, in the context of sequence space. This goal was pursued with the help of high-density random sequence peptide microarrays and neural network algorithms.

In order to elucidate a comprehensive relationship between sequence and molecular recognition (binding), monoclonal antibodies were chosen as model systems whose interactions were studied. The high specificity and the well-characterized binding interactions of a monoclonal antibody towards a particular target sequence makes it an ideal candidate for studying molecular recognition. Despite their high specificity, the monoclonal antibodies do show some cross-reactivity with other sequences that are not cognates. The goal of the project was to exploit these different types of binding interactions of the antibodies towards various peptide sequences on a microarray to derive quantitative sequence vs. binding relationships. These relationships could then be extended to any

sequence from the combinatorial space, to predict their binding interaction with the antibodies.

The use of neural networks to predict sequence vs. binding relationship using data from high-throughput random-sequence peptide microarrays was demonstrated by Taguchi et al., (2020). In this work, they illustrate the use of a simple feedforward, backpropagated neural network to describe the sequence vs. binding relationship of nine proteins. However, the molecular interactions of antibodies, especially monoclonal antibodies, differ from other proteins, owing to their affinity and specificity towards their target antigens. Thus, delineating the sequence vs. binding relationships of highly specific interactions such as these is a necessary step towards understanding molecular recognition in terms of sequence space.

In chapter 1, the sequence vs. binding relationships of five monoclonal antibodies (DM1A, p53Ab1, p53Ab8, 4C1, and LNKB2) were explored which have known and well-characterized binding to the sequences present on the peptide microarrays (Legutki et al., 2014; Richer et al., 2015). These antibodies had their cognate sequences represented among the sequence space represented on the microarrays. The five monoclonal antibodies were first assayed on the random-sequence high density peptide microarrays that represented sparsely sampled sequences from the combinatorial sequence space (126,050 unique sequences). Data from these binding assays were then used to train a feedforward, backpropagated neural network model, which was based on the model demonstrated by Taguchi et al., (2020). The cognate sequences of the monoclonal antibodies were deliberately removed from the training set to avoid biasing the algorithm. The hyperparameters of the neural network were optimized to accommodate the binding

interactions observed in case of monoclonal antibodies (Figure 2.3, Table 2.3). After hyperparameter optimization, the neural network was trained 100 times independently for each monoclonal antibody. The comparison between the predicted and measured binding intensities are shown in Figure 2.4. The performances of the models were also validated by projecting the derived sequence vs, binding relationship for each antibody on in silico libraries of randomly sampled peptide sequences from the combinatorial space and analyzing the motif of the highest predicted binders (Tables 2.5 and 2.6, Figure 2.6). From these results it can be seen that the models were accurately able to identify the amino acid residues pertinent to the binding interactions of each monoclonal antibody. The algorithm was successfully able to predict the cognate sequences of all the monoclonal antibodies within the top 0.2% of sequences out of a million peptides. Specificity studies showed that the out of the 126,050 peptide sequences on the arrays, only a few hundred sequences were actually responsible for defining the molecular interactions of these monoclonal antibodies (Figure 2.7). Further analyses (Figure 2.8) indicated that there was a high preference towards some of the cognate residues that were favored by the model in case of each antibody. The residues which were highly favored were resistant towards in silico substitution with other amino acids. It was also found that using a learned encoder for the amino acids resulted in better performance of the algorithm, as opposed to a supplied encoder (Table 2.7). It was observed that using a supplied encoder, such as physicochemical propensities of amino acids, tended to introduce a bias in the algorithm towards favoring residues which were sometimes not relevant to the molecular interactions of the monoclonal antibodies. Contrary to these observations, an encoder which was learned during the training of the algorithm contained no such predisposed bias towards

any particular amino acid, and only favored the residues as learned from the binding interactions observed on the arrays. Additionally, the efficiency of the algorithm when trained on binding data at various concentrations of the monoclonal antibodies was also put to test (Figures 2.9 and 2.10, Tables 2.8 and 2.9). It was shown that the model was able to capture the relevant binding information even at lower concentration of the monoclonal antibodies, provided the binding intensity signals of the peptides were well above the noise from the data. Consecutively, some of the binding data at different concentrations were fitted together to capture an even broader range of interactions for the monoclonal antibodies on the microarray. This approach showed promising results which were comparable to the performance of the model when single antibody concentration was used. Thus, it was established in this chapter that the neural network could successfully predict the molecular interactions of these five monoclonal antibodies, whose binding behavior on the arrays were well-characterized.

The characterizations that were carried out in chapter 2 were evaluated the predictive abilities of the model with respect to the combinatorial chemical space. In chapter 3 the abilities of the algorithm were assessed with regard to the biological sequence space. At first the trained antibody specific models were projected on the sequences of their respective target proteins. It was found that in all of the five cases, the algorithm was able to recognize the cognate regions on the antigen proteins with very high specificity (Figures 3.2, 3.3, 3.4, 3.5, and 3.6). It was also able to identify possible regions on the target proteins that are not exclusive to the cognate region but might be relevant in the interacting with antibodies (Figures 3.7, 3.8, 3.9, 3.10, and 3.11). This was interesting because although these monoclonal antibodies have linear epitopes, their interactions on the protein surface

were indicated to be conformational in nature. However, to make a conclusive claim more structural evidence is needed which can be acquired through experimental methods.

Further, the trained models were projected on the protein sequences from the human proteome (20,361 unique proteins). This was done to see if the models could identify the respective target proteins among the large number of biological sequences represented by the human proteome (Tables 3.2, 3.3, 3.4, 3.5, and 3.6). It was found that the algorithm was able to correctly identify the target proteins of each antibody within the top 10% of the predicted high binders. The model had some challenge in predicting the cognate interactions of the antibody p53Ab1 when projected on the human proteome, because of the valine residues present in its epitope (RHSVV). The valine residues are a commonly occurring hydrophobic motif throughout the proteome and therefore a lot of such similar sequences were presented to the model, therefore, confounding its predictions to a certain extent. However, the model was still able to recognize all the respective target proteins among the top binders among these large pool biological sequences.

The predictive performance of the model so far had been estimated with respect to the antibodies which have known and well-characterized binding to the sequences present on the microarrays. But in order to develop a comprehensive understanding of how sparsely sampled combinatorial peptides can be used to define sequence vs. binding relationship of monoclonal antibodies, one would also need to study the antibodies which do not have well-defined binding to the sequences present on the microarray. This was demonstrated in the chapter 3 of this thesis. Six monoclonal antibodies (3B5, 1D4, 9E10, AU1, Btag, and Htag) whose binding interactions on the microarray were not well-characterized were chosen for this study. Four of these monoclonal antibodies (1D4, 9E10, AU1, and Btag)

have cognate sequences with amino acids that were not represented among the array peptides. It was found that the model was able to predict the sequence vs. binding relationship of one monoclonal antibody (3B5) from the analyzing the binding interactions observed with the array peptides. In this case, the model was accurately able to predict the relevant interacting residues both in the chemical space as well as in the biological space (Figures 4.2, 4.4, and 4.6, Table 4.5). It must be noted that 3B5 did not have any residue in its cognate sequence that were not represented among the array's sequence space. The model failed to predict the interactions of the monoclonal antibodies whose epitopes contained residues that were omitted from the array. A lot of hydrophobic residues were predicted in these cases. It must be mentioned here that the sequence vs. binding relationships that were derived by the algorithm in these cases were representative of the interactions that were observed on the array. So, it can be assumed that, in absence of the residues that were relevant to them, these antibodies bound non-specifically to a lot of other hydrophobic peptide sequences present on the array. Consequently, this also affected the learning of the model and an accurate sequence vs. binding relationship was not obtained in this case. Thus, this study pointed out the limitations of this approach in attempting to define the sequence vs. binding relationship using sparsely sampled combinatorial sequences from the peptide microarray. In future using larger random-sequence peptide arrays with representation of all 20 canonical amino acids would help the neural network achieve better performance with respect to these antibodies.

