Tool Development for Studying and Manipulating Peptide-MHC Interactions in a Globally-Representative Manner

by

Brooke D. Huisman

B.S.E Biomedical Engineering
University of Michigan, 2017

Submitted to the Department of Biological Engineering
in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY IN BIOLOGICAL ENGINEERING
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MAY 2022

Signature of Author: _____
Department of Biological Engineering
April 8, 2022

Certified by: _____
Michael E. Birnbaum, Ph.D.
Associate Professor of Biological Engineering
Thesis Supervisor

Accepted by: _____
Katharina Ribbeck, Ph.D.
Professor of Biological Engineering
Graduate Program Committee Chair, Biological Engineering

THESIS COMMITTEE

Michael E. Birnbaum, Ph.D. (Thesis Supervisor)
Associate Professor of Biological Engineering
Massachusetts Institute of Technology

Forest M. White, Ph.D. (Thesis Chair)
Professor of Biological Engineering
Massachusetts Institute of Technology

Catherine J. Wu, M.D.
Professor of Medicine
Dana-Farber Cancer Institute and Harvard Medical School

Tool Development for Studying and Manipulating Peptide-MHC Interactions in a Globally-Representative Manner

by

Brooke D. Huisman

Submitted to the Department of Biological Engineering
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Biological Engineering

ABSTRACT
Major histocompatibility complex (MHC) proteins play a critical role in the adaptive immune system, presenting peptide fragments on the surface of cells for surveillance by T cells. In this way, T cells are able to sense cellular dysfunction associated with disease, such as the presence of pathogen-derived peptides. The ability to assess and predict peptide-MHC binding is, therefore, an important component of understanding and engineering immune responses. Peptide-MHC binding is complex, in part due to the immense diversity on both sides of the interaction: MHCs are encoded by the most polymorphic genes in the body and can bind to a subset of trillions of potential peptides. Further, MHC alleles have not been uniformly studied, which presents challenges when designing therapies for diverse patient populations. In this work, we develop tools to study and manipulate peptide-MHC interactions in a more globally-representative manner. First, we study highly polymorphic class II MHC alleles, utilizing data from high-throughput yeast display screens to train algorithms for antigen prediction. Next, we adapt the yeast display platform to screen user-defined libraries of peptides and apply the approach for optimizing peptides and profiling whole viral pathogens for MHC binding. To further increase the MHC throughput of these approaches, we develop a second-generation platform that opens the pipeline for MHC alleles. Finally, we take an orthogonal approach to studying peptide-MHC binding in a representative manner, studying the highly conserved, class Ib MHC HLA-E. We characterize the HLA-E peptide repertoire and train prediction algorithms to identify novel proteome-derived binders. Taken together, these works advance our toolset for studying peptide-MHC interactions across patient populations, with applications in infectious disease, cancer, and autoimmunity.

Thesis Supervisor: Michael E. Birnbaum, Ph.D.
Title: Associate Professor of Biological Engineering

ACKNOWLEDGEMENTS

During my time at MIT, I have been fortunate to be surrounded by a collection of amazing mentors, colleagues, collaborators, and friends. I'm grateful for the all opportunities I have had to learn from them.

First, I extend my deepest thanks to my thesis advisor, Michael Birnbaum. I'm grateful for the opportunity to train in Michael's group and under his mentorship as I navigated the path from starting as a new graduate student to writing and defending a PhD thesis. Michael is a brilliant scientist, who was always willing to dig into the nuance of experiments and data and who cares deeply about tackling important challenges to advance our understanding of biology and better human health. He has been incredibly generous with his time and talents, from offering his time on holidays for in-depth conversations about science to making himself available despite occasional twelve-hour time differences. I very much valued his transparency and willingness to share and discuss his decision-making processes and philosophies. In addition to his abilities as a scientist, Michael has been selfless, kind, and generous, caring deeply about his trainees as whole people and advocating tirelessly for them at every juncture. He has been a constant source of support, encouraging me and his other trainees to pursue our interests and passions in and out of the lab. I'm grateful to have had, and continue to have, a mentor like Michael in my journey through graduate school and beyond.

I am also thankful for the care Michael puts into assembling the team of people in his group. The Birnbaum lab, past and present, has been a wonderful environment for learning and scientific exploration. Thank you to Connor Dobson for being my rotation mentor and a trusted sounding board and friend throughout graduate school. Thank you to Garrett Rappazzo for teaching me much about class II MHCs and sharing his reagents and his knowledge with me. Thank you to Connor, Garrett, Beth Grace, Patrick Holec, Khloe Gordon, Christine Devlin, and Taeyoon Kyung for further assisting my start in the lab. Thank you to Nishant Singh for sharing his enthusiasm and knowledge about proteins, and to Lucy Walters for the many exciting and helpful conversations about HLA-E. Thank you to Jon Krog for being a fun and easygoing bench-mate and for many engaging conversations about science and beyond. Thank you to Christine, Anna Reich, Ruth Tadese, and Isadora Deese for continual support and for keeping the lab running. I've been fortunate to mentor two amazing undergraduate students, Ning (Alexa) Guan and Marcos Labrado. Thank you to Alexa for her faith to start working with me when I was only a first-year graduate student. The first summer and the subsequent two years working with Alexa were a lot of fun, and I learned much along the journey; I'm grateful to have her in my life still. Thank you to Marcos for joining me when I was a fourth-year graduate student and for his continual optimism, flexibility, and inquisitiveness. I am excited to see all he does next. Thank you to Pallavi Balivada for her trust in my mentorship and for her enthusiasm to carry forward some of the projects described in this thesis. Thank you to the entirety of the Birnbaum lab, including the folks I haven't listed yet, Stephanie Gaglione, Ellen Kim, Caleb Perez, Blake Smith, Katie Breuckman, Jiayi Dong, and Allison Lemmer, for sharing their enthusiasm about immunology, viruses, computers, CAR-T cells, engineering, and so much more. Thank you for your friendship and support over the years.

I am also immensely grateful for the support of my thesis committee members, Forest M. White and Catherine (Cathy) J. Wu. Thank you to Forest for serving as my committee chair and for continual support in this role and through all of our interactions in the department. Thank you for being a trusted sounding board. Thank you to Cathy for her enduring support and for bringing her scientific inquisitiveness and perspective to my committee meetings. I'm grateful for her generosity towards me with her time and talents. Thank you to Forest and Cathy for their support in science, career, and beyond.

I've been fortunate to engage in many fruitful and exciting collaborations. Thank you to David Gifford and his group, especially Haoyang Zeng and Zheng Dai. Thank you to Geraline Gillespie and Andrew McMichael at Oxford and their groups, especially Lee Garner. Thank you to Chiara Romagnani and Timo Rückert from her group at the Deutsche Rheuma-Forschungszentrum. Thank you to the many wonderful individuals in the BE Academic Office and graduate program and department leaders for making this program what it is and for continually championing students. I am also grateful for financial support, including the NSF Graduate Research Fellowship.

Thank you as well to my undergraduate research advisors and mentors who sparked in me an interest in academic research, including Sundeep Kalantry and Clair Harris at University of Michigan, and Masoud Hoore, Dmitry Fedosov, and Gerhard Gompper at Forschungszentrum Jülich.

I'm grateful to be a member of the MIT Biological Engineering graduate cohort of 2017. This is an outstanding group of people to call colleagues and friends, each a merger of brilliance, warmth, and character. Thank you for the community, friendship, and adventures shared.

I'm grateful to have been part of several other MIT communities during my time here. Thank you to the MIT Women's Volleyball Club for bringing some balance to my life and providing a great venue to connect with talented and interesting people outside of the academic setting. Thank you to the MIT Warehouse residence hall community for providing a welcoming home for my first two years on campus. I was also fortunate to serve as a graduate residential advisor (GRA) in Baker Hall during my last three years at MIT. Thank you to my undergraduate residents for allowing me to serve as their GRA, and to my fellow GRAs and the Baker House leadership team for nurturing this wonderful community.

I've also been fortunate to be surrounded by fantastic friends throughout my time at MIT. I'll include one anecdote here that, for me, really captures their generosity and care. During March 2020, the COVID-19 pandemic brought the world to a halt, including shutting down MIT's campus. It was a time of great uncertainty about what would happen next. Because of my role as a GRA and living on campus, this uncertainty extended to whether I could remain there during the shutdown. Without any prompting or even mention of this concern, I had six MIT friends offer me a place to sleep on their couches, in their spare rooms, or on spot on their floor if I needed it. While I thankfully didn't need to take them up on the offer, these offers meant more than I can say and really characterize the spirit and generosity of the people I've been grateful to be surrounded by and call my friends. Thank you for your unwavering friendship.

Finally, I'd like to thank my family and the friends I consider family. They have been constants throughout my life and sources of perspective and balance. Above all, I thank my parents, Elizabeth and Warren, and sister, Lauren, who have always been my biggest supporters and trusted friends. Words cannot express how thankful I am for them: thank you for always taking my calls no matter the hour, visiting me often, and being the place I call home. Thank you for always leading by example and for teaching me to love learning. We've had incredible adventures together, and I look forward to so many more to come. Thank you for all the good things in my life.

TABLE OF CONTENTS

1.1 T cells are an important component of the adaptive immune system ... 9

1.2 MHCs present peptides for T cell surveillance and are highly polymorphic and diverse among global populations ... 10

1.3 T cell-directed therapeutics are powerful tools to combat disease ... 11

1.4 Thesis overview and motivation ... 12

CHAPTER 2: UTILIZING PEPTIDE-MHC-II DATA FROM HIGH-THROUGHPUT YEAST DISPLAY EXPERIMENTS FOR IMPROVED ANTIGEN PREDICTION ... **14**

2.1 Introduction ... 14

2.2 Yeast display platform for identifying peptide binders of MHC-II proteins ... 15

2.3 Training prediction algorithms ... 18

2.4 Generating test datasets and benchmarking performance ... 19

2.5 Improvements in affinity prediction and prediction of clinically relevant peptides ... 24

2.6 Discussion and outlook ... 26

2.7 Methods ... 27

2.8 Acknowledgements ... 31

CHAPTER 3: MACHINE LEARNING OPTIMIZATION AND YEAST DISPLAY ASSESSMENT OF PEPTIDES FOR ENHANCED MHC-II BINDING ... **32**

3.1 Introduction ... 32

3.2 Approach: model, model training, selection of seed sequences, and definition of optimization tasks ... 33

3.3 Assessment of optimized peptides using yeast display ... 37

3.4 Round survival rates of peptides across optimization tasks ... 37

3.5 Optimization for single or multiple alleles ... 40

3.6 Complexity of interactions between residue positions captured by optimizations ... 42

3.7 Effects of invariant flanking residues on optimized peptides ... 44

3.8 Sequence motifs of optimized peptides ... 46

3.9 Discussion and outlook ... 47

3.10 Methods ... 47

3.11 Acknowledgements ... 51

3.12 Supplemental Figures ... 53

CHAPTER 4: A HIGH-THROUGHPUT YEAST DISPLAY APPROACH TO PROFILE PATHOGEN PROTEOMES FOR MHC-II BINDING ... **58**

4.1 Introduction ... 58

4.2 Generation of yeast display libraries for profiling the SARS-CoV-2 proteome ... 59

4.3 Strategies for selecting defined libraries ... 62

4.4 Analysis of selection data ... 62

4.5 Sequence motifs of enriched peptides and considerations for epitope identification experiments ... 66

4.6 Examining relationships between MHC-specific binding and spike proteins from SARS-CoV-2 and SARS-CoV ... 69
4.7 Identifying peptide binders missed by computational prediction ... 71
4.8 Comparing whole dengue serotype proteomes for common MHC-binding peptides ... 74
4.9 Discussion and outlook ... 75
4.10 Acknowledgements ... 81
4.11 Supplemental Figures ... 82

CHAPTER 5: SECOND-GENERATION PLATFORM FOR INCREASING MHC-II THROUGHPUT IN YEAST DISPLAY SCREENS ... **86**
5.1 Introduction ... 86
5.2 Design of peptide display libraries and experiments ... 87
5.3 Application of peptide display platform to screen HLA-DR401 ... 91
5.4 Assessing additional MHCs, including HLA-DP and HLA-DQ alleles ... 96
5.5 Adding exogenous peptide as a competitor ... 98
5.6 Discussion and outlook ... 98
5.7 Methods ... 100
5.8 Acknowledgements ... 103

CHAPTER 6: CHARACTERIZING THE PEPTIDE REPERTOIRE OF HLA-E AND NATURAL KILLER CELL RECEPTORS VIA YEAST DISPLAY ... **104**
6.1 Introduction ... 104
6.2 Design and validation of HLA-E for yeast display ... 106
6.3 Selection of HLA-E library ... 107
6.4 Analysis of library-enriched peptides ... 110
6.5 Peptide motifs that differ from the dominant motif ... 115
6.6 Training prediction algorithms and predicting human- and pathogen-derived ligands ... 116
6.7 Validation of HLA-E peptides ... 118
6.8 Discussion and outlook ... 122
6.9 Methods ... 123
6.10 Acknowledgements ... 128

CHAPTER 7: CONCLUSIONS AND FUTURE DIRECTIONS ... **129**

REFERENCES ... **131**

CHAPTER 1: INTRODUCTION

1.1 T cells are an important component of the adaptive immune system

The human immune system is remarkable, capable of responding to diverse pathogens encountered over a lifetime, including a range of bacteria, viruses, fungi, and parasites. To couple speed with specificity and memory, the immune system can be partitioned into two subsets: the innate and adaptive immune systems (Chaplin, 2010). The innate immune system relies on pattern recognition of a conserved set of molecular signals indicative of disease, allowing for rapid deployment to fight pathogens. In contrast, the adaptive immune system, while slower to react, provides a more specific response and can provide immunological memory, such that repeated infections with a given pathogen will provide a more robust and rapid response (Bonilla and Oettgen, 2010; Chaplin, 2010).

In order to recognize and respond to diverse threats with specificity, the adaptive immune system is characterized by a high degree of molecular diversity (Birnbaum et al., 2012; Bonilla and Oettgen, 2010). Two main cell types comprise the adaptive immune system, B cells and T cells, which each detect pathogenic ligands via a repertoire of diverse receptors, called B cell receptors (BCRs) and T cell receptors (TCRs), respectively. BCRs and TCRs are generated through recombination of germline-encoded genes to produce a diverse repertoire of receptors (Chaplin, 2010). B cells can undergo a process of somatic hypermutation to increase their receptors' affinity for their targets, and B cells can secrete their receptors as antibodies (Bonilla and Oettgen, 2010). T cells, in contrast, are unable to affinity mature or secrete their TCRs. Instead, via their TCRs, T cells recognize peptides presented on the cell surface by the specialized immune proteins, Major Histocompatibility Complex (MHC) proteins (Wucherpfennig, 2010), allowing them to react to dangers within a cell. T cells further play a role in coordinating and tuning immune responses (Gonzalez et al., 2018; Swain et al., 2012).

T cells balance opposing needs for specificity and cross-reactivity and are challenging to study in part because of their sequence diversity and low affinities. T cells bind their peptide-MHC (pMHC) ligands with affinities on the order of 1-100 μM, which is much weaker than antibody affinities for their ligands (Gee et al., 2018a; Matsui et al., 1991). Despite this low affinity, even few interactions between TCR molecules on a given cell and its pMHC ligands can elicit a response (Irvine et al., 2002). The potential sequence space of TCRs is massive, with an estimated 10^{15} possible sequences (Birnbaum et al., 2012; Davis and Bjorkman, 1988). Of these, a given person has on the order of ten million unique TCR clones (Arstila et al., 1999; Robins et al., 2010). Each TCR in turn recognizes thousands to millions of peptide sequences (Birnbaum et al., 2014; Wooldridge et al., 2012), and the frequency of TCRs in a person changes over time, as T cell clones expand after antigen exposure and activation (Dash et al., 2017; Kumar et al., 2018).

T cells can be grouped into subsets based on the molecules on their surface. Firstly, T cells can be partitioned based on the heterodimeric TCR chains they express. The majority of T cells express αβ TCRs, although a subset express γδ TCRs which recognize non-peptide antigens presented on non-classical MHC molecules and MHC-related molecules (Bonilla and Oettgen, 2010). αβ T cells are subdivided into $CD4^+$ and $CD8^+$ T cells, based on expression of the CD4 or CD8 co-receptor, each of which recognizes a different class of MHC protein (Chaplin, 2010). Namely, CD4 can bind to class II MHCs (MHC-II) and CD8 can bind to class I MHCs (MHC-I). $CD8^+$ T cells

canonically act as 'killer' T cells, with cytotoxic activity against infected or transformed cells. Conversely, $CD4^+$ T cells act as 'helper' T cells, helping coordinate and modulate immune responses, with a subset playing a critical role in enabling antibody affinity maturation (Bonilla and Oettgen, 2010; Chaplin, 2010; Crotty, 2019).

Another class of cells, called Natural Killer (NK) cells, primarily act as innate immune cells, but are thought to serve as a bridge between innate and adaptive immune responses (Sun and Lanier, 2009). NK cells express germline encoded receptors (Carrillo-Bustamante et al., 2016; Chaplin, 2010), including NKG2A, NKG2C, NKG2E, and NKG2D, as well as killer immunoglobulin-like receptors (KIRs) (Carrillo-Bustamante et al., 2016). These receptors confer the ability to bind to a more constrained set of ligands, and NK cells are capable of recognizing missing MHC expression (Carrillo-Bustamante et al., 2016). Finally, NK-T cells share some characteristics of both NK cells and T cells, typically bearing NK cell markers, as well as an αβ TCR that recognizes glycolipids presented by non-classical MHC-like molecules (Balato et al., 2009).

1.2 MHCs present peptides for T cell surveillance and are highly polymorphic and diverse among global populations

MHCs, which present peptides on the cell surface for T cell surveillance, are highly diverse, contributing a further to the molecular diversity of the adaptive immune system. MHCs in humans, referred to as human leukocyte antigens (HLAs), are encoded on chromosome 6 and represent the most variable region in the human genome (Blackwell et al., 2009). This large degree of variation is crucial for presentation of diverse peptides from potential threats, which is critical to prevent gaps in presentation that could otherwise render T cells oblivious to an ongoing infection or pathogenesis.

The two classes of MHC proteins, MHC-I and MHC-II, are capable of being recognized by $CD8^+$ and $CD4^+$ T cells, respectively. These classes of MHC proteins are present on different populations of cells. Essentially all nucleated cells in the body present MHC-I proteins. In contrast, MHC-II molecules are primarily on more specialized antigen presenting cells, such as dendritic cells, macrophages, and B cells (Piertney and Oliver, 2006; Sommer, 2005). MHC-I and MHC-II also follow distinct processing pathways to reach the cell surface, resulting in presentation of peptides drawn from different pools of peptides, processed in different ways. Broadly, and with a number of exceptions, MHC-I proteins present fragments from inside of the cell, and MHC-II proteins present peptides sampled from outside of the cell, such as through endocytosis (Sommer, 2005; Vyas et al., 2008).

Peptides bind to MHC proteins in a defined manner. Specifically, MHCs have two parallel alpha helices, and the groove between them accommodates peptide binding and is referred to as the peptide binding groove (Wieczorek et al., 2017). The MHC-I and MHC-II grooves are distinct. The MHC-I groove is made of the α1 and α2 domains of the MHC protein heavy α-chain (Wieczorek et al., 2017). MHC-I utilizes an invariant structural component, called Beta-2-microglobulin (β2M), for folding and stability. The ends of the MHC-I groove are closed, with the peptide sitting within the groove; peptide lengths typically range from 8-10 amino acids (Wieczorek et al., 2017). In contrast, the MHC-II groove is a composite of two chains, α and β, and is open at either ends of the groove. The open groove accommodates longer peptides, typically

in the range of 13-25 amino acids, with a 9 amino acid 'core' sitting in the groove, which is the primary determinant of binding affinity (Jones et al., 2006; Stern, 1994; Wieczorek et al., 2017).

Humans each express a set of at least six canonical HLA genes from each copy of chromosome 6. In humans, canonical MHC-I proteins are HLA-A, HLA-B, and HLA-C. Additional non-canonical MHCs include HLA-E, HLA-F, and HLA-G, which display limited polymorphism (Heinrichs and Orr, 1990; O'Callaghan and Bell, 1998; Robinson et al., 2015). Canonical MHC-II proteins are HLA-DR, HLA-DP, and HLA-DQ (Neefjes et al., 2011). MHC-II proteins are comprised of an α and β chain. Every chromosome 6 expresses at least one HLA-DR β chain, HLA-DRB1, which pairs with a conserved α domain, HLA-DRA (Marsh et al., 1999). HLA-DRA can also pair with β chains encoded in the HLA-DRB3, HLA-DRB4, or HLA-DRB5 loci, although not every individual has these genes (Marsh et al., 1999). HLA-DQ forms from gene products of HLA-DQA1 and HLA-DQB1, and HLA-DP from HLA-DPA1 and HLA-DPB1 (Choo, 2007). Unlike the HLA-DR α chain, HLA-DP and HLA-DQ α chains are polymorphic. HLA-DQ and -DP molecules can form from α and β chains pairing from the same chromosome or from the opposite chromosomes, further increasing the number of possible HLA molecules in a person. Across these genes, there are over thirty thousand known and named HLA alleles (Robinson et al., 2015), underscoring the diversity across HLA genes and possible HLA haplotypes across individuals.

From large-scale datasets surveying HLA types, we can observe variability in HLA allele frequencies and HLA haplotypes across global regions and populations (Baek et al., 2021; Bhattacharya et al., 2007; Gonzalez-Galarza et al., 2020; Jawdat et al., 2020; Tokić et al., 2020). However, HLA alleles across populations have not historically been equally represented in MHC datasets, such as datasets assessing peptide-MHC binding rules, as described by Abelin et al (Abelin et al., 2019). For example, HLA-DRB1*04:07 is highly represented in certain South and Central America populations (Gonzalez-Galarza et al., 2020), but has only 169 deposited peptide binders in the central Immune Epitope Database, as of March 2022, while a more highly-studied allele HLA-DRB1*01:01 has over 18,000 known binders (Vita et al., 2019). This issue has been especially acute for MHC-II alleles, as our overall understanding of MHC-II alleles has lagged behind MHC-I (Abelin et al., 2019).

Given the critical role of MHC proteins in the adaptive immune response, there are some striking relationships that can be seen between alleles and disease. For example, HLA-B*57 is related to long term non-progressive control of HIV infection (Altfeld et al., 2003). Similarly, in genome-wide association studies, MHC alleles have some of the highest odds-ratios in describing risk for autoimmune diseases (Karnes et al., 2017).

1.3 T cell-directed therapeutics are powerful tools to combat disease

Given the central role of MHCs and T cells in adaptive immunity, there has been great interest in developing therapeutics to target or modulate their activity. One such therapeutic is neoantigen peptide vaccines for treating cancer (Hu et al., 2018). This approach seeks to enhance the response of tumor antigen-specific T cells by vaccinating with peptides encoding cancer-specific mutations, and it has been utilized to treat cancers including melanoma and glioblastoma (Hu et al., 2021; Keskin et al., 2019; Ott et al., 2017; Sahin et al., 2017). Identification and selection of neoantigen targets involves RNA sequencing and/or whole-exome sequencing to identify somatic mutations, HLA-typing to identify a patient's HLA type, and computational prediction to identify mutation-

containing peptides capable of binding a patient's HLAs (Keskin et al., 2019; Ott et al., 2017; Sahin et al., 2017). As such, a critical component of peptide vaccine design is the ability to predict which peptides can bind to a patients' HLAs, underscoring a need for pan-allele data and predictors. After selecting peptide sequences, long peptides are synthesized and administered in peptide pools (Keskin et al., 2019; Ott et al., 2017). Resulting patient outcomes have been promising, including years after disease (Hu et al., 2021).

A second application in T cell-directed therapeutics is checkpoint immunotherapy, in which negative regulators of T cell function are blocked to promote T cell antitumor function. Two prominent examples include anti-cytotoxic T lymphocyte-associated protein 4 (anti-CTLA-4) and anti-programmed cell death 1 (anti-PD-1) therapies (Ribas and Wolchok, 2018). Checkpoint inhibitors have become routinely used in the clinic for certain cancers (Robert et al., 2015). Compared to peptide vaccines, checkpoint inhibition acts to more indiscriminately increase T cell function, which resultingly requires less tailoring for individual patients, making it more broadly accessible. A related example is anti-NKG2A checkpoint inhibition, which disinhibits T and NK cell effector functions, and an anti-NKG2A antibody is currently being assessed in a clinical trial for head and neck carcinoma (André et al., 2018).

A final example of T cell-directed therapies is adoptive transfer of autologous tumor-infiltrating lymphocytes (TIL). In this approach, TILs are expanded *ex vivo* with cytokines such as IL-2 and adoptively transferred back into patient, with the goal of enhancing existing T cell responses (Wu et al., 2012). This approach highlights questions about what the tumor-associated expanded T cells are recognizing. Identifying and understanding their ligands could open new treatment options.

There has been a great deal of interest in understanding T cell function and pMHC targets in the context of disease, with the potential to identify new treatment avenues. Such work has spanned from cancer (Gee et al., 2018b; Grace et al., 2022; Oliveira et al., 2021), to autoimmunity (Jelcic et al., 2018; Kent et al., 2017; Planas et al., 2018), and infectious disease (Ma et al., 2021; Pan et al., 2021). Further, there has been sustained interest in developing new technologies to study peptide-MHC-TCR interactions in greater detail and throughput (Abelin et al., 2017, 2019; Birnbaum et al., 2014; Dobson et al., 2021; Guo and Elledge, 2022; Joglekar et al., 2019; Kula et al., 2019; Reynisson et al., 2020).

1.4 Thesis overview and motivation

In this thesis, I present our work to utilize and build tools to study peptide-MHC binding and their interactions with immune receptors. Given the 1) vast degree of HLA diversity, 2) clinical relevance of studying peptide-MHC interactions, and 3) non-uniform understanding of binding across alleles, we seek to develop tools to study and manipulate peptide-MHC interactions and to do this in a more globally-representative manner.

First, we approach this goal by studying highly polymorphic MHC-II alleles. We utilize data from high-throughput yeast display screens to train prediction algorithms to extrapolate the data for better antigen prediction (**Chapter 2**). These yeast display-trained algorithms improve peptide-MHC affinity prediction and predictions on a systemically-missed subset of peptides, and demonstrate utility in candidate antigen identification.

Next, we adapt the MHC-II platform for designing better MHC-binding peptides, to engineer stronger binders and more compact sets of peptides, with implications in peptide vaccine design (**Chapter 3**). Further, we extend this approach to directly assess pathogen-derived peptides for MHC binding (**Chapter 4**), identifying SARS-CoV-2-derived peptides capable of binding to an MHC of interest that were missed by other state-of-the-art approaches. To increase our throughput of MHC-II alleles, we develop a second-generation platform for peptide-MHC-II binding (**Chapter 5**), with applications in identification of human autoantigens.

Finally, we take an orthogonal approach to understanding peptide-MHC interactions in a globally-representative manner: we investigate the highly-conserved non-canonical MHC HLA-E (**Chapter 6**). We characterize the peptide repertoire of this allele and its cognate NK cell receptors. Utilizing these data, we perform computational predictions to identify novel HLA-E-binders.

Taken together, these works provide new data and advance our suite of tools for studying peptide-MHC interactions across patient populations, with applications in infectious disease, cancer, autoimmunity, and beyond.

CHAPTER 2: UTILIZING PEPTIDE-MHC-II DATA FROM HIGH-THROUGHPUT YEAST DISPLAY EXPERIMENTS FOR IMPROVED ANTIGEN PREDICTION

This chapter contains work conducted in collaboration with C. Garrett Rappazzo and adapted under the Creative Commons license from: Rappazzo, Huisman*, and Birnbaum (2020). "Repertoire-scale determination of class II MHC peptide binding via yeast display improves antigen prediction". Nat. Commun. A full acknowledgements statement is included as* ***Chapter 2.8****.*

Abstract

$CD4^+$ T cells play an important role in the immune response to infection and cancer by recognizing antigenic peptides presented by class II major histocompatibility complexes (MHC-II) on the cell surface. Predicting peptide-MHC binding is a critical component of probing this interaction. While many computational algorithms for predicting peptide binding to MHC-II proteins have been reported, their performance varies greatly. Here we present a yeast display approach for the identification of over an order of magnitude more unique MHC-II binders than comparable approaches and the application of these data for training prediction algorithms, which improves the prediction of peptide-binding affinity and the identification of pathogen- and tumor-associated peptides. Yeast display-generated MHC-II-binding peptide datasets can be used to improve the accuracy of MHC-II binding prediction algorithms, and potentially enhance our understanding of $CD4^+$ T cell recognition.

2.1 Introduction

Computational prediction algorithms are crucial tools for inferring binding between highly polymorphic MHC molecules and peptides drawn from a diverse pool of possible binders. Recent advances have described improvements of computational methods (Chen et al., 2019; O'Donnell et al., 2020; Racle et al., 2019; Reynisson et al., 2020; Zeng and Gifford, 2019) and training data (Abelin et al., 2017, 2019; Sarkizova et al., 2020). While these advances have benefited antigen prediction for both class I (MHC-I) and class II MHCs (MHC-II), there has been sustained interest in improving the performance of MHC-II prediction algorithms (2017), which frequently under-perform their MHC-I counterparts (Andreatta et al., 2018; Jensen et al., 2018; Lin et al., 2008; Nielsen et al., 2010; Wang et al., 2008; Zhao and Sher, 2018).

The under-performance of MHC-II prediction algorithms has been at least partially due to a relative paucity of peptide-binding data (Vita et al., 2019), as under-performance is particularly pronounced for MHC-II alleles with few reported binders (Nielsen et al., 2010; Zhao and Sher, 2018). However, peptide binding predictions for even well-characterized MHC-II alleles have under-performed their MHC-I counterparts (Wang et al., 2008; Zhao and Sher, 2018). This is likely due to challenges inherent to class II MHCs, which have more degenerate peptide-binding motifs than their class I counterparts (Alvarez et al., 2018), and an open peptide-binding groove that requires an added algorithmic step of peptide-register determination (Jones et al., 2006; Nielsen and Lund, 2009; Nielsen et al., 2010; Stern, 1994). Additionally, publicly available MHC-II-binding peptide datasets contain redundant nested peptide sets and single amino-acid variants of well-characterized peptides, potentially limiting their effective depth and generalizability (Rammensee et al., 1999; Vita et al., 2019). Therefore, we hypothesize that the under-performance

of MHC-II prediction algorithms has been driven by deficiencies in their underlying training data, and can be ameliorated by higher-quality peptide datasets.

Here, we describe a yeast display-based platform to screen 10^8 peptides for MHC-II binding, generating over an order of magnitude more unique binders than comparable approaches for two human MHC-II alleles, and apply these data for training prediction algorithms. Yeast display-trained models perform comparably with other state-of-the-art models, but outperform on subsets of peptides systematically missed by other algorithms. Additionally, yeast display-trained models improve the prediction of peptide-binding affinity for pathogen- and tumor-associated peptides, even when compared to recently described mass spectrometry-based approaches. Collectively, these data show the importance of large datasets of unique peptide binders to improve MHC-II binding prediction, and suggest yeast display-generated data can potentially facilitate better understanding of $CD4^+$ T cell recognition and enhance patient benefit from antigen-targeted therapeutics.

2.2 Yeast display platform for identifying peptide binders of MHC-II proteins

Yeast-displayed MHC-II constructs have been previously reported (Birnbaum et al., 2014, 2017), and we modify these constructs for direct assessment of peptide and MHC-II binding (Rappazzo et al., 2020). Constructs for HLA-DR401 (HLA-DRA1*01:01, HLA-DRB1*04:01) and a lesser-studied allele, HLA-DR402 (HLA-DRA1*01:01, HLA-DRB1*04:02), were adapted to determine peptide-MHC interactions. In the adapted constructs, a 3C protease site and a Myc epitope tag are introduced into the flexible linker that connects the peptide to the HLA β chain (**Figure 1A**). When yeast are incubated with 3C protease, the linker is cleaved, allowing unbound peptides to freely disassociate. Yeast are then incubated at low pH in the presence of a high-affinity competitor peptide and the peptide-exchange catalyst HLA-DM (**Figure 1B**), emulating the native endosomal environment of MHC-II peptide loading (Roche and Furuta, 2015). Yeast encoding binding or non-binding peptides are then differentiated with a fluorescently-labeled antibody directed against the peptide-proximal epitope tag.

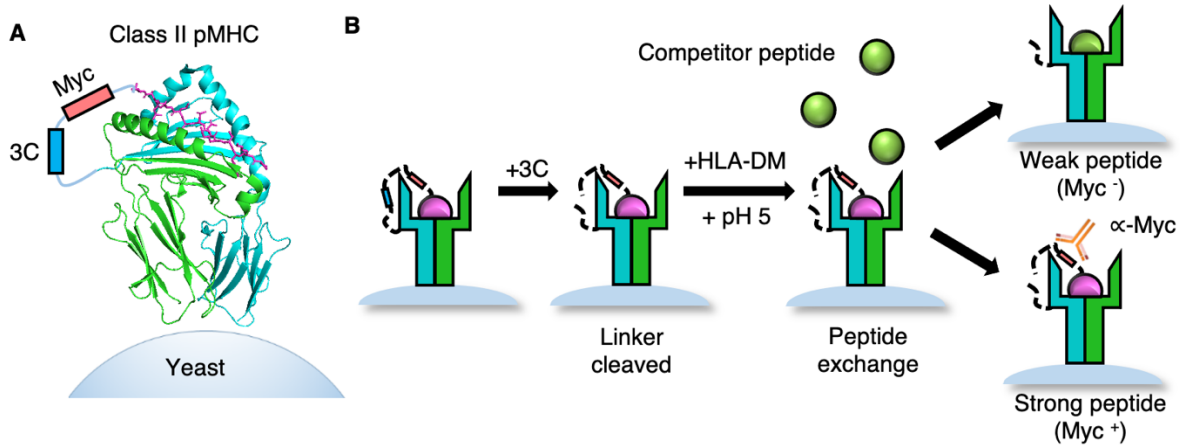


Figure 1. Design and validation of a yeast display platform to identify peptide binding to a co-expressed class II MHC. A) Structural representation of HLA-DR401 (PDB 1J8H) modified to encode a 3C protease cleavage site and Myc epitope tag within the linker connecting the peptide and MHC β1 domain. **B)** Schematic of validation protocol, including linker cleavage with 3C, peptide exchange at low pH in the presence of HLA-DM and high-affinity competitor peptide, and quantification of remaining bound peptide with an anti-Myc antibody.

Repertoire-scale investigations of HLA-DR401 and HLA-DR402 were conducted on libraries encoding $1x10^8$ random peptides. Because the ends of the MHC-II peptide-binding groove are open, the peptide-binding groove can accommodate peptides in many possible registers. To favor peptides binding in a single register and simplify downstream analysis, peptides were designed as randomized 9mers flanked by constant residues (Andreatta et al., 2018; Lin et al., 2008), namely N-terminal Ala-Ala and C-terminal Trp-Glu-Glu (Rappazzo et al., 2020). The library was subjected to iterative rounds of linker cleavage, peptide exchange, and selection for epitope tag retention (**Figure 2A**), resulting in a pool of yeast encoding strong binders after five rounds. The enriched peptides were highly diverse, consisting of 81,422 unique peptides in the expected register for HLA-DR401 and 7,692 unique peptides in the expected register for HLA-DR402. Sequence logos of the resulting enriched peptides were generated using Seq2Logo (Thomsen and Nielsen, 2012) and are shown in **Figure 2B** and **2C**.

Enriched peptides show strong amino acid preferences at MHC anchor peptide positions P1, P4, P6, and P9, which orient directly into MHC surface pockets (**Figure 2F**), largely matching previously reported binders and polymorphisms between HLA-DR401 and HLA-DR402 (**Figure 2D** and **2E**) (Abelin et al., 2019; Dessen et al., 1997; Hammer et al., 1993, 1994; Racle et al., 2019; Scally et al., 2017; Sette et al., 1993). Notably, the observed enrichment of P9 Cys has not been previously reported (Abelin et al., 2019; Racle et al., 2019; Scally et al., 2017). However, binding of the cysteine-containing peptides was specific, as two allele-mismatched cysteine-containing peptides did not exhibit binding, as assessed through a fluorescence polarization competition binding assay (**Figure 3A**).

To investigate the effect of positions outside of the groove on peptide binding, selections were also performed on a randomized 13mer HLA-DR401 library. After selections, 15,147 unique peptides were identified. Through register deconvolution by Gibbs Cluster (Andreatta et al., 2017), 3,374 of these peptides were identified as occupying the central register where positions P(-2) through P11 are diversified (**Figure 3B**). Position P10 displayed a mild preference for aromatic residues, consistent with previous findings (Zavala-Ruiz et al., 2004), and depletion of both Gly and Glu. We also observed depletion of hydrophobic residues and enrichment of acidic residues at positions P(-2) and P(-1). Positional preferences between positions P1 and P9 were consistent with the original library (**Figure 2B, D**), suggesting our motif was not influenced by the fixed peptide flanking residues in our original design.

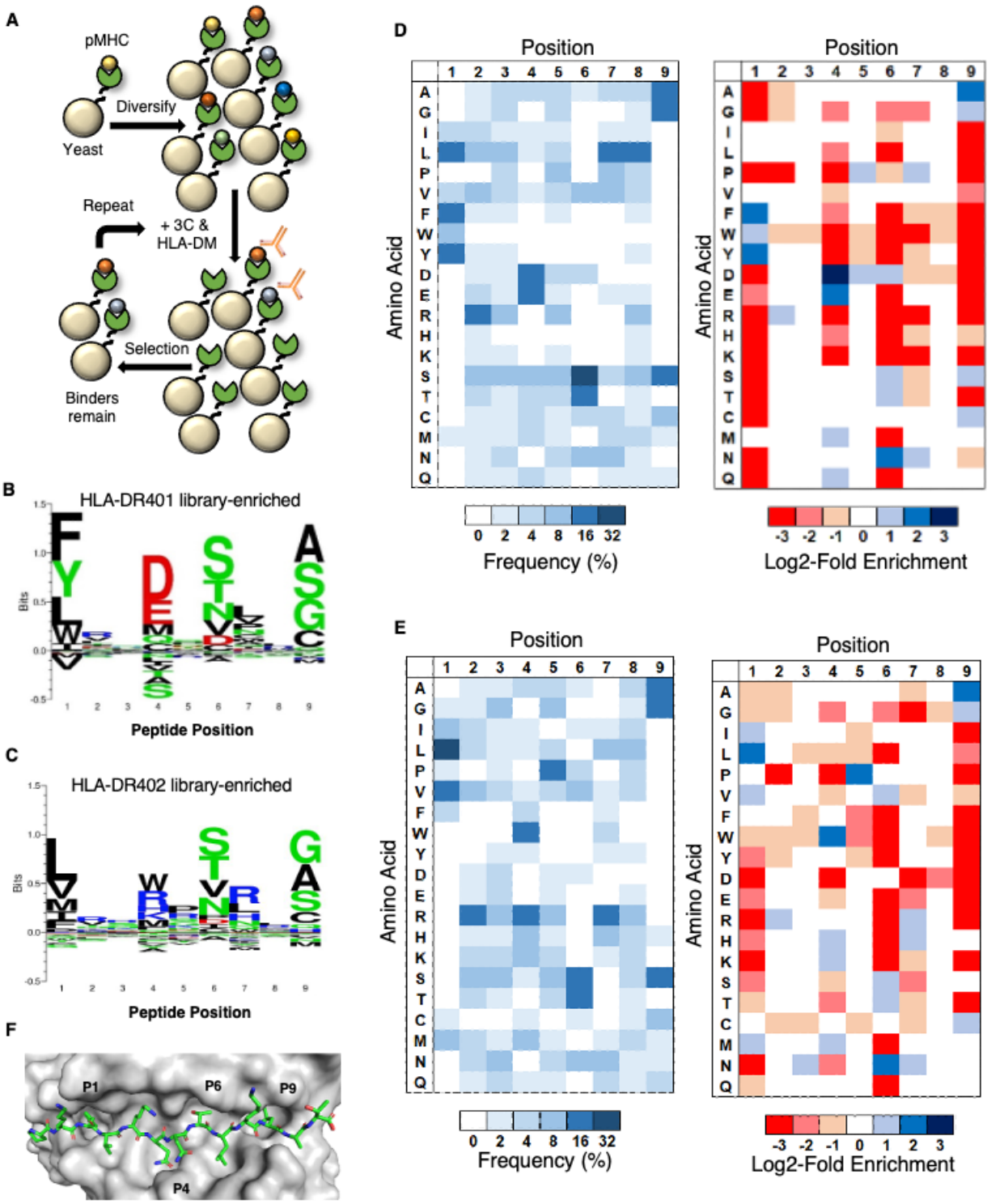


Figure 2. Selection and analysis of a yeast-displayed MHC-II randomized peptide libraries. A) Schematic of sequential rounds of library selection to eliminate non-binding peptides and enrich binders. **B, C)** Kullback-Leibler relative entropy motifs of the core nine amino acids of MHC-binding peptides for **B)** HLA-DR401 and **C)** HLA-DR402. **D, E)** Unweighted heatmaps of positional percent frequency and log2-fold enrichment of each amino acid in round 5 of selection for **D)** HLA-DR401 and **E)** HLA-DR402. **F)** Structure of $HA_{306-318}$ peptide in the HLA-DR401 peptide-binding groove (PDB 1J8H), with primary peptide MHC anchor positions denoted in bold.

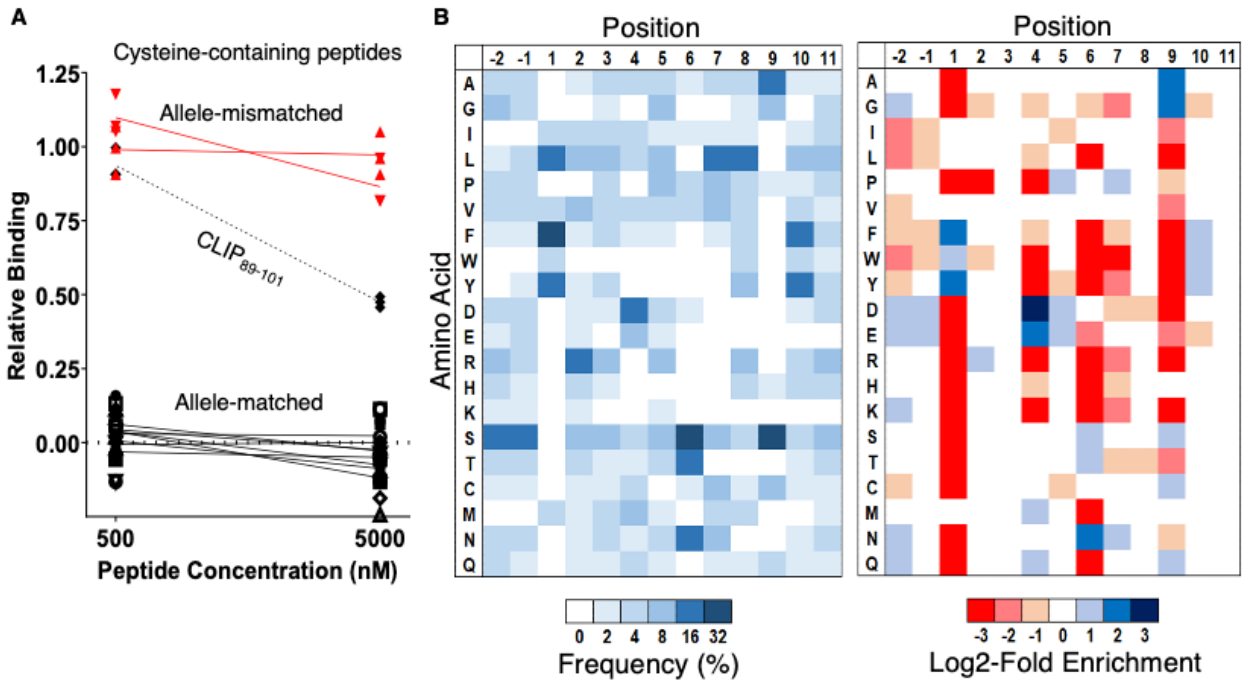


Figure 3. Discovery of preferences at TCR contacts and positions outside the peptide core and validation of cysteine-containing peptides. A) Relative binding of cysteine-containing peptides found in round 5 of selection of either the randomized 9mer HLA-DR401 (allele-matched) or HLA-DR402 (allele-mismatched) libraries, tested at two concentrations with HLA-DR401 in a fluorescence polarization competition assay. Curves are fit to N = 3 technical replicates per condition. **B)** Unweighted heat maps of log2-fold enrichment and/or positional percent frequency of each amino acid for all peptides in round five of selection of a randomized 13mer HLA-DR401 library determined to bind in the third peptide register (N = 3,374 unique peptides).

2.3 Training prediction algorithms

We hypothesized that yeast display-derived peptide binding data could be used to improve algorithmic prediction of MHC binding, due to the dataset's large size and origin from unbiased random libraries. To address this hypothesis, we trained prediction algorithms with our yeast display library data using NNAlign 2.0, an artificial neural network framework, and the architecture underlying the NetMHCII and NetMHCIIpan family of algorithms (Nielsen and Andreatta, 2017). Utilizing NNAlign facilitates direct comparison of the effect of the training data versus that of the prediction algorithm architecture.

First, to determine the identity of yeast-expressed peptides, plasmid DNA was extracted from 5 x 10^7 yeast from each round of selection, including the unselected library. The peptide-encoding region was sequenced via Next-Generation Sequencing to determine peptide identities, and peptides were sorted to determine the final round of selection in which they were observed. Next, NNAlign was trained on sequence data and quantitative target values; up to 80,000 sequenced peptides were assigned a target value commensurate with the final round of selection in which they were observed, between 0 and 1, with increasing target values for observation in later rounds. Specifically, target values for rounds 0 through 5, ascending were assigned as follows: [0, 0.05, 0.2, 0.8, 0.95, 1], to account for convergence seen following the third round of selection. As peptides from the pre-selection library were randomly generated, sequences observed in the pre-selection library but not in subsequent rounds serve as negative examples. To maintain a class-balanced training set, peptides were selected to be evenly distributed among rounds of last

appearance. Yeast display-trained models were generated using data from the 9mer peptide libraries for HLA-DR401 and HLA-DR402, as well as the 13mer peptide library for HLA-DR401. When training the NNAlign algorithm, the 9mer library data was used for training with default settings for 'MHC class II ligands', excepting expected peptide length set to 9 amino acids and expected PFR (peptide flanking residue) length set to 0 amino acids. The 13mer library data was used for trained with default settings, excepting expected peptide length set to 13 amino acids.

Recently, high-quality data has been generated for many MHC-I and MHC-II alleles utilizing eluted ligand mono-allelic mass spectrometry (MS) (Abelin et al., 2017, 2019; Sarkizova et al., 2020; Scally et al., 2017). In mono-allelic MS, antigen-presenting cells are engineered to express only a single MHC-II allele, eliminating the ambiguity in allelic assignment encountered in conventional poly-allelic MS eluted ligand datasets (Abelin et al., 2017; Alvarez et al., 2018). While these datasets are over an order of magnitude smaller than those generated by yeast display in terms of unique peptide cores, their motifs are largely consistent with yeast display, though absent P9 Cys. One of these datasets (Abelin et al., 2019) underlies the recently published MHC-II prediction algorithm NeonMHC2. To provide further comparison on the effect of training data versus the underlying algorithmic architecture, we generated an additional prediction algorithm from this data, again using NNAlign. Because of the open peptide-binding groove, MHC-II-derived mass spectrometry datasets contain overlapping nested peptides, and the mass spectrometry-generated peptide sets were curated to filtered minimum epitopes to remove redundant training examples and assigned a target value of 1. In order to prevent the algorithm from conflating altered amino acid frequencies arising from MS data collection with peptide-binding preferences, each peptide was scrambled to generate negative instances and assigned a target value of 0, in line with previously published recommendations (Abelin et al., 2019). These mass spectrometry algorithms were trained with default 'MHC class II ligands' settings.

For all NNAlign yeast display- and mass spectrometry-trained models, percentile ranks were established by comparing prediction values to the distribution of prediction values generated by apply each model to 50,000 computationally generated random 15mer peptides.

2.4 Generating test datasets and benchmarking performance

We next set out to comprehensively benchmark the predictive performance of our yeast display-trained algorithms as compared to a large array of other described approaches, including NNAlign trained on MS data. We identified a second mono-allelic MS peptide-binding dataset for each allele that were not represented in most current prediction training datasets (Scally et al., 2017). These data were utilized to generate a test dataset to facilitate independent evaluation. The filtered minimum core epitopes (as done in **Chapter 2.3**) were classified as positive instances. Since MS approaches generate positive examples only, we computationally generated negative decoy peptides to be length- and expression-matched, as previously described (Abelin et al., 2019). For each source protein observed within the dataset, we tiled across its sequence with peptide lengths randomly selected from the length distribution of the observed peptides, starting at the first amino acid in the protein and allowing an eight amino acid overlap between subsequent proteins. If the length of the final peptide extended beyond the end of the protein, we randomly shifted the starting amino acid such that the starting amino acid of the first peptide and last amino acid of the final peptide were all within the protein. We randomly selected decoy peptides from this set such that the length distribution of decoy peptides matched that of the positive instances, and that there was

no 9mer sequence match with the other decoys or positive instances, to avoid inadvertently adding redundant sequences to the test set.

We chose two metrics to assess predictive performance: the area under the receiver operating characteristic curve (AUC) and the positive predictive value (PPV). AUC is a commonly used metric in binary classification which illustrates the trade-off between the true positive rate and false positive rate as the threshold for classification is changed (Hanley and McNeil, 1982). A 1:1 ratio of positive instances and decoy peptides was used to generate receiver operating characteristic curves, where AUC was calculated with scikit-learn version 0.20.3. Because AUC is determined across thresholds and classifies a test set with equal numbers of positive and negative examples, it does not accurately capture a relevant ratio of true binders among examples and includes information from non-relevant threshold cutoffs. Therefore, as a second metric, we also calculated PPV. To calculate PPV, a test set with a more biologically-relevant 1:19 ratio of positive instances and decoy peptides was generated, and PPV was calculated as the fraction of true instances observed in the top 5% of predicted peptides for each algorithm (Abelin et al., 2019).

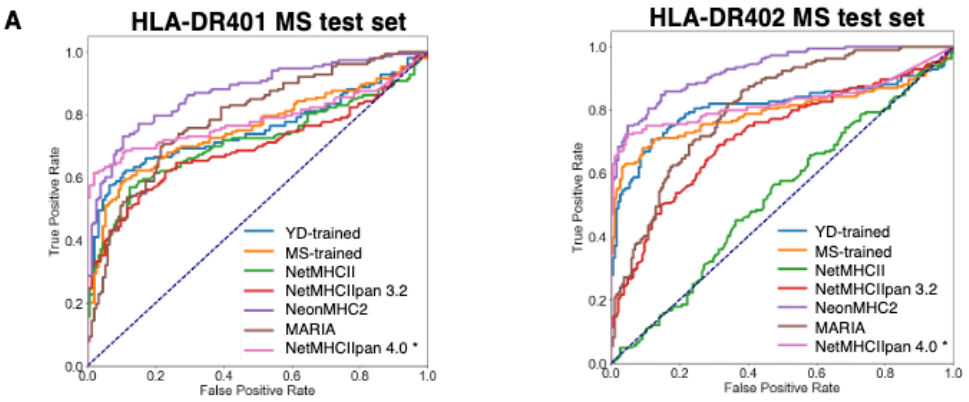
Each algorithm was applied to the allele-matched dataset (Scally et al., 2017), with length- and expression-matched decoy peptides. While the MS- and 9mer yeast display-trained models performed comparably to one another, the overall predictive performance of each of these algorithms was initially relatively low, with a maximum AUC of 0.81 (**Figure 4A**), suggesting a disparity between the training and evaluation sets. Unsupervised clustering of each MS-derived evaluation set with Gibbs Cluster (Andreatta et al., 2017) revealed that a substantial portion of each set (26% for HLA-DR401, 19% for HLA-DR402) were outliers, including peptides with long stretches of Gly or Pro, which have been previously reported to nonspecifically populate eluted ligand datasets (Heyder et al., 2016).

Removal of these outliers universally improved prediction performance (**Figure 5A**). For both alleles, the MS- and yeast display-trained algorithms performed comparably in AUC (0.92-0.94), and outperformed NetMHCII 2.3 and NetMHCIIpan 3.2, which are also built on NNAlign. This outperformance was more pronounced in PPV, with the yeast display-trained algorithm reaching 67% PPV for HLA-DR401. While the recently released NetMHCIIpan 4.0 EL (Reynisson et al., 2020) greatly outperformed its predecessors, its training set included our evaluation set, and therefore this algorithm could not be evaluated equitably. NeonMHC2 demonstrated strong performance for both alleles via AUC (0.96-0.97) and PPV (64-69%). As NeonMHC2 is built upon the same underlying data as the MS-trained algorithms, its improved performance may be due to the incorporation of peptide processing information, such as peptide cleavage preferences (Abelin et al., 2019). In addition, the recently described MixMHC2Pred (Racle et al., 2019), which is trained on conventional poly-allelic eluted ligand MS data, displayed comparable performance to NeonMHC2 on a subset of HLA-DR401 peptides (**Figure 4B**), but could not be fully compared due to peptide length constraints and the absence of an HLA-DR402 predictor. Nearly all algorithms evaluated outperformed another recently released poly-allelic eluted ligand MS-trained algorithm, MARIA (**Figure 5**) (Chen et al., 2019).

Importantly, however, the use of a MS-derived test set in evaluating predictive performance may not fully capture false negatives that might arise due to gaps in MS-derived data, such as those arising from systemic under-sampling of cysteine-containing peptides (Abelin et al., 2019; Scally

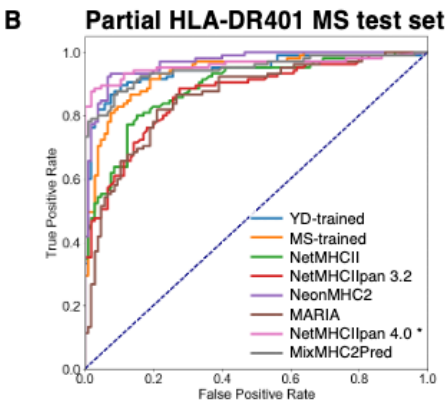
et al., 2017). Therefore, we further evaluated the predictive performance of each algorithm on our 13mer HLA-DR401 library data. For our yeast display test set, a randomly selected size-matched set of peptides found enriched in round 5 of selection were classified as positive instances, and decoy peptides were randomly selected from peptides only observed in their respective unselected library. We observed comparable performance between NetMHCII 2.3, NetMHCIIpan 3.2, NetMHCIIpan 4.0 EL, and the MS-trained NNAlign algorithm (AUC 0.79-0.82, PPV 27-30%) (**Figure 5B**). NeonMHC2 slightly underperformed its NNAlign-based counterpart, even though it was used in 'tiling mode' which ignores peptide cleavage preferences, further suggesting that the incorporation of peptide cleavage preferences underlie its previously noted outperformance on MS-derived data. The yeast display-trained model clearly outperformed each alternative algorithm, with an AUC of 0.92 and a PPV of 55%, and prediction performance only minimally improved by the removal of outlier peptides (**Figure 4C**).

Overall, our yeast display-trained algorithm performed comparably to current state-of-the-art approaches such as NeonMHC2 and NetMHCIIpan 4.0 on MS-derived data, while performing better on yeast display-derived data. These results suggest the presence of *bona fide* peptide motifs in yeast display data that are not adequately sampled in MS-derived data. Direct comparison of the MS- and yeast display-trained algorithms at a positional level revealed a significantly ($p < 0.05$) more stringent P9 preference in the yeast display-trained algorithm for both alleles (**Figure 6**). Furthermore, consistent with its under-representation in MS-derived data, Cys was significantly over- or under-represented at multiple positions and the MS-trained algorithms had a greater preference for small hydrophobic residues Ile, Leu, and Val at multiple positions.

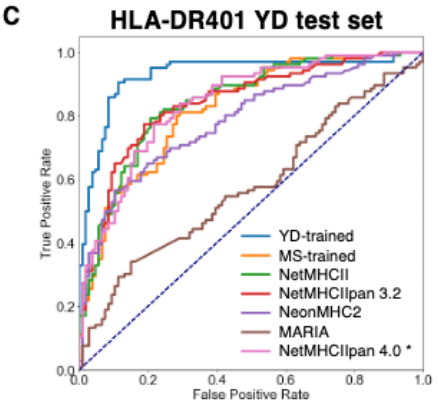


Prediction	AUC	PPV (%)
YD-trained	0.75	50
MS-trained	0.76	45
NetMHCII 2.3	0.71	33
NetMHCIIpan 3.2	0.70	33
NeonMHC2	0.87	49
MARIA	0.79	31
NetMHCIIpan 4.0 *	0.78	58

Prediction	AUC	PPV (%)
YD-trained	0.81	43
MS-trained	0.79	45
NetMHCII 2.3	0.53	5
NetMHCIIpan 3.2	0.72	21
NeonMHC2	0.93	59
MARIA	0.81	25
NetMHCIIpan 4.0 *	0.82	64

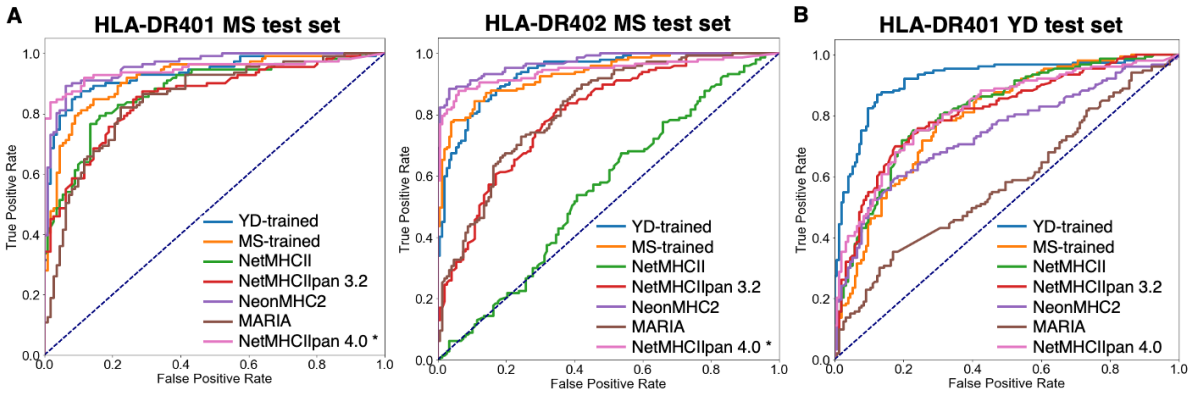


Prediction	AUC	PPV (%)
YD-trained	0.94	70
MS-trained	0.94	61
NetMHCII 2.3	0.89	47
NetMHCIIpan 3.2	0.87	42
NeonMHC2	0.97	68
MARIA	0.86	42
NetMHCIIpan 4.0 *	0.96	77
MixMHC2Pred	0.95	72



Prediction	AUC	PPV (%)
YD-trained	0.93	58
MS-trained	0.82	29
NetMHCII 2.3	0.84	33
NetMHCIIpan 3.2	0.84	32
NeonMHC2	0.79	32
MARIA	0.58	9
NetMHCIIpan 4.0	0.84	28

Figure 4. Benchmarking MHC-II prediction algorithm performance on eluted ligand mass spectrometry or yeast display data. Receiver operating characteristic (ROC) curves for prediction with published prediction algorithms, or algorithms trained on our 9mer yeast display library (YD-trained) or mono-allelic MS (MS-trained) data, for either **A)** mono-allelic MS data for HLA-DR401 and -DR402, with expression-matched decoy peptides, **B)** length restricted (12-24mer) outlier-removed mono-allelic MS data for HLA-DR401, with expression-matched decoy peptides or **C)** outlier-removed round five 13mer yeast display library data with decoys from the naïve library. For each dataset, the area under the ROC curve (AUC) and positive predictive value (PPV) of each prediction are shown. Asterisks indicate algorithms that contain the evaluation set in their training data.



