Inference of T cell antigen specificity with yeast peptide-MHC display libraries, statistics and machine learning



Alexander Obolensky Hudson

St. Hugh's College
Fernandes Group
CAMS Oxford Institute

Supervisors:
Ricardo A. Fernandes, Charlotte Deane, Mark Basham

Thesis submitted for the degree of Doctor of Philosophy
Michaelmas Term, 2025

Summary

Thymic (T) lymphocytes play a fundamental role in the ability of jawed vertebrates to recognise and respond to molecular signals of infection, tissue damage, and malignancy (*antigens*). These recognition events hinge on the interaction between T cell receptors (TCRs) and antigens presented by proteins of the Major Histocompatibility Complex (MHC). A reliable computational mapping from the nucleotide or polypeptide sequence of a TCR to its corresponding antigenic targets could enable rapid isolation of new immune biomarkers of disease, characterisation of the drivers of vaccine response, and development of new, safer TCR therapeutics. Accurate prediction of TCR specificity is therefore regarded as a "holy grail" of quantitative immunobiology.

In this DPhil project, we first review the state-of-the-art in predicting T cell specificity, identifying gaps in the availability of representative data, failure to account for the underlying complexity of the TCR-antigen interface, and methodological limitations as drivers of historically poor computational performance. Next, we turn to clustering models, comparing the performance of algorithms that group TCRs on sequence similarity as a means of inferring specificity. Here, too, we find significant room for improvement: performance being highly sensitive to the dataset, the TCR sequences provided per antigen, and the antigen of interest.

Having thus identified fundamental limitations in the ability of computational models to predict specificity for any given TCR-antigen pair, we harness deep yeast peptide-MHC display libraries to make high-confidence predictions of antigen specificity for a specific receptor of interest. Applying a simple statistical model to screens of billions of random synthetic peptides, the resulting pipeline accurately retrieves the targets of engineered and naturally occurring immune receptors, performing robustly across receptors and MHC types. The system performs well "off-the-shelf" with no receptor-specific parameter optimisation. By incorporating affinity data from yeast display and T cell stimulation data from co-culture experiments, we find that the pipeline can be applied in cycles of computational search and experimental validation. In this way, novel "self" antigens are discovered for a well-studied therapeutic TCR. Finally, we conceive of new measures of promiscuity that enable prioritisation of candidate therapeutic TCRs, facilitating the discovery of a new potential treatment for melanoma, and demonstrate the applicability of the pipeline to characterisation methods beyond yeast display.

This research provides novel insights into the limitations of computational inference of receptor specificity, generates a suite of tools for immune receptor characterisation, supports the identification of a novel TCR therapeutic candidate, and illustrates the power of integrating diverse data sources for the characterisation of protein-protein interactions.

Acknowledgements

This research was made possible by support from the Oxford BBSRC Interdisciplinary Bioscience Doctoral Training Programme (DTP); the Fernandes group at the CAMS Oxford Institute; the Oxford Protein Informatics Group (OPIG); the Basham group at the Rosalind Franklin Institute, and the MRC Weatherall Institute of Molecular Medicine. The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141.Z.16.Z and the NIHR Oxford Biomedical Research Cluster.

I would like to thank my supervisory team for their support, pastoral care, and advice throughout the project: Drs Ricardo A. Fernandes and Mark Basham, Professors Charlotte M. Deane and Esther Becker, and Drs Matthew Raybould and Alex Lubbock. My profound gratitude is also due to the experimental teams and collaborators who so kindly shared their data and advice, provided invaluable immunological context, and suggested novel features and applications for the models developed: to Dr Julia McCarthy, Kristen Koopmans, and Yunkai Yang of the Fernandes group; to Professor Michael Birnbaum at the Massachusetts Institute of Technology; Dr Xinbo Yang at Memorial Sloane Kettering; Professor Omer Dushek, Anna Huhn, and Jose C. Caballero at the Sir William Dunn School of Pathology; and Drs Stephen Harper and Paula Dobrinic of Immunocore, Ltd.

Productive research blossoms in a nurturing, creative environment, and I would like to thank the Nuffield Department of Medicine, the CAMS Oxford Institute and the Fernandes group for providing a home-from-home for the past three years: Professor Tao Dong; Drs Christina Heroven; Victoria Junghans; Caitlin O'Brien-Ball and Martin Fellermeyer; Khuluud Hussein; Raul Cioaca; Mai Vuong; Xiaonan Zheng; Yu (Hugo) Yan; and Sam Farrar in particular. Thanks are also due to Drs Alex Greenshields-Watson and Alissa Hummer; and to Nele Quast and Lewis Chinery of OPIG, among too many others to name, for their advice over the years.

Professors Simon Davis, Max Crispin, and Lynne Cox were responsible for my foundational training: my thanks to them for inspiring a life in science, and for encouraging my return to academia after a ten-year hiatus in industry.

Lastly, I would not have been able to complete this work without the support and understanding of my family. My love and thanks to my parents and to my brother Nick for their enduring patience and care. To my wife Jasmin in particular, who forsook the sunny hills of Tuscany for dreary winters in the Thames valley: thank you.

Table of Contents

Summary ***2***

Acknowledgements ***3***

Abbreviations ***7***

Chapter 1: Introduction ***8***

T cells in human immunity **8**

An overview of the human immune system 8

T lymphocytes form a central part of the adaptive immune system 8

TCR diversity is greatest in its antigen-binding loops 9

T cells recognise antigen in the context of the Major Histocompatibility Complex 10

Coreceptors influence T cell restriction, sensitivity and effector function 11

T cells undergo positive and negative selection during development 12

T cell activation depends on effective antigen presentation by APCs 13

Activated naïve cells differentiate into effector and memory T lymphocytes 14

Characterisation and prediction of TCR specificity **14**

The motivation for TCR de-orphanisation 14

Methods in TCR de-orphanisation 15

Key challenges in computational TCR de-orphanisation 25

Chapter 2: A comparison of clustering models for inference of TCR antigen specificity ***29***

Overview **29**

Introduction **29**

Methods **31**

Datasets 31

Models 32

Metrics 35

Statistics 35

Results **36**

Discussion **45**

Limitations and further work **46**

Chapter acknowledgements **46**

Chapter 3: Statistics and machine learning for inference of T cell antigen specificity from yeast peptide-MHC display libraries ***47***

Overview **47**

Introduction **47**

Methods **50**

Datasets 50

Models 51

Experimental assays 53

Integrating binding and stimulation data 54

Results **54**

Discussion **66**

Limitations and further work **67**
Chapter acknowledgements **68**
Chapter 4: Quantification of cross-reactivity for prioritisation of TCR therapeutic variants **69**
Overview **69**
Introduction **69**
Methods **71**
Datasets 71
Fastq2pep 72
Visualisation of pMHC screen data 73
Cross-reactivity measures 73
Graph network clustering 74
Results **75**
Discussion **88**
Limitations and further work **90**
Chapter acknowledgements **91**
Chapter 5: Deconstructing the predictive power of binding and functional data in TCR-directed pMHC screens **92**
Overview **92**
Introduction **92**
Methods **93**
Datasets 93
Heat maps 95
Predicted hits 95
Results **96**
Discussion **108**
Limitations and further work **109**
Chapter acknowledgements **109**
Chapter 6: Discussion **110**
Generalisable inference of TCR specificity remains unsolved **110**
Improving modelling of the underlying immunology 110
Harnessing new data sources 111
Exploiting advances in structure prediction 112
Simple sequence-driven clustering outperforms more complex models **112**
The Orpheus pipeline performs well in varied contexts **113**
Multi-modal data approaches boost performance of the predictor 114
We propose novel metrics of cross-reactivity for therapeutic candidate prioritisation 115
Limitations and further work **115**
Limitations of yeast display for inference of specificity 115
Limitations in modelling approach 118

Further work..... 119
Bibliography..... 121

Abbreviations

Abbreviation	Meaning
APC	Antigen presenting cell
BCR	B cell receptor
CAR	Chimaeric antigen receptor
CDR	Complementarity Determining Region
CR50	Cumulative rank at 50% saturation
DC	Dendritic cells
DNN	Deep neural network
GPE	Global positional entropy
HLA	Human Leukocyte Antigen
IL	Interleukin
LPE	Local positional entropy
pMHC	Peptide-MHC
MHC	Major Histocompatibility Complex
NK	Natural killer
PWM	Positional Weight Matrix
SPM	Supervised predictive model
TCR	T cell receptor
TCR-T	TCR therapeutic
tSNE	Triplet stochastic neighbourhood embedding
UCM	Unsupervised clustering model
UMAP	Uniform Manifold Approximation and Projection
XSM	X-scanning mutagenesis

Chapter 1: Introduction

T cells in human immunity

An overview of the human immune system

We are not alone. Each surface and compartment of the human body contains a host of thousands of foreign organisms. Many of these play a beneficial role in our ability to derive nutrients from the outside world. However, the relationship between host and microorganism is not always mutually advantageous. Diseases caused by pathogenic microorganisms (*pathogens*) were a leading cause of global mortality prior to the advent of waste sanitation, vaccines, and antibiotics. Their impact can still be felt in the burden of infectious disease across economically developing nations, and worldwide in a modern era of emerging pandemic threats. In a prolonged evolutionary arms race, jawed vertebrates have evolved a complex and dynamic *immune system* to counter these continued threats.

The human immune system is made up of two densely interconnected branches: those of *innate* and *adaptive* immunity. The innate immune system comprises anatomic barriers including the skin and related biochemical defence systems; cellular sentinels including macrophages, granulocytes, and dendritic cells (DC; together, the *phagocytes*); the protein effectors of the complement system; and innate-like agents including natural killer (NK) cells. The biochemical and cellular components of the innate system serve to respond rapidly to conserved signals of infection, cancer, and tissue damage. However, whilst this response may be mounted within minutes or hours of detection, the innate immune system broadly lacks the capacity to learn from and recognise novel signals of harm, and to modulate its intensity in response to repeated exposure. In contrast, the "T" (thymus-dependent) and "B" (bone marrow) lymphocytes of the adaptive immune system can mount a highly specific response to a given molecular signal (*antigen*), whether it has been encountered before or not, in a manner whose strength increases with repeated exposure.

T lymphocytes form a central part of the adaptive immune system

Under considerable selective pressure to do so, many pathogens have evolved mechanisms to evade the structural, protein, biochemical, and cellular components of the innate immune system. Among these mechanisms, error-prone replication enables the emergence of novel protein variants that avoid immune detection whilst preserving function, for example host cell entry. A pertinent recent example is the "spike" glycoprotein of the SARS-CoV-2 betacoronavirus, the mutation of which was observed in successive variants of concern during

the 2020-2023 global COVID-19 pandemic (1). A competent immune system must, therefore, be able to recognise and respond to varied and novel signals of pathogenesis. This is largely achieved by recognising antigens through lymphocyte antigen receptors, the *T cell receptors* (*TCRs*), *B cell receptors* (*BCRs*), and *antibodies* - soluble relatives of the BCR. Patients unable to produce antigen receptors, as a result of defects in the cellular recombination machinery, develop severe combined immunodeficiency (SCID): a condition characterised by increased susceptibility to infection (2). In the present work, we focus on human TCRs, drawing comparison to the structurally related but functionally distinct BCR where relevant to the mechanisms of T lymphocyte development and TCR antigen specificity.

TCR diversity is greatest in its antigen-binding loops

The TCR is a membrane-bound glycoprotein heterodimer of α and β or γ and δ chains, each containing membrane-distal variable (V) and membrane-proximal constant (C) domains of the immunoglobulin superfamily, anchored to the cell surface via closely packed transmembrane alpha helices (**Fig. 1-1**) (3-5). TCR signalling depends on interaction with CD3, the disulphide-linked TCRαβ heterodimer associating in a 1:1:1:1 ratio with CD3-εγ, -εδ, and -ζζ dimers to form the ternary TCR complex (5, 6). An individual T cell expresses approximately 10^5 copies of a single TCR (its *clonotype*), with as few as 3-400 TCRs being required for robust T cell activation (7). T lymphocytes expressing αβ TCRs are more common than γδ T cells in humans, and make up 90-95% of the circulating population in healthy adults (8, 9). $\gamma\delta^+$ T lymphocyte cells play distinct and varied immunological roles, including in defence against mycobacteria, preventing allergic sensitisation to environmental antigens, and activation of naïve B cells (10).

TCRs bind their cognate antigens through V regions, productive TCR-antigen engagement being commonly mediated through hypervariable loops known as the complementarity determining regions (CDRs) 1-3 (11, 12) (**Fig. 1-1b**). CDRs 1 and 2 are the products of highly polymorphic V genes at TRA and TRB gene loci on chromosomes 14 and 7 in humans, respectively. The random somatic combination of V, diversity (D, for β but not α chains), junctional (J), and constant (C) regions during T lymphocyte development generates a combinatorial diversity of over 10^{15} possible receptors (13). This process, mediated by RAG1/RAG2 recombinases, is known as *VDJ recombination* (14).

Unlike heritable CDR1 and CDR2 segments, CDR3 loops are generated through the random addition and deletion of nucleotides at the junctions between exons encoded by V, D, and J genes. Consequently, CDR3 loops vary in length: a recent study of all known paired chain TCRs finding a median CDR3 length of 12 amino acids for both α and β chains (15). This supplementary junctional diversity increases the possible TCR repertoire, however the effective number is narrowed by the action of positive and negative selection in the thymus to produce the 10^6-10^8 unique TCRs found circulating in a healthy adult (16).

T cells recognise antigen in the context of the Major Histocompatibility Complex

Whereas antibodies may bind to antigenic peptides, whole proteins, carbohydrates or even lipid antigens, and can engage directly with either linear or non-contiguous patches (*epitopes*) on a target protein, TCRs are typically more constrained. αβ TCRs most commonly recognise protein fragments (*peptides*) presented in the context of the Major Histocompatibility Complex (MHC) (**Fig. 1-1b**) following proteasomal cleavage (*antigen processing*), loading onto MHC, and display (*presentation*) by antigen presenting cells (APCs). αβ T cell antigens are typically short peptides of eight or more residues, the presentation of which is dictated in large part by the structural preferences of the MHC allele (17).

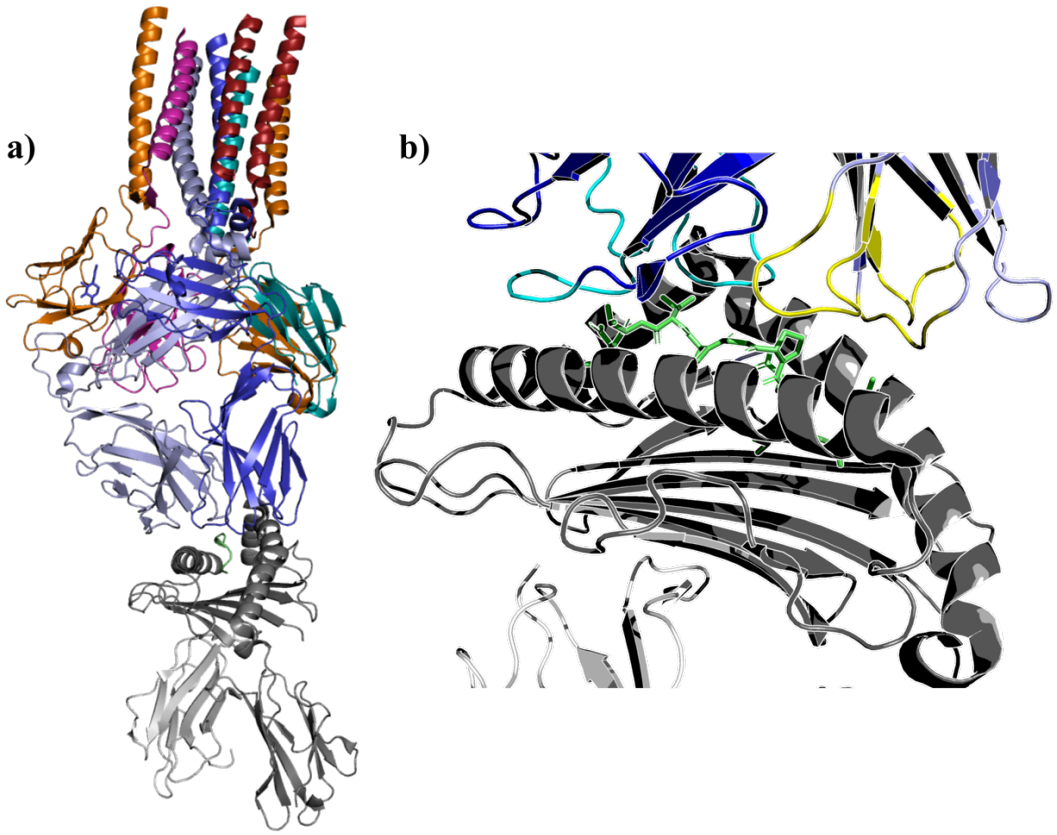


Figure 1-1: Structure of the αβ TCR. a | Three-dimensional structure of the GPa3b17-gp100-HLA-A*02:01 TCR-CD3-peptide-MHC ternary complex. MHC domains $\alpha_{1\text{-}3}$ are shown in dark grey and β2 microglobulin in white, TCR α and β chains are shown in light and dark blue, gp100 in lime; and CD3-δ, -ε, -γ and -ζ chains in turquoise, orange, pink and red, respectively. **b |** CDR loops from TCR α (light blue, CDRs in yellow) and β (dark blue, CDRs in cyan) chains engaging with gp100:HLA-A*02:01. Figure produced with PyMol from RCS PDB 7PHR (5, 18).

Over 40,000 human MHC alleles have been identified to date (19), each having distinct preferences for peptides of a given length and carrying *anchor* residues through which the peptide interacts with the MHC peptide binding groove (20). MHC can be broadly ascribed to one of two major classes, referred to in humans as the human leukocyte antigens (HLA). Classical class I HLA are encoded by the HLA-A, B, and C gene loci and consist of α_1, α_2 and α_3 domains noncovalently associated with β2-microglobulin. These present short (8-10) residue peptides produced from intracellular proteins, which are loaded onto the complex in the endoplasmic reticulum following proteasomal cleavage in the cytosol. Presentation of cytosolic peptides, including viral antigens, represents a critical immune surveillance mechanism that underlies expression of classical class I MHC in most nucleated cells (21). Nonclassical class I HLA types include HLA-E, F, G and H, as well as the MR1 and CD1 proteins which present derivatives of riboflavin metabolism and lipid antigens, respectively (17, 22, 23). Class II MHC (HLA-DR, DP, DQ and the non-classical DM and DO) present longer fragments of extracellular proteins (peptides of 13 or more residues) via an open peptide binding groove formed between the two alpha helices of membrane distal α_1 and β_1 domains (24, 25). For the sake of variety, hereafter we shall refer to antigen-MHC and peptide-MHC (pMHC) interchangeably where appropriate, however processing and presentation of peptides is currently better understood than non-peptide antigens.

TCR interaction with the antigen-MHC complex is structurally constrained, such that the CDR1 and 2 loops commonly interact with MHC in a diagonal binding mode that positions the CDR3 loops over the peptide binding groove: a study of the solved crystal structures of 60 non-redundant TCR-antigen-MHC complexes identified a constrained docking topology with a mean rotation of 46° (26). Whilst alternative docking geometries have been reported, it is still unclear whether TCRs binding in a less conventional mode are capable of provoking functional T cell responses (27). γδ TCR recognition is less well understood, with reported antigen specificities spanning from lipid antigens presented by the non-classical MHC CD1d, to fully MHC-independent binding (10).

Coreceptors influence T cell restriction, sensitivity and effector function

Mature T cells express the transmembrane coreceptors CD4 and CD8 in a manner closely coupled to the effector phenotype, MHC restriction, and antigen specificity of the cell (28). Expression of these proteins greatly impacts antigen sensitivity, coreceptor-positive T cells being capable of recognising a single antigen-MHC complex (29-31). CD8-positive ($CD8^+$) cytotoxic T cells respond to cells expressing antigen presented by class I MHC, the αβ CD8 heterodimer engaging with the α_3 subunit of MHC I (32). Productive pMHC engagement by $CD8^+$ T cells provokes direct killing of cells bearing cognate antigen through release of cytolytic granules or the Fas/FasL apoptotic pathway, and indirectly through the release of tumour necrosis factor (TNF)α and interferon (IFN)γ (33-36). $CD8_{\alpha\alpha}$ homodimers are

encountered in innate-like T lymphocytes including NK cells (37) and mucosa-associated invariant T (MAIT) cells (38) that recognise derivatives of riboflavin metabolism presented in the context of MR1 (22, 39). $CD4^+$ T cells detect antigens presented by MHC class II, the CD4 monomer binding to the interface of the α and β subunits with an affinity approaching 2mM (compared with 50-200 μM for CD8 to MHC class I), one of the weakest known interactions of the immunological synapse (40). Class II MHC are commonly expressed by professional APCs such as DC, B cells, and macrophages in the secondary lymphoid tissues, and by the epithelial cells of the thymus. Activated $CD4^+$ class II-restricted T cells may differentiate to form helper T cells including subsets Th_1, Th_2, or Th_{17}, which direct the activation of other immune effectors including DC and B cells through the release of subset-specific mediators (41).

T cells undergo positive and negative selection during development

T lymphocytes originate in the bone marrow as common lymphoid progenitor cells, migrating to the thymus where they undergo differentiation into thymocytes under the direction of interleukin (IL) and Notch signalling from the thymic epithelium (42, 43). These early *double negative* (DN) thymocytes express CD3 but lack coreceptors CD4 and CD8, entering the thymus at the cortico-medullary junction (44). As DN thymocytes migrate from entry venules deeper via the into the thymic cortex, interaction with cells of the cortical stroma promotes VDJ recombination at the TRB gene locus (45, 46). Successful expression of the pre-TCR complex by DN lymphocytes triggers MHC-independent intracellular signalling pathways, expression of both CD4 and CD8 coreceptors, proliferation and differentiation into *double positive* (DP) thymocytes.

DP thymocytes migrate through the densely connected thymic epithelial cortex (TEC), whose cells express both class I and class II MHC (47). DP thymocytes undergo iterative VDJ recombination at the TRA gene locus, and are subjected to the opposing forces of positive and negative selection, reviewed in (48). Only those cells expressing a productive TCR capable of binding class I or class II MHC are rescued from anergic cell death, in a manner driven by coreceptor-dependent TCR signalling. In this way, T cell maturation in the thymus imposes MHC restriction on the T lymphocyte population. Recognition of non-self MHC can provoke an immune reaction to grafted organs (*alloreactivity*), leading to rejection (49) or, in the case of the allogeneic lymphocytes transferred from the donor organ, to graft versus host disease (50). A small proportion of DP thymocytes receiving a survival signal retain expression of either CD4 or CD8, differentiating to become single positive ($CD4^+$/$CD8^+$) mature T cells and migrating to the thymic medulla.

Thymic selection also imposes an important negative constraint, namely that mature T cells should not recognise harmless “self” antigens. Medullary epithelial cells can express *ectopic*

antigens typically encountered in specific organs, under the direction of the Aire promoter (51). T lymphocytes recognising self-pMHC in the thymus are either removed from the population by apoptosis, or differentiate to become T regulatory cells (T_{regs}) that may play an important role in inflammatory homeostasis (52). Whilst the factors influencing positive selection (rescue from anergy) and negative selection (deletion) remain an active area of research, it is thought that the affinity of the TCR for host antigen-MHC complexes is key (48).

Mature MHC-restricted T cells surviving anergy or deletion in the thymus enter the blood stream as *naïve* T lymphocytes. However, the selective processes governing T cells are not restricted to the thymus. T cells encountering self antigen presented by tolerogenic DC in the lymph nodes or short-lived tertiary lymphoid tissues may also undergo anergy or apoptotic deletion (53).

T cell activation depends on effective antigen presentation by APCs

Naïve T cells recirculate between the blood, lymph nodes, spleen, and peripheral lymphoid tissues by means of the lymphatic system. This allows them to sample antigen-MHC complexes presented by APCs such as macrophages, DC and B lymphocytes. DC and macrophages are examples of phagocytic cells, innate cells that are able to take up extracellular pathogens, virally infected or damaged host cells from the extracellular milieu either through the action of complement, Fc and C-type lectin receptors (*phagocytosis*) or via nonspecific *macropinocytosis* (54). Engulfed pathogens are digested by powerful intracellular reagents including lysozyme, endocytic granules and reactive oxygen species present in intracellular compartments. The resulting antigenic fragments may then be loaded onto class II MHC in endocytic vesicles, and exported to the cell surface for presentation to $CD4^+$ T lymphocytes. As with many other cell types, phagocytes may be directly infected by intracellular pathogens such as viruses and so present viral-derived cytosolic antigens directly through class I MHC to $CD8^+$ T cells. Importantly, phagocytes may also take up virally infected host cells and so present viral peptides present in other cell types: a process known as *cross-priming*.

Dendritic cells in particular play a central role to priming naïve lymphocytes, tissue resident DC becoming activated through pattern recognition receptors that respond to signals of infection or tissue damage, and/or by pro-inflammatory cytokines or chemokines present at the site of inflammation. Activated DC migrate to the T cell zones of the lymphoid tissues, presenting long-lived pMHC complexes along with costimulatory receptor ligands CD80 and CD86 that are upregulated along with molecules implicated in T cell-APC adhesion including CD58, ICAM-1 and ICAM-2 (55).

Activated naïve cells differentiate into effector and memory T lymphocytes

Activation of naïve T cells by cognate pMHC is driven by three distinct and synergistic signals: TCR-pMHC binding, costimulatory receptor engagement, and cytokine-dependent differentiation (56). The αβ TCR alone is not capable of productive signalling, which is mediated through the reversible phosphorylation of ten intracellular tyrosine activation motifs (ITAMs) distributed across CD3-ε, -δ, -γ and -ζ chains by members of the *Src* protein kinase family including *Lck,* which associates with the cytoplasmic tails of the TCR coreceptors CD4 and CD8 (6, 57). Engagement of individual TCR molecules to pMHC may be sufficient to trigger T cell activation (58), ITAM phosphorylation setting off a chain reaction driven by SH2-containing cytosolic proteins including ZAP70 (59). The precise mechanism by which TCR-antigen engagement translates into productive signalling remains hotly debated, and include force-induced conformational changes, mechanosensing, and the kinetic segregation model, reviewed in (60, 61). TCR-pMHC binding is necessary but not sufficient for optimal T cell activation, which also requires antigen-independent engagement of costimulatory molecules such as ITAM-containing CD28 with its B7 ligands (CD80/CD86) on the opposing surface of the APC (62). Finally, T cell activation, survival and differentiation is strongly influenced by the cytokine milieu: including proliferative signalling in response to IL2, the pro-inflammatory cytokines IL12 and IFNα or β in the case of $CD8^+$ T cells, and IL1 in the case of $CD4^+$ T cells (56). Activated naïve T cells differentiate into either effector T cells or long-lived memory T cells, the proportion of naïve, effector and memory cells being dynamic and dependent on both age and the disease status of the individual (63, 64).

Characterisation and prediction of TCR specificity

"Orphan" receptors are those with no known ligand. The challenge of *TCR de-orphanisation* can therefore be cast as a mission to solve two related problems: first, given a TCR of unknown antigen specificity, which antigen-MHC complexes is it most likely to bind; and second, given an antigen-MHC complex, which are the most likely cognate TCRs? In the following section, drawing from our recent published review (65), we first describe the motivation for accurate TCR de-orphanisation, before outlining relevant experimental approaches to do so. We then switch to computational inference of TCR specificity, the focus of the present research, setting out the general requirements of predictive algorithms, highlighting critical barriers to progress, and reviewing the latest modelling advances.

The motivation for TCR de-orphanisation

The ability to accurately predict the cognate ligand of a given TCR or antigen-MHC complex has important implications for the design of new therapies and vaccines, and our understanding

of the biological role of T cells in health and disease (66). In oncology, T cell antigen recognition has become the focus of new drug development efforts, including checkpoint inhibitors; chimeric antigen receptor (CAR)-T cells; endogenous or affinity-enhanced TCRs, and cancer vaccines (67). Cross-reactivity in TCR-based T cell therapies has presented a major roadblock to the development of safe interventions, and gaps in pre-clinical screening have led to tragedies in the clinic (68) which we will explore further in **Chapter 3**. Accurate and generalisable specificity inference could provide an additional safety net to robust experimental screens, predicting likely autoreactivity for a given patient population in oncology and beyond (69).

Beyond the implications for new medicines development, there is significant potential to use predictive tools to dissect the fundamental role of T cells in the surveillance of malignancy. For example, there are reports of the accumulation of clones with driver mutations in sun-exposed skin (70), but the extent to which mutational burden is reflected in TCR repertoires is not well understood (71). Exhausted cytotoxic $CD8^+$ T cells have long been known to be a hallmark of inefficient antitumour immune response (72-74). However, although early data are emerging (75, 76), we do not yet fully know whether T cells with a given antigen specificity are more or less likely to be exhausted.

For infectious diseases such as SARS-CoV-2, predictors of T cell specificity could be of great use in understanding the magnitude and dynamics of antigen-specific T cell responses to the disease (77) and to vaccination (78). However, there remains a significant opportunity to improve open-source systems immunology tools for confident linkage of T cell antigen specificity to differential vaccine-induced response. Linkage of expanded effector T cell populations to their cognate self antigens will also provide vital diagnostic clues as to disease aetiology of autoimmune conditions. This is exemplified by a longitudinal study that demonstrated an association between Epstein-Barr virus infection and the incidence of multiple sclerosis, unlocking new vaccine strategies (79).

Methods in TCR de-orphanisation

From deepening our mechanistic understanding of disease, to providing routes for accelerated development of safer, personalised vaccines and therapies, the case for constructing a complete map of TCR antigen interactions is therefore compelling. We will now explore the landscape of experimental and computational approaches to TCR de-orphanisation.

Experimental approaches

Experimental methods of TCR de-orphanisation can be broadly categorised as *antigen-directed assays*, wherein the reactivity of a given T cell population is tested against a limited panel of

known antigens, and *TCR-directed assays*, in which TCR-pMHC interaction is tested for a single receptor to multiple known or unknown antigens (**Fig. 1-2**) (80). Antigen-directed methods include binding assays in which TCR-pMHC interaction is mapped using fluorescence (81), heavy metal ion (82), or DNA barcode (83)-labelled multimers (84), and *functional readouts* from T cell activation, proliferation, cytokine release, fluorescence reporter, or cell killing assays, reviewed in (80).

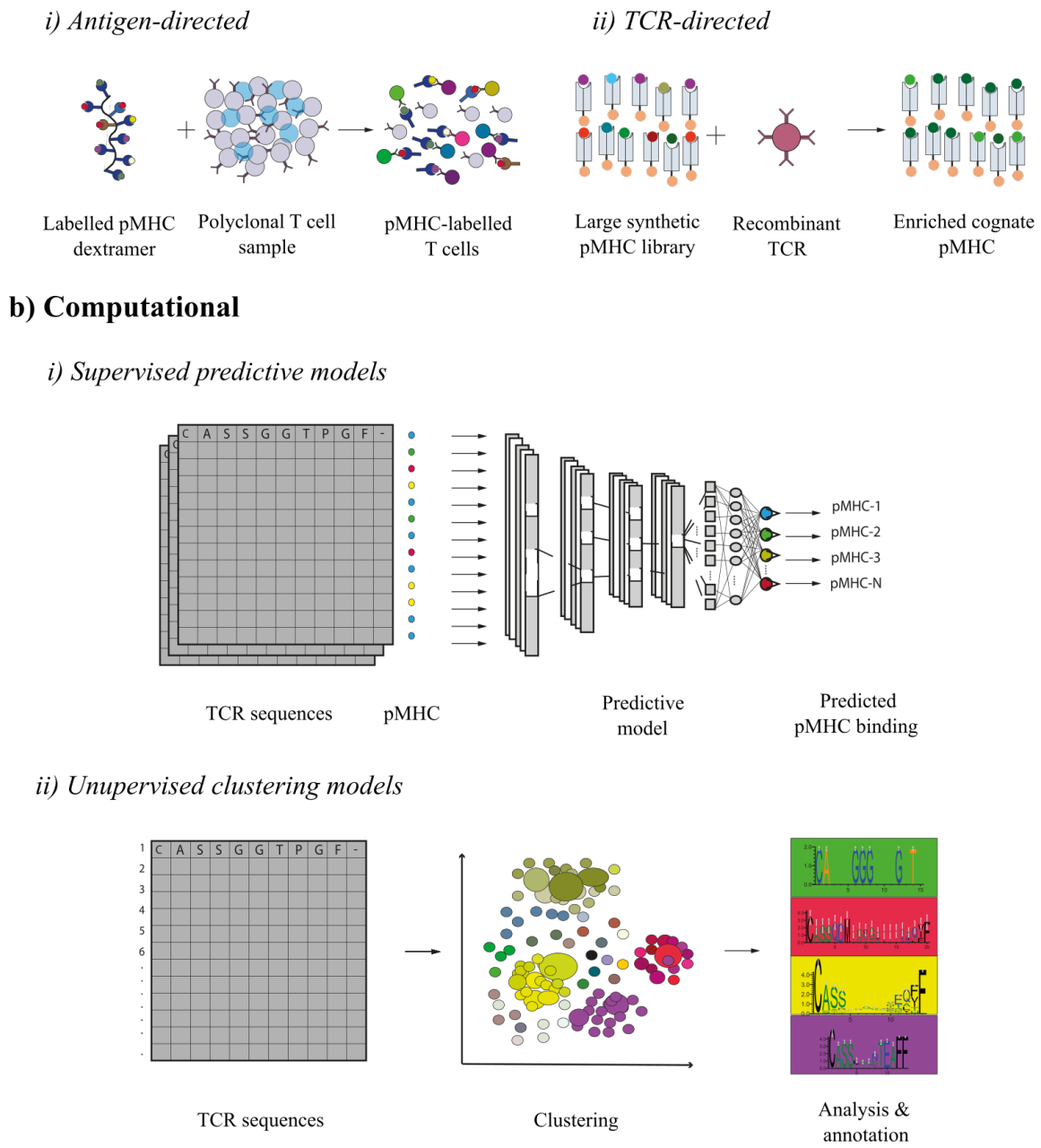


Figure 1-2: Approaches to characterisation of TCR-antigen specificity. a | Experimental methods: showing (i) Multiplex analysis of TCR-pMHC antigen specificity; and (ii) synthetic peptide library

screens for interrogation of peptide specificity of a single TCR. **b | Computational methods:** Showing (i) a representation of a deep neural network (DNN) for supervised prediction of TCR antigen specificity, and (ii) unsupervised clustering analysis of TCR antigen specificity with (*centre*) an example clustering visualisation aiming to assign TCRs to groups (colours) sharing antigen specificity; and (*right*) CDR3 sequence logos for the corresponding clusters.

