Dynamics of Antigen Processing and Presentation in the Context of Cancer

By

Benjamin R. Weeder

MPH, Oregon State University, 2018
B.S., Pacific University, 2016

A DISSERTATION

Presented to the Department of Biomedical Engineering of the Oregon Health & Science University School of Medicine in partial fulfillment of the requirements for the degree of Doctor of Philosophy

September 2023

Table of Contents

Acknowledgements

The work described within would not have been possible without the consistent and loving support of my friends, family, and coworkers. Without their encouragement throughout the PhD process, and well before, I would not be where I am today.

Specifically, I want to acknowledge the support of my parents, John and Tracie Weeder, as well as the support of my step-mom Monica Weeder; all of who fostered my inquisitive nature and love of science, supported me in my times of need, and believed in me throughout the challenges I've faced. I also want to thank my roommates Kevin Jian and Genee Morden who have provided support and encouragement throughout my PhD journey and the Covid-19 pandemic.

My friends and co-workers have also provided a tremendous amount of support and feedback throughout my journey. I'd like to thank Nick and Michelle Calistri, Mary Wood, Austin Nguyen, and Maya Williams-Young for all the good food, camping trips and calming laughs shared throughout the years, hopefully with many more to come. The entire Thompson Lab also deserves acknowledgement for the helpful feedback they've provided throughout my dissertation and PhD processes.

I also want to thank my mentor, Reid Thompson, as well as our collaborators Abhi Nellore and Amy Moran, who have all provided invaluable advice and supported my development as a scientist. While there are too many people to fully list, I also cherish the interactions I've had with other collaborators and students throughout my schooling and the many engaging conversations we've shared that have expanded my views and helped me develop an appreciation for science well beyond my own work and focus.

Dissertation Abstract

Immunotherapeutic approaches, aimed at modulating the immune system's response to cancer and other diseases, have garnered increasing attention over the past decade. For some cancers such as melanoma, non-small cell lung cancer, and B-cell lymphomas, immunotherapy options have been transformative and some patients respond even in the most aggressive and refractory cases. Despite the optimism surrounding various immune-based approaches, many patients still don't respond to treatment and the reasons why are poorly understood. The disparity in responses across patients and cancer types has highlighted a dire need for a better mechanistic understanding of the factors involved in therapeutic response to immunotherapies; supported by tools that help define which patients will benefit from these novel treatment options. While a variety of approaches exist in the immunotherapy space, some with systemic affects and others focused on specific immune targets, all approaches rely heavily on the adaptive immune system and response to class I antigens.

In chapter I, pepsickle rapidly and accurately predicts proteasomal cleavage sites for improved neoantigen identification, I present my peer reviewed work highlighting `pepsickle`, an open-source tool for predicting cleavage sites during protein degradation. The process of peptide cleavage by the proteasome is a fundamental precursor to class I antigen presentation and ultimately shapes the pool of available targets for immune recognition. Unfortunately, at the time of publication few cleavage prediction tools existed and those that did were heavily outdated. Through the work highlighted in this chapter, we demonstrated that cleavage sites can be accurately characterized through deep-learning approaches, and further show that by filtering target candidates based on cleavage likelihood we can enrich the candidate pool for truly immunogenic peptides.

In chapter II, Predicting T-cell Cross-Reactivity using Paired Peptide Data: A Multi-Layer Perceptron Approach, I take a look at the risks associated with targeted immunotherapeutic approaches through the lens of T-cell cross-reactivity. Degeneracy in T-cell receptors, or the ability for T-cell receptors to recognize more than one presented class I epitope, is a fundamental part of T-cell evolution but can also have dire consequences if not properly considered during the target selection process. While there are many emerging approaches aimed at better characterizing when and how T-cell cross reactivity occurs, current tools largely rely on in-depth sequencing of T-cell receptors in conjunction with cognate antigen identification. Although these approaches help give fundamental insight into the mechanics of T-cell cross-reactivity, the direct use of T-cell sequences severely limits the application scope of such tools. Instead, I demonstrate a proof-of-concept approach that leverages cross-reactive epitope pairs to infer key features of cross-reactivity without the direct use of TCR sequences. Using paired epitope data allows for broad application of cross-reactivity predictions to a variety of important contexts where T-cell receptors cannot be exhaustively sequenced such as during the development of mRNA vaccine-based approaches. This method has broad applications outside of cancer, including for viral vaccine development and in the investigation of autoimmune disorders.

Finally, in chapter III, Deciphering the Prostate Tumor Microenvironment: Transcriptional Insights into Therapy Response following Androgen Axis Blockade and Immune Checkpoint Inhibition, I look at the tumor microenvironment of prostate cancer through the lens of single-cells transcriptomics. Using temporal samples from previously unpublished clinical trial data, we compare treatment naïve patient samples to paired samples taken after a treatment course of androgen axis-inhibition and immune checkpoint therapy. We identify treatment responses in identified malignant cells and characterize increases in class I antigen presentation

after treatment. We further leverage single cell techniques to characterize immune populations residing in tumor and non-tumor tissues and understand how they contribute to pro-inflammatory signaling and potential tissue dysregulation after treatment.

While chapters I and II focus on specific mechanisms of antigen presentation and recognition, chapter III takes a step back and contextualizes the environment in which these focused mechanisms occur. Each step in the antigen processing, presentation, and recognition pathway is complex and multi-faceted. While computational approaches can further our understanding of biological processes and help narrow our focus for potential follow up, broader context is often neglected. The milieu of activating and suppressing factors found in biological tissues ultimately shape the broader immune response in cancer and during other host challenges. Taken together, this work details the development of novel tools that help us better characterize immunological processes and provides insights into how malignant cells and surrounding immune populations respond and interact in the context of the broader tumor microenvironment.

Introduction

1.1 Class I Antigen processing, presentation, and relevance in cancer

1.1.1 The importance of class I antigen presentation and adaptive immune response

The adaptive immune system plays a fundamental role in our ability to respond to host challenges, from viral infection to tumor clearance and more. As viruses evolve and tumors mutate, our adaptive response is what allows us to keep up with the ever-changing landscape of threats to our wellbeing and physical survival. As such, our ability to efficiently and accurately present antigens, or processed protein fragments, to the immune system for surveillance is instrumental in mounting a protective response.

While there are multiple ways of presenting foreign antigens, the presentation of antigens via class I major histocompatibility complexes (MHC's) is ubiquitous across nucleated cells[1]. The presentation of class I antigens starts with the marking of intra-cellular proteins for degradation, followed by cleavage of proteins into fragments by the proteasome, then transport to the endoplasmic reticulum, trimming, and mounting of antigens on MHC class I complexes (Figure 1.1)[2]. These mounted antigens, called class I epitopes, and their bound MHC class I complex are subsequently trafficked to the cell surface where they are presented for interaction with surrounding immune cells[3]. Although this process presents foreign and mutated peptides, normal self-peptides are also processed and make up the majority of presented epitopes, allowing immune cell surveillance of both normal and abnormal epitopes[4].

Furthermore, a variety of factors can influence which epitopes are ultimately presented on the cellular surface for further surveillance. Individuals co-expresses up to 6 different classical MHC class I complexes, each with a unique binding preference that alters which epitopes are ultimately bound[5]. Ubiquitination, one key process by which proteins are marked for degradation, also plays a key role in determining which proteins are processed and at what rate[6]. Cleavage preference by the proteasome can further affect what fragments of proteins ultimately become epitopes; a process that can be complicated by the expression of IFN-gamma which can induce the expression of alternative proteasomal subunits shifting cleavage preferences and resulting in the assembly of the immuno-proteasome[7].

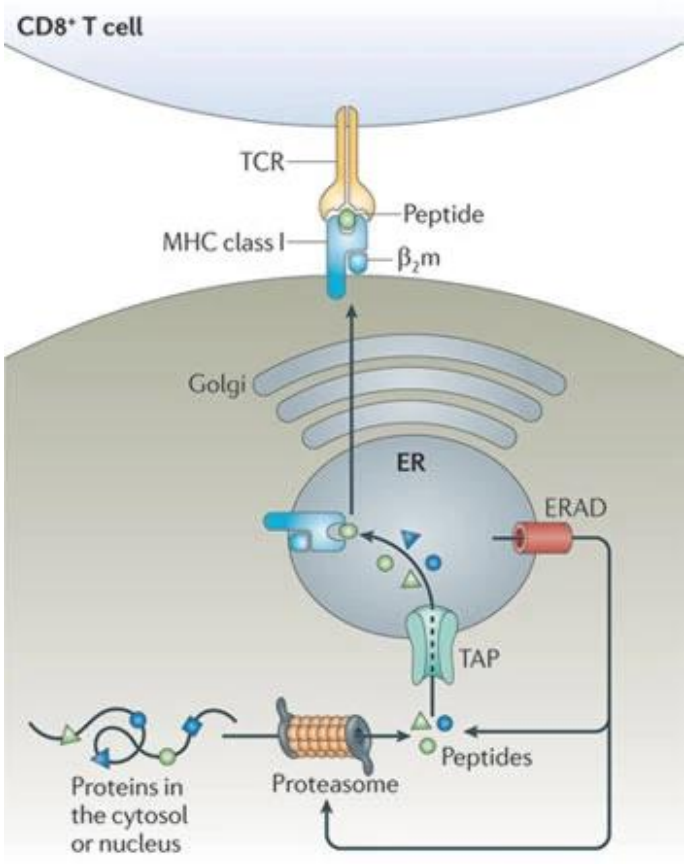


Figure 1.1. Overview of antigen presentation. *The presentation of intracellular antigenic peptides by MHC class I molecules is the result of a series of reactions. First, antigens are degraded by the proteasome. Then, the resulting peptides are translocated via transporter associated with antigen presentation (TAP) into the endoplasmic reticulum (ER) lumen and loaded onto MHC class I molecules. Peptide–MHC class I complexes are released from the ER and transported via the Golgi to the plasma membrane for antigen presentation to CD8+ T cells. Figure and caption adapted from Neefjes et al. (2011)*[2].

1.1.2 Antigen presentation in the context of cancer

Antigen presentation also has an important role in the context of cancer. As mutations accumulate due to DNA damage or replication error, cells must not only circumvent apoptotic processes, but also avoid detection by the immune system. While resistance to apoptosis has long been understood to be a hallmark of cancer as proposed by Hanahan and Weinberg over 20 years ago, immune evasion has only recently become appreciated for its core role in the

development and progression of cancer (Figure 1.2)[8]. This core contribution to cancer development is demonstrated by the fact that tumors consistently evolve multiple mechanisms for facilitating immune evasion, including creation of a suppressive tumor microenvironment through recruitment of fibroblasts and other immunosuppressive cells, downregulation the MHC expression, and upregulation inhibitory molecules that impede T-cell activation[9]–[11]. Under normal conditions, the downregulation of classical MHC molecules often seen in cancer development can induce natural killer (NK) mediated killing of tumor cells, however tumor adaptations such as expression of non-classical MHC's and shedding of inhibitory molecules such as MIC A and MIC B can inhibit killing by NK cells even in the when low levels of classical MHC molecules are expressed[12]. This concurrent down regulation of presented targets and upregulation of inhibitory molecules works in conjunction with the immuno-suppressive microenvironment to minimize adaptive immune responses to cancer even when potentially reactive immune cells are present.

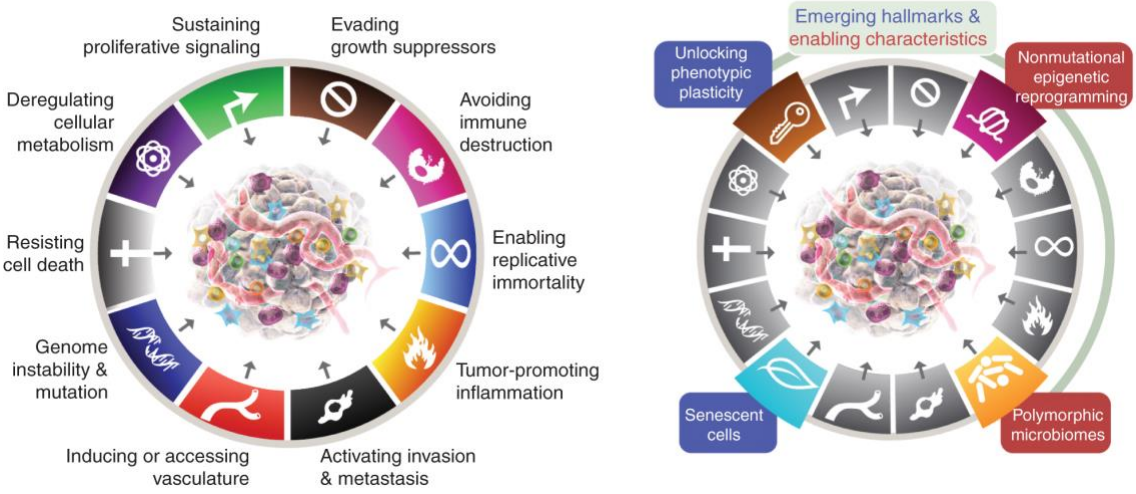


Figure 1.2. The Hallmarks of Cancer, circa 2022*. Left, the Hallmarks of Cancer currently embody eight hallmark capabilities and two enabling characteristics. In addition to the six acquired capabilities—Hallmarks of Cancer—proposed in 2000 (1), the two provisional "emerging hallmarks" introduced in 2011 (2)—cellular energetics (now described more broadly as "reprogramming cellular metabolism") and "avoiding immune destruction"—have been sufficiently validated to be considered part of the core set. Figure and caption adapted from Hanahan (2022)*[8].

The clonal evolution of cancer cells can also impact the landscape of presented antigens. Early in tumor development clonal populations that are highly immunogenic can be weeded out by early immune responses[13]. While it's possible for full tumor clearance to occur at this stage, the preferential removal of highly immunogenic clones can also create space for less immunogenic clones to thrive, even if their general fitness was lower than other clones previously present in the tumor microenvironment[14]. Continued pressure by the immune system can then ultimately lead to the evolution of immune resistant clones that continue to progress and result in further tumor growth and disease advancement. These steps, often termed tumor elimination, tumor equilibrium, and tumor escape, constitute the key components of tumor immuno-editing; one of the fundamental processes that contributes to hallmark immune evasion (Figure 1.3)[15].

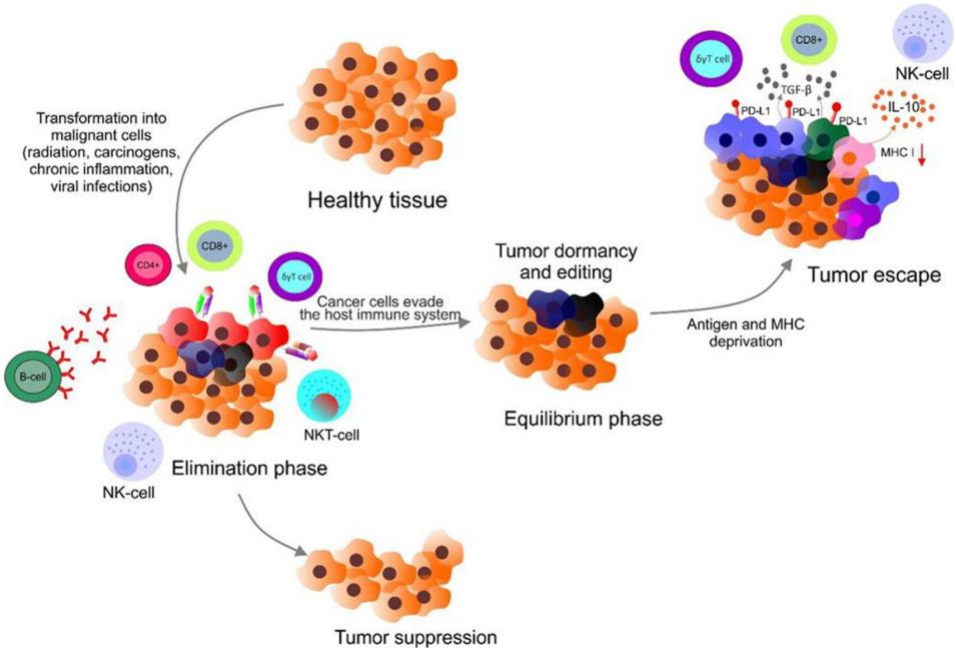


Figure 1.3 Mechanism of immune evasion by tumors. *Transformed cells are eliminated in the first phase by the host immune system (CD4+, CD8+, NK cells, B cells, NK-T cells, γδ T cells, etc.). Equilibrium phase represents surviving tumor cells in dormancy stage. During the second phase, the tumor cells incur editing. Escape phase represents the population of immunologically sculpted tumor cells with specific immunosuppressive mechanisms, including overexpression of PD-L1, production of TGFβ and IL-10, and decreased levels of MHC-I expression. Caption and figure adapted from Samec et al. (2020)[15].*

1.1.3 Metrics of antigen presentation and related prediction tools

Improving our understanding of antigen presentation and refinement of antigen predictions in the context of cancer have both been key research focuses[16]. Early attempts to incorporate antigen presentation as a prognostic signal in cancer relied primarily on bulk metrics such as total tumor mutational burden (TMB) or tumor variant burden (TVB), where whole genome or whole exome sequences were used to identify tumor-specific mutations[17]. These measures relied heavily on the assumptions that 1) number of mutations correlates closely with number of presented epitopes, and 2) epitope quality is less important than epitope quantity. Although initial approaches using TMB and other bulk metrics seemed promising, further

experimentation has shown mixed results[18]–[21]. This is reiterated by work in our own lab which demonstrates that bulk-tumor metrics are only weakly prognostic in some circumstances and highly variable between cancer types[22]. More recent comprehensive analysis across multiple cancer types also supports the conclusion that bulk tumor metrics are not adequate prognostic indicators of patient outcome[23].

More granular approaches have since been attempted that incorporate additional information beyond just simple mutation status. Often termed “neoepitope prediction tools”, these software and pipelines usually incorporate germline variant information from the patient, the MHC haplotype of the individual, and some form of peptide-MHC binding affinity to help determine which mutations are likely to result in presentable epitopes[24]. Additionally, newer prediction tool iterations often include some form of “immunogenicity score” intended to quantify the likelihood of predicted epitopes to elicit an immune response, however the exact meaning of these scores can vary widely between tools and vary greatly in their efficacy when applied to cancer contexts[25].

1.1.4 Challenges with current approaches to antigen identification

Although neoepitope predictions take a step beyond bulk metrics such as TMB, there are still multiple challenges that remain largely unaddressed. Early approaches often ignored key aspects of epitope generation such as mutational haplotyping, which is important for differentiating same-strand from cross-strand mutations in close proximity, and ultimately generate different presented peptide sequences. This in part prompted the creation of our lab’s own tool `neoepiscope`[26]. We have also shown that alternative RNA splicing events can generate peptide sequences in cancer that are often mis-represented as novel peptides by RNA based neoepitope prediction tools, despite expression in other normal or developmental tissues

within the body[27]. Furthermore, neoepitope predictions often focus on epitope binding and recognition, but lack emphasis on the processing steps that precede presentation such as protein degradation by the proteasome and trimming by endoplasmic reticulum associated aminopeptidases (ERAP's). These steps have been shown to play an important role in determining and shaping the final epitope landscape despite their exclusion in many prediction pipelines[28]. In Chapter I: pepsickle rapidly and accurately predicts proteasomal cleavage sites for improved neoantigen identification, I will take an in depth look at current tools available for predicting early-stage protein processing steps, their shortcomings, how we can improve upon previous approaches, and ultimately how proteasomal cleavage predictions can be applied to improve our identification of immunogenic peptides when approaching clinical data.

1.2 Recognition of Class I antigens and TCR degeneracy

1.2.1 Antigen recognition and TCR diversity

While the presentation of class I antigens is a complex multi-step process and imperative for immune surveillance, successful presentation of an antigen does not mandate an immune response. In fact CD8 T-cells, the primary responder to class I antigens, require multiple signals to initiate a true antigen response. The accumulation of signals from T-cell receptor (TCR) ligation with a cognate antigen and co-stimulatory molecules such as CD28 provide activation signals one and two, while cytokine signaling provides the third activation signal[29]. The process of T-cell activation is also complicated by the addition of checkpoint molecules which can further influence the overall signaling landscape through the inhibitory signaling[30]. Ultimately a combination of TCR ligation, co-stimulation, cytokine signaling, and lack of inhibitory signals determine the response of a CD8 T-cell to a potential antigen.

The first step in activation, the process of receptor ligation, is complex in its own right. Unlike many peptide interactions, the interactions that occur between T-cell receptors and presented antigens can be relatively weak[31]. These short interactions are essential as the T-cell itself must continuously move from one presented complex to another for effective surveillance of the presented antigen landscape to take place. During these weak interactions, multiple factors play a pivotal role in whether ligation between the TCR and presented antigen occurs, including both the epitope and TCR sequences. While the presented epitope sequence itself is important, studies have also demonstrated that the T-cell receptor interfaces with both the presented epitope, and components of the presenting MHC complex[32]. This means that T-cell receptors show a preference not only for specific epitopes, but also for specific MHC complexes presenting them.

Additionally, TCR sequences vary immensely due to a developmental process called V(D)J recombination, shaping the landscape of receptor sequences available for antigen interaction[33]. Unlike most genomic regions which stay stable throughout the life of an organism, early in lymphocyte development somatic recombination is induced within the variable (V), joining (J) and diversity (D) regions that eventually contribute to the generation of full receptor sequences (Figure 1.4)[34]. In particular, a set of regions called the complementarity determining regions (CDR's) within the variable domain interface heavily with presented epitopes. CDR loops have been shown to play a particularly important role in determining the antigen specificity of a given receptor, with CDR3 representing the most important and variable region of the three CDR's[35]. These uniquely recombined regions in each lymphocyte generate novel amino acid sequences and greatly expand the lymphocyte receptor repertoire, with studies estimating ~10^{15} possible receptor combinations[36]. Estimates on the number of possible receptor identities greatly outnumber the total T-cell count in the

host, and vastly outnumber estimates of unique T-cell clones in any given individual; suggesting that T-cell clonal diversity and not receptor sequence is the limiting factor in determining the pool of recognizable antigens by a given individual[37].

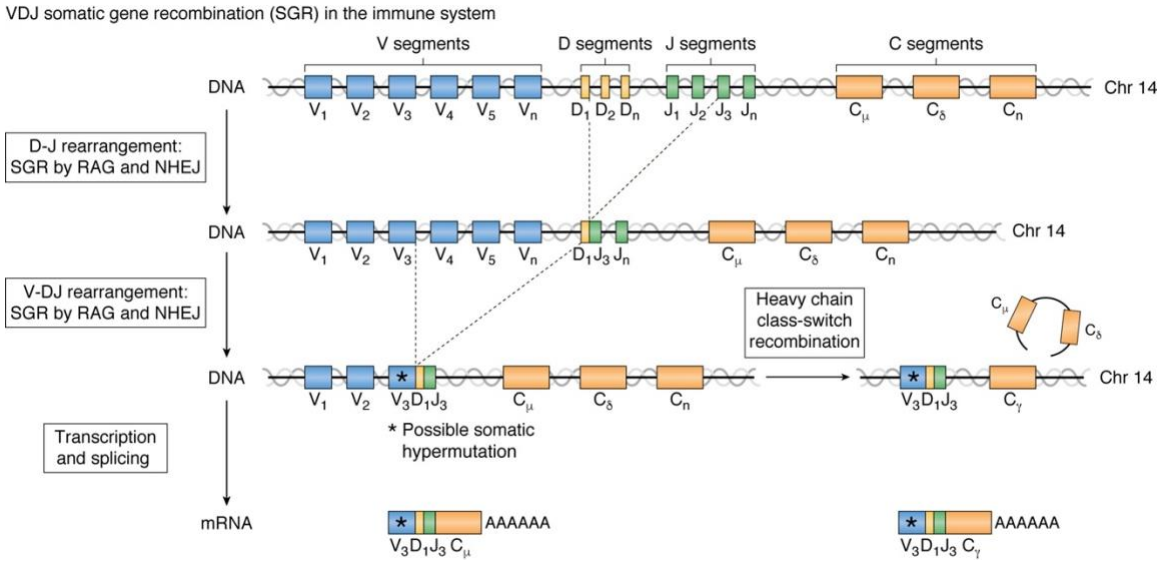


Figure 1.4. VDJ recombination in the immune system. *VDJ recombination requires the RAG complex and non-homologous end joining (NHEJ) enzymes to break and recombine the native genomic locus in a multistep process. Somatic hypermutation may occur. Immunoglobulin heavy-chain class-switch recombination results in isotype replacement of the expressed C region for another downstream C region involving distinct* cis *elements and enzymology. Recombined loci are then transcribed and spliced for translation into antibody proteins. Related VDJ processes affect immunoglobulin light chains (producing VJ joining) and T-cell receptors. Caption and figure adapted from Kaeser and Chun (2020)*[34].

Each unique TCR sequence in conjunction with the surface of the peptide-MHC (pMHC) interface ultimately defines the strength of the bonds, if any, that form in the context of a TCR-pMHC interaction (Figure 1.5)[38]. However, bond affinity alone doesn't determine TCR signal transduction either. Studies examining 3D structure and the binding dynamics of TCR's with their cognate antigens have demonstrated that high affinity bonds which do not induce signaling occur with high frequency in the human T-cell repertoire[39]. Instead, the formation of catch-bonds prolong TCR-pMHC interactions under sheer force and are pivotal for signal

transduction[39]. These dynamic interactions between the T-cell receptor and peptide-MHC complex required for signal transduction complicate our understanding and ability to accurately predict receptor-antigen pairs and emphasize the need for complex models to accurately capture TCR-pMHC interaction dynamics.

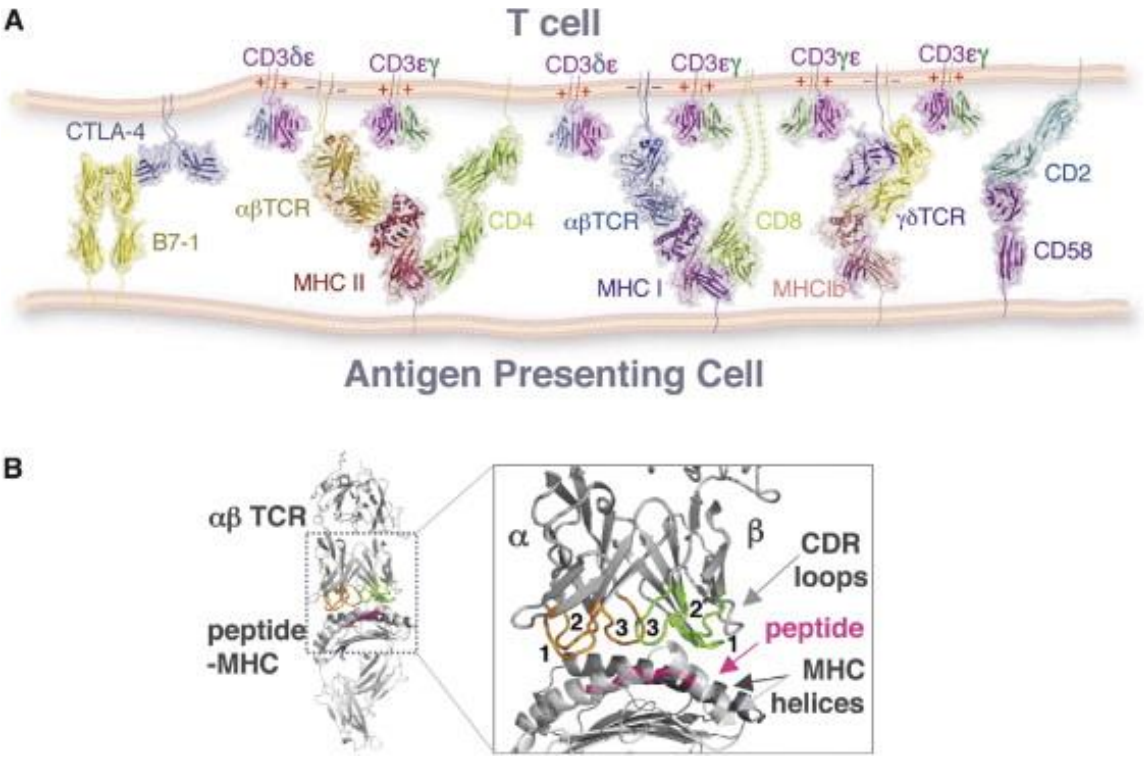


Figure 1.5. Low- and High-Resolution Views of T Cell Recognition *(A) Model of extracellular complex architectures within a T cell/antigen presenting cell interface based on known structural information of the respective receptor ligand complexes. Trimolecular complexes of TCR/MHC/CD8 and TCR/MHC/CD4 have been modeled based on superposition of the MHC in each of the respective TCR/pMHC and MHC/CD8 and MHC/CD4 binary complexes. The transmembrane segments of the CD3 and CD3 subunits have been drawn in with the charges indicated necessary for assembly with the TCR chains. (B) A TCR/pMHC complex (left) and a closeup of the interface (right) showing the "germline" CDR1 and 2 TCR loops contacting the MHC helices, while the centrally located and genetically recombined CDR3 contact the antigenic peptide bound to the MHC. Caption and figure adapted from Garcia et al. (2005)*[38].

1.2.2 TCR degeneracy and its evolutionary importance

The complexity of TCR signal transduction, receptor sequence diversity, and MHC complex preferences together imply that T-cells are designed to be highly specific in their binding selection. While this is true, T-cells have somewhat counterintuitively been shown to frequently respond to multiple cognate antigens; with the exact number of recognizable epitopes varying greatly between individual T-cell clones[36], [40]. This ability to recognize multiple epitopes, often termed TCR promiscuity or TCR degeneracy, can best be conceptualized through the lens of evolutionary host immune challenge. Although we've discussed the vast theoretical diversity of TCR sequences, we have yet to compare this to the variety of class I antigens theoretically in existence. Evidence suggests that class I epitopes most frequently range from 8-11 amino acids long, although exceptions that are both shorter and longer have been reported in literature[41]. Using the standard 20 amino acids, a 9 amino acid long peptide could have over 500 billion (5.12 x 10^11) possible combinatorial identities alone. While the chance that all possible sequences are biologically relevant or even exist within the proteome is quite low, this potential target space still vastly outweighs the number of T-cells, let alone the number of T-cell clones in any given host organism (Figure 1.6)[41]. This imbalance in the ratio of host TCR clones compared to potential antigens that may arise from host immune challenge, highlights the necessity for TCR degeneracy in the context of immune evolution. As viruses and other pathogens continue to evolve, the ability for T-cells to identify multiple potential targets is clearly a necessity.

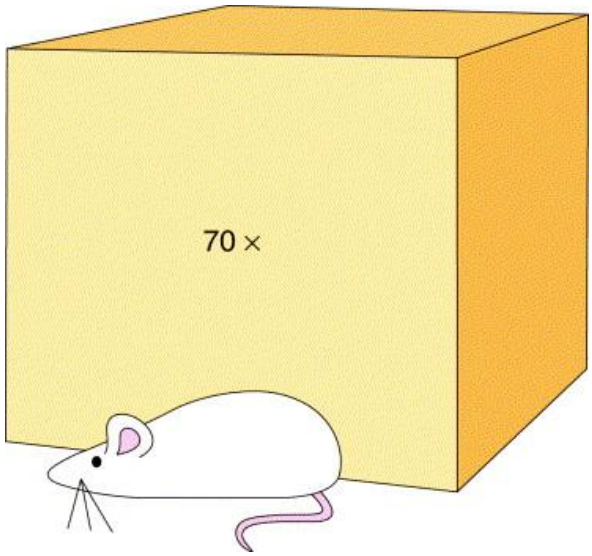


Figure 1.6. Relative proportion of required monospecific T-cells. *If T cells were monospecific, a mouse would require a lymphoid system that was 70 times larger than the cube shown if it possessed just one naive T cell for each possible MHC-associated 11-mer peptide. It is assumed that a quarter of all lymphocytes are naive CD4+ T cells and that 109 lymphocytes occupy a volume of 1 ml. Caption and figure adapted from Mason (1998)*[41].

1.2.3 Impacts of TCR degeneracy and its role in cancer immunotherapy

Despite the evolutionary importance of TCR degeneracy, there are clear drawbacks to TCR recognition of multiple antigens. In the context of auto-immune disorders the recognized epitope is a self-antigen expressed on otherwise normal cells. Although it's possible for these self-epitopes to be the primary target of a given T-cell, resulting from issues with initial T-cell development, many auto-immune disorders are associated with increased prevalence of specific pathogens in patients; leading to the hypothesis that TCR degeneracy is an important contributor[42]. This has been emphasized by studies that demonstrate *Vibrio cholerae* infection can exacerbate lupus symptoms and that patients with rheumatoid arthritis have increased loads of Epstein-Barr Virus compared to normal controls[43], [44].

Off-target toxicities have also been reported in the context of cancers, specifically during the use of experimental next-generation treatments that target specific tumor antigens[45]. One particularly striking example occurred via the use of affinity enhanced TCR's to target a myeloma and melanoma derived epitope generated from the gene *MAGE-A3*. These TCR's also recognized and were unintentionally activated by a cardiac protein titin, which resulted in the death of two patients before termination of the trial[46]. Retrospective analyses confirmed that both proteins were recognized by and activated the affinity enhanced TCR clones used during the experimental trial, highlighting the safety risks posed when cross-reactivity is not taken into account[47].

Although cross-reactivity poses a clear risk in the cancer context, we still lack a strong understanding of how to predict cross-reactive events[48]. Some tools have attempted to use epitope similarity as a proxy for likelihood of cross-reactivity, arguing that epitopes with high similarity are most likely to cross-react[49]. There is some evidence to suggest that similarity plays a role in cross-reactivity, however surveillance of widespread cross-reactivity data demonstrates that cross-reactive peptides can be highly varied and that similarity alone is an insufficient predictor (Table 1.1)[50]. Alternative approaches have relied heavily on T-cell receptor sequences and structure-based predictions to infer a pool of recognized epitopes and subsequently predict cross-reactivity[51]. Sequence based approaches have highlighted key receptor features that impact cross-reactive potential, such as length of the CDR3 loop, however the scope of sequence based approaches is heavily limited by the availability of paired TCR and cognate antigen based datasets[52]. In addition, the use of TCR sequencing data also limits application in many of the pivotal use cases for cross-reactivity predictions where extensive sequencing of T-cell receptors isn't feasible. While some next generation immunotherapeutic approaches rely on expansion of specific T-cell clones exogenously or by the engineering of

chimeric antigen receptors (CAR's) in which case the TCR sequence can be identified, other approaches such as cancer mRNA vaccines rely on endogenous TCR's which can't be exhaustively or feasibly assayed for TCR sequences[53].

Table 1.1. Reported examples of cross-reactive peptides in literature.

MHC	Initial	Subsequent	Initial	Subsequent	Overlap	Id	S_O	S_E
H2-Kd	LCMV NP	PV NP	**YTVKYPNL**	**YTVKFPNM**	**YTVK.PN.**	6/8	0.92	0.81
H2-Kd	LCMV NP	VV P1	**YTVKYPNL**	**YNSLYPNV**	**Y...YPN.**	4/8	0.71	0.68
H2-Kd	LCMV NP	VV P10	**YTVKYPNL**	**STLNFNNL**	**.T....NL**	3/8	0.58	0.48
H2-Kd	LCMV NP	VV A11R	**YTVKYPNL**	**AIVNYANL**	**..V.Y.NL**	4/8	0.61	0.58
H2-Kd	LCMV NP	VV A11R	**AVYNFATC**	**AIVNYANL**	**A..N.A..**	3/8	0.61	0.45
H2-Kd	LCMV NP	VV A11R	**ISHNFCNL**	**AIVNYANL**	**...N..NL**	3/8	0.53	0.48
H2-Kd	RSV5 M2-82	RSV M2-71	**SYIGSINNI**	**EYALGVVGV**	**.Y.....**	1/9	0.46	0.36
H2-Kd	CTL agonist (APL)	IGRP206-214	**KYNKANWFL**	**VYLKTNVFL**	**.Y.K.N.FL**	5/9	0.6	0.64
H2-Kd	Dengue 2 NS3-298	Dengue 3 NS3-299	**GYISTRVEM**	**GYISTRVGM**	**GYISTRV.M**	8/9	0.9	0.92
HLA-A2	EBV BMLF1-280	FLU A M1-58	**GLCTLVAML**	**GILGFVFTL**	**G....V..L**	3/9	0.53	0.47
HLA-A2	EBV BMLF1-280	FLU A NP-85	**GLCTLVAML**	**KLGEFYNQM**	**.L.....**	1/9	0.38	0.28
HLA-A2	EBV BMLF1-280	EBV LMP2	**GLCTLVAML**	**LLWTLVVLL**	**.L.TLV..L**	5/9	0.62	0.59
HLA-A2	EBV BMLF1-280	EBV BRLF1	**GLCTLVAML**	**YVLDHLIVV**	**.....**	0/9	0.32	0.22
HLA-A2	FLU A NA-231	HCV NS3-1073	**CVNGSCFTL**	**CVNGVCWTV**	**CVNG.C.T.**	6/9	0.83	0.78
HLA-A2	FLU A M1-58	EBV EBNA3A-596	**GILGFVFTL**	**SVRDRLARL**	**.....L**	1/9	0.38	0.29
HLA-A2	HPV 16 E7-11	Coronavirus NS2-52	**YMLDLQPET**	**TMLDIQPED**	**.MLD.QPE.**	6/9	0.79	0.76
HLA-A2	HIV ENV GP-120	M. tuberculosis	**VPTDPNPPEV**	**VLTDGNPPEV**	**V.TD.NPPEV**	8/10	0.79	0.8
HLA-B62	Dengue 2 NS3-71	Dengue 3 NS3-71	**DVKKDLISY**	**SVKKDLISY**	**.VKKDLISY**	8/9	0.92	0.9
HLA-A1	Hantaanvirus (Sin)	Hantaanvirus (Seoul)	**ISNQEPLKL**	**ISNQEPMKL**	**ISNQEP.KL**	8/9	0.97	0.93

The columns are as follows: 1) MHC restriction, 2) source pathogen and protein for initial infection, 3) source pathogen and protein for subsequent infection, 4) original epitope of initial infection, 5) cross-reactive epitope for subsequent infection, 6) Sequence overlap between the cross-reactive epitopes, 7) sequence identity (Id), 8) observed peptide similarity (So) and expected peptide similarity (Se). So and Se represent similarity scores based on BLOSUM35 scoring of the cross-reactive peptide and a randomized peptide, respectively. Table and caption adapted from Frankilde et al. (2008).[50]

While the broader environmental and treatment related factors that impact immune activation won't be addressed until the last chapter, Chapter II: Predicting T-cell Cross-Reactivity using Paired Peptide Data: A Multi-Layer Perceptron Approach, will highlight an alternative method for cross-reactivity prediction that does not rely on TCR specific sequencing. The

aggregation of large-scale database data to identify cross-reactive paired epitope examples leverages existing information to help better understand the dynamics of cross-reactive epitopes, even in the absence of receptor sequencing. While this approach falls short of incorporating other key factors in T-cell cross-activation such as co-stimulation and cytokine signaling, it provides important insight into TCR interactions with presented peptides and when those interactions may create a fundamental risk of cross-reactive events.

1.3 Impact of the tumor landscape on immune interactions

1.3.1 Cellular components of the TME

While previous sections have painted antigen presentation and subsequent T-cell activation as well defined - albeit complicated - processes, herein we will acknowledge the messiness that arises when we leave the petri dish and consider isolated components of biology in a more holistic manner. Ultimately TCR signal transduction, co-stimulation, and checkpoint inhibition all exist on a spectrum where the strength of each can ultimately tip the balance towards activation or tolerance. However, cellular interactions with other immune and stromal cells can also greatly influence this balance and whether or not cytotoxic cells ultimately carry out their goal of killing a recognized target cell. This external influence is especially relevant in the context of the tumor micro-environment where a wide range of cells are recruited by the developing tumor or subsequently arrive as part of the mounted immune response.