Prediction	AUC	PPV (%)
YD-trained	0.93	67
MS-trained	0.92	54
NetMHCII 2.3	0.88	42
NetMHCIIpan 3.2	0.85	41
NeonMHC2	0.96	64
MARIA	0.85	34
NetMHCIIpan 4.0 *	0.95	74

Prediction	AUC	PPV (%)
YD-trained	0.94	47
MS-trained	0.93	53
NetMHCII 2.3	0.55	6
NetMHCIIpan 3.2	0.80	24
NeonMHC2	0.97	69
MARIA	0.82	29
NetMHCIIpan 4.0 *	0.94	75

Prediction	AUC	PPV (%)
YD-trained	0.92	55
MS-trained	0.79	27
NetMHCII 2.3	0.82	30
NetMHCIIpan 3.2	0.81	27
NeonMHC2	0.74	29
MARIA	0.58	12
NetMHCIIpan 4.0	0.82	29

Figure 5. Benchmarking performance of yeast display-trained algorithms. Receiver operating characteristic (ROC) curves for prediction with existing prediction algorithms, or algorithms trained on our 9mer yeast display library (YD-trained) or eluted ligand mono-allelic mass spectrometry (MS-trained) data, on either **A)** outlier-removed eluted ligand MS data for HLA-DR401 and -DR402, with expression-matched decoy peptides, or **B)** yeast display 13mer HLA-DR401 library data, with naïve library decoys. For each dataset, the area under the ROC curve (AUC) and positive predictive value (PPV) of each prediction are shown. Asterisks indicate algorithms that contain the evaluation set in their training data.

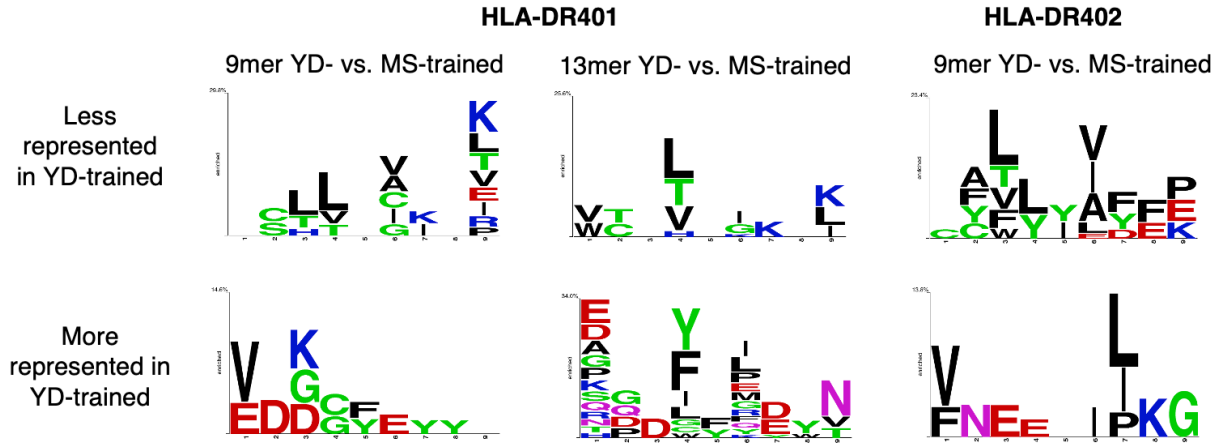


Figure 6. Comparison of peptide motifs derived from eluted ligand mono-allelic mass spectrometry and yeast display library datasets. Amino acids significantly ($p < 0.05$) more or less represented at each position within the core 9 amino acids of HLA-DR401 or -DR402-binding peptides, as determined by algorithms trained on our yeast display libraries (YD-trained), relative to algorithms trained on eluted ligand mono-allelic MS data (MS-trained). Displayed size of residues correlates with statistical significance of deviation and significance was determined by two-sided unweighted binomial test for $p < 0.05$, with a Bonferroni correction for multiple hypothesis testing.

2.5 Improvements in affinity prediction and prediction of clinically relevant peptides

To investigate the effect algorithmic differences may have on the prediction of clinically relevant peptides, we performed antigen prediction for HLA-DR401 with NeonMHC2 and the 9mer yeast display-trained algorithm on two datasets: the proteome of Influenza A virus (IAV), and expression-validated mutations from human lung adenocarcinoma patients (Cai et al., 2018). From these datasets, the 9mer yeast display-trained model differentially classified – relative to NeonMHC2 – 5 IAV-derived peptides as strong or non-binders, and differentially classified 13 adenocarcinoma-derived peptides as potential cancer neoantigens. Interestingly, these algorithms displayed non-overlapping algorithmic misses (**Table 1**, **Figure 7**), suggesting that there are peptide motifs unique to both the MS- and yeast display-derived training data that contribute to improved peptide prediction performance.

Peptide	WT NeonMHC2 Rank (%)	WT YD-trained Rank (%)	Mut. NeonMHC2 Rank (%)	Mut. YD-trained Rank (%)	Measured IC50 (nM)
LLPVKQSVVPLAKGTLITHC	7.7	5.4	14	1.8	307
ASDSSYRNECPMAEKEDTQM	30	22	22	1.8	47
MKRRPVWTDINPKAVTNDEL	14	4.3	11	0.8	21
PIPVIFHRIATELRKTNDIN	0.8	28	< 0.1	0.4	698
VYGWATLVSERSKNGMQRIL	0.4	27	0.3	0.5	245
YVGDMLAWLHQSTASEKEHL	1.0	46	0.5	34	406
EVKYCTFSKDRSKPIPGMTL	1.3	32	1.5	18	> 50,000
WRAPSYILSPELTQRLFSAA	2.1	39	1.8	24	248
RKDLIVMLMDTDVNKQDKQK	1.8	30	1.0	19	3173
YCDLPQLFRLSCSSTQLNEL	2.1	31	2.0	16	39
FVGNIAEDLCLDITKLSARG	2.4	43	1.6	19	86
ESSISDKNYWKTVSNAFSVI	5.8	25	2.0	12	25
YSIEVLLVLDDSVVRFHGKE	1.4	30	1.2	30	> 50,000
FYHKCDNECMESVRNGTYDY	51	1.7	-	-	7273
LFQNWGVEPIDNVMGMIGIL	13	1.6	-	-	2524
EGIPLYDAIKCMRTFFGWKE	0.1	17	-	-	61
ARQMVQAMRTIGTHPSSSAG	1.7	32	-	-	3994
MFLYVRTNGTSKIKMKWGME	0.9	43	-	-	6102

Table 1. Prediction of pathogen- and tumor-associated peptides with MHC-II algorithms trained on yeast display library or eluted ligand mass spectrometry datasets. Mutant peptides from human lung adenocarcinomas differentially predicted as neoantigens for HLA-DR401, or peptides derived from Influenza A virus differentially predicted as strong- versus non-binders, by NeonMHC2 or a model trained on our 9mer yeast display library data, with measured IC_{50} values for peptides that NeonMHC2 and the yeast display-trained algorithm disagreed, derived from fluorescence polarization competition assays.

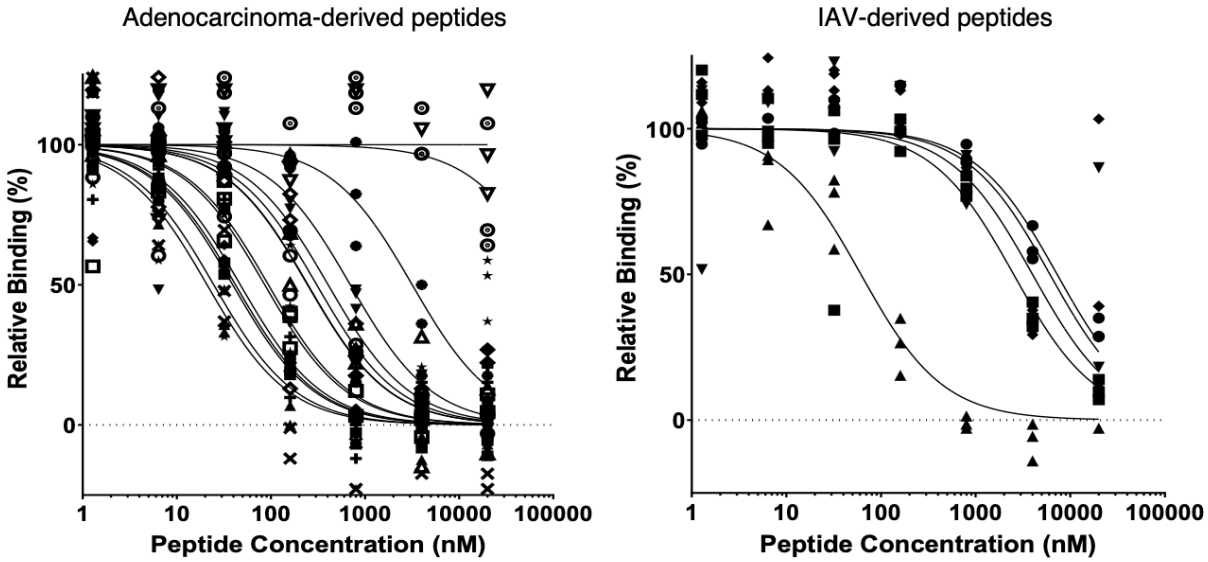


Figure 7. Prediction of pathogen- and tumor-associated peptides with MHC-II algorithms trained on yeast display library or eluted ligand mass spectrometry datasets. Binding curves for HLA-DR401 from fluorescence polarization competition assays for peptides found mutated in human lung adenocarcinomas and differentially predicted as neoantigens for HLA-DR401, or peptides derived from influenza A virus differentially predicted as strong- versus non-binders, by NeonMHC2 or a model trained on our 9mer yeast display library data. Curves are fit to N = 3 technical replicates per condition.

In the course of this study, we have generated data in-house on binding affinity to HLA-DR401 for 55 peptides (Rappazzo et al., 2020). When measured and predicted affinities for all 55 peptides assayed for binding were considered, current eluted ligand MS-trained algorithms NeonMHC2, NetMHCIIpan 4.0 EL, MARIA, and our own MS-trained model displayed little to no correlation with measured IC_{50} (R^2 = 0.08-0.19), indicative of poor peptide affinity prediction performance (**Figure 8**). In addition, NetMHCIIpan 4.0 BA, which is trained exclusively on peptide binding affinity data (Reynisson et al., 2020), failed to show correlation with measured IC_{50} for these peptides (R^2 = 0.01) However, our 9mer yeast display trained model algorithm displayed notably improved correlation with measured IC_{50} (R^2 = 0.47), and consistent with our findings on peptide flanking residues, the predictions of the 13mer yeast display-trained model displayed even greater correlation (R^2 = 0.62).

Overall, our results demonstrated that both eluted ligand MS- and yeast display-derived peptide datasets improved the performance of MHC-II prediction algorithms relative to legacy datasets, and both identified unique peptide motifs. However, we find that yeast display provided much larger datasets than eluted ligand MS, and provided notably improved performance in predicting peptide affinity.

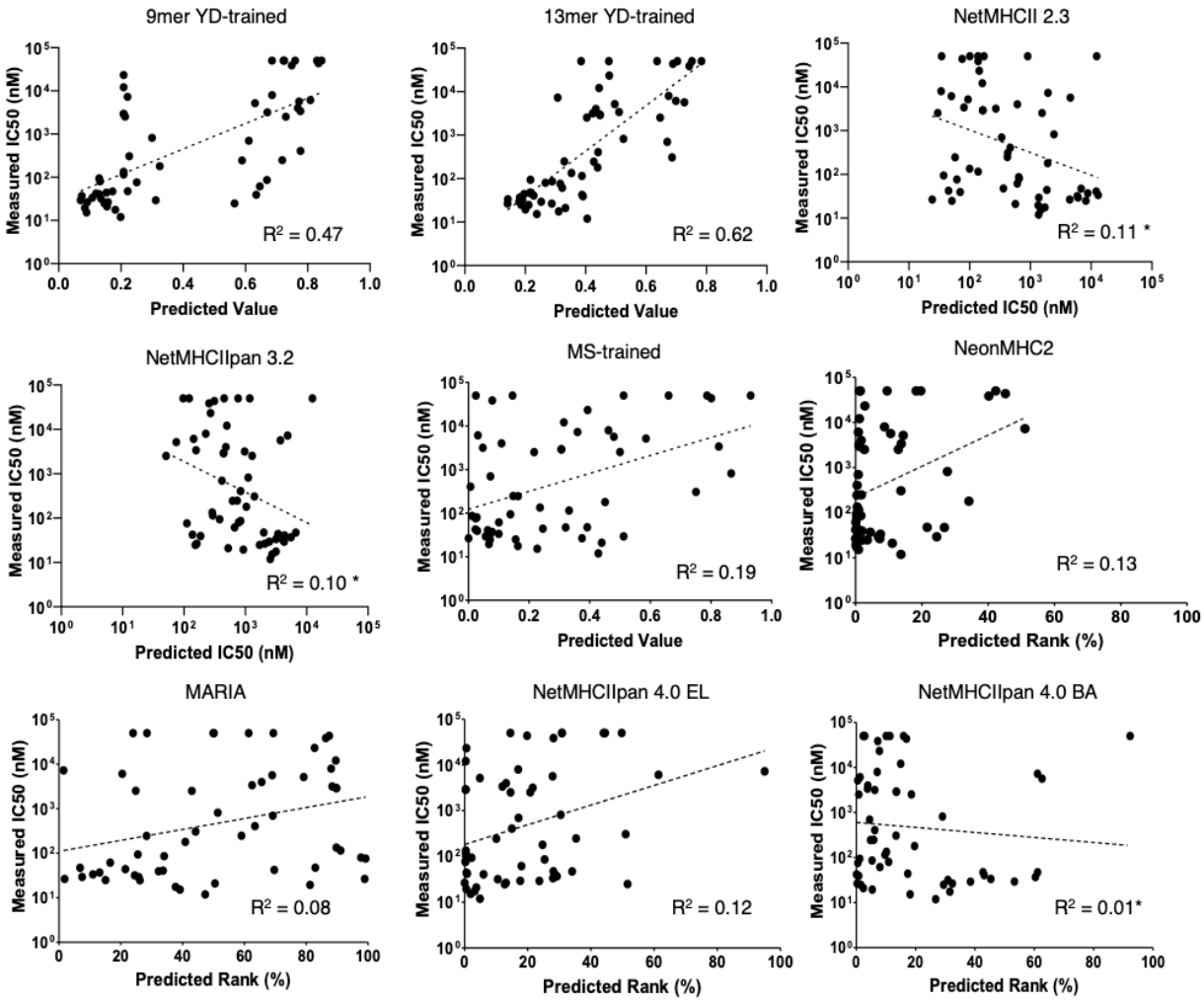


Figure 8. Benchmarking MHC-II algorithm performance for prediction of peptide-binding affinity. Scatterplots of algorithmic predictions versus measured IC_{50} values for 55 peptides assayed for binding to HLA-DR401 in fluorescence polarization competition assays, with lines of best fit and their associated coefficients of determination (R^2). Asterisk denotes R^2 values of negative correlations.

2.6 Discussion and outlook

The central role of $CD4^+$ T cells across infection, cancer, autoimmunity, and allergy motivates a need to predict which peptide antigens can be presented by MHC-IIs. However, MHC-II prediction algorithms can suffer from consequential gaps and inaccuracies in coverage, especially for less characterized alleles (Andreatta et al., 2018; Jensen et al., 2018; Lin et al., 2008; Nielsen et al., 2010; Wang et al., 2008; Zhao and Sher, 2018). Here, we present a platform for large-scale identification of diverse MHC-II-binding peptides, generating over an order of magnitude more unique data than comparable approaches for two human MHC-II alleles and identifying subsets of peptides missed by frequently used prediction algorithms. We utilized this yeast display-generated data to train existing algorithmic architectures and used these algorithms to discover *bona fide* peptide binders that are not predicted by other prediction algorithms.

Analysis of the training data underlying previously described prediction algorithms revealed multiple sources of underperformance. For both alleles studied, we found large numbers of nested

and single amino acid variant peptides within curated training sets. While training algorithms account for redundant information from nested sets (Nielsen and Lund, 2009), their presence diminishes the functional size of the training set. Single amino acid variants, however, are considered unique peptides, and can therefore impart biases. Furthermore, a systemic absence of cysteine in training sets resulted in substantial algorithmic false negatives for cysteine-containing peptides. This is likely due to an aversion to working with cysteine-containing peptides, as well as the difficulties inherent to sampling them in mass-spectrometry (Barra et al., 2018). A systemic underrepresentation of acidic residues in the Immune Epitope Database (IEDB) has also been reported (Abelin et al., 2019).

By using our data to train prediction algorithms and benchmark their performance against existing algorithms, we identified several key considerations for MHC-II antigen prediction. First, our results demonstrate that high-quality training data improves the performance of MHC-II prediction algorithms without alteration of underlying training algorithm architectures, especially for less characterized alleles (**Figure 5A**). However, there are important opportunities for algorithmic improvement, such as increased focus on peptide flanking residues. Second, we find that each source of data has non-overlapping strengths and weaknesses for improving prediction performance. Therefore, an ideal MHC-II prediction algorithm may be trained on both high-quality datasets that reflect native processing (Barra et al., 2018), such as eluted ligand MS datasets, as well as large and diverse peptide datasets, such as those generated by our yeast display platform. Third, we highlight the importance of the choice of validation sets for benchmarking prediction algorithms, as frequently used metrics of prediction performance underestimate false negatives due to gaps in test sets, allowing entire classes of peptides to be missed without impacting performance metrics (**Figure 5A, 5B**). Finally, we find that yeast display-trained algorithms are superior at predicting peptide affinity, which is a crucial consideration in identifying peptides suitable for antigen-targeted therapeutics (Backert and Kohlbacher, 2015; Hu et al., 2018; Patronov and Doytchinova, 2013). The non-binary nature of yeast display data, which is trained on peptides from five rounds of selection, possibly accounts for this key disparity.

This work also presents opportunities for utilizing yeast display for high-throughput, direct assessment of binding between MHC-IIs and defined peptides of interest, which we explore further in **Chapter 3** and **Chapter 4**. And, as this platform does not require allele-specific reagents, we believe it can generate high-quality repertoire-scale data for many additional MHC-II alleles, even those with few curated binders, greatly increasing its applicability. Extensions of the yeast display platform for additional MHC-II alleles are explored in **Chapter 5**. With this in mind, we believe this technology can greatly benefit the field of MHC-II antigen prediction, and therefore the study and application of $CD4^+$ T cell recognition across pathogen infection, cancer, and immune disorders.

2.7 Methods

Yeast-displayed pMHC design

Full-length yeast-displayed HLA-DR401 (HLA-DRA1*01:01, HLA-DRB1*04:01) with a cleavable peptide linker was based upon a previously described HLA-DR401 construct optimized for yeast display with the mutations Mα36L, Vα132M, Hβ33N, and Dβ43E to enable proper folding without perturbing either TCR- or peptide-contacting residues (Birnbaum et al., 2017). The alpha and beta chain ectodomains were expressed as a single transcript connected by a self-

cleaving P2A sequence. The peptide was joined through a flexible linker to N-terminus of MHC β1 domain. This construct was further modified to express a 3C protease site (LEVLFQ/GP) and Myc epitope tag (EQKLISEEDL) within the flexible linker, for a total of 32 amino acids between the peptide and β1 domain. HLA-DR402 (HLA-DRA1*01:01, HLA-DRB1*04:02) was generated by modification of this construct with each native HLA-DRβ polymorphism of HLA-DR402. All yeast display constructs were produced on the pYAL vector as N-terminal fusions to AGA2. All yeast strains were grown to confluence at 30°C in pH 5 SDCAA yeast media then subcultured into pH 5 SGCAA media at OD_{600} = 1.0 for a 48 hour induction at 20°C (Chao et al., 2006).

Library design and selection

Randomized peptide yeast libraries were generated by polymerase chain reaction (PCR) of the pMHC construct with primers encoding NNK degenerate codons. To ensure only randomized peptides expressed within the library, the template peptide region encoded multiple stop codons. Randomized 9mer libraries were designed as [AAXXXXXXXXXWEEG…] to constrain peptide register and randomized 13mer libraries were designed as [AXXXXXXXXXXXXXG…]. Randomized pMHC PCR product and linearized pYAL vector backbone were mixed at a 5:1 mass ratio and electroporated into electrically competent RJY100 yeast (Van Deventer et al., 2015) to generate libraries of at least 1 x 10^8 transformants.

Libraries were subjected to 3C cleavage with 1 μM 3C protease in PBS pH 7.4 at a concentration of 2 x 10^8 yeast/mL for 45 minutes at room temperature. After linker cleavage, yeast expressing the pMHC were washed into pH 5 citric acid saline buffer (20 mM citric acid, 150 mM NaCl) at 1 x 10^8 yeast/mL with 1 μM HLA-DM and a high-affinity competitor peptide at 4°C to catalyze peptide exchange. HLA-DR401-expressing yeast were incubated with 1 μM $HA_{306-318}$ (PKYVKQNTLKLAT) and HLA-DR402-expressing yeast were incubated with 5 μM $CD48_{36-53}$ (FDQKIVEWDSRKSKYFES) (Genscript; Piscataway NJ). Libraries were subjected to 3C cleavage and peptide exchange for 16-18 hours, and were selected for peptide-retention via binding of ∝-Myc-AlexaFluor647 antibody and magnetic ∝-AlexaFluor647 magnetic beads (Miltenyi Biotec; Bergisch Gladbach, Germany). Selected yeast were re-cultured, induced, and selected for an additional four rounds, for five total rounds of selection.

Generation and comparison of peptide motifs

Kullback-Leibler relative entropy motifs were generated with Seq2Logo-2.0 (Thomsen and Nielsen, 2012). For yeast display data, the core 9mers of round 5 sequences were input with background amino acid frequencies derived from their average in their matched unselected library.

Benchmarking and comparison of prediction algorithms

Prediction algorithms were benchmarked against independently generated allele-specific eluted ligand mono-allelic MS or yeast display library data, with matched decoy peptides. AUC and PPV values are provided for the 1-log50k(aff) output of NetMHCII 2.3 and NetMHCIIpan 3.2, and was comparable to the performance of the %Rank output. For NetMHCIIpan 4.0, %Rank_EL was provided, and performs comparably to the Score_EL output. Reported prediction values for NNAlign models are the inverse of model output prediction values (1-value) for ease of comparison to other prediction algorithms.

Prediction algorithms were compared at a positional level by Two Sample Logo (Vacic et al., 2006). For each comparison, the two algorithms were applied to a common set of 50,000 computationally-generated 15mer peptides. The predicted core 9mer of peptides that rank within the 90th percentile or higher of predicted value for only one algorithm were evaluated against the cores of peptides that rank within the 90th percentile or higher of predicted value for both algorithms. Significance was determined by two-sided unweighted binomial test for $p < 0.05$, with a Bonferroni correction for multiple hypothesis testing.

Library deep sequencing and analysis

Libraries were deep sequenced to determine the peptide repertoire at each round of selection. Plasmid DNA was extracted from 5 x 10^7 yeast from each round of selection with the Zymoprep Yeast Miniprep Kit (Zymo Research; Irvine CA), according to manufacturer's instructions. Amplicons were generated by PCR with primers designed to capture the peptide encoding region through the polymorphic region that differentiates HLA-DR401 from HLA-DR402. An additional PCR round was then performed to add i5 and i7 paired-end handles with inline sequencing barcodes unique to each library and round of selection. Amplicons were sequenced on an Illumina MiSeq (Illumina Incorporated; San Diego, CA) with the paired-end MiSeq v2 500bp kit at the MIT BioMicroCenter.

Paired-end reads were assembled via FLASH (Magoč and Salzberg, 2011) and processed with an in-house pipeline that filtered for assembled reads with exact matches to the expected length, polymorphic sequences, and 3C protease cleavage site, then sorted each read based on its inline barcode and extracted the peptide-encoding region. To ensure only high-quality peptides were analyzed, reads were discarded if any peptide-encoding base pair was assigned a Phred33 score less than 20, or did not match the expected codon pattern at NNK sites (N = any nucleotide, K = G or T). To account for PCR and read errors from high-prevalence peptides, reads were discarded if their peptide-encoding regions were Hamming distance > 1 from any more prevalent sequence, Hamming distance > 2 from a sequence 100 times more prevalent, or Hamming distance > 3 from a sequence 10,000 times more prevalent within the same round, in line with previously published analysis methods (Christiansen et al., 2015). Unique DNA sequences were translated by Virtual Ribosome (Wernersson, 2006) and filtered for peptides not encoding a stop codon.

Heat map visualization of library peptide preferences

Heat maps were generated from filtered sequences from indicated round to visually represent positional preferences. For each round, the unweighted prevalence of each amino acid at each position was calculated as a percentage. This positional percent prevalence was compared to its matched value in the unselected library to generate log2-fold enrichment values.

For randomized 9mer libraries, these log2-fold enrichment values were used to generate 20x9 position-specific scoring matrices (PSSMs) that were used to identify out-of-register peptides in round 5 of selection. Each 15mer peptide was scored in each of its seven possible 9mer registers by the PSSM, without positional weighting. Peptides which scored highest in a shifted register, regardless of score, were deemed out-of-register. For the randomized 13mer library, peptide register was determined by Gibbs Cluster 2.0 (Andreatta et al., 2017), with settings imported from 'MHC class I ligands of the same length', a motif of 13 amino acids, no discarding of outlier peptides, and background amino acid frequencies derived from the data. This allowed visualization

of each peptide register independently, without collapsing to a common 9mer motif. The number of unique clusters was determined by maximum Kullback-Leibler distance. Results were comparable between both methods of register determination for the 9mer peptide data.

Analysis of peptide data from external data sources

External MHC-binding peptide data was curated from two previously-published eluted ligand mono-allelic mass-spectrometry (MS) datasets (Abelin et al., 2019; Scally et al., 2017). Eluted ligand mono-allelic MS peptide data was analyzed as previously recommended (Scally et al., 2017), the minimum epitope of nested peptide sets were filtered for those that did not map to immunoglobulin or HLA proteins. Each dataset was clustered by Gibbs Cluster 2.0 (Andreatta et al., 2017) with default settings for 'MHC class II ligands', excepting the default removal of outlier peptides, and amino acid frequencies 'from data', to identify the core 9mer of each peptide. In each case, Kullback-Leibler distance was maximized for one cluster. For identification of outlier peptides, the default removal of outlier peptides was enabled.

Recombinant protein production

Recombinant soluble HLA-DM, HLA-DR401, and HLA-DR402 were produced in High Five (Hi5) insect cells (Thermo Fisher) via a baculovirus expression system, as previously described for other MHC-II proteins (Birnbaum et al., 2014). Ectodomain sequences of each chain followed by a poly-histidine purification site were cloned into pAcGP67a vectors. For each construct, 2 μg of plasmid DNA was transfected into SF9 insect cells with BestBac 2.0 linearized baculovirus DNA (Expression Systems; Davis, CA) using Cellfectin II reagent (Thermo Fisher; Waltham, MA). Viruses were propagated to high titer, co-titrated to maximize expression and ensure 1:1 MHC heterodimer formation, then co-transduced into Hi5 cells and grown at 27°C for 48-72 hours. Proteins were purified from the pre-conditioned media supernatant with Ni-NTA resin and size purified via size exclusion chromatography using a S200 increase column on an AKTAPURE FPLC (GE Healthcare, Chicago IL). HLA-DRB1*04:01 and HLA-DRB1*04:02 chains were expressed with $CLIP_{81\text{-}101}$ peptide connected by a 3C protease-cleavable flexible linker to the MHC N-terminus to improve protein yields.

Peptide competition assays and IC_{50} determination

The IC_{50} of characterized peptides was quantified with a protocol modified from Yin, L. and Stern, L.J. (2014) (Yin and Stern, 2014). Relative binding values were generated at each concentration according to the equation $(FP_{sample} - FP_{free})/(FP_{no\_comp} - FP_{free})$, where FP_{free} is the polarization value of the fluorescent peptide before addition of MHC, FP_{no_comp} is the polarization value with added MHC but no competitor peptide, and FP_{sample} is the polarization value with added MHC and competitor peptide. Relative binding curves were generated and fit by Prism 8.0 (GraphPad Software Inc; San Diego CA) to the equation $y = 1/(1+[pep]/IC_{50})$, where [pep] is the concentration of competitor peptide, to determine the IC_{50} of each peptide, its concentration of half-maximal inhibition.

For each 200 μL assay, 100 nM soluble MHC was combined with 25 nM of fluorescently-modified peptide in pH 5 binding buffer and incubated at 37°C for 72 hours in black 96-well flat bottom plates (Greiner Biotech; Kremsmünster, Austria). Modified $HA_{306\text{-}308}$ peptide [APRFV{Lys(5,6 FAM)}QNTLRLATG] was used for HLA-DR401. N=3 replicates were performed for each unlabeled peptide (Genscript; Piscataway, NJ) concentration, ranging in five-fold dilutions from

20 μM to 1.28 nM. Plates were read on a Tecan M1000 (Tecan Group Ltd., Morrisville NC) with 470 nm excitation, 520 nm emission, optimal gain, and a G-factor of 1.10. An important modification of our protocol is the presence of the MHC-linked CLIP peptide that was released by incubation with 3C protease at a 1:100 molar ratio at room temperature for 1 hour prior to dilution into plates. Residual cleaved CLIP peptide at 100 nM is not expected to alter peptide binding.

Lines of best fit between predicted and measured affinity for characterized peptide, and associated coefficients of determination (R^2), were generated in Prism 8.0.

Data availability
All deep sequencing data was deposited on the sequence read archive (SRA) with accession code PRJNA647875 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA647875/].

Code availability
All scripts used for data processing and analysis, as well as all NNAlign model files, are publicly available at https://github.com/birnbaumlab/Rappazzo-et-al-2020.

2.8 Acknowledgements
The material covered in this chapter has been published as Rappazzo*, Huisman*, and Birnbaum (2020). "Repertoire-scale determination of class II MHC peptide binding via yeast display improves antigen prediction". Nat. Commun. (Rappazzo et al., 2020). Excerpts from the publication were reproduced or adapted here under the Creative Commons license. This MHC-II work was started by C. Garrett Rappazzo, including adapting the yeast display approach for studying peptide-MHC-II binding, and generating the library and fluorescence polarization data referenced in this chapter. Michael E. Birnbaum also contributed to the conceptualization of this work and provided guidance on experimental design and analysis. As such, Garrett and Michael have contributed significantly to the content, text, and figures of this chapter.

We thank K. Christopher Garcia (Stanford University) for providing the plasmids encoding 3C protease and yeast-displayed HLA-DR401, Patrick Holec and Robert Wilson for their assistance in analyzing computational data, Lauren Stopfer and Forest White for their insight and feedback, and Isadora Deese for reviewing the manuscript. We would also like to thank Anthony W. Purcell (Monash University) and Michael S. Rooney (Neon Therapeutics) for generously providing the mono-allelic mass spectrometry data for HLA-DR401 and HLA-DR402. This work was supported by the staff of the Koch Institute Swanson Biotechnology Center, especially the staff of the flow cytometry, biopolymers and proteomics, high-throughput sciences, and genomics cores. This work was also supported by the National Cancer Institute (P30-CA14051), the Packard Foundation, Schmidt Futures, the V Foundation and the AACR-TESARO Career Development Award for Immuno-oncology Research [17-20-47-BIRN] for M.E.B.

CHAPTER 3: MACHINE LEARNING OPTIMIZATION AND YEAST DISPLAY ASSESSMENT OF PEPTIDES FOR ENHANCED MHC-II BINDING

This chapter contains work conducted in collaboration with Zheng Dai and adapted under the Creative Commons license from: Dai, Huisman*, et al (2021). "Machine learning optimization of peptides for presentation by class II MHCs". Bioinformatics. A full acknowledgements statement is included as* ***Chapter 3.11****.*

Abstract
T cells play a critical role in cellular immune responses to pathogens and cancer and can be activated and expanded by MHC-presented antigens contained in peptide vaccines. We present a machine learning method to optimize the presentation of peptides by class II MHCs by modifying their anchor residues. Our method first learns a model of peptide affinity for a class II MHC using an ensemble of deep residual networks, and then uses the model to propose anchor residue changes to improve peptide affinity. We use a high throughput yeast display assay to show that anchor residue optimization improves peptide binding.

3.1 Introduction

Machine learning holds great promise for improving therapeutic molecules, and here we show how it can be applied to enhance the display of peptides that invoke cellular immune responses to pathogens and cancer. T cells surveille peptides displayed on the cell surface by Major Histocompatibility Complexes (MHCs), or Human Leukocyte Antigens (HLAs) in humans, and T cell-mediated killing is initiated by recognition of a foreign peptide bound to an MHC. Specifically, $CD8^+$ cytotoxic T cells recognize peptides presented by class I MHCs (MHC-I), and $CD4^+$ helper T cells recognize peptides presented by class II MHCs (MHC-II) (Hennecke and Wiley, 2001). MHC-Is have a closed peptide-binding groove and typically present peptides of 8-11 amino acids; MHC-IIs have an open binding groove and typically present longer peptides, with a 9 amino acid core binding within the groove and the ends protruding from the groove. Prediction algorithms have been utilized to predict peptide-MHC binding. Recent strides in algorithmic performance have been enabled by advances in computational methods (Chen et al., 2019; O'Donnell et al., 2020; Racle et al., 2019; Reynisson et al., 2020; Zeng and Gifford, 2019) and the development of new methodologies for generating training data, such as mono-allelic mass spectrometry (Abelin et al., 2017, 2019; Sarkizova et al., 2020) and yeast display (Rappazzo et al., 2020). With the help of these tools, peptide vaccines with constituent peptides computationally selected for the ability to be displayed by MHCs have been utilized to amplify T cell responses and proven clinically successful for patients with cancer after eliciting $CD8^+$ and $CD4^+$ T cell responses (Abelin et al., 2017; Hu et al., 2018; Ott et al., 2017).

Engineered peptides with modified residues can further improve the effectiveness of such interventions. It has been observed that peptides with modified peptide anchor residues can improve the tumor cell killing response of the adaptive immune system (van Stipdonk et al., 2009). For peptides presented by MHC-I, not all modifications to the antigen sequences improve the recognition of peptides by the immune system, likely due to subtle structural changes that alter the TCR-binding interface (Cole et al., 2010). However, in contrast to MHC-Is, MHC-IIs have open grooves in which presented peptides are displayed in an extended conformation, resulting in peptides binding in a highly conserved manner. The peptide side chains at positions P1, P4, P6,

and P9 are completely buried within binding pockets in the groove and are considered anchor positions (Jones et al., 2006). These four anchor residues are key determinants of peptide-MHC binding affinity. Because of the highly conserved conformation of peptides within the MHC-II binding groove (Jones et al., 2006), changing the identities of the MHC-II-binding anchor residues will allow us to alter binding affinity without changing binding conformation or the T cell receptor interface.

Traditional approaches to identifying good peptide modifications is a complicated process (van Stipdonk et al., 2009), necessitating the development of computational approaches. A rule-based approach, EpitOpimizer, has been used to design modified peptides for MHC-I binding with anchor position changes that resulted in improved adaptive immune system response (Houghton et al., 2007). EpitOptimizer uses a limited sequence context for its suggestions, and each MHC-I molecule has a different set of rules. By contrast, PeptX (Knapp et al., 2011) uses a genetic algorithm to determine the peptides most likely to be displayed by a specific MHC-I allele, which may provide helpful information for the subsequent design of a vaccine. The performance of PeptX was not experimentally evaluated.

We introduce a model-based approach to optimize peptide-MHC-II binding by optimizing the peptide anchor residues of disease-associated peptides. We optimize peptide-MHC-II affinity by enumerating all possible changes to the anchor positions of a peptide, then scoring them against an objective function in silico and choosing the best ones. This is computationally tractable due to the limited number of anchor positions on a given peptide. We adapt a yeast display platform (described in **Chapter 2**) to test our improved peptide sequences for binding to MHC-II molecules.

For our objective function, we use predictions from the PUFFIN peptide-MHC binding model (Zeng and Gifford, 2019) trained on peptide binding data from a MHC-II yeast display platform (described in **Chapter 2**) (Rappazzo et al., 2020). PUFFIN uses an ensemble of deep residual networks to quantify its uncertainty about its predictions, while achieving state of the art performance on MHC-II binding prediction tasks and allowing us to perform optimization with objective functions which incorporate affinity prediction in addition to prediction uncertainty metrics (Zeng and Gifford, 2019). We show that our method generates peptide modifications that improve peptide binding affinity for two MHC-IIs.

3.2 Approach: model, model training, selection of seed sequences, and definition of optimization tasks

We utilized peptide-MHC binding data from the yeast display platform introduced in **Chapter 2** for training prediction algorithms (Rappazzo et al., 2020) (**Figure 1A** illustrates this step and the overall study). In this platform, MHC-IIs are covalently linked to a query peptide with a flexible linker which contains a 3C protease cleavage site. When the linker is cleaved, unbound peptides can be displaced from the MHC in the presence of a high-affinity competitor peptide. The linker also contains a peptide-proximal epitope tag, which we use to enrich yeast that maintain peptide-MHC binding. **Figure 1B** and **Figure 1C** illustrate the yeast display construct and selection strategy. Data was collected over multiple iterative rounds of selection. After each round of selection, deep sequencing was carried out on the enriched yeast. We filter the deep sequencing results for reads that match the invariant portions of the construct, from which we extract the

peptide sequence. The resulting dataset assigns to every observed peptide its read count for each round of the yeast display assay.

We utilized data from two MHC-II alleles: HLA-DR401 (HLA-DRA1*01:01, HLA-DRB1*04:01) and HLA-DR402 (HLA-DRA1*01:01, HLA-DRB1*04:02). This ensures that our results are not an allele specific artifact and allows us to study optimization for multiple alleles.
For training the machine learning (ML) model, we utilized enrichment data from a library consisting of 10^8 random 9mer peptides flanked by invariant peptide flanking residues (IPFR) which encourages binding in a single register and simplifies identification of anchor residues (Rappazzo et al., 2020). This dataset contains orders of magnitude more peptide binders than comparable approaches and can be used for improved peptide affinity prediction (Rappazzo et al., 2020), which will be beneficial when scoring and selecting optimized sequences.

For each MHC-II allele, we train a neural network-based ML model (PUFFIN) (Zeng and Gifford, 2019) that takes a 9 residue peptide sequence as input, and predicts a measure of the strength of the peptide-MHC interaction. Rather than the NNAlign model architecture utilized in **Chapter 2**, we selected the PUFFIN architecture because it outputs uncertainty estimates in addition to affinity predictions, which allows us to compute Bayesian acquisition functions. Utilizing the yeast display data and PUFFIN framework, we trained two predictors for each allele. The first predictor models the enrichment as a continuous value and outputs a Gaussian distribution, while the second predictor models the enrichment as categoricals and outputs a probability distribution over the categories. In both cases, the enrichment value of a given 9mer is based on the last round of its appearance in the yeast display experiment.

To utilize these prediction tools for optimization, we leverage the relatively small space of 20^4-1 possible anchor substitutions to evaluate an objective function over each substitution based on the output of the model. We then output the 10 substitutions that score the highest as the proposed optimizations. The use of a neural network-based model along with the complete enumeration of the anchor substitution space allows our optimizations to take more complex interactions between residues into account.

We proposed anchor optimizations to 9mers drawn from the proteomes of the zika, HIV, and dengue viral proteomes, which we refer to as seed sequences. We selected three sets of sequences on which to evaluate three different optimization tasks. These sets of sequences are:

1. 82 seed sequences that have some affinity for HLA-DR401, which we optimize for affinity to HLA-DR401.
2. 87 seed sequences that have some affinity for HLA-DR402, which we optimize for affinity to HLA-DR402.
3. 44 seed sequences that have high affinity for HLA-DR402 and some affinity for HLA-DR401, which we optimize for affinity to both MHC alleles.

We use predictions from the categorical ML predictor as a surrogate for affinity in this context. PUFFIN was designed to characterize the uncertainty of its predictions by outputting a variance. This allows us to use various Bayesian acquisition functions as our objectives. For this study, we chose to study point estimate (PE) which is just the enrichment, and upper confidence bound (UCB) which adds the enrichment and the standard deviation of the prediction. For our third task

of optimizing for both alleles, the objectives were computed for each allele individually and then added to produce the combined objective.

Optimizations using PE and UCB were performed with both the Gaussian and categorical models. Optimization with a given objective was done by enumerating and evaluating all possible anchor substitutions with that objective and selecting the top 10 scoring substitutions, giving a total of 40 optimized sequences for each seed. For each seed, 10 random anchor substitutions were also generated as a random control. We add these sequences to a new yeast display library for testing our designs.

Instead of adding the 9mers directly, we first flanked with IPFR so the sequences would resemble those from the original randomized library (Rappazzo et al., 2020). Therefore, the sequences will take the form “AAXXXXXXXXXWEEG”, where “X” denotes any residue. As a further control, we also flanked the 9mers with their wild type peptide flanking residues (WPFR), which were defined as the 3 residues that flanked the seed 9mer in the source proteome. Finally, we sampled some sequences that performed well and some sequences that performed poorly in the training data and added them as positive and negative controls, respectively.

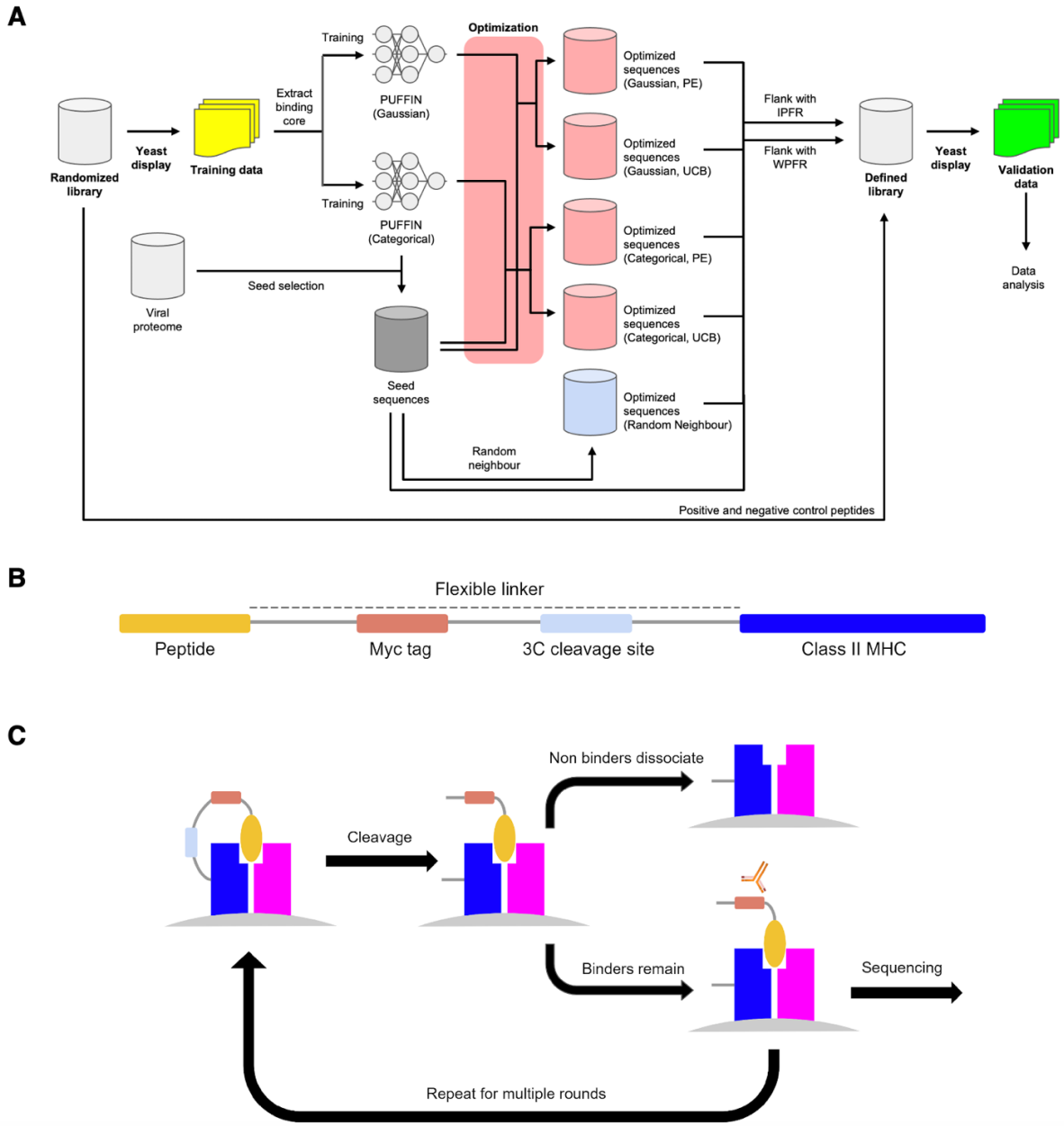


Figure 1. Generation of training and validation data. A) Details on the generation of the training and validation data. The initial randomized library is used to generate the training data. The training data is used to train two variants of PUFFIN. One of these was used to select seeds from viral proteomes. These seeds were then optimized and combined with some peptides from the original randomized library to produce the defined library. Yeast display was done on the defined library to produce the validation data, on which most of our analysis is done. **B)** Schematic of the construct used in the yeast display assay. **C)** The overall process for yeast display. First, the peptide-MHC is expressed on the surface of yeast, and then the linker between peptide and the MHC molecule is cleaved. Peptide exchange is catalyzed, and yeast are selected which retain the Myc epitope tag. The resulting population is then sequenced and carried on to the next round.

3.3 Assessment of optimized peptides using yeast display

In order to test our optimized sequences, we adapted the yeast display platform and workflow from randomized peptide libraries to presentation of user-defined peptides. We designed a 36,000-member defined library containing our optimized sequences and controls from **Chapter 3.2**. Each peptide sequence was encoded in DNA space by randomly selecting an encoding codon, weighted by probability matching yeast codon frequency (GenScript Codon Usage Frequency Table; GenScript, Piscataway, NJ). This helps avoid encoding similar peptide sequences with the same codon encoding, and thus decreases the probability of template switching between similarly encoded sequences. The defined sequences were synthesized by Twist Bioscience as a single-stranded oligonucleotide pool with a maximum length of 120 nucleotides. Utilizing this defined oligonucleotide pool, we generated a second yeast display library encoding these peptides, linked to MHCs.

To better assess enrichment, the HLA-DR401 and HLA-DR402 defined peptide libraries were doped into a 20-million-member randomized peptide library containing stop codons, at a ratio of approximately 1:500 so that each unique peptide was represented at similar starting frequency. Doping into this library provides a null set of peptides over which real binders must enrich, to increase the stringency of selections. Four iterative rounds of selection were performed (**Figure 1C**).

To compare affinities between given peptides, for each peptide we estimated the proportion of that peptide which survives between rounds. This value is determined by fitting a geometric progression to the concentration of each peptide. We assume that read counts are drawn from a Poisson distribution that is parametrized by the concentration of the peptide multiplied by a constant that is dependent on the overall population being sequenced, and we fit the maximum likelihood estimate.

By fitting a geometric progression, we can extract the proportion between successive values in the progression, which we can interpret to be the proportion of peptides that survive a single round of the experiment. The constant of the sequenced population is undetermined, so the proportion we extract is a scalar multiple of the true proportion. Therefore, this proportion is unnormalized, so we refer to it as a round survival rate (RSR). While RSR are not unique, we find that it is able to provide a highly consistent ranking to the peptides. We use RSR as a surrogate for affinity for the remainder of this text. RSR values of replicate selections of the defined library are concordant with the first replicate (Pearson and Spearman correlation coefficients 0.81-0.84; **Supplemental Figure 1**), suggesting selections and RSR determination is reproducible. A subset of sequences is absent from a single replicate due to stochastic dropout, which likely occurs in the initial rounds of selection when each member of the library is present at low frequency.

3.4 Round survival rates of peptides across optimization tasks

We first examine the overall RSR distribution of the following groups of sequences for each allele: sequences optimized for that allele with PE under the Gaussian model, UCB under the Gaussian model, PE under the categorical model, UCB under the categorical model, sequences with random anchor mutations (negative control), seed sequences, sequences from the training data which were not present after round 2 (negative control), and sequences from the training data which were present after round 2 (positive control). We find that the groups of optimized sequences exhibit

higher RSRs for the alleles they were optimized for than either of the negative controls (**Figure 2**). The improvements are statistically significant, with $p \leq 1.58e\text{-}23$ between any optimized set and negative control for either allele by the two-sided Mann-Whitney U test.

We find that no optimization method significantly outperforms the others. Between any two groups of optimized sequences, $p \geq 0.4$ under the two-sided Mann-Whitney U test for both alleles. This can be explained by the overlaps observed in the optimizations proposed by different methods. When optimizing for HLA-DR401 affinity, if we compare the proposals generated by two optimization methods there are at most 2 seed sequences out of 82 whose proposed optimizations did not include any common sequences. Likewise, for HLA-DR402 for any two optimization methods there are at most 3 seed sequences out of 87 that had no overlap. This suggests that the specific objective does not significantly affect the quality of proposals. We compared the consistency of point estimate optimization proposals where uncertainty estimates are not required between NetMHCIIpan 4.0 (Reynisson et al., 2020) and PUFFIN. We found that they proposed overlapping optimized sequences (**Table 1**).

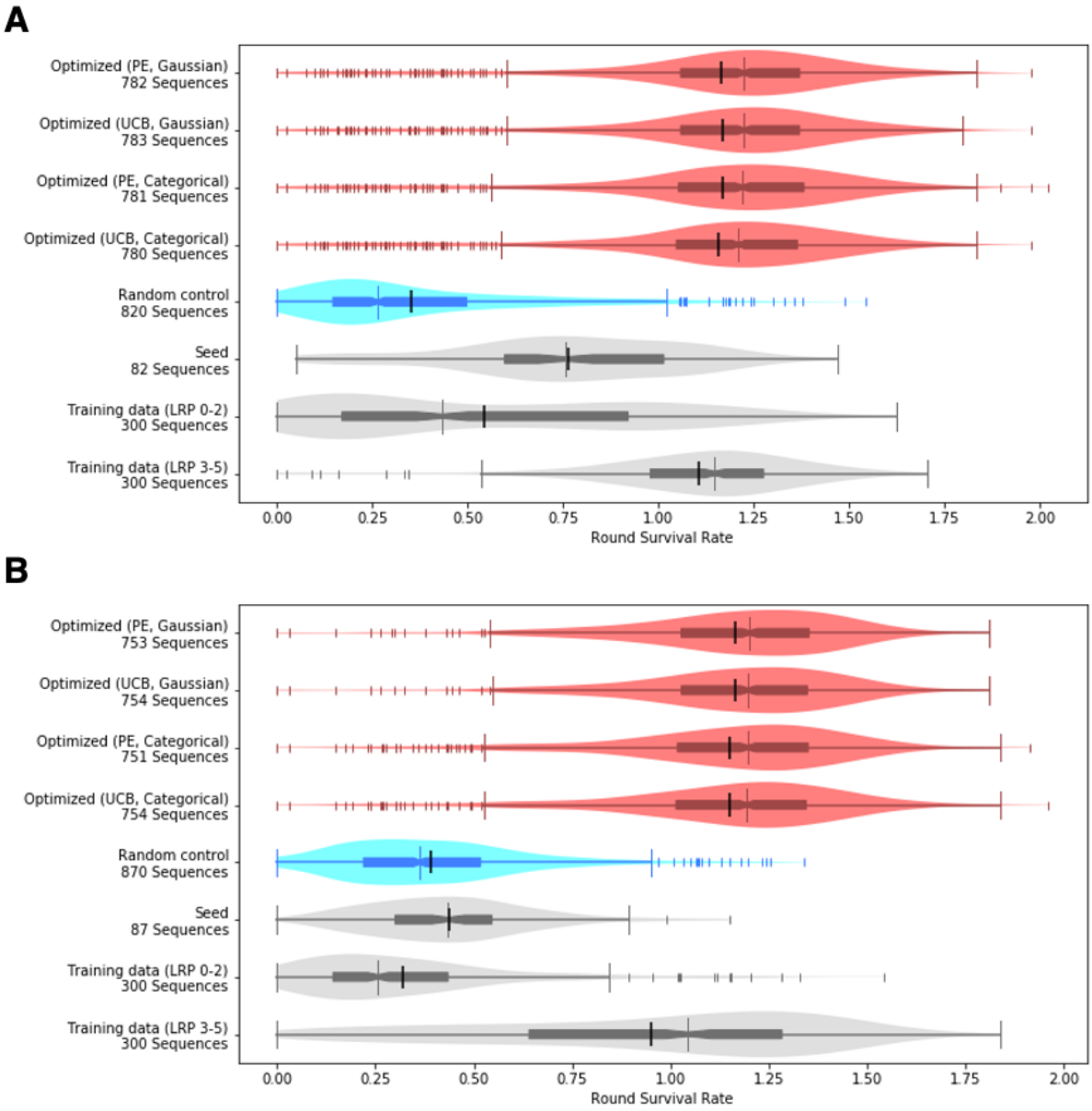


Figure 2. Anchor optimization improves round survival rate. The distributions of RSR for **A)** HLA-DR401 and **B)** HLA-DR402 is plotted for the optimized and control groups. The sequences from the training data are split into two groups: "Training data (LRP 0-2)" is composed of sequences which did not appear after round 2 in the initial yeast display assay and is shown as a negative control, while "Training data (LRP 3-5)" is composed of sequences that did appear after round 2 and is shown as a positive control. The differences between the optimized groups and the negative controls (Random control, Seed and Training data (LRP 0-2)) are significant for both alleles, with $p \leq 1.58e\text{-}23$ under the two-sided Mann-Whitney U test. Each plot is a combination of a box plot and a violin plot, where the distribution is shown by the violin plot in a lighter color, and the box plot shows the middle quartiles in a darker color along with the median. The mean is indicated by a black vertical line. Flier points are marked with the "|" symbol.

Optimized Set	PE, Gaussian	UCB, Gaussian	PE, Categorical	UCB, Categorical	NetMHCII pan 4.0	Random Neighbor
HLA-DR401 PE, Gaussian	82/82	82/82	82/82	82/82	57/82	0/82
HLA-DR401 UCB, Gaussian	82/82	82/82	82/82	82/82	50/82	0/82
HLA-DR401 PE, Categorical	82/82	82/82	82/82	80/82	65/82	0/82
HLA-DR401 UCB, Categorical	82/82	82/82	80/82	82/82	57/82	0/82
HLA-DR401 NetMHCIIpan 4.0	57/82	50/82	65/82	57/82	82/82	0/82
HLA-DR401 Random Control	0/82	0/82	0/82	0/82	0/82	82/82
HLA-DR402 PE, Gaussian	87/87	87/87	86/87	85/87	45/87	0/87
HLA-DR402 UCB, Gaussian	87/87	87/87	84/87	85/87	44/87	0/87
HLA-DR402 PE, Categorical	86/87	84/87	87/87	87/87	52/87	0/87
HLA-DR402 UCB, Categorical	85/87	85/87	87/87	87/87	49/87	0/87
HLA-DR402 NetMHCIIpan 4.0	45/87	44/87	52/87	49/87	87/87	0/87
HLA-DR402 Random Control	0/87	0/87	0/87	0/87	0/87	87/87
Joint PE, Gaussian	44/44	44/44	44/44	44/44	27/44	0/44
Joint UCB, Gaussian	44/44	44/44	44/44	44/44	25/44	0/44
Joint PE, Categorical	44/44	44/44	44/44	44/44	27/44	0/44
Joint UCB, Categorical	44/44	44/44	44/44	44/44	26/44	0/44
Joint NetMHCIIpan 4.0	27/44	25/44	27/44	26/44	44/44	0/44
Joint Random Control	0/44	0/44	0/44	0/44	0/44	44/44

Table 1. Overlapping predictions from alternate prediction algorithms. Each column indicates a proposed optimization method, and each row indicates a proposed optimization. Each entry indicates the number of seeds with overlapping proposed peptides between optimization methods (out of 82, 87, and 44 seed sequences for HLA-DR401, HLA-DR402, and joint optimization, respectively). "PE, Gaussian", "UCB, Gaussian", "PE, Categorical", and "UCB, Categorical" are the main optimization methods we analyzed which use PUFFIN predictions, "NetMHCIIpan 4.0" uses the EL score predicted by NetMHCIIpan 4.0 as an objective function but otherwise operates identically to the other optimization methods, and "Random Control" draws anchor substitutions randomly.

