ABSTRACT

Title of Dissertation: HIGH RESOLUTION MODELING OF ANTIBODY AND T CELL RECEPTOR RECOGNITION USING DEEP LEARNING

Rui Yin, Doctor of Philosophy, 2024

Dissertation directed by: Associate Professor Brian G. Pierce, Department of Cell Biology and Molecular Genetics, Institute for Bioscience and Biotechnology Research

Antibodies and T cell receptors (TCRs) are crucial for the immune system's ability to recognize and combat pathogens and cancer cells. High resolution structures of antibody-antigen complexes and TCR-peptide-MHC (TCR-pMHC) complexes provide key insights into their targeting. This knowledge has enabled the structure-based design of vaccines against viruses and pathogens, and therapeutics against cancer, immunological disorders, and viral infection. However, the vast diversity of the immune repertoire, along with limited resources and time

constraints, makes experimentally determining the structures of most antibody-antigen and TCR-pMHC interactions challenging. To support these experimental efforts, computational approaches have been developed to model the structures of these protein-protein interactions. Despite decades of development, an accurate predictive understanding of the structural basis of antibody and TCR targeting remains a challenge. Recently, deep learning algorithms have shown major promise in the field of molecular modeling, due to their ability to analyze and learn complex non-linear features underlying molecular systems. For my research, I harnessed the power of deep learning tools toward predictive modeling of antibody and TCR recognition. First, I examined the structural and physiochemical features underlying antibody-antigen recognition for antibodies that interact with the SARS-CoV-2 receptor-binding domain (RBD). Then, as a critical step toward the development of highly accurate modeling tools, I conducted a thorough benchmarking of the state-of-the-art deep learning algorithm, AlphaFold, in modeling protein-protein complexes. Focusing on antibody-antigen complexes, I identified critical areas where AlphaFold's modeling capabilities could be enhanced. Next, I developed improvements of AlphaFold to perform accurate modeling of TCR-pMHC complexes, leading to the TCRmodel2 algorithm, which is available to the community as a public web server. This was followed by an effort to explore the use of increased sampling to improve AlphaFold success, which generated near-native predictions for approximately half of antibody-antigen test cases and nearly all TCR-pMHC test cases. These advances in modeling accuracy constitute a leap forward in our predictive understanding of immune recognition and can serve as a step toward successful design of more effective vaccines and therapeutics.

HIGH RESOLUTION MODELING OF ANTIBODY AND T CELL RECEPTOR RECOGNITION USING DEEP LEARNING

by

Rui Yin

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2024

Advisory Committee:
Dr. Brian G. Pierce, Chair
Dr. Roy A. Mariuzza
Dr. John Moult
Dr. David J. Weber
Dr. John P. Dickerson

Dedication

To my mother Jinfeng Li, my father Genquan Yin, and my fiancé Jiacheng Song: I could not have done this without any of you.

Acknowledgements

First and foremost, I extend my deepest gratitude to my academic advisor, Dr. Brian Pierce. Dr. Pierce is an exemplary advisor. He is the one who introduced me to the field of computational structural biology, providing me with the necessary tools and perspective to navigate this complex area. His kindness, commitment, intelligence, patience, empathy, professionalism, encouragement, emotional stability, and unwavering support have been fundamental to the completion of this work. Throughout my doctoral studies, Dr. Pierce has been a constant source of inspiration. His open-door policy, always welcoming discussions - often extending well beyond the initially requested two minutes - provided me with invaluable insights and motivation, especially during moments of doubt or stress. His ability to understand and address problems efficiently is nothing short of remarkable. It is impossible to envision my journey through this program without his mentorship. I am profoundly thankful for having his guidance and for the positive impact he has had on my academic and personal growth.

I am immensely grateful to my committee members, Dr. Roy A. Mariuzza, Dr. John Moult, Dr. David J. Weber, and Dr. John P. Dickerson, for their steadfast support and insightful feedback on my projects, which shaped my research direction and methodology. I am also grateful to Dr. Paul Robbins for co-advising me on an NCI-UMD seed project on the structural prediction of uncharacterized TCRs. The constructive discussions I had with my committee members and Dr. Robbins were instrumental in enhancing my work and boosting my confidence.

I am thankful for all current and past members of the Pierce Lab: Dr. Helder V. Ribeiro-Filho, Dr. Johnathan Guest, Dr. Ragul Gowthaman, Melyssa Cheung, Nathaniel Felbinger, Minjae Park, Shayana Saravanakumar, Valerie Lin, Jessica Lee, Arjun Rakheja, Arnan Huang, Aaron Lewis, Sivan Kaufman, Dr. Dongxiu Zhang Spiering, Jane Quackenbush, Dr. Ghazaleh

Taherzadeh, Dr. Stefan Ivanov, Ipsa Mittra, for your helpful discussions, much appreciated encouragements, intellectual and emotional support. I would also like to express my profound gratitude to Dr. Brandon Yushan Feng, Dr. Feng's Ph.D. advisor, Dr. Amitabh Varshney and Michael Weihao Song for discussions and support during our collaboration on machine learning projects.

I am also grateful to the information technology group at the Institute for Bioscience and Biotechnology Research, including Gale Lane and Christian Presley who provided valuable assistance with high performance computing resources.

This dissertation was supported National Institutes of Health grants R01 GM126299 and R35 GM144083 to Dr. Brian Pierce, and National Cancer Institute (NCI)-University of Maryland (UMD) Partnership for Integrative Cancer Research.

Last but not least, I am profoundly thankful to my parents and my fiancé, whose encouragement and unwavering support were the catalysts for my pursuit of this journey. Their constant presence and readiness to provide support whenever needed have been a source of immense comfort and happiness. The unconditional love they have bestowed upon me instilled the confidence and courage necessary to navigate this challenging path. I am eternally grateful for their love and support, which I will cherish forever.

Table of Contents

List of Abbreviations

ACE2	Angiotensin-Converting Enzyme 2
AUC	Area Under the Curve
BLAST	Basic Local Alignment Search Tool
BLOSUM62	BLOcks SUbstitution Matrix 62
BM5.5	Protein-Protein Docking Benchmark version 5.5
BSA	Buried Surface Area
CAPRI	Critical Assessment of PRedicted Interactions
CASP	Critical Assessment of Structure Prediction
CD-HIT	Cluster Database at High Identity with Tolerance
CD1	Cluster of Differentiation 1
CDR	Complementarity Determining Region
Cryo-EM	Cryogenic Electron Microscopy
Fab	Fragment antigen-binding region
FAPE	Frame-Aligned Point Error (a loss term in AlphaFold)
Fc	Fragment crystallizable region
FFT	Fast Fourier Transform
fnat	Fraction of native contacts
FR	Framework region
GPU	Graphics Processing Unit
HCV	Hepatitis C virus
HETATMs	Heteroatoms
HIV	Human Immunodeficiency Virus
I-pLDDT	Interface predicted Local Distance Difference Test
I-RMSD	Interface Root-Mean-Square Deviation
Ig	Immunoglobulin
IgA	Immunoglobulin A
IgD	Immunoglobulin D
IgE	Immunoglobulin E

IgG	Immunoglobulin G
IgM	Immunoglobulin M
Ig-NAR	Immunoglobulin New Antigen Receptor
IPA	Invariant Point Attention module in AlphaFold
ipTM	Interface predicted TM-score
IRAD	Integration of residue- and atom-based potentials for docking
L-RMSD	Ligand Root-Mean-Square Deviation
MHC	Major Histocompatibility Complex
MR1	MHC Class I-Related protein 1
MSA	Multiple Sequence Alignment
N-glycosylation	N-linked glycosylation
N_{eff}	Number of Effective Sequences
PAE	Predicted Alignment Error
PDB	Protein Data Bank
pMHC	Peptide-MHC
pTM	Predicted TM-score
RBD	Receptor Binding Domain
REF15	Rosetta Energy Function 2015
RFAA	RoseTTAFold All-Atom
RMSD	Root Mean Square Deviation
ROC	Receiver Operating Characteristic
SAbDab	The Structural Antibody Database
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SHM	Somatic Hypermutation
TCR	T Cell Receptor
TCR-pMHC	T Cell Receptor-Peptide-Major Histocompatibility Complex
V(D)J	Variable(Diversity)Joining recombination
VHH	Variable Heavy chain domain of Heavy chain antibodies
VLR	Variable Lymphocyte Receptor
Vα	Variable α chain

Vβ	Variable β chain
X-ray	X-ray crystallography
ZDOCK	Zlab Docking
ZRANK	Zlab Rerank
ZRANK2	Zlab Rerank, verison 2

Chapter 1: Introduction

An accurate predictive understanding of antibody-antigen and T cell receptor-peptide-major histocompatibility complex (TCR-pMHC) interactions plays a crucial role in advancing our knowledge of immune recognition and can serve as the foundation for the development of disease therapeutics and vaccines. My dissertation is focused on developing highly accurate algorithms for modeling antibody-antigen and TCR-pMHC complexes. Specifically, I utilize deep learning models which are adept at handling complex datasets, to approach this problem.

I begin with a comprehensive review of the literature in this field, detailed in **Chapter 2**. In that chapter, I examine the pivotal developments and highlight both the achievements and the challenges faced in accurate modeling of immune recognition.

In **Chapter 3**, I provide an in-depth analysis of the interaction characteristics between antibodies and the receptor binding domain (RBD) of SARS-CoV-2. This analysis details a comprehensive view of the recognition features between the SARS-CoV-2 RBD and antibodies. It also offers valuable insights into the development of strategies to target escape variants effectively.

A critical aspect of this research is the assessment of current modeling tools. In **Chapter 4**, I present a thorough benchmarking analysis of AlphaFold, RoseTTAFold, and ZDOCK for protein-protein complex predictions. This study highlights AlphaFold's limitations in protein-protein complex predictions, particularly when co-evolutionary information is absent. Despite this, the “fold-and-dock” approach, which AlphaFold exemplifies, showed some promise in the context of antibody-antigen modeling.

Chapter 5 focuses on antibody-antigen complexes, where I examine the modeling efficacy of AlphaFold in this specific area. The findings suggest that while AlphaFold underperformed in antibody-antigen complexes, it exhibited relatively higher success compared to traditional rigid-body docking algorithms. This chapter also emphasizes the importance of accurately modeling antibody and antigen chains in their bound conformations to enhance complex modeling accuracy. Moreover, I show that an interface-focused scoring function is effective in discriminating accurate versus inaccurate predictions.

Building on the insights gained from the benchmarking studies, **Chapter 6** introduces an innovative TCR-pMHC complex modeling algorithm, TCRmodel2. This algorithm, an advancement over AlphaFold, offers enhanced accuracy and speed, making it a significant contribution to the field. The algorithm's accessibility through our lab's webserver further extends its impact and utility.

Chapter 7 discusses the application of massive sampling techniques to generate improved models of antibody-antigen and TCR-pMHC complexes. These approaches represent a leap forward in our ability to accurately immune recognition.

In **Chapter 8**, I examine the analysis and findings presented in earlier chapters and provide a comprehensive understanding of the implications of these findings and identify areas for further improvement.

Chapter 2: Literature Review

This chapter reviews the studies on high resolution modeling of antibody and T cell receptor (TCR) recognition using deep learning, as well as structural immunology and computational structural biology areas that serve as the foundation for this research. I will start by discussing the molecular structure and biological functions of antibodies and TCRs. I will then discuss the broader field of protein-protein complex modeling, highlighting the unique challenges and recent breakthroughs in modeling the intricate interactions between antibodies, TCRs, and their antigenic targets. Particular emphasis is placed on the transformative impact of deep learning-based algorithms, which have revolutionized the prediction of protein structures. Synthesizing current knowledge and identifying gaps in the existing literature, this chapter lays the groundwork for the subsequent analysis and development of algorithms to model immune recognition.

2.1 Antibody and TCR molecular recognition

2.1.1 Antibody and TCR function and importance

The highly diverse and specific nature of adaptive immunity protects us from a broad spectrum of foreign pathogens (e.g., bacteria, viruses, and parasites) and abnormal cells. Vital to the specificity and diversity of vertebrates' adaptive immune systems are two classes of molecules: antibodies and TCRs. Antibodies are produced by B cells and are the first class of molecules involved in immune recognition to be identified and characterized by scientists, with their initial discovery tracing back to 1890 by von Behring and Kitasato [1]. Antibodies are

usually found in solution or at the B cell surface. They can recognize and bind intact soluble or surface-exposed antigens, including proteins, peptides, non-protein molecules, and combinations of these molecules [2]. Upon binding the antigen, antibodies can neutralize pathogens by blocking their function. They can also coat the antigens, marking them for destruction by macrophages and neutrophils. Antibody binding to antigens can additionally activate the complement system for enhanced killing of pathogens.

Unlike antibodies, TCRs are found at the surface of T lymphocytes. They typically engage fragments of antigens, presented on the surface of cells by the major histocompatibility complex (MHC). The antigen fragments can be produced by intracellular proteolysis, endocytic uptake, or cross-presentation [3]. TCRs play a vital role in orchestrating the cellular immune response to combat virally infected cells. Thanks to their ability to recognize antigens presented by intracellular proteolysis, TCRs' functionality is broadened to target the comprehensive proteome of cancer cell [4]. This broad targeting potential makes TCRs a versatile tool in the immune system's arsenal against cancer.

Depending on whether an agonist, partial agonist, or antagonist peptide is presented, the subsequent cellular responses differ significantly. When presented by MHC molecules, agonist peptides typically trigger T cell activation, proliferation, and the release of cytokines, leading to an effective immune response [5]. In contrast, partial agonist and antagonist peptides fail to induce a complete activation signal. This results in partial or no T cell activation, potentially leading to anergy or tolerance [6]. T cells can also recognize and bind lipids and metabolites presented by cluster of differentiation 1 (CD1) and MHC class I-related molecule (MR1). Moreover, some γδ TCRs can recognize antigens without MHC, CD1, or MR1 [7].

Studying antibodies and TCRs has refined our knowledge of immunology and revolutionized our approach to combat and treat diseases. By November 2023, the number of therapeutic antibodies that have been approved or are in the process of regulatory evaluation is close to 200, with over 1000 currently in various stages of clinical studies [8]. Despite the fact that fewer TCR therapies have been developed than antibody-based treatments, TCR-based therapies possess enormous potential due to their ability to target intracellular targets. In 2021, the U.S. Food and Drug Administration (FDA) approved its first TCR therapeutic, Tebentafusp [9, 10], for the treatment of metastatic uveal melanoma. Multiple TCR therapies are presently undergoing clinical trials [11], offering hope for further breakthroughs in the realm of therapeutic developments.

2.1.2 Architecture of antibodies and TCRs

In humans, antibodies are categorized into five classes—Immunoglobulin G (IgG), M (IgM), A (IgA), E (IgE), and D (IgD). IgGs form a Y-shaped structure with two antigen-binding fragments (Fab) and a crystallizable fragment (Fc) [12]. Typically, the antibody "Y" is composed of two pairs of heavy and light chains with a disulfide bond linking the heavy and the light chain pair. The variable regions (Fabs) of the antibodies are found at the tips of "Y" and are responsible for recognizing and binding antigens with specificity and avidity. These regions have hypervariable loops (Complementarity Determining Regions, abbreviated CDRs) that form the primary antigen-binding site. Located at the base of the "Y" is the constant region, Fc, that determines the antibody class and mediates interactions with the complement system proteins.

A subclass of antibodies, nanobodies, also known as single-domain antibodies (sdAbs), are exceptions to the commonly observed antibody architecture described in the previous

paragraph. Nanobodies are derived from heavy-chain-only antibodies (HCAbs) found in camelids. Despite having only the heavy chains, nanobodies are capable of binding to their antigenic targets with high affinity that is in the same range as their traditional heavy-light chain counterparts [13]. Remarkably, HCAbs are not only found in camelids but also in cartilaginous fish species (e.g., sharks), which possess structurally similar antigen-binding proteins called Ig-NARs. Ig-NARs and HCAbs demonstrate convergent evolution; they have different origins despite having similar functions [14].

Variable Lymphocyte Receptors (VLRs) serve as another example, illustrating that diverse evolutionary paths can lead to similarly effective solutions in immune response. VLRs are a unique class of antigen receptors found exclusively in jawless fish, such as lampreys and hagfish. In fact, they are believed to be the most ancient adaptive immune receptors [15-17]. VLRs differ from antibodies in their structure; they are constructed from Leucine-Rich Repeat (LRR) modules and adopt a horseshoe-shaped structure [15, 16]. In a VLR-hen egg white lysozyme (HEL) protein complex, researchers demonstrate that VLR binds to HEL via their concave surface ridges [18]. As such, VLRs, while structurally different from antibodies, achieve a similar outcome: specific recognition of antigens.

TCRs are heterodimers that, in most cases, consist of TCRα and TCRβ chains. In an estimated 0.5-5% of cases, TCRs comprise TCRγ and TCRδ chains [19]. Like antibodies, they have variable regions with CDR loops for antigen recognition and interaction. Their constant domain is linked to the transmembrane domain, which anchors the TCR to the surface of the T cell.

Each variable and constant domain of antibody and TCR exhibits a characteristic immunoglobulin fold (Ig fold) – a β-sandwich structure with antiparallel β sheets. The variable

domains of antibody and TCR typically have nine β strands arranged in two sheets. These β strands provide a scaffold for the CDR loops. CDR loops exhibit immense sequence diversity, generated through V(D)J recombination and, in antibodies, somatic recombination. Aside from genetic hypervariability, the CDR loops are structurally flexible and adopt different conformations between unbound and bound states. Together, these loops' immense sequence and structural diversity allow for the recognition of a vast array of antigens.

Antibody and TCR CDR loops can differ in length, structure, and underlying sequence pattern. Wong et al. showed that antibody CDR3 loops are longer and more variable, compared to TCR CDR3 loops [20]. Moreover, most antibody heavy chain CDR3 loops have a “kinked torso” [20, 21], whereas most TCRβ CDR3 loops have an “extended torso” [20].

2.1.3 Sequence diversity

V(D)J recombination greatly shapes the sequence diversity of antibodies and TCRs. V(D)J recombination occurs during the development of B cells and T cells. It involves the splitting and joining of three types of gene segments: Variable (V), Diversity (D), and Joining (J) [22]. Importantly, the D gene is only used to construct antibody heavy chain, TCRβ and TCRδ chain.

There are two primary ways V(D)J recombination fuels diversity. First, random combinations of V, D, and J gene segments create combinatorial diversity. This diversity is further extended by the pairing of heavy and light chains in antibodies and of α and β chains in TCRs. Second, during the V, D, and J gene recombination, random addition and deletion of nucleotides can take place at the gene junction, which further diversifies the sequence of antibodies and TCRs.

Additionally, antibodies can undergo somatic hypermutation (SHM) after encountering the antigen. SHM introduces further mutations in CDRs, which changes the binding affinity of antibodies to the antigen [23]. Subsequently, in a process known as affinity maturation, antibodies with high binding affinities to the antigen are selected to proliferate. In this process, the mutations these antibodies carry are preserved and antibody sequence diversity increased.

2.1.4 Structure of antibody-antigen complexes

Detailed structural analysis of antibody-antigen interactions at high resolution can provide valuable insights into antibody recognition of antigenic targets [24-26]. These insights shape our strategy for combating viral infections for SARS-CoV-2 [27-29], influenza hemagglutinin [26, 30-32], HIV Env [33-35]. Furthermore, these insights guide the rational engineering of the antibody [35, 36] and the rational design of the immunogens [36, 37].

Several studies have focused explicitly on general antibody-antigen interactions [2, 38-40], while another has concentrated on the antibody-peptide interface [41]. These studies offered significant insights into antibody-antigen interactions. Studies found that antibody-antigen complexes can bury a surface area from 600 $Å^2$ to 2400 $Å^2$ [41, 42]. Depending on the size of the antigen, the shape of the antibody-antigen binding interface can change, with flatter interfaces found between antibodies and larger antigens and more concave interfaces found between antibodies and smaller antigens [42]. Moreover, researchers found that in addition to CDRs, other parts of the antibody, such as the framework regions (FRs) and the constant domains, also play a role in antigen binding [2, 39]. Importantly, the epitope, which is the region on the antigen that binds to antibodies, does not differ from the rest of surface-exposed antigen residues in terms of sequence features such as residue composition [38].

Nanobodies, a subclass of antibodies, bind to the antigenic targets in a manner that is different from the traditional heavy-light chain antibodies. Research reveals that nanobodies tend to possess longer CDR3 loops compared to heavy-light chain antibodies [43]. Researchers also found that nanobody paratopes contain fewer aromatic amino acid residues [44], and are more diverse in terms of residue types and residue position, often involving framework residues as paratope residues [45]. Moreover, the nanobody epitopes can be characterized by structural rigidity and concaveness, which is also distinct from the heavy-light chain antibodies epitopes [44].

2.1.5 Structure of αβ TCR-peptide-MHC complexes

The first TCR-pMHC complex structure was determined through X-ray crystallography in 1996 [46]. Since then, structures of over 300 TCR-pMHC complexes have been determined [47]. These high resolution 3D structures of TCR-pMHC complexes have provided major insights into effective immune responses [48], autoimmune diseases [49], and therapeutic TCRs [50], the basis of off-target toxicity in the clinic [51]. Recently, researchers studied the structural feature of TCR-pMHC complexes and revealed how TCRs specifically recognize cancer neoantigens [52].

Unlike antibody-antigen complexes, αβ TCR-pMHC complexes adopt a relatively conserved binding mode, which can be characterized by TCRs sitting diagonally over the pMHC [53]. In this binding mode, TCR CDR1 and CDR2 loops primarily contact the MHC helices, while the CDR3 primarily contacts the peptide [54]. Despite the relatively conserved binding mode, TCR-pMHC complexes can show significant structural diversity. While the general diagonal arrangement of TCR chains over peptides is common, the crossing angles [53] vary

widely by as much as over 100 degrees [47]. Moreover, a highly tilted docking mode has been observed in an autoimmune TCR with a Class II MHC [55]. A systematic review of existing TCR-pMHC experimentally resolved structures reveals that TCR-pMHC complexes can also exhibit a high degree of variation in the incident angle (tilt) [47]. These variations underscore the complexity of TCR-pMHC interactions and the challenge of modeling them accurately.

2.2. Protein-protein docking

2.2.1 Traditional protein-protein docking algorithms

Prior to the development of AlphaFold, traditional techniques for protein-protein docking were employed to predict how proteins structurally interact. These methods fall into two categories based on the availability of homologous templates. Template-based methods rely on 3D protein structures deposited into the Protein Data Bank (PDB) [56]. These methods identify templates based on the sequence similarity to existing protein complex structures in PDB and generate protein complex models using the templates identified [57-59]. Sequence similarity has a notable influence on the results of template-based modeling. In situations where sequence similarity falls below 30%, often described as the "twilight zone" for protein-protein docking templates, the quality of the models produced by this method tends to decrease drastically [60]. When high-homology templates are unavailable, *ab initio* docking algorithms are employed to predict complex structures. These docking algorithms use unbound or modeled subunit structures as inputs and perform an extensive search of binding poses [61-68]. To strike a balance between the resolution and speed, structural approximations were used in Fast Fourier Transform (FFT)-

based approaches [61, 62, 69], geometric hashing [68], or other computation-efficient techniques.

However, accurate *ab initio* docking of protein-protein complexes is challenging. Finding the correct binding site is computationally expensive. Moreover, the side chains and backbones of protein structures in unbound vs. bound states can exhibit large conformational changes. In a previous study, researchers employed a global docking algorithm ClusPro and its protocol designed for antibody-antigen modeling [70] to predict the antibody-antigen complex structures of 67 test cases [71]. Using the Critical Assessment of PRedicted Interactions (CAPRI) evaluation criteria, which assess the predictions' accuracy, it was found that out of 67 antibody-antigen cases tested, only 3 were had Medium or higher accuracy top-ranked predictions [71].

To improve docking success, researchers have explored clustering [67, 72] or rescoring [73-76] algorithms to re-rank and select docking predictions, as well as the explicit consideration of side chain [77] and backbone flexibility during docking [78] to enhance modeling success. Researchers have also explored using data obtained from various experiments as constraints to improve docking performance [66, 79-81]. Apart from experimental constraints, researchers experimented with computationally derived co-evolutionary information as docking constraints [82, 83].

Despite those developments, success in protein-protein docking has generally remained limited, as shown by the Critical Assessment of Structure Prediction (CASP) and CAPRI blind docking prediction experiment [84-86] and protein-protein docking benchmarks [71, 87, 88]. However, the recent advancements in deep learning-based protein complex modeling algorithms, exemplified by the AlphaFold algorithm, brought a major revolution and greatly elevated the success of protein-protein structure prediction [86, 89].

2.2.2 AlphaFold and RoseTTAFold

Due to their capability to learn non-linear hidden information underlying protein folding, deep learning techniques have been increasingly harnessed toward the molecular modeling [90-99]. Traditionally developed for image recognition and processing, deep learning has shown great promise in molecular dynamics simulation [100], computational chemistry [101], and drug discovery [102-108] in recent years.

To date, researchers have utilized a variety of deep learning algorithms, such as convolutional neural networks, residual networks, and graph neural networks, in the field of protein structure prediction. Most efforts in this direction have been focused on intra-chain contact and tertiary structure prediction that leverage co-evolution information from multiple sequence alignments (MSAs) to refine the prediction of inter-residue distances and torsional angles [91, 109-114]. This includes AlphaFold (AlphaFold v.2.0) [111], an end-to-end deep neural network that generates structural models from the sequence, which showed unprecedentedly high modeling accuracy and substantially surpassed the performance of other teams in the CASP round 14 [112]. Beyond MSA-based deep learning algorithms, language model-based algorithms for protein structure prediction also received a lot of attention in the past few years [115, 116].

Key innovations in AlphaFold include the learning of pairwise residue-residue features through blocks of attention neural networks from a joint embedding of MSAs and pair representations derived from the PDB [56]. These pairwise features are used as distance constraints to construct three-dimensional protein structures through a novel equivariant attention neural network.

The use of deep learning algorithms in protein-protein contact and quaternary structure prediction has gained increasing popularity in recent years [91, 96, 117-122]. One such algorithm is AlphaFold-Multimer [118], which predicts protein complex structures from the sequence. Like AlphaFold [111], AlphaFold-Multimer infers pairwise residue contact information from MSAs through iterative refinement steps and extensive use of attention mechanisms.

Another important program in deep learning-based protein structure prediction is RoseTTAFold [91]. RoseTTAFold uses a "three-track" approach and uses biaxial attention. However, AlphaFold performs better than RoseTTAFold [123]. Inspired by AlphaFold, RoseTTAFold2 was released [123]. It maintains the three-track design (1D, 2D, 3D) while increasing network depth and MSA size.

RoseTTAFold All-Atom (RFAA) [124] extends RoseTTAFold's capability further with updated molecular representations to model complex molecular assemblies that include not just proteins but also ligands, small molecules, ions and post-translational modifications. RFAA uses different tracks to represent and process different levels of information, such as element type, bond type, and molecule chirality. RFAA also explored and used graphs, attention mechanisms, and iterative refinement processes.

Recently, the Google DeepMind AlphaFold team and Isomorphic Labs reported progress on developing the next generation of AlphaFold that is also capable of modeling molecular assemblies, including proteins, ligands, ions and post translational modifications [125]. The report mentions that this version of AlphaFold achieves markedly better performance in modeling antibody-antigen complexes. However, as details of this version remain to be released, and this report remains unreviewed, its accuracy needs to be confirmed and further investigated.

2.2.3 Other deep learning-based protein-protein docking algorithms

Recently, developments have been made to combine deep learning with protein-protein docking, including regression-based methods such as EquiDock [126] and DockGPT [127]. Equidock is a rigid-body docking algorithm and does not account for protein backbone conformational changes. DockGPT, on the other hand, allows for flexibility of protein backbone structure and can take interface constraints as input. Apart from regression-based methods, diffusion-based methods for protein-protein docking have also been explored [128, 129].

As the success of AlphaFold-Multimer and RoseTTAFold is built upon MSAs, it is unclear whether they can reliably predict the protein complex structures with little co-evolutionary information, including antibody-antigen complexes and TCR-pMHC complexes. The new generation of docking algorithms that leverage the benefits of deep learning while being less reliant on co-evolutionary signals across protein-protein interfaces harbors new hope for immune recognition modeling.

2.3. Antibody and TCR modeling.

Many programs have been developed to predict antibody and TCR structures from sequence. These programs often perform template-based modeling of framework regions and CDR loop grafting based on templates, followed by loop refinement and energy minimization [130-136]. These methods, while generally successful in predicting structures on the framework regions, often fail to predict CDR loop conformations. The lack of accuracy in modeling CDR loops of these template-based methods is detailed in an assessment published in 2014, which shows that programs could produce overall high-quality antibody models (1.1 ± 0.2 Å average backbone root mean square deviation (RMSD) between prediction and experimentally

determined Fv regions) but are unable to accurately recapitulate CDR loop conformations, especially for the more diverse CDRH3 loops (2.8 Å average backbone RMSD) [137].

In a recent study [138], authors compared the antibody, nanobody, and TCR modeling capacity between AlphaFold (and AlphaFold-Multimer) versus ABodyBuilder [135, 138], ABlooper [139], IgFold [140], TCRBuilder [141], RepertoireBuilder [142], and MOE [142]. They found that while all algorithms generated accurate framework regions, achieving high accuracy CDR loop modeling is difficult considering the high sequence and structural variability. Focusing on the CDR3 loop accuracy, AlphaFold, AlphaFold-Multimer, and ImmuneBuilder all generate similar success for the CDRH3 loop and CDR3β loop of antibody and TCR, which are the most challenging loop of all loops, and the three are generally more accurate than the rest of the methods. Even with these AI-empowered algorithms, CDRH3 loops were still modeled with a mean RMSD of 2.8 Å or worse [138].

While the prediction of antibody or TCR structures has been generally successful, accurate prediction of antibody and TCR in complex with their interacting proteins remains a challenge. General protein-protein docking methods have been applied to model antibody-antigen complex structures with limited success [71, 143]. The challenges in modeling antibody-antigen complexes mainly arise from the need to accommodate for the mobility of key CDR loops and the large-scale search of antigen surface. Despite their relatively conserved binding mode, TCR-pMHC complexes are challenging to predict. This can be attributed to the high mobility of CDR loops and variability in TCR-pMHC crossing angle and docking angle [47, 144]. As such, additional algorithms have been developed specifically for antibody-antigen [79, 145-147] or TCR-pMHC [148-150] complex prediction problems. In spite of these advances, accurate structural prediction of antibody and TCR recognition remains a challenge.

In the recent CASP/CAPRI experiment, antibody-antigen complexes proved challenging for participating teams and algorithms [86]. In the experiment, there are five nanobody-antigen targets with nanobodies binding to different regions of a mammalian CNPase phosphodiesterase domain (T205-T209) and three antibody-antigen targets with antibodies binding to various regions of the SARS-CoV-2 nucleoprotein (T216-T218). Participants successfully generated near-native accuracy predictions for four out of the five nanobody-antigen targets, but only one of the three antibody-antigen targets. Overall, this highlights the need for continued development of modeling methods for antibody-antigen complexes.

2.4 Summary

In this literature review chapter, I provided an overview of the biological roles and structural features of antibodies and TCRs, elucidating their paramount importance in vertebrate adaptive immunity. I then delved into the structural features of antibody and TCR recognition, leveraging the wealth of information provided by high resolution structures of antibody-antigen and TCR-pMHC complexes. Recognizing the pivotal role of these structures, coupled with the practical limitations in experimentally determining all antibody-antigen and TCR-pMHC complexes, I highlighted the need to develop computational methods to model these interactions. However, modeling these interactions with high accuracy remains a challenge, notably due to the structural flexibility inherent in antibody-antigen and TCR-pMHC interactions.

Deep learning algorithms have emerged as a powerful tool in molecular modeling thanks to their capacity to understand and interpret the complex, non-linear characteristics of molecular interactions. I provided an overview of existing docking algorithms, with a particular emphasis on the integration of protein complex modeling methods and the transformative impact of deep

learning. The advancements in protein complex modeling, represented by the AlphaFold algorithm, offer promising avenues for high resolution modeling of immune recognition.

However, as the AlphaFold algorithm relies on co-evolutionary information, it is unclear whether modeling antibody-antigen interactions with this new generation of protein complex modeling tools is successful. In the chapters to follow, I will continue investigating the structural and energetic features underlying antibody recognition of antigenic targets, assess the utility of AlphaFold for modeling these interactions, identify areas of improvement, and propose developments to improve the modeling accuracy for immune recognition.

Chapter 3: Structural and energetic profiling of SARS-CoV-2 antibody recognition and the impact of circulating variants

3.1 Abstract

The SARS-CoV-2 pandemic highlights the need for a detailed molecular understanding of protective antibody responses. This is underscored by the emergence and spread of SARS-CoV-2 variants, including B.1.1.7, P.1, and B.1.351, some of which appear to be less effectively targeted by current monoclonal antibodies and vaccines. Here we report a high resolution and comprehensive map of antibody recognition of the SARS-CoV-2 spike receptor binding domain (RBD), which is the target of most neutralizing antibodies, using computational structural analysis. With a dataset of nonredundant experimentally determined antibody-RBD structures, we classified antibodies by RBD residue binding determinants using unsupervised clustering. We also identified the energetic and conservation features of epitope residues and assessed the capacity of viral variant mutations to disrupt antibody recognition, revealing sets of antibodies predicted to effectively target recently described viral variants. This detailed structure-based reference of antibody RBD recognition signatures can inform therapeutic and vaccine design strategies.

3.2 Introduction

The SARS-CoV-2 pandemic has resulted in a massive and growing global death toll and disease burden. A number of vaccines [151], monoclonal antibodies [152], and small molecule

therapies [153] that target SARS-CoV-2 have been developed. However, viral variants have raised the possibility of viral escape from, or reduced efficacy of, vaccines and therapeutics [154-159].

Several recent studies have used in vitro experimental approaches to test human sera [158, 160] and sets of monoclonal antibodies [155, 158, 161, 162] to profile SARS-CoV-2 antibody resistance. The rapidly expanding set of experimentally determined structures of antibodies targeting the spike glycoprotein provides the opportunity to use computational biology tools to map key features of antibody-spike recognition. At the same time, the impact of viral variability can be predicted, which can provide insights into effective targeting and neutralization of SARS-CoV-2 and enable selection and engineering of anti-spike therapeutics and vaccines.

In this section, we report detailed structural analysis of a large set of high resolution antibody-spike complexes that have been collected in our database, CoV3D [163]. Structure-based mapping of antibody footprints on the receptor binding domain (RBD) and unsupervised clustering led to the identification of four major antibody groups based on their recognition signatures. These antibody-spike complexes were assessed for key energetic features using computational alanine mutagenesis of all RBD interface residues to identify shared and distinct binding hotspots on the RBD. The structure-based antibody clusters were also assessed both for residue conservation with SARS-CoV-1, and predicted effects of individual RBD substitutions from circulating SARS-CoV-2 variants, showing substantial differences between groups of RBD-targeting antibodies. These structural features and clusters can serve as a reference for rational vaccine design and therapeutic efforts, and updated antibody cluster information is available to the community on the CoV3D site: https://cov3d.ibbr.umd.edu/antibody_classification.

3.3 Materials and Methods

3.3.1 Structure assembly and curation

Structures of antibody-RBD complexes were downloaded from the CoV3D database [163], which identifies antibody-RBD structures in the Protein Data Bank [164] on a weekly basis through sequence similarity to coronavirus reference protein sequences in conjunction with identification and annotation of antibody chains. The set of antibody-RBD structures (downloaded in February 2021) was filtered for antibody nonredundancy based on antibody name and sequence identity, as well as resolution (< 4.0 Å). In cases of an antibody present in multiple antibody-RBD complex structures, the structure with highest resolution was selected for analysis. For consistency among antibody-RBD complex structures, and to facilitate calculations, antibodies were truncated to include variable domains, and full spike glycoproteins were truncated to include only RBD residues (residues 333-527) of the sole or primary target of the antibody. To provide uniform input structures for atomic contact and other calculations, non-amino acid HETATMs were removed prior to structural analysis, and to resolve double occupancies and add missing side chain atoms, structures were pre-processed by the "score" application in Rosetta version 3.12 [165]. Two complexes with missing side chain atoms in the experimental PDB coordinates were processed using the FastRelax protocol in Rosetta [166], to perform constrained local minimization and to resolve unfavorable energies due to clashes from rebuilt side chains (antibodies DH1047, C104; PDB codes 7LD1, 7K8U). Parameter flags used in FastRelax ("relax" executable in Rosetta 3.12) are:

-relax:constrain_relax_to_start_coords

-relax:coord_constrain_sidechains

-relax:ramp_constraints false

-ex1

-ex2aro

-no_optH false

-flip_HNQ

-renumber_pdb F

-nstruct 1

The set of pre-processed structures, aligned to a common RBD reference frame, is available through the CoV3D site [163], at: https://cov3d.ibbr.umd.edu/download ("Nonredundant RBD-antibody complex structures" link).

Information regarding neutralization of SARS-CoV-2 and SARS-CoV-1 was obtained from the CoV-AbDab site [167], as well as references from the literature for certain antibodies, where noted in **Table 3.1**.

3.3.2 Computational structural analysis

RMSD values between antibody heavy chain or nanobody orientations were determined by superposition of RBDs from two complexes using least-squares fitting of backbone atoms, followed by superposition of one antibody variable domain onto another using least squares fitting of framework residue backbone atoms, and calculation of backbone RMSD between superposed and non-superposed variable domain. RBD residues used for superposition (present in all structures in this set) are 338-356, 375-382, 397-442, 448-454, 462-467, 490-501, and 503-514. Antibody variable domain framework residues used for superposition and RMSD calculations are 3-7, 21-24, 41-46, 52-57, 78-82, 89-93, 102-108, and 141-144, based on the AHo numbering system [168]. Interface contacts are defined as inter-atomic distance between non-hydrogen atoms of less than 5 Å, and antibody-RBD residue contact maps were generated based on the total number

of antibody atom contacts with each RBD residue. Hierarchical clustering of antibody RMSDs was performed in R version 4.0.3 (www.r-project.org) with the distance matrix of RMSDs as input, and Ward's minimum variance method ("ward.D2" method in hclust). Hierarchical clustering of antibodies and RBD positions based on contact data was performed in R, using Manhattan distance to compute differences in contact profiles between antibodies or RBD positions, and Ward's minimum variance method for clustering. Hierarchical clustering of RBD positions based on hydrogen bond or calculated ΔΔG values, for the respective heatmap figures, was likewise performed in R, using Manhattan distances and Ward's clustering algorithm. RBD residue dimension reduction for representation in main heatmap (**Figure 3.4**) was performed by selecting exemplar residues from 100 hierarchical clusters, which removed residues with highly similar contact profiles (based on Manhattan distance) with respect to those shown in the heatmap. The pvclust method [169], as implemented in R, was used to calculate bootstrap confidence of contact-based hierarchical clusters of antibodies, using 20,000 bootstrap replicates. Principal component analysis of antibody-RBD contact profile data was performed with the scikit-learn Python module.

Buried surface areas (BSAs) were calculated using the naccess program (v. 2.1.1) [170], subtracting the solvent accessible surface area of the antibody-RBD complex structure from the total solvent accessible surface area of the separate antibody and RBD structures, dividing by two to avoid double-counting interface area and to make BSA values commensurate with those from other tools including PISA (http://www.ebi.ac.uk/pdbe/prot_int/pistart.html). Antibody-RBD interface hydrogen bonds were calculated using the hbplus program (v. 3.15) [171], with default parameters.

Structure-based calculations of antibody blocking of ACE2 binding to RBD were calculated using the ACE2-RBD complex structure (PDB code 6LZG) [172]. After superposition of ACE2-RBD and antibody-RBD complexes by RBD, the number of inter-atomic clashes, defined as non-hydrogen atom pairs with distances < 2.5 Å, was calculated between ACE2 and each antibody structure. Antibodies with > 20 atomic clashes with ACE2 were classified as likely to block ACE2 binding. Structure-based calculations of antibody binding to the closed spike structure were performed using the SARS-CoV-2 closed spike structure reported by Walls et al. (PDB code 6VXX) [173]. Antibodies with < 100 atomic clashes with spike atoms outside of the target RBD structure and chain after superposition of the antibody-RBD complex onto the 6VXX structure were classified as predicted to bind the closed spike. Clash thresholds were selected based on agreement with structures and experimental data regarding ACE2 blocking and closed spike binding, when available. Four antibodies that engaged the closed spike and exhibited cross-protomer binding, as confirmed by inspection of antibody-spike complex structures (S2M11, C144, mNb6, LY-CoV555; PDB codes 7K43, 7K90, 7KKL, 7L3N) [27, 174-176], were annotated accordingly in the contact heatmap.

3.3.3 Computational mutagenesis

Computational modeling and prediction of antibody binding energy changes (ΔΔGs) for alanine substitutions and other residue substitutions was performed using Rosetta version 2.3 [177], Rosetta version 3.12 [165], and FoldX version 4 [178]. Benchmarking of computational alanine scanning predictive performance was performed using a subset of the AB-Bind dataset [179] that contains alanine point substitutions with quantified experimental ΔΔG measurements and known wild-type complex structures (347 mutants and ΔΔG values). A larger set with all point

substitutions (including non-alanine substitutions) was also tested (531 mutants and ΔΔG values). Pearson correlation coefficients (PCC) between measured and predicted ΔΔG values, and receiver operating characteristic area under the curve (AUC) values for prediction of hotspot residues (measured ΔΔG for alanine residue substitution > 1 kcal/mol), were calculated using scipy and scikit-learn (sklearn) Python libraries, respectively.

Rosetta 2.3 ΔΔG calculations were performed using the "interface" protocol [177, 180]. An example command line is:

rosetta.mactel -interface -intout pdb.ddgs.out -ignore_unrecognized_res -safety_check -skip_missing_residues -mutlist pdb.muts.txt -extrachi_cutoff 1 -ex1 -ex2 -ex3 -constant_seed -jran 12 -yap -s input.pdb

The input files specified on the command line denote the input PDB file ("input.pdb") and the list of mutations ("pdb.muts.txt"). The default protocol only models the mutant residue for ΔΔG calculation ("Ros2.3_norepack" in **Table 3.2**), and additional flags were used on the command line to perform minimization of mutation-proximal side chains ("-min_interface -int_chi" flags; "Ros2.3_minint_chi" in **Table 3.2**), minimization of mutation-proximal side chains and backbone ("-min_interface -int_bb -int_chi" flags; "Ros2.3_minint_bb_chi" in **Table 3.2**), and rotamer-based packing of mutation-proximal side chains ("-repack" flag, "Ros2.3_repack" in **Table 3.2**).

Rosetta 3 ΔΔG calculations were performed with two available computational mutagenesis protocols. One Rosetta 3 computational alanine scanning protocol was downloaded from a public resource containing benchmarks and Rosetta tools [181], and represents a separate implementation of the Rosetta 2.3 mutagenesis protocol noted above [177, 180]. This protocol was recently used to predict TCR-peptide-MHC interface ΔΔG values [182]. In addition to the default protocol that

does not repack neighboring side chains ("Ros3_norepack" in **Table 3.2**), we also tested this protocol with repacking of neighboring side chains ("Ros3_repack" in **Table 3.2**).

An example command line for running this protocol is:

rosetta_scripts.static.linuxgccrelease -s input.pdb -parser:protocol alascan.xml -parser:view -inout:dbms:mode sqlite3 -inout:dbms:database_name rosetta_output.db3 -no_optH true -parser:script_vars pathtoresfile=input.resfile chainstomove=1,2 -ignore_zero_occupancy false

We additionally performed alanine scanning using the flex ddG protocol, which was developed recently in Rosetta 3 [183]. This protocol uses the backrub algorithm [184] to sample protein backbone conformations at the interface. We tested two sets of ΔΔG scores that are output by flex ddG, representing different scoring functions reported by the authors [183]; they are shown as "flex_ddG-fa_talaris2014" and "flex_ddG-fa_talaris2014-gam" in **Table 3.2**.

An example command line used for flex ddG calculations in this study is:

rosetta_scripts.linuxgccrelease -s input.pdb -parser:protocol flexddg.xml -parser:script_vars chainstomove=1,2 mutate_resfile_relpath=input.resfile number_backrub_trials=35000 max_minimization_iter=5000 abs_score_convergence_thresh=1.0 backrub_trajectory_stride=7000 -restore_talaris_behavior -in:file:fullatom -ignore_unrecognized_res -ignore_zero_occupancy false -ex1 -ex2

For ΔΔG calculations in FoldX [178], complex structures were pre-processed using the FoldX RepairPDB protocol, and ΔΔG values were calculated using the FoldX PSSM protocol.

Prior to running ΔΔG calculations in Rosetta for alanine and non-alanine substitutions, antibody-RBD complex structures were pre-processed using Rosetta's FastRelax protocol [166], using the FastRelax flags noted above, to perform constrained backbone and side chain

minimization to resolve unfavorable energies and anomalies that would bias energetic calculations, and to normalize such effects due to the differing resolutions of the experimentally determined structures.

3.3.4 Sequence conservation

Assessment of sequence conservation of SARS-CoV-2 RBD positions in the SARS-CoV-1 sequence was performed using SARS-CoV-2 (GenBank: QHD43416) and SARS-CoV-1 (GenBank: AAP13441) spike reference sequences aligned with BLAST [185]. The epitope residues of each antibody were defined as any SARS-CoV-2 residue within 5 Å of any antibody residue. An in-house Perl script was used to analyze SARS-CoV-2 antibody-antigen interfaces and calculate epitope conservation.

3.3.5 Figures

Figures of structures were generated using PyMOL version 1.8 (Schrodinger, Inc.). Boxplots and dendrograms were generated using the ggplot2 [186] and factoextra [187] packages in R, and heatmaps were generated using the ComplexHeatmap package [188] in R.

3.4 Results

3.4.1 Clustering of antibody-RBD interaction modes

To identify common recognition modes and key features of antibody recognition of the spike glycoprotein, we analyzed a set of high resolution structures of antibody-spike complexes from the CoV3D database [163], which were originally obtained from the Protein Data Bank [164]. We focused on the SARS-CoV-2 RBD, which is the primary target of neutralizing antibodies [189] and is the target of the vast majority of structurally characterized SARS-CoV-2 antibodies.

Structures were filtered by resolution (< 4.0 Å) and nonredundancy, resulting in 70 antibody-RBD complex structures, representing different antibody formats (heavy-light antibody, nanobody) and a range of IGHV genes (**Table 3.1**). As noted in **Table 3.1**, all structures were obtained by X-ray diffraction or cryogenic electron microscopy (cryo-EM), and while the cryo-EM structures had significantly lower resolutions ($p < 0.001$), as expected, antibody-RBD interface size and number of inter-molecular atomic contacts were also somewhat lower for cryo-EM structures, albeit with less significance (**Figure 3.1**). The complex structures in this set include multiple therapeutic monoclonal antibodies that have been under clinical investigation: REGN10933 and REGN10987 (casirivimab/imdevimab; REGN-COV2) [190], LY-CoV555 (bamlanivimab) [191], and S309 which is the basis for VIR-7831 (GSK4182136; sotrovimab) [192].

Table 3.1 Antibody-spike and antibody-RBD complex structures analyzed in this chapter.

Antibody name	PDB code	Type[1]	Species[2]	IGHV gene[2]	Neut[3]	Resolution (Å)[4]	Structure Method[4]	Release Date[4]
Ab2-4	6XEY	ab	human	IGHV1-2	Y	3.25	EM	7/21/20
BD23	7BYR	ab	human	IGHV7-4-1	Y	3.84	X-ray	6/9/20
B38	7BZ5	ab	human	IGHV3-66	Y	1.84	EM	5/12/20
BD-236	7CHB	ab	human	IGHV3-53	Y	2.4	X-ray	9/15/20
BD-368-2	7CHF	ab	human	IGHV3-23	Y	2.67	X-ray	9/15/20
BD-604	7CHF	ab	human	IGHV3-53	Y	2.67	X-ray	9/15/20
BD-629	7CH5	ab	human	IGHV3-53	Y	2.7	X-ray	9/15/20
C105	6XCM	ab	human	IGHV3-53	Y	3.42	EM	6/30/20
CB6	7C01	ab	human	IGHV3-66	Y	2.88	X-ray	5/26/20
CC12.1	6XC3	ab	human	IGHV3-53	Y	2.7	X-ray	7/7/20
CC12.3	6XC4	ab	human	IGHV3-53	Y	2.34	X-ray	7/7/20
COVA2-04	7JMO	ab	human	IGHV3-53	Y	2.36	X-ray	8/25/20

					Y (Brouwer			
COVA2-39	7JMP	ab	human	IGHV3-53	et al., 2020)	1.71	X-ray	8/25/20
CR3022	6YLA	ab	human	IGHV5-51	N	2.42	X-ray	4/14/20
CV30	6XE1	ab	human	IGHV3-53	Y	2.75	X-ray	6/30/20
				IGHV3-30-				
EY6A	6ZCZ	ab	human	3	Y	2.65	X-ray	6/23/20
				IGHV1-69-				
H014	7CAI	ab	human	2	Cross	3.49	EM	9/22/20
H11-D4	6YZ5	nano	llama	IGHV3-3	Y	1.8	X-ray	6/2/20
H11-H4	6ZH9	nano	llama	IGHV3-3	Y	3.31	X-ray	9/1/20
MR17	7C8W	nano	alpaca	IGHV3S53	Y	2.77	X-ray	6/23/20
				IGHV4-38-				
P2B-2F6	7BWJ	ab	human	2	Y	2.85	X-ray	6/2/20
REGN10933	6XDG	ab	human	IGHV3-11	Y	3.9	EM	6/23/20
REGN10987	6XDG	ab	human	IGHV3-30	Y	3.9	EM	6/23/20
S309	7JX3	ab	human	IGHV1-18	Cross	2.65	X-ray	10/14/20
SR4	7C8V	nano	alpaca	IGHV3-3	Y	2.15	X-ray	6/23/20
Ty1	6ZXN	nano	alpaca	IGHV3-48	Y	2.93	X-ray	9/22/20
S2M11	7K43	ab	human	IGHV1-58	Y	2.6	X-ray	10/7/2020
S2E12	7K4N	ab	human	IGHV1-2	Y	3.3	EM	10/7/2020
S2A4	7JVC	ab	human	IGHV3-7	Y	3.3	EM	10/14/20
S2H13	7JV6	ab	human	IGHV3-7	Y	3	EM	10/14/20
COVA1-16	7JMW	ab	human	IGHV1-46	Cross	2.89	X-ray	10/14/20
CV07-250	6XKQ	ab	human	IGHV1-18	Y	2.55	X-ray	10/14/20
CV07-270	6XKP	ab	human	IGHV3-11	Y	2.72	X-ray	10/14/20
S304	7JX3	ab	human	IGHV3-13	Cross	2.65	X-ray	10/14/20
S2H14	7JX3	ab	human	IGHV3-15	Y	2.65	X-ray	10/14/20

C144	7K90	ab	human	IGHV3-53	Y	3.24	EM	10/21/20
C135	7K8Z	ab	human	IGHV3-30	Y	3.5	EM	10/21/20
C121	7K8X	ab	human	IGHV1-2	Y	3.9	EM	10/21/20
C119	7K8W	ab	human	IGHV1-46	Y	3.6	EM	10/21/20
C110	7K8V	ab	human	IGHV5-51	Y	3.8	EM	10/21/20
C002	7K8T	ab	human	IGHV3-30	Y	3.4	X-ray	10/21/20
C102	7K8M	ab	human	IGHV3-53	Y	3.2	X-ray	10/21/20
Sb23	7A29	nano	alpaca	IGHV3-3	Y	2.94	X-ray	10/21/20
C104	7K8U	ab	human	IGHV4-34	Y	3.8	X-ray	10/21/20
298	7K9Z	ab	human	IGHV1-2	Y	2.95	X-ray	10/28/20
52	7K9Z	ab	human	IGHV1-69	Y	2.95	X-ray	10/28/20
mNb6	7KKL	nano	alpaca	IGHV3S53	Y	2.85	X-ray	11/11/20
P2C-1A3	7CDJ	ab	human	IGHV3-11	Y	3.4	X-ray	11/18/20
P2C-1F11	7CDI	ab	human	IGHV3-66	Y	2.96	X-ray	11/18/20
P4A1	7CJF	ab	human	IGHV3-53	Y	2.11	X-ray	11/11/2020
C1A-B12	7KFV	ab	human	IGHV3-53	Y	2.1	EM	12/2/2020
Nb20	7JVB	nano	alpaca	IGHV3-3	Y	3.29	X-ray	12/2/2020
2H2	7DK4	ab	mouse	IGHV2-5-1	Y	3.8	EM	12/2/2020
P17	7CWN	ab	human	IGHV1-69-2	Y	3.2	X-ray	12/16/2020
STE90-C11	7B3O	ab	human	IGHV3-66	Y	2	X-ray	12/16/2020
CR3014-C8	7KZB	ab	human	IGHV3-72	N	2.83	X-ray	2/3/2021
Sb16	7KGK	nano	alpaca	IGHV3S53	Y	2.6	X-ray	2/3/2021
Sb45	7KGJ	nano	alpaca	IGHV3S53	N	2.3	X-ray	2/3/2021
DH1047	7LD1	ab	human	IGHV1-46	Y	3.4	X-ray	1/27/2021
LY-CoV481	7KMI	ab	human	IGHV3-53	Y	1.73	X-ray	1/27/2021
LY-CoV488	7KMH	ab	human	IGHV3-53	Y	1.72	X-ray	1/27/2021

LY-CoV555	7KMG	ab	human	IGHV1-69	Y	2.16	X-ray	1/27/2021
W	7KN7	nano	alpaca	IGHV3-3	Y	2.73	X-ray	1/20/2021
V	7KN6	nano	alpaca	IGHV3S1	Y	2.55	X-ray	1/20/2021
E	7KN5	nano	alpaca	IGHV3-3	Y	1.87	X-ray	1/20/2021
U	7KN5	nano	alpaca	IGHV3-3	Y	1.87	X-ray	1/20/2021
CT-P59	7CM4	ab	human	IGHV2-70	Y (Kim et al., 2021)	2.71	X-ray	1/20/2021
2-15	7L5B	ab	human	IGHV1-2	Y	3.18	X-ray	2/10/2021
15033-7	7KLH	ab	human	IGHV3-23	Y	3	X-ray	2/10/2021
Sb68	7KLW	nano	alpaca	IGHV3S53	Y	2.6	X-ray	2/3/2021

[1]Antibody type. “ab”: heavy-light chain antibody, “nano”: nanobody/VHH.
[2]Species and IGHV gene name determined by ANARCI from antibody heavy chain or nanobody sequence.
[3]Measured SARS-CoV-2 neutralization, from CoV-AbDab [167] or the literature, where specified by a reference. N: does not neutralize SARS-CoV-2; Y: neutralizes SARS-CoV-2; Cross: neutralizes SARS-CoV-2 and SARS-CoV-1.
[4]Resolution, structure determination method and release date of structure in the Protein Data Bank (PDB) [164]. EM: electron microscopy, X-ray: x-ray diffraction.

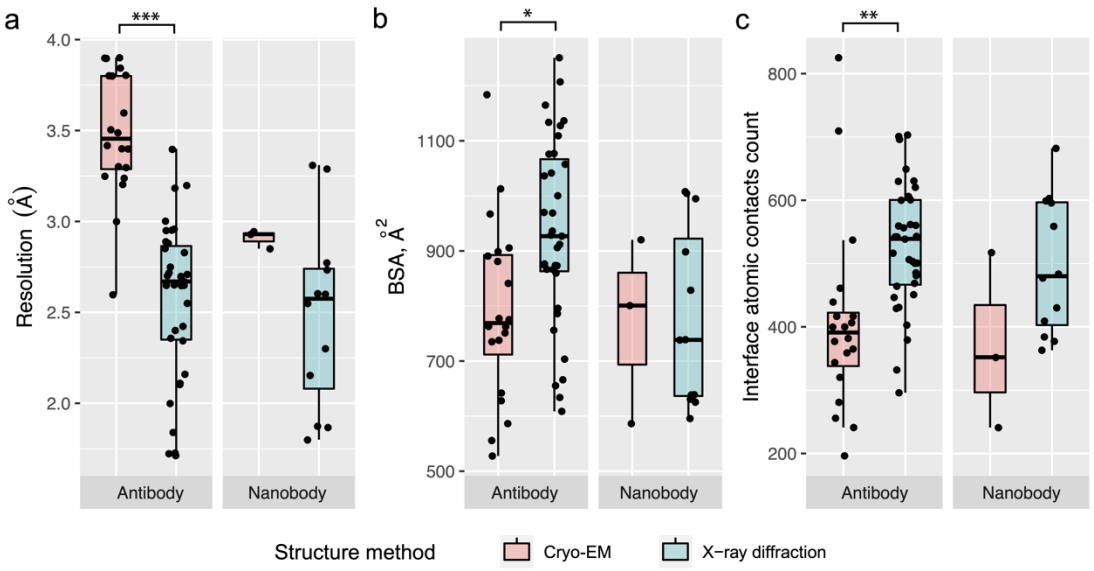

Figure 3.1. Comparison of structural determination methods. (a) Resolution, **(b)** interface buried surface area (BSA), and **(c)** number of interface atomic contacts between antibody and RBD within a 5 Å distance cutoff were compared for structures obtained by cryo-EM and X-ray diffraction. Structures containing antibodies and nanobodies were separated to avoid possible bias in interface size due to smaller size of nanobodies. Statistical significance (Wilcoxon rank-sum test) between properties of cryo-EM and X-ray antibody-RBD structures is indicated at top (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$). Due to small number of values for nanobody cryo-EM complex structures (N = 3), statistical comparisons were not performed for the nanobody-containing structures.

To assess prevalent or shared binding modes in antibody-RBD recognition, pairwise root-mean-square-distances (RMSDs) between antibody heavy chain and nanobody chain orientations were calculated after superposition of RBD coordinates into a common reference frame, and the RMSDs were input to hierarchical clustering analysis (**Figure 3.2**). This analysis identified a set of 17 complexes with a common binding mode and shared heavy chain germline genes (IGHV3-53, IGHV3-66), a feature that has been noted in previous studies describing SARS-CoV-2 antibody-RBD complex structures [193, 194]. Other sets of co-clustered antibodies within the 8 Å RMSD cutoff were limited to antibody pairs, with the exception of a set of five antibodies, of which three (2-15, Ab2-4, C121) share the IGHV1-2 heavy chain germline gene, suggestive of another germline-mediated binding mode. However, other antibodies possessing the IGHV1-2

germline gene exhibited distinct binding modes based on the clustering analysis (298, S2E12), indicating that the heavy chain CDR3 sequence and light chain are relevant factors for that orientation. An example of co-clustered antibodies based on this analysis is shown in **Figure 3.2b**, showing a shared RBD binding mode (heavy chain orientation RMSD: 2.2 Å) for neutralizing antibodies S304 [195] and EY6A [196], and additional examples of co-clustered pairs are shown in **Figure 3.3**.

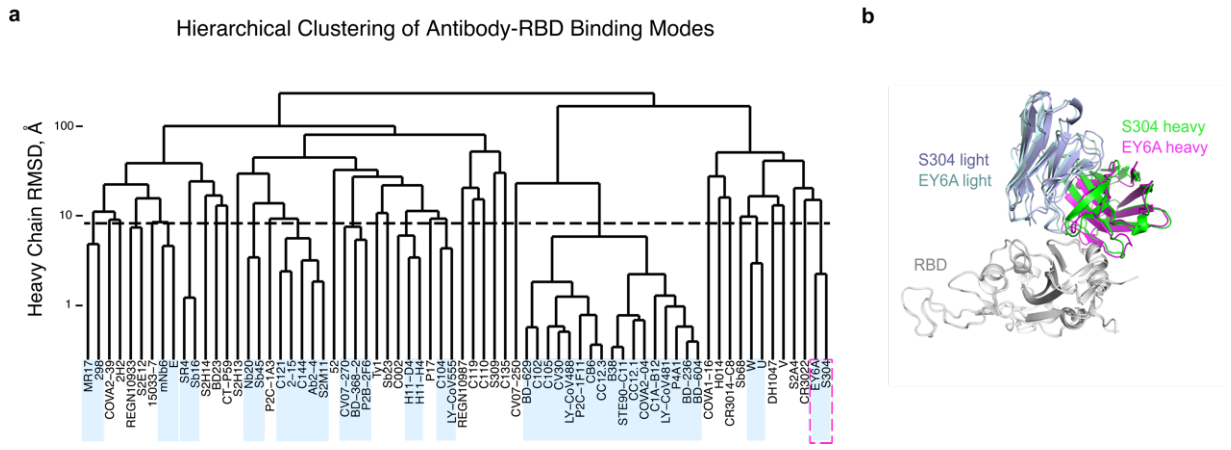

Figure 3.2. Hierarchical clustering of SARS-CoV-2 RBD antibody binding modes. (a) Pairwise root mean square distances (RMSDs) between heavy chain or nanobody binding orientations were determined for 70 antibody-RBD complex structures and used to perform hierarchical clustering. Boxes denote clusters containing multiple antibodies at distance cutoff of 8 Å (shown as dashed horizontal line). (**b**) Example of co-clustered antibodies S304 (PDB code 7JX3) [195] and EY6A (PDB code 6ZCZ) [196] with a shared RBD binding mode (2.2 Å heavy chain orientation RMSD; far right cluster in panel **(a)**, highlighted with dashed magenta lines around the box). Structures are superposed by RBD (gray), and S304 and EY6A heavy and light chains are colored separately as indicated.

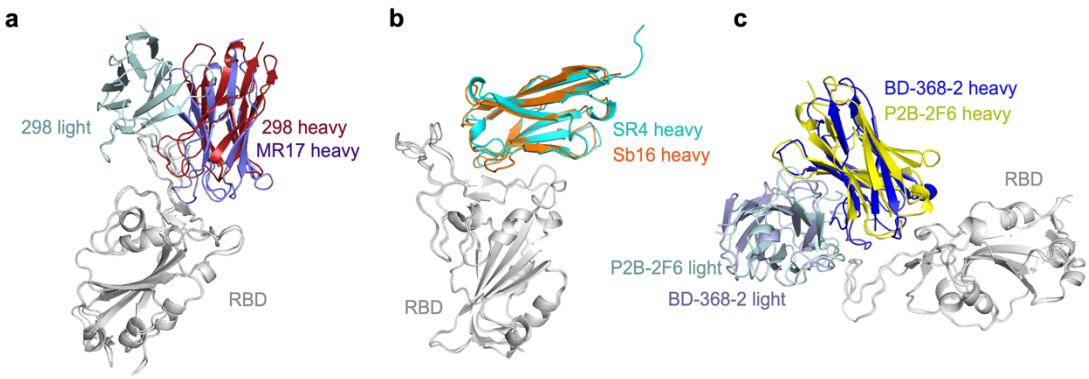

Figure 3.3. Examples of co-clustered antibodies with shared RBD binding modes. (a) The antibody pair MR17 (PDB code 7C8W) [197] and 298 (PDB code 7K9Z) [198] shares a 4.7 Å heavy chain orientation RMSD. **(b)** The antibody pair SR4 (PDB code 7C8V) [197] and Sb16 (PDB code 7KGK [199]) shares a 1.2 Å heavy chain orientation RMSD. **(c)** The antibody pair BD-368-2 (PDB code 7CHF) [200] and P2B-2F6 (PDB code 7BWJ) [201] shares a 5.2 Å heavy chain orientation RMSD. The antibody pair Structures are superposed by RBD (gray), and colored separately as indicated.

3.4.2 High resolution antibody footprinting and clustering analysis

To further delineate features underlying antibody-RBD recognition, we analyzed detailed antibody footprints on the RBD with unsupervised clustering, using the number of atomic contacts by an antibody to each RBD residue as input. Individual antibody footprints and resultant clusters are shown in **Figure 3.4** (a more detailed heatmap including more RBD residues is given in **Figure 3.5**), along with calculated and previously reported properties of the antibodies for reference, including interface buried surface area (BSA), neutralization (SARS-CoV-2 neutralization or SARS-CoV-1/SARS-CoV-2 cross neutralization), ACE2 blocking, and capability to bind the RBD in the context of the closed (or down) spike conformation. This separated the antibodies into four main clusters; these are similar but not identical to previously described SARS-CoV-2 antibody classifications described by Barnes et al. [27], which are shown as the "BBclass" colored sidebar in **Figure 3.5**. Inspection of the heatmap indicates that Clusters 1 and 4 are most distinct, which is supported by high bootstrap confidence levels (100% and 99% respectively; **Figure 3.6**), while

Clusters 2 and 3 are more diverse, and have bootstrap confidence levels of 87% and 83% (**Figure 3.6**). Due to the moderately lower bootstrap confidence, it is possible that some antibodies from Clusters 2 and 3, particularly those proximal to the inter-cluster boundaries and including some cryo-EM structures that have poorer resolutions (**Figure 3.1**), could have potential ambiguity in Cluster 2 versus Cluster 3 assignments. Visualization of the distribution of the antibody positions on the RBD surface (**Figure 3.4**) shows that Clusters 1 and 2 are spatially proximal and overlap with the ACE2 binding site, and the relatively constrained positions of Cluster 1 antibodies are reflective of our RMSD-based analysis (**Figure 3.2**) and known conserved binding mode of that set. Cluster 3 extends to the RBD hinge and N-glycan at RBD position N343, while Cluster 4 occupies a distinct region of the RBD. Principal component analysis using the antibody atom contact data as input enabled visualization of the antibody distributions along the first two principal components, which collectively represent approximately 50% of the data (**Figure 3.7**), and generally supports the hierarchical clustering.

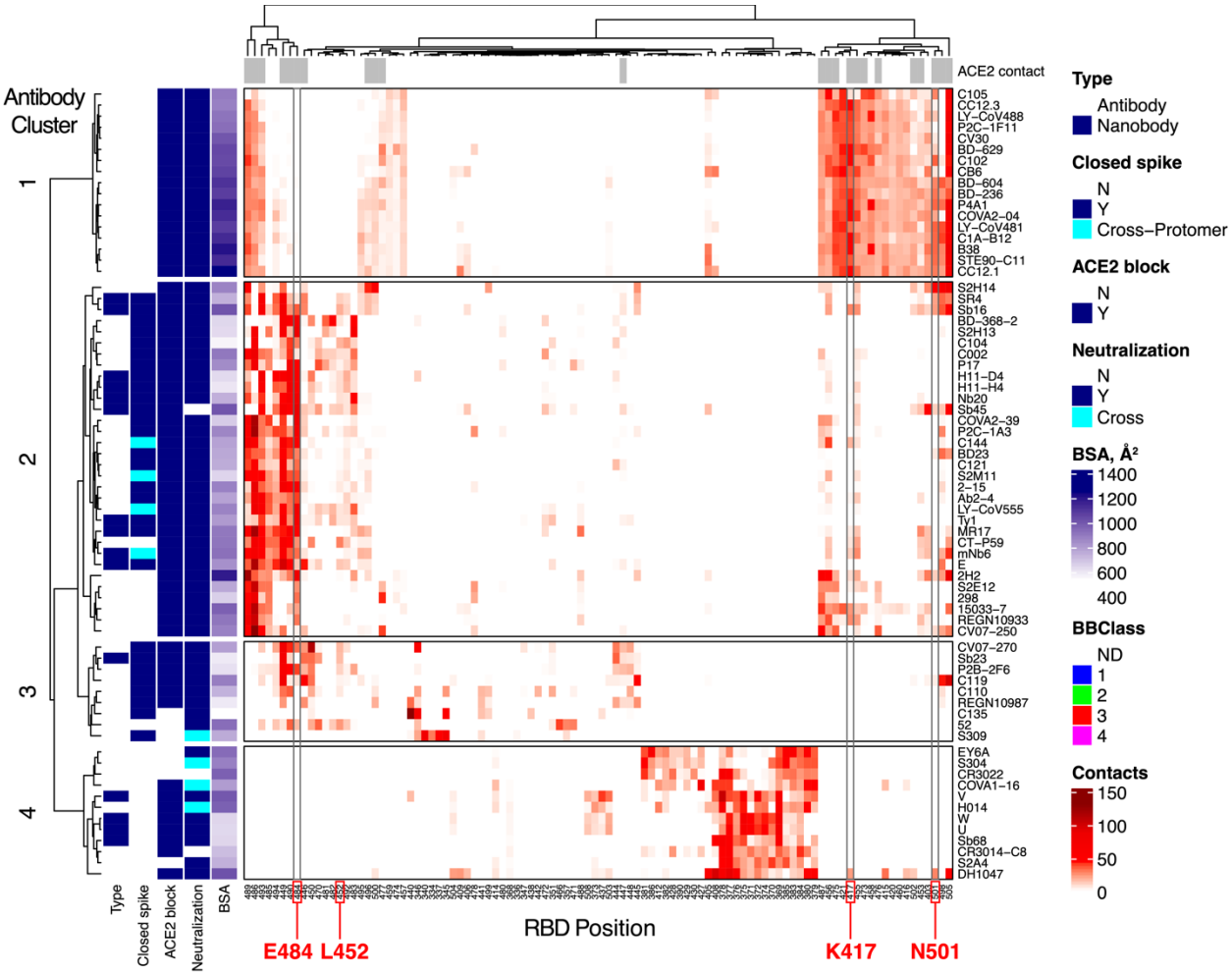

Figure 3.4. High resolution mapping and clustering of SARS-CoV-2 RBD antibody binding. RBD residue contact profiles were generated for each antibody based on number of antibody atomic contacts for each RBD residue within a 5 Å distance cutoff. RBD residues and antibodies are ordered using hierarchical clustering analysis, with dendrograms shown on top and left. The antibodies are separated into four major clusters based on contact profiles, and cluster numbers (1-4) are indicated on left. Contacts in heatmap are colored by number of RBD residue antibody atomic contacts, as indicated in the key. For reference, antibody type (Antibody: heavy-chain antibody, Nanobody: single-chain antibody), binding to RBD-closed spike conformation (Closed spike), ability to block ACE2 binding (ACE2 block) and SARS-CoV-2 neutralization or SARS-CoV-2/SARS-CoV-1 cross-neutralization ("Y" and "Cross", respectively, under Neutralization), interface buried surface area (BSA, Å2) are shown on the left sidebars. Closed spike binding and ACE2 blocking were calculated based on the structures, as described in the Methods. The top bar above the heatmap indicates RBD residues contacted by ACE2 (5 Å distance cutoff) in an ACE2-RBD complex structure (PDB code 6LZG) [172]. For clarity, 100 RBD residues are shown in heatmap; a heatmap with the full set of 139 contacted RBD residues which was used to cluster the antibodies in this figure is shown in **Figure 3.1.** RBD residues that are mutated in SARS-CoV-2 variants of concern (K417, L452R, E484, N501) are labeled at bottom and highlighted with gray boxes in heatmap.

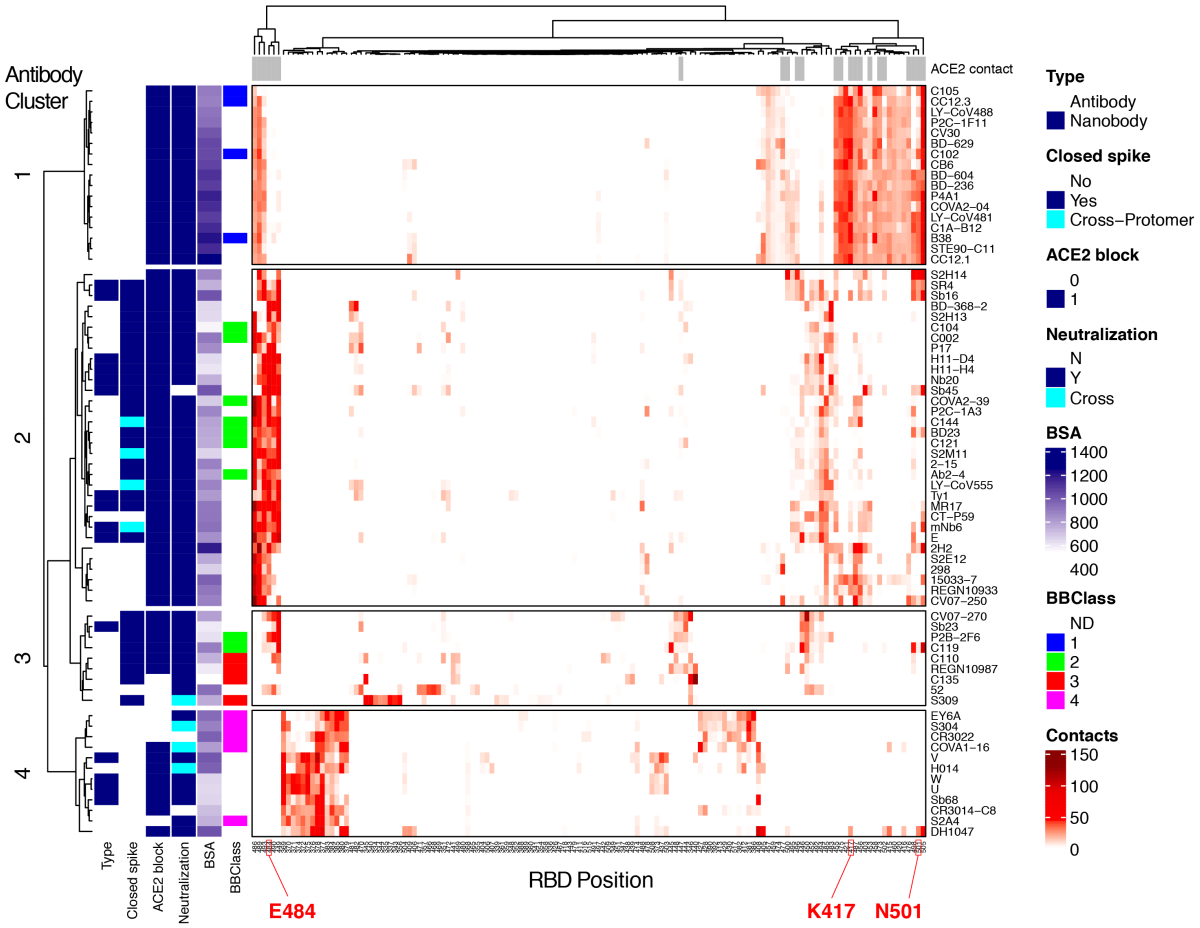

Figure 3.5. Heatmap of antibody-RBD contacts, representing the number of antibody atom contacts with each RBD residue and the full set of 139 contacted RBD positions. Labels and annotations are in accordance with the corresponding labels/annotations in Figure 5.5, and antibodies (rows) and RBD positions (columns) are ordered by hierarchical clustering in R. Antibodies in the heatmap are separated by the four major hierarchical clusters, which are labeled on left.