Although our understanding of cells associated with the tumor microenvironment (TME) is ever expanding, fibroblasts, mast cells, macrophages, dendritic cells, myeloid derived suppressor cells (MDSC's), T-regulatory cells (T-regs), CD8 cytotoxic cells, natural killer (NK) cells, and helper T-cells (Th) are just some of the well-recognized components known to contribute to the overall immune landscape of the tumor (Figure 1.7)[54], [55]. Subsets of dendritic cells and

Th cells help support cytotoxic T-cell activation and generally tip the balance towards tumor killing by priming cytotoxic T-cells[56]. Some macrophages, often termed tumor-suppressive or M1-like, have also been shown to help support tumor clearance through phagocytosis and secretion of stimulatory molecules such as IL-6, and IL-12 [57]. In contrast, T-regs, myeloid derived suppressor cells, and pro-tumorigenic (M2-like) macrophages all inhibit cytotoxic activity through suppression of CD8 T-cell activation and exacerbation of the inflammatory TME[57]–[59]. Fibroblasts and Mast cells contribute through the secretion of inflammatory signals which can also serve to inhibit effective tumor killing, though this is only one of many roles that these cells and inflammation itself can play in the context of the tumor microenvironment[60]. The distribution of varying immune cell populations and overall amount of immune infiltration are sometimes used to define tumors as immune “cold” or immune “hot” based on the overall level of tumor-immune activity[61].

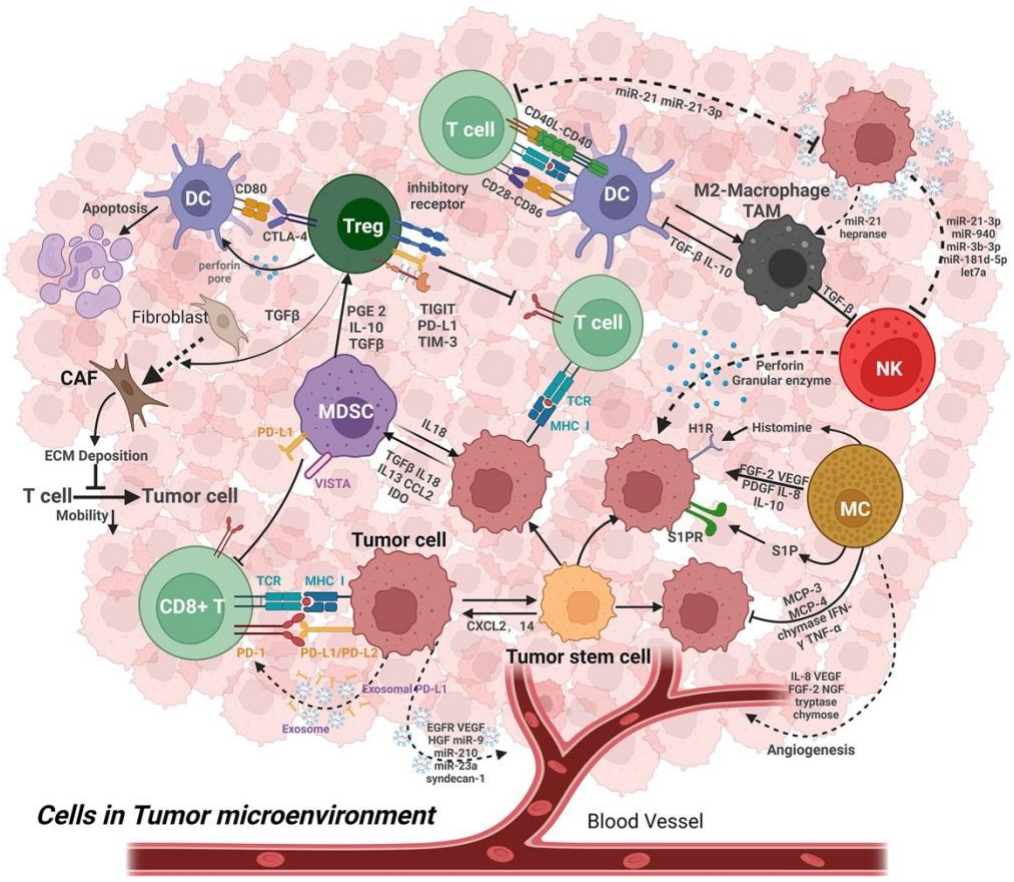


Figure 1.7. Common cellular and structural components in the TME. *Mast cells (MCs), natural killer cells (NKs), dendritic cells (DCs), myeloid derive suppressor cells (MDSCs), tumor associated macrophages (TAMs), and a variety of T-cell populations are all common immune components of the tumor microenvironment and interplay with tumor cells and other resident populations such as cancer associated fibroblasts (CAFs). Figure adapted from Wang et al. (2023)*[55]

1.3.2 The TME and T-cell exhaustion

The Milieu of inhibitory signals present in the tumor can ultimately lead to a phenomenon called exhaustion in which T-cells cells that would otherwise support tumor clearance show greatly diminished cytotoxic abilities[62]. A consensus definition of exhaustion is hard to come by and varies from expert to expert, however, exhaustion generally describes the process by which T-cells lose expression of effector and cytokine molecules due to chronic antigen and/or inflammatory exposure[63]. The loss of effector signals in T-cells undergoing

exhaustion also coincides with the expression of inhibitory markers including but not limited to PD-1, CTLA-4, TIGIT, LAG3 and TIM3[62]. Concurrent expression of these inhibitory molecules alongside the loss of cytotoxic and effector cytokine expression can ultimately be used to define functional exhaustion in the cancer context[64].

Based on our understanding of exhaustion, an immunotherapeutic approach called immune checkpoint inhibitor (ICI) therapy has arisen[65]. ICI therapies aim to prevent exhaustion phenotypes from forming and potentially even reverse the phenotype of some functionally exhausted cells in the TME by targeting inhibitory molecules such as PD-1, PDL-1, and CTLA-4 with monoclonal antibodies[65]. Therapeutic approaches using ICI have demonstrated profound benefits for some patients, even for those whose cancer progressed on more traditional treatments[66]. Despite these promising results, many other patients do not respond to ICI therapy, further highlighting the complexity or the immune response and important role of other cell types beyond those directly targeted by ICI[67].

1.3.3 Single cell approaches to TME profiling

The highly variable response to ICI therapies between patients has emphasized the need for a better understanding of the tumor microenvironment and improved profiling of all components therein. While traditional techniques for profiling the TME have generated a plethora of findings about immune interactions within the greater tumor landscape, novel technologies are also drastically reshaping our understanding of tumor-immune interactions as well. In particular single-cell RNA sequencing has facilitated the simultaneous transcriptomic profiling of thousands of cells in unison; giving a unique perspective into their cell states and unique responses to the greater microenvironment (Figure 1.8)[68].

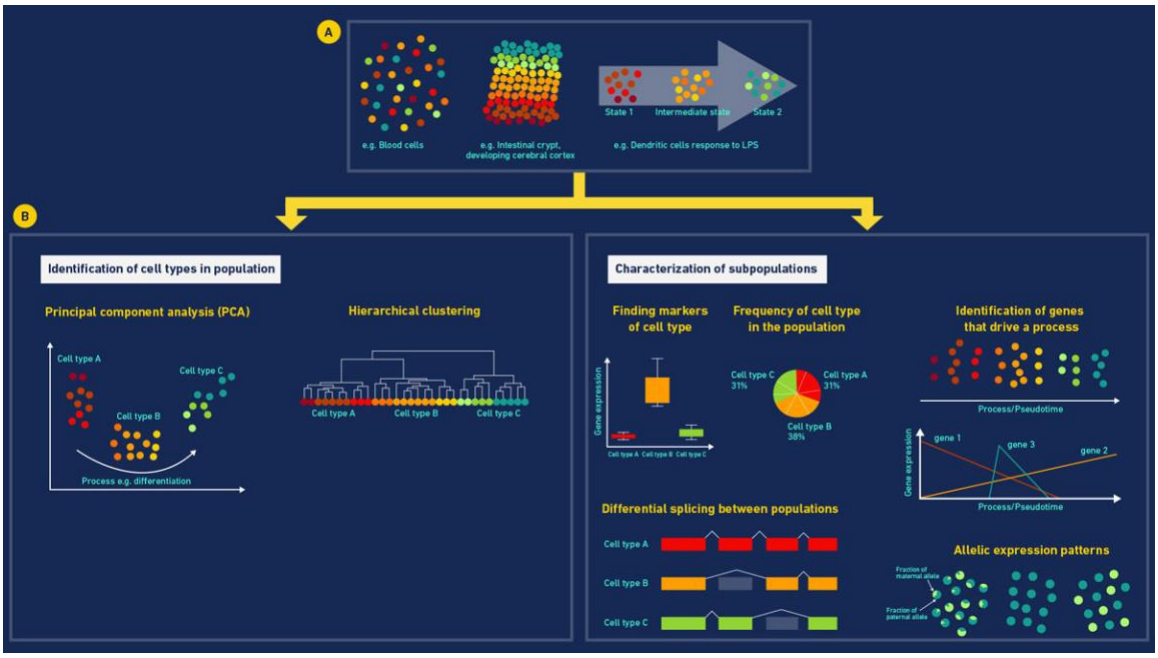


Figure 1.8. Single-cell technology applications. *Using single cell sequencing technologies, we can identify novel subpopulations or cellular states within a seemingly homogeneous cellular population. A) shows different ways in which a cell population may exhibit heterogenicity. B) shows how cell types within populations may be identified and characterized. Caption and figure adapted from Vaga (2022)*[68].

Single-cell transcriptomic analysis relies on a multi-step process, starting with mechanical tissue dissociation, followed enzymatic degradation of matrix proteins, and further agitation that ultimately results in the generation of a single cell suspension. Suspended cells are then combined with unique oligonucleotide barcodes using a microfluidics-based (e.g. 10X) approach, giving rise to a cellular emersion[69]. Each oligonucleotide has both a unique cellular barcode and a unique molecular identifier that allows for downstream tracking of both cell and unique transcript[70]. Cells, in emersion with unique oligonucleotides, are then lysed so that reverse transcription and amplification can occur[70]. Samples can then be pooled and sequenced with the total throughput of cells varying based on sequencing platform and desired sequencing depth.

Other approaches have also been used to characterizing single cells such as immunohistochemistry (IHC), flow cytometry, and cyclic immunofluorescence (CyCIF), however these techniques rely on antibodies and are limited in the number of targets that can be queried per cell at a single time. In contrast, single-cell RNA sequencing can identify hundreds or even thousands of unique transcripts in a mostly unbiased fashion. Although single-cell RNA-seq captures many more targets at a given time, this comes at the cost of selecting specific targets of interest. Because specific targets cannot be selected, single cell RNA-seq also relies heavily on computational approaches to identify specific cell types of interest and leverages group-based analyses to better profile identified cells.

While Single-cell RNAseq technologies have already provided insight into the inner workings of the TME, to date these approaches have rarely examined patients across their treatment course. In Chapter III: Deciphering the Prostate Tumor Microenvironment: Transcriptional Insights into Therapy Response following Androgen Axis Blockade and Immune Checkpoint Inhibition, we will use a paired single cell transcriptomic approach to survey the tumor and immune landscape of patients receiving therapy for the treatment of primary prostate cancer tumors. These patients have been sampled in both a completely treatment-naive context, and after an extensive treatment course with a combination of therapeutics. This approach gives a unique look into both the baseline microenvironment and a look into how therapy shifts the cellular landscape as a whole. Although the work in Chapter III does not yet contain clinical follow up data, the analysis provided within highlights dramatic changes occurring in both tumor and non-tumor tissues due to the administration of treatment and provides further insight into key factors and populations present in the tumor-immune landscape.

Chapter I: `pepsickle` Rapidly and Accurately Predicts Proteasomal Cleavage Sites for Improved Neoantigen Identification

Benjamin R. Weeder[1,2], Mary A. Wood[3] , Ellysia Li[4] , Abhinav Nellore[1,2,5] and Reid F. Thompson[1,2,6,7,8,*]

[1]Computational Biology Program, Oregon Health & Science University, Portland, OR 97239, USA,

[2]Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR 97239, USA

[3]Phase Genomics Inc., Seattle, WA 98109, USA

[4]Pacific University, Forest Grove, OR 97116, USA

[5]Department of Surgery, Oregon Health & Science University, Portland, OR 97239, USA,

[6]Department of Radiation Medicine, Oregon Health & Science University, Portland, OR 97239, USA

[7]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR 97239, USA

[8]Division of Hospital and Specialty Medicine, VA Portland Healthcare System, Portland, OR 97239, USA

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Adapted from Weeder et al. (2021)[71]
Originally published in *Bioinformatics*, Volume 37, Issue 21, November 2021

2.1 Abstract

Motivation: Proteasomal cleavage is a key component in protein turnover, as well as antigen processing and presentation. Although tools for proteasomal cleavage prediction are available, they vary widely in their performance, options, and availability.

Results: Herein, we present `pepsickle`, an open-source tool for proteasomal cleavage prediction with better *in vivo* prediction performance (area under the curve) and computational speed than current models available in the field and with the ability to predict sites based on both constitutive and immunoproteasome profiles. *Post hoc* filtering of predicted patient neoepitopes using `pepsickle` significantly enriches for immune-responsive epitopes and may improve current epitope prediction and vaccine development pipelines.

Availability and implementation: `pepsickle` is open source and available at https://github.com/pdxgx/pepsickle.

2.2 Introduction

The constitutive proteasome is a multimeric protein complex best known for its role in the cleavage and recycling of cellular proteins marked for degradation[72]. The proteasome also generates cleaved peptide fragments (epitopes) for immune surveillance via the major histocompatibility complex (MHC) class I antigen presentation pathway[73]. This immune presentation functionality is critical for antiviral and other antimicrobial responses, and has particular relevance both in the setting of vaccine development and in a cancer context with the advent of immune checkpoint[74]–[78].

Structurally, the proteasome consists of multiple subunits, a 20S barrel core housing the catalytic domains of the proteasome, and two 19S caps which aid in the unfolding of ubiquitin-

tagged proteins[72]. The barrel shape of the 20S core is derived from the fusion of four heptameric rings, the inner two of which contain a β1, β2 and β5 catalytic domain responsible for the cleavage of peptide bonds[79]. Although all tissues express the constitutive proteasome, hematopoietic-lineage cells can also express the alternative catalytic domains β1i, β2i and β5i in response to IFN-γ, which replace their analogues in the constitutive heptameric ring to form the immunoproteasome (Fig. 2.1)[80]. Previous studies support the presence of preferred cleavage motifs and differences in cleavage preferences between the immuno- and constitutive proteasomes; however, our understanding of how these preferences manifest is still not well defined[81], [82].

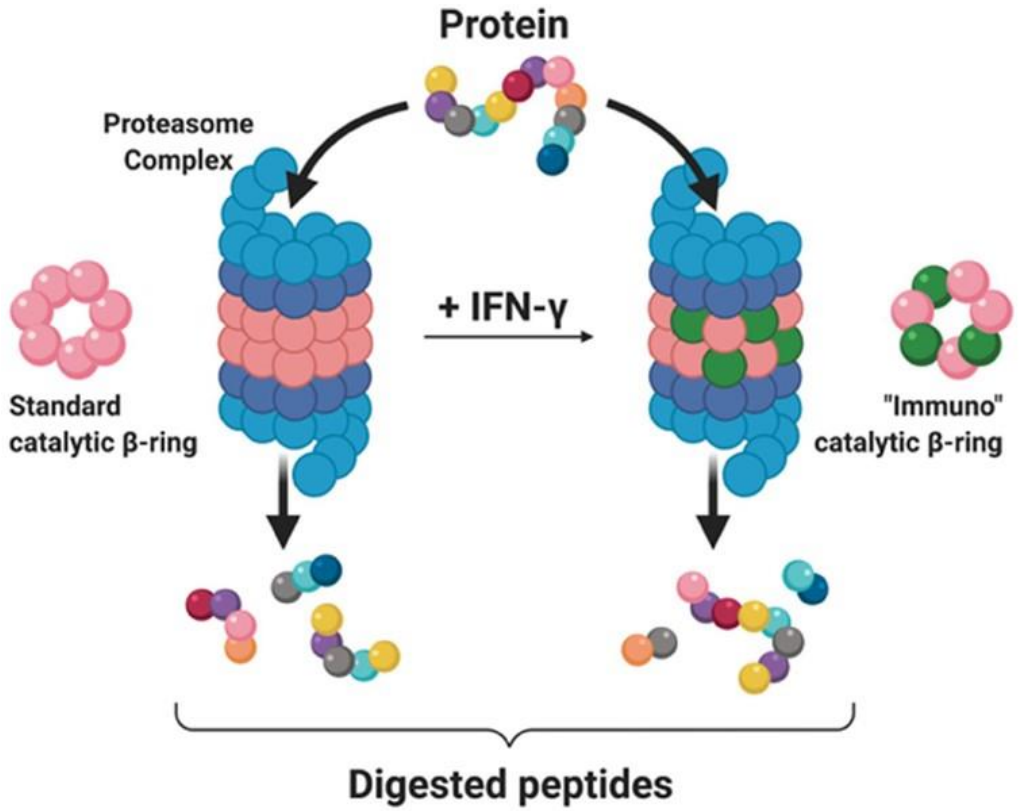


Figure 2.1. Protein degradation by the constitutive and immunoproteasome. *Proteins trafficked to the proteasome complex are fed into the main 20S barrel, with the assistance of 19S caps (blue) that aid with unfolding and linearization. The catalytic domains of the standard β-rings (pink) constituting the 20S barrel cleave the protein sequence and generate the resulting digested peptide fragments. In select tissues, exposure to interferon gamma (IFN-γ) results in replacement of the standard catalytic domains by alternative 'immuno' catalytic domains*

(green). This transition in catalytic domain usage constitutes the construction of the immunoproteasome and may alter cleavage site preference. The differential digestion pattern of a single protein sequence (multi-colored) is depicted below the corresponding proteasome complex.

While existing tools to predict proteasomal cleavage sites are now widely adopted, each has significant limitations affecting the accuracy and/or scope of its predictions, with the potential for real world consequences (Fig. 2.2). For example NetChop 3.1, the most cited proteasomal cleavage tool, does not differentiate between constitutive and immunoproteasomal cleavage when generating predictions[81]. Further, tools such as the proteasomal cleavage prediction server (PCPS) provide options for predicting cleavage by the immunoproteasome but show poor model performance compared to NetChop when benchmarked[81]. Finally, many available tools are either proprietary or otherwise unavailable to the public, complicating their use in both academic and industry analysis pipelines.

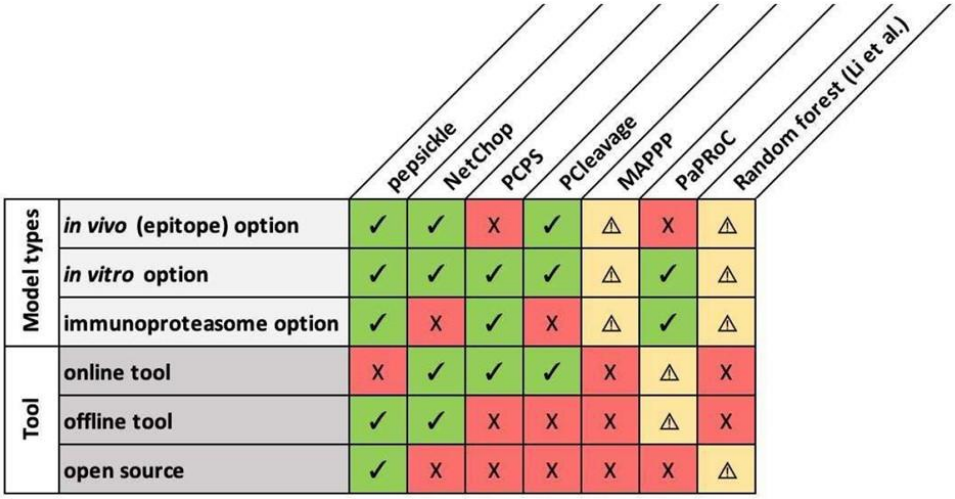


Figure 2.2. Comparison matrix of available proteasomal cleavage tools and their features*. Eight proteasomal cleavage tools are shown (columns) along with their corresponding features (rows). Specific tools are as follows: pepsickle (presented here), NetChop 3.1*[81]*, the Proteasomal Cleavage Prediction Server (PCPS)*[28]*, PCleavage*[83]*, MAPPP*[84]*, PAProC*[85] *and the random forest-based model described in Li et al.*[86]*.Check marks (green) represent available features for each tool while X's (red) represent unavailable features. Warning signs (yellow) represent missing information, or features that are mentioned but not currently available. For MAPPP, the referenced web server is no longer available and therefore we were unable to confirm tool features. For PaPRoC, we were unable to obtain the model despite repeated*

requests. For the random forest model proposed by Li et al., model weights for the proposed model are given, but source code is not available and the type of cleavage sites used (in vivo versus in vitro) are undefined.

By leveraging a comprehensive set of proteasomal cleavage data and an ensemble-based deep learning approach, we developed a set of models that consistently produce more accurate cleavage predictions than existing tools regardless of proteasomal context. We have deployed these models as an open-source command-line tool (`pepsickle`) for broad reuse and application.

2.3 Materials and methods

2.3.1 Collection and processing of in vitro digestion map data for training and testing

We performed a literature search for all studies containing publicly available primary data from *in vitro* digestion experiments using 20S proteasomes. As the proteasome is highly conserved among mammalian species, digestion product results from non-human mammalian proteasomes were also included along with human-specific datasets[82], [87]. The search terms used were 'proteasome', 'proteasomal', 'cleavage', 'digestion', 'immunoproteasome', '20S', 'i20S', both alone and in various combinations. Ultimately, we identified 35 studies with relevant data (Table 2.1), from which we manually extracted individual cleavage sites, along with the parent peptide sequences from which they were derived[88]–[122]. Proteasome types present in the observed system (constitutive, immunoproteasome or mixed) were also annotated for each cleavage experiment. Data from six 20S studies with unique source proteins were held out for downstream validation, while the remaining data (from 29 studies) were aggregated for model training and testing (Table 2.1). For *in vitro* digestion peptide fragments, both the N-terminal and C-terminal cleavage sites were used as cleavage examples. For each cleavage

example, a context window was generated with the cleavage residue (C-terminus of the peptide fragment) as the 'central' amino acid plus an equal number of upstream and downstream amino acids (Fig. 2.3). Independent datasets with window sizes of 7 amino acids (3 upstream and 3 downstream from the central cleavage residue) and 21 amino acids (10 upstream and 10 downstream) were generated to allow for model optimization based on window size. Only unique cleavage windows were retained, yielding a total of 1758 windows that are 7 amino acids in length and 1819 windows that are 21 amino acids in length.

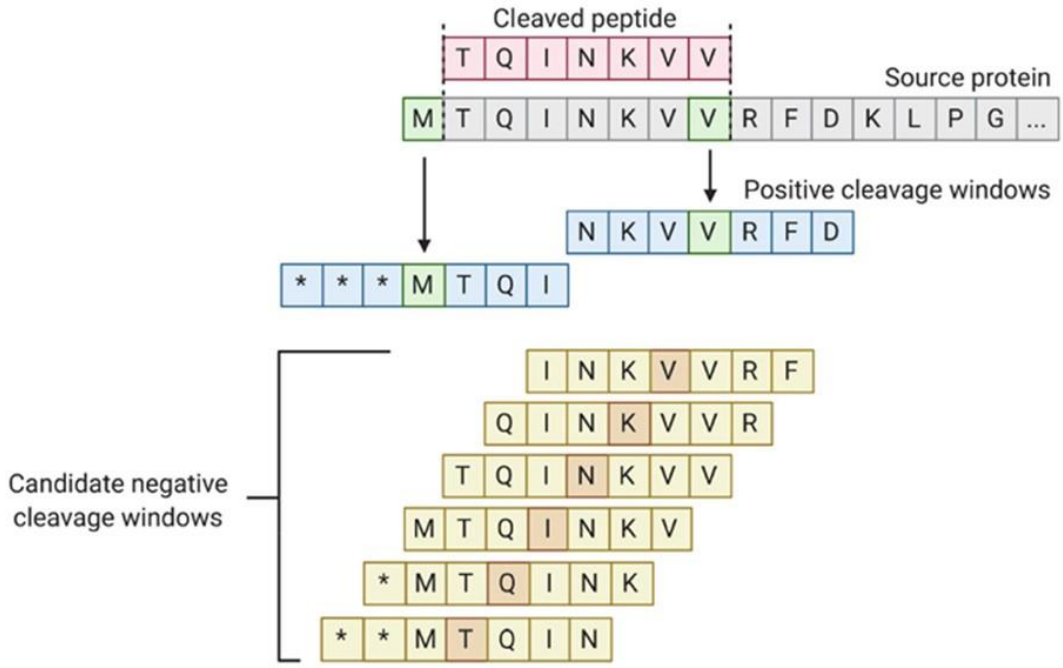


Figure. 2.3. Generation of the in vitro dataset. *Each identified cleaved peptide fragment (red) was mapped back to its source sequence (gray). Using the C-terminus of the fragment, as well as the amino acid prior to the N-terminus of the fragment as cleavage sites (green, with each of their respective downstream bonds cleaved by the proteasome), cleavage windows (blue) were generated using three amino acids upstream and downstream of the cleavage sites identified. Candidate non-cleavage windows (yellow) were generated using the same windowed approach on internal amino acids within the epitope. Before candidate negatives were included in the dataset, they were screened against all positive identified cleavage sites from both N- and C-termini of reported fragments. Note that * indicates the lack of an amino acid (i.e. amino acid position is beyond the peptide terminus).*

Table 2.1. Summary of *in vitro* data.

Source	Dataset	Fragments	Proteasome types
Toes *et al.* (2001)	Train/test	251 (1)	Constitutive, Immuno
Berko *et al.* (2012)	Train/test	236 (2)	Constitutive
Sesma *et al.* (2003)	Train/test	184 (4)	Mixed
Lucchiari-Hartz *et al.* (2003)	Train/test	145 (1)	Mixed
Tenzer *et al.* (2004)	Train/test	129 (1)	Constitutive, Immuno
García-Medel *et al.* (2012)	Train/test	121 (1)	Constitutive
Guillaume *et al.* (2012)	Train/test	103 (10)	Constitutive, Immuno
Chapiro *et al.* (2006)	Train/test	96 (3)	Constitutive, Immuno
Niedermann *et al.* (1996)	Train/test	91 (3)	Mixed
Ehring *et al.* (1996)	Train/test	84 (1)	Mixed
Pinkse *et al.* (2005)	Train/test	81 (2)	Constitutive, Immuno
Emmerich *et al.* (2000)	Train/test	63 (1)	Constitutive
Niedermann *et al.* (1995)	Train/test	56 (8)	Immuno
Kessler *et al.* (2001)	Train/test	47 (4)	Immuno
Hassainya *et al.* (2005)	Train/test	35 (1)	Immuno
Paradela *et al.* (2000)	Train/test	34 (1)	Mixed
Lucchiari-Hartz *et al.* (2000)	Train/test	30 (1)	Constitutive
Theobald *et al.* (1998)	Train/test	24 (2)	Immuno
Popović *et al.* (2011)	Train/test	23 (2)	Immuno
Alvarez-Castelao *et al.* (2014)	Train/test	22 (1)	Constitutive

Source	Dataset	Fragments	Proteasome types
Marcilla *et al.* (2007)	Train/test	21 (1)	Immuno
Dick *et al.* (1996)	Train/test	20 (3)	Mixed
Bruder *et al.* (2006)	Train/test	18 (2)	Constitutive
Morel *et al.* (2000)	Train/test	17 (1)	Constitutive, Immuno
Asemissen *et al.* (2006)	Train/test	14 (1)	Constitutive, Immuno
Michaux *et al.* (2014)	Train/test	13 (1)	Constitutive
Macconi *et al.* (2009)	Train/test	12 (1)	Constitutive
Kimura *et al.* (2005)	Train/test	8 (2)	Mixed
Vigneron *et al.* (2004)	Train/test	6 (1)	Constitutive
Wada *et al.* (2018)	Validation	334 (11)	Immuno
Ayyoub *et al.* (2002)	Validation	49 (1)	Constitutive
Zimbwa *et al.* (2007)	Validation	48 (1)	Immuno
Alvarez-Castelao *et al.* (2012)	Validation	32 (1)	Constitutive
Strehl *et al.* (2008)	Validation	16 (4)	Constitutive, Immuno
Warren *et al.* (2006)	Validation	16 (1)	Immuno

Note: All data used for training, testing and validating in vitro models is summarized above. Fragments represent the number of cleavage by-products reported in each primary literature source, with the number in parentheses representing the number of whole proteins or pre-digestion protein fragments used in each study. Proteasome type(s) denotes what proteasome was queried during experimentation with 'constitutive' and/or 'immuno' denoting isolated contexts, while 'mixed' denotes testing in a non-isolated/heterogenous proteasomal context.

To generate companion non-cleavage examples for modeling, internal sites from each reported peptide fragment were considered as candidates. As above, 7 amino acid and 21 amino acid windows were generated for each non-cleavage example, and subsequently filtered

to remove any duplicate sequences or overlaps with the set of non-cleavage examples. Before these candidates were included as null examples in the dataset, they were further filtered against all positive windows generated across all other studies, controlling for proteasome type, so that positive and negative cleavage examples were mutually exclusive.

2.3.2 Collection and processing of epitope data for training and testing

To study *in vivo* cleavage sites, we extracted endogenously processed T-cell epitopes from three independent public databases: The Immune Epitope Database (IEDB)[123], AntiJen[124] and SYFPEITHI[125] as well as two primary literature sources[126], [127]. Data from Bassani-Sternberg *et al.* was maintained separately and used for downstream validation, while all other sources were aggregated for training and testing purposes. We restricted attention to mammalian endogenously processed and presented peptide ligands of the MHC class I pathway using the flags: assay_type.category = 'Naturally Processed', mhc_allele_restriction.class = 'I' and organism_finder_host_ancestry.obi_id = 'http://purl.obolibrary.org/obo/NCBITaxon_40674'. Epitopes were filtered to retain only those with an unambiguous position among the known source protein sequence(s). Centered windows were generated around each C-terminal cleavage example as above but using the full series of balanced window sizes from 7 amino acids to 21 amino acids, given the larger scale of the data and its accompanying power to detect significant differences in model performance. Only unique cleavage window sequences were retained, resulting in a total of 357,253 unique epitopes with C-terminal cleavage events. Note that epitope N-termini were not processed as cleavage examples due to the uncertainty resulting from N-terminal trimming by endoplasmic reticulum aminopeptidases (ERAP)[128].

To perform non-cleavage site inference, we first sampled internal amino acids for each epitope (Fig. 2.4). Because the overwhelming majority of proteasomal digestion products have a length of at least two amino acids[129]–[131] and because peptides may be threaded into the proteasome in either the N- or C-terminal directions[92], we excluded 1 N-terminal and 1 C-terminal amino acid residue of each epitope from consideration in the potential non-cleavage data. As above, windows were generated for each remaining amino acid position for each potential window size. These potential non-cleavage windows were then filtered to remove any identical sequence matches within the set of positive cleavage examples from above. To additionally account for uncertainty in N-terminal cleavage position(s) due to ERAP[128], we removed any candidate sequence that matched with the set of windows generated by upstream positions from the N-terminus of an epitope up to 16 amino acids upstream of each epitope's C-terminus. Only unique non-cleavage window sequences were retained.

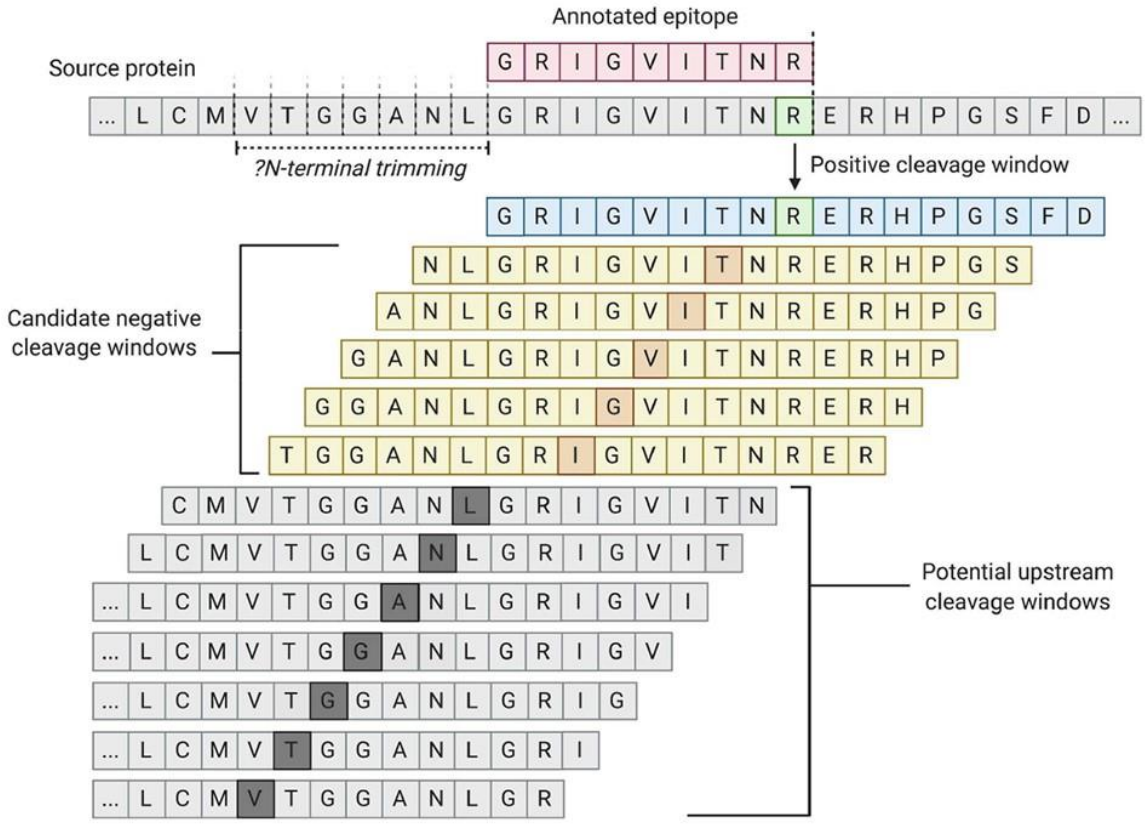


Figure. 2.4. Generation of the epitope dataset. *Each identified epitope (red) was mapped back to its source sequence. Using the C-terminus of the epitope as the cleavage site (green), with the downstream bond as the one cleaved by the proteasome), cleavage windows (blue) were generated using eight amino acids up and downstream from the site identified. Candidate non-cleavage windows (yellow) were generated using the same windowed approach on internal amino acids within the epitope, with the exclusion of the first two and last two amino acids which served as a buffer region to account for minimum proteasomal fragment size. Before candidate negatives were included in the dataset, they were screened against all positive identified cleavage sites as well as against a set of potential upstream cleavage sites (gray); generated by using the same windowed approach on the upstream window that could encapsulate the N-terminal cleavage site prior to ERAP trimming.*

2.3.3 Feature encoding

A vector of features was generated for each amino acid across windows in the cleavage and non-cleavage example sets (Appendix A, Table 6.1). Amino acid identity was one-hot encoded as a bit vector of size 20, with each bit representing one of the standard amino acids.

The null character (*) was used for padding, with all values as zero, while ambiguous amino acids were encoded as relevant combinations of non-zero values corresponding to their ambiguous components (e.g. B represents either aspartic acid or asparagine). Physical properties of amino acids were encoded as follows: side chain polarity was recorded as its isoelectric point (pI)[132], the molecular volume of each side chain was recorded as its partial molar volume at 37°C[133], the hydrophobicity of each side chain was characterized by its simulated contact angle with nanodroplets of water[134] and conformational entropy was derived from peptide bond angular observations among protein sequences without observed secondary structure (e.g. alpha helix)[135]. Proteasomal context was also included where relevant as a single categorical feature with 'C' representing the constitutive proteasome, 'I' representing the immunoproteasome and 'M' representing mixed systems with both proteasome types expressed.

2.3.4 Gradient boosted decision tree structure and training

All gradient boosted classification models were implemented using the Scikit-learn package (v0.22.1)[136] for Python version 3.7. The aggregated positive and negative cleavage examples were randomly split to retain 80% of the examples for training and the remaining 20% for model testing. For each model, inversely balanced class weights were used, and the `'RandomizedSearchCV'` class was used to determine the best option for the `'max_features'` parameter (chosen from `'auto'`, `'sqrt'` or `'log2'`) and the `'n_estimators'` parameter (chosen from values of 100–1000, by 100) of the `'GradientBoostingClassifier'` class. Randomized 10-fold cross validation was run for all combinations of parameters, and the best model (as determined by the

'best_estimator_' attribute) was retained. Model performance was evaluated based on Area under the curve (AUC), using the pROC library (v1.16.2)[137] in R version 4.0.2.

Two distinct classification models were trained. One was based on the one hot encoded amino acid sequence identities, and another on the physical/chemical property encodings as previously described, normalized using the 'fit_transform' method of Scikit-learn's 'MinMaxScaler' class. For epitope data, models were trained on amino acid window sizes 7 through 21 in length and compared to each other to identify the model with optimal performance. For *in vitro* data, only the minimum (7) and maximum (21) length window libraries were assessed, also accounting for constitutive versus immunoproteasomal context.

2.3.5 Neural network structure and training

All neural network models were implemented using the PyTorch package (version 1.3.1)[138] for Python version 3.7. The aggregated positive and negative cleavage examples were randomly split to retain 80% of the examples for training and the remaining 20% for model testing. We next trained two distinct cleavage classification models based on the proteasome type and either (i) amino acid identity encodings, or (ii) amino acid physical property encodings as described previously. Each model consisted of an input layer, two hidden layers and an output layer (Appendix A, Fig. 6.1). For all non-output layers, we applied batch normalization and a 20% dropout layer during each successive forward pass to improve model training and reduce overfitting (for layer sizes, see appendix A, Table 6.2 and 6.3, respectively). ReLU activation functions were employed at each step except for the output layer, where a softmax function was applied prior to final output. For the physical property-based model an additional convolutional layer (1D convolution with a three amino acid window and one amino acid step size) was applied to each physical property independently prior to passing values to the rest of

the model. Cross entropy loss was used for backpropagation during training, with inverse class weights to account for class imbalance in the training set. Both models were trained for 36 epochs before training was halted. AUC assessed on a new subset of the test data after each epoch and compared to the performance at the previous epoch, with the best performing model saved for downstream analysis.

For the two best-performing models (one identity-based and one based on physical properties), final testing performance was then assessed using a consensus approach, where the predicted probability of a test window representing a cleavage site was taken as the average probability across both models. For epitope data, models were trained on window sizes of 7 through 21 amino acids and subsequently compared to identify the window size with optimal performance. Due to the relatively small size of the *in vitro* dataset, only the minimum (7) and maximum (21) length window libraries were assessed, with additional information on constitutive versus immunoproteasomal context included as in the same manner used for the gradient boosted approach.

2.3.6 Analysis of sampled feature space

Because models were studied using variable window lengths up to 21 amino acids, we evaluated all data in a unified feature space, noting that the largest feature space would be guaranteed to provide the most robust assessment across all window lengths. To qualitatively assess how well our training data represented the broader space of possible peptides, we therefore identified all unique 21 amino acid windows within the human proteome (https://www.uniprot.org/proteomes/UP000005640). Using these windows as background, we compared the shared UMAP space calculated with the first 10 principal components across the human proteome, as well as both *in vitro* and *in vivo* training sets using the four chemical

properties at each amino acid position within the window, described previously, as the input feature set (Fig. 2.5). Furthermore, we compared the sampling density for both datasets to the human background set across the first 4 principal components to demonstrate the distribution of sampling in our training sets (Appendix A, Fig. 6.2). In addition to plots comparing the sampling space based on chemical properties, we also generated logo plots based on the amino acid frequencies for positive and negative examples in each training set (Appendix A, Fig. 6.3, 6.4). These plots were generated using the ultimate window sizes retained in modeling; 7 amino acids for *in vitro* data and 17 amino acids for *in vivo* data.