The development of recombinant antigen-MHC multimer assays in particular (84) has proved transformative in the analysis of TCR antigen specificity, enabling researchers to track and study T cell populations under various conditions and disease settings (85-87). Antigen-MHC multimers may be exploited to determine TCR specificity using bulk (pooled) T cell populations, or newer single-cell methods. Bulk methods are widely used and relatively inexpensive, but do not provide information on αβ TCR chain pairing or function. As a result of this and of the increased combinatorial diversity of β over α TCR chains, single chain TCRβ sequences predominate in public TCR-antigen datasets (15). However, both TCRα and TCRβ chains contribute to antigen recognition and specificity (88, 89). We shall discuss the implications of this disparity for modelling approaches in *Computational Methods* and in **Chapter 6 (Discussion)**. Multimodal single-cell technologies provide insight into chain pairing, transcriptomic and phenotypic profiles at cellular resolution, but remain prohibitively expensive, return fewer TCR sequences per run than bulk experiments, and show significant bias towards TCRs with high specificity (90-92). The appropriate experimental protocol for the reduction of non-specific multimer binding, validation of correct folding and computational improvement of signal-to-noise ratios, remain active fields of debate (91, 92). Indeed, concerns over non-specific binding have led some computational studies to exclude data derived from a 10X study of four healthy donors (93).

Whilst bulk and single-cell methods are limited to a modest number of antigen-MHC complexes per run, the advent of technologies such as lentiviral transfection assays (94, 95) provide scalability to up to 96 antigen-MHC complexes through *library-on-library* screens: Experimental screens that permit analysis of the binding between large libraries of (for example) pMHC complexes and various TCRs. These methods include lentiviral reprogramming techniques such as RAPTR (94), pMHC bead display (96), modified antigen presenting cell systems including SABR (97) and T-SCAN (98), and reporter transfer via trogocytosis (99). However, prior knowledge of the antigen-MHC complexes of interest is still required for many of these methods. This precludes epitope discovery in unknown, rare, sequestered, non-canonical and/or non-protein antigens.

As a result of these barriers to scalability, only a minuscule fraction of the total possible sample space of TCR-antigen pairs has been validated experimentally. At the time of writing, fewer than a million unique TCR-epitope pairs are available from VDJdb, McPas-TCR, the Immune Epitope Database (IEDB) and the MIRA dataset (100-102) (**Fig. 1-3**). Just 4% of these instances contain complete chain pairing information (**Fig. 1-3a**). 97% of all antigens reported as binding a TCR are of viral origin (**Fig. 1-3b**), and a group of just 100 antigens makes up

70% of TCR–antigen pairs (**Fig. 1-3b** and **c**). Where the HLA context of a given antigen is known, the training data are dominated by antigens presented by a handful of common alleles (**Fig. 1-3d**). Many antigens have only one known cognate TCR (**Fig. 1-3e**).

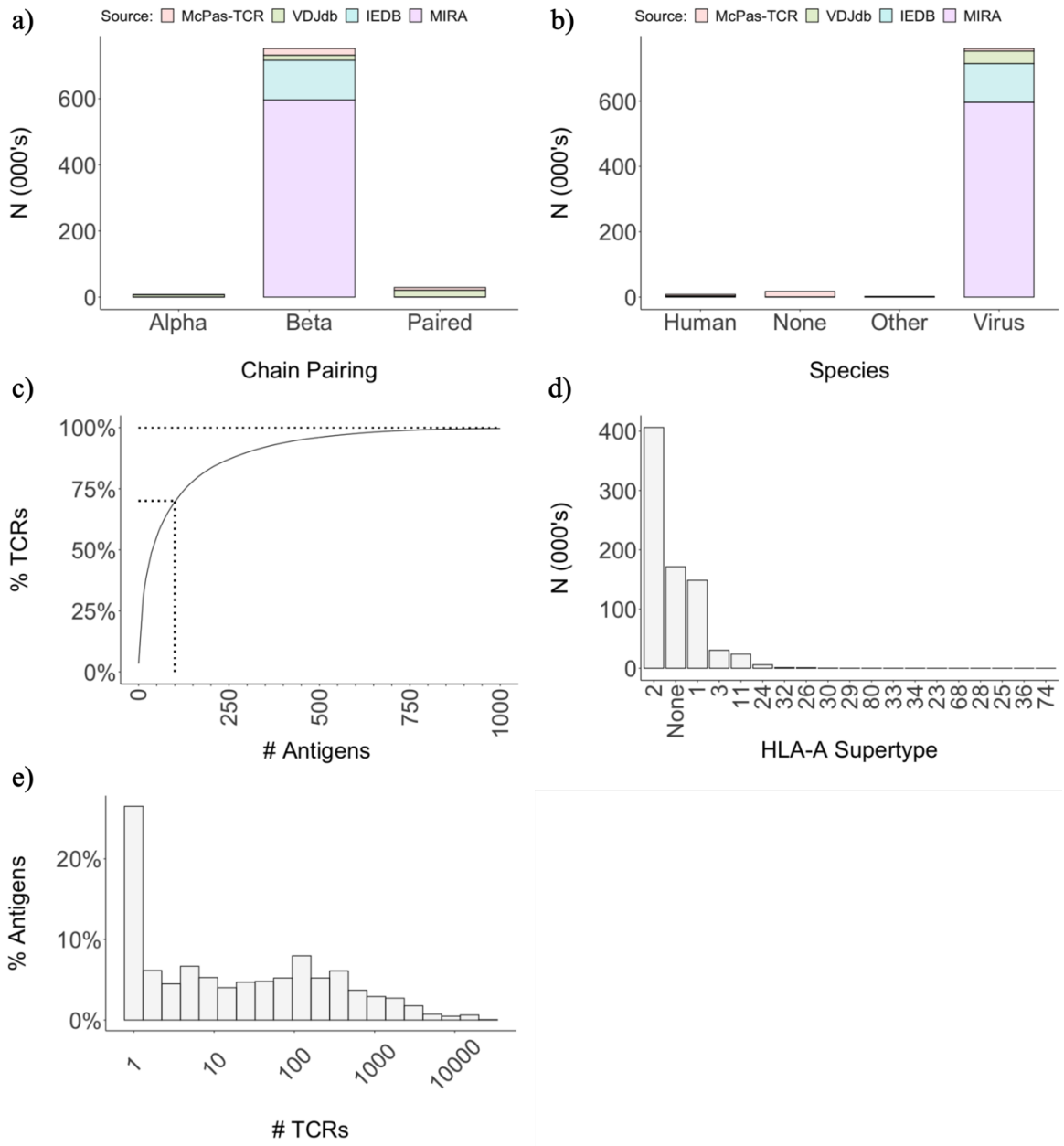


Figure 1-3: **The current landscape of known TCR-antigen pairs**. **a |** Number of TCRs containing α, β or paired chains, showing variation in numbers according to dataset (McPas-TCR, VDJdb, Immune IEDB and MIRA) (100-103). **b |** Number of TCRs per antigen species of origin, showing that most antigens reported as binding a TCR are of viral origin. **c |** Cumulative frequency of antigens, showing that a group of 100 antigens makes up 70% of TCR–antigen pairs. **d |** Number of TCRs by HLA-A type, showing that known antigens are reported in complex with only a few common HLA alleles. **e |** Frequency histogram showing that most antigens have only one known cognate TCR in the combined dataset.

The advent of synthetic peptide display libraries (**Fig. 1-2a**) has permitted the extension of binding analysis to hundreds of thousands of peptides per TCR (104-107). Such TCR-directed strategies can be divided broadly into *binding screens* using bacteriophage, baculoviral or yeast display (104, 105, 107); and *functional screens* including cytokine production, fluorescence reporter, proliferation and killing assays, reviewed in (80). Of these, yeast pMHC display libraries (**Chapter 3**) provide a low cost, scalable option for TCR specificity mapping that has been applied to characterise TCRs implicated in cancer and autoimmunity (104, 105, 108-110), and to engineer receptors with improved specificity (104, 111). Peptides are typically displayed on the surface of yeast as a single chain trimer of peptide, MHC, and β_2 microglobulin. Affinity-driven selection of yeast pMHC complexes using a soluble TCR, followed by deep sequencing of the enriched peptides, provides a quantitative view of receptor binding preferences. Screening large, fully randomised libraries in this way provides a more representative view of the total sequence space than does analysis of peptides drawn from the human proteome alone, or by positional scanning of a reference peptide (**Table 1-1**). As we shall explore in **Chapter 3**, statistical representations of TCR screen data, including Positional Weight Matrices (PWMs) or more complex machine learning models (112, 113), can be developed using selection data and subsequently be used to scan reference proteomes for likely predicted binders (104, 105, 108-111).

Table 1-1: Random peptide libraries. Unique peptide pool size by randomisation method for 9-mer and 10-mer peptides drawing from 19 common natural amino acids (excluding cysteine), using HLA-A*01:01 and HLA-A*02:01 as exemplars.

Randomisation method	**HLA-A*01 9-mer**	**HLA-A*02 10-mer**
Fully random	$19^9 = 3.2 \times 10^{11}$	$19^{10} = 6.1 \times 10^{12}$
Anchor assumption	P3 D/E; P9 Y	P2 L/M; P10 L/M/V/A
Random excluding anchors	$19^7 \times 2 \times 1 = 1.8 \times 10^9$	$19^8 \times 2 \times 4 = 1.6 \times 10^{10}$
Uniref human proteome (114)	1.1×10^7	1.1×10^7
X-scanning mutagenesis	19 x 9 = 171	19 x10 = 190
X-scanning mutagenesis excluding anchors	(19 x 7) + 3 =136	(19 * 8) + 6 = 158

High-throughput library screens such as these provide opportunities for improved screening of the antigen-MHC space, but limit analysis to individual TCRs and rely on TCR-MHC binding instead of measures of functional activation. There remains a need for high-throughput linkage of antigen specificity and T cell function, for example through mammalian or bead display (96, 99, 115).

Antigen- or TCR- directed methods are consequently either limited to a handful of known antigens, or constrained to single TCRs, respectively. The research community has therefore turned to *machine learning* models, a broad family of computational and statistical methods that aim to identify statistically conserved patterns within a data set without being explicitly programmed to do so, as a means of predicting the antigen specificity of orphan TCRs.

A critical requirement of models attempting to predict TCR specificity is that they should be able to make accurate predictions for any combination of TCR and antigen-MHC. These should cover both "seen" pairs included in the data on which the model was trained, and novel or "unseen" TCR-epitope pairs to which the model has not been exposed (116). Significant advances have been made for specificity inference of seen epitopes in particular disease contexts. For example, clusters of TCRs having common antigen specificity have been identified for *Mycobacterium tuberculosis* (117) and SARS-CoV-2 (118), providing possible avenues for new vaccine and pharmaceutical development. However, as we will discuss later, performance for seen epitopes wanes beyond a small number of immunodominant viral epitopes, and is generally poor for unseen epitopes (116, 119). This matters because many epitopes encountered in nature will not have an experimentally validated cognate TCR, particularly those of human or non-viral origin (**Fig. 1-3b**). In the text to follow, we will refer to the case for *generalisable* TCR-antigen specificity inference, meaning prediction of binding for both seen and unseen antigens in any MHC context.

An important distinction must also be made between the related tasks of predicting TCR specificity and antigen immunogenicity. The former is the prediction of binding between sets of TCRs and antigen-MHC complexes. The latter can be described as predicting whether a given antigen will induce a functional T cell immune response: a complex chain of events spanning antigen expression, processing and presentation, TCR binding, T cell activation, expansion and effector differentiation. Whilst great strides have been made in improving prediction of antigen processing and binding to common HLA alleles, the nature and extent to which presented peptides trigger a T cell response is yet to be elucidated (120). A significant gap also remains for the prediction of T cell activation for a given peptide (121, 122), and the parameters that influence pathological peptide or neoantigen immunogenicity remain under intense investigation (123). Multi-modal datasets including binding and activation data provide a potential bridge between these worlds, one that we explore in **Chapters 3** and **5**.

A comprehensive survey of computational models for TCR specificity inference is beyond the scope intended here, but can be found in the following helpful reviews (66, 122, 124-127). Broadly speaking, current models can be divided into two categories, which we dub *supervised predictive models* (SPMs, **Fig. 1-2b**) and *unsupervised clustering models* (UCMs, **Fig. 1-2c**) on account of their respective use of supervised learning and unsupervised learning. Supervised

learning models learn a mathematical function mapping from an input to a predicted label (such as the cognate epitope of an orphan TCR), given some data set containing both input data and associated labels. Common supervised tasks include regression, where the label is a continuous variable, and classification, where the label is discrete. Unsupervised learning models learn to assign input data to clusters having similar features, or otherwise to characterise the underlying statistical patterns of the data. Unlike supervised models, unsupervised models do not require (but may make use of) labels. Common unsupervised techniques include clustering algorithms such as K-Means; anomaly detection models; and dimensionality reduction techniques such as principal component analysis (PCA) (128), triplet stochastic neighbourhood embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) (129). A non-exhaustive summary of open-source SPMs and UCMs that have been deployed for inference of T cell specificity can be found in **Table 1-2**.

Table 1-2: A non-exhaustive list of supervised and unsupervised models for inference of TCR epitope specificity. BFD, Big Fantastic Database; DNN, deep neural network; PIRD, pan-immune repertoire database; SPM, supervised predictive model; TCGA, The Cancer Genome Atlas; TCR, T cell receptor; UCM, unsupervised clustering model. *Not all UCMs are explicitly trained, and so datasets reported for non-DNN UCMs are those on which the models were evaluated.

Model	**TCR chain input**	**Training data**	**Method**	**Availability**
ATM-TCR (130)	Single	IEDB (101) McPas-TCR (103) VDJdb (100)	DNN-SPM	https://github.com/Lee-CBG/ATM-TCR
ImmuneML (131)	Paired	Heikkila (132) VDJdb	DNN-SPM	https://immuneml.uio.no/
NetTCR2 (133)	Single or paired	IEDB VDJdb 10X (134)	DNN-SPM	https://services.healthtech.dtu.dk/service.php?NetTCR-2.0
SwarmTCR (135)	Single or paired	IEDB VDJdb Private	SPM	https://github.com/thecodingdoc/SwarmTCR
ImRex (116)	Single	VDJdb Dean (136)	DNN-SPM	https://github.com/pmoris/ImRex
Luu et al. (137)	Single	IEDB McPas-TCR PIRD (138) VDJdb	DNN-SPM	https://github.com/jssong-lab/TCR-Epitope-Binding
TCRGP (139)	Single or paired	Dash (140) VDJdb	SPM	https://github.com/emmijokinen/TCRGP
TcellMatch (141)	Paired	IEDB VDJdb 10X	DNN-SPM	https://github.com/theislab/tcellmatch

SETE (142)	Single	Dash VDJdb	SPM	https://github.com/wonanut/SETE
ClusTCR (143)	Single or paired	VDJdb Emerson (89)	UCM	https://github.com/svalkiers/clusTCR
TCRdist3 (118)	Single or paired	Dash Nolan (102) Snyder (144) VDJdb	UCM	https://github.com/kmayerb/tcrdist3/
GIANA (145)	Single	Dash Glanville (87) IEDB VDJdb Zhang (146)	UCM	https://github.com/s175573/GIANA
GLIPH2 (117)	Single or paired	Private VDJdb	UCM	http://50.255.35.37:8080/
iSMART (147)	Single	Emerson VDJdb TCGA	UCM	https://github.com/s175573/iSMART
TCRDock (148)	Paired	BFD (149) PDB (150)	Pre-trained DNN and DNN-SPM	https://github.com/phbradley/TCRdock
TCR-BERT (151)	Single	PIRD VDJdb TCRdb (152)	Pre-trained DNN and DNN-SPM	https://huggingface.co/wukevin/tcr-bert
TITAN (119)	Single	BindingDB (153) VDJdb ImmuneCODE (154)	Pre-trained DNN and DNN-SPM	https://github.com/PaccMann/TITAN
pMTNet (155)	Single	Chen (156) Huth (157) Joglekar (97) PIRD McPas-TCR VDJdb Zhang 10X	Pre-trained DNN and DNN-SPM	https://github.com/tianshilu/pMTnet
ERGO-II (158)	Single or paired	Kanakry (159) McPas-TCR VDJdb Zhang	Pre-trained DNN-UCM	https://github.com/IdoSpringer/ERGO-II

			and DNN-SPM	
ICON-TCRAI (91)	Single or paired	McPas-TCR VDJdb 10X	Pre-trained DNN and DNN-SPM	https://github.com/regeneron-mpds/ICON
TCRMatch (160)	Single	IEDB 10X	Unsupervised: nearest neighbour	https://www.github.com/IEDB/TCRMatch
DeepTCR (161)	Single or paired	Dash Glanville Sidhom 10X Multiple public sources: see DeepTCR github repository	DNN-UCM or DNN-SPM	https://github.com/sidhomj/DeepTCR

Supervised predictive models

SPMs are those which attempt to learn a function that will correctly predict the cognate epitope for a given input TCR of unknown specificity, given some training dataset of known TCR-epitope pairs. The training data set serves as an input to the model from which it learns some predictive or analytical function. *Neural networks* are a family of machine learning models, inspired by the synaptic connections of the brain, that are made up of stacked layers of simple interconnected models. Whilst each component of the network may learn a relatively simple linear predictive function, the combination of many predictors allows neural networks to perform arbitrarily complex tasks and to infer non-linear relationships from datasets of millions or billions of instances. Neural networks may be trained using supervised or unsupervised learning, or a combination of the two, and may deploy a wide variety of different model architectures. The last five years have witnessed an acceleration of publications aiming to address the challenge of SPM development with *deep neural networks* (DNNs): Neural networks that have more than one intermediate layer (see **Fig. 1-2b**). Whilst there are many possible approaches to comparing SPM performance, among the most consistently used is the area under the receiver operating characteristic curve (ROC-AUC). ROC-AUC and the area under the precision-recall curve (PR-AUC) are measures of model tendency to different classes of error. These plots are produced for classification tasks by changing the threshold at which a model prediction falling between zero and one is assigned to the positive label class, for

example predicted binding of a given TCR-antigen pair. ROC-AUC can be computed as the area under the line described by a plot of the True Positive Rate and False Positive Rate, and is typically more appropriate for problems where positive and negative labels are proportionally represented in the input data. PR-AUC can be determined as the area under the line described by a plot of model precision against model recall, and is typically more appropriate for problems in which either the positive or the negative label is less frequently observed than the other class. One would expect to observe 50% ROC-AUC from a random guess in a binary (binding or non-binding) task, assuming a balanced proportion of negative and positive pairs.

Performance now surpasses 80% ROC-AUC for a handful of seen immunodominant viral epitopes presented by MHC class I (93, 116, 137). However, representation is not a guarantee of performance: 60% ROC-AUC has been reported for HLA-A*02:01–CMV-NLVPMVATV (133), possibly due to the recognition of this immunodominant antigen by diverse CDR3 motifs. Critically, few models explicitly evaluate the performance of trained predictors on unseen epitopes using comparable datasets. Weber and colleagues (119) achieved an average of 62 ± 6% ROC-AUC for TITAN, compared with 50% for ImRex on a reference data set of unseen epitopes from VDJdb and COVID-19 datasets. Values of 56 ± 5% and 55 ± 3% were reported for TITAN and ImRex respectively in a subsequent paper from the Meysman group (162). Other groups have published unseen epitope ROC-AUC values ranging from 47% to 97%, however many of these values are reported on different datasets (**Table 1-2**), lack confidence estimates following validation (141, 151, 155, 163), and/or have not been consistently reproducible in independent evaluations (164).

Together, these results highlight a critical need for thorough, independent benchmarking studies conducted across models on datasets prepared and analysed in a consistent manner (93). Until reliable generalisable inference is achieved, newer models may be applied with reasonable confidence to the prediction of binding to immunodominant viral epitopes by common HLA alleles. However, SPMs should be used with caution when extending to prediction of any epitope, as performance is likely to drop the further the epitope is in sequence from those in the training set (116).

Unsupervised Clustering Models

Unlike SPMs, UCMs do not depend on the availability of labelled data, learning instead to produce groupings of the TCR or antigen input that reflect the underlying statistical variations of the data (87, 140) (**Fig. 1-2b**). Applied to TCR repertoires, UCMs take as their input single or paired TCR CDR3 amino acid sequences, with or without gene usage information, and return a mapping of sequences to unique clusters. Clustering is achieved by determining the similarity between input sequences, using either “hand-crafted” features such as sequence distance or enrichment of short sub-sequences (k-mers), or by comparing abstract features learned by DNNs (**Table 1-2**).

Clustering provides multiple paths to specificity inference for orphan TCRs (124-126). Epitope specificity can be predicted by assuming that if an unlabelled TCR is similar in sequence to a receptor of known specificity, it will bind to the same epitope (160). One may therefore co-cluster unlabelled and labelled TCRs, and assign the modal or most enriched epitope to all sequences that cluster together (145). Alternatively, DNNs can be used to generate embeddings or *protein fingerprints*, simple fixed-length numerical representations of complex variable input sequences that may serve as a direct input for a second supervised model, or as an alternative input for clustering (91, 161).

As we shall explore in **Chapter 2**, quantitative assessment of the relative merits of hand-crafted and neural network-based UCMs for TCR specificity inference remains limited to the proponents of each new model. Whilst some DNN-UCMs allow for the integration of paired chain sequences and even transcriptomic profiles (141), they are susceptible to the same training biases as SPMs and are notably less easy to implement than established clustering models such as GLIPH and TCRdist (87, 140). However, these established clustering models scale relatively poorly to large datasets compared with newer releases (143, 145). Recent analyses suggest that there is little to differentiate commonly used UCMs from simple sequence distance measures (93, 161).

Key challenges in computational TCR de-orphanisation

Despite the exponential growth of unlabelled immune repertoire data and the recent unprecedented breakthroughs in the fields of data science and artificial intelligence, quantitative immunology still lacks a framework for the systematic and generalisable inference of T cell antigen specificity of orphan TCRs. Among the most plausible explanations for these failures are incomplete modelling of the underlying immunology, limitations in the data, and methodological gaps.

TCR antigen recognition is complex

At a conservative estimate of five million unique TCRs per individual at a given time, a global population of eight billion sharing 11% of their TCRs would represent a unique TCR pool of 3.6×10^{15} (165). This figure excludes recognition of antigens from over 1,400 pathogens known to be capable of infecting humans (166) , binding to self, non-canonical and neoantigens, and presentation of antigens in over 40,000 HLA contexts (19). The universe of feasible TCR-antigen-MHC combinations is therefore likely to be orders of magnitude higher, even when accounting for degeneracy in TCR antigen recognition.

Not only is the TCR-antigen interface diverse at the level of the TCR, the antigen, and MHC, but a given TCR may also bind multiple targets (105). A key challenge to generalisable TCR

specificity inference is therefore that TCRs are on the one hand specific for antigens bearing particular motifs, and on the other capable of considerable promiscuity (13, 167). This contradiction might be explained through specific interaction of conserved *hotspot* residues in the TCR CDR loops with corresponding 2-3 residue clusters in the antigen, balanced by a greater tolerance of variations in amino acids at other positions (168). TCRs may also bind different antigen-MHC complexes using alternative docking topologies (169).

The study of so-called T cell *polyspecificity* or *cross-reactivity* has a long history, reviewed in (125, 170). Departure from the "one receptor, one target" dogma of clonal selection theory emerged in 1977 as hypotheses to explain alloreactivity in murine T cells (171). Don Mason, writing in 1998, described the physical implausibility of clonal selection as applied to TCRs, requiring as it would a lymphoid mass of 500kg per individual to recognise the typical spectrum of extant pathogens (167). Indeed, heterologous immunity - in which priming of the immune system by one pathogen confers protection against another strain or species - has been shown to play an important role in broad antiviral defence, reviewed in (172). Viral priming may also play a role in cross-protection from malignant disease (173).

However, whilst TCR cross-reactivity may have conferred an evolutionary advantage, it poses a significant barrier to the development of new therapies targeting diseased tissue via engineered TCRs. Firstly, receptor mutagenesis to increase affinity towards a disease-associated antigen may introduce a significant risk of severe adverse safety events due to off-target recognition (68, 174-176). This, as introducing single amino acid changes on either side of the TCR-pMHC interface may dramatically impact TCR engagement, as evidenced by alloreactivity towards HLA-B*44:02/03 during organ transplantation (177). Second, if a problematic off-target hit is identified during development, candidates designed to reduce reactivity towards that peptide may inadvertently introduce autoreactivity to one or more other self proteins. This complexity, which we explore in **Chapters 3** and **4**, is among the reasons why, despite the characterisation of the role of the TCR pMHC complex in 1988, the first approval of a TCR therapeutic (TCR-T) was not until 2022 (178, 179).

TCR antigen binding datasets are limited

In the face of the challenges described above, the single most significant limitation to model development is the availability of high-quality TCR and antigen-MHC pairs. The need is most acute for under-represented antigens, for those presented by less frequent HLA alleles, and for linkage of epitope specificity and T cell function (**Fig. 1-3**). Meanwhile, single-cell multimodal technologies have given rise to hundreds of millions of unlabelled TCR sequences (15, 180), linked to transcriptomic, phenotypic and functional information. However, these unlabelled data are not without significant limitations. Notably, biological factors such as age, sex, ethnicity and disease setting vary between studies, and are likely to influence immune repertoires. Differences in experimental protocol, sequence pre-processing, total variation

filtering (denoising) and normalisation between laboratory groups are also likely to have an impact (181). Therefore, thoughtful approaches to data consolidation, noise correction, batch correction, processing and annotation are likely to be crucial in advancing state-of-the-art predictive models.

Historic modelling approaches are incomplete

Despite improvements made in the prediction of "seen" epitopes, models whose TCR sequence input is limited to the use of β chain CDR3 loops and categorical VDJ gene codes (**Table 1-2**) are only ever likely to tell part of the story of antigen recognition, and the extent to which single chain information is sufficient to describe TCR antigen specificity remains an open question. Structural (169) and statistical (182) analyses suggest that α and β chains contribute equally to specificity, and incorporating both chains has improved predictive performance (133). However, chain pairing information is largely absent (**Fig. 1-3a**), and many state-of-the-art SPMs and UCMs rely on single chain information alone (**Table 1-2**). Whilst CDR3 loops may be primarily responsible for antigen recognition, residues from CDR1, CDR2 and even the framework region of both α and β chains may be involved (169). Subtle compensatory changes in interaction networks between pMHC and TCR, altered binding modes, and conformational flexibility in both TCR and MHC may underpin TCR cross-reactivity (168, 183). Explicit encoding of structural information for specificity inference has until recently been limited to studies of a limited set of crystal structures (87, 184). However, the advent of automated protein structure prediction with software programmes such as RoseTTaFold, ESMFold and AlphaFold-Multimer provide potential opportunities for large-scale sequence and structure interpretations of TCR epitope specificity (185-187). This has been illustrated in a recent study in which a modified version of AlphaFold-Multimer was used to identify the most likely binder to a given TCR, achieving mean ROC-AUC of 82% on a small pool of eight seen epitopes (148).

To train robust models, balanced sets of negative and positive samples are required. In the absence of reliable true experimental negatives, negative instances may be produced by *shuffling* or by drawing randomly from healthy donor repertoires (116). Shuffling is the act of assigning a given TCR drawn from the set of known TCR-antigen pairs to an epitope other than its cognate ligand, and labelling the randomly generated pair as a negative instance. However, these approaches assume on the one hand that TCRs do not cross-react and, on the other, that the healthy donor repertoires do not include sequences reactive to the epitopes of interest. A recent study from Jiang and colleagues (188) provides interesting new strategies to address this challenge.

Finally, despite the known potential for promiscuity in the TCR, the pre-processing approaches deployed in many SPM or UCM pipelines assume that a given TCR has only one cognate epitope, and report only the most probable peptide as the predicted binder. A probability

distribution over a subsample of the most likely cognate antigen-MHC complexes is more likely to give a more physiologically accurate picture.

Having thus reviewed the mechanisms of T cell recognition, current experimental and computational methods for characterising T cell antigen specificity, we now proceed to an independent benchmarking exercise examining the potential of UCMs in TCR de-orphanisation.

Chapter 2: A comparison of clustering models for inference of TCR antigen specificity

Overview

The vast potential sequence diversity of TCRs and their ligands has presented a historic barrier to computational prediction of TCR epitope specificity, a holy grail of quantitative immunology. One common approach is to cluster sequences together, on the assumption that similar receptors bind similar epitopes. In this published work (189), we provide the first independent evaluation of widely used clustering algorithms for TCR specificity inference, observing some variability in predictive performance between models, and marked differences in scalability. Despite these differences, we find that different algorithms produce clusters with high degrees of similarity for receptors recognising the same epitope. Our analysis strengthens the case for use of clustering models to identify signals of common specificity from large repertoires, whilst highlighting the scope for improvement of complex models over simple sequence mismatch comparators.

Introduction

UCMs do not require receptor-ligand pairs as an input, but rather group similar TCRs together on the broad-based assumption that receptors having similar sequences share antigen specificity (**Fig. 2-1**) (87, 140). This is of particular use in an era when bulk and single-cell sequencing experiments can yield thousands of unique TCRs per sample; by applying UCMs to group receptors into clusters sharing similar sequence, structural, or physico-chemical properties, one can shortlist TCRs of interest for later experimental de-orphanisation. Such approaches have been successfully applied to identify and characterise TCRs associated with mycobacterial and viral infection, cancer, and autoimmune disease (87, 117, 147, 190-192).

UCMs take as their input single or paired TCR CDR3 nucleotide or amino acid sequences, with or without V and J gene usage information, and return a mapping of sequences to unique clusters. This has historically been achieved using some form of distance measure, typically either direct sequence similarity and/ or the frequency enrichment of short sequence snippets (*k-mers*) compared to a reference dataset. Recent approaches leverage DNNs to generate a compressed numeric representation of the input TCR as a precursor to clustering (91, 161, 163, 193). We refer to such models as DNN-UCMs to differentiate them from “traditional” distance-based UCMs such as GLIPH and tcrdist, both released in 2017 (87, 140). A critical advantage of these traditional UCMs is that they do not require large volumes of training data, a hallmark and limitation of DNNs.

When some or all of the cognate epitopes of a given TCR dataset are known, UCMs can theoretically be used to infer epitope specificity (**Fig. 2-1b**). However, there has to date been no independent benchmarking study of UCMs as predictors of TCR specificity, despite their widespread use in the field. In the present work (**Fig. 2-2**), we provide the first confirmation of the direct predictive capacity of UCMs by co-clustering TCRs of known specificity and comparing their performance using F1-score, a balanced measure of precision and recall. We evaluate five commonly used UCMs on sets of known TCR-epitope pairs, investigating the impact of the number of TCR representatives provided per epitope, the receptor chain, the epitope studied, and the input dataset. We find that all five models outperform CDR3 length, V-gene, and random baselines, however relative predictive performance is sensitive to each of these four factors, and we identify no conclusive winner across tasks. A simple Hamming distance model performs surprisingly well on large repertoires and runs faster than even the most scalable new models. Extending the analysis to a qualitative comparison of the clusters formed, we observe that the five models can produce similar cluster motifs for common viral epitopes. Taken together, we conclude that whilst UCMs are of considerable value in de-orphanisation pipelines, there remains significant room for improvement in performance over simple comparators, and that in practice clustering algorithms and predictive models should be combined to maximise the likelihood of identifying true epitope-specific clusters.

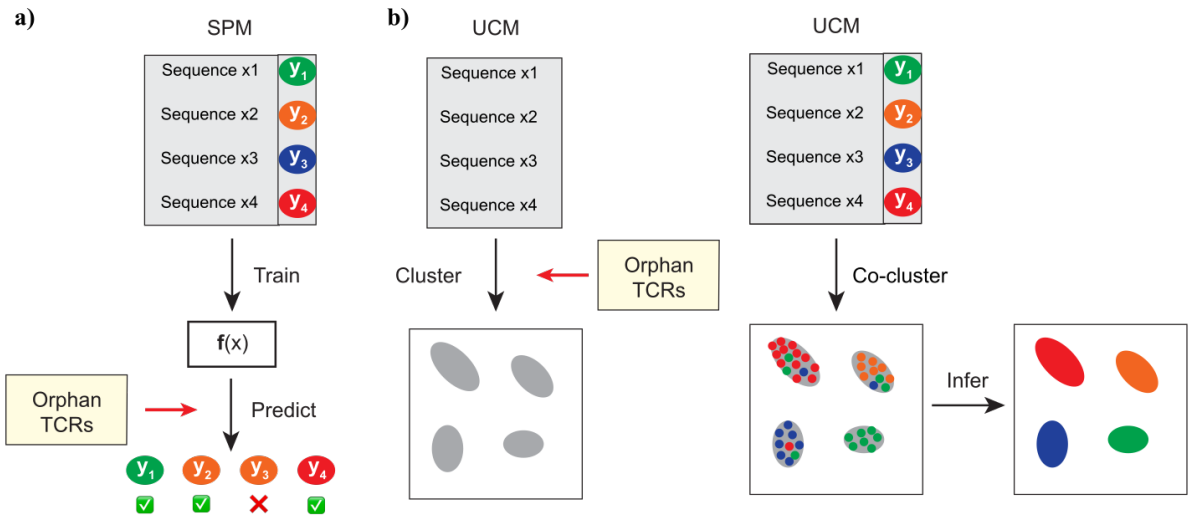


Figure 2-1: Supervised and unsupervised learning in T cell epitope specificity inference. a | SPMs fit a predictive function f(x) to training data having an independent variable X (TCR sequences and other features) and dependent variable y (epitopes or pMHC complexes). This function may then be applied to predict the cognate epitopes of orphan TCRs. **b** | Application of UCMs to de-orphanise TCRs by co-clustering. UCMs generate a mapping from TCR sequences to a cluster allocation, such that each TCR is assigned to one or more clusters having common epitope specificity.