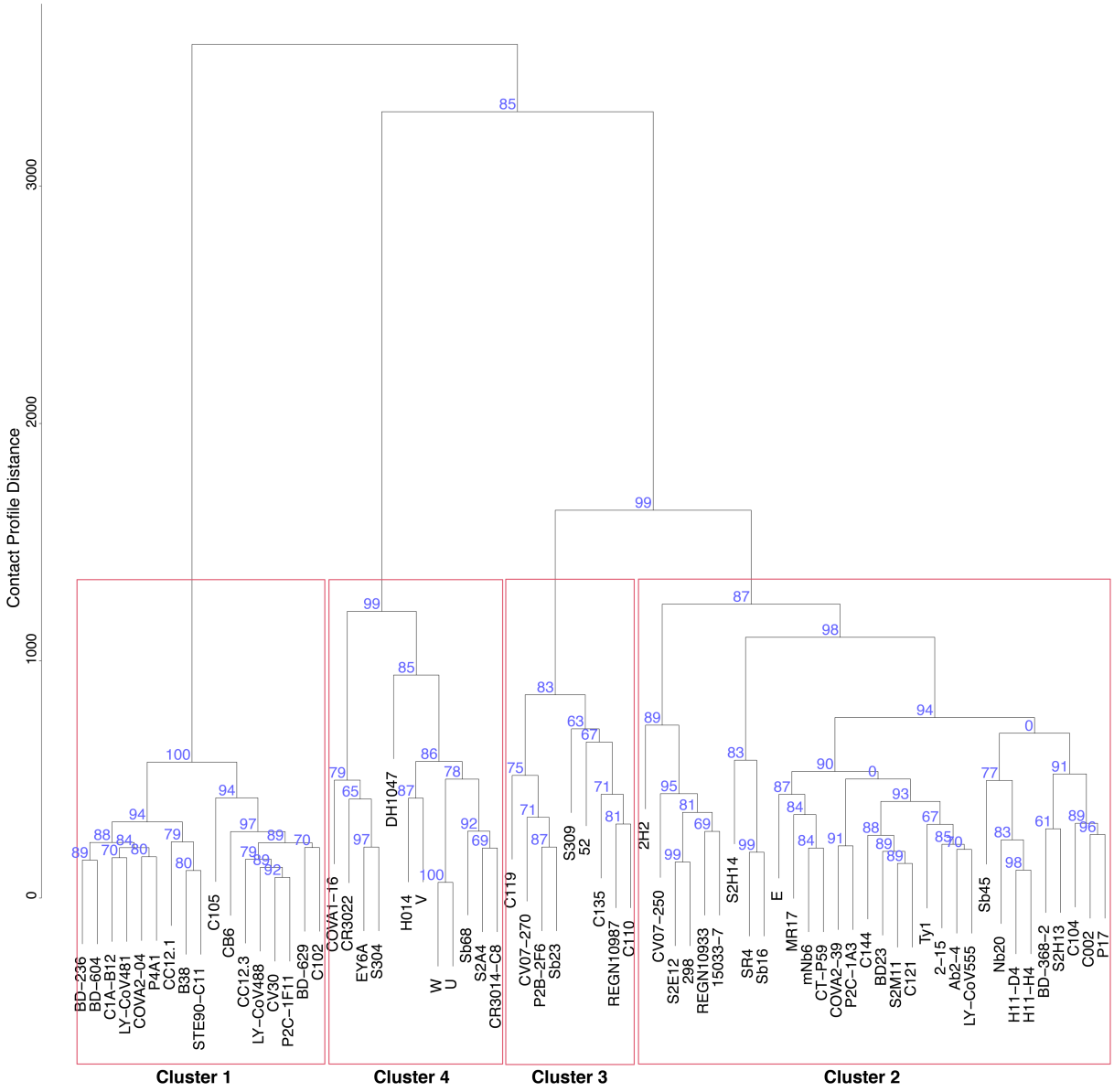

Figure 3.6. Antibody hierarchical clustering bootstrap confidence values, calculated by pvclust [169]. Multiscale bootstrap resampling was performed in pvclust in R, with the antibody-RBD contact data and 10,000 replicates. Values at each node denote the Approximately Unbiased (AU) bootstrap confidence, and red boxes delineate the four major clusters noted in this study, labeled accordingly.

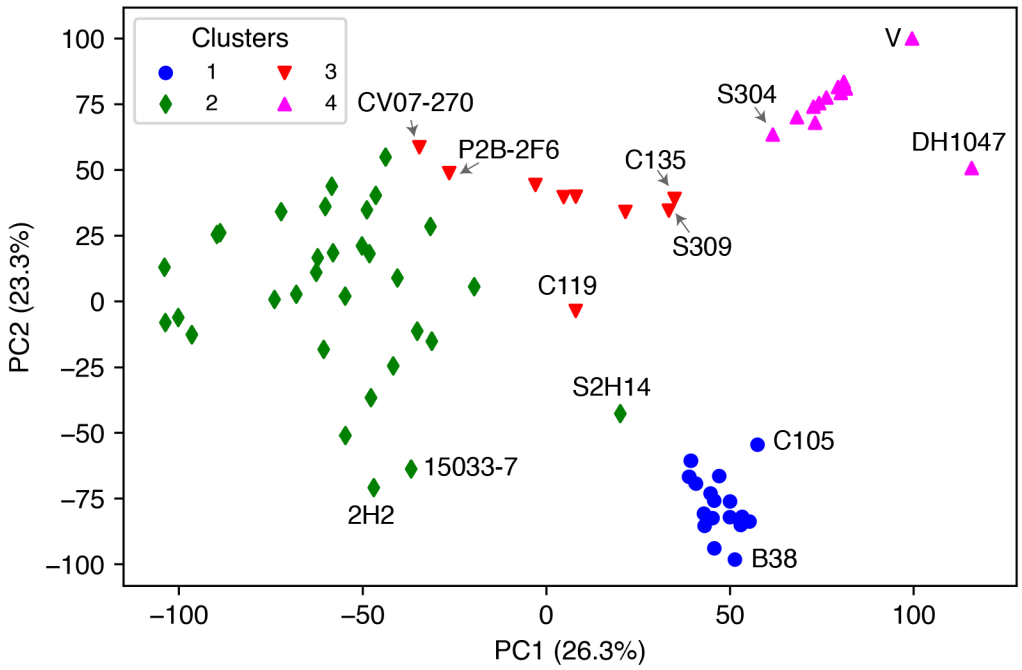

Figure 3.7. Principal component analysis of antibody-RBD residue footprint data. The x and y axes represent the first two principal components (PC1, PC2), with percentage of data variance represented by each principal component shown in parentheses. The 70 antibodies are shown as points, with colors and shapes representing Clusters 1-4, which were determined by hierarchical clustering analysis of antibody-RBD contact profiles. Selected points representing antibodies that are located on the periphery of cluster distributions are labeled by corresponding antibody names.

The contact-based clusters in **Figure 3.4** highlight several notable features within and between sets of RBD-targeting antibodies. Cluster 1 antibodies all neutralize SARS-CoV-2, block ACE2 binding, can only bind the spike in its open conformation, and have relatively high RBD interface buried surface area (BSA). Cluster 2 contains antibodies that can bind the closed spike, some of which can engage multiple RBDs in that context, and all are predicted or confirmed to block ACE2 binding. Cluster 3 is dominated by antibodies that can bind the closed spike, and most Cluster 3 antibodies are predicted to block ACE2 binding through steric hindrance and/or binding site overlap. In Cluster 4, which is mapped closer to the N- and C-termini and the hinge that

connects the RBD to the spike (**Figure 3.8**), multiple antibodies are confirmed to be cross-neutralizing between SARS-CoV-2 and SARS-CoV-1 [195, 202, 203], and no antibodies are predicted to recognize spike in the RBD-closed conformation. The mapped antibody footprints show varying degrees of overlap with ACE2 binding site residues (gray bars at top of **Figure 3.4**) among the clusters. Residues highlighted in **Figure 3.4** that are associated with viral variants of concern (E484, L452, K417, N501) show that Cluster 2 is primarily associated with E484 engagement, Cluster 1 is associated with engagement of K417 and N501, while residue L452, which is mutated in the Delta variant, contacts many of the antibodies in Clusters 2 and 3. Antibodies in Cluster 4 exhibit few or no contacts with those residues, suggesting that they are less susceptible to binding disruption and viral resistance due to variability at those sites.

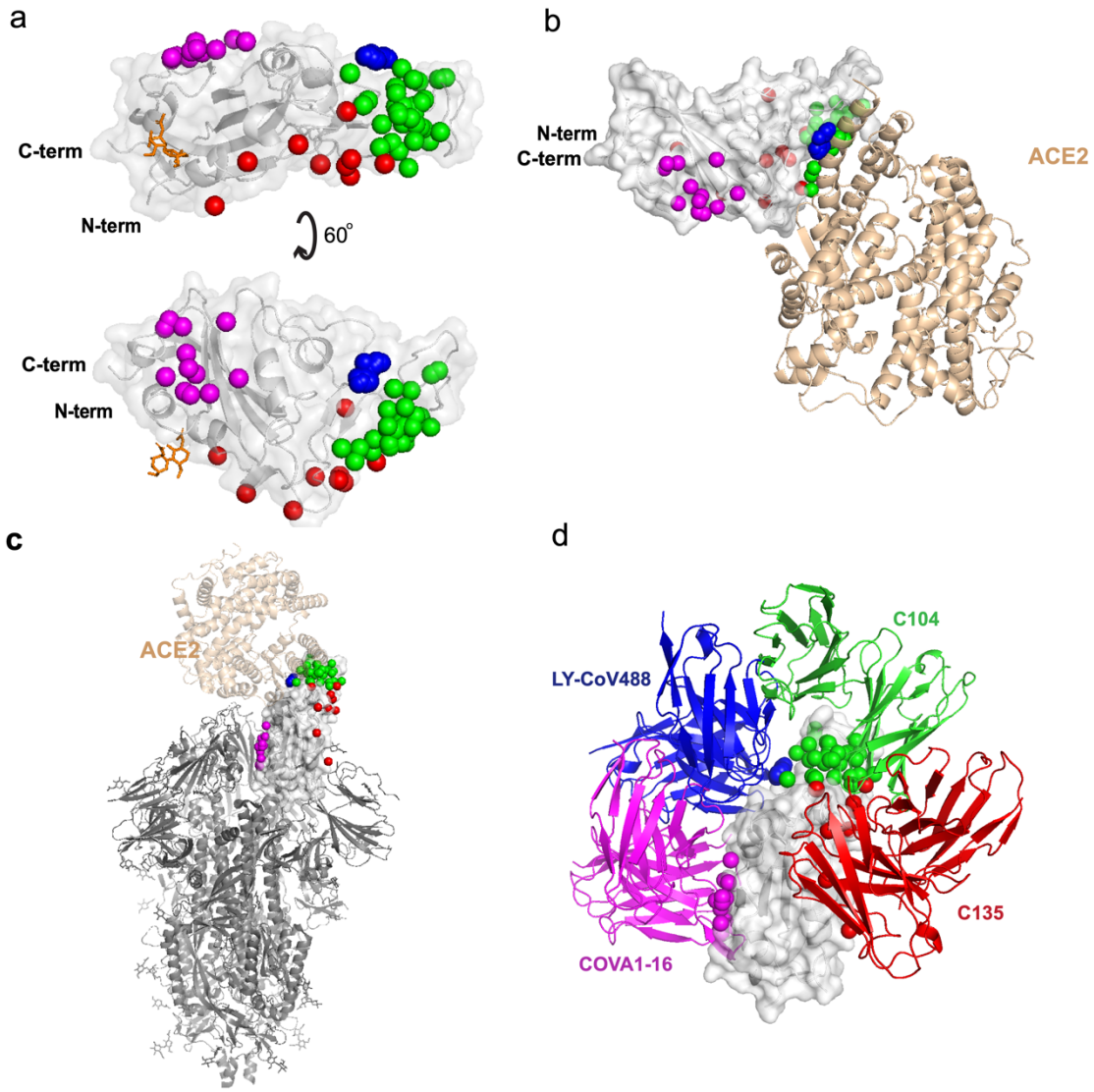

Figure 3.8. Distribution of antibody clusters on the receptor binding domain. (a) Each antibody is represented as a sphere at the paratope center (centroid of all non-hydrogen atoms within 5 Å of the RBD), and colored by contact-based antibody cluster (1: blue, 2: green, 3: red, 4: magenta). A representative RBD structure (from PDB code 7KN5) is shown in gray, and the N-glycan at residue N343 from that structure is shown as orange sticks. **(b)** RBD structure with antibody clusters and superposed ACE2 receptor (tan cartoon; PDB code 6LZG [172]). **(c)** RBD antibody clusters shown in the context of the spike glycoprotein (light blue cartoon; PDB code 6VYB [204]) with the RBD in an open state. **(d)** Representative antibodies from each cluster, labeled by antibody name and colored by cluster, superposed onto the RBD.

3.4.3 Binding energetic features and hotspots

To provide a more detailed and comprehensive view of key residues and energetic features underlying antibody-RBD recognition, all interface structures were analyzed for hydrogen bonds

with RBD residues (**Figure 3.9**) and energetically important RBD residues based on computational alanine scanning (**Figure 3.10**). Hydrogen bonding patterns in RBD-targeting antibodies (**Figure 3.9**) showed clear preferences for hydrogen bond RBD residue interactions among Cluster 1 antibodies, with frequently observed interactions with residues R403, K417, D420, Y421, N487, and Y505. Many Cluster 2 antibodies exhibit hydrogen bond interactions with residue E484 and/or Q493, whereas antibodies from Clusters 3 and 4 have limited shared RBD residues involved in hydrogen bond interactions.

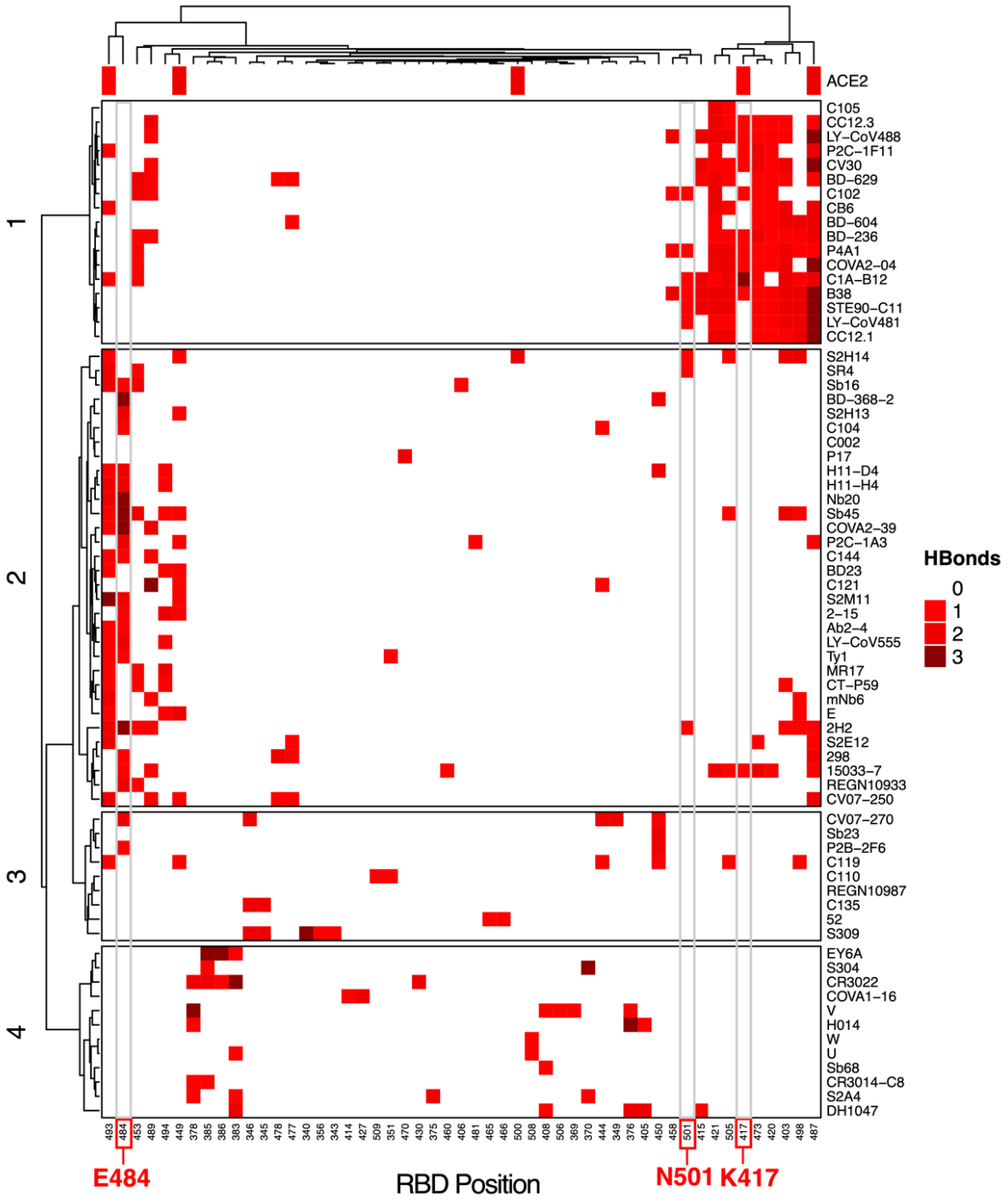

Figure 3.9. RBD hydrogen bond contacts of SARS-CoV-2 antibodies. Hydrogen bonds to RBD residue side chains were calculated for all antibody-RBD complexes using the hbplus program [171]. Each hydrogen bond contact is colored by number of hydrogen bond interactions, as indicated on the key, and RBD positions are ordered by hierarchical clustering based on hydrogen bond profile similarities, with corresponding dendrogram shown at top. Antibodies (rows) are ordered and clustered as in **Figure 3.4**, based on the RBD contact profile similarities, and RBD hydrogen bond contacts with ACE2 (PDB code

6LZG) are shown in the top bar. RBD residues that are mutated in SARS-CoV-2 variants of concern (K417, E484, N501) are labeled at bottom and highlighted with gray boxes in heatmap.

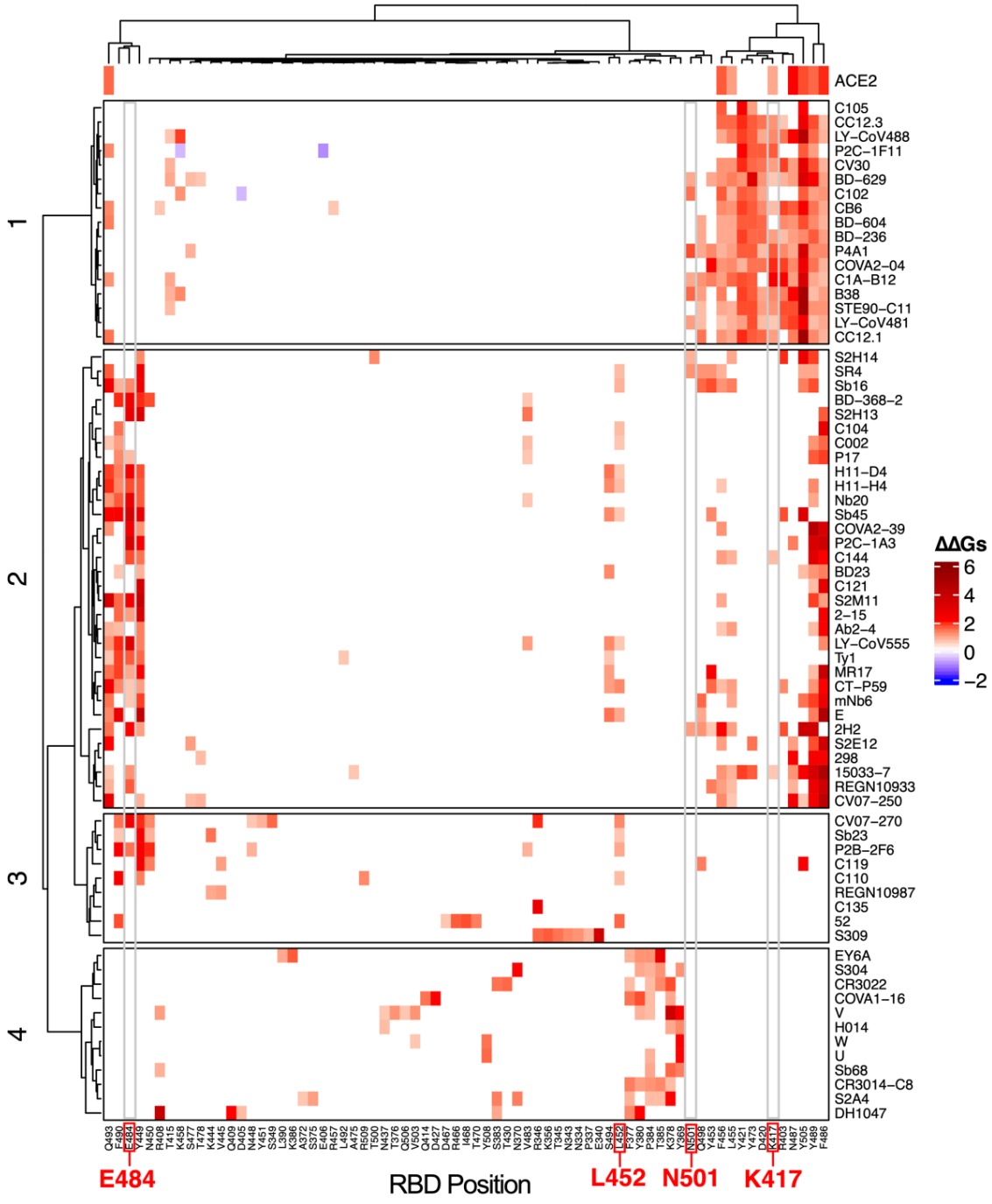

Figure 3.10. Computational mapping of SARS-CoV-2 RBD hotspot residues. Computational alanine scanning of RBD residues in antibody-RBD interfaces was performed using Rosetta [177], to generate binding energy change (ΔΔG) values for alanine substitutions at each RBD position based on modeling of residue substitutions and scoring using an energy-based function. ΔΔG values are in Rosetta Energy Units

(REU) which are comparable to energies in kcal/mol. Alanine residues in the native complex were mutated to glycine for ΔΔG calculations, and glycine RBD residues were omitted from the analysis. In order to highlight substantial predicted binding energy changes, only ΔΔGs with absolute values > 0.5 REU are represented. RBD residues are ordered by hierarchical clustering based on ΔΔG profile similarities, with corresponding dendrogram shown at top. Antibodies (rows) are ordered and clustered as in **Figure 3.4**, based on the RBD contact profile similarities. For reference, ΔΔGs for ACE2 binding based on the ACE2-RBD complex structure (PDB code 6LZG) are shown in the top bar. RBD residues that are mutated in SARS-CoV-2 variants of concern (K417, L452, E484, N501) are labeled at bottom and highlighted with gray boxes in heatmap.

To map key RBD sites and energetic hotspots in the set of antibody-RBD interfaces, we performed computational alanine scanning (**Figure 3.10**) using a mutagenesis protocol in Rosetta [177]. The protocol used for this analysis was selected based on predictive performance from benchmarking of nine computational methods using approximately 350 experimentally determined alanine mutant ΔΔG values for antibody-antigen interfaces (**Table 3.2**). While many energetically important residues identified by this analysis are reflective of the key residues identified by hydrogen bond analysis, including residues N487 and E484 (Cluster 1) and E484 (Cluster 2), numerous hydrophobic RBD residues were additionally identified as important for binding within antibody clusters. These residues include Y505 (Cluster 1), F486 and Y489 (Clusters 1 and 2), and Y449 and F490 (Clusters 2 and 3). As with the analysis of RBD residue contacts, analysis of hydrogen bonds and computational alanine scanning support the overall importance of N417 and Y501 for Cluster 1 antibodies, and E484 for Cluster 2 antibodies. While residue L452 is present in **Figure 3.4**, it is not present in **Figure 3.9** as the hydrophobic leucine residue does not form antibody hydrogen bonds.

Table 3.2 Performance of computational alanine scanning ΔΔG prediction for antibody-antigen interfaces

Method	Correlation[1]	AUC[2]
FoldX	0.49	0.67
Ros2.3_norepack	0.52	0.71
Ros2.3_minint_bb_chi	0.50	0.70
Ros2.3_minint_chi	**0.53**	**0.72**
Ros2.3_repack	0.51	0.70
Ros3_norepack	0.44	0.68
Ros3_repack	0.45	0.69
flex_ddG-fa_talaris2014	0.48	0.66
flex_ddG-fa_talaris2014-gam	0.52	0.69
Including non-alanine mutants		
Ros2.3_minint_chi[3]	0.50	0.70

Computational predictions of binding affinity changes were computed for a subset of the AB-Bind dataset of measured antibody-antigen ΔΔG values [179] with alanine point substitutions, available wild-type complex structures, and quantified ΔΔG measurements (347 total ΔΔGs). Rosetta version 2.3 (Ros2.3) [177], Rosetta version 3.12 (Ros3) [165], and FoldX version 4 [178] were used to predict ΔΔG values using different modeling and scoring protocols, as detailed in the Methods. Protocol selected for RBD alanine scanning based on performance comparison is shown in **bold**.

[1]Pearson correlation coefficient between predicted and experimentally determined ΔΔG values.

[2]ROC AUC value for prediction of hotspot (experimental ΔΔG > 1 kcal/mol) versus non-hotspot residues based on predicted ΔΔG values.

[3]Predictive performance for larger AB-Bind set that includes non-alanine point substitutions (531 mutants and ΔΔG values).

3.4.4 Epitope conservation and targeting of escape variants

To assess the degree to which antibodies of different classes can target sites that are conserved among sarbecoviruses, we calculated the fraction of RBD epitope residues conserved between SARS-CoV-2 and SARS-CoV-1 for each antibody-RBD interface (**Figure 3.11**). Antibodies in Clusters 1-3 exhibit limited conservation (approximately 50% or lower conserved antibody contact residues), with the exception of S309, which shows over 80% epitope residue conservation; this result is in accordance with the observed cross-neutralizing capability for that antibody [28]. In contrast with the other antibody clusters, antibodies in Cluster 4, which includes three confirmed cross-neutralizing antibodies (**Figure 3.4**), exhibit markedly higher epitope conservation, with all values 78% or higher. This suggests that the highly conserved site targeted by Cluster 4 antibodies, which is inaccessible in the closed spike conformation, is potentially important in conferring immunity across sarbecoviruses.

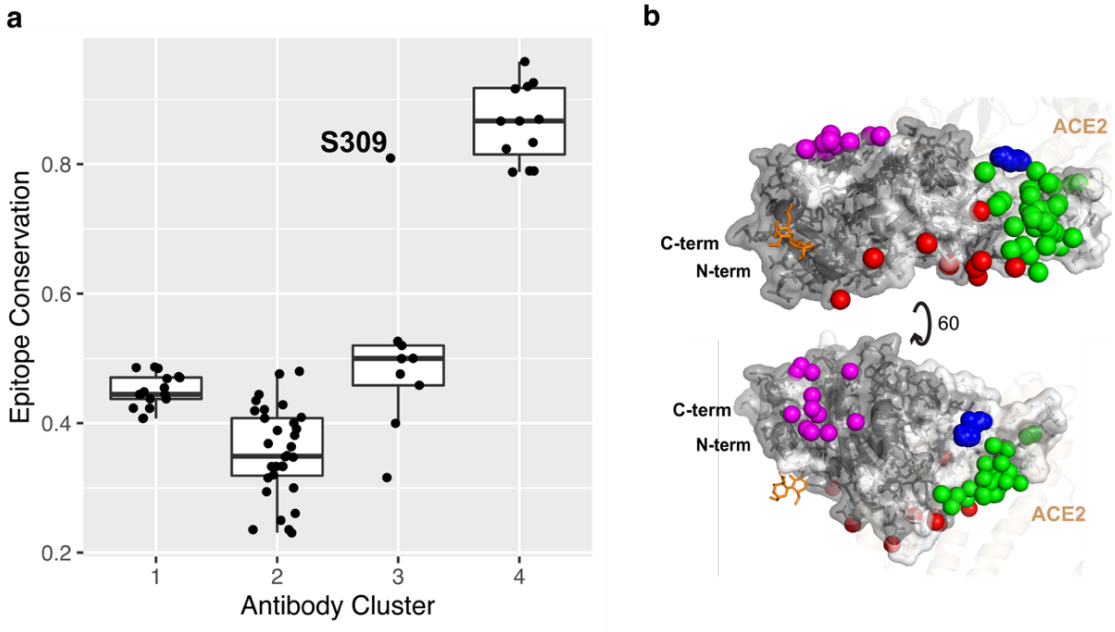

Figure 3.11. Epitope residue conservation in SARS-CoV-1 by antibody cluster. (a) Epitope conservation, defined as the fraction of RBD epitope residues (< 5 Å distance to antibody) conserved between SARS-COV-1 and SARS-COV-2, was calculated for 70 antibody-RBD complex structures, and

conservation values are shown as a boxplot grouped by antibody clusters, with all conservation values shown as points. The outlier point for Cluster 3 (S304 antibody) is labeled, and the total numbers of points are 17 (Cluster 1), 32 (Cluster 2), 9 (Cluster 3), and 12 (Cluster 4). (**b**) Conserved RBD residues are highlighted on the RBD structure, with conserved RBD residues shown as orange and non-conserved residues gray, and represented as in **Figure 3.8a** with antibody cluster paratopes as spheres.

To directly assess the effects of RBD mutations present in recently described SARS-CoV-2 variants of concern, we performed computational mutagenesis to gauge whether antibody binding affinities are predicted to be disrupted by individual RBD substitutions, as well as effects on ACE2 binding. For initial simulations, we utilized the same protocol that was used for computational alanine scanning; we found this method to have similar predictive performance for point residue substitutions to all residue types in comparison with performance for alanine-only substitutions (Pearson Correlation Coefficient (PCC) with experimental ΔΔGs of 0.5 for all residues, versus 0.53 for alanine-only; **Table 3.2**). RBD substitutions K417N, K417T, L452R, S477N, T478K, E484K, E484Q, and N501Y were modeled in all interfaces and assessed for antibody and ACE2 ΔΔGs; these substitutions are collectively represented in variants of concern Alpha (B.1.1.7; N501Y), Beta (B.1.351; K417N, E484K, N501Y), Gamma (P.1; K417N, E484K, N501Y), and Delta (B.1.617.2; L452R, T478K), and variant of interest Kappa (B.1.617.1; L452R, E484Q). Comparison of predicted ΔΔGs (**Figure 3.12**) indicates that K417N, K417T, and to a lesser extent N501Y, are predicted to predominantly affect antibodies in Cluster 1, whereas disruptive effects of E484K and E484Q are primarily observed for antibody Cluster 2. Cluster 3 antibodies with predicted ΔΔG values of over 1 Rosetta Energy Unit (REU) were observed for E484 substitutions, but were very limited (two antibodies for E484K, one antibody for E484Q). In contrast, antibodies in Cluster 3 and 4 exhibit little overall predicted effects from the variant RBD point substitutions considered here, and other variants substitutions did not show marked predicted effects on antibody binding. The binding affinity for ACE2 was predicted to decrease for

substitutions K417N and K417T, and increase for N501Y, while remaining the same for other substitutions. This is in accordance with recently reported ACE2-RBD binding measurements, where N501Y led to a 2-fold improvement in ACE2 binding, K417N led to a 7-fold loss in ACE2 binding, and E484K maintained ACE2 binding (< 2-fold affinity change) [205]. We also tested predicted binding effects using a different modeling tool (FoldX), which uses a distinct modeling and scoring protocol from Rosetta, and found similar trends among antibody classes for the effects of the variants (**Figure 3.13**). However, there are some differences between FoldX and Rosetta ΔΔG predictions, such as the L452R RBD variant, for which FoldX predicted more antibody binding disruptions than Rosetta.

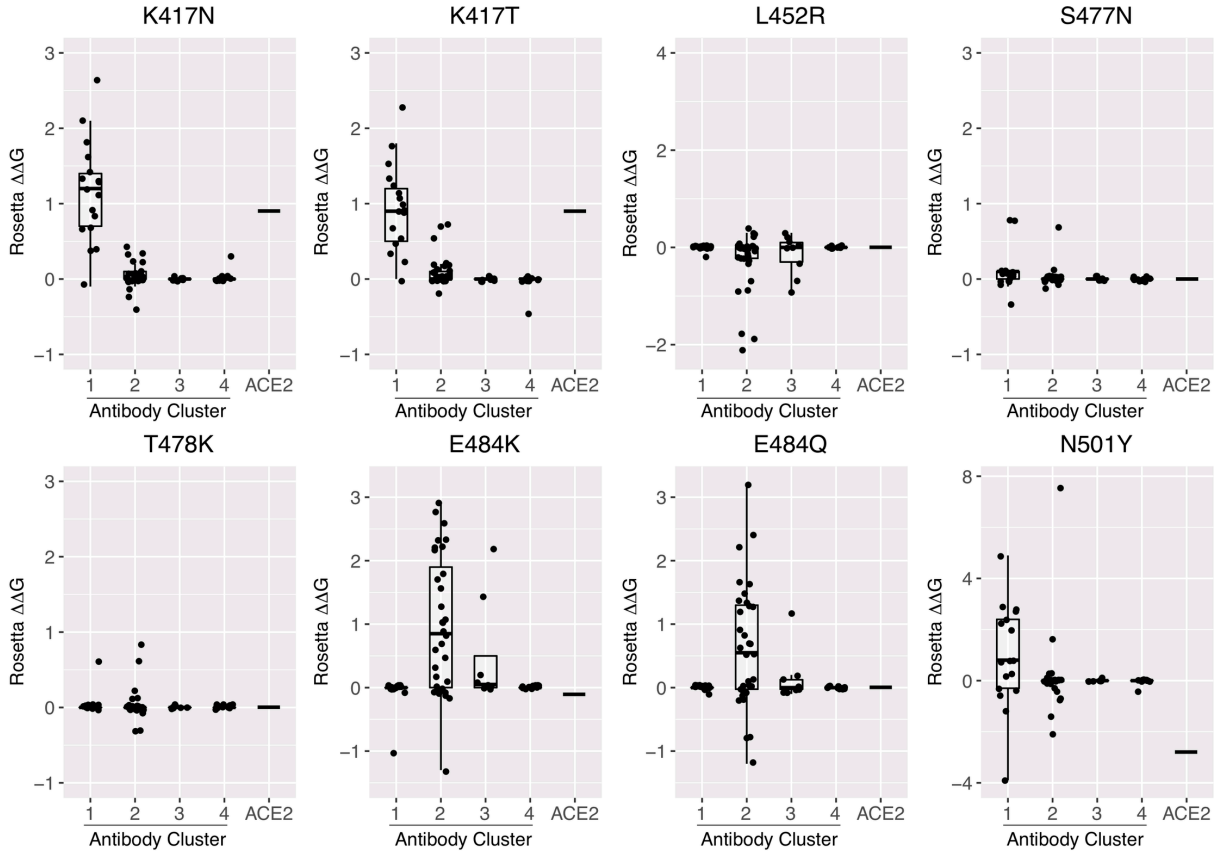

Figure 3.12. **Profiling antibody and receptor binding effects of RBD point substitutions from circulating SARS-CoV-2 variants.** Computational mutagenesis in Rosetta [177] was used to predict

binding affinity effects (ΔΔGs) of RBD variant substitutions K417N, K417T, L452R, S477N, T478K, E484K, E484Q, and N501Y for 70 antibodies that target the RBD, as well as the ACE2 receptor. ΔΔG values are shown as boxplots grouped by antibody clusters, with all antibody ΔΔG values shown as points, and the ACE2 ΔΔG value represented as a horizontal bar in each boxplot. ΔΔG values are in Rosetta Energy Units (REU), which are comparable to energies in kcal/mol.

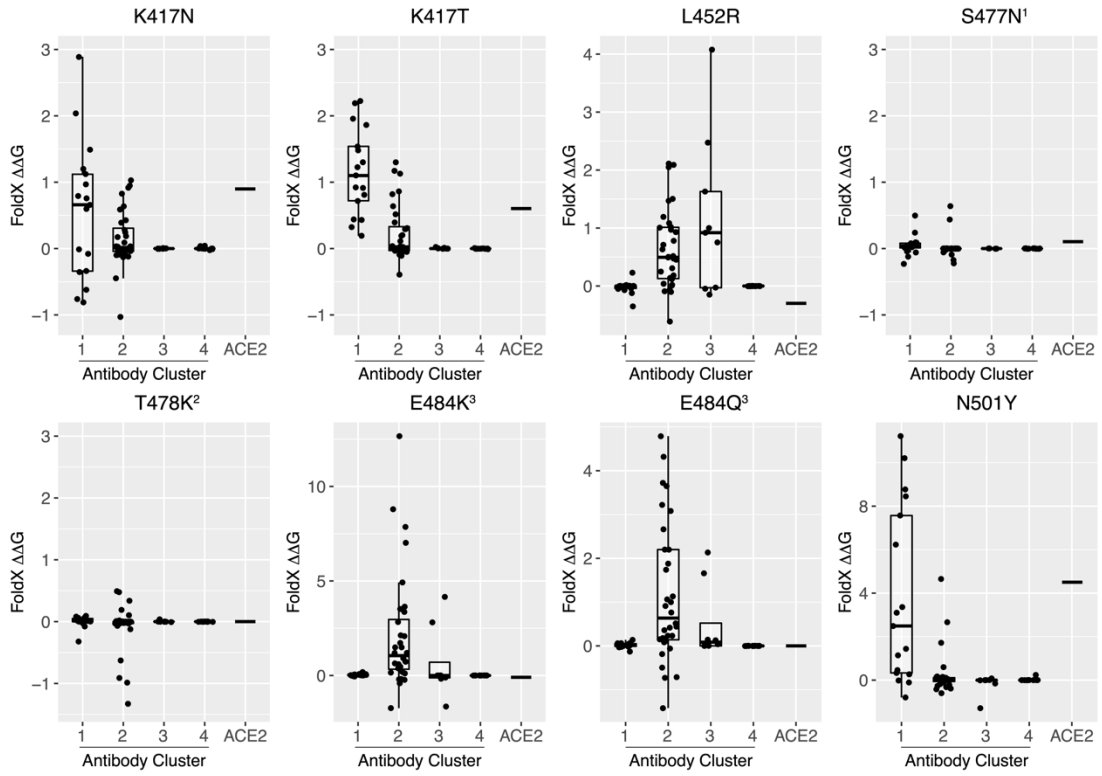

Figure 3.13. Computational assessment of antibody and ACE2 receptor ΔΔG values for RBD variants using FoldX. FoldX [178] was used to simulate and compute binding affinity changes (ΔΔGs, in units of kcal/mol) for RBD point substitutions in 70 antibody-RBD complex structures and the ACE2-RBD complex structure (PDB code 6LZG). ΔΔG values for each RBD substitution are shown as a separate boxplot, with antibodies grouped by contact based cluster (1-4). The ACE2 ΔΔG value for each RBD point substitution is shown as a horizontal bar.

To assess the effects of SARS-CoV-2 variants more directly on antibody binding, we calculated ΔΔGs for combinations of RBD substitutions found in variants of concern Beta, Gamma, and Delta (**Figure 3.14**). Binding effects for Alpha, which is equivalent to N501Y in **Figure 3.12** as it contains the same RBD substitution, are also shown in **Figure 3.14** for reference. Based on comparison of ΔΔG predictions with recently published experimentally measured

neutralization results for variants and monoclonal antibodies overlapping with the set in this study (**Tables 3.3 and 3.4**), FoldX was included along with Rosetta in **Figure 3.14**, as the former showed a modest improvement in sensitivity over the Rosetta ΔΔG protocol, detecting one more antibody-variant pair with loss of neutralization in each of **Tables 3.3 and 3.4**. Overall, the comparison of measured neutralization changes and predicted ΔΔG values shows that the structure-based affinity predictions can in most cases reflect neutralization effects. Predicted ΔΔGs for the antibody clusters from Rosetta (**Figure 3.14a**) indicated that the Alpha, Beta, and Gamma variants are disruptive for Cluster 1 antibodies, and Beta and Gamma variants are disruptive for Cluster 2 antibodies; those results are generally in agreement with FoldX (**Figure 3.14b**). However, in contrast with Rosetta which predicted minor effects from the Delta variant on antibody recognition, FoldX predicted that the Delta variant would markedly disrupt antibody binding in Clusters 2 and 3. Given its modestly higher performance in the comparison with experimentally determined variant neutralization effects (**Tables 3.3 and 3.4**), the predictions of disruption from the FoldX protocol seem more likely to reflect the antibody binding effects for that variant.

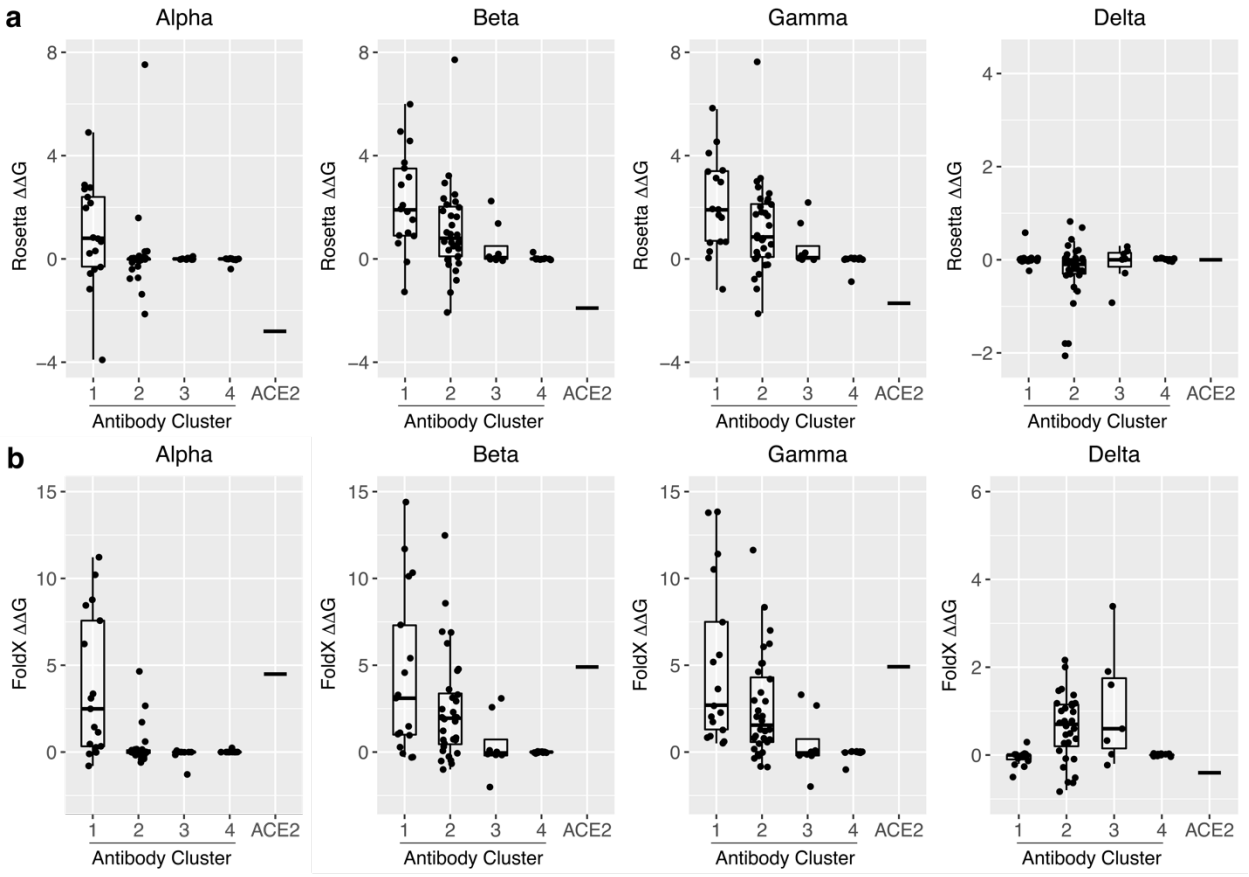

Figure 3.14. **Profiling antibody and receptor RBD binding effects for circulating SARS-CoV-2 variants.** Computational mutagenesis was used to predict binding affinity effects (ΔΔGs) of SARS-CoV-2 variants of concern Alpha (B.1.1.7; RBD substitution N501Y), Beta (B.1.351; RBD substitutions K417N, E484K, N501Y), Gamma (P.1; RBD substitutions K417N, E484K, N501Y), and Delta (B.1.617.2; RBD substitutions L452R, T478K), using **(a)** Rosetta and **(b)** FoldX. ΔΔG values are shown as boxplots grouped by antibody clusters or ACE2 receptor, with all antibody ΔΔG values shown as points, and the ACE2 ΔΔG value represented as a horizontal bar in each boxplot. Both Rosetta and FoldX ΔΔG values are commensurate with energies in kcal/mol.

Table 3.3 Comparison of ΔΔG predictions with measured monoclonal antibody neutralization of SARS-CoV-2 variants from Wang et al. [156].

	N501Y			K417N			E484K			Alpha			Beta		
Antibody	Exp[1]	Ros[2]	Fold X[3]	Exp[1]	Ros[2]	Fold X[3]	Exp[1]	Ros[2]	Fold X[3]	Exp[1]	Ros[2]	Fold X[3]	Exp[1]	Ros[2]	Fold X[3]
2-15	1.5	0.0	0.1	3.3	0.0	-0.1	N	0.6	0.2	-3	0.0	0.1	N	0.6	-0.5
LY-CoV555	-1	0.0	0.0	8.4	0.1	-0.1	N	2.9	12.7	-2.8	0.0	0.0	N	2.9	12.5
REGN10933	-1.4	-0.1	-0.6	-13.1	0.1	1.0	-10.5	1.6	1.2	1	-0.1	-0.6	N	1.6	1.8
C121	1.5	0.0	0.1	1.2	0.0	-0.1	N	0.0	-0.2	4	0.0	0.1	N	0.1	0.1
REGN10987	1.3	0.0	0.1	-1.2	0.0	0.0	-1.1	0.0	0.0	1	0.0	0.1	-3.5	0.0	0.1
S309	1.2	0.0	0.0	1.6	0.0	0.0	2.5	0.0	0.0	-4	0.0	0.0	-2.2	0.0	0.0
COVA1-16	-1.4	0.0	0.0	3.3	0.0	0.0	-1	0.0	0.0	3.4	0.0	0.0	1	0.0	0.0

[1]Experimentally determined neutralization of viral variant, from Wang et al. [206] (Fig 2a in that study). Values reflect fold change in antibody neutralization (IC_{50}) for variant versus wild-type virus, with negative values indicating lower neutralization (higher IC_{50}). "N" indicates unquantifiable neutralization (< -1000) in Wang et al.. Measurements of greater that 10-fold loss of neutralization (< -10, or "N") are highlighted by red cells.
[2]Rosetta ΔΔG for viral variant based on mutagenesis of RBD, in Rosetta Energy Units (REU) which are comparable to energies in kcal/mol. Predicted disruptive effects (ΔΔG > 1.0) have cells shaded red.
[3]FoldX ΔΔG for viral variant based on mutagenesis of RBD, in units of kcal/mol. Predicted disruptive effects (ΔΔG > 1.0) have cells shaded red.

Table 3.4 Comparison of ΔΔG predictions with measured monoclonal antibody neutralization of SARS-CoV-2 variants from Planas et al. [157]

	Alpha			Beta			Delta		
Antibody	**Exp[1]**	**Ros[2]**	**FoldX[3]**	**Exp[1]**	**Ros[2]**	**FoldX[3]**	**Exp[1]**	**Ros[2]**	**FoldX[3]**
LY-CoV555	Y	0	0	N	2.9	12.5	N	-0.2	1.5
REGN10933	Y	-0.1	-0.6	N	1.6	1.8	Y	0	-0.8
REGN10987	Y	0	0.1	Y	0.1	0.1	Y	0	-0.2

[1]Experimentally determined neutralization of viral variant, from Planas et al. [207] (Fig 1 in that study). Y: antibody neutralization; N: low or no antibody neutralization (cells shaded red).
[2]Rosetta ΔΔG for viral variant based on mutagenesis of RBD, in Rosetta Energy Units (REU) which are comparable to energies in kcal/mol. Predicted disruptive effects (ΔΔG > 1.0) have cells shaded red.
[3]FoldX ΔΔG for viral variant based on mutagenesis of RBD, in units of kcal/mol. Predicted disruptive effects (ΔΔG > 1.0) have cells shaded red.

3.5 Discussion

Utilizing a curated set of experimentally determined antibody-RBD complex structures, we have performed detailed mapping of antibody recognition determinants on the SARS-CoV-2 RBD, which were used to identify antibody clusters that exhibit distinct structural and energetic signatures. Notably, these clusters exhibited different destabilizing effects for RBD substitutions found in circulating variants, expanding upon previous observations by others on the effects of specific substitutions such as E484K for specific groups of antibodies [27, 205, 206]. We found that Cluster 2 antibodies, which overlap with Class 2 antibodies reported by Barnes et al. [27], are susceptible to resistance from SARS-CoV-2 variants with the E484K substitution, which include Beta (B.1.351) and Gamma (P.1), whereas other antibodies are not likely to be affected by that substitution. In contrast, substitutions at residues K417 and N501, which are found in several variants of concern, were primarily associated with binding disruption for Cluster 1 antibodies

based on our computational mutagenesis. Given that the E484K substitution appears specifically associated with viral escape, as noted by others [208] and supported by recent studies of monoclonal and polyclonal antibody neutralization of variant viruses and specific mutants [157, 158], our work highlights the relative importance of Cluster 2 antibodies in the neutralizing response against SARS-CoV-2 due to natural infection or immunization.

Our analysis highlights the ability of computational structure-based protocols to rapidly predict and profile resistance for new and emerging SARS-CoV-2 variants. This is exemplified by our results for the Delta variant, which was designated a variant of concern (VOC) in May 2021 and is responsible for a recent global rise in COVID-19 cases. We found that the Delta variant is predicted to be resistant to antibodies in Clusters 2 and 3, and this is likely driven by the L452R RBD substitution. This resistance is corroborated by recent reports of monoclonal antibody resistance [209] and lower neutralization in vaccinated individuals [210] for the Delta variant, which can lead to breakthrough infections in some cases [211]. Based on our predictions for the effects of K417N (Figs 7 and S6), the "Delta plus" variant which includes that mutation would likely exhibit resistance to additional antibodies, including antibodies in Cluster 1, albeit with possibly reduced ACE2 binding.

This study is distinguished from other recently described structure-based [27] and binding competition-based [195, 212] reports to compare and classify antibodies that target SARS-CoV-2, as we directly assessed detailed antibody binding signatures, generated using atomic contact counts to RBD residues, and used unsupervised clustering with these features to generate the resultant classes. Furthermore, we generated an energetic map of RBD antigenicity based on comprehensive computational alanine scanning mutagenesis. To provide an updated reference to the community, we report these clusters on our CoV3D site of coronavirus protein structures [163]

(https://cov3d.ibbr.umd.edu/antibody_classification), which includes the 70 complexes reported in this study as well as newly reported complexes. We also provide a prototype interface on the CoV3D site for the community to input new experimentally determined structures or models of antibody-RBD or protein-RBD complexes to characterize binding footprints and assign contact-based clusters.

Certain elements of our analysis of antibody binding determinants can be expanded in future studies. The calculation of antibody contacts and energetic determinants on the RBD did not include non-protein atoms, such as water molecules and N-glycans, and in some cases, certain residues were disordered in the experimentally determined structures. Water molecules, which could mediate hydrogen bonds between antibody and RBD, were not included here, to avoid bias due to varying experimental structural resolutions which in many cases could not resolve water molecules, necessitating modeling of explicit water molecules which would lead to additional uncertainties in subsequent calculations [213]. Likewise, the N-glycans of the RBD, specifically the glycan at residue N343, has varying occupancies in experimentally determined structures. Though this glycan is contacted by the S309 antibody [28], such glycan contacts appear to be rare in antibody-RBD complex structures, at least for structurally characterized neutralizing antibodies, of which most compete with ACE2 binding and thus target sites that are not proximal to that N-glycan. Modeling of missing N-glycans, water molecules, and any missing residues may still provide possible insights into recognition features, as well as simulations of interface molecular dynamics, or docking simulations of separated antibody and RBD molecules to assess binding energy funnels [71]. Predictive computational docking and template-based modeling can also be used to generate antibody-RBD complex models for antibodies with sequences available but no known structure, enabled in part by databases containing sequences of RBD-targeting antibodies

[167]. An additional avenue for expansion would be the analysis of antibodies that target other regions of the spike glycoprotein, including the N-terminal domain (NTD), which have been described in recent structural and antigenic mapping studies [214, 215]. We currently represent this set as the "NTD" antibody class on the CoV3D site, and may perform a more detailed energetic and footprinting analysis of this set in the future.

In addition to providing a view of the detailed landscape of antibody-RBD recognition determinants and key sites, our results indicate that certain sets of antibodies are less susceptible to resistance from variants and have higher average epitope sequence conservation with SARS-CoV-1. Furthermore, several of the antibodies in Cluster 4 have been experimentally confirmed cross-neutralize SARS-CoV-1 and SARS-CoV-2. Recently reported broadly reactive RBD-binding antibodies that recognize human and zoonotic SARS-like coronaviruses (sarbecoviruses) [216-218] can provide additional structural data to map these key conserved regions and epitopes. Prospective structure-based antigen design studies could potentially focus the antibody response to the corresponding epitopes of the SARS-CoV-2 RBD, versus the epitopes collectively targeted by antibodies in Clusters 1 and 2. As binding of Cluster 4 antibodies is prevented in the context of the closed-RBD spike conformation, open spike antigen designs or RBD-only antigens would likely facilitate elicitation of these antibodies. Several recent studies have reported success using RBDs displayed on self-assembling nanoparticles [219-221], and structure-guided RBD optimization in the context of such a platform could lead to improved elicitation of antibodies associated with a cross-sarbecovirus response. Such antigen design efforts could result in an effective vaccine that provides protection against SARS-CoV-2 variants as well as future emerging coronaviruses.

Chapter 4: Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants

4.1 Abstract

High resolution experimental structural determination of protein-protein interactions has led to valuable mechanistic insights, yet due to the massive number of interactions and experimental limitations there is a need for computational methods that can accurately model their structures. Here we explore the use of the recently developed deep learning method, AlphaFold, to predict structures of protein complexes from sequence. With a benchmark of 152 diverse heterodimeric protein complexes, multiple implementations and parameters of AlphaFold were tested for accuracy. Remarkably, many cases (43%) had near-native models (Medium or High CAPRI accuracy) generated as top-ranked predictions by AlphaFold, greatly surpassing the performance of unbound protein-protein docking (9% success rate for near-native top-ranked models), however AlphaFold modeling of antibody-antigen complexes within our set was unsuccessful. We identified sequence and structural features associated with lack of AlphaFold success, and we also investigated the impact of multiple sequence alignment input. Benchmarking of a multimer-optimized version of AlphaFold (AlphaFold-Multimer) with a set of recently released antibody-antigen structures confirmed a low rate of success for antibody-antigen complexes (11% success), and we found that T cell receptor-peptide-MHC complexes are likewise not accurately modeled by that algorithm, showing that adaptive immune recognition poses a challenge for the current AlphaFold algorithm and model. Overall, our study demonstrates that end-to-end deep learning can accurately model many transient protein complexes, and highlights

areas of improvement for future developments to reliably model any protein-protein interaction of interest.

4.2 Introduction

Protein-protein interactions are the basis of many critical and fundamental cellular and molecular processes, including inhibition or activation of enzymes, cellular signaling, and recognition of antigens by the adaptive immune system. High resolution structural characterization of these interactions provides insights into their molecular basis, as well as structure-guided design of binding affinities and identification of inhibitors. However, structures for large numbers of molecular interactions remain undetermined experimentally, due to limitations in resources, and the challenges of structural determination techniques.

In response to this need, numerous predictive computational methods to model structures of protein-protein complexes have been developed over several decades, including protein docking methods that use unbound or modeled component structures as input to perform rigid-body global searches in six dimensions[61-65], and template-based modeling methods that generate models of complexes based on known structures[57, 58]. Challenges for docking algorithms include side chain and backbone conformational changes between unbound and bound structures, large search spaces, and inability to capture key energetic features in grid-based and other rapidly computable functions, leading to false positive models among top-ranked models or lack of any near-native models within large sets of predicted models. Developments such as explicit side chain flexibility during docking searches[77], use of normal mode analysis to represent protein flexibility[69, 222], clustering[67, 72] or re-scoring[73-76] docking models to improve ranking of near-native models, and use of experimental data as restraints for docking[66] have led to some improvement in docking success, and examples of these and other advances specifically designed to address the

challenge posed by protein backbone flexibility are highlighted in a recent review [78]. However, the Critical Assessment of Predicted Interactions (CAPRI) blind docking prediction experiment[84] and several protein docking benchmarks[87, 88], which have enabled the systematic assessment of predictive docking performance, revealed persistent shortcomings of current computational docking approaches. Several protein-protein complex targets had no accurate model generated by any teams in a set of recent CAPRI rounds[223], while benchmarking of multiple docking algorithms in 2015 showed no accurate models within sets of top-ranked predictions for many of the test cases[87]. A more recent benchmarking study with 67 antibody-antigen docking test cases highlighted the limited success for current global docking approaches, which was more pronounced for cases with more conformational changes between unbound and bound structures[71].

The recently developed AlphaFold algorithm (AlphaFold v.2.0) performs end-to-end modeling with a deep neural network to generate structural models from sequence[111], showing unprecedentedly high modeling accuracy and substantially surpassing the performance of other teams in the most recent Critical Assessment of Structural Prediction (CASP) round (CASP14)[112]. An important element of the AlphaFold algorithm is the combinatorial use of row-wise, column-wise and triangle self-attention to iteratively infer residue contact and evolutionary information from multiple sequence alignments (MSAs), building on previous work demonstrating the use of co-evolution in contact prediction[224, 225]. The resulting feature representations are further processed by a geometry-aware attention-based structure module that rotates and translates each residue to produce a 3D protein structure prediction. After the remarkable success of AlphaFold in CASP14, a separate team of researchers developed RoseTTAFold [91], which likewise takes MSAs as input, and outputs 3D structural predictions,

using attention-based deep learning architecture. Unlike AlphaFold, RoseTTAFold utilizes a "three-track" approach, allowing for concurrent updates within and in-between 1D amino acid sequence, 2D pairwise distances and orientations between residues, and 3D structural coordinates.

The reported capability to model homomultimers[111], as well as a recently reported adaptation of AlphaFold to enable modeling of heteroprotein assemblies[226], raises the question of how accurately AlphaFold can model transient heteroprotein complexes, including classes of complexes that have challenged previously developed and currently available docking approaches. As the AlphaFold deep learning model was trained using experimentally determined structures of individual protein chains[111], and its accuracy was partly enabled by residue contacts within tertiary structures inferred from multiple sequence alignment, it is not clear whether it can reliably generate protein-protein interface structures, particularly for transient protein complexes which have distinct physicochemical properties than protein interiors[227] and obligate protein-protein interfaces[228, 229], as well as a lack of explicit MSA signal from pairs of residues across the protein-protein interface in the sequences.

Here we report a systematic assessment of the accuracy of AlphaFold in performing end-to-end modeling of transient protein complexes, using 152 heterodimeric test cases from Protein-Protein Docking Benchmark version 5.5 (BM5.5)[71, 87] which represent three previously established docking difficulty levels, and classes of interactions including enzyme-containing complexes, antibody-antigen complexes, as well as a range of other complex types. Comparison of AlphaFold performance with the performance of a global protein docking algorithm, ZDOCK[230] showed remarkable and superior accuracy across the benchmark, even with only five models generated per test case. Determinants of modeling success were assessed by case category and other features, and a number of scoring functions, in addition to pTM (predicted TM-

score[231], corresponding to overall topological accuracy) and pLDDT (predicted Local Difference Distance Test[232], corresponding to local structural accuracy) scores generated by AlphaFold, were tested to find optimal scoring criteria to identify correct docking models from AlphaFold. We also tested a recently released version of AlphaFold, named AlphaFold-Multimer, that was specifically trained to model protein-protein complexes[118]. These results illustrate that while not successful for all cases and complex types, AlphaFold is a powerful tool for complex modeling, showing the power and advantage of end-to-end deep learning versus previous docking approaches. Our results also highlight areas for future optimization and developments in this framework, or other end-to-end deep learning frameworks, to effectively and reliably model most or all transient protein-protein complexes.

4.3 Materials and Methods

4.3.1 Protein-protein complex benchmark and additional antibody-antigen test cases

A test set of heterodimeric protein complexes was obtained from Protein-Protein Docking Benchmark 5.5 (BM5.5) [71, 87]. BM5.5 is a set of structures of non-redundant transient protein-protein complexes from the PDB [164], assembled for testing of predictive protein-protein complex modeling algorithms. By filtering for heterodimeric protein-protein complexes in BM5.5, we obtained a total of 152 cases, consisting of 12 antibody-antigen complexes, 72 enzyme-containing complexes and 68 other types of protein complexes. Docking difficulty classifications for test cases were obtained from the BM5.5 site (https://zlab.umassmed.edu/benchmark/), and are based on the extent of binding conformational changes for each complex[87]. Annotations of

protein source organisms were obtained from the PDB, and confirmed by manual inspection. Buried surface area for each complex interface was obtained from the BM5.5 site.

Twenty additional antibody-antigen modeling test cases (**Table 4.5**) were selected from antibody-antigen complex structures in the SAbDab database[233], screened by resolution (< 3.25 Å) and nonredundancy with any BM5.5 test cases by antibody chain sequences (< 90% antibody variable domain sequence identity) or antigen chain sequence (no hit to antigen chains using default parameters) using the "blastp" executable from the BLAST+ suite[234]. VLR-antigen complex structure test cases (**Table 4.7**) were identified from the PDB and inspected manually for nonredundancy. Recently released antibody-antigen docking benchmark cases obtained from a preliminary update of BM5.5 (**Table 4.9**) and antibody-antigen complex structures identified from the SAbDab database (**Table 4.10**), filtered to remove complexes redundant with any complex structures that were released in the PDB before May 2018, were assembled for AlphaFold-Multimer testing. As an additional nonredundancy check for the latter set of cases, we removed any antibody-antigen complexes with antigen BLAST hits (E-value cutoff 5, and ≥ 40% identity) to antibody-antigen complex structures from pre-May 2018, along with similar docking orientation (< 5 Å RMSD for heavy chain variable domain orientation after superposition of antigens using FAST structural alignment[235]) and > 70% sequence identity for heavy chain variable domain, light chain variable domain, or combined CDR sequences. For modeling efficiency, recently released and additional (non-BM5.5) antibody structures were modeled with variable domains only.

T cell receptor-peptide-MHC (TCR-pMHC) complex structures for AlphaFold-Multimer benchmarking were assembled from Class I TCR-pMHC complex structures in the TCR3d database[47], which were originally obtained from the PDB. TCR-pMHC complex structures were

selected from structures released in the PDB after April 2018, with no redundancy (< 90% TCR variable domain sequence identity, in addition to < 95% sequence identity to any individual TCR α or β variable domain) with any Class I TCR-pMHC complex structures released before May, and no redundancy with any of the complex structures in the benchmark set. Complexes with non-canonical or modified amino acids in peptides were excluded, and a resolution cutoff of 3.25 Å was applied (in accordance with the other benchmarks in this study), except for the 7RM4 TCR-pMHC complex structure which was retained due to its resolution being close to the cutoff (3.33 Å). TCR α, TCR β, peptide, and MHC chains were input as separate sequences to AlphaFold-Multimer. For efficiency, only TCR variable domain sequences, and MHC peptide-binding domains (α1 and α2), were used for modeling.

4.3.2 Complex modeling with AlphaFold

AlphaFold was downloaded from Github (https://github.com/deepmind/alphafold) and installed on a local computing cluster. Sequences of protein chains for the protein-protein complexes were obtained from the PDB "seqres" file and used as input for each complex modeling job. Raw MSAs were prepared for each chain with the downloaded published AlphaFold pipeline[111], querying the full databases (UniRef90 version 2020_01, MGnify version 2018_12, Uniclust30 version 2018_08 and BFD). The resulting raw MSAs of the interacting chains were subsequently combined to form the unpaired MSA inputs for complex structure prediction. To generate MSA lines of the same length, gaps equal to the length of the interacting chain were added before or after each sequence. To avoid implicitly biasing the complex structure predictions with knowledge of individual bound protein chain conformations, the use of structural templates was disabled in this study.

To introduce chain breaks, a residue index shift of 200 was added to the junction of interacting chains, as recently implemented in ColabFold [226]. Following the published AlphaFold pipeline, AlphaFold generated five models for each complex, which were ranked in this study by pTM score (predicted TM-score, which is a measure of predicted structure accuracy generated by AlphaFold). After structural predictions were generated, model relaxation by the Amber program [236], which as reported by Jumper et al. [111] was used to ameliorate minor structural defects without impacting accuracy, was replaced by the constrained FastRelax protocol in the Rosetta program [165], as detailed below. To test the impact of varying the ensembling ($N_{ensemble}$) and recycling (N_{cycle}) parameters on complex modeling accuracy, we increased $N_{ensemble}$ or N_{cycle} by modifying those parameters in the AlphaFold Python code, while keeping the input MSAs and sequences/features the same.

Three out of the 152 test cases failed to complete in the AlphaFold pipeline due to GPU memory limits during structure prediction, or errors during feature preparation: 1ZM4, 2OZA, and 1B6C. AlphaFold structure prediction runs were performed on an NVIDIA Titan RTX or NVIDIA Quadro 6000 GPU.

4.3.3 Complex modeling with ColabFold and RoseTTAFold

Protein-protein complex predictions were generated in ColabFold [226] using its "advanced" interface (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/beta/AlphaFold_advanced.ipynb). Input protein sequences were identical to those used for AlphaFold modeling. The MMseqs2 method was selected on ColabFold to generate the MSAs, and Amber relaxation of

models was disabled. The unpaired MSA predictions are generated between August 20th and August 24th, 2021.

ColabFold enables users to pair alignments for different protein sequences based on UniProt accession numbers; this is a selectable option on the ColabFold interface [226]. Since the protocol is designed to pair prokaryotic protein sequences, MSA pairing was only performed on a subset of cases where both protein chains of the heterodimer complex come from the same prokaryotic organism. Prefiltering of MSAs was enabled prior to pairing, with the minimum coverage with query of 50% and minimum sequence identity with query of 20%. Structural predictions based on paired MSAs were generated on September 4, 2021. The resulting paired MSAs were also used as input to RoseTTAFold [91] on the Robetta web server (https://robetta.bakerlab.org/submit.php) to generate complex models.

All models generated with AlphaFold, ColabFold and RoseTTAFold are made available to the public at: https://piercelab.ibbr.umd.edu/af_complex_models.html.

4.3.4 Complex modeling with AlphaFold-Multimer

AlphaFold-Multimer[118] modeling was performed with AlphaFold downloaded from https://github.com/deepmind/alphafold on November 2, 2021, and local ColabFold downloaded from https://github.com/sokrypton/ColabFold on Jan 12, 2021. Input MSA features were generated by either the AlphaFold-Multimer pipeline described in Evans et al. [237], or by local ColabFold [226] using the "MMseqs2 (Uniref+Environmental)" MSA mode. By default, the MSAs constructed contain both unpaired (per-chain) and paired sequences. To generate AlphaFold-Multimer predictions using alternative MSA pairing modes ("unpaired only", "paired only", or "no MSA"), local ColabFold was used. Specifically, MSA pair mode was set to "Paired"

to generate “paired only” predictions. The MSA pair mode was set to “unpaired+paired” to generate “unpaired only” predictions, and “paired” to generate “no MSA” predictions, after modifying the “get_msa_and_templates” function in “batch.py” of local ColabFold, so the list variable “paired_a3m_lines” contains only the query sequences, instead of paired sequences generated by MMSeqs2. While “no MSA” (a.k.a. “single sequence”) and “unpaired” options are available in the ColabFold Google Colab interface, we found the above modification necessary in the version of the ColabFold code that we downloaded at the time. To avoid implicitly biasing the structural predictions with knowledge of known conformations, a template release date cutoff of “2018-04-30” was applied when the use of templates was enabled.

4.3.5 Docking model generation with ZDOCK

To enable comparison against a rigid-body docking algorithm, we generated docking models using ZDOCK version 3.0.2 [230]. Unbound protein structures from BM5.5, with HETATMs removed, were used as inputs to ZDOCK. Dense rotational sampling was used, generating 54000 predictions per complex. The integration of residue- and atom-based potentials for docking (IRAD) [238] scoring function was used to rank the ZDOCK output models.

4.3.6 Docking model accuracy assessment

Docking models were assessed using the Critical Assessment of Predicted Interactions (CAPRI) criteria [223] using custom scripts. Based on the structural similarity between docking models and native structures, docking models were classified into four accuracy classes: “High”, “Medium”, “Acceptable” and “Incorrect”. Such structural similarity is assessed by a combination

of interface RMSD (I-RMSD), ligand RMSD (L-RMSD), and fraction of native interface residue contacts (fnat). Backbone atoms were used in the I-RMSD and L-RMSD calculations.

4.3.7 Interface pLDDT and interface PAE calculation

To calculate the interface pLDDT, we averaged the per-residue pLDDT of interface residues. Interface residues are defined as residues with atomic contacts across the interface within the specified distance cutoff. Interface PAE was calculated by averaging the PAE of cross-interface residue pairs with atomic contacts within a given distance cutoff. The interface distance cutoffs tested range from 4 Å to 10 Å. An interface pLDDT score of 0 and an interface PAE score of 35 was assigned to models without any interface contacts within the distance cutoff.

4.3.8 Structure relaxation using Rosetta

To resolve possible unfavorable geometries or clashes in experimentally determined complex structures and AlphaFold models, the Rosetta FastRelax protocol [166] was applied to the predicted structures prior to scoring of the models using interface analysis protocols (IRAD, ZRANK2, and Rosetta). Parameter flags used in FastRelax ("relax" executable in Rosetta 3.12 [165]) are:

-relax:constrain_relax_to_start_coords

-relax:coord_constrain_sidechains

-relax:ramp_constraints false

-ex1

-ex2

-use_input_sc

-no_optH false

-flip_HNQ

-nstruct 1

4.3.9 Complex and docking model scoring with IRAD, ZRANK2, and Rosetta InterfaceAnalyzer

Post-relax complex structures were used as inputs to obtain IRAD [238], ZRANK2 [75], and Rosetta [165] InterfaceAnalyzer protocol scores. IRAD and ZRANK2 scores were obtained from the downloaded "irad" executable program. InterfaceAnalyzer scores were obtained using the. "InterfaceAnalyzer" executable in Rosetta v. 3.12, with default parameters; the InterfaceAnalyzer protocol computes and outputs interface energetic scores using the Rosetta REF15 function [239], along with REF15 component terms and other interface structure metrics.

4.3.10 Number of effective sequences

The number of effective sequences (N_{eff}) is used as a measure of the MSA depth. N_{eff} score is defined as the number of clusters after the raw MSA inputs were clustered at the 62 % sequence identity using CD-HIT [240] with the word length of 4, as used previously[241].

4.3.11 TM-score calculations

TM-scores were calculated using TM-score executable [231] by comparing the structural similarity between experimentally determined structures and AlphaFold models. Residues that

were unresolved in experimentally determined structures were removed from AlphaFold models before the calculation of TM-scores.

4.3.12 Figures, statistical analysis, and AUC calculations

Figures of structures were generated using PyMOL version 2.4 (Schrodinger, Inc.). Box plots, line plots and bar plots were generated with the ggplot2 package [186] in R (r-project.org), and heatmaps were generated with the ComplexHeatmap package [188] in R. Pearson correlations and their p-values were calculated with ggpubr package in R. Wilcoxon rank-sum test was performed using ggsignif package in R. Binary and multi-class ROC curves with AUC values were calculated with the "pROC" and "multiROC" packages, respectively, in R. 95% confidence interval values for binary ROC AUC values were calculated using the "ci.auc" function in pROC, with 2000 stratified boostrap replicates, and multi-class ROC AUC confidence interval values were calculated using the "boot" R package [242] with the adjusted bootstrap percentile (BCa) method. Calculations of possible score thresholds for model selection were performed using the "cutpointr" package [243], with maximization of the sum of the sensitivity and specificity, based on discrimination of Incorrect versus Medium/High CAPRI accuracy models.

4.4 Results

4.4.1 Performance of AlphaFold on protein-protein complex prediction

To assess the accuracy of AlphaFold in predicting structures of transient protein-protein complexes, we used Protein-Protein Docking Benchmark 5.5 (BM5.5) [71, 87], which contains complexes spanning many classes of interactions that were identified from the Protein Data Bank [164] using an automated pipeline followed by manual inspection and curation. All heterodimeric

protein-protein complexes from that benchmark were identified for this analysis, corresponding to 152 test cases (**Table 4.1**). Based on levels of binding conformational changes and previously defined criteria [244], the cases had unbound docking difficulty classifications of Rigid (95 cases), Medium difficulty (34 cases) and Difficult (23 cases). Sequences of the two chains from each test case were input to AlphaFold, which generated structural models of the protein complexes using unpaired MSAs, without the use of templates. Additionally, the "advanced" interface in ColabFold [226] was utilized to generate protein complex models using the AlphaFold framework. ColabFold uses different databases and a different MSA generation algorithm, but its speed and web accessibility make it a useful alternative to a locally installed full AlphaFold pipeline. To permit comparison with a current docking approach, the rigid-body docking program ZDOCK (version 3.0.2) [230] and the IRAD scoring function [238] were used to perform global docking and rank models for all complexes, using unbound protein structures as input.

Table 4.1. Heterodimeric protein-protein complexes from BM5.5 tested in this study.