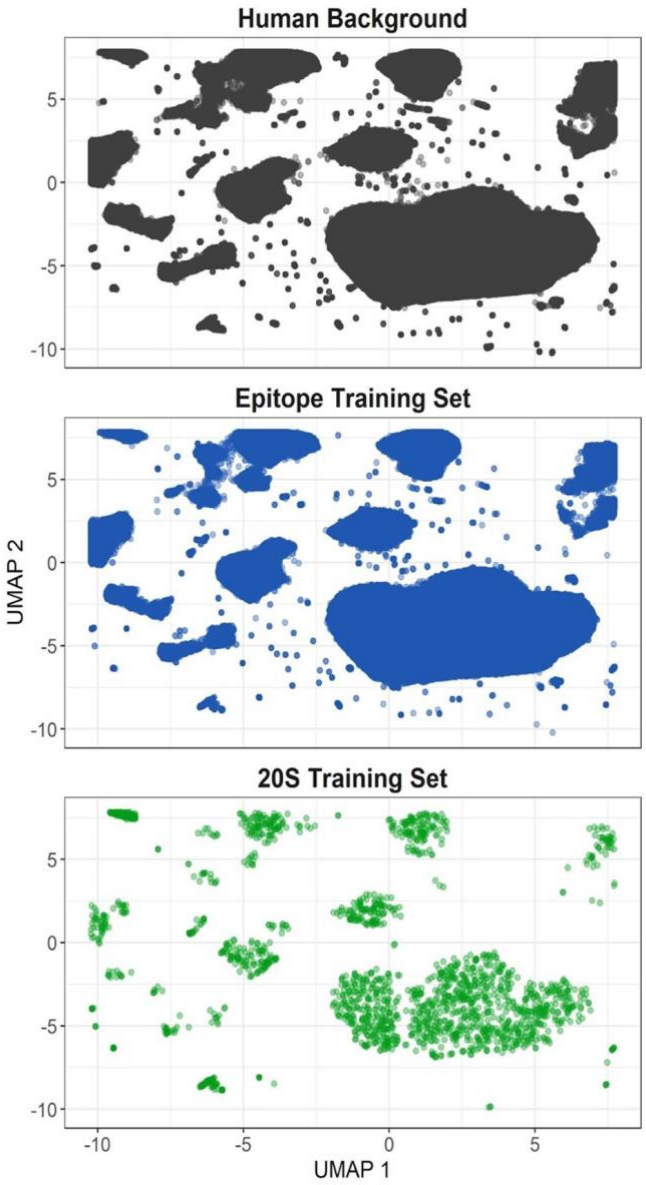


Figure. 2.5. Training set projections in UMAP space. *Amino acid windows (21 residues long) were generated for the whole human proteome (gray), all epitope training examples (blue) and all 20S training examples (green). Principle components were generated from the physical properties of each amino acid at each window position. UMAP projections were generated from the first 10 principal components.*

2.3.7 Collection and processing of in vitro digestion map data for validation

Data from six 20S studies was held out from previous steps to be used in validation. In order to accommodate the analysis of performance for models using multiple different window sizes, window lengths of 21 amino acids were generated for each validation set cleavage example using reported peptide fragments and their source protein contexts. Companion non-cleavage windows were generated in the same way as before, with the exception that only one internal site was sampled at random during non-cleavage window generation to create an initially balanced set of positives and negatives. For validation windows, the additional step of filtering all validation windows with a non-unique interior window of 7 amino acids (smallest centered window size for the models to be assessed) was also taken to ensure no redundant windows were present in the validation set for any training window sizes from 7 to 21 amino acids. These windows were then screened against all training and testing examples to only retain unique, never before seen, entries in their respective sets. Ultimately this generated 171 constitutive 20S cleavage windows and 54 immunoproteasome 20S cleavage windows.

2.3.8 Collection and processing of epitope data for validation

Data from Bassani-Sternberg *et al.* (2016)[126] was held out from previous steps to be used in validation. As described above, 21 amino acid windows were generated using reported epitopes and their source protein contexts. These windows were then screened to only retain unique entries in their respective sets and companion non-cleavage examples were generated as described previously, with the exception that only one internal site was sampled at random during non-cleavage window generation to create an initially balanced set of positives and negatives as with the in-vitro validation set. For validation windows, the additional step of

filtering all validations windows with a non-unique interior window of 7 amino acids (smallest centered window size for the models to be assessed) was also taken to ensure no redundant windows were present in the validation set for any training window sizes from 7 to 21 amino acids. Finally, each window was also screened against those used in the training or testing sets to ensure none had been previously seen by the trained models. Ultimately, this generated 7951 cleavage windows for validation.

2.3.9 Model implementation and availability

Our *in vivo and in vitro* cleavage models were implemented in Python version 3.7. All deep learning models were generated using PyTorch version 1.3.1[138], while all machine learning models were generated using Scikit-learn version 0.22.1[136]. The full instructions and code for replication of the analyses contained herein can be found at https://github.com/pdxgx/pepsickle-paper, while the fully deployed command line version of `pepsickle`, along with relevant installation instructions can be found at https://github.com/pdxgx/pepsickle. To better handle a variety of use cases, including analysis of long and short peptides, the deployed version of `pepsickle` recognizes two different padding characters: (*) for terminal sequences with no adjacent amino acids and (X) for sequences flanked by unknown amino acid residues (see `pepsickle` readme for full details). `pepsickle` is open source and available under the MIT user license.

2.3.10 Comparison of cleavage prediction tools

A literature search and browser query were performed to identify currently available tools for proteasomal cleavage prediction (search terms included ‘cleavage prediction’, ‘proteasomal prediction’, ‘cleavage prediction tool’ and ‘proteasomal cleavage prediction’).

Through this search, six tools were identified including: NetChop 3.1[81], the Proteasomal Cleavage Prediction Server (PCPS)[28], PCleavage[83], MAPPP[84], PAProC[85] and the random forest-based model described in Li *et al.* (2012)[86]. NetChop version 3.1 was downloaded from http://www.cbs.dtu.dk/services/NetChop/ and installed as a command line tool on a Linux server running CentOS 7.7.1908 after which cleavage windows for both *in vitro* and *in vivo* validation examples were given in FASTA format to their respective model types. Predictions were saved and assessed only for the point of potential cleavage in each window. PCPS was run via its web server implementation at http://imed.med.ucm.es/Tools/pcps/ with both constitutive proteasome and immunoproteasome options selected. For each *in vitro* data type, the model corresponding to the proteasome type was used, with only the midpoint of each window reported and recorded as described above. For *in vivo* epitope windows, both models were assessed in the same fashion, with results reported for the model achieving the best AUC. PCleavage was also run via its web server implementation at http://crdd.osdd.net/raghava/pcleavage/, however validation assessment was only performed for *in vivo* epitope windows using the default threshold of 0.3. For both constitutive proteasome and immunoproteasome data; PCleavage did not accept windows that spanned the C- or N-termini of a given source protein, which reduced the validation set size substantially and prevented paired comparisons with the other models available. Additionally, we were unable to confidently reproduce the training dataset(s) for NetChop 3.1, PCPS, and PCleavage, and thus could not guarantee our validation set was mutually exclusive with training data for these tools across all trained models. Three cleavage prediction tools were ultimately not functional in our hands: the MAPPP server is no longer available, and we were unable to locate a publicly downloadable version of the tool, we were unable to obtain a working copy of PAProC II despite

repeated requests, and we were unable to locate any web server or public tool implementing the model from Li *et al.*

Computational performance was assessed for all tools not reliant on a web server (i.e. NetChop 3.1, `pepsickle`). Using a dedicated node (Intel Xeon E5-2697 v2 2.70 GHz, single thread mode) on a Linux server running CentOS 7.7.1908, both *in vitro* and epitope-based models for NetChop 3.1 and `pepsickle` were applied to a performance test set consisting of all proteins in the human proteome (https://www.uniprot.org/proteomes/UP000005640). Total CPU times were calculated as the 'user' time + 'sys' time for each prediction model (Appendix A, Table 6.4).

2.3.11 Model cross-comparison assessments

The cross performance of both constitutive-based and immuno-based *in vitro* models were assessed using the same *in vivo* validation set used for our epitope trained model (Appendix A, Table 6.5). Cleavage predictions were generated using the internal 7 amino acid window centered within each larger 21 amino acid validation window. Predictions were reported at the same central amino acid with the default prediction probability threshold of 0.5 used for determining cleavage versus non-cleavage predictions. We reasoned that positive epitope cleavage examples should be predicted with good accuracy while negative examples would not be due to the complex selectivity of the downstream antigen processing and presentation pathway (e.g. MHC binding); therefore, only the percentage of correctly captured positive cleavage examples was assessed (sensitivity). This removes the possibility of misreporting true cleavage events that are filtered during post-cleavage processing as model misclassifications.

2.3.12 Collection of patient-derived immune response data and model application

Three primary literature articles including patient specific predicted tumor neoepitopes and epitope-specific immune responses were identified for model application, including: (i) the Ott *et al.* patient-specific melanoma vaccine study[139], (ii) the MuPeXI neoepitope prediction study[140] and (iii) a large scale neoepitope prediction comparison from the Tumor Neoantigen Selection Alliance (TESLA)[141]. From sources 1 and 2 where gene/protein sources for each predicted epitope were provided, each mutated candidate was mapped back to its original proteomic position to retrieve upstream (10 amino acids) and downstream (10 amino acids) contexts. For predicted neoepitopes within 10 amino acids from the start or end of the protein, positions were buffered using '*' prior to model input. For predicted neoepitopes reported in the TESLA study the original source proteins were not provided. Instead, candidate neoepitope sequences were queried against the human reference proteome using the BLAST[142] command line tool with the following parameters: `'-matrix BLOSUM62'`, `'-evalue 200000'`, `'-comp_based_stats F'`. Only ungapped alignments were retained, allowing for a singular mismatch at the mutated position with exact matches at all other positions. Protein contexts around each candidate neoepitope were generated as described for the other two studies, however all candidate neoepitopes resulting in more than one unique context window were filtered out to remove any candidate neoepitopes with an ambiguous source in the proteome. All predicted neoepitopes across the three studies were also annotated as 'responsive' or 'non-responsive' based on the reported patient specific immune response. This resulted in 762 candidate neoepitopes, of which 45 (5.9%) were reported as inducing a patient-specific immune response.

After all context windows were collected, our *in vivo* `pepsickle` model was applied to the C-terminal position of each proposed neoepitope candidate, returning the predicted C-terminal cleavage probability. Median cleavage probabilities for predicted neoepitopes that elicited a patient specific immune response were compared to those that were predicted but for which an immune response was not verified using a Wilcoxon ranked sum test. Additionally, the use of cleavage probability as a classification threshold was assessed using the 25th percentile of predicted cleavage probabilities across all candidate neoepitopes as a cutoff. The proportion of responsive versus non-responsive neoepitopes that were properly identified using this thresholding approach was assessed using a Chi-square test for independence.

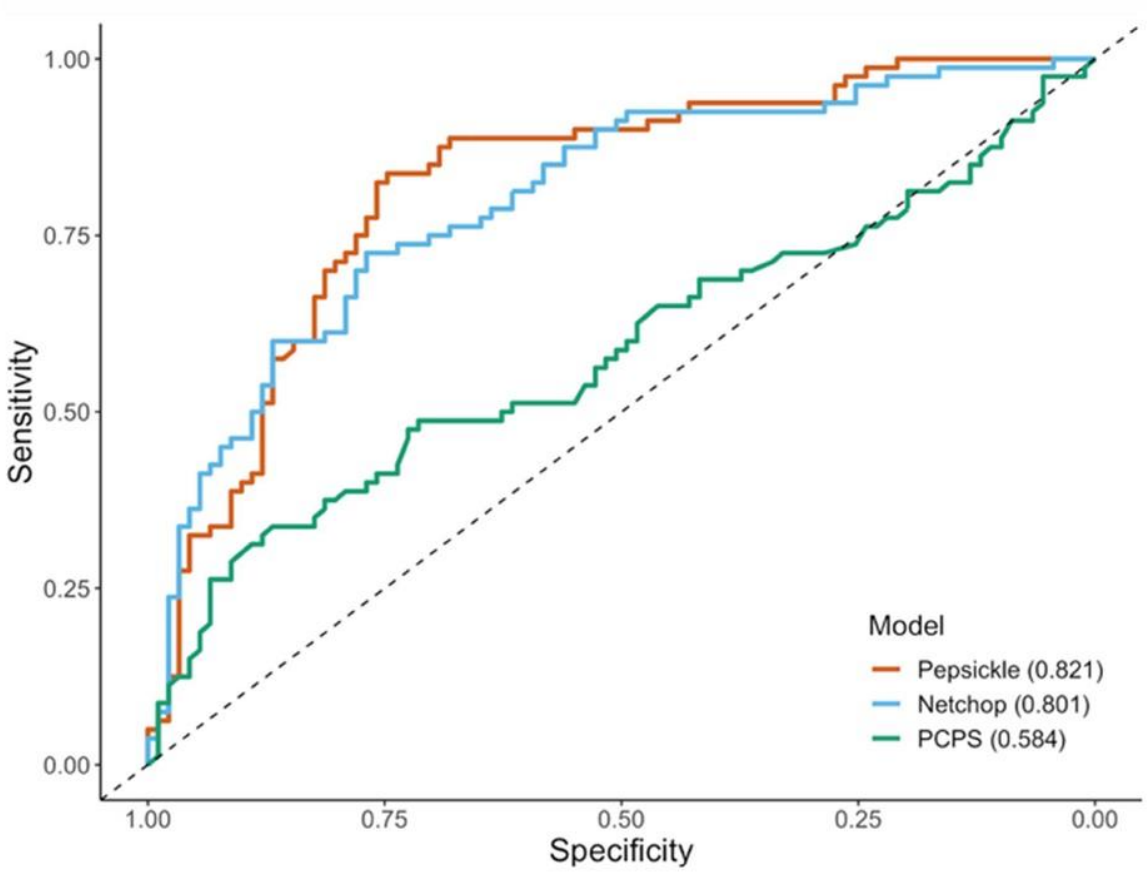
2.4 Results

2.4.1 In vitro digestion-based cleavage prediction

We identified 35 publicly available *in vitro* digestion datasets, constituting both 20S constitutive and 20S immunoproteasomal cleavage experiments (Table 2.1). From these, six studies were reserved for external validation (validation set), while the rest were aggregated to generate a training and testing dataset containing cleavage information from 1984 peptide fragments generated across constitutive, immuno- and mixed proteasomal contexts. We then trained a gradient boosted classifier based on windows of seven amino acids in length centering on each cleaved site. Residues within the window were encoded as the physical properties (polarity, molecular volume, hydrophobicity, and conformational entropy) of each amino acid at each given position in the window. Using annotated proteasome types, the model was trained to differentiate between sites cleaved by the immunoproteasome and those cleaved by the constitutive proteasome, returning the probability of cleavage at the center of each window. This model achieved a test set AUC of 0.759 (Appendix A, Table 6.6). We explored whether

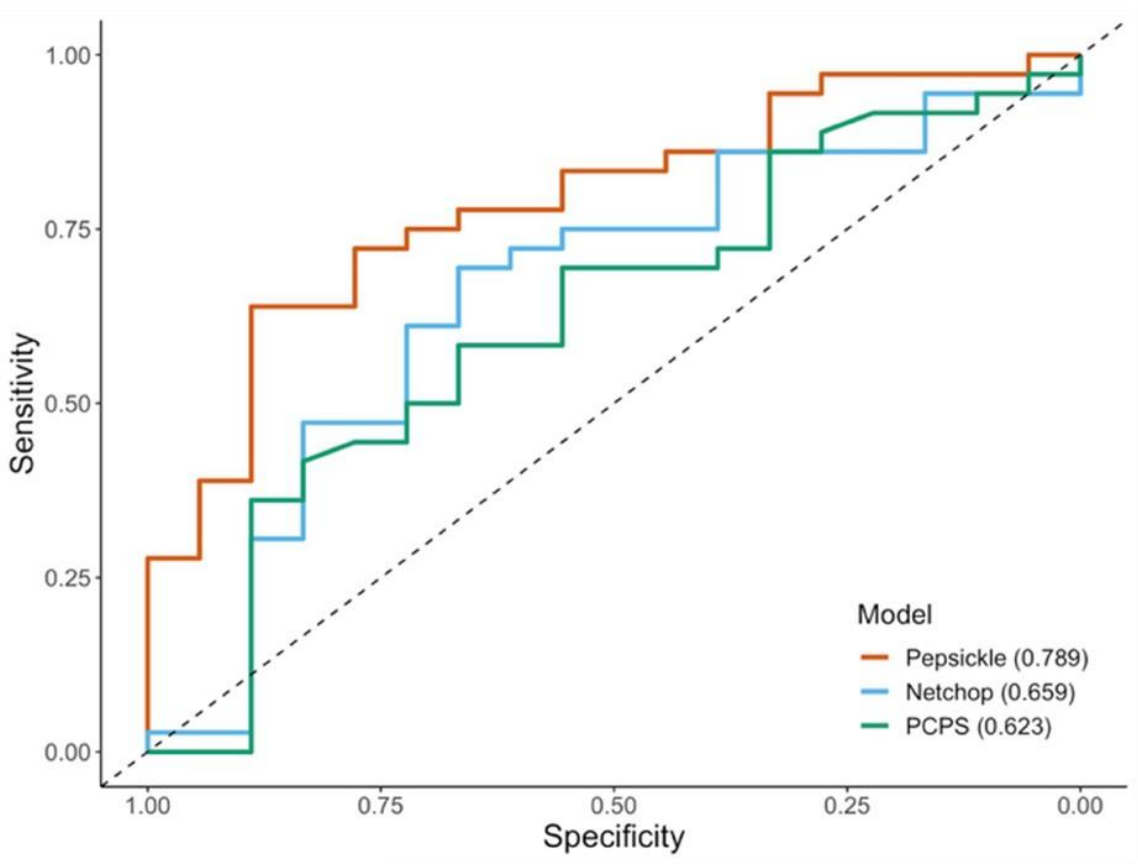
additional peptide context around each cleavage site (21 amino acid windows) improved model performance; however, a comparison of models trained with both 7 amino acid and 21 amino acid window sizes showed no increase in AUC when applied to testing the data (DeLong's T-test, $P = 0.558$). Similarly, we assessed whether a fully connected feed-forward deep learning model could improve cleavage predictions over the initial machine learning approach. The use of this feed-forward network model also did not appear to increase performance significantly ($P = 0.558$). We therefore report and discuss the results from our seven amino acid gradient boosted classifier (`pepsickle`) hereafter.

We next assessed *in vitro* `pepsickle` performance on an independent validation set, consisting of 171 constitutive proteasome and 54 immunoproteasome examples, respectively. Our model achieved an AUC of 0.821 on the constitutive proteasome validation set and 0.789 on the immunoproteasome validation set, respectively. Using the same validation sets, we assessed the corresponding performance of existing tools including NetChop 3.1 and the Proteasomal Cleavage Prediction Server (PCPS) (Fig. 2.6). Note that PCleavage was omitted from these *in vitro*-based comparisons due to its inability to process cleavage sites whose context windows span a peptide fragment's N- or C-termini (54.4% of the constitutive and 81.5% of the immuno validation data respectively). We found that `pepsickle` has significantly higher predictive performance on constitutive proteasomal data compared to PCPS, but similar performance compared to NetChop 3.1 (Fig. 2.6). When applied to immunoproteasomal data, our model compared similarly to both PCPS and NetChop 3.1, acknowledging limited statistical power to detect a difference given the small sample size (Fig. 2.7; Appendix A, Table 6.7).



Tool Comparison	Proteasome Type	Z-score	Adjusted p-vaue
pepsickle v. NetChop	Constitutive	23.25	0.823
pepsickle v. PCPS	Constitutive	28.24	**<0.001**

Figure 2.6. Performance comparison of cleavage prediction models on constitutive proteasome data. *Receiver operating characteristic (ROC) curves are shown for each of three cleavage prediction models, as denoted in legend, with corresponding area under the curve (AUC) values reported in parentheses. Sensitivity (y-axis) and specificity (x-axis) were both evaluated using a validation set (n = 171) consisting of 80 cleavage and 91 non-cleavage in vitro examples not seen during the training or testing of our models (see Methods). For pepsickle (our model) and PCPS, the constitutive proteasome models with default settings were used. For NetChop 3.1, the in vitro model was used with default settings (no specification is available for proteasome type). PCleavage was omitted from this comparison due to restrictions on window sizes and the inability to process the full set of validation examples. Statistical pairwise comparisons of ROC curves (Delong's tests) are shown in corresponding table values (Z-score), with significance reported as P-values after Benjamini–Hochberg correction for multiple comparisons.*



Tool Comparison	Proteasome Type	Z-score	Adjusted p-vaue
pepsickle v. NetChop	Constitutive	23.25	0.116
pepsickle v. PCPS	Constitutive	28.24	0.116

Figure 2.7. Performance comparison of cleavage prediction models on immunoproteasome data. *Receiver operating characteristic (ROC) curves are shown for each of three cleavage prediction models, as denoted in legend, with corresponding area under the curve (AUC) values reported in parentheses. Sensitivity (y-axis) and specificity (x-axis) were both evaluated using a validation set (n = 54) consisting of 36 cleavage and 18 non-cleavage in-vitro examples not seen during the training or testing of our models (see Methods). For pepsickle (our model) and PCPS, the immunoproteasome models with default settings were used. For NetChop 3.1, the in vitro model was used with default settings (no specification is available for proteasome type). PCleavage was omitted from this comparison due to restrictions on window sizes and the inability to process the full set of validation examples. Statistical pairwise comparisons of ROC curves (Delong's tests) are shown in corresponding table values (Z-score), with significance reported as P-values after Benjamini–Hochberg correction for multiple comparisons.*

2.4.2 Epitope-based cleavage prediction

To better interrogate *in vivo* proteasomal cleavage, we identified 357 253 naturally processed human and mammalian class I epitopes from publicly available data (Table 2.2). Using a deep learning framework, we trained a consensus-based neural network on amino acid sequence and physical properties to predict epitope C-terminal cleavage events, independent of proteasome type (Appendix A, Fig. 6.1). Performance of our deep learning model was compared across all odd window sizes ranging from 7 amino acids to 21 amino acids, with windows centered on the cleavage site as described above (Appendix A, Table 6.8). When applied to testing data, the model trained on 17 amino acid windows performed significantly better than the model trained on 7 amino acid windows, however increasing window size beyond 17 amino acids did not improve performance further (Appendix A, Fig. 6.5). This is consistent with the average cleaved peptide length of 8 amino acids[92] and the potential for bi-directional proteasome entry and processing[143]. We additionally studied the influence of model complexity, finding that the consensus-based deep learning approach performed better than a more simplistic random forest model trained using the same window size (DeLong's T-test, $P = 0.021$). We therefore report and discuss results from the consensus-based model (`pepsickle`) using 17 amino acid windows hereafter.

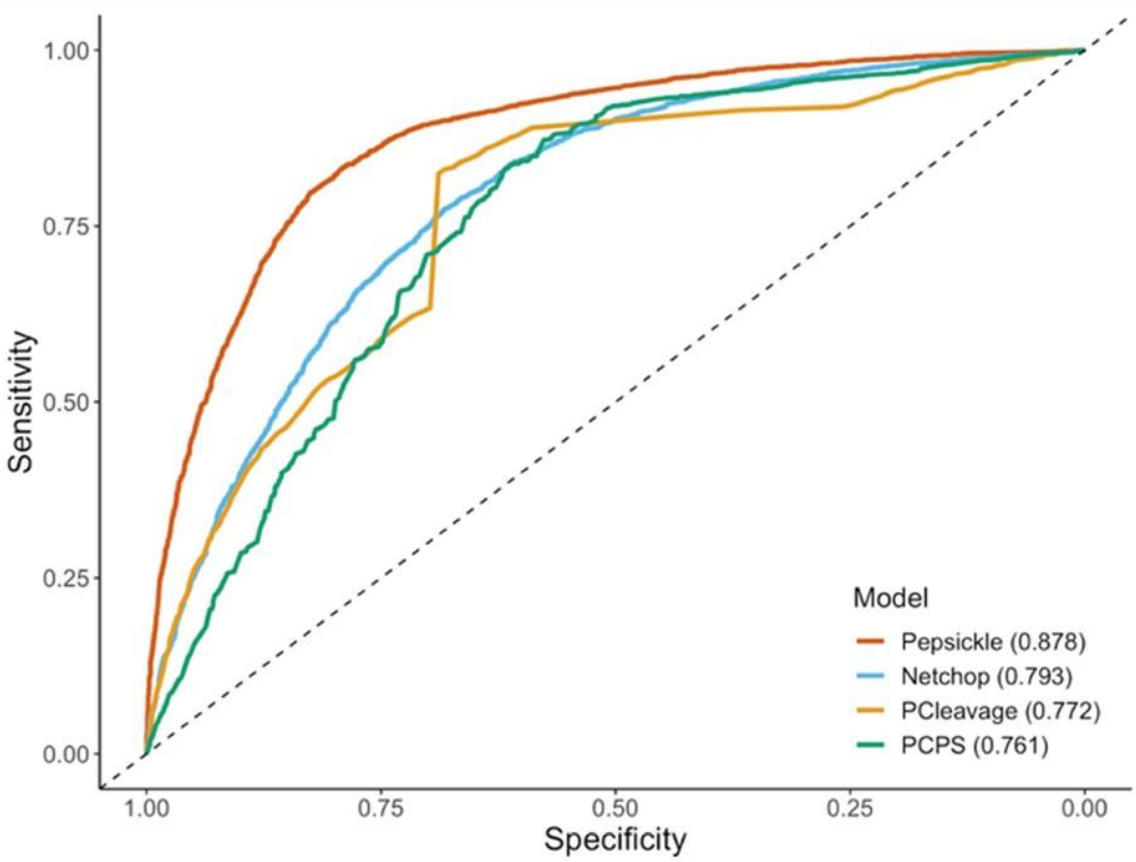
We next assessed performance of the `pepsickle` epitope model on an independent validation set. This dataset consisted of 7951 examples not present in either the training or testing datasets used by our study. Notably, since training sets for all other tools were not explicitly available, we cannot guarantee that our validation set was entirely unseen by other models during their respective training processes. Nonetheless, when applied to this validation data, our deep learning-based ensemble net achieved an AUC of 0.878, representing a

significant improvement in AUC over the corresponding performances of existing tools including NetChop 3.1, PCPS and PCleavage collectively (Fig. 2.8). In addition, `pepsickle` showed better recall and F-1 score than the other models compared (Appendix A, Table 6.9).

Table 2.2 Summary of epitope data sources.

Source	Dataset	Fragments reported
Immune epitope database (IEDB) (Köhler *et al.*, 2001)	Train/test	4 98 419
SYFPEITHI database (Nussbaum *et al.*, 1998)	Train/test	4433
Rozanov *et al.* (2018)	Train/test	3254
AntiJen database (Kisselev *et al.*, 1999	Train/test	1492
Bassani-Sternberg et al. (2016)	*Validation*	*99 356*

Note: Sources of epitope data used for training, testing and validation.



Tool Comparison	Proteasome Type	Z-score	Adjusted p-vaue
pepsickle v. NetChop	-	23.25	**<0.001**
pepsickle v. PCPS	-	28.24	**<0.001**
pepsickle v. PCleavage	-	21.3	**<0.001**

Figure 2.8. Performance comparison of cleavage prediction models on epitope data. *Receiver operating characteristic (ROC) curves are shown for each of four cleavage prediction models, as denoted in legend, with corresponding area under the curve (AUC) values reported in parentheses. Sensitivity (y-axis) and specificity (x-axis) were both evaluated using a validation set (n = 7951) consisting of 3566 cleavage and 4385 non-cleavage epitope examples not seen during the training or testing of our models (see Methods). Default epitope-based models were used for pepsickle (our model), Netchop 3.1 and PCleavage predictions, while constitutive model predictions from the default model 1 were used for PCPS (PCPS immunoproteasome predictions were inferior and therefore omitted). Statistical pairwise comparisons of ROC curves (Delong's tests) are shown in corresponding table values (Z-score), with significance reported as P-values after Benjamini–Hochberg correction for multiple comparisons*

2.4.3 Computational performance of `pepsickle`

In addition to predictive ability, we also assessed the computational speed of `pepsickle`, for both *in vitro* and epitope-based cleavage predictions. Using a list of all protein sequences in the human proteome as a benchmark dataset (*n* = 113 576, including all isoforms and computationally predicted sequences), `pepsickle` was able to achieve a total processing time of 154 m 46 s for *in vitro* predictions (approximately 124 ms per 1000 predictions) and 158 m 21 s for epitope predictions (approximately 127 ms per 1000 predictions) (Appendix A, Table 6.4). These times were compared to NetChop 3.1 run in an identical controlled computing environment. We found that `pepsickle` is 68.5% faster than NetChop 3.1 for *in vitro* cleavage predictions (154 m 46 s versus 260 m 50 s) and 242% faster for epitope-based predictions (158 m 21 s versus 542 m 40 s).

2.4.4 *In vitro* digestion and *in vivo* epitope-based models differ in prediction performance, but with similar feature importance

Despite the substantial differences between sources and structure of training data for both *in vitro* digestion and *in vivo* epitope-based models that hinder the creation of a unified model, we sought to evaluate commonalities in the learned feature sets by evaluating cross-performance of our *in vitro* model on epitope validation data. Acknowledging that epitope-based data is implicitly subject to multiple components of the antigen processing pathway following proteasomal cleavage[144], we evaluated the accuracy of the *in vitro* model exclusively on positive cleavage examples from the *in vivo* epitope validation set (i.e. all positive examples which must have necessarily undergone proteasomal cleavage). Based on this metric, our *in vitro* constitutive model was able to correctly identify 69.9% of the cleavage events observed in the epitope validation set, while our immunoproteasome model was able to

correctly identify 54.5%. Performance by both *in vitro* models on this data is substantially lower than the performance of the original epitope-based model, which was able to capture 82.8% of true cleavage events. However, we acknowledge that the epitope validation data contains an unknown mixture of both constitutive and immunoproteasome-based cleavage, which may contribute to the relatively lower performance of both *in vitro* models in this case.

Because cross-data assessments for both *in vitro* models represent a substantial performance decrease compared to assessment on like-kind data, we sought to further qualify the distinct commonalities and differences between *in vitro* digestion and *in vivo* epitope-based datasets. Using 21 amino acid windows, we compared both training sets to a set of all possible 21 amino acid windows from the human proteome. By overlapping UMAP projections of the windows sampled in the *in vivo* epitope set with those generated by the human proteome, we were able to visually demonstrate that the majority of the sample space constituted by the human proteome was sampled; however substantial portions were under sampled in the *in vitro* dataset compared to the *in vivo* data (Fig. 2.5). Similarly, the underlying density distribution for samples in the *in vitro* dataset differed substantially from that seen in both the *in vivo* dataset and human proteome background (Appendix A, Fig. 6.2).

To further investigate whether differences in the training set representations altered the learned features for each prediction model, we plotted the feature importance for both our *in vivo and in vitro* models (Appendix A, Fig. 6.6, 6.7). Acknowledging that model weights are not directly comparable, we found that similar patterns of amino acid physical properties identified cleavage sites across both models: in particular, low molecular volume and low hydrophobicity were important at the C-terminal amino acid, along with low conformational

entropy at the '1' position and high polarity at the '2' position compared to other features at the same locations.

2.4.5 Proteasomal cleavage helps predict epitope-specific immune responses

We next assessed the potential additive contribution of our model to predicting epitope-specific immune responses in real-world patient data. We identified 762 candidate epitopes from three studies with extensive immunoprofiling data: (i) the Ott *et al.* patient-specific melanoma vaccine study[139], (ii) the MuPeXI neoepitope prediction study[140] and (iii) a large scale neoepitope prediction benchmarking effort from the Tumor Neoantigen Selection Alliance (TESLA)[141]. From these studies, we identified 45 epitopes that elicited an immune response, as well as 717 non-responsive epitopes.

Using the `pepsickle` epitope-based cleavage model, we predicted C-terminal cleavage probability for all predicted epitopes regardless of corresponding immune response status (Fig. 2.9). We demonstrated that the median terminal cleavage probability is significantly higher for immune responsive epitopes compared to those that were predicted but did not elicit an immune response (Wilcoxon ranked sum test, $P = 0.036$). Despite the heavy pre-selection of these epitopes using a collection of predictive methodologies, we find that `pepsickle`-based cleavage thresholding (≥25th percentile threshold) significantly enriched the proportion of immune responsive epitope candidates with 40% of responsive versus 24.4% of non-responsive candidates falling in the top quartile ($\chi_1^2 = 4.86$, $P = 0.027$). This represents a 59.6% increase in the positive predictive value after cleavage-based filtering. Notably, we find that cleavage predictions for two studies follow the trend seen in the aggregate data[139], [141], but the third study does not[140]. While this heterogeneity warrants further investigations, these findings

suggest that even when used as a *post hoc* filter, `pepsickle`-based cleavage predictions may help improve the identification of patient-specific, immune-responsive neoepitopes.

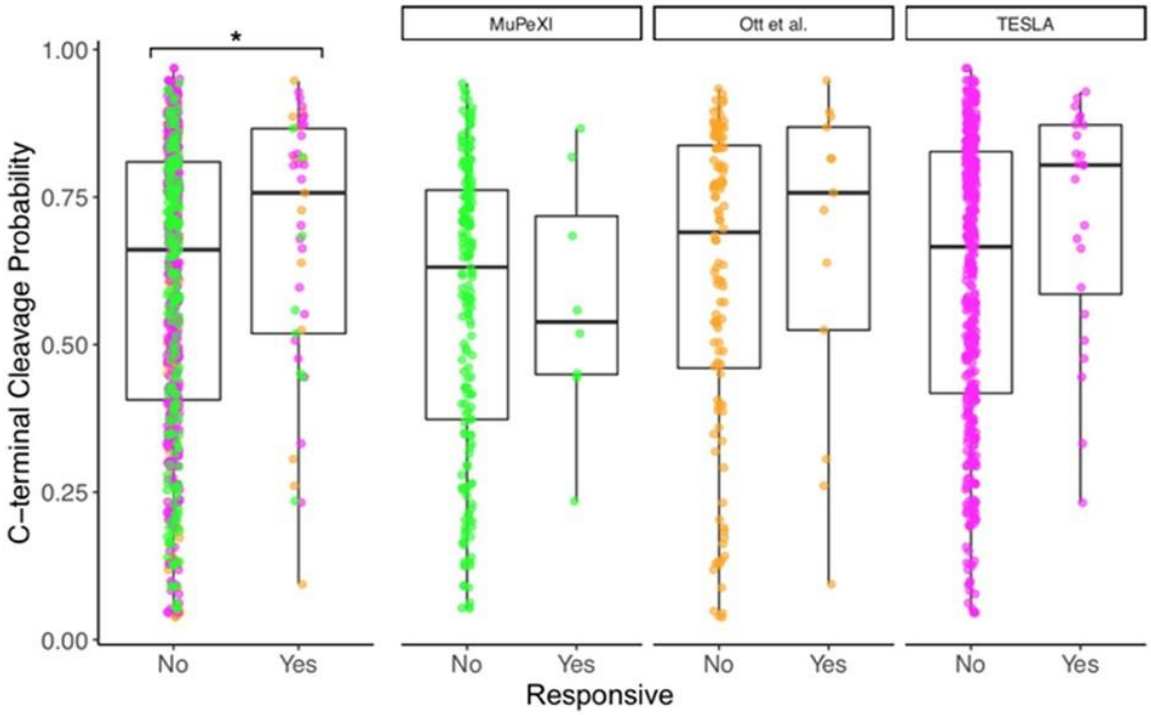


Figure 2.9. C-terminal cleavage predictions on patient neoepitopes. *Predicted neoepitopes from three studies were accumulated. Cleavage predictions at the C-terminus of each epitope were generated using pepsickle and plotted, with predicted epitopes divided based on whether or not they elicited an immune response. Box plot results are shown in aggregate (left) as well as on a per-study basis (right) with values indicating median (horizontal black lines), 25–75%ile (box) and range ('whiskers'), and with colors corresponding to the study origin (green = MuPeXI, orange = Ott et al., magenta = TESLA). For the aggregated dataset, the median C-terminal cleavage probability is significantly higher (*) for responsive epitopes compared to non-responsive epitopes (Wilcoxon ranked sum test, P = 0.036).*

2.5 Discussion

To the best of our knowledge, the data aggregated for this study represents the largest compilation of *in vitro and in vivo* cleavage events to date. Applying machine and deep learning techniques to this data, we have improved upon the current state of the field by developing an *in vivo* model of proteasomal cleavage prediction with improved performance (AUC) over

currently available tools. In addition, we have created an *in vitro* model with performance comparable to the current best-in-class model, NetChop 3.1, but with significantly decreased computational costs and the ability to differentiate between immunoproteasome and standard proteasome cleavage profiles. Although further investigation is needed, application of our *in vivo* model to patient-derived neoepitope data suggests that including cleavage information in the epitope prediction process may improve novel target identification when applied as an additional filter and may be a key component missing from the majority of current prediction tools. This is consistent with recent evidence demonstrating the value of incorporating proteasomal cleavage predictions into epitope prediction pipelines[145].

Despite `pepsickle`'s promising performance using both *in vivo* and *in vitro* models, we note several limitations to our work. The primary challenge given the structure of the *in vivo* data, is that non-cleavage events must be determined heuristically. Although we use stringent filtering criteria throughout our pipeline, accurate negative examples are reliant on sufficient sampling of true cleavage events and may be biased by lack of reporting for less studied portions of the proteome. We also note that differences in the definition of non-cleavage training sites between tools may affect some comparison statistics such as precision and F1-score. While this issue could potentially be addressed by re-training other model architectures using a consistent non-cleavage site definition, this is not possible with closed source tools such as NetChop 3.1. The inability to re-train closed source models and a lack of specific details on model architectures also means that we could not delineate the effects of a larger training set size from differences in model design across tools. Similarly, we did not assess the relative performance differences of explicitly encoded amino acid physical properties that we used versus the implicitly encoded physical properties (e.g. BLOSUM matrices) used in other tools.

While our *in vivo* model performs substantially better than our *in vitro* model on its respective validation set, we note that *in vivo* data was used exclusively for C-terminal cleavage event prediction leaving N-terminal events largely unrepresented in this context. We suspect the difference in optimal window sizes between the *in vivo* and *in vitro* models is a reflection of the shorter peptide fragment inputs to *in vitro* cleavage experiments compared to full protein contexts in *in vivo* datasets but note there may be stochastic or additional latent technical or biological explanations for this distinction. We also note that immunoproteasomal training data was particularly limited due to the scarcity of source data and may help explain the poorer immunoproteasomal cross-performance on *in vivo* epitope data.

We did not see evidence that our models learned features unrelated to cleavage, such as MHC binding[146], but it remains possible that these and other latent biological features may have been partially learned by our models in addition to true cleavage-specific features (Appendix A, Fig. 6.6, 6.7). Additionally, our *in vitro* models are based on relatively small datasets with heterogeneous experimental methodologies, and only a small subset cleanly evaluate the respective roles of the constitutive and immunoproteasomes on the same source proteins. For both *in vivo and in vitro* data, poor sampling from some regions of the proteome is also of concern, due at least in part to a scientific focus on proteins relevant for cancer and autoimmunity, as well as the experimental limitations of mass spectrometry[147]. While the data suggests our model should perform well on previously unseen data, the discrepancies seen in model application to per-study immune response data raise questions of broader generalizability in certain applied contexts. Ultimately, epitope immunogenicity relies on additional immune context and not epitope sequence alone. The data presented herein demonstrates that using `pepsickle` as an additional filtering step can enrich for

immunogenic peptides, however the true immunogenicity of a given epitope relies on more than just sequence and cleavage profile alone.

`pepsickle` provides a promising, open-source, tool for proteasomal cleavage prediction, which may be implemented on its own or otherwise integrated into existing epitope prediction pipelines. Given the recent successes and increasing emphasis on developing and deploying mRNA-based vaccines for individual patients[148] and whole populations[76], [149], any concrete improvements in the accuracy of these epitope prediction pipelines could carry transformative clinical value. We also note that an improved capacity to predict proteasomal cleavage could contribute to our understanding of protein turnover and recycling in healthy and diseased contexts[150], [151] and lead to improvements in rational protein design[152].