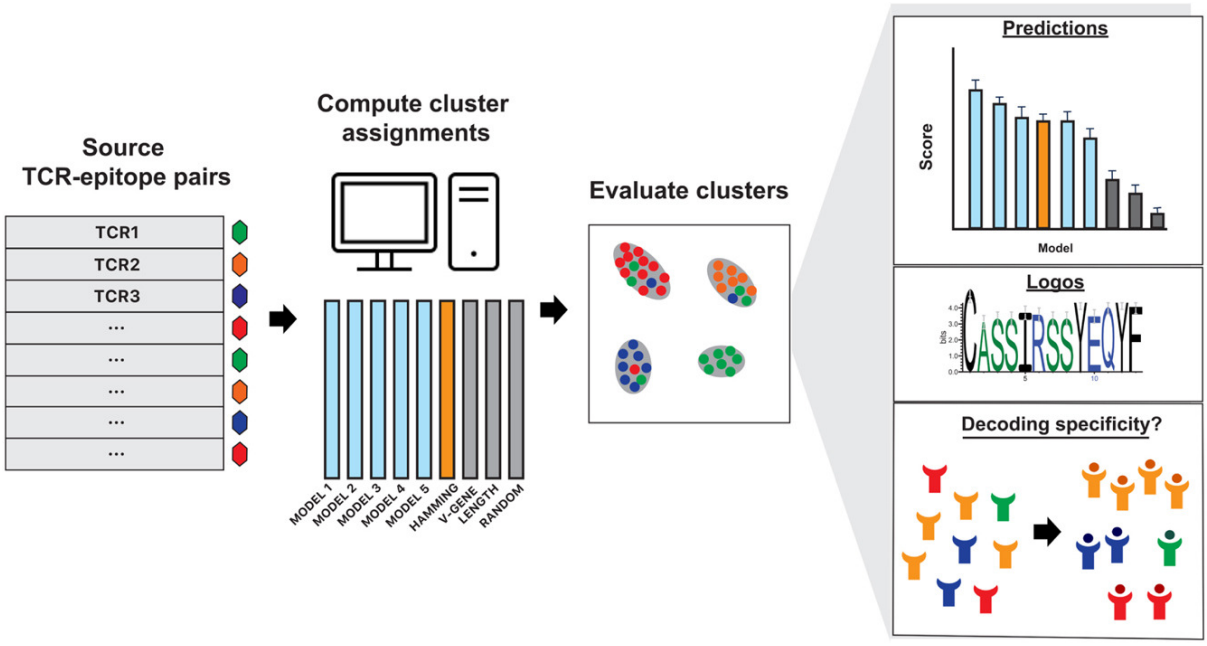


Figure 2-2: Graphical abstract of the methods used in this study. TCR-epitope pairs sourced from publicly available datasets were clustered using one of nine approaches, and the resulting clusters analysed using predictive metrics, qualitative measures, or as correlates of epitope specificity.

Methods

Datasets

A consolidated dataset of paired αβ TCR amino acid sequences of human origin was developed using instances drawn from VDJdb (100) for initial model comparison. TCR-epitope pairs from McPas-TCR (103) and from the MIRA dataset (102) were prepared as separate non-overlapping test sets (**Table 2-1**). Sequences derived from a 10X study of healthy donors (134), and CDR3 sequences containing non amino acid symbols, were removed from the input data. V and J gene codes were processed for consistency with IMGT reference sequences (19). Duplicates were removed within and between datasets using CDR3-V-J bio-identities for both α and β chains, such that a given TCR was encoded in the format CDR3α-TRAV-TRAJ-CDR3β-TRBV-TRBJ. Only those TCRs having V or J genes included in the reference IMGT alleles of the tcrdist module of the CoNGA Conda package (v.0.1.1) (194) were retained, to ensure that consistent numbers of sequences were provided to each model.

Benchmarking experiments were performed on VDJdb data after chain selection and down-sampling as described in the results section below. Neither McPas-TCR nor MIRA datasets

were down-sampled, to give a representative view of model performance on imbalanced datasets that are likely more representative of real-world inference tasks.

Synthetic TCRs were obtained as a common set of 100,000 TCRβ sequences produced using OLGA (195), which was subsampled to inspect the impact of increasing noise on model performance for dataset V1000, such that each model received an identical set of decoy sequences.

Table 2-1: TCR epitope instances per dataset.

Dataset	Min. TCRs / epitope	# Unique epitopes	N total
VDJdb			
V10	10	76	760
V50	50	25	1,250
V100	100	16	1,600
V500	500	5	2,500
V1000	1,000	4	4,000
McPas-TCR	NA	24	509
MIRA	NA	796	59,388

Table 2-2: Frequency of TCR representatives per epitope in V1000 prior to down-sampling.

Epitope	Organism	Source	# Instances
KLGGALQAK	CMV	IE1	13,522
GILGFVFTL	Influenza A	Matrix protein 1	1,830
AVFDRKSDAK	EBV	EBNA-4	1,143
RAKFKQLL	EBV	BLZF1	1,120

Models

Literature search

A search of PubMed was conducted on 30/01/22 with the aim of identifying models for inference of TCR epitope specificity, supplemented by a manual search of the literature making use of Google Scholar and arXiv.

Search terms:

(T cell receptor OR TCR) AND (antigen OR peptide) AND (algorithm OR (cluster* OR predict*)) NOT (chimeric antigen receptor OR CAR OR CAR-T)

Inclusion criteria:

• Primary papers released from 2017 onwards
• Methods for clustering T cell receptors according to antigen specificity

Exclusion criteria:

• BCR clustering / antigen predictions
• Papers published before 01 Jan 2017
• Review papers
• SPMs and deep neural networks

Overview of shortlisted models

ClusTCR (143), GIANA (145), GLIPH2 (117), iSMART (147) and tcrdist3 (118) were shortlisted for analysis based on the availability of open-source Python packages or executable files. ALICE (196) was excluded as it is more appropriately applied to the identification of expanded clones in individual patient repertoire data. To the five test models we added Hamming distance, V-gene, length and random baseline comparators, described below.

ClusTCR (143) makes use of a two-step approach to clustering, in which an N x M matrix of CDR3 amino acid sequence and physicochemical properties is sorted into superclusters using the Facebook AI similarity search (Faiss) library, and the resulting embeddings are clustered with K-Means. A graph network of distances is then produced from these superclusters based on Hamming distances between length-sorted CDR3 sequences. Final cluster assignments are made by applying Markov Clustering (MCL) to the network graph. The ClusTCR Python package (v1.0.2) was imported with Conda and implemented using default settings. Benchmarking of ClusTCR was conducted with the CPU version for fair comparison with non-parallelisable models, and V gene inputs enabled.

GIANA (145) applies multidimensional scaling (MDS) to produce matrix representations of TCR CDR3 sequences that approximate BLOSUM62 physicochemical properties, such that the Euclidean distance between two sequences represented with MDS is equivalent to the Smith-Waterman alignment between the BLOSUM representations of those sequences. MDS vectors are pre-sorted on length, and the resulting superclusters are then sorted into subclusters using the Faiss library before clustering on Smith-Waterman distances between k-mers. GIANAv4.1 was downloaded from GitHub with an IMGT TRBV reference and implemented in CPU mode using default settings following the framework developed for iSMART (147).

GLIPH2 (117) is an update to GLIPH (87) which combines global and local cluster analyses. Global distance is defined as sequence mismatches in CDR3 sequences differing at a given position according to a BLOSUM62 substitution matrix, with shared TRBV gene usage and identical length. Local distance is computed as a statistically significant k-mer frequency enrichment in residues predicted to contact pMHC, compared to a sample population. GLIPH2 was downloaded from the developers' website and run using a combined CD4/CD8 reference, otherwise using default parameters. Where a given sequence was assigned to more than one putative cluster, absolute cluster assignments were made to the cluster having the greatest probability in the output.

iSMART (147) incorporates CDR3 and (optionally) V gene usage information, pre-sorting CDR3 sequences according to length and imposing a gap penalty for length-mismatched CDR3s related by a single insertion. Alignment scores are computed for a subset of the CDR3 sequences using a BLOSUM62 substitution matrix, and output clusters are assigned based on a threshold alignment score. iSMART was implemented as in (143) except that V gene usage was included by default.

tcrdist3 (118) is the latest iteration of tcrdist (140), which makes use of a BLOSUM62 mismatch distance between CDR1, CDR2, "CDR2.5" (an MHC-facing loop), and CDR3 sequences. Non-CDR3 sequences are inferred from a reference database, a gap penalty is applied to account for sequence insertions/deletions, and a combined similarity score is computed that assigns greater weighting to CDR3 sequences. The resulting distance matrix may then be clustered. tcrdist3 (v0.2.2) was installed with PyPI and called with a Python script making use of sparse distance matrices and chunking for large datasets. As tcrdist3 generates a distance measure but does not explicitly cluster instances, a scikit-learn implementation of DBSCAN (197) was used to group distance matrices produced with tcrdist3, consistent with (143) and following comparison of model performance with different model implementations and clustering approaches (**Fig. 2-3**). Amino acid distance matrices were generated with the default meta-clonotype radius of 50. A faster C++ implementation of tcrdist is available as part of the CoNGA package (194), however a steep drop-off in epitope-specific performance was observed when combining this model with DBSCAN (**Fig. 2-3**). Greedy clustering, used in the original tcrdist publication (140) and evaluated in (143), was excluded from the analysis due to prohibitively slow runtimes.

Baseline models

A Hamming distance model was adapted from a version published in the ClusTCR repository (143) that makes use of sequence hashing for efficient CDR3 comparison, first grouping CDR3 sequences by length and then sorting into subclusters differing by only one amino acid. Length, V-gene, and random baseline models were added which assign TCRs to clusters based on CDR3 length, V-gene codes, or random shuffling, respectively.

Compute

All benchmarking experiments were run on a single remote Intel(R) Xeon(R) Gold 6126 CPU at 2.60GHz to ensure fair comparison of algorithms with and without parallel processing capability.

Metrics

Performance of each model was initially analysed using a broad panel of UCM-specific, entropy, and predictive measures. Clustering metrics included purity, consistency and retention, as described previously in (143, 198), where *purity* is the proportion of TCRs in a given cluster having the modal epitope; *consistency* is the proportion of epitope-specific TCRs that are assigned to a single cluster; and *retention* is the proportion of TCRs successfully assigned to a cluster. Entropy was calculated using a scikit learn implementation of adjusted mutual information (AMI) that accounts for class imbalance in computation of the mutual information of true and predicted labels (197). To these were added the predictive metrics accuracy, precision, recall, and f1-score, as described in **Eq. 2-1**. Each of these measures was weighted to adjust for per-epitope class imbalance and computed for all input instances to account for differences in retention between models.

Equation 2-1: Predictive performance metrics.

a. Accuracy = (TP + TN) / (TP + TN + FP + FN)

Where TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

b. Precision = TP / (TP + FP)

c. Recall = TP / (TP + FN)

d. F1-score = (2 ∗ Precision ∗ Recall) / (Precision + Recall)

Statistics

Statistical significance was determined using a two-tailed Mann-Whitney test, consistent with (199). Statistical analyses were performed with GraphPad Prism (200). Plots were produced with GraphPad, Matplotlib, and Seaborn Python libraries (200-202).

Sequence logos were produced using β chain selections from dataset V1000, by retaining TCRs having the modal length from the largest cluster for each of the four epitopes present.

Sequences were aligned with MUSCLEv5.1 (203), and logos were produced from the resulting multiple sequence alignments with WebLogo v3.7.12 (204).

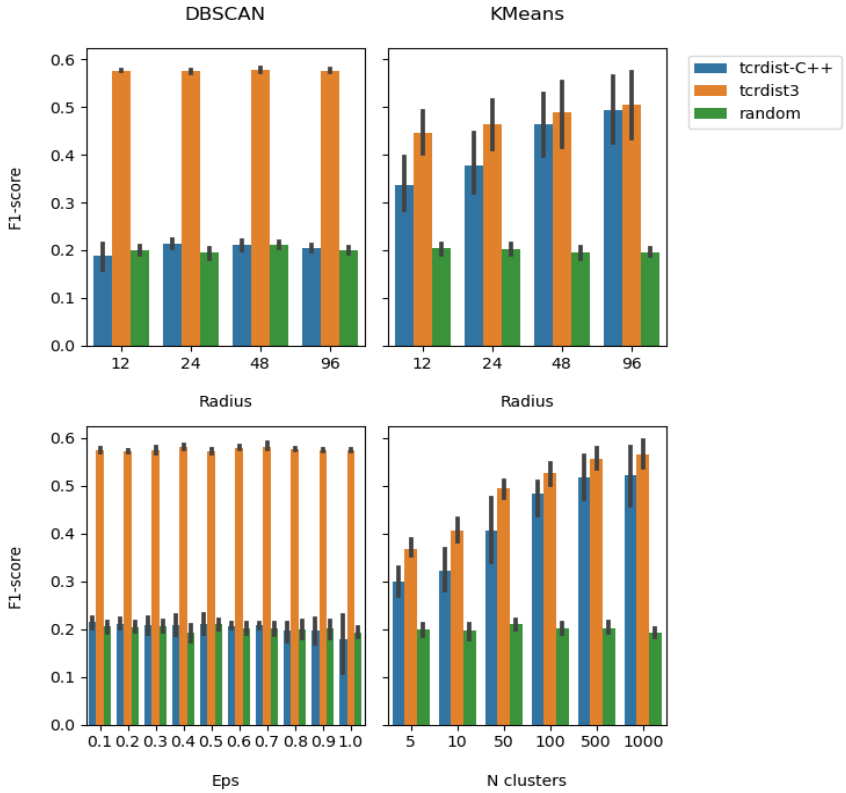


Figure 2-3: **Selecting tcrdist hyperparameters for C++ (tcrdist) or Python (tcrdist3) implementations of tcrdist** (118, 194). **Top row** | Performance as a function of tcrdist meta-clonotype radius. **Bottom row** | Performance as a function of clustering algorithm hyperparameters. Analyses made using K-Means or DBSCAN applied to dataset V1000, β chain selections.

Results

Five open-source UCMs were identified from the literature for which a Python implementation was readily available: ClusTCR, GIANA, GLIPH2, iSMART, and tcrdist3 (117, 118, 143, 145, 147). Four baseline models were added: A Hamming distance model that grouped together sequences having identical length and differing by not more than one amino acid, CDR3 length, V-gene clustering, and random baselines.

Benchmarking analyses were performed on paired αβ TCRs data drawn from VDJdb, a large, public, curated source of TCRs of known epitope specificity and which has been used as a reference for all five models selected (100). However, past studies have presented comparisons using a single pre-processing strategy, for example by retaining only a subset of TCRs binding

selected epitopes (117) or meeting a specific quality threshold (143). To explore whether selective sub-setting of data might mask inherent differences between models, performance was analysed for data subsets generated by retaining epitopes having 10, 50, 100, 500, or 1000 cognate TCRs (datasets V10, V50, V100, V500, and V1000 respectively, **Tables 2-1** and **2-2**). Instances were randomly downsampled after pre-processing, such that each experimental run was performed on the same number of TCR sequences per epitope.

All models were applied to α or β chain selections from the same set of paired TCRs. Sampling was repeated 25 times to account for variance between random TCR selections. We present performance on α and β chain selections independently (see **Discussion**). UCMs have historically been evaluated using cluster quality metrics such as purity, consistency, diversity, and retention (see **Methods**). By applying each model to sets of TCRs of known specificity, we were able to supplement these metrics with direct measures of predictive capacity following the schema depicted in **Fig. 2-1b**.

Observing a positive correlation between performance according to cluster purity, consistency, adjusted mutual information, precision, recall, and F1-score (**Fig. 2-4**), we present results for F1-score alone, weighting scores to account for epitope-specific class imbalance and computing performance across all input TCRs to allow for differences in retention between models.

Differential model performance was first inspected at a global level for datasets V10 through V1000 combined, grouping over α and β chain selections and over all epitopes (**Fig. 2-5a** and **Table 2-3**). All study models outperformed length, V-gene, and random baselines ($p<0.0001$). GIANA and iSMART performed equally well ($p=0.67$), surpassing both a Hamming distance and the other models studied ($p<0.0006$). V-gene usage (*vcluster*) was better predictive of epitope specificity than CDR3 length, both of which outperformed a random baseline.

However, significant variation in relative performance was observed for each model, sensitive to the choice of chain selection (**Fig. 2-5b** and **Table 2-4**), the number of representatives per epitope (**Fig. 2-6a** and **Table 2-5**), the epitope (**Fig. 2-6b** and **Table 2-6**), and the data source (**Fig. 2-6c** and **Table 2-7**), with no one model achieving superiority across all factors. Performance was lower for α than β chain selections for all except V-gene, length and random baselines (**Fig. 2-5b** and **Table 2-4**), whilst predictive capacity of all models was improved when provided with more TCR instances per epitope (**Fig. 2-6a** and **Table 2-5**).

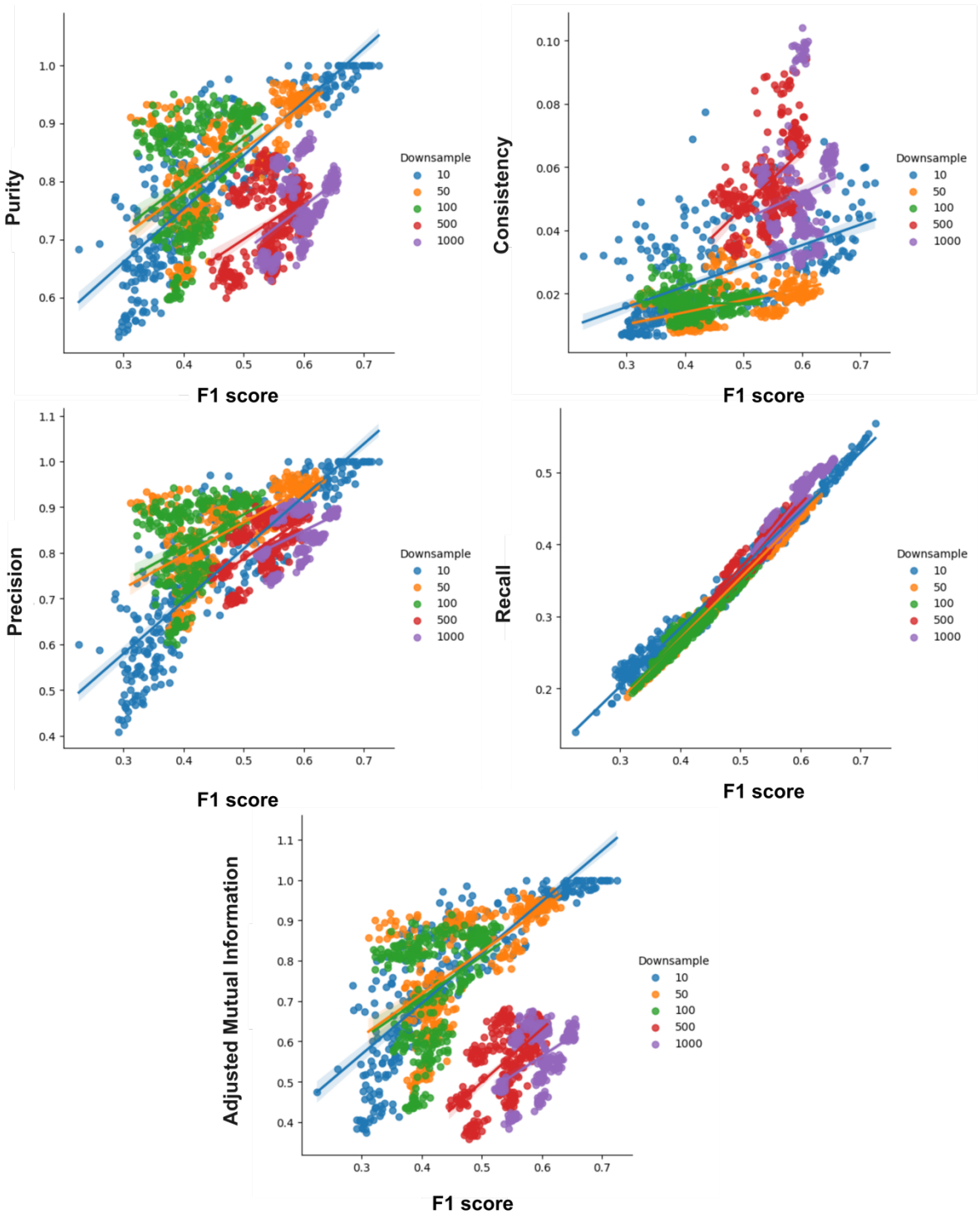


Figure 2-4: Correlation of UCM metrics, datasets V10, V50, V100, V500 and V1000, α and β chain selections combined. Results shown for all models except length, V-gene and random baselines after 25 repeats.

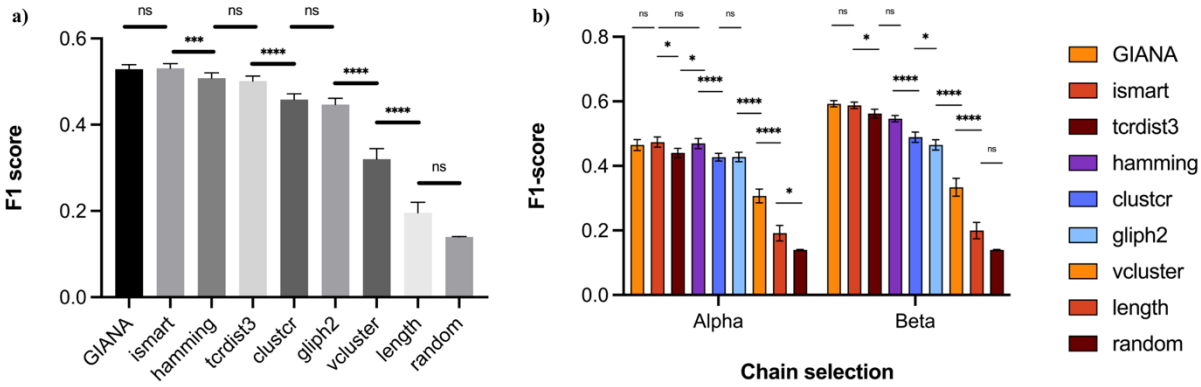


Figure 2-5: Comparison of model performance across datasets. Weighted F1-scores shown for V10-V1000. **a |** α and β chain selections combined. **b |** Split by chain selection. Significance values: ****, $p \leq 0.0001$; ***, $p \leq 0.001$; **, $p \leq 0.01$; *, $p \leq 0.05$, n. s. = not significant.

Table 2-3: Global UCM performance (F1-score). Showing mean values as percentages ± 95% confidence for datasets V10, V50, V100, V500 and V1000 combined, incorporating α and β chain selections and after 25 repeats.

Model	F1-score (%)
GIANA	52.9 ± 1.1
iSMART	53.1 ± 1.1
Hamming	50.8 ± 1.2
tcrdist3	50.1 ± 1.2
ClusTCR	45.8 ± 1.4
GLIPH2	44.7 ± 1.5
vcluster	32.0 ± 2.4
length	19.6 ± 2.5
random	13.9 ± 0.1

Table 2-4: UCM performance (F1-score) by chain selection. Showing mean values as percentages ± 95% confidence for datasets V10, V50, V100, V500 and V1000 combined, after 25 repeats.

Model	α	β
GIANA	46.5 ± 1.7	59.2 ± 0.9
iSMART	47.4 ± 1.6	58.8 ± 1.0
tcrdist3	44.0 ± 1.4	56.2 ± 1.3
Hamming	46.9 ± 1.5	54.6 ± 1.0
ClusTCR	42.7 ± 1.2	48.9 ± 1.6
GLIPH2	42.8 ± 1.5	46.5 ± 1.6
vcluster	30.7 ± 2.1	33.4 ± 2.7
length	19.2 ± 2.4	20.0 ± 2.6
random	14.0 ± 0.2	13.9 ± 0.1

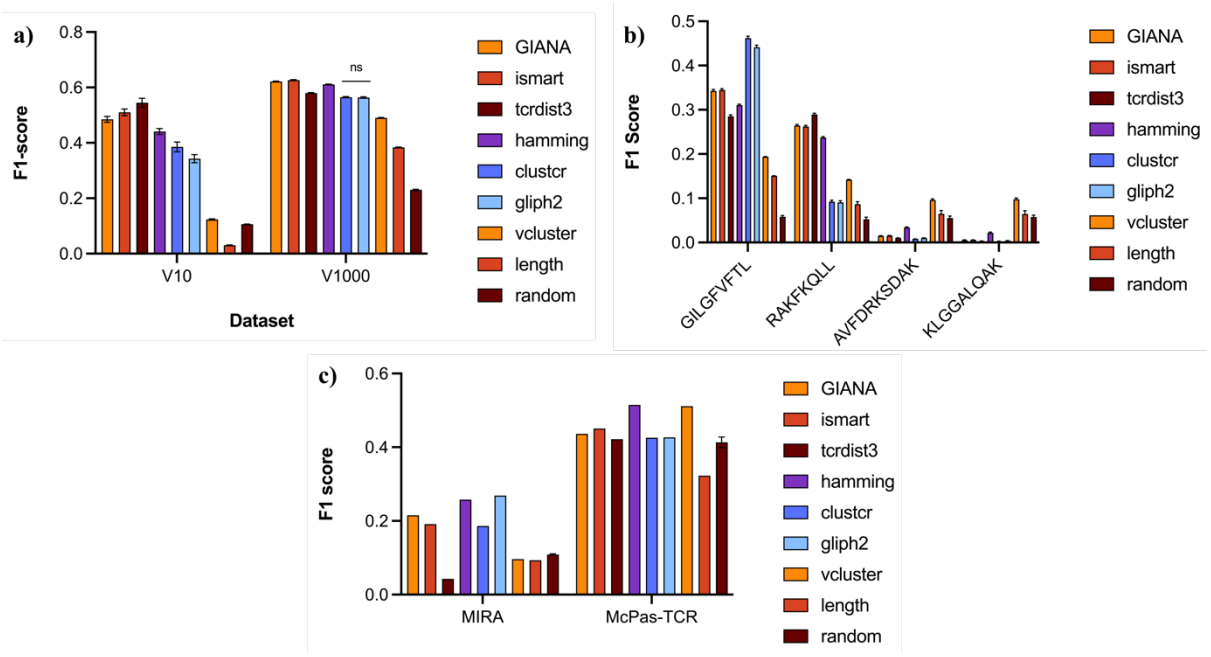


Figure 2-6: Comparison of model performance. a | According to the number of TCRs provided per epitope, α and β chains combined. All comparisons reached statistical significance ($p<0.05$) except for those between ClusTCR and GLIPH2 for V1000. **b** | V1000, β chain selection only, split by epitope. **c** | MIRA (102) and McPas-TCR (103), β chain only selections.

Table 2-5: UCM performance (F1-score) by dataset. Showing mean values as percentages ± 95% confidence for V10 and V1000, α and β chain selections combined, after 25 repeats.

Model	V10	V1000
GIANA	54.5 ± 1.7	58.0 ± 0.9
iSMART	51.1 ± 1.6	62.7 ± 1.0
tcrdist3	48.5 ± 1.4	62.1 ± 1.3
Hamming	44.1 ± 1.5	61.1 ± 1.0
ClusTCR	38.5 ± 1.2	56.5 ± 1.6
GLIPH2	34.3 ± 1.5	56.4 ± 1.6
vcluster	12.3 ± 2.1	49.0 ± 2.7
length	3.0 ± 2.4	38.4 ± 2.6
random	10.6 ± 0.2	23.0 ± 0.1

Table 2-6: UCM performance (F1-score) by epitope. Showing mean values as percentages ± 95% confidence for V1000, β chain selections only, after 25 repeats.

Model	AVFDRKSDAK	GILGFVFTL	KLGGALQAK	RAKFKQLL
GIANA	0.8 ± 0.1	46.2 ± 0.5	0.3 ± 0.1	9.2 ± 0.3
iSMART	1.5 ± 0.1	34.3 ± 0.3	0.5 ± 0.1	26.4 ± 0.2
tcrdist3	1.0 ± 0.1	44.2 ± 0.5	0.4 ± 0.1	9.1 ± 0.4
Hamming	3.4 ± 0.1	31.1 ± 0.3	2.2 ± 0.2	23.7 ± 0.2
ClusTCR	1.5 ± 0.1	34.5 ± 0.3	0.5 ± 0.1	26.2 ± 0.2
GLIPH2	6.5 ± 0.8	15.0 ± 0.0	6.4 ± 0.7	8.7 ± 0.6
vcluster	5.5 ± 0.4	5.8 ± 0.3	5.8 ± 0.4	5.2 ± 0.5
length	1.0 ± 0.1	28.5 ± 0.3	0.3 ± 0.1	28.9 ± 0.2
random	9.6 ± 0.3	19.4 ± 0.1	9.8 ± 0.3	14.2 ± 0.1

Table 2-7: UCM performance (F1-score) for test datasets. Showing mean values as percentages ± 95% confidence, β chain selections only and no down-sampling, after 5 repeats.

Model	MIRA	McPas-TCR
GIANA	21.5 ± 1.7	43.6 ± 0.9
iSMART	19.1 ± 1.6	45.1 ± 1
tcrdist3	4.3 ± 1.4	42.2 ± 1.3
Hamming	25.8 ± 1.5	51.5 ± 1
ClusTCR	18.6 ± 1.2	42.6 ± 1.6
GLIPH2	26.9 ± 1.5	42.7 ± 1.6
vcluster	9.6 ± 2.1	51.1 ± 2.7
length	9.3 ± 2.4	32.3 ± 2.6
random	10.9 ± 0.2	41.3 ± 0.1

Whilst the size and epitope distribution of clusters varied between models (**Fig. 2-7**), CDR3β motifs for the largest clusters associated with each of these epitopes were strikingly similar for all except length, V-gene and random baselines, and were near-identical when applied to GILGFVFTL or RAKFKQLL (**Fig. 2-8**). Motifs were more varied for the other two epitopes in the V1000 dataset (KLGGALQAK and AVFDRKSDAK).

Finally, we compared UCM computational speeds by adding synthetic TCRs, produced with OLGA (195), to V1000 (**Fig. 2-9**). ClusTCR and GIANA were the most scalable of the five models tested, completing their cluster assignments within 10s for 100,000 TCRs, outpaced only by the baseline models. Time constraints prevented extension of the present analysis to comparative performance over parallel CPUs, and to the GPU-enabled versions of ClusTCR and GIANA. A batch-clustering implementation of tcrdist3 was used to avoid CPU overloading during calculation of sequence distances (see **Methods**), tcrdist running slowly compared to

the other test models. Model predictions were negatively impacted by the addition of 100,000 synthetic TCR sequences, with F1-scores dropping for all but tcrdist3 (**Fig. 2-10**).

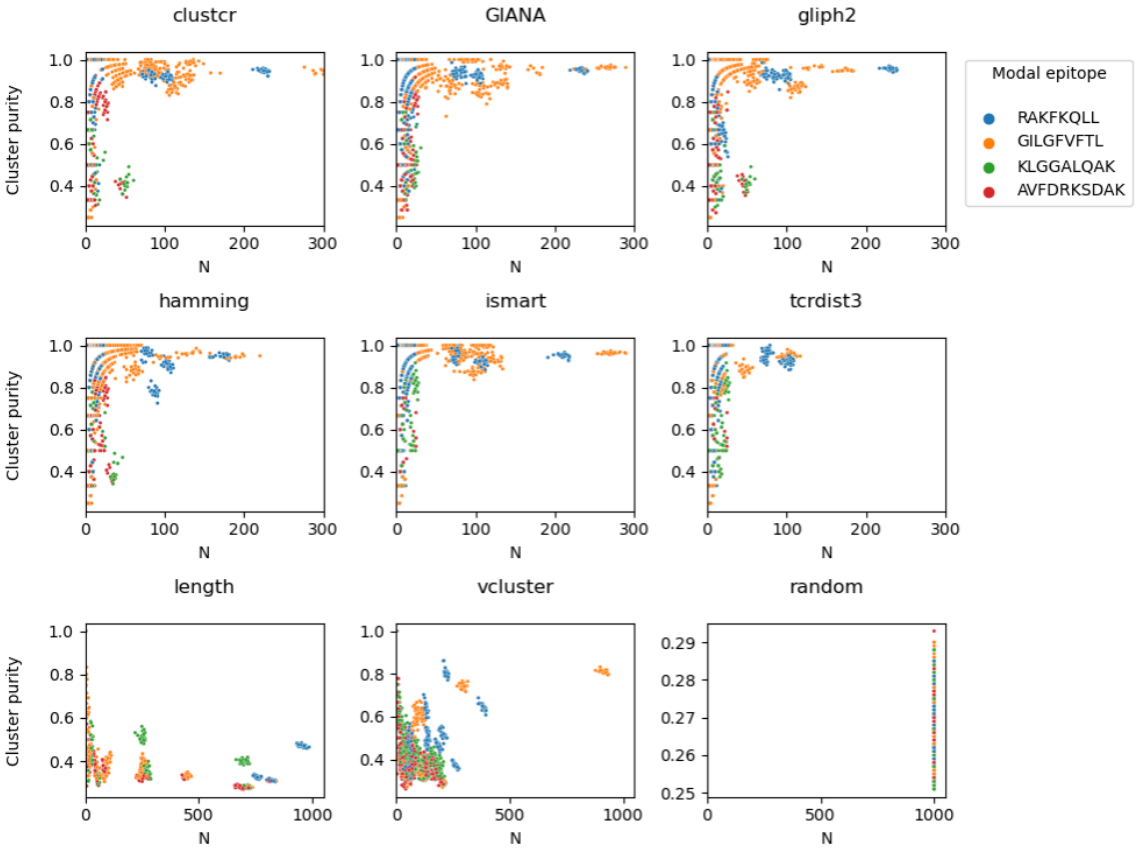


Figure 2-7: Cluster size and epitope purity. Distribution evaluated on V1000 with α and β chain selections combined.