PDB code	Receptor	Ligand	Complex Category[a]	BSA (Å^2)[b]	Protein source[c]
1ACB	Chymotrypsin	Eglin C	EI	1544	ME
1AK4	Cyclophilin	HIV capsid	OX	1029	MM
1ATN	Actin	Dnase I	OX	1774	ME
1AVX	Porcine trypsin	Soybean trypsin inhibitor	EI	1585	ME
1AY7	RNase Sa	Barstar	EI	1237	MB
1B6C	FKBP binding protein	TGFbeta receptor	OX	1752	SE
1BKD	Ras GTPase	Son of sevenless	OG	3163	SE
1BUH	CDK2 kinase	Ckshs1	EI	1324	SE

1BVN	Alpha-amylase	Tendamistat	EI	2222	MM
1CGI	Bovine chymotrypsinogen	PSTI	EI	2053	ME
1CLV	Alpha-amylase	Alpha-amylase inhibitor	EI	2087	ME
1D6R	Bovine trypsin	Bowman-Birk inhibitor	EI	1408	ME
1DFJ	Ribonuclease A	Rnase inhibitor	EI	2582	ME
1E6E	Adrenoxin reductase	Adrenoxin	ES	2315	SE
1E96	Rac GTPase	p67 Phox	OG	1179	SE
1EAW	Matriptase	BPTI	EI	1866	ME
1EFN	HIV-1-NEF protein	SH3 domain	OX	1254	MM
1EWY	Ferredoxin reductase	Ferredoxin	ES	1502	SB
1F34	Porcine pepsin	Ascaris inhibitor 3	EI	3038	ME
1F6M	Thioredoxin reductse	Thioredoxin 1	ES	1821	SB
1FFW	Chemotaxis protein CheY	Chemotaxis protein CheA	OX	1166	SB
1FLE	Elastase	Elafin	EI	1763	ME
1FQ1	CDK2 kinase	CDK inhibitor 3	ES	1832	SE
1FQJ	Gt-alpha	RGS9	OG	1806	ME
1GCQ	GRB2 C-ter SH3 domain	Vav N-ter SH3 domain	OX	1208	ME
1GHQ	Complement C3	Epstein-Barr virus receptor CR2	OR	800	SE
1GL1	Alpha-chymotrypsin	Protease inhibitor LCMI II	EI	1591	ME
1GLA	Glycerol Kinase	Glucose specific phosphocarrier	ER	1304	SB
1GPW	HISF protein	Amidotransferase HISH	OX	2097	SB
1GRN	CDC42 GTPase	CDC42 GAP	OG	2332	SE
1GXD	proMMP2 type IV collagenase	Metalloproteinase inhibitor 2	EI	2445	SE
1H1V	Actin	Gelsolin	OX	2071	ME

1H9D	Runx1 domain of CBFAlpha1	Dimerisation domain of CBF-Beta	OX	2121	SE
1HE1	Rac GTPase	Pseudomonas toxin GAP dom.	OG	2113	MM
1HE8	Ras GTPase	PIP3 kinase	OG	1305	SE
1I2M	Ran GTPase	RCC1	OG	2779	SE
1IBR	Ran GTPase	Importin beta	OG	2270	SE
1IRA	Interleukin-1 receptor	Interleukin-1 receptor antagonist protein	OR	3367	SE
1J2J	Arf1 GTPase	GAT domain of GGA1	OG	1209	ME
1JIW	Alkaline metalloproteinase	Proteinase inhibitor	EI	1997	SB
1JK9	CCS metallochaperone	SOD1 superoxide dismutase	ES	2130	SE
1JTD	BLIP-II	TEM-1 beta-lactamase	EI	2180	MB
1JTG	Beta-lactamase inhibitor protein	Beta-lactamase TEM-1	EI	2600	MB
1KAC	Adenovirus fiber knob protein	Adenovirus receptor	OR	1456	MM
1KTZ	TGF-Beta	TGF-Beta receptor	OR	989	SE
1KXP	Actin	Vitamin D binding protein	OX	3341	ME
1LFD	Ras	RalGDS Ras-interacting domain	OG	1167	ME
1M10	Von willebrand factor dom. A1	Glycoprotein IB-Alpha	ER	2097	SE
1MAH	Acetylcholinesterase	Fasciculin	EI	2145	ME
1MQ8	ICAM-1 domain 1-2	Integrin Alpha-L I domain	OX	1253	SE
1NW9	Capase-9	BIR3-XIAP	ER	2112	SE
1OC0	Plasminogen activator inhibitor-1	Vitronectin Somatomedin B domain	ER	1313	SE

1OPH	Alpha-1-antitrypsin	Trypsinogen	EI	1360	ME
1OYV	Subtilisin Carlsberg	Two-headed tomato inhibitor-II	EI	1930	MM
1PPE	Bovine trypsin	CMTI-1 squash inhibitor	EI	1688	ME
1PVH	IL6 receptor Beta chain D2-D3 domains	Leukemia inhibitory factor	OR	1403	SE
1PXV	Cystein protease	Cystein protease inhibitor	EI	2336	SB
1QA9	CD2	CD58	OX	1353	SE
1R0R	Subtilisin carlsberg	OMTKY	EI	1409	MM
1R6Q	Clp protease subunit ClpA	Clp protease adaptor protein ClpS	ER	1651	SB
1R8S	Arf1 GTPase	Sec 7 domain	OG	2986	ME
1RKE	Vinculin head	Vinculin tail	OX	2614	SE
1S1Q	UEV domain	Ubiquitin	OX	1288	SE
1SBB	T-cell receptor Beta	Staphylococcus enterotoxin B	OR	1064	MM
1SYX	Spliceosomal U5 15 kDa protein	CD2 receptor binding protein 2 C-ter fragment	OX	1293	SE
1T6B	Anthrax protective antigen	Anthrax toxin receptor	OR	1948	MM
1TMQ	alpha-amylase	RAGI inhibitor	EI	2401	ME
1UDI	Uracyl-DNA glycosylase	Glycosylase inhibitor	EI	2022	Viral
1US7	Heat shock protein 82 N-ter domain	HSP 90 co-chaperone CDC37 C-ter domain	ER	1095	ME
1WQ1	Ras GTPase	Ras GAP	OG	2913	SE
1XD3	UCH-L3	Ubiquitin	OX	2281	SE
1XQS	HspBP1	Hsp70 ATPase domain	OX	2350	SE
1Y64	Actin	BNI1 protein	OX	2745	ME
1YVB	Falcipain 2	Cystatin	EI	1743	ME

1Z0K	Rab4A GTPase	RAB4 binding domain of Rabenosyn	OG	1787	SE
1Z5Y	N-term of DsbD	E.coli CCMG protein	ES	1346	SB
1ZHH	Autoinducer 2-binding periplasmic protein LuxP	Autoinducer 2 sensor kinase/phosphatase LuxQ	OR	2189	SB
1ZHI	BAH domain of Orc1	Sir Orc-interaction domain	OX	1322	SE
1ZLI	Carboxypeptidase B	Tick carboxypeptidase inhibitor	EI	2084	ME
1ZM4	Elongation factor 2	Diphtheria toxin A catalytic domain	ES	1554	MM
2A1A	Eukayotic translation initiation factor 2-alpha kinase 2	eIF2 alpha subunit	ES	1186	ME
2A5T	NMDA receptor R1-4A subunit ligand-binding core	NMDA receptor R2A subunit ligand-binding core	OX	1892	ME
2A9K	Ras-related protein Ral-A	Mono-ADP-ribosyltransferase C3	ES	1751	MM
2ABZ	Carboxypeptidase A1	Leech carboxypeptidase inhibitor	EI	1443	ME
2AJF	ACE2	SARS spike protein receptor binding domain	OR	1704	MM
2AYO	Ubiquitin carboxyl-terminal hydrolase 14	Ubiquitin	ER	3027	SE
2B42	Xylanase	Xylanase inhibitor	EI	2520	MM
2BTF	Actin	Profilin	OX	2063	SE
2C0L	PTS1 and TRP region of PEX5	SCP2	OX	2013	SE
2CFH	BET3	TPC6	OX	2384	SE
2FJU	Phospholipase Beta 2	Rac GTPase	OG	1245	SE

2G77	GTPase-activating protein GYP1	Ras-related protein Rab-33B	OG	2524	ME
2GAF	Poly(A) polymerase VP55	Vaccinia protein VP39	ER	3368	Viral
2GTP	Alpha-1 subunit Guanine nucleotide-binding protein G(I), alpha-1 subunit	RGS1	OG	1442	SE
2H7V	Rac GTPase	YpkA	OG	1574	MM
2HLE	Ephrin B4 receptor	Ephrin B2 ectodomain	OR	2116	SE
2HQS	TolB	Pal	OX	2333	SB
2HRK	Glutamyl-t-RNA synthetase	GU-4 nucleic binding protein	OX	1595	SE
2I25	Shark single domain antigen receptor	Lysozyme	AA	1425	ME
2I9B	Urokinase plasminogen activator surface receptor	Urokinase-type plasminogen activator	OR	2371	SE
2IDO	DNA polymerase III Œµ exonuclease domain	HOT protein (P1 phage)	ES	1953	MM
2J0T	MMP1 Intersitial collagenase	Metalloproteinase inhibitor 1	EI	1477	SE
2J7P	SRP GTPase Ffh	Cell division protein FtsY	OX	3008	SB
2NZ8	Rac GTPase	DH/PH domain of TRIO	ER	2599	SE
2O3B	NucA nuclease	NuiA nuclease inhibitor	EI	1675	SB
2O8V	PAPS reductase	Thioredoxin	ES	1619	SB
2OOB	Ubiquitin ligase	Ubiquitin	ES	808	ME
2OT3	Rab21 GTPase	Rabex-5 VPS9 domain	ER	2306	SE

2OUL	Falcipain 2	Chagasin	EI	1933	ME
2OZA	MAP kinase 14	MAP kinase-activated protein kinase 2	OX	6248	ME
2PCC	Cyt C peroxidase	Cytochrome C	ES	1141	SE
2SIC	Subtilisin	Streptomyces subtilisin inhibitor	EI	1617	MB
2SNI	Subtilisin	Chymotrypsin inhibitor 2	EI	1628	MM
2UUY	Trypsin	Tryptase inhibitor from tick	EI	1280	ME
2VDB	Serum albumin	Peptostreptococcal albumin-binding protein	OX	1798	MM
2X9A	TolA C-terminal domain	G3P TolA binding domain	OR	1571	MM
2YVJ	Ferredoxin reductase BPHA4	Biphenyl dioxygenase ferredoxin subunit	ER	1377	SB
2Z0E	Cysteine protease Atg4B	Microtubule-associated proteins 1A/1B light chain 3B	ER	2478	ME
3A4S	SUMO-conjugating enzyme UBC9	NFATC2-interacting protein SLD2 ubiquitin-like domain	EI	1116	ME
3AAD	Double bromodomain	Histone chaperone ASF1	OX	1654	SE
3BIW	Neuroligin-1	Neuroligin-1-beta	OX	1191	SE
3BX7	Lipocalin 2	CTLA-4 extracellular domain	OX	2349	SE
3CPH	Rab GDP-dissociation inhibitor	Ras-related protein Sec4	OG	1685	SE
3D5S	Complement C3d fragment	Fibrinogen-binding protein C-ter domain	OX	1620	MM
3DAW	Alpha actin	Twinfilin-1 C-terminal domain	OX	2323	ME
3F1P	HIF2 alpha C-terminal PAS domain	ARNT C-terminal PAS domain	OX	1919	SE

3FN1	UQ_con domain from NEDD8-conjugating enzyme UBE2F	NEDD8-activaating enzyme E1 catalytic subunit	ER	1897	SE
3H2V	Vinculin tail domain	Raver1 RRM1 domain	OX	1263	SE
3K75	DNA polymerase beta	Reduced XRCC1, N-terminal domain	ER	1195	ME
3PC8	DNA repair protein XRCC1	DNA ligase III-alpha BRCT domain	ER	1240	ME
3RJQ	A12	C1086 HIV gp120	AA	1734	MM
3S9D	IFNAR2	IFNa2	OR	1841	SE
3SGQ	Streptogrisin B	Ovomucoid inhibitor third domain	EI	1211	MM
3VLB	EDGP	Xyloglucan-specific endo-beta-1,4-glucanase A	EI	2020	ME
4CPA	Carboxypeptidase A	Potato carboxypeptidase inhibitor	EI	1175	ME
4FZA	MO25 alpha	Serine/threonine-protein kinase MST4	ER	1695	SE
4H03	Iota toxin component IA	Alpha actin	ES	1474	MM
4IZ7	Non-phosphorylated ERK	PEA-15 Death Effector Domain	EI	1202	ME
4M3K	cAb-H7S	B. licheniformis beta-lactamase	AA	1588	MM
4M76	C3D	Integrin alpha-M CD11B A-domain	OR	1046	SE
4POU	VHHmetal	bovine RNase A	AA	1313	ME
4Y7M	nb25	E coli TssM CTD	AA	1103	MM
5E5M	H11	mouse CTLA-4	AA	1341	ME
5HGG	Nb4	uPA	AA	1969	ME
5JMO	Nb14	Furin	AA	1394	ME
5SV3	A3C8	Ricin	AA	1294	ME
5VNW	Nb.b201	human serum albumin	AA	967	MM

6CWG	A9	Ricin	AA	1151	ME
6DBG	R303	Listeria monocytogenes internalin B	AA	1525	MM
7CEI	Colicin E7 nuclease	Im7 immunity protein	EI	1384	SB
BAAD	Double bromodomain	Histone chaperone ASF1	OX	1461	SE
BOYV	Subtilisin Carlsberg	Two-headed tomato inhibitor-II	EI	1280	MM

[a] Complex category: Antibody-antigen (AA), enzyme-inhibitor (EI), enzyme-substrate (ES), enzyme complex with a regulatory or accessory chain (ER), others, G-protein containing (OG), others, receptor-containing (OR); others, miscellaneous (OX).

[b] Interface buried surface areas.

[c] Based on the source organisms for receptor and ligand protein chains in the Protein Data Bank (PDB) [164], each case is classified as SE ("Single, Eukaryotic", denoting proteins from the same eukaryotic organism), SB ("Single, Bacterial", denoting proteins from the same bacterial organism), ME ("Multiple, Eukaryotic", denoting proteins from different eukaryotic organisms), MB ("Multiple, Bacterial", denoting proteins from different bacterial organisms), Viral (denoting proteins from viruses), or MM ("Multiple, Mix", denoting proteins from mixed origins

The performance of AlphaFold, ColabFold, and ZDOCK was assessed by comparison of models with experimentally determined structures of the bound complexes; overall success rate comparisons are shown in **Figure 4.1a**, for top 1 (T1) and top 5 (T5) ranked models for each test case, with per-case performance shown in **Figure 4.2**. Models were assessed as Acceptable, Medium, or High accuracy, or Incorrect, based on CAPRI criteria [223], based on comparison of models with corresponding experimentally determined structures using ligand root mean square distance (L-RMSD), interface residue root mean square distance (I-RMSD), and fraction of native interface residue contacts (f_{nat}) metrics. While Acceptable accuracy models can include moderate deviation from known structures (including models with up to 10 Å L-RMSD), Medium and High

accuracy models are more reflective of previously utilized model accuracy cutoffs, such as the 2.5 Å I-RMSD cutoff used for near-native models by Chen and Weng [245]; accordingly, multiple studies have used the Medium accuracy cutoff to identify near-native models [148, 246]. Remarkably, AlphaFold was able to generate models with Acceptable or higher accuracy for approximately half (51%) of the 149 test cases for which models were generated, and for many of those cases, Medium or better accuracy (43%) or High accuracy (21%) models were generated. Additionally, the top-ranked model (T1) based on AlphaFold pTM score often represented the highest accuracy level for each case, and only a modest improvement in success was observed when allowing five predicted models per case (T5) (54%, 44%, and 23% success rates for Acceptable accuracy or better, Medium accuracy or better, or High accuracy, respectively). The success rate for ColabFold was similar to the success of AlphaFold, indicating that the different sequence databases and MSA procedure did not reduce or otherwise alter the capability of the AlphaFold deep learning model to generate near-native complex models. Inspection of per-case performance (**Figure 4.2**) confirmed that ColabFold and AlphaFold success was highly correlated across the test cases. Rigid-body global docking success from ZDOCK was considerably lower than AlphaFold and ColabFold, particularly for Medium and High accuracy models (13% Acceptable or higher accuracy, 9% Medium or higher accuracy, 1% High accuracy success for top-ranked models), although a subset of cases was successful for ZDOCK while not successfully predicted by AlphaFold or ColabFold (**Figure 4.2**). A representative successfully modeled complex from AlphaFold is shown in comparison with the experimentally determined structure of the complex (PDB code 2X9A; E coli TolA/Phage G3B complex) in **Figure 4.1b**, demonstrating modeling of a virus-host protein-protein interaction with atomic-level accuracy. As that complex structure was released in 2010, it is possible that one or both of the component proteins were part

of the AlphaFold training set, however the protein-protein interface and binding orientation were not.

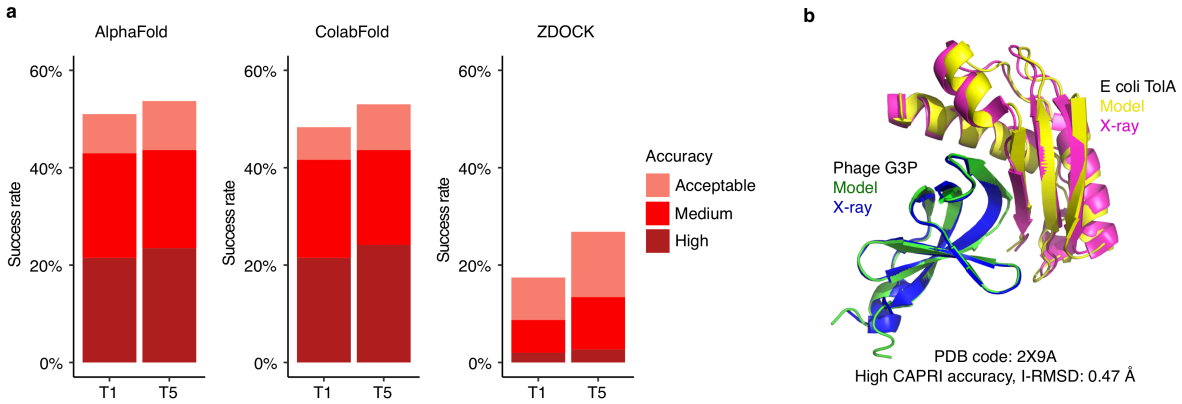

Figure 4.1. Transient protein-protein complex structure prediction success by AlphaFold, ColabFold and ZDOCK. End-to-end modeling using AlphaFold[111] and ColabFold [226] was performed on 152 complex test cases (details in Figure 5.2). AlphaFold failed to generate predictions for three complexes, thus AlphaFold predictions were obtained for 149 complexes; these 149 test cases were used to calculate success rates in this figure. Docking models were also generated with ZDOCK [230], using unbound protein structures as input. All sets of models were assessed for near-native predictions using CAPRI criteria for High, Medium and Acceptable accuracy. **(a)** Complex prediction success of AlphaFold, ColabFold, and ZDOCK for the top 1 (T1) and top 5 (T5) models considered. AlphaFold and ColabFold models were ranked by AlphaFold pTM scores, and ZDOCK models were ranked by IRAD scores [238]. The percentage success was calculated as the percentage of test cases with a given model accuracy from the top N models considered. Bars are colored according to the CAPRI quality classes. (**b**) Example of an accurately predicted complex structure (PDB code: 2X9A) by AlphaFold. This model has High accuracy by CAPRI criteria (I-RMSD = 0.47) and has the highest pTM score (pTM = 0.77) of all 5 models generated for this complex. Structures are superposed by Phage G3P, with the model and the X-ray structure chains are colored separately as indicated. For clarity, regions modeled by AlphaFold but unresolved in the X-ray structure are not shown in the figure.

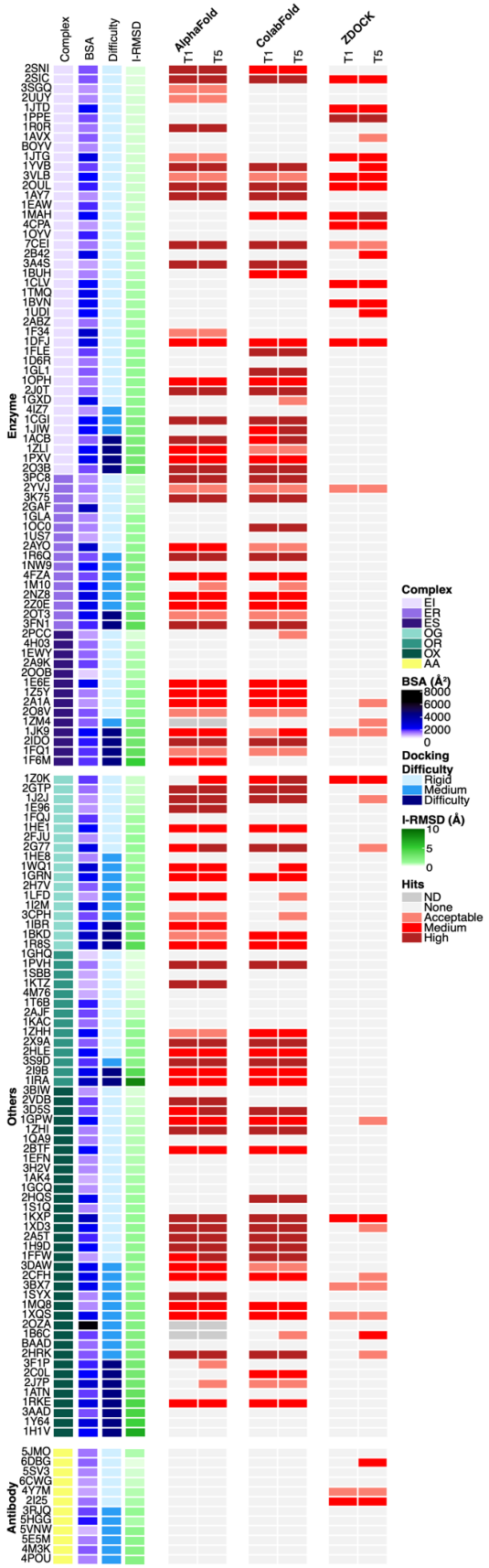

Figure 4.2. Predictive modeling success of AlphaFold, ColabFold, and ZDOCK. Cases are grouped and sorted by complex type. For reference, complex category, buried surface area (BSA, Å^2), docking difficulty based on binding conformational changes, and binding interface root mean square distance (I-RMSD) are indicated on the left. Complex category: Enzyme-inhibitor (EI), enzyme complex with a regulatory or accessory chain (ER), enzyme-substrate (ES), others, G-protein containing (OG), others, receptor-containing (OR); others, miscellaneous (OX), antibody-antigen (AA). Success for top 1 and top 5 (T1, T5) ranked predictions is shown, colored by CAPRI model accuracy as indicated in the key on right (Hits). Three cases failed to complete in AlphaFold and are shown as dark gray cells ("ND" under Hits).

4.4.2 Determinants of successful and unsuccessful AlphaFold performance

To investigate the determinants of successful performance for AlphaFold, we compared performance across subsets of cases divided by various biological and structural properties (**Figure 4.3**). As expected, previously assigned test case difficulty classifications, which are based on binding conformational change between unbound and bound structures [71, 87], did not markedly impact the success of AlphaFold; for Acceptable or higher accuracy predictions in the set of five models, success rates for AlphaFold were found to be 47%, 55%, and 78% for Rigid, Medium, and Difficult docking difficulty categories, respectively (**Figure 4.3a**). The increase in AlphaFold success for cases in the Difficult docking case category relative to the other two categories was less pronounced or not observed for more stringent model accuracy criteria of Medium and High accuracy. AlphaFold Medium or higher model accuracy success rates for the difficulty categories were 39% (Rigid), 48% (Medium), and 57% (Difficult), while High accuracy model success rates were 27% (Rigid), 16% (Medium), and 17% (Difficult). For the docking algorithm ZDOCK, which unlike AlphaFold used unbound protein structures as input, success rates for the top 5 ranked models were 36% (Rigid), 19% (Medium), and 0% (Difficult) for Acceptable or higher accuracy models. This reduced success of ZDOCK for progressively higher docking difficulty categories is in accordance with previous benchmarking studies with ZDOCK and other methods that use unbound structures as input [71, 87]. While the "fold-and-dock" approach in AlphaFold is likely at least partly responsible for improved modeling context-specific conformations versus the reliance of unbound structures for rigid-body docking, it remains possible, as noted above, that some bound conformations of individual protein components in Benchmark 5.5 are part of the AlphaFold training set, which would provide an additional advantage for AlphaFold versus the use of the unbound structures, or models of unbound structures, as input for complex assembly.

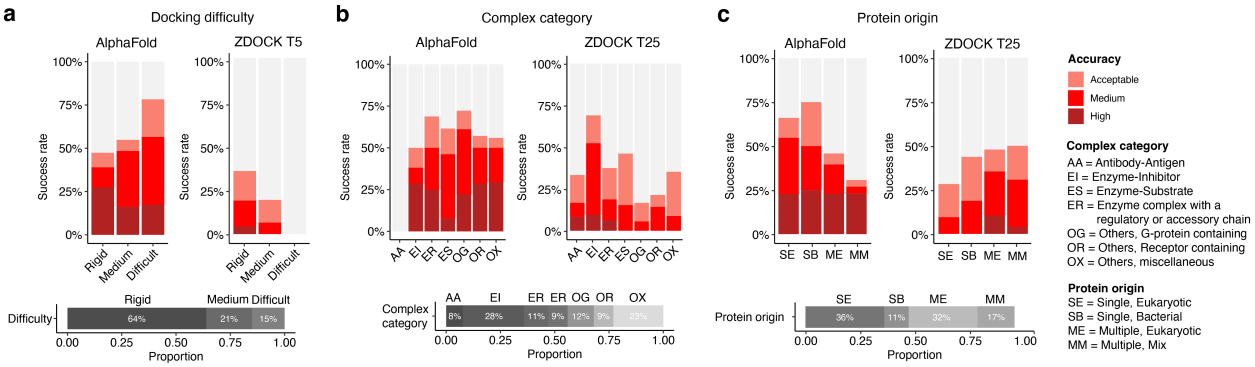

Figure 4.3. Determinants of successful performance. (a) Prediction success of AlphaFold and ZDOCK, grouped by docking difficulty. Based on binding conformational changes as defined by BM5.5 [71, 87], cases are categorized into "Rigid", "Medium", and "Difficult" docking difficulty levels. To evaluate the success rate, all 5 models from AlphaFold and top 5 ZDOCK models were considered. **(b)** Prediction success of AlphaFold and ZDOCK grouped by complex category. To evaluate the success rate, all 5 models from AlphaFold and top 25 ZDOCK models were considered. The number of ZDOCK models was increased to 25 to allow for sufficient success rates to show its relative performance among the categories. **(c)** Prediction success of AlphaFold and ZDOCK grouped by protein source organism(s). To evaluate the success rate, all 5 models from AlphaFold and top 25 ZDOCK models were considered. Based on the source of subunit proteins in the complex structures, each case is classified as either "Single, Eukaryotic" (SE, denoting proteins from the same eukaryotic organism), "Single, Bacterial" (SB, denoting proteins from the same bacterial organism), "Multiple, Eukaryotic" (ME, denoting proteins from different eukaryotic organisms), or "Multiple, Mix" (MM, denoting proteins from mixed origins). Two additional protein source classes, corresponding to proteins from different bacterial organisms, and proteins of viral origin, were omitted from the success plot due to limited representation in each category (4 and 2 cases in those classes, respectively). Bars are colored by model accuracy as indicated in **(a)**. The horizontal stacked bars below each success rate plot denote the composition of the categories by class.

Performance across benchmark cases was also assessed by complex category, as well as protein source (**Figure 4.3b, 4.3c**). Notably, the antibody-antigen complexes had no successfully generated models, while other complex categories considered all showed approximately commensurate levels of AlphaFold performance. There was no major difference observed in AlphaFold success for prediction of complexes with proteins from eukaryotic or bacterial organisms, and while there was a slight reduction in overall success when the two proteins in a complex came from different organisms (which theoretically could impact a cross-interface signal of an MSA), the success for High quality models was approximately the same (~25%) regardless

of single versus multiple source organism, or source organism type.

We performed analysis of a series of geometric and other protein complex properties to identify possible relationships with AlphaFold modeling success. Computed interface features were assessed for association with incorrectly modeled cases versus cases with near-native AlphaFold complex models (Medium and/or High CAPRI accuracy) (**Figure 4.4**, **Table 4.2**). Greater interface size, measured by buried surface area (BSA), was found to be associated with AlphaFold success for Incorrect vs. Medium/High accuracy cases ($p = 0.007$; **Table 4.2**), and Incorrect vs. Medium accuracy cases ($p \leq 0.001$; **Figure 4.4**), yet this trend was not observed when comparing Incorrect vs. High accuracy cases. To account for possible bias from antibody-antigen features and their pronounced lack of AlphaFold success noted above, comparisons were made with antibody-antigen cases excluded (**Figure 4.5**), yielding essentially the same results as with all cases (**Figure 4.4**). Limited multiple sequence alignment depth for either or both partner proteins was explored as a possible factor in poor predictive performance, but it was not found to have a significant impact (**Figure 4.4**). Among the case features analyzed, we found that larger protein sizes, and a relatively small interface in comparison to protein size (measured either by number of residues or solvent accessible surface area), were most associated with poor complex modeling performance (**Figure 4.4, Table 4.2**).

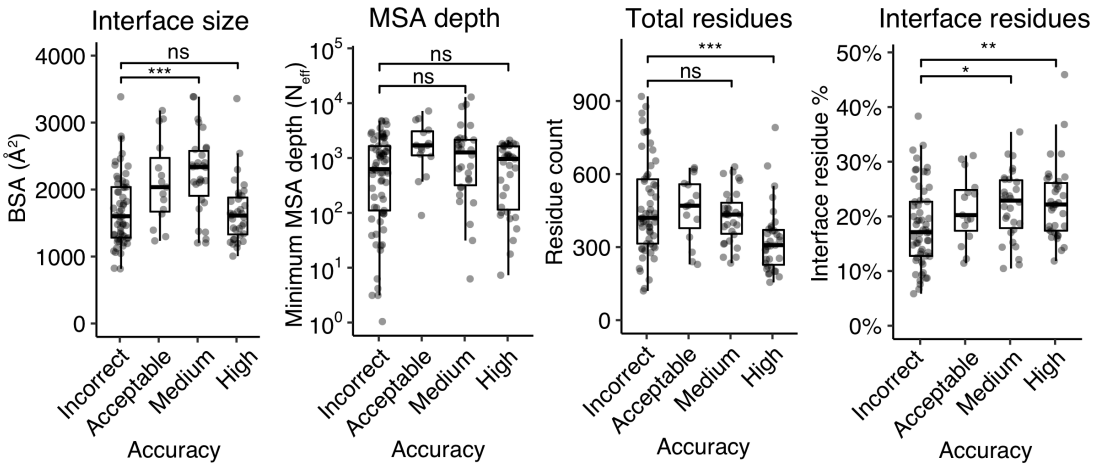

Figure 4.4. Assessing test case features associated with AlphaFold success. Protein complex and MSA feature values were computed for all cases, which are shown according to AlphaFold success (best AlphaFold model accuracy in the five models for that case). Features shown are interface buried surface area (BSA), MSA depth (N_{eff}) for the ligand or receptor (minimum value of the two), total number of residues, and percent of total residues in the protein-protein interface. Statistical significance values (Wilcoxon rank-sum test) were calculated between feature values for sets of cases with Incorrect vs. Medium and Incorrect vs. High CAPRI accuracy, as noted at top (ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$).

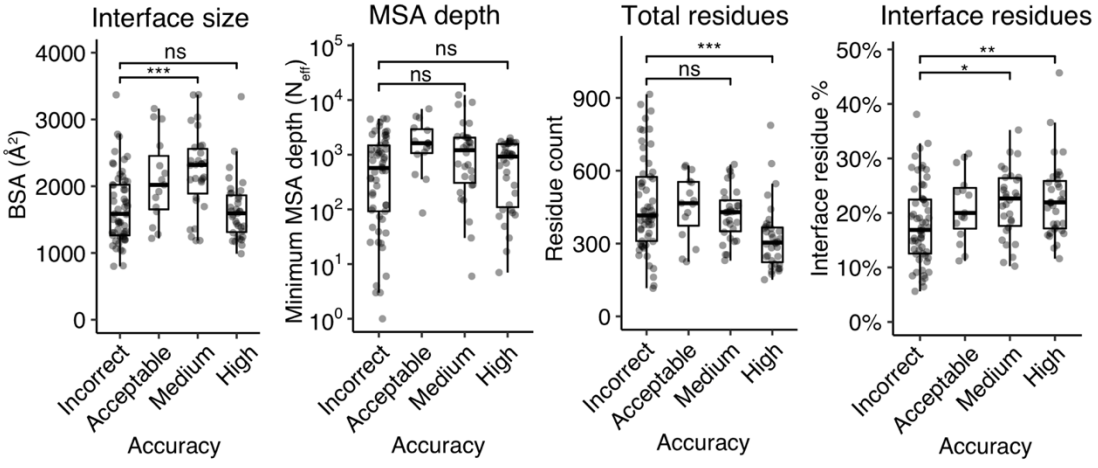

Figure 4.5. Assessment of features associated with AlphaFold success or lack of success, with antibody complexes excluded. Protein complex and MSA feature values were computed for all cases except for antibody-antigen complexes, shown according to AlphaFold success (best AlphaFold model accuracy in the five models for that case). Features shown are interface buried surface area (BSA), MSA depth (N_{eff}) for the ligand or receptor (minimum value of the two), total number of residues, and percent of total residues in protein-protein interface. Statistical significance values (Wilcoxon rank-sum test) were calculated

between feature values for sets of cases with Incorrect vs. Medium and Incorrect vs. High CAPRI accuracy, as noted at top (ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$).

Table 4.2. Analysis of geometric and energetic interface features associated with AlphaFold success or failure.

Interface Feature[a]	p-value Incorrect vs. Medium/High[b]	p-value Incorrect vs. High[b]
nres_tot	**0.02001**	**0.00022**
SASA_Tot	0.21257	**0.00253**
nres_percent	**0.00091**	**0.00412**
BSA_percent	**0.00038**	**0.00605**
per_residue_energy_int	0.53604	**0.04757**
hbond_E_fraction	0.19513	0.12584
sc_value	0.34752	0.18198
dG_cross	**0.00432**	0.22077
packstat	0.94851	0.34572
nres_int	0.07936	0.58908
ZRANK2	**0.00195**	0.66473
BSA	**0.00736**	0.84187
delta_unsatHbonds	0.19421	0.92855
hbonds_int	0.17700	0.99173

[a]Interface geometric or energetic feature, calculated using experimentally determined complex structures with the InterfaceAnalyzer protocol in Rosetta [165], with the exception of ZRANK2, SASA_Tot, BSA, BSA_percent, and nres_percent values. ZRANK2 scores were calculated with the ZRANK executable [75], SASA_Tot and BSA represent the total accessible surface area of the individual proteins and the buried interface surface area, calculated by naccess [170]. BSA_percent and nres_precent represent the proportion of total surface area or total residues in the interface, calculated using BSA and SASA_Tot (BSA_percent), or nres_tot and nres_int (nres_percent).

[b]P-values were calculated using Wilcoxon rank-sum test for sets of BM5.5 cases based on AlphaFold accuracy in the five models generated for that complex (all Incorrect, at least one model with Medium or High accuracy, at least one model with High accuracy). Significant p-values ($p < 0.05$) are shown in bold.

We also explored the accuracy of individual chain structural modeling and MSA depth (**Figure 4.6**); while a range of chain alignment depths (N_{eff}) were observed, in most cases the individual ligand and receptor chains were modeled accurately (backbone RMSD with bound component chains < 2.5 Å). Protein complex model accuracies based on interface residue RMSD (I-RMSD) and CAPRI criteria did not show a relationship with maximum subunit chain RMSD (**Figure 4.6c**), indicating that incorrect binding mode, versus inaccurate chain folding, was the primary cause of incorrect AlphaFold complex models. AlphaFold models representing incorrect binding mode and inaccurate chain folding are shown in **Figure 4.6d** and **Figure 4.6e**, respectively**.**

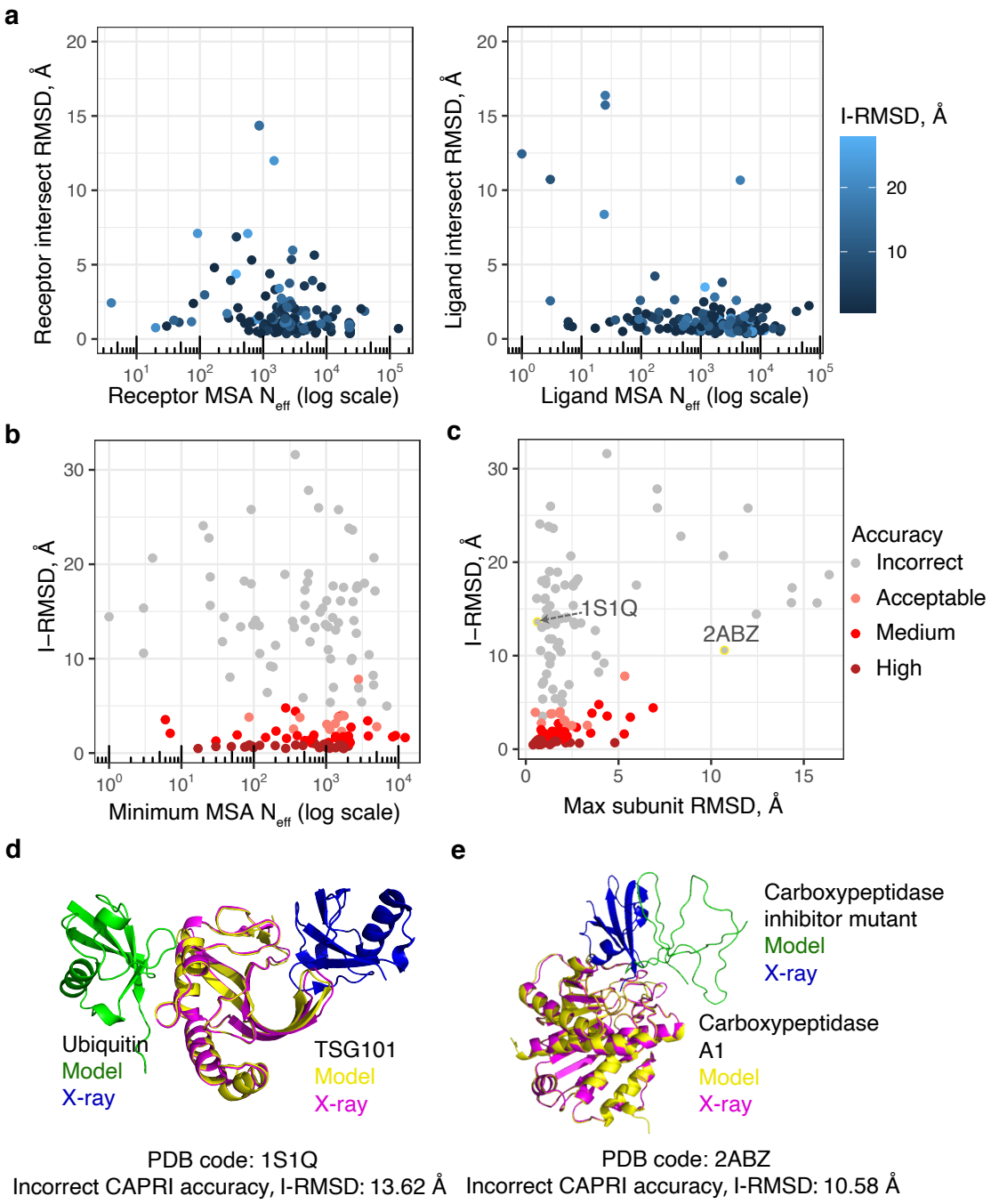

Figure 4.6. Comparison of MSA depth, subunit accuracy, and AlphaFold model quality. (a) Scatter plots depicting the relationship between ligand or receptor intersect RMSD and MSA depth (log scale; measured by N_{eff}, see Materials and Methods for details). Each point represents the top pTM model from each of the 149 cases modeled by AlphaFold. Points are colored by I-RMSD of the docking model. Scatter plot depicting the relationship between the docking model I-RMSD and **(b)** minimum ligand and receptor MSA depth (log scale, measured by N_{eff}), or **(c)** Maximum subunit (ligand or receptor) RMSD from X-ray structure. Points are colored by CAPRI accuracy and represent the top-ranked model (by pTM score) from each of the 149 cases modeled by AlphaFold. **(d)** Native and top-ranked AlphaFold model (pTM = 0.89) for PDB 1S1Q, superposed by TSG101. **(e)** Native and top-ranked AlphaFold model (pTM = 0.81) for PDB

2ABZ, superposed by Carboxypeptidase A1. Model and the X-ray structure chains are colored separately as indicated. Unresolved regions modeled by AlphaFold were omitted from the figures.

4.4.3 Impact of alternative AlphaFold parameters and input

Given the success of AlphaFold with unpaired MSAs, consisting of individual MSAs for each protein, we tested the impact of the use of paired sequences, which represent both chains as a single sequence in the MSA, within the input MSAs. Due to its capability to provide a co-evolution signal between protein residues across an interface, which can then be inferred as cross-interface contacts, use of paired sequences in MSAs has shown promise previously for protein complex structure prediction [91, 109, 247]. MSAs with paired sequences were obtained from the ColabFold Google Colab site (on September 4, 2021, using the MSA pairing protocol described by RoseTTAFold [91]) [226]; these sequence pairs were generated with an automated algorithm intended for prokaryotic proteins, thus a set of 17 cases from BM5.5 was tested that contain two prokaryotic proteins from the same organism. As shown in **Figure 4.7**, the addition of paired sequences did not appear to improve AlphaFold performance over use of unpaired MSAs as input, while use of paired sequences alone was detrimental to successful complex modeling in some cases. One notable exception was test case 1F6M, which had a relatively high number of paired sequences in the MSA. When paired sequences alone were used, High accuracy models were obtained for test case 1F6M, whereas no hits were obtained when unpaired sequences were included. For comparison, the same paired-only MSAs were input to RoseTTAFold [91], which according to its authors can utilize paired MSAs to predict complex structures; while some accurate models were obtained, we observed lower overall success for the models generated with that method.

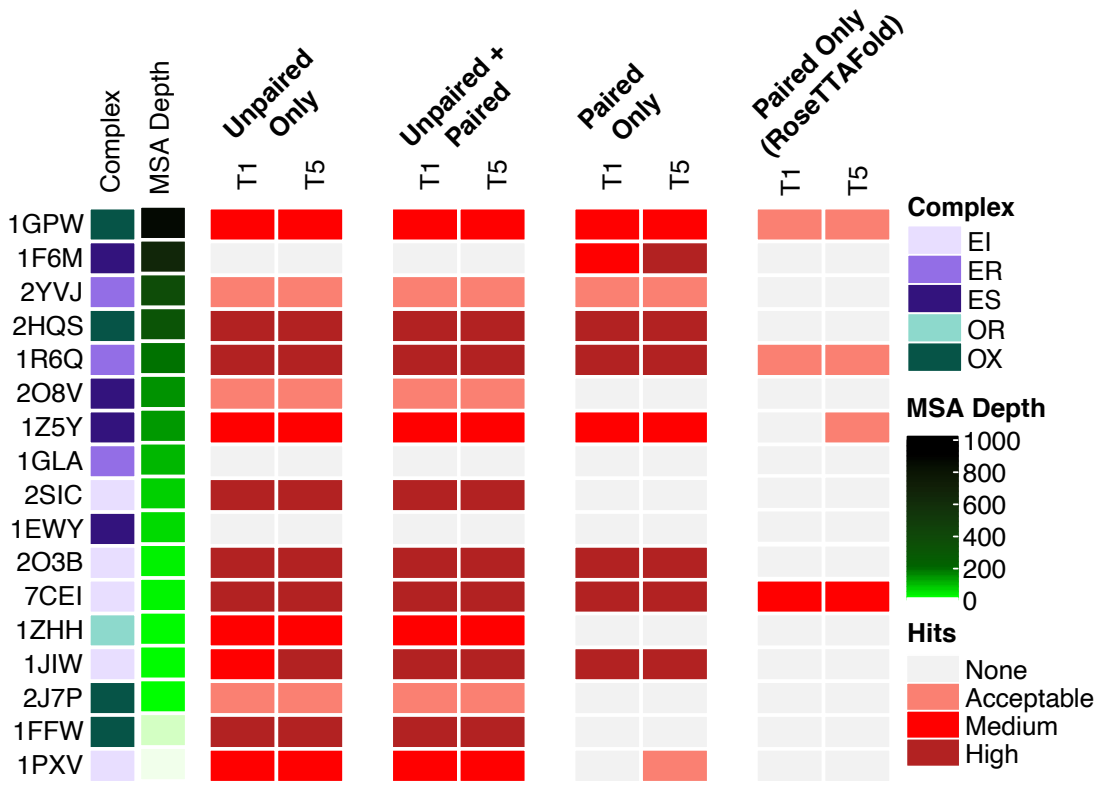

Figure 4.7. The impact of MSA pairing on prediction accuracy. MSAs were generated using MMseqs2 using the "advanced" interface of ColabFold [226]. Pairing was performed in ColabFold on a total of 17 cases whose ligand and receptor proteins come from the same prokaryotic organism. Cases in the heatmap were sorted by the paired MSA depth (N_{eff}; see Material and Methods for details) from the largest to the smallest values. Structural predictions were generated with the "advanced" interface of ColabFold, and RoseTTAFold [91] (through the Robetta server). All models were assessed for near-native predictions within the top-ranked (T1) and top 5 (T5) models using CAPRI criteria. Complex category: Enzyme-inhibitor (EI), enzyme complex with a regulatory or accessory chain (ER), enzyme-substrate (ES), others, receptor-containing (OR); others, miscellaneous (OX).

We separately compared the use of paired sequences, unpaired sequences, or both as MSA input for AlphaFold-Multimer [248], which was trained specifically to model protein-protein complexes (**Figure 4.8**). The paired-only results showed accuracy improvements in some cases versus the unpaired-only baseline, as well as unpaired+paired inputs (e.g. 1F6M, 1ZHH), while loss of near-native models for paired-only was observed for two cases with very low paired MSA depths (1FFW, 1PXV). Thus it seems possible that AlphaFold-Multimer can better utilize paired-

only inputs (with sufficient sequences) for complex modeling than AlphaFold, however it should be noted that the overlap of the set of complexes in this test set with the AlphaFold-Multimer training set (both interfaces and component proteins) may mask comparative differences among MSA inputs and likely leads to high overall baseline performance in **Figure 4.8**.

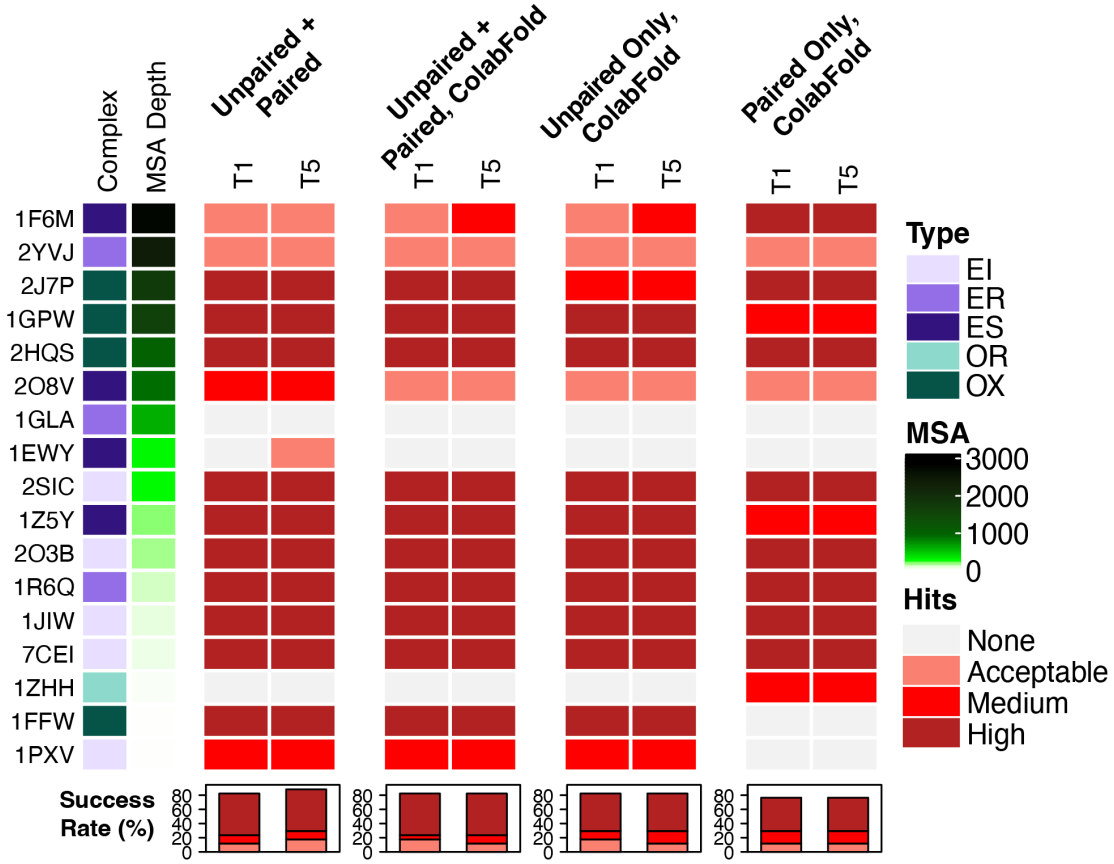

Figure 4.8. Testing the impact of MSA modes on AlphaFold-Multimer accuracy. Models were generated with AlphaFold-Multimer without templates on a set of 17 cases in from BM5.5 possessing chains from same prokaryotic species. Modeling was performed locally using AlphaFold-Multimer from the DeepMind Github downloaded protocol, or AlphaFold-Multimer in ColabFold, with the latter noted in column headers. All models were assessed for CAPRI accuracy, and model quality of top 1 and top 5 predictions (ranked by AlphaFold-Multimer score, 0.8*ipTM + 0.2*pTM) is shown and colored by CAPRI model accuracy as indicated in the key on right (Hits). Cases are sorted by the depth of the paired MSA (N_{eff}), generated by the AlphaFold-Multimer protocol in ColabFold, from the largest to the smallest values.

We also tested altered parameters for the number of iterative refinement cycles (N_{cycle}) and

MSA ensemble size ($N_{ensemble}$) in AlphaFold, for a subset of the docking test cases selected to represent the antibody, enzyme, and "other" protein complex types, and observed very little effect on predictive performance (**Figure 4.9**).

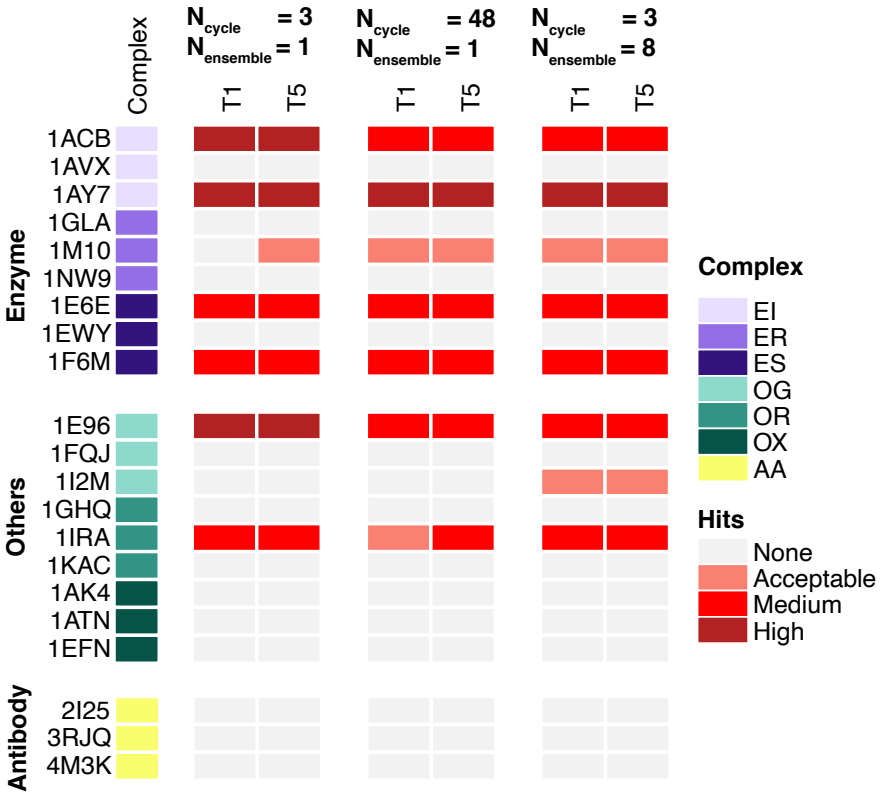

Figure 4.9. Testing the impact of alternative parameters on AlphaFold accuracy. Models were generated with AlphaFold for 21 test cases, representing three cases from each of the 7 complex categories, chosen by random. By default, N_{cycle} is set to 3 and $N_{ensemble}$ is set to 1. To test the impact of using alternative parameters, N_{cycle} and $N_{ensemble}$ were respectively increased to 48 and 8, while all other components of AlphaFold pipeline were kept constant. All models were assessed for near-native predictions using CAPRI criteria for High, Medium and Acceptable accuracy in the top-ranked (T1) and top 5 (T5) models. Complex category: Enzyme-inhibitor (EI), enzyme complex with a regulatory or accessory chain (ER), enzyme-substrate (ES), others, G-protein containing (OG), others, receptor-containing (OR); others, miscellaneous (OX), antibody-antigen (AA).

4.4.4 Docking model discrimination by scoring metrics

Given the reported success of AlphaFold in predicting the quality of its monomeric protein

models through scores representing local accuracy (pLDDT) and global accuracy (pTM)[111], we tested the discriminative capabilities of these values in the context of protein complex modeling (**Figure 4.10**). Average pLDDT scores and pTM scores for AlphaFold complex models were both found to discriminate Incorrect versus higher model accuracy classifications, with pTM scores performing moderately better (**Figure 4.10a**). Comparison of pTM with complex model TM-scores [231] showed a relatively strong correlation of the predicted with the calculated accuracy value ($r = 0.82$; $p < 0.001$; **Figure 4.10b**), while pTM exhibited a significant, though moderately weaker, correlation with I-RMSD of AlphaFold models ($r = -0.55$, $p < 0.001$; **Figure 4.10c**).

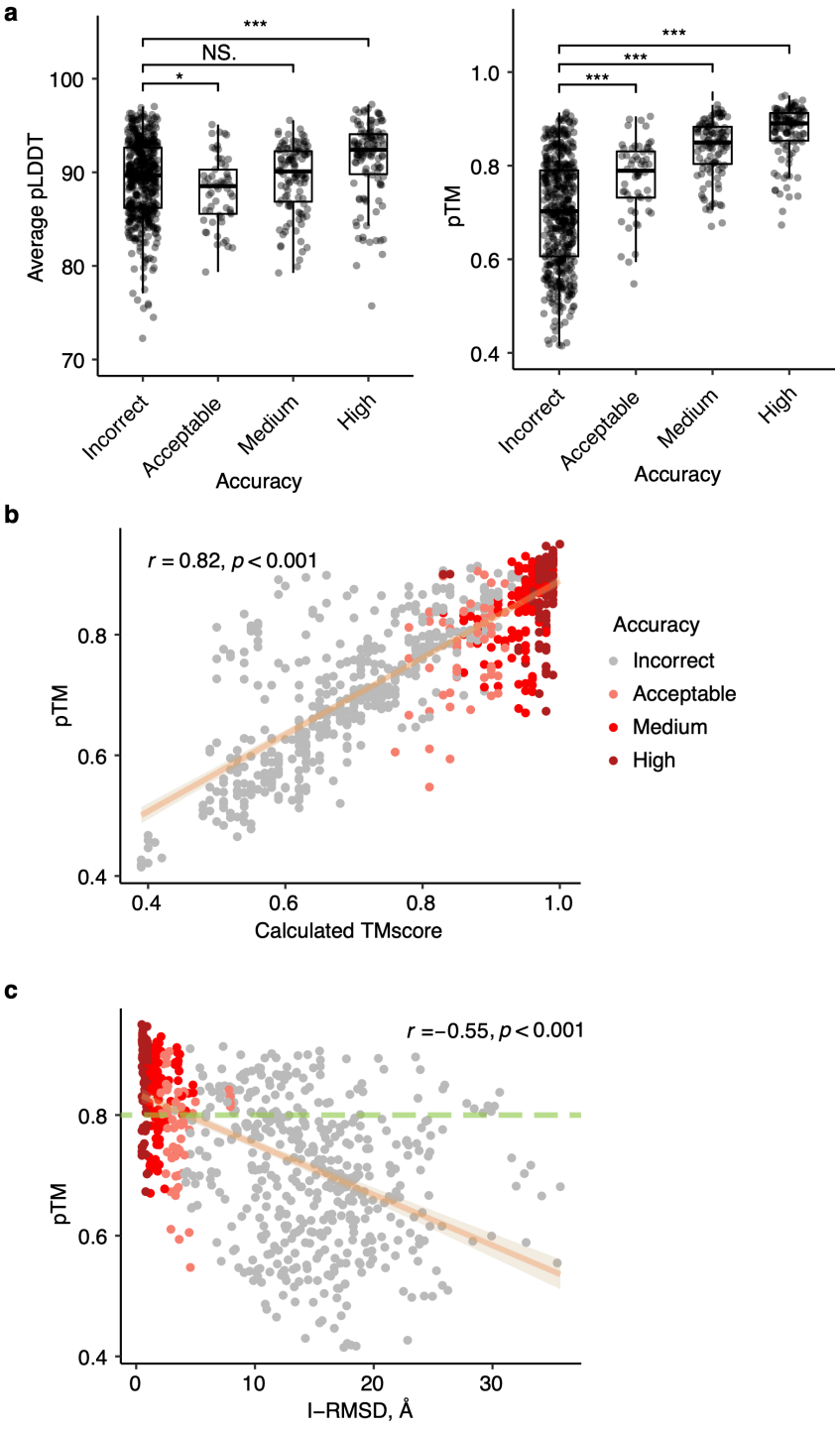

Figure 4.10. Association between AlphaFold predicted scores and docking model quality. (a) Average pLDDT and pTM per CAPRI criteria. Statistical significance (Wilcoxon rank-sum test) between average pLDDT or pTM of Incorrect vs. Acceptable, Incorrect vs. Medium and Incorrect vs. High CAPRI criteria is indicated at the top (ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$). (**b**) Comparisons between pTM and calculated TM score and (c) between pTM and I-RMSD are shown as scatter plots. All 5 models for 149 cases are shown as points, colored by model quality by CAPRI criteria. Linear regression is shown

along with the 95% confidence interval (orange area), and Pearson's correlation coefficients and correlation p-values are denoted in (**b**) and (**c**). In (**c**), the dashed green line indicates a possible pTM score cutoff (pTM = 0.8) for selection of accurate docking models, based on optimization of sensitivity and specificity for Incorrect versus Medium and High model discrimination.

While pTM and pLDDT showed some capability to identify correct versus incorrect complex structural models, the overlap in scores between accuracy categories (**Figure 4.10a**) led us to explore additional scoring functions to predict the structural quality of AlphaFold models (**Figure 4.11**, **Table 4.2**). Given the likely importance of interface residue contacts and packing, versus the folding accuracy of interface-distal protein regions, in discrimination of correct versus incorrect docking models, we tested two residue-level predicted accuracy metrics from AlphaFold, PAE (Predicted Aligned Error, corresponding to expected error in the position of one residue with respect to another residue in a model [249]) and pLDDT, for predicted protein-protein interface residues alone, to assess model discrimination capabilities. Alternative formulations of these metrics were tested with more permissive interface definitions, versus the originally tested 4 Å interface cutoff, but no major difference in model assessment accuracy was observed (**Figure 4.12, Table 4.4**). Interface PAE and interface pLDDT values showed major improvement compared with average pLDDT and pTM from AlphaFold in discriminating accurate complex models, based on receiver operating characteristic area under the curve (AUC) metrics (**Table 4.5**), particularly for the discrimination of models in the most populous and divergent Incorrect and High model accuracy categories (AUCs of 0.93 and 0.97 for interface PAE and interface pLDDT, respectively). Relatively high AUC values were also observed for previously reported docking model ranking methods ZRANK2 [75] and IRAD [238] (**Table 4.5**), while an interface energy score from Rosetta [165] (Cross-interface binding energy) resulted in the highest model classification accuracy, based on the binary classification AUC metrics (**Table 4.5**). However, estimated 95% confidence

intervals (included in **Table 4.5**) showed overlap between AUC value ranges for ZRANK2, IRAD, and Rosetta cross-interface binding energy for Incorrect versus Medium or High accuracy models, indicating that their performance is essentially equivalent for that model accuracy discrimination. Based on discrimination of Incorrect versus Medium and High accuracy models and maximization of sensitivity and sensitivity, possible score cutoffs for model selection are pTM = 0.8 (shown as dashed line in **Figure 4.10c**), interface pLDDT = 84, IRAD = -128, and Rosetta cross-interface binding energy = -16. While performing lower than the Rosetta binding energy score, some relatively simple protein interface assessments, such as the number of interface hydrogen bonds, showed some capability to classify the accuracy of AlphaFold models.

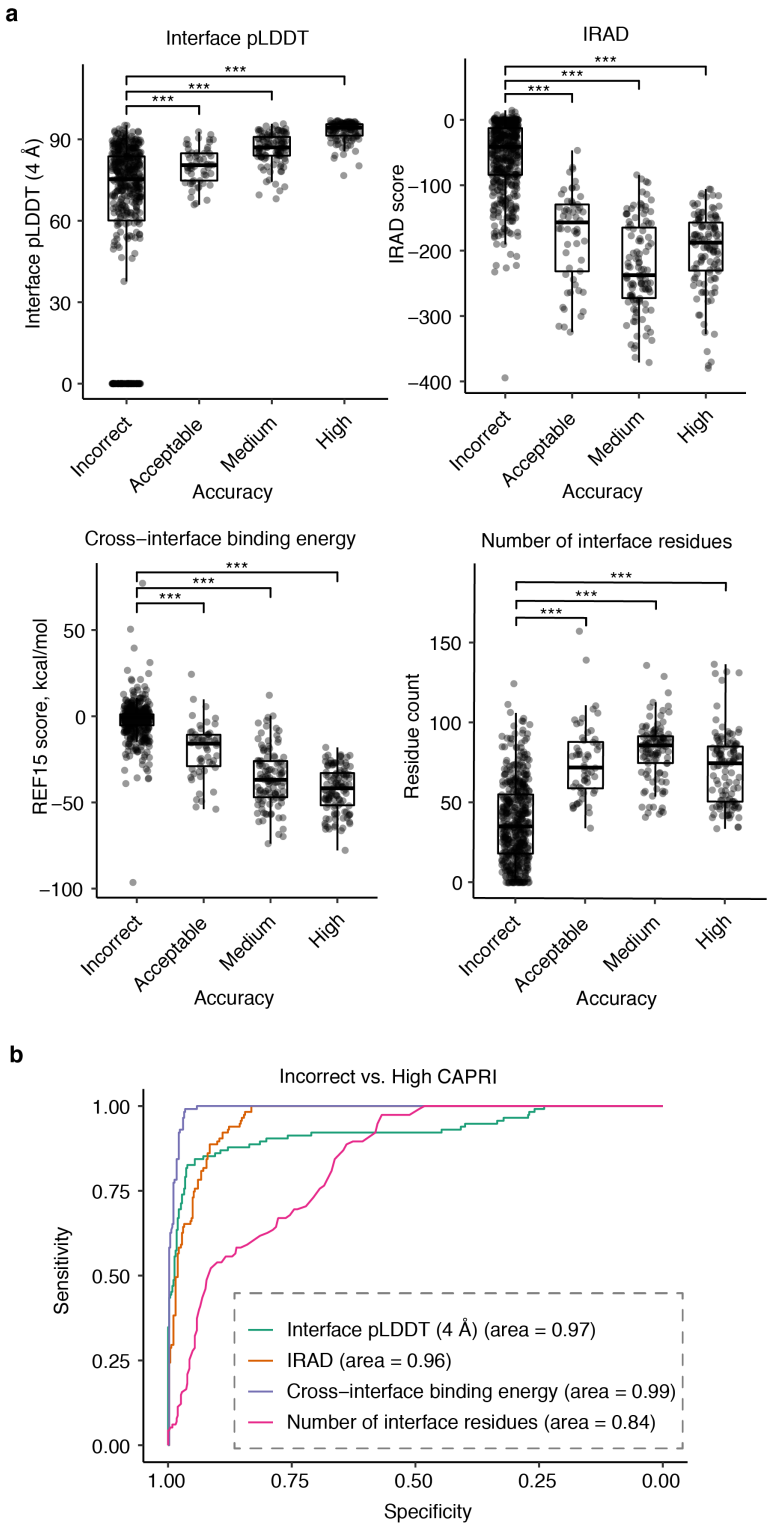

Figure 4.11. Association between alternative scoring metrics and docking model quality. (a) Distributions of interface pLDDT (4 Å), IRAD, Rosetta cross-interface binding energy, and number of interface residues for AlphaFold models grouped by CAPRI criteria. An interface pLDDT score of 0 was

assigned to models without any interface contacts within the distance cutoff (4 Å). Constrained local minimization was performed using Rosetta FastRelax [166] to resolve unfavorable local geometries or clashes in models, and post-relaxation models were scored with IRAD and the Rosetta [165] InterfaceAnalyzer protocol, with the latter used to calculate cross-interface binding energy scores (based on the Rosetta REF15 energy function [239]) and the number of interface residues. Statistical significance values (Wilcoxon rank-sum test) between scores of Incorrect vs. Acceptable, Incorrect vs. Medium, Incorrect vs. High CAPRI criteria are indicated at the top of each plot (***: $p \leq 0.001$). Each point corresponds to one AlphaFold model, and all 5 AlphaFold models for 149 test cases are represented. (**b**) ROC curves among the scoring metrics for classifying Incorrect vs High accuracy models by CAPRI criteria, with corresponding AUC values denoted in parentheses.

Table 4.3. Analysis of geometric and energetic interface features associated with AlphaFold success or failure.

Interface Feature[a]	p-value Incorrect vs. Medium/High[b]	p-value Incorrect vs. High[b]
nres_tot	**0.02001**	**0.00022**
SASA_Tot	0.21257	**0.00253**
nres_percent	**0.00091**	**0.00412**
BSA_percent	**0.00038**	**0.00605**
per_residue_energy_int	0.53604	**0.04757**
hbond_E_fraction	0.19513	0.12584
sc_value	0.34752	0.18198
dG_cross	**0.00432**	0.22077
packstat	0.94851	0.34572
nres_int	0.07936	0.58908
ZRANK2	**0.00195**	0.66473
BSA	**0.00736**	0.84187
delta_unsatHbonds	0.19421	0.92855
hbonds_int	0.17700	0.99173

[a]Interface geometric or energetic feature, calculated using experimentally determined complex structures with the InterfaceAnalyzer protocol in Rosetta [165], with the exception of ZRANK2, SASA_Tot, BSA, BSA_percent, and nres_percent values. ZRANK2 scores were calculated with the ZRANK executable [75], SASA_Tot and BSA represent the total accessible surface area of the individual proteins and the buried interface surface area, calculated by naccess [170]. BSA_percent and nres_precent represent the proportion of total surface area or total residues in the interface, calculated using BSA and SASA_Tot (BSA_percent), or nres_tot and nres_int (nres_percent).
[b]P-values were calculated using Wilcoxon rank-sum test for sets of BM5.5 cases based on AlphaFold accuracy in the five models generated for that complex (all Incorrect, at least one model with Medium or High accuracy, at least one model with High accuracy). Significant p-values ($p < 0.05$) are shown in bold.

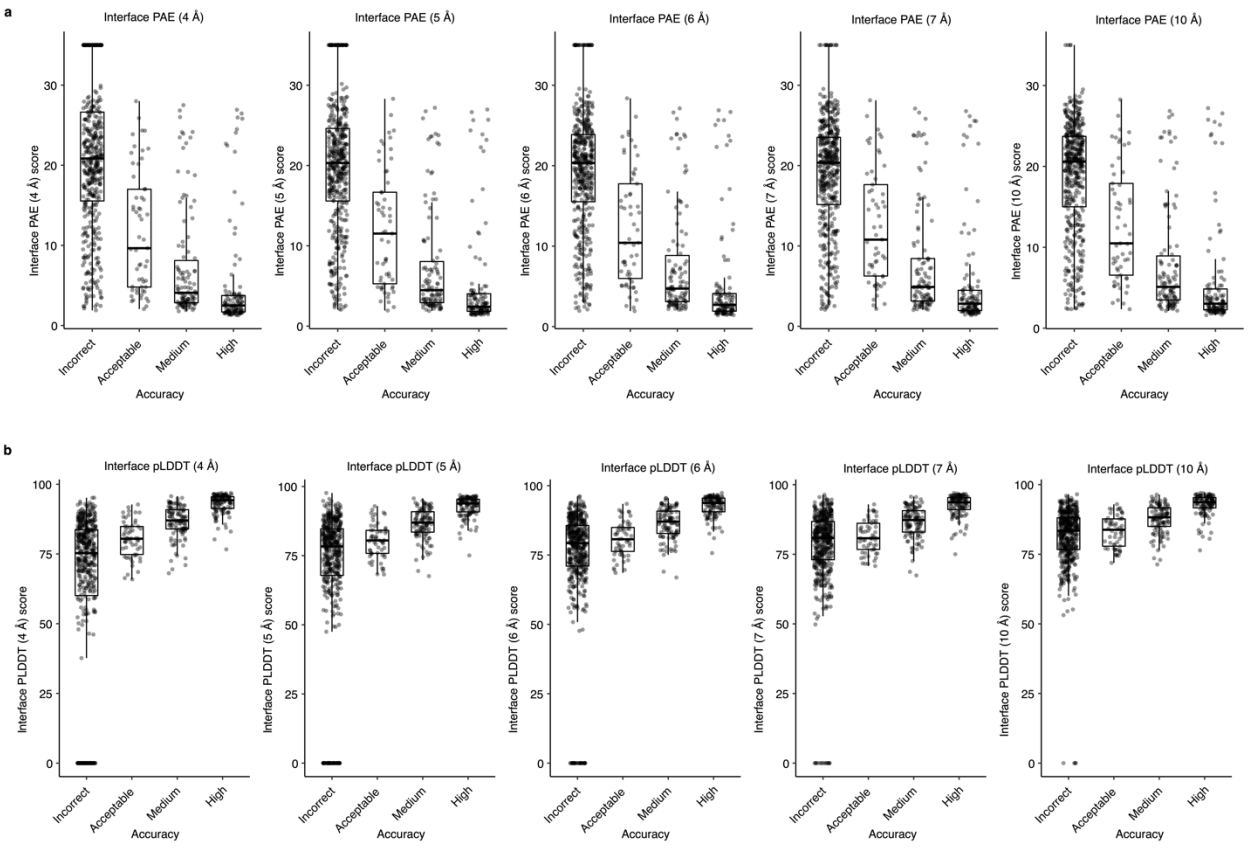

Figure 4.12. Alternative formulation of interface accuracy metrics calculated from AlphaFold predictions (PAE, pLDDT). (a) Interface PAE within a distance cutoff of 4 Å, 5 Å, 6 Å, 7 Å and 10 Å grouped by docking model accuracy. (b) Interface pLDDT within a distance cutoff of 4 Å, 5 Å, 6 Å, 7 Å and 10 Å, grouped by docking model accuracy. An interface pLDDT score of 0 and an interface PAE score of 35 are assigned to models without interface contacts within the distance cutoff specified. See also Table 6.2 for AUC values for interface pLDDT and interface PAE as classifiers for docking model accuracy.

Table 4.4. Area under the ROC curve (AUC) value for protein quality classes as a function of interface scores calculated from AlphaFold predictions.

Score[a]	Binary classification		Multi-class classification
	Incorrect vs. High	Incorrect vs. Medium and High	
Interface PAE (4 Å)	0.93	0.90	0.81
Interface PAE (5 Å)	0.92	0.89	0.81
Interface PAE (6 Å)	0.91	0.89	0.80
Interface PAE (7 Å)	0.91	0.88	0.80
Interface PAE (10 Å)	0.91	0.88	0.80
Interface pLDDT (4 Å)	0.97	0.90	0.84
Interface pLDDT (5 Å)	0.95	0.88	0.82
Interface pLDDT (6 Å)	0.94	0.85	0.80
Interface pLDDT (7 Å)	0.93	0.83	0.77
Interface pLDDT (10 Å)	0.91	0.81	0.77

[a] Scoring methods. “Interface PAE”: the average PAE score of pairs of interface residues within the interface distance cutoff specified in parenthesis; “Interface pLDDT”: the average pLDDT of interface residues within the interface distance cutoff specified in parenthesis.

Table 4.5. Area under the ROC curve (AUC) values and 95% confidence intervals (shown in parentheses) for protein quality classes as a function of different scoring metrics.

Score[a]	Binary classification[b]		Multi-class classification[c]
	Incorrect vs. High	Incorrect vs. Medium and High	
average pLDDT	0.66 (0.60 to 0.72)	0.59 (0.54 to 0.63)	0.64 (0.58 to 0.67)
average resolved pLDDT	0.81 (0.76 to 0.85)	0.69 (0.64 to 0.72)	0.69 (0.65 to 0.73)
pTM	0.92 (0.89 to 0.95)	0.89 (0.86 to 0.91)	0.80 (0.76 to 0.82)
interface PAE (4 Å)	0.93 (0.89 to 0.96)	0.90 (0.87 to 0.92)	0.81 (0.77 to 0.83)
interface pLDDT (4 Å)	0.97 (0.95 to 0.98)	0.90 (0.87 to 0.92)	0.84 (0.82 to 0.85)
IRAD	0.96 (0.95 to 0.98)	0.97 (0.95 to 0.97)	0.80 (0.76 to 0.82)
ZRANK	0.95 (0.93 to 0.96)	0.95 (0.94 to 0.96)	0.77 (0.73 to 0.79)
cross-interface binding energy	0.99 (0.99 to 1.00)	0.97 (0.95 to 0.98)	0.84 (0.81 to 0.86)
interface area	0.90 (0.88 to 0.93)	0.92 (0.90 to 0.94)	0.77 (0.73 to 0.79)
number of interface hydrogen bonds	0.96 (0.95 to 0.98)	0.94 (0.92 to 0.95)	0.81 (0.78 to 0.83)
number of interface residues	0.84 (0.80 to 0.87)	0.87 (0.84 to 0.89)	0.73 (0.66 to 0.74)
shape complementarity	0.91 (0.88 to 0.93)	0.85 (0.81 to 0.87)	0.79 (0.76 to 0.81)

[a]Scoring methods. “average resolved pLDDT”: average pLDDT on the resolved region, “interface PAE (4 Å)”: average PAE of pairs of interface residues within 4 Å distance cutoff, “interface pLDDT (4 Å)”: average pLDDT of interface residues within 4 Å distance cutoff. “cross-interface binding energy”, “interface area”, “number of interface hydrogen bonds”, “number of interface residues” and “shape complementarity” were calculated using the Rosetta InterfaceAnalyzer (see Methods for details).
[b] The AUC values of the binary classification were calculated using the pROC package [250] in R. The 95% confidence intervals were calculated by pROC.
[c] The AUC values of the multi-class classification were calculated with multiROC package [251, 252] in R. The 95% confidence intervals of multi-class AUC values were calculated with the boot package [242] in R with adjusted bootstrap percentile (BCa) interval.