3.5 Optimization for single or multiple alleles

Since our method of generating proposals performs comparably under the different objectives we tested, we will focus the rest of our analysis on point estimate optimization under the Gaussian model for simplicity. We include an analogous analysis of the other optimization methods, which are similar (**Supplemental Figures 2-5**).

We find that most optimized sequences outperform their unaltered seed sequences (**Figure 3A, 3C**). For HLA-DR401, for 44 out of the 82 seed sequences, all of the proposed optimizations performed better, while for HLA-DR402 this was the case for 72 out of the 87 seed sequences. In

sequences where optimization was less effective, we find that generally the seed sequence already performs well via round survival rate (**Figure 3B, 3D**).

We find that sequences that were optimized for both alleles were generally able to improve their RSR for HLA-DR401 while maintaining their RSR for HLA-DR402 (**Figure 4**). Out of a total of 44 seed sequences, there were 35 in which all proposed optimizations had a higher RSR for HLA-DR401. For 23 seed sequences, all proposed sequence optimizations outperformed the seeds on HLA-DR401 and achieved greater than 80% of the seed sequence RSR for HLA-DR402. For 13 seed sequences, all proposed optimizations outperformed the seeds on both HLA-DR401 and HLA-DR402.

If we instead consider seeds where the optimization criterion was reached for at least 8 out of the 10 proposed sequences, these values rise to 42, 35, and 18 respectively. For the random controls, they are 2, 1, and 1 (**Figure 4**).

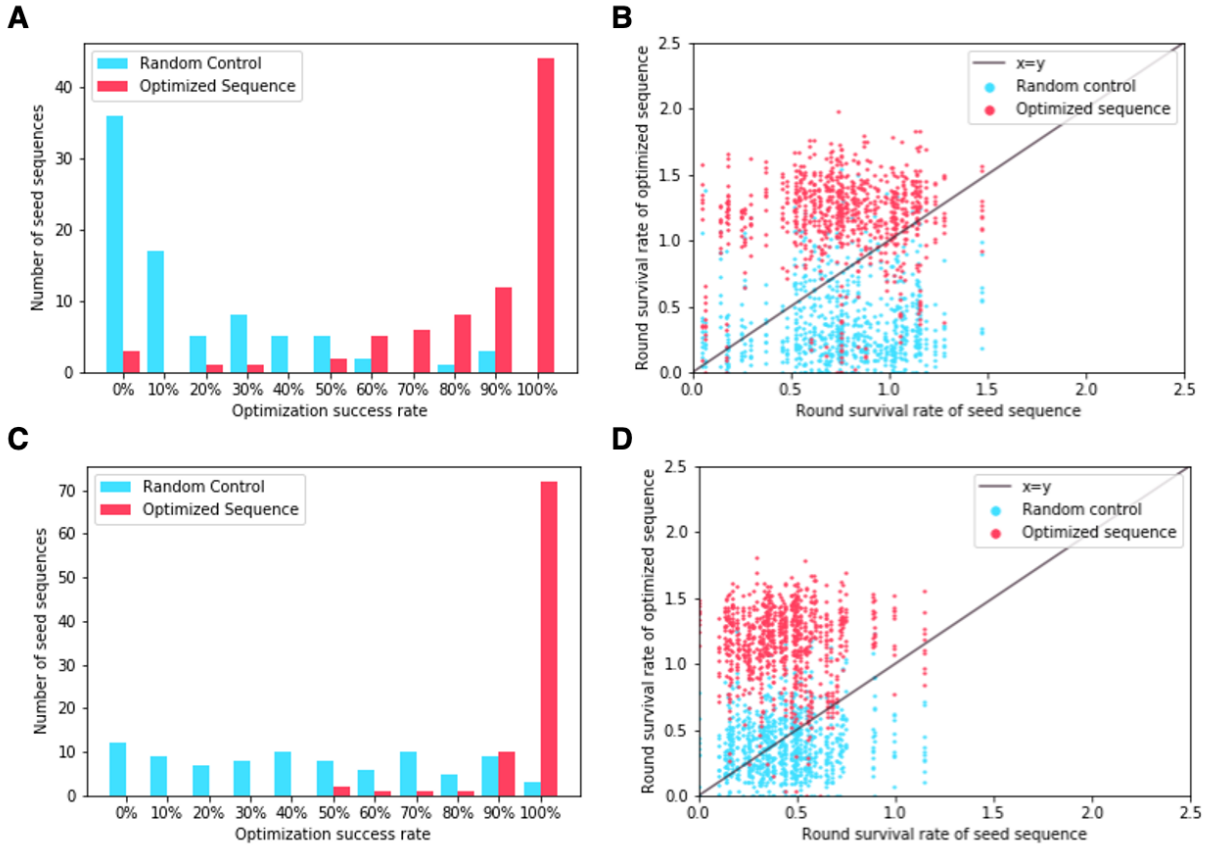


Figure 3. Number of sequences that exhibit improvement after optimizing with the point estimate objective under the Gaussian model. **A)** For each seed sequence, we calculate the number of proposed optimizations that achieve a RSR for HLA-DR401 that is higher than that of the seed. We then take that as a percentage of the number of proposals to obtain the optimization success rate. We plot the distribution of these rates for both sequences optimized for HLA-DR401 affinity and the randomly perturbed sequences. **B)** For each sequence optimized for HLA-DR401 affinity and randomly perturbed sequence, we plot their RSR for HLA-DR401 against the RSR of the seed sequence they derive from. **C)** We calculate the distribution of optimization success rates for sequences optimized for HLA-DR402 using RSR for HLA-DR402. **D)** We plot the RSR for HLA-DR402 of sequences optimized for HLA-DR402 against the RSR of their seed sequence.

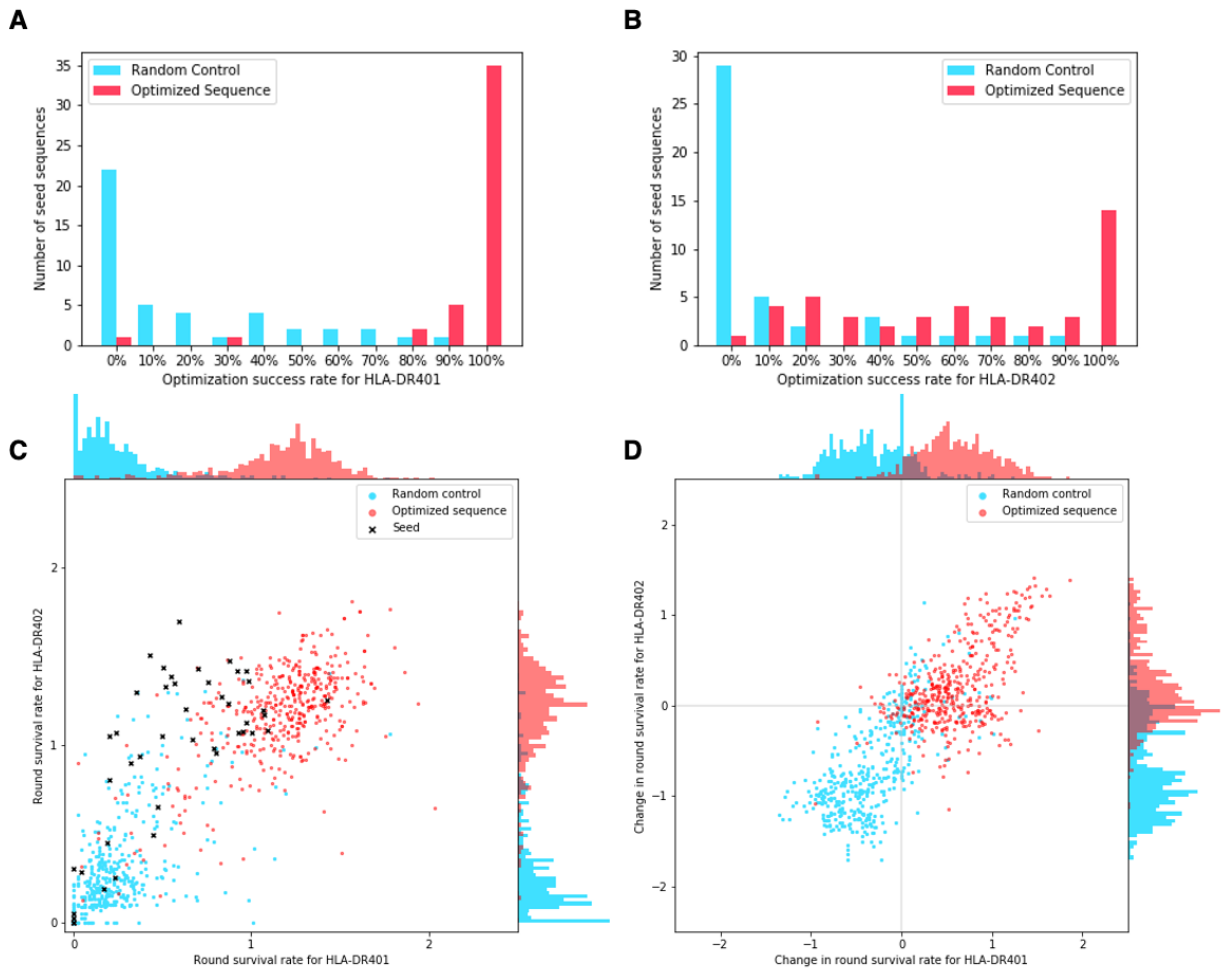


Figure 4. Number of sequences that exhibit improvement for multiple alleles after optimizing with the point estimate objective under the Gaussian model. A) For each seed sequence, we calculate the number of proposed optimizations that achieve a RSR for HLA-DR401 that is higher than that of the seed. We then take that as a percentage of the number of proposals to obtain the optimization success rate. We plot the distribution of these rates for both sequences optimized for HLA-DR401 and HLA-DR402 affinity and the randomly perturbed sequences. **B)** We produce the same distribution but with optimization success rates based on HLA-DR402 affinity. The seed sequences were selected to have high HLA-DR402 affinity. **C)** For each optimized, random control, and seed sequence, we plot their RSR for both alleles. **D)** For each optimized and random control sequence, we take their RSR and subtract the RSR of the seed sequence they derive from to obtain the changes in their RSR.

3.6 Complexity of interactions between residue positions captured by optimizations

By analyzing our training data, we find that the identity of residues outside of the primary anchor residues can have a significant impact on which anchor residues will improve affinity. As an example (**Figure 5**), for HLA-DR401 if a sequence contains a threonine (T) at the non-anchor position P7, then having an aspartic acid (D) at anchor position P6 tends to increase the RSR. However, if the sequence contains a D at the non-anchor position P7, then having a D at P6 tends to decrease the RSR instead. Higher order effects can be seen between other anchor and non-anchor positions as well, so these relationships are not limited to adjacent positions nor to residues at P7, which can be considered an auxiliary anchor because of its contacts with the MHC groove (Jones et al., 2006).

The dependency between anchor positions and non-anchor positions can be observed in the proposals generated by our method. Out of the 820 sequences proposed using PE under the Gaussian model for HLA-DR401, 20 (2%) have a D at anchor position P6. Six of our seed sequences have a T at P7; of our proposed optimizations for these seeds, 17/60 (28%) have a D at P6. Conversely, nine of our seed sequences have a D at P7, and none of their 90 proposed optimizations have a D at P6. This demonstrates the advantages of enumerating the full anchor residue landscape as it allows the capture of these higher order effects.

Given the presence of the higher order effects between peptide positions, including non-anchor positions, it seems unlikely that a more naïve approach to anchor optimization could be as successful. In particular, it is unlikely that there exists a set of anchor residues that would optimize affinity in all non-anchor contexts. As further support for this, we find that there are no sets of anchor residues that were proposed for all seed sequences for any optimization task, even when combining the proposed optimizations across all 4 of our optimization methods. For HLA-DR401 optimization, the most frequently proposed set is Y, D, T, A at anchor positions P1, P4, P6, P9 (respectively), which was proposed for 54 out of the 82 seeds. For HLA-DR402, the most frequently proposed set is L, W, T, A at P1, P4, P6, P9 (respectively), which was proposed for 44 out of the 87 seeds. For optimization for both alleles, the most frequently proposed set is F, M, N, A at P1, P4, P6, P9 (respectively), which was proposed 34 out of 44 times.

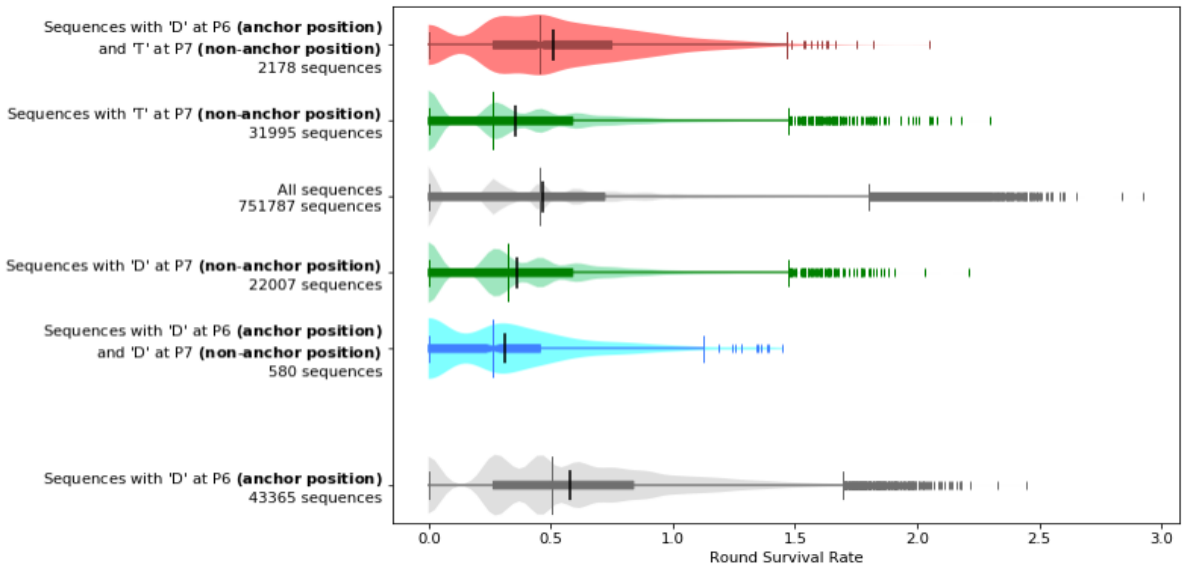


Figure 5. Whether a given anchor residue improves affinity can depend on non-anchor residues. Peptides with aspartic acid at P7 tend to have lower round survival rates when compared to all peptides, and peptides that additionally have another aspartic acid at anchor position P6 tend to have even lower round survival rates than the peptides that just have an aspartic acid at P7. In contrast, although peptides with threonine at non-anchor position P7 also tend to have lower round survival rates when compared to all peptides, peptides that additionally have an aspartic acid at anchor position P6 tend to have higher round survival rates instead even when compared to all peptides. The differences found in these comparisons are significant, with $p \leq 4.968e\text{-}5$ between any two groups mentioned above under the Mann-Whitney two-sided U test. The sequences plotted and used for computing significance are from the training data. Each plot is a combination of a box plot and a violin plot, where the distribution is shown by the violin plot in a lighter color, and the box plot shows the middle quartiles in a darker color along with the median. The mean is indicated by a black vertical line. Flier points are marked with the "|" symbol.

3.7 Effects of invariant flanking residues on optimized peptides

All the optimized peptides we have presented so far have been flanked with IPFR, which were used to train the model. If we replace the IPFR with WPFR, we observe that the optimized sequences still outperform the seed and random controls **(Figure 6)**. The improvement is still significant, with $p \leq 5.51e\text{-}5$ when comparing the optimized sequences to the random control or seed sequences for either allele under the two-sided Mann-Whitney U test. However, the optimized sequences with WPFR significantly underperform their IPFR counterparts ($p \leq 1.36e\text{-}22$ for either allele under the two-sided U test).

The reduction in improvement is only observed in the optimized sequences for WPFR. In the case of the seed and random control groups, the WPFR sequences either do not display any significant difference or mildly outperform the IPFR counterparts ($0.0018 \leq p \leq 0.92$ under the two-sided Mann-Whitney U test).

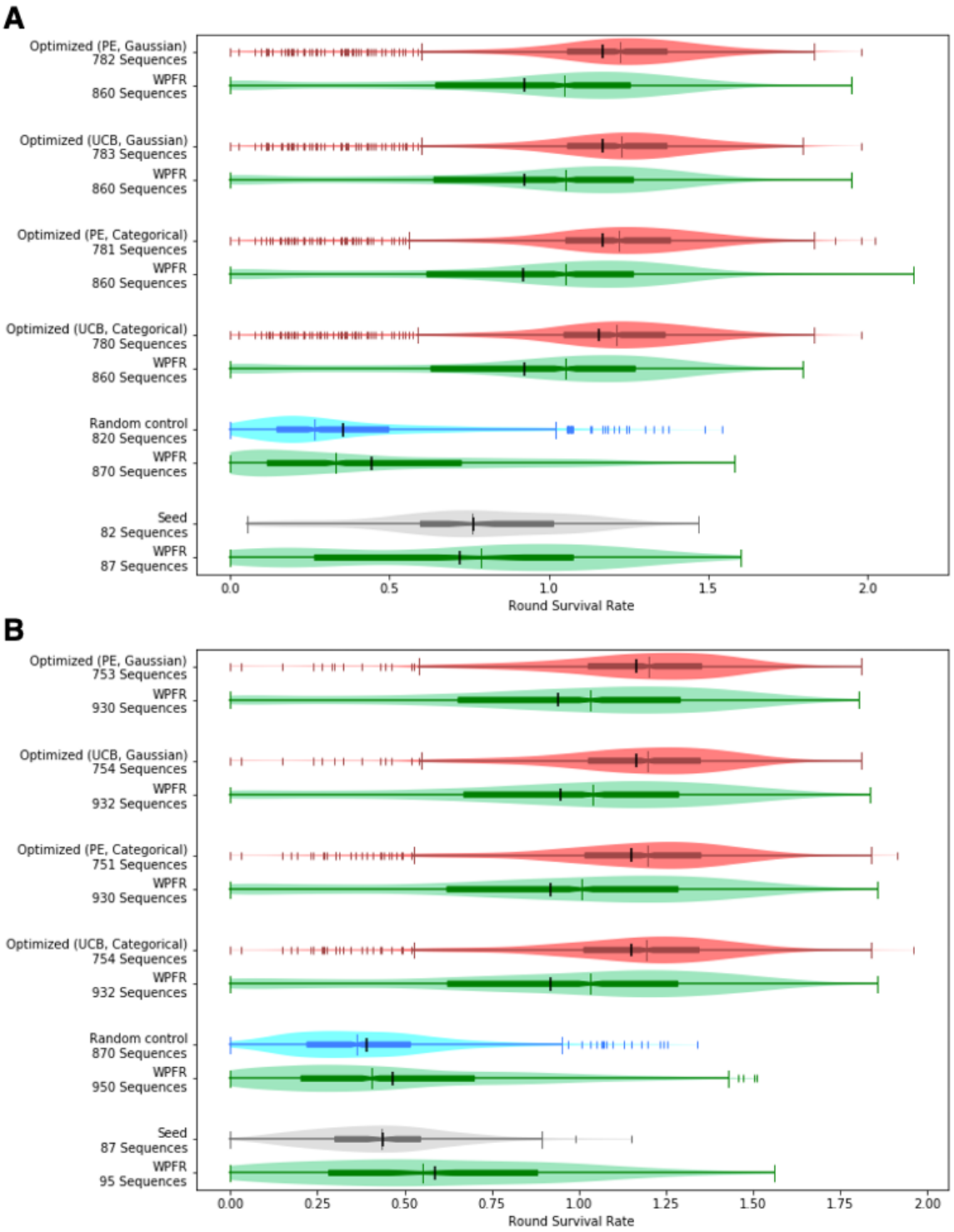


Figure 6. Comparing round survival rate of various groups with invariant peptide flanking residues and wild type peptide flanking residues. From top to bottom, the first four pairs of groups depicted in that figure are optimized sequences, and the pairs below them are sequences with random anchor mutations and seed sequences. The upper group of each pair contain IPFR flanked sequences and are the same as those in **Figure 2**, while the lower group in green contain WPFR flanked sequences. The distributions of RSR for **A)** HLA-DR401 and **B)** HLA-DR402 are plotted for these groups. Each plot is a combination of a box plot and a violin plot, where the distribution is shown by the violin plot in a lighter color, and the box plot shows the middle quartiles in a darker color along with the median. The mean is indicated by a black vertical line. Flier points are marked with the "|" symbol.

3.8 Sequence motifs of optimized peptides

The peptide optimizations made by our machine learning models are consistent with the structures and peptide-binding motifs of HLA-DR401 and HLA-DR402 in our training data. The polymorphisms between HLA-DR401 and HLA-DR402 affect the P1 and P4 binding pockets. Both alleles prefer hydrophobic amino acids in the P1 pocket, although HLA-DR401 prefers larger amino acids, while the truncated HLA-DR402 pocket prefers smaller amino acids. In the P4 pocket, HLA-DR401 prefers acidic residues, and HLA-DR402 prefers basic residues and large hydrophobic residues. The conserved P6 and P9 binding pockets prefer polar and small amino acids, respectively. The preference for each allele is reflected in MHC allele-specific peptide optimization, shown for the optimization with PE objective under the Gaussian model as an example (**Figure 7A**). Joint MHC optimization is also consistent with these preferences: P1 and P4 amino acids are mutually preferred between both alleles, such as F/I/L at P1 and increased usage of M at P4. P6 and P9 amino acids are consistent with usage in individual allele-optimized peptides. Amino acid frequency in the seed sequences is also shown for reference (**Figure 7B**).

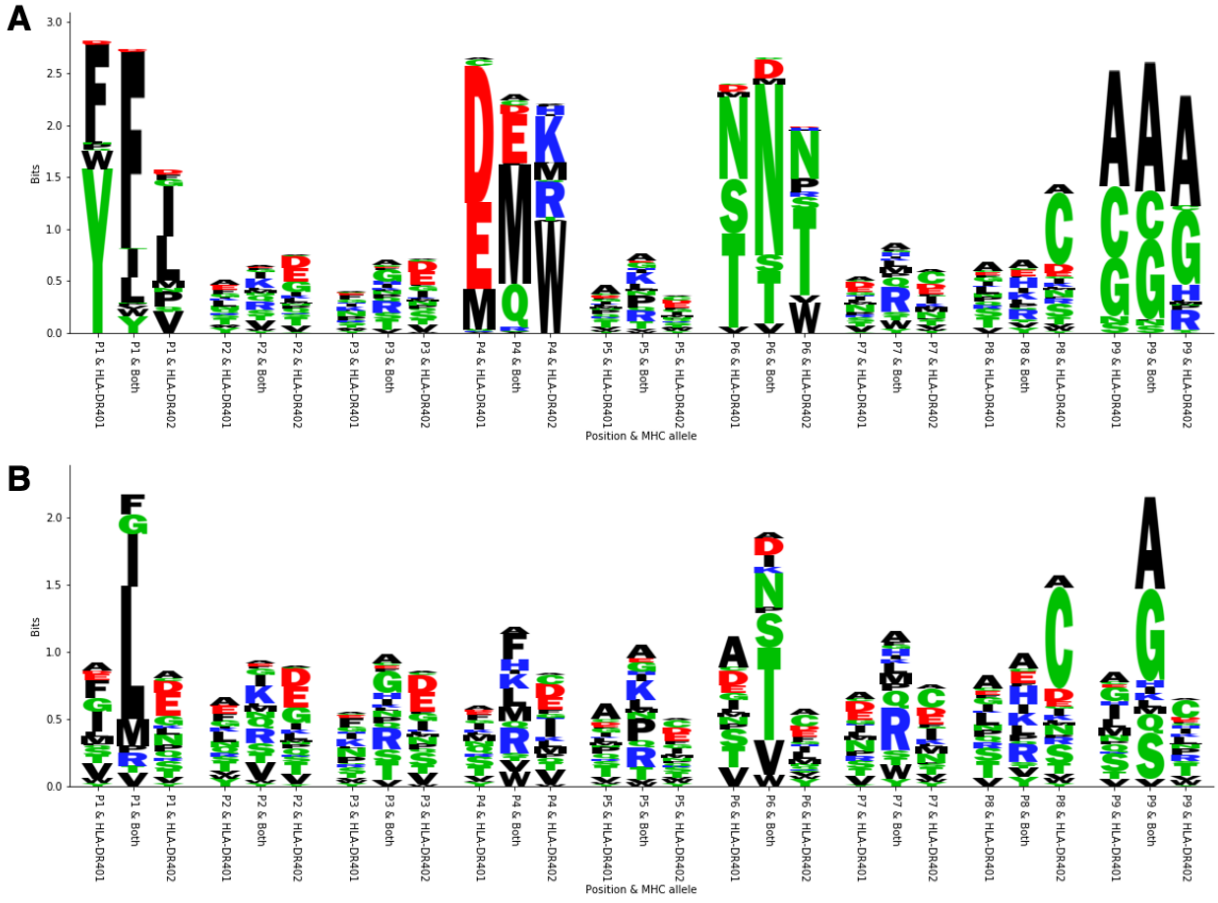


Figure 7. Motifs arising from optimization with the PE objective under the Gaussian model. A) Sequence logos depicting the motifs present in the optimized sequences. For each position, 3 different residue distributions are shown. The first one shows the distribution for the sequences optimized for HLA-DR401, the last one shows the distribution for the sequences optimized for HLA-DR402, and the one in the middle shows the distribution for the sequences optimized for both alleles. Sequence logos were generated with a custom script. **B)** Sequence logos depicting the motifs present in the original seed sequences for comparison with the same setup.

3.9 Discussion and outlook

In this work, we introduced our method for optimizing the affinity of peptide sequences for MHC-IIs by replacing their anchor residues with more optimal residues generated with the help of a machine learning model. We validated this technique on two different MHC-II alleles, and showed that it is possible to optimize a single sequence for multiple alleles simultaneously. We have developed a high-throughput yeast display-based pipeline to test our optimized sequences, and we introduced the notion of a round survival rate which allows us to compare the results of the assay.

We demonstrated that our method leverages deep learning models in a way that allows the proposed optimizations to capture complex interactions between residues. Our ability to optimize sequences for two alleles simultaneously suggests that the method can generalize to even more complicated objectives. These contributions improve our ability to engineer peptides for therapeutic purposes, and allows us to develop more robust cocktails by allowing their constituent peptides to fulfill multiple objectives.

We note that the method for generating proposals is independent of the specific predictor used. Our results indicate that taking uncertainty into account does not significantly improve the quality of the proposals over using a simple point estimate, so a model that quantifies its uncertainty is not strictly necessary. Therefore, substituting PUFFIN for another algorithm that performs comparably to PUFFIN should yield similar results.

As a caveat, we note that our optimization is less effective if we allow arbitrary flanking residues. The reduction in improvement when we change from IPFR to WPFR is only observed in the optimized sequences and is not observed in seed or sequences with random anchor residues. Therefore, it is likely that the drop in performance is due to the predictor being trained on IPFR data, so the predictor is unable to take the effects of flanking residues or register shifts into account. The IPFRs also contain preferred amino acids in the flanking sequences, such as the tryptophan at position P10. Aromatic residues at P10 have been shown to bolster binding and may impact the superior performance of IPFR peptides compared to WPRF peptides (Rappazzo et al., 2020; Zavala-Ruiz et al., 2004). As noted above, since our method is independent of the specific underlying predictor, we should be able to address this issue by replacing our current predictor with one that takes the flanking residues into account. More generally, the quality of our optimization should improve as the quality of predictors available continues to improve.

Our future work will extend our method to incorporate wild type flanking residue information in our optimization, and will seek to characterize the effect of anchor optimization on peptide immunogenicity. Future work can also utilize the yeast display defined library approach for other objectives, including assessing antigenic peptides of interest for MHC-binding in high-throughput, which is explored further in **Chapter 4**.

3.10 Methods

Collecting enrichment data using a high throughput yeast display assay

We utilize peptide-MHC binding data from a yeast display library of 10^8 random 9mer peptides, as also described in **Chapter 2** (**Figure 1**) (Rappazzo et al., 2020). The peptides are flanked by invariant peptide flanking residues (IPFR), which encourages binding in a single register and simplifies identification of anchor residues. The IPFR consists of "AA" on the N-terminus and

"WEEG" on the C-terminus. Paired-end sequencing reads (Rappazzo et al., 2020) were assembled via FLASH (Magoč and Salzberg, 2011) and filtered for correct length and 3C cut site sequence.

We adapted the yeast display platform for testing our optimized sequences, encoding the peptide sequences in DNA and ordering as a defined oligonucleotide pool from Twist Bioscience. The oligo pool was amplified with low cycle number PCR then amplified with construct DNA using overlap extension PCR. This longer DNA product was assembled with the linearized pYAL vector in yeast at a 5:1 mass ratio of insert:vector and electroporated into electrocompetent RJY100 yeast. To increase the stringency of selections, the defined peptide libraries were doped into a 20-million-member randomized peptide library containing stop codons so that each unique peptide was represented at similar starting frequency. The diverse null library had the peptide encoded as "NNNTAANNNNNNNNNTAGNNNNNNNNNNNNTGANNNNNN", where N indicates any nucleotide.

For each round of selection, yeast were washed into PBS, with competitor peptide (HLA-DR401: $HA_{306-318}$, 1 μM; HLA-DR402: $CD48_{36-53}$, 5 μM) and 1 μM 3C protease, then incubated for 45 minutes at room temperature. After incubation, yeast were washed into cold acid saline (20 mM pH 5 citric acid, 150 mM NaCl) with competitor peptide (same concentration as first incubation) and 1 μM HLA-DM, then incubated overnight at 4°C. Negative selections for non-specific binders was performed with ∝-AlexaFluor647 magnetic beads (Miltenyi Biotec; Bergish Gladbach, Germany), followed by a positive selection consisting of incubation with ∝-Myc-AlexaFluor647 antibody (1:100 volume:volume) and positive selection with ∝-AlexaFluor647 magnetic beads. The first round was conducted on 400 million yeast for 20x coverage of peptides and incubations were conducted in 2 mL PBS and 4 mL acid saline. For subsequent rounds, 25 million yeast were selected; incubations were conducted in 250uL PBS, 500uL acid saline. Four iterative rounds of selection were performed and repeated in duplicate. Between rounds, yeast were grown to confluence at 30°C in SDCAA (pH 5) yeast media and sub-cultured into SGCAA (pH 5) media at OD_{600}=1 for two days at 20°C (Chao et al., 2006).

Following selections, plasmid DNA was isolated from 10 million yeast from each round using a Zymoprep Yeast Miniprep Kit (Zymo Research; Irvine, CA). Amplicons were generated to capture the peptide through the 3C protease site. Unique barcodes were added for each library and round of selection and i5 and i7 anchors added through two rounds of PCR. Amplicons were sequenced on an Illumina MiSeq (Illumina; San Diego, California) at the MIT BioMicroCenter, with a paired-end MiSeq v2 300nt kit.

Forward and reverse reads are assembled using PandaSeq. Data were processed using in-house scripts to extract peptide sequences with correctly encoded constant flanking regions. Peptides were filtered for exact matches to the defined sequences ordered from Twist and those matching the DNA encoding of the randomized null library.

HLA-DM was recombinantly expressed as previously described (Rappazzo et al., 2020). In brief, the ectodomains of the alpha and beta chains were followed by a poly-histidine purification site and encoded in pAcGP67a vectors. Plasmids for each chain were separately transfected into SF9 insect cells with BestBac 2.0 baculovirus DNA (Expression Systems; Davis, CA) and Cellfectin II reagent (Thermo Fisher; Waltham, MA). Cells were propagated to high virus titer, co-titrated to

ensure an equal ratio of alpha and beta expression, and co-transduced into Hi5 cells. Following 48-72 hours of incubation, proteins were purified with Ni-NTA resin and purified with size exclusion chromatography on an AKTAPURE FPLC S200 increase column (GE Healthcare; Chicago, IL).

Training a neural network-based ML model to predict the enrichment category of a peptide
We trained a neural network-based ML model (PUFFIN) (Zeng and Gifford, 2019) to predict the enrichment label of a given peptide. The predictor takes a 9-residue peptide sequence as input, and outputs an enrichment label. We use the final round where a peptide is observed in the yeast display assay as our enrichment label for both training and prediction. For example, if a sequence appears in the sequencing reads for round 3 but fails to appear in round 4 or any other future rounds, it receives the label "3". To improve the granularity, round 5 presence was further split up into 3 categories, where "5" indicates round 5 presence with less than 10 read counts in round 5, "6" indicates round 5 presence with a read count between 10-99 inclusive, and "7" indicates round 5 presence with a read count of 100 or more. A label of 0 is given to sequences that only appear before any enrichment is performed. This gives a total of 8 enrichment categories.

PUFFIN is an ensemble of deep residual neural networks that is regularized by dropout and controlled for overfitting with validation data. Each component model consists of one convolutional layer, five residual blocks, and one output layer. Each residual fits the difference between the input and the output of a residual block with two convolutional layers. Each convolutional layer has 256 convolutional filters and is followed by a batch-norm layer. ReLU is used as non-linearity throughout the network.

For each allele, we trained two predictors to predict the enrichment labels. The first predictor assumes the enrichment labels 0-7 are realizations of a continuous random variable taken from a Gaussian distribution, and was trained to output a mean and variance. The second predictor models the labels as categoricals, and outputs a discrete probability distribution over the 8 labels 0-7. For regularization, dropout (Srivastava, 2014) is used in the output layer with a dropout probability of 0.2. We randomly hold out 10% of the data for validation, and the rest is used for training. We use Adam (Kingma and Ba, 2014) to minimize the negative log-likelihood of the observed enrichment under the probability distribution parameterized by the output of the neural network. We train for 50 epochs and select the model from the epoch where validation loss is minimized.

While the outputs of each predictor naturally characterize aleatoric uncertainty, we also characterize the epistemic uncertainty through ensemble methods (Lakshminarayanan et al., 2016). Specifically, we generate 10 training and validation splits of our data and train 2 separate predictors for each split, giving us an ensemble of 20 predictors. When performing predictions, we run each predictor 50 times with dropout turned on (Gal and Ghahramani, 2015), resulting in a total of 1000 predictions for each input. The final output is then characterized by a mean and variance, where the mean is the average of the distribution means over all 1000 trials, and the final variance is the average of the distribution variances for each trial plus the variance of the distribution means.

We also ran PUFFIN predictions on a published orthogonal 13mer peptide yeast display test set (Rappazzo et al., 2020), to compare to existing models. We observe comparable performance between PUFFIN and current state of the art (AUC of 0.91, 0.91, and 0.82, for Gaussian PUFFIN,

Categorical PUFFIN, and NetMHCIIpan 4.0 %rank eluted ligand mode, respectively, on a dataset with a 1:1 ratio of binders to non-binders; and positive predictive values of 0.57, 0.57, and 0.29, for Gaussian PUFFIN, Categorical PUFFIN, and NetMHCIIpan 4.0 %rank eluted ligand mode, respectively, on dataset with 1:19 ratio of binders to non-binders), where PUFFIN-identified cores are the maximum-scoring 9mer within the 13mer.

Using an ML model to compute an objective function for anchor optimization
For both the Gaussian and the categorical predictors, we considered two different objective functions for scoring 9mers: point estimate (PE) and upper confidence bound (UCB). In both functions, we first run our predictor over the 9mer to obtain a predicted mean and variance. Then to compute the PE objective, we simply return the mean. To compute UCB, we return the sum of the mean and the standard deviation, which we take to be the square root of the variance.

This gives us a total of 4 methods for scoring 9mers. For each method, given an input 9mer to optimize, we enumerate all possible residue substitutions at positions 1, 4, 6, and 9 (sequences are 1-indexed). For each substitution, we compute its score using our objective function, and in the end, we output the 10 sequences that score the highest as proposed optimizations.

Designing a validation library to test the efficacy of anchor optimization
We tested the efficacy of anchor optimization on three tasks using a yeast display approach. Our evaluation was conducted using viral peptides selected from the zika, HIV, and dengue viral proteomes. The 9mers in the candidate proteomes have no overlap with the peptides in our random peptide training library. We selected seeds for optimization from zika, HIV, and dengue based on the predictions of the categorical predictor. We first filter the sequences by removing all whose PUFFIN prediction has a predicted variance higher than the median predicted variance. For the seeds for Task 1 and Task 2, we selected peptides with a predicted enrichment mean between 2 and 3, yielding 82 seeds for HLA-DR401 and 87 seeds for HLA-DR402. For the seeds for Task 3, we selected peptides with a predicted HLA-DR401 enrichment mean below 3 and a predicted HLA-DR402 enrichment mean above 5, resulting in 44 seeds.

For each seed sequence, we ran each of our 4 optimization methods over it for each allele, giving 10 optimized sequences for each method. As a control, we also proposed 10 random anchor residue mutations for each seed.

We then take the seed, optimized, and random sequences and flank them with IPFR. As a control, we also produce a second set of sequences from the same 9mers but flanked with WPFR, defined to be the 3 residues that flank the original seed sequence in the original proteome. This forms the basis of the library.

As a further control, we added sequences from the original training data to the library. For each allele, we sampled 300 sequences that had no presence after the second round of selections, and 300 sequences that had presence after the second round of selections, giving us 1200 sequences overall.

Calculating round survival rate as a representation of enrichment

The enrichment information reported in the yeast display assay comes in the form of a vector of read counts indexed by round. In order to compare enrichment between different peptides, we assign to each peptide a value that can be interpreted as an unnormalized proportion of that peptide that survives between rounds of enrichment. We will refer to this quantity as a round survival rate (RSR), where a higher RSR will be indicative of higher enrichment.

To calculate a peptide's RSR, we consider a simplified model where the peptide has a starting concentration drawn from a given prior, and the dominating event is peptide dissociation from the MHC-II. Additionally, we assume that we can treat the entire experiment as though it was happening in one solution, and everything that occurs between rounds can be captured by a scaling factor. Finally, we suppose that read counts follow a Poisson distribution parameterized by the concentration multiplied by a scaling factor.

$$R_{S,i} \sim \text{Poisson}(a_i c_s p_S^i)$$
$$\ln(c_s) \sim \text{Gaussian}(0,1)$$

For all peptides S and all rounds i, where $R_{S,i}$ is the read count of peptide S in round i, c_S is the starting concentration of the peptide S, p_S is an unnormalized proportion of peptide S that survive to the next round, and a_i is a round-specific constant. The prior for constraining c_S is for regularization purposes, and a log normal distribution was selected for its interpretability as the result of geometric Brownian motion.

We then define the RSR for peptide S as the maximum a posteriori (MAP) estimate of p_S. This value is not unique, as an adequate scaling in the a_i values can give the same probabilities with different p_S values. However, such a transformation preserves the ratio between p_S, and in practice we find that the estimates converge reliably. We estimate these values by iteratively optimizing each variable individually for 500 rounds. 23 rounds were carried out with random initializations, where p_S were drawn from Uniform(0.1,1) and $\ln(c_S)$ were drawn from Gaussian(0,1).

For experiments conducted over the defined library, we use the null library to construct a baseline model where the read count in each round follows its own Poisson distribution. The lambda parameter for each distribution was estimated by average read counts (with added pseudocounts for peptides which don't show up in any round added to make the variance of the distribution in the zeroth round match the mean). When performing MAP estimation, an additional parameter is given to each peptide which indicates whether it comes from this baseline distribution or from the model described above to filter out noise.

Data and code availability

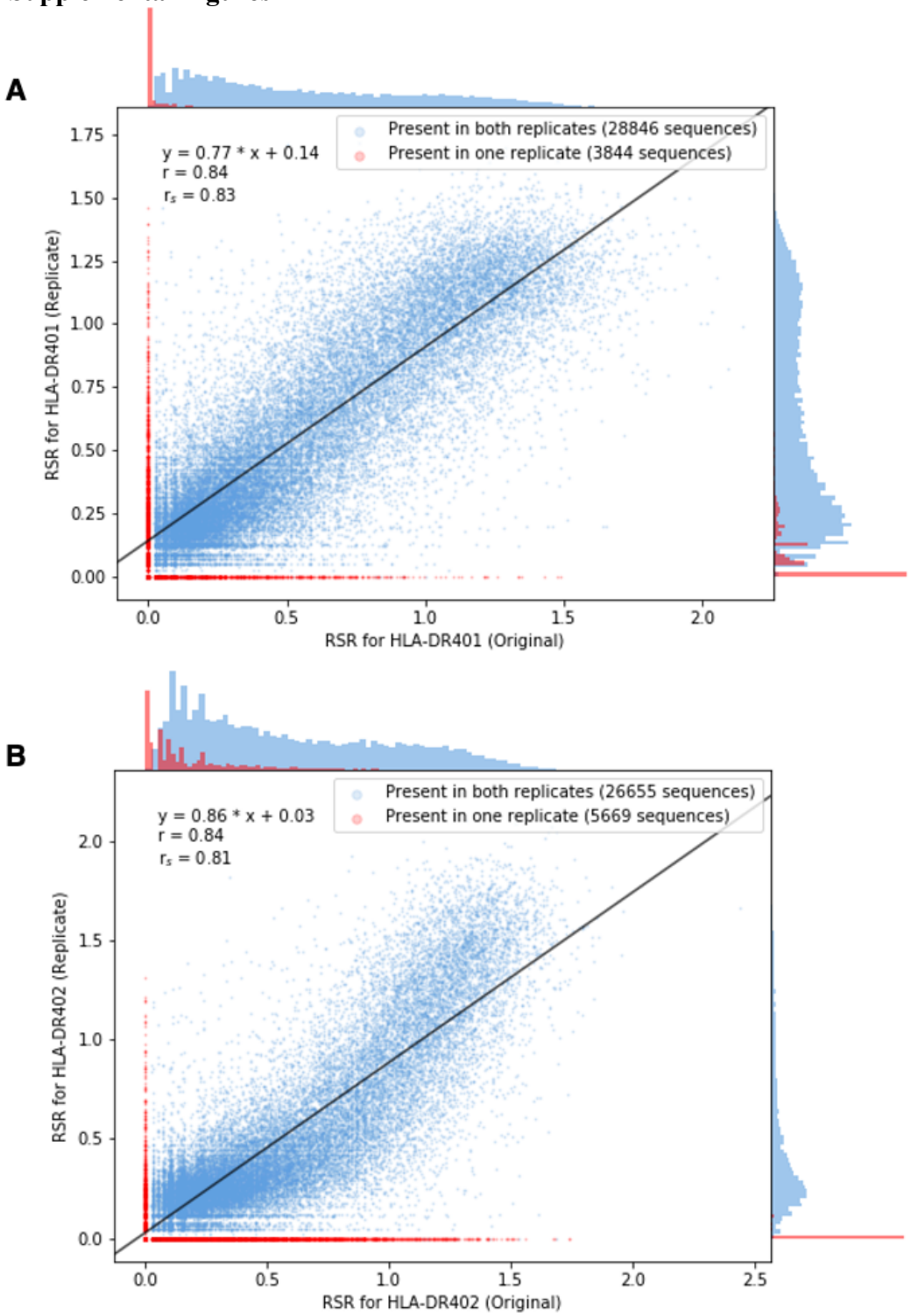
Code utilized in this study can be accessed on GitHub at https://github.com/zheng-dai/MHC2-optimization. New deep sequencing data are available on the NCBI Sequence Read Archive with accession code PRJNA708266 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA708266).

3.11 Acknowledgements

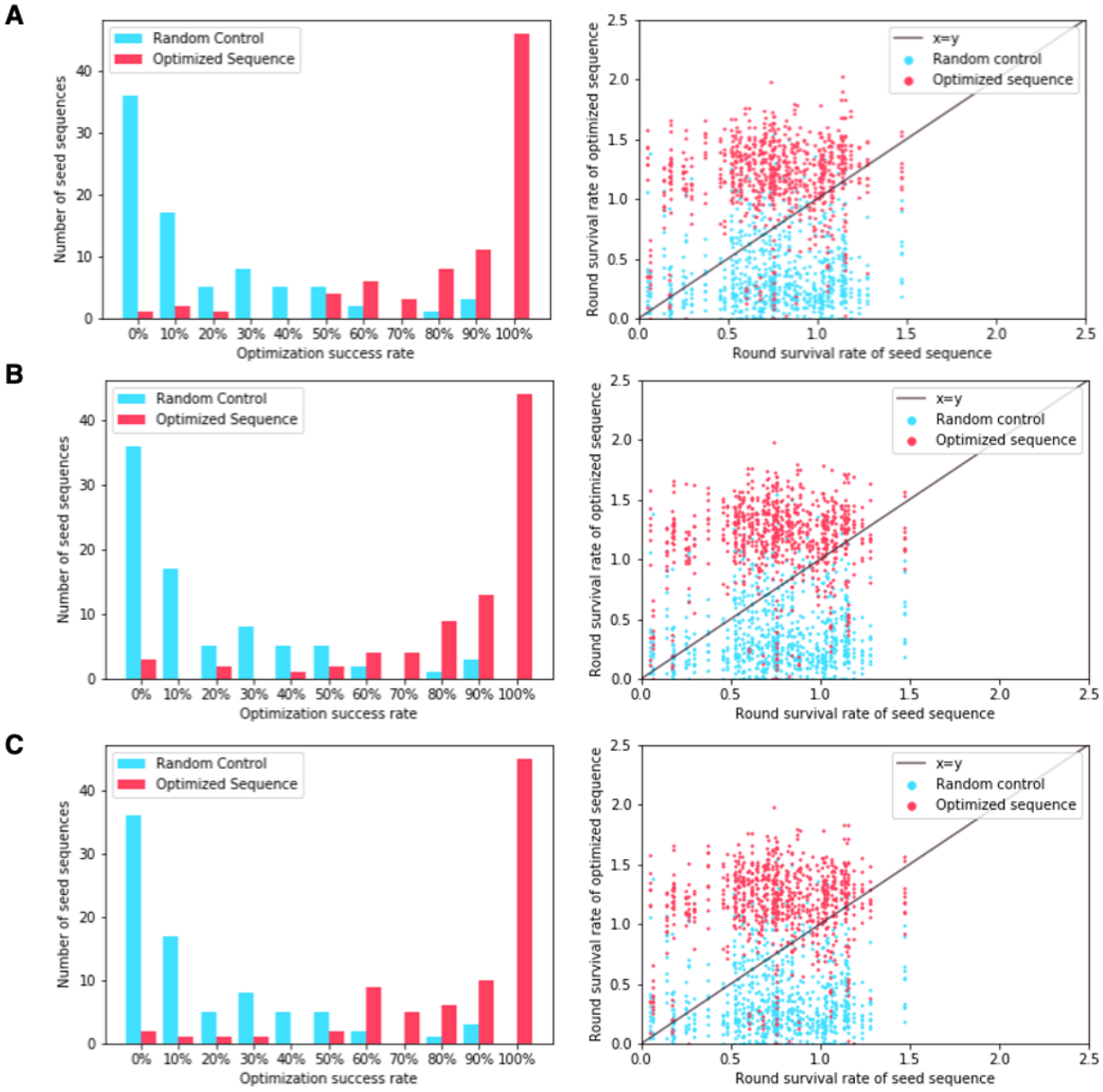
The material covered in this chapter has been published as Dai*, Huisman*, et al (2021). "Machine learning optimization of peptides for presentation by class II MHCs". Bioinformatics (Dai et al., 2021). Excerpts from the publication were reproduced or adapted here under the Creative

Commons license. Zheng Dai generated the RSR metric and RSR values and performed analysis of optimization performance. Haoyang Zeng trained the PUFFIN model and generated optimized candidates. Brandon Carter and Siddhartha Jain provided insightful discussion and prelimary explorations of the analysis. David K. Gifford and Michael E. Birnbaum contributed to the conceptualization of this project, supervision of work, and final manuscript. As such, my paper co-authors, especially Zheng Dai, have contributed significantly to the content, text, and figures of this chapter. This work was supported in part by a Schmidt Futures grant to D.K.G. and M.E.B., a Melanoma Research Alliance Grant to M.E.B., and the National Institutes of Health (R01CA218094 to D.K.G. and P30CA14051 to M.E.B.).

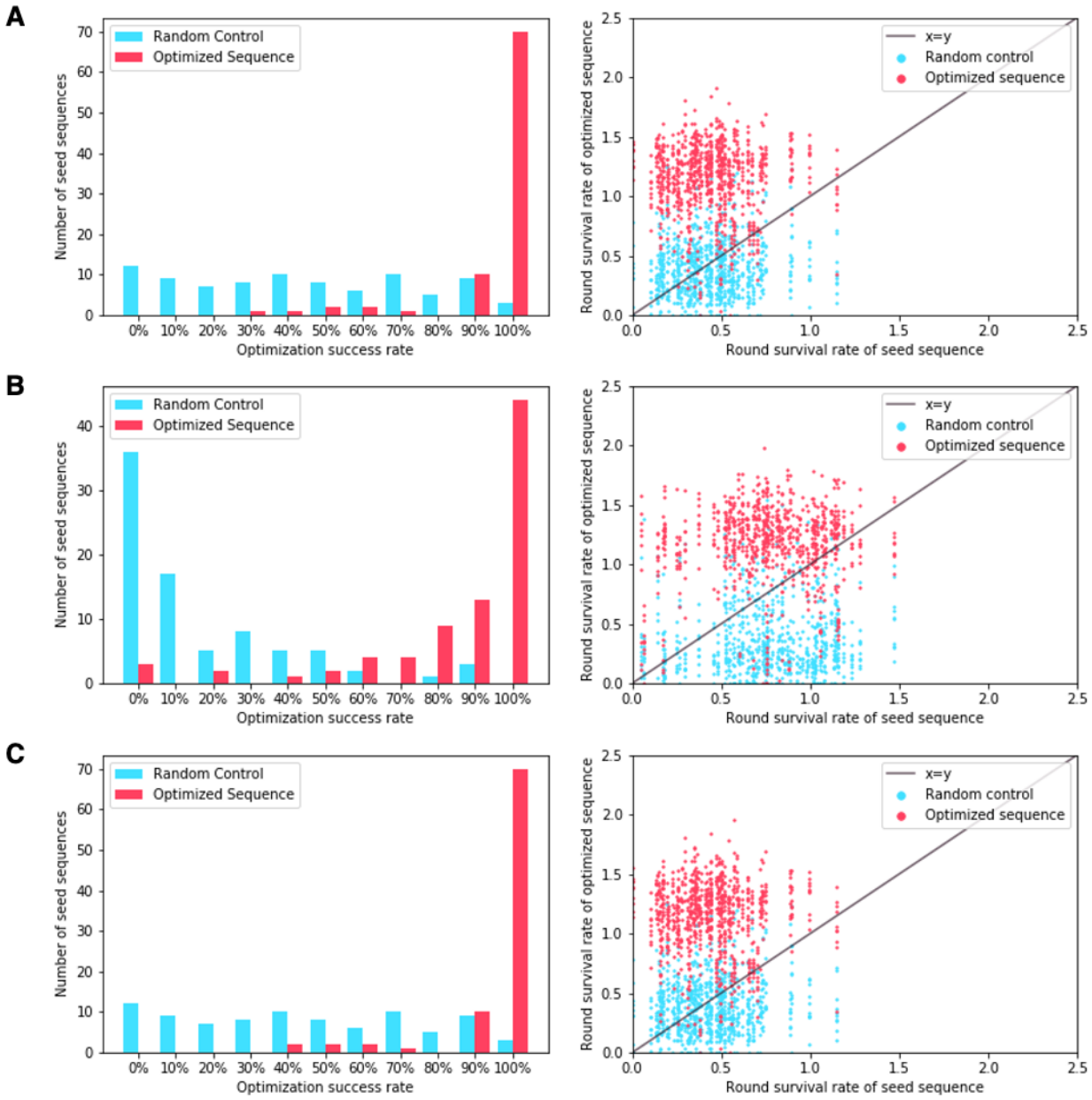
3.12 Supplemental Figures



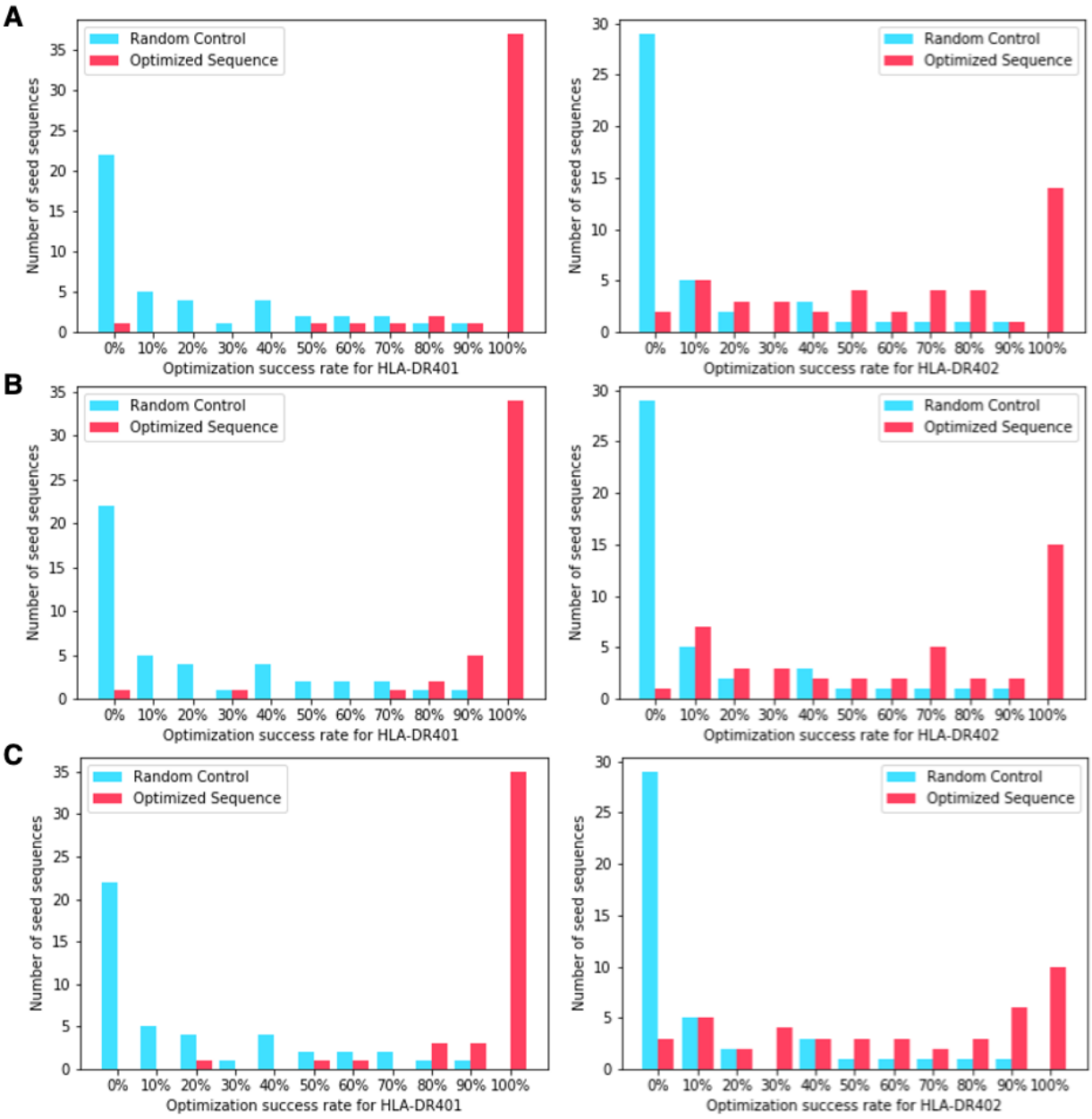
Supplemental Figure 1. Validation through replicates. We validated the RSR values from the validation round by performing a second replicate. **A)** We plot the RSR values between the original and replicate rounds for HLA-DR401. If a point is not present in one of the experiments, it is given a value of 0 and marked in red. The line of best fit is obtained from linear regression for points that were present in both experiments, and is shown alongside the Pearson correlation coefficient r and the Spearman correlation coefficient r_s. **B)** We plot the RSR values between the original and replicate rounds for HLA-DR402.