The performance and potential of `pepsickle` described in this text are encouraging, however many questions remain unanswered. The heterogeneity of C-terminal cleavage profiles seen in our study specific `pepsickle` application raises the question of whether cleavage prediction is universally helpful in target identification, or if specific study or design contexts are required to see benefit from cleavage predictions. In addition, whether or not there is an impact of using proteasomal prediction ad-hoc (via integration with existing neoepitope prediction pipelines[26], [153]) versus *post hoc* remains to be seen. Application of `pepsickle` to more patient derived data in the future will help us better understand the broader potential of applying cleavage prediction in this space, with the potential for broad implications in the research and clinical communities.

2.6 Data availability

Source code is available at https://github.com/pdxgx/pepsickle-paper under the Massachusetts Institute of Technology (MIT) license, including data extracted from primary literature at https://github.com/pdxgx/pepsickle-paper/tree/master/data/raw and scripts for parsing public databases mentioned herein at https://github.com/pdxgx/pepsickle-paper/tree/master/scripts/database_pulls.

2.7 Funding and conflicts of interest

This work was supported by VA Career Development Award (1 IK2 CX002049-01) and the Sunlin & Priscilla Chou Foundation [to R.F.T.].

Conflict of Interest: none declared.

Disclaimer: The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government.

Chapter II: Predicting T-cell Cross-Reactivity Using Paired Peptide Data: A Multi-Layer Perceptron Approach

Benjamin R. Weeder[1] and Reid F. Thompson[1,2,3,4*]

[1]Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR 97239, USA

[2]Department of Radiation Medicine, Oregon Health & Science University, Portland, OR 97239, USA

[3]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR 97239, USA

[4]Division of Hospital and Specialty Medicine, VA Portland Healthcare System, Portland, OR 97239, USA

*To whom correspondence should be addressed.

This manuscript draft is in preparation for submission.

3.1 Abstract

Individual T-cell receptors have the potential to recognize multiple antigens, enabling broad immunological surveillance and powerful adaptation to evolving threats. However, this potential cross-reactivity poses a simultaneous risk for the development of autoimmunity and immune-related adverse events. While we have a growing understanding of how TCR sequences may influence their degeneracy, predictive methods relying on receptor sequence have limited applicability for most real-world use cases. Here we present `crossreactor`, a multi-layer perceptron approach to cross-reactivity modeling that leverages paired peptide antigens mined from large-scale databases. By relying on epitope intrinsic features and a paired dataset structure, we demonstrate how cross-reactive events can be accurately modeled and predicted in novel application contexts. This proof-of-concept approach allows for application of cross-reactive modeling to much broader contexts, including auto-immunity, infectious disease, mRNA vaccination, and cancer immunotherapy.

3.2 Introduction

The mammalian adaptive immune system has evolved to defend against a vast array of potentially harmful microorganisms. In particular, the ability for nucleated cells to present self and foreign antigens via class I major histocompatibility complexes (MHCs) is a highly conserved and essential mechanism for fighting viral infection and even aiding in the clearance of malfunctioning host cells[154], [155]. These class I antigen-MHC complexes are then surveyed by CD8 T-cells to differentiate normal from abnormal presented peptides.

To facilitate their ability to recognize a wide array of targets, T-cells undergo a complex process during early development called V(D)J recombination[156]. In this process the variable (V), diversity (D) and joining (J) regions of the genome, which are ultimately used to construct the T-cell receptor (TCR), undergo genomic rearrangement resulting in unique receptor identities. These randomly rearranged receptor sequences are then screened during further T-cell development in the thymus via both positive and negative selection. During positive selection T-cells and their presented TCR are screened for functionality against host MHC complexes, while negative selection screens against host auto-reactivity[157]. Ultimately, CD8+ T-cells that recognize host MHC's, but don't react to self-antigens in the thymus continue through the full maturation process.

Given the incredible number of potential unique TCR sequences possible (~$1x10^{15}$), and the intense selection process during thymic development, T-cells are often described as highly specific[37]. Despite this, TCR's are regularly reported as being activated by multiple epitope sequences[158]. While the concept of highly specific TCR's that demonstrate sequence degeneracy seems initially counterintuitive, this phenomenon is consistent with an evolutionary understanding of T-cell biology. Combinatorial estimates show that using all 20 canonical amino acids, over 500 billion possible 9 amino acid long epitopes can be constructed. While it's unlikely that all possible combinations of 9 amino acid epitopes are biologically relevant, studies have shown that class I epitopes regularly range from 8-11 amino acids long - the biological combinations of which dramatically dwarf the number of possible T-cells in a given host[36]. Although T-cell specificity is highly desired, the evolutionary necessity of TCR degeneracy to address a disproportionate ratio of host T-cells to potential targets is clear.

Beyond its role in expanding the possible pool of recognizable peptides, cross-reactivity also plays a key role in combating ever-evolving viral strains. Viral studies have demonstrated

conferred immunity to novel viral sequences through exposure to similar viral strains in the same family[159]. Given the disparity in evolutionary time span between humans and viruses, this functionality is pivotal in host defense. Unfortunately cross-reactivity can also have negative consequences, seen in some examples of autoimmunity where T-cells mis-recognize self-antigens despite passing negative thymic selection. It is thought that similarities between self-peptides and pathogenic peptides that have elicited a successful immune response may be key drivers in many of these autoimmune events[160].

As novel therapeutics arise that facilitate the direct targeting of specific antigens, regardless of the disease context, it's important to consider the role of TCR degeneracy in the target selection process. Notable examples of off-target toxicity such as cross-reactivity with the cardiovascular protein TITIN while targeting melanoma epitope from *MAGE-A3* highlight both the safety risk surrounding improper target selection and the importance of understanding cross-reactivity during the therapeutic design process[46]. While progress has been made to improve TCR binding predictions and better understand the mechanics of TCR activation, our ability to computationally screen for cross-reactive candidates remains lacking[48]. Sequence similarity alone is not sufficient to predict cross-reactive peptides[50]. In addition, many newer approaches to understanding TCR recognition also rely heavily on TCR sequencing techniques which can't be applied in some therapeutic contexts such as the use of mRNA vaccines to induce a response from endogenous T-cells[51].

Herein we leverage paired sets of epitopes to help identify epitope-intrinsic features that drive T-cell cross-reactivity. Given that epitope sequences determine peptide-MHC binding and ultimately define the pMHC-TCR interface, we hypothesize that leveraging paired sequence identities in the training of a deep learning model will provide key features for the identification of cross-reactive epitopes without the need for complete TCR sequences or fragment sequences

therein. In addition, this approach will allow for the application of cross-reactive predictions to additional contexts where the collection of TCR sequences is not feasible such as the induction of endogenous T-cells through vaccination.

3.3 Methods

3.3.1 Identification of T-cells and associated epitopes from database data

Examples of cross-reactive T-cells were aggregated from three unique database sources: the immune epitope database (IEDB)[123], VDJdb[161], and the pan immune repertoire database (PIRD)[162]. For all sources the search space was restricted to human entries that included class I antigen presentation (Figure 3.1).

Data was downloaded from IEDB on 05/05/2021. All entries with a curated receptor ID available were queried for multiple epitope entries. This resulted in a total of 8,700 epitope entries and 3,902 unique TCR identities. Data was downloaded from VDJdb on 06/22/2021. Unique identifiers were constructed using reported CDR3, V, and J annotations for both alpha and beta chains where relevant. Unique identifiers with more than one reported epitope entry were retained as examples of cross-reactive T-cells, resulting in 3,153 total entries across 1,330 unique TCR identifiers. Data from PIRD was frozen on 11/30/2021. As with data from VDJdb, unique identifiers were constructed from reported elements of both alpha and beta chains, then used to identify TCR's that were reported to respond to more than one epitope. This resulted in 38 entries across 19 unique TCR sequences.

For all databases entries were checked for duplicates and missing data, then merged resulting in a total of 11,891 entries for further analysis (Appendix B table 7.1).

3.3.2 Construction of positive and negative epitope pairs across TCR examples

Modeling data was formatted as pairs of epitopes where the initial epitope is always an example that elicits a TCR response, and the second epitope is either cross-reactive for positive examples (also an epitope that elicited a response from the same TCR) or a non-reactive epitope for negative examples. For each unique TCR identity within the aggregated dataset, all combinatorial pairs of listed reactive epitopes were generated to constitute the positive example space. After pairs were constructed, each was duplicated in reverse order to create a "mirrored" positive dataset.

For both the mirrored and non-mirrored datasets, negative examples were selected by querying the human proteome as a reference background. For each positive epitope, a random human protein was chosen, and an accompanying fragments were selected based on the length of the positive reference, with a 20% chance of length mismatch (+/- 1aa). Epitopes were then compared with the associated positive for similarity and scored based on BLOSUM-62 similarity weights compared to an ideal match, retaining negatives within at least 90% of the exact match score with no length mismatch penalty. This process was repeated until negative epitopes matching the search criteria were identified and paired with the given positive epitope, resulting in an initial 2:1 ratio of negative to positive examples.

After all positive and negative epitope pairs were constructed, the dataset was queried for duplicate pairs. All duplicated positive pairs were dropped from the data to prevent overrepresentation of specific epitopes. Furthermore, duplicate negative pairs and generated negative pairs that were shared with the positive sample space were removed to prevent contradictory example pairs.

3.3.3 Padding and feature encoding of paired sequences

To accommodate variable length epitope targets, all epitopes within pairs were padded to a length of 11 amino acids using a generic null character (*). Paired epitope sequences were then concatenated and physical properties were encoded as numerical vectors based on side chain polarity[132], molecular volume[133], hydrophobicity[134], and conformational entropy[135] as described in our previous publication[71]. This resulted in an ultimate feature matrix of 22 by 4 for each paired epitope entry.

3.3.4 Model architecture and training schema

A multi-layer perceptron (MLP) was used for cross-reactivity predictions and implemented using the PyTorch package (v1.12.1) in Python (v3.9.12). In total, the MLP consists of an input layer (88 nodes), two internal layers (32 nodes and 16 nodes, respectively) and an output layer; all of which are fully connected. All layers except for the final layer underwent batch normalization, followed by 30% dropout, and use a ReLU activation function. The final layer uses log(SoftMax) to output a relative log probability of cross-reactivity for each given sample.

For model training, data was split into training and validation sets based on unique T-cell identifiers instead of unique entries. Examples for 80% of unique T-cell identifiers were used for training, while the remaining 20% were held out for validation data to avoid training spillover. Training data was pre-normalized using the scikit-learn (v1.1.2) normalizer function, with the fit normalizer saved for future application to validation data and cross-dataset applications. Post normalization, training data was next injected with gaussian noise to mitigate overtraining potential. Training occurred across 36 epochs with negative log likelihood used to calculate loss based on inversely weighted class imbalance and Adam optimization of parameters with a

learning rate of 0.001 and weight decay of 0.01. The most performant model across epochs was captured by retaining the model with the highest AUC for previous epochs and comparing to the most recent epoch performance.

3.3.5 Identification and processing of comprehensive cross-reactivity data

Manuscripts surveilling the cross-reactive landscape of TCR's were identified through literature search using the terms 'cross-reactivity', 't-cell degeneracy', and 'PS-SCL assay', 'peptide scan', and 'T-cell specificity', alone or in combinations thereof. Candidate studies were further filtered to identify those that used comprehensive techniques such as combinatorial peptide libraries (Figure 3.1). These manuscripts were not restricted by disease of study, species type, or the MHC restriction of presented epitopes. Ultimately data from 8 studies were identified as having usable data of TCR clones with reported cross-reactive epitopes. Positive and negative epitope pairs were constructed and featurization was performed using the same approach used for database data, without mirroring for positive pairs.

3.3.6 Model assessment on comprehensive data

The fully trained MLP model was independently applied to data from each of the 8 comprehensive studies previously extracted[163]–[170] . Predicted cross-reactive probabilities were then aggregated across metadata contexts including model species and assay type to determine differences in performance by metadata label and in order to determine sensitivity, specificity and AUC for each metadata label as well as for independent data sources.

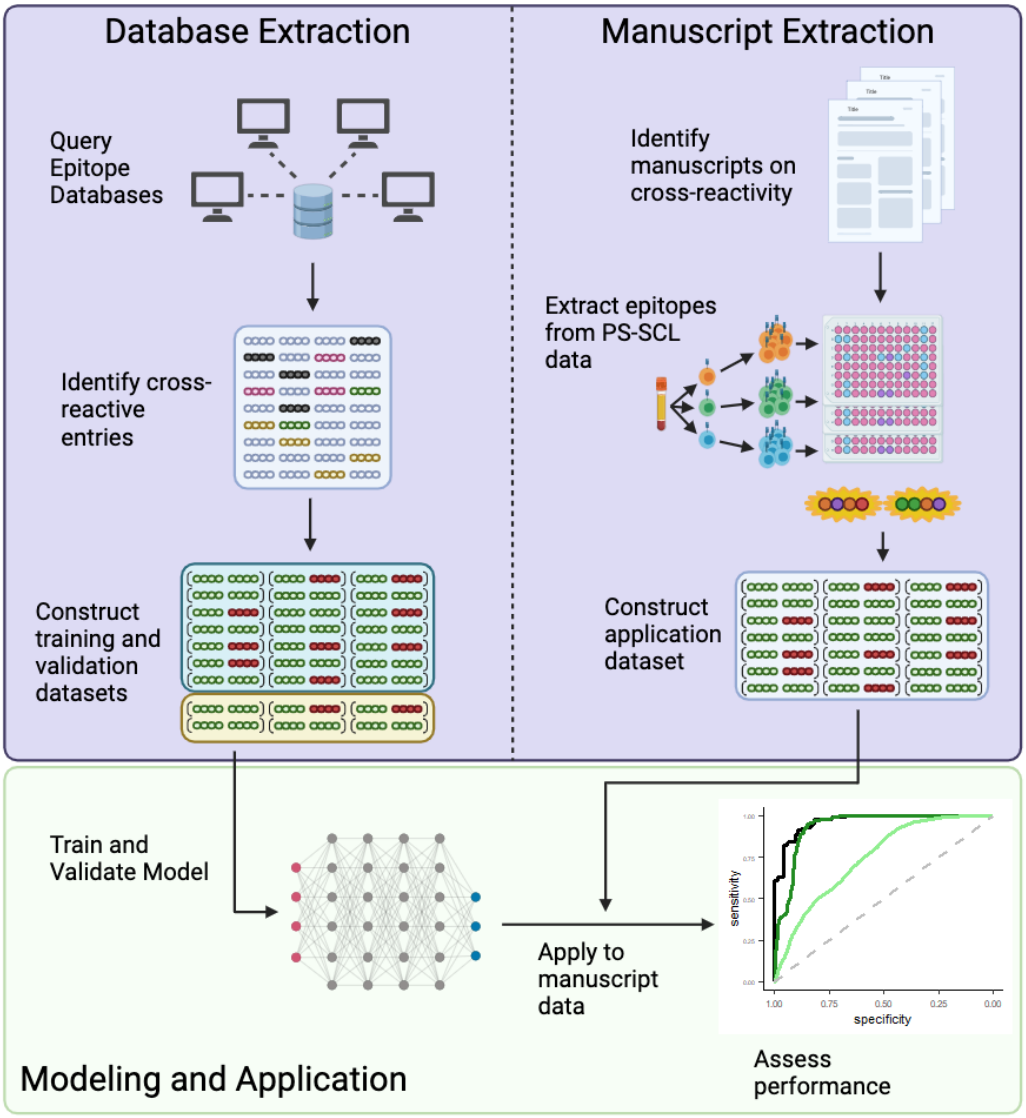


Figure 3.1 Extraction and processing of cross-reactive epitope data. *The database extraction workflow is shown at left, with data from multiple large-scale TCR-epitope databases aggregated and subsequently mined for cross-reactive epitope examples (two or more distinct cognate epitopes reported for the same TCR identity, indicated in this schema by shared color). Training and validation datasets were constructed by integrating cross-reactive examples (green/green epitope pairs) with non cross-reactive examples from the background human proteome (green/red epitope pairs; see methods) to create a positive and negative dataset. The manuscript extraction workflow is shown at right, with a set of studies using PS-SCL assays to assess cross-reactivity identified via literature search. Data from the PS-SCL assay results in each study were extracted to define lists of cross-reactive epitope pairs, then combined with constructed negative pairs (see methods). The modeling and application workflow is shown at bottom, with training and validation data serving as input for a multi-layer perceptron (MLP) model (note that depicted network connectivity is not reflective of actual MLP architecture; see methods). Once fully trained, the MLP was applied to held-out data extracted from PS-SCL assays.*

3.4 Results

3.4.1 `crossreactor` accurately predicts cross-reactivity from large-scale database data

Our first goal was to train and assess a cross-reactivity prediction model on epitopes derived from large scale databases. These entries represent pre-reported examples of cross-reactivity across a wide array of studies and sample contexts. Using a total of 5,251 aggregated TCR clones that showed signs of cross-reactivity, we constructed a total of 20,412 pairs of known cross-reactive epitopes and 40,824 pairs of inferred and cross-referenced non-reactive epitopes for model training, validation and testing. After featurization of the dataset, a multi-layer perceptron was trained to differentiate pairs of cross-reactive epitopes from pairs that contained one activating epitope but did not constitute cross reaction.

After model training and selection, we applied our trained model to a set of like-kind data that had not been seen during training or model validation. In total cross-reactivity predictions were made for 102 examples of cross-reactive pairs and 125 constructed pairs that don't constitute cross reaction. Overall AUC of predictions made for the 227 examples was 0.967, with an F1-score of 0.88 (Figure 3.2). Notably, the model showed high sensitivity, identifying 93% of cross-reactive pairs in the test data set.

3.4.2 `crossreactor` demonstrates strong performance on comprehensive human datasets

While database derived examples of cross-reactivity represent a wide array of epitope sources, interaction contexts, and assays used, these examples often don't encompass a comprehensive approach to epitope screening. To address this issue, we next applied our trained model to cross-reactivity examples derived from two manuscripts using comprehensive

screening approaches to assess human TCR clones via PS-SCL chromium-release assays (Figure 3.2). Application of our `crossreactor` to 1,096 extracted epitope pairs resulted in an AUC of 0.936, with comparable performance to our database test set performance (AUC = 0.967). These results demonstrate that `crossreactor` is capable of accurately detecting a wide variety of cross-reactive epitopes recognized by human T-cells.

3.4.3 Model generalizability is dependent on host system and assay type

Although performance on comprehensive human datasets remains strong, we wanted to assess the generalizability of our model outside of the original training context. To further investigate performance we applied `crossreactor` to six other manuscript derived datasets that annotate cross-reactive epitopes from a variety of host and assay contexts (TABLE 3.1). Strikingly, model performance on datasets that used IFN-gamma secretion as an assay readout were substantially lower than for all other data sources, highlighted by low sensitivity and AUC relative to all other extracted contexts (Figure 3.2B). In addition, model performance was lower on aggregated mouse data than on human derived datasets, however performance partially recovered when IFN-gamma based assays were excluded from mouse data aggregation (Appendix B figure 7.1). Given that model training data is restricted to human derived examples and largely based on chromium release assays these findings coincide with our expectations, while also highlighting that our model can be used to reasonably extrapolate cross-reactive predictions within other host contexts.

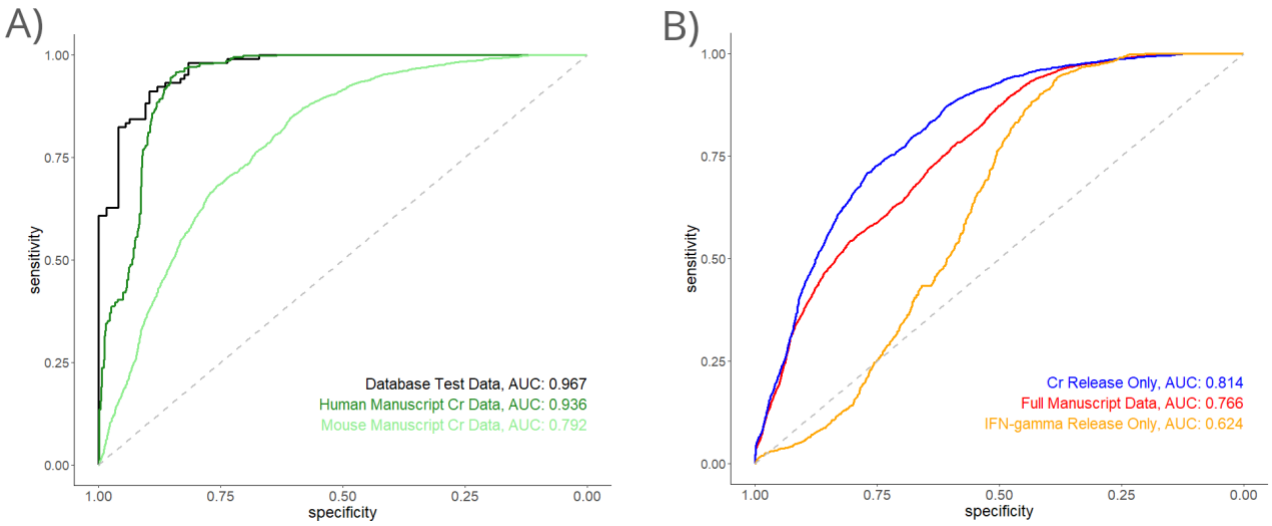


FIGURE 3.2. crossreactor performance on database-derived and comprehensive datasets. *A) Our model (crossreactor) was applied to a held out set of cross-reactive epitope examples from database data (n= 227), as well as PS-SCL data derived from 2 manuscripts using human T-cells from chromium (Cr) release assays (see table 3.1; n=1096), and 3 mouse/humanized mouse models using Cr-release assays(see table 3.1; n=8910). AUC's for each respective subset are reported. Model performance on mouse-only data deteriorated substantially when compared with human database and PS-SCL data. B) performance of crossreactor was assessed on PS-SCL data separated on Cr-release assay (see table 3.1; n=7810) vs. IFN-gamma release assay (see table 3.1; n=2779) readout, as well as aggregated across all comprehensive manuscript data. Performance on studies using IFN-gamma secretion as an assay readout was substantially lower than performance on chromium (Cr) release assays.*

Table 3.1 crossreactor performance on comprehensive manuscript data.

PubMed ID	Total Entries	Host Model	MHC Allele	Assay Type	Sensitivity	Specificity	AUC
16932807	84	Mouse	H-2Ld	IFN-gamma Secretion	0.0%	81.0%	0.813
16424210	2112	Humanized Mouse	A2/Kb	IFN-gamma Secretion	3.9%	96.0%	0.688
29567312	583	Human	A*02:01	IFN-gamma Secretion	16.5%	80.3%	0.588
8621898	1526	Mouse	H-2Kb	Cr Release	45.1%	92.0%	0.811
8699011	1404	Mouse	H-2Kb	Cr Release	63.4%	87.6%	0.845
18481808	3784	Humanized Mouse	A2/Kb	Cr Release	69.9%	66.5%	0.753
11431354	616	Human	A*02:01	Cr Release	96.0%	81.9%	0.959
11169449	480	Human	B*8	Cr Release	100.0%	68.8%	0.919

3.5 Discussion

Herein we have demonstrated an alternative approach to cross-reactivity prediction, leveraging the use of paired epitope data. By using a multi-layer perceptron model to predict cross-reactive epitope pairs, we have circumvented the need for in-depth TCR sequence data while retaining high predictive accuracy of cross-reactive events. This proof-of-concept highlights how epitope intrinsic features can capture fundamental aspects required for determining epitope recognition. Although our approach is limited in scope, approaches that are agnostic to receptor sequencing could play a key role in target selection for methods that use endogenous T-cell priming and don't easily facilitate characterization of all cells that will interface with a selected target.

Other architectures for modeling cross-reactive data were not exhaustively tested or compared and therefore a more optimized modeling approach may exist. Modeling approaches such as natural language processing (NLP) and recurrent architectures have shown promising results in other applications using amino acid sequences and may also perform well in the context of epitope cross-reactivity predictions[171], [172]. Additionally, approaches such as variable ablation or regularization were not attempted, either of which may result in a more simplified model if featurized inputs contain redundant properties.

Although model performance remains relatively high when applied to mouse derived datasets, high variance in performance depending on assay approach highlights the need for additional investigation. While studies using Cr-release as a readout for cytotoxicity performed well, our model performed poorly on data derived from IFN-gamma based readouts. This may have to do with differences in the functional thresholds represented by each assay type. While chromium release marks direct killing of a target cell, IFN-gamma is only correlatively associated with killing efficiency and has been shown to be an insufficient indicator of target clearance in

some *in vivo* settings[173]. The direct vs. indirect nature of these two readouts may be responsible for drastic differences in model performance and warrants further investigation to determine how our model can best leverage heterogeneous data sources.

In addition to challenges with application to disparate data types, our approach is also limited by the availability of data related to cross-reactions as a whole. By extracting examples of cross-reactivity from databases traditionally structured for other uses we were able to expand the pool of usable cross-reactive examples, however the total pool of reported cross reactive peptides still remains small. Furthermore, many cross-reactive peptides of interest are restricted to a few specific disease contexts like Type I diabetes[174], HIV[175], and CMV[44] among others. This focus on a limited pool of diseases, and by proxy limited pool of peptides, may also limit the broader application of all models designed for cross-reactivity prediction. This restricted focus is also exacerbated by database HLA bias, with disproportionate representation of popular HLA's like HLA-A:02:01; a problem that plagues many other immune related predictions such as MHC binding predictions and immunogenicity predictions as well.

Ultimately, *in vitro* approaches to cross-reactivity prediction also neglect to account for the role of co-stimulatory molecules and other immune cells that contribute to overall T-cell activation when antigens are encountered. Effective priming by antigen presenting cells and the presence of T-regulatory and helper T-cell populations can ultimately shift CD8 T-cell responses towards activation or tolerance by contributing to overall T-cell stimulation[176], [177]. The importance of co-stimulation and cytokine signaling in T-cell activation are therefore not appropriately captured when modeling cross-reactions via *in vitro* data.

While TCR sequence-based approaches to T-cell epitope identification and cross-reactivity prediction play a fundamental role in improving our understanding of T-cell dynamics, the applications of such approaches are fundamentally limited. The ability to apply our paired

epitope-based approach to contexts where TCR sequencing isn't feasible highlights a key use case not addressed by most tools that are currently available. Furthermore, strong model performance on human derived comprehensive data highlights the ability of `crossreactor` to make meaningful predictions even with minimal available input and adapt to a variety of applications.

Chapter III: Deciphering the Prostate Tumor Microenvironment: Transcriptional Insights into Therapy Response following Androgen Axis Blockade and Immune Checkpoint Inhibition

Benjamin R. Weeder[1], Reed Hawkins[2,3], Sushil Kumar[2,3], Ryan Kopp[4], Mark Garzotto[4], Reid F. Thompson[1,4,5,6], Amy Moran[2,3]

[1]Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR 97239, USA

[2]Department of Cell, Development & Cancer Biology, Oregon Health & Science University, Portland, Oregon, USA

[3]Knight Cancer Institute, Oregon Health & Science University, Portland, Oregon, USA

[4]VA Portland Healthcare System, Portland, OR 97239, USA

[5]Department of Radiation Medicine, Oregon Health & Science University, Portland, OR 97239, USA

[6]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR 97239, USA

This manuscript draft is part of an ongoing collaboration with the Moran Lab.

4.1 Abstract

Aggressive prostate cancers often exhibit progression and recurrence, including evolving resistance to therapy. Unfortunately, we lack a detailed understanding of the cellular composition and cell-cell interactions within high-risk prostate tumor microenvironments and, in particular, how populations change in response to therapy. In this study we perform single-cell transcriptomic profiling of paired temporal samples from subjects receiving androgen axis inhibition in combination with immune checkpoint inhibition. By leveraging single cell approaches, we show how treatment leads to a significant reduction in malignant cells and results in concurrent reduction in AR activity and upregulation of antigen presentation machinery within the malignant population, even in the absence of IFN-gamma response. We also highlight how mast cells and other components in the tumor microenvironment contribute to tissue dysregulation through angiogenic and immune-suppressive signaling.

4.2 Introduction

Many prostate cancer patients have lower risk disease that can be managed through active surveillance or local therapies, however patients with more aggressive disease have a higher risk of progression even after standard treatment options. Approaches such as androgen deprivation therapy (ADT) are often met with initial success but may result in eventual hormonal resistance and continued disease progression[178]. The treatment-refractory nature of prostate cancer in some cases highlights the need for novel solutions.

Immune checkpoint inhibition (ICI) therapy has been considered as one potential option for patients with aggressive disease, however clinical trials using ICI in prostate cancer have shown minimal success in the metastatic setting[179]. It is hypothesized that this may be due to

the generally immune “cold” environment of prostatic tissue. Studies have shown that prostate tumors have generally low T-cell infiltration and often harbor pro-inflammatory and pro-tumorigenic immune populations such as myeloid derived suppressor cells and M2-like macrophages, especially after long periods of ADT therapy[180], [181]. In addition to a suppressive tumor microenvironment, late-stage prostate cancers are also characterized by high levels of T-cell exhaustion, a mechanism by which cytotoxic cells lose their cytokine expression and upregulate key checkpoint molecules such as PD-1 and CTLA-4[182]. Although ICI therapy can help mitigate the impact of T-cell exhaustion, the lack of response in many patients who have received ICI suggests that exhaustion may not be the sole barrier to ICI response in prostate cancer patients.

This highlights the fundamental need for a better understanding of the prostate tumor landscape and how its intrinsic properties and cellular interactions may support resistance to current therapy options. Unfortunately, the dynamics of the prostate tumor microenvironment (TME) and its changes in response to treatment remain incompletely characterized, with previous single cell RNA-sequencing studies often limited by a lack of temporal sampling. Understanding how prostate cancer responds, adapts, and develops resistance to therapy remains a critical area of research with implications for therapy design and selection and ultimately patient outcomes.

Here, we analyze both tumor and non-tumor specimens from subjects with localized but aggressive prostate cancer, both prior to and after initial course of treatment. This gives a unique look into the naïve tumor landscape and how prostatic tumors fundamentally change after androgen axis blockade alongside the anti-PD-1 immune checkpoint inhibitor pembrolizumab. Transcriptional analysis, in conjunction with inference of structural variations has allowed us to confidently annotate malignant cells and better understand the direct effects

of treatment on the tumor epithelium. Unenriched transcriptional profiling in conjunction with receptor-ligand cross-talk analysis also provides fundamental insight into broader interactions within the TME and their potential implications in therapeutic response or resistance.

4.3 Results

4.3.1 Subject sampling and overview

Recent work has demonstrated that T cell intrinsic androgen receptor (AR) activity limits the effectiveness of PD-1 targeted immunotherapy in metastatic castration resistant prostate cancer patients[183].To gain insight into the effects of AR on the prostate cancer immune landscape, we performed single cell RNA sequencing (scRNA-seq) on cells isolated from both tumor and non-tumor specimens from subjects enrolled in a Phase II single-arm, open label, neoadjuvant hormonal plus immunotherapy clinical trial for high-risk localized prostate cancer (NCT03753243).

Longitudinal specimens were collected prior to the initiation of androgen axis inhibition (GNRH agonist therapy plus AR inhibition) and pembrolizumab, and at the time of prostatectomy after sixteen weeks of treatment (Fig. 4.1A). In total, 47 prostate specimens were collected across 18 individuals (Fig. 4.1B). After scRNA-seq and processing of these specimens, we successfully recovered single cell transcriptomes for 158,838 cells with 108,317 cells (69.5%) passing all filtering criteria.

Integration of subject samples and unsupervised clustering revealed 27 distinct groups of cells. These clusters were merged into 8 super groups shared broadly across samples and annotated using canonical markers and top differentially expressed genes, with resulting cell type identities including: fibroblasts, mast cells, B-cells & plasma cells, endothelial cells,

epithelial cells, myeloid cells, and T-cells & NK cells (Fig. 4.1C, 1D). Notably, epithelial cells also sub-clustered distinctly based on prostate specific antigen (PSA) expression status giving distinct PSA-high and PSA-low subsets of epithelial cells (Fig. 4.1C, 1D).

Across samples, epithelial, myeloid, and NK & T cells were consistently the most abundant cell types captured, however we also observed high variability in cell type abundance between subjects, and even for samples within the same subject (Fig. 4.1E; Appendix C Fig. 8.1).

4.3.2 Neoadjuvant androgen axis inhibition with aPD1 therapy changes both tumor and non-tumor cellular compositions

Previous work analyzing the effects of androgen deprivation therapy has highlighted the immunosuppressive characteristics of androgen and characterized increases in both T-cell abundance and transient proliferative ability under androgen suppression[184]. Other reports have also described increased proportions of tumor-associated macrophages and T-regulatory cells in biopsies from subjects on androgen deprivation therapy (ADT) compared to those on other treatment protocols[185].

To assess the effect of androgen axis inhibition alongside anti-PD1 therapy, we assessed relative abundance of each cell type between tumor and non-tumor labeled specimens, and across pre- and post-treatment timepoints. Surprisingly, the baseline proportion of each cell group was largely consistent between tumor and non-tumor tissues, with no significant differences in overall proportions between locations (Appendix C, Fig. 8.2). In contrast, the administration of neoadjuvant androgen axis inhibition with aPD-1 was associated with multiple changes in the cellular landscape of the prostate, including a significant drop in PSA-high epithelial cells, consistent with previous reports showing tissue atrophy in androgen sensitive tumor and non-tumor tissues (Fig. 4.1F)[186]. Notably, PSA-low epithelial cells did not change

significantly with treatment and remained consistent in both tumor and non-tumor locations (Fig. 4.1F). We also observed an increase in myeloid cells across all tissues, consistent with previous findings that also show an increase in macrophage populations such as tumor associated macrophages after ADT (Fig. 4.1F)[187]. In addition, we observed a tumor sample specific increase in NK & T-cells following treatment (Fig. 4.1F). Previous work has similarly demonstrated an influx of T-cells in tumor tissues as a response to ADT, driven primarily by increases in CD4+populations[188].

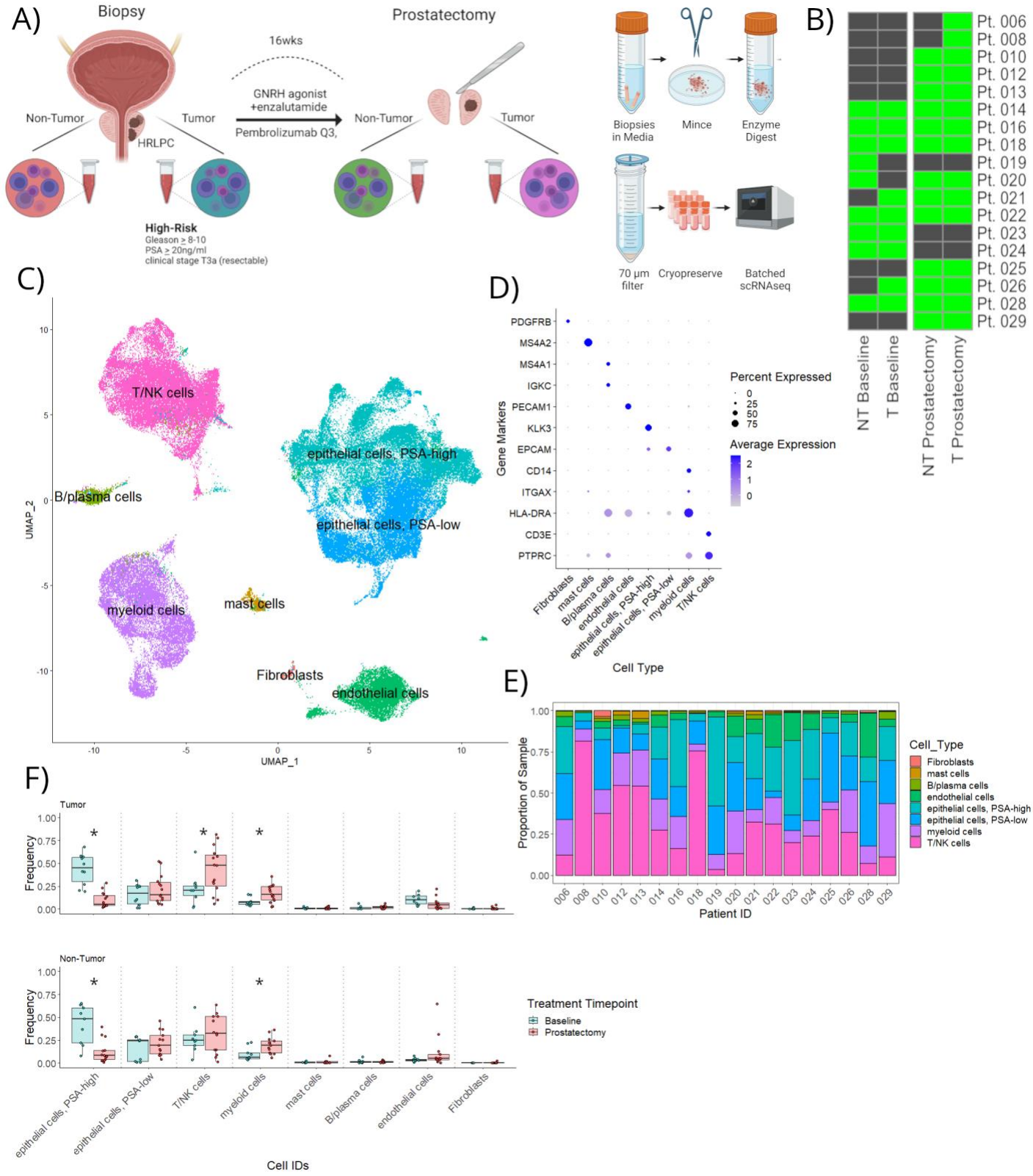


FIGURE 4.1. Sampling overview and heterogeneity of prostate samples. *A) Tumor and non-tumor biopsies were taken from treatment naïve tissues for subjects with stage IIIA high-risk localized prostate cancer (HRLPC), Gleason score ≥ 8*

and PSA ≥20 ng/mL. Subjects underwent 16 weeks of androgen axis inhibition (GNRH agonist + enzalutamide) with concurrent aPD-1 therapy (Pembrolizumab). After drug course, subjects underwent radical prostatectomy where both tumor and non-tumor tissues were again sampled. Single cell suspensions were generated from biopsy and radical prostatectomy samples and sc-RNAseq was performed using an Illumna NovaSeq 6000. B) Heatmap representing the presence or absence of a subjects' samples (y-axis) by timepoint and location (x-axis). Green squares represent subject samples that were successfully collected, processed and sequenced. C) UMAP projection of clustered and annotated cell types: B/plasma cells, T/NK cells, myeloid cells, mast cells, fibroblasts, endothelial cells, epithelial cells (PSA-high and PSA-low). D) Expression of 12 canonical markers (y-axis) was used to annotate each of eight cell types (x-axis). PSA-high and PSA-low cells were categorized based on shared expression of EPCAM with differential KLK3 (PSA) expression. E) Relative cellular composition (y-axis) for each subject (x-axis) is shown as a series of stacked bars, with each color representing an annotated cell type according to the key as shown. F) Baseline versus post-treatment proportions of each cell population within tumor (top) and non-tumor (bottom) specimens are shown as a series of boxplots, where each pair of boxplots corresponds to a specific cell type along the x-axis. Proportions are reported as the relative quantity of each cell type across all cells from a given specimen. Asterisks (*) indicate significant difference (Wilcoxon rank-sum test, p<0.05) in cell composition between baseline and post-treatment timepoints.