4.4.5 Expanded antibody-antigen complex benchmarking

Due to the lack of any successful structural prediction of 11 antibody-antigen complexes from the Benchmark 5.5 set, we assembled a set of 20 additional nonredundant antibody-antigen complexes with known structures to assess AlphaFold accuracy (**Table 4.6**). These complexes include a variety of antigens, and as the Benchmark 5.5 heterodimer set included only nanobodies, a number of single-chain antibodies with both heavy and light chains represented were selected for the additional cases (comprising 17 out of 20 of the cases), while the remaining three cases include single-domain nanobodies. While AlphaFold modeling of most of those complex structures resulted in no accurate predictions, surprisingly two of the antibody-antigen complexes were modeled accurately, with Medium CAPRI accuracy models ranked #1 for each complex (**Table 4.6**, with models shown in **Figure 4.13a** and **4.13b**).

Table 4.6. Additional antibody-antigen complex test cases and AlphaFold prediction success.

PDB	Antibody	Antigen	Type[a]	AlphaFold T1[b]	AlphaFold T5[b]
5F72	LS146	KEAP1	ab	Incorrect	Incorrect
1DZB	1F9	HEW	ab	Incorrect	Incorrect
3UZE	4E11	Dengue E DIII	ab	Incorrect	Incorrect
6EJM	scFv 5	CD81 LEL	ab	Incorrect	Incorrect
5DFW	K13	CD81 LEL	ab	Incorrect	Incorrect
6I07	MM131	EpCAM	ab	Incorrect	Incorrect
5JYM	TSP11	P-cadherin	ab	Incorrect	Incorrect
4NIK	F5	Gankyrin	ab	**Medium**	**Medium**
5JYL	TSP7	P-cadherin	ab	Incorrect	Incorrect
6TOU	RVC20	Rabies gp	ab	Incorrect	Incorrect
6EK2	scFv 10	CD81 LEL	ab	Incorrect	Incorrect
4YJZ	H2526	H1 HA	ab	Incorrect	Incorrect
3UX9	AIFNα1bScFv01	interferon alpha	ab	Incorrect	Incorrect
6OAN	Antibody 053054	P vivax DBP	ab	**Medium**	**Medium**
6J71	HUA21	HER2	ab	Incorrect	Incorrect

7DET	PR961	SARS-CoV-2 RBD	ab	Incorrect	Incorrect
7DEO	PR1077	SARS-CoV-2 RBD	ab	Incorrect	Incorrect
6WAQ	VHH-72	SARS-CoV RBD	nano	Incorrect	Incorrect
6EY0	NB01	PorM	nano	Incorrect	Incorrect
6OQ7	E3	Clostridium difficile toxin B	nano	Incorrect	Incorrect

[a] Antibody type: "ab": heavy-light chain antibody, "nano": nanobody/VHH.
[b] AlphaFold modeling accuracy in top 1 and top 5 models (ranked by pTM). Model quality was assessed by CAPRI criteria. Medium CAPRI accuracy levels are highlighted with bold font.

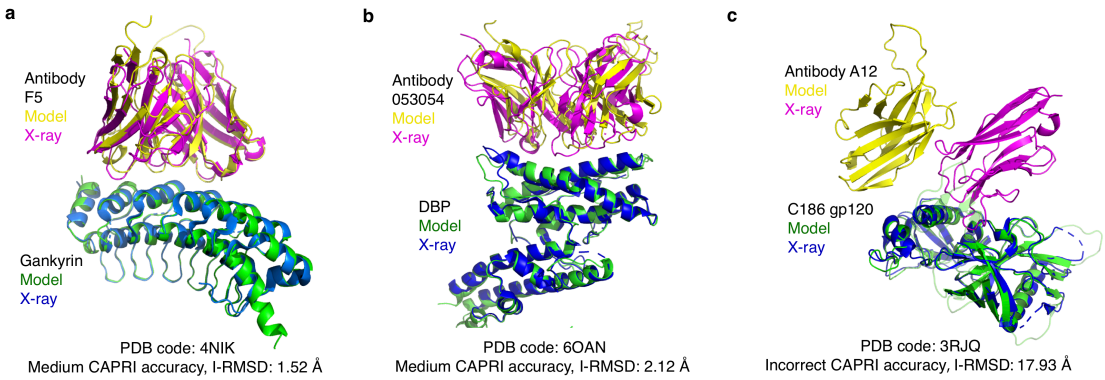

Figure 4.13. Examples of antibody-antigen complex structure predictions by AlphaFold. (a) Native and top-ranked AlphaFold model (pTM = 0.78) for PDB 4NIK (F5 antibody/human gankyrin complex). This model is of Medium accuracy by CAPRI criteria (I-RMSD = 1.52 Å). Modeled and X-ray complex structures are colored as indicated and shown superposed by gankryin. Unresolved regions modeled by AlphaFold are not shown. (b) Native and top-ranked AlphaFold model (pTM = 0.61) for PDB 6OAN (053054 antibody/P vivax DBP complex). This model is of Medium accuracy by CAPRI criteria (I-RMSD = 2.12 Å). Modeled and X-ray structures are colored as indicated, shown superposed by DBP, and unresolved regions modeled by AlphaFold are not shown. (c) Native and top-ranked AlphaFold model (pTM = 0.66) for PDB 3RJQ (A12 nanobody/HIV C186 gp120 complex), superposed by C186 gp120. This AlphaFold model does not have contacting residues between the proteins within a 5 Å distance cutoff. Structures are colored as indicated in the figure, and unresolved regions modeled by AlphaFold on C186 gp120 are shown in light green.

Inspection of AlphaFold models of antibody-antigen complexes indicated that many of the

inaccurate models had few or no contacts between antibody and antigen chains; one example is shown in **Figure 4.13c**). Indeed, analysis of the percentage of models with no atomic contacts between chains showed that antibody-antigen cases had relatively high rates of such models in comparison with the other protein complex categories in the benchmark (**Figure 4.14**).

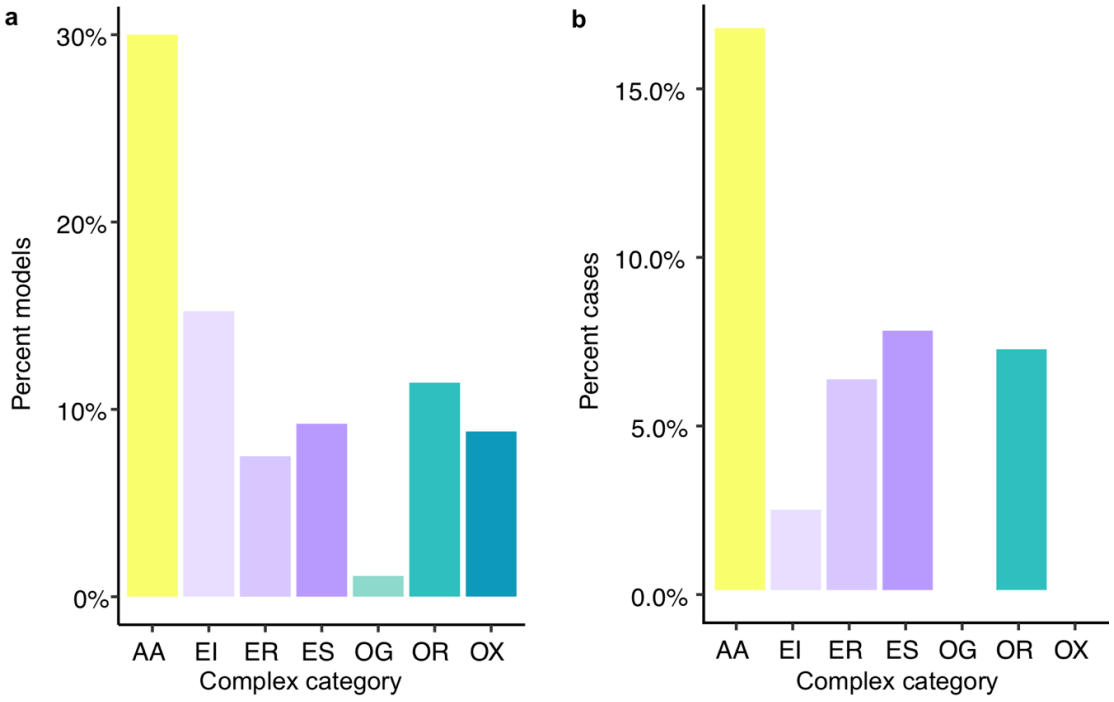

Figure 4.14. Distribution of AlphaFold models without inter-chain atomic contacts by complex category. (a) Percentage of AlphaFold models without inter-chain atomic contacts within 4 Å distance cutoff per complex category. (b) Percentage of cases in which the top-ranked model (ranked by pTM) predicted by AlphaFold has no inter-chain atomic contacts within 4 Å distance cutoff in each complex category. Complex category: Antibody-antigen (AA), enzyme-inhibitor (EI), enzyme complex with a regulatory or accessory chain (ER), enzyme-substrate (ES), others, G-protein containing (OG), others, receptor-containing (OR); others, miscellaneous (OX).

4.4.6 AlphaFold performance for non-immunoglobulin antibody-antigen complexes

After confirming the limited success of AlphaFold in predicting antibody-antigen complex structures, we performed additional modeling assessments in AlphaFold to identify factors

responsible for that performance. While smaller interface size and larger complex structure size were found to be associated with lower AlphaFold success for the overall set of cases (**Figure 4.4, Table 4.2**), additional features specific to antibody-antigen complex structures or sequences likely reduce AlphaFold performance for that class. To assess whether the immunoglobulin architecture shared by the antibodies impacted AlphaFold performance, we modeled a set of complexes containing non-immunoglobulin receptors in complex with protein targets (**Table 4.7**). These receptors correspond to variable lymphocyte receptors (VLRs), which are adaptive immune receptors found in jawless vertebrates (e.g. sea lampreys), and recognize protein and non-protein antigens with leucine-rich repeat architectures [18]. Three complex structures with VLR-based receptors, referred to as repebodies [253], were also included in this set of cases. Only one out of the seven VLR and repebody complexes tested had any correct models from AlphaFold (**Table 4.7**), indicating that the immunoglobulin architecture was not responsible for the observed limited AlphaFold success for antibody-antigen complexes.

Table 4.7. AlphaFold modeling success for VLR-antigen and repebody-antigen complexes.

PDB	Receptor	Antigen	Release Date[a]	AlphaFold T1[b]	AlphaFold T5[b]
3G3A	VLRB.2D	HEL	6/23/09	Incorrect	Incorrect
3M18	VLRA.R2.1	HEL	6/30/10	Incorrect	Incorrect
6BXC	VLR9	Zebrafish TLR5	5/9/18	Incorrect	Incorrect
6BXA	VLR2	Zebrafish TLR5	5/9/18	**High**	**High**
5B4P	r-3E8 repebody	Human C5a	4/12/17	Incorrect	Incorrect
6HTF	rF10 repebody	Human Btk SH2	5/20/20	Incorrect	Incorrect

6LBX	Rb-H2 repebody	HER2 Domain IV	11/18/20	Incorrect	Incorrect

[a]Release date of the experimentally determined complex structure in the Protein Data Bank.
[b]AlphaFold modeling accuracy in top 1 (T1) and top 5 (T5) models (ranked by pTM). Model quality was assessed by CAPRI criteria. High CAPRI accuracy levels are highlighted with bold font.

4.4.7 AlphaFold-Multimer performance for antibody-antigen and T cell receptor complex modeling

After observing limited success of AlphaFold for modeling of antibody-antigen complexes, we tested modeling of that class of complexes with AlphaFold-Multimer [248]. As AlphaFold-Multimer training included protein-protein interfaces from structures released before May 2018 [248], the test set included only antibody-antigen complexes from May 2018-present that are not redundant with pre-May 2018 structures, generated as part of an update to our recently reported set of antibody-antigen docking test cases [71]. Additionally, to investigate the impact of MSAs on antibody-antigen performance, we modeled the antibody-antigen complex structures with and without MSA input. In this context, we allowed the use of structural templates for each chain, in order to focus on complex modeling accuracy without reduction of tertiary structure fidelity due to the lack of MSAs Out of seven antibody-antigen complexes in the test set (**Table 4.8**), one complex (6U54) contains a nanobody. For comparison, recently released non-antibody complex structures from the “Benchmark 2” set described by Ghani et al. [254] were also tested.

Table 4.8. AlphaFold-Multimer performance for recently released antibody-antigen and non-antibody complex structures, with and without multiple sequence alignment input.

Set	Case	With MSA[a]		No MSA[a]	
		T1	T5	T1	T5
Antibody-antigen complexes	6A4K	Incorrect	Incorrect	Incorrect	Incorrect
	6HX4	Medium	Medium	Incorrect	Incorrect
	6P50	Incorrect	Incorrect	Incorrect	Incorrect
	6PXH	Incorrect	Incorrect	Incorrect	Incorrect
	6Q0O	Acceptable	Acceptable	Incorrect	Incorrect
	6U54	Medium	Medium	Medium	Medium
	6ZTR	Incorrect	Incorrect	Incorrect	Incorrect
Other complexes (non-antibody) from Ghani et al. [254]	5ZNG	Incorrect	Incorrect	Incorrect	Incorrect
	6A6I	Acceptable	Acceptable	Incorrect	Incorrect
	6GS2	Medium	Medium	Incorrect	Incorrect
	6H4B	Medium	High	Incorrect	Incorrect
	6IF2	Medium	Medium	Incorrect	Incorrect
	6II6	High	High	Incorrect	Incorrect
	6ONO	Medium	Medium	Incorrect	Incorrect
	6PNQ	Incorrect	Incorrect	Incorrect	Incorrect
	6Q76	High	High	High	High
	6U08	High	High	Incorrect	Incorrect
	6ZBK	High	High	Acceptable	Acceptable
	7AYE	High	High	Acceptable	Acceptable
	7D2T	High	High	Incorrect	Incorrect
	7M5F	Medium	Medium	Incorrect	Incorrect
	7N10	High	High	Incorrect	Incorrect
	7NLJ	Incorrect	Incorrect	Incorrect	Incorrect

	7P8K	Medium	Medium	Incorrect	Incorrect

[a]Modeling was performed using AlphaFold-Multimer [248], with multiple sequence alignment ("With MSA") or without multiple sequence alignment (single sequence, "No MSA") feature input. Shown are CAPRI model accuracy levels for top-ranked model (T1) and five models (T5) for each case, with Medium and High accuracy levels highlighted with light red and dark red cell shading, respectively. Structural templates for subunits were enabled for all runs, to allow for accurate modeling of individual chains in the absence of MSAs, with a date cutoff of 4/30/2018 to avoid use of the bound complex subunit structures as templates.

Results from this assessment, shown in **Table 4.8**, highlight a major difference in overall predictive success for standard AlphaFold-Multimer (with MSA input) between non-antibody success (13/17 cases, or 76%, with a Medium/High accuracy model ranked #1) versus antibody-antigen case success (2/7 cases, or 29%, with a Medium/High accuracy model ranked #1). Furthermore, our results indicate that while the lack of MSA input and corresponding residue co-evolutionary signal has a pronounced impact on near-native accuracy (CAPRI Medium/High models) for non-antibody complexes, it appears to have less of impact on antibody-antigen complex structure prediction. Additional AlphaFold-Multimer modeling of the same set of antibody-antigen complexes with no subunit templates and with MSA input did not affect predictive performance (**Table 4.9**). Taken together with the results for the non-immunoglobulin (VLR) cases, it appears that the limited success of antibody-antigen complex modeling AlphaFold and AlphaFold-Multimer is largely due to the lack of co-evolution signal, demonstrated by the lack of effect of MSA input, versus structural or geometric features of those interfaces. Of relevance, others have recently noted the importance of MSAs and co-evolution signals in AlphaFold's global conformational search [255]. As the AlphaFold-Multimer algorithm generates an interface pTM score (ipTM) which is used in conjunction with pTM to compute model scores

[248], we examined the use of ipTM score alone in model accuracy discrimination for models from the set of cases in **Table 4.8**, and ipTM alone showed promising model discrimination accuracy, with an ipTM score threshold of approximately 0.75 corresponding to a possible model confidence cutoff (**Figure 4.15**).

Table 4.9. AlphaFold-Multimer modeling success for recently released antibody-antigen complex structures.

PDB	Antibody	Antigen	Release Date[a]	T1[b]	T5[b]
6A4K	32D6 Fab	H1N1 hemagglutinin	3/27/19	Incorrect	Incorrect
6HX4	1D9 Fab	Alpha-1-antitrypsin	10/30/19	**Medium**	**Medium**
6P50	4A10 Fab	Interleukin-7 receptor subunit alpha	9/4/19	Incorrect	Incorrect
6PXH	G2 Fab	MERS-CoV spike NTD	9/25/19	Incorrect	Incorrect
6Q0O	H2227 Fab	H1N1 hemagglutinin	12/18/19	Acceptable	Acceptable
6U54	ZC nanobody	Ebolavirus nucleoprotein	11/6/19	**Medium**	**Medium**
6ZTR	CQY684 Fab	Cadherin-3	5/5/21	Incorrect	Incorrect

[a]Release date of the experimentally determined complex structure in the Protein Data Bank.
[b]AlphaFold-Multimer modeling accuracy in top 1 (T1) and top 5 (T5) models (ranked by AlphaFold-Multimer score). AlphaFold-Multimer was run with MSA input and without the use of structural

templates. Model quality was assessed by CAPRI criteria. Medium CAPRI accuracy levels are highlighted with bold font.

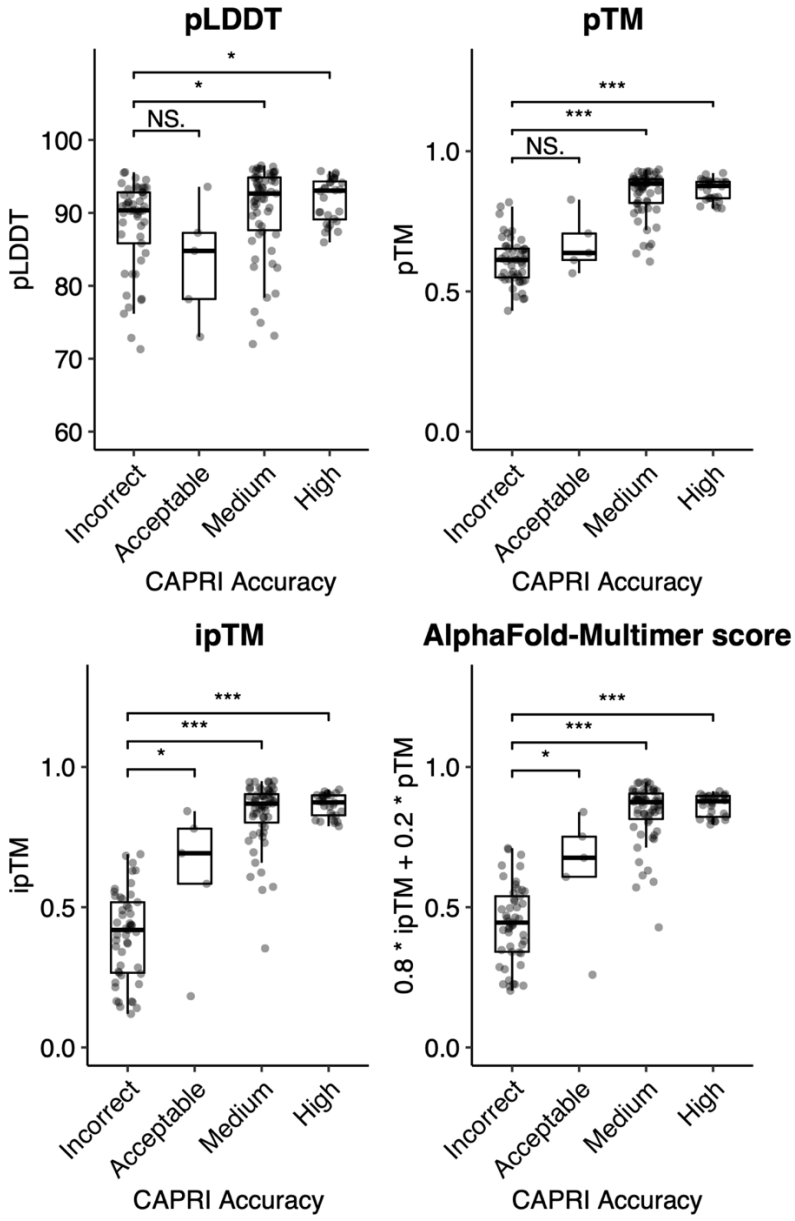

Figure 4.15. Discrimination of AlphaFold-Multimer model accuracies with pLDDT, pTM, ipTM, and AlphaFold-Multimer model scores. A benchmark set of recently released complex structures (from Table 2) was modeled with AlphaFold-Multimer, and all five models from AlphaFold-Multimer per test case were assessed for accuracy using CAPRI criteria. Accuracy-binned models are shown in comparison with pLDDT, pTM, ipTM, and AlphaFold-Multimer (0.8*ipTM + 0.2*pTM) scores. Statistical significance values (Wilcoxon rank-sum test) are shown for group comparisons (NS: $p > 0.05$, *: $p \leq 0.05$, ***: $p \leq 0.001$).

Due to the relatively limited number of antibody-antigen cases tested initially with AlphaFold-Multimer (**Table 4.8**), we assembled a larger set of 100 recently released antibody-antigen complex structures for benchmarking AlphaFold-Multimer for predictive performance with that class of complexes. All of the structures were recently released (May 2018 or later), and they contain complexes with heavy-light chain antibodies (72 complexes) and nanobodies (28 complexes) (**Table 4.10**). All complexes were modeled with AlphaFold-Multimer, and success rates for this set were found to be low, with 6% of cases with Medium or High accuracy models ranked #1, and 11% of cases with Medium or High accuracy models in one or more of the five models generated per case (**Figure 4.16**). This is in accordance with the limited success for antibody-antigen complex modeling briefly noted by the AlphaFold-Multimer developers in a preprint [248], and is similar to the success observed for the non-antibody-antigen cases noted above without MSA input (1 out of 17 cases, or 6% success; **Table 4.8**).

Table 4.10. Expanded set of recent antibody-antigen complex structures used for benchmarking AlphaFold-Multimer.

PDB[a]	Antibody	Antigen	Nanobody[b]	Release Date[c]
6a0z	Fab 13D4	Hemagglutinin,Envelope glycoprotein	0	2018-06-20
6dfj	anti-Zika antibody Z021	Dengue 1 Envelope DIII domain	0	2018-10-24
6ddv	Anti-MICA Fab fragment clone 6E1	MHC class I chain-related protein A	0	2018-10-24
6mej	antibody HEPC3	E2 glycoprotein	0	2018-11-21
6e63	TB31F Fab	Pf48/45	0	2018-11-28
6h70	Nano-62	Capsid protein VP1	1	2018-12-19
6iuv	3C11	Hemagglutinin	0	2019-01-16

6a77	Anti-human Robo1 antibody B5209B Fab	Roundabout homolog 1	0	2019-01-30
6iek	Fab R12	NSmGnGc	0	2019-04-10
6dkj	hGIPR-Ab Fab	Gastric inhibitory polypeptide receptor	0	2019-05-08
6gku	Fab 6F5	Galectin-10	0	2019-06-05
6iap	Fab NKp46-1	Natural cytotoxicity triggering receptor 1	0	2019-06-12
6hhu	nanobody AF684	S-layer protein sap	1	2019-07-24
6nmv	Fab 218 anti-SIRP-alpha antibody	Tyrosine-protein phosphatase non-receptor type substrate 1	0	2019-08-07
6oy4	7A1	Der p 2 variant 3	0	2019-08-28
6hhc	FXIA ANTIBODY Fab	Coagulation factor XI	0	2019-09-11
6mg7	CH54 Fab	clade A/E 93TH057 HIV-1 gp120 core	0	2019-09-25
6phc	2544 Fab	25 kDa ookinete surface antigen	0	2019-10-02
6hjx	nanobody 72	Cys-loop ligand-gated ion channel	1	2019-10-09
6j15	GY-5 Fab	Programmed cell death protein 1	0	2019-11-06
6ir2	mCherry's nanobody LaM2	MCherry fluorescent protein	1	2019-11-13
6ir1	mCherry's nanobody LaM4	MCherry fluorescent protein	1	2019-11-13
6ppg	Fab MCAF5352A	Interleukin-17F	0	2019-12-11
6k0y	mAb059c	Programmed cell death protein 1	0	2019-12-11
6ion	anti-C4.4A antibody 11H10	Ly6/PLAUR domain-containing protein 3	0	2020-01-15
6o9i	Fab 2	Gastric inhibitory polypeptide receptor	0	2020-01-22
6vel	66E8 Fab	Ubiquitin-like protein SMT3,Cadherin-1	0	2020-01-29
6ssi	NANOBODY 22	Cys-loop ligand-gated ion channel	1	2020-02-12

6os2	Nanobody Nb.AT110i1_le	Type-1 angiotensin II receptor,Soluble cytochrome b562 BRIL fusion protein	1	2020-02-19
6uft	JLK-G12	Botulinum neurotoxin type B	1	2020-03-04
6ui1	ciA-B5	BoNT/A	1	2020-03-04
6lz9	t8E4 Fab	Hepatocyte growth factor	0	2020-03-11
6tej	Syb_NL5	Drug ABC transporter ATP-binding protein	1	2020-04-01
6obo	VHH antibody V6A6	Ricin	1	2020-04-01
6ute	Z032 Fab	Envelope domain III	0	2020-04-15
6xw7	Nanobody NB-5829	Capsid protein	1	2020-04-22
6lr7	Nanobody LaG16	Green fluorescient protein uv	1	2020-04-29
6u9s	5A6 Fab	CD81 antigen	0	2020-05-13
6p3r	Fab H5.31	Hemagglutinin	0	2020-05-27
7bwj	P2B-2F6	Spike protein S1	0	2020-06-03
6m3b	c25k23 Fab	Vitamin K-dependent protein C	0	2020-07-08
7jmp	COVA2-39	Spike protein S1	0	2020-08-26
6xkr	Sasanlimab Fab	Programmed cell death protein 1	0	2020-09-09
6wh9	1D10	KR1	0	2020-09-09
6w7s	2G10 Fab	EryAI	0	2020-09-09
7chf	BD-368-2 Fab	Spike protein S1	0	2020-09-16
6xzw	Fab 4B3	Lipoprot_C domain-containing protein	0	2020-10-14
7a29	Sb23	Spike glycoprotein	1	2020-10-21
7kkl	Synthetic nanobody mNb6	Spike glycoprotein	1	2020-11-11
7kfv	C1A-B3 Fab	Spike glycoprotein	0	2020-12-02
7dm2	Fab p4-170	Phosphate-binding protein PstS 1	0	2020-12-23
6ye3	UFKA-20	Interleukin-2	0	2020-12-30
7chz	IgG26A	Interleukin-1 beta	0	2021-01-13

7kn6	VHH V	Spike protein S1	1	2021-01-20
7kn5	VHH E	Spike protein S1	1	2021-01-20
7kgj	Sb45	Spike glycoprotein	1	2021-02-03
7kzb	CR3014-C8 antibody	Spike glycoprotein	0	2021-02-03
7bei	COVOX-150	Spike glycoprotein	0	2021-03-03
7lab	DH1052	Spike glycoprotein	0	2021-03-10
7czw	P5A-2G7	Spike glycoprotein	0	2021-03-10
7lbg	Fab 13H11	Envelope glycoprotein H	0	2021-03-10
6was	GN1_PA8 Fab	1FD6 16055 V1V2 scaffold	0	2021-03-31
7lm8	COVA1-16 Fab	Spike protein S1	0	2021-03-31
7lsg	T025 Fab	Core protein	0	2021-04-07
7c88	JS003	Programmed cell death 1 ligand 1	0	2021-04-14
7dr4	anti-human IL-2 antibody, mouse Ig G,	Interleukin-2	0	2021-04-14
6ztr	CQY684 Fab	Cadherin-3	0	2021-05-05
7m7w	Monoclonal antibody S2X259 Fab	Spike protein S1	0	2021-05-05
7mjh	VH ab8	Spike glycoprotein	1	2021-05-12
7cho	antibody P5A-1D2	Spike protein S1	0	2021-05-19
7mfu	Synthetic Nanobody #14 (Sb14)	Spike protein S1	1	2021-06-02
7djz	MW01	Spike protein S1	0	2021-06-09
7mhy	3H02 Fab	Protein-cysteine N-palmitoyltransferase HHAT	0	2021-06-16
7o9s	Fab nnHTN-Gn2	Envelope polyprotein	0	2021-06-23
6ze1	G11a nanobody	Cystic fibrosis transmembrane conductance regulator	1	2021-06-30
7apj	NB41	RAC-alpha serine/threonine-protein kinase	1	2021-08-25
7kyo	F100S PsaBC-Fab	Manganese ABC transporter, ATP-binding protein	0	2021-08-25
7o06	Camelid nanobody 10Z	Centrosomal protein of 164 kDa	1	2021-09-08

7e5o	NT-193	Spike protein S1	0	2021-09-08
7rah	M2B10 Fab	Bifunctional adenylate cyclase toxin/hemolysin CyaA	0	2021-09-15
7kn3	S-B8 Fab	Spike protein S1	0	2021-09-22
7mzm	PDI 215	Spike protein S1	0	2021-10-06
7mzj	PDI 93	Spike protein S1	0	2021-10-06
7anq	VHH P1.40 minibody anti-Cter PCSK9	Proprotein convertase subtilisin/kexin type 9	1	2021-10-20
7daa	anti-basigin Fab	Isoform 2 of Basigin	0	2021-10-20
7e72	Fab 3H7	Angiopoietin-1 receptor	0	2021-11-10
7bnv	ION-300 Fab	Surface glycoprotein	0	2021-11-17
7seg	anti-CD16A Fab	Low affinity immunoglobulin gamma Fc region receptor III-A	0	2021-11-24
7dk2	MW07	Spike protein S1	0	2021-12-08
7ps2_1	Beta-53	Spike protein S1	0	2021-12-15
7ps2_2	Beta-29 Fab	Spike protein S1	0	2021-12-15
7ps6	Beta-54 Fab	Spike protein S1	0	2021-12-15
7ps0	Beta-24	Spike protein S1	0	2021-12-15
7lzp_1	JPU-G11	Botulinum neurotoxin A	1	2021-12-22
7q0i	Beta-43	Spike protein S1	0	2021-12-22
7na9	JSG-C1	Botulinum neurotoxin type B	1	2021-12-22
7q0g	Beta-49 Fab	Spike protein S1	0	2021-12-22
7lzp_2	JPU-B9	Botulinum neurotoxin A	1	2021-12-22
7bbj	mAb19	5'-nucleotidase	0	2021-12-29
7t5f	JLJ-G3	Botulinum neurotoxin type B	1	2021-12-29

[a]Protein Data Bank (PDB) code of the antibody-antigen complex structure. For complexes with two distinct antibody-antigen interfaces in the same PDB structure (two antibodies in complex with the same antigen), the names are differentiated with "_1" or "_2" after the PDB code.
[b]Antibody is a nanobody (1) or heavy-light chain antibody (0).
[c]Release date of the structure in the PDB.

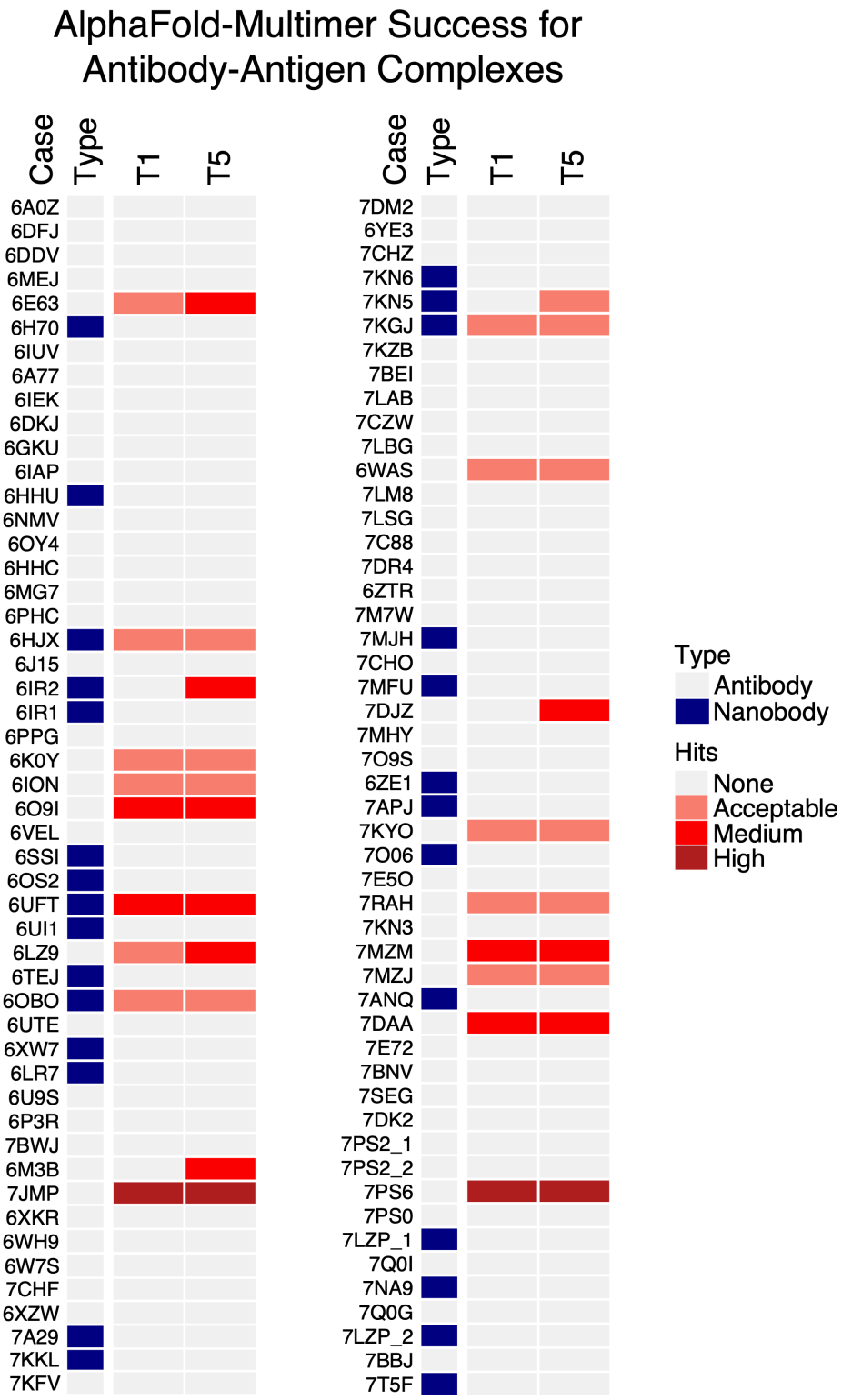

Figure 4.16. AlphaFold-Multimer modeling success for an expanded set of recently released antibody-antigen complex structures. A benchmark set of 100 recently released antibody-antigen complex structures was modeled with AlphaFold-Multimer, and all five models from AlphaFold-

Multimer per test case were assessed for accuracy using CAPRI criteria. AlphaFold-Multimer was run with MSA input and with the use of structural templates released before 4/30/2018, and models were ranked by pTM score. Success for top 1 and top 5 (T1, T5) ranked predictions is shown, colored by CAPRI model accuracy as indicated in the key on right (Hits). "Type" distinguishes complexes containing heavy-light chain antibodies ("Antibody") and single-chain nanobodies ("Nanobody").

Having determined the predictive performance of AlphaFold-Multimer for antibody-antigen complexes, we tested that algorithm for its capability to predict T cell receptor-peptide-MHC (TCR-pMHC) complexes, to further delineate its modeling accuracy for adaptive immune recognition. Although most TCRs share a general binding site and orientation over the pMHC [256], their diversity of pMHC recognition modes, mediated by flexible and variable complementarity determining region loops, pose a challenge for predictive modeling methods, of which several have been developed based on docking [148] and template-based assembly [149, 150]. We assembled a set of 14 Class I TCR-pMHC complexes with known structures that were released in May 2018 or later, and modeling of those complexes with AlphaFold-Multimer showed a success rate of 2 out of 14 complexes (14%) with near-native (Medium or High CAPRI accuracy) ranked at #1 or within the five models for each case (**Table 4.11**). This highlights another class of complexes that is challenging for the current implementation of AlphaFold-Multimer, likely in part due to the limited co-evolution signal in the interface. While there is evidence that TCR genes have co-evolved with MHC genes to promote TCR-pMHC interactions [257], the critical peptide-MHC and TCR-peptide interfaces in TCR-pMHC complexes are not guided by co-evolution, and the accurate modeling of the bound peptide as well as the correctly docked TCR presents a clear challenge in a fold-and-dock scenario.

Table 4.11. AlphaFold-Multimer modeling performance for T cell receptor-peptide-MHC complexes.

PDB	Release date[a]	TCR name	T1[b]	T5[b]
6l9l	11/18/2020	1D4	Incorrect	Acceptable
6mtm	2/13/2019	EM2	**Medium**	**Medium**
6r2l	2/26/2020	NYBR1	Incorrect	Incorrect
6rpa	1/15/2020	NYE_S2	Incorrect	Incorrect
6rpb	1/15/2020	NYE_S1	Incorrect	Incorrect
6rsy	4/15/2020	a7b2	Incorrect	Incorrect
6tro	6/24/2020	GVY01	Acceptable	Acceptable
6uk4	8/19/2020	302 TIL	Acceptable	Acceptable
6uln	5/27/2020	9d	Acceptable	Acceptable
6vrm	6/17/2020	12-6	Incorrect	Incorrect
6vrn	6/17/2020	38-10	Incorrect	Incorrect
7n1e	7/28/2021	pRLQ3	Incorrect	Acceptable
7n1f	7/28/2021	pYLQ7	**Medium**	**Medium**
7rm4	2/9/2022	6-11	Incorrect	Incorrect

[a]Release date of the structure in the PDB.
[b]AlphaFold-Multimer modeling accuracy in top 1 (T1) and top 5 (T5) models (ranked by AlphaFold-Multimer score). Model quality was assessed by CAPRI criteria. AlphaFold-Multimer was run with MSA input and with the use of structural templates for individual component chains, excluding structures released after 4/30/2018 as chain templates. Medium CAPRI accuracy levels are highlighted with bold font.

4.4 Discussion

Our extensive testing of AlphaFold performance on a nonredundant benchmark of protein-protein complexes indicates that AlphaFold is largely successful at predicting binary transient

protein complex structures. However, some complexes were not successfully modeled, most notably antibody-antigen and other adaptive immune interactions, while other categories of test cases showed upper limits of success as well. The limited number of antibody-antigen complex structures that were successfully modeled show that antibody complex modeling can be performed in some cases with the AlphaFold framework, while many of the incorrect models seem to be readily identifiable based on AlphaFold metrics of predicted interface residues, or previously developed energy-based scoring functions.

While protein complex and interface size showed some associations with AlphaFold success for complexes in general, we found that the lack of useful co-evolution signals in the MSAs for antibody-antigen complexes was likely responsible for the limited success of those cases, as shown by the lack of effect on performance when removing the MSA input signal. We tested this using a recently optimized version of AlphaFold for protein-protein interfaces, named AlphaFold-Multimer [248], and the antibody-antigen modeling success with that version of AlphaFold was reflective of our results with the previous AlphaFold release (AlphaFold2) [111], and in accordance with the observation that AlphaFold-Multimer is "generally not able to predict" antibody-antigen complex structures noted by the AlphaFold-Multimer authors [248]. While some antibody-antigen interfaces have been reported to undergo co-evolution in vivo, in particular with evolving viral antigens [258, 259], it is unlikely that corresponding sets of sequences are available for many antibody-antigen pairs in the AlphaFold and ColabFold sequence databases, or in general. However, as noted above, the geometric and structural elements of the AlphaFold and AlphaFold-Multimer framework appear sufficient to construct some antibody-antigen complex structures with high accuracy, and through further training and optimization, success can potentially be improved for such complexes.

Although protein-protein interfaces were not used for training of AlphaFold v.2.0, it is possible that individual chain structures from Benchmark 5.5 complex structures were part of the AlphaFold training set, which could influence the predicted conformations of the subunits and indirectly influence the complex structure models. Benchmarking of this AlphaFold model with recently released complex structures that have no related complexes released prior to the AlphaFold training date addresses this concern, and results with such a set were reported by Evans et al. in the AlphaFold-Multimer study [248], as well as this study. As most complexes in our Benchmark 5.5 set are classified as Rigid-body (64%), with minimal conformational change between unbound and bound structure, the knowledge and possible use of the bound conformation, if used for training of the AlphaFold model, may have little effect or bias versus use of the unbound or accurately modeled unbound structure for that large subset of cases. Benchmarking of traditional docking methods with unbound-bound cases (with one input protein taken from the bound complex structure rather than an unbound structure) could better reflect the use of knowledge of at least one bound component.

While this manuscript was under review, a separate study [237] was published that also reported benchmarking of AlphaFold with a set of protein-protein complexes. While the authors used a distinct test set of 216 protein complexes from the Dockground protein docking benchmark [88], and employed a different MSA-generation method, they reported a 63% success rate for Acceptable or higher model quality, which is similar to our observed 51% success for models of that quality from AlphaFold. Furthermore, as in our study, the authors found that larger interface sizes were associated with improved AlphaFold success, and that interface residue-based pLDDT scores were useful in model selection. However, Bryant et al. noted that size of paired MSAs resulted in improved AlphaFold success, as well as a possible greater dependence of protein source

on AlphaFold success; those differences with our observations may be in part due to their use of their own optimized paired MSAs as AlphaFold input [237], while we obtained paired MSAs from ColabFold [260].

AlphaFold's end-to-end modeling approach represents a major advance and performance improvement over traditional protein-protein docking methods, serving as a proof of concept and a possible framework for optimization to accurately model most or all protein-protein interactions. While optimization of AlphaFold for protein complexes (AlphaFold-Multimer) was recently reported and released by the DeepMind team [248] and was tested in this study, another team showed that combination of AlphaFold with a previously developed protein docking method was able to achieve an improvement in docking success [254], and others have shown the effects of optimized MSAs in AlphaFold complex modeling performance [237]. Notably, a recent study used a combination of AlphaFold and RoseTTAFold to model structures of a large set of eukaryotic protein complexes [261]. Prospective developments that build upon and optimize the AlphaFold framework, or utilize other geometric deep learning methods, can bring the field closer to solving the longstanding challenge of predictive protein docking.

Chapter 5: Evaluation of AlphaFold Antibody-Antigen Modeling with Implications for Improving Predictive Accuracy

5.1 Abstract

High resolution antibody-antigen structures provide critical insights into immune recognition and can inform therapeutic design. The challenges of experimental structural determination and the diversity of the immune repertoire underscore the necessity of accurate computational tools for modeling antibody-antigen complexes. Initial benchmarking showed that despite overall success in modeling protein-protein complexes, AlphaFold and AlphaFold-Multimer have limited success in modeling antibody-antigen interactions. In this study, we performed a thorough analysis of AlphaFold's antibody-antigen modeling performance on 427 nonredundant antibody-antigen complex structures, identifying useful confidence metrics for predicting model quality, and features of complexes associated with improved modeling success. Notably, we found that the latest version of AlphaFold improves near-native modeling success to over 30%, versus approximately 20% for a previous version, while additional AlphaFold sampling gives approximately 50% success. With this improved success, AlphaFold can generate accurate antibody-antigen models in many cases, while additional training or other optimization may further improve performance.

5.2 Introduction

Antibodies are a key component of the immune system, defending the host from viruses and other pathogens through specific recognition of protein and non-protein antigens. Typically, antibodies engage their antigenic targets using the hypervariable complementarity determining region (CDR) loops within the variable domain [262], which are stabilized by the β-sandwich structure of the framework region [39]. Despite sharing a conserved immunoglobulin structure, antibodies collectively exhibit a remarkable ability to recognize and bind to a wide array of antigens with high specificity. The highly specific and diverse nature of antibody-antigen interactions makes antibodies highly useful as therapeutics as well as a consideration in vaccine development efforts [263-266].

High resolution structures of antibody-antigen complexes have refined our knowledge of immunity [267], revealed molecular basis of antibody recognition of viral epitopes [26, 27, 35], and guided the effective design of antibodies [268, 269] and immunogens [36]. However, due to the challenges of experimental structure determination, resource and time constraints, as well as the highly diverse nature of the immune repertoire [270, 271], experimental characterization of most antibody-antigen complex structures is impractical. Therefore, computational tools have been developed and applied to bridge this gap. General protein-protein docking methods have been applied to model antibody-antigen complex structures with limited success [87], due in part to the need to account for the mobility of key CDR loops, as well as the size of certain antigens. To address this, algorithms have been developed specifically for antibody-antigen complex modeling [79, 145-147]. However, accurate structural prediction of antibody-antigen complexes remains a challenge [71, 87].

Recently, the scientific community saw a major breakthrough with AlphaFold (v.2.0), which uses an end-to-end deep neural network to predict protein structures from sequence [111]. AlphaFold iteratively infers and refines pairwise residue-residue evolutionary and geometric information from multiple sequence alignments (MSA) and has achieved unprecedented success in protein structure prediction [111, 112]. Its capabilities were expanded by the development of AlphaFold-Multimer [118] (released in AlphaFold v.2.1), an updated implementation of AlphaFold that was designed to predict protein-protein complex structures. The overall architecture of AlphaFold-Multimer is similar to the previous version of AlphaFold, with changes including cross-chain MSA pairing, adjusted loss functions, and training on protein-protein interface residues.

Previously our benchmarking revealed that, while generally successful in protein-protein complex structure prediction, AlphaFold was less successful in modeling antibody-antigen complexes, and adaptive immune recognition in general [272]. This lack of success in antibody-antigen structure prediction was also noted by the developers of AlphaFold-Multimer [118]. However, some highly accurate antibody-antigen complex models were generated by AlphaFold [272], which shows potential for success of the "fold-and-dock" approach for antibody-antigen structure prediction. While recent studies have assessed the predictive performance of AlphaFold for modeling unbound antibodies [138, 273], or optimization of AlphaFold's ability to predict protein complexes in general [237, 274], studies have not focused on AlphaFold's performance in antibody-antigen recognition, particularly in light of updated versions of AlphaFold and its multimer model (v2.2, v2.3) [275], since our initial test of AlphaFold v2.1 on a set of 100 antibody-antigen complexes [272]. Thus, there is a need for an updated and expanded

benchmarking and analysis of AlphaFold performance on this challenging and important class of complexes.

Here we report a comprehensive benchmarking of AlphaFold for antibody-antigen complex structure modeling. With a dataset of over 400 high resolution and non-redundant antibody-antigen complexes, representing a major increase over the 100 complexes that were used previously [272], we investigated factors contributing to modeling successes and failures, including antibody class and subunit accuracy. The default AlphaFold model confidence score was found to be well correlated with antibody-antigen model accuracy, while residue-level confidence for interface residues was likewise correlated with model accuracy. Interestingly, we found that recent optimization of AlphaFold led to notably higher antibody-antigen accuracy, while use of a "massive sampling" strategy with large sets of pooled AlphaFold models for each complex [274] led to even better performance. Our study presents a thorough analysis of AlphaFold's ability to predict antibody-antigen complexes, yielding valuable insights for interpreting model accuracy, identifying obstacles in the modeling process, and highlighting potential areas for improvement.

5.3 Methods

5.3.1 Antibody-antigen benchmark assembly

We assembled two nonredundant sets of high resolution structures to benchmark AlphaFold, following the general protocol that we described previously [272]. To obtain an initial list of antibody-antigen complexes from the PDB, we downloaded the full SAbDab antibody structure dataset in January 2022. The antibody-antigen complex dataset for AlphaFold

v2.2 benchmarking was assembled using the following criteria: 1) structure resolution <= 3.0 Å, 2) protein antigen in the structure (based on SAbDab annotation), and 3) nonredundant with antibody-antigen complexes with structural resolution <= 9.0 Å released before April 30, 2018 (AlphaFold v2.2 training sample cutoff date) based on sequence criteria. Sequence criteria for nonredundancy are: 1) heavy chain variable domain sequence ID < 90% and full variable domain sequence ID < 90%, or 2) no match between antigen chain sequences (no hit detected using BLAST [234] with default parameters). Pairwise sequence alignments were performed using the "blastp" executable in the BLAST suite [234]. Structural nonredundancy criteria were then applied to the set. We removed antibody-antigen structures with < 5 Å heavy chain Cα atom RMSD, after superposition of antigens using the FAST structure alignment program [235], and > 70 % identity between heavy chain variable domain, light chain variable domain, or concatenated CDR loop sequences. To avoid modeling antigen chains with large regions that are not resolved in the experimentally determined structures, we additionally removed structures with PDB "seqres" file sequence annotation and resolved region sequence length difference > 70%, or sequence length difference > 35% and resolved antigen length > 500 aa. We also removed non-canonical antibody-antigen complex cases (e.g. with antibody-tetramerization, dimeric sdAb, or constant domain binding), and we removed cases with incomplete antigen chain annotations by SAbDab, identified through manual inspection of the PDB bioassembly structure.

To benchmark AlphaFold v.2.3, we identified a subset of 41 antibody-antigen complexes within the v.2.2 benchmarking set. These antibody-antigen complexes were released after September 30, 2021, and are not redundant with structures released before that date based on the

sequence criteria detailed above. AlphaFold v.2.2 and v.2.3 generated models for 39 out of 41 of those complexes without runtime errors.

The AlphaFold v2.2 and v2.3 benchmarking cases are shown in **Table 5.1**.

5.3.2 AlphaFold antibody-antigen modeling

Sequences input to AlphaFold were obtained from the PDB "seqres" file. Antibody sequences were processed by ANARCI to remove non-variable domain sequence regions. We downloaded and installed AlphaFold v2.2 from Github (https://github.com/deepmind/alphafold) in May 2022 and v.2.3 in February 2023. Both versions of AlphaFold were installed on a local computing cluster. During the structure prediction or feature preparation step in the AlphaFold pipeline, 15 cases failed to complete because of GPU and memory limitations out of a total of 444 test cases.

For generating unbound antibody and antigen structures, we employed AlphaFold in Multimer setting when the input consisted of a heavy-light chain antibody or a multimeric antigen. Alternatively, the Monomer setting was utilized when the input was a single chain. A template date cutoff of April 30, 2018 was applied to avoid template overlap with benchmarking set.

To generate AlphaFold predictions without the use of MSAs (corresponding to single-sequence modeling), we modified "all_seq_msa_features" variable of chain features, to include only the query sequence. To use custom templates, we adapted the template featurization function from Motmaen et al. [276] (https://github.com/phbradley/alphafold_finetune/blob/main/predict_utils.py).

AlphaFold modeling in ColabFold [226] was performed with ColabFold version 1.3.0 (commit 26de12d3afb5f85d49d0c7db1b9371f034388395), installed on a local computing cluster using scripts from Github (https://github.com/YoshitakaMo/localcolabfold). During ColabFold AlphaFold modeling, MSA was built by querying the MMseqs2 MSA server using unpaired and paired MSA. To generate a total of 25 predictions per complex, modifications were made to "load_models_and_params" function, utilizing a different random seed for each prediction, producing five predictions per AlphaFold model parameter.

Unless otherwise specified, a template date cutoff of April 30, 2018 was applied for benchmarking AlphaFold v.2.2 and ColabFold, and a template date cutoff of September 30, 2021 was applied for benchmarking of AlphaFold v.2.3, to avoid using bound structures as template.

AlphaFold and ColabFold modeling runs were performed using NVIDIA Titan RTX and Quadro 6000 GPUs.

5.3.3 Complex model accuracy assessment

We assessed antibody-antigen complex model accuracy using DockQ [277], which was downloaded from GitHub (https://github.com/bjornwallner/DockQ). Antibody-antigen complex model accuracy was computed by DockQ using the experimentally determined antibody-antigen complex structures obtained from the PDB. DockQ calculates interface backbone RMSD (I-RMSD), ligand backbone RMSD (L-RMSD), fraction of native contacts (fnat), DockQ score, as well as the CAPRI (Critical Assessment of PRediction of Interactions) accuracy level, which assigns the model into one of four discrete accuracy classes: incorrect, acceptable, medium, and high, based on the model's similarity to the native structure [223].

5.3.4 Interface pLDDT calculation

To determine the interface pLDDT (I-pLDDT), we computed the average pLDDT value for all residues at the antibody-antigen interface. Interface residues were defined as any residue with a non-hydrogen atom within 4.0 Å of the binding partner. An I-pLDDT score of 30 was assigned to predictions with no antibody-antigen interface residues.

5.3.5 CDR loop accuracy analysis

The complementarity determining regions (CDRs) and the framework regions of antibodies were identified by AHo numbering [168], assigned using ANARCI software [278]. The CDR loops were defined as residues 24-42 (CDR1), 57-76 (CDR2), and 107-138 (CDR3), as in previous work [41].

ProFit v 3.1 [279] was used to calculate backbone RMSDs between modeled and experimentally determined CDR loop structures, after superposing the modeled antibody structures onto the experimentally resolved structures by the framework residues.

5.3.6 Figures and statistical analysis

PyMOL version 2.4 (Schrodinger, Inc.) was used to generate structural figures. The ggplot2 [186] package in R (r-project.org) was utilized to generate box plots, line plots, and bar plots. Pearson correlations and their corresponding p values were calculated using the ggpubr package in R, while the Wilcoxon rank-sum test was performed using the ggsignif package in R. Binary and multi-class ROC AUC values were calculated using the pROC [250] and multiROC [251] packages in R, respectively.

5.3.7 Antibody-antigen complex scoring and native complex relaxation

The "InterfaceAnalyzer" executable in Rosetta [165] (v.3.12) was employed to calculate interface energetic scores, using the Rosetta Energy Function 2015 (REF15) scoring function [239] and default parameters. Prior to scoring, structural relaxation was performed on native antibody-antigen complex obtained from PDB using the FastRelax protocol [280] ("relax" executable) in Rosetta using the following flags:

-ignore_unrecognized_res

-relax:constrain_relax_to_start_coords

-relax:coord_constrain_sidechains

-relax:ramp_constraints false

-ex1

-ex2

-use_input_sc

-no_optH false

-flip_HNQ

-overwrite

-nstruct 1

5.3.8 Rigid-body docking with ZDOCK and IRAD

To establish a rigid-body docking baseline, ZDOCK version 3.0.2 [230] was used to generate antibody-antigen docking models using unbound or bound input structures. Unbound antibody and antigen input structures for the ZDOCK algorithm were generated by AlphaFold,

using the above-mentioned protocol. Bound antibody and antigen input structures were extracted from experimentally resolved antibody-antigen complex structures downloaded from PDB, with HETATMs removed. Through dense rotational sampling ("-D" flag in ZDOCK), a total of 54,000 predictions per complex were generated. Subsequently, the docking poses were scored and reranked using the IRAD scoring function [281]. Predictions were assessed with the DockQ program [282] to determine the CAPRI model accuracy based on comparison with the known complex structure.

5.3.9 Rigid-body docking with ClusPro

To establish a baseline for a rigid-body docking algorithm specific to antibody-antigen complexes, the ClusPro web server (https://cluspro.bu.edu/) was utilized in its antibody mode [146]. Unbound antibody and antigen input structures (same input for unbound ZDOCK docking) were generated by AlphaFold as described above. Within the ClusPro interface, the options 'Use Antibody Mode' and 'Automatically Mask non-CDR regions' were selected, in accordance with the server's recommendations. Docking poses were ranked using ClusPro's default method, based on cluster size from large to small. All docking poses were assessed with the DockQ program [282] to assess CAPRI model accuracy based on comparison with the known complex structure.

5.3.10 TM-score calculation

To provide an assessment of antigen prediction accuracy, we employed the TM-score executable [231] to calculate the TM-score, comparing the structural similarity between the antigen chain(s) in the experimentally resolved antibody-antigen complex structure, and the antigen prediction in the antibody-antigen complex predictions. Prior to running TM-score on the

antigen chains, residues that were not experimentally resolved, or absent from the experimentally resolved structure, were removed from the antigen prediction.

5.3.11 MSA depth calculation

The MSA of the antibody-antigen complex was retrieved from the 'msa' key value in the feature dictionary (“feature_dict” variable). Given that the MSA values of the ‘msa’ key are encoded in AlphaFold residue IDs, we converted the amino acids back to the one-letter amino acid type using “ID_TO_HHBLITS_AA” dictionary, and replaced gaps (denoted by “-“) by “U” for downstream MSA depth calculation. Number of effective sequences (N_{eff}) was calculated by CD-Hit [240]. For consistency with the AlphaFold MSA N_{eff} calculation scheme [111], we used an identity cutoff of 80% in CD-HIT to calculate nonredundant sequence clusters. MSA depth was successfully calculated for 426 out of 427 antibody-antigen complexes, with one failure due to a technical issue.

5.3.12 Hetero-atoms at the interface

Glycans and ligands at the antibody-antigen interface were identified through inspection of HETATM records in experimentally resolved structure coordinates. Asymmetric unit structures (defined by the PDB entry) of the antibody-antigen complexes were inspected, and cases with hetero-atoms matching the "saccharide" classification identified from PDB ligands summary pages (from wwPDB's Chemical Component Dictionary), as well as hetatoms in the category of lipids and nucleotides within 6 Å of the antibody and the antigen were identified as positive hits.

Glycan clusters were also accounted for, in order to appropriately group individual glycan residues into a single N-glycan during analysis. Glycan clusters are identified by checking

the distance between the C and the O atoms of glycan HETATM residues “MAN”, “BMA”, “NAG”, and “FUC”. Glycans were considered to be covalently linked if their C atom to another glycan residue’s O atom has a distance < 2.0 Å. One cluster of glycans is formed by pooling together all glycan residues that are covalently attached. One cluster of glycans were considered in the interface between antibody and antigen, if the cluster has glycan residues < 6 Å to non-hydrogen atoms of antibody chains, and has glycan residues < 6 Å to non-hydrogen atoms of antigen chains.

Since not all antigen glycosylation is experimentally resolved, we predict the possible presence of antigen glycosylation near antibody binding site based on the source species of the antigen and the presence of surface-accessible glycosylation motifs. We first determined the source species of each antigen in our dataset to determine if the antigen can be glycosylated. This assessment was based on whether proteins from these species are known to undergo glycosylation either intrinsically or through the hijacking of host machinery, as observed in enveloped viruses that infect eukaryotic hosts. In our set, such species include eukaryotic species (excluding Plasmodia, Butyriboletus subregius, Atractiella rhizophila, Aequorea victoria, Betula pendula due to uncertain, less predictable, or lack of N-glycosylation for those organisms) and enveloped viruses. For antigens identified as possible to be N-glycosylated based on source organism, we next examined their sequences for antibody-proximal glycosylation motifs. Specifically, we focused on N-glycosylation sequons, defined as the sequence motif N-X-S/T (where X is any amino acid except proline). We then employed Naccess v2.1.1 [170] to evaluate the surface accessibility of the asparagine residue in each sequon. We consider those with a relative accessibility of the side chain above 15% as surface-exposed asparagine residues. Finally, we measured the proximity of these candidate asparagine residues to the antibody in the

antibody-antigen complex structure, classifying those within 12 Å as potential sites of antigen glycosylation near the antibody binding site.

5.3.13 AFsample antibody-antigen modeling

AFsample was downloaded from GitHub (https://github.com/bjornwallner/alphafoldv2.2.0). The modeling protocol described in Wallner, 2023 [274] was followed. A total of 6,000 predictions were generated per case, including 2,000 predictions generated using AlphaFold v.2.1 and v.2.2 models with templates and full dropout, 2,000 predictions generated using v.2.1 and v.2.2 models without templates and with dropout in Evoformer but not structure module, 1,000 predictions generated with v.2.1 models without templates, with maximum number of 21 recycles, with dropout in the Evoformer but not the structure module, and 1,000 predictions generated with v.2.2 models without templates, with maximum number of 9 recycles, with dropout in the Evoformer but not the structure module. When templates were used, a template date cutoff of September 30, 2021 was applied. The top 5 predictions per case were relaxed. Modeling runs were performed using NVIDIA Titan RTX, Quadro 6000 and RTX A100 GPUs.

5.3.14 Identification of experimentally resolved antigens bound to other antibodies

Antigen chains bound to distinct antibodies were identified through a systematic approach. Each antigen sequence from the complexes was queried against the PDB SEQRES database using BLAST [234] to find PDB chains with at least 95% sequence identity, covering a minimum of 90% of the query antigen sequence and structural resolution <= 3.0 Å. Subsequently, the SAbDab [233] database was utilized to determine if the identified PDB chains were complexed with antibodies. The hits were further examined for antibody distinctness from

the test case. Antibodies complexed with the antigen hit were considered distinct if the heavy chain V domain sequence identity with the query antibody heavy chain was < 90%, or the full antibody V domain sequence identity was < 90%. We additionally removed antigen hits where the experimentally resolved sequence covered < 70% of the query antigen sequence. Finally, all qualifying hits were ranked by their resolution, from lowest to highest, and the top hit was selected as the template for modeling.

5.3 Results

5.3.1 AlphaFold antibody-antigen complex modeling accuracy

To perform a comprehensive and detailed assessment of AlphaFold's ability to model antibody-antigen complexes, we assembled a set of over 400 nonredundant antibody-antigen complexes released after April 30, 2018 (**Table 5.1**). The date cutoff was selected to avoid overlap with the training set of the tested version of AlphaFold (v2.2.0, hereafter denoted as v2.2 for brevity). Nonredundancy and additional test case selection criteria are described in the Methods section. For efficiency, we only utilized the variable domains of the antibody sequences for modeling. The accuracy of antibody-antigen complex predictions was evaluated using Critical Assessment of Predicted Interactions (CAPRI) criteria [223], which classify predictions as incorrect, acceptable, medium, or high based on a combination of interface root-mean-square distance (I-RMSD) and ligand root-mean-square distance (L-RMSD) from the experimentally determined complex structure, as well as the fraction of experimentally observed interface residue contacts in the model.

Table 5.1. Antibody-antigen structures used for AlphaFold benchmarking.

PDB	Heavy chain	Light chain	Antigen chain(s)	Release Date	v.2.3.0 set[1]	100 subset[2]
6was	H	L	J	3/31/21		X
6p50	H	L	C	9/4/19		
6urm	H	L	F	9/16/20		
6umg	h	l	cr	2/12/20		
6a0z	H	L	A	6/20/18		X
7daa	H	L	A	10/20/21	X	X
6s8j	P	O	C	2/12/20		
6s8i	P	O	C	2/12/20		
7lf7	A	B	M	8/4/21		
6urh	H	L	C	3/18/20		
7lfb	H	L	X	8/4/21		
6ktr	A	B	C	7/8/20		
6meh	H	L	C	11/21/18		
6vel	H	L	C	1/29/20		X
6y9b	I	M	A	5/20/20		
6xkq	H	L	A	10/14/20		
6o9h	H	L	D	1/22/20		
6yio	H	L	B	11/11/20		
7l7r	B	A	G	12/1/21	X	
7l7r	D	C	G	12/1/21	X	X
6svl	A	B	C	11/27/19		
6okm	H	L	R	8/28/19		
6z3p	H	L	CAB	9/2/20		
6u36	H	L	B	11/6/19		
6jbt	H	L	F	6/19/19		
6wbv	H	L	A	9/9/20		
7jmp	H	L	A	8/26/20		X
6wo5	G	I	F	8/19/20		X
7lfa	B	D	A	8/4/21		
7np1	H	L	A	11/17/21		
6hig	H	L	B	6/5/19		
6vyh	D	C	A	11/11/20		
6k65	H	L	A	8/14/19		
6umx	h	l	B	2/26/20		
7nx3	B	C	F	10/27/21	X	
7jv6	C	D	B	10/14/20		

6xqw	H	L	E	3/3/21	
7jtg	A	B	E	3/10/21	X
6o9i	D	E	C	1/22/20	X
7e7x	H	L	A	6/9/21	
6wo5	A	B	E	8/19/20	
6mvl	H	L	A	10/23/19	
7kqg	B	C	A	12/16/20	
6glw	H	L	A	6/5/19	
6gku	H	L	A	6/5/19	
7lue	H	L	A	6/16/21	
7k9j	J	N	C	9/29/21	
7n8h	C	B	A	7/14/21	
6wo3	H	L	E	8/19/20	
6wmw	M	N	B	7/15/20	
6wmw	H	L	B	7/15/20	
7czx	I	M	B	3/10/21	
7czw	H	L	A	3/10/21	
7rah	B	A	E	9/15/21	
7rah	D	C	E	9/15/21	
7kn4	M	N	B	9/22/21	
6nyq	H	L	C	1/22/20	
6xkr	H	L	P	9/9/20	X
6jjp	D	E	F	10/30/19	
7lxy	N	O	J	4/14/21	
7lr4	H	L	D	12/15/21	
7r89	C	D	BA	9/8/21	
6ywc	D	E	F	10/7/20	
6xxv	D	E	F	4/22/20	
7ly2	N	O	J	4/14/21	
6vvu	G	I	D	12/30/20	
6k0y	A	B	C	12/11/19	X
6v4o	H	L	N	10/7/20	
6wgl	A	B	C	9/16/20	
6xsw	D	E	F	7/21/21	
6hx4	H	L	B	10/30/19	
6wzk	A	B	E	11/25/20	
7lxz	H	J	A	4/14/21	
7ceb	C	D	B	6/23/21	
6lxi	C	D	B	12/2/20	
6osv	H	L	K	4/1/20	

7lbg	H	G	A	3/10/21		
7lbg	F	E	A	3/10/21		X
6a3w	A	B	C	10/10/18		
7bnv	H	L	A	11/17/21	X	X
7lm9	H	L	A	3/31/21		
7o52	H	L	U	7/28/21		
7s4s	H	L	A	9/22/21		
7lop	X	Y	Z	3/3/21		
6xkp	M	N	B	10/14/20		
6h2y	H	L	D	8/14/19		
6phc	C	D	E	10/2/19		X
6osh	H	L	K	4/1/20		
7n3d	H	L	C	7/7/21		
7e7y	A	B	R	6/9/21		
7chf	A	B	R	9/16/20		X
7lm8	H	L	A	3/31/21		X
7kn3	M	N	B	9/22/21		X
7n0u	H	L	C	8/11/21		
7joo	H	L	C	10/14/20		
7dk2	D	E	F	12/8/21		X
6sni	H	L	X	3/11/20		
7c88	A	B	C	4/14/21		X
7dc8	B	A	C	1/13/21		
7mmo	D	E	F	5/12/21		
7kfv	F	G	E	12/2/20		X
7k9z	B	A	E	10/28/20		
6mlk	H	L	A	10/17/18		
6mg7	H	L	G	9/25/19		X
7l0l	H	L	BA	11/3/21	X	
6ogx	C	D	G	7/10/19		
7lcv	A	B	C	6/9/21		
6pi7	F	E	D	7/24/19		
7ps6	H	L	E	12/15/21	X	X
6rlo	G	H	L	5/12/21		
7lab	Y	X	B	3/10/21		
7cm4	H	L	A	1/20/21		
6gv4	H	L	BA	11/21/18		
6j14	A	B	G	11/6/19		
6ohg	C	B	A	6/17/20		
6v4n	D	E	W	10/7/20		

7e5o	H	L	A	9/8/21		X
6p67	A	B	K	9/4/19		
7ean	H	L	A	3/31/21		X
7eam	H	L	A	3/17/21		
6ppg	B	A	G	12/11/19		X
6nmr	J	K	M	8/7/19		
7djz	A	B	C	6/9/21		X
6m3b	C	B	A	7/8/20		X
6o1f	H	L	AI	10/16/19		
7lr3	H	L	D	12/15/21		
6r8x	C	B	A	4/10/19		X
6wxl	F	E	DC	6/9/21		
6a67	H	L	A	8/29/18		
6mej	H	L	C	11/21/18		X
6mej	A	B	C	11/21/18		
6udj	H	I	J	1/29/20		
7cho	B	C	A	5/19/21		X
7bwj	H	L	E	6/3/20		X
7kqb	H	L	A	5/26/21		
6kz0	K	L	J	5/27/20		
7vux	H	L	A	11/17/21	X	
7q0i	H	L	C	12/22/21	X	X
7orb	E	F	X	7/7/21		
6nmu	B	A	C	8/7/19		
7s0b	C	D	E	10/6/21		
6rps	H	L	A	11/13/19		
7l7e	O	P	K	9/1/21		
7orb	H	L	R	7/7/21		
7r6w	A	B	R	7/21/21		
7r6w	H	L	R	7/21/21		
7kyl	H	L	E	4/7/21		
6iut	H	L	A	1/16/19		
6nmt	B	A	C	8/7/19		
6dkj	A	B	D	5/8/19		X
7c61	H	L	A	7/29/20		
7m3n	H	L	A	7/28/21		
7s13	H	L	D	10/20/21		
7m7w	C	D	S	5/5/21		
6nmv	H	L	S	8/7/19		X
6icc	H	L	A	2/13/19		

7m7w	H	L	R	5/5/21		X
6j15	A	B	C	11/6/19		X
6whk	C	B	A	4/14/21		
7kpb	H	L	AC	1/13/21		
6pe8	H	L	T	8/14/19		
7ket	A	B	C	6/9/21		
6ion	H	L	A	1/15/20		X
6nms	B	A	C	8/7/19		
6lz9	H	L	B	3/11/20		X
6p3r	C	D	E	5/27/20		
7coe	B	C	D	8/4/21		
7ps4	H	L	E	12/15/21	X	
7ps1	A	B	E	12/15/21		X
6phb	D	C	I	10/2/19		
6lgw	C	D	F	2/19/20		
7s11	I	M	D	11/3/21	X	
6ieb	E	F	B	4/10/19		
6e63	H	L	P	11/28/18		X
7seg	H	L	C	11/24/21		X
6wh9	E	F	D	9/9/20		
7ps0	H	L	E	12/15/21	X	X
6xlq	B	C	A	9/2/20		
7dm1	D	C	A	12/23/20		
7dm2	H	L	A	12/23/20		X
6j5d	H	L	A	2/6/19		
7o9s	H	L	A	6/23/21		X
7kmg	D	E	F	1/27/21		
6q0e	H	L	A	12/18/19		
7cr5	H	L	A	3/24/21		
7q0g	A	B	E	12/22/21	X	X
6h3t	I	M	B	2/27/19		
6s5a	H	L	DA	9/25/19		
7dha	C	B	A	9/22/21		
6ddm	B	A	C	10/24/18		
7mzm	H	L	A	10/6/21	X	X
6nha	H	L	AB	12/25/19		
6xpx	B	C	A	5/19/21		
6xq0	E	F	D	5/19/21		
6e56	H	J	A	5/22/19		
7ps2	H	L	G	12/15/21	X	X

7ps2	A	B	G	12/15/21		X
6tyb	H	L	G	10/2/19		
6u9s	D	E	F	5/13/20		X
7bq5	H	L	A	3/24/21		
7dr4	A	B	J	4/14/21		X
6m58	C	D	B	4/29/20		
6ss2	H	L	A	6/10/20		
7kzb	H	L	C	2/3/21		X
6vug	D	C	B	2/17/21		
6e4x	Z	Y	B	5/22/19		X
6otc	H	L	A	6/5/19		
6oy4	D	C	A	8/28/19		X
7nx8	H	L	E	4/7/21		X
7bel	C	D	X	3/3/21		
7mzg	H	L	A	10/6/21		
7bek	H	L	E	3/3/21		
7mf1	H	L	A	5/12/21		
7bei	H	L	E	3/3/21		X
7bel	E	F	X	3/3/21		
6iuv	C	D	B	1/16/19		X
6iea	H	L	A	4/10/19		
7e3o	H	L	R	9/15/21		
7ahu	B	A	EF	7/7/21		
6q18	H	L	A	12/18/19		
7kmh	A	B	C	1/27/21		
7mzk	N	M	B	10/6/21	X	
7or9	H	L	E	7/7/21		X
6oz2	H	L	G	8/19/20		
6xzw	H	L	D	10/14/20		X
6iek	B	C	A	4/10/19		X
7rks	I	M	S	9/22/21		
6w7s	H	L	A	9/9/20		X
7mzi	H	L	A	10/6/21		
7neh	H	L	E	3/3/21		
7mzj	H	L	A	10/6/21	X	X
6i8s	E	I	A	2/13/19		
6oor	H	L	A	7/17/19		
6vy6	H	L	A	1/6/21		X
7bbj	H	L	A	12/29/21	X	
6mfp	C	D	A	9/18/19		

6wm9	E	F	D	1/27/21		
7mzh	H	L	E	10/6/21		
7bep	A	B	E	3/3/21		
7chz	H	L	I	1/13/21		X
6qig	H	L	A	9/4/19		
6hhc	H	L	A	9/11/19		X
6woz	K	L	J	1/27/21		
7lsg	H	L	C	4/7/21		X
7jx3	C	D	R	10/14/20		
7jx3	H	L	R	10/14/20		X
6j6y	E	F	D	8/7/19		
7kq7	H	L	B	4/7/21		
6ztr	A	B	J	5/5/21		X
7ce2	Z	B	A	4/7/21		
6wtu	E	F	D	1/27/21		
6ocb	H	L	A	5/29/19		
6wds	H	L	CAB	7/15/20		
7kyo	H	L	B	8/25/21		
6o39	B	A	C	4/3/19		
6ba5	F	E	O	6/13/18		
6n6b	K	L	A	7/3/19		
6ye3	G	H	I	12/30/20		X
7bsc	H	L	A	12/23/20		
6kyz	B	C	A	5/27/20		
7ec5	E	F	BAC	3/31/21		
6mhr	A	B	C	11/21/18		
6pzf	F	E	B	12/4/19		
6pze	H	L	A	12/4/19		
6z2l	C	B	A	7/22/20		
6e3h	H	L	BA	9/26/18		
6cxy	H	L	C	4/10/19		
6nz7	H	L	BA	5/8/19		
6vy4	C	D	B	12/30/20		
7e72	C	D	F	11/10/21	X	X
7n4j	H	L	A	10/6/21	X	
6e62	H	L	P	11/28/18		
6q20	H	L	A	10/23/19		
6vc9	H	L	A	11/11/20		
6lyn	H	L	D	2/24/21		
6ivz	H	L	A	2/13/19		

6id4	C	D	EF	2/6/19	
6hxw	C	D	B	8/28/19	
7d85	E	F	D	4/7/21	
7r8l	H	L	E	8/4/21	
7klh	H	L	A	2/10/21	
7mhy	O	P	A	6/16/21	X
7mhy	M	N	A	6/16/21	
6hga	H	L	B	3/18/20	
6pxh	H	L	B	9/25/19	
7cj2	K	L	B	7/14/21	
6j11	F	G	B	7/24/19	
7jie	E	F	A	6/30/21	
6n5e	G	F	B	6/5/19	
6u6u	H	L	R	4/22/20	
6iap	E	D	A	6/12/19	
6iap	H	L	A	6/12/19	X
7n3c	H	L	C	7/7/21	
7e9b	H	L	C	7/28/21	
7kpg	H	L	S	12/16/20	
6jep	H	L	E	5/15/19	
6dfj	H	L	E	10/24/18	X
6ute	C	D	S	4/15/20	X
6ddr	B	A	C	10/24/18	X
6ddv	B	A	C	10/24/18	X
6a77	H	L	A	1/30/19	
6rvc	D		A	10/2/19	
6gju	C		A	6/26/19	
6gjq	B		A	6/19/19	
6r7t	A		B	5/1/19	
7l1v	S		R	2/10/21	
7my3	E		A	6/16/21	
7kkl	E		C	11/11/20	
6v7y	F		A	9/16/20	
6x04	H		G	12/9/20	X
7rnn	C		D	8/11/21	
7p77	A		B	8/4/21	
6u54	A		B	11/6/19	
7mjh	F		C	5/12/21	
7my2	H		E	6/16/21	
7kn6	C		A	1/20/21	X

7kjh	A	C	2/3/21		
7o06	A	C	9/8/21		X
6zxn	E	B	9/23/20		
7a29	E	B	10/21/20		
6ze1	B	A	6/30/21		X
7d6y	B	A	10/6/21		
6rnk	B	A	8/14/19		
6v7z	F	AB	9/16/20		
7kn5	E	A	1/20/21		
7kn5	C	A	1/20/21		X
6u55	A	B	11/6/19		
6yu8	B	A	2/17/21		
7apj	B	A	8/25/21		
6x05	K	A	12/9/20		
7olz	B	A	8/11/21		
7olz	C	A	8/11/21		
6qup	B	A	8/5/20		X
6qgw	B	A	6/26/19		
6z20	D	C	9/23/20		
6rqm	B	A	7/8/20		
6oyh	E	A	7/10/19		
6os2	D	A	2/19/20		X
6qgx	B	A	6/26/19		
6qgy	B	A	6/26/19		
6z1v	B	A	9/23/20		
7o0s	A	B	9/15/21		
7r98	F	C	7/7/21		
7c8v	A	B	6/24/20		
6ir1	B	A	11/13/19		X
7cz0	E	A	9/8/21		
6x07	B	A	12/9/20		X
7nfq	C	A	12/1/21	X	
7nfr	B	A	12/1/21	X	
6lz2	B	A	12/23/20		
6gs4	H	A	1/30/19		
6gk4	F	D	6/19/19		
6o3c	B	A	7/3/19		
7nx0	D	C	10/27/21	X	
6yz5	F	E	6/3/20		
6z6v	G	BC	6/10/20		

6rtw	B	A	10/2/19		
7k84	B	A	10/14/20		
6uc6	D	B	3/4/20		
7lzp	G	A	12/22/21		X
6oq6	D	A	7/10/19		
6ir2	B	A	11/13/19		X
6gkd	B	A	6/19/19		
7d30	A	B	2/17/21		
7lzp	F	D	12/22/21		X
7lzp	E	D	12/22/21	X	
6mxt	N	A	11/14/18		
6dbg	C	B	7/18/18		
6vbg	D	B	11/25/20		
7azb	B	A	11/25/20		
6gwn	B	A	1/1/20		
6gwn	C	A	1/1/20		
6zg3	E	AI	3/3/21		
7a0v	B	A	12/30/20		
6ssi	J	E	2/12/20		
6gjs	C	A	6/26/19		
6gjs	B	A	6/26/19		
6lr7	B	A	4/29/20		X
7vnb	A	B	11/24/21	X	
7ldj	G	C	5/5/21		X
7anq	B	A	10/20/21	X	X
6h02	B	A	8/29/18		
6hhu	G	A	7/24/19		X
6hhu	H	A	7/24/19		
6uft	B	A	3/4/20		X
7e53	B	A	10/13/21	X	
6oca	C	A	4/1/20		
7m1h	G	A	12/22/21	X	
7m1h	E	A	12/22/21	X	
7m1h	F	A	12/22/21	X	
7kdu	C	BA	8/4/21		
6fyu	C	BA	11/7/18		
6h6y	G	C	12/19/18		
7na9	D	A	12/22/21	X	X
6ui1	D	A	3/4/20		X
7t5f	E	D	12/29/21	X	

7kc9	F	ED	8/4/21		
7lvw	I	D	3/24/21		
7aqy	C	B	11/3/21	X	
6zrv	B	A	8/26/20		
7t5f	C	A	12/29/21	X	X
7kbk	C	AB	8/4/21		
7kd2	C	BA	8/4/21		
6ul6	C	A	3/4/20		
6ul6	B	A	3/4/20		
6i8g	B	A	10/2/19		
6h16	B	A	1/30/19		
7kd0	C	BA	8/4/21		
7aqz	D	A	11/3/21		
7l6v	B	A	12/22/21	X	
7l6v	D	A	12/22/21	X	
7l6v	C	A	12/22/21	X	
7l6v	F	A	12/22/21	X	
6uht	C	A	3/4/20		
7n0r	D	B	6/2/21		
6oq7	C	A	7/10/19		
6rah	C	B	7/31/19		
7ar0	B	A	11/3/21	X	
6xw4	C	A	4/22/20		
6h15	D	B	1/30/19		
6h72	C	A	12/19/18		
7aqx	D	B	11/3/21		
6sge	B	A	9/25/19		
6waq	A	B	4/1/20		
7n0i	L	GH	6/9/21		
7d2z	A	B	2/17/21		
6ocd	B	A	4/1/20		
6tej	C	B	4/1/20		
6oq8	D	A	7/10/19		
6b20	F	A	5/30/18		
6obe	B	A	4/1/20		
6obc	B	A	4/1/20		
6obo	C	A	4/1/20		X
6i6j	C	A	2/27/19		X
7mfu	B	A	6/2/21		X
7mfu	F	D	6/2/21		

7kgj	B	A	2/3/21	X
7kgk	B	A	2/3/21	
7now	C	D	4/7/21	X
7nqa	D	A	7/21/21	
6xzu	A	B	8/12/20	
7czd	A	B	7/14/21	

[1]Subset of cases used for benchmarking AlphaFold v.2.3.
[2]Subset of 100 cases used for benchmarking the use of bound input template structures.