Following the predictive performance of the neural network with respect to monoclonal antibodies, the changes in sequence vs. binding relationship as observed across different platforms, both predictive and experimental, was also probed. The predicted

protein-peptide binding relationship was compared to the observed binding relationship that was acquired using assay experiments on the arrays and the SPR. The three proteins whose molecular interactions were investigated in this study are diaphorase, FNR, and ferredoxin. The predictive relationship between these proteins and a number of peptides that were randomly sampled in silico was experimentally assessed. It was shown how the molecular interactions were affected by the surface of the assaying platform as well as the chemical state of the analyte (Figure 5.8). This highlighted that fact that context is important in the definition of any molecular interactions Furthermore, the predictions as well as the assay results indicated a hydrophobic sequence motif among the peptides (KEYQWFYWFD, QERWFYYEF) that was favored for binding to diaphorase. Sequence motifs that were predicted for the higher or lower binders, given the amino acid residues, were generally found to match the experimental outcomes, both during rapid binding screens as well as during the measurement of dissociation constant (Figures 5.4, 5.8, 5.9, and 5.10). However, there were technical limitations during this study that did not allow the scope to probe these relationships even further. Further a combination of predictive and experimental approaches was chosen to demonstrate how the binding interactions differ across local sequence space (substitutions of a peptide, Table 5.7) and how it can be used to optimize peptide selectivity (Figures 5.13 and 5.14).

In conclusion, this thesis highlights that it possible to derive comprehensive sequence vs. binding relationship between proteins and peptides by analyzing the interactions of sparsely sampled peptides from the combinatorial chemical space using a neural network. This relationship can then be extended to predict the interactions between the protein and any sequence from the combinatorial space. As of now, the major limitation

of this approach lies in the fact that the arrays only use 16 of the 20 canonical amino acids. Also, the sequence space represented among the arrays (126,050 peptides) is relatively smaller compared to the combinatorial space (10^{12} sequences). If one were to use larger arrays, then possibly more binding information could be gathered from the sparsely sampled random peptides. Furthermore, one needs to consider how molecular interactions vary depending on the physicochemical context of binding and take that into account while developing predictive in silico models. Incorporation of information about conformational sequences will also help strengthen the predictive capabilities of such an algorithm. In future, using such a predictive tool in conjunction with existing experimental methodologies will help identify and discover molecular interactions of proteins and antibodies with greater speed and accuracy. Such a combined approach would prove beneficial in the field of diagnostics and therapeutics research.

REFERENCES

Afonin, P. V., Fokin, A. V., Tsygannik, I. N., Mikhailova, I. Y. U., Onoprienko, L. V., Mikhaleva, I. I., Ivanov, V. T., Mareeva, T. y. Y. U., Nesmeyanov, V. A., & Li, N. (2001). Crystal structure of an anti-interleukin-2 monoclonal antibody Fab complexed with an antigenic nonapeptide. *Protein Science*, *10*(8), 1514-1521.

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol*, *33*(8), 831-838.

Andresen, H., Zarse, K., Grotzinger, C., Hollidt, J. M., Ehrentreich-Forster, E., Bier, F. F., & Kreuzer, O. J. (2006). Development of peptide microarrays for epitope mapping of antibodies against the human TSH receptor. *J Immunol Methods*, *315*(1-2), 11-18.

Ansar, W., & Ghosh, S. (2013a). Monoclonal Antibodies: A Tool in Clinical Research. *Indian Journal of Clinical Medicine*, *4*.

Ansar, W., & Ghosh, S. (2013b). Monoclonal antibodies: a tool in clinical research. *Indian Journal of Clinical Medicine*, *4*, IJCM-S11968.

Ansari, H. R., & Raghava, G. P. S. (2013). In silico models for B-cell epitope recognition and signaling. *In silico Models for Drug Discovery*, 129-138.

Arkin, M. R., Randal, M., DeLano, W. L., Hyde, J., Luong, T. N., Oslob, J. D., Raphael, D. R., Taylor, L., Wang, J., & McDowell, R. S. (2003). Binding of small molecules to an adaptive protein–protein interface. *Proceedings of the National Academy of Sciences*, *100*(4), 1603-1608.

Bailey, T. L. (2021). STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, *37*(18), 2834-2840.

Baker, C. C., & Howley, P. M. (1987). Differential promoter utilization by the bovine papillomavirus in transformed cells and productively infected wart tissues. *The EMBO journal*, *6*(4), 1027-1035.

Barlow, D. J., Edwards, M. S., & Thornton, J. M. (1986). Continuous and discontinuous protein antigenic determinants. *Nature*, *322*(6081), 747-748.

Beck, A., Haeuw, J.-F., Wurch, T., Goetsch, L., Bailly, C., & Corvaïa, N. (2010). The next generation of antibody-drug conjugates comes of age. *Discovery medicine*, *10*(53), 329-339.

Beiko, R. G., Chan, C. X., & Ragan, M. A. (2005). A word-oriented approach to alignment validation. *Bioinformatics*, *21*(10), 2230-2239.

Benyamini, H., & Friedler, A. (2010). Using peptides to study protein–protein interactions. *Future Medicinal Chemistry*, *2*(6), 989-1003.

Berggard, T., Linse, S., & James, P. (2007). Methods for the detection and analysis of protein-protein interactions. *Proteomics*, *7*(16), 2833-2842.

Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H., & Westbrook, J. (2000). The Protein Data Bank and the challenge of structural genomics. *Nature structural biology*, *7*(11), 957-959.

Blythe, M. J., & Flower, D. R. (2005). Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Science*, *14*(1), 246-248.

Bradbury, A. R. M., Trinklein, N. D., Thie, H., Wilkinson, I. C., Tandon, A. K., Anderson, S., Bladen, C. L., Jones, B., Aldred, S. F., & Bestagno, M. (2018). When monoclonal antibodies are not monospecific: hybridomas frequently express additional functional variable regions.

Breitling, F., & Little, M. (1986). Carboxy-terminal regions on the surface of tubulin and microtubules epitope locations of YOL1/34, DM1A and DM1B. *Journal of Molecular Biology*, *189*(2), 367-370.

Brennan, F. R., Morton, L. D., Spindeldreher, S., Kiessling, A., Allenspach, R., Hey, A., Müller, P., Frings, W., & Sims, J. (2010). Safety and immunotoxicity assessment of immunomodulatory monoclonal antibodies.

Budach, S., & Marsico, A. (2018). pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks. *Bioinformatics*, *34*(17), 3035-3037.

Burnet, F. M. (1957). A modification of Jerne's theory of antibody production using the concept of clonal selection. *Australian Journal of Science*, *20*(3), 67-69.

Cafarelli, T. M., Desbuleux, A., Wang, Y., Choi, S. G., De Ridder, D., & Vidal, M. (2017). Mapping, modeling, and characterization of protein-protein interactions on a proteomic scale. *Curr Opin Struct Biol*, *44*, 201-210.