4.3.3 Malignant epithelial cells show responsiveness to androgen deprivation

Given that the goal of androgen deprivation therapy (ADT) is to starve androgen dependent tumor cells of a necessary hormone, we sought to confirm the 'on target' drug effect of androgen blockade. We focused on the epithelial cell population, interrogating nuances among five distinct sub-populations including: two different luminal cell clusters (high *KLK2*, *KLK3*, and *KLK4)*, basal cells (*KRT5*, *KRT14*, and *TP63 positive)*, and two other epithelial clusters (OE1 and OE2) with less defined cellular characteristics but clear pan-epithelial marker expression (*EPCAM*, *CDH1*, and *CEACAM1 positive)* (Fig. 4.2A, 4.2B). To clarify the cell identities of both the OE1 and OE2 clusters, all epithelial subsets were scored using gene signatures for club cells; a population previously identified in normal prostatic as well as lung tissues[189], [190]. The OE1 subset had a significantly higher club signature score than all other identified subsets, suggesting this cluster may consist of predominantly club-like epithelial cells (Fig. 4.2C). OE2 cells expressed genes associated with neuroendocrine cells but could not be confidently annotated.

To identify a malignant subset within the tumor-specific epithelial cell population, we interrogated expression of a malignancy signature constructed from genes previously defined in literature (see methods)[191]. One subset of luminal cells, denoted as luminal (1), expressed a high malignancy signature (Fig. 4.2D) compared to the second subset, subsequently referred to as luminal (2) cells. (Fig. 4.2D). Moreover, the prevalence of luminal (1) cells was heavily enriched in subject tumor samples compared to non-tumor samples and similarly enriched at baseline prior to treatment when compared to samples taken post-treatment at time of prostatectomy (Fig. 4.2E). In contrast, luminal (2) cells did not show significant changes in prevalence across tissues or sampling timepoints, suggesting that changes in luminal cell proportions are specific to cells expressing high malignancy signature. We also observed a significant increase in the proportion of club-like OE1 cells with treatment (Fig. 4.2E). We further confirmed the malignant nature of luminal (1) cells via copy number variation (CNV) analysis using inferCNV. Inferred aberrations for subjects with paired tumor and non-tumor biopsies, demonstrated that luminal (1) populations harbored unique subject-specific aberrations, while other epithelial populations did not.

While the specific decrease in the proportion of malignant epithelial cells after treatment suggests on-target therapeutic effects, we also investigated the effects of androgen axis inhibition directly on epithelial cell androgen pathway signaling. Average androgen response scores for each subject showed a notable reduction in AR signaling with treatment, primarily driven by luminal populations (Appendix C, Fig. 8.3) and reflecting on-target androgen axis inhibition (Fig. 4.2G). This result was consistent across tumor and non-tumor tissues with no significant difference in response signatures between sampling locations. Repeating this analysis specifically for malignant epithelial cells confirmed downregulation of androgen response signature (Fig. 4.2G). Taken in conjunction with proportional differences, this suggests that all

luminal populations are responding to androgen axis inhibition, however population loss is preferentially occurring in malignant luminal cells.

4.3.4 Antigen presentation machinery is upregulated with treatment and correlated with androgen response

Previous literature has also shown that inhibition of AR activity in prostate tumors is associated with increased antigen presentation and may aid with improved immune-driven tumor clearance[192]. To investigate whether we see similar responses in our data we scored all epithelial cells using an antigen processing and presentation score using genes defined in the Reactome database[193]. Pseudobulk comparisons of antigen presentation scores showed significant increases in antigen processing and presentation across subjects following treatment (Fig. 4.2G). This upregulation of antigen presentation signature with treatment was also seen in malignant cells, suggesting this a common treatment response across both malignant and non-malignant epithelial cells. Additional analysis also demonstrated a significant inverse correlation between androgen pathway score and antigen presentation score across subjects (r=-0.49, p=0.013).

Although AR activity has been shown to directly modulate antigen pathway expression *in vitro*[194], one alternative explanation for increased antigen presentation *in vivo* is through intrinsic IFN-gamma response[195]. We therefore assessed an IFN-gamma response score using a signature constructed from epithelial response to exogenous IFN-gamma application in epithelial cells from previous literature[196]. Although we see indications of IFN-gamma response in pseudobulk data, IFN-gamma response scores in malignant cells alone do not change significantly with treatment (Fig. 4.2G). Notably we also see higher levels of FOXA1 expression in malignant epithelial cells compared to combined non-malignant epithelial sub-

populations, a gene associated with inhibition of IFN response in previous literature (log2FC=0.37, p<1e-16)[197]. This suggests that while IFN-gamma may play a role in the overall upregulation of antigen presentation in bulk epithelial cells, it may not contribute to the same antigen upregulation in malignant subsets.

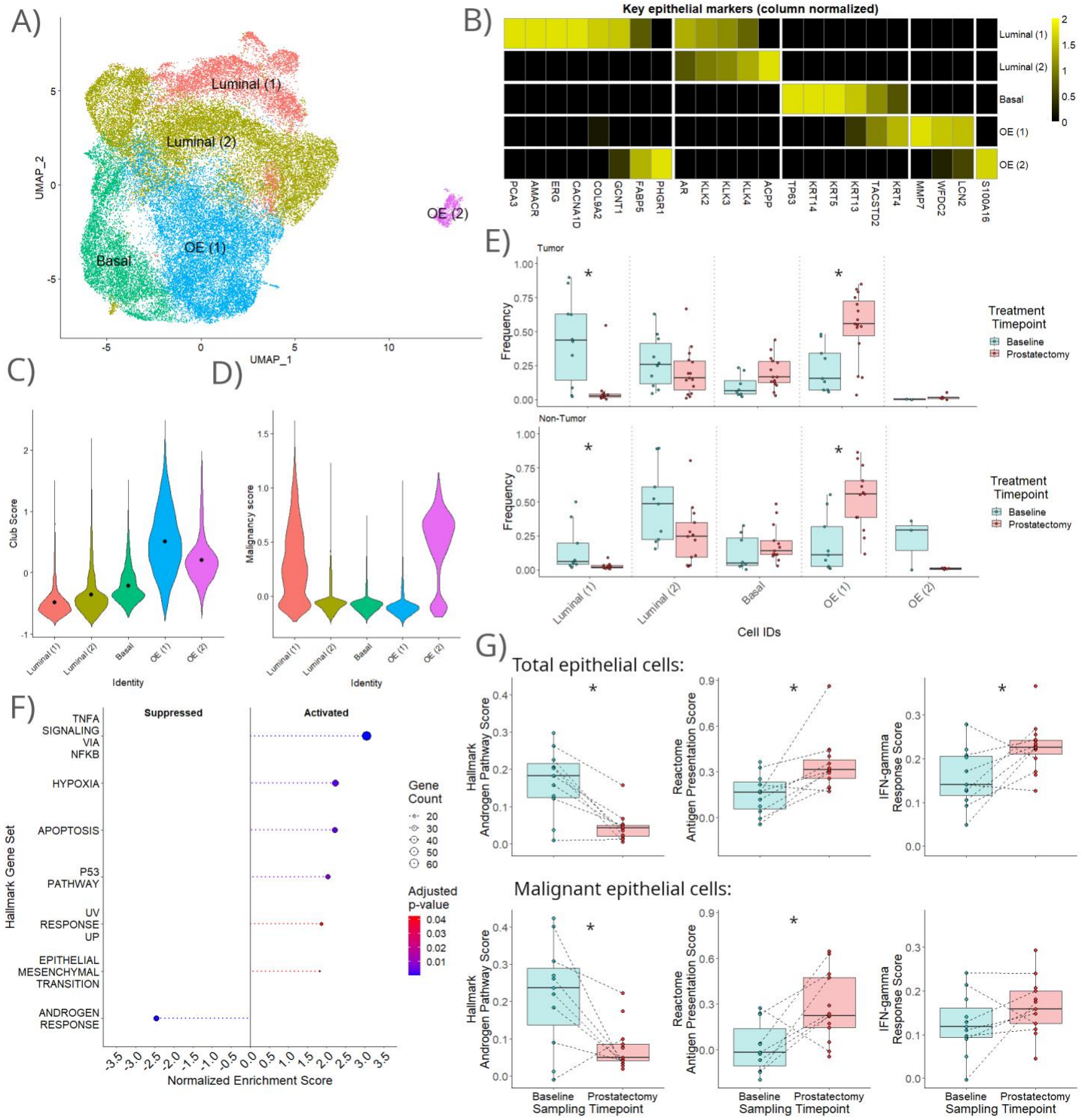


Figure 4.2. Epithelial cell identities and responses to treatment. *A) Identified epithelial cells were re-clustered into five subtypes (two luminal, one basal, and two other epithelial cells [OE]) and projected into UMAP space. B) Heatmap demonstrates relative expression of malignancy associated genes (PCA3, AMACR, ERG, CACNA1D, COL9A2, GCNT1, GABP5, PHGR1) and other canonical markers (x-axis) among each of the five epithelial sub-populations (y-axis). C) Violin plot demonstrating the distributions of relative club cell signature expression*[191] *(y-axis) among epithelial subpopulations as labeled (x-axis). D) Violin plot demonstrating the distributions of relative malignancy signature*

expression (y-axis; see methods) among epithelial subpopulations as labeled (x-axis). E) Baseline versus post-treatment (prostatectomy) proportions of each epithelial cell subpopulation within tumor (top) and non-tumor (bottom) specimens are shown as a series of boxplots, where each pair of boxplots corresponds to a specific cell type along the x-axis. Proportions are reported as the relative quantity of each cell type across all epithelial cells from a given specimen. Asterisks (*) indicate significant difference (Wilcoxon rank-sum test, p<0.05) in cell composition between baseline and post-treatment timepoints. F) Hallmark gene set enrichment analysis (GSEA), with normalized GSEA enrichment score (x-axis) depicted for each of seven hallmark gene sets (y-axis). Size of points represents the number of genes differentially expressed within an altered gene set. All presented gene sets are significantly altered (Wilcoxon rank-sum test, adj. p < 0.05). G) Androgen pathway, antigen presentation, and IFN-gamma response scores (see methods) are plotted as aggregate values, either across total epithelial cells or identified malignant cells per subject sample. Dotted lines connect subjects with paired baseline and prostatectomy samples. Asterisks (*) indicate significant changes in the proportion of a given cellular population between baseline and prostatectomy (Wilcoxon rank-sum test, p < 0.05).

4.3.5 Evidence of activated CD8 T-cell influx after neoadjuvant androgen axis inhibition with aPD1 therapy.

In previous studies investigating T-cell exhaustion, ICI therapy has been shown to increase the expression of proliferative markers such as Ki67 and is matched with restoration of cytotoxic signaling through re-invigoration of T-cells[198]. Since subjects received treatment that also includes aPD-1 therapy, we sought to investigate exhaustion and potential aPD-1 response among the T-cell populations present in tumor samples.

Upon re-clustering CD3+ NK and T cells into distinct subtypes (Fig. 4.3A, 4.3B), we identified multiple CD8 T-cell populations at varying stages in the activation spectrum according to relative expression of early effector genes (*IFN-gamma, IL2, TNF*) compared to activated cytokine expression (*GZM* family genes); these cell types included: naïve CD8 T-cells, recently activated CD8 effector cells, cytotoxic T-cells, and terminally differentiated T-cells (Fig. 4.3D). Other populations were also characterized and annotated in detail (see methods). Interestingly, we saw a significant increase in what we defined as recently activated CD8-effector cell populations across both tumor and non-tumor tissues with response to treatment (Fig. 4.3C). In tumor, CD8-effector changes were in conjunction with a proportional decrease in

CD4-naïve and gamma-delta T-cell populations. In non-tumor tissues, we saw a similar loss of naïve and gamma-delta populations, as well as an increase in terminally differentiated CD8 T-cells and Tregs (Fig. 4.3C).

With the abundance of literature highlighting the relevance of T-cell exhaustion in the context of aPD-1 treatment, we also wanted to assess any signs of functional exhaustion in our data. Interestingly, we found negligible expression of PD-1, LAG3, and TIGIT (Appendix C, Fig. 8.4), all markers generally associated T-cell exhaustion[62]. Only one cluster, annotated as CD8-cytotoxic-2, expressed any level of exhaustion associated markers, however these cells retained expression of *IFN-gamma, FASLG*, and multiple *GZM* family genes normally lost in functional exhaustion phenotypes (Fig 4.3D). While PD-1, LAG3, and TIGIT are associated with exhaustion in T-cells, they are also important markers of cognate antigen encounter. Unfortunately, we are unable to differentiate early exhaustion from antigen encounter in CD8-cytotoxic-2 cells given the available data.

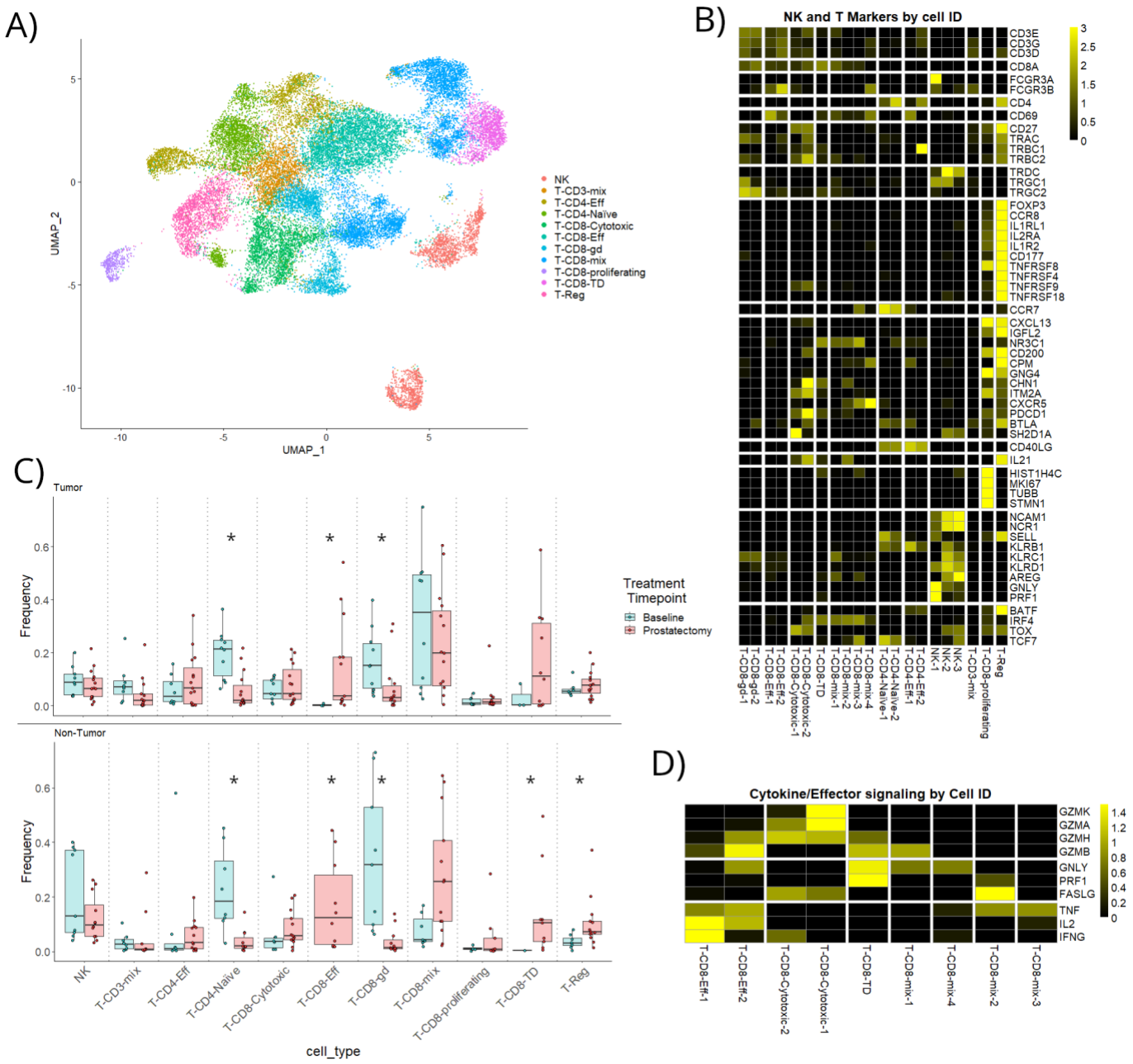


FIGURE 4.3. Neoadjuvant androgen axis inhibition with aPD1 therapy results in an influx of recently activated T-cells. *A) UMAP projection of 11 broadly annotated NK and T-cell clusters as labeled. B) Row-normalized heatmap showing expression of canonical and functional T-cell markers (y-axis) by annotated high-resolution cluster identities merged for broad cell annotations (x-axis). C) Boxplot of proportional changes in annotated NK & T cell sub-types across tumor (top) and non-tumor (bottom) with treatment. Baseline versus post-treatment (prostatectomy) proportions of NK&T cell subpopulations, grouped by broader cell type, within tumor (top) and non-tumor (bottom) specimens are shown as a series of boxplots, where each pair of boxplots corresponds to a specific cell type along the x-axis. Proportions are reported as the relative quantity of each cell type across all NK&T cells from a given specimen. Asterisks (*) indicate significant difference (Wilcoxon rank-sum test, p<0.05) in cell composition between baseline and post-treatment timepoints. D) Row-normalized heatmap showing expression of cytokine and effector genes (y-axis) in identified CD8 T-cell populations (x-axis).*

4.3.6 Myeloid sub-population proportions remain constant across treatment

Not considered canonical aPD1 targets, myeloid cells nonetheless have a complex and important role in both suppressing and assisting with tumor growth and metastasis. Previous studies have shown that M2-like macrophages and MDSC's can alter the balance of pro- vs. anti-tumorigenic signaling and assist in tumor resistance to therapy[185], [199]. We therefore investigated the distribution of distinct myeloid subpopulations across our samples, including immature myeloid suppressor-like cells (iMSCs), dendritic cells, innate lymphocytic cells (ILC's), patrolling monocytes, resident macrophages, and tumor associated macrophage (TAM)-like cells with high expression of APOE (*APOC1, APOE high*) and alternately with low expression of APOE (Fig. 4.4A, Fig. 4.4B).

While the global proportion of myeloid cells across subject samples increases with treatment, we do not observe significant proportional changes in the vast majority of myeloid subsets, with the exception of non-tumor APOE-high TAM-like cells (Fig. 4.4C). This suggests that treatment stimulated influx of myeloid populations occurs in a non-specific manner and not through recruitment of specific sub-populations. Notably, we see relatively stable proportions of iMSCs across treatment in both tumor and non-tumor tissues, a population characterized by high their expression of genes associated with MDSC's (Fig. 4.4D). These cells are often associated with resistance to therapy and tumor metastasis in late-stage disease, however we observe their presence even in treatment naïve biopsy samples, consistent with previous treatment naïve observations from single cell prostate cancer biopsies[191].

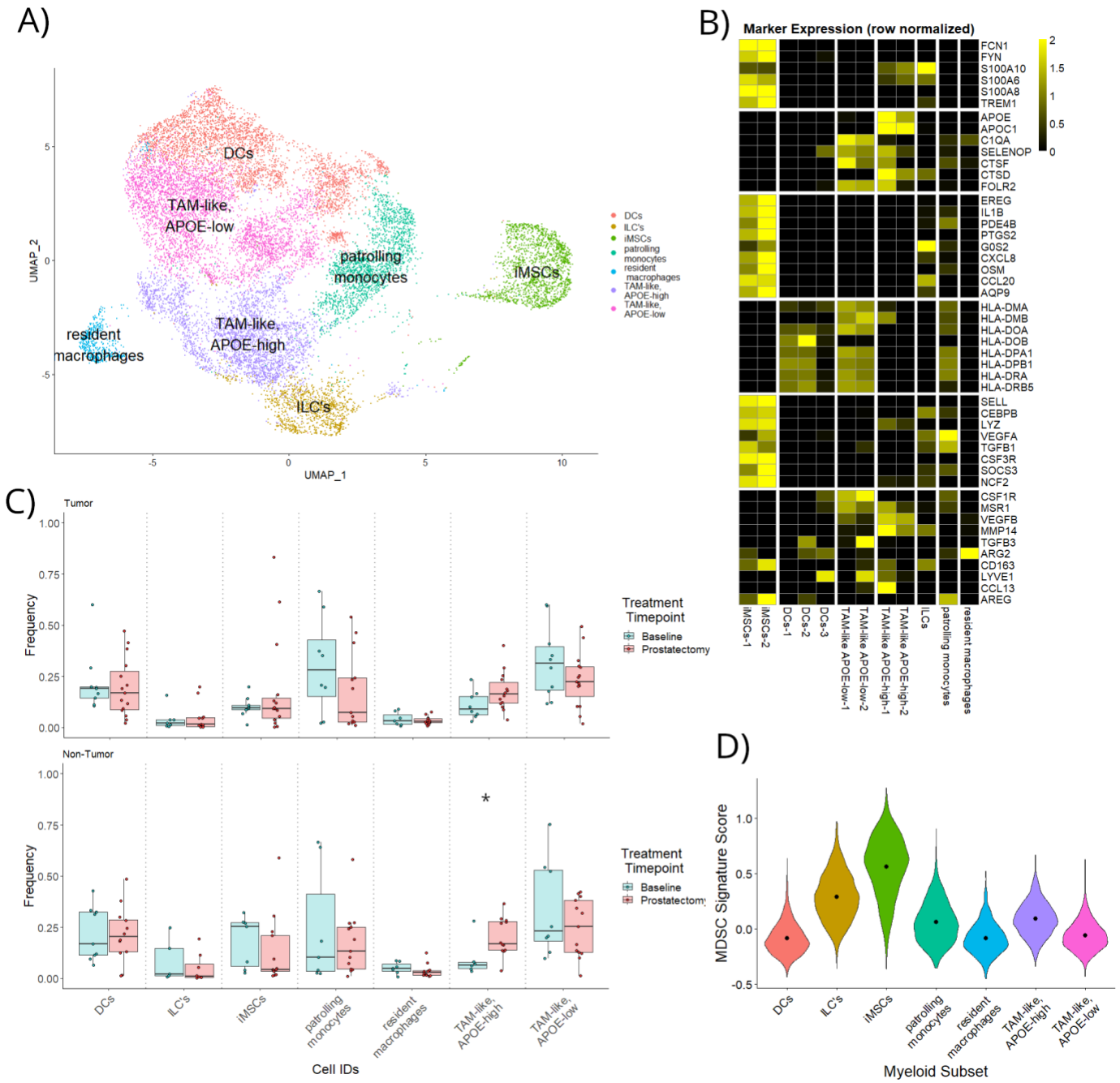


Figure 4. Inflammatory myeloid subsets are present in treatment naïve tissues and remain constant with treatment. *A) UMAP projection of re-clustered myeloid populations with broad annotations. B) Heatmap represents expression of common myeloid markers (y-axis) across 12 high-resolution myeloid cell clusters ultimately merged for broad annotations (x-axis). C) Baseline versus post-treatment (prostatectomy) proportions, after grouping by broader myeloid cell type, of 7 myeloid subpopulations within tumor (top) and non-tumor (bottom) specimens are shown as a series of boxplots, where each pair of boxplots corresponds to a specific cell type along the x-axis. Proportions are reported as the relative quantity of each cell type across all myeloid cells from a given specimen. D) Gene signature score for MDSC-like phenotype was constructed and applied to cells aggregated by sub-population annotation (see methods). iMSCs had significantly higher levels of MDSC-score than other myeloid populations (Wilcoxon rank-sum test, adj. p < 0.05).*

4.3.7 Angiogenesis, inflammation and wound healing are upregulated with treatment

Cellular processes such as angiogenesis and inflammation have been well described in the context of various cancers, including prostate cancer[60], [200]–[202]. In particular, inflammatory signaling has been emphasized as a key driver of immunosuppression and dysregulated angiogenesis has been linked with tumor growth and metastasis[203]. In prostate cancer specifically, androgen deprivation therapy has been associated with increased angiogenesis and studies have suggested a role for anti-angiogenic targeting in conjunction with aPD1 therapy as a potential treatment for further investigation[204], [205].

Leveraging gene set enrichment analysis (GSEA) of pre- versus post-treatment pseudobulk samples, we identified suppression of androgen response and activation of apoptotic signaling, in keeping with our prior observations among isolated epithelial cells. Additionally, we saw activation of signals associated with inflammation and tissue dysregulation (i.e. TNF-alpha signaling, hypoxia, and IFN-gamma response), consistent with previous reports demonstrating the prevalence of chronic inflammation and oxidative stress in human prostate cancer (Appendix C, Fig. 8.5)[206], [207].

To further investigate how specific cell types might be driving tissue dysregulation in the tumor microenvironment we scored each annotated cell type using an inflammatory signaling score, as well as an angiogenic signaling score through the application of gene sets previously defined in Hirz et al.[191]. Although myeloid cells contributed the most to overall inflammatory signaling, consistent with our identification of pro-inflammatory myeloid populations, we did not observe an increase in myeloid inflammatory signature after treatment. In contrast, PSA-high epithelial cells, endothelial cells, B/plasma cells, mast cells, and T&NK cells all showed

relatively low levels of baseline inflammation but signs of increased inflammatory signaling at time of prostatectomy, indicating broad upregulation of inflammatory signaling across the majority of cell types with treatment (Fig. 4.5A). Similarly, we observed broad upregulation of angiogenic signaling with treatment across a variety of cell types including both PSA-high and PSA-low epithelial cells, myeloid cells, T&NK cells, and mast cells (Fig. 4.5B). Importantly, these phenomena were highly consistent across all subjects (Fig. 4.5C, 4.5D). Notably, we did not observe any changes in endothelial proportions despite changes in angiogenic signaling.

Alongside upregulation of angiogenic and inflammatory signaling pathways, we also saw upregulation of multiple genes associated with growth factor signaling and wound healing responses (*VEGF, EGF, S100 family genes, MMP family genes*)[208]. To investigate the contributions among different cell populations, we next analyzed inferred receptor-ligand interactions via CellChat[209]. Most notably, cross-talk inference highlighted mast cells as a key contributor of VEGF signaling to endothelial cells and EGF signaling to epithelial cells at time of prostatectomy (Fig. 4.5E, 4.5F). In particular, EGF signaling was primarily driven by interactions between EGFR and AREG (Appendix C, Fig. 8.6). Notably, AREG expression has been previously associated with cancer migration and metastasis and is also upregulated in other cell populations including myeloid cells and B/plasma cells at time of prostatectomy[210]. Although we do not see signs of strong cross-talk between AREG expressing cells and T-cells, expression of AREG in literature has also been shown to upregulate Treg activity and further facilitates their suppressive function[211]. For mast cells specifically, their presence and expression of VEGF has also been associated with increased micro vessel density and resistance to anti-angiogenic therapies, as well as resistance to aPD-1 therapies[212], [213].

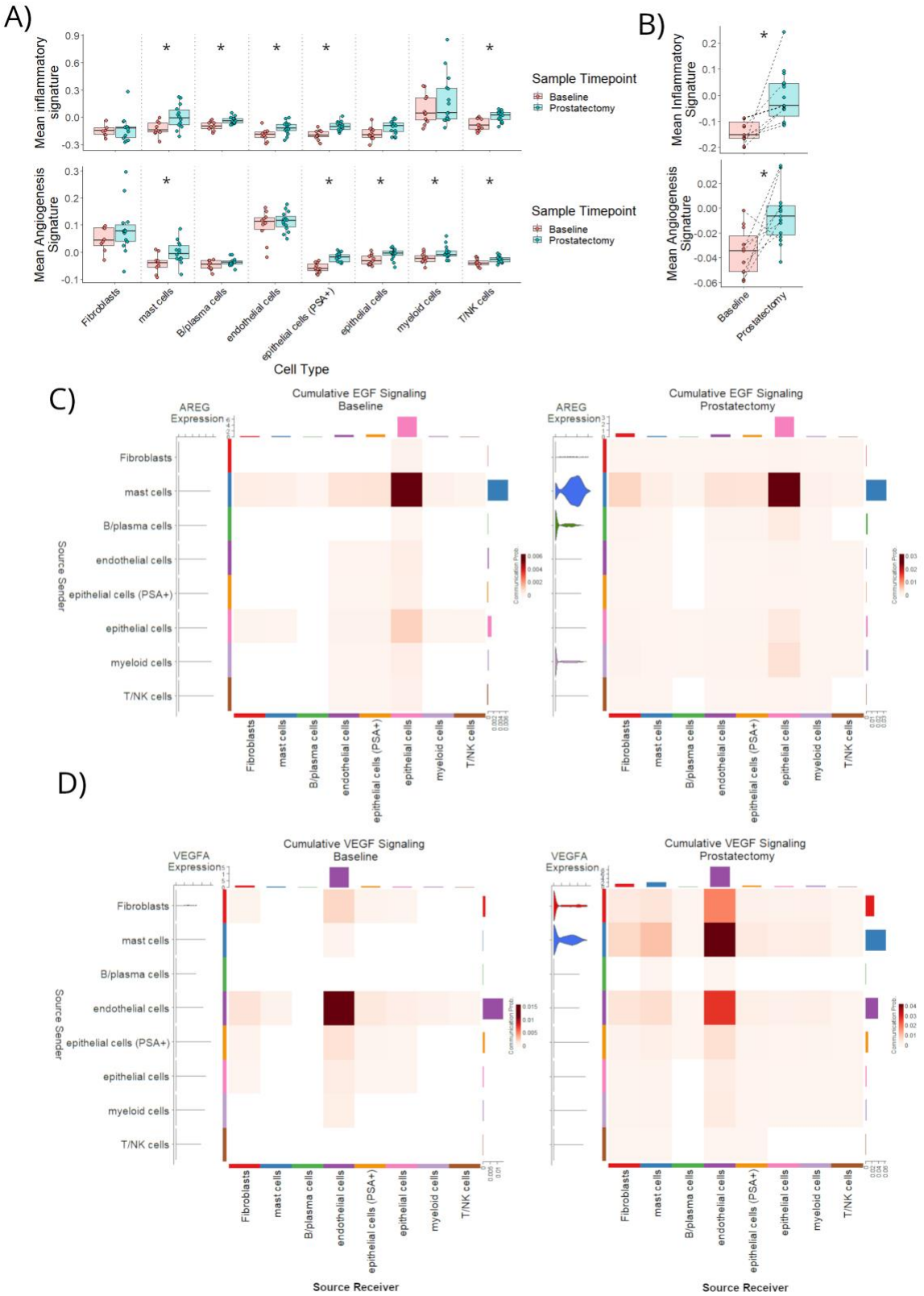


Figure 4.5. Receptor-ligand interactions highlight mast cells as key drivers of growth factor signaling. *A) Baseline versus prostatectomy inflammatory (top) and angiogenesis (bottom) signatures adapted from Hirz et al.*[191] *are shown as a series of boxplots, where each pair of boxplots corresponds to a specific cell type*

along the x-axis. B) Inflammatory (top) and angiogenesis (bottom) signatures (see methods) are plotted as aggregate values across all cells per subject sample. Dotted lines connect subjects with paired baseline and prostatectomy samples. Epidermal growth factor (EGF) signaling (C) and vascular endothelial growth factor (VEGF) signaling (D) were assessed through receptor-ligand interactions using CellChat[209]*. Cell types initiating pathway signaling are represented on the left-hand side of each plot (y-axis), while cell types receiving pathway signal are represented on the bottom of each plot (x-axis). Within the heatmap, each value represents the relative pathway signaling strength of each sender-receiver pairing. Bars on the right and top of each plot represent the cumulative signal sent and cumulative signal received for each cell type, respectively. For both EGF and VEGF, plots on the left side represent baseline receptor-ligand interactions while the right hand represents receptor-ligand interactions at prostatectomy.*

4.4 Discussion

To the best of our knowledge, this represents the first single-cell transcriptomic study to report on temporal sampling of paired naïve and post-treatment samples in prostate cancer. We characterize compositional and transcriptional changes throughout the prostate cancer cellular landscape, including malignant epithelial cells and key immune cell populations.

Across subjects, we were able to identify a sharp and consistent decrease in PSA-high epithelial cells consistent with on-target treatment response. While proportional changes suggest a response to therapy, re-clustering of epithelial cells allowed for the identification of malignant cells and direct analysis of treatment response in both malignant cells and psuedobulk epithelial cells. Using this approach, we were able to identify a significant drop in androgen pathway activity indicative of androgen axis inhibition and a correlated increase in expression of antigen presentation machinery. Although we cannot comment on a causal link between androgen pathway expression and antigen presentation, the observed correlation is consistent with previous *in vitro* findings that show androgen receptor activity can directly modulate the expression of genes related to antigen processing and presentation through upstream androgen response elements[194] . While there are other known mechanisms that upregulate antigen presentation, such as response to IFN-gamma, malignant epithelial cells in

our dataset do not show signs of increased IFN-gamma response with treatment and subsequently show increased expression of FOXA1, a gene linked with IFN-gamma resistance in previous literature[197].

Within the tumor, we also see an increase in NK & T-cells with treatment, primarily driven by CD8+ T-cells with an effector like phenotype. Previous studies on the effects of ADT in prostate cancer patients have demonstrated similar increases in T-cell proportions after treatment, however these changes driven primarily by changes CD4+ populations in a tissue agnostic manner, instead of CD8+ in tumor tissues specifically [188]. While NK & T cell proportions increased in tumor tissue, we saw a much broader influx of myeloid cells across both tumor and non-tumor tissues. Unlike with NK & T-cell populations this shift appeared to be driven by a broad influx across a variety of cell types already present in baseline samples. Additionally we observed the presence of both iMSC's and TAM-like populations at both baseline and prostatectomy time points, both of which have also been identified in previous single-cell prostate cancer analyses and associated with poor prognosis in bulk tumor analyses[185], [199].

Although our analysis of T-cell populations indicates signs of activation, the presence of multiple immune-suppressive myeloid populations highlights the importance of the tumor microenvironment and complex interplay between immune populations therein. This is further emphasized by post-treatment upregulation of both inflammatory and angiogenic signaling across a variety of different cell types. Interestingly, myeloid populations did not show a significant change in inflammatory signaling, but instead showed consistently high expression of inflammation related genes both pre- and post- treatment while other cell subsets contributed to an overall increase in inflammatory signaling. In particular, mast cell expression of both VEGF and AREG highlights the broad ability for immune cells in the TME to influence other cell

populations. VEGF is perhaps the best known regulator of endothelial growth and associated with both normal and tumor-associated angiogenesis[214]. Similarly, under normal conditions AREG is associated with wound healing and tissue normalization, however in tumor contexts increased AREG expression has also been associated with tumor migration and resistance to therapy[210], [215]. The association of AREG and VEGF expression with poor disease prognosis in other studies, and the treatment-related upregulation of cross-talk related to both receptor-ligand pathways in our data warrants further investigation.

Our study has several limitations. Given current clinical trial is under ongoing subject accrual and additional time needed to reach study endpoints, we are unable at this time to report on clinical outcomes or other subject-level clinical details. This unfortunately limits our ability to link observed proportional and transcriptional changes with clinical response. In addition, we are blinded to measurable tumor burden, potentially reducing our power to detect differences between tumor and non-tumor specimens, as some subjects were reported to have diffuse prostatic disease which may have impacted the ability to take purely non-tumor biopsies. In addition, while we attempted to use objective clustering approaches with minimal bias, single cell annotations are inherently subjective and clustering itself may misattribute cell identities for intermediate or outlier cell states. Our interpretations of results are based on a mix of both canonical and differential genes, but ultimately constrained by gaps in current biological knowledge. Our results are further constrained by the current limits of scRNA-seq, including limited per-cell sampling depth, transcriptional dropout, and variance in read quality, among other phenomena. In future studies we plan to link clinical outcomes with transcriptional biomarkers to identify potential prognostic markers of response or early indicators of resistance

4.5 Methods

4.5.1 Subject samples

Treatment-naïve, high-risk prostate cancer subjects with Gleason grade > 8-10 enrolled on clinical trial NCT03753243 underwent biopsy of a primary tumor lesion and non-tumor tissue prior to treatment. Following 14 to 16 weeks of treatment with neoadjuvant pembrolizumab (QW3) with anti-androgen therapy (enzalutamide plus GNRH inhibitor), subjects received a radical prostatectomy procedure. At the time of prostatectomy, biopsies from tumor lesions and paired non-tumor tissue were obtained.

4.5.2 Biopsy processing

Fresh biopsy specimens were collected immediately following the biopsy or prostatectomy procedure and processed same-day. Biopsies were mechanically dissociated using forceps and scissors into pieces that could be pipetted in phosphate buffered saline (PBS, Hyclone #SH30028FS) using a serological pipette. Biopsies were further dissociated by shaking at 300 rpm for 30 minutes at 37°C in PBS containing 30 U/mL DNAase I (Roche #04536282001), hyaluronidase (Sigma # H6254-500MG), and 1 mg/mL collagenase IV (Sigma #C5138-1G). Tissue digests were then filtered through 70 µm mesh filters (BD Biosciences ##352350) to obtain single cell suspensions. Samples were then cryopreserved in 90% FBS + 10% DMSO for later batch processing.

4.5.3 10x Genomics Library Preparation and Sequencing

Single cell capturing and library preparation were performed using the Chromium Next GEM Single Cell 3’ v3.1 kit (10X Genomics, PN-1000128) according to the manufacturer’s

instructions. Cryopreserved single cell suspensions of biopsies were thawed and filtered through a 30 µm filter prior to loading up to 30,000 cells per sample onto the Next GEM chip. Libraries were pooled and sequenced using an Illumina NovaSeq 6000 with 2 x 100 bp paired-end sequencing. Raw sequencing reads were aligned to the human reference genome GRCh38 and quantified using CellRanger (10x Genomics, v6.2.1).

4.5.4 Sample pre-processing and integration

Unless otherwise specified, all single cell analysis was performed using R v.4.2.2 and Seurat v.4.3.0[216]. Initial samples were filtered to remove ambient RNA contamination using SoupX[217] (https://github.com/constantAmateur/SoupX) with default recommended settings. After ambient de-contamination, individual samples were filtered to keep only cells with greater than 500 features and less than 25% of reads aligning to mitochondrial genes. Filtered samples next underwent doublet identification and prediction using DoubletFinder[218] (https://github.com/chris-mcginnis-ucsf/DoubletFinder) with default settings and an expected doublet formation rate of 7.5%, filtering out all droplets with a high doublet likelihood. Samples were then merged and normalized by batch based on the 2000 most variable genes using the Seurat ScaleData() function. Merged data was then integrated across batches by performing principal component analysis and using Harmony[219] v0.1 on the first 30 principal components (PC's).

4.5.5 Initial clustering and cell identification

Clustering was performed using the Louvain algorithm after calculating nearest neighbors using the first 10 Harmony components as input to the FindNeighbors function. A clustering resolution of 0.68 was selected after optimizing to reduce the average root mean square deviation (RMSD) of clusters, then fine tuning resolution for cluster stability using the Clustree package v0.5.0, ultimately identifying 27 clusters. Positive differentially expressed markers for each of the identified clusters were determined using Seurat's FindAllMarkers function and manual cell type annotation was performed based on the top markers for each cluster as well as canonical markers. Ultimately, clusters with shared canonical markers were merged to into 8 broad supergroups. Gene set enrichment analysis (GSEA) was performed using the GSEA() function from cluster profiler to compare psuedobulk baseline to prostatectomy samples. Only genes with a log2FC threshold of at least +/- 0.5 between baseline and prostatectomy were used and all hallmark gene sets were queried for enrichment.

4.5.6 Epithelial re-clustering and identification

Identified epithelial cells from the initial clustering step were re-normalized and re-clustered using the same approach as above, but instead using the 5000 most variable features during normalization and re-integration. A clustering resolution of 0.35 was ultimately selected, again based on RMSD minimization and clustering stability. Positive differentially expressed markers for each of the identified clusters were determined using Seurat's FindAllMarkers function and manual cell type annotation was performed based on the top markers for each of 15 graph based clusters, as well as using key genes from previous literature[191], [220]. Grouping based on shared gene signatures resulted in 5 broader epithelial groups including two subsets of luminal cells, one subset of basal cells, and two subsets of other epithelial cells that did not directly fit pre-defined cell identities. Luminal subsets were differentiated based on a

shared luminal gene signature including (*KLK2*, *KLK3*, and *KLK4, AR)*, in conjunction with expression of defined malignancy genes including (*PCA3, AMACR, ERG, CACNA1D, COL9A2, GCNT1, FABP5, and PHGR1)* that were consistently expressed in the luminal (1) subset but not luminal (2) cells. Scores used to assess club and malignancy phenotype in epithelial cells were defined using the genes defined in Hirz et al.[191] and scoring cells using the AddModuleScore function in seurat. Gene set enrichment analysis (GSEA) was performed using the GSEA() function from clusterProfiler to compare psuedobulk baseline to prostatectomy samples. Only genes with a logFC threshold of at least +/- 0.5 between baseline and prostatectomy were used and all hallmark gene sets were queried for enrichment.