AlphaFold generated acceptable or higher accuracy models as top-ranked predictions for 26% of the 427 test cases for which models were generated (**Figure 5.1a**). medium or higher accuracy models, which we refer to as near-native predictions, were generated as top-ranked predictions for 18% of the cases, and high accuracy models were generated for 5% of the test cases. Success rates increased when all 25 predictions per complex were taken into consideration, leading to 37% of the cases achieving acceptable or higher accuracy predictions, 22% achieving medium or higher accuracy predictions and 6% achieving high accuracy predictions.

Representative models generated by AlphaFold are shown in **Figure 5.1b** (PDB code 6nmv; antibody/SIRP-alpha complex) and **Figure 5.1c** (PDB code 6j15; antibody/PD-1 complex). Both models are top-ranked predictions for the respective complex. The model in **Figure 5.1b** has high CAPRI accuracy, and an interface root-mean squared distance (I-RMSD) value of 0.68 Å, indicating a low level of structural deviation of this modeled antibody-antigen complex from the native complex. **Figure 5.1c** shows an acceptable CAPRI accuracy prediction with an I-RMSD of 3.55 Å. While the antibody engages the correct site of the antigen in this

example, a deviation in positioning of the antibody on the antigen, with respect to the experimentally determined structure, is observed.

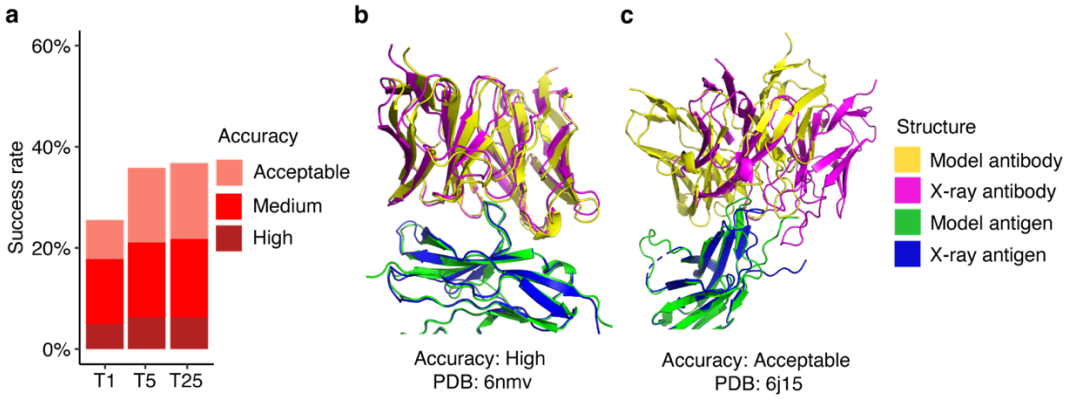

Figure 5.1. Antibody-antigen modeling accuracy of AlphaFold. a) Benchmarking of AlphaFold (v.2.2, AlphaFold-Multimer model) was performed on 427 antibody-antigen complexes. For each complex, 25 predictions were generated, and ranked by AlphaFold model confidence score. Antibody-antigen predictions were evaluated for complex modeling accuracy using CAPRI criteria for high, medium, and acceptable accuracy. The success rate was calculated based on the percentage of cases that had at least one model among their top N ranked predictions that met a specified level of CAPRI accuracy. Bars are colored by CAPRI accuracy level. **b**) Example of a near-native prediction by AlphaFold, in comparison with the experimentally determined structure (PDB 6nmv; antibody/SIRP-alpha complex). This model has high CAPRI accuracy (I-RMSD=0.68 Å) and has the highest model confidence of all 25 predictions of this complex (model confidence=0.88). **c**) An example of an acceptable accuracy complex model from AlphaFold, in comparison with the experimentally determined structure (PDB: 6j15; antibody/PD-1 complex). This model has acceptable CAPRI accuracy (I-RMSD=3.35 Å), and has the highest model confidence of all 25 predictions of this complex (model confidence=0.75). Complex structures in b and c are superposed by antigen with the model and the X-ray structure components colored separately as indicated on right.

We additionally assessed antibody-antigen modeling accuracy of AlphaFold in ColabFold [226]. For fairness of the comparison with the full AlphaFold pipeline's results, ColabFold was modified to generate 25 predictions per complex. ColabFold's modeling success was similar to that of AlphaFold for 426 cases for which both algorithms were able to generate models, with slightly lower success observed for ColabFold (**Figure 5.2)**. The difference in success may be due to factors such as different MSAs or structural templates, as ColabFold and

AlphaFold employ distinct approaches for building and pairing multiple sequence alignments, and utilize different sequence and template databases.

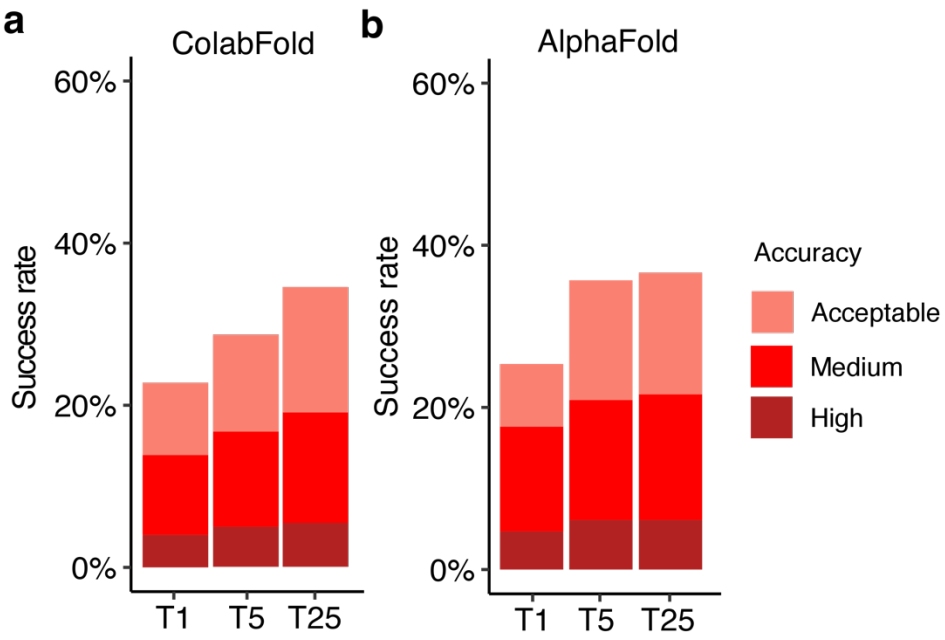

Figure 5.2. Antibody-antigen modeling success by ColabFold and AlphaFold. Antibody-antigen modeling success comparison of **a)** ColabFold and **b)** AlphaFold on 426 antibody-antigen complexes for which both algorithms successfully generated predictions. Structures released on or before April 30, 2018, were allowed as templates during modeling. For each complex, 25 predictions were generated, and were ranked by AlphaFold model confidence score. Antibody-antigen predictions were evaluated for complex modeling accuracy using CAPRI criteria for high, medium, and acceptable accuracy. The success rate was calculated based on the percentage of cases that had at least one model among their top N predictions that met a specified level of CAPRI accuracy. Bars are colored by CAPRI accuracy criteria.

We observed higher success in modeling antibody-antigen complexes by AlphaFold here versus our previous benchmarking study, in which fewer than 10% of cases had top-ranked predictions with near-native accuracy [272] (versus 18% here, as noted above). This difference is likely due to the newer version of AlphaFold used in this study (v2.2 vs. v2.1), which uses a retrained multimer model, as well as different sets of test cases, with the current study representing a substantial expansion over the cases used previously.

5.3.2 Comparing AlphaFold with rigid-body docking algorithms

To compare antibody-antigen modeling performance with a previously developed docking approach, we utilized the global rigid-body docking algorithm, ZDOCK (version 3.0.2) [230], with the IRAD (Integration of Residue- and Atom-based Potentials for Docking) reranking function which was developed to improve the ranking of near-native ZDOCK models [281]. Since many of the antibody-antigen complexes in our benchmark set do not have experimentally determined unbound antibody and/or antigen structures, AlphaFold was employed to generate unbound antibody and unbound antigen inputs for ZDOCK, using templates released on or before April 30, 2018. We selected the top-ranked prediction from AlphaFold as the input for ZDOCK, and only performed ZDOCK docking with unbound structures having a minimum average pLDDT score over 80, in order to exclude cases with likely low quality input modeled structures. In total, 389 complexes met the criteria for minimum average pLDDT score cutoff. Using ZDOCK with IRAD reranking, the majority of test cases did not yield highly accurate predictions (1% medium or higher accuracy, **Figure 5.3a**) as the top-ranked prediction. In contrast, AlphaFold-generated antibody-antigen complex models have a higher percentage of cases (19%) with medium or higher accuracy top-ranked predictions within this set (**Figure 5.3b**). The success rate increases when we consider top 25 predictions generated by ZDOCK (7% of complexes have medium or higher accuracy predictions, **Figure 5.3a**), yet the success is still lower than AlphaFold, which produced medium or higher accuracy predictions for 23% of cases when all 25 predictions were considered. This ZDOCK success is similar to the unbound antibody-antigen docking success in Guest et al. [71], although the difference in test cases, inputs (unbound versus modeled structures), and ZDOCK sampling and model scoring do not support a direct comparison of success rates.

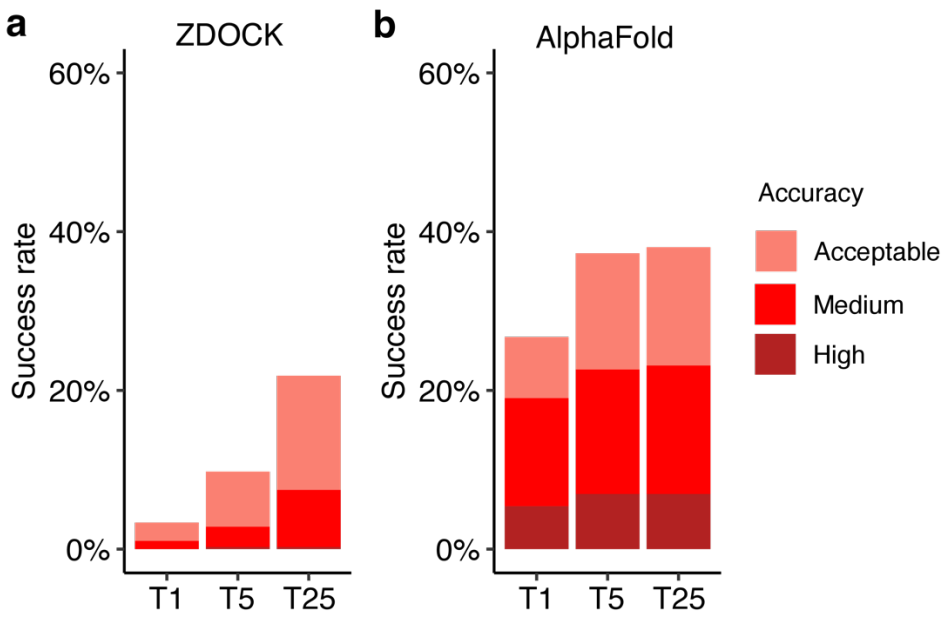

Figure 5.3. Antibody-antigen modeling success by ZDOCK and AlphaFold. Antibody-antigen modeling success comparison of **a** ZDOCK (version 3.0.2) [230] **b** AlphaFold on 389 antibody-antigen complexqes. AlphaFold was used to generate unbound antibody and antigen structure inputs for ZDOCK. Templates released on or before April 30, 2018, were allowed during modeling. Only the top-ranked prediction, based on AlphaFold's model confidence score, was used as input. ZDOCK docking was performed only on inputs with a minimum antibody or antigen pLDDT score above 80 to ensure input model quality. ZDOCK dense sampling was employed, generating in 54,000 predictions per complex, which were subsequently ranked using IRAD [281] scores. Antibody-antigen predictions were evaluated for complex modeling accuracy using CAPRI criteria for high, medium, and acceptable accuracy. For each complex, 25 predictions were generated, and were ranked by AlphaFold model confidence score. The success rate was calculated based on the percentage of cases that had at least one model among their top N predictions that met a specified level of CAPRI accuracy. Bars are colored by CAPRI accuracy criteria.

To compare AlphaFold with a rigid-body docking algorithm that is specialized in antibody-antigen docking, we additionally generated antibody-antigen docking predictions using ClusPro in antibody mode [146] on a subset of 100 cases randomly subsampled from the 389 ZDOCK test set cases. As with ZDOCK, the input for ClusPro consisted of the unbound antibody and antigen models produced by AlphaFold, which met the quality criteria as previously described. The default ranking method of the ClusPro server, which is based on the cluster size, was employed for evaluating the docking poses. ClusPro generated comparable

results as ZDOCK. On this set, AlphaFold generated medium or higher accuracy top-ranked predictions for 19% of the test cases (**Figure 5.4a**), whereas ZDOCK achieved this in 1% of cases (**Figure 5.4b**), and ClusPro did not produce any such top-ranked predictions (**Figure 5.4c**). When top 20 predictions were considered, the success rates rose to 24% for AlphaFold (**Figure 5.4a**), 6% for ZDOCK (**Figure 5.4b**) and 9% for ClusPro (**Figure 5.4c**).

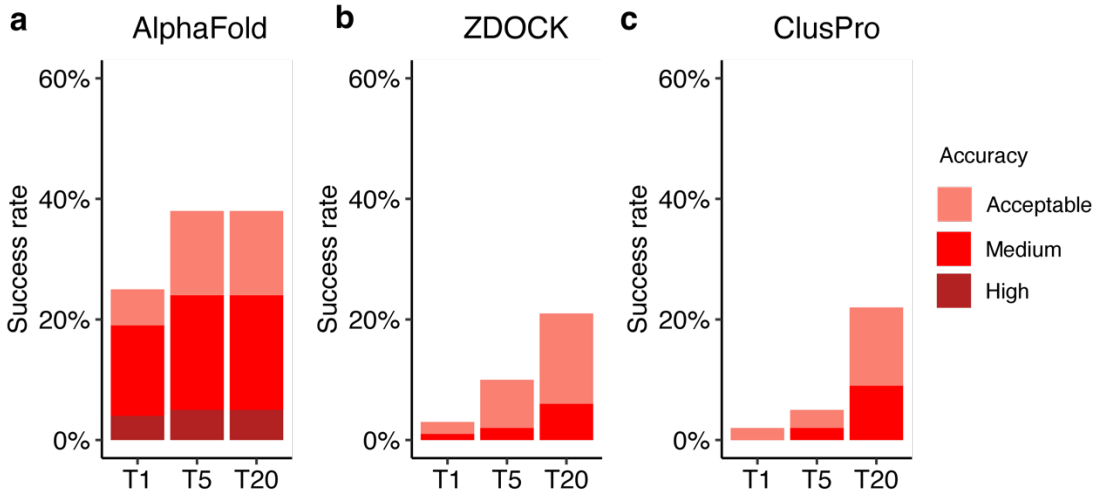

Figure 5.4. Antibody-antigen modeling success by AlphaFold, ZDOCK and ClusPro. Antibody-antigen modeling success comparison of **a** AlphaFold **b** ZDOCK (version 3.0.2) [230] and **c** ClusPro (antibody mode) [146] on 100 antibody-antigen complexes. For each complex, 25 predictions were generated by AlphaFold, and were ranked by AlphaFold model confidence score. AlphaFold was also used to generate unbound antibody and antigen structure inputs for ZDOCK and ClusPro. Templates released on or before April 30, 2018, were allowed during modeling. Only the top-ranked prediction, based on AlphaFold's model confidence score, was used as input. To ensure input quality, only antibody or antigen predictions with minimum pLDDT score above 80 were used. ZDOCK dense sampling was employed, generating in 54,000 predictions per complex, which were subsequently ranked using IRAD [281] scores. ClusPro antibody mode and automatic masking of non-CDR regions were enabled. Modeling success up to top 20 predictions were shown in the figure, as ClusPro generated 20-30 predictions for most cases, except for case 7aqy, where ClusPro produced only 14 predictions. Antibody-antigen predictions were evaluated for complex modeling accuracy using CAPRI criteria for high, medium, and acceptable accuracy. The success rate was calculated based on the percentage of cases that had at least one model among their top N predictions that met a specified level of CAPRI accuracy. Bars are colored by CAPRI accuracy criteria.

Challenges for the rigid-body docking algorithms likely responsible for the relatively low observed success rates include possible local inaccuracies in the unbound models, binding conformational changes (e.g. in CDR loops) even in the case of ideal unbound models, as well as modeled flexible protein terminal regions that are not resolved in some of the structures, which can lead to false positive binding sites for ZDOCK and ClusPro models.

5.3.3 Antibody-antigen modeling accuracy determinants

To identify possible factors associated with modeling outcome, we analyzed properties of the native antibody-antigen complexes in relation to predictive modeling success. As non-protein hetero-atoms (referred to as hetatoms for brevity) such as glycans, lipids, and nucleotides are not modeled by AlphaFold and can play a pivotal role in antibody-antigen interactions, the subset of complexes with such interface hetatoms in our set was identified to assess how their presence in the native complex impacts the accuracy of AlphaFold prediction. In our set, large hetatoms (versus water, ions, or other smaller-sized hetatoms) found at the antibody-antigen interface included glycans, cholesterols, lipid antigens presented by CD1 molecules, and ATP (adenosine triphosphate). Our analysis showed that the presence of large interaction hetatms at the native antibody-antigen interface is associated with lower modeling success (**Figure 5.5a**). Among a total of 49 cases with large interaction hetatoms at the interface, which are primarily cases with interface N-glycans (45 out of 49 cases), the top-ranked predictions of medium accuracy were produced in only 8% of the cases, and no high accuracy top-ranked predictions were produced. In contrast, for cases not belonging to this category, 19% generated top-ranked predictions of medium or higher accuracy. Thus, the lack of explicit consideration of interaction hetatoms may reduce modeling accuracy for some antibody-antigen complexes. Nonetheless, AlphaFold was

able to accurately model a single-domain antibody-antigen complex with a glycosphingolipid antigen α-galactosylceramide (α-GalCer) in the binding interface as shown in **Figure 5.5b** (PDB code 6v7y; single-domain antibody/CD1d α-GalCer complex). The model, a top-ranked prediction for the complex, has medium CAPRI accuracy, and an I-RMSD value of 1.02 Å, indicating that AlphaFold accurately captured the antibody-antigen docking conformation despite the absence of an explicit representation of the glycosphingolipid antigen at the binding interface. Relatedly, implicit accounting for non-protein molecules has been noted previously for AlphaFold models of monomeric proteins with bound-like small molecule binding and catalytic sites [111, 283].

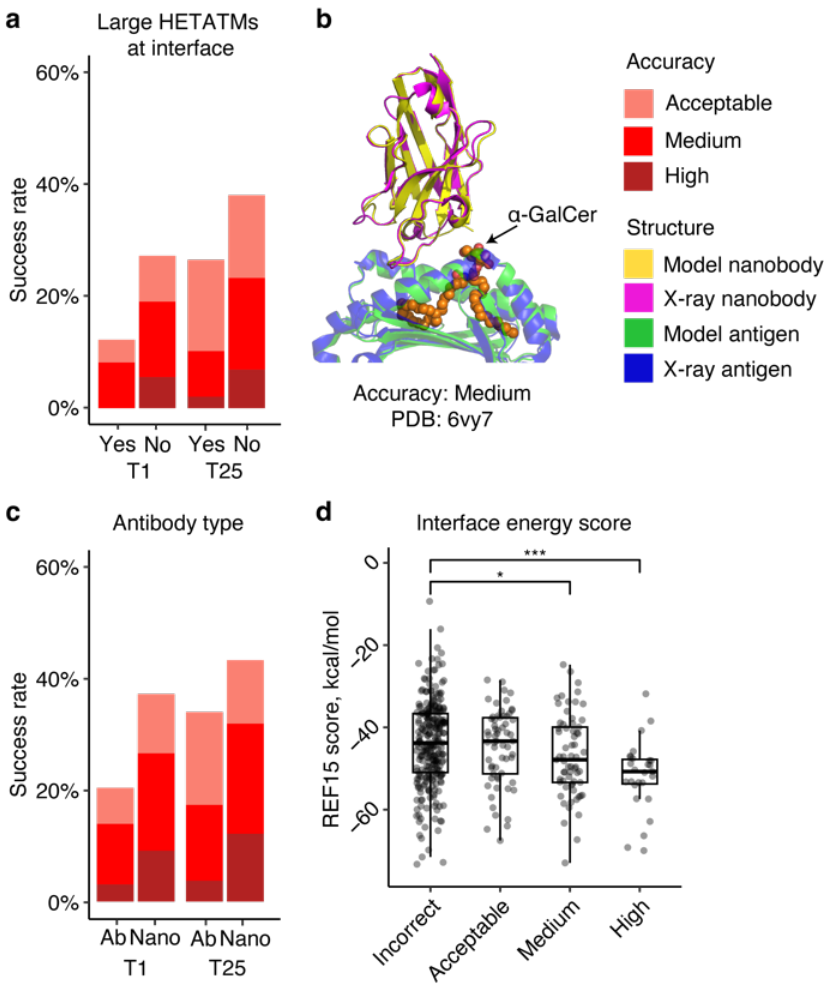

Figure 5.5. Properties associated with antibody-antigen modeling success. a Presence of interaction hetatoms (non-protein atoms) at the antibody-antigen interface. Complexes are classified as either "Yes" (N=49) or "No" (N=378) to indicate whether hetatoms are present or absent in the antibody-antigen interface. **b** An example of a medium accuracy complex model from AlphaFold for an interface hetatom complex, in comparison with the experimentally determined structure (PDB: 6vy7; single-domain antibody/CD1d α-GalCer complex). This model has medium CAPRI accuracy (I-RMSD=1.02 Å), and has the highest model confidence of all 25 predictions of this complex (model confidence=0.85). The complex structure is superposed by antigen with the model and the X-ray structure components colored separately as indicated on right. The α-GalCer glycolipid from the X-ray structure is colored orange. **c** Type of antibody in the complex. Complexes were classified as “Ab” (heavy-light antibody, N=295), or “Nano” (nanobody/VHH, N=132) based on antibody type. T1 and T25 denote AlphaFold modeling accuracy in top 1 (ranked by AlphaFold model confidence score) and in all 25 predictions of the complex. Bars were colored by CAPRI criteria. **d** Distribution of Interface energy score calculated by Rosetta InterfaceAnalyzer [284] protocol (based on Rosetta REF15 energy function [239]) grouped by AlphaFold modeling accuracy. The modeling accuracy is defined as the highest CAPRI criteria prediction in the complex, considering all 25 predictions. Statistical significance values (Wilcoxon rank-sum test) were calculated between interface energy scores for sets of cases with incorrect versus medium and incorrect versus high CAPRI accuracy predictions, as noted at top (* : $p \leq 0.05$, *** : $p \leq 0.001$).

Antigen glycosylation can be an important component in antibody-antigen recognition, with many cases of glycans contacted directly by antibodies [285]. The importance of glycans, as well as the prevalence of antigen N-glycosylation in our dataset (45 out of 49 hetatom interface complexes, as noted above), prompted us to examine antigen glycosylation in the set further. As some X-ray or cryo-EM structures used for analysis may lack resolved glycan atoms, or naturally occurring glycans can be removed enzymatically or via mutation to enable structural characterization, it is possible that some members of the non-interface hetatom set (N=378) may have interface-proximal glycans in vivo. Based on the analysis of antigen source organism and proximity of surface-exposed N-glycosylation motif to the interacting antibody, a subset of N=91 cases were identified to have possible antigen N-glycosylation near antibody binding site. Within the set of non-interface hetatom cases, the predicted antibody-proximal antigen glycosylation subset (N=91) showed a lower modeling success, with medium or higher accuracy top-ranked predictions generated in 16% of cases, compared to 20% for the cases without likely N-glycosylation (N=287) (**Figure 5.6**). Despite the overall lower modeling success, a higher proportion of high accuracy top-ranked predictions were generated in the predicted glycosylation subset (9%, compared to 5% in the aglycosylated interface subset). However, it is important to consider the possibility of false positives in our method. Although the proteins of this organism can be glycosylated, there is a chance that the actual antigen protein is not glycosylated.

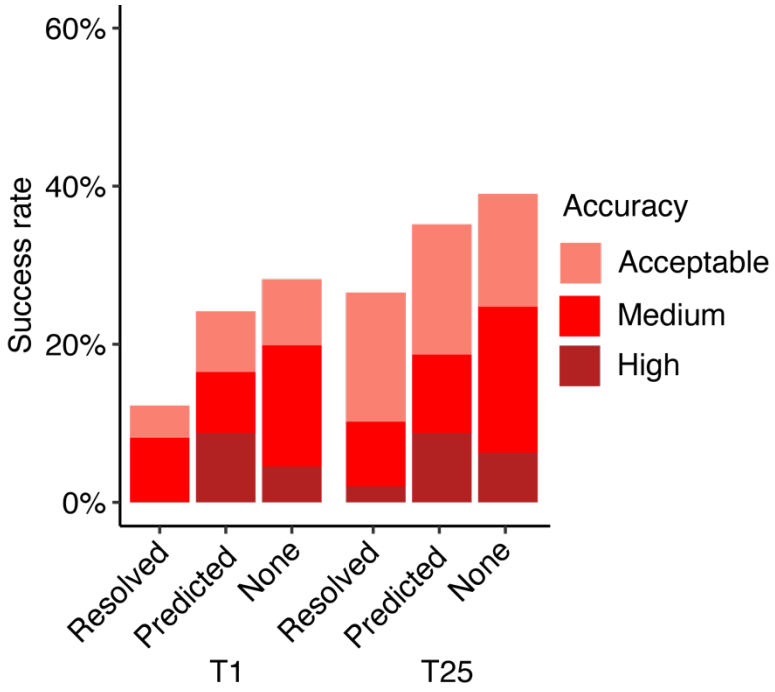

Figure 5.6. Presence of experimentally resolved and predicted interface glycans on modeling success. Complexes were classified as either "Resolved" (glycans or ligands found in antibody-antigen chain interface in experimentally resolved structures, N=49), "Predicted" (glycans not found in antibody-antigen interface of experimentally resolved structures but the antibody-antigen complex was predicted to have antibody-proximal antigen glycosylation, N=91), and "None" (none of the above, N=287). Bars were colored by CAPRI criteria.

We also investigated whether antibody-antigen complexes containing single-chain antibodies (or nanobodies) are more successfully modeled compared to the heavy-light chain only counterparts (**Figure 5.5c**). For nanobody-antigen complexes (N=132), 27% of cases had medium or higher accuracy top-ranked predictions, versus 14% of cases with medium or higher accuracy top-ranked predictions for heavy-light chain antibody-antigen complexes (N=297). To understand the pronounced difference in modeling nanobody-antigen complexes versus antibody-antigen complexes, we investigated the difference in MSA depth of the two types of complexes. We hypothesized that the single-chain variable domains in nanobodies may simplify construction of cross-chain MSAs for nanobody-antigen complexes, as opposed to the more

complex heavy-light chain antibodies. However, after analyzing the MSA depth, we found no statistically significant difference in the number of effective sequences (N_{eff}, a measure of the effective sequence count in an MSA [111]) between the two types of complexes (**Figure 5.7**). This suggests that other factors, such as fewer CDR loops and a smaller search space, may contribute to the observed difference in modeling success. Unlike heavy-light chain antibodies, which possess six CDR loops, the variable domain of nanobodies contains three loops only, thus it is possible that the lower complexity and size of the receptor component of the complex may play a role in the observed improved modeling performance for AlphaFold.

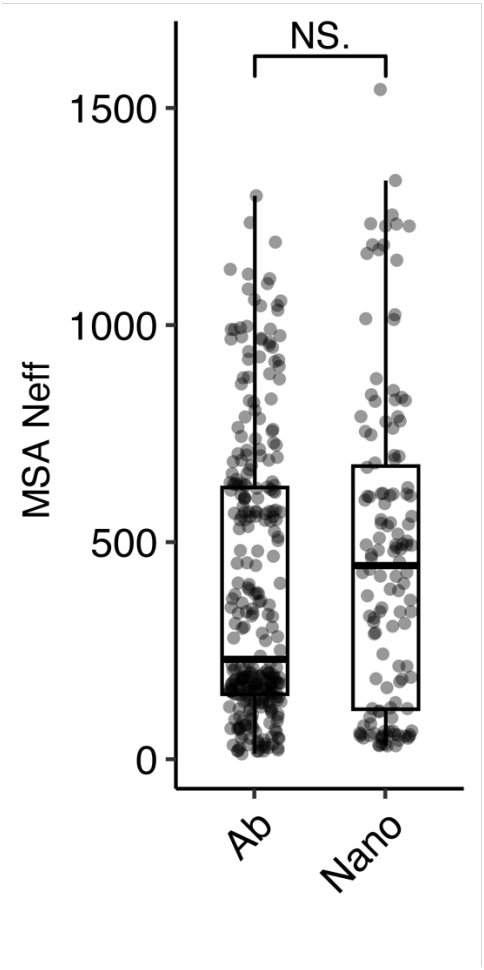

Figure 5.7. Distribution of MSA depth (Neff) grouped by antibody type. Based on the antibody type, complexes are categorized into heavy-light chain antibody-antigen complexes (Ab, N=294), or

nanobody/VHH-antibody complexes (Nano, N=132). Statistical significance values (Wilcoxon rank-sum test) were calculated between MSA depth for antibody targets versus nanobody targets, as noted at top (NS.: not significant, $p > 0.05$).

To investigate whether more favorable antibody-antigen interfaces are more successfully predicted by AlphaFold, we compared antibody-antigen interface energy, computed from the bound complex structure using Rosetta [165], with modeling success considering all 25 predictions of each case (**Figure 5.5d**). We found that more negative interface energies, indicative of more energetically favorable protein-protein interactions, are associated with higher AlphaFold modeling success. The difference in distribution of interface energy scores between complexes is statistically significant between incorrect vs medium accuracy prediction ($p \leq 0.05$), and incorrect vs high accuracy complexes ($p \leq 0.001$), based on Wilcoxon rank-sum test.

While all complexes were modeled with antigen chains comprising the full epitope, a subset of cases did not include additional chains from the full antigen multimeric assembly (based on PDB "bioassembly") in the modeling, due to computational limitations and for modeling efficiency. We examined whether these "partial antigen assembly" cases (N=49) were not as successfully modeled as the non-multimer or full assembly cases (N=378, **Figure 5.8a**), due for instance to non-native surface regions that are normally buried in the full antigen assembly being engaged by antibodies in the models. Indeed, the partial antigen assembly cases exhibited lower modeling success versus the remainder of the cases. Additionally, consistent with our previous benchmarking study [272], success analysis on the current set of antibody-antigen complexes shows that larger complexes are generally more difficult to predict (**Figure 5.8b**).

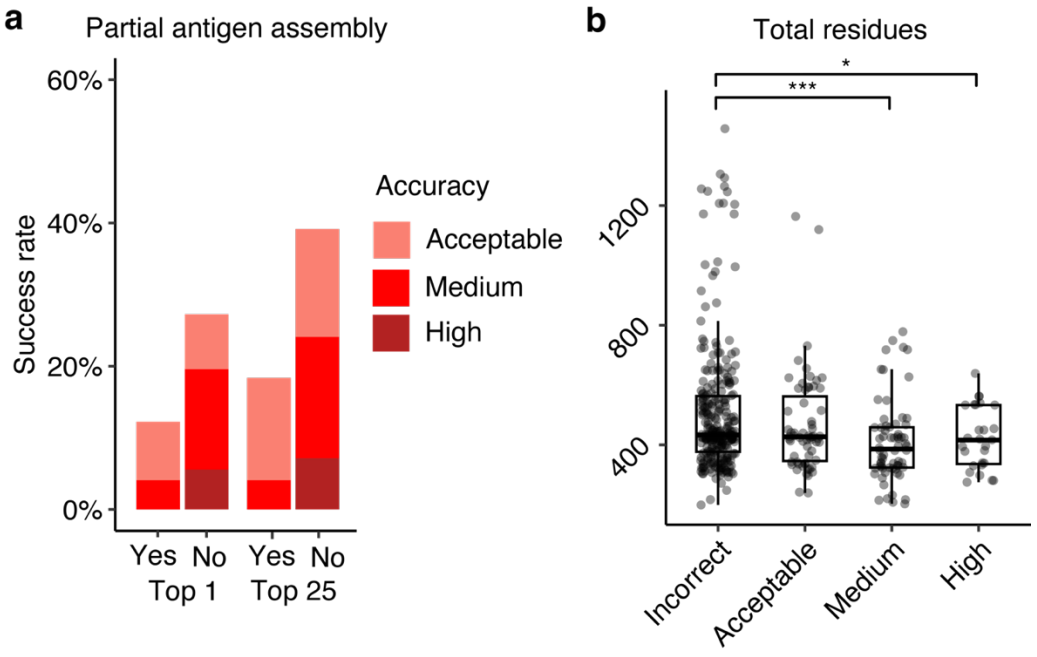

Figure 5.8. Antibody-antigen modeling success determinants. (a) Partial versus full antigen assembly input. Complexes were classified as either "Yes" (N=49) or "No" (N=378) to indicate whether a partial antigen assembly was modeled, meaning that the antigen was modeled without additional chains that are present in the full PDB bioassembly. T1 and T25 denote AlphaFold modeling accuracy in top 1 (ranked by AlphaFold model confidence score) and in all 25 predictions of the complex. Bars are colored by CAPRI criteria. (**b**) Total number of residues in the complex grouped by AlphaFold modeling success. The modeling success is defined as the highest CAPRI criteria prediction in the complex, considering all 25 predictions. Statistical significance values (Wilcoxon rank-sum test) were calculated between total residue counts for sets of cases with incorrect versus medium and incorrect versus high CAPRI accuracy predictions, as noted at top (*$p \leq 0.05$, ***$p \leq 0.001$).

We also examined the accuracy of six complementarity determining region (CDR) loops in the modeled complexes, computing CDR loop RMSDs in the models with respect to the corresponding CDR loop from the experimentally determined complex structure (**Figure 5.9**). Interestingly, the CDRH3 loop exhibited differences in accuracy across sets of models with different CAPRI accuracy levels, with median CDRH3 RMSD of 0.6 Å for high accuracy models, 1.2 Å for medium accuracy models, 2.2 Å for acceptable accuracy models, and 2.4 Å for incorrect models, indicating that accurate prediction of CDRH3 loops is associated with near-native modeling accuracy in antibody-antigen complex prediction in AlphaFold. However, it

should be noted that CDR loop accuracy (among other features of the modeled interfaces) can play a role in the CAPRI antibody-antigen complex accuracy assessments themselves, with high CAPRI accuracy models likely requiring relatively low CDRH3 RMSDs to closely reflect the native interface.

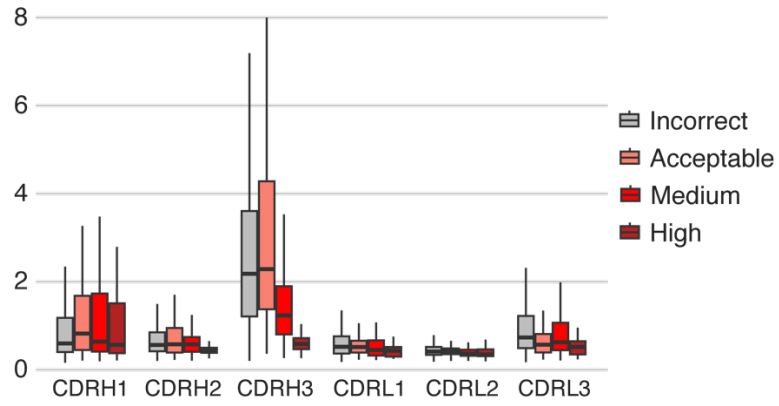

Figure 5.9. Distribution of CDR modeling accuracy, grouped by CDR type and by AlphaFold modeling accuracy. AlphaFold modeling accuracy is defined as the accuracy of highest CAPRI criteria prediction in the complex, considering all 25 predictions. For CDRH1, CDRH2 and CDRH3, the numbers of data points in each category are 261, 60, 62, 25 for incorrect, acceptable, medium and high success groups. For CDRL1, CDRL2 and CDRL3, the numbers of data points in each category are 192, 45, 39, 10 for incorrect, acceptable, medium, and high success groups. Bars are colored by CAPRI model accuracy.

We also examined possible failures of antigen structure modeling as factors in antibody-antigen modeling success (**Figure 5.10**). The analysis of the antigen accuracy in top-ranked complex predictions revealed that while only 12 complexes out of 427 complexes have antigen predictions with TM-score [231] values below 0.7 with respect to the experimentally determined antigen, indicating a relatively lower level of structural similarity [286], the majority of complex predictions included relatively accurate modeling of the antigen subunit. This suggests that most of the failed predictions can be primarily attributed to incorrect docking poses or local structural perturbations rather than inaccurate antigen subunit predictions.

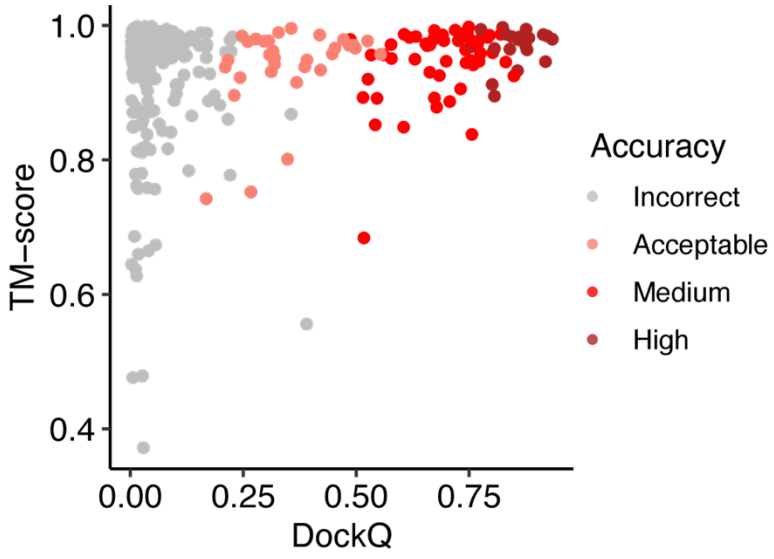

Figure 5.10. Relationship between antigen modeling accuracy and complex prediction accuracy. Top-ranked predictions of 427 complexes generated by AlphaFold are represented as data points. The antigen accuracy is measured by TM-score of the top-ranked prediction, and the complex model accuracy measured by DockQ score. If the antigen has multiple chains, the minimum TM-score of all antigen chains is selected as the antigen TM-score. Data points are colored by CAPRI model accuracy.

5.3.4 Model confidence score comparison

The reported success of model accuracy scores produced by AlphaFold [118, 272] led us to evaluate the ability of those scores, or adaptations thereof, to discriminate between accurate vs. incorrect antibody-antigen predictions. We assessed AlphaFold's model confidence score, which is a linear combination of pTM and ipTM [118] scores, as well as interface pLDDT (I-pLDDT), which is based on residue-level confidence scores for antibody-antigen interface residues (4 Å distance cutoff), as used in previous studies [237, 272], for discrimination of correct antibody-antigen models (**Figure 5.11**). While both exhibited significant correlations with DockQ score [277], which is a continuous measure of complex model accuracy, I-pLDDT was marginally superior (**Figure 5.11a-b**); this was also evident for comparison of the scores with CAPRI accuracy levels (**Figure 5.11c-d**).

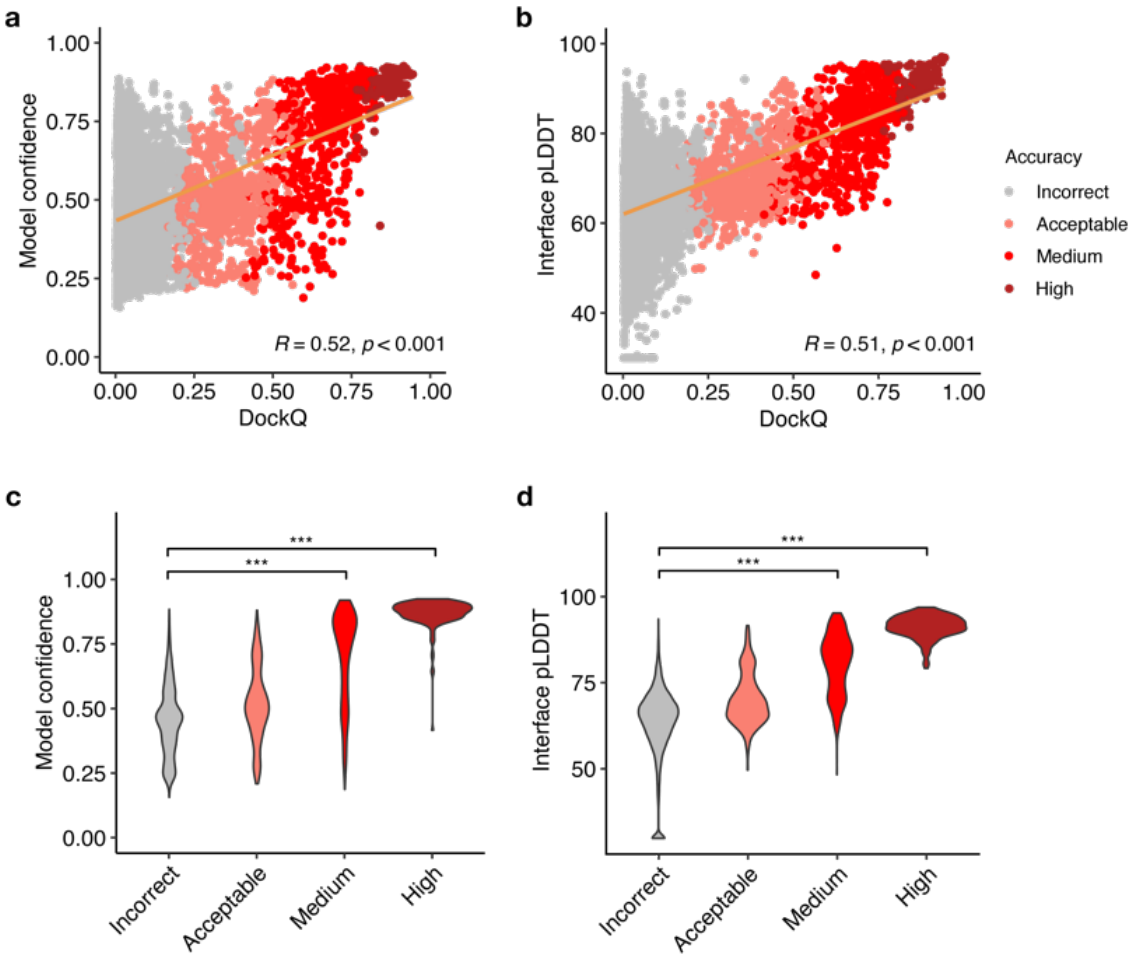

Figure 5.11. AlphaFold model confidence scores and model accuracy. Scatter plots compare **a)** model confidence and **b)** interface pLDDT score with model accuracy, with accuracy assessed by DockQ score. In the scatter plots, all 25 models representing 427 complexes are depicted as data points, with their colors indicating the model quality according to CAPRI criteria. The orange line represents the linear regression, and the lower right corner of the scatter plots displays the Pearson's correlation coefficients and correlation p-values. Distribution of **c** model confidence and **d** interface pLDDT score, grouped by the CAPRI criteria of AlphaFold predictions. Interface pLDDT score is defined as the mean of pLDDT scores of residues within 4 Å of the antibody-antigen interface. Complexes without contacts within 4 Å of antibody-antigen interface is assigned a pLDDT score of 30. Statistical significance values (Wilcoxon rank-sum test) were calculated between model scores for sets of predictions with incorrect versus medium and incorrect versus high CAPRI accuracy, as noted at top (*** : $p \leq 0.001$).

I-pLDDT also provided outstanding discrimination between incorrect vs. medium or higher accuracy models based on receiver operating characteristic (ROC) area under the curve (AUC) metrics (AUC=0.92), which is higher than that of the model confidence (AUC=0.88;

Table 5.2). We also tested the individual components of the model confidence scores (pTM and ipTM) (**Figure 5.12**), which did not yield improved correlations with DockQ scores versus model confidence. When excluding minterface (for which I-pLDDT was set to an arbitrary minimum value in **Figure 5.11b** and in the corresponding correlation calculation), the correlation between the interface pLDDT and DockQ increased to r=0.57 (**Figure 5.13a**), which demonstrates a more significant difference compared to the correlation between the model confidence and DockQ (r=0.53, **Figure 5.13b**).

Table 5.2. Area under the ROC curve (AUC) value for protein model quality classes as a function of different scoring metrics.

Score[a]	Binary classification ROC AUC[b]		Multi-class classification[b]
	Incorrect vs. High	Incorrect vs. Medium and High	
Interface pLDDT	1.00	0.92	0.88
Model confidence	0.99	0.88	0.85
ipTM	0.99	0.87	0.85
pTM	0.99	0.88	0.84

[a]Scoring methods. Model confidence, ipTM, and pTM are confidence scores from AlphaFold. Interface pLDDT is the average AlphaFold pLDDT score of antibody-antigen interface residues within 4 Å distance cutoff. Models without antibody-antigen interface contacts were assigned an interface pLDDT value of 30.
[b]The ROC AUC values of binary classification and multi-class classification were calculated using the R pROC[250] and multiROC[251] packages, with classes defined by model CAPRI accuracy, which assigned antibody-antigen models into incorrect (n=9,062), acceptable (n=773), medium (n=684) and high (n=156) accuracy categories.

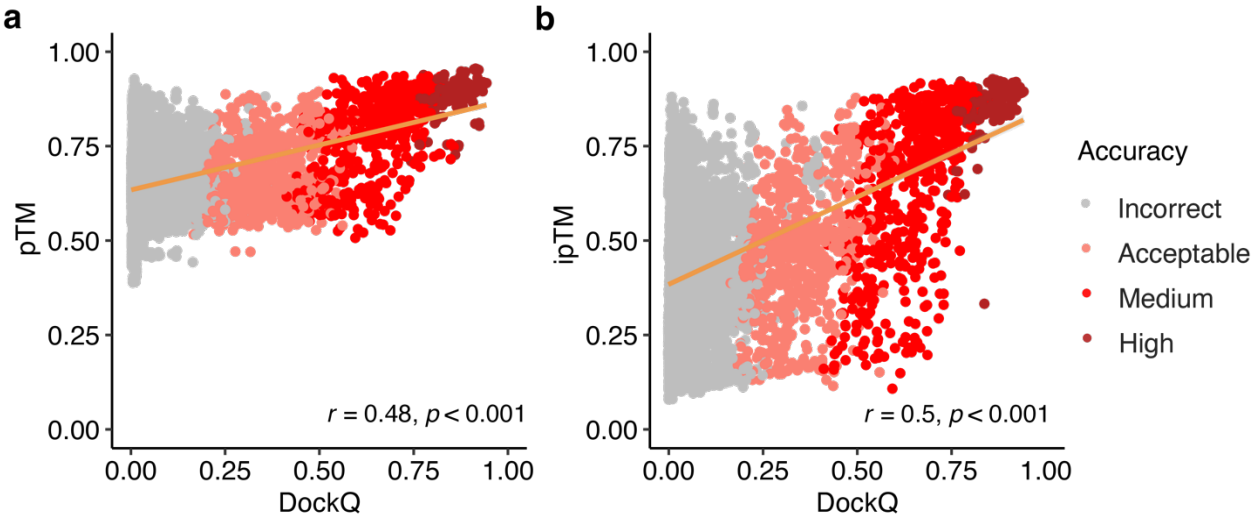

Figure 5.12. **Relationship between model confidence scores and model accuracy.** Scatter plots depicting the association between the **a)** pTM, **b)** ipTM scores and DockQ scores. In the scatter plots, all 25 models representing 427 complexes are depicted as data points, with their colors indicating the model quality according to CAPRI criteria. The orange line represents the linear regression, and the lower right corner of the scatter plots displays the Pearson's correlation coefficients and correlation p-values.

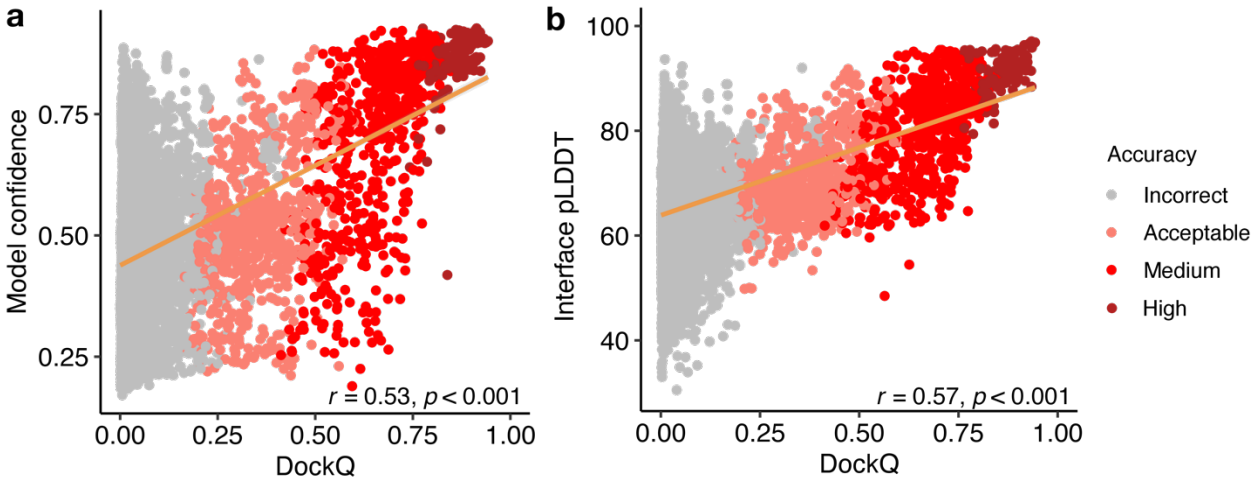

Figure 5.13. **Relationship between model scores and model accuracy.** Scatter plots depicting the association between the **(a)** model confidence, (**b)** interface pLDDT scores and the DockQ scores. A total of 10,236 data points were present in each scatter plot, which include all 25 models representing 427 complexes, excluding models without side-chain contacts within 4 Å across the antibody-antigen interface. Data points are colors indicating the model quality according to CAPRI criteria. The orange line

represents the linear regression, and the lower right corner of the scatter plots displays the Pearson's correlation coefficients and correlation p-values.

One advantage of I-pLDDT over ipTM and model confidence (which primarily consists of ipTM) is that it is specifically focused on the antibody-antigen interface, whereas ipTM is calculated across all inter-chain interfaces of complex models, including heavy-light and multiple antigen chains, thus the latter scores may be influenced by less relevant elements of the complex. Overall, these results support the use of I-pLDDT as a primary metric in assessing the quality of AlphaFold antibody-antigen models.

5.3.5 Progressive improvements over recycling iterations

Recycling is a critical component of the AlphaFold algorithm[111, 118], wherein each model is input back to the system for further optimization. To improve our understanding of the impact of recycling iterations on AlphaFold modeling of antibody-antigen complexes, we modified the AlphaFold pipeline in ColabFold. ColabFold was preferable to utilize in this context versus the default AlphaFold pipeline due to its speed, in order to enable output and analysis of the antibody-antigen complex predictions at each recycling iteration. Our analysis demonstrates an increase in model accuracy as recycling iterations progress (**Figure 5.14a**). In fact, approximately 50% of predictions of medium or higher accuracy after the third recycle iteration were incorrect models before recycling iterations (**Figure 5.15**).

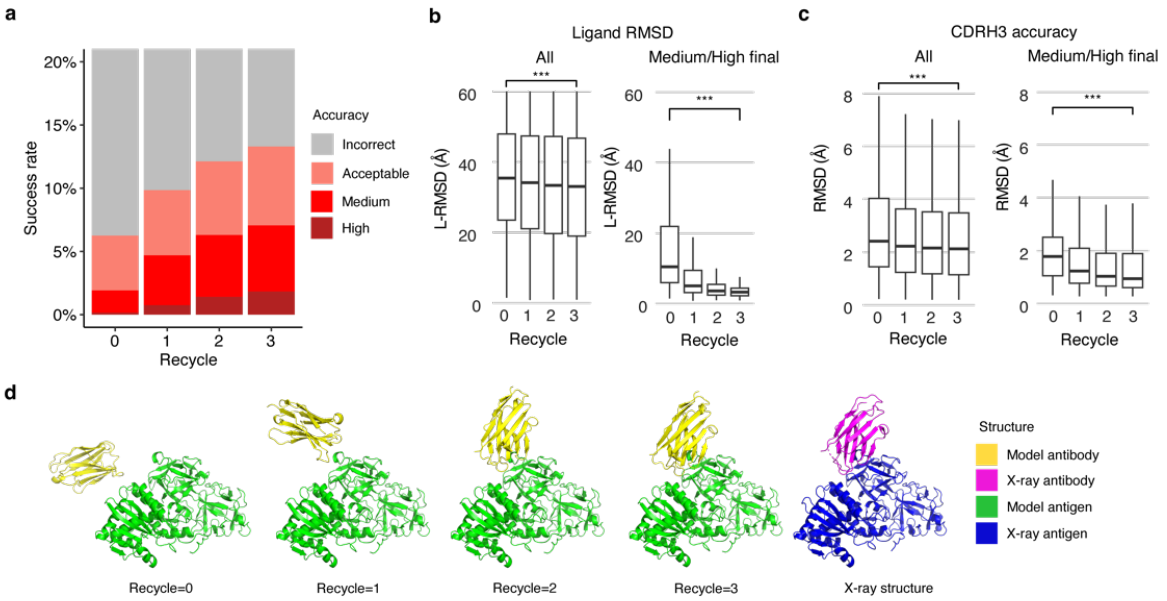

Figure 5.14. Analysis of antibody-antigen predictions across recycling iterations. (a) The accuracy of antibody-antigen complex predictions across up to three recycling iterations. Complex prediction accuracy across recycling iterations (up to three recycles, denoted by the x-axis). Success rate is defined as the proportion of predictions of specific level of CAPRI criteria in a total of 25 prediction per complex, 426 complexes total, at the given recycle. Recycle=0 denotes the state of the prediction before recycling iterations begin. (**b)** Distribution of the ligand RMSD (L-RMSD, Å) of antibody-antigen prediction at each recycling iteration (denoted by the x-axis), of all predictions (25 predictions * 426 complexes, left panel) or a subset of predictions of medium or high CAPRI accuracy at recycle=3 (106 predictions, right panel). (**c)** Distribution of the CDRH3 accuracy of antibody-antigen prediction at each recycling iteration (denoted by the x-axis), of all or a subset of predictions of medium or high CAPRI accuracy at recycle=3. CDRH3 accuracy is defined as the change in RMSD of the CDRH3 region, when superposing the predicted antibody (in the antibody-antigen complex prediction) onto the experimentally resolved antibody (in the antibody-antigen complex) using the antibody framework region. Statistical significance values (Wilcoxon rank-sum test) were calculated between RMSD values for sets of predictions at the outset of recycling iterations (recycle=0) vs at recycle=3, as noted at top (***$p \leq 0.001$). **d** Example of a prediction across recycling iterations (PDB 7kd2; nanobody/Ricin complex). This prediction's CAPRI accuracy level across recycles was incorrect at recycle=0 (I-RMSD=17.98 Å), incorrect at recycle=1 (I-RMSD=10.90 Å), acceptable at recycle=2 (I-RMSD=2.52 Å), and medium at recycle=3 (I-RMSD=1.45 Å). The CDRH3 RMSD of the prediction across recycling iteration 0, 1, 2, and 3 was 1.39 Å, 1.19 Å, 1.27 Å, and 1.17 Å, respectively. The L-RMSD of the prediction across recycling iteration 0, 1, 2, 3 was 49.95 Å, 24.68 Å, 5.42 Å, 4.25 Å, respectively. Antibody and antigen chains of the predictions and x-ray structure were colored as indicated. Predictions were generated with ColabFold due to its faster model generation speed, compared to AlphaFold.

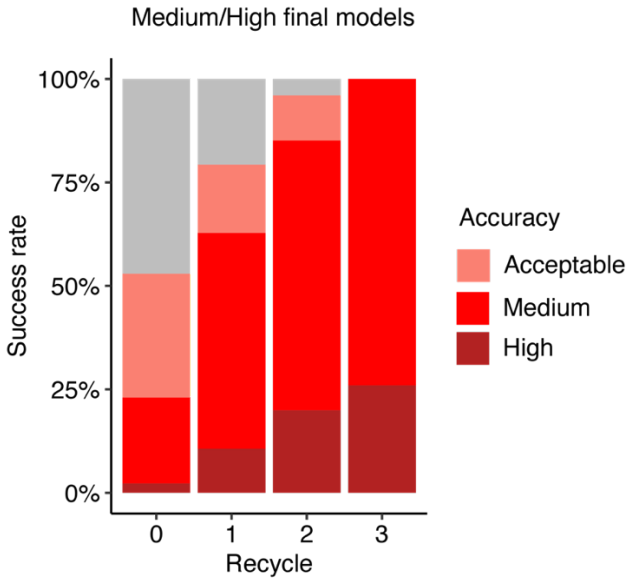

Figure 5.15. Analysis of antibody-antigen predictions accuracy across recycling iterations. The modeling success of complexes at each recycle focusing on a subset of predictions that reached medium or higher accuracy after 3 recycling iterations (N=106). Recycle=0 denotes the state of the prediction before recycling iterations begin.

Next, we analyzed specific changes in antibody-antigen model across recycling iterations, identifying notably enhanced features and those that are unchanged. Features that were improved highlights the strength of AlphaFold, whereas the lack of improvement may highlight areas of difficulty or suggest that these features were already optimal at the start and did not require further refinement. We analyzed both the accuracy of antibody positioning on the antigen and the quality of the highly variable CDR loop of the antibody. Given the high variability in CDRH3 RMSD (**Figure 5.8**), compared to the RMSD of other CDR loops, we focused our analysis of CDR loops on the CDRH3. Considering all predictions, we observed a marginal yet significant improvement in both the antibody-antigen binding conformation as measured by ligand RMSD (L-RMSD) (**Figure 5.14b**, left panel) and the CDRH3 loop accuracy (**Figure 5.14c**, left panel). Upon examining the subset of cases with medium or higher accuracy at recycle 3, we observed that the antibody-antigen binding conformation score L-RMSD exhibited a pronounced and

significant improvement (**Figure 5.14b**, right panel), while the improvement in CDRH3 loop RMSD was significant but not as pronounced (**Figure 5.14c**, right panel), indicating that for models to attain high accuracy at the end of the recycling iteration, it is helpful for AlphaFold to accurately predict the CDRH3 loop relatively accurately before recycling iterations begin.

The capability of AlphaFold to perform rigid-body protein movements over recycling iterations, is shown in **Figure 5.14d** (nanobody/Ricin complex). This prediction was of incorrect accuracy before recycling iterations and was improved to a medium accuracy prediction at recycle 3. Over recycling iterations, the L-RMSD of this prediction exhibited a substantial degree of improvement, from 49.95 Å before recycling, to 4.25 Å at recycle 3. Unlike L-RMSD, the CDRH3 loop of this prediction was accurately predicted (CDRH3 RMSD=1.39 Å) before the recycling iterations.

The importance of CDRH3 loop accuracy for complex modeling success was further explored by the analysis of CDRH3 loop conformations of modeled unbound structures. Unbound antibody structures were generated with AlphaFold with a template date cutoff of April 30, 2018, and the CDR loops of the unbound antibody models were compared to those of the antibodies in the antibody-antigen complexes. The RMSD between CDR loops of the unbound models and the antibody in the bound is compared against the complex modeling success of top-ranked antibody-antigen models generated by AlphaFold in **Figure 5.16**. Although the relatively small numbers of high accuracy cases limits this comparison, the accuracy of the CDRH3 modeling in unbound antibody structures for high antibody-antigen models was found to be significantly higher than that of the incorrect accuracy models ($p \leq 0.05$), suggesting that antibodies with unbound models that more closely resemble the bound loop conformation are likely to be more accurately modeled in the form of antibody-antigen complexes.

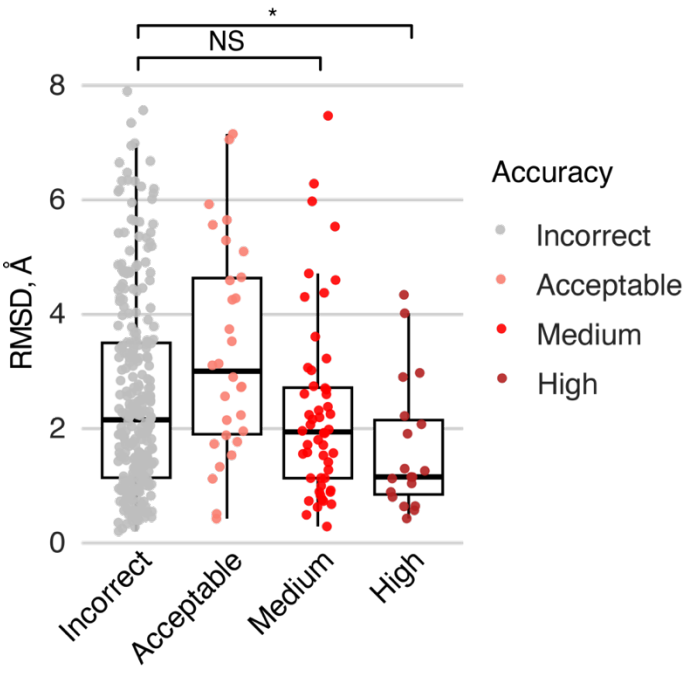

Figure 5.16. The distribution of CDRH3 modeling accuracy in top-ranked unbound antibody model grouped by top-ranked complex modeling success. The modeling success is CAPRI criteria of top-ranked complex prediction generated by AlphaFold. The CDRH3 RMSD measures the RMSD of the top-ranked unbound antibody prediction generated by AlphaFold. Numbers of data points in incorrect, acceptable, medium, and high categories are 304, 31, 52, 19. Statistical significance values (Wilcoxon rank-sum test) were calculated between RMSD values for sets of cases with incorrect versus medium and incorrect versus high CAPRI accuracy predictions, as noted at top (NS: $p > 0.05$, *$p \leq 0.05$).

5.3.6 Input of subunit chains in bound conformation enables higher success

To better understand the factors that can enhance the success rate of the AlphaFold antibody-antigen modeling, we utilized native antibody-antigen chains as templates within the AlphaFold modeling pipeline, to gauge whether AlphaFold can better assemble the complex structures given the bound subunit chains. Modifications were made to the AlphaFold pipeline to optionally input specific selected PDB templates for each chain. To test performance, we randomly selected 100 cases from the full antibody-antigen benchmark that do not have observed glycans at the antibody-antigen interface and do not belong to the partial antigen assembly category, due to observed change in performance for those sets of cases (**Figure 5.5**, **Figure 5.8**). On this subset

of 100 cases, the use of default templates identified from the AlphaFold pipeline resulted in 18% success in generating near-native (medium or high accuracy) top-ranked predictions (**Figure 5.17a**), which is similar to the performance on the full benchmark (**Figure 5.1a**).

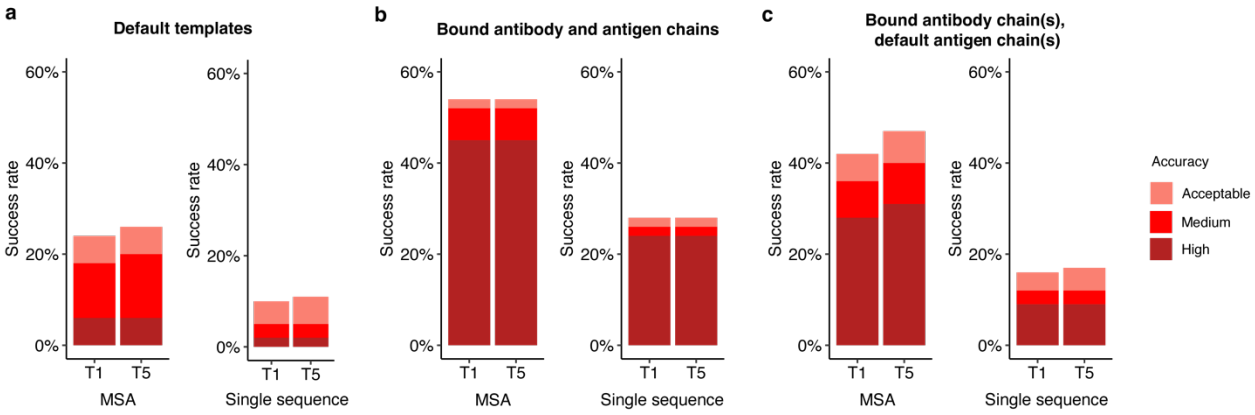

Figure 5.17. Improved subunit modeling enhances antibody-antigen complex modeling success. Antibody-antigen modeling success of AlphaFold by utilizing **(a)** templates identified through the default template search protocol, (**b)** bound antibody and antigen chains as templates, **c** bound antibody and default antigen chains (identified by the default search protocol) as templates. Benchmarking was performed on a total of 100 antibody-antigen complexes. The success rate was calculated based on the percentage of cases that had at least one model among their top N predictions that met a specified level of CAPRI accuracy. Bars are colored by CAPRI accuracy criteria.

A substantial improvement in accuracy was observed when experimentally determined antibody-antigen chains were used as individual chain templates, in which case the success in generating near-native top-ranked predictions was 52% (**Figure 5.17b**). Analysis of the top-ranked prediction success determinants shows that distribution of interface energy score (**Figure 5.18a**) and change in solvent-accessible surface area (ΔSASA) for hydrophobic part of the antibody-antigen interface (**Figure 5.18b**) are significantly different ($p \leq 0.01$) between complexes that have incorrect versus high accuracy top-ranked predictions, indicating that despite using bound template structures, AlphaFold has difficulty predicting the complex structure for antibody-antigen interactions with less favorable computed interface energies and with smaller hydrophobic interface area.

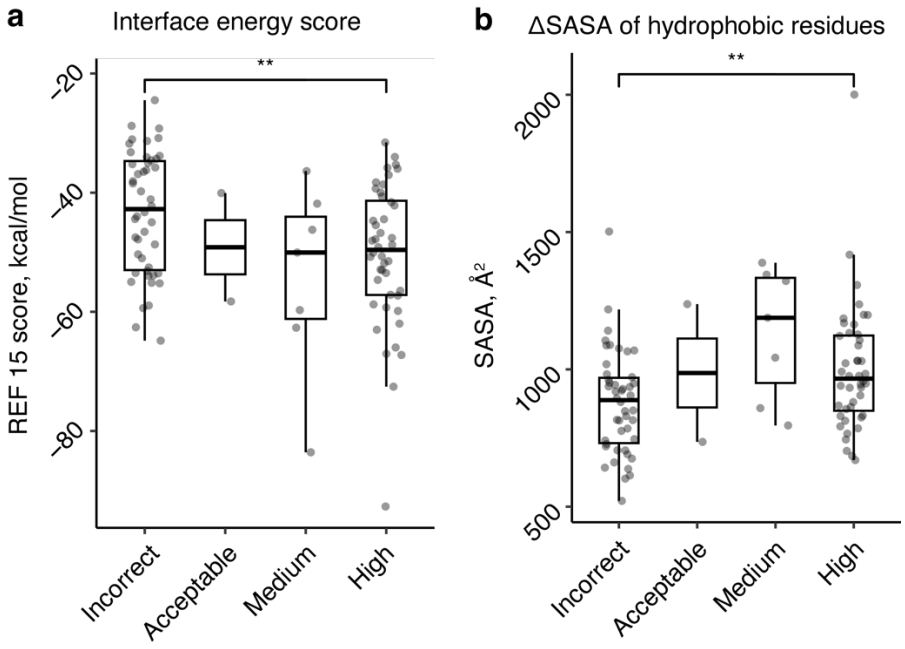

Figure 5.18. Antibody-antigen modeling success determinants of AlphaFold with bound antibody and antigen structures as templates. Distribution of **(a)** interface energy score and (**b)** Change in solvent-accessible surface area (ΔSASA) of hydrophobic part of the antibody-antigen interface by complex modeling success by AlphaFold when using bound antibody and antigen structures as templates. The modeling success is defined as the highest CAPRI criteria prediction in the complex, considering all 5 predictions. Numbers of data points in incorrect, acceptable, medium and high categories are 46, 2, 7 and 45. Statistical significance values (Wilcoxon rank-sum test) were calculated between scores for sets of cases with incorrect versus high CAPRI accuracy predictions, as noted at top (**$p \leq 0.01$).