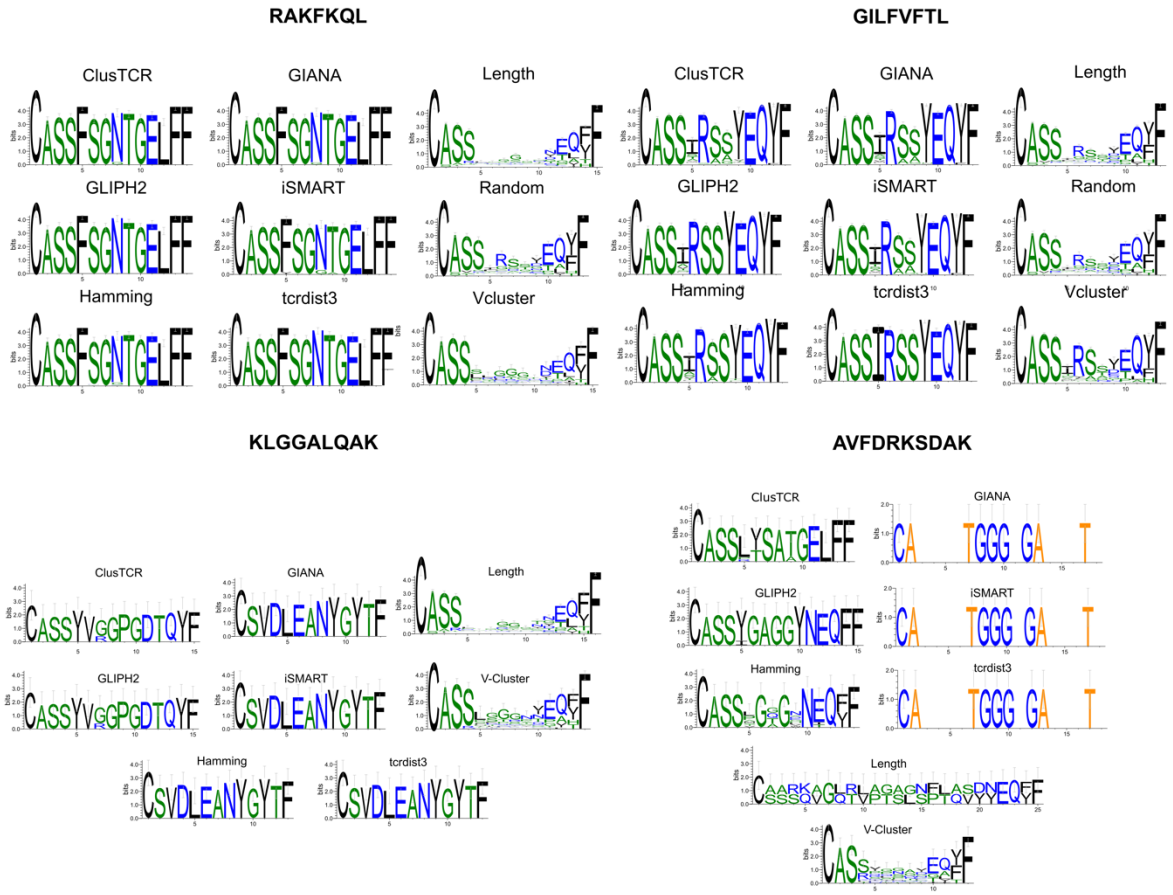


Figure 2-8: Clustering algorithms identify similar logos. CDR3β sequence logos for the largest clusters produced per epitope, using dataset V1000. Logos were produced with WebLogo (204) for TCRs in the largest cluster produced for a given epitope per model following sequence alignment with MUSCLE (203).

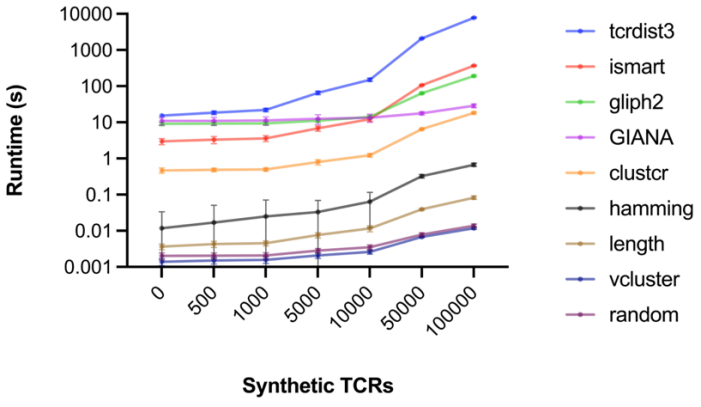


Fig. 2-9: Investigating model scalability. Comparing model runtimes as a function of the number of synthetic TCR sequences introduced with OLGA (195). All experiments conducted on dataset V1000 (β chain selection only).

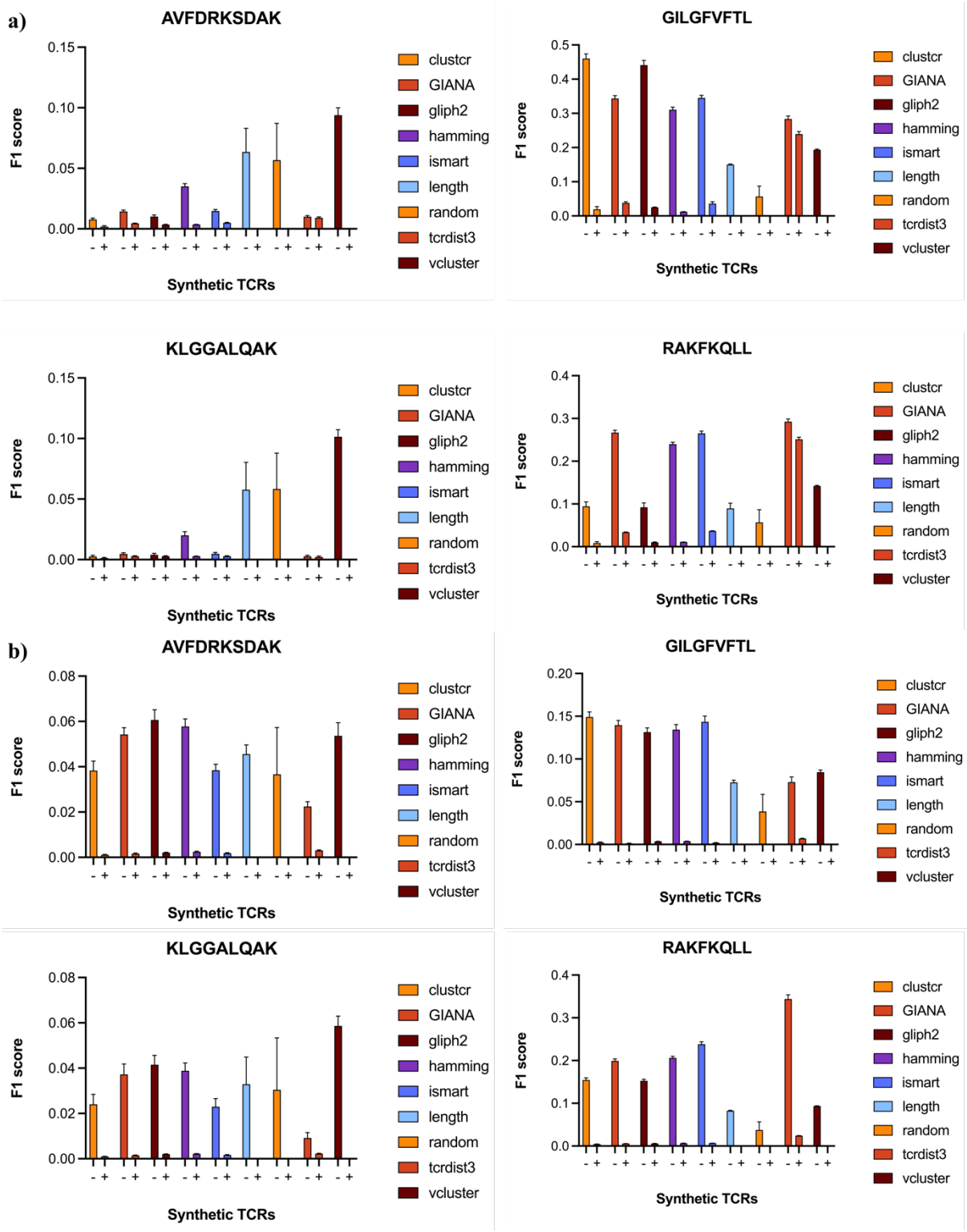


Figure 2-10: Model sensitivity to noise. Showing F1-scores per epitope in the presence or absence of 100,000 synthetic TCR sequences, produced with OLGA (195). All experiments conducted on dataset V1000. **a** | β chains. **b** | α chains.

Discussion

Despite the exponential growth of orphan TCR datasets, and the widespread use of UCMs in de-orphanisation pipelines, an independent comparison of the predictive capacity of clustering models has thus far been missing from the field. Here, we present an initial attempt to address this need. We demonstrate conclusively that clustering models can indeed be used to infer epitope specificity, by co-clustering TCR sequences having known and unknown cognate epitopes and assigning the most common epitope as the predicted binder for a given TCR cluster. Whilst this analytical strategy has been deployed elsewhere (124), we report performance using metrics that can be applied equally to clustering and predictive models, over a wide variety of datasets and preprocessing strategies, which therefore give a more balanced view of true model performance. The comparative analytical framework and model datasets are made freely available at www.github.com/hudsondan/TCR-SCAPES.

We find that model ranking is sensitive to the number of TCRs provided per epitope, to the epitope studied, and to the dataset tested. Whilst GIANA and iSMART perform well on VDJdb instances (**Fig. 2-5**), GLIPH2 provides superior predictions when applied to the large MIRA SARS-CoV-2 dataset. Conversely, Hamming distance and V-gene clustering baselines outperform the study models for McPas-TCR data (**Fig. 2-6**). Only tcrdist3 was able to retain predictive performance in the presence of many synthetic TCRs, for some but not all epitopes (**Fig. 2-10**). This is of particular relevance when applied to the task of extracting expanded antigen-specific clones from a mixed repertoire of effector and antigen-naive T cells in which effector cells are often significantly outnumbered.

Despite these differences, different models produced similar CDR3 motifs for TCRs recognising common viral epitopes, suggesting that one might reach a similar immunological conclusion irrespective of the model used. These mixed results suggest that there is considerable room for improvement in the design of UCMs for inference of TCR specificity.

Finally, although recent structural (26, 205, 206), statistical (182), and predictive (133) analyses suggest that both polypeptide chains play an important role in epitope recognition, we observed consistently lower F1-scores for α compared to β chain selections (**Table 2-4**). One possible explanation is that the default model hyperparameters have been optimised for β chain data, which make up the majority of published TCR-epitope pairs (116). Alternatively, the β chain may simply contribute more to determination of overall epitope specificity, as a product of its increased diversity relative to α chains (16), however this warrants further investigation. For example, modelling strategies that permit integration of α and β chain pairing with transcriptomic and phenotypic information, including graph network approaches such as CoNGA (194), may help efforts to decode the relative contribution of chain pairing to epitope specificity at single cell resolution.

Limitations and further work

The significant scientific and economic potential of a generalisable solution to prediction of TCR epitope specificity has encouraged the development of a multitude of new SPMs and UCMs, summarised in **Chapter 1** and in (65). However, the scope of the present study is limited to a handful of commonly used UCMs, on the basis of their widespread use and relative freedom from training data bias as compared to DNN-UCMs and SPMs. Nonetheless, an independent comparison of UCMs, DNN-UCMs and SPMs would be of great use to the community. We also limit our analysis to performance in an inference task, overlooking a host of useful features such as motif extraction, computation of probability of generation, and visualisation tools, which will likely influence user choice for a given immunological question.

There is also growing evidence that inclusion of both α and β chains improves predictive performance in SPMs and DNN-SPMs (133). However, whilst ClusTCR, tcrdist3, and GLIPH2 may all theoretically be applied to paired chain data, cluster assignments are produced for a given CDR3 independent of the other chain for all but tcrdist3. An investigation of whether the integration of α and β chain information improves performance equally across models might reveal the relative merits of each, when applied to large orphan repertoires. Finally, comparison of the generation probabilities of TCRs drawn from true repertoires and those produced with OLGA have concluded that the latter produces TCRs that are too distant from native repertoires to represent a true background (191). It may be that the unique inclusion of an intermediate pMHC-facing "CDR2.5" loop in tcrdist3 distance calculations, or of a more permissive edit distance than other models, allows the model to differentiate between synthetic and input TCRs and so to cluster TCRs effectively with or without noise. Use of a mixed background (191) or of true orphan repertoires such as those available in (15, 89, 180), might confirm this and so test the robustness of the models to other sources of noise.

Having thus identified limitations in the generalisability of predictive models for inference of TCR specificity (reviewed in **Chapter 1**), and variability in the performance of clustering models (**Chapter 2**), we turn in **Chapter 3** to TCR-directed synthetic display libraries as a rich source of training data for receptor-specific predictions.

Chapter acknowledgements

Our comparative framework borrowed from that developed by the Meysman group for ClusTCR (github.com/svalkiers/clusTCR). Thanks are also due to Professor Hashem Koohy, Dr Mark Basham, Alex Lubbock, Dr Ricardo A. Fernandes and Sam Farrar for critical review, and Andreas Wilm for code testing.

Chapter 3: Statistics and machine learning for inference of T cell antigen specificity from yeast peptide-MHC display libraries

Overview

An individual TCR may interact with a wide variety of peptides in the context of the major histocompatibility complex. This promiscuity presents significant safety concerns for TCR-T development and has hindered previous attempts to predict specificity from sequence data alone. Here, we combine experimental and computational methods to characterise on- and off-target pMHC binding for a diverse panel of natural and engineered immune receptors. The resultant statistical model is capable of accurately identifying agonist peptides from a large reference proteome, and discovery of novel agonists of extensively characterised TCRs, when provided with binding and / or functional pMHC screens for a given receptor.

Introduction

T lymphocytes are capable of recognising antigens derived from intracellular or extracellular antigen pools. This has encouraged use of engineered T cells and TCRs in a wide range of novel therapeutic strategies, and has driven concerted efforts to decode the rules by which their antigens are recognised (65). Indeed, a new generation of TCR-Ts directly incorporate TCRs or TCR-like antibodies, either as soluble bispecific “T cell engagers” or by surface expression, reviewed in (207). However, a critical limitation of these treatments is their reliance on receptors that can recognise multiple distinct pMHC complexes, a property known as *polyspecificity* or *cross-reactivity* (13, 125, 167, 168, 208). Off-target tissue destruction by cross-reactive TCR-Ts has had tragic consequences, including fatal toxicity, in a clinical setting (68, 175, 209). These cases have highlighted the need for reliable experimental TCR-T de-risking strategies, which can be divided broadly into *binding screens* using pMHC multimers (84), bacteriophage or yeast display (104, 105, 107); and *functional screens* including cytokine production, fluorescence reporter, proliferation and killing assays, reviewed in (80). Whilst computational predictors abound, open-source algorithms are currently limited in their capacity to reliably infer the specificity of, or *de-orphanise*, TCRs beyond a narrow subset of well-characterised epitopes and MHC alleles from publicly available datasets (65, 93).

In this chapter, we turn to yeast pMHC display to profile the binding specificity of a given therapeutic candidate, using affinity-driven selection with randomised libraries of over 1 billion peptides in a fixed HLA context. This study builds on previous work using yeast display to

map the binding profiles of cross-reactive TCRs (104, 105, 109, 112, 210), as summarised in **Fig. 3-1**. This methodology was originally conceived by Brophy, Holler, and Kranz (104), who demonstrated that murine pMHC complexes expressed as single polypeptide chains fused to yeast Aga2 protein were recognised by a high-affinity TCR (2C) in a manner that directly activated naïve $2C^{+}$ T cells. Later, Michael Birnbaum and colleagues used Ob.1A12 and Ob.2F3 TCRs isolated from patients with multiple sclerosis to screen transgenic yeast expressing peptides bound to HLA-DR15 (105). The output of synthetic display screens such as these is typically represented as either *heat maps* (two-dimensional representations of the observed frequencies of amino acids at a given position, see **Fig. 3-1**); as *peptide motifs* (for example E-S/V-D-P-I-x-x-x-Y for receptors binding the MAGE-A3 peptide EVDPIGHLY in the context of HLA-A*01); or as *logos* in which the height of a given residue is proportional to its frequency in the input data (204). However, whilst these representations provide insight into the binding preferences of a given receptor, they do not provide a direct prediction of binding to target peptides from a reference proteome. To solve this problem, the frequency of amino acids at a given position in the enriched peptides can be recorded as a two-dimensional L x A grid (see **Methods**), wherein L is the length of the peptide and A the number of commonly occurring natural amino acids (twenty). The resultant matrix can then be used to generate scores for each peptide in a reference proteome (see **Eq. 3-1**), producing a ranked list of predicted binders. Birnbaum and colleagues used this methodology to identify six self-peptides that activated Ob.1A12 and Ob.2F3 TCRs *in vitro*, including a self-ligand derived from Myelin Basic Protein (MBP).

The use of synthetic display libraries as an input for the prediction of immune receptor specificity is now well-established (69, 104, 105, 107, 109, 112, 210, 211). Historic approaches with yeast selection data include the use of simple statistical methods such as PWMs (105, 109, 210), and (more recently) deep learning models such as NNalign (112). Binding motifs derived from phage display screens have been leveraged to deorphanise TCRs specific to NY-ESO-1 (107) and KRAS (212) cancer antigens. Other studies have deployed shallow peptide libraries produced with X-scanning mutagenesis (XSM) to explore the reactivity profiles of TCR-Ts using binding and/or functional screens, applying either statistically inspired predictors or web tools to scan a reference proteome (69, 211).

Despite this progress, to date there has been no analysis of the relative capacity of PWMs or machine learning models to retrieve known functional agonists of well-characterised receptors, nor the capacity of the models developed to perform "off-the-shelf": each model being developed to make predictions from a given dataset. This is potentially problematic for researchers seeking tools for de-orphanisation of novel receptors of unknown specificity, as the capacity of models to perform across varied datasets remains unknown. Finally, existing open-source models make use of affinity-driven selection data from large, random peptide libraries, or functional data for a narrower set of peptides. However, in practice either or both may be available, and the utility of binding and/or functional data remains an open question in the field.

In the present work, we exploit large random yeast pMHC libraries of over one billion peptides as a source of rich, debiased binding information for a given TCR (**Fig. 3-1**). These deep datasets inform the development of the Orpheus computational pipeline, which is capable of accurately retrieving known agonists of a panel of diverse receptors from a large reference proteome with no need for model parameter optimisation. Applying the pipeline to affinity-driven selection data reveals that binding information is sufficient to retrieve the agonists of known TCR-Ts, however we find that functional data can be integrated to further improve model performance.

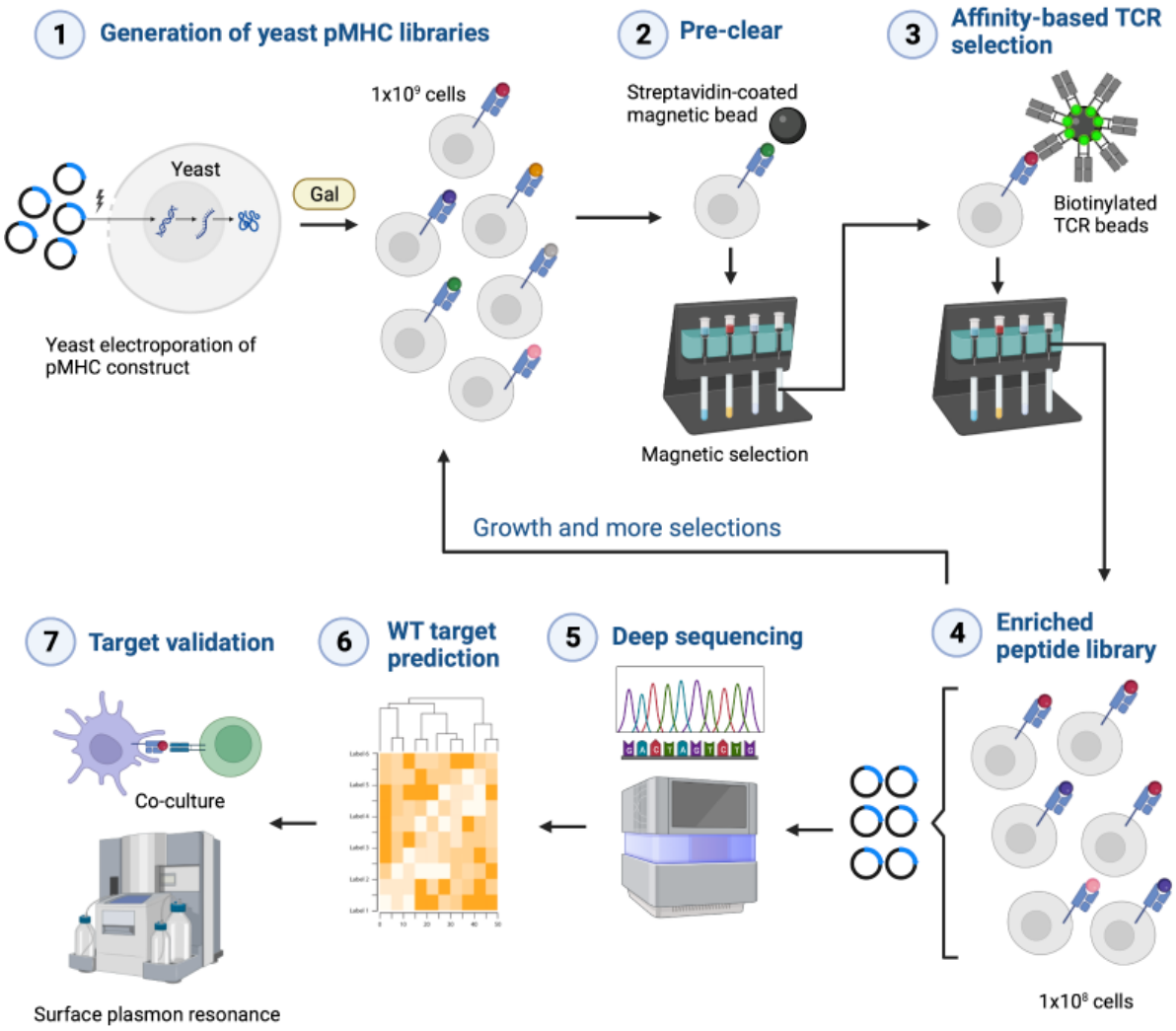


Figure 3-1: Yeast pMHC screens for experimental TCR de-orphanisation. Details of the yeast library preparation and selection protocol are provided in **Methods**.

Methods

Datasets

This study explores the capacity of predictive models to retrieve the known antigens of well-studied self-reactive immune receptors whose binding preferences have been characterised using yeast display screens. This includes selection data for the a3a (68, 213) and DMF5 (174) receptors generated by the Fernandes group as described below, and published datasets for DMF5 (109); the HLA-E specific CD94/nkg2c activatory NK cell receptor (112); a wildtype HLA-B*27 restricted TCR identified in patients with ankylosing spondylitis (as4.3) (110); and the HLA-DR15 restricted MBP-specific TCRs Ob.1A12 and Ob.2F3, characterised in (105). Raw data for as4.3 and Ob TCRs were kindly provided by Dr Xinbo Yang (Memorial Sloane Kettering Institute, US) and Professor Michael Birnbaum (Massachusetts Institute of Technology, US), respectively.

Yeast peptide MHC selection data were generated within the Fernandes lab for the a3a and DMF5 receptors as follows, with further methodological detail reported previously (214). HLA-A*01:01 and HLA-A*02:01 single chain trimer constructs were produced for peptides of a fixed length and assuming fixed anchor residues as described in **Table 3-1**, otherwise randomising to any of the 20 most common natural amino acids except cysteine. *S. cerevisiae* were transfected by electroporation with digested pYal plasmid and DNA encoding for the peptide, B2M, and MHC in a single synthetic polypeptide joined with flexible linkers. Transfected yeast were grown at 30°C in SDCAA media, induced at 20°C in SGCAA media for ~24-48 hrs and selected using recombinantly purified and biotinylated TCR bound to streptavidin-coated microbeads. The expression of the peptide MHC construct was confirmed by fluorescence-activated cell sorting, staining with anti-cMyc antibodies. After three to four consecutive selection rounds with microbeads, DNA was extracted from all rounds for deep sequencing to determine the abundance of each peptide.

Table 3-1: Comparing receptors and library preparation methods. Total reads: summed read counts for all peptides at the third round of selection; Max reads: the maximum read count for any enriched peptide in the third round of selection; Unique hits: peptides having five or more reads in the third round of selection.

Receptor	a3a	DMF5	nkg2c
HLA	A*01:01	A*02:01	E*01:03
Peptide length	9	10	9
Primary self-agonist	EVDPIGHLY MAGE-A3	EAAGIGILTV MART-1	VMAPRTLFL VL9-1
K_D	2μM	11μM	24μM
Total reads	794415	52307	1394002
Max reads	11541	3337	103978
Unique hits	4200	1283	3590
Assumed anchors	P3 D/E P9 Y	P2 L/M P10 L/M/V/A	N/A

Receptor	as4.3	Ob.A12	Ob.F3
HLA	B*27:05	DR15	
Peptide length	9	14	
Primary self-agonist	GQVMVVAPR RNASEH2B	ENPVVHFFKNIVTP MBP	
K_D	15μM	Not reported	
Total reads	335619	226708	362543
Max reads	1008	34723	114216
Unique hits	4716	686	614
Assumed anchors	P2 R/Q/K, P9 R/K/M/I/L/ Y/F/W/V	P1 I/L/V P4 F/Y	

Models

PWMs

The Orpheus pipeline takes tables of peptide abundance in *CSV* format as its input. Peptides having five or more reads in the naïve library were excluded from input data, as were those whose nucleotide sequence matched known contaminants (see **Chapter 4**). Enriched peptides were defined as those having five or more reads in the third selection round, consistent with (109). PWMs were extracted from enriched peptides in the resulting dataset as described in **Eq. 3-1**. Anchor switching was achieved by substitution of the PWM values at MHC allele-specific anchor positions (**Table 3-1**) for the frequencies of amino acids observed in elution mass

spectrometry data obtained for that HLA allele from a combination of two datasets (20, 215), allele-specific PWMs being produced as described in **Eq. 3-1a**. A consensus human proteome was sourced from UniProt (114) and split into 9-mer, 10-mer or 14-mer libraries containing 1.12 x 10^7, 1.13 x 10^7 or 1.46 x 10^7 unique peptides, respectively. Scores were obtained for each peptide in the reference proteome as described in **Eq. 3-1b,** returning a list of peptides in order of decreasing probability of recognition.

Performance was evaluated based on the absolute ranking of agonists reported for a3a (108, 213, 216), DMF5 (109, 217, 218), nkg2c (112), as4.3 (110), Ob.A12 or Ob.F3 (105) receptors. All experiments were run on a single remote Intel(R) Xeon(R) Gold 6126 CPU at 2.60GHz. Plots were generated using the Matplotlib (201) and Seaborn (202) Python packages.

Equation 3-1: Position Weight Matrices.

a. $\mathbf{PWM} = \{\mathbf{p}(\mathrm{a}, \mathrm{l})\}\ \forall\ \mathrm{a}\ \in \mathrm{A}, \mathrm{l} \in \mathrm{L}$, where

$$\mathbf{p}(\mathrm{a}, \mathrm{l}) = \frac{\mathbf{p}(\mathrm{a} \mid \mathrm{l})}{\mathbf{p}(\mathrm{a})}$$

$$\mathbf{p}(\mathrm{a} \mid \mathrm{l}) = \frac{|\mathrm{a_i},\ \mathrm{l}|}{\sum_{\mathrm{j}} |\mathrm{a_j}, \mathrm{l}|} \forall\ \mathrm{a} \in \mathrm{A}$$

$$\mathbf{p}(\mathrm{a}) = \frac{|\mathrm{a_i}|}{\sum_{\mathrm{j}} |\mathrm{a_j}|} \forall\ \mathrm{a} \in \mathrm{A}$$

Where **PWM** is a positional weight matrix; **p**(a, l) is the frequency of amino acid *a* belonging to the set of A common amino acids at position *l* in a peptide of length L; **p**(a | l) is the probability of observing amino acid *a* at position l; p(a) is the probability of observing amino acid a at any position; and |a| is the count of amino acid *a* in the dataset; such that a given PWM has dimensionality A x L.

b. $\mathbf{Score}(\mathrm{p}) = \prod_{\mathrm{l}=1}^{\mathrm{L}} \mathbf{PWM}(\mathrm{a}_l, l)$

Where p is a peptide of length L drawn from a reference proteome

c. $\mathbf{PWM_b} = \left\{\mathbf{p}(a_i, l) + \left(\sum_{i,j} \mathbf{p}(a_i, l) * norm(B62\left(a_i, a_j\right))\right)\right\} \forall\ a_i\ \in A, a_j \in A_{allowed},\ l \in L$

Where $\mathbf{PWM_b}$ is a PWM adjusted for evolutionary coincidence using a BLOSUM62 (B62) lookup matrix; norm is a minmax normalisation function such that the values of a B62 substitution matrix are adjusted to fall between 0 and 1; and $A_{allowed}$ is the set of amino acids having B62 scores > 0.

d. $\mathbf{PWM}_{final} = \sum_{c \in C_{tk}} \mathbf{PWM_{cb}} \cdot \sum_{p \in c} \boldsymbol{r_p}$

Where $\mathbf{PWM_{final}}$ is the final optimised PWM; $\mathbf{PWM_{cb}}$ is the PWM_b calculated for a cluster c of peptides having a Hamming distance of < 2 drawn from the set C_{tk} of clusters meeting the threshold number of

reads *t* (default 5) multiplied by the minimum cluster membership *k* (default 2); and *r* is the number of reads associated with peptide *p*.

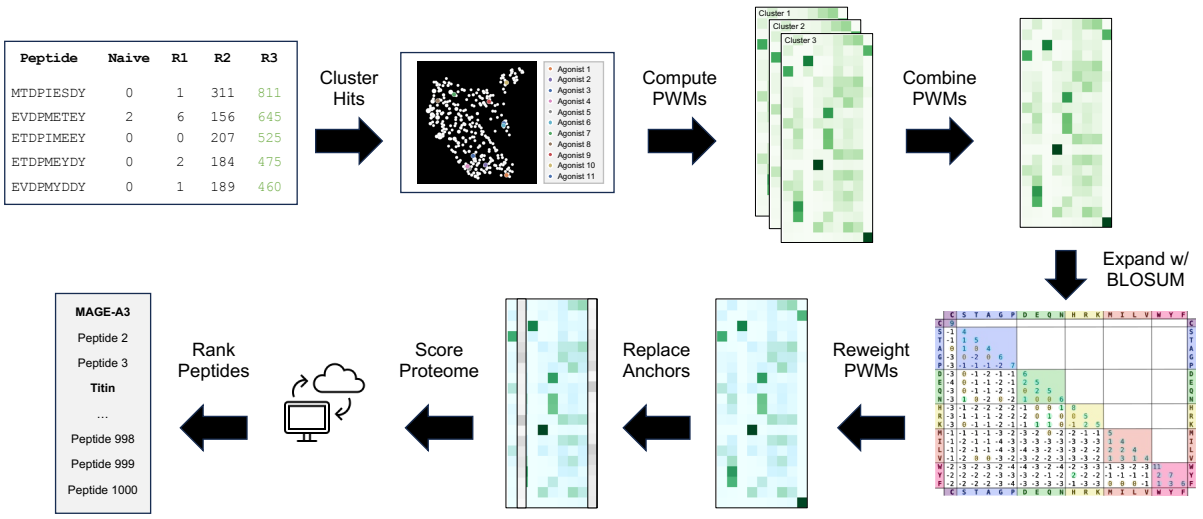


Figure 3-2: Schematic of the Orpheus predictive pipeline. PWM: Positional Weight Matrix.

Machine learning

A grid search was conducted of commonly used machine learning models and sequence representations from the scikit-learn library (see **Results**) (197). Negative instances were defined as peptides absent in the enriched peptide population and having fewer than five reads in the naïve round. Each model was trained on negative and enriched peptides drawn from the yeast screen provided at a 5:1 ratio, and each peptide in the consensus proteome was ranked according to the likelihood of binding predicted by the trained model. Peptides were represented either as one-hot encodings using the scikit learn library (197), integer encodings (produced using a custom Python script), BLOSUM62 encodings, or as a combination of BLOSUM62, isoelectric point and hydrophobicity scores. BLOSUM62, isoelectric point and hydrophobicity scores were generated using the Peptides Python library. Model runs were performed in triplicate to account for random negative sampling, and the final peptide ranking was determined as the mean score (the probability of predicted binding) across runs.

Experimental assays

The stimulatory potential of candidate agonists, identified either as top enriched peptides in yeast pMHC screens or as top-ranked predicted hits for a given receptor, was evaluated by Julia McCarthy and Kristen Koopmans as described in (214). Briefly, $CD8^+$ Jurkat T cells were modified to express the TCR construct of interest using lentiviral transduction. The resulting

a3a or DMF5-positive Jurkats were incubated with HLA-A*01:01^{+} or HLA-A*02:01^{+} K562 APCs pulsed with a single concentration of peptide for 14-16 hours at 37°C. Expression of the αβ TCR and activation markers CD69 and CD45 by the Jurkat T cell population was then analysed using an Attune flow cytometer (Thermo Fisher).

Integrating binding and stimulation data

%CD69 values were normalised by subtracting control values for each known agonist of the a3a TCR, and the resulting values passed through a sigmoid function (**Eq. 3-2**) to favour highly stimulatory peptides and to disfavour weak agonists with profiles close to the negative control. Stimulation data were converted to a synthetic count matrix by multiplying the normalised stimulation values by the maximum read in the third round of selection. This matrix was then concatenated with the raw yeast pMHC data and passed to the Orpheus pipeline as described above.

Equation 3-2: Integration of binding and functional data

a. $P_{tot} = P_{bind} + P_{stim}$

b. $P_{stim} = R * \max(P_{bind;round}) * \sigma\left(S(p) - S(c)\right) \nabla\, p \in P$

c. $\sigma(x; L, k, p_0) = \frac{L}{1+e^{-k(x-x_0)}}$

Where P_{tot} is an updated peptide count table; P_{bind} is the raw count data from the selection screen; P_{stim} is a synthetic count table produced by multiplying a ratio R (default 1) by the maximum count in the target selection round (default round 3), S are stimulation values (% CD69) for a peptide *p* belonging to the set of known agonists P for a TCR, relative to a negative control c; σ is a sigmoid transform having constants L, k and x_0 with default values 100, 0.1 and 50.

Results

Yeast peptide MHC screens

The Orpheus pipeline was developed using yeast pMHC screens of the a3a (213), DMF5 (174), and nkg2c receptors (112), obtained as described in **Methods** and summarised in **Table 3-1**. These proteins were selected firstly as they demonstrate functionally validated activation by multiple distinct pMHCs, and secondly as their agonists span peptides of differing lengths, presented by different MHC subtypes, over a range of solution-state binding affinities reflective of their origin as engineered (a3a, DMF5) or wildtype (nkg2c) receptors. Briefly, the a3a TCR is an affinity-enhanced TCR-T that recognises the MAGE-A3 cancer testis antigen presented

by HLA-A*01 (213). Molecular mimicry of MAGE-A3 (EVDPIGHLY) by Titin (ESDPIVAQY), expressed in beating heart cardiomycocytes, led to severe cardiotoxicity events early in clinical development (68). DMF5 was derived from a MART1 melanoma antigen-specific TCR, which was later found to engage two biochemically and structurally distinct peptide classes in the context of HLA-A*02 (109, 217). The CD94/nkg2c receptor is an activatory NK receptor capable of agonism by diverse peptides presented by HLA-E*01 and which is implicated in overcoming immune evasion by cancer and certain viruses (112). as4.3, a naturally occurring TCR identified from HLA-B*27 positive patients with ankylosing spondylitis (110), as well as Ob.A12 and Ob.F3 MBP-specific TCRs, were included as a discrete test case. The preferences of each receptor for different residues at each position in a peptide are shown in **Fig. 3-3**.