4.5.7 Inference of chromosomal aberrations in epithelial subsets

For subjects with paired tumor and non-tumor samples, chromosomal aberrations were inferred using inferCNV v1.3.3 (Trinity CTAT Project, https://github.com/broadinstitute/inferCNV). For each subject, non-tumor epithelial cells, excluding luminal (1) cells, were used as background reference while all tumor epithelial cells were assessed for chromosomal aberrations. Copy number variants were inferred using the inferCNV "subcluster" mode with a cutoff of 0.1 as recommended for 10X derived data and use of HMM for inference smoothing. After inference and smoothing, CNV's were compared visually for regional amplifications/deletions.

4.5.8 Quantification of androgen response, antigen presentation, and IFN-gamma response in epithelial subsets

Epithelial cells were scored for antigen presentation and androgen pathway expression using the AddModuleScore function in Seurat, and the

“REACTOME_ANTIGEN_PRESENTATION_FOLDING_ASSEMBLY_AND_PEPTIDE_LOADING_OF_CLASS_I_MHC ” and “HALLMARK_ANDROGEN_RESPONSE ” gene sets from the Molecular Signatures Database (MSigDB) respectively. To aggregate scores on a per sample basis, mean expression was calculated across all scored cells, sub-setting by cell type and time point where relevant. Wilcoxon ranked sum tests were used to compare mean scores at baseline and prostatectomy in order to incorporate both paired and unpaired samples. IFN-gamma response signature in epithelial cells was again calculated using Seurat’s AddModuleScore(), using a list of 453 genes differentially upregulated (Log2FC > 0.5, FDR < 0.01) in HCC1143 cells treated with 10ng/mL of IFN-gamma for 72 hours compared to PBS vehicle control[196].

4.5.9 NK and T cell re-clustering and identification

Identified NK and T cells were further analyzed through re-normalization and re-clustering using the approach previously described for epithelial re-clustering with the 5000 most variable features and a cluster resolution of 0.99. Using the same graph-based clustering approach and RMSD minimization, we identified 3 NK and 18 T-cell clusters. Each cluster was identified using a mix of top differentially expressed genes and relative expression of canonical markers. Cells annotated based on activation spectrum including naïve, early/recently activated effector, cytotoxic, and terminally differentiated T-cells were all characterized based on their position on a spectrum of early effector signal expression (*IFN-gamma, IL2, and TNF*) compared to cytotoxic gene expression (*GZMA, GZMB, GZMH, and GZMK*).

4.5.10 Myeloid re-clustering and identification

As with epithelial and NK&T cell subsets, myeloid populations were re-normalized and re-clustered using the same approach RMSD based approach and a clustering resolution of 0.56. In total, we identified 14 myeloid clusters. Using canonical marker expression we identified two clusters of immature myeloid suppressor-like cells (iMSC's) (*S100A8, TREM1, CSF3R positive*), three clusters of dendritic cells, two tumor associated macrophage (TAM)-like clusters with high expression of APOE (*MSR1, APOC1, APOE positive*), and two TAM-like clusters characterized by low expression of APOE. In addition, we identified clusters including innate lymphocytic cells (ILC's), patrolling monocytes (high levels of *CXCL* markers), and resident macrophages. MDSC signature was defined using genes previously detailed in Hirz et al.[191] and applied using the AddModuleScore() function.

4.5.11 Cell-cell communication inference

Cell to Cell communication analysis between initially identified cell types was performed using CellChat with the default receptor-ligand database. The integrated single-cell dataset was split into baseline and prostatectomy subsets and receptor-ligand interactions were estimated using the identifyOverExpressedInteractions() function on each respectively. Cell to cell communication probabilities were assessed using the computeCommunProb() function. Baseline and prostatectomy communication inference objects were then merged, and interactions upregulated with treatment were identified using the rankNet() function with statistical estimation.

Conclusion

5.1 Summary

In this dissertation I've examined tumor-immune interactions across multiple scales, assessing factors involved in both efficacy and safety. I demonstrated how small steps in antigen processing are often ignored and can have a significant impact on the landscape of targetable epitopes; aggregating the largest available dataset of cleavage examples at time of publication and leveraging it to develop an open-source tool for improved proteasomal cleavage predictions. We further showed how this tool, `pepsickle`, outperformed currently available models and could be applied to previous study data investigating novel neoepitopes in order to enrich candidate pools for truly immunogenic targets. Since its release `pepsickle` has seen active use and continued interest, with multiple manuscript citations and consistent package downloads.

Shifting the focus from efficacy to safety, I next highlighted the importance of considering cross-reactivity during novel target selection for directed immunotherapies. Using `crossreator`, I showed how the use of a multi-layer perceptron model and paired epitope structures could be leveraged for accurate cross-reactivity predictions even in contexts that are traditionally challenging. While other approaches to cross-reactivity prediction, relying heavily on T-cell receptor sequencing, can give important insight into the dynamics of cross-reactive interactions, these tools are limited in their application to important treatment contexts. `Crossreactor` demonstrates how unique approaches to leveraging epitope data allow for the application of cross-reactivity models to emerging immunotherapy applications where TCR sequencing isn't feasible, such as mRNA vaccine design.

Lastly, I discuss the importance of considering the broader picture and how *in vivo* environment helps define tumor-immune interactions. While *in silico* tools such as `pepsickle` and `crossreactor` are important in the therapeutic research and development ecosystem, the success or failure of cancer treatment isn't determined by the output of a computer. Leveraging single-cell transcriptomic analysis of temporally sampled prostate cancer specimens, we characterized the treatment naïve landscape of prostate tumors and demonstrated how tumors change in response to a combination therapy with androgen axis inhibition and immune checkpoint inhibition. This work highlights dynamic changes in response to therapy such as broad increases in antigen presentation across patients concurrent with decreases in androgen pathway activity. While the structure of our data does not facilitate causal inference, previous *in vitro* work by our collaborators demonstrates how androgen response elements can directly modulate the expression of genes related to antigen processing and presentation. Furthermore, this work characterizes a post-treatment influx of recently activated CD8-effector cells and details transcriptomic signs of potential antigen encounter.

Together this body of work highlights multiple aspects of class I antigen interactions, from presentation to (mis)recognition, and ultimately modulation by therapeutic and environmental factors. In total, we show how each component of the complex antigen presentation process is important and impactful, but also how each in isolation neglects to tell the full biological story. While much more is required to provide the full picture of antigen processing, presentation, and immune activation, this body of work takes important steps towards filling in the pieces and adds to the vast body of work trying to comprehensively describe and model the complex process of antigen presentation and recognition.

5.2 Future Directions

The work detailed in this dissertation only scratches the surface. While the focus here has been on class I antigen presentation, class II antigens also play an important role in adaptive immune response. Although the underlying mechanisms are different, class II antigens must similarly be cleaved and processed prior to presentation. Data on class II antigen processing is more difficult to come by, however the modeling of class II cleavage motifs is a natural extension of our work on proteasomal cleavage predictions. The process of class II antigen preparation also involves a milieu of multiple different proteases, providing an opportunity for the application of unique modeling architectures and mixture models.

Similarly, our work on cross-reactivity is also focused on class I antigens leaving a natural extension to class II cross-reactive applications open for the future. Beyond this obvious extension, there are also multiple additions can be made within the class I antigen presentation space. As it stands `crossreator` is trained on human data only and while cross-applied performance mouse data is reasonable, we have yet to perform extensive comparisons or assess performance of a model trained on data from combined human and mouse data. Given the relative lack of observations in the cross-reactivity space, leveraging data across multiple model organisms may lead to more robust predictions if done in a thoughtful and thorough manner. Furthermore, the work presented as is does not take an in-depth look at what factors of epitope structure are most important for cross-reactive prediction accuracy. Looks into model learning and key epitope features may help in improving future models or defining key features of cross-reactive epitopes. In addition, examination of MHC allele representation in currently extracted data and assessment across a balanced MHC repertoire might help better define the generalizability of our work to broader contexts.

The analysis of prostate cancer specimens presented here will also benefit from a variety of follow up work. A lack of clinical correlates heavily restricts the conclusions made so far. While the data available is primed for investigation into markers of clinical response, prognostic indicators, and a search for markers of emerging tumor resistance, these approaches all require extended clinical data that has not yet been collected. As more patient data is collected and information on patient response is aggregated, the work presented already will serve as a steppingstone for further investigation. While many consistent responses to therapy were observed across patients, other genes and signatures demonstrated divergent responses that require further investigation. At this time, it's unclear if these divergent signals represent disparate treatment outcomes or simply transcriptomic noise, however marriage of the current computational analysis with future clinical responses will give the opportunity to address this question more clearly.

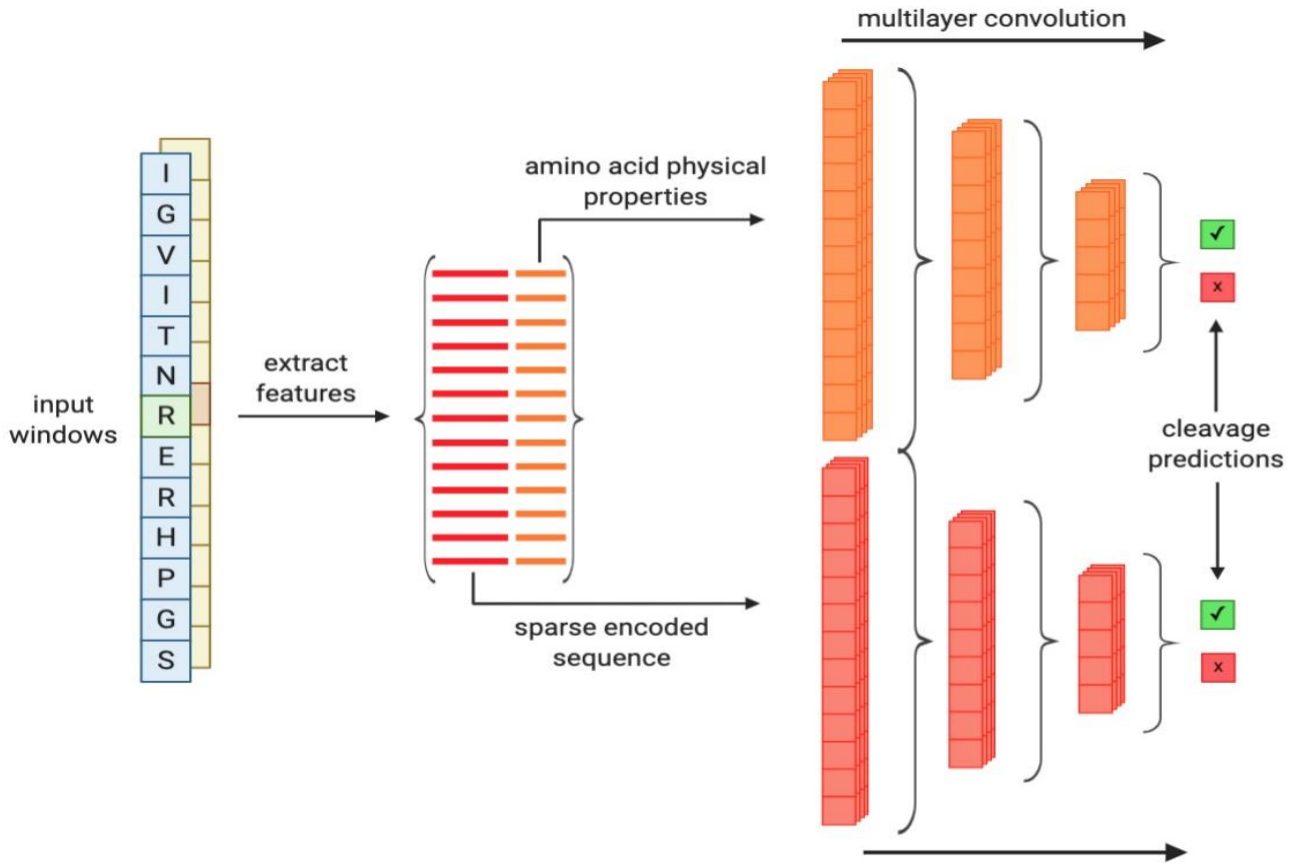
5.3 Concluding Remarks

The field of cancer immunotherapy is still rapidly changing. While the work presented in this dissertation addresses key challenges the field faces, the ever-evolving landscape of treatment targets and methodologies requires continued innovation and support. As we continue to search for more effective therapy options, we must keep in mind patient safety as well. Our immune system is ubiquitous, with unbridled access to almost every tissue in the body. It's not enough to consider how immune modulation will affect a tumor or organ of interest; we must consider how our approaches affect the system as a whole.

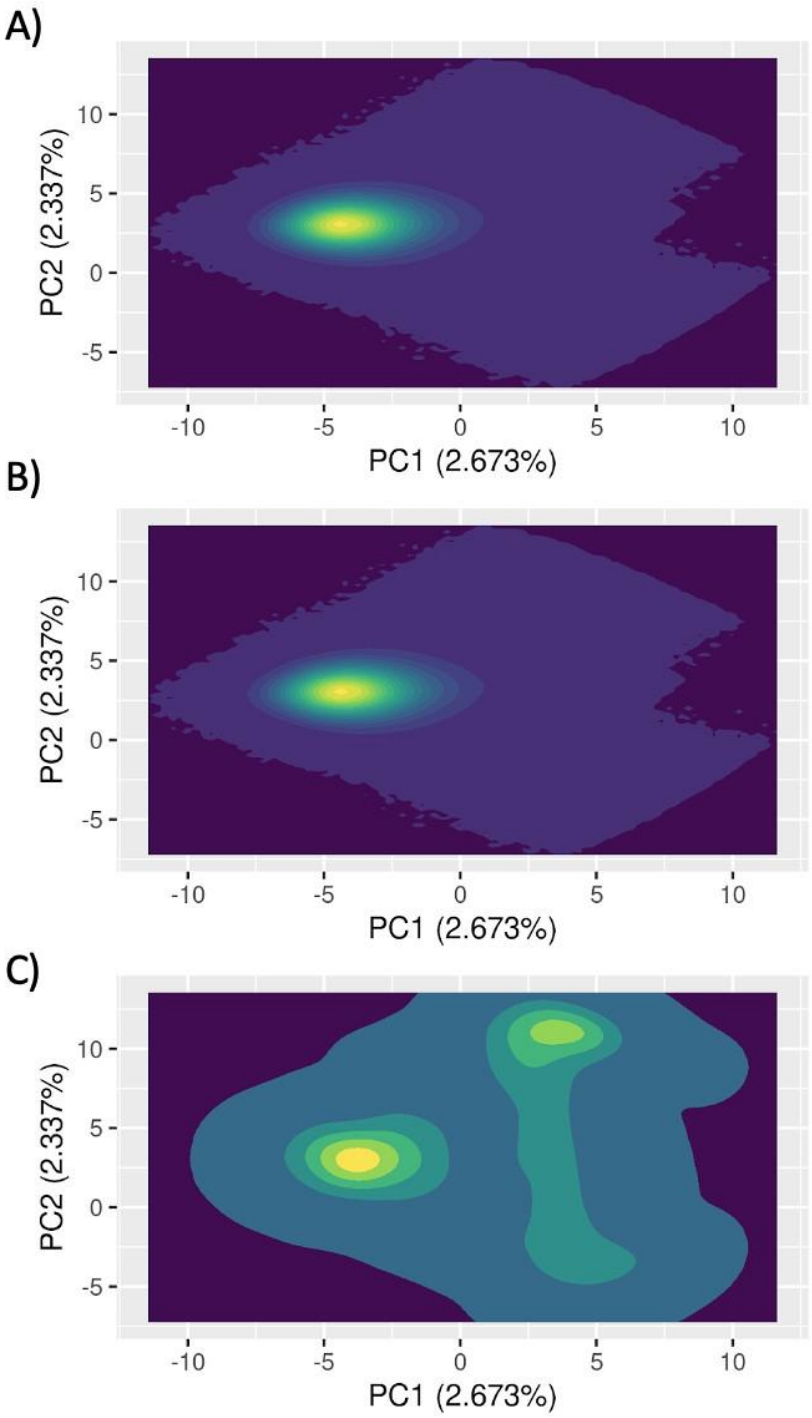
Computational methods have exploded in popularity the past decade and their adoption has accelerated even faster in the past few years. Despite the novelty and promise of deep-learning and high-throughput techniques, we must also remain grounded in relevant biology.

Computational approaches are most impactful when leveraging our foundational understandings of biological processes, and biological approaches can be expedited by intelligent use of *in silico* tools. At the end of the day science is a collaborative effort requiring thought, input, and insight from all.

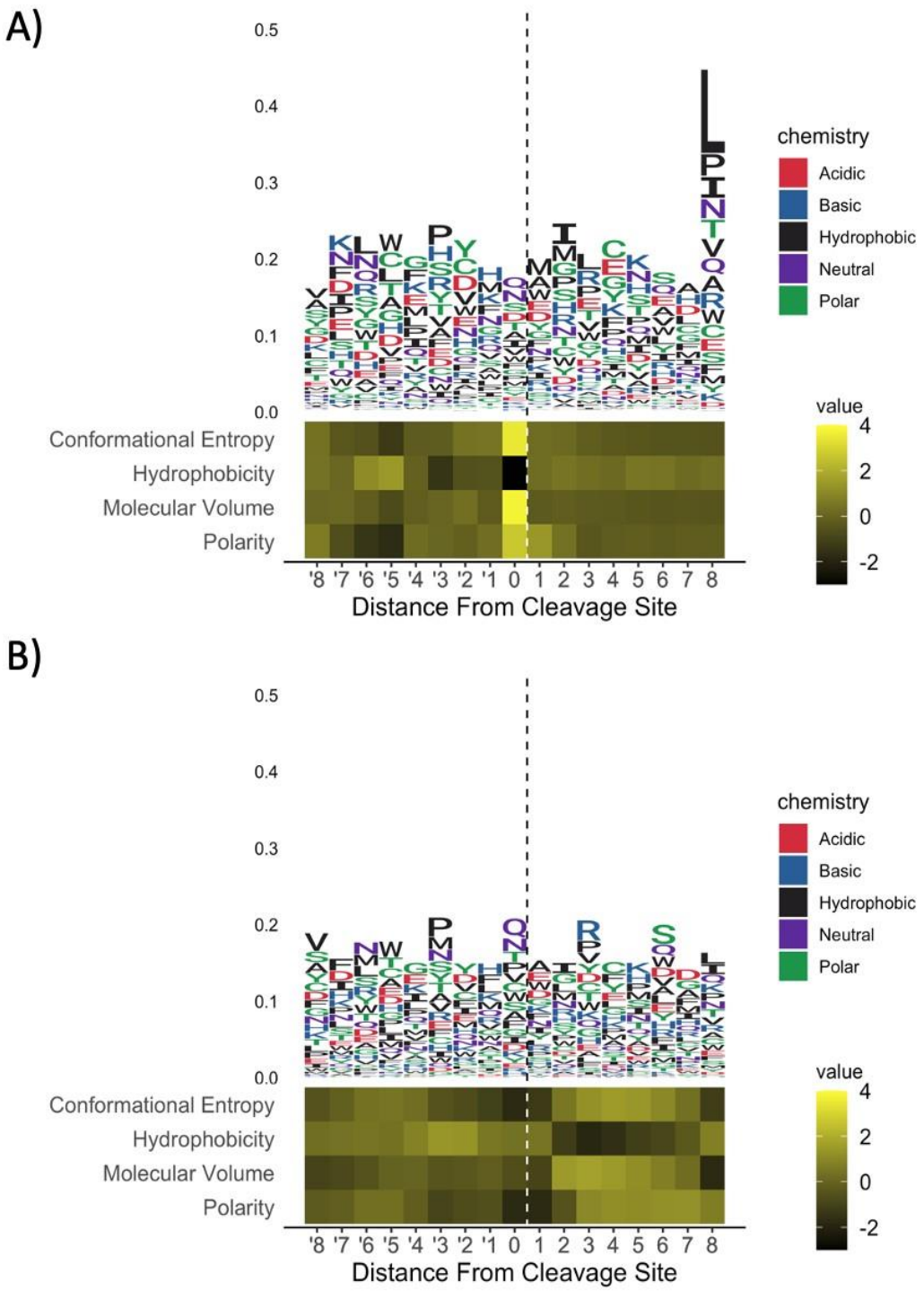
Appendix A: Supplemental Figures and Tables from Chapter I



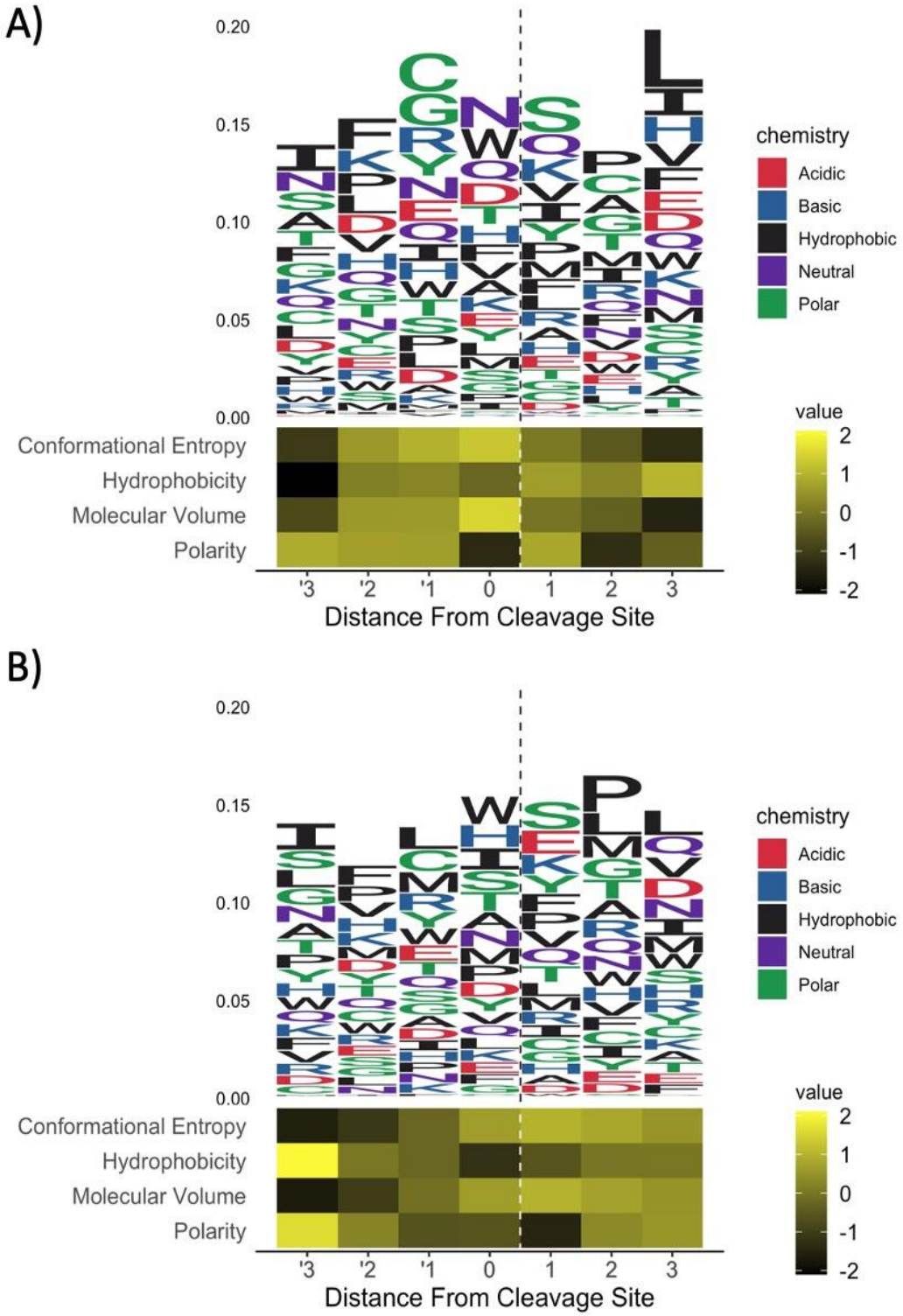
Appendix A, Figure 6.1. Epitope consensus model layout. *Features from amino acid windows in the epitope datasets were extracted to identify the one-hot encoded amino acid sequences, as well as the physical properties at each window position. One-hot encoded sequences were fed directly into the first layer of the deep learning model, while physical properties underwent a 1D convolution (span = 3) across each property prior to first layer input. For each internal layer, ReLU activation functions were used with 20% dropout. For final layers, log(SoftMax) was used to give class probability outputs. For exact layer numbers and sizes based on input window size see* ***Appendix A, Table 6.2 & 6.3.***



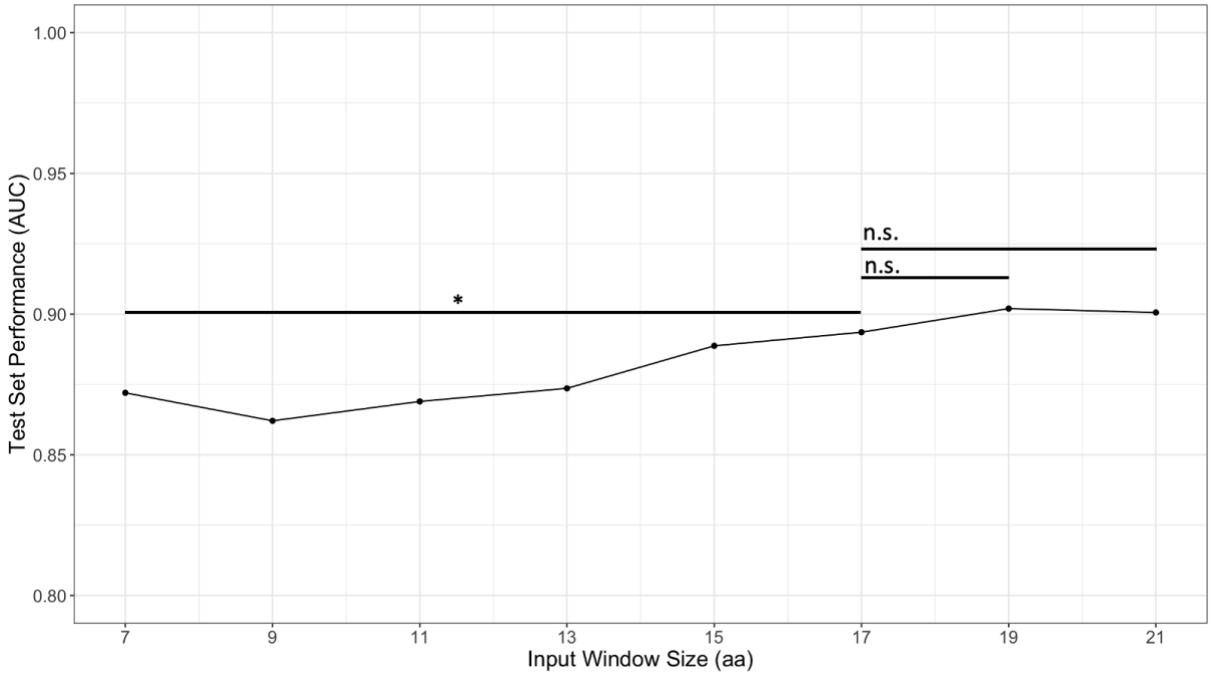
Appendix A, Figure 6.2. Training set sample densities compared to human background. A) *Principle components were constructed using the physical property values across 21 amino acid windows generated from all proteins in the human proteome. Using the first and second principle components (PC1 and PC2, respectively), sample density was calculated and plotted in PCA space.* ***B)*** *The density distribution for all 21 amino acid windows in the epitope based training set are shown using the same encoding and PCA approach.* ***C)*** *The density distribution for all in vitro based training examples are shown based on the same encoding approach.*



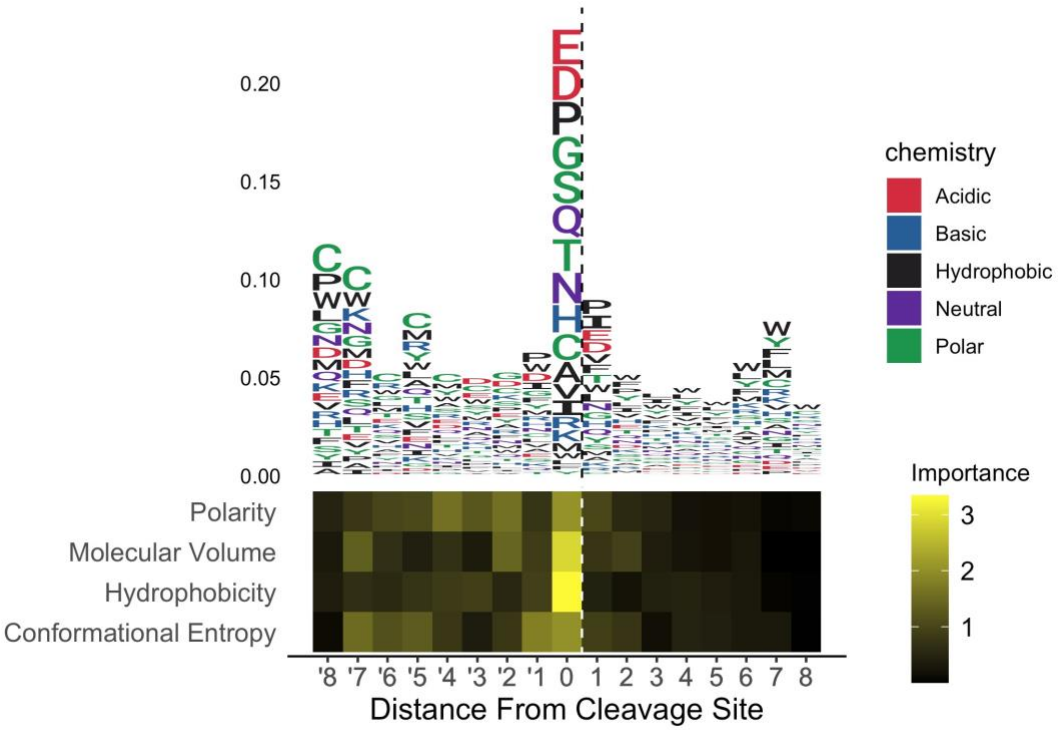
Appendix A, Figure 6.3. Epitope training set features. A) *Amino acid identities (top) and chemical properties (bottom) from positive cleavage windows were plotted as the average frequency (sequence) or average normalized value (chemical properties) across all amino acids at a given position.* ***B)*** *Non-cleavage windows were plotted using the same schema and ranges used for cleavage events.*



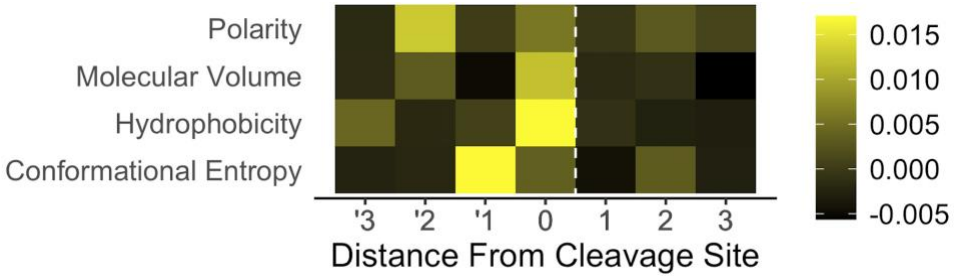
Appendix A, Figure 6.4. Digestion training set features. A) *Amino acid identities (top) and chemical properties (bottom) from positive cleavage windows were plotted as the average frequency (sequence) or average normalized value (chemical properties) across all amino acids at a given position.* ***B)*** *Non-cleavage windows were plotted using the same schema and ranges used for cleavage events.*



Appendix A, Figure 6.5. Effect of window size on in vivo deep-learning model performance. *AUC values for deep learning models (y-axis) trained on window sizes ranging from 7 amino acids to 21 amino acids (x-axis). (*) indicates a significant difference in AUC between models, while n.s. indicates no significant difference. For statistical comparisons of models across window sizes, see table S5.*



Appendix A, Figure 6.6. Epitope consensus model feature importances. *Feature importances were calculated as the absolute values of the model saliencies for the sequence identities (top) and chemical properties (bottom) at each given position in the input window of our 17 amino acid consensus model. For sequences, the total height of each bar corresponds to overall importance of a given position in the model, while the height of each letter corresponds to importance of the corresponding amino acid at that position. Chemical property feature importance is indicated by color gradient from most important (yellow) to least important (black).*



Appendix A, Figure 6.7. Chemical property feature importances for in vitro digestion model. *Feature importances were calculated as the normalized absolute values of the model weights for chemical properties at each given position in the input window of our 7 amino acid digestion based in vitro model. Feature importance is indicated by color gradient from most important (yellow) to least important (black).*

Appendix A, Table 6.1. Amino acid feature matrix.

Amino Acids	One -Hot encoded ID's																					Physical Properties			
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Polarity (pI)	Molecular Volume	Hydrophobicity (cos -theta)	Confromational Entropy
A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	56.15265	-0.495	-2.4
C	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.07	69.61701	0.081	-4.7
D	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.77	70.04515	9.573	-4.5
E	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.22	86.35615	3.173	-5.2
F	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.48	119.722	-0.37	-4.9
G	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.97	37.80307	0.386	-1.9
H	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	7.59	97.94236	2.029	-4.4
I	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	6.02	103.6644	-0.528	-6.6
K	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	9.74	102.7783	2.101	-7.5
L	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	5.98	102.7545	-0.342	-6.3
M	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	5.74	103.928	-0.324	-6.1
N	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	5.41	76.56687	2.354	-4.7
P	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	6.3	71.24858	-0.322	-0.8
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	5.65	88.62562	2.176	-5.5
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	10.76	110.5867	4.383	-6.9
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	5.68	55.89516	0.936	-4.6
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	5.6	72.0909	0.853	-5.1
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	5.96	86.28358	-0.308	-4.6
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	5.89	137.5186	-0.27	-4.8
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	5.66	121.5862	1.677	-5.4
*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7.5	0	1.689157	0
B	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	4.09	73.30601	5.964	-4.6
Z	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	4.44	87.49089	2.675	-5.35
J	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	6	103.2094	-0.426	-6.45
U	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.07	69.61701	0.081	-4.7
X	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	6.008095	88.55829	0.6195	-4.845

Encoded feature matrix including bit vector notation and physical/chemical properties for each standard amino acid and recognized ambiguous amino acids.

Appendix A, Table 6.2 Epitope sequence-based deep learning layer sized by input window size.

	Layers				
Window size	feature size	input layer	internal #2	internal #3	Output
7	140	136	68	34	2
9	180	136	68	34	2
11	220	136	68	34	2
13	260	136	68	34	2
15	300	136	68	34	2
17	340	136	68	34	2
19	380	136	68	34	2
21	420	136	68	34	2

Model layer sizes for sequence based deep learning models based on training window size. For all internal layers a ReLU activation function was used with 20% dropout. For the final output layer, log(Softmax) was used as the activation function.

Appendix A, Table 6.3. ***Epitope chemical-based deep learning layer sizes by input window size.***

	Layers				
Window size	input	1D Convolutional layer	internal #1	internal #2	Output
7	28	20	38	20	2
9	36	28	38	20	2
11	44	36	38	20	2
13	52	44	38	20	2
15	60	52	38	20	2
17	68	60	38	20	2
19	76	68	38	20	2
21	84	76	38	20	2

Model layer sizes for physical property based deep learning models based on training window size. For the initial convolutional layer, a 1D convolution with span=3 and step=1 was used. For all internal layers a ReLU activation function was used with 20% dropout and for the final output layer, log(Softmax) was used as the activation function.

Appendix A, Table 6.4 Command line computational performance.

	Model	
	pepsickle	NetChop 3.1
epitope	158m 21s	542m 40s
in-vitro	154m 46s	260m 50s

Time performances are based on processing time for the whole human proteome (see methods).

Appendix A, Table 6.5. In vitro model performance on epitope validation data.

Proteasome mode	Sensitivity	Specificity	AUC
Constitutive	69.85%	51.49%	0.650
Immuno	54.54%	73.71%	0.679

Comparison of *in vitro* constitutive and immuno- model performance on epitope data.

Appendix A, Table 6.6. In vitro model performances by window size.

Model 1				Model 2				
feature input	Size	AUC		feature input	Size	AUC	delta-AUC	adj-pvalue
chemical	**7**	**0.759**	**vs.**	**sequence**	**7**	**0.723**	**0.036**	**0.002**
chemical	**21**	**0.771**	**vs.**	**sequence**	**21**	**0.743**	**0.028**	**0.012**
chemical	7	0.759	vs.	chemical	21	0.771	-0.012	0.558
sequence	7	0.723	vs.	sequence	21	0.743	-0.020	0.513

Comparison of model performances on in vitro test data based on feature window size used for trianing input.

Appendix A, Table 6.7. Immunoproteasome validation set power analysis.

	Estimated Delta-AUC	# cases	# controls	N	Beta	Alpha
Current Beta, controlled alpha	0.130	36	18	54	0.665	0.05
Target Beta, controlled alpha	0.130	55	27	82	0.800	0.05

Statistical power analysis for `pepsickle` and NetChop 3.1 *in vitro* model comparison on immunoproteasome validation data. Initial estimate represents the actual type II error based on the available *in vitro* immunoproteasome data with type I error controlled at 0.05 and the observed difference in AUC. The second estimate represents the requisite number of cases and controls required to achieve a target type II error (1 - β) of 0.20, using the same case/control ratio and AUC difference observed in the available validation set.

Appendix A, Table 6.8. Epitope model test-set comparisons by window size.

Model Window Size		Model Comparison (large - small)	
Small	Large	AUC difference	Adj. P-value
7	9	-0.010	0.352
7	11	-0.003	0.778
7	13	0.002	0.851
7	15	0.017	0.080
7	**17**	**0.022**	**0.019**
7	**19**	**0.030**	**0.001**
7	**21**	**0.029**	**0.001**
9	11	0.007	0.527
9	13	0.012	0.269
9	**15**	**0.027**	**0.004**
9	**17**	**0.031**	**0.001**
9	**19**	**0.040**	**<0.001**
9	**21**	**0.038**	**<0.001**
11	13	0.005	0.655
11	**15**	**0.020**	**0.034**
11	**17**	**0.025**	**0.007**
11	**19**	**0.033**	**<0.001**
11	**21**	**0.032**	**0.001**
13	15	0.015	0.111
13	**17**	**0.020**	**0.028**
13	**19**	**0.028**	**0.001**
13	**21**	**0.027**	**0.002**
15	17	0.005	0.624
15	19	0.013	0.136
15	21	0.012	0.189
17	19	0.008	0.352
17	21	0.007	0.450
19	21	-0.001	0.851

Epitope models were trained on odd size starting windows between 7 amino acids and 21 amino acids. Each model was applied to the held out test set and assessed based on AUC. Delong's tests were used for pairwise statistical comparisons of the performance for each size of base window in contrast with other window sizes. P-values for comparisons were adjusted using Benjamini-Hochberg p-value correction and model comparisons with significant differences in AUC after correction are denoted in bold.

Appendix A, Table 6.9. Epitope validation performance metrics.

	Precision	Recall	F1
Pepsickle	0.766	**0.828**	**0.796**
NetChop	0.671	0.747	0.707
PCPS	0.656	0.619	0.637
PCleavage	**0.834**	0.182	0.298

Performance statistics for epitope based models available.