Chen, E. Y., Howley, P. M., Levinson, A. D., & Seeburg, P. H. (1982). The primary structure and genetic organization of the bovine papillomavirus type 1 genome. *Nature*, *299*(5883), 529-534.

Chen, J., Liu, H., Yang, J., & Chou, K. C. (2007). Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids*, *33*(3), 423-428.

Chen, W., Guo, W. W., Huang, Y., & Ma, Z. (2012). PepMapper: a collaborative web tool for mapping epitopes from affinity-selected peptides. *PLoS ONE*, *7*(5), e37869.

Chen, W. H., Sun, P. P., Lu, Y., Guo, W. W., Huang, Y. X., & Ma, Z. Q. (2011). MimoPro: a more efficient Web-based tool for epitope prediction using phage display libraries. *BMC bioinformatics*, *12*(1), 1-13.

Chockalingam, K., Blenner, M., & Banta, S. (2007). Design and application of stimulus-responsive peptide systems. *Protein Engineering, Design & Selection*, *20*(4), 155-161.

Chou, M. F., & Schwartz, D. (2011). Biological sequence motif discovery using motif-x. *Current protocols in bioinformatics*, *35*(1), 13-15.

Clarke, B. (1970). Darwinian evolution of proteins. *Science*, *168*(3934), 1009-1011.

Colwell, L. J. (2018). Statistical and machine learning approaches to predicting protein-ligand interactions. *Curr Opin Struct Biol*, *49*, 123-128.

Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A., & DeLisi, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *Journal of Molecular Biology*, *195*(3), 659-685.

Crooks, G. E., Hon, G., Chandonia, J.-M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome research*, *14*(6), 1188-1190.

Cruciani, G., Baroni, M., Carosati, E., Clementi, M., Valigi, R., & Clementi, S. (2004). Peptide studies by means of principal properties of amino acids derived from MIF descriptors. *Journal of Chemometrics*, *18*(34), 146-155.

da Silva, A. D., Bitencourt-Ferreira, G., & de Azevedo Jr, W. F. (2020). Taba: A tool to analyze the binding affinity. *Journal of Computational Chemistry*, *41*(1), 69-73.

Davey, N. E., Shields, D. C., & Edwards, R. J. (2006). SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. *Nucleic Acids Research*, *34*(12), 3546-3554.

de Juan, D., Pazos, F., & Valencia, A. (2013). Emerging methods in protein co-evolution. *Nat Rev Genet*, *14*(4), 249-261.

Deng, J., Yang, Z., Ojima, I., Samaras, D., & Wang, F. (2022). Artificial intelligence in drug discovery: applications and techniques. *Brief Bioinform*, *23*(1).

Dighiero, G., Lymberi, P., Mazie, J. C., Rouyre, S., Butler-Browne, G. S., Whalen, R. G., & Avrameas, S. (1983). Murine hybridomas secreting natural monoclonal antibodies reacting with self antigens. *The Journal of Immunology*, *131*(5), 2267-2272.

Disfani, F. M., Hsu, W. L., Mizianty, M. J., Oldfield, C. J., Xue, B., Dunker, A. K., Uversky, V. N., & Kurgan, L. (2012). MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics*, *28*(12), i75-83.

Dryden, D. T., Thomson, A. R., & White, J. H. (2008). How much of protein sequence space has been explored by life on Earth? *J R Soc Interface*, *5*(25), 953-956.

Eisenberg, D. (1984). Three-dimensional structure of membrane and surface proteins. *Annual review of biochemistry*, *53*(1), 595-623.

El-Manzalawy, Y., & Honavar, V. (2010). Recent advances in B-cell epitope prediction methods. *Immunome research*, *6*(2), 1-9.

Emamzadah, S., Tropia, L., Vincenti, I., Falquet, B., & Halazonetis, T. D. (2014). Reversal of the DNA-binding-induced loop L1 conformational switch in an engineered human p53 protein. *Journal of Molecular Biology*, *426*(4), 936-944.

Engelman, D. M., Steitz, T. A., & Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annual review of biophysics and biophysical chemistry*, *15*(1), 321-353.

Fibriansah, G., Ibarra, K. D., Ng, T.-S., Smith, S. A., Tan, J. L., Lim, X.-N., Ooi, J. S. G., Kostyuchenko, V. A., Wang, J., & de Silva, A. M. (2015). Cryo-EM structure of an antibody that neutralizes dengue virus type 2 by locking E protein dimers. *Science*, *349*(6243), 88-91.

Flores-Moreno, K., Celis-Meneses, J. S., Meneses-Ruiz, D. M., Castillo-Rodal, A. I., Orduña, P., Montiel, B. A., & López-Vidal, Y. (2014). Potential cross-reactivity of monoclonal antibodies against clinically relevant mycobacteria. *Clinical and experimental immunology*, *177*(2), 454-463.

Fout, A., Byrd, J., Shariat, B., & Ben-Hur, A. (2017). Protein interface prediction using graph convolutional networks. *Advances in neural information processing systems*, *30*.

Frank, S. A. (2002). Specificity and cross-reactivity. In *Immunology and evolution of infectious disease*. Princeton University Press.

Freschlin, C. R., Fahlberg, S. A., & Romero, P. A. (2022). Machine learning to navigate fitness landscapes for protein engineering. *Curr Opin Biotechnol, 75*, 102713.

Gabernet, G., Gautschi, D., Müller, A. T., Neuhaus, C. S., Armbrecht, L., Dittrich, P. S., Hiss, J. A., & Schneider, G. (2019). In silico design and optimization of selective membranolytic anticancer peptides. *Scientific reports*, *9*(1), 1-11.

Gao, J., Faraggi, E., Zhou, Y., Ruan, J., & Kurgan, L. (2012). BEST: improved prediction of B-cell epitopes from antigen sequences. *PLoS ONE, 7*(6), e40104.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., & Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research, 31*(13), 3784-3788.

Gellman, S. H. (1997). Introduction: Molecular Recognition. *Chemical Reviews*, *97*(5), 1231-1232.

Georgiev, A. G. (2009). Interpretable numerical descriptors of amino acid space. *J Comput Biol, 16*(5), 703-723.

Getzoff, E. D., Tainer, J. A., Lerner, R. A., & Geysen, H. M. (1988). The chemistry and mechanism of antibody binding to protein antigens. *Advances in immunology*, *43*, 1-98.

Geysen, H. M., Rodda, S. J., & Mason, T. J. (1986). A priori delineation of a peptide which mimics a discontinuous antigenic determinant. *Molecular Immunology, 23*(7), 709-715.

Gligorijevic, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B. C., Fisk, I. M., Vlamakis, H., Xavier, R. J., Knight, R., Cho, K., & Bonneau, R. (2021). Structure-based protein function prediction using graph convolutional networks. *Nat Commun*, *12*(1), 3168.

Gosline, J., Lillie, M., Carrington, E., Guerette, P., Ortlepp, C., & Savage, K. (2002). Elastic proteins: biological roles and mechanical properties. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *357*(1418), 121-132.

Graves, J., Byerly, J., Priego, E., Makkapati, N., Parish, S. V., Medellin, B., & Berrondo, M. (2020). A Review of Deep Learning Methods for Antibodies. *Antibodies (Basel), 9*(2).