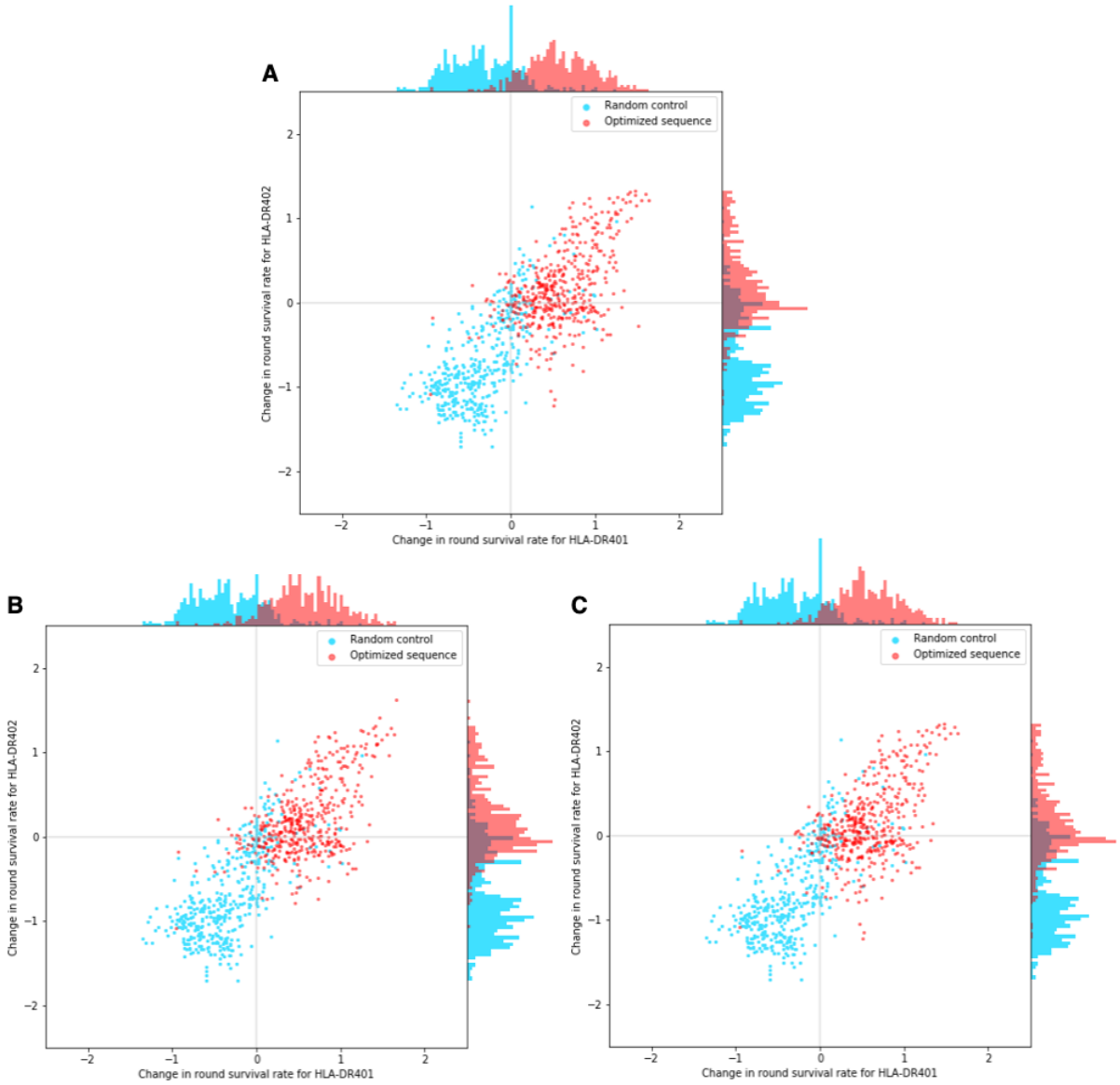
Supplemental Figure 2. Number of sequences that exhibit improvement for other optimization methods on HLA-DR401. These depict the same plots as **Figure 3**, but for different optimization schemes for HLA-DR401. **A)** PE under the categorical model. **B)** UCB under the Gaussian model. **C)** UCB under the categorical model.



Supplemental Figure 3. Number of sequences that exhibit improvement for other optimization methods on HLA-DR402. These depict the same plots as **Figure 3**, but for different optimization schemes for HLA-DR402. **A)** PE under the categorical model. **B)** UCB under the Gaussian model. **C)** UCB under the categorical model.



Supplemental Figure 4. Multiple allele optimization improvements for other optimization methods. The same plots as **Figure 4A**, **4B**, but for different optimization schemes. **A)** PE under the categorical model. **B)** UCB under the Gaussian model. **C)** UCB under the categorical model.



Supplemental Figure 5. RSR changes in multiple allele optimization for other optimization methods. The same plots as **Figure 4D**, but for different optimization schemes. **A)** PE under the categorical model. **B)** UCB under the Gaussian model. **C)** UCB under the categorical model.

CHAPTER 4: A HIGH-THROUGHPUT YEAST DISPLAY APPROACH TO PROFILE PATHOGEN PROTEOMES FOR MHC-II BINDING

The material covered in this chapter has been submitted for review and made available on bioRxiv as: Huisman, Dai, Gifford, Birnbaum (2022). "A high-throughput yeast display approach to profile pathogen proteomes for MHC-II binding". A full acknowledgements statement is included as ***Chapter 4.10****.*

Abstract

T cells play a critical role in the adaptive immune response, recognizing peptide antigens presented on the cell surface by Major Histocompatibility Complex (MHC) proteins. While assessing peptides for MHC binding is an important component of probing these interactions, traditional assays for testing peptides of interest for MHC binding are limited in throughput. Here we present a yeast display-based platform for assessing the binding of tens of thousands of user-defined peptides in a high throughput manner. We apply this approach to assess a tiled library covering the SARS-CoV-2 proteome and four dengue virus serotypes for binding to human class II MHCs, including HLA-DR401, -DR402, and -DR404. This approach identifies binders missed by computational prediction, highlighting the potential for systemic computational errors given even state-of-the-art training data, and underlines design considerations for epitope identification experiments. This platform serves as a framework for examining relationships between viral conservation and MHC binding, and can be used to identify potentially high-interest peptide binders from viral proteins. These results demonstrate the utility of this approach for determining high-confidence peptide-MHC binding.

4.1 Introduction

Major histocompatibility complex (MHC) proteins play a critical role in adaptive immunity by presenting peptide fragments on the surface of cells. Peptide-MHCs (pMHCs) are then surveilled by T cells via their T cell receptors (TCRs), enabling immune cells to sense dysfunction, such as the presence of pathogen-derived peptides (Chaplin, 2010; Hennecke and Wiley, 2001). Class II MHC molecules (MHC-II) are expressed primarily on professional antigen presenting cells, and are recognized by antigen-specific $CD4^+$ T cells that drive the coordination of innate and adaptive immune responses (Chaplin, 2010; Swain et al., 2012). MHC-II molecules have an open peptide-binding groove, allowing for display of long peptides, consisting of a 9 amino acid 'core' flanked by a variable number of additional residues on each side (Jones et al., 2006).

Generating reliable and rapid data on peptide-MHC binding is beneficial for understanding the underlying biology of adaptive immunity and for clinical applications, including for optimized T cell epitopes in vaccine design (Dai et al., 2021; Keskin et al., 2019; Liu et al., 2020, 2021a; Moise et al., 2015; Ott et al., 2017; Patronov and Doytchinova, 2013; Rosati et al., 2021). In fact, therapeutics to generate antigen-specific T cell responses have shown great promise in cancer (Keskin et al., 2019; Ott et al., 2017) and infectious disease (Gambino et al., 2021). Since understanding peptide-MHC binding is critical for identifying and engineering T cell epitopes, there have been sustained efforts to produce high-quality experimental data and predictive algorithms.

Initial experimental methods for determining peptide binding to MHC relied upon the analysis of synthesized candidate peptides via MHC stability or functional assays, and can produce high-confidence data, but can be difficult to scale beyond a small number of candidate peptides (Altmann and Boyton, 2020; Justesen et al., 2009; Mateus et al., 2020; Sidney et al., 2010; Yin and Stern, 2014). More recently, mass spectrometry-based approaches have been demonstrated for determining the MHC-presented peptide repertoire of cells. These approaches include monoallelic mass spectrometry, which allows for the unambiguous assignment of presented peptides to a given MHC allele. However, mass spectrometry-based approaches are not necessarily quantitative measures of presented peptide affinity or abundance, although there have been advances in quantitation using internal standards (Stopfer et al., 2020, 2021). Additionally, the peptides endogenously expressed by a cell can crowd out exogenously examined peptides of interest, and mass spectrometry approaches typically require large numbers of input cells (Abelin et al., 2017, 2019; Parker et al., 2021; Purcell et al., 2019).

A wave of higher throughput approaches have been recently developed for studying peptide-MHC interactions, including yeast display (Jiang and Boder, 2010; Liu et al., 2021b; Rappazzo et al., 2020) and mammalian display-based methods (Obermair et al., 2021). Several of these approaches circumvent the bottlenecks of synthesizing or identifying peptides by utilizing DNA-based inputs and outputs (Jiang and Boder, 2010; Obermair et al., 2021; Rappazzo et al., 2020). These assays rely upon libraries that are often generated via DNA oligonucleotide synthesis, and use peptide stabilization and surface expression (Jiang and Boder, 2010; Liu et al., 2021b; Obermair et al., 2021) or peptide dissociation (Rappazzo et al., 2020) to assess peptide-MHC binding.

In addition to experimental advances, computational approaches for peptide-MHC binding prediction have advanced markedly over the past decade. These developments are due to algorithmic advances (O'Donnell et al., 2020; Racle et al., 2019; Reynisson et al., 2020; Zeng and Gifford, 2019) and the availability of large, high-quality training data (Abelin et al., 2017, 2019; Rappazzo et al., 2020; Reynisson et al., 2020). However, despite the improvements in predicting peptide binding to MHC in a broad sense, the predictive power for individual peptides often remain imperfect relative to experimental measurements (Rappazzo et al., 2020; Zhao and Sher, 2018).

In this chapter, we present a yeast display approach to directly assess peptide-MHC binding for large collections of defined peptide antigens to screen whole viral proteomes for MHC-II binding in high-throughput. We utilize this approach to screen the full proteome of SARS-CoV-2, a present, global threat to public health, and identify SARS-CoV-2-derived MHC binders missed by computational prediction. We additionally apply this approach to screen proteomes from serotypes 1-4 of dengue viruses, in which antibody dependent enhancement results in more severe disease upon second infection with a different dengue virus serotype (Guzman et al., 2016), and thus represents a potential important application area for T cell-directed therapeutics. Our approach enables exploration of peptide binding to MHCs in the context of serotype-specific mutations, identifying homologous, pan-serotype regions of interest that are capable of MHC binding and thus may represent desirable targets for immune interventions.

4.2 Generation of yeast display libraries for profiling the SARS-CoV-2 proteome

Previous studies have reported the use of yeast-displayed MHC-II for characterizing peptide-MHC and pMHC-TCR interactions (Birnbaum et al., 2014, 2017; Rappazzo et al., 2020). We adapted

MHC-II yeast display constructs (Rappazzo et al., 2020) to generate a defined library of peptides that cover the SARS-CoV-2 proteome to assess them for MHC binding. To compare SARS-CoV-2 with a related coronavirus, we also included peptides from the spike and nucleocapsid proteins from SARS-CoV.

Each protein was windowed into peptides of 15 amino acids in length, with a step size of 1 to cover every possible 15mer peptide in the protein (**Figure 1A**). Each peptide was encoded in DNA and cloned in a pooled format into yeast vectors containing MHC-II proteins. The generated library was linked to three MHC-II alleles: HLA-DR401 (HLA-DRA1*01:01, HLA-DRB1*04:01), HLA-DR402 (HLA-DRA1*01:01, HLA-DRB1*04:02), and HLA-DR404 (HLA-DRA1*01:01, HLA-DRB1*04:04). Yeast were formatted with a flexible linker connecting the peptide and MHC, containing a 3C protease site and a Myc epitope tag, which can be used for selections (**Figure 1A**) (Rappazzo et al., 2020). The final library contained 11,040 unique peptides, with 99% of the designed peptides present in each cloned yeast library, as assessed by next-generation sequencing.

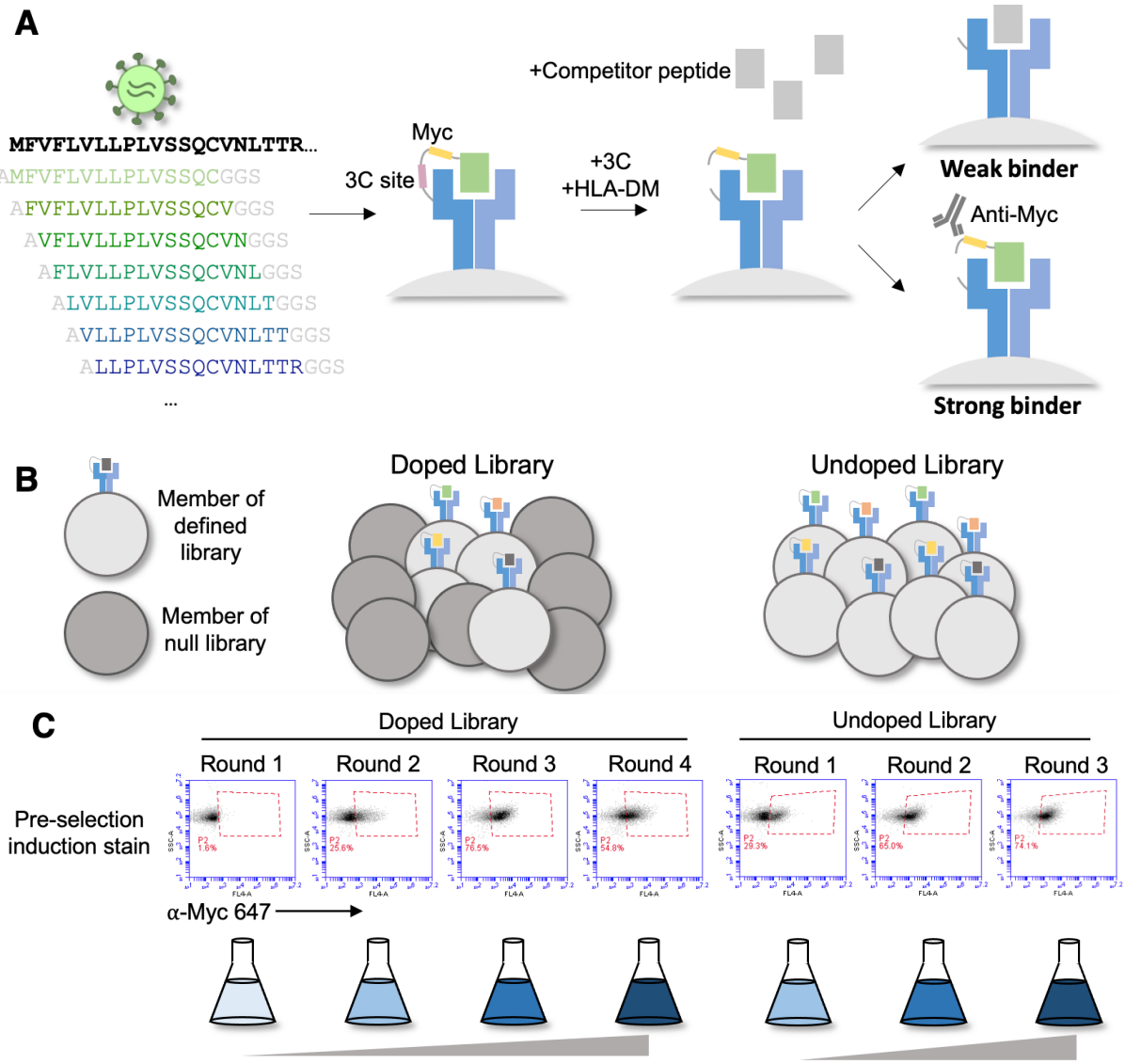


Figure 1. Overview of library and selections. A) The defined library contains pathogen proteome peptides (length 15, sliding window 1). Poor binding peptides are displaced with addition of protease, competitor peptide, and HLA-DM. **B)** Schematic of doped and undoped libraries: in the doped selection strategy, the library is added to a library of null, non-expressing constructs. **C)** Representative flow plots showing enrichment of MHC-expressing yeast over rounds of selection for the library containing SARS-CoV-2 and SARS-CoV peptides on HLA-DR401.

4.3 Strategies for selecting defined libraries

To enrich for peptide binders, iterative selections were performed (**Figure 1A**): the library is first incubated with competitor peptide and 3C protease, which cleaves the covalent linkage between peptide and MHC, followed by the addition of HLA-DM at lower pH. These conditions allow for the encoded peptide to be displaced from the peptide-binding groove. The Myc epitope tag is proximal to the peptide, which can be identified via incubation with an anti-epitope tag antibody followed by enrichment via magnetic bead selection if the yeast-expressed peptide remains bound to the MHC after the peptide exchange reaction.

Three rounds of selection were iteratively performed. Representative enrichment of yeast expressing Myc-tagged peptides can be seen in **Figure 1C** ("undoped library"), for the library displayed by HLA-DR401. Here the pre-selection Myc-positive population starts at 29.3% and quickly converges, with 65.0% positive in the pre-selection Round 2 population and 74.1% in the pre-selection Round 3 population.

Given the rapid convergence of the library, we performed a second set of selections in which we doped the defined library into a randomized, null library to enable a greater degree of enrichment as compared to non-binding peptides. The null library was generated by fully randomizing ten amino acids in the peptide region of the peptide-MHC-II construct while fixing three amino acids to encode stop codons. This library provides a baseline population of yeast which should not express pMHC, and therefore not enrich in our selections. We doped our defined peptide library into a 500-fold excess of null library, such that each peptide member was represented at approximately the same frequency (**Figure 1B**). The null library provides baseline competition, which true binders must enrich beyond, and increases the stringency of the enrichment task.

We performed four rounds of selection on the doped library. Because of the excess of null yeast, the initial pre-selection stain is low (1.6%) compared to the initial undoped library (**Figure 1C**). This staining enriched over the first three rounds of selection, reflective of the stringency of the task and clarity of enrichment. This is in contrast to the initial undoped library, which began with a much higher pre-selection stain, with a lower fold-change in staining over rounds of selection. The low frequency of each member in the starting doped library, however, increases the likelihood of stochastic dropout for any given member.

4.4 Analysis of selection data

After selections, peptide identities were determined through deep sequencing of enriched yeast populations, providing us with a dataset comprised of positive enrichment over four rounds of selection from the doped library and both positive and negative enrichment for three rounds of selection from the undoped library. **Supplemental Figure 1** shows the correlation between defined library members on HLA-DR401 across rounds. As expected, the unselected library correlated poorly with post-selection rounds. Consistent with the observed staining (**Figure 1C**), the doped library essentially converged after Round 3. Similarly, the undoped library appears converged following Round 2.

Next, we established metrics for enrichment for each mode of selection. Given the high starting frequency of members in the undoped library, we classify enrichment based on fold change between Round 1 and Round 2, and we define criteria for enriched yeast in the undoped library as

making up a higher fraction of reads following Round 2 compared to Round 1. In contrast, in the doped library, members start at low frequencies, and we define enrichment based on presence above a threshold in Round 3 of selection, specifically as having greater than or equal to 10 reads following Round 3. **Figure 2B** illustrates the correspondence between enrichment metrics in the doped and undoped library for the library on HLA-DR401. Of the 11,040 peptides in the library, 2,467 enriched in both the doped and undoped libraries displayed by HLA-DR401 (**Figure 2A**). An additional 1,252 enriched in the doped library only and 797 enriched in the undoped library only.

Because the library is designed with a step size of one, we next utilized overlap between adjacent peptides to determine high-confidence binders. This analysis allows us to address the potential that peptide sequences could register shift in such a way that invariant portions of the linker sequences could inadvertently be incorporated into the peptide-binding groove. To do this, we develop and implement a smoothing method, examining overlapping peptides for shared enrichment behavior. Classically, the strongest determinant of peptide affinity for an MHC is the nine amino acid stretch sitting within the peptide-binding groove (Jones et al., 2006; Stern, 1994), although proximal peptide flanking residues can also affect binding (Lovitch et al., 2006; O'Brien et al., 2008; Zavala-Ruiz et al., 2004). In our libraries, a given 9mer is present in seven overlapping 15mer peptides, and we calculate how many of these seven 15mers have enriched. This calculation is shown schematically in **Figure 3A** with toy sequences and applied to enrichment data for SARS-CoV-2 nucleocapsid on HLA-DR401 in **Figure 3B**. Sequences with good 9mer cores should enrich along with neighboring sequences with the same 9mer sequence. In contrast, sequences which enrich spuriously or due to linker sequence in the peptide groove or other stochastic factors should have few neighbor sequences also enriching. Thus, we define a cutoff for high confidence 9mer enrichment of five out of seven 9mer-containing sequences enriching. This cutoff tolerates some stochastic dropout, while still disallowing any cores that may solely enrich by register shifting the Gly-Ser linker residues into the Position 9 pocket, which are favorable for each MHC allele in our study. (Abelin et al., 2019; Rappazzo et al., 2020; Reynisson et al., 2020). Of the 2,467 peptides which enriched in both the doped and undoped libraries for HLA-DR401, 1,791 also contain a 9mer sequence which enriched in five or more peptides of the seven neighboring sequences containing it (**Figure 2A**), with 676 peptides enriching in both doped and undoped libraries but not containing a 9mer core enriched in five or more peptides, and 788 15mers containing a 9mer which enriched in five or more peptides but enriched in zero or one of the doped and undoped libraries. These full relationships are captured in Venn diagrams in **Supplemental Figure 2** for all three MHC alleles studied here.

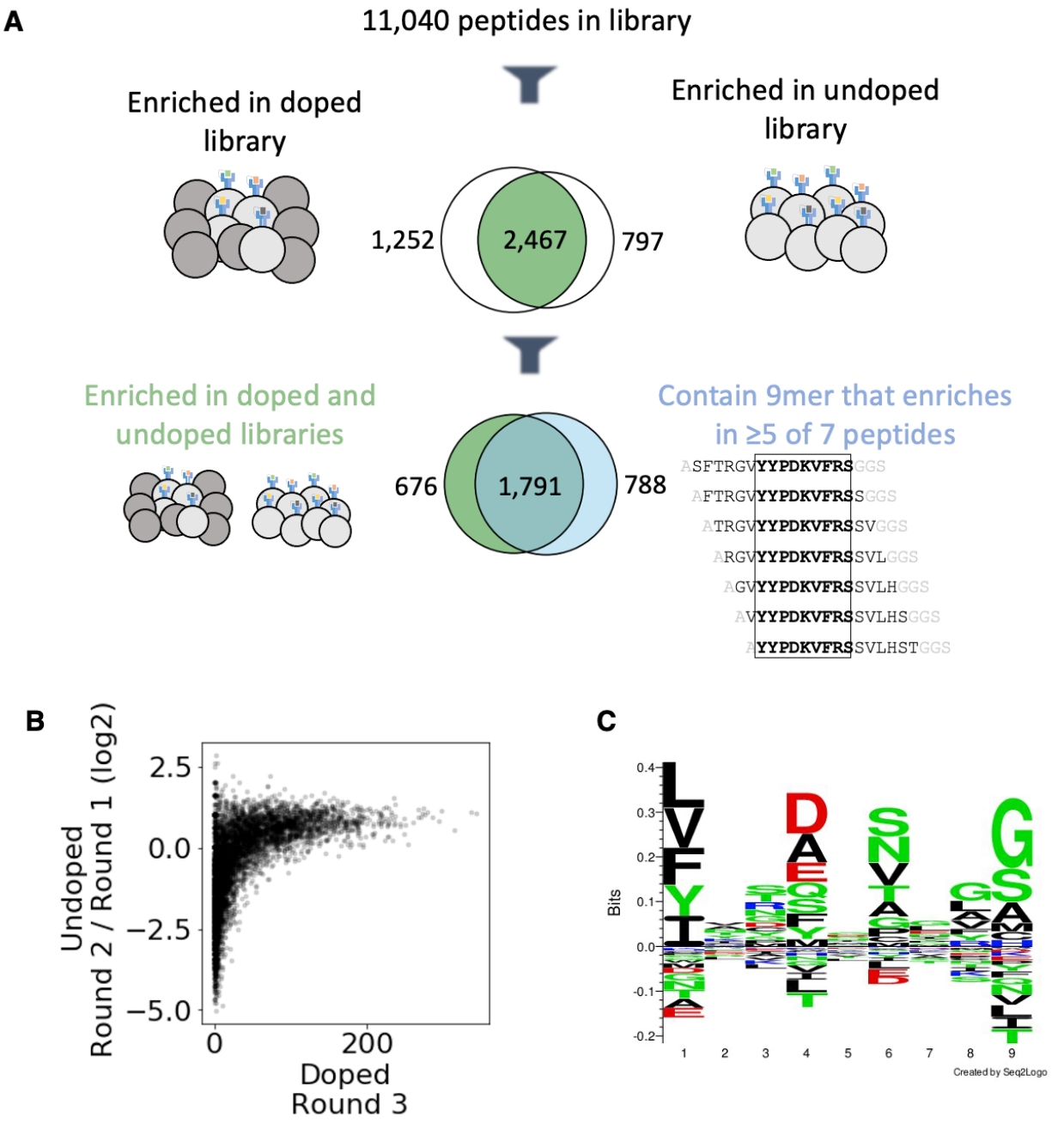


Figure 2. Output of selections and analysis of selection data. A) Overview of filtering peptides and correspondence between selection strategies for SARS-CoV and SARS-CoV-2 library on HLA-DR401. Peptides are filtered for enrichment in both doped and undoped libraries. Further, the relationship between these peptides and peptides which contain a 9mer that is enriched in five or more of the seven peptides containing it is shown. **B)** Relationships between enrichment in doped and undoped libraries. Absolute counts following Round 3 of selection of the doped library are plotted against the log2 fold change between read fraction for peptides in Round 2 and Round 1. Data are shown for the library on HLA-DR401. **C)** Sequence logo of 2,467 peptides that enriched in both doped and undoped selected libraries for HLA-DR401. Registers are inferred with a position weight matrix-based alignment method. Logos were generated with Seq2Logo-2.0.

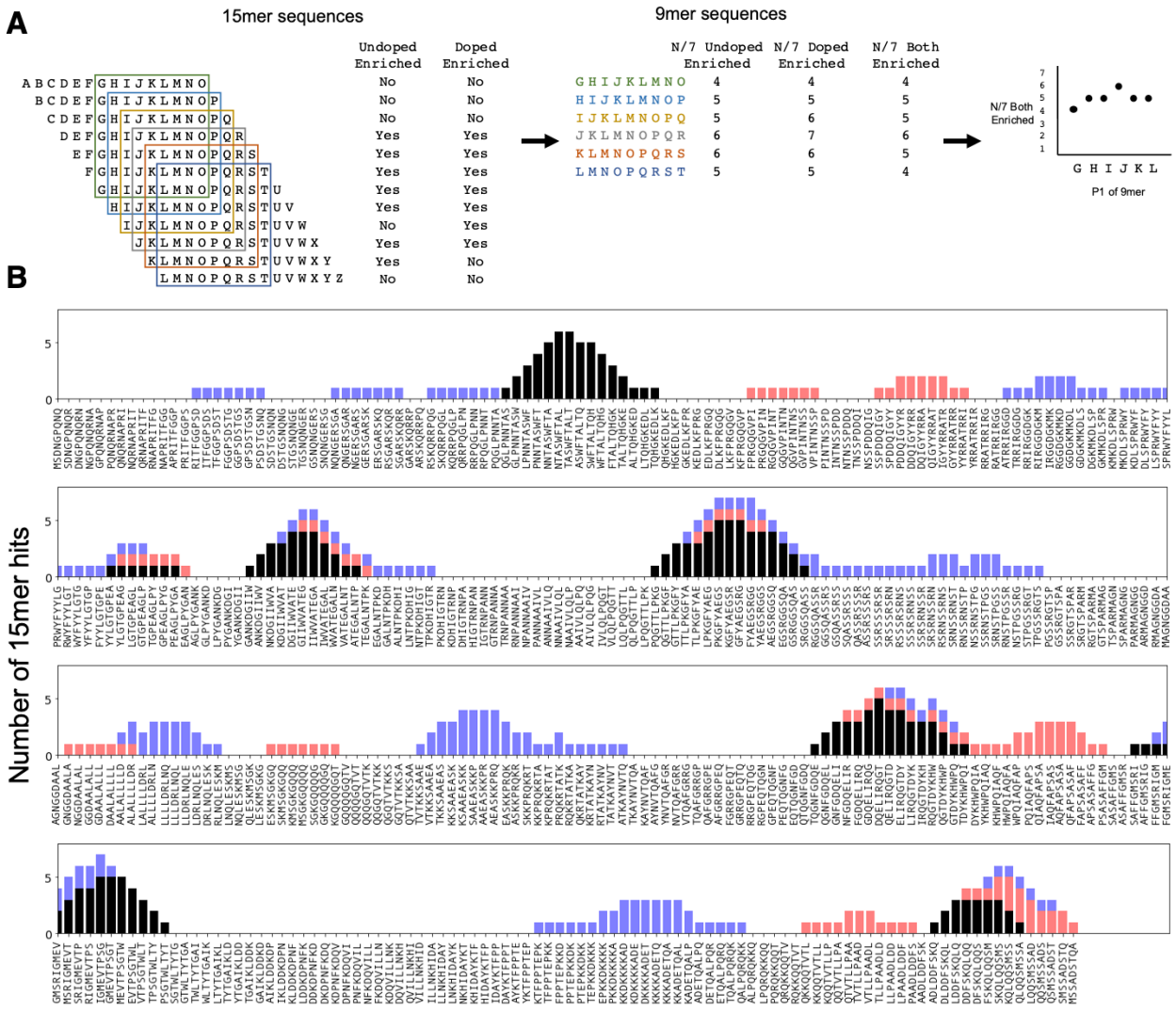


Figure 3. Utilizing overlapping peptides to call high confidence peptides. A) Schematic showing use of overlapping 15mers, containing redundant 9mers. Each 9mer is present in seven 15mers; for each 9mer, we calculate how many of these seven 15mers enriched. **B)** Number of peptides containing a given 9mer that are hits, for SARS-CoV-2 nucleocapsid on HLA-DR401. Black=hits in both undoped and doped libraries; blue=hits in undoped library only; red=hits in doped library only. Enrichment categories are stacked, for a maximum of seven 15mer hits, since each 9mer is present in seven 15mers.

4.5 Sequence motifs of enriched peptides and considerations for epitope identification experiments

To examine the 9mer core motifs of enriched peptides, we utilized a position weight matrix method to infer the peptide register and generated visualizations of the 9mer cores using Seq2Logo (Thomsen and Nielsen, 2012). **Figure 2C** shows a sequence logo of the aligned 9mer cores from the 2,467 15mer peptides which enriched on HLA-DR401 in both doped and undoped libraries. The peptide motif is consistent with previously reported motifs for HLA-DR401 (Abelin et al., 2019; Rappazzo et al., 2020): hydrophobic amino acids are preferred at P1, acidic residues at P4, polar residues at P6, and small residues at P9. We also observe some preference for glycine at P8 in the sequence logo, which is potentially an artifact of non-native registers with linker at P8 and P9.

The other alleles used in the study, HLA-DR402 and HLA-DR404, have polymorphisms in their peptide binding groove sequences as compared to HLA-DR401, which affect binding preferences. HLA-DR401 differs from HLA-DR402 at four amino acids and from HLA-DR404 at two amino acids, with all polymorphisms located in the beta chain. HLA-DR402 and HLA-DR404 share an amino acid distinct from HLA-DR401 affecting the P1 pocket (Gly86Val), resulting in a preference for smaller hydrophobic residues (**Figure 5A**). Three polymorphisms in HLA-DR402 affect P4, P5, and P7 compared to HLA-DR401 (Leu67Ile, Gln70Asp, and Lys71Glu), while HLA-DR404 has only one (Lys71Arg). Sequence logos for HLA-DR402 and HLA-DR404 are consistent with previously reported motifs and MHC polymorphisms (**Figure 4**). For HLA-DR402, we observe less P4 preference compared to the motif of HLA-DR402 binders enriched from a randomized yeast display peptide library (Rappazzo et al., 2020), albeit consistent with mass spectrometry-generated motifs which also showed minimal P4 preference for HLA-DR402 (Abelin et al., 2019).

To explore differences between mass spectrometry, defined libraries, and random libraries, and to probe the differing strengths of P4 peptide preference observed for HLA-DR402 between these modalities, we examined the compositions of randomized and defined libraries. We hypothesized that skewed amino acid abundances in nature, which are reflected in the defined library, could result in an apparent diminished amino acid preference. Indeed, three of the most preferred P4 residues for binding HLA-DR402, Trp, His, and Met (Rappazzo et al., 2020), are all low abundance in the SARS-CoV-2 proteome (Trp 1.1%, His 1.9%, Met 2.2%). In comparison, a randomized peptide library for HLA-DR402 (Rappazzo et al., 2020) had a higher representation of these amino acids (Trp 3.8%, His 2.9%, Met 3.8%). Additionally, the randomized library had approximately nine thousand-fold more members than the defined library, providing more instances of all amino acids. The low abundance and underrepresentation of these amino acids likely underlies the apparent lack of amino acid consensus at P4 in enriched peptides. Interestingly, Arg and Lys, which have also been reported as preferred HLA-DR402 P4 residues, are more abundant than Trp, His, and Met in the SARS-CoV-2 proteome (Arg 3.4% and Lys 5.9%; compare to Arg 9.7%, Lys 4.0% in the random library), but still show less representation at P4 in the defined library enriched peptides compared to the random library-enriched peptides. These differences in motifs between randomized and defined libraries highlight the utility of randomized libraries for downstream applications such as training prediction algorithms. Approaches influenced by amino acid abundance in nature, such as defined libraries and mass spectrometry approaches, could

inadvertently bias against possible binders because of absence of amino acids in their null distribution, rather than true binding preference.

Next, we wanted to examine the distribution of peptides among the possible 9mer registers along each 15 amino acid sequence. Based on our register inference, of the 2,467 enriched peptides from the HLA-DR401 library, 1,610 peptides bound native 9mer cores without using any linker sequence residues in the 9mer core, which is consistent with theoretical ratios of possible native and non-native cores for a given 9mer. The peptides with predicted native 9mer cores were approximately equally distributed between possible registers, with the exception of the N-terminal register, which had one-third fewer peptides. This register had only a single N-terminal flanking residue (a fixed Ala), which is likely disfavored.

Because the library was designed with step size of one, many of the 9mer cores will be repeated among neighboring peptides. Of the 1,610 HLA-DR401 peptides which enriched using a native 9mer core, there are 563 unique 9mer cores identified through register-inference. **Table 1** summarizes enrichment for each protein included in the library, highlighting the number of 15mers which enriched in both the doped and undoped libraries, the number of unique native 9mer cores, and the number of 15mers containing a 9mer enriched in at least five of seven overlapping peptides.

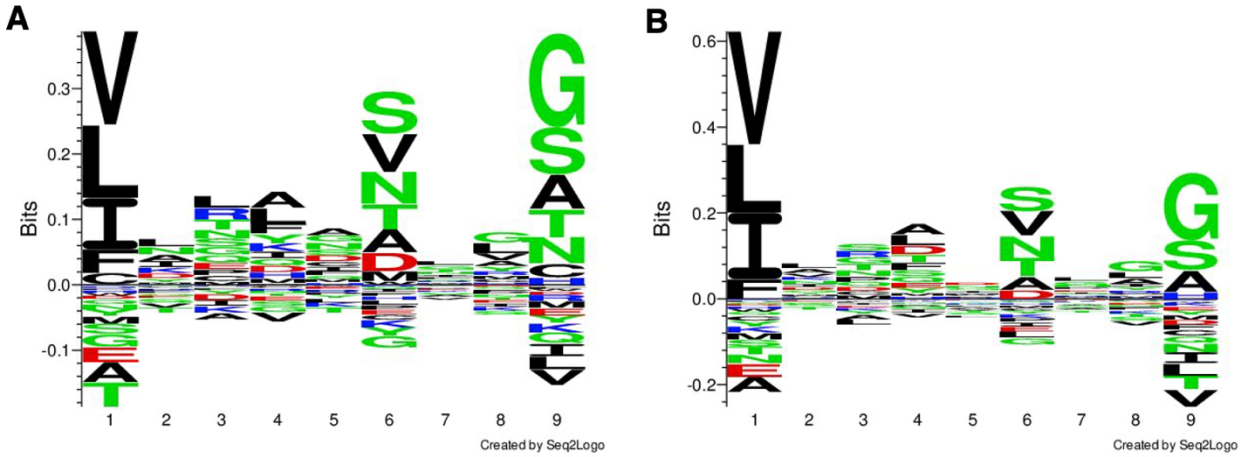


Figure 4. Sequence logo for HLA-DR402 and HLA-DR404. A) HLA-DR402: Sequence logo of 1,690 peptides that enriched in both doped and undoped selections of the SARS-CoV and SARS-CoV-2 library for HLA-DR402. **B)** HLA-DR404: Sequence logo of 2,094 peptides that enriched in both doped and undoped selections of the SARS-CoV and SARS-CoV-2 library for HLA-DR404. Logos were generated with Seq2Logo-2.0.

Virus	Protein	Protein length (# of amino acids)	MHC Allele	# of 15mers	# of 9mer cores	# of smoothed 15mers
SARS-CoV	Spike	1255	HLA-DR401	324	74	221
			HLA-DR402	217	65	110
			HLA-DR404	289	61	193
SARS-CoV	Nucleocapsid	422	HLA-DR401	40	8	34
			HLA-DR402	34	13	12
			HLA-DR404	31	6	20
SARS-CoV-2	Spike	1273	HLA-DR401	305	67	221
			HLA-DR402	230	62	130
			HLA-DR404	290	64	217
SARS-CoV-2	Nucleocapsid	419	HLA-DR401	34	8	24
			HLA-DR402	33	10	15
			HLA-DR404	30	8	18
SARS-CoV-2	Replicase polyprotein 1ab	7096	HLA-DR401	1652	388	1204
			HLA-DR402	1104	325	678
			HLA-DR404	1368	350	890
SARS-CoV-2	Non-structural protein 8	121	HLA-DR401	41	10	32
			HLA-DR402	21	7	17
			HLA-DR404	32	8	19
SARS-CoV-2	Protein 7a	121	HLA-DR401	27	8	18
			HLA-DR402	7	3	0
			HLA-DR404	13	2	6
SARS-CoV-2	Non-structural protein 6	61	HLA-DR401	0	0	0
			HLA-DR402	1	1	0
			HLA-DR404	0	0	0
SARS-CoV-2	Membrane protein	222	HLA-DR401	40	7	29
			HLA-DR402	26	6	19
			HLA-DR404	23	7	21
SARS-CoV-2	Envelope small membrane protein	75	HLA-DR401	6	1	0
			HLA-DR402	7	3	0
			HLA-DR404	6	1	0
SARS-CoV-2	Protein 3a	275	HLA-DR401	22	4	11
			HLA-DR402	13	4	10
			HLA-DR404	10	2	0
SARS-CoV-2	Replicase polyprotein 1a	4405	HLA-DR401	948	228	658
			HLA-DR402	657	196	409
			HLA-DR404	865	222	582
SARS-CoV-2	ORF10 protein	38	HLA-DR401	6	1	6
			HLA-DR402	2	0	0
			HLA-DR404	5	1	5
SARS-CoV-2	Protein non-structural 7b	43	HLA-DR401	0	0	0
			HLA-DR402	0	0	0
			HLA-DR404	0	0	0
SARS-CoV-2	Uncharacterized protein 14	73	HLA-DR401	8	4	6
			HLA-DR402	20	5	16
			HLA-DR404	22	4	21
SARS-CoV-2	Protein 9b	97	HLA-DR401	29	7	27
			HLA-DR402	35	6	31
			HLA-DR404	37	9	34

Table 1. Summary of enriched peptides for each source protein. This includes: the number of unique 15mers which each enriched in both of the doped and undoped libraries; the number of unique 9mer cores identified by register-inference in these enriched 15mers (native cores only, so linker-containing inferred cores excluded); and the number of unique enriched 15mers that contain 9mer sequences enriched in five or more of overlapping neighbors.

4.6 Examining relationships between MHC-specific binding and spike proteins from SARS-CoV-2 and SARS-CoV

To further explore relationships between the MHCs studied here and their virally-derived peptide repertoires, we compared the binding of SARS-CoV-2 and SARS-CoV spike proteins to all three MHC alleles. Sequence alignment of these three MHC alleles is shown in **Figure 5A**, with polymorphic regions highlighted on an HLA-DR401 structure (adapted from PDB 1J8H). Interplay between viral conservation and binding are illustrated in **Figure 5B**, highlighting conserved regions of the proteome in black and binders to each allele in grey, red, and blue. Regions are highlighted where sequences enrich in overlapping peptides; that is, for each 9 amino acid stretch along the proteome, we calculated how many of the seven 15mer peptides enrich in the yeast display assay, and if a 9mer enriched five or more times, it is marked as a hit. Specific examples of these relationships are probed in **Figure 5C, D, and E**, where individually enriched 15mer sequences are represented as horizontal lines above 15mer stretches in the proteome. Bolded 9mers are identified through register inference as consensus binding cores for these peptides. Only 15mers which contain the bolded 9mer are included in this representation. Non-conserved amino acids within this 9mer are highlighted in yellow.

Figure 5C illustrates a region that is not conserved between SARS-CoV-2 and SARS-CoV, where the SARS-CoV-2 peptides containing the core IYQAGSTPC are enriched for binding to all three MHCs, but mutations, including at both P1 and P4 to Proline, discourage binding of the aligned SARS-CoV peptide. **Figure 5E** illustrates a core that is conserved between SARS-CoV and SARS-CoV-2, which can bind only to HLA-DR401, but not to HLA-DR402 or HLA-DR404, likely due to the size of the P1 hydrophobic residue and, for HLA-DR402, the acidic P4 residue. **Figure 5D** illustrates relationships between both viral conservation and MHC preference. In **Figure 5D**, the SARS-CoV peptides containing the core IKNQCVNFN can bind to all three alleles. However, the aligned SARS-CoV-2 peptides containing the core VKNKCVNFN do not bind to HLA-DR401, likely because of the less preferable P1 Valine and basic P4 Lysine, but can bind to HLA-DR402, which prefers these residues. These peptides can bind to HLA-DR404, although only four of the adjacent peptides containing this core enrich, which is below the cutoff of five or more, and since no other adjacent peptides enriched, this would not have been classified as a binder (reflected in **Figure 5B**). This marginal, but below-threshold binding is logical, given that the P4 pocket for HLA-DR404 is similar to HLA-DR401, which does not prefer P4 Lysine, but HLA-DR404 has the same P1 binding pocket as HLA-DR402, which both prefer the P1 Valine in the SARS-CoV-2 peptide.

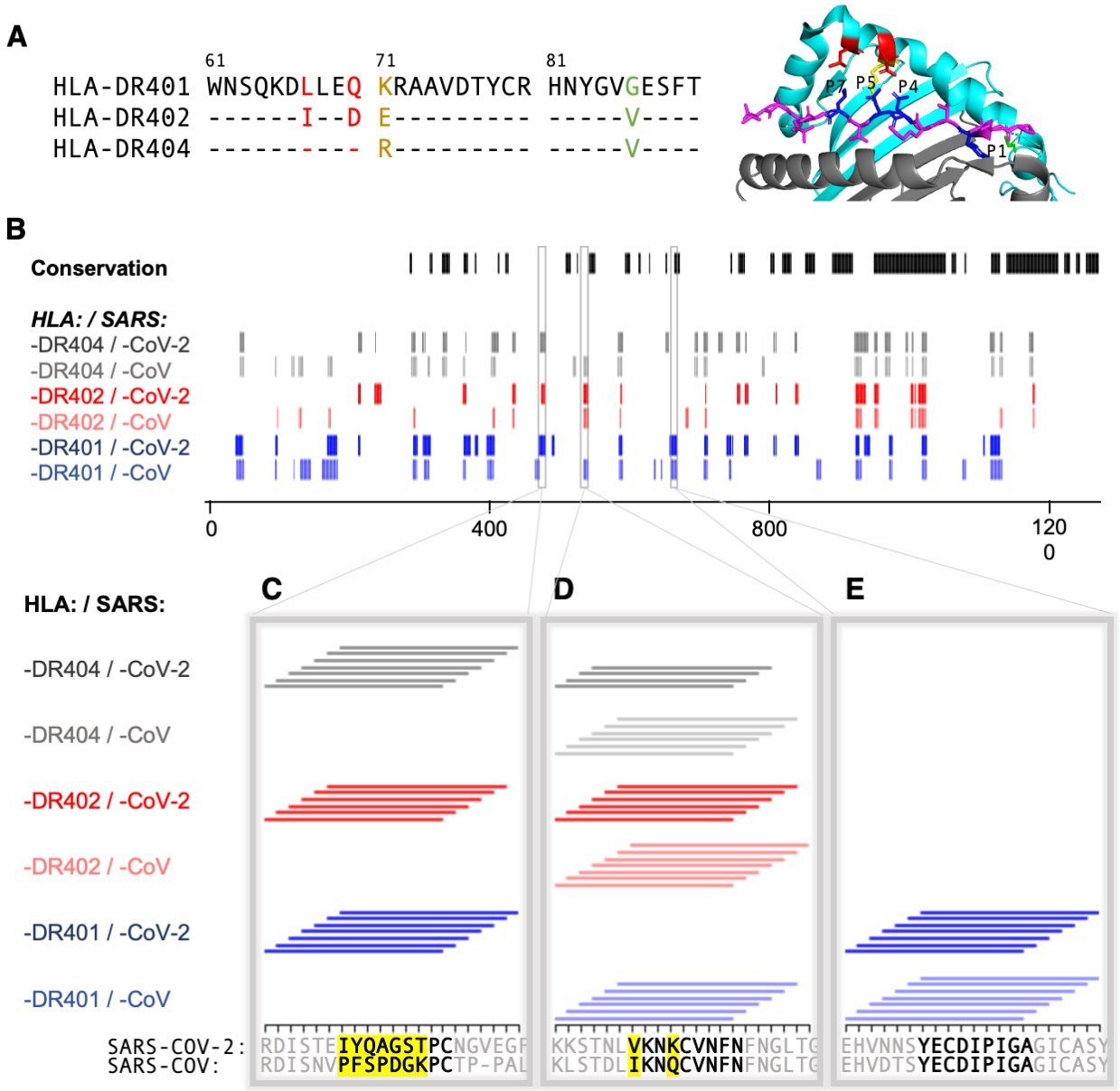


Figure 5. Comparing HLA-DR401, HLA-DR402, and HLA-DR404 for binding to related Spike proteins from SARS-CoV-2 and SARS-CoV. A) Sequence alignment showing sequence differences in HLA-DR402 and HLA-DR404 compared to HLA-DR401 and highlighted on HLA-DR401 structure (PDB 1J8H). Colors are: red for amino acids shared between HLA-DR401 and HLA-DR404, green for amino acids shared between HLA-DR402 and HLA-DR404, and yellow for amino acids different in all 3 alleles. Affected peptide positions (P1, P4, P5, P7) are colored in blue and labeled on the structure. **B)** Conservation and enrichment of 9mer peptides from SARS-CoV-2 and SARS-CoV Spike proteins. Conserved 9mers are indicated in black. If a 9mer along the proteome enriched in 5 or more of the adjacent peptides containing it, its enrichment is indicated with a vertical line with color for allele (HLA-DR401: blue; HLA-DR402: red; HLA-DR404: grey) and opacity for virus (SARS-CoV-2: dark; SARS-CoV: light). **C-E)** Zoomed regions show enrichment of individual 15mer peptides. Only peptides containing the bolded 9mer sequence are shown. Amino acids in the bolded 9mer that are not conserved between SARS-CoV-2 and SARS-CoV are highlighted in yellow.

4.7 Identifying peptide binders missed by computational prediction

Next, we compared our direct experimental assessments with results from computational MHC binding predictions. Prediction algorithms allow for rapid computational screening of potential peptide binders (Abelin et al., 2019; Reynisson et al., 2020), although they can contain systemic biases (Rappazzo et al., 2020). To test the outputs of our direct assessment approach and computational prediction algorithms, we assessed binding of several peptides using a fluorescence polarization competition assay to determine IC_{50} values, as described previously (Rappazzo et al., 2020; Yin and Stern, 2014). Yeast-formatted peptides (Ala+15mer+Gly+Gly+Ser) from SARS-CoV-2 spike protein were run through NetMHCIIpan 4.0 for binding to HLA-DR401, with binders defined as having ≤ 10% Rank (Eluted Ligand mode). Yeast display binders to HLA-DR401 were defined via the stringent criteria of 1) enriching in both in doped and undoped selections, and 2) containing a 9mer that enriched in five or more of the overlapping seven 15mers. 15mers were selected such that they could contain a maximum overlap of 8 amino acids with other selected peptides, to avoid selecting peptides with redundant 9mer cores. An length-matched version of the commonly studied Influenza A $HA_{306-318}$ peptide (APKYVKQNTLKLATG) known to bind HLA-DR401 (Hennecke and Wiley, 2002; Rappazzo et al., 2020) was included as a positive control, along with sequences that yeast display and NetMHCIIpan 4.0 both classified as either binders or non-binders. **Supplemental Figure 3** shows a comparison of yeast-enriched and NetMHCIIpan 4.0 predicted binders, with boxed sequences selected for testing by fluorescence polarization.

The resulting fluorescence polarization IC_{50} data from the native 15mer peptides are shown in **Table 2** and **Supplemental Figure 4**. Peptides which both enriched in yeast display and were predicted by NetMHCIIpan 4.0 to bind ('Agreed Binders') all showed IC_{50} values consistent with binding, each with $IC_{50} < 2.2$ μM. Similarly, peptides which were agreed non-binders showed no affinity for HLA-DR401, with $IC_{50} > 50$ μM.

All 8 'Yeast-Enriched Binders', which enriched in the yeast display assay but were not predicted to bind via NetMHCIIpan 4.0, showed some degree of binding, with IC_{50} values distributed from 14 nM (higher affinity than the HA control peptide) to 18 μM (weak, but measurable, binding). Retrospectively, the weakest two binders appear to be enriching in the yeast display assay using the peptide linker or have a binding core offset from center. Interestingly, NetMHCIIpan 4.0 predictions on the peptides identified via yeast display proved highly sensitive to the length or content of the flanking sequences: if we repeat predictions on only the antigen-derived 15mer sequences without the flanking sequences, NetMHCIIpan 4.0 recovers four of its former false negative peptides (**Table 3**; peptides listed at the top in each section of the table). We will refer to these four peptides as 'flank-sensitive centered peptides', as they each have the consensus 9mer core centered in the peptide.

To further investigate the relationship with flanking residues, we selected five additional peptides ('offset peptides') matching three criteria; these offset peptides were 1) enriched in the yeast display assay, 2) share an overlapping core with the four flank-sensitive centered peptides, but are 3) not predicted by NetMHCIIpan 4.0 to be binders (either with or without invariant flanking sequence added). All five offset peptides have their predicted cores offset by 1-2 amino acids from center, leaving at minimum 1 amino acid on both ends of the 9mer core for each peptide. All five offset peptides exhibit some binding, with IC_{50} values below 13 μM. Each peptide is lower affinity than its overlapping centered counterpart, illustrating effects of flanking residues on peptide

binding, although some over-estimation of these effects in NetMHCIIpan 4.0 predictions are present.

We tested three 'NetMHC-Predicted Binders', which were predicted to bind by NetMHCIIpan 4.0, but were not enriched (nor did any neighboring sequences within an offset of 4 amino acids) in the yeast display assay (**Table 2**). Of these, one bound to HLA-DR401 (IC_{50} 475 nM), while two showed minimal binding with IC50 > 35 μM, which is above the maximum 20 μM concentration tested. All three were predicted by NetMHCIIpan 4.0 to bind with or without the invariant flanking sequences (Eluted ligand mode % Rank: 5.7, 4.1, 8.7 (with flanking residues) and 2.3, 0.6, 7.0 (without flanking residues), for ELDKYFKNHTSPDVD, LQSYGFQPTNGVGYQ, and KTQSLLIVNNATNVV, respectively).

Of the eight 'Yeast-Enriched Binders' in **Table 2**, six contain cysteine residues, which have been shown to be systematically absent from other datasets, including those from mono-allelic mass spectrometry (Abelin et al., 2019; Barra et al., 2018), yet present in yeast display-derived datasets (Rappazzo et al., 2020). To test for non-specific binding due to cysteine, two cysteine-containing 'Agreed Non-Binders' were also tested and showed no affinity for HLA-DR401, suggesting that cysteine itself is not causing non-specific binding. In the fluorescence polarization dataset, the highest affinity binder (14 nM) contained cysteine and was missed by NetMHCIIpan 4.0 predictions (Eluted ligand mode % Rank: 71 (with flanking residues) and 28 (without flanking residues)).

The relationship between measured IC_{50} values and NetMHCIIpan 4.0 predicted values for all 15mer SARS-CoV-2 spike peptides tested is shown in **Figure 6** and **Supplemental Figure 5**.

	Spike Position	Peptide+flank (A+15mer+GGS)	NetMHCIIpan4.0 Predicted Core (A+15mer+GGS)	NetMHCIIpan4.0 %Rank (A+15mer+GGS)	15mer Affinity from FP (IC_{50}, nM)
Agreed Binders	34-48	ARGVYYPDKVFRSSVLGGS	YYPDKVFRS	1.49	15.8
	87-101	ANDGVYFASTEKSNIIGGS	VYFASTEKS	4.28	2117
	303-317	ALKSFTVEKGIYQTSNGGS	FTVEKGIYQ	8.41	396.9
	362-376	AVADYSVLYNSASFSTGGS	YSVLYNSAS	8.36	113.7
	1015-1029	AAAEIRASANLAATKMGGS	IRASANLAA	3.13	105.4
	1112-1126	APQIITTDNTFVSGNCGGS	ITTDNTFVS	7.32	527.0
Yeast-Enriched Binders	165-179	ANCTFEYVSQPFLMDLGGS	YVSQPFLMD	64.83	14,652
	172-186	ASQPFLMDLEGKQGNFGGS	FLMDLEGKQ	20.34	123.2
	286-300	ATDAVDCALDPLSETKGGS	VDCALDPLS	32.68	521.6
	373-387	ASFSTFKCYGVSPTKLGGS	YGVSPTKLG	16.59	18,452
	469-483	ASTEIYQAGSTPCNGVGGS	IYQAGSTPC	18.22	67.7
	580-594	AQTLEILDITPCSFGGGGS	LEILDITPC	62	119.9
	739-753	ATMYICGDSTECSNLLGGS	YICGDSTEC	70.91	14.4
	920-934	AQKLIANQFNSAIGKIGGS	FNSAIGKIG	20.47	1121
NetMHC-Predicted Binders	113-127	AKTQSLLIVNNATNVVGGS	IVNNATNVV	8.74	>50,000
	492-506	ALQSYGFQPTNGVGYQGGS	YGFQPTNGV	4.11	454.7
	1151-1165	AELDKYFKNHTSPDVDGGS	YFKNHTSPD	5.74	35,510
Agreed Non-Binders	534-548	AVKNKCVNFNFNGLTGGGS	FNFNGLTGG	57.13	>50,000
	1079-1093	APAICHDGKAHFPREGGGS	ICHDGKAHF	80.47	>50,000

Table 2. Peptides selected for fluorescence polarization (FP) experiments for binding to HLA-DR401. NetMHCIIpan 4.0 predictions for HLA-DR401 binding were performed on 15mers plus invariant flanking residues (N-terminal Ala, C-terminal Gly-Gly-Ser) and percent rank values generated using Eluted Ligand mode. Fluorescence polarization was performed on native 15mer peptides without invariant flanking residues.

Spike Position	Sequence	NetMHCIIpan4.0 Predicted Core (A+15mer+GGS)	NetMHCIIpan4.0 %Rank (A+15mer+GGS)	NetMHCIIpan4.0 Predicted Core (15mer)	NetMHCIIpan4.0 %Rank (15mer)	15mer Affinity from FP (IC50, nM)
172-186	**SQP**FLMDLEGKQ**GNF**	FLMDLEGKQ	20.34	FLMDLEGKQ	4.1	123.2
173-187	**QP**FLMDLEGKQ**GNFK**	FLMDLEGKQ	27.73	FLMDLEGKQ	12.21	8613
286-300	**TDA**VDCALDPLS**ETK**	VDCALDPLS	32.68	VDCALDPLS	9.8	1154
287-301	**DA**VDCALDPLS**ETKC**	VDCALDPLS	42.42	VDCALDPLS	22.57	4393
469-483	**STE**IYQAGSTPC**NGV**	IYQAGSTPC	18.22	IYQAGSTPC	5.41	67.7
467-481	**DISTE**IYQAGSTPC**N**	IYQAGSTPC	11.47	IYQAGSTPC	12.61	4875
471-485	**E**IYQAGSTPC**NGVEG**	YQAGSTPCN	39.17	YQAGSTPCN	21.81	12519
920-934	**QKL**IANQFNSAI**GKI**	FNSAIGKIG	20.47	IANQFNSAI	7.89	1495
921-935	**KL**IANQFNSAI**GKIQ**	FNSAIGKIQ	18.3	IANQFNSAI	19.79	11937

Table 3. Effects of peptide flanking sequences on NetMHCIIpan 4.0 predictions for HLA-DR401 binding and measured fluorescence polarization (FP) values for overlapping peptides. Yeast display-enriched peptides that are predicted to bind by NetMHCIIpan 4.0 when without flanking residues, plus offset variants of these peptides, which are not predicted to bind, with or without flanking sequence. Yeast display register-inferred consensus cores are highlighted in green. NetMHCIIpan 4.0 percent rank values were generated using Eluted Ligand mode.

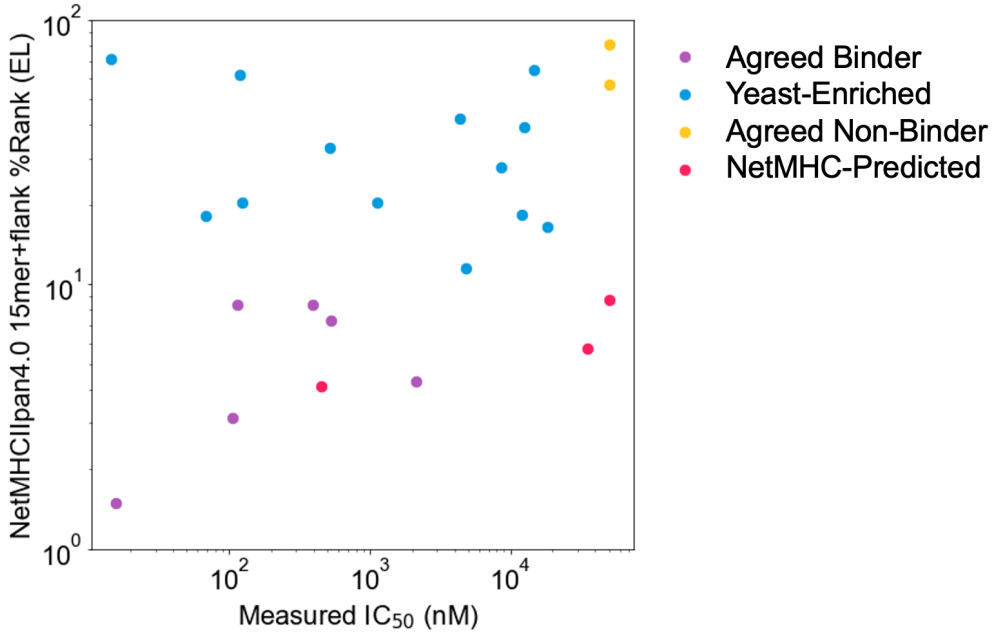


Figure 6. Comparing measured IC_{50} values and computational predictions. Relationship between measured IC_{50} values and NetMHCIIpan 4.0 predicted ranks in Eluted Ligand mode (EL) on invariant-flanked sequences. Data points are colored by label, and IC_{50} values ≥50 μM are set to 50 μM.