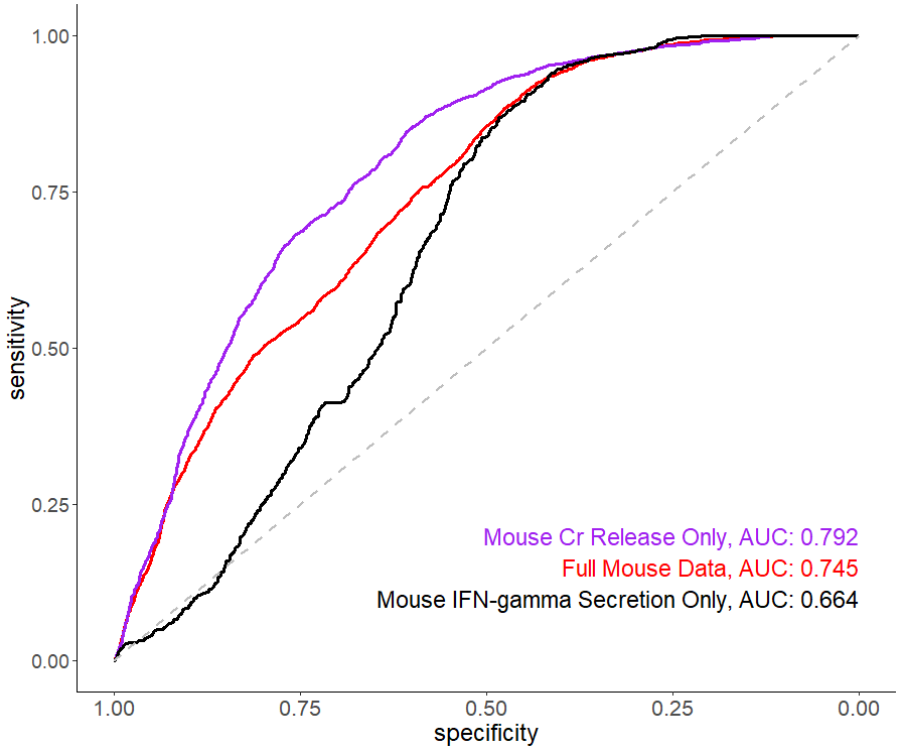
Appendix B: for Supplemental Figures and Tables from Chapter II

Appendix B, Table 7.1.

T-cell entries by database.

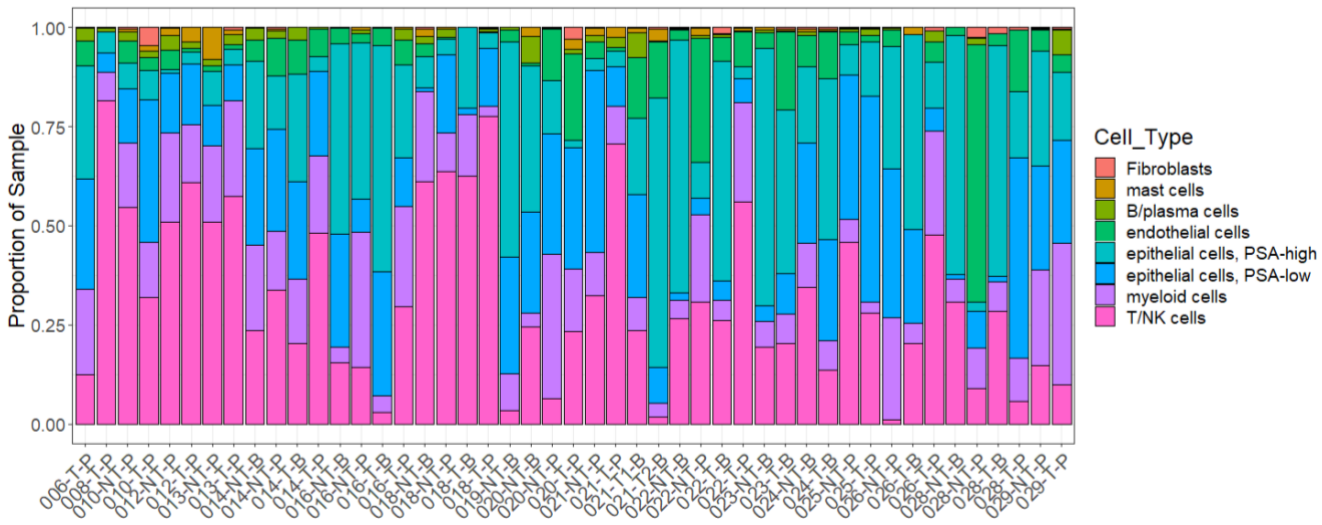
Database	Total Entries Isolated	Unique TCRs Identified	Avg. Epitopes per TCR
IEDB	8,700	3902	2.23
VDJDB	3,153	1330	2.37
PIRD	38	19	2.00

Number of T-cell receptors (TCRs) identified and isolated from each database for further aggregation.

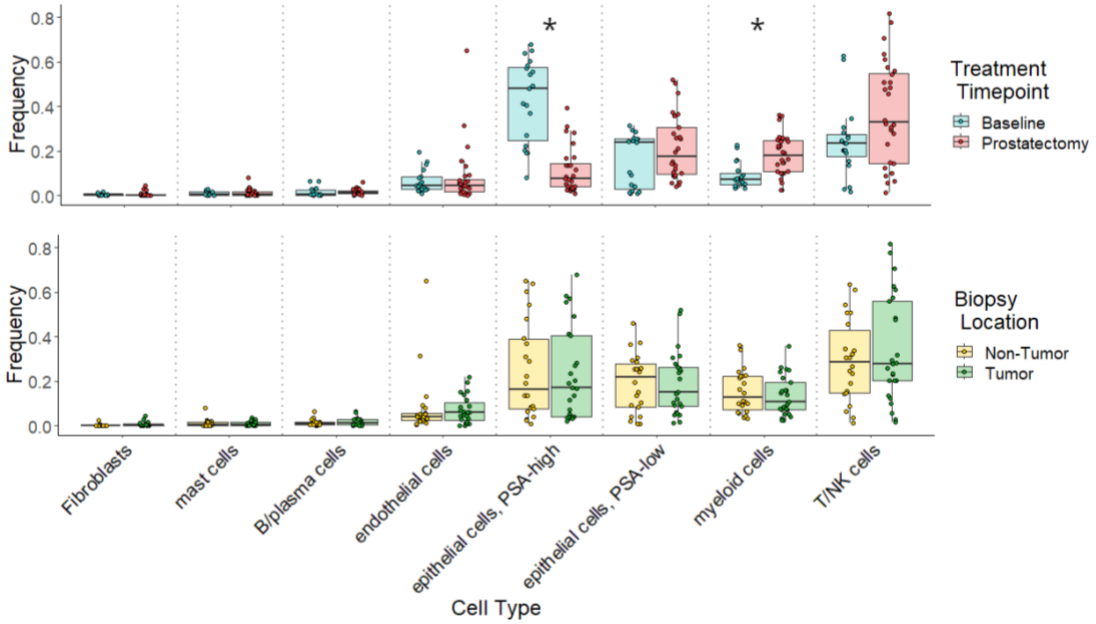


Appendix B, Figure 7.1. Crossreactor performance on manuscript derived mouse data. *Receiver operating characteristic (ROC) curves are shown for mouse data derived from: Chromium-release (Cr release) assays, IFN-gamma release assays,or the aggregation of the two. Sensitivity (y-axis) and specificity (x-axis) were evaluated on paired epitope examples, with a total of 6714 Cr-based examples and 2196 IFN-gamma based examples. Model performance deteriorates substantially on example datasets generated either partially or entirely from IFN-gamma based assays but recovers when restricting to mouse-based Cr-release assay data only.*

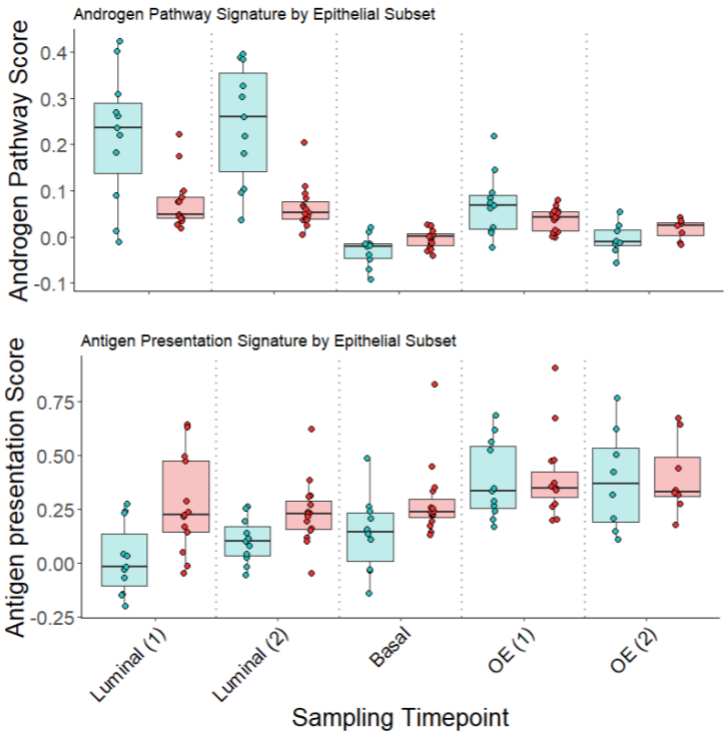
Appendix C: for Supplemental Figures and Tables from Chapter III



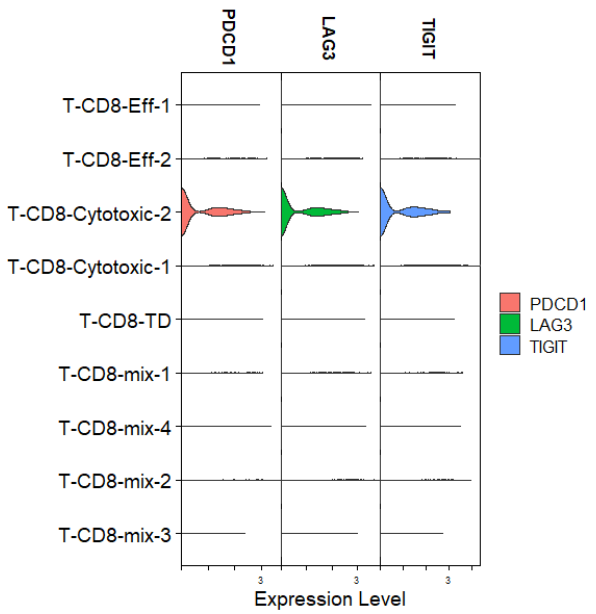
Appendix C Figure 8.1 Distribution of identified cell types across patient samples. *Each bar represents an individual patient sample (x-axis). Colors represent individual cell types identified through single-cell transcriptomic characterization. The proportion of a given sample constituted by a specific cell type is represented from 0-1 on the y-axis.*



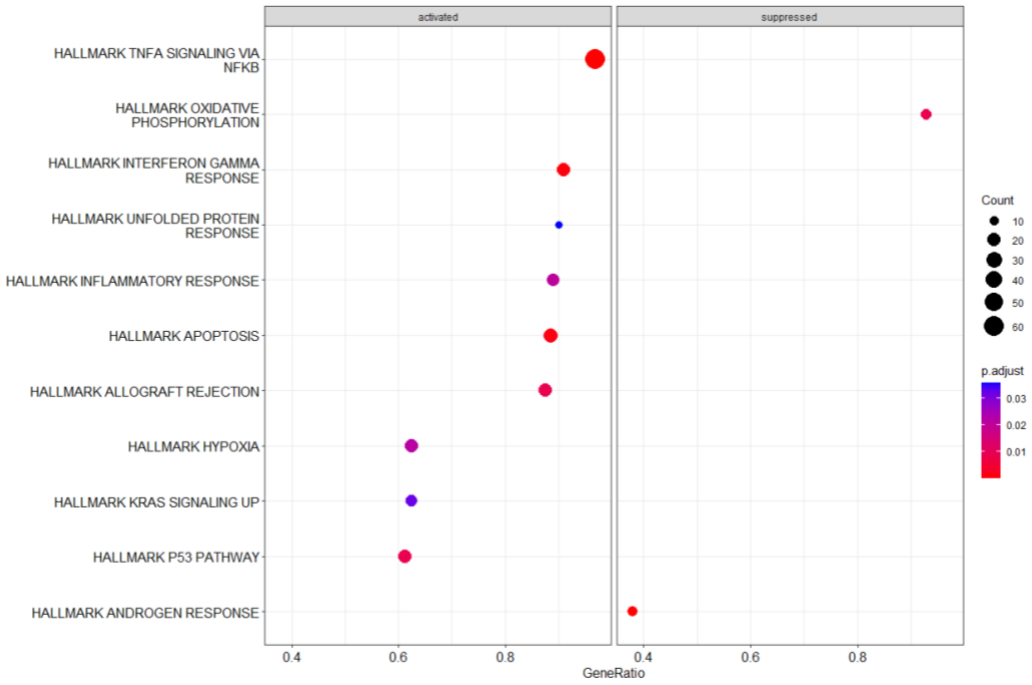
Appendix C Figure 8.2. Cell type proportions aggregated by tissue and time. *Each of the 8 broadly identified cell types is represented by a set of boxes. Each sample is represented by a dot and aggregated by timepoint (top) or location (bottom). For time based aggregations samples aggregated at baseline are represented in blue and those aggregated at prostatectomy are represented in red. For location based comparison non-tumor and tumor aggregations are represented by yellow and green respectively. Asterisks (*) indicate significant difference (Wilcoxon rank-sum test, p<0.05) in cell composition between baseline and post-treatment timepoints or tumor and non-tumor locations.*



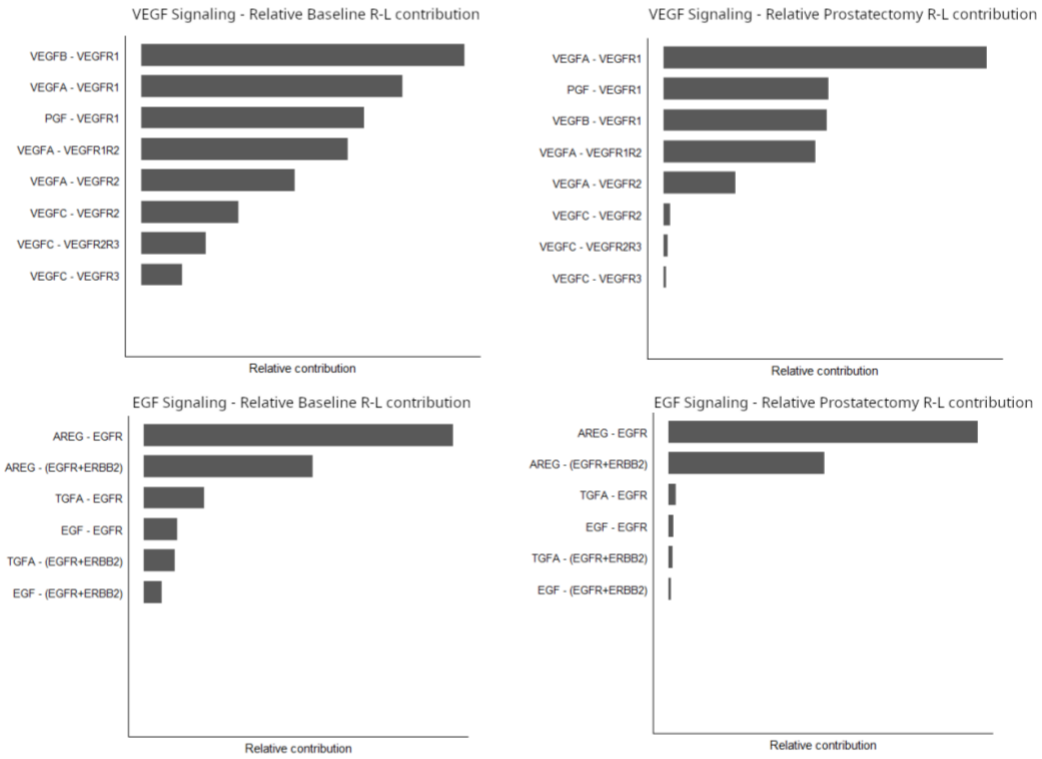
Appendix C Figure 8.3 Treatment based changes in androgen pathway and antigen presentation scores by epithelial sub group. *Androgen pathway score (top) and antigen presentation score (bottom) are compared across baseline (blue) and prostatectomy (red) for each sub-population of epithelial cells.*



Appendix C Figure 8.4 CD8 T-cell expression of exhaustion markers. *Y-axis shows each of the identified CD8 T-cell populations. X-axis represents normalized expression of each exhaustion-associated gene: PDCD1 (PD-1), LAG3, and TIGIT.*



Appendix C Figure 8.5. Gene set enrichment analysis of hallmark pathways in psuedobulk tissue. *Enrichment of hallmark pathways was assessed using differentially expressed genes identified through psuedobulk comparison of all cell types in aggregate between baseline and prostatectomy tissues.*



Appendix C. Figure 8.6. Relative receptor-ligand contributions to pathway signaling. Receptor-ligand pairs (y-axes) active in VEGF signaling (top) and EGF signaling (bottom) are each represented by a bar. The relative size of bars within each plot represent the relative contribution of pairs within and not overall strength of receptor-ligand signaling. Left plots represent receptor-ligand pairs at basline and right plots represent receptor-ligand pairs at prostatectomy

References

[1] E. W. Hewitt, “The MHC class I antigen presentation pathway: strategies for viral immune evasion,” *Immunology*, vol. 110, no. 2, pp. 163–169, Oct. 2003, doi: 10.1046/j.1365-2567.2003.01738.x.

[2] J. Neefjes, M. L. M. Jongsma, P. Paul, and O. Bakke, “Towards a systems understanding of MHC class I and MHC class II antigen presentation,” *Nat. Rev. Immunol.*, vol. 11, no. 12, Art. no. 12, Dec. 2011, doi: 10.1038/nri3084.

[3] J. G. Donaldson and D. B. Williams, “Intracellular Assembly and Trafficking of MHC Class I Molecules,” *Traffic Cph. Den.*, vol. 10, no. 12, pp. 1745–1752, Dec. 2009, doi: 10.1111/j.1600-0854.2009.00979.x.

[4] J. S. Blum, P. A. Wearsch, and P. Cresswell, “Pathways of Antigen Processing,” *Annu. Rev. Immunol.*, vol. 31, pp. 443–473, 2013, doi: 10.1146/annurev-immunol-032712-095910.

[5] P. V. Markov and O. G. Pybus, “Evolution and Diversity of the Human Leukocyte Antigen(HLA),” *Evol. Med. Public Health*, vol. 2015, no. 1, p. 1, Jan. 2015, doi: 10.1093/emph/eou033.

[6] G. M. Cooper, “Protein Degradation,” in *The Cell: A Molecular Approach. 2nd edition*, Sinauer Associates, 2000. Accessed: Aug. 30, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK9957/

[7] D. A. Ferrington and D. S. Gregerson, “Immunoproteasomes: Structure, Function, and Antigen Presentation,” *Prog. Mol. Biol. Transl. Sci.*, vol. 109, pp. 75–112, 2012, doi: 10.1016/B978-0-12-397863-9.00003-1.

[8] D. Hanahan, “Hallmarks of Cancer: New Dimensions,” *Cancer Discov.*, vol. 12, no. 1, pp. 31–46, Jan. 2022, doi: 10.1158/2159-8290.CD-21-1059.

[9] T. L. Whiteside, “The tumor microenvironment and its role in promoting tumor growth,” *Oncogene*, vol. 27, no. 45, Art. no. 45, Oct. 2008, doi: 10.1038/onc.2008.271.

[10] A. M. Cornel, I. L. Mimpen, and S. Nierkens, “MHC Class I Downregulation in Cancer: Underlying Mechanisms and Potential Targets for Cancer Immunotherapy,” *Cancers*, vol. 12, no. 7, p. 1760, Jul. 2020, doi: 10.3390/cancers12071760.

[11] J.-H. Cha, L.-C. Chan, C.-W. Li, J. L. Hsu, and M.-C. Hung, “Mechanisms Controlling PD-L1 Expression in Cancer,” *Mol. Cell*, vol. 76, no. 3, pp. 359–370, Nov. 2019, doi: 10.1016/j.molcel.2019.09.030.

[12] K. Dhatchinamoorthy, J. D. Colbert, and K. L. Rock, “Cancer Immune Evasion Through Loss of MHC Class I Antigen Presentation,” *Front. Immunol.*, vol. 12, 2021, Accessed: Oct. 17, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fimmu.2021.636568

[13] M. M. Gubin and M. D. Vesely, “Cancer Immunoediting in the Era of Immuno-oncology,” *Clin. Cancer Res.*, vol. 28, no. 18, pp. 3917–3928, Sep. 2022, doi: 10.1158/1078-0432.CCR-21-1804.

[14] M. Angelova *et al.*, “Evolution of Metastases in Space and Time under Immune Selection,” *Cell*, vol. 175, no. 3, pp. 751-765.e16, Oct. 2018, doi: 10.1016/j.cell.2018.09.018.

[15] M. Samec *et al.*, “The role of plant-derived natural substances as immunomodulatory agents in carcinogenesis,” *J. Cancer Res. Clin. Oncol.*, vol. 146, no. 12, pp. 3137–3154, Dec. 2020, doi: 10.1007/s00432-020-03424-2.

[16]E. S. Borden, K. H. Buetow, M. A. Wilson, and K. T. Hastings, “Cancer Neoantigens: Challenges and Future Directions for Prediction, Prioritization, and Validation,” *Front. Oncol.*, vol. 12, 2022, Accessed: Aug. 30, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fonc.2022.836821
[17]M. J. Fusco, H. (Jack) West, and C. M. Walko, “Tumor Mutation Burden and Cancer Treatment,” *JAMA Oncol.*, vol. 7, no. 2, p. 316, Feb. 2021, doi: 10.1001/jamaoncol.2020.6371.
[18]N. J. Birkbak *et al.*, “Tumor Mutation Burden Forecasts Outcome in Ovarian Cancer with BRCA1 or BRCA2 Mutations,” *PLOS ONE*, vol. 8, no. 11, p. e80023, Nov. 2013, doi: 10.1371/journal.pone.0080023.
[19]N. A. Rizvi *et al.*, “Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer,” *Science*, vol. 348, no. 6230, pp. 124–128, Apr. 2015, doi: 10.1126/science.aaa1348.
[20]J. H. Strickler, B. A. Hanks, and M. Khasraw, “Tumor Mutational Burden as a Predictor of Immunotherapy Response: Is More Always Better?,” *Clin. Cancer Res.*, vol. 27, no. 5, pp. 1236–1241, Mar. 2021, doi: 10.1158/1078-0432.CCR-20-3054.
[21]D. L. Jardim, A. Goodman, D. de Melo Gagliato, and R. Kurzrock, “The Challenges of Tumor Mutational Burden as an Immunotherapy Biomarker,” *Cancer Cell*, vol. 39, no. 2, pp. 154–173, Feb. 2021, doi: 10.1016/j.ccell.2020.10.001.
[22]M. A. Wood, B. R. Weeder, J. K. David, A. Nellore, and R. F. Thompson, “Burden of tumor mutations, neoepitopes, and other variants are weak predictors of cancer immunotherapy response and overall survival,” *Genome Med.*, vol. 12, no. 1, p. 33, Mar. 2020, doi: 10.1186/s13073-020-00729-2.
[23]G. Carino, T. Dina, I. Maxim, L. L. J, and M. L. A, “Is tumor mutational burden predictive of response to immunotherapy?,” *eLife*, vol. 12, Jun. 2023, doi: 10.7554/eLife.87465.
[24]G. Fotakis, Z. Trajanoski, and D. Rieder, “Computational cancer neoantigen prediction: current status and recent advances,” *Immuno-Oncol. Technol.*, vol. 12, p. 100052, Dec. 2021, doi: 10.1016/j.iotech.2021.100052.
[25]P. R. Buckley *et al.*, “Evaluating performance of existing computational models in predicting CD8+ T cell pathogenic epitopes and cancer neoantigens,” *Brief. Bioinform.*, vol. 23, no. 3, p. bbac141, May 2022, doi: 10.1093/bib/bbac141.
[26]M. A. Wood, A. Nguyen, A. J. Struck, K. Ellrott, A. Nellore, and R. F. Thompson, “neoepiscope improves neoepitope prediction with multivariant phasing,” *Bioinformatics*, vol. 36, no. 3, pp. 713–720, Feb. 2020, doi: 10.1093/bioinformatics/btz653.
[27]J. K. David, S. K. Maden, B. R. Weeder, R. F. Thompson, and A. Nellore, “Putatively cancer-specific exon–exon junctions are shared across patients and present in developmental and other non-cancer cells,” *NAR Cancer*, vol. 2, no. 1, p. zcaa001, Mar. 2020, doi: 10.1093/narcan/zcaa001.
[28]M. Gomez-Perosanz, A. Ras-Carmona, E. M. Lafuente, and P. A. Reche, “Identification of CD8+ T cell epitopes through proteasome cleavage site predictions,” *BMC Bioinformatics*, vol. 21, no. 17, p. 484, Dec. 2020, doi: 10.1186/s12859-020-03782-1.
[29]J.-R. Hwang, Y. Byeon, D. Kim, and S.-G. Park, “Recent insights of T cell receptor-mediated signaling pathways for T cell activation and development,” *Exp. Mol. Med.*, vol. 52, no. 5, Art. no. 5, May 2020, doi: 10.1038/s12276-020-0435-8.
[30]Y. Zhang and J. Zheng, “Functions of Immune Checkpoint Molecules Beyond Immune Evasion,” *Adv. Exp. Med. Biol.*, vol. 1248, pp. 201–226, 2020, doi: 10.1007/978-981-15-3266-5_9.

[31]M. H. Gee *et al.*, “Stress-testing the relationship between T cell receptor/peptide-MHC affinity and cross-reactivity using peptide velcro,” *Proc. Natl. Acad. Sci.*, vol. 115, no. 31, pp. E7369–E7378, Jul. 2018, doi: 10.1073/pnas.1802746115.
[32]M. G. Rudolph and I. A. Wilson, “The specificity of TCR/pMHC interaction,” *Curr. Opin. Immunol.*, vol. 14, no. 1, pp. 52–65, Feb. 2002, doi: 10.1016/S0952-7915(01)00298-9.
[33]D. Jung and F. W. Alt, “Unraveling V(D)J Recombination: Insights into Gene Regulation,” *Cell*, vol. 116, no. 2, pp. 299–311, Jan. 2004, doi: 10.1016/S0092-8674(04)00039-X.
[34]G. Kaeser and J. Chun, “Brain cell somatic gene recombination and its phylogenetic foundations,” *J. Biol. Chem.*, vol. 295, no. 36, pp. 12786–12795, Sep. 2020, doi: 10.1074/jbc.REV120.009192.
[35]E. P. Rock, P. R. Sibbald, M. M. Davis, and Y. H. Chien, “CDR3 length in antigen-specific immune receptors,” *J. Exp. Med.*, vol. 179, no. 1, pp. 323–328, Jan. 1994, doi: 10.1084/jem.179.1.323.
[36]A. K. Sewell, “Why must T cells be cross-reactive?,” *Nat. Rev. Immunol.*, vol. 12, no. 9, pp. 669–677, 2012, doi: 10.1038/nri3279.
[37]G. Lythe, R. E. Callard, R. L. Hoare, and C. Molina-París, “How many TCR clonotypes does a body maintain?,” *J. Theor. Biol.*, vol. 389, pp. 214–224, Jan. 2016, doi: 10.1016/j.jtbi.2015.10.016.
[38]K. C. Garcia and E. J. Adams, “How the T Cell Receptor Sees Antigen—A Structural View,” *Cell*, vol. 122, no. 3, pp. 333–336, Aug. 2005, doi: 10.1016/j.cell.2005.07.015.
[39]L. V. Sibener *et al.*, “Isolation of a Structural Mechanism for Uncoupling T Cell Receptor Signaling from Peptide-MHC Binding,” *Cell*, vol. 174, no. 3, pp. 672-687.e27, Jul. 2018, doi: 10.1016/j.cell.2018.06.017.
[40]L. Wooldridge *et al.*, “A single autoimmune T cell receptor recognizes more than a million different peptides.,” *J. Biol. Chem.*, vol. 287, no. 2, pp. 1168–77, Jan. 2012, doi: 10.1074/jbc.M111.289488.
[41]D. Mason, “A very high level of crossreactivity is an essential feature of the T-cell receptor,” *Immunol. Today*, vol. 19, no. 9, pp. 395–404, Sep. 1998, doi: 10.1016/S0167-5699(98)01299-7.
[42]C. Münz, J. D. Lünemann, M. T. Getts, and S. D. Miller, “Antiviral immune responses: triggers of or triggered by autoimmunity?,” *Nat. Rev. Immunol.*, vol. 9, no. 4, pp. 246–258, Apr. 2009, doi: 10.1038/nri2527.
[43]G.-M. Deng and G. C. Tsokos, “Cholera toxin B accelerates disease progression in lupus-prone mice by promoting lipid raft aggregation,” *J. Immunol. Baltim. Md 1950*, vol. 181, no. 6, pp. 4019–4026, Sep. 2008.
[44]N. Balandraud *et al.*, “Epstein-Barr virus load in the peripheral blood of patients with rheumatoid arthritis: accurate quantification using real-time polymerase chain reaction,” *Arthritis Rheum.*, vol. 48, no. 5, pp. 1223–1228, May 2003, doi: 10.1002/art.10933.
[45]D. J. Pallin, C. W. Baugh, M. A. Postow, J. M. Caterino, T. B. Erickson, and G. H. Lyman, “Immune-Related Adverse Events in Cancer Patients,” *Acad. Emerg. Med. Off. J. Soc. Acad. Emerg. Med.*, vol. 25, no. 7, pp. 819–827, Jul. 2018, doi: 10.1111/acem.13443.
[46]G. P. Linette *et al.*, “Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma,” *Blood*, vol. 122, no. 6, pp. 863–871, Aug. 2013, doi: 10.1182/blood-2013-03-490565.
[47]B. J. Cameron *et al.*, “Identification of a Titin-Derived HLA-A1–Presented Peptide as a Cross-Reactive Target for Engineered MAGE A3–Directed T Cells,” *Sci. Transl.*

Med., vol. 5, no. 197, pp. 197ra103-197ra103, Aug. 2013, doi: 10.1126/scitranslmed.3006034.
[48]C. H. Lee, M. Salio, G. Napolitani, G. Ogg, A. Simmons, and H. Koohy, “Predicting Cross-Reactivity and Antigen Specificity of T Cell Receptors,” *Front. Immunol.*, vol. 11, p. 565096, Oct. 2020, doi: 10.3389/fimmu.2020.565096.
[49]S. Hall-Swan, J. Slone, M. M. Rigo, D. A. Antunes, G. Lizée, and L. E. Kavraki, “PepSim: T-cell cross-reactivity prediction via comparison of peptide sequence and peptide-HLA structure,” *Front. Immunol.*, vol. 14, p. 1108303, 2023, doi: 10.3389/fimmu.2023.1108303.
[50]S. Frankild, R. J. de Boer, O. Lund, M. Nielsen, and C. Kesmir, “Amino Acid Similarity Accounts for T Cell Cross-Reactivity and for ‘Holes’ in the T Cell Repertoire,” *PLOS ONE*, vol. 3, no. 3, p. e1831, Mar. 2008, doi: 10.1371/journal.pone.0001831.
[51]E. Jokinen *et al.*, “TCRconv: predicting recognition between T cell receptors and epitopes using contextualized motifs,” *Bioinformatics*, vol. 39, no. 1, p. btac788, Jan. 2023, doi: 10.1093/bioinformatics/btac788.
[52]Y. Tsuchiya, Y. Namiuchi, H. Wako, and H. Tsurui, “A study of CDR3 loop dynamics reveals distinct mechanisms of peptide recognition by T-cell receptors exhibiting different levels of cross-reactivity,” *Immunology*, vol. 153, no. 4, pp. 466–478, Apr. 2018, doi: 10.1111/imm.12849.
[53]C. Pan *et al.*, “Next-generation immuno-oncology agents: current momentum shifts in cancer immunotherapy,” *J. Hematol. Oncol.J Hematol Oncol*, vol. 13, no. 1, p. 29, Apr. 2020, doi: 10.1186/s13045-020-00862-w.
[54]A. Labani-Motlagh, M. Ashja-Mahdavi, and A. Loskog, “The Tumor Microenvironment: A Milieu Hindering and Obstructing Antitumor Immune Responses,” *Front. Immunol.*, vol. 11, 2020, Accessed: Sep. 01, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fimmu.2020.00940
[55]Q. Wang *et al.*, “Role of tumor microenvironment in cancer progression and therapeutic strategy,” *Cancer Med.*, vol. 12, no. 10, pp. 11149–11165, 2023, doi: 10.1002/cam4.5698.
[56]C. Fu and A. Jiang, “Dendritic Cells and CD8 T Cell Immunity in Tumor Microenvironment,” *Front. Immunol.*, vol. 9, 2018, Accessed: Sep. 02, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fimmu.2018.03059
[57]J. Liu, X. Geng, J. Hou, and G. Wu, “New insights into M1/M2 macrophages: key modulators in cancer progression,” *Cancer Cell Int.*, vol. 21, no. 1, p. 389, Jul. 2021, doi: 10.1186/s12935-021-02089-2.
[58]F. Veglia, E. Sanseviero, and D. I. Gabrilovich, “Myeloid-derived suppressor cells in the era of increasing myeloid cell diversity,” *Nat. Rev. Immunol.*, vol. 21, no. 8, Art. no. 8, Aug. 2021, doi: 10.1038/s41577-020-00490-y.
[59]K. Kondělková, D. Vokurková, J. Krejsek, L. Borská, Z. Fiala, and A. Ctirad, “Regulatory T cells (TREG) and their roles in immune system with respect to immunopathological disorders,” *Acta Medica (Hradec Kralove)*, vol. 53, no. 2, pp. 73–77, 2010, doi: 10.14712/18059694.2016.63.
[60]L. Yang and P. C. Lin, “Mechanisms that drive inflammatory tumor microenvironment, tumor heterogeneity, and metastatic progression,” *Semin. Cancer Biol.*, vol. 47, pp. 185–195, Dec. 2017, doi: 10.1016/j.semcancer.2017.08.001.
[61]P. Bonaventura *et al.*, “Cold Tumors: A Therapeutic Challenge for Immunotherapy,” *Front. Immunol.*, vol. 10, 2019, Accessed: Sep. 02, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fimmu.2019.00168

[62]E. J. Wherry and M. Kurachi, “Molecular and cellular insights into T cell exhaustion,” *Nat. Rev. Immunol.*, vol. 15, no. 8, pp. 486–499, Aug. 2015, doi: 10.1038/nri3862.
[63]C. U. Blank *et al.*, “Defining ‘T cell exhaustion,’” *Nat. Rev. Immunol.*, vol. 19, no. 11, Art. no. 11, Nov. 2019, doi: 10.1038/s41577-019-0221-9.
[64]J. S. Yi, M. A. Cox, and A. J. Zajac, “T-cell exhaustion: characteristics, causes and conversion,” *Immunology*, vol. 129, no. 4, pp. 474–481, Apr. 2010, doi: 10.1111/j.1365-2567.2010.03255.x.
[65]Y. Shiravand *et al.*, “Immune Checkpoint Inhibitors in Cancer Therapy,” *Curr. Oncol.*, vol. 29, no. 5, pp. 3044–3060, Apr. 2022, doi: 10.3390/curroncol29050247.
[66]C. Robert, “A decade of immune-checkpoint inhibitors in cancer therapy,” *Nat. Commun.*, vol. 11, no. 1, Art. no. 1, Jul. 2020, doi: 10.1038/s41467-020-17670-y.
[67]X. Liu, G. D. Hogg, and D. G. DeNardo, “Rethinking immune checkpoint blockade: ‘Beyond the T cell,’” *J. Immunother. Cancer*, vol. 9, no. 1, p. e001460, Jan. 2021, doi: 10.1136/jitc-2020-001460.
[68]S. Vaga, “Understanding Single Cell Sequencing, How It Works and Its Applications,” Genomics Research from Technology Networks. Accessed: Sep. 02, 2023. [Online]. Available: http://www.technologynetworks.com/genomics/articles/understanding-single-cell-sequencing-how-it-works-and-its-applications-357578
[69]X. Wang, Y. He, Q. Zhang, X. Ren, and Z. Zhang, “Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2,” *Genomics Proteomics Bioinformatics*, vol. 19, no. 2, pp. 253–266, Apr. 2021, doi: 10.1016/j.gpb.2020.02.005.
[70]P. See, J. Lum, J. Chen, and F. Ginhoux, “A Single-Cell Sequencing Guide for Immunologists,” *Front. Immunol.*, vol. 9, 2018, doi: 10.3389/fimmu.2018.02425.
[71]B. R. Weeder, M. A. Wood, E. Li, A. Nellore, and R. F. Thompson, “pepsickle rapidly and accurately predicts proteasomal cleavage sites for improved neoantigen identification,” *Bioinformatics*, vol. 37, no. 21, pp. 3723–3733, Nov. 2021, doi: 10.1093/bioinformatics/btab628.
[72]A. Ciechanover, “The ubiquitin-proteasome proteolytic pathway,” *Cell*, vol. 79, no. 1, pp. 13–21, Oct. 1994, doi: 10.1016/0092-8674(94)90396-4.
[73]P.-M. Kloetzel and F. Ossendorp, “Proteasome and peptidase function in MHC-class-I-mediated antigen presentation,” *Curr. Opin. Immunol.*, vol. 16, no. 1, pp. 76–81, Feb. 2004, doi: 10.1016/j.coi.2003.11.004.
[74]R. Carretero *et al.*, “Regression of melanoma metastases after immunotherapy is associated with activation of antigen presentation and interferon-mediated rejection genes,” *Int. J. Cancer*, vol. 131, no. 2, pp. 387–395, 2012, doi: 10.1002/ijc.26471.
[75]S. Ostrand-Rosenberg, “Tumor immunotherapy: the tumor cell as an antigen-presenting cell,” *Curr. Opin. Immunol.*, vol. 6, no. 5, pp. 722–727, Jan. 1994, doi: 10.1016/0952-7915(94)90075-2.
[76]N. Pardi, M. J. Hogan, F. W. Porter, and D. Weissman, “mRNA vaccines — a new era in vaccinology,” *Nat. Rev. Drug Discov.*, vol. 17, no. 4, Art. no. 4, Apr. 2018, doi: 10.1038/nrd.2017.243.
[77]M. T. Bethune *et al.*, “Isolation and characterization of NY-ESO-1–specific T cell receptors restricted on various MHC molecules,” *Proc. Natl. Acad. Sci.*, vol. 115, no. 45, pp. E10702–E10711, Nov. 2018, doi: 10.1073/pnas.1810653115.
[78]E. J. A. M. Sijts and P.-M. Kloetzel, “The role of the proteasome in the generation of MHC class I ligands and immune responses,” *Cell. Mol. Life Sci.*, vol. 68, no. 9, pp. 1491–1502, May 2011, doi: 10.1007/s00018-011-0657-y.
[79]J. Adams, “The proteasome: structure, function, and role in the cell,” *Cancer Treat. Rev.*, vol. 29, pp. 3–9, May 2003, doi: 10.1016/S0305-7372(03)00081-1.