5.3.7 Accurate subunit modeling and antibody-antigen prediction success

For comparison of predictive success with bound component inputs, we employed ZDOCK[230] with IRAD[281] scoring to perform global rigid-body docking and ranking with the bound antibody and antigen structures (with randomized initial orientations). This approach led to a relatively high proportion (62%) of cases having top-ranked predictions with medium or higher accuracy (**Figure 5.19a**), indicating that both traditional docking as well as AlphaFold can successfully assemble over half, antibody-antigen complex structures with bound components, while still unable to assemble a sizable fraction of them.

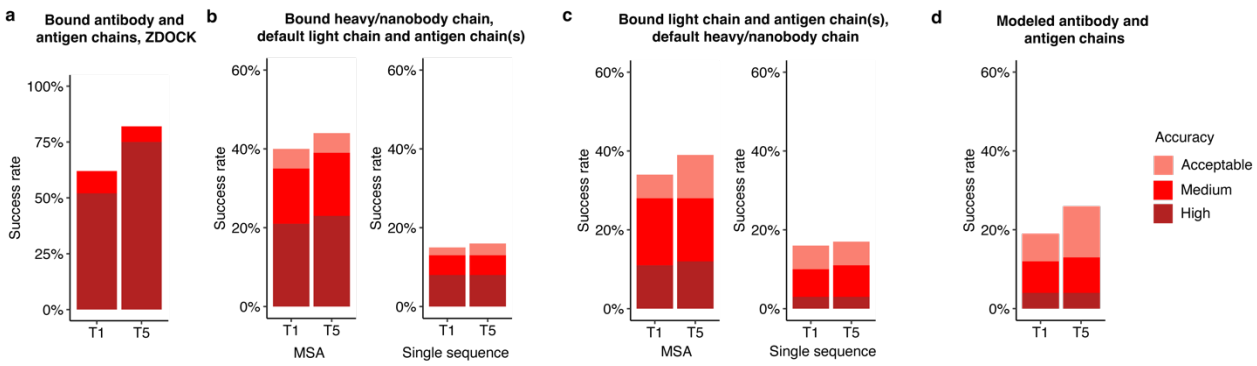

Figure 5.19. Antibody-antigen modeling success using varying template inputs. Antibody-antigen modeling success of **(a)** ZDOCK (version 3.0.2) with IRAD re-ranking of dense-sampling predictions (54,000 predictions per complex), utilizing bound antibody and bound antigen chains as docking input, and of AlphaFold by utilizing (**b**) bound heavy/nanobody chain, and default light chain and antigen chains as templates, (**c)** bound light chain, and default heavy/nanobody chain and antigen chains as templates, (**d)** antibody and antigen chains modeled by AlphaFold as templates. A template date cutoff of 2018-04-30 was applied to identify default templates. Benchmarking was performed on a total of 100 antibody-antigen complexes. The success rate was calculated based on the percentage of cases that had at least one model among their top N predictions that met a specified level of CAPRI accuracy. Bars are colored by CAPRI accuracy criteria.

Compared to using all experimentally determined chains as templates, utilizing only certain bound chains as templates resulted in decreased model accuracy. Specifically, with default antigen templates, and experimentally determined antibody heavy and light chains provided as templates, 36% of top-ranked predictions were of medium or higher accuracy (**Figure 5.17c**). A similar near-native success rate (35%) was observed when using only bound antibody heavy chains as templates (**Figure 5.19b**), while light chain and antigen bound chain templates led to 28% of cases with near-native top-ranked predictions (**Figure 5.19c**). Overall, the resulting success rates of these bound chain template scenarios are higher than those obtained using the default templates, in which case 18% of the complexes have medium or higher accuracy top-ranked predictions. While not reflective of actual predictive modeling scenarios due to the use of bound structure information, these results indicate the potential impact of accurate

and bound-like subunit modeling, as well as its theoretical maximal effect, for AlphaFold antibody-antigen modeling success.

Inspired by the findings, we tested using modeled subunit structures as templates in antibody-antigen modeling, as we hypothesized that if these models are more accurate than the default pipeline's selected template structures, they could potentially improve AlphaFold's performance over the default pipeline and templates. However, this yielded lower success than observed when using default templates (12%, versus 18% for near-native modeling success) (**Figure 5.19d**), showing that inaccuracies or deviations from the bound components for the unbound models led to far different behavior than when using actual bound components.

5.3.8 Utilizing antigen structures bound to other antibodies as templates

Given the practical limitations in obtaining antigens bound to the target antibodies as modeling templates, we evaluated the impact of using antigens that are bound to distinct antibodies as templates. The criteria for selecting such antigens are outlined in the Supplementary Methods. In the subset of 100 antibody-antigen complexes, we conducted such experiment on 73 antibody-antigen complexes with qualifying antigen templates identified. Results indicated that use of alternative antigen templates, compared to the original antigen template, decreased modeling success. Percentage of cases with top-ranked medium or higher accuracy predictions were 49% using bound antibody and bound original antigen (**Figure 5.20a**), 41% using bound antibody and alternative bound antigen (**Figure 5.20a**), 18% using default AlphaFold antibody templates and bound original antigen (**Figure 5.20b**), 14% with default antibody and alternative bound antigen (**Figure 5.20b**). When models were generated with bound antibody templates, the use of alternative bound antigens resulted in higher modeling success (41% cases with medium or

higher top ranked predictions, **Figure 5.20a**) compared to the default AlphaFold antigen templates (35%, **Figure 5.20a**). Conversely, when default AlphaFold antibody templates were employed, the success rate with alternative bound antigens (14%, **Figure 5.20b**) was comparable to that achieved with default AlphaFold antigen templates (15%, **Figure 5.20b**). This suggest that while alternative bound antigens can enhance model accuracy when paired with their corresponding bound antibodies, their advantage diminishes when used with generic antibody templates. Taken together, these results highlight the importance of the antibody template choice in modeling outcomes, and more broadly, deviations in modeling template from bound form led to decrease in modeling success.

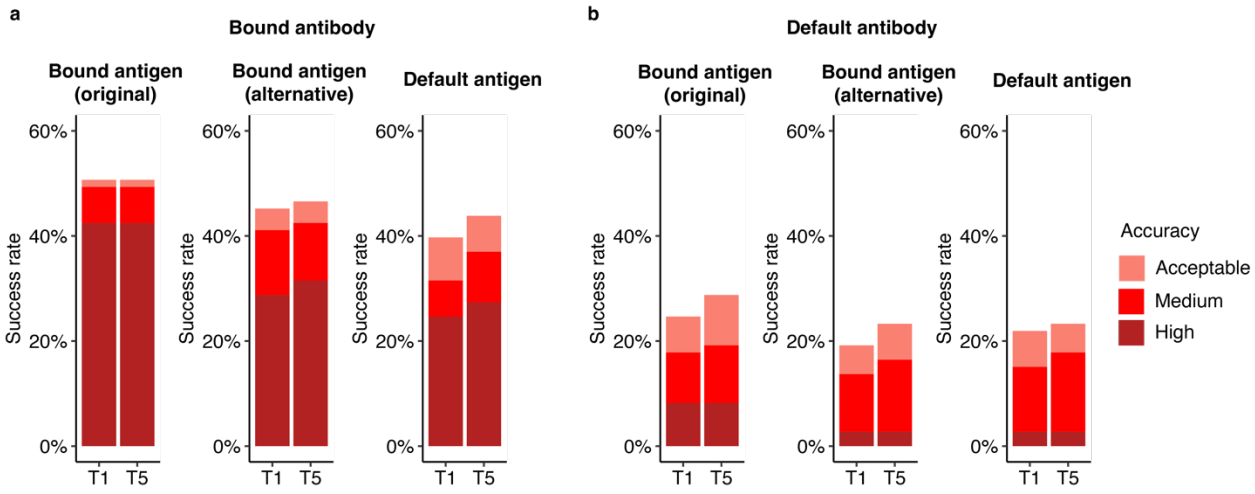

Figure 5.20. Antibody-antigen modeling success using antigens bound to alternative antibodies as template inputs. AlphaFold generated antibody-antigen complexes using **(a)** bound antibody and (**b)** default antibody, coupled with bound antigen (original), bound antigen (alternative) and default antigen as templates. A template date cutoff of 2018-04-30 was applied to identify default templates. Benchmarking was performed on a total of 73 antibody-antigen complexes. The success rate was calculated based on the percentage of cases that had at least one model among their top N predictions that met a specified level of CAPRI accuracy. Bars were colored by CAPRI accuracy criteria.

5.3.9 MSA provides important information for accurate prediction of complexes

We also evaluated the performance of AlphaFold without MSAs, to test the impact on complex assembly when subunit structures are known (thus MSA would not in principle be needed for subunit structure modeling), given the likely lack of direct co-evolutionary information present in antibody-antigen MSAs. The removal of MSAs was implemented through modifications to the AlphaFold pipeline, as noted in the Methods. Our results indicated a notable decrease in accuracy when MSA was disabled, as compared to the with-MSA counterparts (**Figure 5.17** and **Figure 5.18b**,**c**). This prompted us to investigate the possible association between the depth of MSA and the modeling outcome by AlphaFold.

We investigated the impact of MSA depth on modeling success by the full AlphaFold protocol, grouping the complexes by prediction accuracy and comparing distributions of MSA depth (N_{eff}) (**Figure 5.21**). The distribution of N_{eff} was found to be statistically significant between incorrect and medium accuracy classes ($p \leq 0.01$), and between incorrect and high accuracy classes ($p \leq 0.01$). We also compared the docking model quality (DockQ score) for all cases when binned by MSA depth levels (**Figure 5.22**). A slight trend was observed indicating that a greater MSA depth is associated with higher DockQ scores (higher model accuracies), suggesting that compared to a shallow MSA, predictions with a deeper MSA are more likely to be of higher accuracy. Thus, it is possible that increasing MSA depth, particularly for antibody-antigen complexes with very shallow MSAs, could lead to some improvement in overall modeling performance.

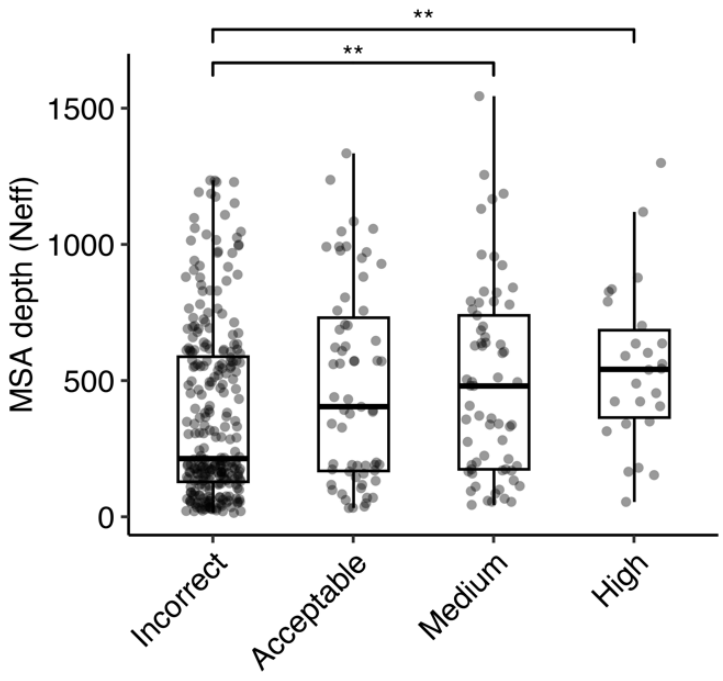

Figure 5.21. Comparison of MSA depth and modeling success. The distribution of MSA depth (number of effective sequences, N_{eff}), calculated using CD-HIT [240] with an identity cutoff of 80%, is shown for antibody-antigen complexes grouped by AlphaFold modeling accuracy. The modeling accuracy is defined as the highest CAPRI criteria prediction in the complex, considering all 25 predictions. Numbers of data points in incorrect, acceptable, medium and high categories are 272, 63, 65 and 26. Statistical significance values (Wilcoxon rank-sum test) were calculated between interface energy scores for sets of cases with incorrect versus medium and incorrect versus high CAPRI accuracy predictions, as noted at top (**$p \leq 0.01$).

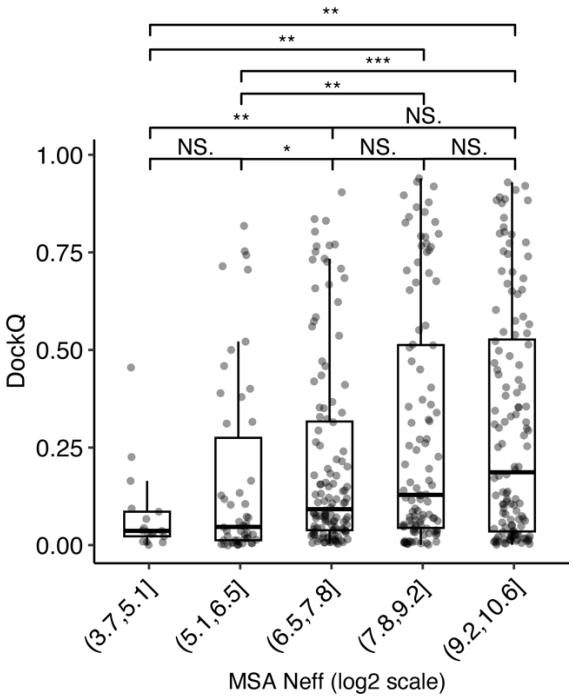

Figure 5.22. Distribution of DockQ scores grouped by ranges of AlphaFold MSA depth. The DockQ scores selected for each individual data points are the highest DockQ score of all 25 complex predictions generated by AlphaFold. Numbers of data points in (3.69,5.08], (5.08,6.46], (6.46,7.84], (7.84,9.21] and (9.21,10.6] Neff ranges are 17, 50, 124, 107, 128 respectively. Statistical significance values (Wilcoxon rank-sum test) were calculated between DockQ scores for sets of cases with varying ranges of MSA depth, as noted at top (NS.: $p > 0.05$, *$p \leq 0.005$, **$p \leq 0.001$, ***$p \leq 0.001$).

5.3.10 Modeling accuracy of AlphaFold v.2.3.0

Recently, an updated version of AlphaFold (v.2.3.0, hereafter denoted as v.2.3) was released, with modifications to the pipeline and deep learning model[275]. Compared with the previous version, this version was trained on PDB structures released until September 30, 2021, resulting in a 30% increase in training data. This version also increased the maximum number of recycles, from 3 recycles in v.2.2 to 20 recycles in v.2.3, with early stopping, and utilized larger interface regions (crops) and more chains during training. To benchmark its performance, we assembled a test set of 41 nonredundant antibody-antigen complexes released after the September 30, 2021 training date (**Table 5.1**). On a total of 39 cases for which models were successfully produced by

both versions of AlphaFold, v.2.3 generated medium or higher accuracy models as top-ranked predictions for 36% of the test cases, notably higher than the 23% generated by v.2.2 (**Figure 5.23**), with no significant difference in antibody CDR loop accuracy.

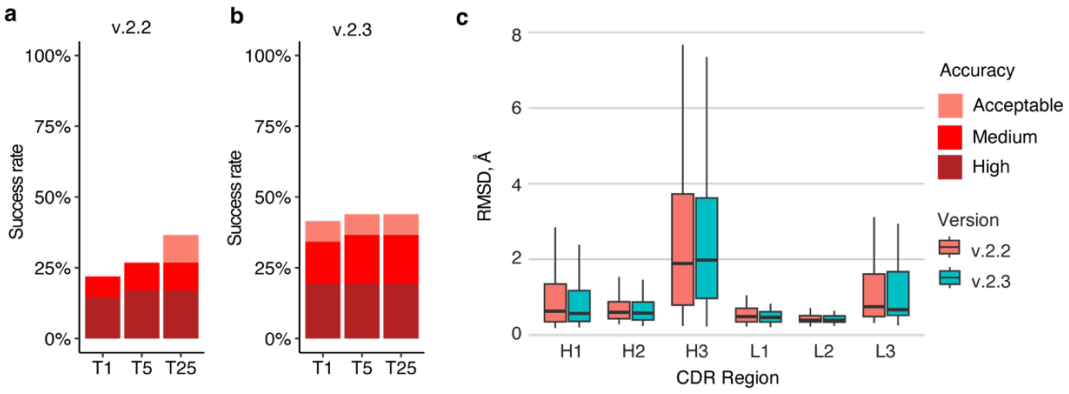

Figure 5.23. Antibody-antigen modeling success by AlphaFold v.2.3. Modeling success of **(a)** AlphaFold v.2.2 and **(b)** AlphaFold v.2.3 on 41 antibody-antigen complexes. The success rate was calculated based on the percentage of cases that had at least one model among their top N predictions that met a specified level of CAPRI accuracy. Bars were colored as per CAPRI accuracy criteria. **(c)** Distribution of the CDR loop prediction accuracy of AlphaFold v.2.2 (denoted by salmon color) vs v.2.3 (denoted by cyan color). CDR loop accuracy is defined as the change in RMSD of the CDR regions, when superposing the predicted antibody (in the antibody-antigen complex prediction) onto the experimentally resolved antibody (in the antibody-antigen complex) using the antibody framework region.

Compared to v.2.2, v.2.3 predictions were produced with a higher number of recycles, with 95% of v.2.3 predictions generated using more than three recycles. To investigate the impact of recycles, we reduced the maximum number of recycles in v.2.3 from 20 to 5 (**Figure 5.24a**) or 3 (**Figure 5.24b**). Compared to the default setting with a maximum number of recycles of 20, the antibody-antigen complex prediction success remained identical for generating near-native (medium or higher accuracy) top-ranked predictions, regardless of the recycle limit. This suggests that the observed difference in the number of recycles between v.2.3 and v.2.2 is not the main factor contributing to the increased success, and that the updated and expanded training of the deep learning model training may be responsible.

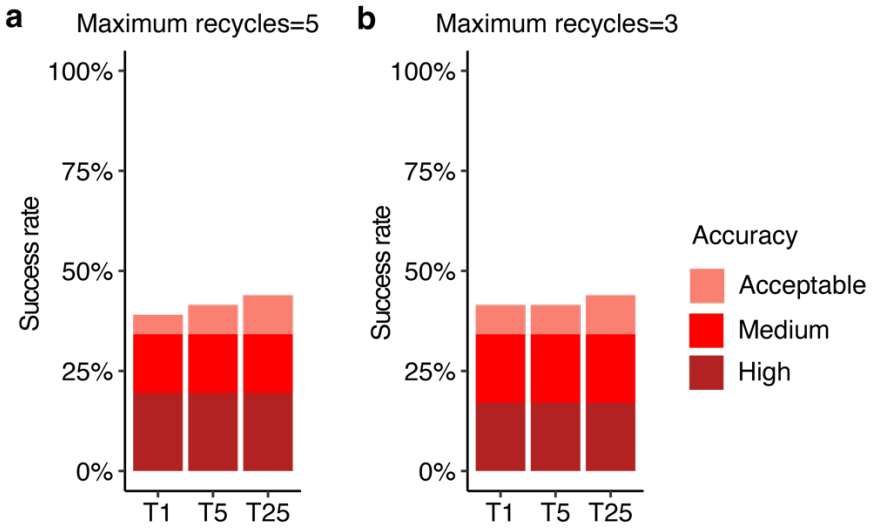

Figure 5.24. Antibody-antigen modeling success comparison of AlphaFold (v.2.3) using a varying number of maximum recycles. The maximum number of recycling iterations were set to **(a)** 5 and (**b)** 3, and tested on 41 antibody-antigen complexes. Templates released on or before September 30, 2021, were allowed during modeling. For each complex, 25 predictions were generated, and were ranked by AlphaFold model confidence score. Antibody-antigen predictions were evaluated for complex modeling accuracy using CAPRI criteria for high, medium and acceptable accuracy. The success rate was calculated based on the percentage of cases that had at least one model among their top N predictions that met a specified level of CAPRI accuracy. Bars were colored by CAPRI accuracy criteria.

Recently, Wallner demonstrated that by introducing stochastic perturbations through activating dropout during AlphaFold inference and employing extensive sampling, the modeling success of AlphaFold can be improved[274, 287]. Using this technique, named AFsample, Wallner group ranked among the top predictors at CASP15 competition. Anecdotally, A sample demonstrated enhanced accuracy specifically for nanobody antigen targets. In light of this, we applied the AFsample protocol to model our benchmarking set of antibody-antigen complexes to assess its performance on a broader dataset. On a total of 37 cases for which all models were successfully generated, AFsample genereated medium or higher accuracy top-ranked predictions for 51% of the test cases, which is significantly higher than 35% for AlphaFold v.2.3, and 24% AlphaFold v.2.2 (**Figure 5.25**). When top 25 predictions are considered, AFsample's medium or

higher success rate rose to 59%. The jump in success indicates potential for improvement in the default ranking method, which is based on model confidence. In summary, our findings indicate that the application of stochastic perturbation in the inference model, combined with extensive sampling, provides a significant advantage over the standard protocol. This approach resulted in enhanced modeling accuracy for antibody-antigen models, surpassing even the performance of the updated AlphaFold version, v.2.3.

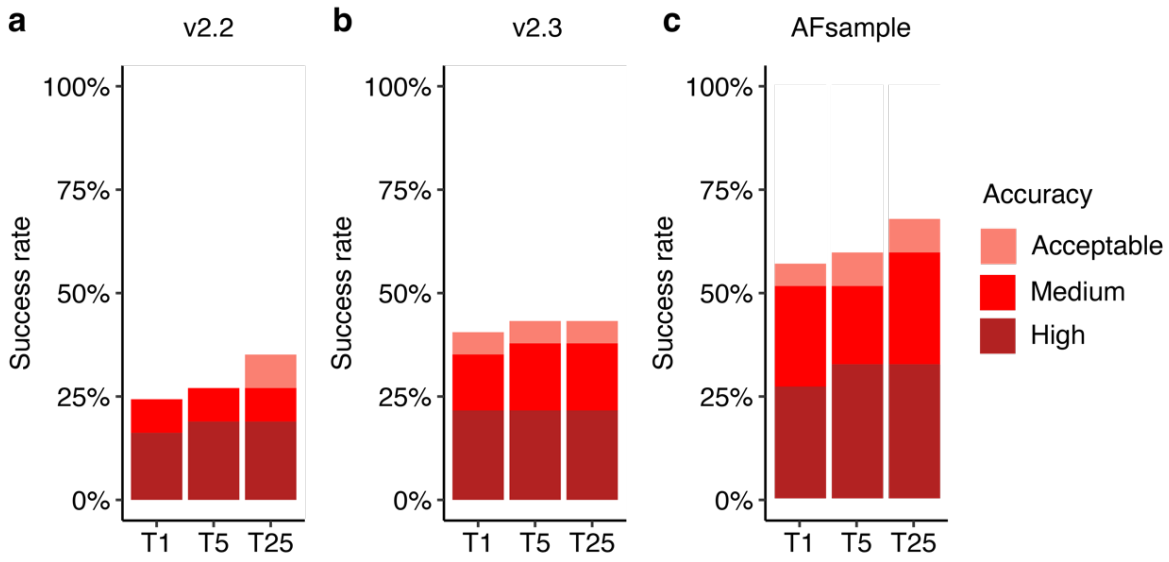

Figure 5.25. Antibody-antigen modeling success by AlphaFold v.2.2, v2.3 and AFsample. Modeling success of **(a)** AlphaFold v.2.2, **(b)** AlphaFold v.2.3 and **(c)** AFsample on 37 antibody-antigen complexes. The success rate was calculated based on the percentage of cases that had at least one model among their top N predictions that met a specified level of CAPRI accuracy. Bars are colored by CAPRI accuracy criteria.

5.4 Discussion

Using a set of over 400 nonredundant antibody-antigen complexes, we benchmarked and evaluated AlphaFold's ability to model antibody-antigen complexes. On this set, we observed a limited yet higher success in the prediction of antibody-antigen structures by AlphaFold,

compared to our previous benchmarking that used an older AlphaFold version accessed via ColabFold, and was based on a limited set of 100 antibody-antigen cases[272]. Analyses of factors that could influence the prediction outcome showed that AlphaFold was less able to accurately predict antibody-antigen structures with glycans at the antibody-antigen interface, which highlights AlphaFold's limitation in handling complexes with post-translational modifications. We also found that AlphaFold is more successful at modeling nanobody-antigen complexes and has difficulty predicting the structure of larger antibody-antigen complexes. An analysis of prediction accuracy at each recycling iteration, as well as the bound antibody-antigen template tests shows the importance of accurate subunit modeling for success in predicting the antibody-antigen complex. Relatedly, the ability to accurately predict CDRH3 loops is important for overall docking success.

Our benchmarking also shows that the latest version of AlphaFold (v.2.3) exhibits improved success in predicting antibody-antigen structures versus the previous AlphaFold version (v.2.2), likely due at least in part to the model training on an updated and expanded set of complex structures from the PDB [275]. It is possible that success can be improved further through additional optimization or other adaptations of the AlphaFold framework or model. Recent work has shown that higher structural diversity in AlphaFold predictions can be achieved by via enabling dropout [226, 277, 288]. When this technique is coupled with enhanced sampling, the quality of the generated predictions can be improved, as recently described [288], while additional strategies that have been explored for protein conformational sampling in AlphaFold, as noted in a recent review [289], could likewise be tested for antibody-antigen complexes. The observed strong association between antibody-antigen model accuracies and multiple confidence scores indicates that such increased sampling may be a worthwhile

approach. Another potential avenue for elevating the accuracy of AlphaFold predictions is demonstrated by the recent development of fully trainable AlphaFold implementations [276, 290, 291], which enable researchers to adapt and refine the model to specific datasets or domains of interest, opening up new possibilities for customization and optimization of the AlphaFold network.

Despite the lack of explicit co-evolutionary signal, our data show that the inclusion of diverse sequence information in MSAs is helpful for maintaining AlphaFold's modeling success of antibody-antigen complexes. As such, curation or optimization of MSAs could be another avenue for improving the accuracy of AlphaFold predictions. Previous work showed that AlphaFold prediction of protein-protein complexes can be augmented with improved MSA cross-chain pairing [237], while others have developed alternative MSA methods such as DeepMSA2 [292], which was part of a successful pipeline in a recent CASP/CAPRI complex structure prediction round. Recent work leveraging protein language models shows promise in constructing diversified and informative MSAs for enhancing accuracy in AlphaFold protein complex prediction [293], while it may be possible to replace or augment the MSA in AlphaFold with language model representations, potentially building on recent language models developed for antibodies [294, 295] or proteins in general [296].

Our results also demonstrate that accurate subunit prediction is associated with higher antibody-antigen complex prediction success. Recent work has shown improved accuracy in antibody prediction, particularly in the context of CDR loops, leveraging elements of AlphaFold architecture, especially the structure module, with modifications [138, 273]. Incorporating such advances into the prediction pipeline may enable the prediction of more accurate antibody-antigen complexes.

While it is possible or even likely that antibody-antigen modeling success may ultimately be improved in AlphaFold or related deep learning frameworks, the current success of AlphaFold and version 2.3 in particular, in conjunction with the observed confidence scoring accuracy, indicates that AlphaFold may potentially be of practical use to researchers in modeling this important and challenging class of complexes, and can complement or assist experimental structural determination methods.

Chapter 6: TCRmodel2: high resolution modeling of T cell receptor recognition using deep learning

6.1 Abstract

The cellular immune system, which is a critical component of human immunity, uses T cell receptors (TCRs) to recognize antigenic proteins in the form of peptides presented by major histocompatibility complex (MHC) proteins. Accurate definition of the structural basis of T cell receptors and their engagement of peptide-MHCs can provide major insights into normal and aberrant immunity, and can help guide the design of vaccines and immunotherapeutics. Given the limited amount of experimentally determined TCR-peptide-MHC structures, and the vast amount of TCRs within each individual as well as antigenic targets, accurate computational modeling approaches are needed. Here we report a major update to our web server, TCRmodel, which was originally developed to model unbound TCRs from sequence, to now model TCR-peptide-MHC complexes from sequence, utilizing several adaptations of AlphaFold. This method, named TCRmodel2, allows users to submit sequences through an easy-to-use interface, and shows similar or greater accuracy than AlphaFold and other methods to model TCR-peptide-MHC complexes based on benchmarking. It can generate models of complexes in 10 minutes, and output models are provided with confidence scores and an integrated molecular viewer. TCRmodel2 is available at: https://tcrmodel.ibbr.umd.edu.

6.2 Introduction

T cell immunity is a key component of immune protection from viruses and pathogens [22], such as SARS-CoV-2 [297]. Additionally, T cells and T cell receptors (TCRs) often play a role in autoimmunity [298, 299], and T cell receptors are increasingly being utilized as therapeutics in clinical and pre-clinical studies [300-302]. Understanding the structural basis of TCR recognition of peptide-MHC (pMHC) complexes can yield major mechanistic insights [51, 298, 299, 303] and provide the means to perform structure-based design of TCR specificity or affinity [304, 305]. While several hundred high resolution structures of TCR-pMHC complexes have been determined experimentally and are available in the Protein Data Bank (PDB) [164], this represents only a small fraction of TCRs (with millions of TCRs in each human repertoire [306]), and high throughput sequencing and screening technologies are enabling large sets of antigen-specific TCR sequences to be routinely identified [307]. The capability to perform accurate computational modeling of TCRs and TCR-pMHC complex structures would be highly useful, effectively bridging the gap between TCR sequence and 3D structural information. Such algorithms and models could be used for structure-based TCR design, or generalizable prediction of "unseen" TCR epitopes, which represents a major challenge in computational biology [308] that may potentially be addressed through structure-based methods [309, 310].

Several algorithms have been developed to perform modeling of unbound TCRs [130, 131, 141], and TCR-pMHC complexes [148-150] from sequence or unbound structures, primarily through template-based modeling combined with energy minimization. These approaches can often be unsuccessful due to limitations of templates coupled with the flexibility and diversity of TCR complementarity determining region (CDR) loops, and the wide range of TCR-pMHC docking orientations. Recently, deep learning-based structure prediction methods,

and particularly AlphaFold [111], have proven remarkably successful in predicting structures of monomeric proteins [111] and multimeric proteins [248] from sequence. While our own initial benchmarking of AlphaFold for modeling TCR-pMHC complexes showed limited success (2 out of 14 cases with near-native model accuracy) [272], its success in some cases showed that it is possible in principle to "fold-and-dock" TCR-pMHC complexes with deep learning, and a recent study demonstrated that AlphaFold can be fine-tuned and optimized to model TCR-pMHC complexes [309].

Here we describe the development of TCRmodel2, which is a major update of our previously released TCR modeling web server, TCRmodel [130]. While the previous version used template-based modeling and Rosetta [165] to generate unbound TCR structural models from sequence, TCRmodel2 uses AlphaFold to generate models of TCR-pMHC complexes, with several modifications to improve its speed and accuracy. TCRmodel2 can also generate models of unbound TCRs using the same AlphaFold-based framework. Based on benchmarking, TCRmodel2 generates models of TCR-pMHC complexes with greater accuracy than AlphaFold and previously developed TCR-pMHC modeling methods, and it is over 10 times faster than the default AlphaFold protocol. To enable progress in structural immunology, we provide TCRmodel2 to the community as a web server, with user-friendly features such as multiple sequence input options, interactive structural visualization, and model confidence scores.

6.3 Methods

6.3.1 TCRmodel2 algorithm

The TCRmodel2 modeling pipeline was generated through several modifications of the AlphaFold pipeline and database, as noted below. These changes were separately implemented in the AlphaFold v2.2 and v2.3 code (both downloaded from the AlphaFold Github repository), to enable comparative performance of the two AlphaFold models in the context of TCRmodel2.

6.3.1.1 Multiple Sequence Alignment database

Given that the AlphaFold feature selection stage includes sequence searches against large databases containing a large variety of proteins for each input chain, we reduced the databases to contain prospective TCR and MHC hits to speed up the multiple sequence alignment (MSA) building step. The AlphaFold pipeline was run with four representative human and murine TCR and MHC sequences, using the reduced database option. All hits from the TCR and MHC sequence searches against the Small BFD (Big Fantastic Database [111]), Uniref90, and Uniprot AlphaFold databases were combined into new database files in FASTA format, replacing the full database files. The resultant databases have sizes of 450 (Small BFD), 43638 (Uniref90), and 145999 (Uniprot) sequences, with the sequences collectively comprising 52096 TCR-related sequences and 137991 MHC-related sequences.

6.3.1.2 TCR templates

The AlphaFold template search, which utilizes an MSA built from the input sequence to search against PDB sequences to identify templates for each input chain [111], was found to identify non-TCR immunoglobulin structures as templates for TCR chains, versus TCR chain

structures with closer identity to the input sequence (e.g. a human TCR with the same germline gene). As this was due to the use of the MSA in the query against the PDB sequences (which is useful when distant orthologs may need to be detected as candidate templates), we modified AlphaFold to only utilize the input TCR sequences rather than MSAs to search against the PDB.

6.3.1.3 Peptide-MHC structural templates

AlphaFold was modified to utilize peptide-MHC complex structures as input templates by representing peptide and MHC as a single structure, with a chain break between peptide and MHC given by a residue index shift as used by ColabFold [226]. Template featurization of peptide-MHC templates from PDB structures was conducted using the AlphaFold modification described by Motmaen et al. [276]. To obtain peptide-MHC templates, peptide-MHC structures with resolution ≤ 3.5 Å were obtained from TCR3d. For Class II peptide-MHC structures, peptides were trimmed to include the 9-mer core sequence, and due to diversity of length and conformation of Class I peptides, additional unbound Class I peptide-MHC structures within the resolution cutoff were identified from the PDB and included in the set. To account for peptide structural heterogeneity while limiting redundancy, up to two structures with identical peptide-MHC sequences were retained. In total, our template set includes 884 Class I and 44 Class II peptide-MHC template structures. At the peptide-MHC template selection stage, structures containing peptides with the same length as query sequence are identified, and ranked first by MHC similarity score, and then by peptide similarity score if multiple identical MHCs are identified. Similarity scores for MHC and peptide sequences are calculated using BLOSUM62 in the Bio.Align Biopython package [311], with a large gap penalty (-100) for peptide alignments to ensure ungapped peptide alignments and template scoring.

6.3.1.4 Model scoring

For TCRmodel2 models, we provide AlphaFold-generated confidence scores, specifically the average predicted local difference distance test (pLDDT, corresponding to local structural accuracy), predicted TM-score (pTM, corresponding to overall topological accuracy), ipTM (pTM calculated for inter-chain interfaces) score, and model confidence, which is a linear combination of pTM and ipTM (0.2*pTM + 0.8*ipTM) [118]. Additionally, for TCR-pMHC complex models we modified AlphaFold to calculate TCR-pMHC ipTM, which corresponds to the interface pTM score calculated only across the interface between TCR and pMHC, versus the default ipTM which is calculated between all chains (e.g. peptide-MHC, TCR α and β chain). The TCR-pMHC ipTM score is calculated by modifying the chain IDs of the TCR-pMHC complex predictions in AlphaFold at the time of ipTM calculation to represent the TCR and pMHC each as one chain. TCRmodel2 also calculates average pLDDT score for each of the CDR3 loops, to enable users to specifically view the confidence levels of the CDR3 loops in the models.

6.3.2 Web server implementation

The TCRmodel2 web server interface was developed using Python3 and the Flask framework (https://flask.palletsprojects.com/). Users can choose to model a TCRpMHC complex (Class 1 or Class II) or an unbound TCR. In both run modes, the users can either provide the target sequences or build the sequences on-the-fly by selecting the target TCR or MHC genes. The latter option makes possible the modeling of sequences by the information collected from databases such as VDJdb [312] without the need to manually build TCR and MHC sequences. As an additional option, users can input TCR, peptide, and MHC sequences in a FASTA format

file. Input TCR sequences are preprocessed using the ANARCI tool (2) to identify and keep only its variable domain. In the case of MHC Class II modeling, peptide sequences are truncated to 9-mer core sequences using the NetMHCIIPan program (5). Reference TCR and MHC protein sequences were obtained from the IMGT database (3). Modeling jobs are submitted to a computing cluster with queue management and processed on a dedicated GPU node. Output models are post processed by renumbering TCR sequences following the Aho numbering scheme [168] using ANARCI, renaming chains according to TCR3d scheme and aligning models to the top ranked model by the pMHC chains as reference.

6.3.3 Benchmarking

6.3.3.1 Benchmark assembly

Experimentally determined TCR-pMHC complex structures used for benchmarking were obtained from the TCR3d database [47]. TCR-pMHC complex structures were selected as benchmark cases based on these criteria: 1) Release date after the selected cutoff date (April 30, 2018 or September 30, 2021) 2) Structure resolution of 3.25 Å or better, 3) No redundancy with any TCR-pMHC complex structure from on or before the respective cutoff date (April 30, 2018 or September 30, 2021), 4) No redundancy with other structures within the benchmark. Redundancy between a pair of complexes was defined as TCR Vα or Vβ sequence identity of ≥ 95%, or V domain sequence identity of ≥ 92%, with a complex of the same class (Class I or II). Additionally, complexes containing peptides with modified amino acids (e.g. citrullination, lipopeptide) were excluded from the benchmark sets.

6.3.3.2 Accuracy assessment

TCR-pMHC models were assessed using Critical Assessment of Predicted Interactions (CAPRI) criteria [313], which are based on a combination of fraction of native interface residue contacts present in the model (Fnat), interface backbone root-mean-square distance (RMSD) between model and native structure (I-RMSD), and ligand RMSD between model and native structure (L-RMSD). Those metrics were calculated by the DockQ program [282]. Additionally, we separately assessed the peptide-MHC interface in models using DockQ and CAPRI peptide docking criteria [223] to identify models with partially or fully displaced peptides (corresponding to CAPRI peptide Incorrect or Acceptable peptide-MHC accuracy). Assessment of unbound TCR models was performed through calculation of backbone RMSD between model and native CDR loop residues after superposition of framework residues using the ProFit program [279]. CDR loop residue ranges were based on TCR3d loop definitions.

6.3.3.3 Correlations and ROC AUC calculations

Pearson correlations and their p-values were calculated with ggpubr package in R (r-project.org), and ROC AUC calculations were performed using the pROC package [250] in R.

6.3.4 Other modeling servers and tools

All other modeling tools used for comparison were run with default parameters, using sequences or gene names as input, as required for the respective programs. TCRFlexDock was run using the published Rosetta-based pipeline [148], generating 1000 models per complex from individually modeled TCR and pMHC structures (generated by TCRmodel2 with a April 30, 2018 template date cutoff), as input. ImmuneScape [150] and TCRpMHCmodels [149] were run from the respective web server interfaces. TCRDock [309] was downloaded from Github and run

locally; based on its pipeline, three TCRDock models were generated per complex, and models were ranked by their PAE scores to select the top model. AlphaFold versions 2.2 and 2.3 were run locally with default parameters and databases for multimer protein prediction, except with five models generated per complex in order to compare with TCRmodel2, and a template date cutoff corresponding to the benchmark set being tested (April 30, 2018 or September 30, 2021).

6.4 Results

6.4.1 TCRmodel2 interface

6.4.1.1 Overview

TCRmodel2 allows users to submit TCR, peptide, and MHC sequences to model TCR-pMHC complex structures through its main server interface, and it is able to model Class I and Class II complex structures. As with the original TCRmodel interface [130], users can enter all sequences directly, or generate TCR and MHC sequences from sets of human and mouse genes. As noted in the Materials and Methods, the TCRmodel2 algorithm is based on an adaptation of AlphaFold2, with a focused database of TCR and MHC sequences to speed up MSA feature building, optimization of the TCR template selection, and utilization of peptide-MHC complex structures as templates to improve AlphaFold's peptide-MHC modeling accuracy. Users have the option of performing Amber relaxation of models in AlphaFold, which, as noted in the AlphaFold publication, can improve local geometries in some models (e.g. remove side chain clashes) but will not markedly affect overall model accuracies [111]. Currently TCRmodel2 supports models of TCR complexes with peptide-MHC, and not TCR complexes with MHC-like

molecules CD1 and MR1, due to the small molecule and lipid antigens presented by those molecules [314] which are not supported in AlphaFold.

6.4.1.2 Timing

The TCRmodel2 server TCR-pMHC modeling jobs take approximately 15 minutes on average, using a dedicated NVIDIA Titan RTX GPU and generating five ranked TCR-pMHC models. Modeling of unbound TCRs requires approximately 12 minutes to generate five models. Use of Amber relaxation for the models, which can remove clashes but does not impact overall model accuracy, takes approximately 1-2 minutes (included in the above times). In contrast, generation of five TCR-pMHC models using the standard AlphaFold2 pipeline on the same computer cluster takes approximately 5-7 hours, of which over 90% of the time is spent in the feature generation and MSA building stage.

6.4.2 TCRmodel2 modeling accuracy

6.4.2.1 Initial benchmarking

To benchmark the TCR-pMHC modeling accuracy of TCRmodel2, we assembled a set of nonredundant TCR-pMHC structures from TCR3d that were released after April 30, 2018, with the date cutoff selected to avoid overlap with the AlphaFold (v2.2) model training set. Nonredundancy and other criteria for benchmark case selection are detailed in the Materials and Methods. In total, we identified 48 test cases, including 32 Class I complexes and 16 Class II complexes (**Table 6.1**). Comparison of modeling accuracy of TCRmodel2 with the AlphaFold 2.2 model against AlphaFold2.2 (**Figure 6.1A**, **Table 6.1**) shows that TCRmodel2 has higher accuracy, achieving a Medium or High CAPRI accuracy model for over 50% of cases. For several cases, such as 6R0E, 6R2L, 6ULN, and 7L1D, TCRmodel2 outperformed AlphaFold 2.2

where the latter method improperly modeled the peptide in the interface (**Table 6.1**), indicating that the pMHC structure templates used by TCRmodel2 likely enabled improved accuracy. Both AlphaFold-based methods outperformed the previously developed template-based TCR-pMHC modeling methods ImmuneScape [150] (**Figure 6.1A**, **Table 6.1**) and TCRpMHCmodels [149] (which only generates Class I TCR-pMHC models; **Table 6.1**, **Figure 6.2**), as well as the TCR-pMHC docking algorithm, TCRFlexDock [148] (**Table 6.1**). Details regarding the CDR loop accuracies of the TCR-pMHC models and individual accuracy metrics are in **Table 6.2** and **Table 6.3**, respectively.

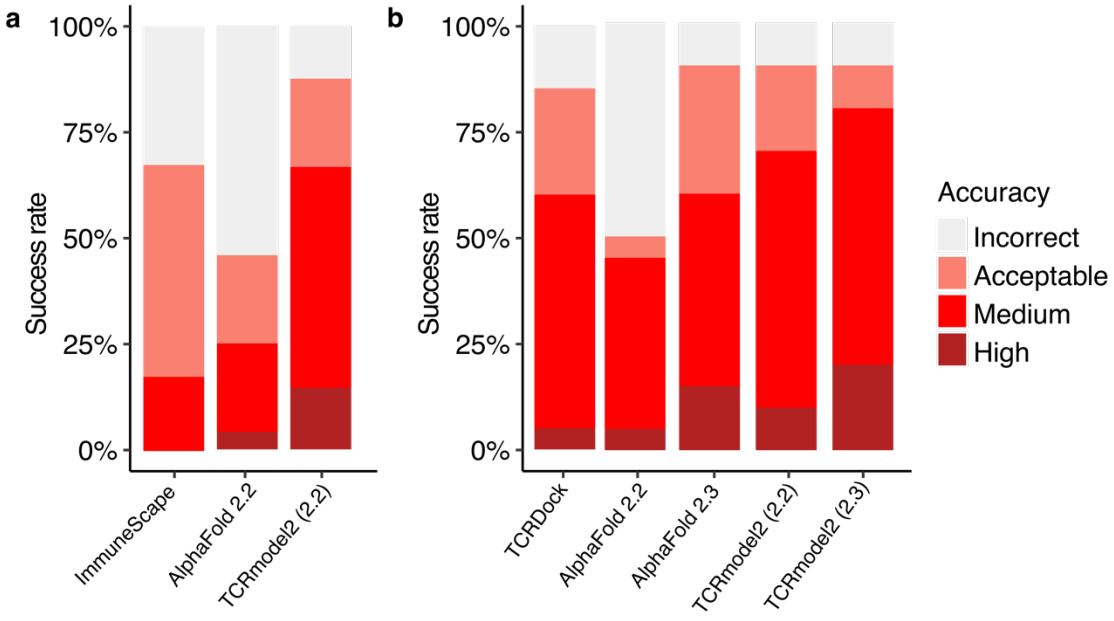

Figure 6.1. Success rate of TCRmodel2 and comparison with other modeling algorithms. (a) Modeling success comparison of AlphaFold 2.2, TCRmodel2 with the AlphaFold 2.2 model (2.2), and ImmuneScape on the initial set of 48 TCR-pMHC benchmarking cases. A template date cutoff of 2018-04-30 was applied. Due to an overlap with the modeling template, benchmarking of ImmuneScape and success rate calculation is for a subset of 47 cases. **(b)** Modeling success comparison of AlphaFold 2.2, AlphaFold 2.3, TCRmodel2 with the AlphaFold 2.2 model (2.2), TCRmodel2 with the AlphaFold 2.3 model (2.3), and TCRDock on a recently released set of 20 TCR-pMHC structures. Modeling success denotes the success rate of top-ranked prediction if multiple predictions were produced by the modeling algorithm. Three predictions were generated per case by TCRDock, and were ranked by AlphaFold predicted aligned error (PAE) score. For AlphaFold and TCRmodel runs, one prediction per deep learning

model was generated, resulting in 5 predictions per case, ranked by model confidence score. All models were assessed by CAPRI criteria of Incorrect, Acceptable, Medium and High accuracy.

Table 6.1. Modeling accuracy of AlphaFold, TCRmodel2, TCRpMHCmodels, ImmuneScape and TCRFlexDock on the benchmarking set of 48 TCR-pMHC complexes.

PDB	MHC class	Vα %[a]	Vβ %[a]	AlphaFold 2.2[b]		TCRmodel2 (v2.2)[b]		TCRpMHC models[c]	Immune Scape[d]	TCRFlex Dock[e]
				T1	T5	T1	T5			
6l9l	1	91	93	1	1	1	1	1	1	2
6mtm	1	86	92	2	2	2	2	1	2	2
6r2l	1	**87**	**82**	-1	-1	2	2	0	0	0
6rpa	1	**65**	**57**	-1	1	1	1	-	2	1
6rpb	1	87	91	0	0	0	1	2	0	1
6rsy	1	**84**	**89**	2	2	2	3	1	1	1
6tro	1	51	92	0	2	3	3	0	1	1
6uk4	1	**88**	**89**	-1	-1	1	1	1	2	1
6uln	1	88	90	-1	-1	2	2	1	1	2
6vrm	1	89	95	-1	-1	1	2	0	1	0
6vrn	1	86	91	-1	-1	0	0	0	0	0
7n1e	1	52	95	0	0	0	0	-	1	1
5nht	1	95	89	1	2	3	3	2	2	-
6rp9	1	88	90	1	1	2	2	0	0	0
6vmx	1	92	91	-1	2	3	3	1	1	0
6vm8	1	73	95	1	1	1	1	1	1	-
7n6e	1	91	89	1	1	1	1	1	1	1

6zkw	1	89	92	1	2	1	1	0	0	-
7dzm	1	**88**	**88**	2	2	0	0	0	1	-
7ndq	1	89	91	2	2	2	2	0	-	-
7l1d	1	**89**	**89**	-1	-1	2	2	-	1	-
7rrg	1	88	92	0	0	2	2	0	0	-
7na5	1	**62**	**54**	0	1	0	1	0	0	-
7ow5	1	83	90	-1	-1	2	2	1	1	-
7qpj	1	**88**	**88**	3	3	3	3	2	2	-
7phr	1	87	94	3	3	3	3	1	1	-
7nme	1	90	94	2	2	2	2	1	1	-
8d5q	1	66	93	2	2	2	2	1	0	-
8gvb	1	90	90	2	2	2	2	1	0	-
7rk7	1	**88**	**85**	-1	-1	1	1	-	0	-
7n2n	1	91	93	0	0	0	1	-	0	-
8gom	1	91	88	-1	0	1	1	0	1	-
6cql	2	**52**	**55**	-1	2	1	1	-	-	-
6dfx	2	89	93	1	1	2	2	-	0	2
6py2	2	92	58	1	1	2	2	-	0	1
6r0e	2	92	61	-1	-1	3	3	-	2	2
6u3n	2	92	76	-1	2	2	2	-	2	2
6xc9	2	45	93	2	2	2	2	-	1	1
6xco	2	89	94	2	2	3	3	-	1	1
6dfs	2	**67**	**38**	1	1	2	2	-	-	2
6dfw	2	66	93	-1	1	2	2	-	-	1
6px6	2	51	95	0	2	2	2	-	1	-
7rdv	2	64	92	-1	-1	2	2	-	-	-

7sg0	2	93	90	-1	2	2	2	-	-	-
7sg1	2	91	57	-1	-1	2	2	-	-	-
7z50	2	**67**	**59**	-1	-1	2	3	-	-	-
7t2c	2	92	92	2	2	2	2	-	1	-
7t2b	2	91	94	1	2	2	2	-	1	-

All predictions were assessed by CAPRI criteria for Incorrect (denoted by "0"), Acceptable (denoted by "1"), Medium (denoted by "2"), and High (denoted by "3") modeling accuracy. Peptide displacement from the MHC was assessed using CAPRI criteria, and denoted with "-1". Cases without predictions generated are denoted with "-".

[a]Maximum α chain variable domain (Vα) and β chain variable domain (Vβ) sequence identity with any TCR from known TCR-pMHC structures of the same class from on or before the 2018-04-30 date cutoff. Values in bold denote complexes with both chains below a 90% sequence identity level, indicative of structures with limited germline and/or CDR3 sequence match to previously described known structures for both chains.

[b]Five models were generated for each complex, and predictions were ranked by model confidence score. A template date cutoff of 2018-04-30 was applied to disable the use of recently released structures (overlapping with this benchmark set) as templates.

[c]TCRpMHCmodels predicts TCR-pMHC complexes with MHC Class I molecules. Some Class I cases resulted in job failures and are noted with "-".

[d]ImmuneScape job failures, or cases for which the appropriate MHC allele was not found, are noted as "-". Additionally, complex 6cql was excluded from ImmuneScape benchmarking, due to its overlap with the template used by that server.

[e]TCRFlexDock was run on a subset of cases from an early version of the benchmark. Unbound modeled TCR, and pMHC structures extracted from the top-ranked bound TCRmodel2 (2.2 AlphaFold model) predictions were used as input.

Table 6.2. CDR loop modeling accuracy of TCRmodel2 (2.2 AlphaFold model) and TCRmodel2 (2.3 AlphaFold model) for TCR-pMHC structures.

PDB	MHC class	TCRmodel2 (2.2)						TCRmodel2 (2.3)					
		1α	**2α**	**3α**	**1β**	**2β**	**3β**	**1α**	**2α**	**3α**	**1β**	**2β**	**3β**
6zkw	1	0.57	0.44	0.56	0.41	0.37	0.98	0.53	0.44	0.35	0.39	0.36	1.09
7na5	1	0.35	0.39	0.78	0.59	1.16	1.07	0.36	0.37	0.77	0.57	1.18	0.93
7ow5	1	2.16	2.16	3.82	0.60	0.48	6.67	2.23	0.74	4.37	0.66	0.45	5.86
7qpj	1	0.40	0.61	2.85	0.43	0.50	1.38	0.49	0.60	2.47	0.31	0.35	1.39
8gvb	1	0.77	0.40	0.77	0.61	0.81	1.51	0.74	0.46	0.69	0.59	1.03	1.37
7phr	1	0.77	0.71	1.44	0.76	1.38	2.04	0.76	0.61	1.65	0.42	1.31	2.00
7ndq	1	0.56	0.64	1.12	0.36	0.36	1.28	0.54	0.64	1.01	0.32	0.30	1.36
7rrg	1	2.16	0.74	3.82	0.37	0.66	0.47	0.50	0.43	1.07	0.36	0.65	0.77
7l1d	1	0.43	0.51	0.52	0.37	0.31	1.00	0.39	0.48	0.54	0.32	0.27	0.87
7dzm	1	0.59	0.85	5.03	0.84	0.42	4.51	0.59	0.88	3.82	0.87	0.43	3.87
7sg1	2	0.52	0.45	2.36	0.26	0.56	0.45	0.45	0.38	4.86	0.32	0.53	1.79
7rdv	2	1.11	0.30	2.45	0.33	0.56	1.21	1.15	0.35	2.79	0.29	0.51	1.05
7z50	2	0.57	0.59	2.08	0.55	0.89	2.31	0.40	0.50	1.91	0.52	0.90	3.18

Modeling accuracy is shown as backbone root-mean-square distance (RMSD) in Ångstroms of the respective CDR loop between model and native structures, after superposition of framework regions.

Table 6.3. TCR-pMHC modeling accuracy metrics and scores of TCRmodel2 (2.2 AlphaFold model) on the benchmarking set of 48 TCR-pMHC complexes.

PDB	TCR-pMHC complex accuracy						Model confi-dence	pLDDT	pTM	ipTM	TCR-pMHC ipTM[a]	Peptide-MHC accuracy[b]
	DockQ	Fnat	I-RMSD	L-RMSD	Fnonnat	CAPRI						
6l9l/ranked_0	0.50	0.60	2.38	6.48	0.43	Acceptable	0.85	94.27	0.88	0.85	0.82	High
6l9l/ranked_1	0.06	0.00	11.02	19.95	1.00	Incorrect	0.57	87.99	0.66	0.55	0.37	High
6l9l/ranked_2	0.14	0.02	5.83	11.59	0.98	Incorrect	0.56	89.45	0.65	0.53	0.33	High
6l9l/ranked_3	0.05	0.00	10.18	21.55	1.00	Incorrect	0.51	88.28	0.61	0.48	0.27	High
6l9l/ranked_4	0.06	0.00	10.58	19.23	1.00	Incorrect	0.49	88.81	0.60	0.46	0.24	High
6mtm/ranked_0	0.68	0.71	1.64	3.16	0.41	Medium	0.89	93.63	0.91	0.88	0.88	Medium
6mtm/ranked_1	0.66	0.67	1.65	3.33	0.44	Medium	0.88	92.85	0.90	0.88	0.87	Acceptable
6mtm/ranked_2	0.68	0.64	1.57	2.71	0.49	Medium	0.88	92.97	0.90	0.87	0.86	Medium
6mtm/ranked_3	0.76	0.67	1.14	1.48	0.34	Medium	0.88	93.33	0.90	0.87	0.86	Medium
6mtm/ranked_4	0.65	0.59	1.64	2.80	0.49	Medium	0.86	92.55	0.88	0.85	0.83	Medium
6r2l/ranked_0	0.75	0.67	1.07	2.37	0.22	Medium	0.76	91.27	0.80	0.75	0.67	High
6r2l/ranked_1	0.06	0.08	11.39	27.84	0.88	Incorrect	0.62	90.09	0.70	0.60	0.44	High
6r2l/ranked_2	0.25	0.20	3.43	10.40	0.64	Acceptable	0.50	88.89	0.60	0.48	0.24	High
6r2l/ranked_3	0.06	0.07	10.94	27.71	0.80	Incorrect	0.48	88.31	0.59	0.45	0.21	High
6r2l/ranked_4	0.01	0.00	25.32	69.32	1.00	Incorrect	0.46	86.32	0.57	0.43	0.16	High
6rpa/ranked_0	0.41	0.40	3.33	6.16	0.61	Acceptable	0.84	92.82	0.86	0.83	0.80	High

6rpa/ranked_1	0.04	0.04	14.55	27.78	0.94	Incorrect	0.63	89.99	0.71	0.61	0.46	High
6rpa/ranked_2	0.04	0.02	14.38	27.54	0.96	Incorrect	0.63	89.62	0.70	0.61	0.45	High
6rpa/ranked_3	0.09	0.11	10.52	21.17	0.86	Incorrect	0.62	90.37	0.70	0.60	0.46	High
6rpa/ranked_4	0.05	0.05	13.10	25.46	0.93	Incorrect	0.56	88.82	0.65	0.54	0.36	High
6rpb/ranked_0	0.22	0.29	4.98	13.77	0.59	Incorrect	0.66	92.83	0.73	0.65	0.50	High
6rpb/ranked_1	0.24	0.32	4.76	12.64	0.54	Incorrect	0.58	92.21	0.67	0.56	0.37	High
6rpb/ranked_2	0.28	0.24	3.55	9.25	0.53	Acceptable	0.57	91.91	0.66	0.55	0.34	Medium
6rpb/ranked_3	0.20	0.15	4.59	11.44	0.55	Incorrect	0.51	91.07	0.61	0.48	0.24	Medium
6rpb/ranked_4	0.23	0.21	4.24	11.42	0.36	Incorrect	0.51	91.02	0.61	0.48	0.25	Medium
6rsy/ranked_0	0.78	0.77	1.02	2.97	0.21	Medium	0.89	94.42	0.91	0.89	0.88	High
6rsy/ranked_1	0.77	0.74	1.04	3.05	0.17	Medium	0.88	94.02	0.90	0.88	0.86	High
6rsy/ranked_2	0.73	0.69	1.13	3.37	0.26	Medium	0.87	93.68	0.89	0.87	0.84	High
6rsy/ranked_3	0.81	0.78	0.97	2.24	0.12	High	0.86	93.37	0.88	0.85	0.83	High
6rsy/ranked_4	0.01	0.00	23.49	55.40	1.00	Incorrect	0.56	89.19	0.66	0.54	0.36	High
6tro/ranked_0	0.78	0.78	0.99	3.40	0.21	High	0.89	96.24	0.91	0.89	0.87	High
6tro/ranked_1	0.81	0.82	0.93	3.15	0.14	High	0.88	95.90	0.90	0.88	0.86	High
6tro/ranked_2	0.82	0.82	0.94	2.78	0.14	High	0.86	95.21	0.88	0.85	0.82	High
6tro/ranked_3	0.79	0.78	0.97	3.09	0.22	High	0.81	94.10	0.84	0.80	0.73	Medium
6tro/ranked_4	0.79	0.81	1.07	3.12	0.21	Medium	0.76	92.91	0.80	0.75	0.66	High
6uk4/ranked_0	0.46	0.55	2.58	7.44	0.51	Acceptable	0.86	94.37	0.89	0.85	0.85	High
6uk4/ranked_1	0.44	0.52	2.63	7.56	0.48	Acceptable	0.84	93.55	0.87	0.83	0.83	High
6uk4/ranked_2	0.44	0.55	2.65	7.80	0.47	Acceptable	0.81	93.07	0.86	0.80	0.80	High

6uk4/ranked_3	0.14	0.08	6.77	13.64	0.92	Incorrect	0.69	91.29	0.75	0.67	0.56	High
6uk4/ranked_4	0.20	0.14	5.39	10.73	0.86	Incorrect	0.59	90.00	0.67	0.57	0.44	High
6uln/ranked_0	0.54	0.49	2.01	4.61	0.68	Medium	0.85	92.94	0.88	0.85	0.82	Medium
6uln/ranked_1	0.50	0.45	2.31	5.04	0.70	Acceptable	0.85	93.32	0.87	0.84	0.81	High
6uln/ranked_2	0.46	0.47	2.66	5.93	0.71	Acceptable	0.84	92.65	0.87	0.84	0.80	Medium
6uln/ranked_3	0.52	0.49	2.18	5.05	0.67	Acceptable	0.83	92.76	0.86	0.83	0.79	High
6uln/ranked_4	0.06	0.10	14.75	29.29	0.87	Incorrect	0.57	87.71	0.66	0.55	0.36	High
6vrm/ranked_0	0.45	0.50	2.56	6.98	0.43	Acceptable	0.66	92.34	0.73	0.64	0.50	Medium
6vrm/ranked_1	0.52	0.52	2.13	5.65	0.48	Acceptable	0.63	89.91	0.71	0.61	0.47	High
6vrm/ranked_2	0.49	0.44	2.35	5.10	0.47	Acceptable	0.61	90.07	0.69	0.59	0.42	High
6vrm/ranked_3	0.42	0.43	2.37	7.55	0.41	Acceptable	0.58	91.05	0.67	0.56	0.40	Medium
6vrm/ranked_4	0.62	0.52	1.67	3.12	0.22	Medium	0.56	90.00	0.65	0.54	0.34	High
6vrn/ranked_0	0.17	0.19	6.02	14.59	0.83	Incorrect	0.80	92.87	0.84	0.79	0.74	High
6vrn/ranked_1	0.16	0.15	6.06	14.29	0.88	Incorrect	0.79	92.61	0.83	0.78	0.71	High
6vrn/ranked_2	0.16	0.13	5.79	13.33	0.88	Incorrect	0.67	91.41	0.74	0.65	0.50	High
6vrn/ranked_3	0.16	0.13	5.69	13.65	0.87	Incorrect	0.65	91.12	0.72	0.64	0.48	High
6vrn/ranked_4	0.15	0.11	6.02	14.17	0.89	Incorrect	0.56	89.75	0.65	0.53	0.33	High
7n1e/ranked_0	0.11	0.03	7.53	14.31	0.97	Incorrect	0.68	90.52	0.74	0.67	0.54	High
7n1e/ranked_1	0.01	0.00	22.98	59.15	1.00	Incorrect	0.54	88.96	0.64	0.52	0.32	High
7n1e/ranked_2	0.01	0.00	22.86	59.03	1.00	Incorrect	0.50	88.01	0.61	0.48	0.30	High
7n1e/ranked_3	0.01	0.00	22.98	58.78	1.00	Incorrect	0.48	87.91	0.59	0.46	0.24	High
7n1e/ranked_4	0.25	0.05	4.06	7.25	0.73	Incorrect	0.42	88.16	0.53	0.39	0.14	High

5nht/ranked_0	0.78	0.67	0.95	1.78	0.24	High	0.90	94.27	0.91	0.89	0.88	High
5nht/ranked_1	0.79	0.68	0.93	1.67	0.27	High	0.89	94.17	0.91	0.89	0.88	High
5nht/ranked_2	0.79	0.67	0.93	1.53	0.28	High	0.89	94.07	0.91	0.89	0.88	High
5nht/ranked_3	0.73	0.61	1.06	2.33	0.25	Medium	0.88	94.08	0.90	0.88	0.85	High
5nht/ranked_4	0.76	0.64	1.02	1.78	0.28	Medium	0.88	93.61	0.90	0.87	0.85	High
6rp9/ranked_0	0.61	0.54	1.69	3.65	0.54	Medium	0.82	94.02	0.85	0.81	0.75	Medium
6rp9/ranked_1	0.61	0.54	1.69	3.63	0.53	Medium	0.73	93.02	0.78	0.71	0.60	Medium
6rp9/ranked_2	0.09	0.10	9.17	20.98	0.91	Incorrect	0.61	91.02	0.69	0.59	0.41	Medium
6rp9/ranked_3	0.29	0.22	3.75	8.22	0.74	Acceptable	0.52	89.39	0.62	0.49	0.27	Medium
6rp9/ranked_4	0.10	0.07	8.17	17.98	0.90	Incorrect	0.48	88.74	0.59	0.45	0.20	Medium
6vmx/ranked_0	0.83	0.74	0.83	1.45	0.25	High	0.87	94.19	0.89	0.86	0.83	Medium
6vmx/ranked_1	0.82	0.72	0.84	1.57	0.23	High	0.86	93.85	0.88	0.85	0.82	Medium
6vmx/ranked_2	0.81	0.74	0.91	1.46	0.25	High	0.82	93.29	0.85	0.81	0.76	Medium
6vmx/ranked_3	0.81	0.72	0.84	1.88	0.26	High	0.79	92.57	0.83	0.78	0.71	Medium
6vmx/ranked_4	0.73	0.63	1.07	3.01	0.20	Medium	0.78	92.43	0.82	0.77	0.69	Medium
6vm8/ranked_0	0.27	0.25	3.59	9.99	0.64	Acceptable	0.85	93.94	0.88	0.84	0.81	High
6vm8/ranked_1	0.30	0.32	3.50	10.09	0.61	Acceptable	0.80	93.37	0.83	0.79	0.73	High
6vm8/ranked_2	0.27	0.29	3.67	10.73	0.59	Acceptable	0.79	93.39	0.83	0.78	0.72	High
6vm8/ranked_3	0.31	0.35	3.51	10.07	0.55	Acceptable	0.78	92.95	0.82	0.77	0.70	High
6vm8/ranked_4	0.01	0.00	20.01	70.42	0.00	Incorrect	0.37	87.62	0.51	0.34	0.08	High
7n6e/ranked_0	0.53	0.57	2.26	5.63	0.36	Acceptable	0.86	94.29	0.88	0.86	0.82	High
7n6e/ranked_1	0.40	0.47	3.09	7.95	0.41	Acceptable	0.82	93.40	0.85	0.81	0.77	High

7n6e/ranked_2	0.56	0.65	2.35	5.17	0.37	Acceptable	0.81	92.84	0.84	0.80	0.77	High
7n6e/ranked_3	0.50	0.52	2.29	5.60	0.41	Acceptable	0.76	92.53	0.80	0.74	0.68	High
7n6e/ranked_4	0.01	0.00	24.05	58.59	1.00	Incorrect	0.58	91.31	0.67	0.56	0.37	High
6zkw/ranked_0	0.53	0.69	2.65	6.21	0.32	Acceptable	0.85	94.83	0.88	0.84	0.79	Medium
6zkw/ranked_1	0.44	0.48	2.49	7.44	0.53	Acceptable	0.72	91.83	0.78	0.70	0.58	Medium
6zkw/ranked_2	0.51	0.52	2.14	5.97	0.44	Acceptable	0.67	90.73	0.74	0.65	0.50	Medium
6zkw/ranked_3	0.40	0.41	2.81	7.36	0.54	Acceptable	0.65	90.90	0.72	0.63	0.47	Medium
6zkw/ranked_4	0.40	0.28	2.35	6.37	0.45	Acceptable	0.50	88.63	0.61	0.47	0.25	Medium
7dzm/ranked_0	0.04	0.02	8.84	35.06	0.98	Incorrect	0.65	91.27	0.72	0.63	0.49	High
7dzm/ranked_1	0.06	0.00	7.92	20.09	1.00	Incorrect	0.61	88.14	0.69	0.59	0.43	High
7dzm/ranked_2	0.04	0.02	9.05	29.02	0.96	Incorrect	0.50	89.17	0.60	0.47	0.23	High
7dzm/ranked_3	0.01	0.00	23.41	67.91	1.00	Incorrect	0.38	86.52	0.51	0.35	0.08	High
7dzm/ranked_4	0.01	0.00	19.63	58.44	1.00	Incorrect	0.36	85.61	0.51	0.33	0.14	High
7ndq/ranked_0	0.55	0.53	2.23	4.35	0.55	Medium	0.87	92.68	0.89	0.86	0.84	Acceptable
7ndq/ranked_1	0.51	0.45	2.36	4.49	0.59	Medium	0.87	92.29	0.89	0.86	0.84	Acceptable
7ndq/ranked_2	0.50	0.45	2.56	4.50	0.60	Medium	0.87	92.68	0.89	0.86	0.84	Acceptable
7ndq/ranked_3	0.48	0.42	2.39	5.19	0.58	Acceptable	0.82	91.75	0.85	0.81	0.76	Acceptable
7ndq/ranked_4	0.49	0.47	2.36	5.50	0.56	Acceptable	0.77	91.14	0.81	0.76	0.68	Acceptable
7l1d/ranked_0	0.56	0.49	2.52	2.04	0.52	Medium	0.88	91.52	0.89	0.88	0.87	Medium
7l1d/ranked_1	0.41	0.35	3.84	4.79	0.64	Medium	0.86	91.06	0.88	0.85	0.83	High
7l1d/ranked_2	0.48	0.42	2.88	4.29	0.57	Medium	0.84	90.76	0.87	0.84	0.81	Medium
7l1d/ranked_3	0.51	0.43	3.37	2.50	0.58	Medium	0.81	89.86	0.84	0.81	0.76	Medium

7l1d/ranked_4	0.46	0.41	3.00	4.61	0.59	Medium	0.80	90.61	0.84	0.80	0.74	High
7rrg/ranked_0	0.55	0.43	1.87	3.86	0.58	Medium	0.79	91.94	0.83	0.78	0.72	Medium
7rrg/ranked_1	0.58	0.44	1.77	3.37	0.49	Medium	0.73	90.84	0.78	0.71	0.61	Medium
7rrg/ranked_2	0.54	0.41	1.99	3.62	0.57	Medium	0.69	90.61	0.75	0.68	0.55	Medium
7rrg/ranked_3	0.08	0.08	9.79	20.34	0.92	Incorrect	0.67	90.62	0.73	0.65	0.52	Medium
7rrg/ranked_4	0.11	0.11	8.48	17.85	0.86	Incorrect	0.57	88.85	0.66	0.55	0.36	Medium
7na5/ranked_0	0.03	0.02	15.39	30.73	0.98	Incorrect	0.82	92.88	0.85	0.81	0.76	High
7na5/ranked_1	0.05	0.06	15.38	30.78	0.95	Incorrect	0.80	92.45	0.84	0.79	0.73	High
7na5/ranked_2	0.34	0.33	3.43	7.81	0.68	Acceptable	0.80	92.73	0.83	0.79	0.74	High
7na5/ranked_3	0.05	0.06	15.32	30.66	0.95	Incorrect	0.79	92.30	0.83	0.78	0.71	High
7na5/ranked_4	0.05	0.06	17.33	31.80	0.93	Incorrect	0.63	89.32	0.70	0.61	0.47	High
7ow5/ranked_0	0.69	0.64	1.26	3.80	0.35	Medium	0.80	92.84	0.84	0.79	0.71	Acceptable
7ow5/ranked_1	0.10	0.02	8.59	14.69	0.98	Incorrect	0.69	91.37	0.75	0.67	0.54	Medium
7ow5/ranked_2	0.05	0.03	12.29	24.80	0.93	Incorrect	0.56	90.03	0.65	0.53	0.30	Medium
7ow5/ranked_3	0.05	0.03	12.24	24.49	0.92	Incorrect	0.55	90.00	0.65	0.53	0.30	Medium
7ow5/ranked_4	0.05	0.02	12.06	24.43	0.96	Incorrect	0.49	87.36	0.60	0.47	0.21	Medium
7qpj/ranked_0	0.87	0.75	0.60	1.09	0.18	High	0.89	95.31	0.90	0.88	0.86	High
7qpj/ranked_1	0.87	0.78	0.62	1.18	0.17	High	0.88	94.99	0.90	0.88	0.85	Medium
7qpj/ranked_2	0.88	0.77	0.56	0.99	0.19	High	0.88	95.01	0.90	0.88	0.85	High
7qpj/ranked_3	0.91	0.87	0.55	0.84	0.14	High	0.88	95.26	0.90	0.87	0.85	High
7qpj/ranked_4	0.33	0.44	3.86	10.16	0.47	Acceptable	0.82	94.07	0.85	0.81	0.75	Medium
7phr/ranked_0	0.87	0.77	0.64	1.21	0.13	High	0.88	93.71	0.90	0.88	0.87	High

7phr/ranked_1	0.85	0.77	0.72	1.22	0.13	High	0.88	93.85	0.90	0.88	0.87	High
7phr/ranked_2	0.86	0.79	0.75	1.41	0.17	High	0.88	93.81	0.90	0.87	0.87	High
7phr/ranked_3	0.85	0.76	0.73	1.35	0.16	High	0.88	93.78	0.90	0.87	0.87	High
7phr/ranked_4	0.85	0.76	0.70	1.26	0.11	High	0.85	92.99	0.87	0.84	0.83	High
7nme/ranked_0	0.62	0.71	1.88	4.63	0.33	Medium	0.90	94.91	0.92	0.90	0.89	Medium
7nme/ranked_1	0.62	0.64	1.78	4.31	0.35	Medium	0.88	94.28	0.90	0.87	0.85	High
7nme/ranked_2	0.62	0.72	1.89	4.76	0.32	Medium	0.88	94.38	0.90	0.87	0.85	High
7nme/ranked_3	0.64	0.72	1.78	4.47	0.36	Medium	0.88	94.51	0.90	0.87	0.85	High
7nme/ranked_4	0.04	0.05	17.46	32.89	0.88	Incorrect	0.50	89.20	0.60	0.48	0.23	High
8d5q/ranked_0	0.62	0.55	1.73	3.21	0.62	Medium	0.87	92.01	0.89	0.87	0.83	Acceptable
8d5q/ranked_1	0.61	0.51	1.75	3.11	0.62	Medium	0.87	92.00	0.89	0.86	0.83	Acceptable
8d5q/ranked_2	0.63	0.55	1.73	2.87	0.60	Medium	0.87	91.91	0.89	0.86	0.83	Acceptable
8d5q/ranked_3	0.62	0.53	1.75	2.81	0.57	Medium	0.86	91.84	0.89	0.86	0.82	Acceptable
8d5q/ranked_4	0.61	0.51	1.55	3.96	0.57	Medium	0.72	89.52	0.77	0.70	0.58	Medium
8gvb/ranked_0	0.53	0.53	2.33	4.63	0.45	Medium	0.89	94.15	0.91	0.89	0.87	High
8gvb/ranked_1	0.52	0.52	2.31	4.88	0.45	Medium	0.88	93.81	0.90	0.88	0.86	High
8gvb/ranked_2	0.53	0.52	2.25	4.71	0.47	Medium	0.88	93.40	0.90	0.87	0.85	High
8gvb/ranked_3	0.48	0.42	2.39	5.20	0.52	Acceptable	0.84	92.85	0.87	0.83	0.79	High
8gvb/ranked_4	0.01	0.00	25.35	69.60	1.00	Incorrect	0.45	88.34	0.57	0.42	0.14	High
7rk7/ranked_0	0.50	0.62	2.23	7.61	0.41	Acceptable	0.85	91.21	0.87	0.84	0.82	High
7rk7/ranked_1	0.47	0.56	2.29	7.67	0.42	Acceptable	0.82	90.84	0.85	0.81	0.77	High
7rk7/ranked_2	0.38	0.42	2.75	8.90	0.56	Acceptable	0.75	89.55	0.79	0.74	0.66	High