4.8 Comparing whole dengue serotype proteomes for common MHC-binding peptides

Defined yeast display libraries can generate data for diverse objectives. Dengue viruses typically cause most severe disease after a second infection with a serotype different from the first infection, due to antibody dependent enhancement (Guzman et al., 2016), which makes T cell-directed therapeutics a potentially attractive means of combatting disease. To profile and compare MHC binding across serotypes, we generated libraries containing 12,672 dengue-derived peptides, covering the entire proteomes of dengue serotypes 1-4. These libraries were on HLA-DR401 and HLA-DR402 and had coverage of 98% and 96% of the dengue library members after construction, respectively.

Peptides from homologous regions of the four dengue serotypes have different MHC binding ability, as illustrated in **Figure 7A** for binding to HLA-DR401. The proteins encoded in the dengue genome are indicated along the horizontal axis (C: capsid; M: membrane; E: envelope; NS: nonstructural proteins). Peptides that enriched in the yeast display assay are marked by a line (serotype 1 in blue, serotype 2 in purple, serotype 3 in red, and serotype 4 in grey). The proteome is smoothed to 9 amino acid stretches (as in **Figure 5B**), with a given 9 amino acid region marked as a hit if five or more of the seven adjacent peptides enrich. For each 9mer, the maximum number of serotypes with a conserved identical 9mer at that position is indicated at the top in black.

These data can reveal relationships between conservation and binding ability. **Figure 7B-D** shows enrichment data for individual 15mer peptides, with consensus inferred 9mer cores in bold and non-conserved amino acids in these cores highlighted in yellow, as in **Figure 5C-E**. Conserved cores which show binding ability (**Figure 7C**) may be ideal T cell targets. However, the permissiveness of the binding groove allows for peptides to bind that have mutations at the anchors, such as in NS5 (**Figure 7D**), where P4 Asn and P4 Met both allow binding. Interestingly, the serotype 3 core (LASNAICSA) only enriched in four peptides, which is below our described cutoff for high-confidence peptide cores. However, three adjacent peptides enriched and register-inference for these peptides identifies the non-native, linker-containing version of the LASNAICSA core as binding in the MHC-binding groove. This results in an adjacent 9mer being highlighted as a binder in this region (**Figure 7A**) because overlapping 15mers enrich in five or more of the seven adjacent peptides. With this in mind, care must be taken for core identification in enriched regions and can be aided by coupling enrichment with register-inference of enriched peptides. Further, we can also see relationships between conservation and binding in non-conserved regions, such as in the envelope protein (**Figure 7B**) with the mutations in serotype 3 enabling binding.

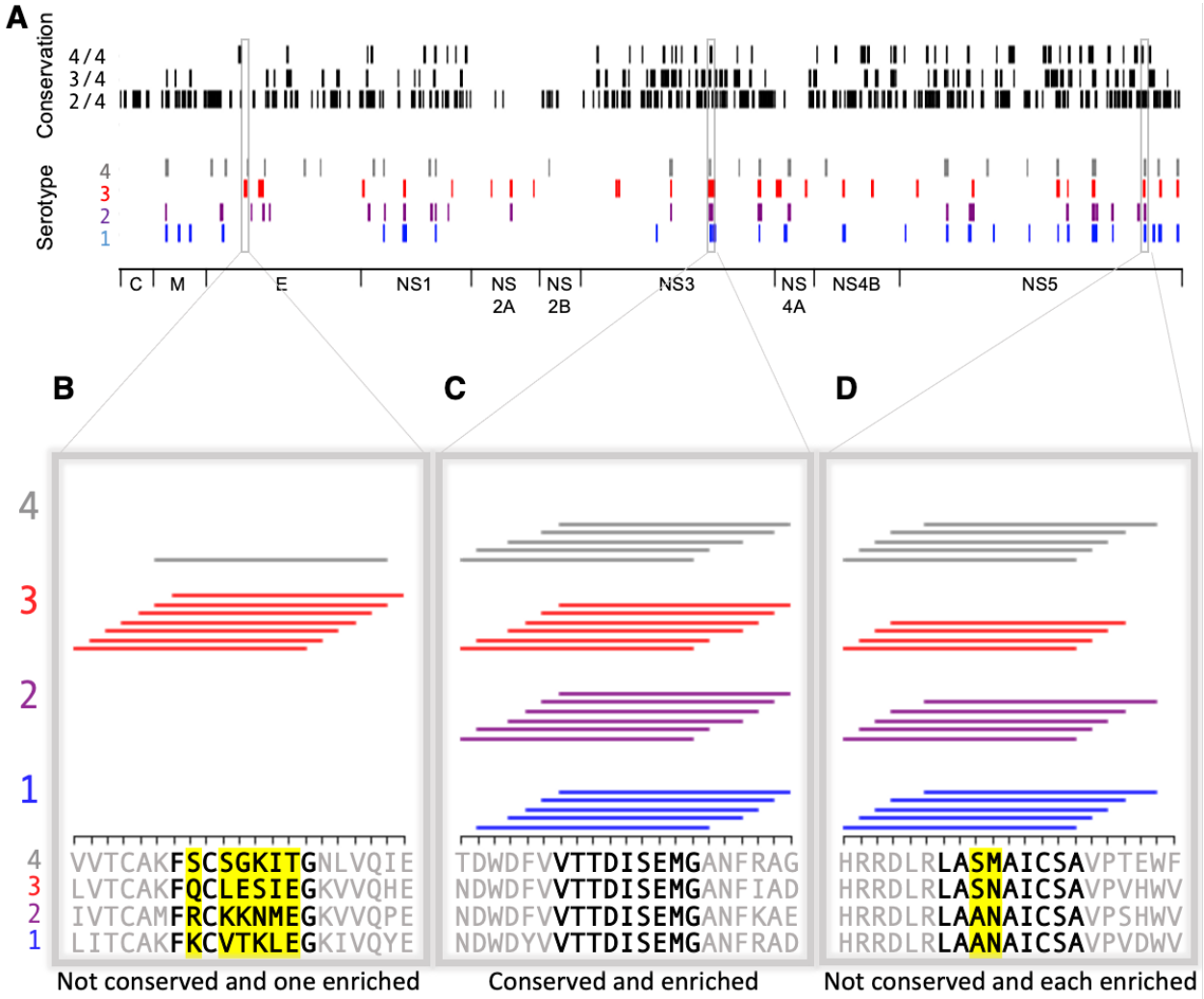


Figure 7. Conservation and enrichment of dengue virus serotypes 1-4. A) Conservation and enrichment of 9mer peptides along four aligned dengue serotypes. All stretches of 9 amino acids are compared across the four serotypes and conservation is indicated with a black vertical line (i.e. 2, 3, or 4 of 4 serotypes conserved). 9mers which enriched on HLA-DR401 are also indicated, colored by virus serotype. **B-D)** Zoomed regions, showing enrichment for individual 15mer peptides to HLA-DR401. Only peptides which contain the bolded 9mer sequence are shown. Amino acids in the bolded 9mer that are not conserved between serotypes are highlighted in yellow. Insets show regions which are differently conserved and enriched: **B)** non-conserved sequences with peptides from one serotype enriched; **C)** conserved sequences enriched across all serotypes; **D)** non-conserved sequences which are enriched.

4.9 Discussion and outlook

$CD4^+$ T cell responses play important roles in infection, autoimmunity, and cancer. By extension, understanding peptide-MHC binding is critical for identifying and engineering T cell epitopes. Here we present an approach to directly assess defined libraries of peptides covering whole pathogen proteomes for binding to MHC-II proteins. We examine alternative modes of selection and utilize overlapping peptides to determine high-confidence binders. We demonstrate the utility of this approach by identifying binders that are missed by prediction algorithms, highlighting a prediction algorithm bias against cysteine-containing peptides and sensitivity to peptide flanking residues (**Table 2** and **Table 3**). Finally, this approach can be utilized for different objectives, including comparing binding to multiple MHC alleles (**Figure 5**) or comparing peptides from

related pathogen sequences for MHC-II binding (**Figure 7**). Whole protein- or proteome-scale analysis across related viruses provides insight into relationships between conserved epitopes and MHC binding (**Figure 5B, 7A**) and specific examples validate the consistency with the underlying biophysics of peptide-MHC binding (**Figures 5C-E** and **7B-D**).

This approach for direct assessment shows benefit compared to prediction algorithms for identifying binders, particularly for finding weak peptide binders. The overlapping peptides in our library were useful for identifying enriched cores, especially when combined with our register inference to identify consensus cores shared between these overlapping peptides. NetMHCIIpan 4.0 exhibits a sensitivity to length and register, which may cause users to miss binders, albeit potentially of lower affinity. Of the overlapping peptides we tested to study this phenomenon, NetMHCIIpan 4.0 correctly ranked the affinities of the overlapping peptides (**Table 3**), but missed binders. **Supplemental Figure 3** also highlights the sensitivity of NetMHCIIpan 4.0 to flanking sequences, where neighboring peptides with shared cores often are not predicted to bind, resulting in fewer clusters of peptides in **Supplemental Figure 3**.

Our work reveals insights on the design of epitope identification experiments, including the utility of overlapping peptides and considerations for comparing libraries of unbiased and proteome-derived peptides. Design of defined libraries with sources of redundancy, such as overlapping peptides, was critical for determining binders with higher degrees of confidence and allowed us to apply stringent cutoffs for individual peptides. Overlapping peptides allowed us to account for construct-specific confounding effects, such as the peptides binding using non-native residues in the linker. Future iterations can change the sequence of the linker, such as defining favorable P(-1) and P10 anchors to fix the register (Rappazzo et al., 2020), although these adaptations would likely require MHC-specific knowledge in advance and may need to be altered for different MHCs. Additionally, the engineered redundancy and multiple modes of selection result in hyperparameters that can be tuned to meet users' stringency requirements, such as defining different thresholds for calling individual 15mer binders or alternative integration of overlapping binders. Additionally, our comparison of unbiased and proteome-derived libraries highlights how aggregate motifs may be affected by underlying amino acid preferences found in protein sequences themselves, which may inadvertently disfavor sequences that can bind strongly to MHC molecules yet consist of amino acid covariates that are not as commonly found in proteins.

Further, this approach can be used to study MHC binding between similar viruses, as done with the dengue proteomes and the spike proteins from SARS-CoV-2 and SARS-CoV, highlighting regions where mutations disrupt binding as well as regions where binding is unperturbed. This method can also be rapidly adapted to study future sequences if pathogens evolve over time.

As experimental approaches and computational approaches continue to co-develop, they present complementary benefits. Though this platform allows for rapid assessment of peptide-MHC binding, the speed of computational prediction surpasses experimental approaches. NetMHCIIpan 4.0 prediction and yeast display selections identified sets of non-overlapping misses, highlighting a utility for both. Additionally, all agreed binders and non-binders matched fluorescence polarization results, suggesting a consensus of yeast display enrichment and algorithmic prediction provide high-confidence results. Approaches such as yeast display assessment can be used to complement computational approaches, such as for identifying cysteine-containing peptides which

are still under-predicted by algorithms. Similarly, prediction algorithms can be trained using large, quality datasets to account for biases. In another application, our platform to assess peptide-MHC binding can be used to design high-throughput assays to test peptide immunogenicity in clinical samples (Klinger et al., 2015; Snyder et al., 2020).

Defined yeast display peptide libraries can also be readily applied to identification of T cell ligands and present an opportunity for identifying unknown ligands from orphan TCRs known to respond to a proteome of interest (Birnbaum et al., 2014; Gee et al., 2018b). Indeed, as DNA synthesis and sequencing continue to advance, defined peptide libraries expanding beyond viral proteomes to covering whole bacterial or human proteomes will be possible, and could present opportunities for investigating autoimmune diseases, which frequently have strong MHC-II associations (Karnes et al., 2017). Such tools would be rich resources for identifying both peptide-MHC binders and TCR ligands.

4.9 Methods

Library design and creation

Yeast display libraries were designed to cover all 15mer sequences within a given proteome, with step size one. Reference proteomes used in creating defined libraries were accessed from Uniprot, with the following Proteome IDs. SARS-CoV-2: UP000464024, SARS-CoV: UP000000354, dengue serotype 1: UP000002500, dengue serotype 2: UP000180751, dengue serotype 3: UP000007200, dengue serotype 4: UP000000275. The dengue proteome is expressed as a single polypeptide, and peptides were generated from that contiguous stretch.

Each library peptide is encoded in DNA space, with specific codons selected randomly from possible codons, with probabilities matching yeast codon usage (GenScript Codon Usage Frequency Table). The DNA-encoded peptide sequences were flanked by invariant sequences from the yeast construct for handles in amplification and cloning, and the DNA oligonucleotide sequences were ordered from Twist Bioscience (South San Francisco, CA), with maximum length of 120 nucleotides. The DNA oligo pool was amplified in low cycle PCR, followed by amplification with construct DNA using overlap extension PCR. This extended product was assembled in yeast with linearized pYAL vector at a 5:1 insert:vector via electroporation with electrocompetent RJY100 yeast.

HLA-DR401 and HLA-DR402 libraries were generated using previously described vectors (Rappazzo et al., 2020) which contain mutations from wild type Metα36Leu, Valα132Met, Hisβ33Asn, and Aspβ43Glu to enable proper folding without disrupting TCR or peptide contact residues (Birnbaum et al., 2017). HLA-DR404 was generated using the same stabilizing mutations. As previously described (Rappazzo et al., 2020), the peptide C-terminus is connected to the MHC construct via a Gly-Ser linker (**Figure 1A**), and the N-terminus of the peptide includes an extra alanine to ensure consistent cleavage between the construct and its signal peptide.

The previously described null library (Dai et al., 2021) was generated with a peptide encoded as "NNNTAANNNNNNNNNTAGNNNNNNNNNNNTGANNNNNN", where "N" indicates any nucleotide and encodes ten random amino acids and three stop codons. This library was similarly generated in yeast using electrocompetent RJY100 yeast.

Peptide visualizations and predictions

Data visualizations of viral conservation and enrichment were generated using custom scripts. For each 9mer stretch in a protein of interest, there are seven 15mer sequences that overlap and contain that 9mer. We calculate how many of these seven 15mers enriched in both the doped and undoped libraries. If five or more of the seven 15mers enriched, that stretch is marked as a 'hit'. To examine conservation between viruses, viral proteins are aligned using ClustalOmega (Madeira et al., 2019). Aligned 9mer stretches are compared between viruses and identical stretches are considered conserved. Hits are determined individually for each virus before merging, such that gaps in sequence alignments do not affect calculations of enrichment for a given virus.

Representations of 15mer hits (as in **Figure 5**, **Figure 7** and **Supplemental Figure 3**) were generated using in-house scripts, such that a 15mer that enriched in both the doped and undoped library was marked as a horizontal line above the relevant 15mer sequence. Only 15mers containing the bolded 9mer in **Figure 5** and **Figure 7** were included.

NetMHCIIpan 4.0 webserver was used for computational predictions (Reynisson et al., 2020), where a binder is defined as having a predicted percent rank ≤ 10%, as defined in the webserver instructions.

Yeast library selections

Library selections were consistent with previous peptide-MHC-II yeast display dissociation studies (Dai et al., 2021; Rappazzo et al., 2020). Yeast were washed into pH 7.2 PBS with 1 μM 3C protease and incubated at room temperature for 45 minutes. Yeast were then washed into 4°C acid saline (150 mM NaCl, 20 mM citric acid, pH 5) with 1 μM HLA-DM and incubated at 4°C overnight. Each step takes place in the presence of competitor peptide (HLA-DR401: $HA_{306-318}$ PKYVKQNTLKLAT, 1 μM; HLA-DR402: $CD48_{36-51}$ FDQKIVEWDSRKSKYF, 5 μM; HLA-DR404: NKVKSLRILNTRRKL, 5 μM (Vita et al., 2019)). Non-specific binders are removed by incubating yeast with ∝-AlexaFluor647 magnetic beads and flowed over a magnetic Miltenyi column at 4°C (Miltenyi Biotec; Bergisch Gladbach, Germany). A positive selection follows, comprised of incubation with ∝-Myc-AlexaFluor647 antibody (1:100 volume:volume) and ∝-AlexaFluor647 magnetic beads (1:10 volume:volume) and flowed over a Miltenyi column on a magnet at 4°C, such that yeast with bound peptide are retained on the column. These yeast are eluted, grown to confluence in at 30°C in SDCAA media (pH 5), and sub-cultured in at 20°C SGCAA media (pH 5) at OD600=1 for two days. The first round of selections of doped libraries were conducted on 180 million yeast (SARS-CoV-2 library) or 400 million yeast (dengue library) to ensure at least 20-fold coverage or peptides. Subsequent rounds of doped library selection, and all rounds of undoped library selections, were performed on 20-25 million yeast.

Library sequencing and analysis

Libraries were deep sequenced to determine their composition after each round of selection. Plasmid DNA was extracted from ten million yeast from each round of selection using the Zymoprep Yeast Miniprep Kit (Zymo Research), following manufacturer instructions. Amplicons were generated through PCR, covering the peptide sequence through the 3C cut site. A second PCR round was performed to add i5 and i7 sequencing handles and in-line index barcodes unique to each round of selection. Amplicons were sequenced on an Illumina MiSeq using paired-end MiSeq v2 300bp kits at the MIT BioMicroCenter.

Paired-end reads were assembled using PandaSeq (Masella et al., 2012). Peptide sequences were extracted by identifying correctly encoded flanking regions, and were filtered to ensure they matched designed members of the library or the randomized null construct encoding, providing a stringent threshold for contamination and PCR and read errors.

The resulting data are analyzed for convergence, as described in the main text. Once a library has converged, it is likely that changes in subsequent rounds of selection are due to stochastic variation rather than improved binding.

Register inference and sequence logos

The 9mer core of enriched sequences was inferred using an in-house alignment algorithm. In this approach, we utilize a 9mer position weight matrix (PWM), which we assess at different offsets along the peptide. We one-hot encode sequences and pad with zeros on the C-terminus of the peptide; to assess seven native registers and four non-native registers, we pad the peptides with four zeros. Three of the non-native registers utilize the linker at the P9 anchor but not the P6 anchor, and the addition of a fourth register captures a minority set of peptides which utilize Gly-Gly-Ser-Gly of the linker at P6 through P9 in the groove. Register-setting is performed with zero-padded 15mers, rather than 15mers flanked by invariant flanking residues, because the PWM would otherwise align all sequences to the invariant region.

At the start, we randomly assign peptides to registers and generate a 9mer PWM. Over subsequent iterations, peptides are assigned to new registers and the PWM was updated. Assignments are random but biased, such that clusters corresponding to registers that match the PWM are favored. Specifically, at each assignment we first take out the sequence under consideration from the PWM. The PWM then defines an energy value for each register shift of a given peptide, which is then used to generate a Boltzmann distribution from which we sample the updated register shift. The stochasticity is decreased over time by raising the inverse temperature linearly from 0.05 to 1 over 60 iterations, simulating 'cooling' (Andreatta et al., 2017). A final deterministic iteration was carried out, where the distribution concentrates entirely on the optimal register shift.

After register inference, sequence logo visualizations of the 9mer cores were generated using Seq2Logo-2.0 with default settings, except using background frequencies from the SARS-CoV-2 proteome and SARS-CoV spike and nucleocapsid proteins (Thomsen and Nielsen, 2012). For registers with the C-terminus utilizing the C-terminal linker, the relevant linker sequence was added to achieve a full 9mer sequence for visualizing the full 9mer core. For HLA-DR401, distribution among registers, starting from N-terminally to C-terminally aligned in the peptide, is: 161, 237, 227, 238, 231, 279, 237, 266, 271, 202, 118.

Recombinant protein expression

HLA-DM and HLA-DR401 were expressed recombinantly in High Five insect cells (Thermo Fisher; Waltham, MA) using a baculovirus expression system, as previously described (Birnbaum et al., 2014; Rappazzo et al., 2020). Ectodomain sequences of each chain were formatted with a C-terminal poly-histidine purification tag and cloned into pAcGP67a vectors. Each vector was individually transfected into SF9 insect cells (Thermo Fisher) with BestBac 2.0 linearized baculovirus DNA (Expression Systems; Davis, CA) and Cellfectin II Reagent (Thermo Fisher), and propagated to high titer. Viruses were co-titrated for optimal expression to maximize balanced

MHC heterodimer formation, co-transduced into Hi5 cells, and grown for 48-72 hours at 27°C. The secreted protein was purified from pre-conditioned media supernatant with Ni-NTA resin and purified via size exclusion chromatography with a S200 increase column on an AKTA PURE FPLC (GE Healthcare). To improve protein yields, the HLA-DRB1*04:01 chain was expressed with a $CLIP_{87\text{-}101}$ peptide (PVSKMRMATPLLMQA) connected to the N-terminus of the MHC chain via a flexible, 3C protease-cleavable linker.

Fluorescence polarization experiments for peptide IC_{50} determination

Peptide IC_{50} values were determined following a protocol modified from Yin & Stern (Yin and Stern, 2014), as in Rappazzo et al (Rappazzo et al., 2020). In the assay, recombinantly expressed HLA-DR401 is incubated with fluorescently labelled modified $HA_{306\text{-}318}$ (APRFV{Lys(5,6 FAM)}QNTLRLATG) peptide and a titration series for each unlabeled competitor peptide is added (1.28 nM – 20 μM). A change in polarization value resulting from displacement of fluorescent peptide from the binding groove is used to determine IC_{50} values.

Relative binding at each concentration is calculated as $(FP_{sample} - FP_{free})/(FP_{no\_comp} - FP_{free})$. Here, FP_{free} is the polarization value for the fluorescent peptide alone with no added MHC, FP_{no_comp} is polarization value for MHC with no competitor peptide added, and FP_{sample} is the polarization value with both MHC and competitor peptide added. Relative binding curves were then generated and fit in Prism 9.3 (GraphPad Software Inc; San Diego CA) to the equation $y = 1/(1+[pep]/IC_{50})$, where [pep] is the concentration of un-labelled competitor peptide, in order to determine the concentration of half-maximal inhibition, the IC_{50} value.

Each assay was performed at 200 μL, with 100 nM recombinant MHC, 25 nM fluorescent peptide, and competitor peptide (GenScript; Piscataway, NJ). This mixture co-incubates in pH 5 binding buffer at 37°C for 72 hours in black flat bottom 96-well plates. Competitor peptide concentrations ranged from 1.28 nM to 20 μM, as a five-fold dilution series. Three replicates are performed for each peptide concentration. Fluorescent peptide-only, no competitor peptide, and binding buffer controls were also included. Our MHC was expressed with a linked CLIP peptide, so prior to co-incubation, the peptide linker is cleaved by addition of 3C protease at 1:10 molar ratio at room temperature for one hour; the residual cleaved 100 nM CLIP peptide is not expected to alter peptide binding measurements.

Measurements were taken on a Molecular Devices SpectraMax M5 instrument. G-value was 1.1 for each plate, as calculated per manufacturer instructions for each plate based on fluorescent peptide-only wells minus buffer blank wells, with 35 mP reference for 5,6FAM (Fluorescein setting). Measurements were made with 470 nm excitation and 520 nm emission, 10 flashes per read, and default PMT gain high.

Data Availability

All deep sequencing data are deposited on the Sequence Read Archive (SRA), with accession codes PRJNA806475 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA806475] and PRJNA708266 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA708266]

Code Availability
Scripts used for data processing and visualization are publicly available at https://github.com/birnbaumlab/Huisman-et-al-2022.

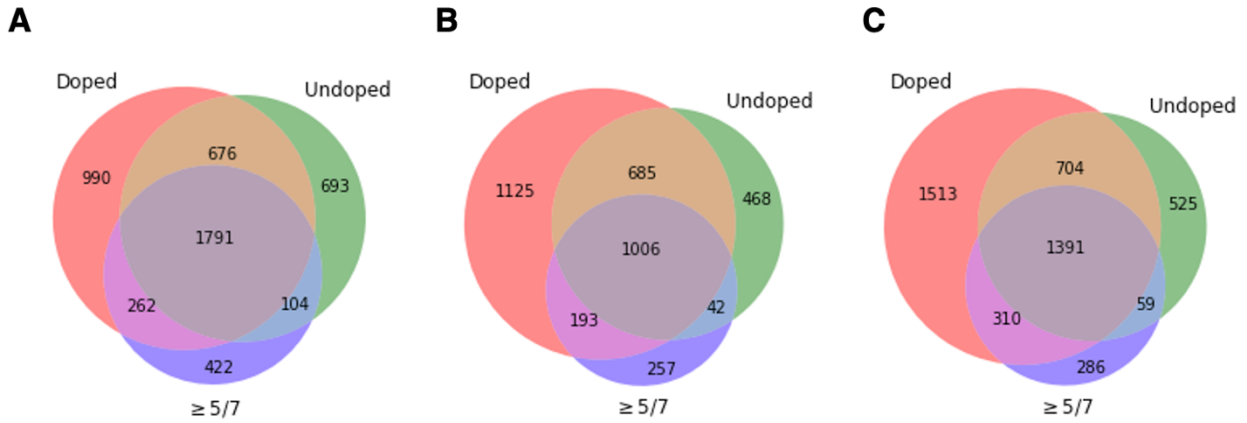
4.10 Acknowledgements

The material covered in this chapter has been submitted for review as Huisman, Dai, Gifford, Birnbaum (2022). "A high-throughput yeast display approach to profile pathogen proteomes for MHC-II binding". (Huisman et al., 2022). Zheng Dai contributed to the development of the register inference algorithm. Zheng Dai and David K, Gifford provided critical feedback on the manuscript. Michael E. Birnbaum contributed to the conceptualization of the work, guided experimental design and contributed to the data analysis and manuscript writing. Thank you to Patrick Holec for feedback on the manuscript, and the MIT BioMicro Center for library sequencing. This work was supported in part by the Koch Institute Support (core) Grant P30-CA14051 from the National Cancer Institute. This work was supported by National Institute of Health (U19-AI110495), the Packard Foundation to M.E.B., a Schmidt Futures grant to D.K.G. and M.E.B., and a National Science Foundation Graduate Research Fellowship to B.D.H.

4.11 Supplemental Figures

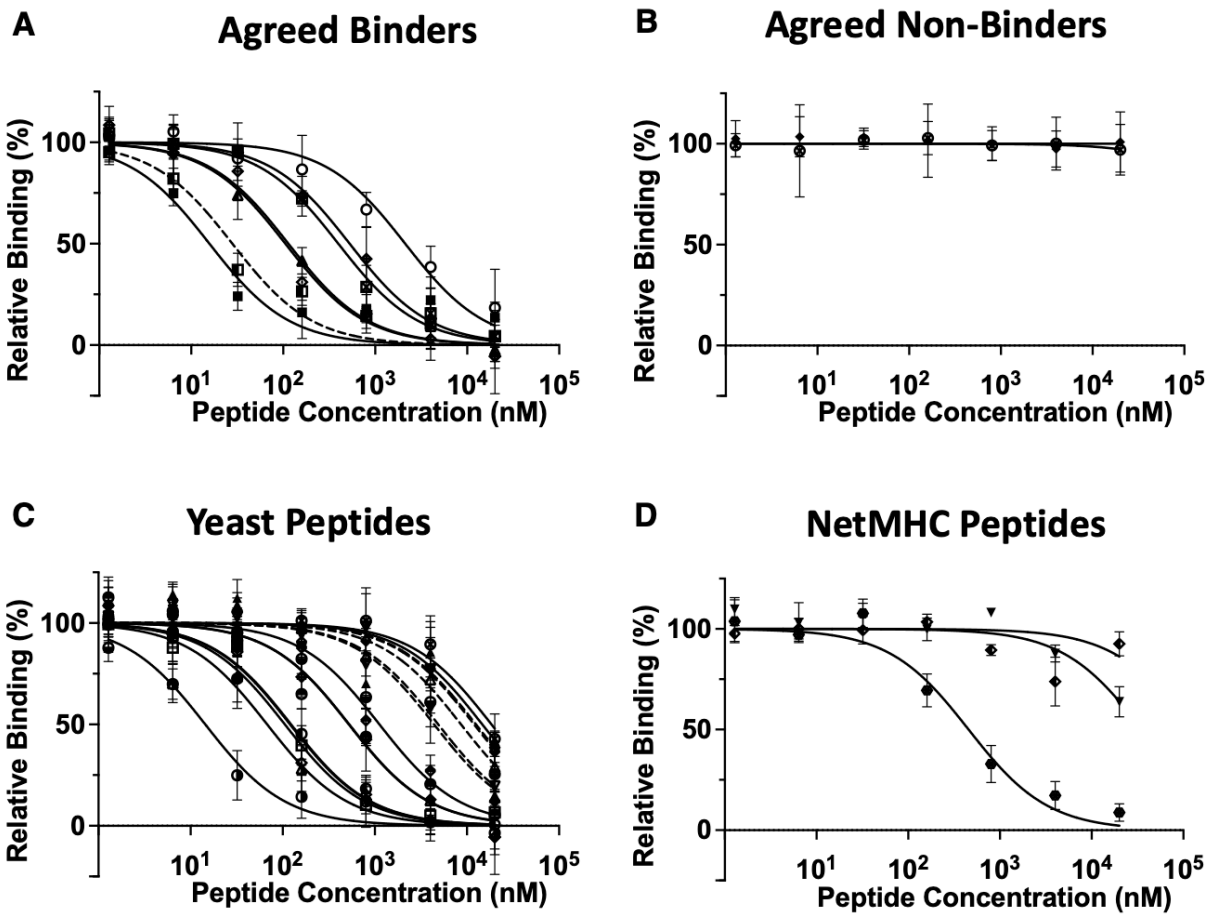
		Doped					Undoped						
		R0	R1	R2	R3	R4	R0	R1+	R1-	R2+	R2-	R3+	R3-
Doped	R0	1.00	0.16	0.16	0.16	0.16	0.09	0.18	-0.03	0.18	0.11	0.16	0.16
	R1	0.16	1.00	0.91	0.86	0.75	0.24	0.76	-0.30	0.78	0.42	0.72	0.72
	R2	0.16	0.91	1.00	0.95	0.85	0.21	0.79	-0.35	0.86	0.37	0.81	0.76
	R3	0.16	0.86	0.95	1.00	0.94	0.17	0.75	-0.37	0.89	0.25	0.88	0.69
	R4	0.16	0.75	0.85	0.94	1.00	0.13	0.65	-0.34	0.84	0.11	0.90	0.53
Undoped	R0	0.09	0.24	0.21	0.17	0.13	1.00	0.36	0.41	0.21	0.42	0.16	0.27
	R1+	0.18	0.76	0.79	0.75	0.65	0.36	1.00	-0.27	0.86	0.66	0.76	0.87
	R1-	-0.03	-0.30	-0.35	-0.37	-0.34	0.41	-0.27	1.00	-0.40	0.04	-0.38	-0.32
	R2+	0.18	0.78	0.86	0.89	0.84	0.21	0.86	-0.40	1.00	0.34	0.94	0.81
	R2-	0.11	0.42	0.37	0.25	0.11	0.42	0.66	0.04	0.34	1.00	0.18	0.62
	R3+	0.16	0.72	0.81	0.88	0.90	0.16	0.76	-0.38	0.94	0.18	1.00	0.66
	R3-	0.16	0.72	0.76	0.69	0.53	0.27	0.87	-0.32	0.81	0.62	0.66	1.00

Supplemental Figure 1. Correlations between selection rounds. Pearson correlation for HLA-DR401 SARS-CoV-2 and SARS-CoV defined library members (+/- signs indicate enriched (+) or not enriched (-) yeast in undoped library; rounds of selection are indicated e.g. "R1" indicates "Round 1", and "R0" is the unselected "Round 0" library).

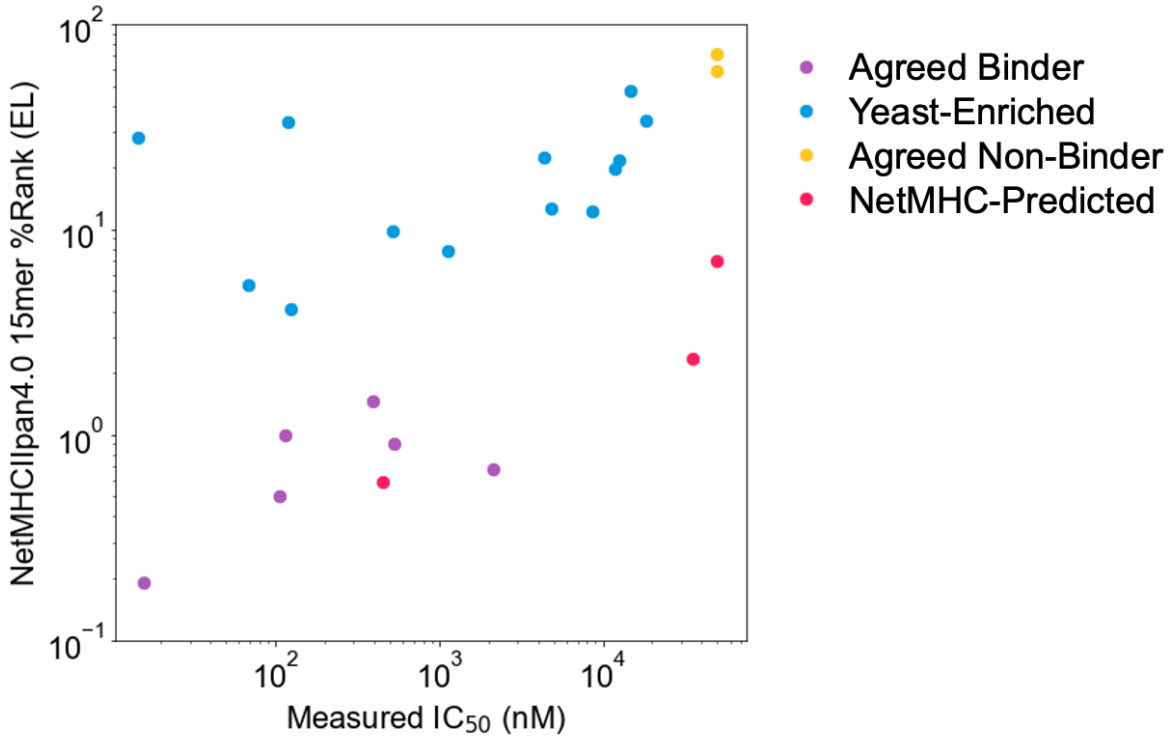


Supplemental Figure 2. Full Venn diagrams. Full Venn diagrams showing relationships between peptides which enriched in the doped library ("Doped"), and undoped library ("Undoped"), and contained a 9mer peptide which enriched in five or more of the seven 15mers containing it ("≥5/7"), for **A)** HLA-DR401, **B)** HLA-DR402, and **C)** HLA-DR404.

Supplemental Figure 3. Comparing defined library selections with algorithmic predictions: SARS-CoV-2 spike protein. 15mer peptides which enriched for binding to HLA-DR401 in both the doped and undoped libraries are indicated with horizontal lines above the enriched 15mer sequence (blue). NetMHCIIpan 4.0 predicted binders (rank ≤ 10%) on yeast-formatted peptides are shown in red. Boxed sequences are tested in subsequent fluorescent polarization experiments, and colored as indicated in the legend.



Supplemental Figure 4. Titration curves for peptides tested via fluorescence polarization for binding to HLA-DR401, by category. A) Agreed binder peptides which are predicted to bind by NetMHCIIpan 4.0 and enriched in yeast display experiments. Dashed line is the positive control HA peptide. **B)** Agreed non-binder peptides which did not enrich in yeast display experiments and were not predicted to bind by NetMHCIIpan 4.0. **C)** Yeast enriched peptides from Table 2 and Table 3. Offset variants from Table 3 are dashed lines. **D)** NetMHCIIpan 4.0 predicted peptides which are not enriched in the yeast display library.



Supplemental Figure 5. Comparing measured IC_{50} values and predictions. Relationship between measured IC_{50} values and NetMHCIIpan 4.0 predicted ranks in Eluted Ligand mode (EL) on unflanked (native) 15mer sequences. Data points are colored by label, and IC_{50} values ≥50 μM are set to 50 μM.

CHAPTER 5: SECOND-GENERATION PLATFORM FOR INCREASING MHC-II THROUGHPUT IN YEAST DISPLAY SCREENS

Abstract

Yeast display serves as a powerful tool for assessing peptide-MHC (pMHC) binding and pMHC-TCR binding. This tool is often limited, however, by the need to optimize MHC proteins for yeast surface expression, which can be laborious and time-consuming, with no guarantee that screens for stabilizing mutations will yield productive results. Here we present a second-generation yeast display platform for class II MHC molecules (MHC-II) which decouples MHC-II expression from yeast-expressed peptides, referred to as "peptide display". Peptide display obviates the need for yeast-specific MHC optimizations and increases the throughput of MHC-II alleles in yeast display screens. Because MHC identity is separated from the peptide library, a further benefit of this platform is the ability to assess a single library of peptides against any MHC-II. We demonstrate the utility of the peptide display platform across MHC-II proteins, screening HLA-DR, HLA-DP, and HLA-DQ alleles. We further explore parameters of selections, including reagent dependencies, MHC avidity, and use of competitor peptides. This approach presents an advance in throughput and accessibility of screening peptide-MHC-II binding, and we explore the potential implications of this technology in identifying disease-relevant peptides, such as in autoimmune diseases.

5.1 Introduction

Yeast display assessment of peptide-MHC-II binding has proven useful for generating large, high-quality datasets of peptide binders, useful for training prediction algorithms (**Chapter 2**) (Jiang and Boder, 2010; Liu et al., 2021b; Rappazzo et al., 2020), as well as assessing user-defined libraries of peptides for peptide optimization (**Chapter 3**) (Dai et al., 2021) and pathogen screening (**Chapter 4**). In traditional peptide-MHC-II yeast display approaches, peptide and MHCβ are covalently linked (**Figure 1A**), in a mini (Birnbaum et al., 2014; Fernandes et al., 2020) or full length format (Rappazzo et al., 2020) and cloned into the pYAL vector. While theoretically any MHC allele can be presented by yeast display, the simple machinery for protein-folding in yeast can cause some MHCs, especially MHC-IIs, to not fold properly, despite detectable surface expression (Birnbaum et al., 2014; Fernandes et al., 2020; Rappazzo et al., 2020). This necessitates each MHC be validated for fold, such as through binding of a recombinantly expressed TCR (Birnbaum et al., 2014; Fernandes et al., 2020). In the absence of detectable TCR binding, error-prone mutagenesis of the MHC is required to install stabilizing mutations away from the peptide- and receptor-binding surfaces (Birnbaum et al., 2014; Fernandes et al., 2020). This process can be time and effort intensive and requires availability of reagents such as a known MHC-restricted TCR.

To circumvent the need for individual optimization of MHC proteins, we have developed a second-generation yeast display approach for assessing peptide-MHC-II binding which does not require yeast-specific MHC-II optimizations. In this approach, peptide and MHC-II are decoupled, with a library of peptides displayed on the surface of yeast and MHC-II protein expressed recombinantly. Because of this decoupling and exclusive presentation of peptides on the yeast surface, we refer to this approach as "peptide display".

In this chapter, we outline the system, explore experimental conditions, and compare the results with existing yeast display datasets. We apply this approach to study HLA-DR alleles, followed by HLA-DP and HLA-DQ alleles, which have to this point largely been inaccessible in conventional yeast display approaches (Liu et al., 2021b). The original pMHC yeast display platform further requires generation of a new library for each new MHC being screened (Birnbaum et al., 2014; Dai et al., 2021; Fernandes et al., 2020; Gee et al., 2018b; Rappazzo et al., 2020). In contrast, because the peptide library is decoupled from MHC, a single peptide display library is extensible to any MHC-II without requiring additional library cloning and generation. Because of its functionality across MHCs, ongoing work applies this platform to antigen discovery in autoimmunity through comparison of the peptide repertoires of disease-predisposing and disease-protective MHC-II alleles.

5.2 Design of peptide display libraries and experiments

For the peptide display approach, we adapted the pCT302 vector for protein fusions to the C-terminus of Aga2 (Midelfort et al., 2004). We left unchanged the N-terminus of the construct, which consists of Aga2 fused to an epitope tag and a subsequent C-terminal linker sequence. To the C-terminus of this linker, we connected the N-terminus of our peptide sequence of interest. In order to readily assess frameshift mutations in our peptide sequence, we linked the peptide to a Myc epitope tag via a short linker.

Rather than co-expressing MHC-II proteins on the surface of yeast, MHC-II protein was expressed recombinantly, which circumvents yeast protein folding concerns. MHC-II proteins were stabilized with a linked class II-associated Ii peptide (CLIP) (Busch et al., 2005), specifically $CLIP_{87\text{-}101}$ (PVSKMRMATPLLMQA), connected to the N-terminus of MHC-IIβ chain via a linker containing a 3C protease cleavage site. MHC-II protein was site specifically biotinylated, allowing for easy functionalization and visualization with fluorescent streptavidin.

To assess peptide-MHC-II binding, yeast were co-incubated with fluorescent tetramerized MHC-II protein, 3C protease, and HLA-DM (**Figure 1B**). 3C protease cleaves the CLIP-MHC linker, enabling the stabilizing peptide to dissociate. The peptide exchange catalyst HLA-DM is available to assist with dissociation of CLIP and binding to the yeast-expressed peptide (Jiang et al., 2015). If the MHC is able to bind to the peptide of interest, we can detect fluorescence via flow cytometry-based assessment of the yeast population. We see a clear fluorescent population for yeast expressing a version of the well-studied Influenza A $HA_{306\text{-}318}$ peptide (APKYVKQNTLKLAT) known to bind HLA-DR401 (Hennecke and Wiley, 2002; Rappazzo et al., 2020), with minimal binding to an off-target $CD48_{36\text{-}51}$ peptide (FDQKIVEWDSRKSKYF) (**Figure 1C**) when utilizing tetramerized HLA-DR401 (HLA-DRA1*01:01 / DRB1*04:01). $CD48_{36\text{-}51}$ binds to a related allele, HLA-DR402 (HLA-DRA1*01:01 / DRB1*04:02) (Rappazzo et al., 2020), but is likely dis-preferred for HLA-DR401 binding because of the large aromatic Trp residue at the expected P4 pocket in the optimal register.

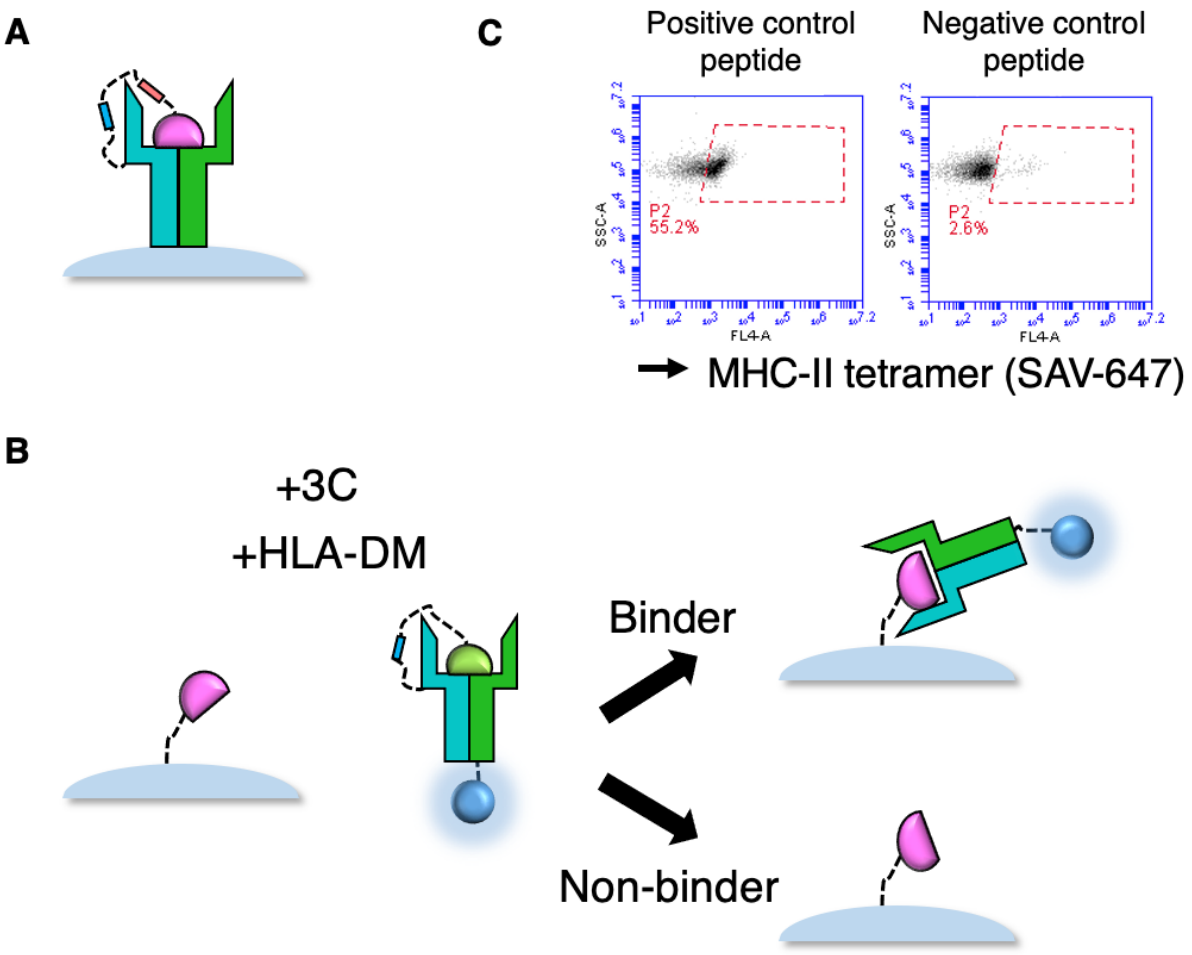


Figure 1. Overview of peptide display formatting and experiments. A) Formatting of pMHC construct for original MHC-II platform. **B)** New peptide display platform formatting and selection strategy. Peptides are expressed on the surface of yeast, and MHC is expressed recombinantly, with $CLIP_{81-101}$ peptide for stability. Upon addition of HLA-DM and 3C protease to cleave the CLIP-MHC linkage, the MHC can bind to yeast-expressed peptides. Binders can be manipulated using a handle on the MHC, such as a fluorescent streptavidin. **C)** Flow cytometry analysis of tetramerized HLA-DR401 with binder (HA-derived) and non-binder (CD48-derived) peptides. MHC is tetramerized with streptavidin-AlexaFluor647 (SAV-647).

To explore the constraints of the system and reliance on each component of the selection reaction mix, we separately titrated HLA-DR401 tetramers, HLA-DM, and 3C protease and assessed peptide-MHC-II binding. As expected, there was MHC tetramer concentration dependence, with binding signal decreasing as MHC concentration decreased (**Figure 2A**). We similarly observe a dependence on the presence of HLA-DM, with largely unchanged signal across concentrations from 0.5 μM – 3.5 μM, and with signal lost at the lowest concentration tested, 75nM (**Figure 2B**). The clear dependence on HLA-DM presence stands in contrast to the minimal HLA-DM reliance in the original, dissociation-based pMHC-II platform (**Figure 1A**), which has previously shown similar results in library selections with or without HLA-DM (Rappazzo et al., 2020). These differences are likely due to the need to dissociate CLIP peptide and bind to a yeast-expressed peptide, compared to approaches in which the peptide of interest begins in the groove and dissociation is assessed (Rappazzo et al., 2020). Next, we were curious if the peptide exchange

reaction could proceed without cleaving the linker between recombinant MHC-II and the stabilizing CLIP peptide. We observe signal is lost when we decrease 3C concentrations, illustrating a need to freely allow the stabilizing peptide to dissociate to enable loading of yeast-expressed peptide (**Figure 2C; Figure 1C** is a subset of this figure). Across all experiments, the negative control peptide showed minimal signal with the concentrations tested.

From these experiments, we selected 75 nM MHC-II tetramer concentration, 0.5 μM HLA-DM concentration, and 1 μM 3C concentration to utilize in large-scale experiments. To fill in the concentration gap in HLA-DM titration (**Figure 2B**), an additional experiment (data not shown) assessing HLA-DM concentrations of 0.5 μM, 0.4 μM, 0.25 μM, and 0.1 μM showed a decrease in signal over concentrations, supporting proceeding with 0.5 μM HLA-DM.

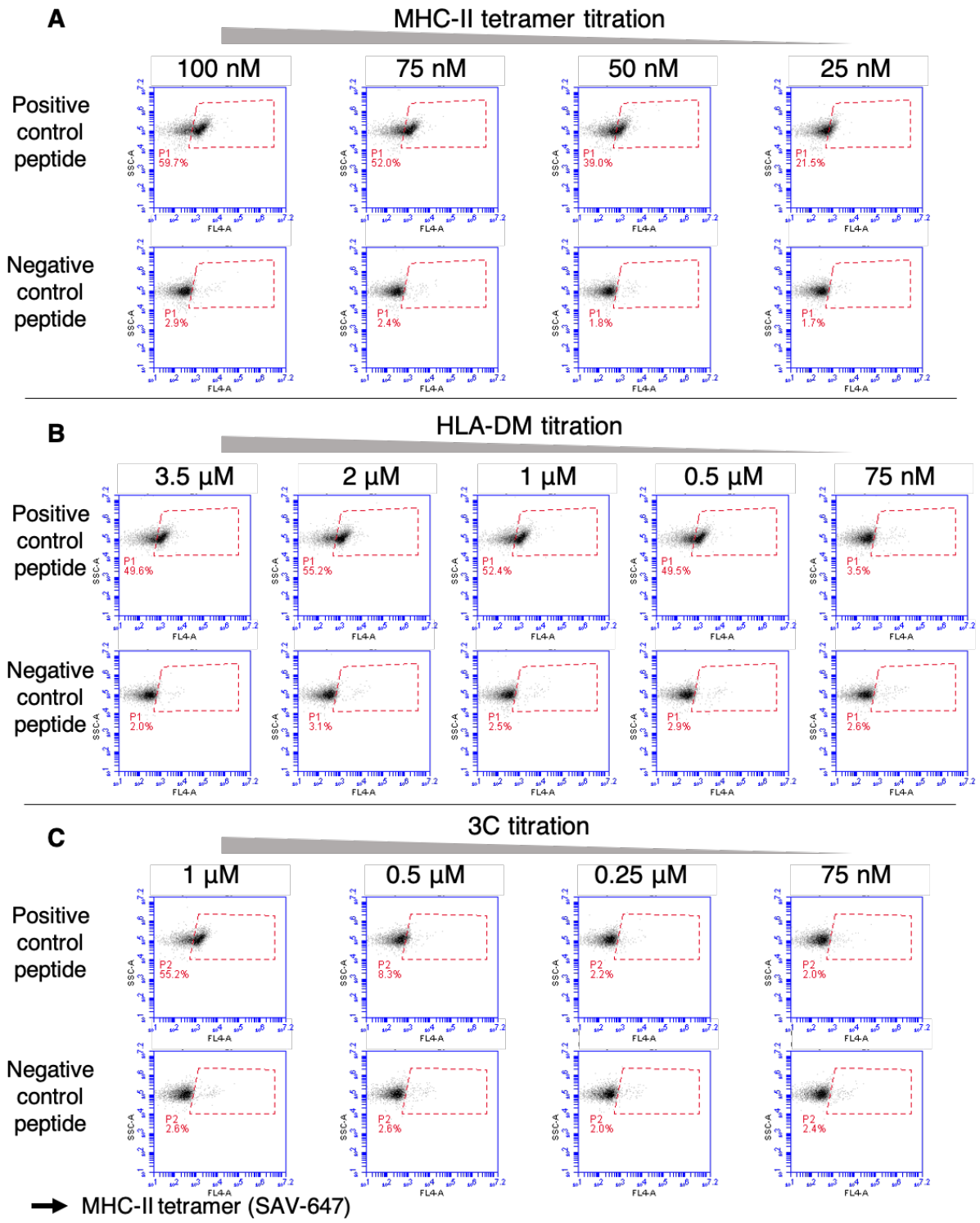


Figure 2. Testing experimental conditions. Reliance on each component of the selection reaction mix is assessed with titrations of **A)** HLA-DR401 tetramers, **B)** HLA-DM, and **C)** 3C protease with binder (HA-derived) and non-binder (CD48-derived) peptides. MHC is tetramerized with streptavidin-AlexaFluor647 (SAV-647).

5.3 Application of peptide display platform to screen HLA-DR401

To utilize the peptide display approach to screen peptides at a repertoire-scale, we next generated a library of fully randomized 13mer peptides, containing 100 million members. We performed selections with HLA-DR401, which would enable comparison with existing pMHC datasets. Yeast were incubated with fluorescent tetramerized HLA-DR401, with 3C protease, and HLA-DM in phosphate buffered saline (PBS) for 45 minutes. The neutral pH allows 3C protease to cleave the linker which connects MHC-II to the stabilizing CLIP peptide. To mimic acidic endosomal conditions, yeast were then moved to an excess of pH 5 acidic saline for an additional 45 minutes. Next, yeast were incubated with anti-fluorophore magnetic beads, and yeast which express peptides which bound to HLA-DR401 were enriched utilizing a magnetic enrichment strategy.

We performed three iterative rounds of selection, growing yeast to confluence between each round. In each round, we sampled yeast after the first incubation in PBS, the second incubation in acidic saline, and after elution from the magnetic column and examine the fluorescent population via flow cytometry. Over subsequent rounds, the population of fluorescent MHC-bound yeast increased at each of these steps (**Figure 3**).

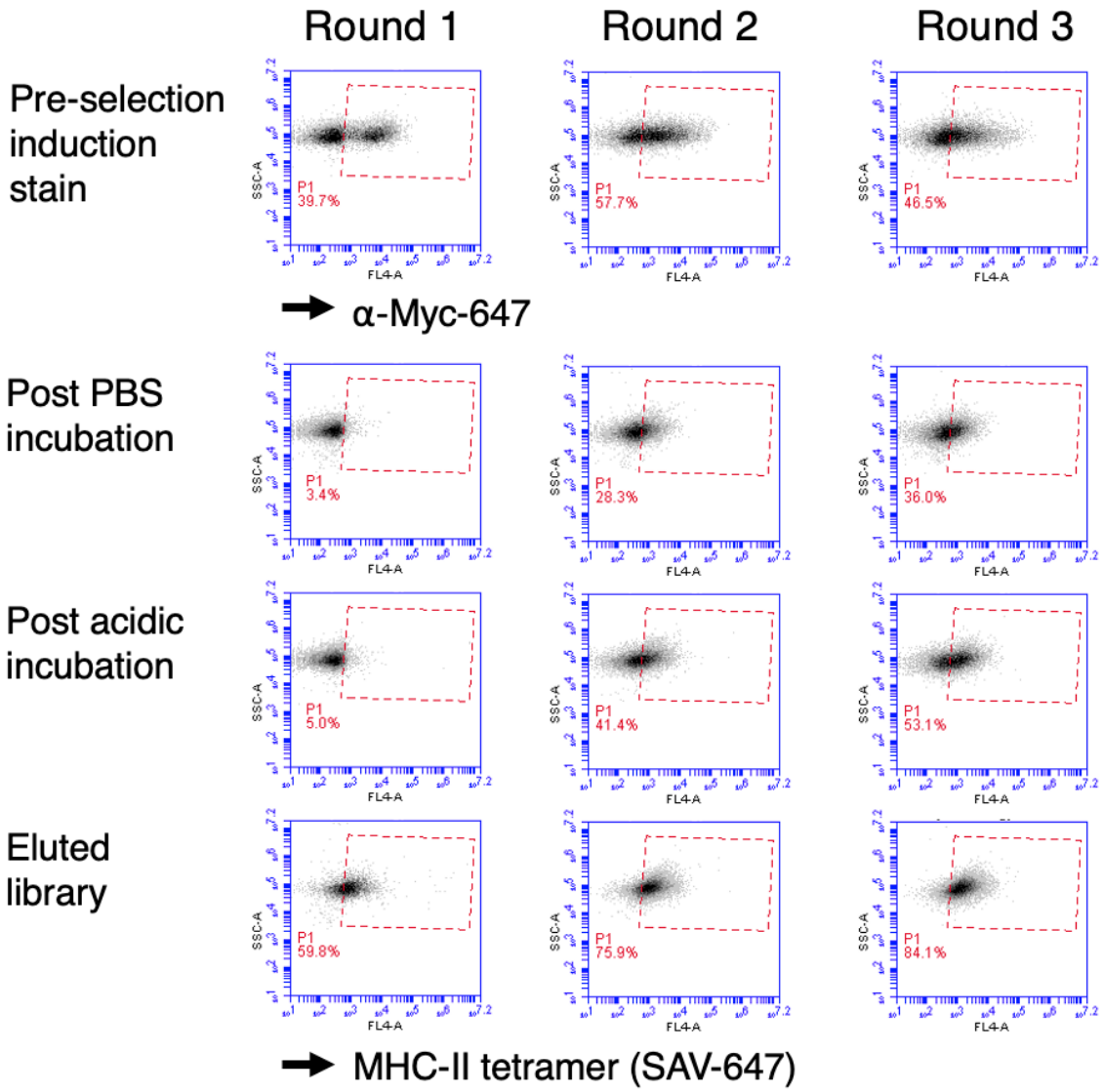


Figure 3. Dynamics of library selections for HLA-DR401 tetramer-based library selections. Pre-selection induction staining with anti-epitope tag antibody and tetramer staining at timepoints throughout each selection, including after incubation in pH 7.2 PBS, after subsequent incubation in acidic saline, and after elution from magnetic enrichment column. MHC is tetramerized with streptavidin-AlexaFluor647 (SAV-647).