[80]L. Budenholzer, C. L. Cheng, Y. Li, and M. Hochstrasser, “Proteasome Structure and Assembly,” *J. Mol. Biol.*, vol. 429, no. 22, pp. 3500–3524, Nov. 2017, doi: 10.1016/j.jmb.2017.05.027.
[81]M. Nielsen, C. Lundegaard, O. Lund, and C. Keşmir, “The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage,” *Immunogenetics*, vol. 57, no. 1–2, pp. 33–41, Apr. 2005, doi: 10.1007/s00251-005-0781-7.
[82]P. Fort, A. V. Kajava, F. Delsuc, and O. Coux, “Evolution of Proteasome Regulators in Eukaryotes,” *Genome Biol. Evol.*, vol. 7, no. 5, pp. 1363–1379, May 2015, doi: 10.1093/gbe/evv068.
[83]M. Bhasin and G. P. S. Raghava, “Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences,” *Nucleic Acids Res.*, vol. 33, no. Web Server issue, pp. W202-207, Jul. 2005, doi: 10.1093/nar/gki587.
[84]H. G. Holzhütter and P. M. Kloetzel, “A kinetic model of vertebrate 20S proteasome accounting for the generation of major proteolytic fragments from oligomeric peptide substrates,” *Biophys. J.*, vol. 79, no. 3, pp. 1196–1205, Sep. 2000, doi: 10.1016/S0006-3495(00)76374-0.
[85]A. K. Nussbaum, C. Kuttler, K. P. Hadeler, H. G. Rammensee, and H. Schild, “PAProC: a prediction algorithm for proteasomal cleavages available on the WWW,” *Immunogenetics*, vol. 53, no. 2, pp. 87–94, Mar. 2001, doi: 10.1007/s002510100300.
[86]B.-Q. Li, Y.-D. Cai, K.-Y. Feng, and G.-J. Zhao, “Prediction of Protein Cleavage Site with Feature Selection by Random Forest,” *PLOS ONE*, vol. 7, no. 9, p. e45854, Sep. 2012, doi: 10.1371/journal.pone.0045854.
[87]R. J. Tomko and M. Hochstrasser, “Molecular Architecture and Assembly of the Eukaryotic Proteasome,” *Annu. Rev. Biochem.*, vol. 82, no. 1, pp. 415–445, 2013, doi: 10.1146/annurev-biochem-060410-150257.
[88]B. Alvarez-Castelao, C. Muñoz, I. Sánchez, M. Goethals, J. Vandekerckhove, and J. G. Castaño, “Reduced protein stability of human DJ-1/PARK7 L166P, linked to autosomal recessive Parkinson disease, is due to direct endoproteolytic cleavage by the proteasome,” *Biochim. Biophys. Acta BBA - Mol. Cell Res.*, vol. 1823, no. 2, pp. 524–533, Feb. 2012, doi: 10.1016/j.bbamcr.2011.11.010.
[89]B. Alvarez-Castelao, M. Goethals, J. Vandekerckhove, and J. G. Castaño, “Mechanism of cleavage of alpha-synuclein by the 20S proteasome and modulation of its degradation by the RedOx state of the N-terminal methionines,” *Biochim. Biophys. Acta BBA - Mol. Cell Res.*, vol. 1843, no. 2, pp. 352–365, Feb. 2014, doi: 10.1016/j.bbamcr.2013.11.018.
[90]A. M. Asemissen *et al.*, “Identification of a Highly Immunogenic HLA-A*01-Binding T Cell Epitope of WT1,” *Clin. Cancer Res.*, vol. 12, no. 24, pp. 7476–7482, Dec. 2006, doi: 10.1158/1078-0432.CCR-06-1337.
[91]M. Ayyoub *et al.*, “Proteasome-Assisted Identification of a SSX-2-Derived Epitope Recognized by Tumor-Reactive CTL Infiltrating Metastatic Melanoma,” *J. Immunol.*, vol. 168, no. 4, pp. 1717–1722, Feb. 2002, doi: 10.4049/jimmunol.168.4.1717.
[92]D. Berko *et al.*, “The Direction of Protein Entry into the Proteasome Determines the Variety of Peptide Products and Depends on the Force Needed to Unfold Its Two Termini,” *Mol. Cell*, vol. 48, no. 4, pp. 601–611, Nov. 2012, doi: 10.1016/j.molcel.2012.08.029.

[93]D. Bruder *et al.*, “Multiple synergizing factors contribute to the strength of the CD8+ T cell response against listeriolysin O,” *Int. Immunol.*, vol. 18, no. 1, pp. 89–100, Jan. 2006, doi: 10.1093/intimm/dxh352.
[94]J. Chapiro *et al.*, “Destructive Cleavage of Antigenic Peptides Either by the Immunoproteasome or by the Standard Proteasome Results in Differential Antigen Presentation,” *J. Immunol.*, vol. 176, no. 2, pp. 1053–1061, Jan. 2006, doi: 10.4049/jimmunol.176.2.1053.
[95]T. P. Dick *et al.*, “Coordinated dual cleavages induced by the proteasome regulator PA28 lead to dominant MHC ligands,” *Cell*, vol. 86, no. 2, pp. 253–262, Jul. 1996, doi: 10.1016/s0092-8674(00)80097-5.
[96]B. Ehring, T. H. Meyer, C. Eckerskorn, F. Lottspeich, and R. Tampé, “Effects of Major-Histocompatibility-Complex-Encoded Subunits on the Peptidase and Proteolytic Activities of Human 20S Proteasomes,” *Eur. J. Biochem.*, vol. 235, no. 1–2, pp. 404–415, 1996, doi: https://doi.org/10.1111/j.1432-1033.1996.00404.x.
[97]N. P. N. Emmerich *et al.*, “The Human 26 S and 20 S Proteasomes Generate Overlapping but Different Sets of Peptide Fragments from a Model Protein Substrate,” *J. Biol. Chem.*, vol. 275, no. 28, pp. 21140–21148, Jul. 2000, doi: 10.1074/jbc.M000740200.
[98]N. Garcia-Medel, A. Sanz, E. Barnea, A. Admon, and J. A. L. de Castro, “The origin of proteasome-inhibitor resistant HLA class I peptidomes: a study with HLA-A*68:01,” *Mol. Cell. Proteomics*, Jan. 2011, doi: 10.1074/mcp.M111.011486.
[99]B. Guillaume *et al.*, “Analysis of the Processing of Seven Human Tumor Antigens by Intermediate Proteasomes,” *J. Immunol.*, vol. 189, no. 7, pp. 3538–3547, Oct. 2012, doi: 10.4049/jimmunol.1103213.
[100] Y. Hassainya *et al.*, “Identification of Naturally Processed HLA-A2—Restricted Proinsulin Epitopes by Reverse Immunology,” *Diabetes*, vol. 54, no. 7, pp. 2053–2059, Jul. 2005, doi: 10.2337/diabetes.54.7.2053.
[101] J. H. Kessler *et al.*, “Efficient Identification of Novel Hla-A*0201–Presented Cytotoxic T Lymphocyte Epitopes in the Widely Expressed Tumor Antigen Prame by Proteasome-Mediated Digestion Analysis,” *J. Exp. Med.*, vol. 193, no. 1, pp. 73–88, Jan. 2001, doi: 10.1084/jem.193.1.73.
[102] Y. Kimura, T. Gushima, S. Rawale, P. Kaumaya, and C. M. Walker, “Escape Mutations Alter Proteasome Processing of Major Histocompatibility Complex Class I-Restricted Epitopes in Persistent Hepatitis C Virus Infection,” *J. Virol.*, vol. 79, no. 8, pp. 4870–4876, Apr. 2005, doi: 10.1128/JVI.79.8.4870-4876.2005.
[103] M. Lucchiari-Hartz *et al.*, “Cytotoxic T Lymphocyte Epitopes of HIV-1 Nef: Generation of Multiple Definitive Major Histocompatibility Complex Class I Ligands by Proteasomes,” *J. Exp. Med.*, vol. 191, no. 2, pp. 239–252, Jan. 2000, doi: 10.1084/jem.191.2.239.
[104] M. Lucchiari-Hartz *et al.*, “Differential proteasomal processing of hydrophobic and hydrophilic protein regions: Contribution to cytotoxic T lymphocyte epitope clustering in HIV-1-Nef,” *Proc. Natl. Acad. Sci.*, vol. 100, no. 13, pp. 7755–7760, Jun. 2003, doi: 10.1073/pnas.1232228100.
[105] D. Macconi *et al.*, “Proteasomal Processing of Albumin by Renal Dendritic Cells Generates Antigenic Peptides,” *J. Am. Soc. Nephrol.*, vol. 20, no. 1, pp. 123–130, Jan. 2009, doi: 10.1681/ASN.2007111233.
[106] M. Marcilla, J. A. L. D. Castro, J. G. Castaño, and I. Alvarez, “Infection with Salmonella typhimurium has no effect on the composition and cleavage specificity of the 20S proteasome in human lymphoid cells,” *Immunology*, vol. 122, no. 1, pp. 131–139, 2007, doi: https://doi.org/10.1111/j.1365-2567.2007.02624.x.

[107] A. Michaux *et al.*, “A Spliced Antigenic Peptide Comprising a Single Spliced Amino Acid Is Produced in the Proteasome by Reverse Splicing of a Longer Peptide Fragment followed by Trimming,” *J. Immunol.*, vol. 192, no. 4, pp. 1962–1971, Feb. 2014, doi: 10.4049/jimmunol.1302032.
[108] S. Morel *et al.*, “Processing of Some Antigens by the Standard Proteasome but Not by the Immunoproteasome Results in Poor Presentation by Dendritic Cells,” *Immunity*, vol. 12, no. 1, pp. 107–117, Jan. 2000, doi: 10.1016/S1074-7613(00)80163-6.
[109] G. Niedermann *et al.*, “Contribution of proteasome-mediated proteolysis to the hierarchy of epitopes presented by major histocompatibility complex class I molecules,” *Immunity*, vol. 2, no. 3, pp. 289–299, Mar. 1995, doi: 10.1016/1074-7613(95)90053-5.
[110] G. Niedermann *et al.*, “The proteolytic fragments generated by vertebrate proteasomes: structural relationships to major histocompatibility complex class I binding peptides,” *Proc. Natl. Acad. Sci.*, vol. 93, no. 16, pp. 8572–8577, Aug. 1996, doi: 10.1073/pnas.93.16.8572.
[111] A. Paradela *et al.*, “Limited diversity of peptides related to an alloreactive T cell epitope in the HLA-B27-bound peptide repertoire results from restrictions at multiple steps along the processing-loading pathway,” *J. Immunol. Baltim. Md 1950*, vol. 164, no. 1, pp. 329–337, Jan. 2000, doi: 10.4049/jimmunol.164.1.329.
[112] G. Pinkse *et al.*, “Autoreactive CD8 T cells associated with β cell destruction in type 1 diabetes,” 2005, doi: 10.1073/PNAS.0508621102.
[113] J. Popović, L.-P. Li, P. M. Kloetzel, M. Leisegang, W. Uckert, and T. Blankenstein, “The only proposed T-cell epitope derived from the TEL-AML1 translocation is not naturally processed,” *Blood*, vol. 118, no. 4, pp. 946–954, Jul. 2011, doi: 10.1182/blood-2010-12-325035.
[114] L. Sesma, I. Alvarez, M. Marcilla, A. Paradela, and J. A. L. de Castro, “Species-specific Differences in Proteasomal Processing and Tapasin-mediated Loading Influence Peptide Presentation by HLA-B27 in Murine Cells,” *J. Biol. Chem.*, vol. 278, no. 47, pp. 46461–46472, Nov. 2003, doi: 10.1074/jbc.M308816200.
[115] B. Strehl *et al.*, “Antitopes Define Preferential Proteasomal Cleavage Site Usage*,” *J. Biol. Chem.*, vol. 283, no. 26, pp. 17891–17897, Jun. 2008, doi: 10.1074/jbc.M710042200.
[116] S. Tenzer *et al.*, “Quantitative Analysis of Prion-Protein Degradation by Constitutive and Immuno-20S Proteasomes Indicates Differences Correlated with Disease Susceptibility,” *J. Immunol.*, vol. 172, no. 2, pp. 1083–1091, Jan. 2004, doi: 10.4049/jimmunol.172.2.1083.
[117] M. Theobald *et al.*, “The Sequence Alteration Associated with a Mutational Hotspot in p53 Protects Cells From Lysis by Cytotoxic T Lymphocytes Specific for a Flanking Peptide Epitope,” *J. Exp. Med.*, vol. 188, no. 6, pp. 1017–1028, Sep. 1998, doi: 10.1084/jem.188.6.1017.
[118] R. E. M. Toes *et al.*, “Discrete Cleavage Motifs of Constitutive and Immunoproteasomes Revealed by Quantitative Analysis of Cleavage Products,” *J. Exp. Med.*, vol. 194, no. 1, pp. 1–12, Jul. 2001, doi: 10.1084/jem.194.1.1.
[119] N. Vigneron *et al.*, “An Antigenic Peptide Produced by Peptide Splicing in the Proteasome,” *Science*, vol. 304, no. 5670, pp. 587–590, Apr. 2004, doi: 10.1126/science.1095522.
[120] H. Wada, A. Shimizu, T. Osada, Y. Tanaka, S. Fukaya, and E. Sasaki, “Development of a novel immunoproteasome digestion assay for synthetic long peptide vaccine design,” *PLOS ONE*, vol. 13, no. 7, p. e0199249, Jul. 2018, doi: 10.1371/journal.pone.0199249.

[121] E. H. Warren *et al.*, “An Antigen Produced by Splicing of Noncontiguous Peptides in the Reverse Order,” *Science*, vol. 313, no. 5792, pp. 1444–1447, Sep. 2006, doi: 10.1126/science.1130660.
[122] P. Zimbwa *et al.*, “Precise Identification of a Human Immunodeficiency Virus Type 1 Antigen Processing Mutant,” *J. Virol.*, vol. 81, no. 4, pp. 2031–2038, Feb. 2007, doi: 10.1128/JVI.00968-06.
[123] R. Vita *et al.*, “The Immune Epitope Database (IEDB): 2018 update,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D339–D343, 08 2019, doi: 10.1093/nar/gky1006.
[124] H. McSparron, M. J. Blythe, C. Zygouri, I. A. Doytchinova, and D. R. Flower, “JenPep: a novel computational information resource for immunobiology and vaccinology,” *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 4, pp. 1276–1287, Aug. 2003, doi: 10.1021/ci030461e.
[125] H.-G. Rammensee, J. Bachmann, N. P. N. Emmerich, O. A. Bachor, and S. Stevanović, “SYFPEITHI: database for MHC ligands and peptide motifs,” *Immunogenetics*, vol. 50, no. 3, pp. 213–219, Nov. 1999, doi: 10.1007/s002510050595.
[126] M. Bassani-Sternberg *et al.*, “Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry,” *Nat. Commun.*, vol. 7, no. 1, Art. no. 1, Nov. 2016, doi: 10.1038/ncomms13404.
[127] D. V. Rozanov *et al.*, “MHC class I loaded ligands from breast cancer cell lines: A potential HLA-I-typed antigen collection,” *J. Proteomics*, vol. 176, pp. 13–23, 30 2018, doi: 10.1016/j.jprot.2018.01.004.
[128] I. Evnouchidou and P. van Endert, “Peptide trimming by endoplasmic reticulum aminopeptidases: Role of MHC class I binding and ERAP dimerization,” *Hum. Immunol.*, vol. 80, no. 5, pp. 290–295, May 2019, doi: 10.1016/j.humimm.2019.01.003.
[129] A. Köhler, P. Cascio, D. S. Leggett, K. M. Woo, A. L. Goldberg, and D. Finley, “The Axial Channel of the Proteasome Core Particle Is Gated by the Rpt2 ATPase and Controls Both Substrate Entry and Product Release,” *Mol. Cell*, vol. 7, no. 6, pp. 1143–1152, Jun. 2001, doi: 10.1016/S1097-2765(01)00274-X.
[130] A. F. Kisselev, T. N. Akopian, K. M. Woo, and A. L. Goldberg, “The Sizes of Peptides Generated from Protein by Mammalian 26 and 20 S Proteasomes: IMPLICATIONS FOR UNDERSTANDING THE DEGRADATIVE MECHANISM AND ANTIGEN PRESENTATION*,” *J. Biol. Chem.*, vol. 274, no. 6, pp. 3363–3371, Feb. 1999, doi: 10.1074/jbc.274.6.3363.
[131] A. K. Nussbaum *et al.*, “Cleavage motifs of the yeast 20S proteasome subunits deduced from digests of enolase 1,” *Proc. Natl. Acad. Sci.*, vol. 95, no. 21, pp. 12504–12509, Oct. 1998, doi: 10.1073/pnas.95.21.12504.
[132] D. R. Lide, G. Baysinger, S. Chemistry, L. I. Berger, R. N. Goldberg, and H. V. Kehiaian, “CRC Handbook of Chemistry and Physics”.
[133] M. Häckel, H.-J. Hinz, and G. R. Hedwig, “Partial molar volumes of proteins: amino acid side-chain contributions derived from the partial molar volumes of some tripeptides over the temperature range 10–90°C,” *Biophys. Chem.*, vol. 82, no. 1, pp. 35–50, Nov. 1999, doi: 10.1016/S0301-4622(99)00104-0.
[134] C. Zhu *et al.*, “Characterizing hydrophobicity of amino acid side chains in a protein environment via measuring contact angle of a water nanodroplet on planar peptide network,” *Proc. Natl. Acad. Sci.*, vol. 113, no. 46, pp. 12946–12951, Nov. 2016, doi: 10.1073/pnas.1616138113.
[135] F. Fogolari *et al.*, “Distance-Based Configurational Entropy of Proteins from Molecular Dynamics Simulations,” *PLOS ONE*, vol. 10, no. 7, p. e0132356, Jul. 2015, doi: 10.1371/journal.pone.0132356.

[136] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.
[137] X. Robin *et al.*, “pROC: an open-source package for R and S+ to analyze and compare ROC curves,” *BMC Bioinformatics*, vol. 12, no. 1, p. 77, Mar. 2011, doi: 10.1186/1471-2105-12-77.
[138] A. Paszke *et al.*, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” p. 12.
[139] P. A. Ott *et al.*, “An immunogenic personal neoantigen vaccine for patients with melanoma,” *Nature*, vol. 547, no. 7662, Art. no. 7662, Jul. 2017, doi: 10.1038/nature22991.
[140] A.-M. Bjerregaard, M. Nielsen, S. R. Hadrup, Z. Szallasi, and A. C. Eklund, “MuPeXI: prediction of neo-epitopes from tumor sequencing data,” *Cancer Immunol. Immunother.*, vol. 66, no. 9, pp. 1123–1130, Sep. 2017, doi: 10.1007/s00262-017-2001-3.
[141] D. K. Wells *et al.*, “Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction,” *Cell*, vol. 183, no. 3, pp. 818-834.e13, Oct. 2020, doi: 10.1016/j.cell.2020.09.015.
[142] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.
[143] M. Groll *et al.*, “Structure of 20S proteasome from yeast at 2.4Å resolution,” *Nature*, vol. 386, no. 6624, Art. no. 6624, Apr. 1997, doi: 10.1038/386463a0.
[144] K. L. Rock, E. Reits, and J. Neefjes, “Present Yourself! By MHC Class I and MHC Class II Molecules,” *Trends Immunol.*, vol. 37, no. 11, pp. 724–737, Nov. 2016, doi: 10.1016/j.it.2016.08.010.
[145] S. Sarkizova *et al.*, “A large peptidome dataset improves HLA class I epitope prediction across most of the human population,” *Nat. Biotechnol.*, vol. 38, no. 2, pp. 199–209, Feb. 2020, doi: 10.1038/s41587-019-0322-9.
[146] M. Wieczorek *et al.*, “Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation,” *Front. Immunol.*, vol. 8, 2017, doi: 10.3389/fimmu.2017.00292.
[147] Y. Danilova, A. Voronkova, P. Sulimov, and A. Kertész-Farkas, “Bias in False Discovery Rate Estimation in Mass-Spectrometry-Based Peptide Identification,” *J. Proteome Res.*, vol. 18, no. 5, pp. 2354–2358, May 2019, doi: 10.1021/acs.jproteome.8b00991.
[148] Ö. Türeci, M. Vormehr, M. Diken, S. Kreiter, C. Huber, and U. Sahin, “Targeting the Heterogeneity of Cancer with Individualized Neoepitope Vaccines,” *Clin. Cancer Res.*, vol. 22, no. 8, pp. 1885–1896, Apr. 2016, doi: 10.1158/1078-0432.CCR-15-1509.
[149] G. Maruggi, C. Zhang, J. Li, J. B. Ulmer, and D. Yu, “mRNA as a Transformative Technology for Vaccine Development to Control Infectious Diseases,” *Mol. Ther.*, vol. 27, no. 4, pp. 757–772, Apr. 2019, doi: 10.1016/j.ymthe.2019.01.020.
[150] M. A. Smail, J. K. Reigle, and R. E. McCullumsmith, “Using protein turnover to expand the applications of transcriptomics,” *Sci. Rep.*, vol. 11, no. 1, Art. no. 1, Feb. 2021, doi: 10.1038/s41598-021-83886-7.
[151] B. Bibo-Verdugo, Z. Jiang, C. R. Caffrey, and A. J. O’Donoghue, “Targeting proteasomes in infectious organisms to combat disease,” *FEBS J.*, vol. 284, no. 10, pp. 1503–1517, 2017, doi: https://doi.org/10.1111/febs.14029.
[152] I. Coluzza, “Computational protein design: a review,” *J. Phys. Condens. Matter*, vol. 29, no. 14, p. 143001, Feb. 2017, doi: 10.1088/1361-648X/aa5c76.

[153] J. Hundal *et al.*, “pVACtools: A Computational Toolkit to Identify and Visualize Cancer Neoantigens,” *Cancer Immunol. Res.*, vol. 8, no. 3, pp. 409–420, Mar. 2020, doi: 10.1158/2326-6066.CIR-19-0401.
[154] E. W. Hewitt, “The MHC class I antigen presentation pathway: strategies for viral immune evasion,” *Immunology*, vol. 110, no. 2, pp. 163–169, Oct. 2003, doi: 10.1046/j.1365-2567.2003.01738.x.
[155] M. Y. Lee, J. W. Jeon, C. Sievers, and C. T. Allen, “Antigen processing and presentation in cancer immunotherapy,” *J. Immunother. Cancer*, vol. 8, no. 2, p. e001111, Aug. 2020, doi: 10.1136/jitc-2020-001111.
[156] D. B. Roth, “V(D)J Recombination: Mechanism, Errors, and Fidelity,” *Microbiol. Spectr.*, vol. 2, no. 6, p. 10.1128/microbiolspec.MDNA3-0041–2014, Dec. 2014, doi: 10.1128/microbiolspec.MDNA3-0041-2014.
[157] K. M. Ashby and K. A. Hogquist, “A guide to thymic selection of T cells,” *Nat. Rev. Immunol.*, pp. 1–15, Jul. 2023, doi: 10.1038/s41577-023-00911-8.
[158] D. B. Wilson *et al.*, “Specificity and degeneracy of T cells,” *Mol. Immunol.*, vol. 40, no. 14–15, pp. 1047–1055, Feb. 2004, doi: 10.1016/j.molimm.2003.11.022.
[159] R. B. Couch, J. A. Kasel, J. L. Gerin, J. L. Schulman, and E. D. Kilbourne, “Induction of Partial Immunity to Influenza by a Neuraminidase-specific Vaccine,” *J. Infect. Dis.*, vol. 129, no. 4, pp. 411–420, Apr. 1974, doi: 10.1093/infdis/129.4.411.
[160] G. Petrova, A. Ferrante, and J. Gorski, “Cross-Reactivity of T Cells and Its Role in the Immune System,” *Crit. Rev. Immunol.*, vol. 32, no. 4, pp. 349–372, 2012.
[161] D. V. Bagaev *et al.*, “VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium,” *Nucleic Acids Res.*, vol. 48, no. D1, pp. D1057–D1062, Jan. 2020, doi: 10.1093/nar/gkz874.
[162] W. Zhang *et al.*, “PIRD: Pan Immune Repertoire Database,” *Bioinformatics*, vol. 36, no. 3, pp. 897–903, Feb. 2020, doi: 10.1093/bioinformatics/btz614.
[163] H. M. Bijen *et al.*, “Preclinical Strategies to Identify Off-Target Toxicity of High-Affinity TCRs,” *Mol. Ther. J. Am. Soc. Gene Ther.*, vol. 26, no. 5, pp. 1206–1214, May 2018, doi: 10.1016/j.ymthe.2018.02.017.
[164] S. E. Blondelle, R. Moya-Castro, K. Osawa, K. Schroder, and D. B. Wilson, “Immunogenically optimized peptides derived from natural mutants of HIV CTL epitopes and peptide combinatorial libraries,” *Biopolymers*, vol. 90, no. 5, pp. 683–694, 2008, doi: 10.1002/bip.21020.
[165] B. R. Gundlach, K. H. Wiesmüller, T. Junt, S. Kienle, G. Jung, and P. Walden, “Specificity and degeneracy of minor histocompatibility antigen-specific MHC-restricted CTL,” *J. Immunol. Baltim. Md 1950*, vol. 156, no. 10, pp. 3645–3651, May 1996.
[166] B. R. Gundlach, K. H. Wiesmüller, T. Junt, S. Kienle, G. Jung, and P. Walden, “Determination of T cell epitopes with random peptide libraries,” *J. Immunol. Methods*, vol. 192, no. 1–2, pp. 149–155, Jun. 1996, doi: 10.1016/0022-1759(96)00040-3.
[167] T. Linnemann *et al.*, “Mimotopes for tumor-specific T lymphocytes in human cancer determined with combinatorial peptide libraries,” *Eur. J. Immunol.*, vol. 31, no. 1, pp. 156–165, Jan. 2001, doi: 10.1002/1521-4141(200101)31:1<156::aid-immu156>3.0.co;2-p.
[168] J. Lustgarten, A. L. Dominguez, and C. Pinilla, “Identification of cross-reactive peptides using combinatorial libraries circumvents tolerance against Her-2/neu-immunodominant epitope,” *J. Immunol. Baltim. Md 1950*, vol. 176, no. 3, pp. 1796–1805, Feb. 2006, doi: 10.4049/jimmunol.176.3.1796.

[169] R. H. McMahan, J. A. McWilliams, K. R. Jordan, S. W. Dow, D. B. Wilson, and J. E. Slansky, “Relating TCR-peptide-MHC affinity to immunogenicity for the design of tumor vaccines,” *J. Clin. Invest.*, vol. 116, no. 9, pp. 2543–2551, Sep. 2006, doi: 10.1172/JCI26936.
[170] C. Pinilla *et al.*, “Combinatorial peptide libraries as an alternative approach to the identification of ligands for tumor-reactive cytolytic T lymphocytes,” *Cancer Res.*, vol. 61, no. 13, pp. 5153–5160, Jul. 2001.
[171] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, and M. Linial, “ProteinBERT: a universal deep-learning model of protein sequence and function,” *Bioinformatics*, vol. 38, no. 8, pp. 2102–2110, Apr. 2022, doi: 10.1093/bioinformatics/btac020.
[172] A. T. Müller, J. A. Hiss, and G. Schneider, “Recurrent Neural Network Model for Constructive Peptide Design,” *J. Chem. Inf. Model.*, vol. 58, no. 2, pp. 472–479, Feb. 2018, doi: 10.1021/acs.jcim.7b00414.
[173] S. A. Ghanekar, L. E. Nomura, M. A. Suni, L. J. Picker, H. T. Maecker, and V. C. Maino, “Gamma Interferon Expression in CD8+ T Cells Is a Marker for Circulating Cytotoxic T Lymphocytes That Recognize an HLA A2-Restricted Epitope of Human Cytomegalovirus Phosphoprotein pp65,” *Clin. Diagn. Lab. Immunol.*, vol. 8, no. 3, pp. 628–631, May 2001, doi: 10.1128/CDLI.8.3.628-631.2001.
[174] K. T. Coppieters and M. von Herrath, “Antibody cross-reactivity and the viral aetiology of type 1 diabetes,” *J. Pathol.*, vol. 230, no. 1, pp. 1–3, May 2013, doi: 10.1002/path.4174.
[175] S. Swaminathan *et al.*, “Prevalence and pattern of cross-reacting antibodies to HIV in patients with tuberculosis,” *AIDS Res. Hum. Retroviruses*, vol. 24, no. 7, pp. 941–946, Jul. 2008, doi: 10.1089/aid.2007.0211.
[176] E. M. Eisenstein and C. B. Williams, “The Treg/Th17 Cell Balance: A New Paradigm for Autoimmunity,” *Pediatr. Res.*, vol. 65, no. 7, Art. no. 7, May 2009, doi: 10.1203/PDR.0b013e31819e76c7.
[177] A.-K. Hopp, A. Rupp, and V. Lukacs-Kornek, “Self-Antigen Presentation by Dendritic Cells in Autoimmunity,” *Front. Immunol.*, vol. 5, 2014, Accessed: Oct. 17, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fimmu.2014.00055
[178] T. Karantanos, P. G. Corn, and T. C. Thompson, “Prostate cancer progression after androgen deprivation therapy: mechanisms of castrate-resistance and novel therapeutic approaches,” *Oncogene*, vol. 32, no. 49, pp. 5501–5511, Dec. 2013, doi: 10.1038/onc.2013.206.
[179] I. Wang, L. Song, B. Y. Wang, A. Rezazadeh Kalebasty, E. Uchio, and X. Zi, “Prostate cancer immunotherapy: a review of recent advancements with novel treatment methods and efficacy,” *Am. J. Clin. Exp. Urol.*, vol. 10, no. 4, pp. 210–233, Aug. 2022.
[180] C. Han *et al.*, “The Roles of Tumor-Associated Macrophages in Prostate Cancer,” *J. Oncol.*, vol. 2022, p. 8580043, Sep. 2022, doi: 10.1155/2022/8580043.
[181] P. Xu *et al.*, “Androgen receptor blockade resistance with enzalutamide in prostate cancer results in immunosuppressive alterations in the tumor immune microenvironment,” *J. Immunother. Cancer*, vol. 11, no. 5, p. e006581, May 2023, doi: 10.1136/jitc-2022-006581.
[182] Y. Jiang, Y. Li, and B. Zhu, “T-cell exhaustion in the tumor microenvironment,” *Cell Death Dis.*, vol. 6, no. 6, Art. no. 6, Jun. 2015, doi: 10.1038/cddis.2015.162.
[183] X. Guan *et al.*, “Androgen receptor activity in T cells limits checkpoint blockade efficacy,” *Nature*, pp. 1–6, Mar. 2022, doi: 10.1038/s41586-022-04522-6.

[184] A. C. Roden *et al.*, “Augmentation of T Cell Levels and Responses Induced by Androgen Deprivation1,” *J. Immunol.*, vol. 173, no. 10, pp. 6098–6108, Nov. 2004, doi: 10.4049/jimmunol.173.10.6098.
[185] P. Yuri *et al.*, “Increased tumor-associated macrophages in the prostate cancer microenvironment predicted patients’ survival and responses to androgen deprivation therapies in Indonesian patients cohort,” *Prostate Int.*, vol. 8, no. 2, pp. 62–69, Jun. 2020, doi: 10.1016/j.prnil.2019.12.001.
[186] M. Mercader *et al.*, “Early effects of pharmacological androgen deprivation in human prostate cancer,” *BJU Int.*, vol. 99, no. 1, pp. 60–67, 2007, doi: 10.1111/j.1464-410X.2007.06538.x.
[187] A. Erlandsson *et al.*, “Infiltrating immune cells in prostate cancer tissue after androgen deprivation and radiotherapy,” *Int. J. Immunopathol. Pharmacol.*, vol. 37, p. 03946320231158025, Dec. 2023, doi: 10.1177/03946320231158025.
[188] M. Mercader *et al.*, “T cell infiltration of the prostate induced by androgen withdrawal in patients with prostate cancer,” *Proc. Natl. Acad. Sci.*, vol. 98, no. 25, pp. 14565–14570, Dec. 2001, doi: 10.1073/pnas.251140998.
[189] D. T. Montoro *et al.*, “A revised airway epithelial hierarchy includes CFTR-expressing ionocytes,” *Nature*, vol. 560, no. 7718, pp. 319–324, Aug. 2018, doi: 10.1038/s41586-018-0393-7.
[190] G. H. Henry *et al.*, “A Cellular Anatomy of the Normal Adult Human Prostate and Prostatic Urethra,” *Cell Rep.*, vol. 25, no. 12, pp. 3530-3542.e5, Dec. 2018, doi: 10.1016/j.celrep.2018.11.086.
[191] T. Hirz *et al.*, “Dissecting the immune suppressive human prostate tumor microenvironment via integrated single-cell and spatial transcriptomic analyses,” *Nat. Commun.*, vol. 14, no. 1, Art. no. 1, Feb. 2023, doi: 10.1038/s41467-023-36325-2.
[192] A. Ardiani, S. R. Gameiro, A. R. Kwilas, R. N. Donahue, and J. W. Hodge, “Androgen deprivation therapy sensitizes prostate cancer cells to T-cell killing through androgen receptor dependent modulation of the apoptotic pathway,” *Oncotarget*, vol. 5, no. 19, pp. 9335–9348, Sep. 2014.
[193] M. Gillespie *et al.*, “The reactome pathway knowledgebase 2022,” *Nucleic Acids Res.*, vol. 50, no. D1, pp. D687–D692, Jan. 2022, doi: 10.1093/nar/gkab1028.
[194] L. N. Chesner *et al.*, “AR inhibition increases MHC Class I expression and improves immune response in prostate cancer,” *Rev.*.
[195] N. H. Bander *et al.*, “MHC class I and II expression in prostate carcinoma and modulation by interferon-alpha and -gamma,” *The Prostate*, vol. 33, no. 4, pp. 233–239, 1997, doi: 10.1002/(SICI)1097-0045(19971201)33:4<233::AID-PROS2>3.0.CO;2-I.
[196] N. Calistri, “Paclitaxel phenocopies interferon response and alters cell cycle dynamics in triple negative breast cancer,” *Prep.*.
[197] Y. He *et al.*, “FOXA1 overexpression suppresses interferon signaling and immune response in cancer,” *J. Clin. Invest.*, vol. 131, no. 14, p. e147025, doi: 10.1172/JCI147025.
[198] D. L. Barber *et al.*, “Restoring function in exhausted CD8 T cells during chronic viral infection,” *Nature*, vol. 439, no. 7077, Art. no. 7077, Feb. 2006, doi: 10.1038/nature04444.
[199] C. R. Consiglio, O. Udartseva, K. D. Ramsey, C. Bush, and S. O. Gollnick, “Enzalutamide, an Androgen Receptor Antagonist, Enhances Myeloid Cell–Mediated Immune Suppression and Tumor Progression,” *Cancer Immunol. Res.*, vol. 8, no. 9, pp. 1215–1227, Sep. 2020, doi: 10.1158/2326-6066.CIR-19-0371.

[200] R. Lugano, M. Ramachandran, and A. Dimberg, “Tumor angiogenesis: causes, consequences, challenges and opportunities,” *Cell. Mol. Life Sci.*, vol. 77, no. 9, pp. 1745–1770, 2020, doi: 10.1007/s00018-019-03351-7.
[201] R. J. A. van Moorselaar and E. E. Voest, “Angiogenesis in prostate cancer: its role in disease progression and possible therapeutic approaches,” *Mol. Cell. Endocrinol.*, vol. 197, no. 1–2, pp. 239–250, Nov. 2002, doi: 10.1016/s0303-7207(02)00262-9.
[202] K. S. Sfanos and A. M. De Marzo, “Prostate cancer and inflammation: the evidence,” *Histopathology*, vol. 60, no. 1, pp. 199–215, Jan. 2012, doi: 10.1111/j.1365-2559.2011.04033.x.
[203] F. R. Greten and S. I. Grivennikov, “Inflammation and Cancer: Triggers, Mechanisms and Consequences,” *Immunity*, vol. 51, no. 1, pp. 27–41, Jul. 2019, doi: 10.1016/j.immuni.2019.06.025.
[204] Y. Zhang *et al.*, “Androgen deprivation promotes neuroendocrine differentiation and angiogenesis through CREB-EZH2-TSP1 pathway in prostate cancers,” *Nat. Commun.*, vol. 9, no. 1, p. 4080, Oct. 2018, doi: 10.1038/s41467-018-06177-2.
[205] Z. Melegh and S. Oltean, “Targeting Angiogenesis in Prostate Cancer,” *Int. J. Mol. Sci.*, vol. 20, no. 11, Art. no. 11, Jan. 2019, doi: 10.3390/ijms20112676.
[206] K. S. Sfanos, S. Yegnasubramanian, W. G. Nelson, and A. M. De Marzo, “The inflammatory microenvironment and microbiome in prostate cancer development,” *Nat. Rev. Urol.*, vol. 15, no. 1, Art. no. 1, Jan. 2018, doi: 10.1038/nrurol.2017.167.
[207] G. Deep and G. K. Panigrahi, “Hypoxia-Induced Signaling Promotes Prostate Cancer Progression: Exosomes Role as Messenger of Hypoxic Response in Tumor Microenvironment,” *Crit. Rev. Oncog.*, vol. 20, no. 5–6, pp. 419–434, 2015, doi: 10.1615/CritRevOncog.v20.i5-6.130.
[208] K. Deonarine *et al.*, “Gene expression profiling of cutaneous wound healing,” *J. Transl. Med.*, vol. 5, p. 11, Feb. 2007, doi: 10.1186/1479-5876-5-11.
[209] S. Jin *et al.*, “Inference and analysis of cell-cell communication using CellChat,” *Nat. Commun.*, vol. 12, no. 1, p. 1088, Feb. 2021, doi: 10.1038/s41467-021-21246-9.
[210] C. Bolitho, M. Moscova, R. C. Baxter, and D. J. Marsh, “Amphiregulin increases migration and proliferation of epithelial ovarian cancer cells by inducing its own expression via PI3-kinase signaling,” *Mol. Cell. Endocrinol.*, vol. 533, p. 111338, Aug. 2021, doi: 10.1016/j.mce.2021.111338.
[211] S. Wang *et al.*, “Amphiregulin Confers Regulatory T Cell Suppressive Function and Tumor Invasion via the EGFR/GSK-3β/Foxp3 Axis,” *J. Biol. Chem.*, vol. 291, no. 40, pp. 21085–21095, Sep. 2016, doi: 10.1074/jbc.M116.717892.
[212] A. Imada, N. Shijubo, H. Kojima, and S. Abe, “Mast cells correlate with angiogenesis and poor outcome in stage I lung adenocarcinoma,” *Eur. Respir. J.*, vol. 15, no. 6, pp. 1087–1093, Jun. 2000, doi: 10.1034/j.1399-3003.2000.01517.x.
[213] M. Wroblewski *et al.*, “Mast cells decrease efficacy of anti-angiogenic therapy by secreting matrix-degrading granzyme B,” *Nat. Commun.*, vol. 8, no. 1, Art. no. 1, Aug. 2017, doi: 10.1038/s41467-017-00327-8.
[214] M. Shibuya, “Vascular Endothelial Growth Factor (VEGF) and Its Receptor (VEGFR) Signaling in Angiogenesis,” *Genes Cancer*, vol. 2, no. 12, pp. 1097–1105, Dec. 2011, doi: 10.1177/1947601911423031.
[215] D. M. W. Zaiss, W. C. Gause, L. C. Osborne, and D. Artis, “Emerging Functions of Amphiregulin in Orchestrating Immunity, Inflammation, and Tissue Repair,” *Immunity*, vol. 42, no. 2, pp. 216–226, Feb. 2015, doi: 10.1016/j.immuni.2015.01.020.

[216] T. Stuart *et al.*, “Comprehensive Integration of Single-Cell Data,” *Cell*, vol. 177, no. 7, pp. 1888-1902.e21, Jun. 2019, doi: 10.1016/j.cell.2019.05.031.
[217] M. D. Young and S. Behjati, “SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data,” *GigaScience*, vol. 9, no. 12, p. giaa151, Nov. 2020, doi: 10.1093/gigascience/giaa151.
[218] C. S. McGinnis, L. M. Murrow, and Z. J. Gartner, “DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors,” *Cell Syst.*, vol. 8, no. 4, pp. 329-337.e4, Apr. 2019, doi: 10.1016/j.cels.2019.03.003.
[219] I. Korsunsky *et al.*, “Fast, sensitive and accurate integration of single-cell data with Harmony,” *Nat. Methods*, vol. 16, no. 12, Art. no. 12, Dec. 2019, doi: 10.1038/s41592-019-0619-0.
[220] H. Song *et al.*, “Single-cell analysis of human primary prostate cancer reveals the heterogeneity of tumor-associated epithelial cell states,” *Nat. Commun.*, vol. 13, no. 1, Art. no. 1, Jan. 2022, doi: 10.1038/s41467-021-27322-4.