Gunasekaran, K., Tsai, C. J., & Nussinov, R. (2004). Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J Mol Biol*, *341*(5), 1327-1341.

Gustafsson, C., Govindarajan, S., & Emig, R. (2001). Exploration of sequence space for protein engineering. *J Mol Recognit*, *14*(5), 308-314.

Halperin, R. F., Stafford, P., & Johnston, S. A. (2011). Exploring antibody recognition of sequence space through random-sequence peptide microarrays. *Molecular & Cellular Proteomics*, *10*(3).

Haspel, M. V., Onodera, T., Prabhakar, B. S., McClintock, P. R., Essani, K., Ray, U. R., Yagihashi, S., & Notkins, A. L. (1983). Multiple organ-reactive monoclonal autoantibodies. *Nature*, *304*(5921), 73-76.

Heiss, K., Heidepriem, J., Fischer, N., Weber, L. K., Dahlke, C., Jaenisch, T., & Loeffler, F. F. (2020). Rapid response to pandemic threats: immunogenic epitope detection of pandemic pathogens for diagnostics and vaccine development using peptide microarrays. *Journal of proteome research*, *19*(11), 4339-4354.

Hochuli, E., Bannwarth, W., Döbeli, H., Gentz, R., & Stüber, D. (1988). Genetic approach to facilitate purification of recombinant proteins with a novel metal chelate adsorbent. *Bio/technology*, *6*(11), 1321-1325.

Hodges, R. S., Heaton, R. J., Parker, J. M., Molday, L., & Molday, R. S. (1988). Antigen-antibody interaction. Synthetic peptides define linear antigenic determinants recognized by monoclonal antibodies directed to the cytoplasmic carboxyl terminus of rhodopsin. *Journal of Biological Chemistry*, *263*(24), 11768-11775.

Hopp, T. P., & Woods, K. R. (1981). Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences*, *78*(6), 3824-3828.

Hua, C. K., Gacerez, A. T., Sentman, C. L., Ackerman, M. E., Choi, Y., & Bailey-Kellogg, C. (2017). Computationally-driven identification of antibody epitopes. *Elife*, *6*, e29023.

Huang, R. Y. C., & Chen, G. (2014). Higher order structure characterization of protein therapeutics by hydrogen/deuterium exchange mass spectrometry. *Analytical and bioanalytical chemistry*, *406*(26), 6541-6558.

Huang, Y. X., Bao, Y. L., Guo, S. Y., Wang, Y., Zhou, C. G., & Li, Y. X. (2008). Pep-3D-Search: a method for B-cell epitope prediction based on mimotope analysis. *BMC bioinformatics*, *9*(1), 1-17.

Hughes, A. K., Cichacz, Z., Scheck, A., Coons, S. W., Johnston, S. A., & Stafford, P. (2012). Immunosignaturing can detect products from molecular markers in brain cancer. *PLoS ONE, 7*(7), e40201.

Jackups, R., Jr., & Liang, J. (2006). Combinatorial model for sequence and spatial motif discovery in short sequence fragments: examples from beta-barrel membrane proteins. *Conf Proc IEEE Eng Med Biol Soc, 2006*, 3470-3473.

James, L. C., & Tawfik, D. S. (2003). The specificity of cross-reactivity: Promiscuous antibody binding involves specific hydrogen bonds rather than nonspecific hydrophobic stickiness

James Leo, C., Roversi, P., & Tawfik Dan, S. (2003). Antibody Multispecificity Mediated by Conformational Diversity. Science, 299(5611), 1362-1367.

Jansen, S. R., Holman, R., Hedemann, I., Frankes, E., Elzinga, C. R. S., Timens, W., Gosens, R., de Bont, E. S., & Schmidt, M. (2015). Prostaglandin E2 promotes MYCN non-amplified neuroblastoma cell survival via β-catenin stabilization. *Journal of cellular and molecular medicine, 19*(1), 210-226.

Jenson, A. B., Jenson, M. C., Cowsert, L., Ghim, S.-J., & Sundberg, J. P. (1997). Multiplicity of uses of monoclonal antibodies that define papillomavirus linear immunodominant epitopes. *Immunologic Research, 16*(1), 115-119.

Johansson-Åkhe, I., Mirabello, C., & Wallner, B. (2019). Predicting protein-peptide interaction sites using distant protein complexes as structural templates. *Scientific reports, 9*(1), 1-13.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., & Potapenko, A. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583-589.

Kang, Y., Zhou, X. E., Gao, X., He, Y., Liu, W., Ishchenko, A., Barty, A., White, T. A., Yefanov, O., Han, G. W., Xu, Q., de Waal, P. W., Ke, J., Tan, M. H. E., Zhang, C., Moeller, A., West, G. M., Pascal, B. D., Van Eps, N., . . . Xu, H. E. (2015). Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser. *Nature, 523*(7562), 561-567.

Kaplon, H., Muralidharan, M., Schneider, Z., & Reichert, J. M. (2020). Antibodies to watch in 2020.

Katz, C., Benyamini, H., Rotem, S., Lebendiker, M., Danieli, T., Iosub, A., Refaely, H., Dines, M., Bronner, V., Bravman, T., Shalev Deborah, E., Rüdiger, S., & Friedler, A. (2008). Molecular basis of the interaction between the antiapoptotic Bcl-2

family proteins and the proapoptotic protein ASPP2. *Proceedings of the National Academy of Sciences, 105*(34), 12277-12282.

Katz, C., Levy-Beladev, L., Rotem-Bamberger, S., Rito, T., Rüdiger, S. G. D., & Friedler, A. (2011). Studying protein–protein interactions using peptide arrays. *Chemical Society Reviews, 40*(5), 2131-2145.

Kidera, A., Konishi, Y., Oka, M., Ooi, T., & Scheraga, H. A. (1985). Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *Journal of Protein Chemistry, 4*(1), 23-55.

Köhler, G., & Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature, 256*(5517), 495-497.

Kolaskar, A. S., & Tongaonkar, P. C. (1990). A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS letters, 276*(1-2), 172-174.

Kowalsky, C. A., Faber, M. S., Nath, A., Dann, H. E., Kelly, V. W., Liu, L., Shanker, P., Wagner, E. K., Maynard, J. A., & Chan, C. (2015). Rapid fine conformational epitope mapping using comprehensive mutagenesis and deep sequencing. *Journal of Biological Chemistry, 290*(44), 26457-26470.

Krawczyk, K., Dunbar, J., & Deane, C. M. (2017). Computational tools for aiding rational antibody design. In *Computational Protein Design* (pp. 399-416). Springer.

Kricka, L. J. (1988). Molecular and ionic recognition by biological systems. In T. E. Edmonds (Ed.), *Chemical Sensors* (pp. 3-14). Springer Netherlands.

Kukreja, M., Johnston, S., & Stafford, P. (2012). Immunosignaturing microarrays distinguish antibody profiles of related pancreatic diseases. *J. Proteomics Bioinformatics S, 6*(001), 10-4172.

Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology, 157*(1), 105-132.

Legutki, J. B., Magee, D. M., Stafford, P., & Johnston, S. A. (2010). A general method for characterization of humoral immunity induced by a vaccine or infection. *Vaccine, 28*(28), 4529-4537.

Legutki, J. B., Zhao, Z.-G., Greving, M., Woodbury, N., Johnston, S. A., & Stafford, P. (2014). Scalable high-density peptide arrays for comprehensive health monitoring. *Nature Communications, 5*(1), 4785.