Through deep sequencing of the selected libraries, we can determine the identities of peptides enriched in selections. Given the open MHC-II binding groove and 13mer length of peptides assess, we inferred the binding register of each peptide utilizing a position weight matrix approach. Peptides inferred to bind using the central register are represented in **Figure 4**. These representations include fractions of peptides with amino acids at each position (**Figure 4A**), log2 fold change compared to the unselected library (**Figure 4B**), and sequence logos (**Figure 4C**). We observe a strong P1 preference for large and aromatic hydrophobic residues, a preference at P4 for acidic residues and non-aromatic hydrophobic residues. Polar residues are preferred at P6, and small residues are preferred at P9. The auxiliary anchor P7 exhibits some preference for Pro, Leu, and Asn. This motif is consistent with motifs from other high-throughput datasets (Abelin et al., 2019; Racle et al., 2019; Rappazzo et al., 2020; Reynisson et al., 2020; Scally et al., 2017), with a greater representation of P4 hydrophobic residues.

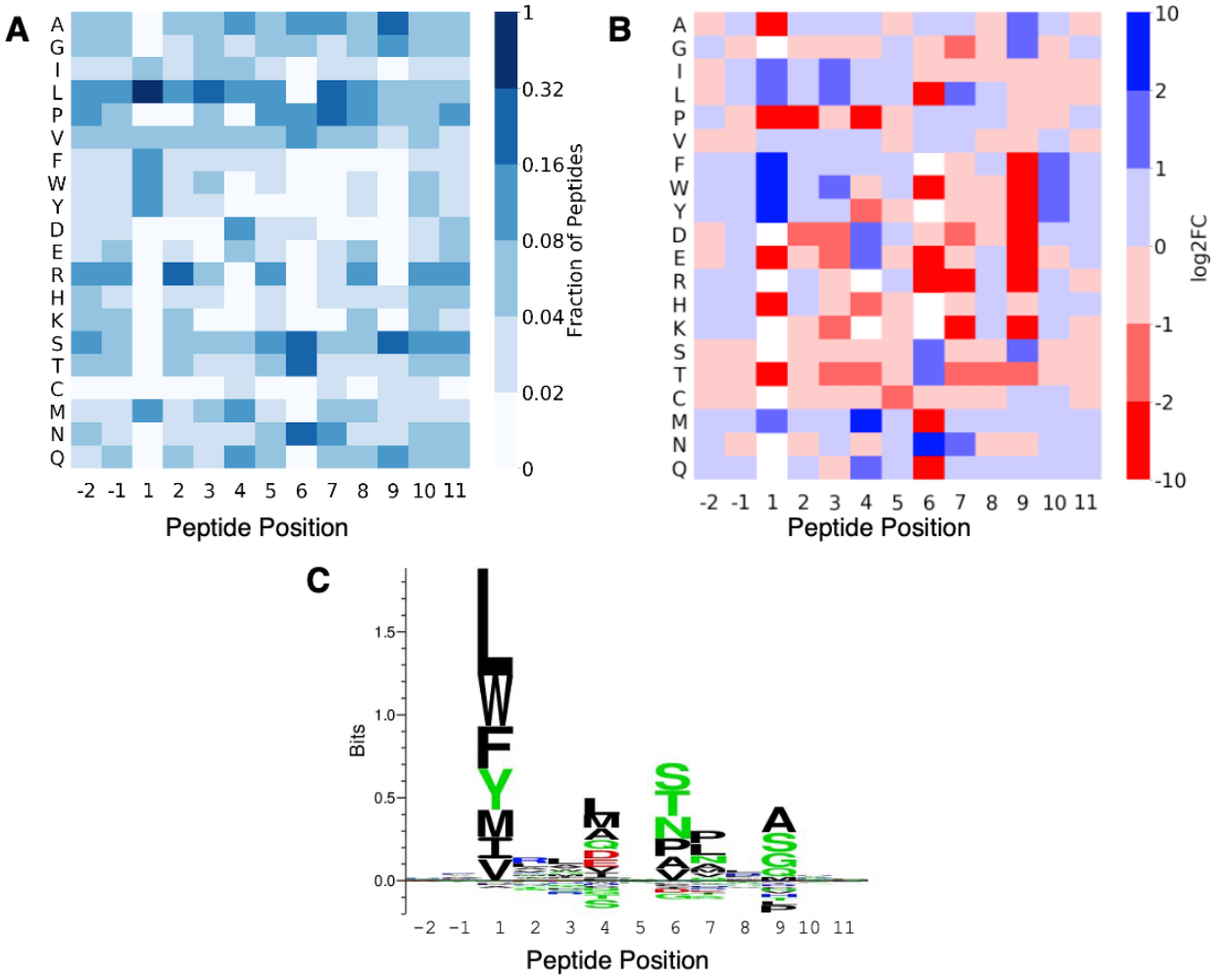


Figure 4. Motifs of enriched peptides in the central register from HLA-DR401 tetramer-based library selections. A) Heatmap highlighting the fraction of peptides with each amino acid at each position. **B)** Heatmap of log2 fold change of positional amino acid frequency compared to unselected library. Amino acids which did not appear in the enriched sequences at a given position are white. **C)** Sequence logo of enriched peptides.

To investigate the relative prevalence of P4 acidic residues and ensure acidic residue-containing peptides were not systemically excluded from MHC-binding in this format, we next assessed yeast expressing individual peptides for their ability to bind to HLA-DR401. These peptides have previously-determined affinity values for HLA-DR401. Performing binding reactions similar to selection experiments, we utilize flow cytometry to examine binding after the first incubation in PBS and the second incubation in acidic saline. We inferred register from similarly with HLA-DR401 peptide motifs and underlined the central 9mer core for each (**Figure 5**). After the incubation in acidic saline, we observe a clear monotonic relationship between measured IC_{50} value and percent MHC-tetramer positive population or mean fluorescence intensity. This relationship is only clear after the second incubation, which suggests that the acidic incubation may be helpful in enabling peptide-MHC binding or establishing the protonation state of acidic residues. The clear binding of peptides containing acidic residues suggests that there is not systematic bias against acidic residues, at least at an individual peptide level.

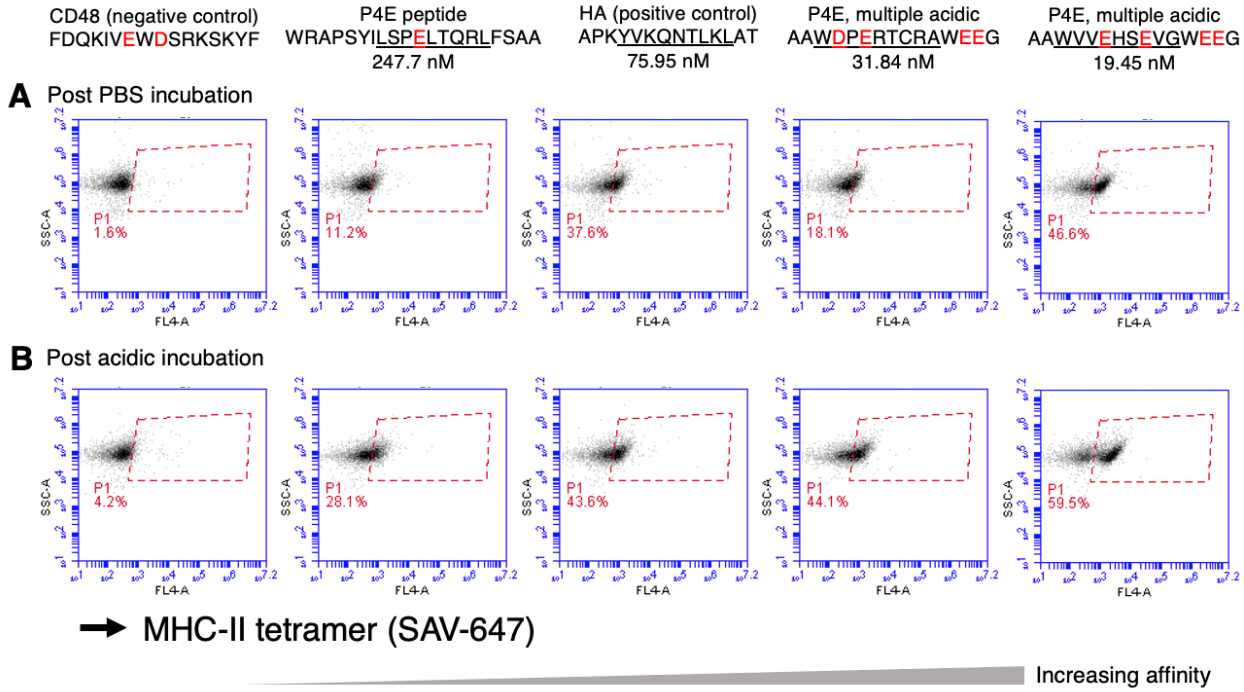


Figure 5. Tests on clonal yeast populations containing acidic residues. HLA-DR401 tetramer staining on individual yeast clones after **A)** incubation in pH 7.2 PBS and **B)** subsequent incubation in acidic saline. CD48-derived peptide was included as a negative control and remaining peptides have previously measured affinities indicated. HA peptide affinity was measured with an additional C-terminal Gly, which is present in the peptide display C-terminal linker. Expected peptide cores are underlined and acidic residues highlighted in red. MHC is tetramerized with streptavidin-AlexaFluor647 (SAV-647).

To explore the need for MHC avidity and alternative modes of selections, we performed selections with monomeric MHC. Performing monomeric selections also provides us a mode to assess if charged AlexaFluor647, which is conjugated to our streptavidin, could be affecting P4 acidic preference. Motifs from monomer selections (**Figure 6**) are largely similar to the tetrameric selections (**Figure 4**), suggesting MHC avidity in tetrameric selection is not necessary and the use of streptavidin-AlexaFluor647 is not biasing peptide binding.

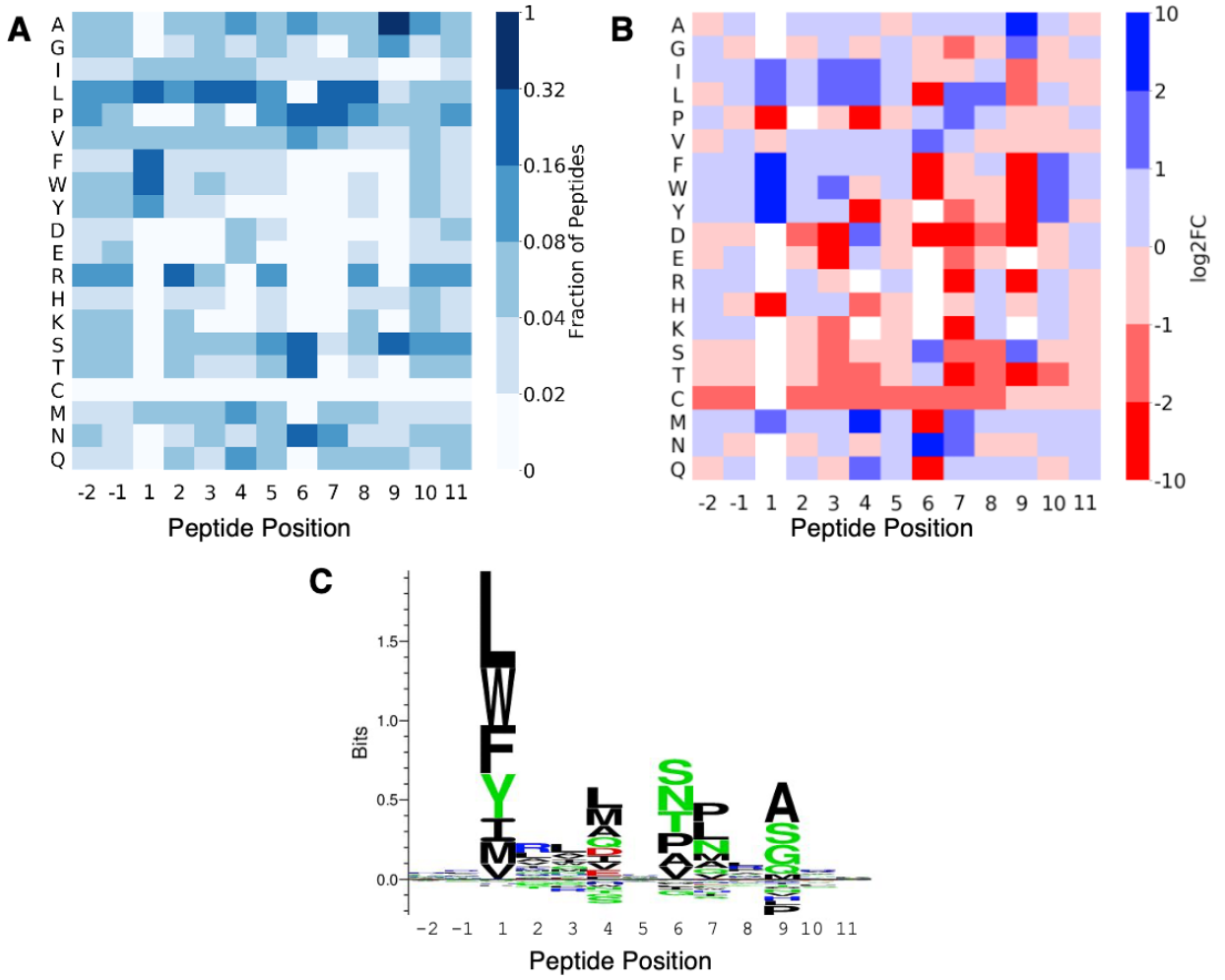


Figure 6. Motifs from peptides in the central register from library selections with HLA-DR401 monomers without fluorescent streptavidin. A) Heatmap highlighting the fraction of peptides with each amino acid at each position. **B)** Heatmap of log2 fold change of positional amino acid frequency compared to unselected library. Amino acids which did not appear in the enriched sequences at a given position are white. **C)** Sequence logo of enriched peptides.

5.4 Assessing additional MHCs, including HLA-DP and HLA-DQ alleles

Given the clear enrichment with HLA-DR401 and extensibility of this approach to other MHC proteins, we expressed HLA-DQ6 (HLA-DQA1*01:02 / DQB1*06:02), HLA-DP401 (HLA-DPA1*01:03 / DPB1*04:01), and HLA-DR15 (HLA-DRA1*01:01 / HLA-DRB1*15:01). These proteins cover alleles expressed by the three canonical MHC-II genes (Neefjes et al., 2011). Further, several of these alleles are implicated in autoimmunity, with HLA-DQ6 and HLA-DR15 comprising a haplotype linked to protection from type 1 diabetes and susceptibility to multiple sclerosis (Karnes et al., 2017; Kaushansky and Ben-Nun, 2014; Pugliese et al., 2016).

For all three alleles, we see clear enrichment, with preferences at anchor residues (**Figure 7**). HLA-DR15 prefers hydrophobic residues at P1; aromatic residues at P4; Asn, Ser, and Gly at P6; and small hydrophobic Leu, Val, Ala at P9 (**Figure 7A**). These preferences closely resemble previously reported motifs (Abelin et al., 2019; Racle et al., 2019).

HLA-DP401 has strong preferences at P1, P6, and P9 for hydrophobic residues, including aromatics at P1 and P6. This motif is similar to those in existing datasets (Abelin et al., 2019; Racle et al., 2019). However, while somewhat present and most visible in the heatmaps (**Figure 7B**), the preference for P4 Glu, Thr, and Ser is less pronounced than in the mass spectrometry comparison datasets (Abelin et al., 2019; Racle et al., 2019).

HLA-DQ6 has clear preference at P1, P4, P6, P7, and P9, alongside greater amino acid preference at non-anchors compared to the other alleles studied here (**Figure 7C**). The anchor preferences and increased preference across the peptide is generally consistent with previously reported peptides from mass spectrometry (Abelin et al., 2019). These data demonstrate the utility of the peptide display approach for extending yeast display assessment of peptide-MHC binding across alleles, especially alleles not previously validated on yeast.

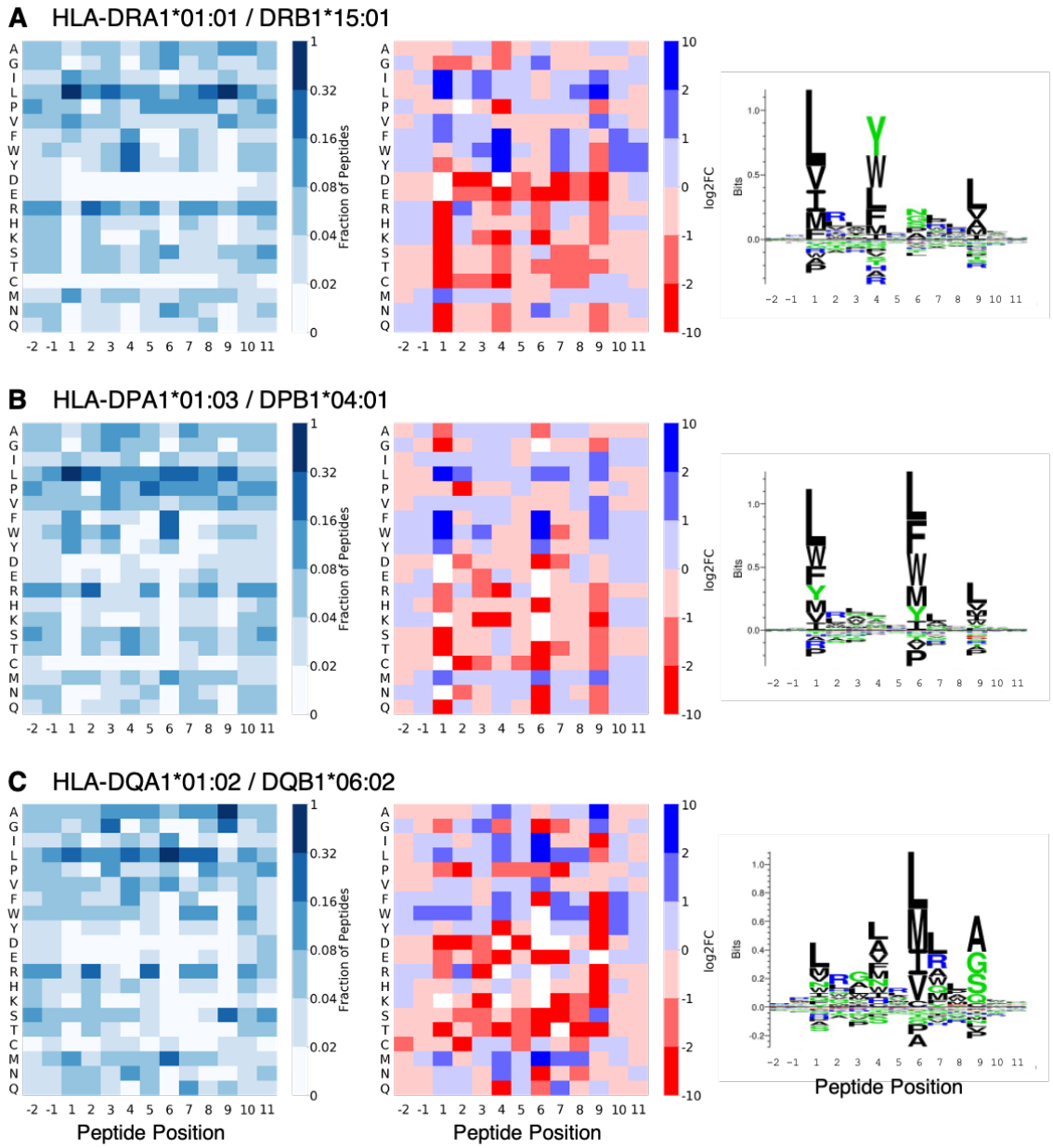


Figure 7. Motifs for peptides enriched for additional MHC-II alleles. Motifs for **A)** HLA-DRB1*15:01 (HLA-DR15), **B)** HLA-DPA1*01:03 / DPB1*04:01 (HLA-DP401), and **C)** HLA-DQA1*01:02 / DQB1*06:02 (HLA-DQ6). **Left)** Heatmaps highlighting the fraction of peptides with each amino acid at each position. **Center** Heatmaps of log2 fold change of positional amino acid frequency compared to unselected library. Amino acids which did not appear in the enriched sequences at a given position are white. **Right)** Sequence logos of enriched peptides.

5.5 Adding exogenous peptide as a competitor

The peptide display approach is highly tunable and parameters can be modified to suit users' downstream needs. As an example, one such tunable parameter is the addition of competitor peptide of known affinity. To examine the effects of adding exogenous peptide on MHC binding to yeast-expressed peptides, we titrated competitor binding and assessed its effects on peptide-MHC-II binding. Our exogenous peptide was used to separately compete with two peptides expressed on the surface of yeast. We expressed $CLIP_{81\text{-}101}$ peptide (LPKPPKPVSKMRMATPLLMQA) on yeast, as well as a synthetic HLA-DP401 binder which was the most frequent peptide in the inferred central binding register from our HLA-DP401 selections (KFFLLLPMCVWCK). The synthetic strong binder matches with the overall motif of enriched peptides, with hydrophobic anchor residues. We assessed binding of HLA-DP401 to each of the yeast-expressed peptides in the presence of varying concentrations of exogenously added $CLIP_{81\text{-}101}$ peptide. We see a clear loss of MHC tetramer signal as we increase the concentration of competitor peptide, with the synthetic strong binder retaining signal at higher concentrations of competitor than the yeast-expressed CLIP peptide. Addition of exogenous peptide such as a weak and pan-MHC-II allele binder like CLIP (Fallang et al., 2009) presents an opportunity to further segregate weak and strong binders in selections and presents another method for tuning stringency of selection experiments.

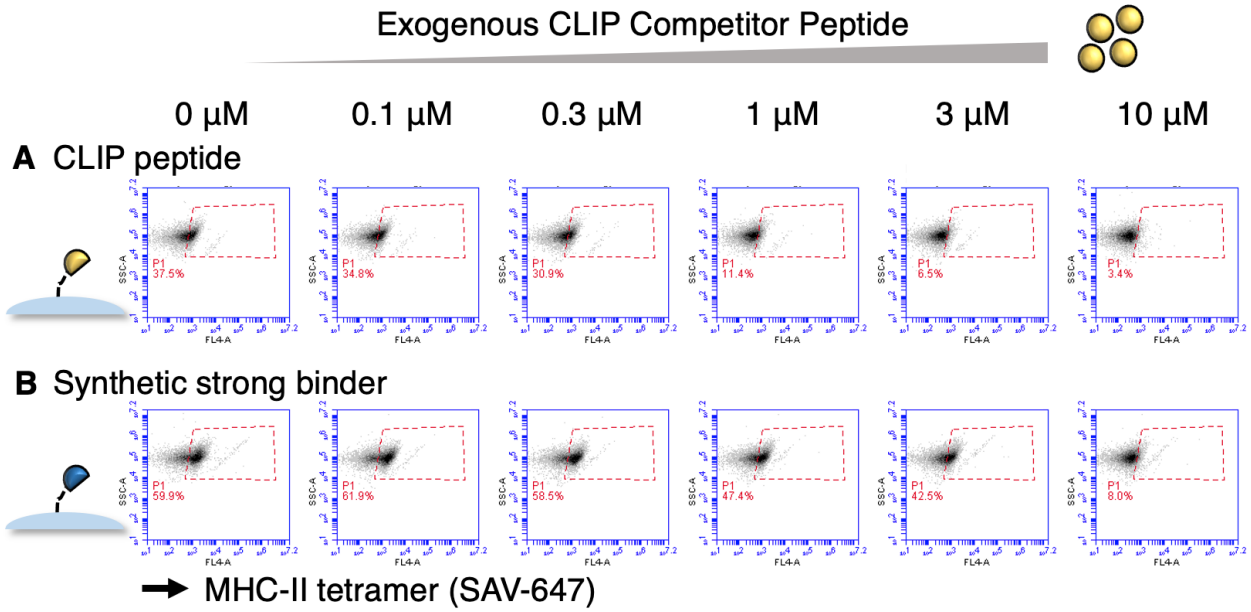


Figure 8. Competition with exogenous competitor peptide. Binding of tetramerized HLA-DP401 to **A)** $CLIP_{81\text{-}101}$ peptide or **B)** synthetic strong binding peptide expressed on yeast, and with exogenous $CLIP_{81\text{-}101}$ peptide added as a competitor, from 0 µM to 10 µM. MHC is tetramerized with streptavidin-AlexaFluor647 (SAV-647).

5.6 Discussion and outlook

We have developed a yeast surface display platform that decouples peptide and MHC expression, enabling users to express MHCs without optimizing them for yeast expression. Libraries of non-structural peptides are expressed on the surface of yeast, and MHCs are expressed recombinantly. MHC proteins can be functionalized such as with fluorescent streptavidin and, when bound to yeast-expressed peptides, detected via flow cytometry, akin to the binding of a cell surface antibody. As an additional benefit, separating the MHC from yeast expression enables users to

make a single library which can be screened against any recombinantly expressed MHC-II protein, so assessing a new MHC does not require generation of a new yeast display library.

In comparison to existing peptide-MHC-II yeast display techniques, the peptide display platform captures additional dynamics. Namely, peptide display requires dissociation of the invariant CLIP peptide from the MHC groove and binding of other peptides. As a result, we capture the interplay of CLIP dissociation and peptide binding to MHC, and potential equilibrium binding and unbinding. While we examined MHC binding at a few timepoints, a thorough assessment of binding over timescales may also yield further insight into the kinetics of peptide-MHC binding. In contrast, the technique described in **Chapter 2** captures only peptide dissociation since peptides of interest begin bound to the MHC-II (Rappazzo et al., 2020).

We demonstrate the utility of this platform by generating datasets for alleles from all three canonical MHC-II genes, HLA-DR, HLA-DP, and HLA-DQ (**Figure 7**). Using a single peptide library and expressing these MHCs recombinantly, we were able to quickly generate quality, large-scale datasets of peptide binders.

Selection strategies using peptide display are versatile and tunable. For instance, we demonstrate the use of tetramerized MHC-II proteins (**Figure 4**) as well as monomeric protein for selections (**Figure 6**), and the use of exogenous competitor peptide for tuning binding stringency (**Figure 8**). The monotonic relationship between measured IC_{50} affinity values and the MHC-tetramer mean fluorescence intensity also suggests that selections could be performed via fluorescence-activated cell sorting, where binning on fluorescence intensity may separate peptides by affinity. Because we do not rely on protein fold for this platform, it is also readily adaptable to other surface display platforms, such as phage display (Bazan et al., 2012), for even greater library sizes. Further, there is great flexibility for MHC generation, including use of bacterial, insect, or other mammalian expression systems. Further, this platform should be readily extensible for peptide-MHC-TCR binding.

We explore constraints of the system, including reliance on 3C protease, HLA-DM, and acidic conditions (**Figure 2** and **Figure 5**). Further exploration of the construct itself can help tune the platform for downstream applications and ensure minimal construct-specific artifacts. For instance, we saw minimal off-target enrichment due to MHCs binding solely to auxiliary material in the construct, such as the epitope tags. This may be due to the binding preferences of the MHCs assessed or steric accessibility or flexibility of the peptide regions. However, other MHCs may have affinity for auxiliary constant peptide flanking sequence and may bind off-target. Since the construct is plasmid-encoded, editing and tuning the construct is readily accessible however. Interestingly, we observed diminished MHC-binding signal when truncating the construct after the peptide and excluding the linker and Myc epitope tag (data not shown), suggesting that there may be length or peptide flanking residue requirements for peptide accessibility. An additional potentially interesting experiment may be to perform selections with an anti-epitope tag antibody and examine if there are any biases in the peptide sequence that allow surface expressed constructs to be more or less accessible. Normalizing our enriched peptides to this epitope-tag enriched library may be preferable to comparing to the unselected library as there may be a layer of peptide accessibility not taken into account here.

The peptide display platform also presents exciting application opportunities. One such application is in autoantigen discovery. MHC-II alleles are highly associated with autoimmune diseases (Karnes et al., 2017), and we could potentially utilize the peptide display approach to determine differential peptide binders of predisposing and protective alleles. We have begun assessing several autoimmunity-associated alleles, including HLA-DR15, HLA-DR401, and HLA-DQ6, which are associated with susceptibility or protection to multiple sclerosis and type 1 diabetes. Future work will extend this set of disease-associated alleles and compare binders between alleles through computational prediction algorithms trained on data from random libraries, as well as direct assessment of binding utilizing defined libraries covering human proteome-derived peptides.

5.7 Methods

Vector formatting for peptide display

pCT302 vector was adapted to fuse a peptide sequence to Aga2. We left unchanged the N-terminus of the construct, in which Aga2 is connected via a linker to an HA epitope tag (YPYDVPDYA). We also leave unchanged the subsequent C-terminal linker sequence (LQASGGGGSGGGGSGGGGSAS). The C-terminus of this linker is connected to the N-terminus of our peptide sequence of interest. To assess frameshift mutations in our peptide sequence, we link the peptide to a Myc epitope tag (EQKLISEEDL) via a short linker (GGSGG). The Myc tag is proceeded by a stop codon.

Peptide display parameter exploration experiments

To optimize peptide display reaction conditions, we generated reaction mixtures containing MHC-II tetramers, HLA-DM, and 3C protease in PBS. Yeast were washed with PBS then resuspended in the reaction mixture made in PBS. Yeast were incubated for 45 minutes at room temperature, rotating, then washed into FACS buffer (0.5% BSA and 2 mM EDTA in 1x PBS) and assessed via flow cytometry (Accuri C6 flow cytometer, BD Biosciences; Franklin Lakes, New Jersey). In the reaction mixture for MHC-II titration, we included 1 μM 3C protease and 3.5 μM of HLA-DM. In the reaction mixture for HLA-DM titration, we included 1 μM 3C protease and 75 nM MHC tetramer. And in the reaction mixture for 3C protease titration, we included 0.5 μM HLA-DM and 75 nM MHC tetramer.

For testing individual peptides with measured affinity values (**Figure 5**), similar conditions were utilized: 75 nM HLA-DR401 tetramer, 1 μM 3C, and 0.5 μM HLA-DM. Yeast were also subsequently incubated in a nine-fold excess of acidic saline and sampled after 45 minutes. Experiments assessing the impacts of CLIP as a competitor peptide (**Figure 8**), utilized the same concentrations, with 75 nM HLA-DP401 tetramer, 1 μM 3C, 0.5 μM HLA-DM, in addition to 10 μM TCEP and 20-minute pre-incubation of yeast with TCEP-containing PBS before incubating with the reaction mixture. Yeast were sampled after a 45-minute room temperature incubation.

MHC tetramers were generated using streptavidin-AlexaFluor647 (SAV-647; made in-house), with a 5-fold molar ratio of MHC to SAV-647, which has four binding sites. Tetramer concentration is given as the concentration of SAV-647, representing the concentration of tetrameric species.

Randomized library generation
A 13mer randomized peptide library was generated by performing polymerase chain reaction (PCR) with a degenerate primer encoding NNK codons (N = A, C, T, or G; K = G or T). We utilized a stop codon-containing template for PCRs to decrease the likelihood of contamination with a strong-binding construct. Libraries were generated by electroporating RJY100 yeast (Van Deventer et al., 2015) with a 5:1 mass ratio of randomized peptide product and linearized pCT302 vector.

Recombinant protein expression
α and β chains from HLA-DR401, HLA-DR15, HLA-DQ6, HLA-DP401, and HLA-DM were cloned into pAcGP67a vectors for expression in insect cells. The ectodomains of each chain were formatted with a poly-histidine purification site and encoded on separate plasmids. The α chains contained an AviTag biotinylation site (GLNDIFEAQKIEWHE). Excepting HLA-DM, the N-termini of β chains were linked to $CLIP_{81-101}$ peptide via flexible Gly-Ser linker containing a 3C protease site (LEVLFQGP). Plasmid DNA was transfected into SF9 insect cells with BestBac 2.0 DNA (Expression Systems; Davis, CA) using Cellfectin II reagent (Thermo Fisher; Waltham, MA). Viruses for α and β chains were co-titrated to determine an optimal ratio for heterodimer formation, co-transduced into High Five (Hi5) insect cells (Thermo Fisher), and incubated for 48-72 hours. Proteins were secreted by Hi5s and purified from pre-conditioned media supernatant using Ni-NTA resin. Proteins were further purified using a S200 increase column on an AKTAPURE FPLC (GE Healthcare; Chicago, IL).

Library selections
Yeast libraries were bottlenecked to 5 x 10^6 yeast to increase coverage of enriched yeast by sequencing and for easier comparison across iterations of selections. An overrepresentation of these yeast were input into Round 1, with 5 x 10^7 yeast as input, measured by optical density. Rounds 2 and 3 used 2.5 x 10^7 – 5 x 10^7 input yeast.

To perform tetrameric selections, yeast were first washed into pH 7.2 PBS with 10 μM TCEP and incubated for 20 minutes while making the reaction mixture. The reaction mixture contained MHC tetramers, generated by pre-incubating 75 nM SAV-647 and 375 nM biotinylated MHC. Tetramers are added to the appropriate volume of pH 7.2 PBS with 10 μM TCEP. This is followed by the addition of 3C protease to 1 μM, then the addition of HLA-DM. Each batch of HLA-DM was titrated for functional effect with measured concentrations ranging from 0.5 - 5 μM utilized; lower functional HLA-DM likely results from the presence of some contaminating unfolded protein, single chains, or αα or ββ homodimers. TCEP was added to prevent disulfide bonds which could diminish the enrichment of peptides containing cysteine. Yeast were spun down and resuspended in reaction mixture at 2 x 10^8 yeast/mL. This reaction was incubated at room temperature for 45 minutes, rotating, covered to decrease risk of photobleaching. Then yeast in the reaction mixture were placed on ice, and a subsample taken for assessment via flow cytometry, and the remaining yeast were moved to a nine-fold excess of pH 5 citric acid saline buffer (20 mM citric acid, 150 mM NaCl). Yeast were incubated in the acid saline at 4°C for 45 minutes, rotating. Yeast were then washed into FACS buffer, again at 2 x 10^8 yeast/mL, and ∝-AlexaFluor647 beads (Miltenyi Biotec, Bergisch Gladbach, Germany) equal to 10% of the FACS buffer volume were added. Yeast were incubated with beads at 4°C, rotating for 15 minutes. Yeast were then pelleted, resuspended in 5 mL FACS buffer, and run over a Miltenyi LS column (Miltenyi Biotec), washed three times

with 3 mL FACS buffer, and eluted with 5 mL FACS buffer. The elution contained the yeast which bound to the MHC. Selected yeast were then washed into SDCAA media and grown to confluence. Yeast were sub-cultured in SGCAA media at OD_{600} = 1 for 48-72 hours at 20°C before each experiment (Chao et al., 2006).

Monomeric selections were performed similarly to tetrameric selections, with a few modifications. Monomeric MHC-II was added in the reaction mixture, absent SAV-647. 300 nM MHC-II was added, equivalent to the amount of MHC-II which binds to the four binding sites of 75 nM SAV-647 in tetrameric selections. Streptavidin beads (Miltenyi Biotec) were added in place of ∝-AlexaFluor647 beads, and selections proceeded in the same manner.

Deep sequencing and data processing
Peptide identities were determined from enriched yeast following selections. Plasmid DNA was extracted from 5 x 10^7 input yeast via the Zymoprep Yeast Miniprep Kit (Zymo Research; Irvine, CA) following manufacturer's instructions. Amplicons were generated by two rounds of PCR to add inline i5 and i7 paired-end handles and inline sequencing barcodes, unique to rounds multiplexed in a single sample. PCR primers were designed to capture the 13mer peptide sequence, priming off of adjacent constant sequences. Amplicons were sequenced with an Illumina MiSeq (Illumina Incorporated; San Diego, CA) at the MIT BioMicroCenter using 300 nucleotide kits.

Paired-end reads were assembled using PandaSeq (Masella et al., 2012). Given the large space of possible NNK-encoded 13mer peptides, single nucleotide variants are likely to be a result of PCR or sequencing errors. Using CDHIT (Fu et al., 2012), single nucleotide-variants were combined and the more frequent sequence chosen as the consensus sequence.

Peptide binding register inference and motif visualization
To infer the register in which the enriched 13mer peptides bound, we utilized a position weight matrix (PWM) method, similar to that in **Chapter 4**. Enriched peptides from round 3 of selection were one-hot encoded and padded on the C-terminus of the peptide. We utilized a 9mer PWM and eight registers and one 'trash' register.

At the start, peptides were randomly assigned to registers and a 9mer PWM was generated. The contribution of a given peptide to the PWM was weighted by its read count. Over successive iterations, peptides were assigned to new registers and the PWM updated, excluding peptides in the trash cluster. Assignment to each register is stochastic, but biased, such that registers which match the PWM were favored. At each assignment, if the peptide was assigned to a non-trash register, we first took out that peptide from the PWM. The PWM then defined an energy value for each register option, which was used to generate a Boltzmann distribution from which the updated register shift was sampled. Stochasticity was decreased over time by raising the inverse temperature linearly from 0.05 to 1 over 20 iterations, simulating cooling (Andreatta et al., 2017). The final iteration was deterministic, where the distribution concentrated solely on the optimal register option.

From selections with HLA-DR401 (tetramers), HLA-DR401 (monomers), HLA-DP401, HLA-DR15, HLA-DQ6, we identify 111,203 / 32,885 / 89,746 / 46,276 / 27,231 enriched peptides in

our sequencing after Round 3, respectively. Of these, from register-inference, we identify 62,421 / 18,845 / 56,977 / 28,167 / 11,584 peptides in native registers, respectively.

Heatmap and sequence logo visualizations were generated for peptides which were inferred to bind in the central register, with two peptide flanking residues on either side of the 9mer core. Heatmaps were generated using a custom script to show amino acid frequency or log2 fold change over the unselected library at each position. Amino acids which did not appear at a given position in the round 3 library appear as white spaces in the log2 fold change heatmap. Sequence logos were generated using Seq2Logo (Thomsen and Nielsen, 2012), with default settings, except background frequencies were generated from the unselected library. In heatmaps and sequence logos visualizations, peptides were not weighted by read counts.

5.8 Acknowledgements
The register inference algorithm utilized here was generated by Zheng Dai in David Gifford's group (as in **Chapter 4**). Marcos Labrado assisted with amplicon generation for several deep sequencing runs. Michael E. Birnbaum contributed to the conceptualization of this work, and provided guidance on plasmid and experimental design and analysis.

CHAPTER 6: CHARACTERIZING THE PEPTIDE REPERTOIRE OF HLA-E AND NATURAL KILLER CELL RECEPTORS VIA YEAST DISPLAY

Ning (Alexa) Guan contributed significantly to work in this chapter. A full acknowledgements statement is included as ***Chapter 6.10****.*

Abstract

HLA-E (Human Leukocyte Antigen E) is a non-classical major histocompatibility complex (MHC) class I (MHC-I) protein involved in innate and adaptive immune recognition. Expressed widely on nucleated cells, HLA-E canonically presents leader peptides from other MHC-I molecules, to inhibit natural killer (NK) cell killing of healthy cells. Recent studies show HLA-E can present diverse peptides to NK cells and T cells, but the HLA-E peptide repertoire is understudied and only a limited number of peptide ligands have been identified. Here we characterize the diversity of the HLA-E peptide repertoire, performing selections on a yeast display library of 100 million peptides using cognate receptors CD94/NKG2A and CD94/NKG2C. We identify ~500 high-confidence unique peptides that bind both HLA-E and CD94/NKG2A or CD94/NKG2C, characterizing preferences of both HLA-E- and NK receptor-facing residues. We find that the inhibitory NKG2A and activating NKG2C receptors have similar peptide preferences, but NKG2A is more permissive across different peptide positions likely due to its higher affinity to HLA-E. Utilizing these data, we train prediction algorithms and identify human and CMV proteome-derived HLA-E and NK receptor ligands, in addition to non-canonical HLA-E-binders, with potential implications for vaccine and therapeutic design.

6.1 Introduction

Human leukocyte antigen E (HLA-E) is a highly conserved, non-canonical, class Ib MHC. In contrast to the highly diverse canonical MHCs, HLA-E has only two alleles (HLA-E*01:01 and HLA-E*01:03) found in worldwide populations, corresponding to one amino acid difference (107 Arg/Gly) outside of the peptide binding groove (Zheng et al., 2015). HLA-E modulates natural killer cell (NK cell) activity by binding to both inhibitory and activating receptors on NK cells (Hammer et al., 2018). HLA-E was previously thought to exclusively present highly conserved VL9 signal peptides of other class I MHCs (MHC-I) for recognition by heterodimeric NK receptors CD94/NKG2A and CD94/NKG2C (CD94/NKG2x) (Carrillo-Bustamante et al., 2016). When pathogens evade antigen presentation by downregulating class I MHC expression, a lack of peptides loading into HLA-E decreases HLA-E surface expression, removing the mainly inhibitory signal to NK cells, allowing NK cell cytotoxic activity. However, recent work indicates HLA-E's peptide repertoire is more diverse than previously thought (Hansen et al., 2016; Lampen et al., 2013), and can bind pathogen-derived peptides with diverse sequences (Walters et al., 2018). Moreover, in certain cases, peptide-HLA-E complexes (pHLA-E) can be recognized by T cells, via T cell receptors (TCRs) and NK cell receptors, in addition to NK cells (Kulski et al., 2019; Mathewson et al., 2021; Pietra et al., 2001). Ongoing studies show that HLA-E acts as a bridge between innate and adaptive immunity (Kraemer et al., 2014).

HLA-E's cognate CD94/NKG2x receptors are expressed by NK cells and certain T cells. NK cells, important in antiviral and anti-tumor responses, use these and other receptors to modulate their cytotoxic activity (Stabile et al., 2017). CD94/NKG2A is a high affinity inhibitory receptor and CD94/NKG2C a lower affinity activating receptor (Brooks et al., 1997). For a given peptide,

CD94/NKG2C generally binds to peptide-HLA-E with a 6-fold lower affinity than CD94/NKG2A (Kaiser et al., 2008). NKG2A is expressed on 20-80% of NK cells (Mahaweni et al., 2018). $NKG2C^+$ NK cell population expands upon CMV infection, and is associated with a lower viral setpoint in HIV infection (Cannon et al., 2010; Gumá et al., 2004; Ma et al., 2017). Co-expression of NKG2A and NKG2C is generally observed in only 1-2% of NK cells (Béziat et al., 2012; Ma et al., 2017) though is highly variable (Angelini et al., 2011). Besides NK cell expression, CD94/NKG2C and CD94/NKG2A are expressed by significant fractions of tumor-infiltrating $CD8^+$ T cells (Mathewson et al., 2021; van Montfoort et al., 2018). NKG2A is a novel immune checkpoint, and anti-NKG2A monoclonal antibodies promote anti-tumor immunity in mouse models and a Phase II head and neck cancer clinical trial (van Hall et al., 2019). In addition to binding CD94/NKG2x receptors, HLA-E, or it's MHC-E homologues, can alternatively be recognized by TCRs on $CD8^+$ T cells. Vaccine strategies against SIV (simian homolog of HIV), HBV, and typhoid, including CMV-vector-based vaccines, elicit MHC-E restricted $CD8^+$ T cell responses (Burwitz et al., 2020; Hansen et al., 2013, 2016; Yang et al., 2021). Overall, HLA-E-CD94/NKG2x interactions modulate anti-viral and anti-tumor immunity of NK cells and T cells.

Peptide-HLA-E can activate or inhibit NK cells and certain $CD8^+$ T cells. To evaluate HLA-E's and CD94/NKG2x's potential in vaccine and cancer treatment, one needs to first understand the HLA-E and CD94/NKG2x peptide repertoire. However, a limited number of HLA-E peptide binders is currently known (~700 total in IEDB) (Vita et al., 2019), which likely do not represent a systematic assessment of possible HLA-E binders. Other studies advance our knowledge about HLA-E-binding preferences in a broad sense, but without identifying individual binders from pooled screens (Ruibal et al., 2020). Current state-of-art MHC motif predictors, NetMHC (Jurtz et al., 2017) and MHCflurry (O'Donnell et al., 2018), are likely limited in HLA-E prediction performance due to this scarcity of available training data, although they have not been extensively benchmarked on HLA-E. More research is necessary to systematically identify both what peptides can be presented by HLA-E and bind to CD94/NKG2x receptors.

Yeast display of peptide-MHC molecules represents a means of assessing large collections of peptides for MHC function. For example, yeast display libraries of MHC linked to diverse peptides, and subsequent selection by binding to a receptor such as a TCR, allow identification of peptide ligands for both the MHC and TCR (Birnbaum et al., 2014, 2017; Fernandes et al., 2020; Gee et al., 2018b; Sibener et al., 2018). A yeast display library allows experimental study of large unbiased libraries of $\sim 10^8$ peptides, although the larger possible peptide search space (20^9 for a 9 amino acid library) calls for computational methods such as neural networks to predict peptide binding. Yeast library generation followed by next-generation sequencing (NGS) of the selected library can be used to identify a large quantity of peptide binders, and improve computational peptide prediction (Rappazzo et al., 2020).

Here, we analyze the diversity of the HLA-E peptide repertoire via yeast display, deep sequencing, and computational analysis. Yeast presenting HLA-E linked to library of 100 million unique peptides were selected for binding to CD94/NKG2A and CD94/NKG2C. The library hits bind both HLA-E and CD94/NKG2x because CD94/NKG2x recognize the peptide-HLA-E complex and this recognition is peptide dependent (Valés-Gómez et al., 1999). Deep sequencing of the HLA-E selection libraries identified ~500 high-confidence unique peptides that bind to HLA-E and CD94/NKG2A or CD94/NKG2C. With these data, we developed computational algorithms to

predict peptide binders to HLA-E and NK receptors. To validate a select set of potential HLA-E peptide binders, we performed biophysical validation, including peptide stability assays and surface plasmon resonance (SPR), and have ongoing work examining peptide effects on NK cell function. This improved understanding of the HLA-E peptide repertoire helps identify peptides from the human and pathogen proteomes that bind to HLA-E and CD94/NKG2x, for potential vaccine design and cancer therapeutics.

6.2 Design and validation of HLA-E for yeast display

To study the diversity of HLA-E peptide repertoire, we utilized a versatile MHC yeast surface display system (Gee et al., 2018b), in which peptide, Beta-2-microglobulin (β2M), and MHC heavy chain are covalently linked. First, we designed a peptide-β2M-HLA-E single chain trimer (pHLA-E SCT) construct with two epitope tags: an HA tag near the C-terminus of the construct and a Myc tag in the peptide-β2M linker (**Figure 1A, 1B**). To accommodate the linker connecting the C-terminus of the peptide to the N-terminus of β2M, we modified HLA-E to include a tyrosine to alanine mutation at position 84 (Y84A). This mutation could affect the binding preferences of the MHC F-pocket, but has been shown to largely recapitulate the endogenous binding properties of wild type class I MHCs (Gee et al., 2018b). The HLA-E single chain trimer construct is well-expressed on the yeast surface, as confirmed via epitope tag staining and anti-HLA-E antibody (clone 3D12) staining (**Figure 1C**).

Since epitope staining assesses construct expression rather than protein fold (Birnbaum et al., 2014; Fernandes et al., 2020; Rappazzo et al., 2020), we next tested pHLA-E fold by staining with its cognate receptors, CD94/NKG2A and CD94/NKG2C. Using a canonical peptide VMAPRTLFL, pHLA-E stained with CD94/NKG2A tetramer showed a 20.9% positive population, indicating some pHLA-E Y84A was correctly folded for receptor recognition (**Figure 1D**).

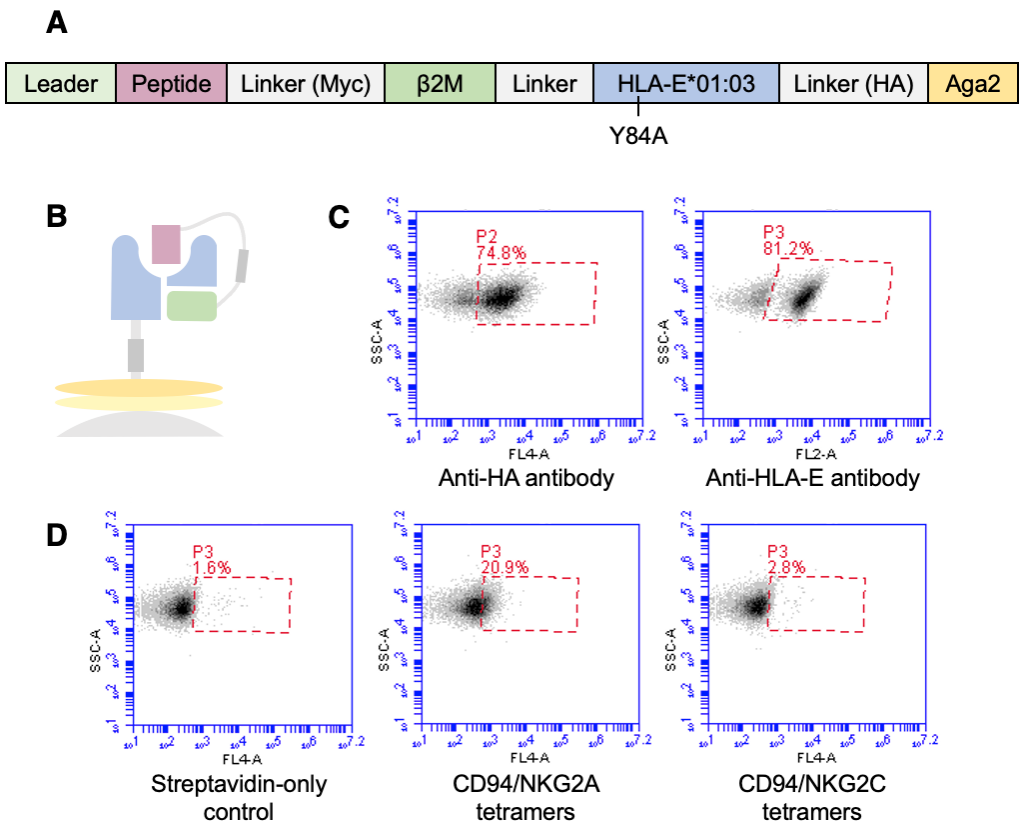


Figure 1. Design and validation of HLA-E on yeast. Representation of HLA-E **A)** sequence and **B)** on the surface of yeast. **C)** Validation staining of HLA-E with VMAPRTLFL peptide by anti-HA-647 and anti-HLA-E-PE antibodies and with **D)** CD94/NKG2A and CD94/NKG2C tetramers made with streptavidin-AlexaFluor647.

6.3 Selection of HLA-E library

Next, we created a 9mer peptide library using the pHLA-E construct, where each peptide position can encode any of the 20 amino acids. To identify peptide binders to HLA-E, we considered library selection via anti-HLA-E antibodies, anti-β2M antibodies, or CD94/NKG2x receptors. Binding of anti-HLA-E and anti-β2M antibodies were not selective for HLA-E constructs containing known HLA-E-binding peptides as compared to negative control peptides (**Figure 2**). Thus, we turned to CD94/NKG2x receptors for library selection of peptides that bind both HLA-E and CD94/NKG2A or CD94/NKG2C.

We performed iterative selections on the randomized peptide library displayed by HLA-E using CD94/NKG2A or CD94/NKG2C. Specifically, yeast that bound to CD94/NKG2x-coated magnetic beads (selection Rounds 1-3) or CD94/NKG2x tetramers (followed by anti-tetramer magnetic beads; Round 4) were enriched using magnetic columns (**Figure 3A**). For Round 4, both iterative selection and cross-selection were performed using CD94/NKG2x tetramers, which were more stringent than receptor-coated beads due to tetramers' lower avidity (**Figure 3B**). Cross-selection was performed to study the peptide motif overlap between NKG2A and NKG2C, as in the ability of NKG2C to bind NKG2A binders, and vice versa. CD94/NKG2A showed similar tetramer staining for yeast previously selected by CD94/NKG2A or CD94/NKG2C for 3 rounds, indicating that the peptide binding motifs of CD94/NKG2A and CD94/NKG2C may have significant similarities (**Figure 3B**).

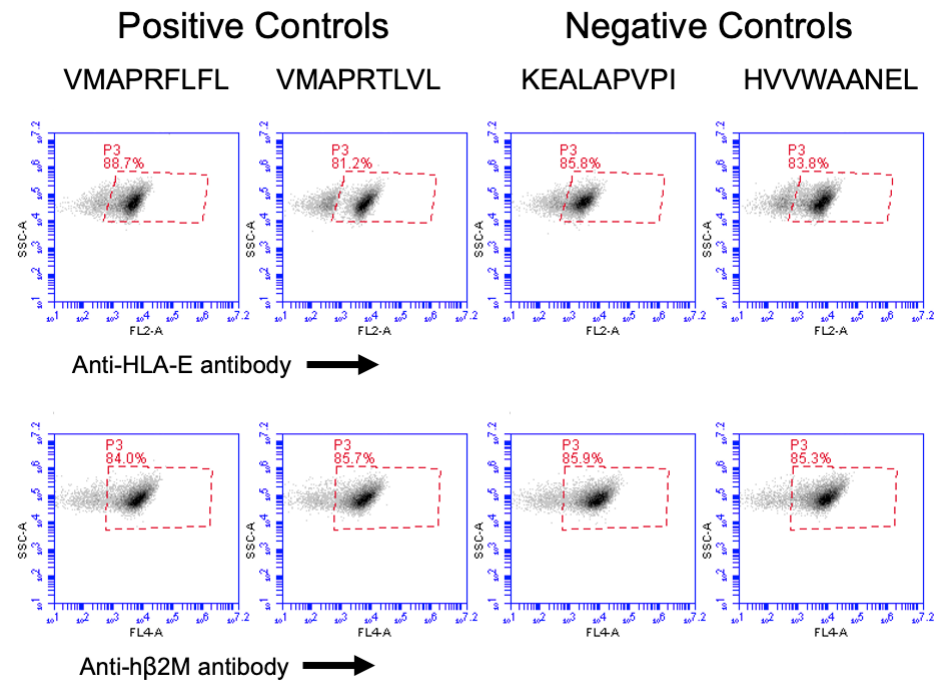


Figure 2. Non-conformation specificity of anti-HLA-E 3D12 antibody and anti-Beta-2-Microglobulin antibodies. Staining of canonical HLA-E binding peptides and negative control peptides expressed as single chain trimers on yeast.

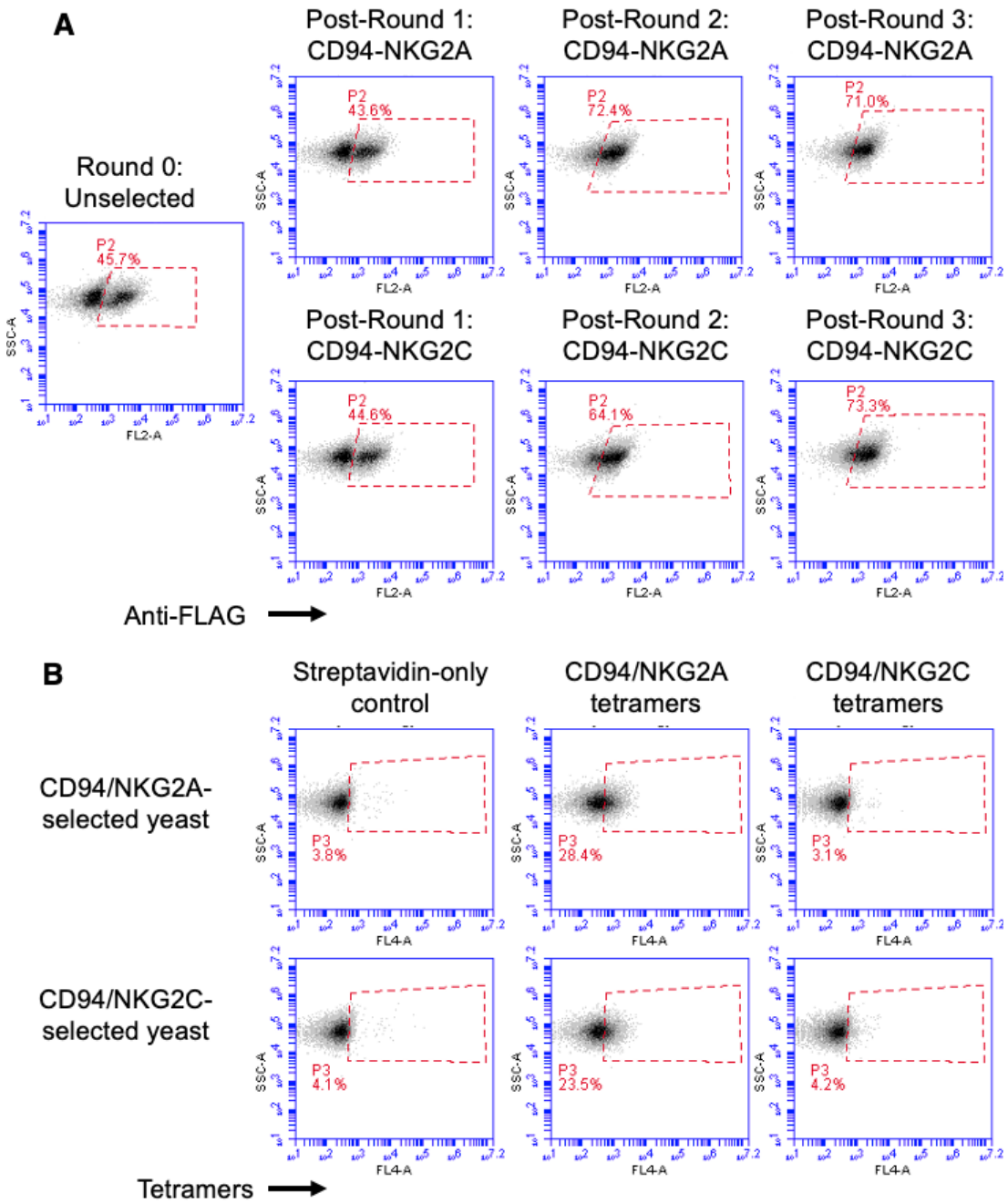


Figure 3. Library dynamics over rounds of selection with CD94/NKG2A or CD94/NKG2C. A) Prior to selections, a sampling of yeast was assessed for FLAG epitope tag expression, which increased over rounds of selection as (gate drawn daily on unstained yeast). **B)** Staining of yeast during Round 4 with CD94/NKG2A or CD94/NKG2C tetramers, on libraries previously selected with CD94/NKG2A or CD94/NKG2C in Rounds 1-3, including negative streptavidin-only control.