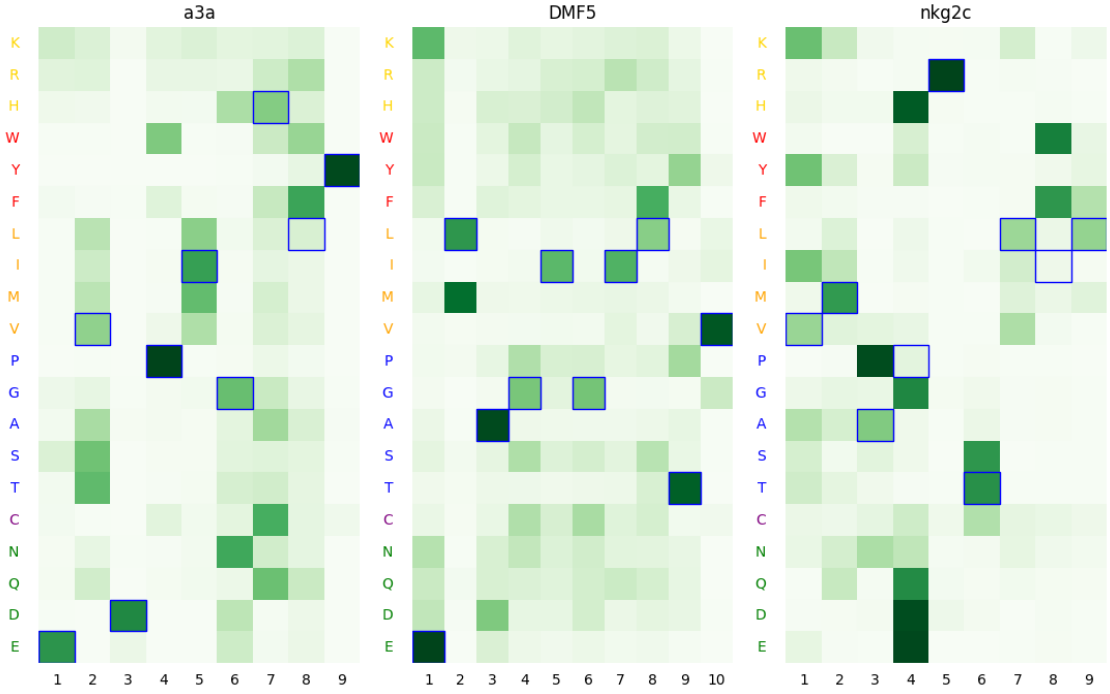


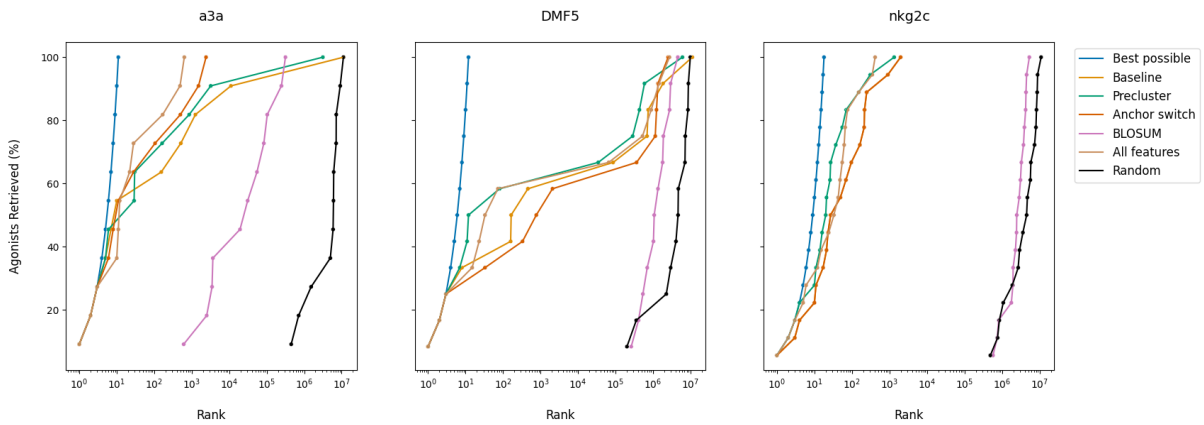
Figure 3-3: Heatmaps showing the positional preferences of the a3a, DMF5, and nkg2c receptors. Positions corresponding to the primary agonists EVDPIGHLY (MAGE-A3), ELAGIGILTV (MART1), and VMAPRTLIL (VL9) are highlighted with blue boxes.

Optimisation of the predictive pipeline

Our baseline algorithm uses a simple PWM in which the normalised frequency of amino acids at each position in a pool of enriched peptides of a given length was recorded in a two-dimensional matrix (**Eq. 3-1a**). The resultant matrix was then used to generate scores for reference peptides as the product of the frequencies of the corresponding residue (**Eq. 3-1b**).

Performance was evaluated in a ranked retrieval task, such that each model was first built with data from a given receptor screen and then used to score and rank all potential hits from a reference proteome of 11 million self-peptides (**Fig. 3-1** and **Methods**) (114). The absolute ranking of experimentally validated agonists was reported for all datasets and model configurations. Assuming that binding is necessary but not sufficient for agonism, an ideal model would retrieve all known functional agonists of a given TCR in as few predicted binders as possible.

The baseline predictor performed well "off-the-shelf", retrieving the primary agonists of a3a, DMF5, and nkg2c at ranks 1, 1, and 3 of over eleven million reference peptides, respectively, with no need for parameter tuning or changes to data preprocessing between datasets (**Fig. 3-3**, **Tables 3-2**, **3-3**, and **3-4**). However, performance was poor for some agonists, particularly those whose anchor residues violated the assumed preferences of the library (**Fig. 3-3**, **Tables 3-1** to **3-4**). The algorithm was therefore adapted to allow for substitution of evolutionarily related amino acids during PWM computation using a BLOSUM62 matrix (**Eq. 3-1c**); to upweight the contribution of less well-represented peptide clusters (**Eq. 3-1d**); and to loosen the rigid anchor assumptions imposed on the library (*anchor switching*, **Fig. 3-1** and **Methods**). These modifications dramatically improved the quality of predictions (**Tables 3-2, 3-3** and **3-4**), the final model retrieving all known agonists of the a3a and nkg2c receptors by rank 634 and 413 respectively, within the top 0.01% of the reference proteome. Whilst BLOSUM62 substitution alone performed poorly, combining all features provided the best results for a3a and ngk2c (**Fig. 3-3**, **Tables 3-2** and **3-3**). Retrieval was less effective for DMF5 (**Table 3-4**), where all MART1-like [I/L]GI[L/M/V] containing agonists were identified by rank 73, but those containing a subdominant DRG motif reported in (109) were retrieved only by rank 2,767,277 (within the top 24%). Performance was improved for DRG-containing peptides when applied to the original study data for DMF5 (109), but to the detriment of the physiological agonist MART1, which was retrieved at rank 4373, albeit still falling within the top 0.04% of the reference proteome (**Table 3-5**). This trade-off suggests that failure in DRG peptide detection may be due to experimental rather than computational constraints.



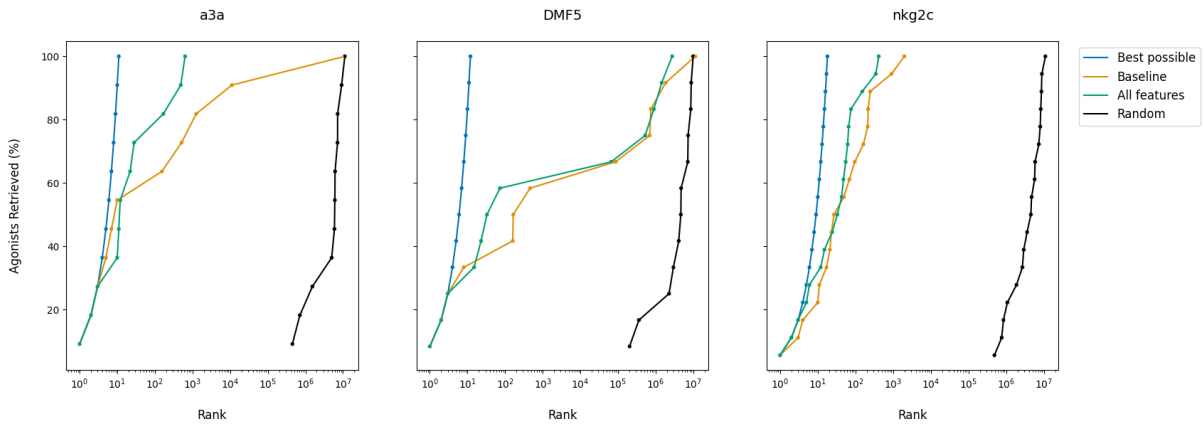


Figure 3-3: **Ranked retrieval of known agonists of the a3a, DMF5 and nkg2c receptors**. Retrieval traces show the proportion of known agonists (y axis) retrieved by a given position (x axis, filled circles show exact rankings) in the reference proteome, compared to the best possible ranking and to a random shuffle of the reference. **Top** | Performance by model architecture. **Bottom** | Baseline and combined (all feature) configurations only.

Table 3-2: Best rank per dataset by model type, a3a. Highlighted epitopes: MAGE-A3 (EVDPIGHLY) and Titin (ESDPIVAQY).

Peptide	Baseline	Pre-cluster	Anchor switch	BLOSUM	All features
EVDPIGHLY	1	1	1	31048	1
EVDPIGHVY	2	2	2	19409	2
EMDPVTQLY	3	3	3	610	3
EVDPIRHYY	5	30	6	3484	11
ESDPIVAQY	10	5	11	3665	12
ETDPVENTY	7	6	8	83875	10
ETDPVNHMV	1253	856	28	83875	22
ETDPLTFNF	NA	3149716	105	247158	28
EGDPLILQY	154	29	498	54998	168
EMDPFHYDY	511	162	2408	103545	487
ELDPILTEV	11120	3216	1515	316324	634

Table 3-3: Performance per dataset by model type, nkg2c. Highlighted epitope: The MHC class I VL9 signal peptide.

Peptide	Baseline	Pre-cluster	Anchor switch	BLOSUM	All features
TLPERTLYL	17	21	17	804464	1
VNPGRSLFL	1	2	1	1931639	2
VMAPRTLFL	4	1	4	3066555	3
ILTDRSLWL	10	4	10	2836281	5
VMPPRTLLL	11	3	11	1751579	6
VMPGRTLCF	3	11	3	4465985	12
TAPARTMFL	49	16	49	1953902	15
TGPWRSLWI	246	151	246	3685752	24
VMAPRTLLL	21	10	21	3238350	33
VMAPRTLIL	27	14	27	3905716	43
VMAPRTVLL	23	20	23	2406843	48
RMPPRSVLL	96	26	96	569687	55
VMAPRTLVL	69	27	69	2443194	62
FLPNRSLLF	211	70	211	4303683	66
TLPKRGLFL	901	55	901	4205934	76
VMAPRALLL	216	37	216	5282220	151
QMPSRSLLF	1973	1329	1973	735599	346
NMPARTVLF	162	307	162	2251320	413

Table 3-4: Best rank per dataset by model type, DMF5 (screen conducted by the Fernandes lab). Highlighted epitope: MART-1.

Peptide	Baseline	Pre-cluster	Anchor switch	BLOSUM	All features
ELAGIGILTV	1	1	1	706363	1
SLANIGILPV	2	2	3	413474	2
EAAGIGILTV	3	3	2	1822061	3
NLSNLGILPV	160	11	332	541691	15
SMAGIGIVDV	8	7	33	1085542	23
SMLGIGIVPV	165	12	773	264321	33
NMGGLGIMPV	462	81	2092	1905577	73
IMEDVGWLNV	86791	34774	367119	1369507	66321
LMFDRGMSLL	697808	288597	1158123	1035009	528578
ILEDRGMGLL	1864605	437858	2556864	4618205	902117
MMWDRGMGLL	751135	596821	1250853	2791186	1458801
MMWDRGLGMM	NA	6158827	1366909	2973600	2767277

Table 3-5: Best rank per dataset by model type, DMF5 (screen conducted by the Garcia lab as reported in (109)). Highlighted epitope: MART-1.

Peptide	Baseline	Pre-cluster	Anchor switch	BLOSUM	All features
LMFDRGMSLL	2	1	4	1360254	1
MMWDRGMGLL	1	2	1	1932672	2
ILEDRGMGLL	9	3	3	4722396	3
MMWDRGLGMM	4	1006	17	1388963	4
SLANIGILPV	43	22	61	2928301	11
SMLGIGIVPV	6	4	82	3376735	66
IMEDVGWLNV	97	9	2278	2735988	320
SMAGIGIVDV	16	6	454	4600221	506
NLSNLGILPV	792	423	6193	2527176	1042
ELAGIGILTV	351	89	1630	6451217	4373
EAAGIGILTV	NA	2244	5721	6311001	17695
NMGGLGIMPV	NA	222	285347	4676048	31865

The Orpheus PWM receives no prior information for a given input screen and so cannot benefit from "overfitting" or leakage between datasets. However, it is possible that the preprocessing and model parameters were over-engineered to suit the three yeast pMHC screens analysed. Independent test sets were therefore obtained from a screen of as4.3, a wildtype HLA-B*27-restricted TCR isolated from patients with ankylosing spondylitis (110), and from class II restricted Ob.A12 and Ob.F3 TCRs (105).

For as4.3, the combined model retrieved the primary self-agonist RNASEH2B by rank 47, and all but two reported agonists within the top 0.004% (**Table 3-6**). Interestingly, anchor switching appeared to damage performance on as4.3 agonists DRQQLPAPK and GRVGDLSPR, despite these peptides conforming with the anchor preferences of HLA-B*27 (**Table 3-1**). The top-ranked predictions of the Orpheus model on this dataset have not yet been functionally validated, which might reveal further clues as to the rules of recognition by this naturally occurring TCR.

A combined model provided with data from the MHC class II-restricted Ob.F3 and Ob.A12 TCRs retrieved MBP at rank 3 and 95 respectively. All reported agonists were retrieved by rank 34259 and 9729, within the top 0.2% of hits. The baseline model performed better for most agonists (**Fig. 3-4**), perhaps because elution mass spectrometry data were not identified for HLA-DR15 and so no anchor switching was performed with these datasets.

Table 3-6: Best rank per dataset by model type, as4.3. The primary human target of the as4.3 TCR, RNASEH2B, is highlighted.

Peptide	Baseline	Pre-cluster	Anchor switch	BLOSUM	All features
GQVMVVAPR	956	1108	130	403624	47
TRLALIAPK	2	2	12	31616	48
LRQMLQAPH	NA	575337	221	120441	87
VQVMLNNPK	1233	972	421	34598	218
TQMPLVAPV	NA	755482	301	986867	276
RRVILRAPY	NA	676613	1126	222975	429
DRQQLPAPK	241	101	2049	1879	30511345
GRVGDLSPR	5769	21834	41559	435685	155647171

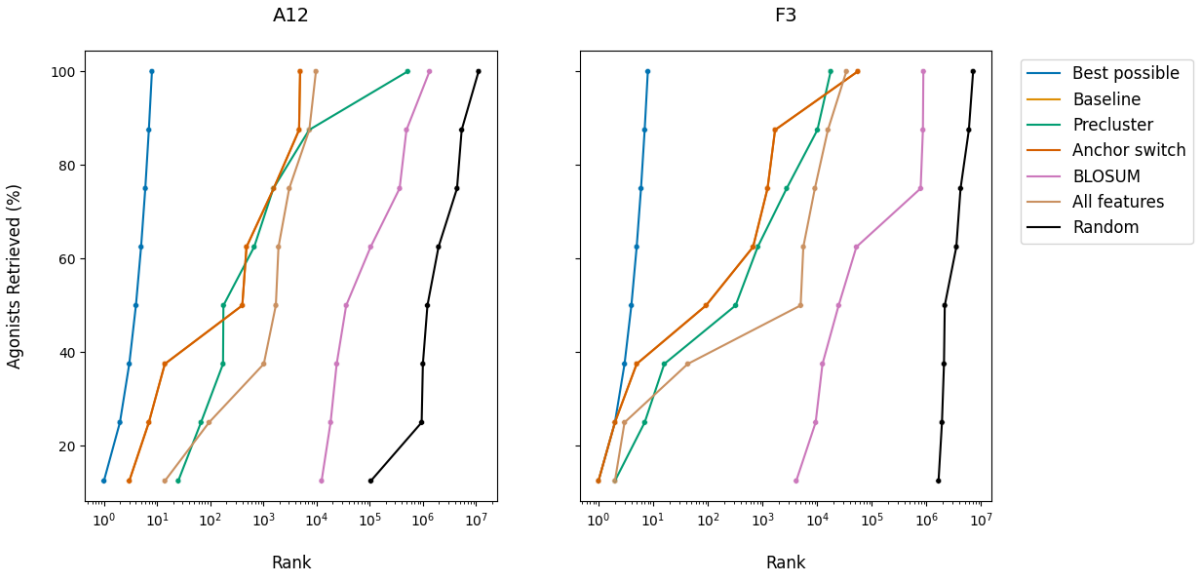


Figure 3-4: Agonist retrieval by DR15 restricted Ob TCRs.

Table 3-7: Best rank per dataset by model type, Ob.A12. The primary human target of the Ob.A12 TCR, MBP, is highlighted.

Peptide	Baseline	Pre-cluster	BLOSUM	All features
NANVLHFLKNIICQ	3	25	12432	14
ENPVVHFFKNIVTP	7	67	23894	95
MNASIHFLKALETY	481	174	18358	1018
IIPAFHFLKSEKGL	401	177	1331101	1714
SANNIHFMRQSEIG	14	1540	104093	1923
FLKKFHFLKGATLC	1562	517623	490986	3060
HMLSFHFWKSRGQT	4908	7248	36068	7307
APLVIHFLKAPPAP	4688	670	364766	9729

Table 3-8: Best rank per dataset by model type, Ob.F3. The primary human target of the Ob.F3 TCR, MBP, is highlighted.

Peptide	Baseline	Pre-cluster	BLOSUM	All features
NANVLHFLKNIICQ	1	2	12538	2
ENPVVHFFKNIVTP	5	16	9399	3
MNASIHFLKALETY	2	7	4148	43
FLKKFHFLKGATLC	674	819	861856	4993
SANNIHFMRQSEIG	1238	10120	52346	5593
APLVIHFLKAPPAP	93	323	878375	9100
IIPAFHFLKSEKGL	1702	2772	782766	15748
HMLSFHFWKSRGQT	55465	17715	24758	34259

Whilst our simple statistical model exhibited impressive retrieval performance across class I and class II restricted receptors, recent years have witnessed widespread use of machine learning tools for predictive modelling tasks. To assess whether such approaches could boost agonist retrieval from yeast display, a grid search was carried out using a suite of off-the shelf predictors available in scikit-learn (**Figs. 3-5** and **3-6**) (197). Each of these models was trained using positive instances (all enriched peptides), and a random selection of negative peptides (see **Methods**). Stochasticity in negative sampling led to variations in model performance when applied to the same datasets, and so peptide rankings were determined as the mean score of a given model after three repeats. After extensive engineering, no machine learning approach was identified that provided consistent improvements in performance over the optimised PWM, either alone (**Fig. 3-5**) or in combination with the best PWM (**Fig. 3-6**).

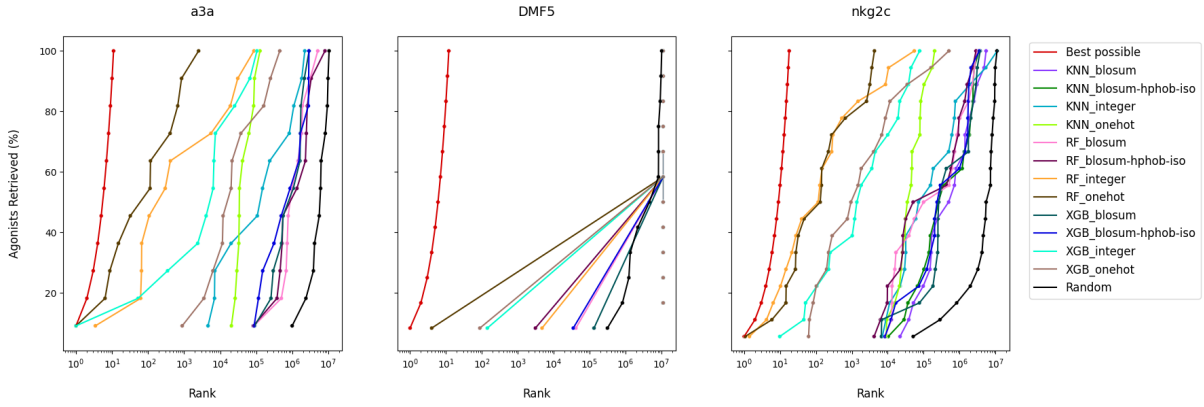


Figure 3-5: Evaluating machine learning models for retrieval of known agonists of the a3a, DMF5 and nkg2c receptors. KNN: K nearest neighbour, RF: Random Forest; XGB: XGBoost; BLOSUM+hphob+iso: a combination of BLOSUM62 encoding, hydrophobicity score, and isoelectric point for a peptide, computed using the Peptides Python library.

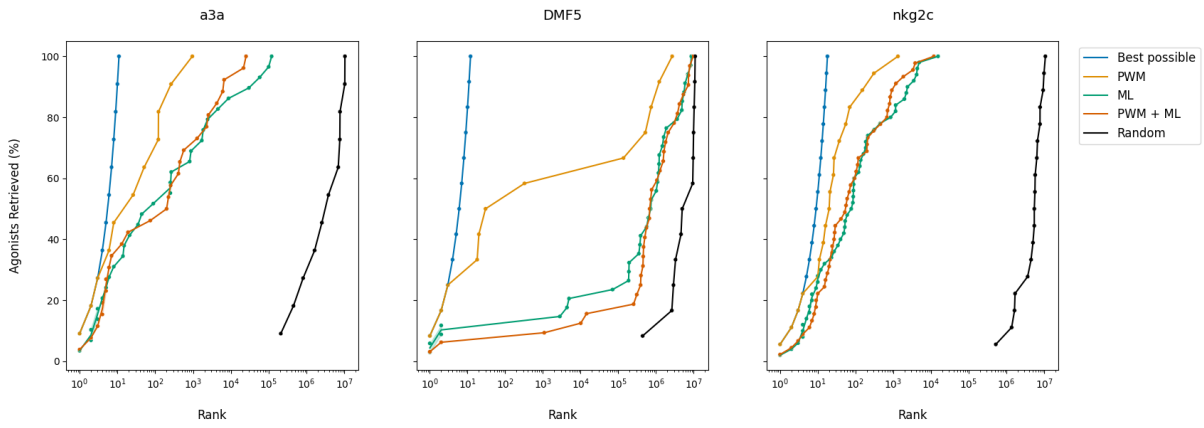


Figure 3-6: Performance of an optimised PWM compared with the best machine learning (ML) model in ranked retrieval of known agonists. "ML": an ensemble of three random forest models trained using the same set of enriched peptides as positives, and a random selection of negatives drawn from the naïve selection round at a 5:1 ratio (see **Methods**).

Validating predicted hits

Having thus demonstrated the capacity of the Orpheus pipeline to retrieve the cognate agonists of diverse receptors, we next evaluated the immunogenic potential of the most enriched peptides and top ranked predicted binders. To do so, the a3a receptor was introduced into TCR-Jurkat T cells through lentiviral transection, before co-culture with peptide-pulsed HLA^{+} K562 cells (see **Methods**). These analyses confirmed stimulation by 7 of the 10 previously reported human agonists (**Table 3-3**), as well as identifying a previously undescribed human agonist of the a3a TCR (EMDPFHYDY, derived from the FXYD6 protein, **Fig. 3-7**) identified using the Orpheus pipeline; and an additional 41 synthetic agonists (or synthetic mimotopes) identified through yeast pMHC display (**Table 3-7**).

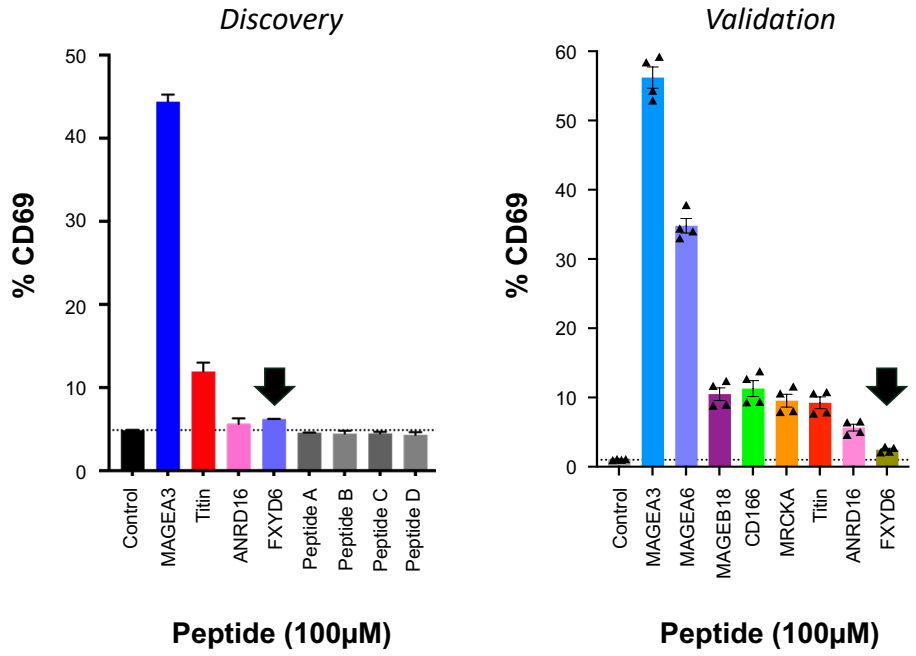


Figure 3-7: Discovery of a novel human agonist of the a3a TCR. Left | Stimulation data for peptides retrieved among the top predicted hits of the Orpheus pipeline when trained on yeast pMHC data for the a3a TCR. **Right |** Stimulation data for known human agonists of the a3a TCR, including the newly discovered FXYD6.

Integrating display and functional data

Whilst peptide presentation by MHC can sometimes be taken as a proxy for immunogenicity, not all enriched peptides were capable of stimulating antigen-specific T cells (**Fig. 3-7**). To explore whether co-culture data could be exploited to improve retrieval of functionally validated agonists, we integrated yeast pMHC selection and stimulation data for the a3a TCR using a simple transform permitting conversion of %CD69 values to a projected binding readout (**Eq. 3-2**). Integration of binding and immunogenicity data in this manner improved ranked retrieval over the best model (**Fig. 3-8** and **Table 3-9**), with all 52 (self and synthetic) agonists of the a3a TCR being retrieved by rank 586 of the reference library, compared with rank 1415 for the best model without stimulation data. Provision of co-stimulation information alone for all agonists was insufficient to achieve this level of performance, achieving a worst rank of 1162. The addition of co-culture data from both human-derived agonists and synthetic mimotopes improved retrieval of both categories, superior to either mimotopes or human agonists alone. Filtering the top peptides for likely presentation on HLA-A*01:01 using NetMHCpan4.1 (215) improved performance of the best model further (**Table 3-9**), with all agonists of the a3a TCR being retrieved by rank 444, and all but seven by rank 73: well within the range of peptides that could be tested for functional response using a single 96-well plate.

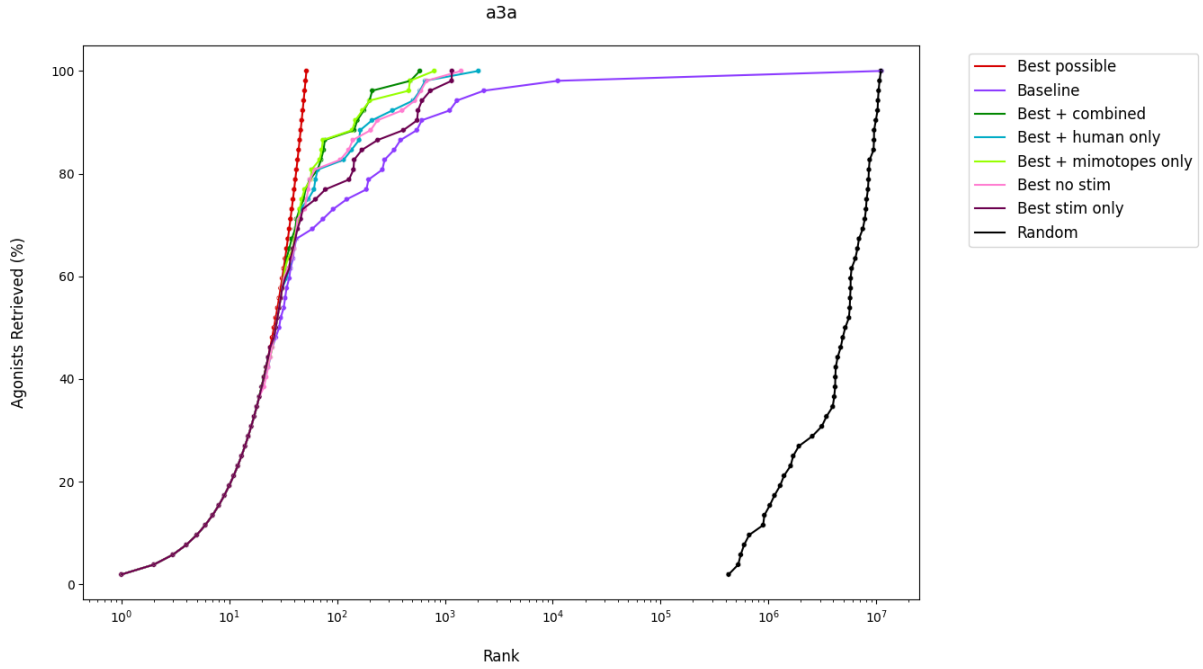


Figure 3-8: Relative contribution of binding and activation to predictive performance. *Baseline*: A baseline model trained with only yeast pMHC binding data; *Best stim only*: Optimised model trained with stimulation data but not binding data (see **Methods**); *Best no stim*: Optimised model trained with yeast pMHC binding data only; *Best + human only*: Optimised model trained with pMHC binding data and human agonist stim data; *Best + mimotopes only*: Optimised model trained with pMHC binding data and synthetic mimotope agonist stim data; *Best + combined*: Optimised model trained with pMHC binding data and all agonist stim data; Random: A random shuffle of the reference proteome.

Table 3-9: Incorporating binding, activation, and presentation data to improve model performance for the a3a TCR. Showing best rank retrieval for all known human or synthetic agonists of the a3a TCR. MAGE-A3 (EVDPIGHLY) and titin (ESDPIVAQY) are highlighted. %CD69: Mean activation values after subtraction of negative control, N=2. *Baseline*: A baseline model trained with only yeast pMHC binding data*; Best stim only*: Optimised model trained with stimulation data but not binding data (see **Methods**); *Best no stim*: Optimised model trained with yeast pMHC binding data only; *Best + human only*: Optimised model trained with pMHC binding data and human agonist stim data; *Best + mimotopes only*: Optimised model trained with pMHC binding data and synthetic mimotope agonist stim data; *Best + combined*: Optimised model trained with pMHC binding data and all agonist stim data; *Best + combined + filter*: Optimised model trained with pMHC binding data and all agonist stim data and filtered for presentation by HLA-A*02:01 using NetMHCpan4.1 (215).

Peptide	Type	%CD69	Baseline	Best no stim	Best + human	Best + mimotopes	Best + combined	Best stim only	Best + combined + netMHCpan
ETDPIGQWY	Synthetic	53.2	1	1	2	1	1	1	1
EVDPIGQYY	Synthetic	49.0	8	2	3	2	2	2	2
EIDPINHYY	Synthetic	38.9	12	4	1	5	3	15	3
EVDPMEQFY	Synthetic	30.8	2	3	5	3	4	4	4
EVDPIGHLY	Human	41.3	4	8	4	7	6	3	6
EVDPMDWFY	Synthetic	44.9	7	5	6	4	5	8	5
ESDPMDMFY	Synthetic	49.7	5	6	7	6	7	19	7
EVDPIGHVY	Human	23.1	9	13	8	15	8	5	8
ETDPMSQQY	Synthetic	25.2	13	7	9	8	9	13	9
ETDPMDQAY	Synthetic	50.5	10	10	13	9	10	7	10
ESDPIDAWY	Synthetic	51.6	3	12	10	11	11	12	11
ETDPIDQMY	Synthetic	51.9	14	14	14	12	12	6	12
ETDPIEWRY	Synthetic	44.6	11	9	12	10	13	23	13
ESDPIDFWY	Synthetic	33.4	6	11	11	13	14	16	14
ESDPMEQMY	Synthetic	44.0	17	16	17	14	15	11	15
EVDPMDNLY	Synthetic	32.5	21	19	19	17	16	14	16
ETDPIEQIY	Synthetic	36.2	15	15	18	16	17	10	17
EMDPVTQLY	Human	11.5	23	17	15	18	18	17	18
EMDPMTDFY	Synthetic	41.4	39	26	26	19	19	22	19
ESDPIDTVY	Synthetic	38.9	19	27	24	22	20	21	20
ETDPMDEQY	Synthetic	48.1	34	21	25	20	21	20	21
EVDPIRHYY	Human	8.6	25	41	16	45	22	9	22
EMDPMDRRY	Synthetic	6.6	16	18	21	21	23	46	23
EMDPLENYY	Synthetic	24.9	33	22	20	27	25	28	25
ESDPIDTTY	Synthetic	44.1	20	32	28	26	26	412	26
ESDPMDAEY	Synthetic	39.2	18	28	30	25	27	33	27
EVDPMETEY	Synthetic	28.0	29	30	31	24	28	31	28

Peptide	Type	%CD69	Baseline	Best no stim	Best + human	Best + mimotopes	Best + combined	Best stim only	Best + combined + netMHCpan
EMDPIDIMY	Synthetic	42.0	37	23	23	30	29	41	29
ETDPMDGIY	Synthetic	53.1	22	29	36	28	30	18	30
EMDPMELIY	Synthetic	47.5	36	24	29	29	31	27	31
EVDPMTPEY	Synthetic	25.7	42	35	40	31	32	29	32
ESDPIVAQY	Human	9.8	32	43	34	40	33	24	33
ETDPMEYDY	Synthetic	43.8	59	31	38	32	34	26	34
EMDPIEVVY	Synthetic	28.8	30	25	27	35	36	63	36
ELDPMDEDY	Synthetic	8.5	74	38	46	33	38	39	38
EVEWLHPLY	Synthetic	35.9	123	42	44	39	41	78	41
ETDPVENTY	Human	8.2	27	40	41	42	42	733	42
ETDPVNHMV	Human	0.0	1293	56	42	58	45	30	45
EMDPMQPDY	Synthetic	32.9	92	54	66	44	49	36	48
ETDPIMEEY	Synthetic	45.4	263	128	160	74	77	37	75
SLDPIDTYY	Synthetic	9.5	40	52	61	47	51	129	50
EVEWLTMDY	Synthetic	32.5	196	50	54	50	56	144	55
EVDPMYDDY	Synthetic	13.7	1107	140	211	57	66	43	64
ETDPLTFNF	Human	0.0	11213514	62	63	72	71	170	69
MLDPIDMYY	Synthetic	26.0	276	107	164	69	75	142	73
EGDPLILQY	Human	5.8	187	205	115	201	144	563	130
EMDPLFDAY	Synthetic	31.3	610	599	587	148	154	48	140
EQDPLMPEY	Synthetic	14.1	388	237	326	138	177	553	156
EMDPFHYDY	Human	5.4	549	526	136	798	212	613	185
EIEWLQDEY	Synthetic	5.8	2301	401	507	172	200	1155	176
ELDPILTEV	Human	0.0	11161	674	657	474	479	1162	381
MTDPIESDY	Synthetic	17.8	338	1415	2046	462	586	238	444

Discussion

We present Orpheus, an open source and fully automated pipeline for prediction of TCR cross-reactivity from binding and/or functional data screens. Exploitation of the statistical and evolutionary patterns present in deep library screens enables retrieval of the known agonists of the a3a, DMF5, nkg2c, as4.3, Ob.A12 and Ob.F3 receptors. Whilst PWMs are commonplace in the interrogation of yeast and other display screens, to our knowledge, this is the first pipeline demonstrating reproducible off-the-shelf performance across multiple datasets. There being no consensus definition of cross-reactivity (discussed in **Chapter 4**), nor of how "accurate" a predictive model might be in this context; here we cast the problem as a retrieval task, observing ranking of many known agonists within a range that could feasibly be validated in a

research setting. The model will be made freely available at https://github.com/hudsondan/Orpheus.