Li, X., Kazan, H., Lipshitz, H. D., & Morris, Q. D. (2014). Finding the target sites of RNA-binding proteins. *Wiley Interdiscip Rev RNA*, *5*(1), 111-130.

Liang, C. T. H., Li, L. C., & Kim, B. S. K. (2004). The Asian American Racism-Related Stress Inventory: Development, Factor Analysis, Reliability, and Validity. *Journal of Counseling Psychology*, *51*(1), 103-114.

Liang, G., Chen, G., Niu, W., & Li, Z. (2008). Factor analysis scales of generalized amino acid information as applied in predicting interactions between the human amphiphysin-1 SH3 domains and their peptide ligands. *Chem Biol Drug Des*, *71*(4), 345-351.

Lim, Y. W., Adler, A. S., & Johnson, D. S. (2022). Predicting antibody binders and generating synthetic antibodies using deep learning. *MAbs*, *14*(1), 2069075.

Lippow, S. M., Wittrup, K. D., & Tidor, B. (2007). Computational design of antibody-affinity improvement beyond in vivo maturation. *Nature Biotechnology*, *25*(10), 1171-1176.

Lord, J. M., Bunce, C. M., & Brown, G. (1988). The role of protein phosphorylation in the control of cell growth and differentiation. *British journal of cancer*, *58*(5), 549.

Madeira, F., Pearce, M., Tivey, A., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., & Lopez, R. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Research*.

Mahmuda, A., Bande, F., Kadhim Al-Zihiry, K. J., Abdulhaleem, N., Majid, R. A., Hamat, R. A., Abdullah, W. O., & Unyah, Z. (2017). Monoclonal antibodies: A review of therapeutic applications and future prospects. *Tropical Journal of Pharmaceutical Research*, *16*(3).

Manieri, T. M., Magalhaes, C. G., Takata, D. Y., Batalha-Carvalho, J. V., & Moro, A. M. (2020). In silico Techniques for Prospecting and Characterizing Monoclonal Antibodies. In *Monoclonal Antibodies*. IntechOpen.

Maynard Smith, J. (1970). Natural selection and the concept of a protein space. *Nature*, *225*(5232), 563-564.

Mei, H., Liao, Z. H., Zhou, Y., & Li, S. Z. (2005). A new set of amino acid descriptors and its application in peptide QSARs. *Biopolymers*, *80*(6), 775-786.

Meng, X., Wei, J., Wang, Y., Zhang, H., & Wang, Z. (2018). The role of peptide microarrays in biomedical research. *Analytical Methods*, *10*(38), 4614-4624.

Mimmi, S., Maisano, D., Quinto, I., & Iaccino, E. (2019). Phage display: an overview in context to drug discovery. *Trends in pharmacological sciences*, *40*(2), 87-91.

Mirny, L. A., Abkevich, V. I., & Shakhnovich, E. I. (1998). How evolution makes proteins fold quickly. *Proceedings of the National Academy of Sciences*, *95*(9), 4976-4981.

Miura, K. (2018). An Overview of Current Methods to Confirm Protein-Protein Interactions. *Protein Pept Lett*, *25*(8), 728-733.

Mohamed, S. A. E. H., Elloumi, M., & Thompson, J. D. (2016). Motif discovery in protein sequences. *Pattern Recognition-Analysis and Applications*.

Morrone Xavier, M., Sehnem Heck, G., Boff de Avila, M., Maria Bernhardt Levin, N., Oliveira Pintro, V., Lemes Carvalho, N., & Filgueira de Azevedo, W. (2016). SAnDReS a computational tool for statistical analysis of docking results and development of scoring functions. *Combinatorial chemistry & high throughput screening*, *19*(10), 801-812.

Murphy, K., & Weaver, C. (2016). *Janeway's immunobiology*. Garland science.

Nealon, J. O., Philomina, L. S., & McGuffin, L. J. (2017). Predictive and Experimental Approaches for Elucidating Protein-Protein Interactions and Quaternary Structures. *International journal of molecular sciences*, *18*(12), 2623.

Nelson, A. L., Dhimolea, E., & Reichert, J. M. (2010). Development trends for human monoclonal antibody therapeutics. *Nature reviews drug discovery*, *9*(10), 767-774.

Nelson, P., Reynolds, G., Waldron, E., Ward, E., Giannopoulos, K., & Murray, P. (2000). Demystified…: monoclonal antibodies. *Molecular pathology*, *53*(3), 111.

Nguyen, H. H., Park, J., Kang, S., & Kim, M. (2015). Surface plasmon resonance: a versatile technique for biosensor applications. *Sensors (Basel)*, *15*(5), 10481-10510.

Nomikou, K., Hughes, J., Wash, R., Kellam, P., Breard, E., Zientara, S., Palmarini, M., Biek, R., & Mertens, P. (2015). Widespread Reassortment Shapes the Evolution and Epidemiology of Bluetongue Virus following European Invasion. *PLoS pathogens*, *11*(8), e1005056-e1005056.

Norman, R. A., Ambrosetti, F., Bonvin, A. M. J. J., Colwell, L. J., Kelm, S., Kumar, S., & Krawczyk, K. (2020). Computational approaches to therapeutic antibody design: established methods and emerging trends. *Briefings in bioinformatics*, *21*(5), 1549-1567.

Notkins, A. L. (2004). Polyreactivity of antibody molecules. *Trends in immunology*, *25*(4), 174-179.

Odorico, M., & Pellequer, J. L. (2003). BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. *Journal of Molecular Recognition*, *16*(1), 20-22.

Papadopoulos, J. S., & Agarwala, R. (2007). COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, *23*(9), 1073-1079.

Part II: Molecular Recognition in Protein Assay. (2019). In G. Li (Ed.), *Nano-Inspired Biosensors for Protein Assay with Clinical Applications* (pp. 113-114). Elsevier.

Pellequer, J.-L., Westhof, E., & Van Regenmortel, M. H. V. (1993). Correlation between the location of antigenic sites and the prediction of turns in proteins. *Immunology letters*, *36*(1), 83-99.

Pellequer, J. L., & Westhof, E. (1993). PREDITOP: a program for antigenicity prediction. *Journal of molecular graphics*, *11*(3), 204-210.

Peng, H.-P., Lee, K. H., Jian, J.-W., & Yang, A.-S. (2014). Origins of specificity and affinity in antibody–protein interactions. *Proceedings of the National Academy of Sciences*, *111*(26), E2656-E2665.

Persch, E., Dumele, O., & Diederich, F. (2015). Molecular recognition in chemical and biological systems. *Angew Chem Int Ed Engl*, *54*(11), 3290-3327.

Petersen, G., Song, D., Hügle-Dörr, B., Oldenburg, I., & Bautz, E. K. F. (1995). Mapping of linear epitopes recognized by monoclonal antibodies with gene-fragment phage display libraries. *Molecular and General Genetics MGG*, *249*(4), 425-431.

Potocnakova, L., Bhide, M., & Pulzova, L. B. (2016). An Introduction to B-Cell Epitope Mapping and In silico Epitope Prediction. *J Immunol Res*, *2016*, 6760830.

Povolotskaya, I. S., & Kondrashov, F. A. (2010). Sequence space and the ongoing expansion of the protein universe. *Nature*, *465*(7300), 922-926.

Rebek, J. (2009). Introduction to the Molecular Recognition and Self-Assembly Special Feature. *Proceedings of the National Academy of Sciences*, *106*(26), 10423-10424.