7rk7/ranked_3	0.45	0.55	2.30	8.24	0.46	Acceptable	0.75	90.18	0.79	0.74	0.66	High
7rk7/ranked_4	0.43	0.50	2.39	8.21	0.41	Acceptable	0.65	88.56	0.72	0.64	0.49	High
7n2n/ranked_0	0.05	0.09	16.25	33.23	0.92	Incorrect	0.84	94.68	0.87	0.84	0.79	High
7n2n/ranked_1	0.04	0.04	17.03	32.79	0.96	Incorrect	0.62	90.51	0.70	0.61	0.45	High
7n2n/ranked_2	0.27	0.21	3.71	9.50	0.73	Acceptable	0.61	91.32	0.69	0.59	0.40	High
7n2n/ranked_3	0.06	0.11	16.41	33.42	0.88	Incorrect	0.56	90.35	0.65	0.54	0.33	High
7n2n/ranked_4	0.36	0.33	3.12	7.59	0.56	Acceptable	0.56	90.48	0.65	0.53	0.34	High
8gom/ranked_0	0.53	0.59	2.54	5.26	0.39	Acceptable	0.87	93.35	0.89	0.87	0.85	High
8gom/ranked_1	0.49	0.53	2.49	5.94	0.40	Acceptable	0.86	93.48	0.88	0.86	0.84	High
8gom/ranked_2	0.08	0.06	6.16	22.65	0.91	Incorrect	0.69	91.41	0.75	0.67	0.54	High
8gom/ranked_3	0.24	0.17	3.06	11.17	0.46	Acceptable	0.50	89.64	0.60	0.47	0.24	High
8gom/ranked_4	0.15	0.06	4.03	13.63	0.83	Incorrect	0.39	87.65	0.52	0.36	0.18	High
6cql/ranked_0	0.35	0.28	2.29	9.26	0.59	Acceptable	0.59	89.71	0.68	0.57	0.38	Medium
6cql/ranked_1	0.29	0.20	2.55	9.91	0.68	Acceptable	0.54	87.82	0.63	0.51	0.28	Medium
6cql/ranked_2	0.27	0.20	2.86	10.30	0.69	Acceptable	0.51	86.49	0.61	0.48	0.25	Medium
6cql/ranked_3	0.17	0.05	4.97	10.65	0.88	Incorrect	0.50	87.02	0.60	0.47	0.22	Medium
6cql/ranked_4	0.19	0.10	3.92	11.56	0.71	Incorrect	0.47	87.27	0.58	0.44	0.19	Medium
6dfx/ranked_0	0.59	0.49	1.58	4.13	0.28	Medium	0.70	91.95	0.76	0.68	0.57	High
6dfx/ranked_1	0.15	0.06	4.94	13.20	0.82	Incorrect	0.50	88.30	0.60	0.47	0.22	High
6dfx/ranked_2	0.36	0.22	2.71	6.52	0.38	Acceptable	0.47	88.70	0.58	0.44	0.21	High
6dfx/ranked_3	0.05	0.00	13.38	21.42	1.00	Incorrect	0.47	88.20	0.57	0.44	0.18	High
6dfx/ranked_4	0.06	0.00	9.88	20.83	1.00	Incorrect	0.46	87.67	0.57	0.43	0.17	High

6py2/ranked_0	0.57	0.45	1.93	3.33	0.36	Medium	0.84	92.40	0.87	0.83	0.79	High
6py2/ranked_1	0.35	0.38	3.80	7.89	0.65	Acceptable	0.73	90.76	0.78	0.72	0.61	High
6py2/ranked_2	0.36	0.37	3.63	7.31	0.63	Acceptable	0.70	89.84	0.76	0.68	0.55	High
6py2/ranked_3	0.33	0.35	3.92	8.26	0.64	Acceptable	0.60	88.82	0.68	0.59	0.41	High
6py2/ranked_4	0.32	0.28	3.52	8.04	0.56	Acceptable	0.58	88.23	0.66	0.56	0.35	High
6r0e/ranked_0	0.86	0.85	0.84	1.59	0.35	High	0.89	95.71	0.91	0.89	0.87	High
6r0e/ranked_1	0.74	0.68	1.19	2.68	0.38	Medium	0.88	94.81	0.90	0.88	0.86	High
6r0e/ranked_2	0.83	0.85	0.93	2.39	0.29	High	0.88	95.56	0.90	0.88	0.85	High
6r0e/ranked_3	0.80	0.72	0.93	1.91	0.37	High	0.88	94.92	0.90	0.87	0.85	High
6r0e/ranked_4	0.72	0.68	1.23	3.32	0.33	Medium	0.83	93.45	0.86	0.82	0.78	High
6u3n/ranked_0	0.68	0.63	1.43	3.15	0.35	Medium	0.86	93.34	0.88	0.86	0.82	High
6u3n/ranked_1	0.62	0.56	1.53	4.26	0.41	Medium	0.86	93.34	0.88	0.86	0.82	High
6u3n/ranked_2	0.53	0.51	1.83	5.67	0.49	Medium	0.86	93.39	0.88	0.85	0.82	High
6u3n/ranked_3	0.60	0.56	1.67	4.48	0.45	Medium	0.86	93.63	0.88	0.85	0.82	High
6u3n/ranked_4	0.66	0.62	1.51	3.41	0.38	Medium	0.85	93.04	0.87	0.84	0.80	High
6xc9/ranked_0	0.58	0.56	1.67	5.16	0.40	Medium	0.87	93.65	0.89	0.86	0.84	High
6xc9/ranked_1	0.61	0.54	1.55	4.09	0.36	Medium	0.85	93.09	0.88	0.85	0.81	High
6xc9/ranked_2	0.67	0.65	1.42	3.98	0.35	Medium	0.85	93.09	0.88	0.85	0.82	High
6xc9/ranked_3	0.01	0.00	24.78	65.64	1.00	Incorrect	0.63	91.74	0.71	0.61	0.45	High
6xc9/ranked_4	0.31	0.26	2.79	9.29	0.46	Acceptable	0.45	86.07	0.56	0.42	0.25	High
6xco/ranked_0	0.88	0.89	0.67	2.52	0.16	High	0.90	95.57	0.92	0.90	0.88	High
6xco/ranked_1	0.91	0.91	0.58	2.23	0.12	High	0.90	95.44	0.91	0.90	0.88	High

6xco/ranked_2	0.91	0.94	0.64	2.33	0.17	High	0.90	95.02	0.91	0.90	0.88	High
6xco/ranked_3	0.03	0.02	17.45	30.62	0.97	Incorrect	0.54	89.00	0.63	0.51	0.29	High
6xco/ranked_4	0.05	0.00	14.86	22.59	1.00	Incorrect	0.46	89.22	0.58	0.44	0.20	High
6dfs/ranked_0	0.63	0.70	1.81	4.51	0.48	Medium	0.78	91.85	0.82	0.77	0.69	High
6dfs/ranked_1	0.56	0.49	1.89	4.39	0.55	Medium	0.65	90.39	0.72	0.64	0.47	High
6dfs/ranked_2	0.37	0.36	2.95	7.65	0.59	Acceptable	0.60	90.35	0.68	0.58	0.38	High
6dfs/ranked_3	0.52	0.45	1.94	5.15	0.33	Medium	0.51	87.78	0.62	0.49	0.24	High
6dfs/ranked_4	0.04	0.00	14.75	22.85	1.00	Incorrect	0.49	87.06	0.60	0.47	0.20	High
6dfw/ranked_0	0.66	0.63	1.42	3.97	0.31	Medium	0.79	91.92	0.82	0.78	0.69	High
6dfw/ranked_1	0.13	0.18	7.73	19.15	0.76	Incorrect	0.75	92.08	0.80	0.73	0.62	High
6dfw/ranked_2	0.11	0.09	6.78	16.95	0.87	Incorrect	0.69	91.30	0.76	0.68	0.53	High
6dfw/ranked_3	0.28	0.25	3.74	9.67	0.71	Acceptable	0.67	90.85	0.74	0.65	0.51	High
6dfw/ranked_4	0.18	0.18	5.26	13.16	0.76	Incorrect	0.64	90.60	0.71	0.62	0.46	High
6px6/ranked_0	0.60	0.61	1.85	4.25	0.36	Medium	0.87	93.77	0.89	0.86	0.84	High
6px6/ranked_1	0.79	0.75	1.05	2.34	0.25	Medium	0.86	93.74	0.88	0.85	0.81	High
6px6/ranked_2	0.75	0.73	1.21	2.69	0.27	Medium	0.85	93.74	0.88	0.85	0.81	High
6px6/ranked_3	0.12	0.01	6.88	12.74	0.94	Incorrect	0.49	87.82	0.61	0.47	0.20	High
6px6/ranked_4	0.15	0.07	6.06	12.05	0.77	Incorrect	0.48	86.40	0.60	0.45	0.20	High
7rdv/ranked_0	0.70	0.59	1.16	2.80	0.41	Medium	0.86	94.33	0.89	0.86	0.84	High
7rdv/ranked_1	0.72	0.61	1.15	2.79	0.40	Medium	0.86	94.51	0.89	0.86	0.84	High
7rdv/ranked_2	0.43	0.40	2.61	6.33	0.51	Acceptable	0.86	94.62	0.88	0.85	0.83	High
7rdv/ranked_3	0.50	0.50	2.29	5.77	0.33	Acceptable	0.86	94.35	0.88	0.85	0.83	High

7rdv/ranked_4	0.71	0.61	1.19	2.81	0.38	Medium	0.83	93.76	0.86	0.83	0.79	High
7sg0/ranked_0	0.64	0.59	1.49	3.89	0.46	Medium	0.88	93.84	0.90	0.88	0.86	High
7sg0/ranked_1	0.64	0.61	1.54	3.98	0.46	Medium	0.88	94.40	0.90	0.88	0.85	High
7sg0/ranked_2	0.64	0.64	1.56	4.22	0.45	Medium	0.88	93.87	0.90	0.87	0.85	High
7sg0/ranked_3	0.65	0.64	1.54	3.94	0.46	Medium	0.87	93.95	0.90	0.87	0.84	High
7sg0/ranked_4	0.61	0.61	1.67	4.55	0.45	Medium	0.86	93.63	0.88	0.85	0.82	High
7sg1/ranked_0	0.65	0.55	1.31	3.83	0.40	Medium	0.88	94.45	0.90	0.88	0.86	High
7sg1/ranked_1	0.36	0.38	3.39	8.16	0.42	Acceptable	0.85	93.38	0.88	0.85	0.82	High
7sg1/ranked_2	0.38	0.44	3.29	7.82	0.39	Acceptable	0.85	93.34	0.88	0.85	0.82	High
7sg1/ranked_3	0.36	0.38	3.30	8.00	0.45	Acceptable	0.84	92.88	0.87	0.83	0.80	High
7sg1/ranked_4	0.41	0.43	2.92	7.04	0.38	Acceptable	0.80	92.34	0.84	0.80	0.76	High
7z50/ranked_0	0.59	0.47	1.50	4.42	0.39	Medium	0.89	94.52	0.91	0.88	0.87	High
7z50/ranked_1	0.73	0.72	1.14	3.56	0.33	Medium	0.88	94.37	0.90	0.87	0.87	High
7z50/ranked_2	0.78	0.73	0.99	2.43	0.37	High	0.84	93.29	0.87	0.83	0.85	High
7z50/ranked_3	0.08	0.05	8.90	20.18	0.90	Incorrect	0.53	88.51	0.63	0.50	0.26	High
7z50/ranked_4	0.01	0.00	23.66	59.85	1.00	Incorrect	0.44	87.97	0.56	0.41	0.14	High
7t2c/ranked_0	0.76	0.69	1.07	2.19	0.29	Medium	0.89	92.54	0.90	0.88	0.86	Medium
7t2c/ranked_1	0.77	0.70	1.08	2.24	0.37	Medium	0.89	93.05	0.90	0.88	0.86	Medium
7t2c/ranked_2	0.77	0.71	1.04	2.36	0.30	Medium	0.88	92.31	0.90	0.87	0.85	High
7t2c/ranked_3	0.73	0.66	1.16	2.73	0.33	Medium	0.88	92.35	0.90	0.87	0.85	High
7t2c/ranked_4	0.54	0.62	2.00	6.25	0.51	Medium	0.85	91.90	0.88	0.85	0.84	Medium
7t2b/ranked_0	0.70	0.62	1.22	2.95	0.37	Medium	0.76	90.36	0.81	0.75	0.68	High

7t2b/ranked_1	0.54	0.52	2.14	4.82	0.48	Medium	0.76	89.75	0.80	0.75	0.66	High
7t2b/ranked_2	0.65	0.60	1.41	3.87	0.41	Medium	0.74	89.88	0.79	0.72	0.63	High
7t2b/ranked_3	0.59	0.53	1.76	4.06	0.45	Medium	0.72	89.56	0.77	0.70	0.59	High
7t2b/ranked_4	0.56	0.52	1.51	6.08	0.48	Medium	0.71	89.63	0.77	0.70	0.60	High

Docking quality metrics DockQ, Fnat, I-RMSD, L-RMSD, CAPRI, Fnonnat, and Peptide-MHC accuracy were calculated by the DockQ program.
[a]The ipTM score was modified to evaluate only the TCR-pMHC interface, instead of including all chains in the default ipTM score.
[b]We evaluated the peptide displacement in the MHC using DockQ with the “capri_peptide” setting, and report the CAPRI accuracy.

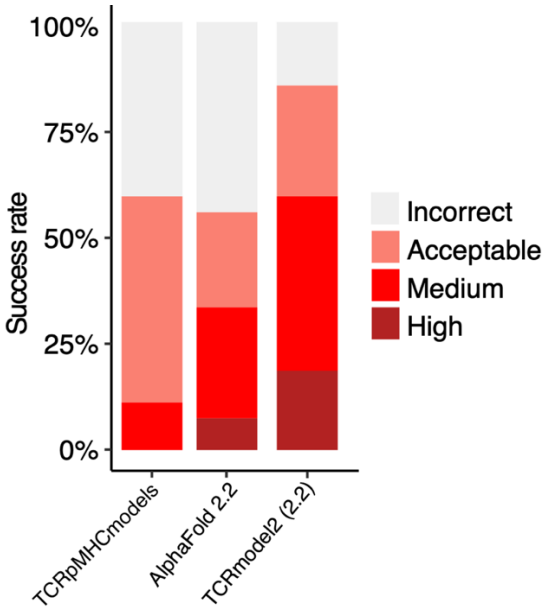

Figure 6.2. Modeling success comparison of TCRpMHCmodels, AlphaFold 2.2, and TCRmodel2 (with the AlphaFold 2.2 model), on the subset of 27 Class I TCR-pMHC benchmarking cases that were modeled by TCRpMHCmodels (from **Table S1**). For AlphaFold and TCRmodel2, 5 predictions were generated per case, and the top-ranked model was used for success calculation. All models were assessed by CAPRI criteria of Incorrect, Acceptable, Medium and High accuracy.

6.4.2.2 Benchmarking updated model

After the recent release of a new AlphaFold model and algorithm (v2.3), which includes an updated training set of structures (up to September 30, 2021) and additional recycling iterations during modeling [275], we implemented TCRmodel2 with the AlphaFold 2.3 model and pipeline to test whether it would lead to accuracy improvement over TCRmodel2 with the AlphaFold 2.2 model. This was benchmarked using 20 TCR-pMHC test cases with release dates after September 2021 (to ensure no overlap with the AlphaFold 2.3 training set), which is a subset of the original benchmark set (14 Class I complexes, 6 Class II complexes; **Table 6.4**). Modeling performance of TCRmodel2.2, TCRmodel2.3, AlphaFold2.2, and AlphaFold2.3 was assessed on this recently released benchmark set, along with TCRdock, which is an AlphaFold-based algorithm to model TCR-pMHC complexes that uses a fine-tuned TCR-pMHC model and

TCR-pMHC complex templates [309] (**Figure 6.1B**, **Table 6.2**). For this set, the AlphaFold and TCRmodel2 methods were permitted to use TCR and MHC structural templates from on or before the September 2021 date cutoff, versus the April 2018 template date cutoff used for the larger benchmark set. Based on this comparison, the AlphaFold 2.3 model and pipeline led to improved performance, with AlphaFold 2.3 outperforming AlphaFold 2.2, and TCRmodel2 with the AlphaFold 2.3 model outperformed the previous TCRmodel2 implementation (with the AlphaFold 2.2 model). TCRmodel2 (AlphaFold 2.3 model) achieved 20% success for High accuracy near-native models, and showed superior modeling accuracy on the benchmark versus AlphaFold 2.3 and TCRDock. One case for which TCRmodel2 and AlphaFold 2.3 outperformed TCRDock is 7RRG (**Table 6.4**); as that complex has an unusual TCR docking orientation (74□ TCR-pMHC crossing angle, according to TCR3d [47]), it may be more amenable to approaches such as TCRmodel2 and AlphaFold that do not utilize TCR-pMHC structural templates for TCR-pMHC orientation, versus TCRDock which uses TCR-pMHC orientations from experimentally determined complex structures as templates. Given its superior modeling performance, TCRmodel2 with the AlphaFold 2.3 model was selected for use in the TCRmodel2 server.

Table 6.4. Modeling success of AlphaFold, TCRmodel2, and TCRDock on a set of 20 recently released TCR-pMHC complexes.

PDB	MHC class	Vα %[a]	Vβ %[a]	AlphaFold 2.2[b]		AlphaFold 2.3[b]		TCRmodel2 (2.2)[b]		TCRmodel2 (2.3)[b]		TCRDock[c]	
				T1	T5	T1	T5	T1	T5	T1	T5	T1	T3
6zkw	1	89	92	0	2	1	2	1	3	3	3	1	1
7dzm	1	**88**	**88**	2	2	1	1	1	1	1	1	2	2
7ndq	1	89	91	2	2	2	2	2	2	2	2	0	1
7l1d	1	**89**	**89**	-1	-1	2	2	2	2	2	2	2	2
7rrg	1	88	92	0	1	2	2	0	2	2	2	0	1
7na5	1	**62**	**54**	0	1	1	1	1	1	0	0	1	1
7ow5	1	83	90	-1	-1	3	3	2	2	2	2	2	2
7qpj	1	**88**	**88**	3	3	2	3	3	3	3	3	2	2
7phr	1	87	94	2	3	3	3	3	3	3	3	2	2
7rk7	1	**88**	**85**	-1	-1	1	1	1	1	1	1	0	1
8d5q	1	66	93	2	2	2	2	2	2	2	2	2	2
8gvb	1	90	90	2	2	2	2	2	2	2	2	2	2
7n2n	1	91	93	-1	0	1	1	0	1	0	1	1	1

8gom	1	91	88	1	1	2	2	2	2	2	2	2	2
7rdv	2	64	92	-1	-1	3	3	2	2	2	3	2	2
7sg0	2	93	90	2	2	-1	2	2	2	2	2	1	1
7sg1	2	91	57	-1	2	1	2	2	2	2	2	2	2
7z50	2	**67**	**59**	-1	-1	-1	-1	2	3	3	3	3	3
7t2c	2	92	92	2	2	2	2	2	2	2	2	2	2
7t2b	2	91	94	2	2	2	2	2	2	2	2	1	1

All predictions were assessed by CAPRI criteria for Incorrect (denoted by "0"), Acceptable (denoted by "1"), Medium (denoted by "2"), and High (denoted by "3") modeling accuracy. Peptide displacement from the MHC was assessed using CAPRI criteria, and denoted with "-1". Cases without predictions generated were denoted with "-".

[a]Maximum α chain variable domain (Vα) and β chain variable domain (Vβ) sequence identity with any TCR from known TCR-pMHC structures of the same class from on or before the 2021-09-30 date cutoff. Values in bold denote complexes with both chains below a 90% sequence identity level, indicative of structures with limited germline and/or CDR3 sequence match to previously described known structures for both chains.

[b]Five predictions per case were generated, and predictions were ranked by model confidence score. A template date cutoff of 2021-09-30 was applied to disable the use of recently released structures (overlapping with the benchmark set) as templates. T1 indicates accuracy for top model, and T5 indicates best accuracy among all five models.

[c]TCRDock generated three predictions per case, and predictions were ranked by PAE score. T1 indicates accuracy for top model, and T3 indicates best accuracy among all three models.

6.4.2.3 Unbound TCRs

We also benchmarked the use of TCRmodel2 to model individual TCR structures (without pMHC), in comparison with TCRmodel (which uses structural templates to generate models) and AlphaFold (**Figure 6.3**, **Table 6.5**). We found that TCRmodel2 showed commensurate accuracy with AlphaFold (v2.2 and v2.3), while both AlphaFold and TCRmodel2 showed superior performance to TCRmodel, particularly for CDR3 loops which are more challenging to model (as observed during the initial TCRmodel benchmarking [130]) yet critical for peptide recognition. This demonstrates that deep learning-based approaches can overcome CDR3 loop modeling challenges faced by approaches that are fully or mostly reliant on structural templates, including CDR3 loop structural diversity, limited structural templates available, and nontrivial relationships between loop sequences and structures (for accurate template identification).

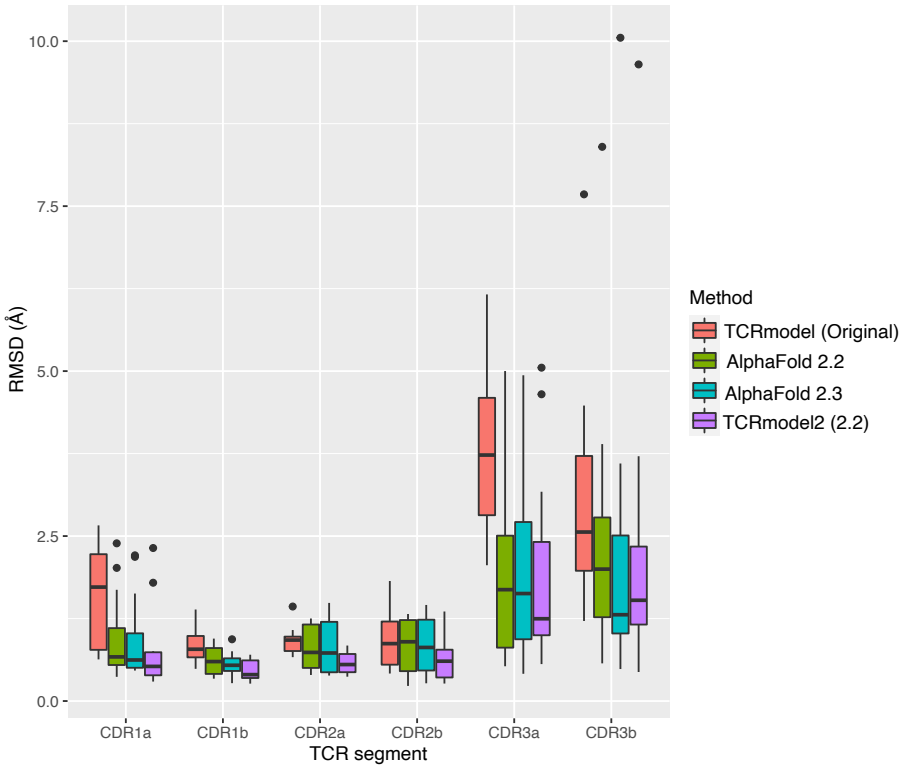

Figure 6.3. CDR loop modeling accuracy for unbound TCR models. Backbone root-mean-square distances (RMSDs) between experimentally determined structure and model were calculated for each CDR, after superposition of framework region of the variable domain. Test cases correspond to 13 recently released nonredundant TCR structures which are a subset of our recent TCR-pMHC benchmark cases for which a model was successfully generated by the original TCRmodel (cases and individual RMSDs are shown in **Table 6.3**). Methods compared are TCRmodel (original), AlphaFold 2.2, AlphaFold 2.3, and TCRmodel2 (with AlphaFold 2.3 model).

Table 6.5. Receiver operating characteristic area under the curve (AUC) values for CAPRI accuracy classes as a function of model confidence metrics for TCRmodel2 predictions on the benchmarking set of 48 cases.

Score	Incorrect vs. Acceptable, Medium, High	Incorrect, Acceptable vs. Medium, High	Incorrect vs. Medium, High
Model confidence	0.91	0.91	0.97
pTM	0.91	0.91	0.97
ipTM	0.91	0.91	0.97
TCR-pMHC ipTM[a]	0.91	0.91	0.97

AUC values were calculated using the pROC package in R [250].
[a]ipTM score considering only the TCR-pMHC interface, rather than the interfaces between all chains as used by the default ipTM score.

6.4.3 Model confidence scoring

Given that AlphaFold outputs model confidence estimates that are generally correlated with model accuracy [111, 118], we tested the use of AlphaFold confidence estimates in discriminating accurate versus inaccurate TCR-pMHC models in TCRmodel2 (**Table 6.6**). To maximize the amount of data for this comparison, the models from TCRmodel2 for the larger set of 48 cases (using TCRmodel2 with AlphaFold 2.2 model), and five models per case, were considered. Based on the receiver operating characteristic (ROC) area under the curve (AUC) values reported in **Table 6.6**, the overall model confidence score, which is a combination of ipTM and pTM scores, was found to provide very good discrimination of Medium and High models versus Incorrect (AUC=0.97). All of the other confidence metrics tested showed similar

AUC values, thus we focused on the model confidence score for further analysis. When comparing the model confidence score versus model accuracy (**Figure 6.4**), a relatively high correlation was observed between model confidence and the model accuracy (denoted by DockQ score) ($r = 0.75$; $p<0.001$). Model confidence scores also showed significant correlations with individual accuracy metrics Fnat, L-RMSD, and I-RMSD (**Figure 6.5**). Based on analysis of model accuracy discrimination using this score and our benchmark, we have determined model confidence score cutoffs of 0.85 and 0.49 for denoting likely accurate models (≥ 0.85) or likely inaccurate models (≤ 0.49) (shown as dashed lines in **Figure 6.4**); these cutoffs can be referred to by TCRmodel2 users to gauge the presence of a likely accurate model in the set of five produced by TCRmodel2. As also used by AlphaFold2, the model confidence score is used by TCRmodel2 to rank the five models for each TCR-pMHC complex.

Table 6.6. CDR loop modeling accuracy of TCRmodel, AlphaFold, and TCRmodel2 (2.3 AlphaFold model) for unbound TCR structures.

PDB	MHC	TCRmodel						AlphaFold 2.2						AlphaFold 2.3						TCRmodel2 (2.3)					
	class	**1α**	**2α**	**3α**	**1β**	**2β**	**3β**	**1α**	**2α**	**3α**	**1β**	**2β**	**3β**	**1α**	**2α**	**3α**	**1β**	**2β**	**3β**	**1α**	**2α**	**3α**	**1β**	**2β**	**3β**
6zkw	1	0.88	0.77	2.82	0.69	0.87	3.84	0.69	0.44	1.70	0.53	0.44	3.60	0.54	0.50	1.69	0.47	0.41	3.89	0.39	0.40	2.40	0.40	0.35	3.71
7na5	1	1.81	0.98	2.06	0.79	1.22	4.48	0.74	0.39	0.65	0.65	0.81	1.38	0.67	0.40	0.55	0.60	0.96	1.33	0.76	0.37	0.69	0.66	0.76	1.53
7ow5	1	0.78	0.99	3.73	0.64	0.55	2.96	0.52	1.49	1.29	0.46	0.46	1.22	0.55	1.17	1.17	0.39	0.41	1.27	0.32	0.61	1.08	0.32	0.36	1.40
7qpj	1	0.76	0.74	2.82	0.55	0.42	2.56	0.47	0.42	0.41	0.46	0.38	1.02	0.37	0.41	0.53	0.41	0.45	0.57	0.42	0.48	0.56	0.31	0.33	0.85
8gvb	1	2.59	0.96	2.78	1.25	1.41	3.23	2.19	0.86	2.71	0.94	1.46	0.94	2.02	0.87	2.44	0.95	1.32	1.37	1.79	0.84	2.02	0.60	0.60	1.17
7phr	1	2.66	1.08	3.06	0.67	0.81	1.21	0.50	0.55	0.69	0.63	0.97	1.12	0.52	0.55	0.78	0.63	0.90	1.17	0.53	0.51	0.71	0.38	0.66	1.16
7ndq	1	1.81	0.93	4.21	1.27	1.21	2.19	0.62	1.20	0.94	0.75	1.33	2.51	0.70	1.25	1.57	0.88	1.23	2.78	0.64	0.67	3.17	0.35	1.36	2.95
7rrg	1	0.97	0.67	4.60	0.66	0.58	3.71	1.03	1.25	4.85	0.45	0.47	3.09	1.11	1.16	4.44	0.40	0.52	3.33	0.62	0.44	4.65	0.35	0.57	1.86
7l1d	1	2.23	0.91	6.16	0.87	0.49	7.68	2.21	0.63	4.94	0.54	0.73	10.05	2.39	0.69	5.00	0.80	0.66	8.40	2.32	0.73	5.05	0.62	0.57	9.65
7dzm	1	2.28	0.92	5.12	0.80	1.03	1.52	0.50	0.41	4.03	0.74	1.23	1.31	1.69	0.41	2.51	0.75	1.27	1.00	0.30	0.38	1.00	0.70	1.17	1.13
7sg1	2	1.73	0.68	5.34	1.39	1.82	1.97	1.63	1.25	1.63	0.62	0.82	0.62	0.89	1.19	0.81	0.88	1.10	2.00	0.74	0.55	1.02	0.62	0.78	1.78
7rdv	2	0.63	1.43	4.50	0.49	0.48	1.31	0.46	0.73	1.25	0.27	0.27	0.49	0.60	0.86	1.94	0.34	0.23	2.57	0.39	0.71	1.25	0.26	0.27	0.44
7z50	2	0.69	0.85	2.73	0.99	1.04	2.39	0.61	0.73	2.70	0.46	1.28	2.44	0.64	0.74	2.74	0.43	1.27	2.54	0.51	0.75	2.41	0.40	1.03	2.34

Modeling accuracy is shown as backbone root-mean-square distance (RMSD) in Ångstroms of the respective CDR loop between model and native structures, after superposition of framework regions.

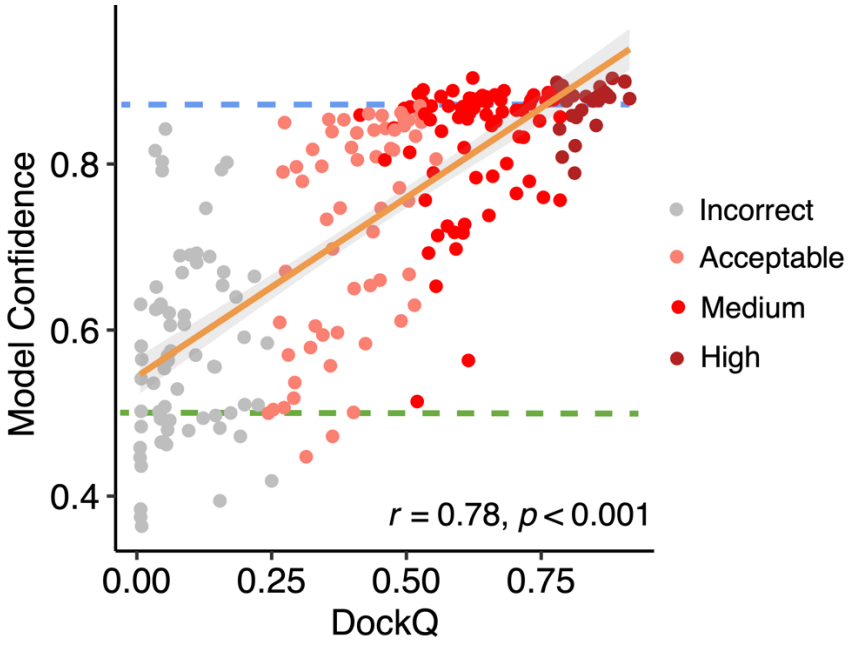

Figure 6.4. Comparison of model confidence score and model accuracy. Model confidence of all 5 models for 48 cases generated by TCRmodel2 are shown in comparison with model accuracy with respect to the experimentally determined structure (DockQ score [282]), with each model represented as a point and colored by CAPRI accuracy level. Pearson's correlation coefficient and the associated p-value are noted on the lower right corner, and the orange line represents the linear fit (with 95% confidence area in gray). The dashed blue line indicates a suggested model confidence cutoff (model confidence = 0.85) for identification of high accuracy predictions, based on maximization of sensitivity and specificity for discriminating Incorrect and Acceptable vs. Medium and High CAPRI accuracy models. The dashed green line indicates the lower bound of the model confidence score for models with Acceptable or higher accuracy (model confidence = 0.49).

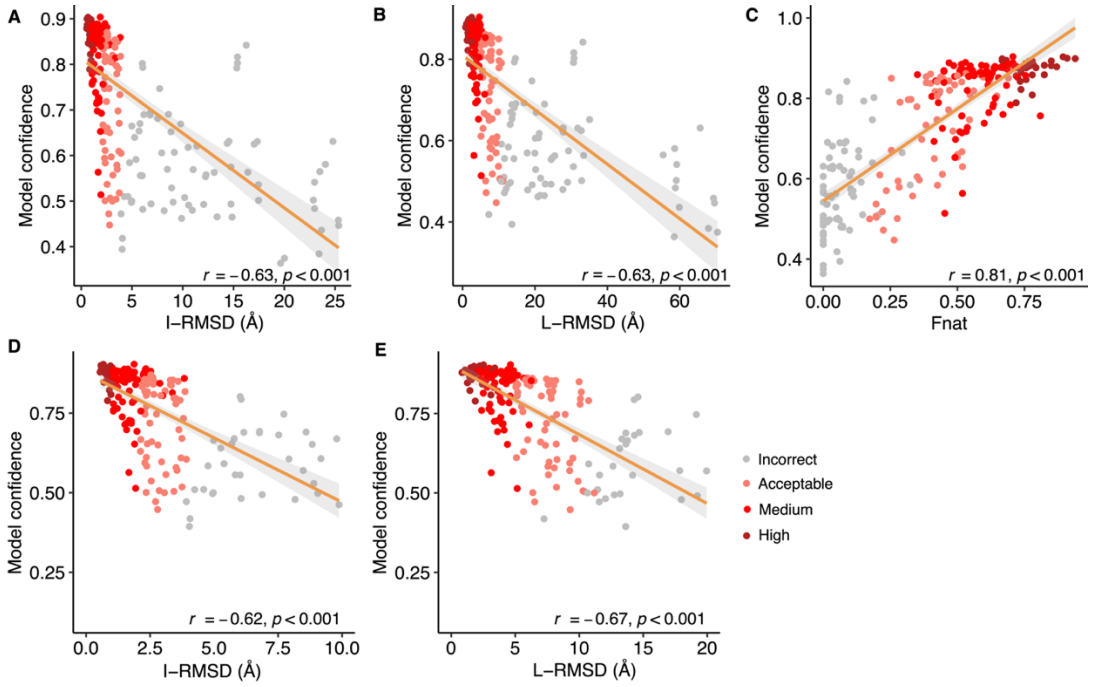

Figure 6.5. Comparison of model confidence score and model accuracy criteria. Model confidence of all 5 models for 48 cases generated by TCRmodel2 are shown in comparison with model accuracy with respect to the experimentally determined structure, with each model represented as a point and colored by CAPRI accuracy level. Pearson's correlation coefficient and the associated p-value are noted on the lower right corner, and the orange line represents the linear fit (with 95% confidence area in gray). Comparisons of model confidence score versus **(A)** I-RMSD, **(B)** L-RMSD, and **(C)** Fnat are shown. Additionally, comparisons of model confidence score versus **(D)** I-RMSD with values less than or equal to 10 Å, and **(E)** L-RMSD values less than or equal to 20 Å are shown.

To further assess expected model confidence for structurally uncharacterized TCR-pMHC complexes, we used TCRmodel2 to model additional Class I and II complexes obtained from the VDJdb database [312]. The distributions of model confidence scores for Class I (N=414) and Class II (N=47) complexes is shown in **Figure 6.6**, indicating that many complexes have top-ranked models in the high confidence range (≥ 0.85 model confidence score), with 30% of the Class I complexes and 47% of the Class II complexes at that level. With a slightly more permissive threshold (model confidence ≥ 0.75), TCRmodel2 generated top-ranked models for 77% and 89% of the Class I and Class II complexes, respectively.

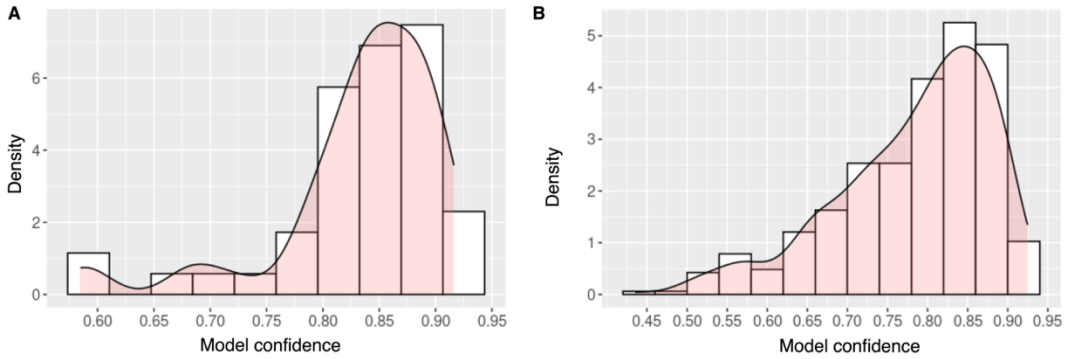

Figure 6.6. Model confidence distribution for TCR-pMHC sequences. A dataset of TCR-pMHC sequences was collected from the VDJdb database [312], consisting of 414 Class I TCR-pMHC and 47 Class II TCR-pMHC sequences that have no redundancy with TCRs from known TCR-pMHC complex structures (cutoff date 9/30/2021) and each other, based on our nonredundancy criteria used for benchmark assembly. Only samples with a VDJdb confidence score >=1 were considered. We then modeled all samples using our TCRmodel2 and obtained the model confidence of the top-ranked model for each complex. The distributions of model confidence scores are shown for **(A)** Class I TCR-pMHC complexes and (B) Class II TCR-pMHC complexes. Plots were generated using the ggplot2 package [186] in R (r-project.org), with a bin width of 0.04.

6.4.4 TCR complex modeling examples

As an example of TCR-pMHC complex modeling in TCRmodel2, the server was used to predict the structure of a human TCR in complex with an immunodominant SARS-CoV-2 nucleocapsid epitope presented by the Class I HLA-B*07:02 MHC. The complex has not been structurally characterized, nor have any complexes with TCRs targeting that epitope, and its sequence is from a set of TCRs from COVID-19 recovered and unexposed donors reported in a recent study to bind that peptide (sequence: SPRWYFYYL) and MHC [315]. Of note, the TCR contains the TRBV27 germline gene and a long CDR3β sequence (18 residues) containing a sequence motif (PxxGxP); those features were found by the authors to be associated with TCRs targeting that epitope. After input of the germline gene and TCR CDR3 sequences reported by the authors (α: TRAV35/TRAJ39, CAGQLNAGNMLTF; β: TRBV27/TRBJ2-4, CASAPLVGAPEAKNIQYF), along with the epitope sequence and MHC, TCRmodel2 was used

to generate five structural models of the complex. The server identified the unbound pMHC structure of the target peptide and MHC (PDB code 7LG0) among its four pMHC templates, and the TCR α and β chain templates with highest sequence identities to the target sequences (each with 89% identity) contain germline genes matching the target (α: 5W1V, TRAV35; β: 6VQO, TRBV27). The top-ranked model (**Figure 6.7**) had a high model confidence score (0.86), and inspection of the predicted interface with pMHC (**Figure 6.7a**) showed extensive interaction of the CDR3β, and the PxxGxP motif residues (PLVGAP) in particular, with the peptide, as well as the TRBV27-encoded germline loops making extensive interactions with the MHC. This provides a possible mechanistic explanation for the observed preference for TRBV27 in TCRs targeting that epitope, as well as the observed CDR3β sequence motif within the long CDR3β loop.

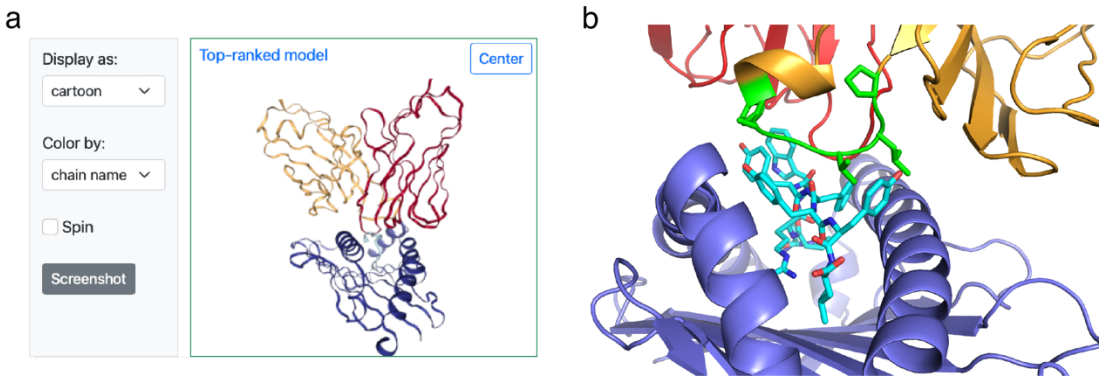

Figure 6.7. Example TCR-pMHC modeling output from TCRmodel2. (a) A TCR-pMHC complex with a human TCR, SARS-CoV-2 nucleocapsid epitope, and HLA-B*07:02 MHC from a recent study [315] was modeled using TCRmodel2. The visualization of the top-ranked model from the Results page is shown, with TCR α chain red, β chain orange, peptide cyan, and MHC blue. (b) The interface between the TCR and pMHC of the top-ranked model is shown, with TCR, peptide, and MHC chains colored as in **(a)**, and shared CDR3β motif residues (sequence: PLVGAP) colored green and shown as sticks. Peptide residues are shown as sticks, and TCR CDR1β and CDR2β residues interacting with the MHC and/or peptide are shown as sticks and circled. Structure visualized using PyMOL (Schrödinger, Inc.).

For a second example, TCRmodel2 was used to model the structure of a Class II TCR-pMHC interaction with a tumor-infiltrating lymphocyte TCR (named 4285-TCR1) that was

found to target the common Class II MHC allele HLA-DRB1*13:01 and a p53 neoantigen with the R175H mutation [316]. While structures of Class I TCR-pMHC complexes with the p53 neoantigen mutation have been reported [182, 317], no Class II structures with that mutation have yet been described. To elucidate that mode of CD4+ T cell recognition of the p53 R175H hotspot mutation, we input the TCR Vα and Vβ sequences into the TCRmodel2 submission page, along with a p53 peptide sequence containing the mutant residue (TEVVRHCPHHERCSD; mutant histidine in bold), and selected the HLA-DRA*01:01 and HLA-DRB1*13:01 MHC genes. The results page from TCRmodel2 included the top-ranked model of that TCR-pMHC complex (Figure 6.8A), which has a high confidence score (0.88). Downloading the PDB structure of the top-ranked model and visualization of its structure indicates that the mutant histidine residue is located directly at the interface with the TCR, engaging both α and β chains, suggesting a possible mechanism for the neoantigen specificity of that TCR (Figure 6.8B).

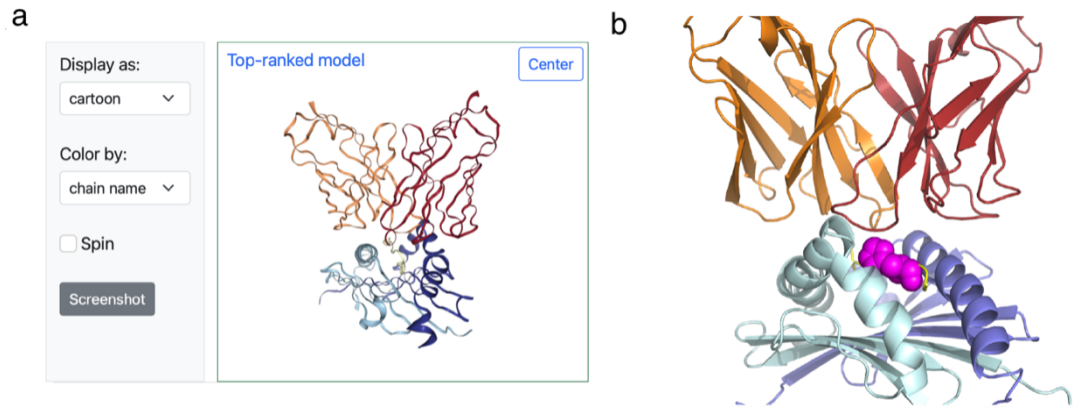

Figure 6.8. Example Class II TCR-pMHC modeling output from TCRmodel2. (a) A previously identified neoantigen-specific TCR [316] was modeled in complex with its target, a p53 R175H neoantigen and Class II HLA-DRB1*13:01 MHC, and the visualized top-ranked model from the TCRmodel2 results page is shown. The model is shown in cartoon representation with TCR α chain (red), TCR β chain (orange), peptide (yellow), MHC α chain (cyan), and MHC β chain (blue) colored separately. (**b**) The PDB file for the top-ranked model, downloaded from the TCRmodel2 results page, is shown visualized in PyMOL (Schrödinger, Inc.). Chains are colored as in **(a)**, and the neoantigen mutant residue (H175) is shown in sphere representation and colored magenta.

6.5 Discussion

The TCRmodel2 web server provides the community with a deep learning method to accurately model structures of TCRs and TCR-pMHC complexes. Its TCR-pMHC accuracy is higher than AlphaFold, it runs faster and without the need for dedicated computing resources, and it provides a submission and output interface designed for TCR and TCR-pMHC modeling. TCRmodel2 is distinguished from another recently reported AlphaFold-based TCR-pMHC modeling method [309], as it does not rely on fine-tuning of the AlphaFold model or TCR-pMHC complex templates, in addition to its availability as a web server versus a command-line program.

Future possible developments of TCRmodel2 can address improving the accuracy of confidence estimates, and increasing the overall success rate, including the generation of near-native (CAPRI High criteria) accuracy models, through additional optimizations of the modeling pipeline. Additional testing and developments may focus on application of TCRmodel2 for related complexes of interest, such as TCR-mimic antibodies [318], which engage pMHC targets and are becoming increasingly of interest as therapeutics. Modeling of such complexes would likely entail limited, if any, adaptations to the current TCRmodel2 framework, including a possible expansion of the MSA database to optimize antibody sequence hits. Given the recent utilization deep learning structure-prediction methods to design new proteins and interactions [319, 320], it may be possible that TCRmodel2 or similar methods can be used in future studies to design and optimize TCRs to target antigens of interest.

Chapter 7: Exploring AlphaFold massive sampling for immune recognition modeling

7.1 Abstract

The massive sampling technique in AlphaFold involves the random dropping of neurons in a neural network, a technique termed "dropout," and extensive sampling to produce more than 1000 predictions for each complex. Our prior investigations revealed that this approach could produce successful predictions in approximately half of the antibody-protein antigen cases examined. However, it is not clear whether the success of this approach for antibody-protein antigen complexes translates to other immune recognition interactions, such as antibody-peptide antigen and TCR-pMHC complexes. This chapter investigates the effectiveness of massive sampling for these additional interaction types on a set of 41 antibody-peptide and 20 TCR-pMHC complexes test cases. Considering the computational cost of the conventional massive sampling method, this chapter also investigates the potential for achieving similar successes with a more resource-efficient sampling strategy. We also provide insights in the model scores, offering practical guidance for the interpretation of prediction quality. Our results underscore the value of massive sampling in augmenting immune recognition modeling accuracy and indicate that obtaining comparable results with lower computational costs is feasible.

7.2 Introduction

In the latest CASP/CAPRI round, Bjorn Wallner's group stood out as one of the top three performers by accurately predicting protein-protein interactions through innovative massive sampling techniques[86, 274, 287]. By activating dropout during the inference stage and conducting extensive sampling to generate up to 6,000 models per target, Wallner's group achieved higher modeling higher success beyond what is achievable with the default AlphaFold configuration. Another top-performing group that achieved notable success was led by Česlovas Venclovas [86]. They also attained high modeling accuracy by implementing an extensive sampling strategy [321]. Specifically, they built custom MSAs by adjusting the MSA search parameters. With the custom MSAs as feature inputs, they generated a large number of predictions and thereby achieved higher modeling success than the default AlphaFold. This approach also underscores the benefits of employing extensive sampling techniques.

In the Wallner group's approach to massive sampling, dropout is activated at inference time [274]. A regularization technique used in training neural networks to prevent overfitting [322], dropout involves randomly setting a subset of neurons and their connections to zero. This forces the network to learn a robust representation that can generalize well to unseen data. It is typically employed only during training, as dropout introduces randomness, making outputs less deterministic. However, research shows that leaving dropout turned on during inference can mimic the behavior of an ensemble of models, with each forward pass being akin to sampling from a slightly different network, and this variation in outputs can be utilized for uncertainty estimation in the model's predictions [323]. Prior to the release of the AFsample study, Johansson-Åkhe and Wallner [277] and Mirdita et al. [226] both showed that activating dropout in AlphaFold (and ColabFold in Mirdita et al. [226]) leads to diversified structure predictions.

In Chapter 5, we presented findings that the use of massive sampling led to an approximate 15% increase in the near-native accuracy success rate for antibody-protein antigen structure predictions. The applicability of this enhanced success rate to other types of protein-protein interactions involved in immune recognition, such as antibody-peptide and TCR-pMHC complexes, remains unexplored in previous research. Given that massive sampling generates up to 6000 predictions per complex, it incurs significant computational demands. Therefore, in this chapter, we aim to explore two key questions: 1) Does the improved success rate observed with antibody-protein antigen complexes extend to other immune recognition interactions, such as antibody-peptide and TCR-pMHC complexes? 2) Is it possible to develop a more efficient strategy that reduces the computational burden of massive sampling while preserving its success rate?

Antibody-peptide interactions are crucial to study for a myriad of reasons. The ability of antibodies to recognize and bind to specific linear peptide epitopes is fundamental to the development of immunotherapies and vaccines against a wide range of diseases, including viruses like hepatitis C virus and various coronaviruses [324-326], as well as for targeting host proteins for therapeutic intervention [327]. These interactions are pivotal for neutralizing pathogens and the design of immunogens to elicit protective immune responses against diseases such as HIV [328] and Respiratory Syncytial Virus (RSV) [329]. Furthermore, our previously published paper suggests nuanced differences between antibody-peptide and antibody-protein interfaces [41], highlighting the importance of specifically considering antibody-peptide interactions. Due to the peptide and CDR loop flexibility in antibody-peptide complexes [41], accurate modeling of antibody-peptide complexes can be challenging for traditional modeling tools that have limited success with flexible docking. Previously, researchers reported success

using AlphaFold for protein-peptide modeling [277, 330]. However, those studies didn't specifically focus on antibody-peptide modeling. To our knowledge, there is a lack of comprehensive studies examining the capability of AlphaFold or massive sampling techniques to model antibody-peptide interactions, indicating a gap in current research. Given this backdrop, studying how well these interactions can be modeled is crucial and will be addressed in the current chapter.

7.3 Methods

7.3.1 Antibody-antigen massive sampling

To produce predictions using the AFsample algorithm for antibody-protein antigen test cases, we first downloaded the AFsample software from its GitHub repository at https://github.com/bjornwallner/alphafoldv2.2.0. We then followed the modeling procedures as outlined by Wallner [274]. For each antibody-protein antigen complex, 6,000 predictions were generated using various configurations of AlphaFold versions v2.1 and v2.2. Specifically, we generated 2,000 predictions with the AlphaFold v.2.1 and v.2.2 models, utilizing templates and applying full dropout. We produced an additional 2,000 predictions using the v.2.1 and v.2.2 models without templates while implementing dropout in the Evoformer. We also generated 1,000 predictions with the v2.1 models, omitting templates, expanding the maximum number of recycles to 21, and applying dropout only in the Evoformer. We additionally generated 1,000 predictions using the v2.2 models without templates, with a maximum of 9 recycles, and with dropout applied solely in the Evoformer. For scenarios where templates were enabled, we employed a template date cutoff of September 30, 2021. Following the AFsample protocol, we

performed Amber relaxation on the top 5 predictions, ranked by model confidence. Modeling was performed on high performance computing clusters using NVIDIA GPUs, including Titan RTX, Quadro 6000, and A100.

As it was initially developed for compatibility with the AlphaFold v2.2.0 framework, we implemented the AFsample protocol in the AlphaFold v2.3.2 code to allow incorporation of v2.3 model parameters in massive sampling.

Three protocols featuring parameters from AlphaFold v2.1, v2.2, and v2.3 were used to generate massive sampling predictions for antibody-peptide complexes. The protocols applied dropout to modules except for the structure module. For each model, 100 predictions were generated. No templates were used during the modeling of antibody-peptide complexes. In total, 1,500 predictions per complex were generated (100 predictions * 5 models * 3 protocols).

7.3.2 TCRmodel2 Massive sampling

The AFsample protocol was incorporated into the TCRmodel2 code. One of TCRmodel2's key innovations was merging peptides and MHC molecules into a single entity and employing templates to regularize the peptide’s geometry in the MHC groove[331]. The massive sampling version of TCRmodel2 retains this essential feature, employing peptide-MHC complex templates. TCR templates were omitted. Peptide-MHC templates released on or before September 30, 2021, were used. Due to resource limitations, 100 predictions were generated per model. A total of 3 protocols were used, featuring parameters from AlphaFold v2.1, v2.2, and v2.3, applying dropout to modules except for the structure module. In total, 1,500 predictions per complex were generated (100 predictions * 5 models * 3 protocols).

7.3.3 Docking model accuracy assessment

The accuracy of antibody-antigen and TCR-pMHC complex models was evaluated with DockQ [277], (https://github.com/bjornwallner/DockQ). DockQ calculates the accuracy of models by comparing them to experimentally determined structures. It calculates interface and ligand backbone RMSD (I-RMSD and L-RMSD, respectively), the fraction of native contacts (fnat), and the DockQ score. Additionally, based on model's similarity to the native structure, DockQ determines the model's CAPRI (Critical Assessment of Prediction of Interactions) accuracy, categorizing models into four distinct classes: "High", "Medium", "Acceptable" and "Incorrect" accuracy classes [223]. For assessment of antibody-peptide modeling accuracy, "-capri_peptide" was added to indicate assessment of protein-peptide interfaces.

7.3.4 Interface pLDDT score

Interface pLDDT (I-pLDDT) score was determined by computing the average pLDDT value for residues within 4.0 Å distance cutoff at the antibody-antigen, antibody-peptide and TCR-pMHC interface. Residues at the interface were defined as those having any non-hydrogen atom located within 4.0 Å of their interacting partner. For complexes lacking interface residues within a 4.0 Å distance of the interacting partners, we assigned an I-pLDDT score of -1.

7.3.5 TCR-pMHC ipTM score

The TCR-pMHC ipTM score is specifically designed to assess the quality of the interaction interface between the TCR and pMHC of TCRmodel2 predictions[331]. To calculate the TCR-pMHC ipTM score, AlphaFold is modified to treat the TCR and pMHC within the

predicted complex as a single chain each at the time of the ipTM score calculation. This score is outputted by TCRmodel2 algorithm.

7.3.6 Benchmarking set assembly

To benchmark antibody-peptide complex modeling accuracy, we construct a non-redundant dataset of antibody-peptide complexes based on the following criteria: 1) The structural resolution <= 3.0 Å, 2) The antigen is a peptide as identified through SAbDab annotation, and 3) The antibody-peptide complex is not redundant with complexes within the dataset, and with previously released antibody-antigen (protein and peptide antigens all considered) complexes that have a structural resolution of 9.0 Å or better, up until September 30, 2021. Nonredundancy is defined as having less than 90% identity in the sequence of the heavy chain variable domain or the full variable domain.

The TCR-pMHC benchmarking set is the TCRmodel2 (AlphaFold v2.3 based) benchmarking set described in Chapter 6. The antibody-protein antigen benchmarking set is the AlphaFold v2.3 benchmarking set described in Chapter 5.

7.3.7 Docking angle and incident angle assessment

A previously developed C++ program was employed to assess the TCR crossing angle and incident angle. The program was developed for TCR angle assessment [40] based on the methodology outlined by Rudolph et al. [46]. It is available via GitHub (https://github.com/piercelab/tcr_docking_angle) and as a public web server (https://tcr3d.ibbr.umd.edu/tcr_angle).

7.3.8 Plotting and statistical analysis

The ggplot2 package in R (r-project.org) was employed for producing plots in this chapter. The "cutpointr" package [243] was used to obtain the optimal cutoffs for distinguishing predictions of Incorrect accuracy from Medium and High accuracy.

7.4 Results

7.4.1 Massive sampling increases antibody-peptide modeling success

Antibody-peptide interactions represent an important class of interactions in immune recognition. On a set of 41 nonredundant antibody-peptide complex structures that we compiled from SAbDab and the PDB (**Table 7.1**), we found that AlphaFold v2.3 generated Medium or higher accuracy predictions as top-ranked predictions for 34% of the test cases (**Figure 7.1a**). This success is comparable to the 35% success rate we observed in AlphaFold v2.3 modeling of 37 antibody-protein antigen complexes test set (**Figure 5.25b**). Overall, our observations indicate that there is no significant difference in AlphaFold's ability to model antibody-protein antigen complexes compared to antibody-peptide antigen complexes, suggesting that AlphaFold's predictive capability extends effectively across both types of antibody-antigen complexes.

Table 7.1 Antibody-peptide complexes analyzed in this chapter.

PDB	Heavy	Light	Antigen	Release date[1]	Resolution[1]
7y8j	H	L	A	7/12/23	1.03
7rpu	A	B	P	7/13/22	1.27
7doh	H	L	I	10/27/21	1.45
8d36	H	L	F	7/27/22	1.45
8dtr	E	F	J	11/23/22	1.5
8dtx	A	B	G	11/23/22	1.6
8bbh	H	L	A	12/21/22	1.62
7q4q	B	A	E	6/15/22	1.65
8duz	C	D	E	7/5/23	1.65
7u0a	H	L	A	8/3/22	1.7
7raq	H	L	P	4/13/22	1.74
8dtt	E	F	J	11/23/22	1.75
8f0l	A	B	Q	3/29/23	1.81
7skz	H	L	A	7/20/22	1.86
8dgu	H	L	A	1/25/23	1.89
7s3n	H	L	A	10/27/21	1.9
7rlx	A	B	P	1/26/22	1.97
7n08	H	L	F	3/30/22	2
8d47	H	L	C	12/21/22	2
7n0x	H	L	G	4/6/22	2
7tcq	H	L	C	8/3/22	2.02
7u0e	H	L	C	8/3/22	2.1
7u09	H	L	A	8/3/22	2.1
8fax	A	B	L	5/3/23	2.1
7sjp	H	L	E	9/7/22	2.1
7s4g	E	F	I	4/13/22	2.2
7uym	H	L	P	11/23/22	2.2
7rqr	A	B	C	3/2/22	2.23
8d6z	A	B	H	7/27/22	2.3
8dgv	H	L	A	1/25/23	2.3
8fg0	A	B	Q	5/10/23	2.36
7um3	I	M	B	9/7/22	2.4
7k7r	B	A	C	1/12/22	2.5
7sl5	A	B	C	7/20/22	2.5
7e6p	H	L	A	1/5/22	2.5
7x9e	C	D	F	5/11/22	2.6
7y3j	H	L	A	8/17/22	2.6

8dao	C	D	I	7/27/22	2.8
8dgw	C	D	J	1/25/23	2.81
8dgx	H	L	C	1/25/23	2.89
8op0[2]	A		B	5/31/23	1.54

[1]Resolution and release date of structure in the Protein Data Bank (PDB) [164].
[2]A nanobody-peptide complex.

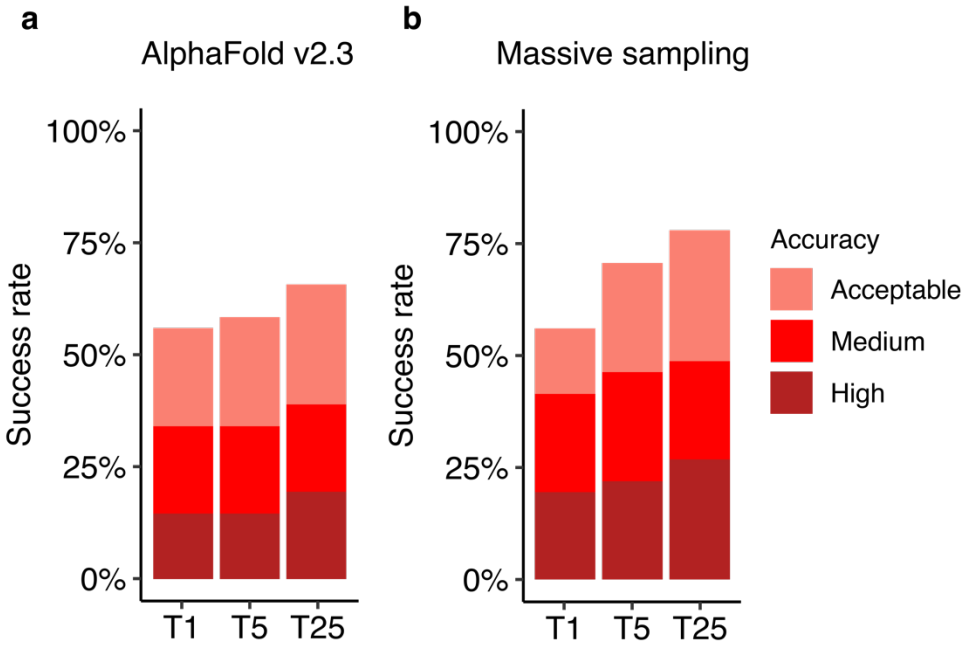

Figure 7.1. Massive sampling can improve antibody-peptide modeling success. (a) Default AlphaFold v2.3 generates 25 predictions per complex. **(b)** Massive sampling generates 1,500 predictions per complex. Ranking is based on model confidence scores. Model quality assessed by DockQ protein-peptide setting matching CAPRI criteria for protein-peptide complexes and colored as indicated.

Next, we investigated whether the use of massive sampling increases the modeling success for antibody-peptide complexes. Based on prior investigation (data not shown), we observed that the use of templates in massive sampling of antibody-antigen complexes didn't yield higher success than the massive sampling protocol that does not use templates, and in some cases, not using templates yielded higher success. Therefore, I chose a no-template approach and

pooled predictions from versions v2.1, v2.2, and v2.3, with 100 predictions per model. This strategy yielded 1,500 predictions per complex.

Massive sampling produced 41% of Medium or higher accuracy predictions as top-ranked predictions (**Figure 7.1b**), which is 7% higher than the default AlphaFold v2.3 (**Figure 7.1a**), corresponding to 3 more successful complexes predicted out of 41. Furthermore, the use of massive sampling in comparison to the default AlphaFold v2.3 resulted in a higher proportion of High accuracy predictions. Specifically, massive sampling produced High accuracy top-ranked predictions for 20% of the test cases (versus 15% with the default protocol), yielding two additional cases with High accuracy top-ranked predictions. With the caveat of small sample size, massive sampling seems to provide a modest increase success for antibody-peptide modeling.

7.4.2 Massive sampling increases TCRmodel2 modeling accuracy

We next applied massive sampling strategy to TCRmodel2 framework. We modified TCRmodel2 for compatibility with massive sampling by integrating changes from AFsample. We generated 100 predictions per model across three protocols based on AlphaFold v2.1, v2.2, and v2.3, applying dropout to all but the structure module, utilizing only pMHC templates released by September 30, 2021, and excluding TCR templates.

We also removed predictions with non-canonical TCR angles (crossing angle > 100°, or incident angle > 50°). The ranges were determined based on an examination of angles of experimentally resolved Class I and Class II TCR-pMHC structures (https://tcr3d.ibbr.umd.edu/complexsplots). The majority of Class I and Class II TCR-pMHC complexes fall within this range, with exceptions where the TCRs are docked on pMHC

with reversed polarity [332, 333]. In total, predictions for each complex having non-canonical TCR angles ranged from 6 to 414 and were excluded.

On a total of 20 TCR-pMHC benchmarking cases, massive sampling generated High accuracy predictions as top-ranked predictions for 30% of test cases (**Figure 7.2b**), which is 10% higher than the TCRmodel2 default protocol (**Figure 7.2a**), with the caveat of a small test set.

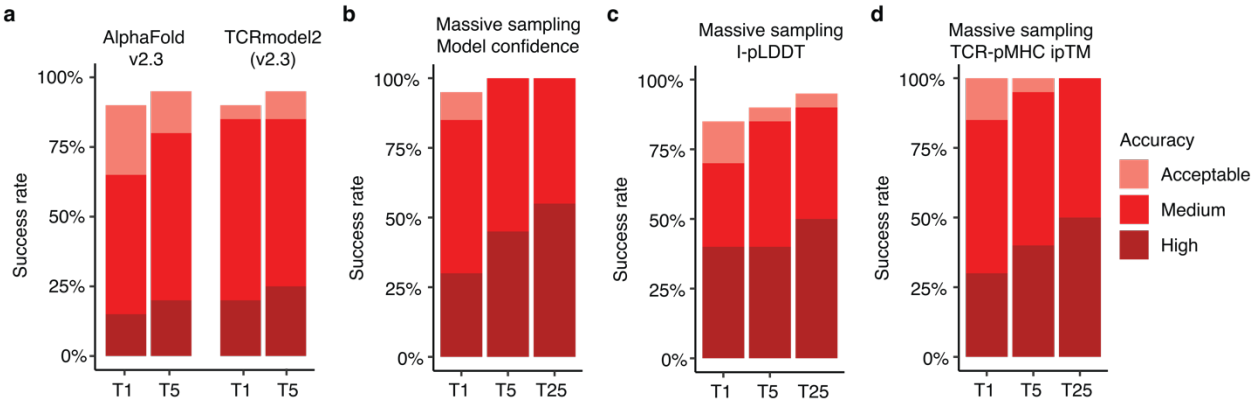

Figure 7.2. Massive sampling can improve TCR-pMHC modeling success. (a) AlphaFold v2.3 generates 25 predictions per complex. TCRmodel2 generates 5 predictions per complex. Massive sampling generates 1,500 predictions per complex. Ranking of massive sampling predictions based on **(b)** model confidence scores, **(c)** interface pLDDT scores, **(d)** TCR-pMHC ipTM scores.

In addition to ranking models based on their confidence scores, we explored the effectiveness of ranking them using interface-focused confidence scores. In Chapters 5 and 6, we introduced the interface pLDDT score (I-pLDDT score) as a scoring function that offers improved specificity compared to the default model confidence scores. Relatedly, TCRmodel2 outputs a TCR-pMHC interface-oriented pTM score, TCR-pMHC ipTM. These scores concentrate solely on the interaction interface, in this case specifically between TCR and pMHC, instead of spanning all interaction interfaces. We investigated whether employing these

interface-focused scores could improve the success of ranking more High accuracy models as top-ranked prediction.

While ranking models by the TCR-pMHC ipTM scores (**Figure 7.2d**) did not affect the near-native success rate in comparison to the success rate ranked by model confidence scores (**Figure 7.2b**), ranking models by their I-pLDDT scores (**Figure 7.2c**) resulted in an increased High accuracy top-ranked success rate to 40%. However, it resulted in a decrease in the success rate for Medium accuracy top-ranked predictions. This underscores the potential of using the I-pLDDT score for identifying High accuracy, near-native predictions but also highlights the need for a composite scoring system that can mitigate the reduction in Medium accuracy ranking success.

7.4.3 Achieving massive sampling success with reduced computational cost

The default AFsample massive sampling protocol is heavily resource intensive. It generates 200 predictions per model, employing variations of v2.1 and v2.2 model parameters, generating a total of 6,000 predictions. For an antibody-antigen complex of 433 amino acids in total length, after generating features, which takes two hours, generating one complex takes 4 minutes on one NVIDIA A100 Tensor Core GPU. For this complex, generating 6,000 predictions would take 400 hours. To make the situation worse, the computational cost drastically increases as the size of the complex increases.

To investigate whether similar success could have been achieved on a relatively smaller scale of sampling, we subsampled within the default AFsample protocol, reducing the number of predictions per model from 200 to 100, 20 and 5. The reduction in sampling number from 200 (**Figure 7.3a**) to 100 (**Figure 7.3b**) did not change the near-native accuracy success rate, while

the drop to 20 (**Figure 7.3c**) predictions per model resulted in a 3% drop (one case) in percentage Medium or higher accuracy top-ranked prediction, and a 5% drop (two cases) in percentage of High accuracy top-ranked prediction produced. Conversely, reduction in sampling number from 200 (**Figure 7.3a**) to 5 (**Figure 7.3d**) predictions per model did not result in drop in percentage Medium or higher accuracy top-ranked prediction, but likewise resulted in a 5% (two cases) drop in percentage of High accuracy top-ranked prediction produced. Collectively, these findings underscore that achieving massive sampling success in generating near-native predictions is possible even with fewer predictions per model generated.

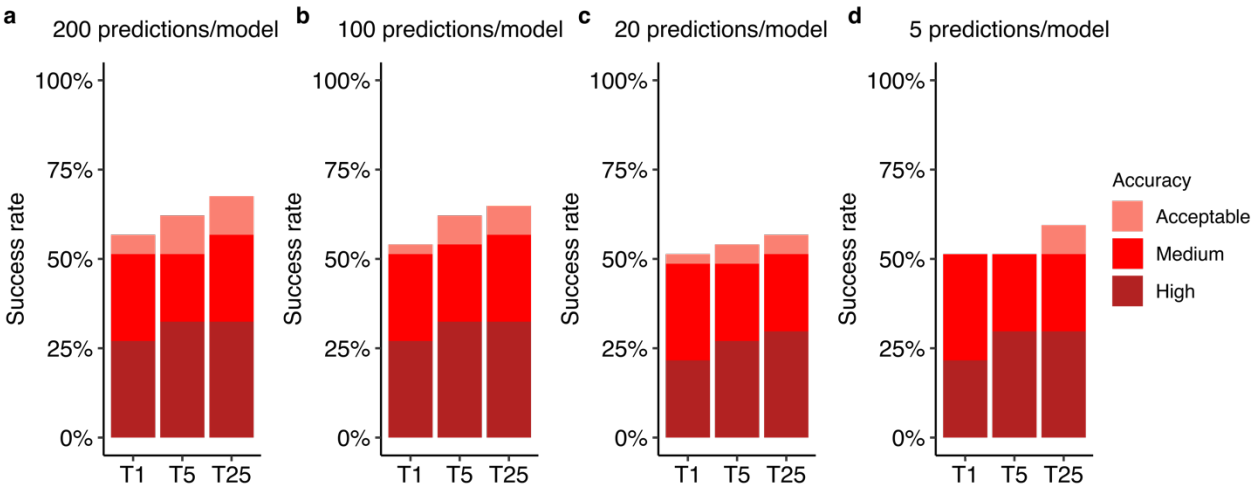

Figure 7.3. Reducing predictions per model maintains massive sampling success in generating near-native predictions as top-ranked predictions. (a) Default AFsample predictions. For each model, 200 predictions were made, leading to a total of 6,000 predictions per complex (5 models × 200 predictions × 6 complexes). **(b)** Subsampling of AFsample predictions, generating 100 predictions per model, leading to a total of 3,000 predictions per complex (5 models × 100 predictions × 6 protocols). **(c)** Subsampling of AFsample predictions, generating 20 predictions per model, leading to a total of 600 predictions per complex (5 models × 20 predictions × 6 protocols). **(d)** Subsampling of AFsample predictions, generating 5 predictions per model, leading to a total of 150 predictions per complex (5 models × 5 predictions × 6 protocols). Model quality assessed by CAPRI criteria and colored as indicated.

7.4.4 Investigating predicted model quality scores

Massive sampling markedly expanded our repertoire of predictions, enabling us to perform a comprehensive analysis of the relationship between various predictive scores and the model quality. For each of the 37 antibody-antigen test cases, 6,000 default AFsample predictions are made per complex. Comparing the predictions' model confidence scores and their accuracies, measured by DockQ, there is a Pearson's correlation coefficient of r=0.7 (**Figure 7.4a**). This number is higher than what we observed previously (**Figure 5.11a**, r=0.52). The increase is possibly attributable to the presence of a massive number of predictions of poor quality within the massive sampling test dataset, disproportionately influencing the correlation measurement, leading to the observed increase.

For an alternative formulation of model quality prediction, I-pLDDT, its correlation with model quality, DockQ score, is 0.51 (**Figure 7.4b**). The correlation between I-pLDDT score and DockQ is not as strong as that of model confidence; however, visual inspection of **Figure 7.4** indicates that I-pLDDT score has less numbers of false negatives relative to model confidence.

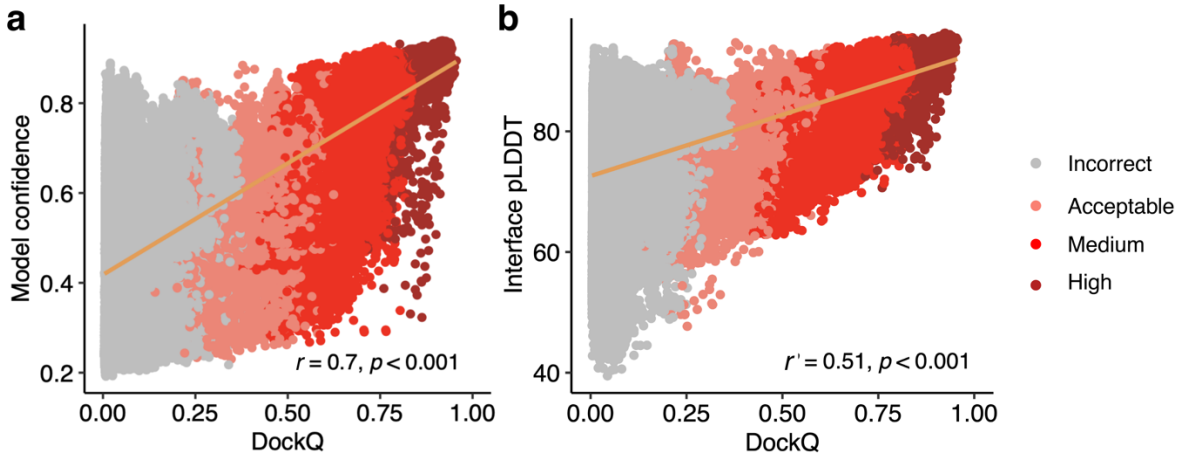

Figure 7.4. Correlation between model confidence, interface pLDDT and model accuracy measured by DockQ. Scatter plots illustrate the relationship between **(a)** model confidence, **(b)** interface pLDDT

scores, and DockQ scores, featuring 6,000 models per complex, on a total of 37 complexes as data points. The color of each point reflects the model quality based on CAPRI criteria. Predictions without contacts within 4 Å across the antibody-antigen interface were given an interface pLDDT score of -1. A total of 10 predictions do not have the interface contacts and were omitted from both plots. An orange line indicating the linear regression is shown, and the lower right corner of each scatter plot contains Pearson's correlation coefficients and their corresponding p-values.