Modifying a PWM to account for evolutionary substitutability, subdominant peptide clusters, and to relax anchor assumptions, provides performance superior to our best machine learning model, despite evaluation of many feature and model combinations (**Fig. 3-5** and **3-6**). These findings are in line with previous comparisons of TCR epitope specificity predictors, where simple distance measures performed comparably with sophisticated neural networks (93). We demonstrate the capacity of the pipeline to identify agonists of both engineered and natural TCRs, discovering a novel agonist of the a3a TCR-T (FXYD6). This nine-residue peptide belongs to the phosphohippolin ion channel regulatory protein expressed in the human brain (219). Whilst the known cardiotoxicity of the a3a TCR halted further clinical development, the identification of multiple additional targets of a widely studied TCR serves to illustrate both the considerable promiscuity of this receptor, and the power of display screens in mining specificity.

Integrating knowledge of known stimulatory peptides improved model predictions even further and may be helping the model to discern canonical from non-canonical binders. This highlights an opportunity to harness multimodal data in iterative cycles of discovery, screening, prediction, experimental validation, and rational engineering both for TCR-T development, and for comprehensive profiling of all possible agonist peptides for a given natural TCR.

Limitations and further work

In the present study, the Orpheus pipeline has been applied to receptor-pMHC libraries spanning prevalent class I and class II HLA alleles by expression as single-chain trimers. The solution state affinities recorded for single chain trimers to cognate peptides are typically weaker than the native complex, which might result in a more permissive binding profile, and these screens are typically limited to receptor-ligand interactions under 100 μM. Nonetheless, we find that screens with these modified constructs are sufficient for accurate ranked retrieval of known TCR-pMHC pairs.

The use of consensus proteomes may also overlook post-translationally modified or other non-canonical peptides, a further source of potential toxicity. Extension of the pipeline to more comprehensive tissue-specific datasets might improve the specificity of predicted hits (see **Chapter 5** and **Discussion**).

Finally, we demonstrate the application of the Orpheus system in the limited context of TCR-T toxicity screening, however the approach is clearly applicable to inference of the antigen

specificity of other wildtype or engineered orphan immune receptors of clinical interest, and to other forms of binding and functional data, which we explore in **Chapters 4** and **5**.

Chapter acknowledgements

The development of the Orpheus pipeline was dependent on yeast pMHC selection and co-culture data generated by Dr Ricardo Fernandes, Dr Julia McCarthy and Kristen Koopmans, and test datasets provided by Professor Michael Birnbaum and Dr Xinbo Yang.

Chapter 4: Quantification of cross-reactivity for prioritisation of TCR therapeutic variants

Overview

Having previously demonstrated the capacity of the Orpheus pipeline to retrieve the cognate agonists of TCR-Ts with well-characterised binding preferences, we now set out to explore its use in the de-orphanisation of novel TCR-T candidates. First, we develop a novel software package (fastq2pep) for the preparation and quality control of multiple deep sequencing datasets. We then conceive of a suite of cross-reactivity metrics for the interrogation of candidate therapeutics, applying them to prioritise variants of the a3a TCR engineered to have reduced reactivity towards titin. Finally, we apply unsupervised machine learning methods to determine the shared and unique binding preferences of priority variants.

Introduction

Unlike most mature antibodies, TCRs bind their cognate antigen with relatively low affinities (K_D 1-100μM) in a manner thought to enable rapid scanning of pMHC, but which may favour off-target binding (105, 167, 208). There is now a growing body of evidence to support widespread promiscuity in TCR antigen recognition, reviewed in (13, 60, 167, 168, 220). Notably, affinity-driven display screens yield hundreds of enriched peptides per receptor (105, 107), and functional analyses have demonstrated that individual TCRs may respond to multiple peptides (221, 222). Such broad receptor degeneracy is thought to be driven by evolutionary pressure to recognise a broad population of foreign peptide antigens with a limited repertoire (16, 167). Structural studies of the TCR-pMHC complex have revealed diverse biophysical mechanisms of cross-reactivity, including molecular mimicry (**Fig. 4-1**), altered TCR binding modes, and conformational flexibility in both the TCR and MHC proteins, reviewed in (60, 168).

Among the most widely studied examples of the polyspecificity problem in TCR-T development, and a focus of this chapter, is the a3a TCR (68, 108, 213, 216, 223). This affinity-enhanced variant of a natural TCR (A3) targets MAGE-A3, a cancer-testis antigen constitutively upregulated in some forms of melanoma (224). Importantly, MAGE-A3 is otherwise not expressed outside of immune-privileged organs. During early clinical trials, off-target binding to heart muscle tissue by the a3a TCR induced fatal cardiac events in two patients within days of adoptive cell transfer (68). Subsequent investigations identified a nine-residue peptide derived from the abundant muscle protein titin, expressed in the smooth muscle of beating cardiomyocytes, as the most likely culprit (213). X-ray crystallographic analysis of a

soluble version of a3a (MAG-IC3) bound to EVDPIGHLY (MAGE-A3) and the titin-derived peptide ESDPIVAQY, revealed that titin acts as a molecular mimic of MAGE-A3 when presented by HLA-A*01:01 (**Fig. 4-1**) (223).

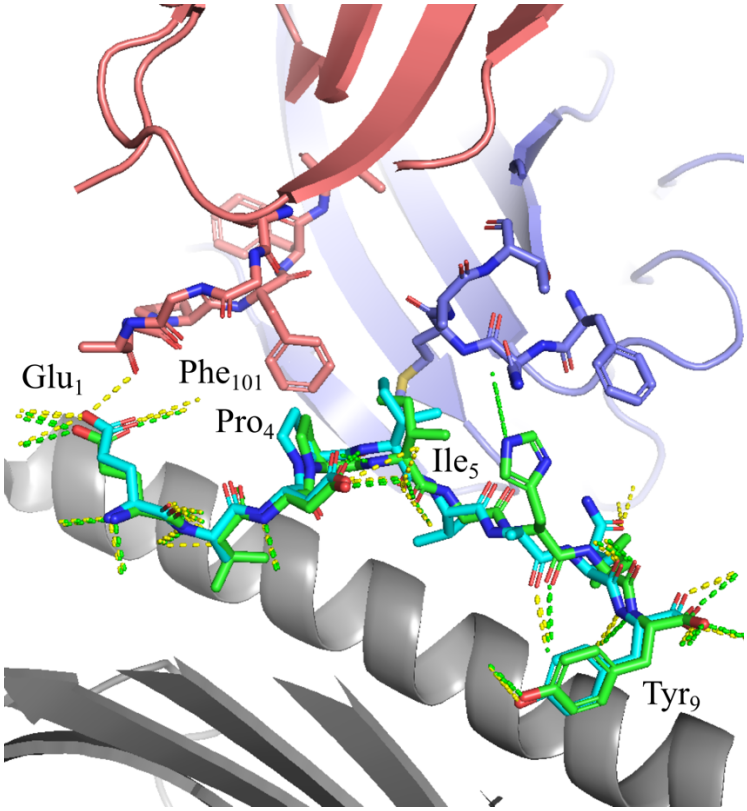


Figure 4-1: Molecular mimicry between Titin and MAGE-A3. X-ray crystal structures solved for titin (cyan sticks, PDB 5BS0), aligned with HLA-A*01:01 from MAGE-A3 (green sticks, PDB 5BRZ). Polar contacts between MAGE-A3 atoms within 4Å of peptides and CDR loops of the 5BRZ a3a TCR alpha (pink sticks) and beta (blue sticks) chains or HLA-A*01 helices (grey cartoon) are shown as green or yellow dashed lines, respectively (150, 223). Figure produced with PyMOL (18).

A fundamental question raised by the a3a TCR and other cross-reactive T cell candidates remains unanswered in the field, namely whether it is possible to define TCR promiscuity, and if so, whether promiscuity can be quantified as a means of comparing receptors. In a 2012 study, Linda Wooldridge and colleagues probed the cross-reactivity of the 1E6 TCR towards 9 x 10^{12} variants of a preproinsulin-derived decamer peptide through *in vitro* functional assays (221). Reactivity was probed through analysis of T-cell responses to pooled combinatorial peptide libraries, each representing a mathematically derived subsample of all possible 10-mers. The authors concluded that 1E6 can respond to over a million distinct peptides. However, such bulk pooling approaches overlook TCR sensitivity to single amino acid changes, and also fail to account for the likelihood of peptide presentation by MHC - resulting in a likely overestimation of true cross-reactivity (105). XSM studies take a narrower view, studying the reactivity of a given receptor to a set of peptides generated through sequential mutation of

amino acids along the length of a known cognate agonist (69, 107, 211). However, such analyses may overly constrain the search space of TCR-pMHC complexes to local minima around the starting peptide and so are likely to underestimate receptor promiscuity.

A 2014 study from the Davis group at Stanford was among the first to probe TCR cross-reactivity using the results of selection-driven analysis from libraries of thousands of pMHC complexes (105). Exploiting diverse yeast pMHC libraries to define the binding of three murine and two human $CD4^+$ TCRs to class II MHC, Michael Birnbaum and colleagues reported hundreds of unique hits per receptor, and tens of unique functional agonists. This and later affinity-driven selection analyses with yeast or bacteriophage display libraries extended description of cross-reactivity to include the presence of clusters of peptides distinct in sequence or in biophysical descriptors (107, 112). Leveraging an increasing number of TCR-pMHC crystal structures (169), recent studies have attempted to extend measures of cross-reactivity to include biophysical or structural correlates, reviewed in (170). However, despite the plethora of experimental tools, there exists no consensus measure of promiscuity for candidate therapeutic TCRs.

Several groups have attempted to engineer novel variants of problematic receptors including a3a (108, 216), whilst others have proposed peptide display screening as a means of testing TCR the safety of new candidates (69). However, none have attempted to balance the need for designing receptors with an increased affinity towards a given target, with the risk of introducing new unintended off-target specificity when departing from a natural TCR repertoire screened by the thymus.

Here, we explore parallel yeast peptide MHC screens of eight novel variants of the a3a TCR, designed to preserve reactivity towards MAGE-A3 whilst reducing off-target toxicity towards titin. Leveraging these rich datasets, we propose novel metrics of TCR cross-reactivity that allow us to prioritise a subset of these variants for further development. Finally, we apply graph network clustering methods to inspect the unique and shared actual and predicted peptide binding preferences of priority variants, which supports progression of a single functionally validated TCR-T candidate.

Methods

Datasets

A panel of eight TCR candidates was produced by the Fernandes group, with CDR2α sequences spanning the sequence space between a3a (hereafter, v1) and wildtype A3 TCRs (v2) (**Fig. 4-2**), to understand whether it is possible to decouple recognition of MAGE-A3 and titin peptides

(214). The binding preferences of each of the ten constructs were then characterised using yeast pMHC display on an HLA-A*01:01 9-mer library, as described previously in **Chapter 3**.

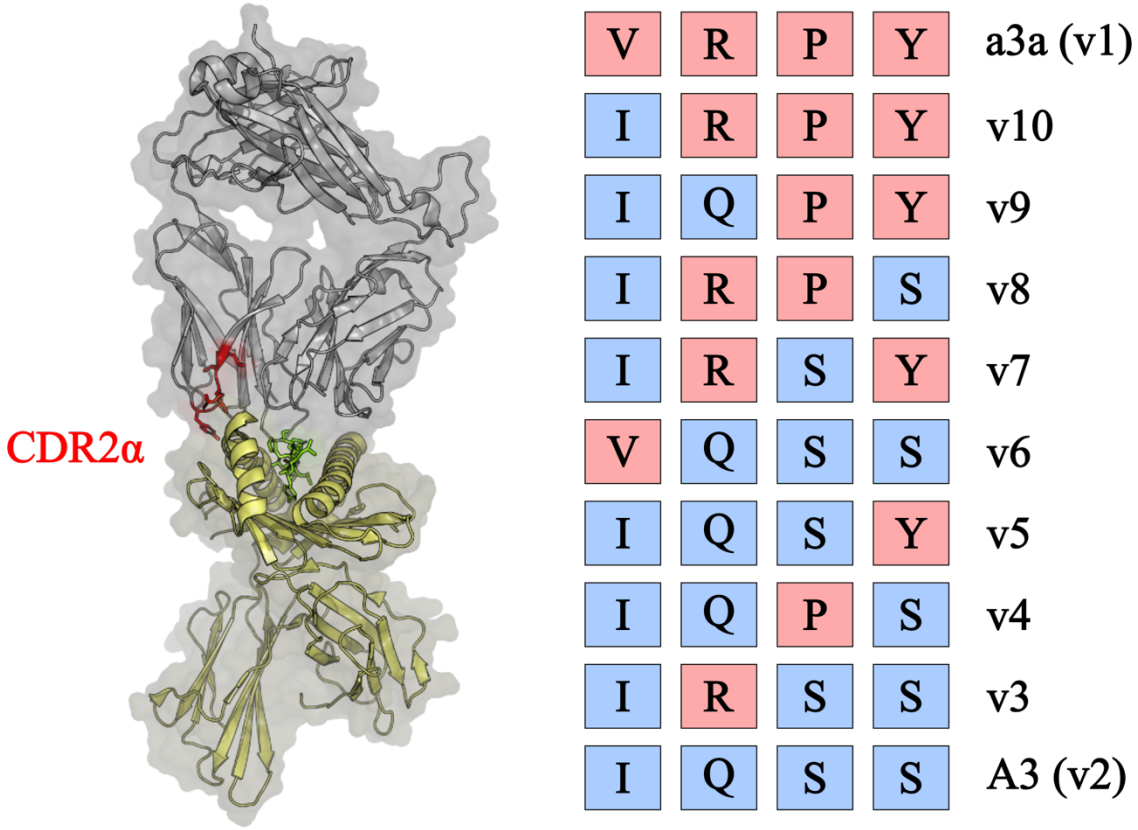


Figure 4-2: Exploring cross-reactivity of the a3a TCR. **Left** | Cartoon of the crystal structure of the a3a TCR bound to MAGE-A3 in the context of HLA-A*01:01. **Right** | Designing a panel of mutants to reduce a3a reactivity towards titin, showing CDR2α sequences of the variants analysed; work led by Julia McCarthy (214).

Fastq2pep

Analysis of yeast pMHC data for the ten constructs of interest first required conversion of large files containing millions of reads into summary peptide count tables as an input to the analytical component of the Orpheus pipeline developed in **Chapter 3**. The fastq2pep pipeline (see **Results**) was written in Python, incorporating PEAR v0.9.2 released on March 26 2014 (225) and the BioPython six_frame_translations module (226). The pipeline was applied to fastq files produced for the A3 wildtype and a3a variant TCRs using a single Intel(R) Xeon(R) Gold 6126 CPU @ 2.60GHz. Contaminant screening was applied to remove the control peptides encoded by the sequences listed in **Table 4-1**.

Table 4-1: Contaminants. Nucleotide sequences of known control peptides used in the Fernandes laboratory and removed from the fastq input files.

Construct	Sequence
MAGE-A3	GAAGTTGACCCAATCGGACATTTGTAC
NY-ESO C9V	TCCCTGTTGATGTGGATAACCCAATGC
NY-ESO	TCCCTGTTGATGTGGATAACCCAATGCGTA
p53	CATATGACGGAGGTGGTCCGCCATTGT
MART-1	GAACTGGCAGGTATTGGTATTTTAACTGTA
gp100	TACCTGGAACCAGGTCCAGTAACCGCT

Visualisation of pMHC screen data

Python code for the cross-reactivity analyses detailed in this study will be made publicly available at http://www.github.com/hudsondan/Orpheus. All plots were generated using the Matplotlib and Seaborn visualisation packages (201, 202). Heat maps were produced using the baseline PWM described in **Eq. 3-1**. Peptide logos were produced using a script adapted from (112). HLA-specific reference peptide lists were generated by drawing elution mass spectrometry data from the NetMHCpan4.1 training dataset (215) and from the MHC motif atlas (20), generating baseline PWMs from the combined datasets, and isolating the top 1000 predicted hits from the human consensus proteome (114). 1000 negative reference peptides were drawn randomly from the UniProt consensus human proteome (114).

Cross-reactivity measures

Comparative analysis of cross-reactivity in the ten CDR2α variants of interest, and in two controls (HLA-A*01:01 specific peptides, and random negatives) was performed using newly conceived cross-reactivity measures: Cumulative Rank at 50% saturation (CR50), Local Positional Entropy (LPE) and Global Positional Entropy (GPE) (**Eq. 4-1**). Each measure was applied to the set of enriched peptides in a yeast pMHC screen, as defined in **Chapter 3**.

Equation 4-1: Local and global measures of receptor cross-reactivity.

a) $\mathbf{CR_{50}} = \boldsymbol{max}\left\{k: \frac{\sum_{i=1}^{k} \boldsymbol{r}(p_i)}{T} < 0.5\right\}$

$\mathbf{T} = \sum_{p \in P} \boldsymbol{r}(p)$

Where k is the rank of peptide *p* belonging to the set of P enriched peptides meeting the criterion of 50% of the cumulative reads of P at position k, and $\mathbf{r}(p_i)$ is the read count associated with p at rank *i.*

b) $\mathbf{LPE}(\mathrm{l}) = -(\mathbf{H}(\mathrm{l}) + \mathbf{H}(\mathrm{random}))$

$\mathbf{H}(\mathrm{l}) = \sum_{\mathrm{aa} \in \mathrm{AA}} \mathbf{p}(\mathrm{a}, \mathrm{l}) \log_2 \mathrm{p}(\mathrm{a}, \mathrm{l})$

$\mathbf{H}(\mathrm{random}) = -\sum_{\mathrm{a} \in \mathrm{A}} \frac{1}{20} \log_2 \frac{1}{20}$

Where $\mathbf{H}(\mathrm{l})$ is the Shannon Entropy of amino acids observed at position *l* in a peptide of *L* residues in length, p(a, l) is the probability of observing an amino acid at position l, give in **Eq. 3-1a**, and *a* is one of the 20 amino acids in the set A.

c) $\mathbf{GPE} = \frac{\sum_{\mathrm{l}=1}^{\mathrm{L}} \mathbf{LPE}(\mathrm{l})}{\mathrm{L}}$

Graph network clustering

To inspect the distribution of enriched peptides across variant screen data, peptide amino acids were converted to biophysical embeddings by concatenating BLOSUM62 encodings, hydrophobicity scores, and isoelectric points generated using the Peptides library v0.2.1. Embeddings were clustered with hdbscan (227), and compressed using the scikit learn implementations of principal component analysis (PCA) - retaining the first five dimensions after elbow plot analysis (**Fig. 4-3**) - and tSNE (197). Graph networks were produced using scikit network (228), each peptide being represented by a node of size proportional to its frequency. Edges were added for peptides whose biophysical embeddings had a Euclidean distance ≤ 0.1 and Hamming distance ≤ 4, determined using a script adapted from (143). Node clusters were identified using the Louvain clustering module within Scikit-network (228). Chord diagrams were produced using PyCirclize, available at https://moshi4.github.io/pyCirclize/. Analysis of predicted hits was performed using the same parameters.

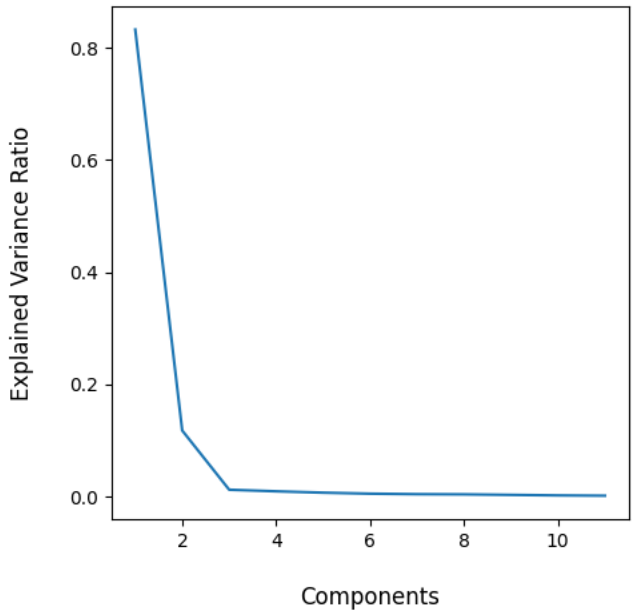


Figure 4-3: Principal component analysis of input yeast pMHC data. Showing explained variance ratio as a function of the components retained for biophysical embeddings for all datasets combined.

Results

The first challenge encountered in the analysis of ten parallel deep yeast pMHC library screens was the need to process millions of yeast DNA sequences, and so generate a count table appropriate for cross-reactivity analysis. Previously, this had been approached by the Fernandes group by translating the DNA sequences identified in each sequencing run into their corresponding amino acids using Geneious (229). This was followed by a manual search for the SVLA/GGGGS linker sequences flanking the peptide component of the single chain trimer, the resulting peptide sequences then being exported to Microsoft Excel to produce count tables. However, this process was laborious even for shallow libraries of thousands of reads, and impractical for the present datasets; no freely available software was identified for automated conversion.

We therefore developed an open-source package (fastq2pep) for the translation of raw sequence reads to peptide count tables (**Fig. 4-4**). The pipeline includes an initial quality control filter with PEAR, an algorithm that matches paired end sequence reads generated using forward or reverse primers, in this case directed to the linker regions flanking the randomised peptide sequence within the pMHC trimer (225). This process produces a fastq file containing a cleaned subset of sequences appearing in both runs (**Table 4-2** and **4-3**). At the core of the fastq2pep package is a simple BioPython script (226) that searches for the linker sequences, translates them into the corresponding amino acid sequence, and then trims out the linkers. The resulting peptides are counted for each round of selection, and read counts are then aggregated to generate a table of counts per round of enrichment. We provide a library-splitting bash script

for processing of large screens that may otherwise cause memory time-outs even on high performance clusters.

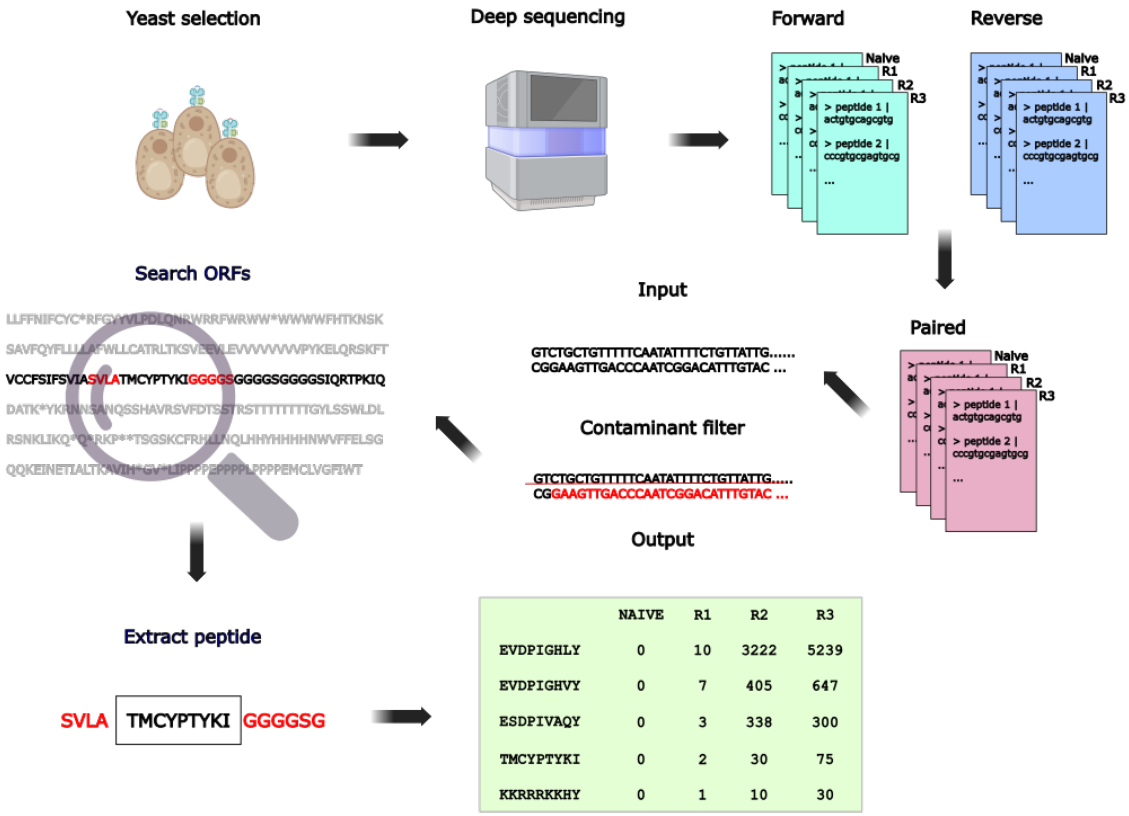


Figure 4-4: The fastq2pep QC and count table production pipeline. Enriched yeast expressing pMHC constructs are deep sequenced with forward and reverse primers to generate fastq files per round of enrichment. PEAR is then applied to produce a consolidated sequence file per round. Contaminants are filtered out by comparison with a reference file, before conversion of nucleotide to amino acid sequences. ORFs containing known flanking sequences are extracted and clipped before consolidation of counts across rounds.

The ability of *S. cerevisiae* to survive and reproduce under diverse conditions renders them a convenient tool for molecular biology, but poses an additional technical challenge of cross-contamination between experiments. This typically manifests as an abundance of copies of a specific yeast pMHC clone used to test TCR multimer binding, such as MAGE-A3:HLA-A*01:01, including in the naïve (preselection) round of a given experiment. However, simple removal of contaminant peptides on amino acid sequence alone might overlook convergent selection on a true signal in later rounds, such that a TCR may enrich non-contaminant peptide MHC constructs having the same amino acid sequence as, but different nucleotide sequences to, a contaminant. To avoid the removal of these true positives from the screen, a module was

added to fastq2pep that permits extraction of all pMHC constructs having an exact nucleotide sequence match to a known contaminant.

Combining PEAR and contaminant filtering with fastq2pep reduced the total unique peptides by 26% for yeast screens of the a3a TCR (**Table 4-2**). However, the impact on total read counts was minimal (**Table 4-3**). Applying the Orpheus predictive pipeline to unmodified, PEAR and/or contaminant-filtered datasets, we observed that PEAR treatment and contaminant filtering had a negligible impact on retrieval of the suite of known a3a agonists (**Fig. 4-5**).

Table 4-2: Unique peptides per selection round, a3a TCR.

Dataset	NAIVE	R1	R2	R3	R4
Untreated	1922942	633461	108552	44386	34780
Untreated + filter	1922936	633458	108549	44383	34777
% untreated	100%	100%	100%	100%	100%
PEAR	1888726	605886	89274	32813	21481
% untreated	98%	96%	82%	74%	62%
PEAR + filter	1888721	605883	89271	32809	21478
% untreated	98%	96%	82%	74%	62%

Table 4-3: Total read counts per selection round, a3a TCR.

Dataset	NAIVE	R1	R2	R3	R4
Untreated	3040036	1795631	1420848	1957216	1862338
Untreated + filter	3031713	1794872	1420061	1956040	1861911
% untreated	100%	100%	100%	100%	100%
PEAR	3029356	1787241	1416324	1943808	1858916
% untreated	100%	100%	100%	99%	100%
PEAR + filter	3020901	1786480	1415534	1942635	1858482
% untreated	99%	99%	100%	99%	100%

Applying PEAR and contaminant filtering to yeast screens for each of the ten variants yielded count data summarised in **Table 4-4**. We next developed a suite of metrics to probe the relative promiscuity of each variant. The first lens applied was the proportion of peptides contributing the most to the overall read counts. Cumulative frequency plots (**Fig. 4-6**) suggested faster saturation by v6, v8 and v9 than other constructs, such that fewer peptides contributed to a greater proportion of the total enriched population at the third round of selection. To quantify this, we first computed the *cumulative rank* of enriched peptides contributing to 50% of the total reads (CR50, **Eq. 4-1a**) for a given dataset, defining enriched peptides as those having 5 or more reads at round 3 and taking 50% as an arbitrary threshold.

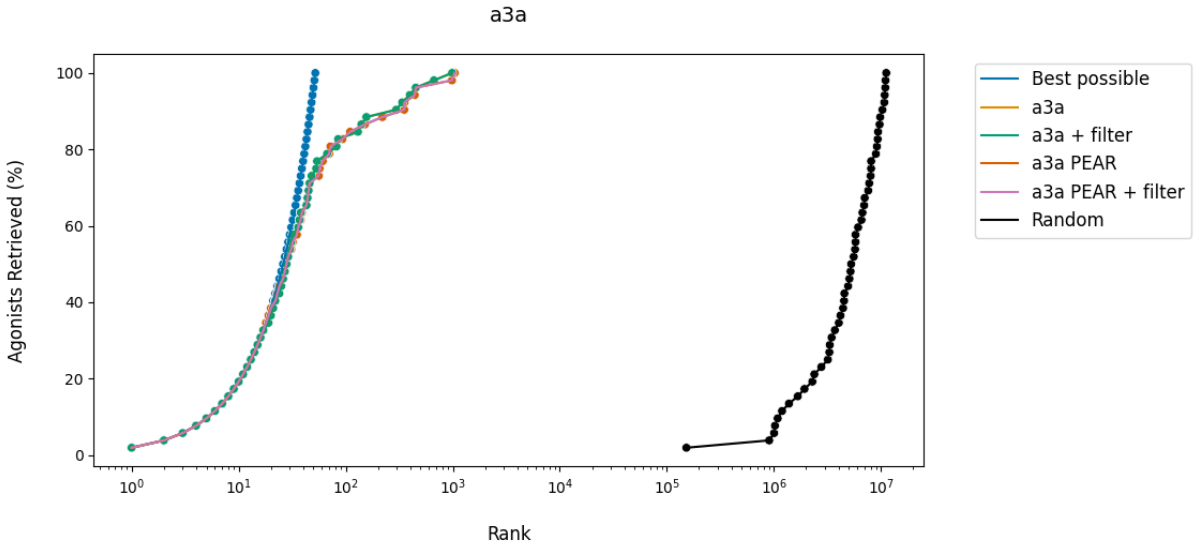


Figure 4-5: Impact of paired end read and contaminant filtering on agonist recall.

CR50 can be otherwise interpreted as the number of enriched peptides contributing half of the total reads, more selective TCRs producing screens dominated by fewer peptides. The a3a TCR achieved lower CR50 than the wildtype A3 TCR - as might be expected from the previously reported relative affinities towards MAGE-A3 of 2.3μM and 500μM, respectively (213). Variants v3 and v5-9 achieved 50% saturation within 38 peptides or less, compared with 97 for the a3a TCR: these constructs and were therefore prioritised for further analysis.

Table 4-4: Yeast pMHC datasets for variants of the a3a TCR.

Dataset	Uniques	Reads (m)	CR50 (rank)
a3a	5743	1.90	97
a3	14528	1.75	243
v3	6846	1.83	20
v4	6769	0.83	200
v5	4953	0.88	24
v6	4183	0.98	16
v7	2239	1.17	30
v8	4315	2.00	22
v9	3354	1.82	38
v10	8547	2.02	689

Whilst they provide some insight into relative receptor promiscuity, saturation curves do not provide information on the diversity of different amino acids tolerated by a given candidate. This is particularly relevant in the context of molecular mimicry, the a3a TCR recognising MAGE-A3 and titin peptides that share just 56% sequence identity (213). An additional metric was therefore devised, *positional entropy*, which gives quantitative insight into promiscuity at

both local (residue) and global (full peptide) scales (**Eq. 4-1b** and **c**). Given some screening dataset for peptides of fixed length, *local positional entropy* (LPE) computes the probability of observing an amino acid at a given position in the set of enriched peptides, relative to a random draw. Peaks in LPE bely a reduced diversity of amino acids at a certain position, which can be interpreted as "hot spots" or contact points in the TCR-pMHC interface that strongly influence specificity (105). Crucially, unlike TCR heat maps or peptide motifs - common representations in the analysis of such datasets (105, 107, 112) - LPE plots allow side-by-side comparison of hotspots between TCR variants recognising peptides of the same length.