Restrepo, L., Stafford, P., Magee, D. M., & Johnston, S. A. (2011). Application of immunosignatures to the assessment of Alzheimer's disease. *Annals of neurology*, *70*(2), 286-295.

Reverberi, R., & Reverberi, L. (2007). Factors affecting the antigen-antibody reaction. *Blood transfusion = Trasfusione del sangue*, *5*(4), 227-240.

Richer, J., Johnston, S. A., & Stafford, P. (2015). Epitope Identification from Fixed-complexity Random-sequence Peptide Microarrays. *Molecular & Cellular Proteomics*, *14*(1), 136-147.

Righetti, P. G. (2004). Determination of the isoelectric point of proteins by capillary isoelectric focusing. *J Chromatogr A*, *1037*(1-2), 491-499.

Rockberg, J., Löfblom, J., Hjelm, B., Uhlén, M., & Ståhl, S. (2008). Epitope mapping of antibodies using bacterial surface display. *Nature Methods*, *5*(12), 1039-1045.

Roggen, E. L. (2006). Recent developments with B-cell epitope identification for predictive studies. *J Immunotoxicol*, *3*(3), 137-149.

Roggen, E. L. (2008). B-cell epitope engineering: A matter of recognizing protein features and motives. *Drug Discovery Today: Technologies*, *5*(2-3), e49-e55.

Rose, G. D., & Wolfenden, R. (1993). Hydrogen bonding, hydrophobicity, packing, and protein folding. *Annual review of biophysics and biomolecular structure*, *22*(1), 381-415.

Rose, S., & Stevens, A. (2003). Computational design strategies for combinatorial libraries. *Current Opinion in Chemical Biology*, *7*(3), 331-339.

Rotem, S., Katz, C., Benyamini, H., Lebendiker, M., Veprintsev, D., Rüdiger, S., Danieli, T., & Friedler, A. (2008). The Structure and Interactions of the Proline-rich Domain of ASPP2. *Journal of Biological Chemistry*, *283*(27), 18990-18999.

Rowe, M., Melnick, J., Gerwien, R., Legutki, J. B., Pfeilsticker, J., Tarasow, T. M., & Sykes, K. F. (2017). An ImmunoSignature test distinguishes Trypanosoma cruzi, hepatitis B, hepatitis C and West Nile virus seropositivity among asymptomatic blood donors. *PLOS Neglected Tropical Diseases*, *11*(9), e0005882.

Russell, R. B., Alber, F., Aloy, P., Davis, F. P., Korkin, D., Pichaud, M., Topf, M., & Sali, A. (2004). A structural perspective on protein-protein interactions. *Curr Opin Struct Biol*, *14*(3), 313-324.

Sanchez-Trincado, J. L., Gomez-Perosanz, M., & Reche, P. A. (2017). Fundamentals and methods for T-and B-cell epitope prediction. *Journal of immunology research*, *2017*.

Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., & Wold, S. (1998). New chemical descriptors relevant for the design of biologically active peptides. A multivariate

characterization of 87 amino acids. *Journal of medicinal chemistry*, *41*(14), 2481-2491.

Sautto, G., Mancini, N., Gorini, G., Clementi, M., & Burioni, R. (2013). Possible future monoclonal antibody (mAb)-based therapy against arbovirus infections. *BioMed Research International*, *2013*.

Schissel, C. K., Mohapatra, S., Wolfe, J. M., Fadzen, C. M., Bellovoda, K., Wu, C.-L., Wood, J. A., Malmberg, A. B., Loas, A., Gómez-Bombarelli, R., & Pentelute, B. L. (2020).

Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, *18*(20), 6097-6100.

Schüchner, S., Behm, C., Mudrak, I., & Ogris, E. (2020). The Myc tag monoclonal antibody 9E10 displays highly variable epitope recognition dependent on neighboring sequence context. *Science signaling*, *13*(616)

Schweitzer, B., Predki, P., & Snyder, M. (2003). Microarrays to characterize protein interactions on a whole-proteome scale. *Proteomics*, *3*(11), 2190-2199.

Sette, A., & Fikes, J. (2003). Epitope-based vaccines: an update on epitope identification, vaccine design and delivery. *Current Opinion in Immunology*, *15*(4), 461-470.

Shen, C., Ding, J., Wang, Z., Cao, D., Ding, X., & Hou, T. (2019). From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. *WIREs Computational Molecular Science*, *10*(1).

Shepherd, P. S., Da Costa, C. R., Cridland, J. C., Gilmore, K. S., & Johnstone, A. P. (1999). Identification of an important thyrotrophin binding site on the human thyrotrophin receptor using monoclonal antibodies. *Molecular and Cellular Endocrinology*, *149*(1-2), 197-206.

Sher, G., Zhi, D., & Zhang, S. (2017). DRREP: deep ridge regressed epitope predictor. *BMC genomics*, *18*(6), 55-65.

Shi, S.-R., Prince, J. B., Jones, C. M., Kalra, K. L., & Tandon, A. K. (1995). Use of monoclonal antibodies in immunohistochemistry. In *Monoclonal Antibody Protocols* (pp. 89-108). Springer.

Shirai, H., Prades, C., Vita, R., Marcatili, P., Popovic, B., Xu, J., Overington, J. P., Hirayama, K., Soga, S., & Tsunoyama, K. (2014). Antibody informatics for drug discovery. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1844*(11), 2002-2015.

Sidhu, S. S., Fairbrother, W. J., & Deshayes, K. (2003). Exploring protein–protein interactions with phage display. *Chembiochem*, *4*(1), 14-25.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., & Söding, J. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, *7*(1), 539.

Singh, H., Ansari, H. R., & Raghava, G. P. S. (2013). Improved method for linear B-cell epitope prediction using antigen's primary sequence. *PLoS ONE*, *8*(5), e62216.

Smith, K., Garman, L., Wrammert, J., Zheng, N.-Y., Capra, J. D., Ahmed, R., & Wilson, P. C. (2009). Rapid generation of fully human monoclonal antibodies specific to a vaccinating antigen. *Nature Protocols*, *4*(3), 372-384.

Sollner, J., & Mayer, B. (2006). Machine learning approaches for prediction of linear B-cell epitopes on proteins. *J Mol Recognit*, *19*(3), 200-208.

Song, J., Zheng, Y., Huang, M., Wu, L., Wang, W., Zhu, Z., Song, Y., & Yang, C. (2020). A Sequential Multidimensional Analysis Algorithm for Aptamer Identification based on Structure Analysis and Machine Learning. *Anal Chem*, *92*(4), 3307-3314.

Sonsare, P. M., & Gunavathi, C. (2019). Investigation of machine learning techniques on proteomics: A comprehensive survey. *Prog Biophys Mol Biol*, *149*, 54-69.

Stadler, C., Rexhepaj, E., Singan, V. R., Murphy, R. F., Pepperkok, R., Uhlén, M., Simpson, J. C., & Lundberg, E. (2013). Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells. *Nature Methods*, *10*(4), 315-323.

Stafford, P., Halperin, R., Legutki, J. B., Magee, D. M., Galgiani, J., & Johnston, S. A. (2012). Physical characterization of the "immunosignaturing effect". *Molecular & Cellular Proteomics*, *11*(4).