6.4 Analysis of library-enriched peptides

Following selections, we performed deep sequencing on library-enriched yeast to determine the identities of the peptides expressed in each construct. We identify HLA-E 9mer library hits with significant diversity compared to the canonical MHCI leader-like ligands (Lampen 2013, Walters 2018). Specifically, deep sequencing showed ~10^4 unique peptides per sample after the third round, and ~5 x10^3 unique peptides per sample after the fourth round of selection, although approximately 500 peptides are dominant among the enriched peptides, each with read counts >100 (**Figure 4**).

A clear motif of enriched peptides emerged in the data after the third round of selection (**Figure 5**, **Figure 6**). **Figure 5** represents the enriched peptides as a heatmap and sequence logos (Crooks et al., 2004), both weighted by read counts from the deep sequencing data, to reflect peptide frequency. The motif, highly conserved between CD94/NKG2A- and CD94/NKG2C-selected libraries, contained significant diversity on the peptide N-terminal side. P2 was not restricted to Met, but rather accommodated most hydrophobic residues. A strong preference for P3 Pro also emerged, indicating P3 may be a secondary HLA-E anchor residue. Conserved residues were consistent with primary anchors at P2 and P9, secondary anchors at P7 and possibly P3 and P6 (Miller et al., 2003; Stevens et al., 2001). We observed generally more conservation on the peptide C-terminus (P5 through P9), corresponding to CD94/NKG2x binding to pHLA-E on the C-terminal side of the peptide. CD94/NKG2x-contact P5 had strong preference for Arg, and P8 for Trp, Phe, or Leu. CD94/NKG2A was more permissive than CD94/NKG2C, potentially due to CD94/NKG2A's higher affinity for HLA-E, consistent with more diversity for the CD94/NKG2A motif, even at positions other than CD94/NKG2x contact residues. While the motifs after Round 4 selection with CD94/NKG2x tetramers were consistent with Round 3 motifs, CD94/NKG2A tetramer selection led to a less diverse motif likely containing higher affinity binders (**Figure 6**).

In addition to the 9mer library, we also generated and performed selections on a 10mer peptide library. Unfortunately, contamination of the CD94/NKG2A-selelcted 10mer library with the 9mer library prohibited our analysis, but we see clear enrichment in the CD94/NKG2C library. Interestingly, the motif in the N-terminal 9 amino acids largely resembles the motif in the 9mer library. The C-terminal amino acid showed less preference than the 9mer library PΩ position, consistent with the 10mer peptides binding as 9mer peptides, with the tenth residue acting as part of the peptide linker, suggesting that 9mer peptides preferentially bind compared to 10mer peptides (**Figure 7**).

Overall, from our library data, we observe that the HLA-E and CD94/NKG2x peptide repertoire is highly diverse, and we identify binders to elucidate the previously uncharacterized repertoire motif. Inhibitory CD94/NKG2A and activating CD94/NKG2C receptors have similar peptide preferences, with CD94/NKG2A more permissive across multiple peptide positions.

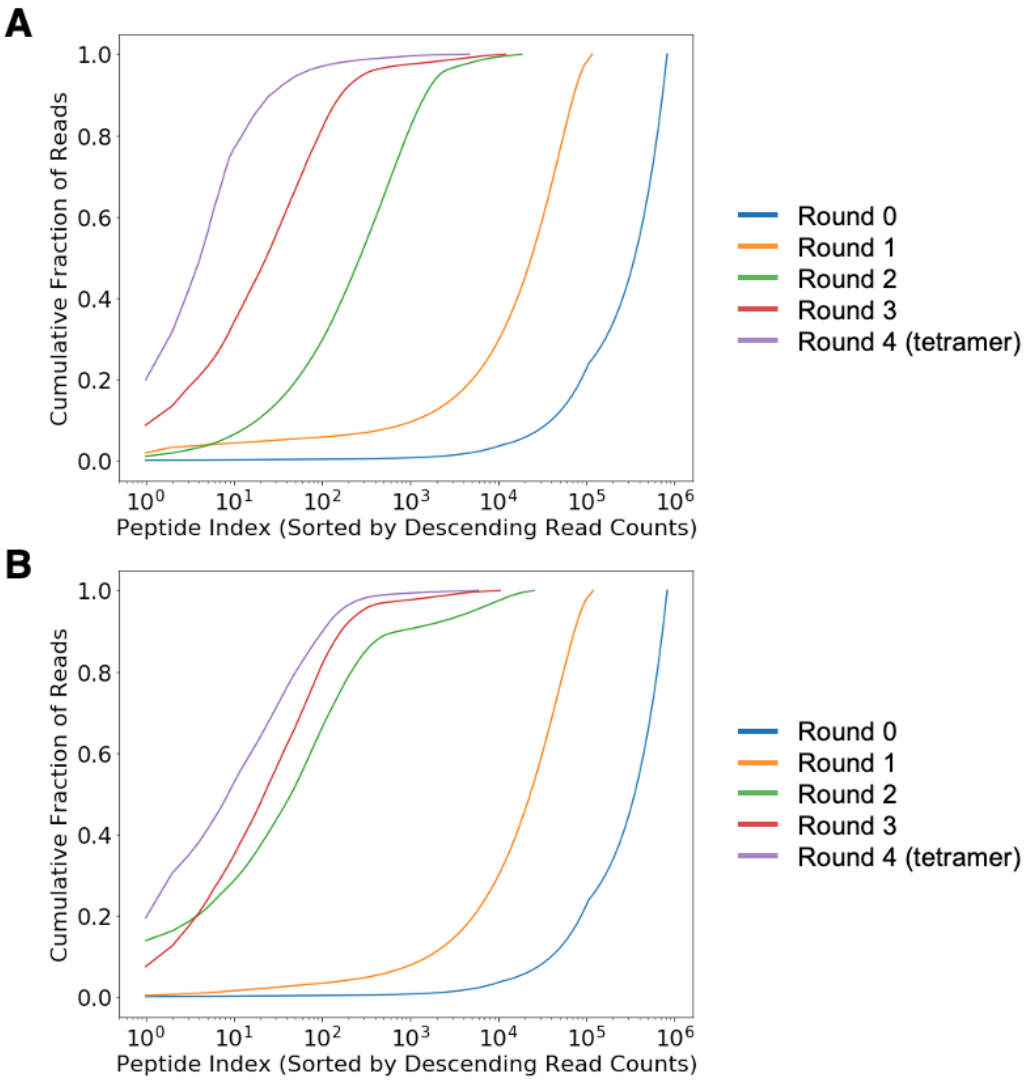


Figure 4. Peptide frequency in deep sequencing data. Cumulative fraction of read counts for peptides selected with **A)** CD94/NKG2A or **B)** CD94/NKG2C.

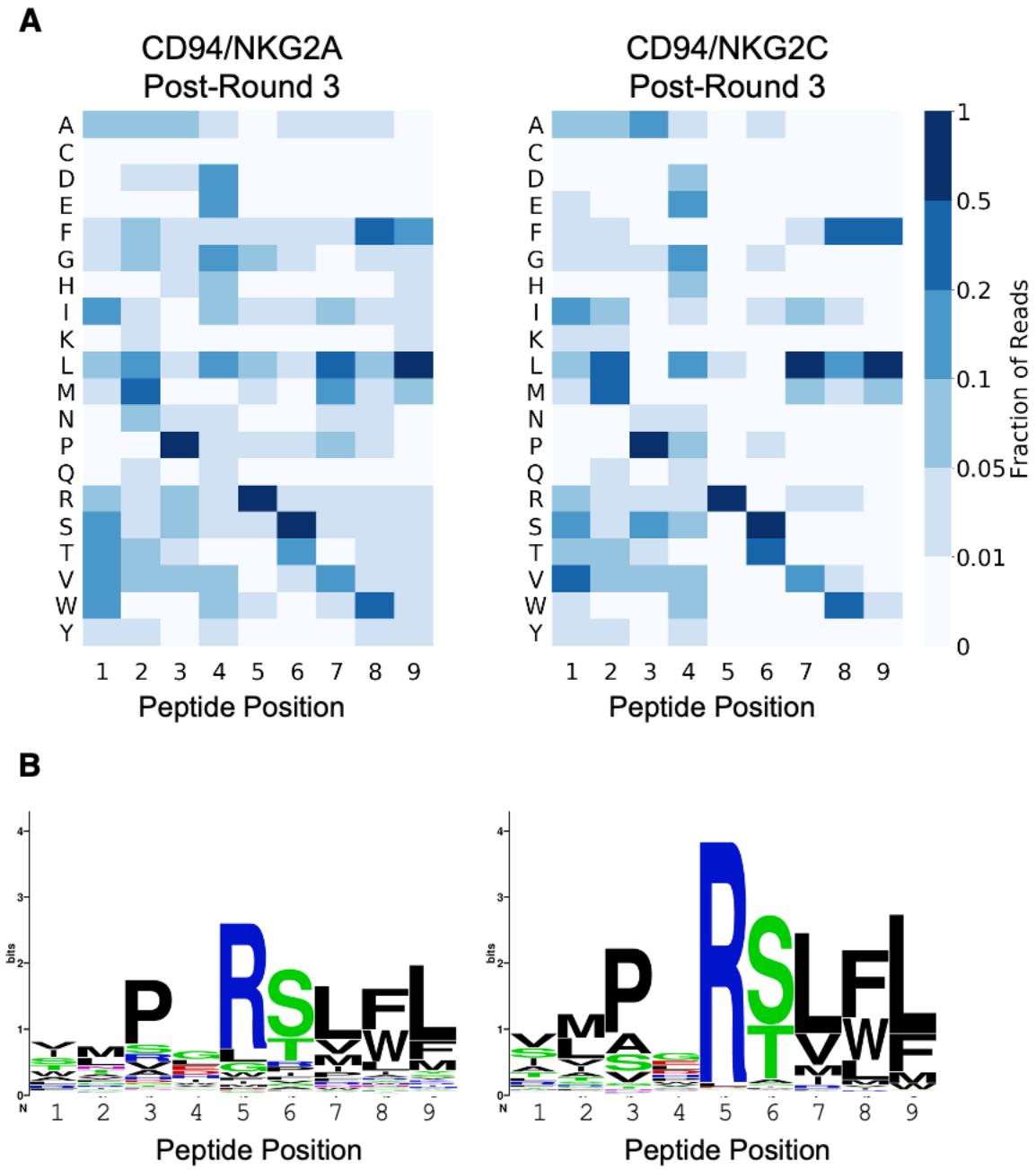


Figure 5. Peptide repertoire of HLA-E with CD94/NKG2A or CD94/NKG2C. **A)** Heatmaps showing peptide positional amino acid preferences for binding HLA-E and CD94/NKG2A or CD94/NKG2C with peptides weighted by read count. **B)** Sequence logos present the same data, also weighted by read count, generated with WebLogo.

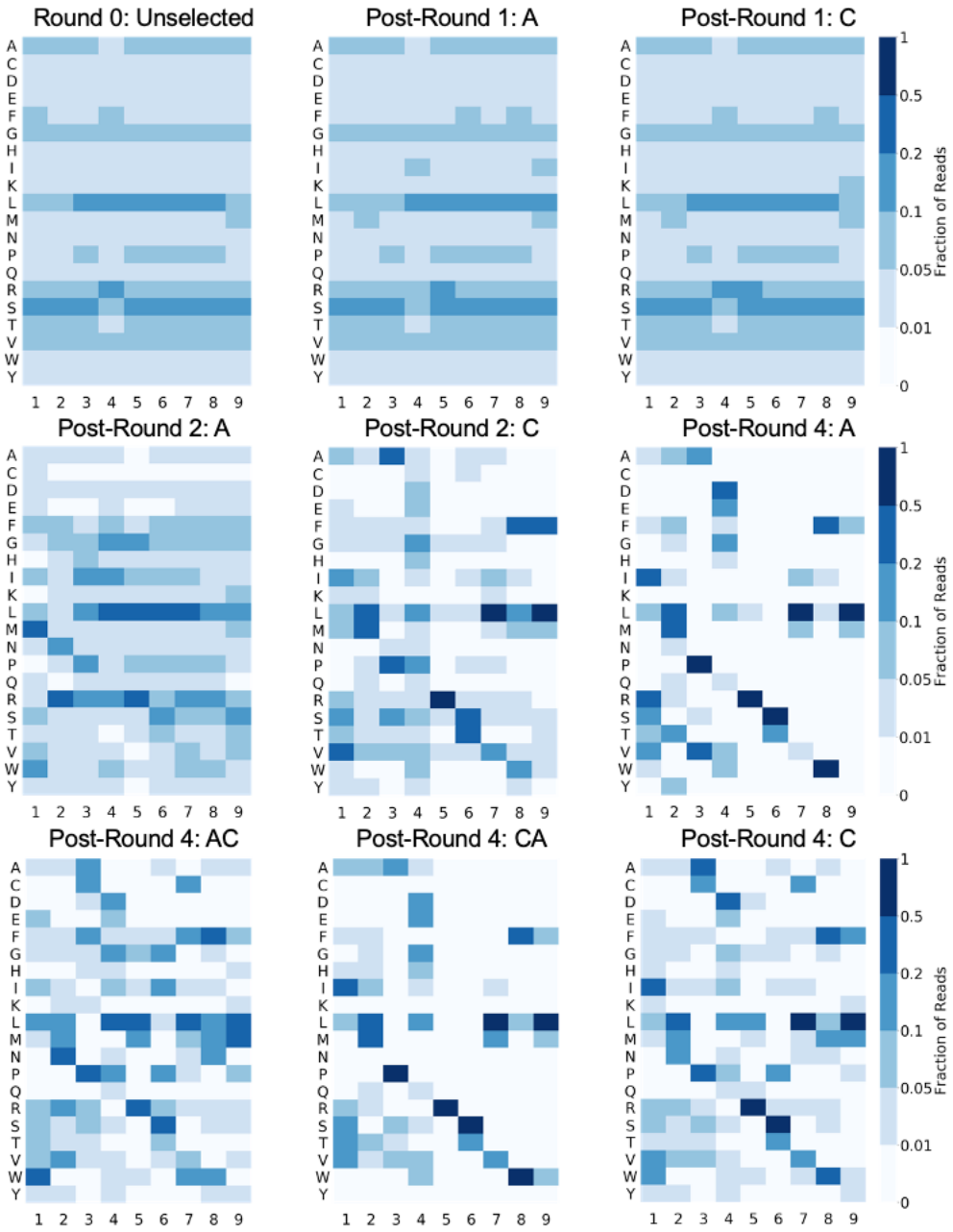


Figure 6. Heatmaps for other rounds of selection. A: Selected with CD94/NKG2A; C: Selected with CD94/NKG2C; AC: yeast previously selected with CD94/NKG2A, cross-selected with CD94/NKG2C in this round; CA: yeast previously selected with CD94/NKG2C, cross-selected with CD94/NKG2A in this round.

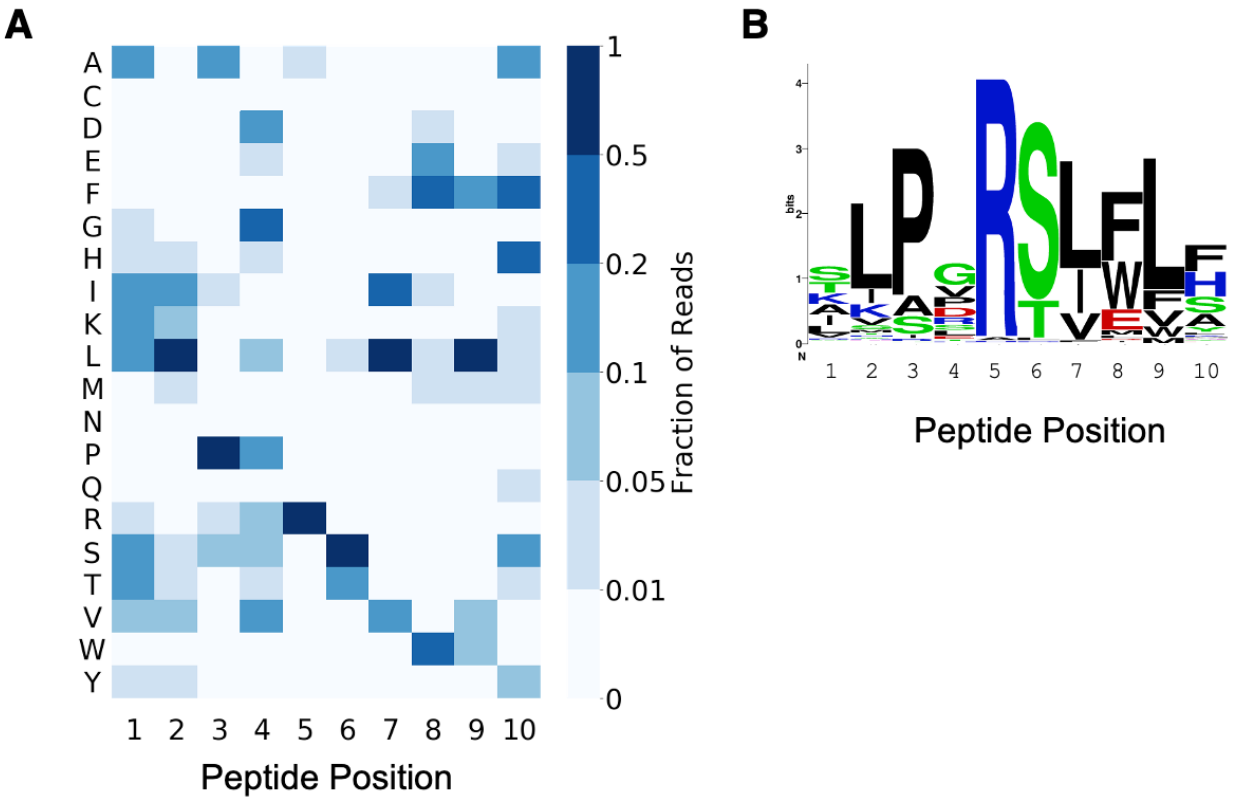


Figure 7. 10mer peptide data A) Heatmap and **B)** sequence logo representation of peptide positional amino acid preferences in 10mer peptides after three rounds of selection with CD94/NKG2C, with peptides weighted by read count.

6.5 Peptide motifs that differ from the dominant motif

We hypothesized that there could be high-confidence peptide binders that differ from the dominant motif but which would be non-obvious when viewing the composite motif. To investigate if there were subdominant binders, we performed clustering on the peptides that enriched in Round 3 of the CD94/NKG2x selections. Utilizing Gibbs Clustering (Andreatta et al., 2017), we interestingly observe a set of peptides characterized by a strong preference for P1 Trp and P2 Asn ("WN peptides") (**Figure 8**). We observe WN peptides with read counts >100 in both Round 3 CD94/NKG2A- and CD94/NKG2C-selected libraries, although there are more instances in the CD94/NKG2A-selected libraries (54 vs 6 instances). Examination of all enriched peptides with P1 Trp and P2 Asn revealed a clear amino acid preference at the N-terminus of the peptide. The peptide may be binding in an altered register, enabled by the Y84A mutation, where the conserved N-terminal residues sit in the middle of the groove, such as with the P3 Arg sitting in the 9mer P5 position and contacting the CD94/NKG2x receptor. This motif differs significantly from the dominant motif, regardless of peptide register and may bind with a partially filled groove (Anjanappa et al., 2020).

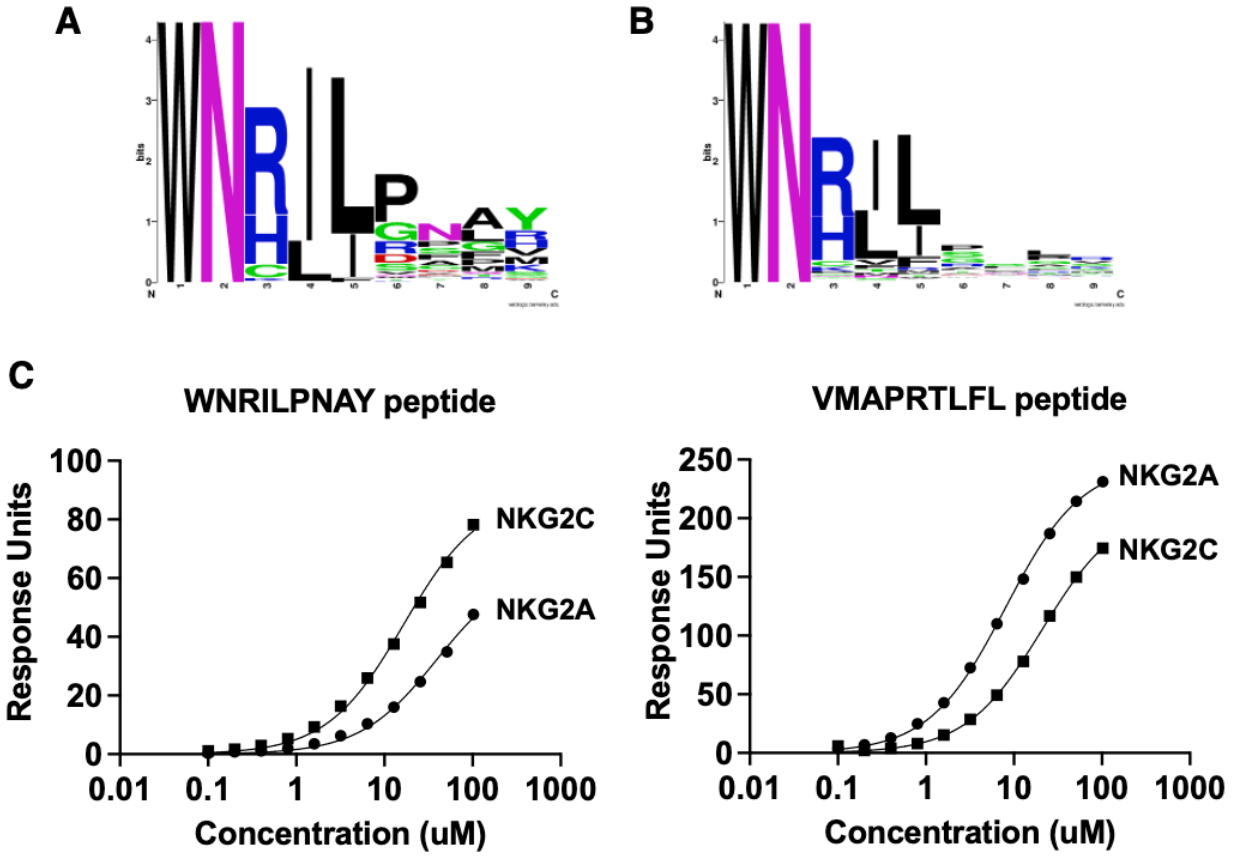


Figure 8. Subdominant motif of WN peptides with P1 Trp and P2 Asn. Sequence motifs of peptides that have P1 Trp and P2 Asn, **A)** weighted by frequency or **B)** not weighted by frequency in Round 3 of selections with CD94/NKG2A. **C)** SPR data for HLA-E single chain trimers containing WN 9mer peptide (left) or VL9 control peptide (right), binding to CD94/NKG2A or CD94/NKG2C.

6.6 Training prediction algorithms and predicting human- and pathogen-derived ligands

Given the large theoretical space (20^9) of 9mer peptides, we were interested in training a prediction algorithm on our yeast display data and predicting putative human- and pathogen-derived peptide binders to HLA-E and CD94/NKG2x. Given recent data suggesting that predictive performance on peptide-MHC binders is often limited by training data rather than architectural framework (Rappazzo et al., 2020), we trained the NNAlign architecture (Nielsen and Andreatta, 2017), which underlies the NetMHC suite of algorithms, on our yeast display data. We trained two models on yeast display data from Round 3 of CD94/NKG2A or CD94/NKG2C selections, with negative examples drawn from the unselected library such that the training set was class-balanced.

We applied this algorithm to 11 million unique 9mers, generated from a sliding window of size 9, step size 1, along a reference human proteome. Given the absence of relevant comparator methods for predicting HLA-E and CD94/NKG2x binding, we examined the rankings of known binders. The ranks of the following canonical VL9 signal peptide-derived HLA-E ligands were examined: VMAPRTLFL, VMAPRTLLL, VMAPRTLIL, VMAPRALLL, VMAPRTVLL, VMAPRTLVL (Hoare et al., 2006, 2008; Kraemer et al., 2015; O'Callaghan et al., 1998; Pietra et al., 2003; Strong et al., 2003; Sullivan et al., 2017; Ulbrecht et al., 2000), with their predicted ranks shown in **Table 1**. By Mann-Whitney U test, the distribution of ranks for these peptides were significantly different than the human proteome as a whole (with *p*-value of 1.1×10^{-5} for either CD94/NKG2A and CD94/NKG2C model predictions). These ranks for known binders support that our model is picking up true signal.

The top ten predicted human proteome-derived peptides are shown in **Table 2**. These binders commonly share features of known binders, including hydrophobic P7-P9 residues and P5 Arg. One of the top predicted binders resembles the N-terminal motif of the subdominant WN peptides. Peptides, such as VNPGRSLFL (derived from a bifunctional apoptosis regulator) and QMPSRSLLF (derived from cyclic AMP-responsive element-binding protein 3-like protein 1), are associated with stress responses or cell death. MHC-Ib have been associated with stress response, and these peptides may present potentially novel subsets of peptides HLA-E can present in states of cellular distress (Sasaki et al., 2016).

In addition to predicting on human-derived peptides, we applied our models to a reference human CMV virus. One mechanism by which CMV evades T cell and NK cell detection is by generating a UL40-derived peptide identical to canonical VL9 signal peptides, which can be loaded onto HLA-E, even when canonical class I processing is disrupted (Pietra et al., 2003; Ulbrecht et al., 2000). We rank this UL40-derived peptide highly among CMV-derived peptides (**Table 3**). Interestingly, however, both CD94/NKG2A and CD94/NKG2C models rank a UL120-derived peptide more highly, which bears the hallmarks of a true binder, including P3 Pro, P5 Arg, P8 Phe, and P9 Leu. The UL120-derived peptide that we predict as a strong binder is from a highly mutated region among different CMV strains. For example, despite a large degree of conservation in the UL120 proteins between strain AD169 and strain Merlin (used in our search), strain AD169 has the divergent peptide SAPLKTRFL at the corresponding position. We hypothesize that these differences could lead to differential HLA-E presentation of CMV-derived peptides and may contribute to different abilities to respond to CMV strains (Hammer et al., 2018; Pietra et al., 2010).

VL9 Peptides	NKG2A Model: Rank	NKG2C Model: Rank
VMAPRTLFL	224	2
VMAPRTLLL	418	12
VMAPRTVLL	547	22
VMAPRTLIL	481	32
VMAPRTLVL	602	48
VMAPRALLL	514	69

Table 1. VL9 peptide predicted ranks. Predicted ranks of eight signal peptides by CD94/NKG2A and CD94/NKG2C models, ranked out of 11 million human-derived 9mer peptides.

Model	Rank	Peptide	Source
NKG2A	1	QMPSRSLLF	Cyclic AMP-responsive element-binding protein 3-like protein 1
NKG2A	2	TLPKRGLFL	A-kinase anchor protein 6
NKG2A	3	TGPWRSLWI	General transcription factor 3C polypeptide 5
NKG2A	4	ILTDRSLWL	F-box only protein 41
NKG2A	5	VNPGRSLFL	Bifunctional apoptosis regulator
NKG2A	6	WNRLFPPLR	Ubiquitin-associated domain-containing protein 2
NKG2A	7	TLPERTLYL	Endothelin-converting enzyme-like 1
NKG2A	8	FLPNRSLLF	Solute carrier family 52, riboflavin transporter, member 3
NKG2A	9	VMGDRSVLY	ER membrane protein complex subunit 1
NKG2A	10	VMADKSIFY	Olfactory receptor 5D14
NKG2C	1	VMPPRTLLL	HLA class I histocompatibility antigen
NKG2C	2	VMAPRTLFL	HLA class I histocompatibility antigen
NKG2C	3	VNPGRSLFL	Bifunctional apoptosis regulator
NKG2C	4	QMPSRSLLF	Cyclic AMP-responsive element-binding protein 3-like protein 1
NKG2C	5	TLPERTLYL	Endothelin-converting enzyme-like 1
NKG2C	6	VMPGRTLCF	Neuromedin-K receptor
NKG2C	7	ILTDRSLWL	F-box only protein 41
NKG2C	8	RMPPRSVLL	Integrator complex subunit 1
NKG2C	9	TAPARTMFL	Phosphatidylserine decarboxylase proenzyme, mitochondrial
NKG2C	10	NMPARTVLF	Exosome RNA helicase MTR4

Table 2. Top ten predicted peptides from both models. Top predicted peptides by CD94/NKG2A and CD94/NKG2C models, with source proteins noted.

NKG2A Model: Rank	NKG2C Model: Rank	Peptide	Source
1	1	VLPHRTQFL	Membrane protein UL120
2	3	TGAARSFFF	DNA helicase/primase complex-associated protein
3	2	VMAPRTLIL	UL40

Table 3. Top predicted hCMV peptides from both models. Top predicted peptides by CD94/NKG2A and CD94/NKG2C models, with source protein noted. Ranks are among 61 thousand CMV-derived peptides.

6.7 Validation of HLA-E peptides

We first set out to validate a subset of our human and CMV proteome-derived predicted peptides in clonal yeast populations. To assess binding, we expressed several peptides individually on the surface of yeast, including two human- and one CMV-derived peptides. Additionally, we assess a top-enriching WN 9mer, the 5mer version of it, and the WN 5mer with P6-9 exchanged for Gly and Ser, to extend the linker length, in the event of steric restrictions. Given the low affinity of tetrameric CD94/NKG2x for even known positive control peptides (**Figure 1D**), we utilized high avidity receptor-coated beads to perform mock selections to enrich for yeast expressing our protein of interest, compared to a competing null population. We stained individual yeast constructs with an anti-epitope antibody then doped them into an excess of uninduced yeast (**Figure 9**, left column), such that the stained, peptide-expressing yeast were infrequent in the pool. This mixed population of yeast were then incubated with CD94/NKG2A-coated magnetic beads and enriched, as in library selections. The post-selection yeast were assessed (**Figure 9**, right column), and for nearly all constructs (excepting VLPHRTQFL, which enriched poorly compared to the others), the stained population enriched clearly, reflecting the pHLA-E construct binding to CD94/NKG2A.

Next, we performed differential scanning fluorimetry to measure melting temperatures reflective of protein stability, utilizing a Prometheus NanoTemper for measurements (**Figure 10**). Melting temperature was highest for VL9 peptide (51.5°C) and lowest for empty HLA-E with DMSO control (33.8°C). Three proteome-derived peptides were tested by Prometheus, with melting temperatures between the VL9 and empty controls, consistent with previously reported melting temperatures for non-VL9 binders (Hoare et al., 2008; Walters et al., 2018). A secondary inflection point around 60°C in most melts is consistent with melting of β2M (Hellman et al., 2016). Since Prometheus measurements utilize intrinsic tryptophan fluorescence, the approach was unfortunately not amenable to assessing binding of the WN peptides because of high tryptophan background.

To more directly test the WN peptides for binding to HLA-E and CD94/NKG2x, we performed surface plasmon resonance (SPR) on HLA-E SCT containing a 9mer WN-peptide with CD94/NKG2A or CD94/NKG2C (**Figure 8C**). Affinity of WN peptide SCT for CD94/NKG2x was lower than VL9 SCT. Excitingly, the WN peptide SCT some showed binding affinity to both CD94/NKG2A and CD94/NKG2C, and higher affinity to CD94/NKG2C than CD94/NKG2A, in contrast to VL9 peptide. The higher affinity for CD94/NKG2C receptor presents exciting questions and opportunities for skewing NK cell function through binding with higher affinity to the activating, rather than inhibitory, NK receptor. These data also leave questions about WN peptide affinity for HLA-E in the absence of the SCT linker sequence.

Finally, we assessed binding to HLA-E of top human- and CMV-derived peptides predicted by our CD94/NKG2x models (**Table 2** and **Table 3**) via their ability to stabilize HLA-E surface expression on RMA-S cell lines engineered to express HLA-E (**Figure 11**) (Hammer et al., 2018). At the two concentrations tested, we see clear stabilization by 12 peptides, including the stress-associated peptide QMPSRSLLF and UL120 CMV peptide VLPHRTQFL that we previously assessed by DSF. We see weak stabilization by the other stress-associated peptide (VNPGRSLFL) that we have tested by DSF. Ten remaining peptides, including WN peptides of variable lengths, showed minimal or no stabilization at the concentrations tested.

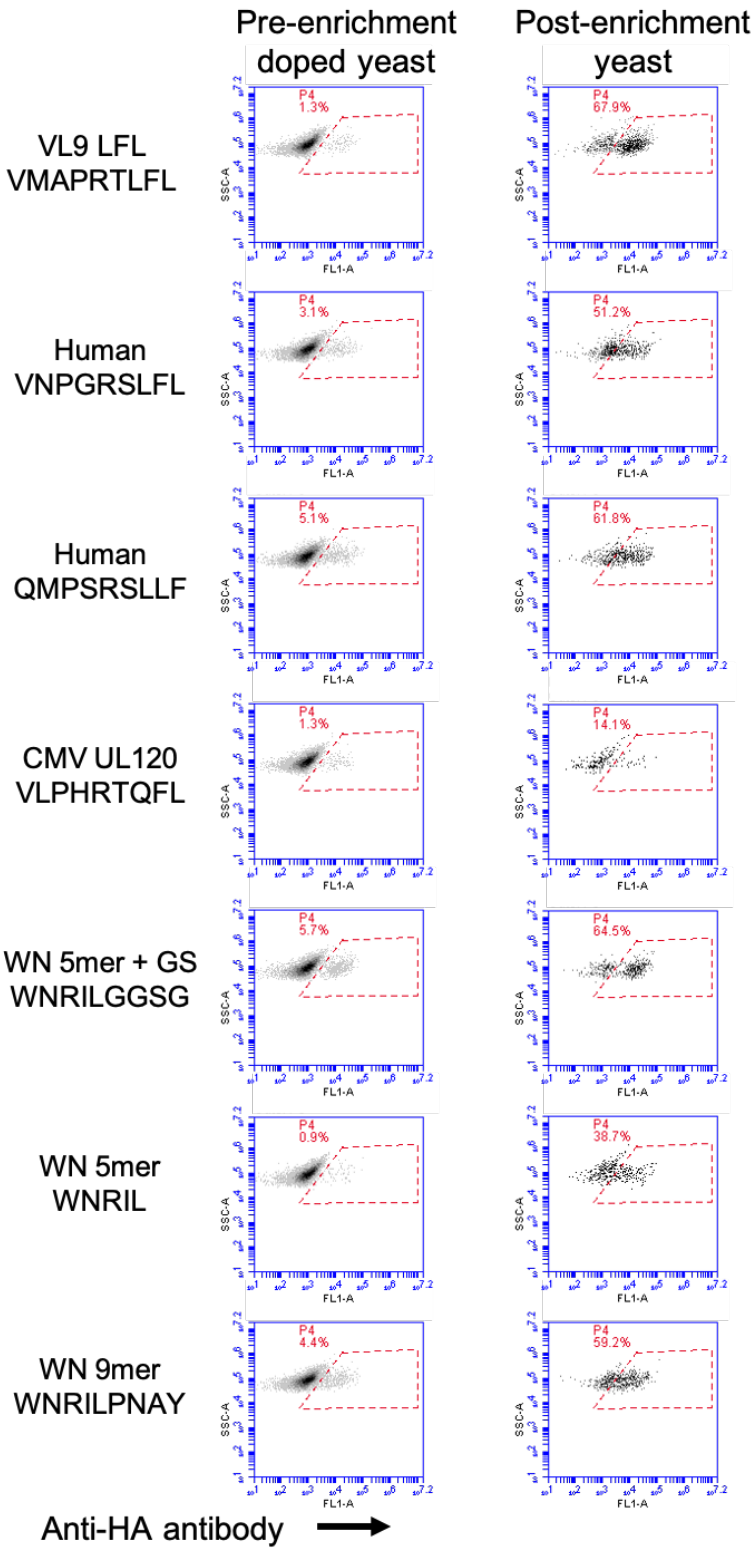
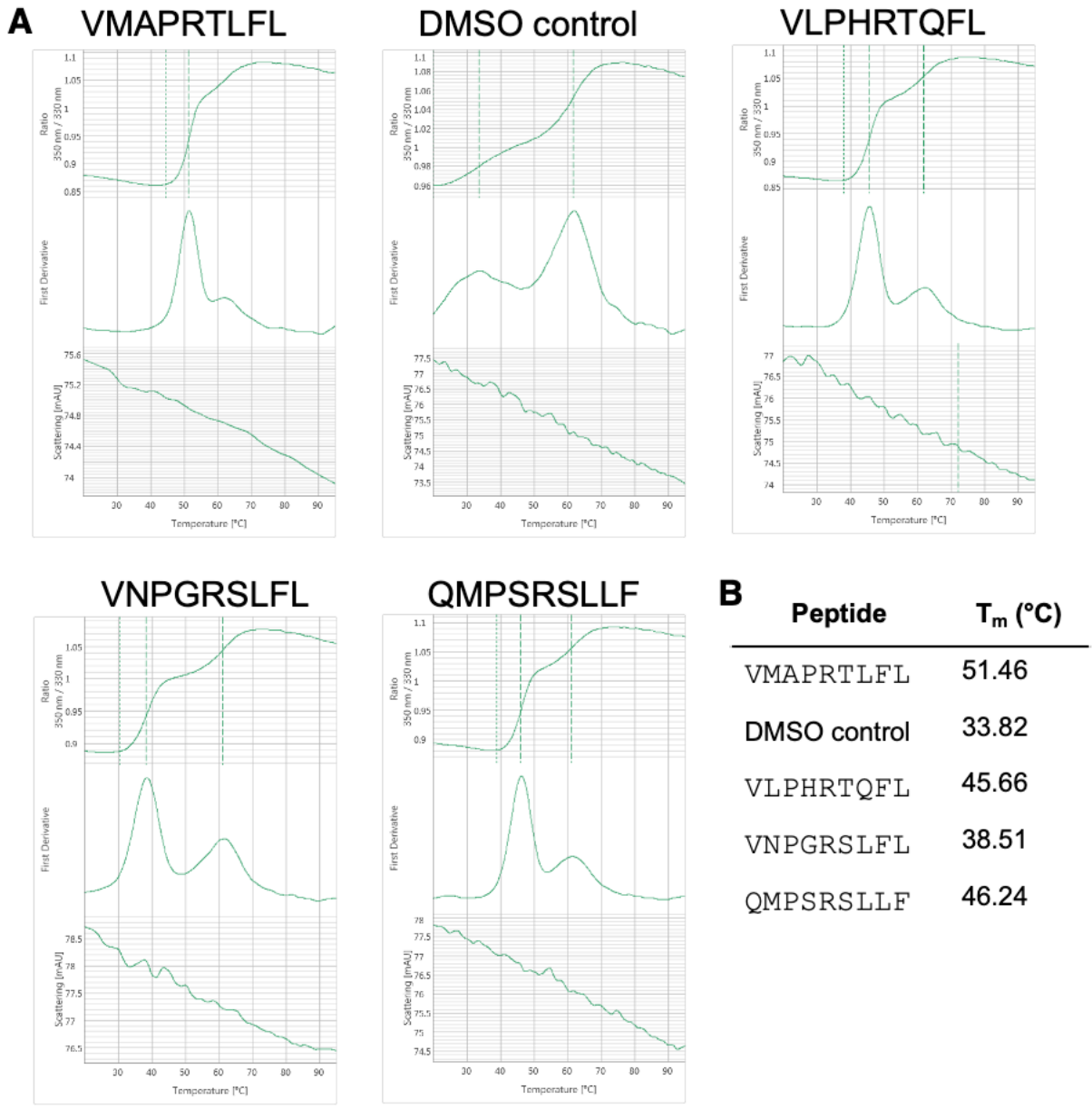


Figure 9. Yeast-based validations for WN and predicted peptides. Mock-selections of clonal yeast expressing pHLA-E constructs of interest that were epitope tag stained and doped into uninduced yeast. Samples of yeast before and after enrichment with CD94/NKG2A were assessed via flow cytometry.



B

Peptide	T_m (°C)
VMAPRTLFL	51.46
DMSO control	33.82
VLPHRTQFL	45.66
VNPGRSLFL	38.51
QMPSRSLLF	46.24

Figure 10. Differential scanning fluorimetry data generated via Prometheus NanoTemper for predicted peptides with HLA-E. A) Raw 350/330 nm ratios and first derivative for samples and **B)** summary of T_m values for the first inflection point called by Prometheus software for these samples, where the second inflection point is consistent with β2M unfolding.

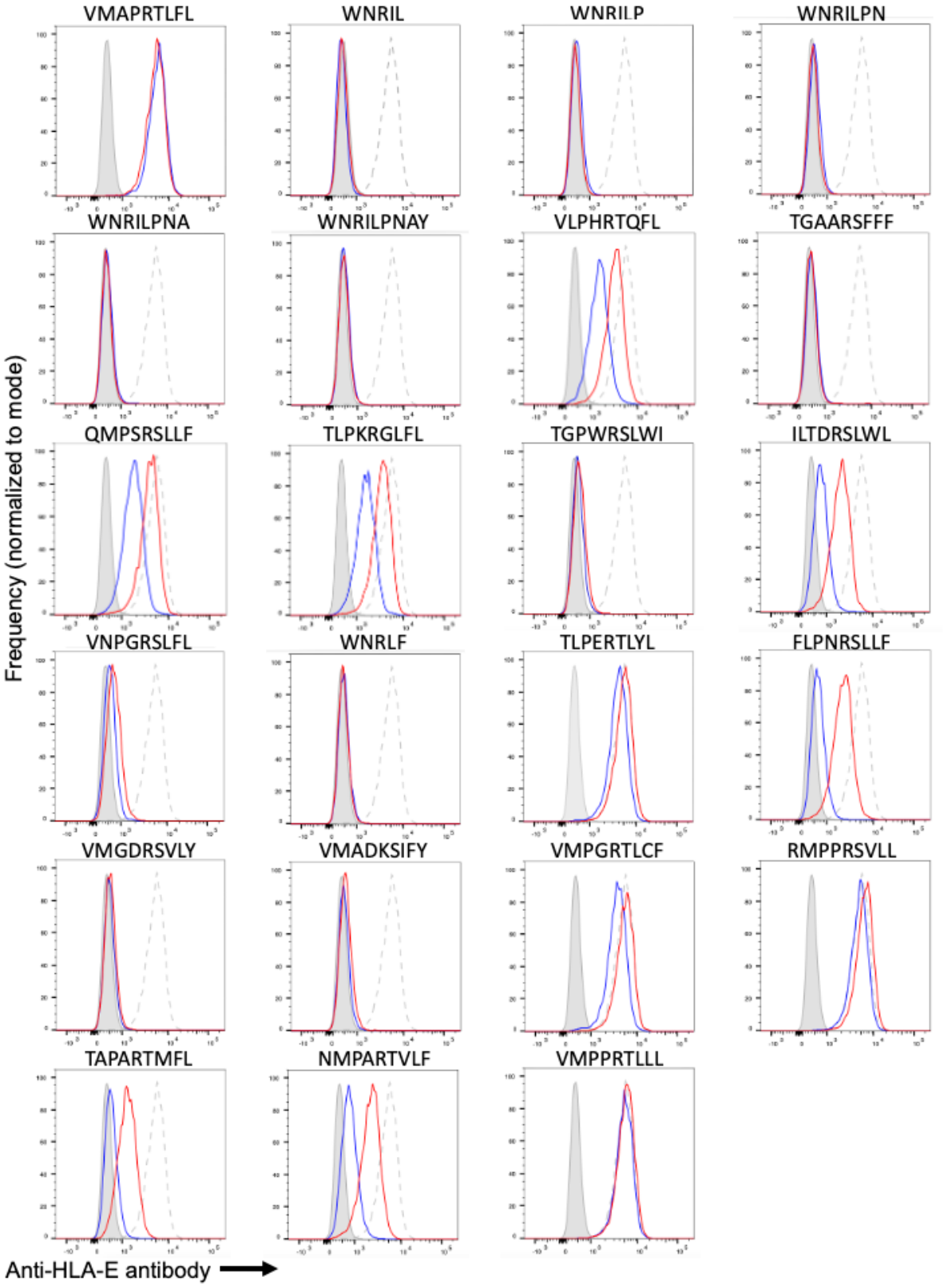


Figure 11. HLA-E peptide stabilization data. HLA-E surface levels are assessed in the presence of exogenously added peptide. DMSO control in grey, 30 μM in red, 3 μM in blue. VL9 positive control is shown as a dashed grey line in other plots.

6.8 Discussion and outlook

HLA-E is a highly promising target for vaccine and therapeutic design, recently shown to present more diverse peptides than previously thought to NK and T cell receptors (Burwitz et al., 2020; Hansen et al., 2016; Yang et al., 2021). Coupled with its conservation across global populations (Zheng et al., 2015), HLA-E along with CD94/NKG2x receptors have become active therapeutic areas of development (André et al., 2018; Burwitz et al., 2020; van Hall et al., 2019; van Montfoort et al., 2018; Yang et al., 2021). However, diverse HLA-E presented ligands (Hansen et al., 2016) highlight questions about the breadth and characteristics of peptides that can bind to HLA-E and cognate receptors. Here we utilize yeast-displayed libraries of 100 million peptides to characterize the repertoire of ligands that can bind to HLA-E and CD94/NKG2A or CD94/NKG2C (**Figure 1**).

We identify significant overlap between CD94/NKG2A and CD94/NKG2C peptide preferences, but more permissive binding across peptide positions for CD94/NKG2A ligands (**Figure 5**). These data reveal P2 hydrophobic preferences beyond Met, a strong P3 Pro preference in addition to an overwhelming P5 Arg preference. P6 prefers small polar residues, and P7-P9 strongly prefer hydrophobic residues. Our data are consistent with structural knowledge about CD94/NKG2x and HLA-E binding (Kaiser et al., 2008), canonical VL9 peptides, and broad preferences (Ruibal et al., 2020), and we further elucidate preferences at primary and secondary anchors and CD94/NKG2x-interacting residues.

In addition to the dominant motif, we identify a subdominant motif characterized by P1 Trp and P2 Asn preferences and minimal constraints at P6-P9, suggesting these may be short peptides. Excitingly, we see higher affinity of a WN peptide for the activating CD94/NKG2C receptor over the inhibitory CD94/NKG2A receptor (**Figure 8**), presenting opportunities for altering NK cell activity setpoints. These peptides appear to be very low affinity, with occupancy of the HLA-E groove only clear when linked to the MHC (**Figure 11**), leaving open questions about the physiological relevance of these subdominant peptides. Future stabilization assays (**Figure 11**) will be conducted at higher peptide concentrations to test for binding at higher concentrations. Given the novelty of these WN peptides and departure from the dominant motif, we are interested in crystallographic approaches to visualize how the peptide interacts with the HLA-E groove. Unfortunately, preliminary sparse screens for generating crystals with HLA-E single chain trimers with a WN peptide did not yield crystals. Future work may yield fruitful conditions for crystal generation or alternative structural approaches may be productive.

We also utilize our yeast display-generated data to train prediction algorithms and identify novel human- and CMV-derived peptides capable of binding to HLA-E and CD94/NKG2x. We assessed two stress-associated peptides in a yeast enrichment assay and with DSF (**Figure 9** and **Figure 10**), and both assays support these peptides binding to HLA-E and CD94/NKG2x. Of the two, VNPGRSLFL had the lower melting temperature and showed less stabilizing effect on surface-expressed HLA-E (**Figure 11**). We also assessed a CMV UL120-derived peptide on yeast (**Figure 9**), by DSF (**Figure 10**), and HLA-E stabilization (**Figure 11**), yielding high melting temperatures and clear HLA-E stabilization. With these data, marginal enrichment in the yeast enrichment assay may suggest this UL120-derived peptide can bind strongly to HLA-E but weakly to CD94/NKG2A. Additional human-derived peptides showed HLA-E-stabilizing effects (**Figure 11**), consistent with HLA-E-binding ability. Unique features of several non-stabilizing peptides could be used to rationalize their lack of binding, such as steric hindrance from P9 Tyr, P7 Phe, or

P4 Trp. We will assess these peptides further for HLA-E stabilization at higher concentrations of peptide.

To further assess our proteome predicted and WN peptides, we are collaborating with Chiara Romagnani's lab to investigate the peptides' functional effects on NK cell activity. We are excited to examine how our identified UL120-derived CMV peptide affects NK cell activity and if it may provide an addition mode for CMV to evade NK cell detection, in addition to the known UL40-derived VL9 peptide (Ulbrecht et al., 2000). Additionally, our UL120 peptide is from a highly mutated region and may underlie differing responses across CMV strains (Hammer et al., 2018; Pietra et al., 2010).

Taken together, we have generated a dataset and prediction tools that serve as tools for both studying underlying HLA-E and NK cell biology and HLA-E and CD94/NKG2x in therapeutic contexts, and whose utility has been demonstrated by identification of novel HLA-E and CD94/NKG2x ligands.

6.9 Methods

Yeast-displayed pMHC design

Yeast-displayed HLA-E*01:03 was generated as a single trimer, covalently linked to a peptide and human β2M. Specifically, the C-terminus of the peptide was linked to N-terminus of β2M via Gly-Ser linker containing a Myc epitope tag (EQKLISEEDL) and protease site (LEVLFQGP). A second Gly-Ser linker connected the C-terminus of β2M to the N-terminus of HLA-E heavy chain α1, α2, and α3 domains. HLA-E contained a Y84A mutation to open the peptide binding groove to accommodate the peptide linker. HLA-E C-terminus is connected to Aga2 via a final Gly-Ser linker containing an epitope tag (HA in clonal constructs: YPYDVPDYA; FLAG in library: DYKDDDDK). Yeast display constructs were generated on the pYAL vector. Yeast strains were grown to confluence at 30°C in pH 5 SDCAA yeast media then subcultured into pH 5 SGCAA media at OD_{600} = 1.0 for 48-72 hours induction at 20°C (Chao et al., 2006).

Tetramer and antibody staining yeast

To stain peptide-HLA-E (pHLA-E) with CD94/NKG2x tetramers, each biotinylated receptor was mixed with streptavidin coupled to AlexaFluor647 (made in-house) at a 5:1 ratio and incubated for 5 min on ice. Yeast were stained with 500 nM tetramer for 2 hours at 4°C in the dark with rotation. Then, the yeast were washed twice with FACS buffer (0.5% BSA and 2 mM EDTA in 1x PBS) before analysis via flow cytometry (Accuri C6 flow cytometer, BD Biosciences; Franklin Lakes, New Jersey).

Antibody staining was performed on yeast washed into FACS buffer, with antibody at a 1:50 or 1:100 volume ratio. Yeast incubated with antibody for at least 20 minutes and excess antibody is removed by washing with FACS buffer. Staining was assessed on the Accuri C6 flow cytometer. Antibody clone 3D12 was used for anti-HLA-E staining. Negative control peptide sequences assessed in **Figure 2** were provided by Simon Brackenridge (Gillespie and McMichael Groups, Oxford).

Library design and selection
Randomized peptide libraries were generated using polymerase chain reaction (PCR) on the pMHC construct with primers encoding NNK degenerate codons (N = any base; K = G or T), to encode 9 or 10 randomized amino acids.

Randomized pMHC PCR product was mixed with linearized pYAL vector backbone at a 5:1 mass ratio and electroporated into electrocompetent RJY100 yeast (Van Deventer et al., 2015). The final 9mer and 10mer libraries contained approximately 1.5×10^8 and 7×10^7 yeast transformants, respectively.

As described above, yeast were subject to selections for CD94/NKG2x binding. Selections were performed using streptavidin-coated magnetic beads (Miltenyi Biotec; Begisch Gladbach, Germany) coupled to biotinylated CD94/NKG2x (provided by Lee Garner from the Gillespie and McMichael labs at Oxford). Biotinylation was confirmed by streptavidin gel shift assay (Fairhead and Howarth, 2015). Yeast were cultured, induced, and selected in three iterative rounds of selection. In a fourth round of selections, yeast were incubated with tetrameric AlexaFluor647-conjugated streptavidin (produced in-house) coupled to biotinylated CD94/NKG2x, and selected with ∝-AlexaFluor647 magnetic beads (Miltenyi Biotec; Bergisch Gladbach, Germany). Each round was preceded by negative selection clearance round with uncoated streptavidin beads (rounds 1-3) or streptavidin and ∝-AlexaFluor647 beads (round 4). In round four, prior to addition of beads, a sampling of CD94/NKG2x tetramer-stained yeast were assessed via flow cytometry on an on an Accuri C6 flow cytometer.

To prevent contamination with clonal yeast, the library template DNA encoded a stop codon in the peptide-encoding region. Additionally, the C-terminal HA tag was swapped for a FLAG tag to readily assess library expression and contamination.

Library sequencing and analysis
Following selections, plasmid DNA from ~10^7 yeast were extracted using Zymoprep II Yeast Miniprep Kit (Zymo Research; Irvine, CA) from each round of selection and the unselected library. Amplicons for deep sequencing were generated by PCR using the purified plasmids from yeast miniprep. Two rounds of PCR were completed to add homology for sequencing primers, i5 and i7 paired-end handles, and sequencing barcodes that are unique to each round of selection and enable pooling of DNA from rounds of selection. Amplicons were sequenced on an Illumina MiSeq (Illumina; San Diego, CA) with 2 x 150 nt paired-end reads at the MIT BioMicroCenter.

Paired-end reads were assembled and filtered for length and correct flanking sequences using PandaSeq (Masella et al., 2012). To correct for PCR or sequencing errors, sequences were clustered with more frequent sequences within Hamming Distance = 1 in DNA space using CDHit (Fu et al., 2012; Li and Godzik, 2006). Peptide sequences were translated from DNA and stop codon-containing sequences removed using an in-house script.

Heat map and sequence logo visualization of library-enriched peptides
Sequence logo visualizations were generated using the sequence logo generator WebLogo (Crooks et al., 2004). Sequences were repeated *int(count/100)* times, which filters possible noise with read

counts <100 and weights sequences by frequency. Sequence logos were generated using WebLogo default settings.

Heatmaps were generated with a custom script, on all enriched peptides from a given round of selection. Frequencies were calculated by weighting each peptide by its read count. Maps were binned as shown to capture the frequency of reads that encode a peptide with a given amino acid at a given position.

Clustering peptides to identify subdominant motifs
We identified the subdominant cluster of WN peptides using various clustering methods, including Gibbs Cluster 2.0 (Andreatta et al., 2017) on peptides enriched in Round 3 of CD94/NKG2A selections (not weighted by read counts), excluding stop codon-containing peptides, using the default method "MHC class I ligands of same length", modified to assess 1-15 clusters, with the trash cluster to remove outliers turned off.

Prediction algorithm generation
CD94/NKG2A- or CD94/NKG2C-specific prediction models were trained on yeast display library data, with positive examples drawn from Round 3 of selections, with counts greater than or equal to 100, each appearing in the training data *int(count/100)* times. Negative examples were selected from the unselected library, with read count equal to 1 and excluding stop codon-containing peptides. The total number of negative peptides were selected such that there were equal numbers of positive and negative examples in the training set (total number, not unique examples). Positive examples were assigned a target value of 1 and negative examples were assigned a target value of 0.

These data were used to train NNAlign 2.0 with MHC Class I defaults, excepting Maximum Length for Deletions, Maximum Length for Insertions, and Length of PFR for Composition Encoding were set to zero; Encode PFR Length was set to -1 (for no encoding); and both Binned Peptide Length Encoding and Length of Peptides Generated from FASTA Entries were set to 9.

Prediction on proteome-derived peptides
With a 9mer window, step size 1, 11,019,710 unique human proteome 9mers were generated from a reference proteome (Uniprot UP000005640), excluding peptides containing selenocysteine. The same was done for human CMV (reference Uniprot proteome UP000000938), generating 61,303 peptides.

Mann Whitney U test was performed on HLA-E-binding VL9 signal peptides from the human proteome compared to other 9mer human proteome peptides, and *p*-values were calculated using scipy version 1.4.1.

Yeast column enrichment competition assay
A yeast column enrichment competition assay was performed to evaluate pHLA-E binding to CD94/NKG2A receptor. Induced yeast expressing pHLA-E SCT were epitope stained with HA-488, then mixed with uninduced, unstained yeast at 1:50 ratio (2×10^5 induced, 10^7 uninduced). A sample of the mixture was assessed via flow cytometry analysis. The remaining mixed population of yeast were incubated with CD94/NKG2A-coated magnetic beads and enriched, as in library

selections. The enriched population was assessed by flow cytometry. The HA-488 epitope stain differentiated pHLA-E-expressing yeast from uninduced competitor yeast.

Recombinant protein production for SPR

Recombinant soluble HLA-E single chain trimers, CD94/NKG2A, and CD94/NKG2C for SPR were produced using a baculovirus expression system with High Five (Hi5) insect cells (ThermoFisher; Waltham, MA). Individual constructs were cloned into pAcGP67a vectors. For each, 2 μg of plasmid DNA was transfected into SF9 insect cells using BestBac 2.0 baculovirus DNA (Expression Systems; Davis, CA) and Cellfectin II reagent (ThermoFisher). Viruses were propagated to high titer and transduced into Hi5 cells, grown at 27°C for 48–72 hours, and purified from pre-conditioned cell media supernatant using Ni-NTA resin and size exclusion chromatography with a S200 column on an AKTAPure FPLC (GE Healthcare; Chicago, IL).

Single chain timers were formatted with the C-terminus of the peptide linked to the N-terminus of human β2M via a flexible linker. In turn a flexible Gly-Ser linker connects the C-terminus of β2M to the N-terminus of the extracellular α1, α2, and α3 domains of HLA-E*01:03 heavy chain, containing a Y84A mutation to accommodate the linked peptide. An AviTag biotinylation tag and poly-histidine tag were connected to the C-terminus of HLA-E heavy chain. Protein was biotinylated overnight before FPLC-based purification.