Plotting LPE for each of the most promising variants identified from saturation curves revealed that all showed similar positional preferences to the a3a TCR, having hot spots at positions 1, 3-4, 6 and 9 (**Fig. 4-7**). These preferences were most pronounced for v8. Overlaying the LPE plot for peptides preferentially recognised by HLA-A*01:01 (see **Methods**) highlighted anchor positions 2, 3 and 9 as determinants of pMHC presentation, consistent with (20). As for a3a, the lowest LPE, or greatest peptide diversity, was observed at positions 2, 7, and 8 across the prioritised constructs. **Fig. 4-8** compares GPE values across constructs: v3, v7, v8 and v9 emerging as the least cross-reactive when averaging over all positions in the peptide. Retaining only those peptides contributing to 50% of the total reads improved resolution of relative cross-reactivity, suggesting greater selectivity for v8 than other variants at both local and global scales.

Comparison of LPE and GPE plots with TCR footprints and the corresponding logos for priority variants (**Fig. 4-9**) revealed the contribution of individual amino acids and biochemically similar groups of residues, to variant specificity. The increase in global selectivity for v7, v8 and v9 was reflected in less noisy heatmaps for these three variants than v3, a3a, and a3 in turn. All priority variants shared a strong preference for glutamic acid (E) at position 1, hydrophobic residues at p2, aspartic acid (D) at p3, proline (P) or tryptophan (W) at p4, hydrophobic residues at p5, histidine (H) and/or glycine (G) at p8, and the anchor tyrosine (Y) at p9. As for the a3a TCR, a greater diversity of amino acids was observed at C-terminal positions 6 and 7, with a preference for methionine (M) and phenylalanine (F), respectively. The lower affinity a3 TCR showed few distinct preferences beyond p3 and p9 anchor residues.

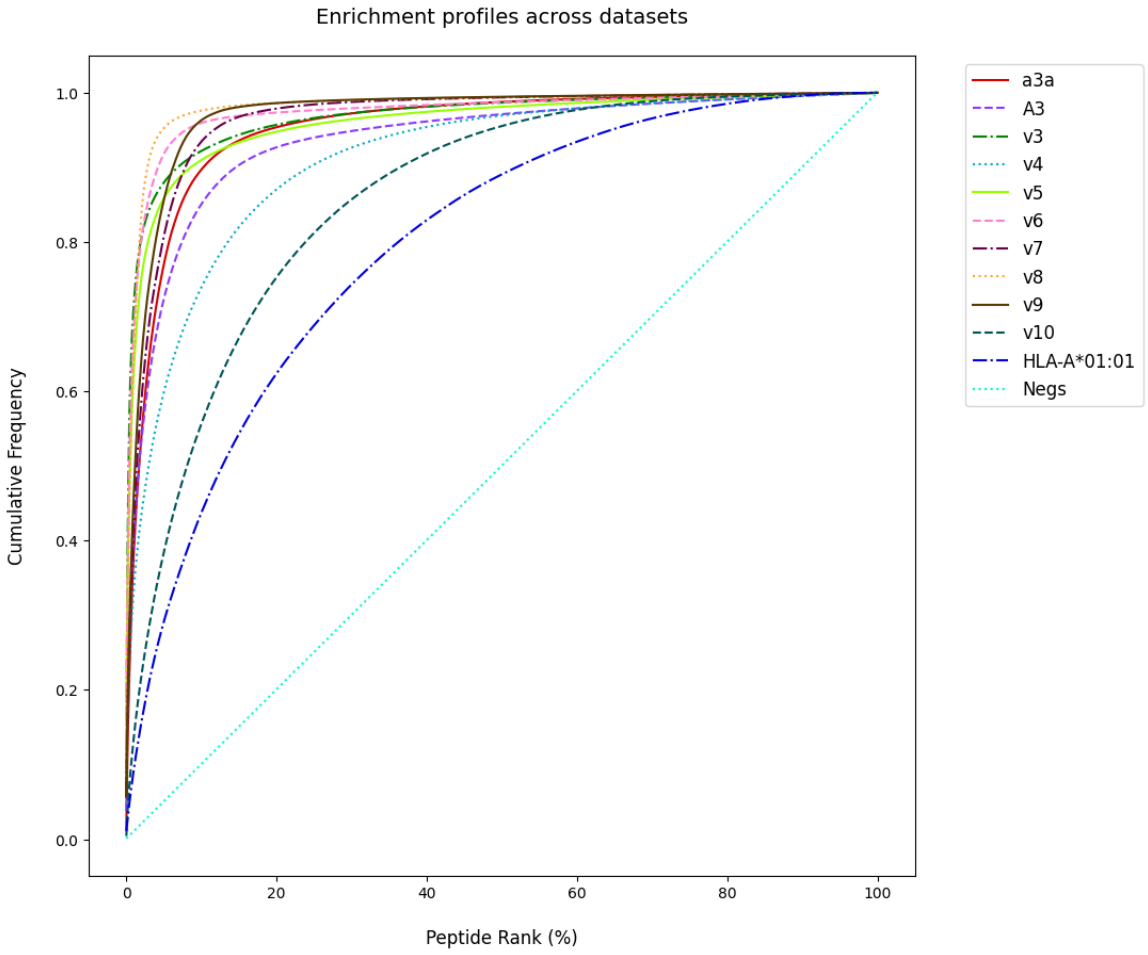


Figure 4-6: Saturation curves for a3a variants, HLA-A*01, and negative reference peptides. Shown for unique hits at the third round of selection.

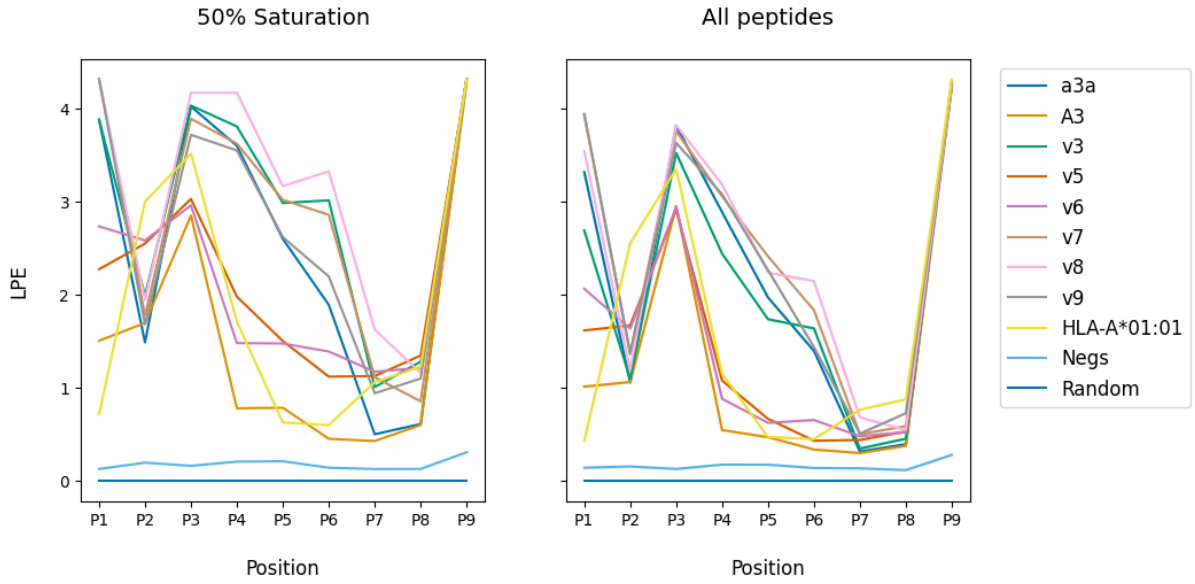


Figure 4-7: Local positional entropy plots for variants of the a3a TCR.

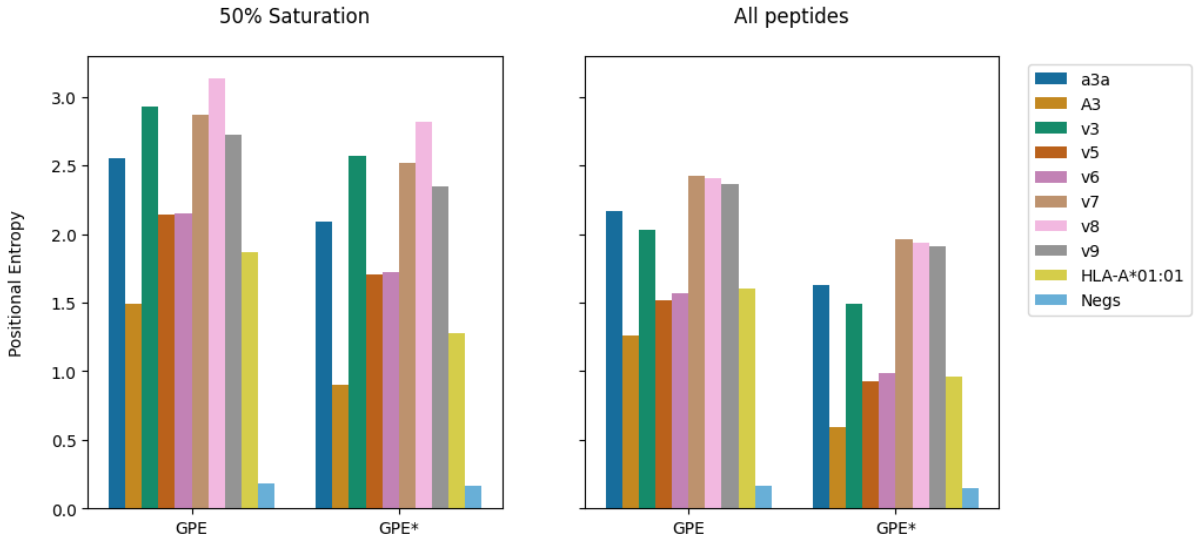
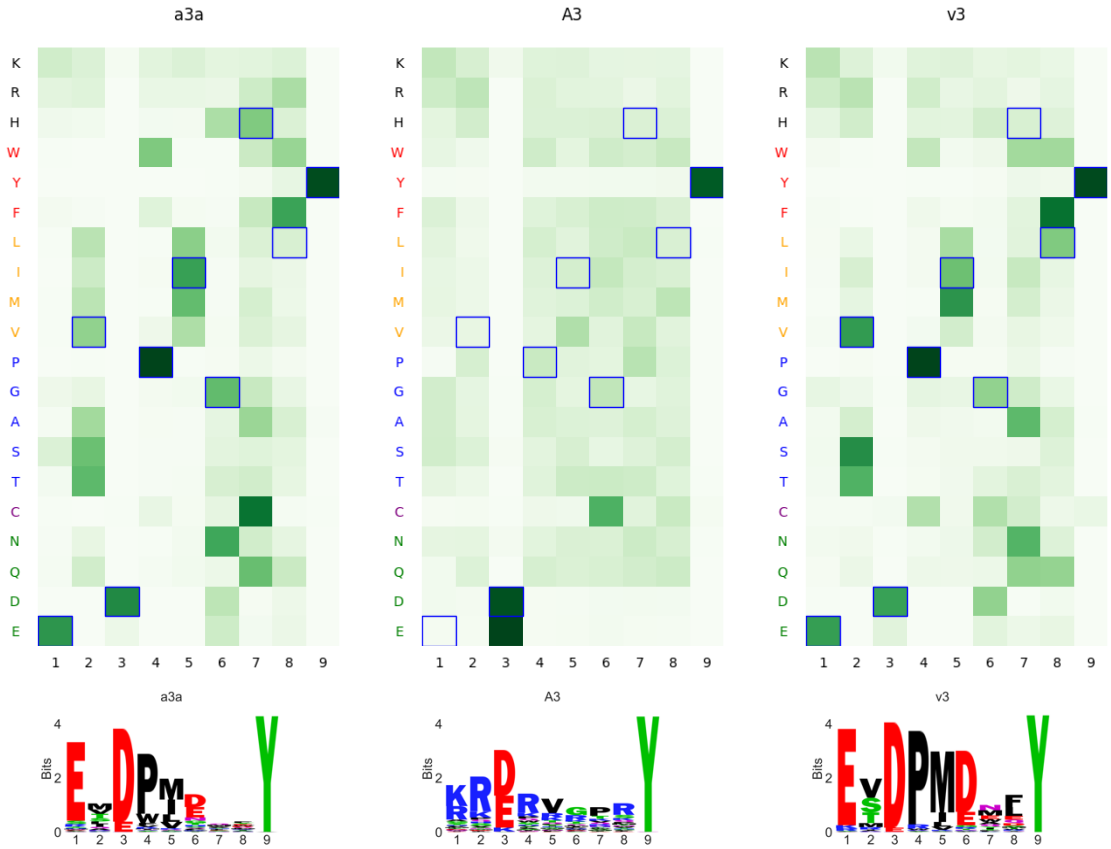


Figure 4-8: Global positional entropy plots for variants of the a3a TCR. GPE scores computed either including (GPE) or excluding (GPE*) LPE scores at anchor positions 3 and 9, computed as described in **Eq. 4-1**.



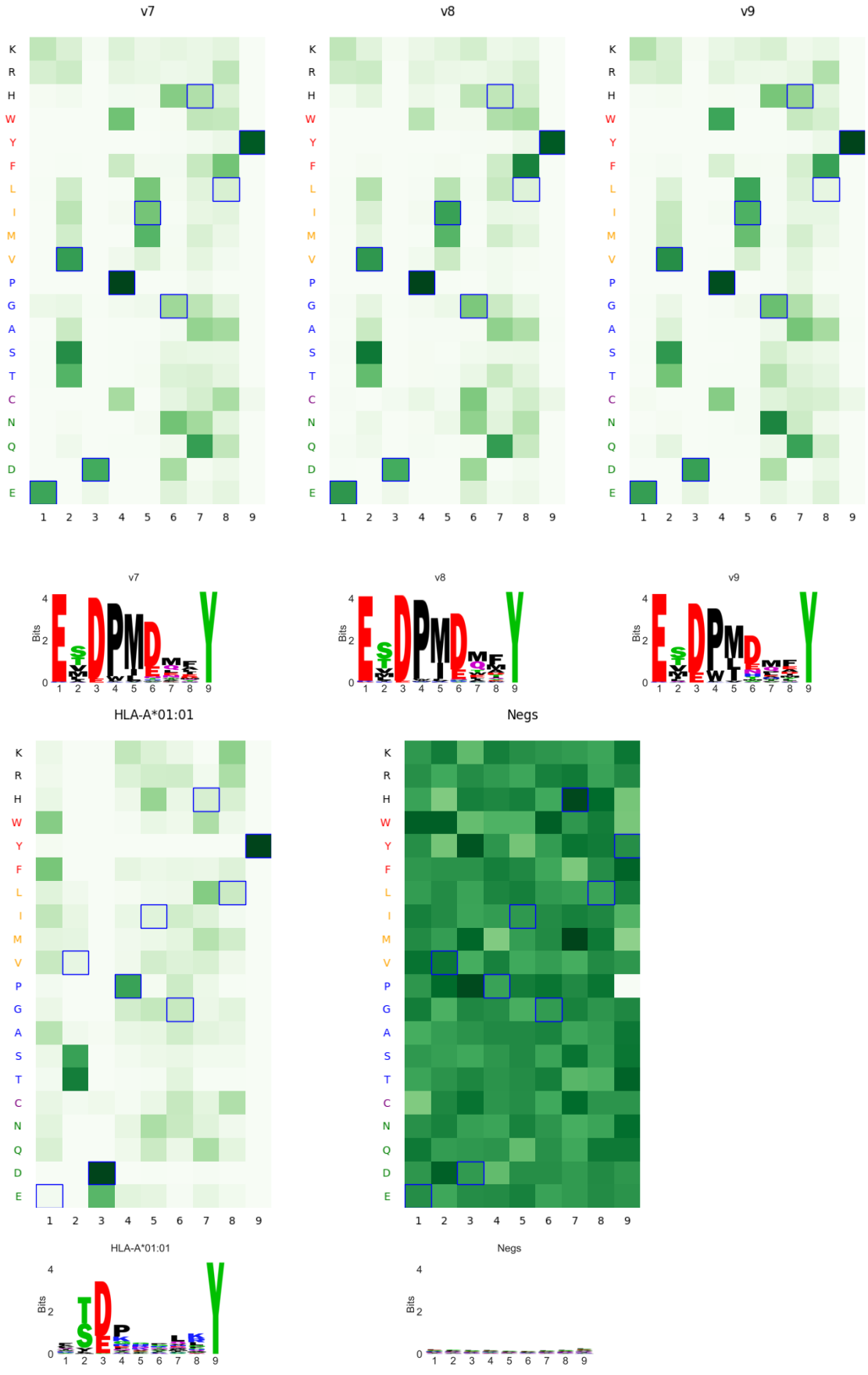

v7
v8
v9
HLA-A*01:01
Negs
Bits

Figure 4-9: TCR footprints and logos for priority variants. Heat maps show the frequency of amino acids at each position in the input library peptides, wherein darker colours described higher frequencies.

Having thus quantified the cross-reactivity and positional preferences of the priority variants, we next set out to characterise whether the observed subtle changes in amino acid preferences would manifest in the enrichment of distinct families of peptides in the yeast display data. **Fig. 4-10** depicts the distribution of enriched peptides for each variant after encoding with a combination of sequence, hydrophobicity and isoelectric point ("*biophysical embeddings*"), compressed into two dimensions using tSNE (230). Enriched peptides for prioritised variants fell predominantly into a single MAGE A3-like cluster (Cluster 0, **Fig. 4-10a**) that contained both MAGE-A3 and titin peptides. Enriched peptides for the a3 TCR, negatives and HLA-A*01:01 were more widely distributed across the plot, including in clusters 1, 2 and 3 that were rich in positive residues and likely representative of non-specific interactions.

To investigate whether the global MAGE-like motifs for priority variants contained sub-dominant groups of distinct binders, enriched peptides were converted to biophysical embeddings (as above for tSNE analysis) and then represented as graph networks in which enriched peptides (nodes) contributing to the top 75% of reads, having a Euclidean distance ≤ 0.1, and a Hamming distance of ≤ 4 amino acids, were connected by edges (**Fig. 4-11**). Each of the variants v7, v8 and v9 returned three Louvain clusters (marked red, blue and green), of which the two largest were joined into a single network. However for v7 and v9 but not v8, the third (green) cluster was distinct from this primary network, suggesting subtle differences in enriched peptide profiles between these variants. Sharing of peptide groups between TCRs is depicted in a chord diagram (**Fig. 4-12)**, indicating that the largest peptide clusters (including the MAGE-A3 containing cluster 1) are recognised by both a3a and each of the priority variants, albeit to differing extents. Motifs produced for these clusters, generated using Louvain network clustering with Scikit-network (228), were similar to MAGE-A3 for the two largest peptide groups but distinct for clusters 3-6.

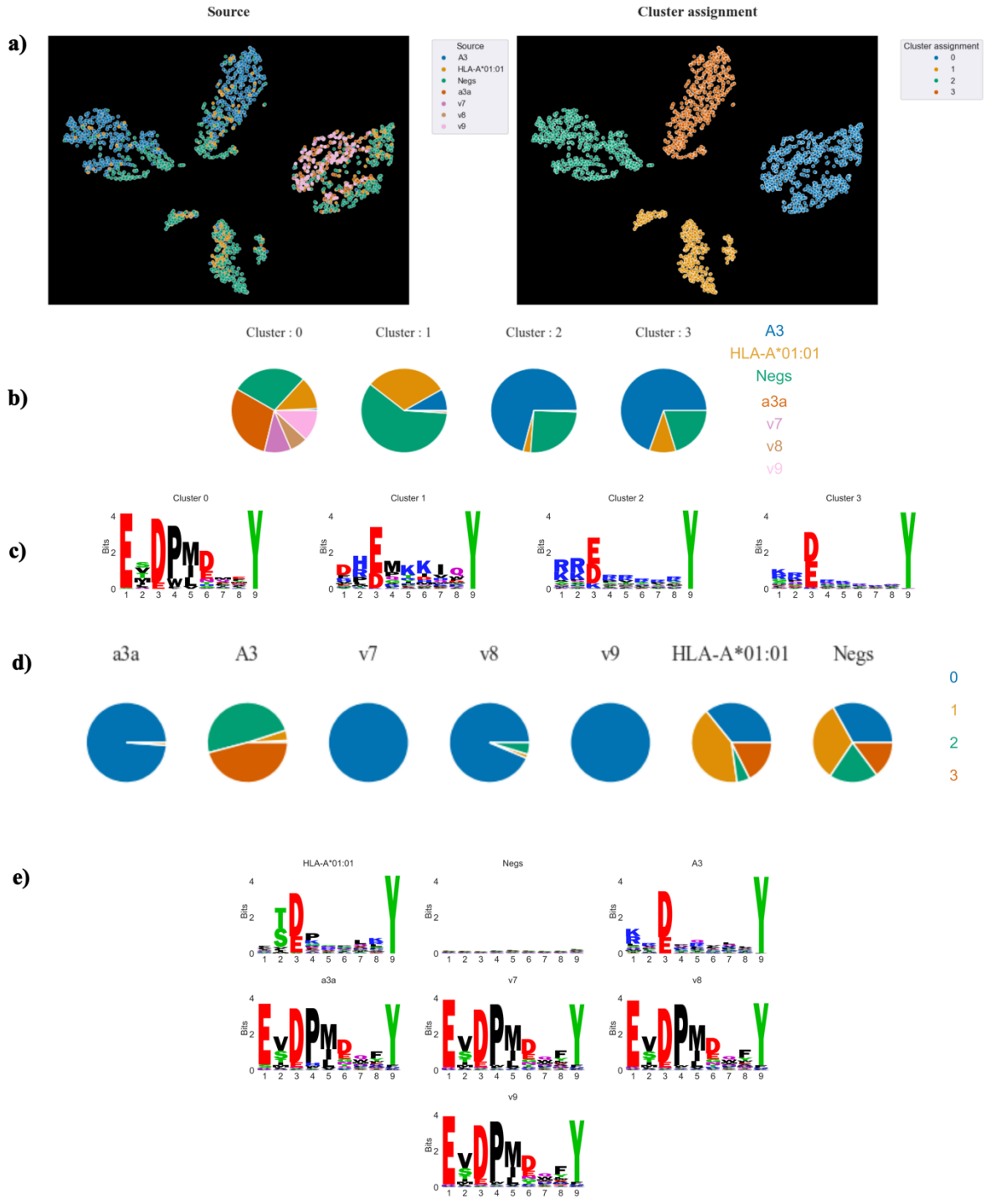


Figure 4-10: tSNE analysis of enriched peptide distribution between variants and controls. a | tSNE projection of biophysical embeddings of enriched peptides found in each dataset. **b |** The proportion of unique peptides contributed to each cluster by dataset. **c |** Logos for each cluster across all datasets. **d |** The proportion of unique peptides contributed to each dataset by cluster. **e |** Logos for each dataset across all clusters.

To confirm whether the patterns observed in the selection data would translate to differences in predicted binders, the optimised Orpheus model was then applied to each set of a3a variant screen data, as described in **Chapter 3**. The relative rankings of known self-agonists of the a3a TCR are shown for the a3 and a3a TCRs, and for priority variants in **Table 4-5**. Despite the increases in specificity suggested by comparisons of positional entropy, the priority variants all returned MAGE-A3 at rank 1, titin by rank 9, and all other off-target self-peptides within the top 1200 results. A chord diagram of the top predicted hits for each variant (**Fig. 4-13**) showed comparable cluster motifs and sharing patterns between constructs to those observed in the input datasets (**Fig. 4-12**), particularly for the largest MAGE-A3 like cluster 1.

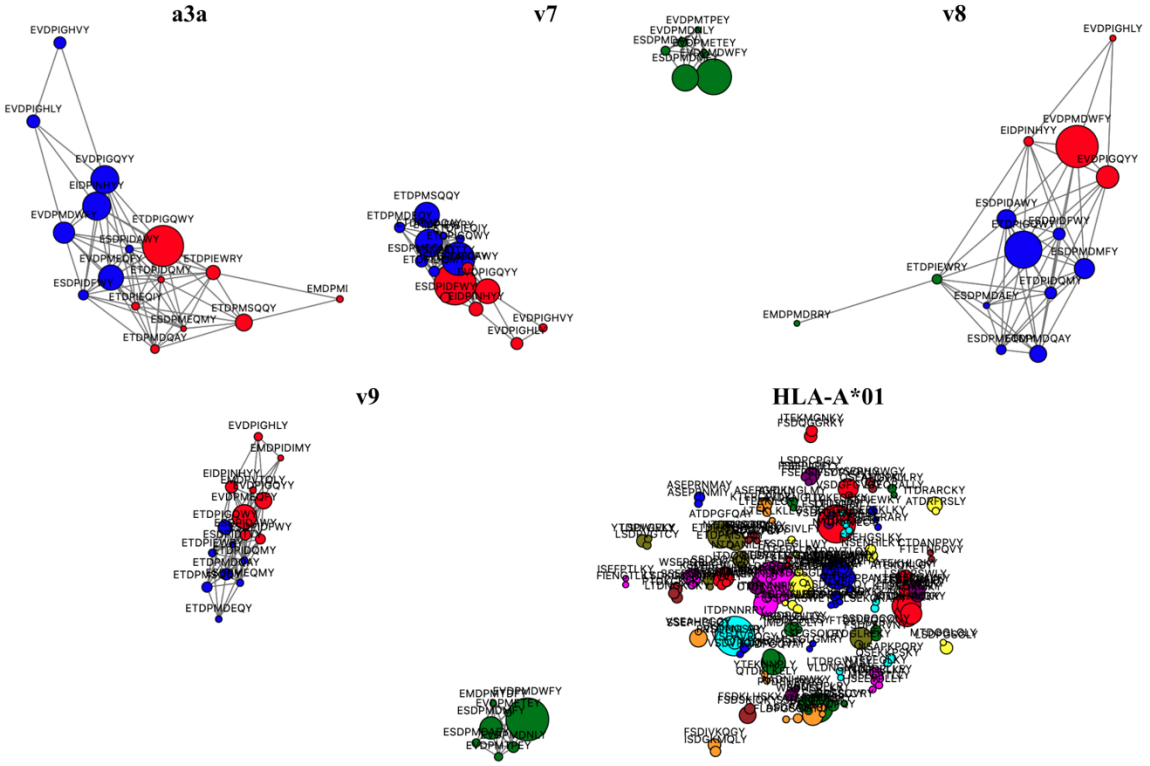


Figure 4-11: Graph network analysis of enriched peptide clusters for priority variants. Plots produced with PyCirclize (see **Methods**), showing peptides connected by Euclidian distance in the biophysical embedding space, and differing by no more than 4 amino acids. Circle sizes are proportional to the number of reads associated with that peptide in the input dataset, and colours represent Louvain cluster membership.

Table 4-5: Predicted ranking of self-agonists of the a3a TCR for priority variants.

Peptide	Protein	A3	a3a	v7	v8	v9
EVDPIGHLY	MAGE-A3	4466	1	1	1	1
EVDPIGHVY	MAGE-A6	3361	2	2	3	6
EVDPIRHYY	MAGE-B18	6505	9	5	2	7
ESDPIVAQY	Titin	3385	11	7	4	9
EMDPVTQLY	CD166	1408	3	4	8	2
ETDPVENTY	MRCKA	13777	15	18	20	12
EGDPLILQY	ANR16	2418	75	34	29	15
ETDPVNHMV	FAT2	102677	25	63	39	133
EMDPFHYDY	FXYD6	17612	402	83	116	40
ELDPILTEV	COG4	475081	1001	624	1156	830
ETDPLTFNF	PLD5	46818	26	31	26	19

To determine whether these predictions, generated using affinity-driven selection data, would translate into functional agonism, the reactivity of TCR-transfected Jurkat T cells towards each agonist was tested in the Fernandes group through co-culture as described in **Chapter 3** and in (214). Binding and co-culture assays confirmed the reactivity of the a3a TCR towards previously reported agonists (108, 213, 216), and weaker but more selective responses for a3 towards the MAGE group (**Fig. 4-14**). Consistent with their relative GPE scores, but not with the predictions of the Orpheus pipeline, constructs v8 and v9 showed improved selectivity towards the MAGE family of peptides and reduced agonism by titin and other self antigens (214).

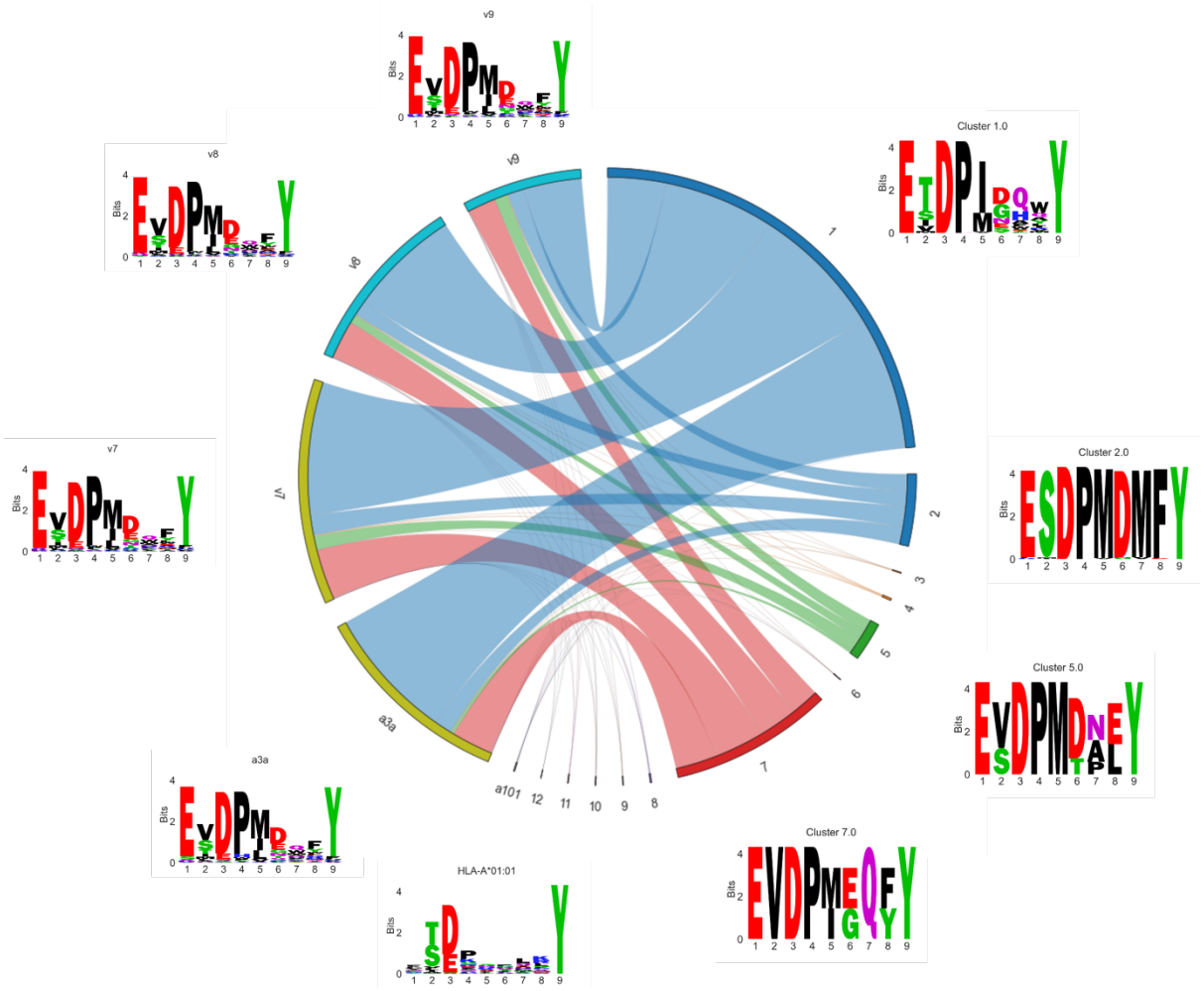


Figure 4-12: Peptide specificity sharing between constructs in selection data. Chord diagram illustrating the extent to which (top) peptide clusters identified through network analysis are shared between variants and (bottom) the corresponding logos for those variants.

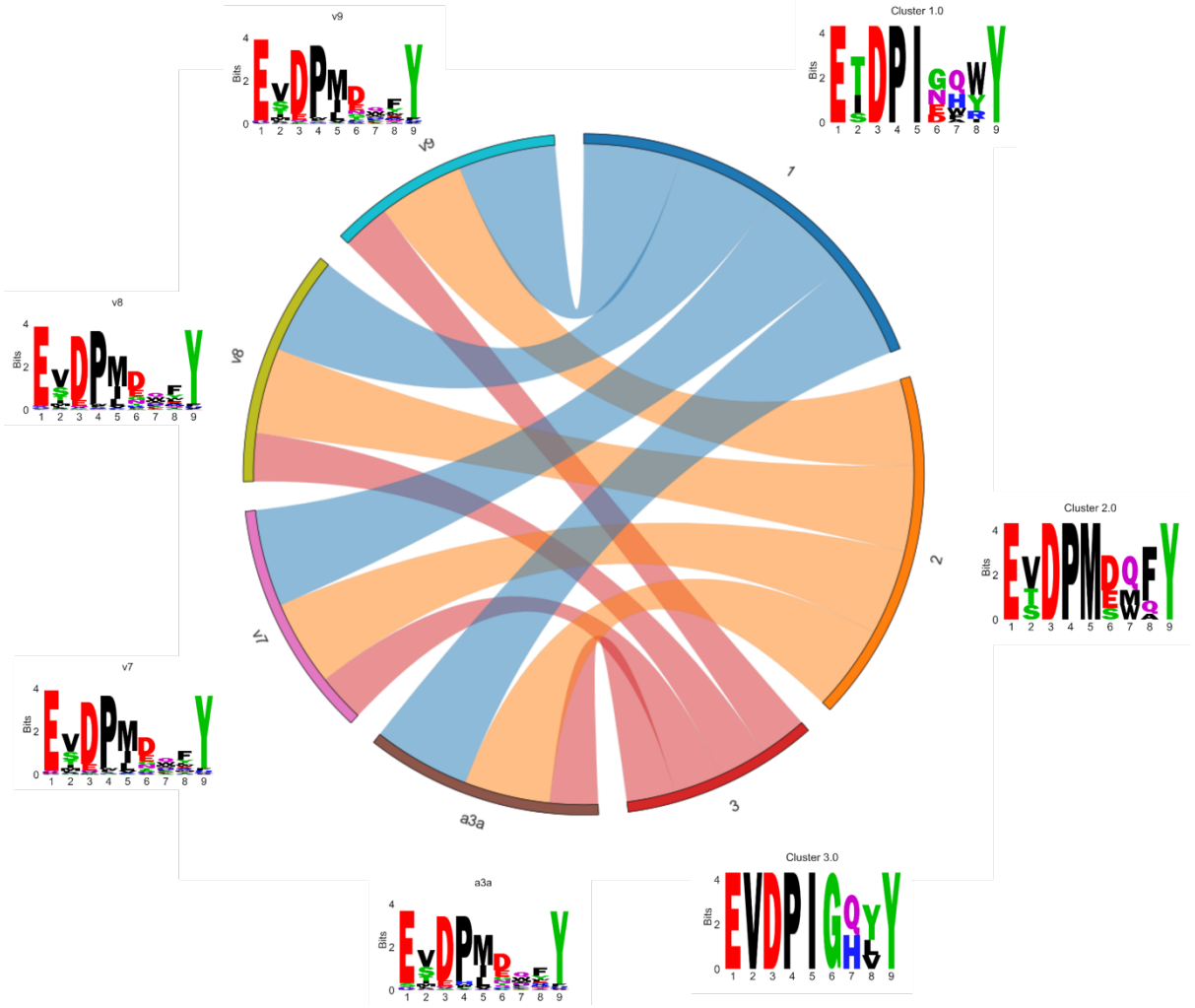


Figure 4-13: **Peptide specificity sharing between constructs, for predicted hits.** Chord diagram for predicted hits, illustrating the extent to which peptide clusters identified through network analysis are shared between variants, and the corresponding logos for those variants.