Statt, A., Casademunt, H., Brangwynne, C. P., & Panagiotopoulos, A. Z. (2020). Model for disordered proteins with strongly sequence-dependent liquid phase behavior. *J Chem Phys*, *152*(7), 075101.

Sundberg, E. J., & Mariuzza, R. A. (2002). Molecular recognition in antibody-antigen complexes. In *Advances in Protein Chemistry* (Vol. 61, pp. 119-160). Academic Press.

Sykes, K. F., Legutki, J. B., & Stafford, P. (2013). Immunosignaturing: a critical review. Trends in Biotechnology, 31(1), 45-51.

Taguchi, A. T., Boyd, J., Diehnelt, C. W., Legutki, J. B., Zhao, Z.-G., & Woodbury, N. W. (2020). Comprehensive Prediction of Molecular Recognition in a Combinatorial Chemical Space Using Machine Learning. *ACS Combinatorial Science*, *22*(10), 500-508.

Terwilliger, N. B. (1998). Functional adaptations of oxygen-transport proteins. *The journal of experimental Biology*, *201*(8), 1085-1098.

Ti, S.-C., Pamula, M. C., Howes, S. C., Duellberg, C., Cade, N. I., Kleiner, R. E., Forth, S., Surrey, T., Nogales, E., & Kapoor, T. M. (2016). Mutations in human tubulin proximal to the kinesin-binding site alter dynamic instability at microtubule plus- and minus-ends. *Developmental cell*, *37*(1), 72-84.

Tian, F., Zhou, P., & Li, Z. (2007). T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. *Journal of Molecular Structure*, *830*(1-3), 106-115.

Tiwari, V. (2016). In vitro engineering of novel bioactivity in the natural enzymes. *Frontiers in chemistry*, *4*, 39.

Torrisi, M., Pollastri, G., & Le, Q. (2020). Deep learning methods in protein structure prediction. *Comput Struct Biotechnol J*, *18*, 1301-1310.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., & Hober, S. (2010). Towards a knowledge-based human protein atlas. *Nature Biotechnology*, *28*(12), 1248-1250.

Ullman, C. G., Frigotto, L., & Cooley, R. N. (2011). In vitro methods for peptide display and their applications. *Briefings in functional genomics*, *10*(3), 125-134.

UniProt: the universal protein knowledgebase in 2021. (2021). *Nucleic Acids Research*, *49*(D1), D480-D489.

van de Vijver, M. J., Peterse, J. L., Mooi, W. J., Wisman, P., Lomans, J., Dalesio, O., & Nusse, R. (1988). Neu-Protein Overexpression in Breast Cancer. *New England Journal of Medicine*, *319*(19), 1239-1245.

van Westen, G. J. P., Swier, R. F., Wegner, J. K., Ijzerman, A. P., van Vlijmen, H. W. T., & Bender, A. (2013). Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets. *Journal of cheminformatics*, *5*(1), 1-11.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., & Laydon, A. (2022). AlphaFold Protein Structure

Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, *50*(D1), D439-D444.

Veltri, D., Kamath, U., & Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. *Bioinformatics*, *34*(16), 2740-2747.

Vickery, H. B. (1950). The origin of the word protein. *The Yale journal of biology and medicine*, *22*(5), 387.

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., & Peters, B. (2019). The immune epitope database (IEDB): 2018 update. *Nucleic Acids Research*, *47*(D1), D339-D343.

Vojdani, A., Vojdani, E., & Kharrazian, D. (2021). Reaction of Human Monoclonal Antibodies to SARS-CoV-2 Proteins With Tissue Antigens: Implications for Autoimmune Diseases [Original Research]. *Frontiers in Immunology*, *11*.

Walport, L. J., Low, J. K. K., Matthews, J. M., & Mackay, J. P. (2021). The characterization of protein interactions - what, how and how much? *Chem Soc Rev*, *50*(22), 12292-12307.

Wang, J., Cao, H., Zhang, J. Z. H., & Qi, Y. (2018). Computational Protein Design with Deep Learning Neural Networks. *Sci Rep*, *8*(1), 6349.

Wang, L.-F., Yu, M., White, J. R., & Eaton, B. T. (1996). BTag: A novel six-residue epitope tag for surveillance and purification of recombinant proteins. *Gene*, *169*(1), 53-58.

Watford, M., & Wu, G. (2018). Protein. *Adv Nutr*, *9*(5), 651-653.

Weiner, L. M., & Carter, P. (2005). Tunable antibodies. *Nature Biotechnology*, *23*(5), 556-557.

Yang, L., Shu, M., Ma, K., Mei, H., Jiang, Y., & Li, Z. (2010). ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues. *Amino Acids*, *38*(3), 805-816.

Zaliani, A., & Gancia, E. (1999). MS-WHIM scores for amino acids: a new 3D-description for peptide QSAR and QSPR studies. *Journal of chemical information and computer sciences*, *39*(3), 525-533.

Zhang, W., & Niu, Y. (2010). Predicting flexible length linear b-cell epitopes using pairwise sequence similarity.

Zhao, S., Xu, W., Jiang, W., Yu, W., Lin, Y., Zhang, T., Yao, J., Zhou, L., Zeng, Y., & Li, H. (2010). Regulation of cellular metabolism by protein lysine acetylation. *Science*, *327*(5968), 1000-1004.

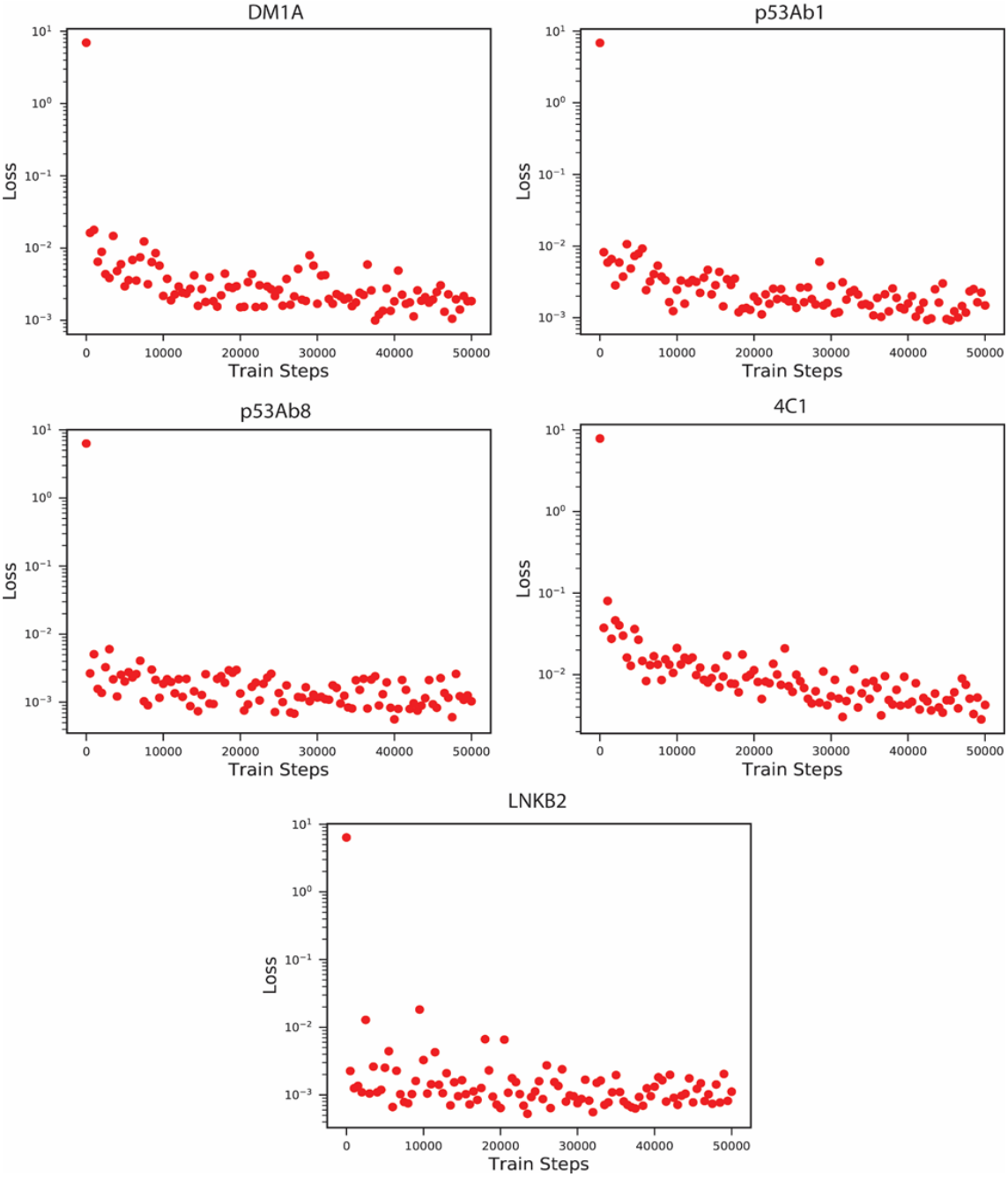
Zhou, M., Li, Q., & Wang, R. (2016). Current Experimental Methods for Characterizing Protein–Protein Interactions, *ChemMedChem*, *11*(8), 738-756.