CD94/NKG2x proteins were expressed as single-chain fusion proteins. The C-terminus of human NKG2A or NKG2C ectodomain was connected via a GGSGGS linker to human CD94 ectodomain. The C-terminus of CD94 is connected to an AviTag biotinylation tag and poly-histidine tag. Protein was used fresh and dialyzed overnight into HBS-EP+ buffer (10 mM HEPES pH 7.4, 150 mM NaCl, 3 mM EDTA and 0.05% v/v Surfactant P20) before use in SPR (Cytiva; Malborough, MA).

Surface plasmon resonance

Steady-state surface plasmon resonance experiments were performed with a Biacore T200 instrument. pMHC was immobilized at 392±5 (for WN peptide SCT) or 400±5 (VL9 peptide SCT) response units on a Series S SA sensor chip (Cytiva). Reference flow cells utilized immobilized HLA-DR (HLA-DRA1*01:01, HLA-DRB1*04:01, linked to CLIP peptide) (Rappazzo et al., 2020), and matched to test cells: 392±5 response units for WN SCT control and 399±5 response units for VL9 SCT control. CD94/NKG2x in HBS-EP+ buffer was injected as analyte in a concentration range of 0.1 μM to 102.4 μM and flow rate of 10 μL/min at 25°C. Data was fit with Prism 9.3 (GraphPad Software Inc; San Diego CA) to a "one site, specific binding" model.

For each non-reference flowcell, data was fit to the B_{max} for the strongest binding analyte. K_D values were estimated as follows and are not exact estimates as they are near to the maximum tested concentrations: WN SCT + CD94/NKG2A: 74.06 μM; WN SCT + CD94/NKG2C: 16.49 μM; VL9 SCT + CD94/NKG2A: 8.015 μM; VL9 SCT + CD94/NKG2C: 30.84 μM.

Prometheus differential scanning fluorimetry experiments

Differential scanning fluorimetry (DSF) experiments were performed with a Prometheus NanoTemper NT.48 (Munich, Germany) to measure intrinsic tryptophan fluorescence, after initial

attempts with SYPRO Orange appeared to destabilize complexes such that they resembled the relatively stable empty HLA-E.

For these experiments, refolded empty HLA-E was generated. Human β2M plus a poly-histidine tag and HLA-E*01:03 extracellular α1, α2, and α3 domains plus an AviTag biotinylation tag were separately codon optimized for *E. coli* and cloned into the bacterial expression vector pET28a. Inclusion bodies for HLA-E and β2M were made in BL21 *E. coli*. Inclusion bodies were purified, homogenized, and dissolved, including denaturation in 8 M urea solution. Before refolding, each IB was mixed with an equal volume of Gdn-Cl solution (6 M Guanidine HCl, 500 mM Tris, 2 mM EDTA, 100 mM NaCl) and incubated overnight at 37°C. β2M was injected dropwise with a 27-gauge needle into refolding buffer (100 mM pH 8.3 Tris, 400 mM L-arginine hydrochloride, 2 mM EDTA, 0.5 mM oxidized glutathione, 5 mM reduced glutathione, 0.2 mM PMSF) and incubated, stirring gently, at 4°C for 1 hour, after which an equal mass of HLA-E heavy chain was similarly injected. This solution incubated at 4°C overnight, followed by two days of dialysis in 10 mM Tris + 50 mM NaCl. Protein was concentrated in 10 kDa molecular weight cutoff centrifugal concentrators and purified by size exclusion chromatography on an AKTAPURE FPLC (GE Healthcare, Chicago IL), using an S200 column followed by a S75 column. Protein was concentrated with size filters as initial attempts to concentrate using the poly-histidine tag and Ni-NTA resin resulted in the protein precipitating out of solution.

Individual DSF reactions were set up with 9 ug of MHC and 100x, 50x, or 10x peptide to MHC ratio by molarity (peptides from GenScript). Final reaction volume was 20 uL and reaction mixtures were incubated at room temperature for 20-30 minutes.

Prometheus excitation power was set such that raw fluorescence counts were between 8000-15000 for each sample. Samples were heated at 1°C per minute, from 20-95°C. DSF measurements were taken in duplicate with high-sensitivity and standard capillaries. Melting temperatures were similar for a given peptide across conditions, and representative data from 50x ratio in high sensitivity capillary are presented.

HLA-E surface stabilization assay

To test peptide binding to HLA-E, peptides were added to RMA-S cell lines expressing HLA-E (Hammer et al., 2018) and incubated overnight. HLA-E stability was assessed by anti-HLA-E surface staining. Peptide concentrations of 30 μM and 3 μM were assessed and DMSO controls were included.

6.10 Acknowledgements
Yeast display experiments were conducted in conjunction with Ning (Alexa) Guan, who was a key driver in completing the yeast display selection experiments. Thank you to Geraldine Gillespie, Andrew McMichael, and Lee Garner at Oxford for their insight and support of this project and generously providing CD94/NKG2x protein for yeast display selections and the map of their CD94/NKG2x construct. Geraldine Gillespie and Max Quastel generously shared their HLA-E Prometheus protocol. Simon Brackenridge provided HLA-E negative control peptide sequences. SPR measurements were taken in conjunction with Nishant Singh, who also provided advice and insight on HLA-E refolds. Thank you to Chiara Romagnani and Timo Rückert at DRFZ Berlin for helpful conversations about effects of HLA-E-binding peptides on NK cell function, for performing the cell-based HLA-E stabilization experiments, and ongoing work to assess impacts of peptides on NK cell activation. This work is being written up as a manuscript; Alexa Guan was an important contributor to the content of this work, especially the content of the introduction.

CHAPTER 7: CONCLUSIONS AND FUTURE DIRECTIONS

In this thesis, I have described our work to develop and apply tools for studying peptide-MHC interactions in a more globally-representative manner. These tools present new opportunities for continued study and application, which are further discussed in this chapter.

To investigate peptide-MHC-II binding, we used a yeast display-based platform for generating high-quality datasets of peptide binders from unbiased libraries and highlight the utility of these data for training prediction algorithms (**Chapter 2**). Yeast display-trained prediction algorithms improve predictions of peptide affinity and predictions on peptides that are underrepresented in other datasets, such as cysteine-containing peptides. The improvement in peptide affinity prediction is likely due to generation of data across multiple rounds of selections, and we are currently investigating methods for adapting the yeast display platform to more directly capture peptide affinity data. We also demonstrate a novel approach for expanding yeast display throughput across alleles in a second-generation platform (**Chapter 5**). Generating additional large-scale peptide-MHC binding datasets will likely be most impactful for augmenting datasets of poorly-studied MHC alleles.

Development of our yeast display platform for peptide-MHC-II binding (**Chapter 2**) (Rappazzo et al., 2020) was concurrent with advances in the field in mono-allelic mass spectrometry (Abelin et al., 2017, 2019) and prediction algorithms (Chen et al., 2019; O'Donnell et al., 2018; Racle et al., 2019; Reynisson et al., 2020; Zeng and Gifford, 2019). Datasets generated through mass spectrometry and yeast display provide complementary advantages, and future work may combine these datatypes to capture the benefits of each. For example, mass spectrometry captures information about peptide processing and binding, but, by its nature, likely contains biases related to peptide ionization ability and native protein expression (Abelin et al., 2019; Barra et al., 2018; Scally et al., 2017; Stopfer et al., 2021). Recent advances in the field incorporate internal mass spectrometry standards to account for these biases, though they are not uniformly employed (Stopfer et al., 2020, 2021). Contrastingly, yeast display directly captures peptide-MHC binding, though does not include information about native peptide processing. Future efforts to combine these complementary data generation techniques will likely provide further improvements in peptide-MHC prediction. Data combination may necessitate concerted advances in computational algorithms, and may require multi-mode or multi-step prediction, such as prediction of peptide-MHC binding and assessment of the likelihood of peptide processing. Relatedly, we identified a high degree of sensitivity in prediction algorithms to peptide flanking sequences, including their length and register (**Chapter 3**), which may need to be tuned, either through algorithmic optimization or improved training data.

To benchmark peptide-MHC binding prediction, careful generation and curation of test sets will also be required, to avoid inadvertently biasing towards non-meaningful dataset features. Further, peptide-MHC binding approaches can be utilized to benchmark or curate data in public repositories, such as the Immune Epitope Database (Vita et al., 2019). Curation would allow for more reliable use of high-confidence data in the repositories, including peptide-MHC affinity data.

We also describe an approach for direct assessment of peptide-MHC binding for defined libraries of peptides using yeast display, with applications such as benchmarking optimized sequences

(**Chapter 3**), assessing pathogen-derived peptides (**Chapter 4**), and discovery of human autoantigens (proposed in **Chapter 5**). In the future, targeted approaches like these can be utilized to augment and complement computational approaches. For example, it may be advantageous to rely on computation for many peptide-MHC pairings and to experimentally assess peptide-MHC binding for peptides containing cysteine and for less-studied MHC alleles. Additionally, high-confidence data can be generated by combining algorithmic and experimental investigations.

Despite advances in peptide-MHC binding assessment and prediction, the field is currently limited in predictions of peptide immunogenicity. Though we presently lack a high-throughput paradigm for assessing immunogenicity, one potential investigative route could assess peptide similarity with human proteins, hypothesizing that difference from self-peptides against which T cells are tolerized could be a determinant of peptide immunogenicity. We are also interested in the immunogenicity of optimized peptides (**Chapter 3**) and if optimized peptides could have differential clinical effects over non-optimized sequences in peptide vaccines. Immunogenicity prediction could also be helpful in identification of human autoantigens, related to work proposed in **Chapter 5**.

In addition to our investigations of MHC-II proteins, we took an orthogonal approach to studying MHCs across populations, investigating the highly conserved class Ib MHC HLA-E (**Chapter 6**). We elucidated the repertoire of peptides capable of binding to HLA-E and its cognate receptors CD94/NKG2A and CD94/NKG2C. From these data, we train prediction algorithms to identify novel proteome-derived peptides capable of binding HLA-E and NK receptors. Among the novel peptides, we identify a CMV-derived peptide capable of binding to HLA-E. Future work on the role of this peptide in native CMV infection could increase our understanding of the pathogen and its immune evasion strategies, potentially revealing vulnerabilities that can be targeted therapeutically.

Recent work has identified highly diverse peptides capable of binding to MHC-E (the pan-species counterpart to HLA-E). These diverse peptides were in turn recognized by $CD8^+$ T cells (Hansen et al., 2013, 2016). Future work investigating how these diverse peptides are binding to MHC-E would be illuminating, and identification of TCR ligands and their binding modes would increase our understanding of MHC-E in disease contexts and its therapeutic potential.

In conclusion, the diversity of the immune system underlies its remarkable ability to respond to the myriad of pathogens encountered over a lifetime. To better study peptide-MHC binding across diverse MHC alleles, we have developed tools to study binding in a more globally-representative manner. These tools open new avenues for generating large, unbiased datasets and combining with computation to probe important immunological questions, across cancer, infectious disease, autoimmunity, and beyond.

REFERENCES

Abelin, J.G., Keskin, D.B., Sarkizova, S., Hartigan, C.R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G.L., Eisenhaure, T.M., et al. (2017). Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate Epitope prediction. Immunity *46*, 315–326.

Abelin, J.G., Harjanto, D., Malloy, M., Suri, P., Colson, T., Goulding, S.P., Creech, A.L., Serrano, L.R., Nasir, G., Nasrullah, Y., et al. (2019). Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer Epitope prediction. Immunity *51*, 766-779.e17.

Altfeld, M., Addo, M.M., Rosenberg, E.S., Hecht, F.M., Lee, P.K., Vogel, M., Yu, X.G., Draenert, R., Johnston, M.N., Strick, D., et al. (2003). Influence of HLA-B57 on clinical presentation and viral control during acute HIV-1 infection. AIDS *17*, 2581–2591.

Altmann, D.M., and Boyton, R.J. (2020). SARS-CoV-2 T cell immunity: Specificity, function, durability, and role in protection. Sci. Immunol. *5*, eabd6160.

Alvarez, B., Barra, C., Nielsen, M., and Andreatta, M. (2018). Computational tools for the identification and interpretation of sequence motifs in immunopeptidomes. Proteomics *18*, 1700252.

André, P., Denis, C., Soulas, C., Bourbon-Caillet, C., Lopez, J., Arnoux, T., Bléry, M., Bonnafous, C., Gauthier, L., Morel, A., et al. (2018). Anti-NKG2A mAb is a checkpoint inhibitor that promotes anti-tumor immunity by unleashing both T and NK cells. Cell *175*, 1731-1743.e13.

Andreatta, M., Alvarez, B., and Nielsen, M. (2017). GibbsCluster: unsupervised clustering and alignment of peptide sequences. Nucleic Acids Res. *45*, W458–W463.

Andreatta, M., Trolle, T., Yan, Z., Greenbaum, J.A., Peters, B., and Nielsen, M. (2018). An automated benchmarking platform for MHC class II binding prediction methods. Bioinformatics *34*, 1522–1528.

Angelini, D.F., Zambello, R., Galandrini, R., Diamantini, A., Placido, R., Micucci, F., Poccia, F., Semenzato, G., Borsellino, G., Santoni, A., et al. (2011). NKG2A inhibits NKG2C effector functions of γδ T cells: implications in health and disease. J. Leukoc. Biol. *89*, 75–84.

Anjanappa, R., Garcia-Alai, M., Kopicki, J.-D., Lockhauserbäumer, J., Aboelmagd, M., Hinrichs, J., Nemtanu, I.M., Uetrecht, C., Zacharias, M., Springer, S., et al. (2020). Structures of peptide-free and partially loaded MHC class I molecules reveal mechanisms of peptide selection. Nat. Commun. *11*, 1314.

Arstila, T.P., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J., and Kourilsky, P. (1999). A direct estimate of the human alphabeta T cell receptor diversity. Science *286*, 958–961.

Backert, L., and Kohlbacher, O. (2015). Immunoinformatics and epitope prediction in the age of genomic medicine. Genome Med. *7*, 119.

Baek, I.-C., Choi, E.-J., Shin, D.-H., Kim, H.-J., Choi, H., and Kim, T.-G. (2021). Allele and haplotype frequencies of human leukocyte antigen-A, -B, -C, -DRB1, -DRB3/4/5, -DQA1, -DQB1, -DPA1, and -DPB1 by next generation sequencing-based typing in Koreans in South Korea. PLoS One *16*, e0253619.

Balato, A., Unutmaz, D., and Gaspari, A.A. (2009). Natural killer T cells: an unconventional T-cell subset with diverse effector and regulatory functions. J. Invest. Dermatol. *129*, 1628–1642.

Barra, C., Alvarez, B., Paul, S., Sette, A., Peters, B., Andreatta, M., Buus, S., and Nielsen, M. (2018). Footprints of antigen processing boost MHC class II natural ligand predictions. Genome Med. *10*, 84.

Bazan, J., Całkosiński, I., and Gamian, A. (2012). Phage display--a powerful technique for immunotherapy: 2. Vaccine delivery. Hum. Vaccin. Immunother. *8*, 1829–1835.

Béziat, V., Dalgard, O., Asselah, T., Halfon, P., Bedossa, P., Boudifa, A., Hervier, B., Theodorou, I., Martinot, M., Debré, P., et al. (2012). CMV drives clonal expansion of NKG2C+ NK cells expressing self-specific KIRs in chronic hepatitis patients. Eur. J. Immunol. *42*, 447–457.

Bhattacharya, T., Daniels, M., Heckerman, D., Foley, B., Frahm, N., Kadie, C., Carlson, J., Yusim, K., McMahon, B., Gaschen, B., et al. (2007). Founder effects in the assessment of HIV polymorphisms and HLA allele associations. Science *315*, 1583–1586.

Birnbaum, M.E., Dong, S., and Garcia, K.C. (2012). Diversity-oriented approaches for interrogating T-cell receptor repertoire, ligand recognition, and function. Immunol. Rev. *250*, 82–101.

Birnbaum, M.E., Mendoza, J.L., Sethi, D.K., Dong, S., Glanville, J., Dobbins, J., Ozkan, E., Davis, M.M., Wucherpfennig, K.W., and Garcia, K.C. (2014). Deconstructing the peptide-MHC specificity of T cell recognition. Cell *157*, 1073–1087.

Birnbaum, M.E., Mendoza, J., Bethune, M., Baltimore, D., and Garcia, K.C. (2017). Ligand discovery for t cell receptors. US20170192011A1.

Blackwell, J.M., Jamieson, S.E., and Burgner, D. (2009). HLA and infectious diseases. Clin. Microbiol. Rev. *22*, 370–385, Table of Contents.

Bonilla, F.A., and Oettgen, H.C. (2010). Adaptive immunity. J. Allergy Clin. Immunol. *125*, S33-40.

Brooks, A.G., Posch, P.E., Scorzelli, C.J., Borrego, F., and Coligan, J.E. (1997). NKG2A complexed with CD94 defines a novel inhibitory natural killer cell receptor. J. Exp. Med. *185*, 795–800.

Burwitz, B.J., Hashiguchi, P.K., Mansouri, M., Meyer, C., Gilbride, R.M., Biswas, S., Womack, J.L., Reed, J.S., Wu, H.L., Axthelm, M.K., et al. (2020). MHC-E-restricted CD8+ T cells target hepatitis B virus-infected human hepatocytes. J. Immunol. *204*, 2169–2176.

Busch, R., Rinderknecht, C.H., Roh, S., Lee, A.W., Harding, J.J., Burster, T., Hornell, T.M.C., and Mellins, E.D. (2005). Achieving stability through editing and chaperoning: regulation of MHC class II peptide binding and expression. Immunol. Rev. *207*, 242–260.

Cai, W., Zhou, D., Wu, W., Tan, W.L., Wang, J., Zhou, C., and Lou, Y. (2018). MHC class II restricted neoantigen peptides predicted by clonal mutation analysis in lung adenocarcinoma patients: implications on prognostic immunological biomarker and vaccine design. BMC Genomics *19*.

Cannon, M.J., Schmid, D.S., and Hyde, T.B. (2010). Review of cytomegalovirus seroprevalence and demographic characteristics associated with infection. Rev. Med. Virol. *20*, 202–213.

Carrillo-Bustamante, P., Keşmir, C., and de Boer, R.J. (2016). The evolution of natural killer cell receptors. Immunogenetics *68*, 3–18.

Chao, G., Lau, W.L., Hackel, B.J., Sazinsky, S.L., Lippow, S.M., and Wittrup, K.D. (2006). Isolating and engineering human antibodies using yeast surface display. Nat. Protoc. *1*, 755–768.

Chaplin, D.D. (2010). Overview of the immune response. J. Allergy Clin. Immunol. *125*, S3-23.

Chen, B., Khodadoust, M.S., Olsson, N., Wagar, L.E., Fast, E., Liu, C.L., Muftuoglu, Y., Sworder, B.J., Diehn, M., Levy, R., et al. (2019). Predicting HLA class II antigen presentation through integrated deep learning. Nat. Biotechnol. *37*, 1332–1343.

Choo, S.Y. (2007). The HLA system: genetics, immunology, clinical testing, and clinical implications. Yonsei Med. J. *48*, 11–23.

Christiansen, A., Kringelum, J.V., Hansen, C.S., Bøgh, K.L., Sullivan, E., Patel, J., Rigby, N.M., Eiwegger, T., Szépfalusi, Z., de Masi, F., et al. (2015). High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum. Sci. Rep. *5*, 12913.

Cole, D.K., Edwards, E.S.J., Wynn, K.K., Clement, M., Miles, J.J., Ladell, K., Ekeruche, J., Gostick, E., Adams, K.J., Skowera, A., et al. (2010). Modification of MHC anchor residues generates heteroclitic peptides that alter TCR binding and T cell recognition. J. Immunol. *185*, 2600–2610.

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

Crotty, S. (2019). T follicular helper cell biology: A decade of discovery and diseases. Immunity *50*, 1132–1148.

Dai, Z., Huisman, B.D., Zeng, H., Carter, B., Jain, S., Birnbaum, M.E., and Gifford, D.K. (2021). Machine learning optimization of peptides for presentation by class II MHCs. Bioinformatics.

Dash, P., Fiore-Gartland, A.J., Hertz, T., Wang, G.C., Sharma, S., Souquette, A., Crawford, J.C., Clemens, E.B., Nguyen, T.H.O., Kedzierska, K., et al. (2017). Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature *547*, 89–93.

Davis, M.M., and Bjorkman, P.J. (1988). T-cell antigen receptor genes and T-cell recognition. Nature *334*, 395–402.

Dessen, A., Lawrence, C.M., Cupo, S., Zaller, D.M., and Wiley, D.C. (1997). X-ray crystal structure of HLA-DR4 (DRA*0101, DRB1*0401) complexed with a peptide from human collagen II. Immunity *7*, 473–481.

Dobson, C.S., Reich, A.N., Gaglione, S., Smith, B.E., Kim, E.J., Dong, J., Ronsard, L., Okonkwo, V., Lingwood, D., Dougan, M., et al. (2021). Antigen identification and high-throughput interaction mapping by reprogramming viral entry.

Fairhead, M., and Howarth, M. (2015). Site-specific biotinylation of purified proteins using BirA. Methods Mol. Biol. *1266*, 171–184.

Fallang, L.-E., Roh, S., Holm, A., Bergseng, E., Yoon, T., Fleckenstein, B., Bandyopadhyay, A., Mellins, E.D., and Sollid, L.M. (2009). Complexes of two cohorts of CLIP peptides and HLA-DQ2 of the autoimmune DR3-DQ2 haplotype are poor substrates for HLA-DM. J. Immunol. *182*, 726.1-726.

Fernandes, R.A., Li, C., Wang, G., Yang, X., Savvides, C.S., Glassman, C.R., Dong, S., Luxenberg, E., Sibener, L.V., Birnbaum, M.E., et al. (2020). Discovery of surrogate agonists for visceral fat Treg cells that modulate metabolic indices in vivo. Elife *9*.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Gal, Y., and Ghahramani, Z. (2015). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning.

Gambino, F., Jr, Tai, W., Voronin, D., Zhang, Y., Zhang, X., Shi, J., Wang, X., Wang, N., Du, L., and Qiao, L. (2021). A vaccine inducing solely cytotoxic T lymphocytes fully prevents Zika virus infection and fetal damage. Cell Rep. *35*, 109107.

Gee, M.H., Sibener, L.V., Birnbaum, M.E., Jude, K.M., Yang, X., Fernandes, R.A., Mendoza, J.L., Glassman, C.R., and Garcia, K.C. (2018a). Stress-testing the relationship between T cell receptor/peptide-MHC affinity and cross-reactivity using peptide velcro. Proc. Natl. Acad. Sci. U. S. A. *115*, E7369–E7378.

Gee, M.H., Han, A., Lofgren, S.M., Beausang, J.F., Mendoza, J.L., Birnbaum, M.E., Bethune, M.T., Fischer, S., Yang, X., Gomez-Eerland, R., et al. (2018b). Antigen identification for orphan T cell receptors expressed on tumor-infiltrating lymphocytes. Cell *172*, 549-563.e16.

Gonzalez, H., Hagerling, C., and Werb, Z. (2018). Roles of the immune system in cancer: from tumor initiation to metastatic progression. Genes Dev. *32*, 1267–1284.

Gonzalez-Galarza, F.F., McCabe, A., Santos, E.J.M.D., Jones, J., Takeshita, L., Ortega-Rivera, N.D., Cid-Pavon, G.M.D., Ramsbottom, K., Ghattaoraya, G., Alfirevic, A., et al. (2020). Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. Nucleic Acids Res. *48*, D783–D788.

Grace, B.E., Backlund, C.M., Morgan, D.M., Kang, B.H., Singh, N.K., Huisman, B.D., Rappazzo, C.G., Moynihan, K.D., Maiorino, L., Dobson, C.S., et al. (2022). Identification of highly cross-reactive mimotopes for a public T cell response in murine melanoma.

Gumá, M., Angulo, A., Vilches, C., Gómez-Lozano, N., Malats, N., and López-Botet, M. (2004). Imprint of human cytomegalovirus infection on the NK cell receptor repertoire. Blood *104*, 3664–3671.

Guo, X.-Z.J., and Elledge, S.J. (2022). V-CARMA: A tool for the detection and modification of antigen-specific T cells. Proc. Natl. Acad. Sci. U. S. A. *119*, e2116277119.

Guzman, M.G., Gubler, D.J., Izquierdo, A., Martinez, E., and Halstead, S.B. (2016). Dengue infection. Nat. Rev. Dis. Primers *2*, 16055.

van Hall, T., André, P., Horowitz, A., Ruan, D.F., Borst, L., Zerbib, R., Narni-Mancinelli, E., van der Burg, S.H., and Vivier, E. (2019). Monalizumab: inhibiting the novel immune checkpoint NKG2A. J. Immunother. Cancer *7*, 263.

Hammer, J., Valsasnini, P., Tolba, K., Bolin, D., Higelin, J., Takacs, B., and Sinigaglia, F. (1993). Promiscuous and allele-specific anchors in HLA-DR-binding peptides. Cell *74*, 197–203.

Hammer, J., Belunis, C., Bolin, D., Papadopoulos, J., Walsky, R., Higelin, J., Danho, W., Sinigaglia, F., and Nagy, Z.A. (1994). High-affinity binding of short peptides to major histocompatibility complex class II molecules by anchor combinations. Proc. Natl. Acad. Sci. U. S. A. *91*, 4456–4460.

Hammer, Q., Rückert, T., Borst, E.M., Dunst, J., Haubner, A., Durek, P., Heinrich, F., Gasparoni, G., Babic, M., Tomic, A., et al. (2018). Peptide-specific recognition of human cytomegalovirus strains controls adaptive natural killer cells. Nat. Immunol. *19*, 453–463.

Hanley, J.A., and McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology *143*, 29–36.

Hansen, S.G., Sacha, J.B., Hughes, C.M., Ford, J.C., Burwitz, B.J., Scholz, I., Gilbride, R.M., Lewis, M.S., Gilliam, A.N., Ventura, A.B., et al. (2013). Cytomegalovirus vectors violate CD8+ T cell epitope recognition paradigms. Science *340*, 1237874.

Hansen, S.G., Wu, H.L., Burwitz, B.J., Hughes, C.M., Hammond, K.B., Ventura, A.B., Reed, J.S., Gilbride, R.M., Ainslie, E., Morrow, D.W., et al. (2016). Broadly targeted $CD8^+$ T cell responses restricted by major histocompatibility complex E. Science *351*, 714–720.

Heinrichs, H., and Orr, H.T. (1990). HLA non-A,B,C class I genes: their structure and expression. Immunol. Res. *9*, 265–274.

Hellman, L.M., Yin, L., Wang, Y., Blevins, S.J., Riley, T.P., Belden, O.S., Spear, T.T., Nishimura, M.I., Stern, L.J., and Baker, B.M. (2016). Differential scanning fluorimetry based assessments of the thermal and kinetic stability of peptide-MHC complexes. J. Immunol. Methods *432*, 95–101.

Hennecke, J., and Wiley, D.C. (2001). T cell receptor-MHC interactions up close. Cell *104*, 1–4.

Hennecke, J., and Wiley, D.C. (2002). Structure of a complex of the human alpha/beta T cell receptor (TCR) HA1.7, influenza hemagglutinin peptide, and major histocompatibility complex class II molecule, HLA-DR4 (DRA*0101 and DRB1*0401): insight into TCR cross-restriction and alloreactivity. J. Exp. Med. *195*, 571–581.

Heyder, T., Kohler, M., Tarasova, N.K., Haag, S., Rutishauser, D., Rivera, N.V., Sandin, C., Mia, S., Malmström, V., Wheelock, Å.M., et al. (2016). Approach for identifying human leukocyte antigen (HLA)-DR bound peptides from scarce clinical samples. Mol. Cell. Proteomics *15*, 3017–3029.

Hoare, H.L., Sullivan, L.C., Pietra, G., Clements, C.S., Lee, E.J., Ely, L.K., Beddoe, T., Falco, M., Kjer-Nielsen, L., Reid, H.H., et al. (2006). Structural basis for a major histocompatibility complex class Ib-restricted T cell response. Nat. Immunol. *7*, 256–264.

Hoare, H.L., Sullivan, L.C., Clements, C.S., Ely, L.K., Beddoe, T., Henderson, K.N., Lin, J., Reid, H.H., Brooks, A.G., and Rossjohn, J. (2008). Subtle changes in peptide conformation profoundly affect recognition of the non-classical MHC class I molecule HLA-E by the CD94-NKG2 natural killer cell receptors. J. Mol. Biol. *377*, 1297–1303.

Houghton, C.S.B., Engelhorn, M.E., Liu, C., Song, D., Gregor, P., Livingston, P.O., Orlandi, F., Wolchok, J.D., McCracken, J., Houghton, A.N., et al. (2007). Immunological validation of the EpitOptimizer program for streamlined design of heteroclitic epitopes. Vaccine *25*, 5330–5342.

Hu, Z., Ott, P.A., and Wu, C.J. (2018). Towards personalized, tumour-specific, therapeutic vaccines for cancer. Nat. Rev. Immunol. *18*, 168–182.

Hu, Z., Leet, D.E., Allesøe, R.L., Oliveira, G., Li, S., Luoma, A.M., Liu, J., Forman, J., Huang, T., Iorgulescu, J.B., et al. (2021). Personal neoantigen vaccines induce persistent memory T cell responses and epitope spreading in patients with melanoma. Nat. Med. *27*, 515–525.

Huisman, B.D., Dai, Z., Gifford, D.K., and Birnbaum, M.E. (2022). A high-throughput yeast display approach to profile pathogen proteomes for MHC-II binding.

Irvine, D.J., Purbhoo, M.A., Krogsgaard, M., and Davis, M.M. (2002). Direct observation of ligand recognition by T cells. Nature *419*, 845–849.

Jawdat, D., Uyar, F.A., Alaskar, A., Müller, C.R., and Hajeer, A. (2020). HLA-A, -B, -C, -DRB1, -DQB1, and -DPB1 allele and haplotype frequencies of 28,927 Saudi stem cell donors typed by next-generation sequencing. Front. Immunol. *11*, 544768.

Jelcic, I., Al Nimer, F., Wang, J., Lentsch, V., Planas, R., Jelcic, I., Madjovski, A., Ruhrmann, S., Faigle, W., Frauenknecht, K., et al. (2018). Memory B cells activate brain-homing, autoreactive CD4+ T cells in multiple sclerosis. Cell *175*, 85-100.e23.

Jensen, K.K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J.A., Yan, Z., Sette, A., Peters, B., and Nielsen, M. (2018). Improved methods for predicting peptide binding affinity to MHC class II molecules. Immunology *154*, 394–406.

Jiang, W., and Boder, E.T. (2010). High-throughput engineering and analysis of peptide binding to class II MHC. Proc. Natl. Acad. Sci. U. S. A. *107*, 13258–13263.

Jiang, W., Strohman, M.J., Somasundaram, S., Ayyangar, S., Hou, T., Wang, N., and Mellins, E.D. (2015). pH-susceptibility of HLA-DO tunes DO/DM ratios to regulate HLA-DM catalytic activity. Sci. Rep. *5*, 17333.

Joglekar, A.V., Leonard, M.T., Jeppson, J.D., Swift, M., Li, G., Wong, S., Peng, S., Zaretsky, J.M., Heath, J.R., Ribas, A., et al. (2019). T cell antigen discovery via signaling and antigen-presenting bifunctional receptors. Nat. Methods *16*, 191–198.

Jones, E.Y., Fugger, L., Strominger, J.L., and Siebold, C. (2006). MHC class II proteins and disease: a structural perspective. Nat. Rev. Immunol. *6*, 271–282.

Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017). NetMHCpan-4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. J. Immunol. *199*, 3360–3368.

Justesen, S., Harndahl, M., Lamberth, K., Nielsen, L.-L.B., and Buus, S. (2009). Functional recombinant MHC class II molecules and high-throughput peptide-binding assays. Immunome Res. *5*, 2.

Kaiser, B.K., Pizarro, J.C., Kerns, J., and Strong, R.K. (2008). Structural basis for NKG2A/CD94 recognition of HLA-E. Proc. Natl. Acad. Sci. U. S. A. *105*, 6696–6701.

Karnes, J.H., Bastarache, L., Shaffer, C.M., Gaudieri, S., Xu, Y., Glazer, A.M., Mosley, J.D., Zhao, S., Raychaudhuri, S., Mallal, S., et al. (2017). Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. Sci. Transl. Med. *9*.

Kaushansky, N., and Ben-Nun, A. (2014). DQB1*06:02-associated pathogenic anti-myelin autoimmunity in multiple sclerosis-like disease: Potential function of DQB1*06:02 as a disease-predisposing allele. Front. Oncol. *4*, 280.

Kent, S.C., Mannering, S.I., Michels, A.W., and Babon, J.A.B. (2017). Deciphering the pathogenesis of human type 1 diabetes (T1D) by interrogating T cells from the "scene of the crime." Curr. Diab. Rep. *17*.

Keskin, D.B., Anandappa, A.J., Sun, J., Tirosh, I., Mathewson, N.D., Li, S., Oliveira, G., Giobbie-Hurder, A., Felt, K., Gjini, E., et al. (2019). Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. Nature *565*, 234–239.

Kingma, D.P., and Ba, J. (2014). Adam: A method for stochastic optimization.

Klinger, M., Pepin, F., Wilkins, J., Asbury, T., Wittkop, T., Zheng, J., Moorhead, M., and Faham, M. (2015). Multiplex Identification of Antigen-Specific T Cell Receptors Using a Combination of Immune Assays and Immune Receptor Sequencing. PLoS One *10*, e0141561.

Knapp, B., Giczi, V., Ribarics, R., and Schreiner, W. (2011). PeptX: using genetic algorithms to optimize peptides for MHC binding. BMC Bioinformatics *12*, 241.

Kraemer, T., Blasczyk, R., and Bade-Doeding, C. (2014). HLA-E: a novel player for histocompatibility. J. Immunol. Res. *2014*, 352160.

Kraemer, T., Celik, A.A., Huyton, T., Kunze-Schumacher, H., Blasczyk, R., and Bade-Döding, C. (2015). HLA-E: Presentation of a broader peptide repertoire impacts the cellular immune response-implications on HSCT outcome. Stem Cells Int. *2015*, 346714.

Kula, T., Dezfulian, M.H., Wang, C.I., Abdelfattah, N.S., Hartman, Z.C., Wucherpfennig, K.W., Lyerly, H.K., and Elledge, S.J. (2019). T-Scan: A genome-wide method for the systematic discovery of T cell epitopes. Cell *178*, 1016-1028.e13.

Kulski, J.K., Shiina, T., and Dijkstra, J.M. (2019). Genomic diversity of the Major Histocompatibility Complex in health and disease. Cells *8*, 1270.

Kumar, B.V., Connors, T.J., and Farber, D.L. (2018). Human T cell development, localization, and function throughout life. Immunity *48*, 202–213.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2016). Simple and scalable predictive uncertainty estimation using deep ensembles.

Lampen, M.H., Hassan, C., Sluijter, M., Geluk, A., Dijkman, K., Tjon, J.M., de Ru, A.H., van der Burg, S.H., van Veelen, P.A., and van Hall, T. (2013). Alternative peptide repertoire of HLA-E reveals a binding motif that is strikingly similar to HLA-A2. Mol. Immunol. *53*, 126–131.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.

Lin, H.H., Zhang, G.L., Tongchusak, S., Reinherz, E.L., and Brusic, V. (2008). Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. BMC Bioinformatics *9 Suppl 12*, S22.

Liu, G., Carter, B., Bricken, T., Jain, S., Viard, M., Carrington, M., and Gifford, D.K. (2020). Computationally optimized SARS-CoV-2 MHC class I and II vaccine formulations predicted to target human haplotype distributions. Cell Syst. *11*, 131-144.e6.

Liu, G., Carter, B., and Gifford, D.K. (2021a). Predicted cellular immunity population coverage gaps for SARS-CoV-2 subunit vaccines and their augmentation by compact peptide sets. Cell Syst. *12*, 102-107.e4.

Liu, R., Jiang, W., and Mellins, E.D. (2021b). Yeast display of MHC-II enables rapid identification of peptide ligands from protein antigens (RIPPA). Cell. Mol. Immunol. *18*, 1847–1860.

Lovitch, S.B., Pu, Z., and Unanue, E.R. (2006). Amino-terminal flanking residues determine the conformation of a peptide-class II MHC complex. J. Immunol. *176*, 2958–2968.

Ma, K.-Y., Schonnesen, A.A., He, C., Xia, A.Y., Sun, E., Chen, E., Sebastian, K.R., Guo, Y.-W., Balderas, R., Kulkarni-Date, M., et al. (2021). High-throughput and high-dimensional single-cell analysis of antigen-specific CD8+ T cells. Nat. Immunol. *22*, 1590–1598.

Ma, M., Wang, Z., Chen, X., Tao, A., He, L., Fu, S., Zhang, Z., Fu, Y., Guo, C., Liu, J., et al. (2017). NKG2C+NKG2A- natural killer cells are associated with a lower viral set point and may predict disease progression in individuals with primary HIV infection. Front. Immunol. *8*, 1176.

Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. *47*, W636–W641.

Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*, 2957–2963.

Mahaweni, N.M., Ehlers, F.A.I., Sarkar, S., Janssen, J.W.H., Tilanus, M.G.J., Bos, G.M.J., and Wieten, L. (2018). NKG2A expression is not per se detrimental for the anti-multiple myeloma activity of activated natural killer cells in an in vitro system mimicking the tumor microenvironment. Front. Immunol. *9*.

Marsh, S.G.E., Parham, P., and Barber, L.D. (1999). The HLA FactsBook (San Diego, CA: Academic Press).

Masella, A.P., Bartram, A.K., Truszkowski, J.M., Brown, D.G., and Neufeld, J.D. (2012). PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformatics *13*, 31.

Mateus, J., Grifoni, A., Tarke, A., Sidney, J., Ramirez, S.I., Dan, J.M., Burger, Z.C., Rawlings, S.A., Smith, D.M., Phillips, E., et al. (2020). Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. Science *370*, 89–94.

Mathewson, N.D., Ashenberg, O., Tirosh, I., Gritsch, S., Perez, E.M., Marx, S., Jerby-Arnon, L., Chanoch-Myers, R., Hara, T., Richman, A.R., et al. (2021). Inhibitory CD161 receptor identified in glioma-infiltrating T cells by single-cell analysis. Cell *184*, 1281-1298.e26.

Matsui, K., Boniface, J.J., Reay, P.A., Schild, H., Fazekas de St Groth, B., and Davis, M.M. (1991). Low affinity interaction of peptide-MHC complexes with T cell receptors. Science *254*, 1788–1791.

Midelfort, K.S., Hernandez, H.H., Lippow, S.M., Tidor, B., Drennan, C.L., and Wittrup, K.D. (2004). Substantial energetic improvement with minimal structural perturbation in a high affinity mutant antibody. J. Mol. Biol. *343*, 685–701.

Miller, J.D., Weber, D.A., Ibegbu, C., Pohl, J., Altman, J.D., and Jensen, P.E. (2003). Analysis of HLA-E peptide-binding specificity and contact residues in bound peptide required for recognition by CD94/NKG2. J. Immunol. *171*, 1369–1375.

Moise, L., Gutierrez, A., Kibria, F., Martin, R., Tassone, R., Liu, R., Terry, F., Martin, B., and De Groot, A.S. (2015). iVAX: An integrated toolkit for the selection and optimization of antigens and the design of epitope-driven vaccines. Hum. Vaccin. Immunother. *11*, 2312–2321.

van Montfoort, N., Borst, L., Korrer, M.J., Sluijter, M., Marijt, K.A., Santegoets, S.J., van Ham, V.J., Ehsan, I., Charoentong, P., André, P., et al. (2018). NKG2A blockade potentiates CD8 T cell immunity induced by cancer vaccines. Cell *175*, 1744-1755.e15.

Neefjes, J., Jongsma, M.L.M., Paul, P., and Bakke, O. (2011). Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nat. Rev. Immunol. *11*, 823–836.

Nielsen, M., and Andreatta, M. (2017). NNAlign: a platform to construct and evaluate artificial neural network models of receptor-ligand interactions. Nucleic Acids Res. *45*, W344–W349.

Nielsen, M., and Lund, O. (2009). NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC Bioinformatics *10*, 296.

Nielsen, M., Lund, O., Buus, S., and Lundegaard, C. (2010). MHC class II epitope predictive algorithms. Immunology *130*, 319–328.

Obermair, F.J., Renoux, F., Heer, S., Lee, C., Cereghetti, N., Maestri, G., Haldner, Y., Wuigk, R., Iosefson, O., Patel, P., et al. (2021). High resolution profiling of MHC-II peptide presentation capacity, by Mammalian Epitope Display, reveals SARS-CoV-2 targets for CD4 T cells and mechanisms of immune-escape (bioRxiv).

O'Brien, C., Flower, D.R., and Feighery, C. (2008). Peptide length significantly influences in vitro affinity for MHC class II molecules. Immunome Res. *4*, 6.

O'Callaghan, C.A., and Bell, J.I. (1998). Structure and function of the human MHC class Ib molecules HLA-E, HLA-F and HLA-G. Immunol. Rev. *163*, 129–138.

O'Callaghan, C.A., Tormo, J., Willcox, B.E., Braud, V.M., Jakobsen, B.K., Stuart, D.I., McMichael, A.J., Bell, J.I., and Jones, E.Y. (1998). Structural features impose tight peptide binding specificity in the nonclassical MHC molecule HLA-E. Mol. Cell *1*, 531–541.

O'Donnell, T.J., Rubinsteyn, A., Bonsack, M., Riemer, A.B., Laserson, U., and Hammerbacher, J. (2018). MHCflurry: Open-source class I MHC binding affinity prediction. Cell Syst. *7*, 129-132.e4.

O'Donnell, T.J., Rubinsteyn, A., and Laserson, U. (2020). MHCflurry 2.0: Improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. Cell Syst. *11*, 42-48.e7.

Oliveira, G., Stromhaug, K., Klaeger, S., Kula, T., Frederick, D.T., Le, P.M., Forman, J., Huang, T., Li, S., Zhang, W., et al. (2021). Phenotype, specificity and avidity of antitumour CD8+ T cells in melanoma. Nature *596*, 119–125.

Ott, P.A., Hu, Z., Keskin, D.B., Shukla, S.A., Sun, J., Bozym, D.J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L., et al. (2017). An immunogenic personal neoantigen vaccine for patients with melanoma. Nature *547*, 217–221.

Pan, Y.-G., Aiamkitsumrit, B., Bartolo, L., Wang, Y., Lavery, C., Marc, A., Holec, P.V., Rappazzo, C.G., Eilola, T., Gimotty, P.A., et al. (2021). Vaccination reshapes the virus-specific T cell repertoire in unexposed adults. Immunity *54*, 1245-1256.e5.

Parker, R., Partridge, T., Wormald, C., Kawahara, R., Stalls, V., Aggelakopoulou, M., Parker, J., Powell Doherty, R., Ariosa Morejon, Y., Lee, E., et al. (2021). Mapping the SARS-CoV-2 spike glycoprotein-derived peptidome presented by HLA class II on dendritic cells. Cell Rep. *35*, 109179.

Patronov, A., and Doytchinova, I. (2013). T-cell epitope vaccine design by immunoinformatics. Open Biol. *3*, 120139.

Piertney, S.B., and Oliver, M.K. (2006). The evolutionary ecology of the major histocompatibility complex. Heredity (Edinb.) *96*, 7–21.

Pietra, G., Romagnani, C., Falco, M., Vitale, M., Castriconi, R., Pende, D., Millo, E., Anfossi, S., Biassoni, R., Moretta, L., et al. (2001). The analysis of the natural killer-like activity of human cytolytic T lymphocytes revealed HLA-E as a novel target for TCR alpha/beta-mediated recognition. Eur. J. Immunol. *31*, 3687–3693.

Pietra, G., Romagnani, C., Mazzarino, P., Falco, M., Millo, E., Moretta, A., Moretta, L., and Mingari, M.C. (2003). HLA-E-restricted recognition of cytomegalovirus-derived peptides by human CD8+ cytolytic T lymphocytes. Proc. Natl. Acad. Sci. U. S. A. *100*, 10896–10901.

Pietra, G., Romagnani, C., Manzini, C., Moretta, L., and Mingari, M.C. (2010). The emerging role of HLA-E-restricted CD8+ T lymphocytes in the adaptive immune response to pathogens and tumors. J. Biomed. Biotechnol. *2010*, 907092.

Planas, R., Santos, R., Tomas-Ojer, P., Cruciani, C., Lutterotti, A., Faigle, W., Schaeren-Wiemers, N., Espejo, C., Eixarch, H., Pinilla, C., et al. (2018). GDP-l-fucose synthase is a CD4+ T cell-specific autoantigen in DRB3*02:02 patients with multiple sclerosis. Sci. Transl. Med. *10*, eaat4301.

Pugliese, A., Boulware, D., Yu, L., Babu, S., Steck, A.K., Becker, D., Rodriguez, H., DiMeglio, L., Evans-Molina, C., Harrison, L.C., et al. (2016). HLA-DRB1*15:01-DQA1*01:02-DQB1*06:02 haplotype protects autoantibody-positive relatives from type 1 diabetes throughout the stages of disease progression. Diabetes *65*, 1109–1119.

Purcell, A.W., Ramarathinam, S.H., and Ternette, N. (2019). Mass spectrometry-based identification of MHC-bound peptides for immunopeptidomics. Nat. Protoc. *14*, 1687–1707.

Racle, J., Michaux, J., Rockinger, G.A., Arnaud, M., Bobisse, S., Chong, C., Guillaume, P., Coukos, G., Harari, A., Jandus, C., et al. (2019). Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. Nat. Biotechnol. *37*, 1283–1286.

Rammensee, H., Bachmann, J., Emmerich, N.P., Bachor, O.A., and Stevanović, S. (1999). SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics *50*, 213–219.

Rappazzo, C.G., Huisman, B.D., and Birnbaum, M.E. (2020). Repertoire-scale determination of class II MHC peptide binding via yeast display improves antigen prediction. Nat. Commun. *11*, 4414.

Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. *48*, W449–W454.

Ribas, A., and Wolchok, J.D. (2018). Cancer immunotherapy using checkpoint blockade. Science *359*, 1350–1355.

Robert, C., Schachter, J., Long, G.V., Arance, A., Grob, J.J., Mortier, L., Daud, A., Carlino, M.S., McNeil, C., Lotem, M., et al. (2015). Pembrolizumab versus ipilimumab in advanced melanoma. N. Engl. J. Med. *372*, 2521–2532.

Robins, H.S., Srivastava, S.K., Campregher, P.V., Turtle, C.J., Andriesen, J., Riddell, S.R., Carlson, C.S., and Warren, E.H. (2010). Overlap and effective size of the human CD8+ T cell receptor repertoire. Sci. Transl. Med. *2*, 47ra64.

Robinson, J., Halliwell, J.A., Hayhurst, J.D., Flicek, P., Parham, P., and Marsh, S.G.E. (2015). The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. *43*, D423-31.

Roche, P.A., and Furuta, K. (2015). The ins and outs of MHC class II-mediated antigen processing and presentation. Nat. Rev. Immunol. *15*, 203–216.

Rosati, E., Pogorelyy, M.V., Minervina, A.A., Scheffold, A., Franke, A., Bacher, P., and Thomas, P.G. (2021). Characterization of SARS-CoV-2 public CD4+ αβ T cell clonotypes through reverse epitope discovery. BioRxivorg.

Ruibal, P., Franken, K.L.M.C., van Meijgaarden, K.E., van Loon, J.J.F., van der Steen, D., Heemskerk, M.H.M., Ottenhoff, T.H.M., and Joosten, S.A. (2020). Peptide binding to HLA-E molecules in humans, nonhuman primates, and mice reveals unique binding peptides but remarkably conserved anchor residues. J. Immunol. *205*, 2861–2872.

Sahin, U., Derhovanessian, E., Miller, M., Kloke, B.-P., Simon, P., Löwer, M., Bukur, V., Tadmor, A.D., Luxemburger, U., Schrörs, B., et al. (2017). Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature *547*, 222–226.

Sarkizova, S., Klaeger, S., Le, P.M., Li, L.W., Oliveira, G., Keshishian, H., Hartigan, C.R., Zhang, W., Braun, D.A., Ligon, K.L., et al. (2020). A large peptidome dataset improves HLA class I epitope prediction across most of the human population. Nat. Biotechnol. *38*, 199–209.

Sasaki, T., Kanaseki, T., Shionoya, Y., Tokita, S., Miyamoto, S., Saka, E., Kochin, V., Takasawa, A., Hirohashi, Y., Tamura, Y., et al. (2016). Microenvironmental stresses induce HLA-E/Qa-1 surface expression and thereby reduce CD8(+) T-cell recognition of stressed cells. Eur. J. Immunol. *46*, 929–940.

Scally, S.W., Law, S.-C., Ting, Y.T., van Heemst, J., Sokolove, J., Deutsch, A.J., Bridie Clemens, E., Moustakas, A.K., Papadopoulos, G.K., van der Woude, D., et al. (2017). Molecular basis for increased susceptibility of Indigenous North Americans to seropositive rheumatoid arthritis. Ann. Rheum. Dis. *76*, 1915–1923.

Sette, A., Sidney, J., Oseroff, C., del Guercio, M.F., Southwood, S., Arrhenius, T., Powell, M.F., Colón, S.M., Gaeta, F.C., and Grey, H.M. (1993). HLA DR4w4-binding motifs illustrate the biochemical basis of degeneracy and specificity in peptide-DR interactions. J. Immunol. *151*, 3163–3170.

Sibener, L.V., Fernandes, R.A., Kolawole, E.M., Carbone, C.B., Liu, F., McAffee, D., Birnbaum, M.E., Yang, X., Su, L.F., Yu, W., et al. (2018). Isolation of a structural mechanism for uncoupling T cell receptor signaling from peptide-MHC binding. Cell *174*, 672-687.e27.

Sidney, J., Steen, A., Moore, C., Ngo, S., Chung, J., Peters, B., and Sette, A. (2010). Divergent motifs but overlapping binding repertoires of six HLA-DQ molecules frequently expressed in the worldwide human population. J. Immunol. *185*, 4189–4198.

Snyder, T.M., Gittelman, R.M., Klinger, M., May, D.H., Osborne, E.J., Taniguchi, R., Zahid, H.J., Kaplan, I.M., Dines, J.N., Noakes, M.T., et al. (2020). Magnitude and dynamics of the T-cell response to SARS-CoV-2 infection at both individual and population levels. MedRxiv.

Sommer, S. (2005). The importance of immune gene variability (MHC) in evolutionary ecology and conservation. Front. Zool. *2*, 16.

Srivastava, N. (2014). Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res *15*, 1929–1958.

Stabile, H., Fionda, C., Gismondi, A., and Santoni, A. (2017). Role of distinct natural killer cell subsets in anticancer response. Front. Immunol. *8*, 293.

Stern, L.J. (1994). Crystal structure of the human class II MHC protein HLA- DR1 complexed with an influenza virus peptide. Nature *368*, 215–221.

Stevens, J., Joly, E., Trowsdale, J., and Butcher, G.W. (2001). Peptide binding characteristics of the non-classical class Ib MHC molecule HLA-E assessed by a recombinant random peptide approach. BMC Immunol. *2*, 5.

van Stipdonk, M.J.B., Badia-Martinez, D., Sluijter, M., Offringa, R., van Hall, T., and Achour, A. (2009). Design of agonistic altered peptides for the robust induction of CTL directed towards H-2Db in complex with the melanoma-associated epitope gp100. Cancer Res. *69*, 7784–7792.

Stopfer, L.E., Mesfin, J.M., Joughin, B.A., Lauffenburger, D.A., and White, F.M. (2020). Multiplexed relative and absolute quantitative immunopeptidomics reveals MHC I repertoire alterations induced by CDK4/6 inhibition. Nat. Commun. *11*, 2760.

Stopfer, L.E., Gajadhar, A.S., Patel, B., Gallien, S., Frederick, D.T., Boland, G.M., Sullivan, R.J., and White, F.M. (2021). Absolute quantification of tumor antigens using embedded MHC-I isotopologue calibrants. Proc. Natl. Acad. Sci. U. S. A. *118*, e2111173118.

Strong, R.K., Holmes, M.A., Li, P., Braun, L., Lee, N., and Geraghty, D.E. (2003). HLA-E Allelic variants. J. Biol. Chem. *278*, 5082–5090.

Sullivan, L.C., Walpole, N.G., Farenc, C., Pietra, G., Sum, M.J.W., Clements, C.S., Lee, E.J., Beddoe, T., Falco, M., Mingari, M.C., et al. (2017). A conserved energetic footprint underpins recognition of human leukocyte antigen-E by two distinct αβ T cell receptors. J. Biol. Chem. *292*, 21149–21158.

Sun, J.C., and Lanier, L.L. (2009). Natural killer cells remember: an evolutionary bridge between innate and adaptive immunity? Eur. J. Immunol. *39*, 2059–2064.

Swain, S.L., McKinstry, K.K., and Strutt, T.M. (2012). Expanding roles for $CD4^+$ T cells in immunity to viruses. Nat. Rev. Immunol. *12*, 136–148.

Thomsen, M.C.F., and Nielsen, M. (2012). Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. Nucleic Acids Res. *40*, W281-7.

Tokić, S., Žižkova, V., Štefanić, M., Glavaš-Obrovac, L., Marczi, S., Samardžija, M., Sikorova, K., and Petrek, M. (2020). HLA-A, -B, -C, -DRB1, -DQA1, and -DQB1 allele and haplotype frequencies defined by next generation sequencing in a population of East Croatia blood donors. Sci. Rep. *10*, 5513.

Ulbrecht, M., Martinozzi, S., Grzeschik, M., Hengel, H., Ellwart, J.W., Pla, M., and Weiss, E.H. (2000). Cutting edge: the human cytomegalovirus UL40 gene product contains a ligand for HLA-E and prevents NK cell-mediated lysis. J. Immunol. *164*, 5019–5022.

Vacic, V., Iakoucheva, L.M., and Radivojac, P. (2006). Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics *22*, 1536–1537.

Valés-Gómez, M., Reyburn, H.T., Erskine, R.A., López-Botet, M., and Strominger, J.L. (1999). Kinetics and peptide dependency of the binding of the inhibitory NK receptor CD94/NKG2-A and the activating receptor CD94/NKG2-C to HLA-E. EMBO J. *18*, 4250–4260.

Van Deventer, J.A., Kelly, R.L., Rajan, S., Wittrup, K.D., and Sidhu, S.S. (2015). A switchable yeast display/secretion system. Protein Eng. Des. Sel. *28*, 317–325.

Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A., and Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. *47*, D339–D343.

Vyas, J.M., Van der Veen, A.G., and Ploegh, H.L. (2008). The known unknowns of antigen processing and presentation. Nat. Rev. Immunol. *8*, 607–618.

Walters, L.C., Harlos, K., Brackenridge, S., Rozbesky, D., Barrett, J.R., Jain, V., Walter, T.S., O'Callaghan, C.A., Borrow, P., Toebes, M., et al. (2018). Pathogen-derived HLA-E bound epitopes reveal broad primary anchor pocket tolerability and conformationally malleable peptide binding. Nat. Commun. *9*.

Wang, P., Sidney, J., Dow, C., Mothé, B., Sette, A., and Peters, B. (2008). A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. PLoS Comput. Biol. *4*, e1000048.

Wernersson, R. (2006). Virtual Ribosome--a comprehensive DNA translation tool with support for integration of sequence feature annotation. Nucleic Acids Res. *34*, W385-8.

Wieczorek, M., Abualrous, E.T., Sticht, J., Álvaro-Benito, M., Stolzenberg, S., Noé, F., and Freund, C. (2017). Major histocompatibility complex (MHC) class I and MHC class II proteins: Conformational plasticity in antigen presentation. Front. Immunol. *8*, 292.

Wooldridge, L., Ekeruche-Makinde, J., van den Berg, H.A., Skowera, A., Miles, J.J., Tan, M.P., Dolton, G., Clement, M., Llewellyn-Lacey, S., Price, D.A., et al. (2012). A single autoimmune T cell receptor recognizes more than a million different peptides. J. Biol. Chem. *287*, 1168–1177.

Wu, R., Forget, M.-A., Chacon, J., Bernatchez, C., Haymaker, C., Chen, J.Q., Hwu, P., and Radvanyi, L.G. (2012). Adoptive T-cell therapy using autologous tumor-infiltrating lymphocytes for metastatic melanoma: current status and future outlook. Cancer J. *18*, 160–175.

Wucherpfennig, K.W. (2010). The first structures of T cell receptors bound to peptide-MHC. J. Immunol. *185*, 6391–6393.

Yang, H., Rei, M., Brackenridge, S., Brenna, E., Sun, H., Abdulhaqq, S., Liu, M.K.P., Ma, W., Kurupati, P., Xu, X., et al. (2021). HLA-E-restricted, Gag-specific CD8+ T cells can suppress HIV-1 infection, offering vaccine opportunities. Sci. Immunol. *6*.

Yin, L., and Stern, L.J. (2014). Measurement of peptide binding to MHC class II molecules by fluorescence polarization. Curr. Protoc. Immunol. *106*, 5.10.1-5.10.12.

Zavala-Ruiz, Z., Strug, I., Anderson, M.W., Gorski, J., and Stern, L.J. (2004). A polymorphic pocket at the P10 position contributes to peptide binding specificity in class II MHC proteins. Chem. Biol. *11*, 1395–1402.

Zeng, H., and Gifford, D.K. (2019). Quantification of uncertainty in peptide-MHC binding prediction improves high-affinity peptide selection for therapeutic design. Cell Syst. *9*, 159-166.e3.

Zhao, W., and Sher, X. (2018). Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes. PLoS Comput. Biol. *14*, e1006457.

Zheng, H., Lu, R., Xie, S., Wen, X., Wang, H., Gao, X., and Guo, L. (2015). Human leukocyte antigen-E alleles and expression in patients with serous ovarian cancer. Cancer Sci. *106*, 522–528.

(2017). The problem with neoantigen prediction. Nat. Biotechnol. *35*, 97–97.