To investigate the two score further, we plotted the distribution of model accuracy at given confidence score ranges. Upon examining the model confidence scores, we observed that predictions with high model confidence scores, greater than 0.85, tend to yield Medium or High accuracy predictions (**Figure 7.5**). This indicates that higher confidence scores are generally reliable indicators of prediction quality. However, the histogram above the bar plot reveals that high model confidence scores are not common occurrences (**Figure 7.5**). This suggests that while high confidence is a good indicator of success, it is achieved infrequently.

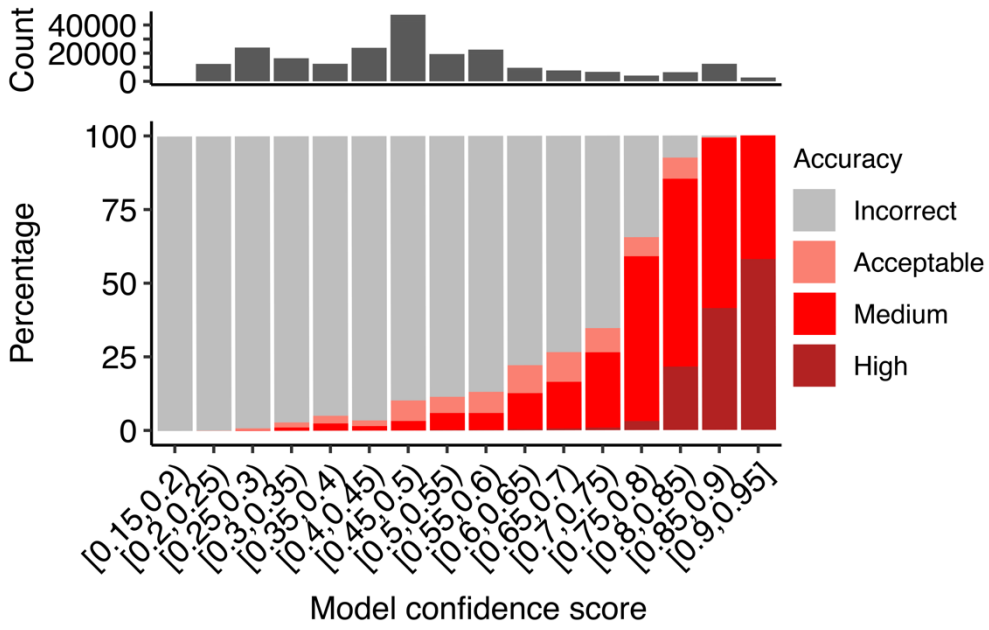

Figure 7.5. Percentage of successful predictions by model confidence intervals. The upper panel shows the distribution of data points at each model confidence interval, denoted on the x-axis. The lower panel shows the percentage of data points at the designated score interval possessing CAPRI accuracy as

indicated by the color. A total of 6000 predictions per complex for 37 complexes were represented on the plot.

Analysis of the I-pLDDT scores likewise shows that high scores, specifically above 90, generally correlate with Medium or High accuracy in predictions (**Figure 7.6**). Notably, when the I-pLDDT score surpasses 95, a substantial proportion of predictions fall into the High accuracy category, emphasizing the score's effectiveness as a predictor of model quality. However, like model confidence scores, the histogram indicates that instances of high I-pLDDT scores are rare (**Figure 7.6**).

Despite the overall trend towards accuracy with higher model confidence and I-pLDDT scores, many near-native predictions do not achieve ideal scores. The challenge ahead lies in figuring out a method to distinguish near-native accuracy predictions from the rest effectively, while also maintaining high sensitivity in the evaluation metrics.

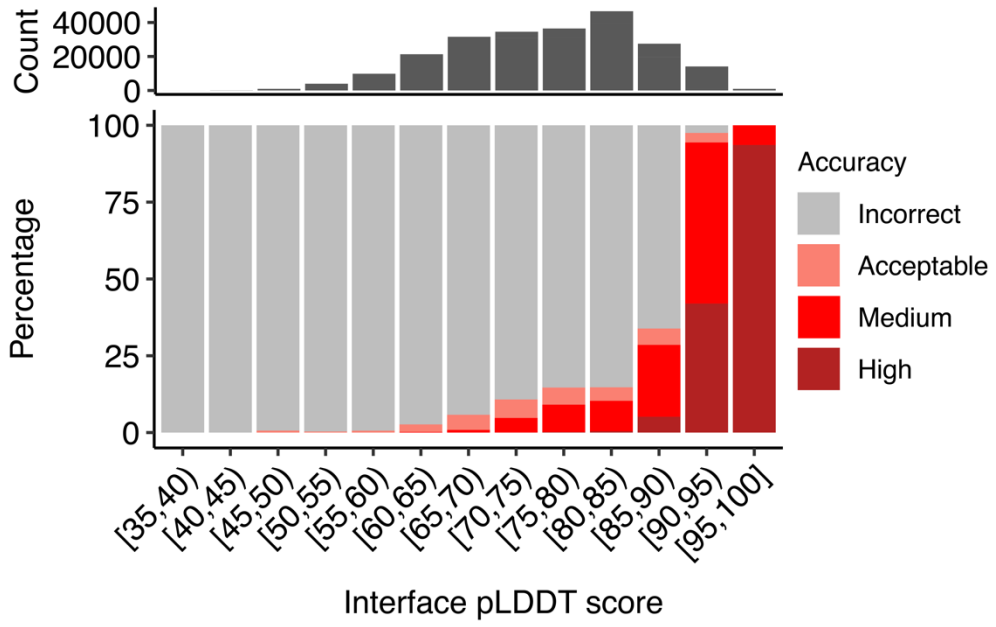

Figure 7.6. Percentage of successful predictions by interface pLDDT intervals. The upper panel shows the distribution of data points at each I-pLDDT score interval, denoted on the x-axis. The lower panel shows the percentage of data points at the designated score interval possessing CAPRI accuracy as

indicated by the color. A total of 6000 predictions per complex for 37 complexes were represented on the plot. Predictions without interface contacts within the 4 Å distance cutoff were given a I-pLDDT score of -1. In total, there were 10 predictions with a I-pLDDT score of -1 and not shown on the plot.

7.4.5 Identifying high accuracy predictions with a combination of model confidence and I-pLDDT score cutoffs

To maximize the number of near-native predictions we can identify and maintain high sensitivity, we explored a composite scoring approach that considers both the model confidence and I-pLDDT instead of considering them separately. Individual score cutoffs that maximize the sensitivity and specificity of discriminating between Incorrect and Medium or higher accuracy predictions were calculated. For model confidence, the cutoff was 0.64 and for I-pLDDT, 86.54.

On their own, the score cutoffs yield 63% and 70% Medium or higher accuracy prediction, respectively (**Figure 7.7**). However, when both conditions were met, there was a remarkable 93% hit rate (**Figure 7.7**). The hit rate for High accuracy prediction even reached 36% (**Figure 7.7**), emphasizing the effectiveness of integrating model confidence with I-pLDDT score cutoffs. This success is likely attributed to the synergy of model confidence scores, which predict global accuracy, and I-pLDDT scores, which predict interface accuracy.

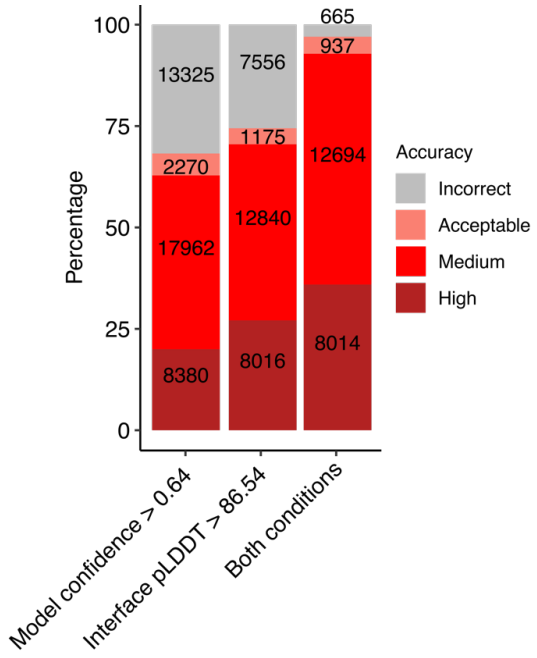

Figure 7.7. A combination of score cutoffs can be used to identify good versus bad predictions. The score cutoffs core model confidence and interface pLDDTs are optimal cutoffs that best maximize the specificity and sensitivity for discriminating predictions of Incorrect and Medium or higher accuracy criteria. Each bar shows the percentage of predictions of Incorrect, Acceptable, Medium and High accuracy criteria when the particular score cutoff is met. "Both conditions" denote the condition in which predictions possess model confidence > 0.64 and I-pLDDT score > 86.54.

7.5 Discussion

In this chapter, we demonstrate that massive sampling approaches improve Medium or higher accuracy modeling success for antibody-peptide by 7%, compared to the default AlphaFold v2.3 baseline. It improves the TCR-pMHC High accuracy modeling success by 10-25% compared to the TCRmodel2 baseline, depending on the ranking scheme. Our finding extends the utility of massive sampling beyond the initial antibody-protein antigen test set. Additionally, we investigated whether the success achieved through massive sampling could be replicated with a reduced number of predictions per model. Our findings reveal that it is indeed possible to achieve a comparable level of near-native modeling success even with a substantial

reduction in the number of predictions per complex. This suggests that enhanced modeling success can be attained without a significant investment in resources. Further research may investigate the applicability of this finding to TCR-pMHC or antibody-peptide complexes.

A promising direction for future research involves exploring the potential for improving the ranking of predictions. Across all test cases examined, including antibody-peptide, TCR-pMHC, and antibody-protein antigen interactions, we observe an increase in success as the number of predictions considered expands. This trend indicates that accurate predictions are not always ranked at the top by the model confidence-based ranking function, suggesting a potential area for refinement in prediction ranking. Our finding that the I-pLDDT score successfully identified a higher number of High accuracy TCR-pMHC predictions as top-ranked suggests the potential value of exploring alternative confidence score formulation or a composite ranking score for reranking of predictions.

Another promising direction for future research involves the fine-tuning of AlphaFold. Currently, our extensive sampling approach utilizes the standard AlphaFold model for antibody-antigen and antibody-peptide complexes, as well as the unmodified TCRmodel2, which is based on AlphaFold v2.3, without any fine-tuning. Previous research has shown that fine-tuning, coupled with adjustments to the input parameters and the inclusion of custom templates, is an effective strategy for enhancing TCR-pMHC modeling success beyond the baseline capabilities of AlphaFold's monomer [309]. This strategy warrants exploration of fine-tuning the AlphaFold-Multimer. The fine-tuning of AlphaFold-Multimer is made possible by recent developments such as OpenFold [291] and UniFold [290].

Future research could explore alternative structural prediction network architectures, such as those utilized by tFold [334], a deep learning algorithm specifically developed for antibody

and antibody-antigen modeling, or DMFold [335], which utilizes a distinctive method to construct MSAs, and is reported to perform better than AlphaFold on nanobody-antigen modeling. By continuously seeking improvements in these areas, we can strive towards more accurate, reliable, and generalizable prediction algorithms in immune recognition modeling.

Chapter 8: Summary and Future Directions

This dissertation investigates the application of deep learning in high resolution modeling of antibody-antigen and TCR-pMHC complexes. In Chapter 2, I reviewed the relevant literature, identifying and highlighting areas for improvement. In Chapter 3, I analyzed the key patterns in antibody recognition and targeting of the SARS-CoV-2 RBD. In Chapters 4 and 5, I evaluated deep learning tools like AlphaFold and classical docking algorithms, including ZDOCK, for their effectiveness in modeling molecular interactions, particularly in immune recognition. Next, I proposed targeted strategies to improve the accuracy of AlphaFold in modeling immune recognition. In Chapter 6, I introduced the TCRmodel2 algorithm, an adaptation of AlphaFold designed to improve accuracy and speed for modeling TCR-pMHC complexes. In Chapter 7, I delved into the potential of increased sampling to bolster AlphaFold and TCRmodel2's predictive success for antibody-protein and peptide antigen and TCR-pMHC interactions.

The insights obtained from Chapter 3 on the interactions between antibodies and the SARS-CoV-2 RBD have been extended to newly discovered antibodies targeting the SARS-CoV-2 RBD. The classification table, which lists antibodies targeting the SARS-CoV-2 virus, is regularly updated to show classifications of newly discovered antibodies and can be publicly accessed at https://cov3d.ibbr.umd.edu/antibody_classification. This extends the analysis provided in Chapter 3 beyond the initial dataset of 70 antibody-RBD complexes. Furthermore, the methodologies and tools developed for this study are versatile and can be utilized in examining other types of antibody-antigen interactions, thereby extending the chapter's significance beyond the SARS-CoV-2 research.

The discoveries from the benchmarking analysis in Chapters 4 and 5 have informed our approach to modeling and analyzing protein-protein complexes using AlphaFold. The

investigation has clarified the capabilities and limitations of AlphaFold, laying the groundwork for interpreting the accuracy of its generated models. Specifically, I-pLDDT, a scoring function focused on the interface between interacting proteins, is now routinely used within the lab to assess the reliability of AlphaFold's predictions for a variety of targets. Building on the findings from our benchmarking study, we have successfully predicted several antibody-antigen targets in collaboration with Dr. Roy Mariuzza's lab. Currently, we are exploring ways to combine structure prediction with sequence design tools to perform antibody design against targets of therapeutic relevance in collaboration with Dr. Yuxing Li's lab.

The TCRmodel2 algorithm, detailed in Chapter 6, has become an essential component of our lab's toolkit for modeling TCR-pMHC interactions. We have utilized this algorithm in collaboration with Dr. Paul Robbins to model previously uncharacterized neoantigen TCR-pMHC complexes, yielding insights into the structural basis for the TCR targeting and specificity. We have also successfully predicted several TCR-pMHC complexes in collaboration with Dr. Roy Mariuzza's lab. Moreover, we have started applying the TCRmodel2 algorithm to predict TCR-mimic antibody-peptide-MHC complexes, extending the toolkit's application beyond αβ TCR-pMHC complexes.

The massive sampling analysis presented in Chapter 7 also informed our approach to generate and evaluate models in various predictive modeling scenarios, including the CAPRI round 55. The application of massive sampling strategies resulted in a Medium accuracy prediction for the antibody-peptide complex target T231. Additionally, it produced a High accuracy prediction for the antibody-MHC target T233, although the prediction was not ranked highly according to our selection criteria. This experience has provided valuable insights, guiding us toward adopting more effective model selection strategies in the future.

The current dissertation delves into alternative formulations of scoring functions for estimating model accuracy and ranking models. However, there is still ample room for a more comprehensive investigation. For instance, the use of machine learning-based protein-protein interface scoring systems such as DOVE [121, 336] or iScore [337], or integrating physics-based methods with machine learning, like ZRANK [74], IRaPPA [76], should be explored as alternative strategies. Additionally, AlphaFold's predicted alignment error (PAE) could be a valuable area for further investigation and score function training. By enhancing the model scoring algorithm, it becomes possible to more accurately estimate model quality, rank near-native accuracy models more effectively, and consequently improve the success of modeling and ranking efforts.

Additionally, we are exploring the potential of leveraging structure prediction tools for specificity prediction. Motmaen and colleagues showed that by fine-tuning AlphaFold and incorporating a scoring function, it is possible to predict the specificity of peptide-MHC binding [276].

The recent release of RoseTTAFold All-Atom (RFAA) [124] marks a significant advancement in protein modeling, offering a potential solution to one of AlphaFold's limitations: its inability to accurately model heteroatoms at the antibody-antigen interface. This development suggests a promising direction for enhancing modeling success beyond the capabilities of the original AlphaFold. Concurrently, Google DeepMind's AlphaFold team and Isomorphic Labs have reported progress in developing a next-generation AlphaFold [125] capable of predicting the structure of proteins, ligands (small molecules), nucleic acids (DNA and RNA), and molecules with post-translational modifications. Moreover, this next-generation AlphaFold was reported to have increased antibody-antigen modeling accuracy, compared to the current version.

However, as mentioned in Chapter 2, this new version of AlphaFold remains unreleased and unreviewed. Nonetheless, the advancement of AlphaFold in its next-generation iteration and RFAA underscores the potential benefits of fine-tuning the algorithm to expand its modeling capabilities and enhance accuracy. Consequently, a promising direction for future research involves investigating methods to optimize both AlphaFold and TCRmodel2 using fine-tuning strategies. To achieve this, existing frameworks such as OpenFold [291] or Uni-Fold [290], designed for fine-tuning AlphaFold, could be utilized.

In the future, I also aim to assess deep learning-based protein complex modeling tools besides AlphaFold. Given AlphaFold's reliance on co-evolutionary information inferred from MSA, alternative algorithms that do not rely on MSA could be explored as effective modeling tools for antibody-antigen and TCR-pMHC interactions. Such tools include EquiDock [126], DockGPT [127], DiffDock-PP[128] and LATENTDOCK [129].

Publication Information

Peer-reviewed studies presented in this dissertation:

Yin R, Guest JD, Taherzadeh G, Gowthaman R, Mittra I, Quackenbush J, Pierce BG. *Structural and energetic profiling of SARS-CoV-2 receptor binding domain antibody recognition and the impact of circulating variants.* PLoS Comput Biol. 2021 Sep 7;17(9):e1009380. Doi: 10.1371/journal.pcbi.1009380. PMID: 34491988; PMCID: PMC8448325.

Yin R, Feng BY, Varshney A, Pierce BG. *Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants.* Protein Sci. 2022 Aug;31(8):e4379. Doi: 10.1002/pro.4379. PMID: 35900023; PMCID: PMC9278006.
**This article has received the most citations among those published in this journal over the past two years.*

Yin R, Ribeiro-Filho HV, Lin V, Gowthaman R, Cheung M, Pierce BG. *TCRmodel2: high-resolution modeling of T-cell receptor recognition using deep learning.* Nucleic Acids Res. 2023 Jul 5;51(W1):W569-W576. Doi: 10.1093/nar/gkad356. PMID: 37140040; PMCID: PMC10320165.

Yin R, Pierce BG. *Evaluation of AlphaFold antibody-antigen modeling with implications for improving predictive accuracy.* Protein Science. 2024 Jan;33(1):e4865. Doi: 10.1002/pro.4865. PMID: 38073135; PMCID: PMC10751731.

Other peer-reviewed publications during my Ph.D. research:

Lensink MF, et al., **Yin R**, Cheung M, Guest JD, Lee J, Pierce BG, et al., Wodak SJ. *Impact of AlphaFold on structure prediction of protein complexes: The CASP15-CAPRI experiment.* Proteins. 2023 Dec;91(12):1658-1683. doi: 10.1002/prot.26609. Epub 2023 Oct 31. PMID: 37905971.

Choy C, Chen J, Li J, Gallagher DT, Lu J, Wu D, Zou A, Hemani H, Baptiste BA, Wichmann E, Yang Q, Ciffelo J, **Yin R**, McKelvy J, Melvin D, Wallace T, Dunn C, Nguyen C, Chia CW, Fan J, Ruffolo J, Zukley L, Shi G, Amano T, An Y, Meirelles O, Wu WW, Chou L, Shen RF, Willis RA, Ko MSH, Liu Y, De S, Pierce BG, Ferrucci L, Egan J, Mariuzza R, Weng NP. *SARS-CoV-2 infection establishes a stable and age-independent CD8+ T cell response against a dominant nucleocapsid epitope using restricted T cell receptors.* Nat Commun. 2023 Oct 23;14(1):6725. doi: 10.1038/s41467-023-42430-z. PMID: 37872153; PMCID: PMC10593757.

Metcalf MC, Janus BM, **Yin R**, Wang R, Guest JD, Pozharski E, Law M, Mariuzza RA, Toth EA, Pierce BG, Fuerst TR, Ofek G. *Structure of engineered hepatitis C virus E1E2 ectodomain in complex with neutralizing antibodies.* Nat Commun. 2023 Jul 5;14(1):3980. doi: 10.1038/s41467-023-39659-z. PMID: 37407593; PMCID: PMC10322937.

Lee JH, **Yin R**, Ofek G, Pierce BG. *Structural Features of Antibody-Peptide Recognition.* Front Immunol. 2022 Jul 7;13:910367. doi: 10.3389/fimmu.2022.910367. PMID: 35874680; PMCID: PMC9302003.

Wu D, Kolesnikov A, **Yin R**, Guest JD, Gowthaman R, Shmelev A, Serdyuk Y, Dianov DV, Efimov GA, Pierce BG, Mariuzza RA. *Structural assessment of HLA-A2-restricted SARS-CoV-2 spike epitopes recognized by public and private T-cell receptors.* Nat Commun. 2022 Jan 10;13(1):19. doi: 10.1038/s41467-021-27669-8. PMID: 35013235; PMCID: PMC8748687.

Lensink MF, et al., Gowthaman R, Guest JD, **Yin R**, Taherzadeh G, Pierce BG, et al., Wodak SJ. *Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment.* Proteins. 2021 Dec;89(12):1800-1823. doi: 10.1002/prot.26222. Epub 2021 Sep 13. PMID: 34453465; PMCID: PMC8616814.

Gowthaman R, Guest JD, **Yin R**, Adolf-Bryfogle J, Schief WR, Pierce BG. *CoV3D: a database of high resolution coronavirus protein structures.* Nucleic Acids Res. 2021 Jan 8;49(D1):D282-D287. doi: 10.1093/nar/gkaa731. PMID: 32890396; PMCID: PMC7778948.

Bibliography

1. von Behring, E. and S. Kitasato, *[The mechanism of diphtheria immunity and tetanus immunity in animals. 1890].* Mol Immunol, 1991. **28**(12): p. 1317, 1319-20.
2. Wilson, I.A. and R.L. Stanfield, *Antibody-antigen interactions: new structures and new conformational changes.* Curr Opin Struct Biol, 1994. **4**(6): p. 857-67.
3. Blum, J.S., P.A. Wearsch, and P. Cresswell, *Pathways of antigen processing.* Annu Rev Immunol, 2013. **31**: p. 443-73.
4. Klebanoff, C.A., et al., *T cell receptor therapeutics: immunological targeting of the intracellular cancer proteome.* Nat Rev Drug Discov, 2023. **22**(12): p. 996-1017.
5. Chakraborty, A.K. and A. Weiss, *Insights into the initiation of TCR signaling.* Nat Immunol, 2014. **15**(9): p. 798-807.
6. Jameson, S.C. and M.J. Bevan, *T cell receptor antagonists and partial agonists.* Immunity, 1995. **2**(1): p. 1-11.
7. Chien, Y.H., C. Meyer, and M. Bonneville, *gammadelta T cells: first line of defense and beyond.* Annu Rev Immunol, 2014. **32**: p. 121-55.
8. Crescioli, S., et al., *Antibodies to watch in 2024.* MAbs, 2024. **16**(1): p. 2297450.
9. Nathan, P., et al., *Overall Survival Benefit with Tebentafusp in Metastatic Uveal Melanoma.* N Engl J Med, 2021. **385**(13): p. 1196-1206.
10. Hua, G., D. Carlson, and J.R. Starr, *Tebentafusp-tebn: A Novel Bispecific T-Cell Engager for Metastatic Uveal Melanoma.* J Adv Pract Oncol, 2022. **13**(7): p. 717-723.
11. Baulu, E., et al., *TCR-engineered T cell therapy in solid tumors: State of the art and perspectives.* Sci Adv, 2023. **9**(7): p. eadf3700.
12. Murphy, K. and C. Weaver, *Janeway's immunobiology*. 9th edition. ed. 2016, New York, NY: Garland Science/Taylor & Francis Group, LLC. xx, 904 pages.
13. Muyldermans, S., *Nanobodies: natural single-domain antibodies.* Annu Rev Biochem, 2013. **82**: p. 775-97.
14. Flajnik, M.F., N. Deschacht, and S. Muyldermans, *A case of convergence: why did a simple alternative to canonical antibodies arise in sharks and camels?* PLoS Biol, 2011. **9**(8): p. e1001120.
15. Pancer, Z., et al., *Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey.* Nature, 2004. **430**(6996): p. 174-80.
16. Alder, M.N., et al., *Diversity and function of adaptive immune receptors in a jawless vertebrate.* Science, 2005. **310**(5756): p. 1970-3.
17. Pancer, Z. and M.D. Cooper, *The evolution of adaptive immunity.* Annu Rev Immunol, 2006. **24**: p. 497-518.
18. Velikovsky, C.A., et al., *Structure of a lamprey variable lymphocyte receptor in complex with a protein antigen.* Nat Struct Mol Biol, 2009. **16**(7): p. 725-30.
19. Zhao, Y., C. Niu, and J. Cui, *Gamma-delta (gammadelta) T cells: friend or foe in cancer development?* J Transl Med, 2018. **16**(1): p. 3.
20. Wong, W.K., J. Leem, and C.M. Deane, *Comparative Analysis of the CDR Loops of Antigen Receptors.* Front Immunol, 2019. **10**: p. 2454.
21. North, B., A. Lehmann, and R.L. Dunbrack, Jr., *A new clustering of antibody CDR loop conformations.* J Mol Biol, 2011. **406**(2): p. 228-56.

22. Janeway, C., *Immunobiology : the immune system in health and disease*. 6th ed. 2005, New York: Garland Science. xxiii, 823 p.
23. Tiller, T., et al., *Autoreactivity in human IgG+ memory B cells.* Immunity, 2007. **26**(2): p. 205-13.
24. Sajadi, M.M., et al., *Identification of Near-Pan-neutralizing Antibodies against HIV-1 by Deconvolution of Plasma Humoral Responses.* Cell, 2018. **173**(7): p. 1783-1795 e14.
25. Oyen, D., et al., *Cryo-EM structure of P. falciparum circumsporozoite protein with a vaccine-elicited antibody is stabilized by somatically mutated inter-Fab contacts.* Sci Adv, 2018. **4**(10): p. eaau8529.
26. Dreyfus, C., et al., *Highly conserved protective epitopes on influenza B viruses.* Science, 2012. **337**(6100): p. 1343-8.
27. Barnes, C.O., et al., *SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies.* Nature, 2020. **588**(7839): p. 682-687.
28. Pinto, D., et al., *Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody.* Nature, 2020. **583**(7815): p. 290-295.
29. Wu, N.C., et al., *An Alternative Binding Mode of IGHV3-53 Antibodies to the SARS-CoV-2 Receptor Binding Domain.* Cell Rep, 2020. **33**(3): p. 108274.
30. Ekiert, D.C., et al., *Antibody recognition of a highly conserved influenza virus epitope.* Science, 2009. **324**(5924): p. 246-51.
31. Corti, D., et al., *A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins.* Science, 2011. **333**(6044): p. 850-6.
32. Lang, S., et al., *Antibody 27F3 Broadly Targets Influenza A Group 1 and 2 Hemagglutinins through a Further Variation in VH1-69 Antibody Orientation on the HA Stem.* Cell Rep, 2017. **20**(12): p. 2935-2943.
33. Zhou, T., et al., *Structural definition of a conserved neutralization epitope on HIV-1 gp120.* Nature, 2007. **445**(7129): p. 732-7.
34. Huang, J., et al., *Broad and potent neutralization of HIV-1 by a gp41-specific human antibody.* Nature, 2012. **491**(7424): p. 406-12.
35. Zhou, T., et al., *Structural Repertoire of HIV-1-Neutralizing Antibodies Targeting the CD4 Supersite in 14 Donors.* Cell, 2015. **161**(6): p. 1280-92.
36. Graham, B.S., M.S.A. Gilman, and J.S. McLellan, *Structure-Based Vaccine Antigen Design.* Annu Rev Med, 2019. **70**: p. 91-104.
37. Sesterhenn, F., J. Bonet, and B.E. Correia, *Structure-based immunogen design-leading the way to the new age of precision vaccines.* Curr Opin Struct Biol, 2018. **51**: p. 163-169.
38. Kringelum, J.V., et al., *Structural analysis of B-cell epitopes in antibody:protein complexes.* Mol Immunol, 2013. **53**(1-2): p. 24-34.
39. Sela-Culang, I., V. Kunik, and Y. Ofran, *The structural basis of antibody-antigen recognition.* Front Immunol, 2013. **4**: p. 302.
40. MacCallum, R.M., A.C. Martin, and J.M. Thornton, *Antibody-antigen interactions: contact analysis and binding site topography.* J Mol Biol, 1996. **262**(5): p. 732-45.
41. Lee, J.H., et al., *Structural Features of Antibody-Peptide Recognition.* Front Immunol, 2022. **13**: p. 910367.
42. Lo Conte, L., C. Chothia, and J. Janin, *The atomic structure of protein-protein recognition sites.* J Mol Biol, 1999. **285**(5): p. 2177-98.

43. Vu, K.B., et al., *Comparison of llama VH sequences from conventional and heavy chain antibodies.* Mol Immunol, 1997. **34**(16-17): p. 1121-31.
44. Zavrtanik, U., et al., *Structural Basis of Epitope Recognition by Heavy-Chain Camelid Antibodies.* J Mol Biol, 2018. **430**(21): p. 4369-4386.
45. Mitchell, L.S. and L.J. Colwell, *Analysis of nanobody paratopes reveals greater diversity than classical antibodies.* Protein Eng Des Sel, 2018. **31**(7-8): p. 267-275.
46. Garboczi, D.N., et al., *Structure of the complex between human T-cell receptor, viral peptide and HLA-A2.* Nature, 1996. **384**(6605): p. 134-41.
47. Gowthaman, R. and B.G. Pierce, *TCR3d: The T cell receptor structural repertoire database.* Bioinformatics, 2019. **35**(24): p. 5323-5325.
48. Chen, G., et al., *Sequence and Structural Analyses Reveal Distinct and Highly Diverse Human CD8+ TCR Repertoires to Immunodominant Viral Antigens.* Cell Rep, 2017. **19**(3): p. 569-583.
49. Petersen, J., et al., *T-cell receptor recognition of HLA-DQ2-gliadin complexes associated with celiac disease.* Nat Struct Mol Biol, 2014. **21**(5): p. 480-8.
50. Borbulevych, O.Y., et al., *TCRs Used in Cancer Gene Therapy Cross-React with MART-1/Melan-A Tumor Antigens via Distinct Mechanisms.* Journal of immunology, 2011. **187**(5): p. 2453-63.
51. Raman, M.C., et al., *Direct molecular mimicry enables off-target cardiovascular toxicity by an enhanced affinity TCR designed for cancer immunotherapy.* Sci Rep, 2016. **6**: p. 18851.
52. Mariuzza, R.A., D. Wu, and B.G. Pierce, *Structural basis for T cell recognition of cancer neoantigens and implications for predicting neoepitope immunogenicity.* Front Immunol, 2023. **14**: p. 1303304.
53. Rudolph, M.G., R.L. Stanfield, and I.A. Wilson, *How TCRs bind MHCs, peptides, and coreceptors.* Annu Rev Immunol, 2006. **24**: p. 419-66.
54. Sharon, E., et al., *Genetic variation in MHC proteins is associated with T cell receptor expression biases.* Nat Genet, 2016. **48**(9): p. 995-1002.
55. Sethi, D.K., et al., *A highly tilted binding mode by a self-reactive T cell receptor results in altered engagement of peptide and MHC.* J Exp Med, 2011. **208**(1): p. 91-102.
56. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-242.
57. Kundrotas, P.J., et al., *Templates are available to model nearly all complexes of structurally characterized proteins.* Proc Natl Acad Sci U S A, 2012. **109**(24): p. 9438-41.
58. Vangaveti, S., et al., *Integrating ab initio and template-based algorithms for protein-protein complex structure prediction.* Bioinformatics, 2020. **36**(3): p. 751-757.
59. Szilagyi, A. and Y. Zhang, *Template-based structure modeling of protein-protein interactions.* Curr Opin Struct Biol, 2014. **24**: p. 10-23.
60. Negroni, J., R. Mosca, and P. Aloy, *Assessing the applicability of template-based protein docking in the twilight zone.* Structure, 2014. **22**(9): p. 1356-1362.
61. Pierce, B.G., et al., *ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers.* Bioinformatics, 2014. **30**(12): p. 1771-3.
62. Kozakov, D., et al., *PIPER: an FFT-based protein docking program with pairwise potentials.* Proteins, 2006. **65**(2): p. 392-406.

63. Gabb, H.A., R.M. Jackson, and M.J. Sternberg, *Modelling protein docking using shape complementarity, electrostatics and biochemical information.* J Mol Biol, 1997. **272**(1): p. 106-20.
64. Ritchie, D.W. and V. Venkatraman, *Ultra-fast FFT protein docking on graphics processors.* Bioinformatics, 2010. **26**(19): p. 2398-405.
65. Katchalski-Katzir, E., et al., *Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques.* Proc Natl Acad Sci U S A, 1992. **89**(6): p. 2195-9.
66. de Vries, S.J., M. van Dijk, and A.M. Bonvin, *The HADDOCK web server for data-driven biomolecular docking.* Nat Protoc, 2010. **5**(5): p. 883-97.
67. Comeau, S.R., et al., *ClusPro: a fully automated algorithm for protein-protein docking.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W96-9.
68. Schneidman-Duhovny, D., et al., *PatchDock and SymmDock: servers for rigid and symmetric docking.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W363-7.
69. de Vries, S. and M. Zacharias, *Flexible docking and refinement with a coarse-grained protein model using ATTRACT.* Proteins, 2013. **81**(12): p. 2167-74.
70. Kozakov, D., et al., *The ClusPro web server for protein-protein docking.* Nat Protoc, 2017. **12**(2): p. 255-278.
71. Guest, J.D., et al., *An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants.* Structure, 2021. **29**(6): p. 606-621 e5.
72. Rodrigues, J.P., et al., *Clustering biomolecular complexes by residue contacts similarity.* Proteins, 2012. **80**(7): p. 1810-7.
73. Cheng, T.M., T.L. Blundell, and J. Fernandez-Recio, *pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking.* Proteins, 2007. **68**(2): p. 503-15.
74. Pierce, B. and Z. Weng, *ZRANK: reranking protein docking predictions with an optimized energy function.* Proteins, 2007. **67**(4): p. 1078-86.
75. Pierce, B. and Z. Weng, *A combination of rescoring and refinement significantly improves protein docking performance.* Proteins, 2008. **72**(1): p. 270-9.
76. Moal, I.H., et al., *IRaPPA: information retrieval based integration of biophysical models for protein assembly selection.* Bioinformatics, 2017. **33**(12): p. 1806-1813.
77. Gray, J.J., et al., *Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations.* J Mol Biol, 2003. **331**(1): p. 281-99.
78. Harmalkar, A. and J.J. Gray, *Advances to tackle backbone flexibility in protein docking.* Curr Opin Struct Biol, 2021. **67**: p. 178-186.
79. Ambrosetti, F., et al., *Modeling Antibody-Antigen Complexes by Information-Driven Docking.* Structure, 2020. **28**(1): p. 119-129 e2.
80. van Noort, C.W., R.V. Honorato, and A. Bonvin, *Information-driven modeling of biomolecular complexes.* Curr Opin Struct Biol, 2021. **70**: p. 70-77.
81. Rodrigues, J.P., E. Karaca, and A.M. Bonvin, *Information-driven structural modelling of protein-protein interactions.* Methods Mol Biol, 2015. **1215**: p. 399-424.
82. Nadaradjane, A.A., R. Guerois, and J. Andreani, *Protein-Protein Docking Using Evolutionary Information.* Methods Mol Biol, 2018. **1764**: p. 429-447.

83. Quignot, C., et al., *InterEvDock2: an expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs.* Nucleic Acids Res, 2018. **46**(W1): p. W408-W416.
84. Janin, J., et al., *CAPRI: a Critical Assessment of PRedicted Interactions.* Proteins, 2003. **52**(1): p. 2-9.
85. Lensink, M.F., et al., *Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment.* Proteins, 2021. **89**(12): p. 1800-1823.
86. Lensink, M.F., et al., *Impact of AlphaFold on structure prediction of protein complexes: The CASP15-CAPRI experiment.* Proteins, 2023. **91**(12): p. 1658-1683.
87. Vreven, T., et al., *Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2.* J Mol Biol, 2015. **427**(19): p. 3031-41.
88. Douguet, D., et al., *DOCKGROUND resource for studying protein-protein interfaces.* Bioinformatics, 2006. **22**(21): p. 2612-8.
89. Wodak, S.J., et al., *Critical Assessment of Methods for Predicting the 3D Structure of Proteins and Protein Complexes.* Annu Rev Biophys, 2023. **52**: p. 183-206.
90. AlQuraishi, M., *Machine learning in protein structure prediction.* Curr Opin Chem Biol, 2021. **65**: p. 1-8.
91. Baek, M., et al., *Accurate prediction of protein structures and interactions using a three-track neural network.* Science, 2021. **373**(6557): p. 871-876.
92. Ovchinnikov, S. and P.S. Huang, *Structure-based protein design with deep learning.* Curr Opin Chem Biol, 2021. **65**: p. 136-144.
93. Wang, B. and E.R. Gamazon, *Modeling mutational effects on biochemical phenotypes using convolutional neural networks: application to SARS-CoV-2.* bioRxiv, 2021.
94. Krishnan, S.R., et al., *Accelerating De Novo Drug Design against Novel Proteins Using Deep Learning.* J Chem Inf Model, 2021. **61**(2): p. 621-630.
95. Andrianov, A.M., et al., *Application of deep learning and molecular modeling to identify small drug-like compounds as potential HIV-1 entry inhibitors.* J Biomol Struct Dyn, 2021: p. 1-19.
96. Sverrisson, F., et al., *Fast end-to-end learning on protein surfaces.* bioRxiv, 2020.
97. Gao, W., et al., *Deep Learning in Protein Structural Modeling and Design.* Patterns (N Y), 2020. **1**(9): p. 100142.
98. Zhang, H., et al., *Deep Learning Based Drug Screening for Novel Coronavirus 2019-nCov.* Interdiscip Sci, 2020. **12**(3): p. 368-376.
99. Goh, G.B., N.O. Hodas, and A. Vishnu, *Deep learning for computational chemistry.* J Comput Chem, 2017. **38**(16): p. 1291-1307.
100. Chmiela, S., et al., *Towards exact molecular dynamics simulations with machine-learned force fields.* Nat Commun, 2018. **9**(1): p. 3887.
101. Keith, J.A., et al., *Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems.* Chem Rev, 2021. **121**(16): p. 9816-9872.
102. Palazzesi, F. and A. Pozzan, *Deep Learning Applied to Ligand-Based De Novo Drug Design.* Methods Mol Biol, 2022. **2390**: p. 273-299.
103. Anighoro, A., *Deep Learning in Structure-Based Drug Design.* Methods Mol Biol, 2022. **2390**: p. 261-271.
104. Xu, Y., *Deep Neural Networks for QSAR.* Methods Mol Biol, 2022. **2390**: p. 233-260.

105. Mao, J., et al., *Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models.* iScience, 2021. **24**(9): p. 103052.
106. Joshi, T., H. Pundir, and S. Chandra, *Deep-learning based repurposing of FDA-approved drugs against Candida albicans dihydrofolate reductase and molecular dynamics study.* J Biomol Struct Dyn, 2021: p. 1-17.
107. Madaj, R., et al., *Target2DeNovoDrug: a novel programmatic tool for in silico-deep learning based de novo drug design for any target of interest.* J Biomol Struct Dyn, 2021: p. 1-6.
108. Chen, W., et al., *Predicting Drug-Target Interactions with Deep-Embedding Learning of Graphs and Sequences.* J Phys Chem A, 2021. **125**(25): p. 5633-5642.
109. Zeng, H., et al., *ComplexContact: a web server for inter-protein contact prediction using deep learning.* Nucleic Acids Res, 2018. **46**(W1): p. W432-W437.
110. Quadir, F., et al., *DNCON2_Inter: predicting interchain contacts for homodimeric and homomultimeric protein complexes using multiple sequence alignments of monomers and deep learning.* Sci Rep, 2021. **11**(1): p. 12295.
111. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
112. Jumper, J., et al., *Applying and improving AlphaFold at CASP14.* Proteins, 2021. **89**(12): p. 1711-1721.
113. Guest, J.D., et al., *An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants.* Structure, 2021.
114. AlQuraishi, M., *End-to-End Differentiable Learning of Protein Structure.* Cell Syst, 2019. **8**(4): p. 292-301 e3.
115. Ruidong, W., et al., *High-resolution de novo structure prediction from primary sequence.* bioRxiv, 2022: p. 2022.07.21.500999.
116. Lin, Z., et al., *Evolutionary-scale prediction of atomic-level protein structure with a language model.* Science, 2023. **379**(6637): p. 1123-1130.
117. Humphreys, I.R., et al., *Computed structures of core eukaryotic protein complexes.* Science, 2021: p. eabm4805.
118. Evans, R., et al., *Protein complex prediction with AlphaFold-Multimer.* bioRxiv, 2021.
119. Liu, Y., et al., *Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks.* Cell Syst, 2018. **6**(1): p. 65-74 e3.
120. Gainza, P., et al., *Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning.* Nat Methods, 2020. **17**(2): p. 184-192.
121. Wang, X., et al., *Protein docking model evaluation by 3D deep convolutional neural networks.* Bioinformatics, 2020. **36**(7): p. 2113-2118.
122. Cao, Y. and Y. Shen, *Energy-based graph convolutional networks for scoring protein docking models.* Proteins, 2020. **88**(8): p. 1091-1099.
123. Baek, M., et al., *Efficient and accurate prediction of protein structure using RoseTTAFold2.* bioRxiv, 2023: p. 2023.05. 24.542179.
124. Krishna, R., et al., *Generalized biomolecular modeling and design with RoseTTAFold All-Atom.* Science, 2024: p. eadl2528.
125. Google DeepMind AlphaFold, T. and T. Isomorphic Labs, *Performance and structural coverage of the latest, in-development AlphaFold model*. 2023.
126. Ganea, O.-E., et al., *Independent se (3)-equivariant models for end-to-end rigid protein docking.* arXiv preprint arXiv:2111.07786, 2021.

127. McPartlon, M. and J. Xu, *Deep learning for flexible and site-specific protein docking and design.* BioRxiv, 2023: p. 2023.04. 01.535079.
128. Ketata, M.A., et al., *Diffdock-pp: Rigid protein-protein docking with diffusion models.* arXiv preprint arXiv:2304.03889, 2023.
129. McPartlon, M., et al., *LATENTDOCK: Protein-Protein Docking with Latent Diffusion.*
130. Gowthaman, R. and B.G. Pierce, *TCRmodel: high resolution modeling of T cell receptors from sequence.* Nucleic Acids Res, 2018. **46**(W1): p. W396-W401.
131. Klausen, M.S., et al., *LYRA, a webserver for lymphocyte receptor structural modeling.* Nucleic Acids Res, 2015. **43**(W1): p. W349-55.
132. Sircar, A., E.T. Kim, and J.J. Gray, *RosettaAntibody: antibody variable region homology modeling server.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W474-9.
133. Marcatili, P., et al., *Antibody modeling using the prediction of immunoglobulin structure (PIGS) web server.* Nat Protoc, 2014. **9**(12): p. 2771-83.
134. Weitzner, B.D., et al., *Modeling and docking of antibody structures with Rosetta.* Nat Protoc, 2017. **12**(2): p. 401-416.
135. Leem, J., et al., *ABodyBuilder: Automated antibody structure prediction with data-driven accuracy estimation.* MAbs, 2016. **8**(7): p. 1259-1268.
136. Yamashita, K., et al., *Kotai Antibody Builder: automated high-resolution structural modeling of antibodies.* Bioinformatics, 2014. **30**(22): p. 3279-80.
137. Teplyakov, A., et al., *Antibody modeling assessment II. Structures and models.* Proteins, 2014. **82**(8): p. 1563-82.
138. Abanades, B., et al., *ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins.* Commun Biol, 2023. **6**(1): p. 575.
139. Abanades, B., et al., *ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation.* Bioinformatics, 2022. **38**(7): p. 1877-1880.
140. Ruffolo, J.A. and J.J. Gray, *Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies.* Biophysical Journal, 2022. **121**(3): p. 155a-156a.
141. Wong, W.K., et al., *TCRBuilder: multi-state T-cell receptor structure prediction.* Bioinformatics, 2020. **36**(11): p. 3580-3581.
142. Schritt, D., et al., *Repertoire Builder: high-throughput structural modeling of B and T cell receptors.* Molecular Systems Design & Engineering, 2019. **4**(4): p. 761-768.
143. Porter, K.A., et al., *What method to use for protein-protein docking?* Curr Opin Struct Biol, 2019. **55**: p. 1-7.
144. Gowthaman, R. and B.G. Pierce, *Modeling and Viewing T Cell Receptors Using TCRmodel and TCR3d.* Methods Mol Biol, 2020. **2120**: p. 197-212.
145. Sircar, A. and J.J. Gray, *SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models.* PLoS computational biology, 2010. **6**(1): p. e1000644.
146. Brenke, R., et al., *Application of asymmetric statistical potentials to antibody-protein docking.* Bioinformatics, 2012. **28**(20): p. 2608-14.
147. Krawczyk, K., et al., *Antibody i-Patch prediction of the antibody binding site improves rigid local antibody-antigen docking.* Protein Eng Des Sel, 2013. **26**(10): p. 621-9.
148. Pierce, B.G. and Z. Weng, *A flexible docking approach for prediction of T cell receptor-peptide-MHC complexes.* Protein science : a publication of the Protein Society, 2013. **22**(1): p. 35-46.

149. Jensen, K.K., et al., *TCRpMHCmodels: Structural modelling of TCR-pMHC class I complexes.* Sci Rep, 2019. **9**(1): p. 14530.
150. Li, S., et al., *Structural Modeling of Lymphocyte Receptors and Their Antigens.* Methods Mol Biol, 2019. **2048**: p. 207-229.
151. Krammer, F., *SARS-CoV-2 vaccines in development.* Nature, 2020. **586**(7830): p. 516-527.
152. Jiang, S., et al., *Neutralizing antibodies for the treatment of COVID-19.* Nat Biomed Eng, 2020. **4**(12): p. 1134-1139.
153. Simonis, A., et al., *A comparative analysis of remdesivir and other repurposed antivirals against SARS-CoV-2.* EMBO Mol Med, 2021. **13**(1): p. e13105.
154. Wang, Z., et al., *mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants.* Nature, 2021.
155. Starr, T.N., et al., *Prospective mapping of viral mutations that escape antibodies used to treat COVID-19.* Science, 2021. **371**(6531): p. 850-854.
156. Liu, Y., et al., *Neutralizing Activity of BNT162b2-Elicited Serum.* N Engl J Med, 2021.
157. Wu, K., et al., *Serum Neutralizing Activity Elicited by mRNA-1273 Vaccine - Preliminary Report.* N Engl J Med, 2021.
158. Wang, P., et al., *Antibody Resistance of SARS-CoV-2 Variants B.1.351 and B.1.1.7.* Nature, 2021.
159. Madhi, S.A., et al., *Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant.* N Engl J Med, 2021.
160. Greaney, A.J., et al., *Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies.* Cell Host Microbe, 2021.
161. Liu, Z., et al., *Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization.* Cell Host Microbe, 2021.
162. Greaney, A.J., et al., *Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition.* Cell Host Microbe, 2021. **29**(1): p. 44-57 e9.
163. Gowthaman, R., et al., *CoV3D: a database of high resolution coronavirus protein structures.* Nucleic Acids Res, 2021. **49**(D1): p. D282-D287.
164. Rose, P.W., et al., *The RCSB Protein Data Bank: redesigned web site and web services.* Nucleic acids research, 2011. **39**(Database issue): p. D392-401.
165. Leman, J.K., et al., *Macromolecular modeling and design in Rosetta: recent methods and frameworks.* Nat Methods, 2020. **17**(7): p. 665-680.
166. Khatib, F., et al., *Algorithm discovery by protein folding game players.* Proc Natl Acad Sci U S A, 2011. **108**(47): p. 18949-53.
167. Raybould, M.I.J., et al., *CoV-AbDab: the Coronavirus Antibody Database.* Bioinformatics, 2020.
168. Honegger, A. and A. Pluckthun, *Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool.* J Mol Biol, 2001. **309**(3): p. 657-70.
169. Suzuki, R. and H. Shimodaira, *Pvclust: an R package for assessing the uncertainty in hierarchical clustering.* Bioinformatics, 2006. **22**(12): p. 1540-2.
170. Hubbard, S.J. and J.M. Thornton, *NACCESS*. 1993, Department of Biochemistry and Molecular Biology, University College London.

171. McDonald, I.K. and J.M. Thornton, *Satisfying hydrogen bonding potential in proteins.* Journal of molecular biology, 1994. **238**(5): p. 777-93.
172. Wang, Q., et al., *Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2.* Cell, 2020. **181**(4): p. 894-904 e9.
173. Walls, A.C., et al., *Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein.* Cell, 2020. **181**(2): p. 281-292 e6.
174. Tortorici, M.A., et al., *Ultrapotent human antibodies protect against SARS-CoV-2 challenge via multiple mechanisms.* Science, 2020. **370**(6519): p. 950-957.
175. Schoof, M., et al., *An ultrapotent synthetic nanobody neutralizes SARS-CoV-2 by stabilizing inactive Spike.* Science, 2020. **370**(6523): p. 1473-1479.
176. Jones, B.E., et al., *LY-CoV555, a rapidly isolated potent neutralizing antibody, provides protection in a non-human primate model of SARS-CoV-2 infection.* bioRxiv, 2020.
177. Kortemme, T., D.E. Kim, and D. Baker, *Computational alanine scanning of protein-protein interfaces.* Sci STKE, 2004. **2004**(219): p. pl2.
178. Schymkowitz, J., et al., *The FoldX web server: an online force field.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W382-8.
179. Sirin, S., et al., *AB-Bind: Antibody binding mutational database for computational affinity predictions.* Protein Sci, 2016. **25**(2): p. 393-409.
180. Kortemme, T. and D. Baker, *A simple physical model for binding energy hot spots in protein-protein complexes.* Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14116-21.
181. S, O.C., et al., *A Web Resource for Standardized Benchmark Datasets, Metrics, and Rosetta Protocols for Macromolecular Modeling and Design.* PLoS One, 2015. **10**(9): p. e0130433.
182. Wu, D., et al., *Structural basis for oligoclonal T cell recognition of a shared p53 cancer neoantigen.* Nat Commun, 2020. **11**(1): p. 2908.
183. Barlow, K.A., et al., *Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation.* J Phys Chem B, 2018. **122**(21): p. 5389-5399.
184. Smith, C.A. and T. Kortemme, *Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction.* J Mol Biol, 2008. **380**(4): p. 742-56.
185. Altschul, S.F., et al., *Basic Local Alignment Search Tool.* J Mol Biol, 1990. **215**(3): p. 403-410.
186. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. 2016: Springer-Verlag New York.
187. Kassambara, A. and F. Mundt, *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. 2020.
188. Gu, Z., R. Eils, and M. Schlesner, *Complex heatmaps reveal patterns and correlations in multidimensional genomic data.* Bioinformatics, 2016. **32**(18): p. 2847-9.
189. Zost, S.J., et al., *Rapid isolation and profiling of a diverse panel of human monoclonal antibodies targeting the SARS-CoV-2 spike protein.* Nat Med, 2020.
190. Weinreich, D.M., et al., *REGN-COV2, a Neutralizing Antibody Cocktail, in Outpatients with Covid-19.* N Engl J Med, 2021. **384**(3): p. 238-251.
191. Chen, P., et al., *SARS-CoV-2 Neutralizing Antibody LY-CoV555 in Outpatients with Covid-19.* N Engl J Med, 2021. **384**(3): p. 229-237.

192. Tuccori, M., et al., *Anti-SARS-CoV-2 neutralizing monoclonal antibodies: clinical pipeline.* MAbs, 2020. **12**(1): p. 1854149.
193. Yuan, M., et al., *Recognition of the SARS-CoV-2 receptor binding domain by neutralizing antibodies.* Biochem Biophys Res Commun, 2021. **538**: p. 192-203.
194. Barnes, C.O., et al., *Structures of Human Antibodies Bound to SARS-CoV-2 Spike Reveal Common Epitopes and Recurrent Features of Antibodies.* Cell, 2020.
195. Piccoli, L., et al., *Mapping Neutralizing and Immunodominant Sites on the SARS-CoV-2 Spike Receptor-Binding Domain by Structure-Guided High-Resolution Serology.* Cell, 2020. **183**(4): p. 1024-1042 e21.
196. Zhou, D., et al., *Structural basis for the neutralization of SARS-CoV-2 by an antibody from a convalescent patient.* Nat Struct Mol Biol, 2020.
197. Li, T., et al., *A potent synthetic nanobody targets RBD and protects mice from SARS-CoV-2 infection.* bioRxiv, 2020.
198. Rujas, E., et al., *Multivalency transforms SARS-CoV-2 antibodies into ultrapotent neutralizers.* Nature Communications, 2021. **12**(1): p. 3661.
199. Ahmad, J., et al., *Synthetic nanobody-SARS-CoV-2 receptor-binding domain structures identify distinct epitopes.* bioRxiv, 2021.
200. Du, S., et al., *Structurally Resolved SARS-CoV-2 Antibody Shows High Efficacy in Severely Infected Hamsters and Provides a Potent Cocktail Pairing Strategy.* Cell, 2020. **183**(4): p. 1013-1023 e13.
201. Ju, B., et al., *Human neutralizing antibodies elicited by SARS-CoV-2 infection.* Nature, 2020. **584**(7819): p. 115-119.
202. Liu, H., et al., *Cross-Neutralization of a SARS-CoV-2 Antibody to a Functionally Conserved Site Is Mediated by Avidity.* Immunity, 2020. **53**(6): p. 1272-1280 e5.
203. Lv, Z., et al., *Structural basis for neutralization of SARS-CoV-2 and SARS-CoV by a potent therapeutic antibody.* Science, 2020. **369**(6510): p. 1505-1509.
204. Wrapp, D., et al., *Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation.* Science, 2020. **367**(6483): p. 1260-1263.
205. Yuan, M., et al., *Structural and functional ramifications of antigenic drift in recent SARS-CoV-2 variants.* Science, 2021.
206. Wang, P., et al., *Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7.* Nature, 2021. **593**(7857): p. 130-135.
207. Planas, D., et al., *Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization.* Nature, 2021. **596**(7871): p. 276-280.
208. Altmann, D.M., R.J. Boyton, and R. Beale, *Immunity to SARS-CoV-2 variants of concern.* Science, 2021. **371**(6534): p. 1103-1104.
209. Planas, D., et al., *Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization.* Nature, 2021.
210. Wall, E.C., et al., *AZD1222-induced neutralising antibody activity against SARS-CoV-2 Delta VOC.* Lancet, 2021. **398**(10296): p. 207-209.
211. Holtgrave, D.R., S.H. Vermund, and L.S. Wen, *Potential Benefits of Expanded COVID-19 Surveillance in the US.* JAMA, 2021.
212. Dejnirattisai, W., et al., *The antigenic anatomy of SARS-CoV-2 receptor binding domain.* Cell, 2021.
213. Lensink, M.F., et al., *Blind prediction of interfacial water positions in CAPRI.* Proteins, 2014. **82**(4): p. 620-32.

214. Cerutti, G., et al., *Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite.* Cell Host & Microbe, 2021.
215. McCallum, M., et al., *N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2.* Cell.
216. Rappazzo, C.G., et al., *Broad and potent activity against SARS-like viruses by an engineered human monoclonal antibody.* Science, 2021. **371**(6531): p. 823-829.
217. Tortorici, M.A., et al., *Broad sarbecovirus neutralization by a human monoclonal antibody.* Nature, 2021.
218. Martinez, D.R., et al., *A broadly neutralizing antibody protects against SARS-CoV, pre-emergent bat CoVs, and SARS-CoV-2 variants in mice.* bioRxiv, 2021.
219. Walls, A.C., et al., *Elicitation of Potent Neutralizing Antibody Responses by Designed Protein Nanoparticle Vaccines for SARS-CoV-2.* Cell, 2020. **183**(5): p. 1367-1382 e17.
220. Zhang, B., et al., *A platform incorporating trimeric antigens into self-assembling nanoparticles reveals SARS-CoV-2-spike nanoparticles to elicit substantially higher neutralizing responses than spike alone.* Sci Rep, 2020. **10**(1): p. 18149.
221. Cohen, A.A., et al., *Mosaic nanoparticles elicit cross-reactive immune responses to zoonotic coronaviruses in mice.* Science, 2021. **371**(6530): p. 735-741.
222. Torchala, M., et al., *SwarmDock: a server for flexible protein-protein docking.* Bioinformatics, 2013. **29**(6): p. 807-9.
223. Lensink, M.F., et al., *Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition.* Proteins, 2020. **88**(8): p. 916-938.
224. Marks, D.S., T.A. Hopf, and C. Sander, *Protein structure prediction from sequence variation.* Nat Biotechnol, 2012. **30**(11): p. 1072-80.
225. Kamisetty, H., S. Ovchinnikov, and D. Baker, *Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era.* Proc Natl Acad Sci U S A, 2013. **110**(39): p. 15674-9.
226. Mirdita, M., et al., *ColabFold: making protein folding accessible to all.* Nat Methods, 2022. **19**(6): p. 679-682.
227. Jones, S., A. Marin, and J.M. Thornton, *Protein domain interfaces: characterization and comparison with oligomeric protein interfaces.* Protein Eng, 2000. **13**(2): p. 77-82.
228. Nooren, I.M. and J.M. Thornton, *Structural characterisation and functional significance of transient protein-protein interactions.* J Mol Biol, 2003. **325**(5): p. 991-1018.
229. Dey, S., et al., *The subunit interfaces of weakly associated homodimeric proteins.* J Mol Biol, 2010. **398**(1): p. 146-60.
230. Pierce, B.G., Y. Hourai, and Z. Weng, *Accelerating protein docking in ZDOCK using an advanced 3D convolution library.* PLoS One, 2011. **6**(9): p. e24657.
231. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins, 2004. **57**(4): p. 702-10.
232. Mariani, V., et al., *lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests.* Bioinformatics, 2013. **29**(21): p. 2722-8.
233. Dunbar, J., et al., *SAbDab: the structural antibody database.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1140-6.
234. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421.

235. Zhu, J. and Z. Weng, *FAST: a novel protein structure alignment algorithm.* Proteins, 2005. **58**(3): p. 618-27.
236. Case, D.A., et al., *The Amber biomolecular simulation programs.* J Comput Chem, 2005. **26**(16): p. 1668-88.
237. Bryant, P., G. Pozzati, and A. Elofsson, *Improved prediction of protein-protein interactions using AlphaFold2.* Nat Commun, 2022. **13**(1): p. 1265.
238. Vreven, T., H. Hwang, and Z. Weng, *Integrating atom-based and residue-based scoring functions for protein-protein docking.* Protein science : a publication of the Protein Society, 2011. **20**(9): p. 1576-86.
239. Alford, R.F., et al., *The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design.* J Chem Theory Comput, 2017. **13**(6): p. 3031-3048.
240. Fu, L., et al., *CD-HIT: accelerated for clustering the next-generation sequencing data.* Bioinformatics, 2012. **28**(23): p. 3150-2.
241. Kandathil, S.M., J.G. Greener, and D.T. Jones, *Prediction of interresidue contacts with DeepMetaPSICOV in CASP13.* Proteins, 2019. **87**(12): p. 1092-1099.
242. Canty, A. and B.D. Ripley, *boot: Bootstrap R (S-Plus) Functions.* 2021.
243. Thiele, C. and G. Hirschfeld, *cutpointr: Improved Estimation and Validation of Optimal Cutpoints in R.* Journal of Statistical Software, 2021. **98**(11): p. 1 - 27.
244. Mintseris, J., et al., *Protein-Protein Docking Benchmark 2.0: an update.* Proteins, 2005. **60**(2): p. 214-6.
245. Chen, R., L. Li, and Z. Weng, *ZDOCK: an initial-stage protein-docking algorithm.* Proteins, 2003. **52**(1): p. 80-7.
246. Mintseris, J., et al., *Integrating statistical pair potentials into protein complex prediction.* Proteins, 2007. **69**(3): p. 511-20.
247. Hopf, T.A., et al., *Sequence co-evolution gives 3D contacts and structures of protein complexes.* Elife, 2014. **3**.
248. Evans, R., et al., *Protein complex prediction with AlphaFold-Multimer.* bioRxiv, 2021: p. 2021.10.04.463034.
249. Varadi, M., et al., *AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models.* Nucleic Acids Res, 2022. **50**(D1): p. D439-D444.
250. Robin, X., et al., *pROC: an open-source package for R and S+ to analyze and compare ROC curves.* BMC Bioinformatics, 2011. **12**: p. 77.
251. Wei, R. and J. Wang, *multiROC: Calculating and Visualizing ROC and PR Curves Across Multi-Class assifications.* 2018.
252. Davison, A.C. and D.V. Hinkley, *Bootstrap Methods and Their Applications*. 1997, Cambridge: Cambridge University Press.
253. Lee, S.C., et al., *Design of a binding scaffold based on variable lymphocyte receptors of jawless vertebrates by module engineering.* Proc Natl Acad Sci U S A, 2012. **109**(9): p. 3299-304.
254. Ghani, U., et al., *Improved Docking of Protein Models by a Combination of Alphafold2 and ClusPro.* bioRxiv, 2021: p. 2021.09.07.459290.
255. Roney, J.P. and S. Ovchinnikov, *State-of-the-Art Estimation of Protein Model Accuracy Using AlphaFold.* Phys Rev Lett, 2022. **129**(23): p. 238101.
256. Rossjohn, J., et al., *T cell antigen receptor recognition of antigen-presenting molecules.* Annu Rev Immunol, 2015. **33**: p. 169-200.

257. Rangarajan, S. and R.A. Mariuzza, *T cell receptor bias for MHC: co-evolution or co-receptors?* Cell Mol Life Sci, 2014. **71**(16): p. 3059-68.
258. Liao, H.X., et al., *Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus.* Nature, 2013. **496**(7446): p. 469-76.
259. Doria-Rose, N.A., et al., *Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies.* Nature, 2014. **509**(7498): p. 55-62.
260. Mirdita, M., S. Ovchinnikov, and M. Steinegger, *ColabFold - Making protein folding accessible to all.* bioRxiv, 2021: p. 2021.08.15.456425.
261. Humphreys, I.R., et al., *Computed structures of core eukaryotic protein complexes.* Science, 2021. **374**(6573): p. eabm4805.
262. Chothia, C. and A.M. Lesk, *Canonical structures for the hypervariable regions of immunoglobulins.* J Mol Biol, 1987. **196**(4): p. 901-17.
263. Nelson, A.L., E. Dhimolea, and J.M. Reichert, *Development trends for human monoclonal antibody therapeutics.* Nat Rev Drug Discov, 2010. **9**(10): p. 767-74.
264. Carter, P.J., *Potent antibody therapeutics by design.* Nat Rev Immunol, 2006. **6**(5): p. 343-57.
265. Scott, A.M., J.D. Wolchok, and L.J. Old, *Antibody therapy of cancer.* Nat Rev Cancer, 2012. **12**(4): p. 278-87.
266. Rappuoli, R., et al., *Reverse vaccinology 2.0: Human immunology instructs vaccine antigen design.* J Exp Med, 2016. **213**(4): p. 469-81.
267. Li, Y., et al., *X-ray snapshots of the maturation of an antibody response to a protein antigen.* Nat Struct Biol, 2003. **10**(6): p. 482-8.
268. Haidar, J.N., et al., *A universal combinatorial design of antibody framework to graft distinct CDR sequences: a bioinformatics approach.* Proteins, 2012. **80**(3): p. 896-912.
269. Hanf, K.J., et al., *Antibody humanization by redesign of complementarity-determining region residues proximate to the acceptor framework.* Methods, 2014. **65**(1): p. 68-76.
270. Georgiou, G., et al., *The promise and challenge of high-throughput sequencing of the antibody repertoire.* Nat Biotechnol, 2014. **32**(2): p. 158-68.
271. Li, Z., et al., *The generation of antibody diversity through somatic hypermutation and class switch recombination.* Genes Dev, 2004. **18**(1): p. 1-11.
272. Yin, R., et al., *Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants.* Protein Sci, 2022. **31**(8): p. e4379.
273. Ruffolo, J.A., et al., *Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies.* Nat Commun, 2023. **14**(1): p. 2389.
274. Wallner, B., *AFsample: improving multimer prediction with AlphaFold using massive sampling.* Bioinformatics, 2023. **39**(9).
275. DeepMind. *AlphaFold v2.3.0 technical note*. 2022; Available from: https://github.com/deepmind/alphafold/blob/main/docs/technical_note_v2.3.0.md.
276. Motmaen, A., et al., *Peptide-binding specificity prediction using fine-tuned protein structure prediction networks.* Proc Natl Acad Sci U S A, 2023. **120**(9): p. e2216697120.
277. Johansson-Akhe, I. and B. Wallner, *Improving peptide-protein docking with AlphaFold-Multimer using forced sampling.* Front Bioinform, 2022. **2**: p. 959160.
278. Dunbar, J. and C.M. Deane, *ANARCI: antigen receptor numbering and receptor classification.* Bioinformatics, 2016. **32**(2): p. 298-300.

279. Martin, A.C. and C.T. Porter. *ProFit Version 3.1*. ProFit Version 3.1. Available online at: http://www.bioinf.org.uk/software/profit/ 2009; Available from: http://www.bioinf.org.uk/software/profit/.
280. Conway, P., et al., *Relaxation of backbone bond geometry improves protein energy landscape modeling.* Protein Sci, 2014. **23**(1): p. 47-55.
281. Vreven, T., H. Hwang, and Z. Weng, *Integrating atom-based and residue-based scoring functions for protein-protein docking.* Protein Sci, 2011. **20**(9): p. 1576-86.
282. Basu, S. and B. Wallner, *DockQ: A Quality Measure for Protein-Protein Docking Models.* PLoS One, 2016. **11**(8): p. e0161879.
283. Akdel, M., et al., *A structural biology community assessment of AlphaFold2 applications.* Nat Struct Mol Biol, 2022. **29**(11): p. 1056-1067.
284. Stranges, P.B. and B. Kuhlman, *A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds.* Protein Sci, 2013. **22**(1): p. 74-82.
285. Kappler, K. and T. Hennet, *Emergence and significance of carbohydrate-specific antibodies.* Genes Immun, 2020. **21**(4): p. 224-239.
286. Xu, J. and Y. Zhang, *How significant is a protein structure similarity with TM-score = 0.5?* Bioinformatics, 2010. **26**(7): p. 889-95.
287. Wallner, B., *Improved multimer prediction using massive sampling with AlphaFold in CASP15.* Proteins, 2023. **91**(12): p. 1734-1746.
288. Björn, W., *AFsample: Improving Multimer Prediction with AlphaFold using Aggressive Sampling.* bioRxiv, 2023: p. 2022.12.20.521205.
289. Sala, D., et al., *Modeling conformational states of proteins with AlphaFold.* Curr Opin Struct Biol, 2023. **81**: p. 102645.
290. Ziyao, L., et al., *Uni-Fold: An Open-Source Platform for Developing Protein Folding Models beyond AlphaFold.* bioRxiv, 2022: p. 2022.08.04.502811.
291. Gustaf, A., et al., *OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization.* bioRxiv, 2022: p. 2022.11.20.517210.
292. Zheng, W., et al., *Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14.* Proteins, 2021. **89**(12): p. 1734-1751.
293. Bo, C., et al., *Improved the Protein Complex Prediction with Protein Language Models.* bioRxiv, 2022: p. 2022.09.15.508065.
294. Ruffolo, J.A., J.J. Gray, and J. Sulam, *Deciphering antibody affinity maturation with language models and weakly supervised learning.* arXiv preprint arXiv:2112.07782, 2021.
295. Olsen, T.H., I.H. Moal, and C.M. Deane, *AbLang: an antibody language model for completing antibody sequences.* Bioinform Adv, 2022. **2**(1): p. vbac046.
296. Hie, B.L., et al., *Efficient evolution of human antibodies from general protein language models.* Nat Biotechnol, 2024. **42**(2): p. 275-283.
297. Moss, P., *The T cell immune response against SARS-CoV-2.* Nat Immunol, 2022. **23**(2): p. 186-193.
298. Yin, Y., Y. Li, and R.A. Mariuzza, *Structural basis for self-recognition by autoimmune T-cell receptors.* Immunological reviews, 2012. **250**(1): p. 32-48.
299. Yang, X., et al., *Autoimmunity-associated T cell receptors recognize HLA-B*27-bound peptides.* Nature, 2022. **612**(7941): p. 771-777.

300. Tran, E., et al., *T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer.* N Engl J Med, 2016. **375**(23): p. 2255-2262.
301. Zacharakis, N., et al., *Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer.* Nat Med, 2018. **24**(6): p. 724-730.
302. Lowe, K.L., et al., *Novel TCR-based biologics: mobilising T cells to warm 'cold' tumours.* Cancer Treat Rev, 2019. **77**: p. 35-43.
303. Song, I., et al., *Broad TCR repertoire and diverse structural solutions for recognition of an immunodominant CD8(+) T cell epitope.* Nat Struct Mol Biol, 2017. **24**(4): p. 395-406.
304. Pierce, B.G., et al., *Computational design of the affinity and specificity of a therapeutic T cell receptor.* PLoS Comput Biol, 2014. **10**(2): p. e1003478.
305. Malecek, K., et al., *Specific increase in potency via structure-based design of a TCR.* J Immunol, 2014. **193**(5): p. 2587-99.
306. Qi, Q., et al., *Diversity and clonal selection in the human T-cell repertoire.* Proc Natl Acad Sci U S A, 2014. **111**(36): p. 13139-44.
307. Pai, J.A. and A.T. Satpathy, *High-throughput and single-cell T cell receptor sequencing technologies.* Nat Methods, 2021. **18**(8): p. 881-892.
308. Hudson, D., et al., *Can we predict T cell specificity with digital biology and machine learning?* Nat Rev Immunol, 2023: p. 1-11.
309. Bradley, P., *Structure-based prediction of T cell receptor:peptide-MHC interactions.* Elife, 2023. **12**.
310. Karnaukhov, V.K., et al., *Predicting TCR-peptide recognition based on residue-level pairwise statistical potential.* bioRxiv, 2022: p. 2022.02.15.480516.
311. Cock, P.J., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-3.
312. Bagaev, D.V., et al., *VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium.* Nucleic Acids Res, 2020. **48**(D1): p. D1057-D1062.
313. Lensink, M.F. and S.J. Wodak, *Docking, scoring, and affinity prediction in CAPRI.* Proteins, 2013. **81**(12): p. 2082-95.
314. Van Rhijn, I., et al., *Lipid and small-molecule display by CD1 and MR1.* Nat Rev Immunol, 2015. **15**(10): p. 643-54.
315. Lineburg, K.E., et al., *CD8(+) T cells specific for an immunodominant SARS-CoV-2 nucleocapsid epitope cross-react with selective seasonal coronaviruses.* Immunity, 2021. **54**(5): p. 1055-1065 e5.
316. Malekzadeh, P., et al., *Neoantigen screening identifies broad TP53 mutant immunogenicity in patients with epithelial cancers.* J Clin Invest, 2019. **129**(3): p. 1109-1114.
317. Wu, D., et al., *T cell receptors employ diverse strategies to target a p53 cancer neoantigen.* J Biol Chem, 2022. **298**(3): p. 101684.
318. Duan, Z. and M. Ho, *T-Cell Receptor Mimic Antibodies for Cancer Immunotherapy.* Mol Cancer Ther, 2021. **20**(9): p. 1533-1541.
319. Anishchenko, I., et al., *De novo protein design by deep network hallucination.* Nature, 2021. **600**(7889): p. 547-552.
320. Dauparas, J., et al., *Robust deep learning-based protein sequence design using ProteinMPNN.* Science, 2022. **378**(6615): p. 49-56.

321. Olechnovic, K., et al., *Prediction of protein assemblies by structure sampling followed by interface-focused scoring.* Proteins, 2023. **91**(12): p. 1724-1733.
322. Srivastava, N., et al., *Dropout: a simple way to prevent neural networks from overfitting.* J. Mach. Learn. Res., 2014. **15**(1): p. 1929–1958.
323. Gal, Y. and Z. Ghahramani, *Dropout as a Bayesian approximation: representing model uncertainty in deep learning*, in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. 2016, JMLR.org: New York, NY, USA. p. 1050–1059.
324. Broering, T.J., et al., *Identification and characterization of broadly neutralizing human monoclonal antibodies directed against the E2 envelope glycoprotein of hepatitis C virus.* J Virol, 2009. **83**(23): p. 12473-82.
325. Zhou, P., et al., *A human antibody reveals a conserved site on beta-coronavirus spike proteins and confers protection against SARS-CoV-2 infection.* Sci Transl Med, 2022. **14**(637): p. eabi9215.
326. Sauer, M.M., et al., *Structural basis for broad coronavirus neutralization.* Nat Struct Mol Biol, 2021. **28**(6): p. 478-486.
327. Basi, G.S., et al., *Structural correlates of antibodies associated with acute reversal of amyloid beta-related behavioral deficits in a mouse model of Alzheimer disease.* J Biol Chem, 2010. **285**(5): p. 3417-27.
328. Ofek, G., et al., *Elicitation of structure-specific antibodies by epitope scaffolds.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(42): p. 17880-7.
329. Correia, B.E., et al., *Proof of principle for epitope-focused vaccine design.* Nature, 2014. **507**(7491): p. 201-6.
330. Tsaban, T., et al., *Harnessing protein folding neural networks for peptide-protein docking.* Nat Commun, 2022. **13**(1): p. 176.
331. Yin, R., et al., *TCRmodel2: high-resolution modeling of T cell receptor recognition using deep learning.* Nucleic Acids Res, 2023. **51**(W1): p. W569-W576.
332. Singh, N.K., et al., *An Engineered T Cell Receptor Variant Realizes the Limits of Functional Binding Modes.* Biochemistry, 2020. **59**(43): p. 4163-4175.
333. Beringer, D.X., et al., *T cell receptor reversed polarity recognition of a self-antigen major histocompatibility complex.* Nat Immunol, 2015. **16**(11): p. 1153-61.
334. Fandi, W., et al., *Fast and accurate modeling and design of antibody-antigen complex using tFold.* bioRxiv, 2024: p. 2024.02.05.578892.
335. Zheng, W., et al., *Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data.* Nat Methods, 2024. **21**(2): p. 279-289.
336. Wang, X., S.T. Flannery, and D. Kihara, *Protein Docking Model Evaluation by Graph Neural Networks.* Front Mol Biosci, 2021. **8**: p. 647915.
337. Geng, C., et al., *iScore: a novel graph kernel-based function for scoring protein-protein docking models.* Bioinformatics, 2020. **36**(1): p. 112-121.