Zolotukhin, S., & Vandenberghe, L. H. (2022). AAV capsid design: A Goldilocks challenge. *Trends Mol Med*, *28*(3), 183-193.

Zuiderweg, E. R. P. (2002). Mapping protein− protein interactions in solution by NMR spectroscopy. *Biochemistry*, *41*(1), 1-7.

APPENDIX A

SUPPLEMENTARY FIGURE FOR CHAPTER 2



Supplementary 2.1. Mean squared loss of the training set against the number of training steps (50,000) for five monoclonal antibodies used in the study. Points are plotted over every 500 steps.

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 5

Supplementary 5.1. List of Peptides Used in the Study

#	Sequence [N]-[NH_2]	Mol. Weight	pI	Diaphorase Predicted Binding	Ferredoxin Predicted Binding	FNR Predicted Binding
1	AYELVLD	820.95	3.55	6830.5	572.5	3918.2
2	PWELYFWRD	1310.51	4.184	7078	919.67	3235.7
3	GGQLVLFDG	904.04	3.75	5578.5	572.67	2752
4	KEYQWFYWFD	1510.7	4.184	7026	576.33	2237.7
5	FAWPAWVAWFE	1408.65	3.85	6770.8	748.33	2691.5
6	FSWFFPWFSE	1378.58	3.85	6599.5	769.33	2432
7	KSQLFEYVYNE	1418.58	4.258	6530.7	624	2823
8	WEFKLYAQHVL	1432.71	7.542	5237.5	880	3279.5
9	PAFRARKLFE	1233.5	10.46	15146	3001.66	50659.3
10	AWYSYPYFG	1152.3	6.084	6313.3	1232.8	3083.5
11	QERWFYYEF	1366.52	6.405	12186.6	2063.85	3945.78
12	PEEYYLYKY	1266.44	3.614	12165.56	919.25	4642.66
13	RYEEYFFPW	1335.51	4.258	6996.18	1397.23	2771.09
14	FWGEYFYPP	1204.37	3.85	5858.36	806.54	1976.26
15	FQYFKVVEE	1187.37	9.298	12034.12	1270.02	4739.42
16	FLGEVYYDK	1132.29	3.85	5937.88	984.65	2367.49
17	GERWVYYEY	1263.4	6.405	10919.78	2058.85	4197.21
18	FEWYELKLV	1225.47	3.614	10162.74	583.38	3981.82
19	FYVQFEFDR	1249.4	3.85	9510.71	1268.05	3532.01
20	PEQYYVYPL	1170.35	3.85	8786.78	985.83	2796.76
21	SFWEVVYDK	1171.33	3.85	5519.06	859.84	2184.04
22	GRYEVYYDY	1226.33	6.398	8306.9	2122.88	3287.88
23	QERWFYYEFF	1513.7	4.44	N/A	N/A	N/A
24	AYELVLDD	96.04	3.13	N/A	N/A	N/A
25	QERWLYYEF	1332.5	6.405	10406.36	2048.23	4029.3
26	FNYFKVVEE	1173.34	9.298	10787.73	1142.68	3768.94
27	GELYFYKEF	1194.37	3.85	10835.63	920.65	4000.47
28	SEQWYYYEF	1313.41	3.85	9424.15	1214.91	3293.48
29	GEKWFFYEF	1251.43	6.412	11458.65	1654.71	4390.72
30	KENYYYYEF	1317.44	6.392	7954.68	1272.94	2887.92
31	GWFEYVYDK	1205.35	3.85	5469.9	843.36	2174.29
32	FFQEYFYPP	1236.41	3.85	6459.2	767.93	2536.78
33	FFFQYEFDK	1269.44	3.85	8329.88	760.7	3021.55
34	NEFYFLKDL	1187.38	3.85	7844.86	879.24	2971.88
35	DEKWFVVFV	1167.38	4.184	26714.14	2002.33	20962.4
36	EERYVVVLV	1104.32	4.258	26376.93	2960.48	34218.4
37	NEPWYVVLV	1117.32	3.85	25086.31	1562.12	12015.6

38	EERWYVVLV	1191.41	4.44	24740.01	2896.66	35892.1
39	PENWVLVVV	1053.27	3.85	26598.99	1807.12	18142.1
40	SEPWVLVVV	1026.25	3.85	25217.19	1771.28	18348.2
41	PEPWVLVLV	1050.32	3.64	26949.62	1885.11	16995.9
42	GDHWVLVLV	1036.25	4.78	29636.08	1980.99	23756
43	EWEKVVVLV	1099.34	4.258	25435.02	1519.9	15889
44	GEPWVLVLV	1010.25	3.85	25478.75	1840.27	16350.1
45	GYHQVEYDR	1165.24	5.364	2145.17	683.2	2126.58
46	ERFEVQYDY	1247.34	4.258	2999.93	672.12	1920.32
47	LRHEVQYDY	1221.35	7.551	2772.74	951.29	2346.36
48	GVEEVLYDR	1078.2	3.614	2912.78	649.64	2304.53
49	LRHEYNYDY	1271.38	7.543	2643.44	776.07	2292.08
50	ANGPRRRYR	1144.33	12.5	67855.91	1085.73	68009.9
51	ANDFRRRKR	1217.42	10.4	63157.07	787.46	46748.4
52	AWVDYRRRF	1267.48	6.33	62395.15	3216.92	53459.4
53	ADGFRRRKR	1160.37	10.4	60656.84	933.89	55421.9
54	ARYRRYRRK	1323.6	11.21	74116.07	727.15	109123

= serial number of peptides. Peptide 1 to 10 were selected from the peptides present on V13 microarray.