Discussion

In the present work, we develop a pipeline for quality control, variant prioritisation, deep specificity analysis, and *in silico* de-orphanisation of candidate TCR-Ts. Our quality control programme, fastq2pep, enables the removal of both spurious peptide sequences arising from deep sequencing errors, and of potential contaminants. We find not only that this cleaning step dramatically reduces the time required to generate count tables from deep sequencing runs, but that it also preserves recall of known agonists of the a3a TCR. This feature is particularly relevant in the context of a pipeline whose objective is the de-orphanisation of candidate TCR-Ts in which off-target hits might have life-threatening consequences. Fastq2pep will be made publicly available at www.github.com/hudsondan/fastq2pep.

Turning to the analysis of the resultant datasets for a panel of variants of the a3a TCR, we propose novel metrics of cross-reactivity, dubbed local and global positional entropy, which permit prioritisation of variants v7, v8, and v9 on the basis of reduced cross-reactivity. To our knowledge, these are the first quantitative measures of TCR promiscuity, or indeed of antibody polyreactivity (231), beyond the number of peptides recognised. Further, unlike conventional modes of analysis such as TCR heatmaps and peptide logos, LPE and GPE plots allow like-for-like quantitative comparison of relative cross-reactivity. Later co-culture data confirm that reduced cross-reactivity, as indicated by GPE, translates into improved functional selectivity towards the MAGE family and away from titin, for prioritised variants v8 and v9.

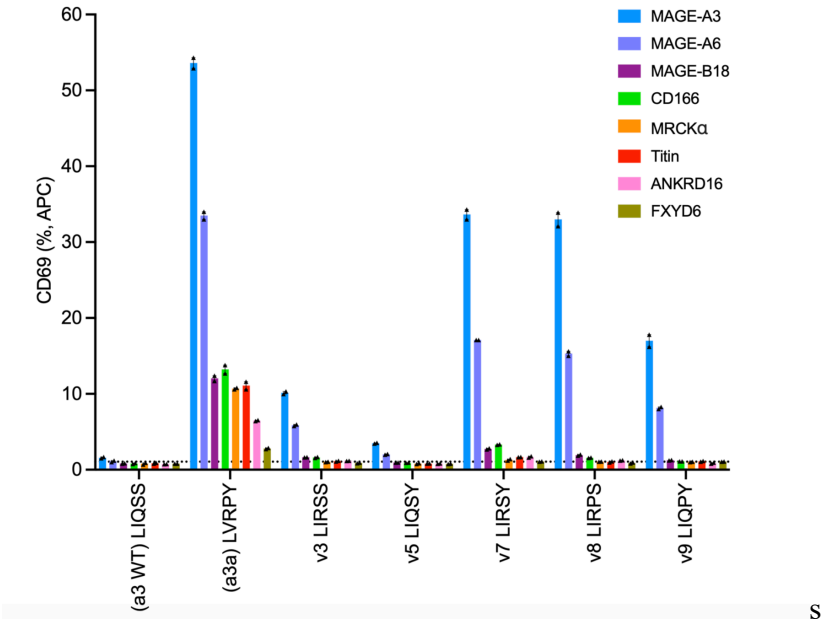


Figure 4-14: Evaluating functional activation of a3a variants by self-peptides. Reactivity of a3a variant TCR-transfected Jurkats towards known self-agonists of the a3a TCR. The polypeptide sequences of the reference agonists are provided in **Table 4-4**.

Applying unsupervised learning methods to dive deeper into the specificity profiles of priority variants, we generate graph networks and chord diagrams to investigate the unique and shared peptide specificity profiles of closely related TCRs. This approach provides not only a convenient representation for inspection of shared specificity but also yields insight into the nuance of TCR epitope binding that is overlooked by existing clustering and inference models in which TCRs are typically assumed to be monospecific (65, 189).

We demonstrate the utility of the Orpheus pipeline in predicting the likely autoreactivity of TCR variants, bridging from analysis of specificity profiles to explicit inference of on and off-target hits. However, not all the predicted hits are functional agonists, as seen for titin with v9. This calls into question the correlation of binding and activation that might be concluded from

the performance of the pipeline in retrieving true agonists of the a3a TCR (**Chapters 3** and **6**). Whilst this remains an open question in the field, in practice the severe consequences of missing off-target true positives during TCR-T screening imply that it is preferable to screen predicted hits and find them to be poor or non-agonists, than to miss them altogether. The potential risk to patients inherent in TCR therapy also signifies that, regardless of future improvements to predictive model performance, comprehensive human or human-like cell-based screening assays will remain an essential part of the discovery process.

Limitations and further work

Time constraints limited the present analysis to ten variants of the a3a TCR. However, these yeast pMHC datasets represent a thorough characterisation of the TCR-pMHC interface comprising thousands of positive and millions of negative TCR-pMHC pairs. These rich datasets could theoretically facilitate construction of receptor-specific binding predictors, and so enable the identification of additional TCR-T candidates. However, the mutations introduced are selectively focused on the CDR2α region as the primary driver of peptide recognition by the a3a TCR: screening of CDR3 variants including those identified in (108, 216) might identify candidates with even greater selectivity towards MAGE-A3.

An arbitrary 50% saturation cutoff was selected as a means of prioritising variants on the basis of improved selectivity, however there is scope to explore the optimal threshold taking account of the characteristics (including depth and skew) of the library. This absolute threshold may also mask differences in the distribution of peptides making up ≤ 50% of the reads, such that a candidate achieving 45% saturation within 5 peptides but 50% by rank 40 may be of greater interest than a receptor with a more gradual rate of saturation. The change in read counts of peptides or peptide clusters over successive rounds (or *velocity*) might yield a more nuanced picture.

The cross-reactivity metrics and graph visualisation methods developed in this study may theoretically be applied to any pMHC dataset. Whilst we have focussed on the a3a TCR, these measures could be applied to other receptors recognising broader families of peptides, including nkg2c, and to other means of characterising TCR cross-reactivity such as phage display and XSM, which we explore in more detail in **Chapter 5**. However, the extent to which the parameters selected for network generation influence the conclusions drawn merit further investigation: including the choice of embedding separation distance and Hamming distance thresholds.

Chapter acknowledgements

The yeast display and functional data for the a3a and variant TCRs included in this study were produced by Julia McCarthy (Fernandes Group).

Chapter 5: Deconstructing the predictive power of binding and functional data in TCR-directed pMHC screens

Overview

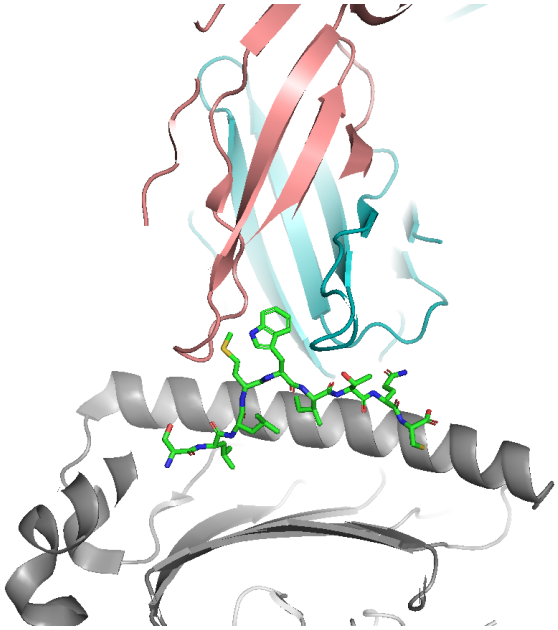
The grand challenge of experimental TCR de-orphanisation may be approached using a broad toolkit of binding or functional assays beyond yeast display. Having previously established that the Orpheus pipeline may be used to infer TCR epitope specificity from yeast pMHC display data, we next set out to determine the extent to which the choice of characterisation method impacts the predictive performance of the model. We find that the pipeline can retrieve the primary agonists of engineered and wildtype receptors using data from diverse characterisation methods, and provides a lens into the relative merits of binding and functional screens, and the role of data completeness, in experimental TCR de-orphanisation.

Introduction

In this Chapter, we assess how the use of different binding measures and functional readouts of TCR-pMHC interaction impacts the predictive capacity of the Orpheus pipeline, using a panel of TCRs recognising the New York oesophageal squamous cell carcinoma 1 antigen (NY-ESO-$1_{157–165}$, SLLMWITQC). The NY-ESO-1 peptide is a human cancer testis antigen preferentially presented by HLA-A*02:01, which was identified through serological analysis of tumour cDNA libraries (232). NY-ESO-1 expression is prognostic of poor response in non-small cell lung cancer (NCSLC), hepatocellular carcinoma (HCC), head and neck cancer, gastrointestinal cancer, multiple myeloma, and malignant melanoma (233). Consequently, NY-ESO-1 and its 9V synthetic analogue SLLMWITQV (234) have been the focus of therapeutic development programmes spanning from cancer vaccines (235), to TCR-based soluble bispecific receptors (236), chimeric antigen receptor (CAR) and TCR-T cell therapies (237, 238), reviewed in (233).

The 1G4 receptor is among the earliest examples of TCR adaptation for therapeutic use, and was originally isolated from NY-ESO-1 specific cytotoxic T cells present in the blood of patients with melanoma (234, 239). Structural studies of the interaction between soluble 1G4 and NY-ESO-1 in the context of HLA-A*02:01 reveal that the interaction hinges on docking of a central Met^4 and Trp^5 protrusion (the “MW peg”) into a pocket formed by 1G4 CDR3α and CDR3β loops (**Fig. 5-1**) (239). Single and double amino acid substitutions of these loops generated affinity-enhanced variants, of which 1G4-α95LY (c259) has entered clinical

development (240-242). We exploit an abundance of binding and functional datasets for natural and affinity-enhanced NY-ESO-1 specific receptors (**Table 5-1**) to determine the generalisability of the Orpheus pipeline beyond yeast pMHC screens.



TCR	CDR3α
1G4	CAVRPTSGGSYIPTF
c259	CAVRPLYGGSYIPTF

Figure 5-1: The 1G4 and c259 TCRs. Left | Interaction between 1G4 with the "MW peg" of NY-ESO-1 (green), showing the pocket formed by 1G4 CDR3α (salmon) and CDR3β (teal). Backbone placement of CDR3α Thr^{95} and Ser^{96} is shown in red, and the flanking helix of the HLA-A*02:01 peptide binding groove in grey. Image produced with PyMol from PDB ID 2BNR (18, 150, 239). **Right** | Amino acid sequences of the CDR3α loops of the 1G4 and c259 TCRs.

Methods

Datasets

Independent datasets were obtained for a panel of five TCRs recognising the NY-ESO-$1_{157–165}$ 9-mer in the context of HLA-A*02:01: 1G4 and its variant c259, and the wildtype human TCRs NYES1-3 (**Table 5-1**) (107, 211). These studies deployed diverse binding and functional characterisation methods, including surface plasmon resonance (SPR), bacteriophage display, tetramer staining, NFAT-GFP reporter and killing assays (69, 107, 211). 1G4 and c259 binding affinity screens were generated by Anna Huhn and Jose Cabezas Caballero using a Carterra instrument and were kindly provided by Omer Dushek (University of Oxford). Raw K_D values were averaged over three repeats, and converted to pseudoreads as described in **Eq. 5-1a-b.**

c259 functional data were obtained from (211), and transformed as described in **Eq. 5-1c-d**. Bacteriophage display data for NYES1-3 derived from (107) were kindly provided by Dr Stephen Harper of Immunocore, Ltd.

Table 5-1: Summary of NY-ESO-1 specific TCR datasets analysed.

Source	TCR	Library	Method	Class	Uniques	K_D (9V)
Dushek lab	c259	XSM	SPR	Binding	136	820 nM
(211)			Tetramer stain	Binding	135	
			NFAT-GFP	Functional	135	
			Killing	Functional	135	
Dushek lab	1G4	XSM	SPR	Binding	136	3.7 μM
(107)	NYES1	Fully random	Phage display	Binding	4832	7.0 μM
	NYES2	Fully random	Phage display	Binding	5415	7.1 μM
	NYES3	Fully random	Phage display	Binding	7053	82-186 μM

Data subsets were produced by deleting peptides absent in either SPR or tetramer binding datasets. Blended datasets were generated by summing counts from all c259 data; binding (SPR) and activation data; or SPR and killing data. Enriched peptides were defined for all datasets as those having 5 or more pseudoreads in the input, following optimisation experiments conducted in **Chapter 3**.

Equation 5-1: Pseudoread transformation. Transforming functional data into pseudoreads, where the input is a solution state dissociation constant (K_D) (**Eq. 5-1a-b**) or a functional readout (**Eq. 5-1c-d**).

a) $Pseudoreads_{sum} = D * k_{norm}$

$$k_{norm} = \frac{k_i^{-1}}{\sum_i^j k_j^{-1}}$$

$$D = \frac{C}{\max(k_{norm})}$$

Where C is a constant (200,000 by default), k = mean K_D

b) $\text{Pseudoreads}_{minmax} = C * (1 - k_{norm})$

$$k_{norm} = \frac{k_i - k_{min}}{k_{max} - k_{min}}$$

c) $Pseudoreads_{sum} = D * x_{norm}$

$$x_{norm} = \frac{x_i}{\sum_i^j x_j}$$

$$D = \frac{C}{\max(x_{norm})}$$

Where C is a constant (200,000 by default), x = mean value over repeats

d) $\text{Pseudoreads}_{minmax} = C * x_{norm}$

$$x_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Heat maps

TCR heat maps were generated from enriched peptides and represented as matrices of normalised frequencies derived as set out in **Eq. 3-1**. Negative control peptide profiles were computed using baseline PWMs trained on a random selection of 1000 self-peptides drawn from the UniProt consensus proteome. HLA-A*02:01 specific peptides were isolated as the top 1000 self-peptide hits predicted by an MHC allele-specific PWM trained on elution mass spectrometry data obtained from (20, 215). All plots were produced using the Matplotlib and Seaborn Python packages (201, 202).

Predicted hits

Predictions were made for a consensus human proteome of 1.1 x 10^7 9-mer peptides derived from UniProt (114), topped up with known agonists of the TCRs being studied, using baseline or model variants as described in **Chapter 3**. No filter was applied to peptides enriched in the naïve library, which was assumed to be zero for all peptides in the case of 1G4 and c259 data. For NYES1-3, the naïve round was simulated as a random draw with replacement, using preselection amino acid frequencies kindly provided by Immunocore, Ltd.

A synthetic library of 5.6 x 10^7 fully random 9-mer peptides was generated with a random draw with a probability of 1/20 for each of the common natural amino acids. For ratios of 1:1 and 1:2 for human wildtype to synthetic peptide mixtures, subsets of the full synthetic library were extracted such that the subsets had length equal to or double the length of the UniProt reference, respectively, merging the resulting peptides with the UniProt reference.

Results

TCR-pMHC SPR data were kindly provided by the Dushek group for 1G4 and c259. Tetramer binding, activation, and killing data were obtained from (211). Each of these studies used of a panel of peptides identified through XSM, in which single residues of NY-ESO-1 9V were sequentially exchanged for any of the 20 common naturally occurring amino acids. The representation of amino acids in each dataset is summarised in **Fig. 5-2** and **Fig. 5-3**. Cysteine substitutions were not included for SPR datasets, as discussed later (**Fig. 5-12**). Bacteriophage selection data for three wildtype NY-ESO-1 specific TCRs derived from healthy donor PBMCs (NYES1-3) were obtained from a fully random pMHC library originally published in (107).

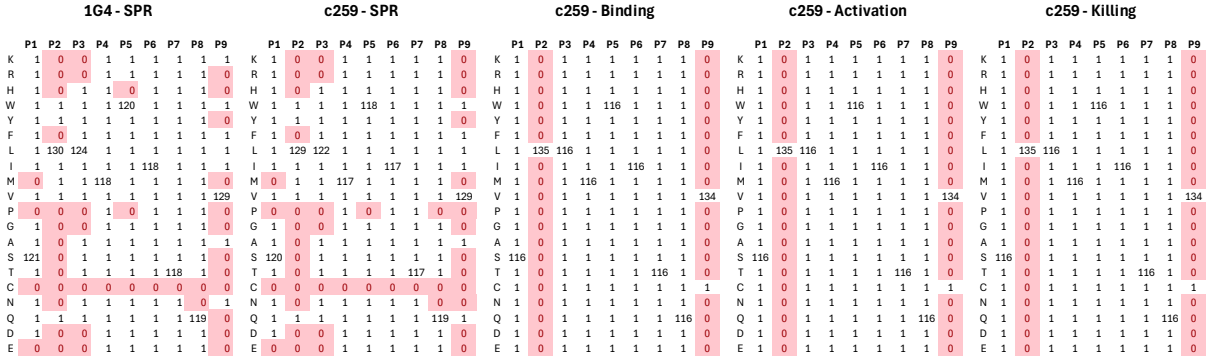


Figure 5-2: Missing data, therapeutic TCRs. Data completeness for 1G4 and c259 datasets, showing dataset counts for each amino acid at each position of the 9-mer SLLMWITQV peptide. Colour scale showing missing values in red.

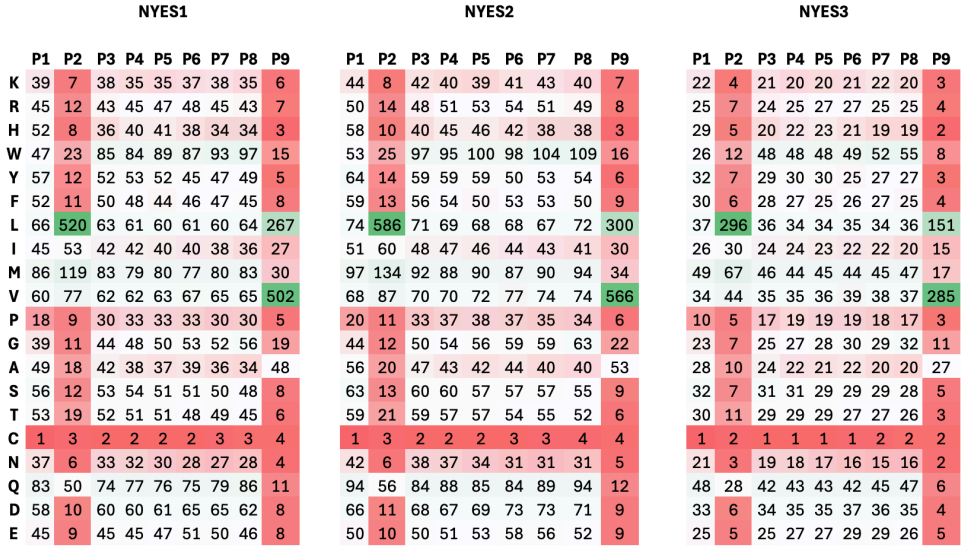


Figure 5-3: **Missing data, natural TCRs**. Data completeness for phage display screens of NYES1-3, showing dataset counts (in 000's) for each amino acid at each position the 9-mer SLLMWITQV peptide. Colour scale shows lowest counts in red and highest counts in green.

Each dataset was first transformed into a common "pseudoread" space suitable for application of the Orpheus model (**Eq. 5-1**). Transformation with a sum denominator (**Eq. 5-1a**, **5-1c** and **Fig. 5-4**) generated datasets with steeper cumulative frequency curves (**Fig. 5-5**), more pronounced heatmaps (**Fig. 5-6** and **5-7**), and improved predictive accuracy (**Fig. 5-8**) compared to those produced by minmax normalisation (**Eq. 5-1b** and **5-1d**). Sum transformation was therefore applied for all later analytical and predictive tasks.

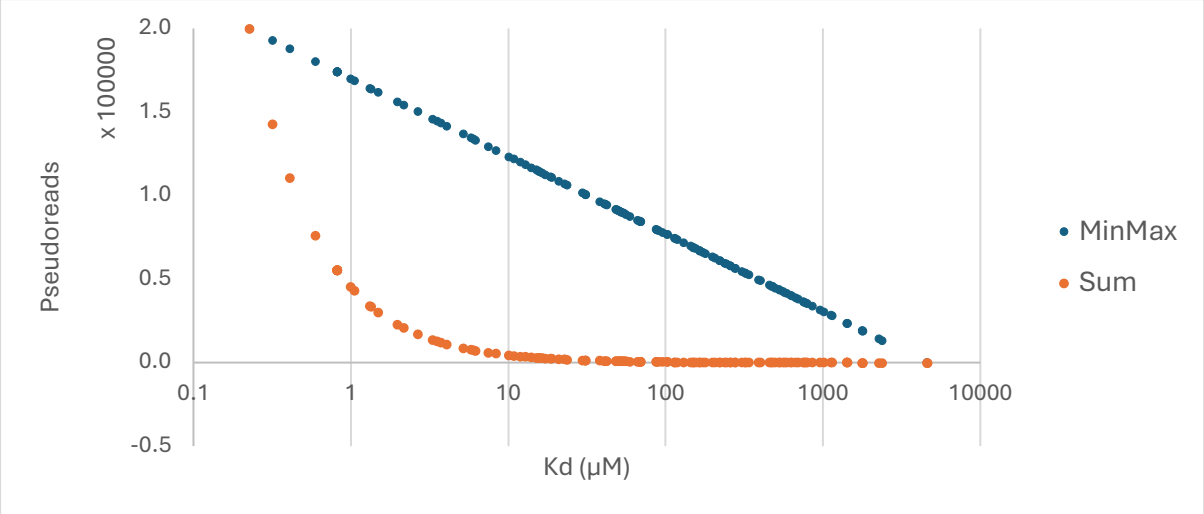


Figure 5-4: Minmax and sum (Entropy) pseudoread transformation of K_D data for c259 (**Eq. 5-1**).

As might be expected from their relative affinities for the NY-ESO-1 9V peptide (820nM and 3.7 μM, respectively, values provided by the Dushek lab), c259 showed steeper saturation in the cumulative frequency of enriched peptides than did 1G4, with fewer peptides contributing a greater proportion of the overall reads (**Fig. 5-5**). Comparison of footprints for 1G4 and c259 SPR data revealed highly similar recognition patterns, with only subtle differences in observed preferences discernible at positions 6 through 8 (**Fig. 5-6** and **5-7**). c259 footprints derived from tetramer and functional analysis (211) showed a greater degree of diversity in tolerated amino acid substitutions at the N terminus of the 9-mer than would be concluded from SPR data alone.

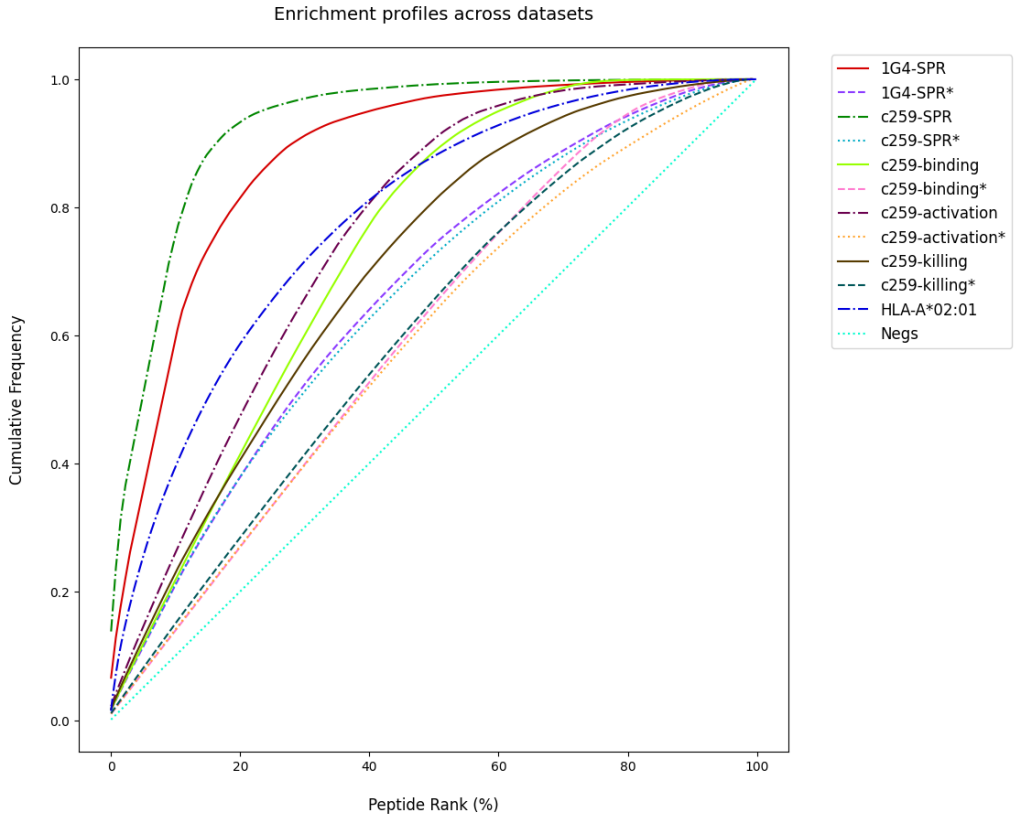


Figure 5-5: Cumulative frequency of enriched peptides with different pseudoread transform methods. Datasets marked with an asterisk are minmax normalised, otherwise produced using a sum transform.

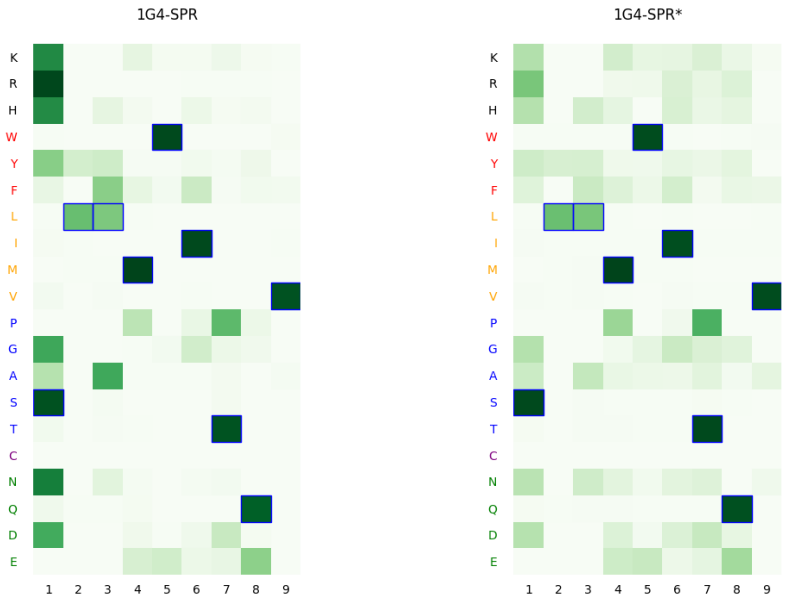


Figure 5-6: Comparing heatmaps for 1G4 with different pseudoread transform methods. Datasets marked with an asterisk are minmax transformed (**Eq. 5-1b** and **5-1d**) otherwise sum transformed (**Eq. 5-1a** and **5-1c**).

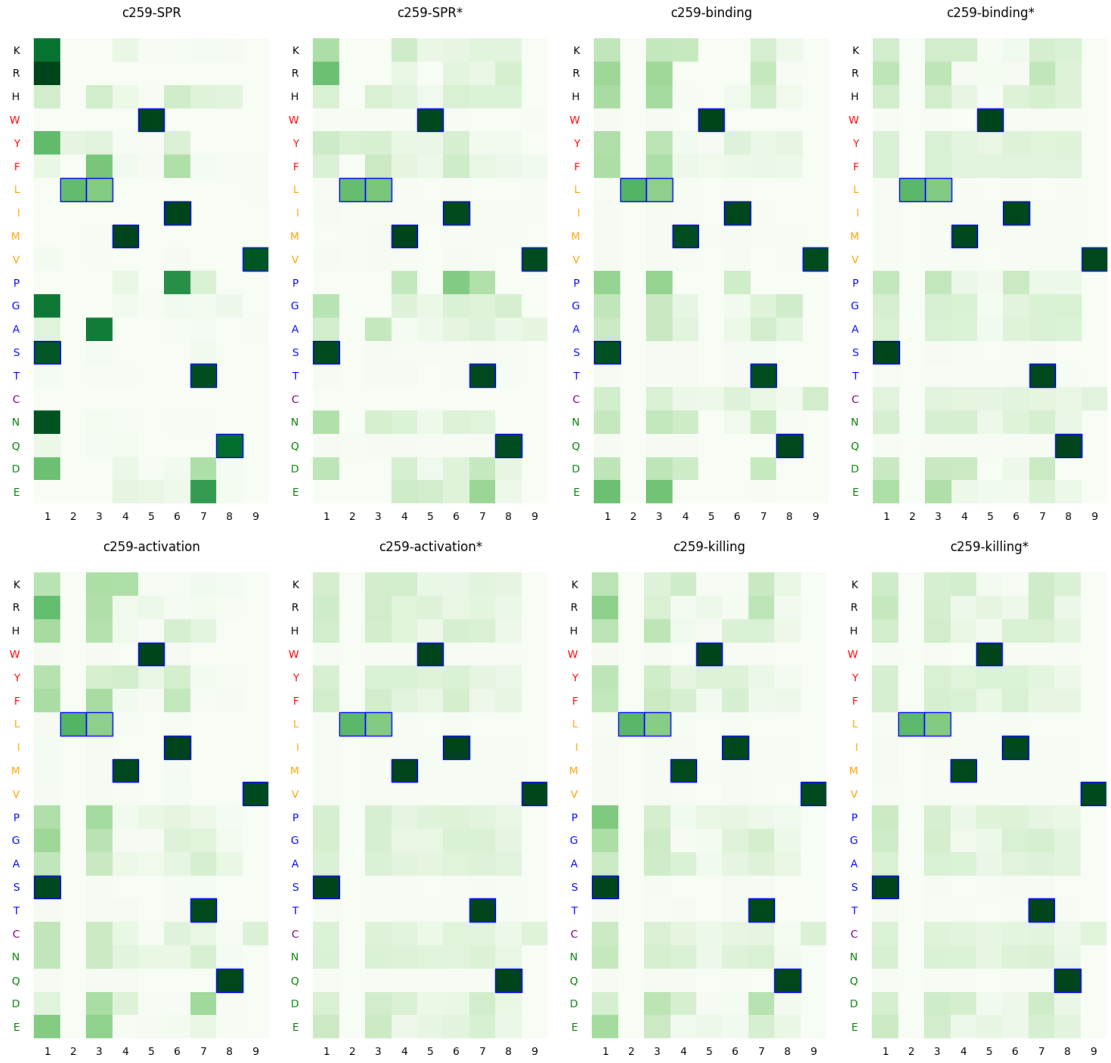


Figure 5-7: Comparing heatmaps for c259 with different pseudoread transform methods. Datasets marked with an asterisk are minmax transformed (**Eq. 5-1b** and **5-1d)** otherwise sum transformed (**Eq. 5-1a** and **5-1c**).

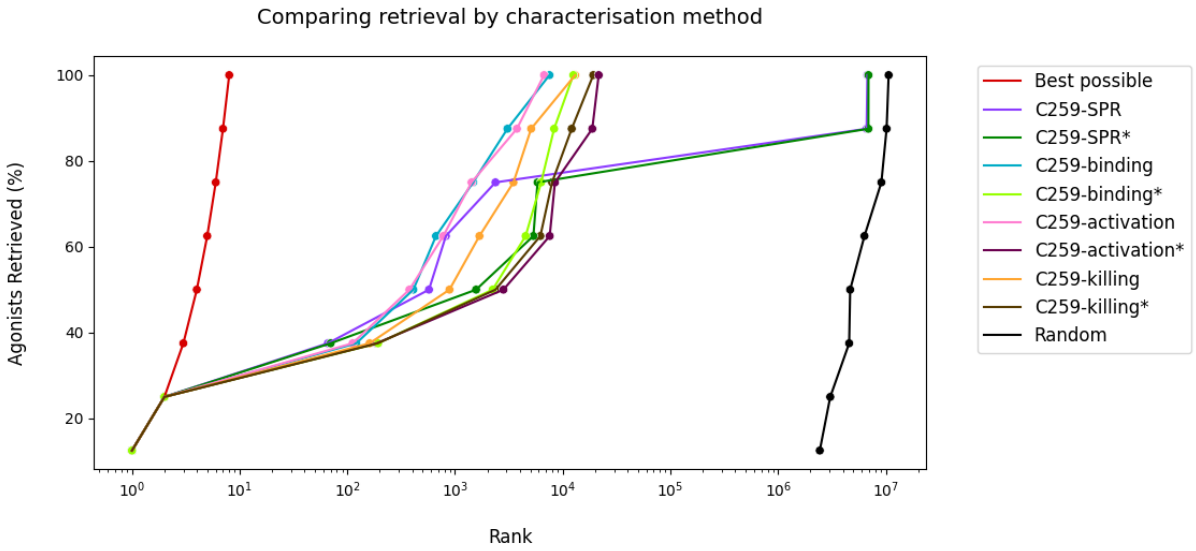


Figure 5-8: Predictive accuracy across transformation methods. Datasets marked with an asterisk are minmax transformed (**Eq. 5-1b** and **5-1d)** otherwise sum transformed (**Eq. 5-1a** and **5-1c**).

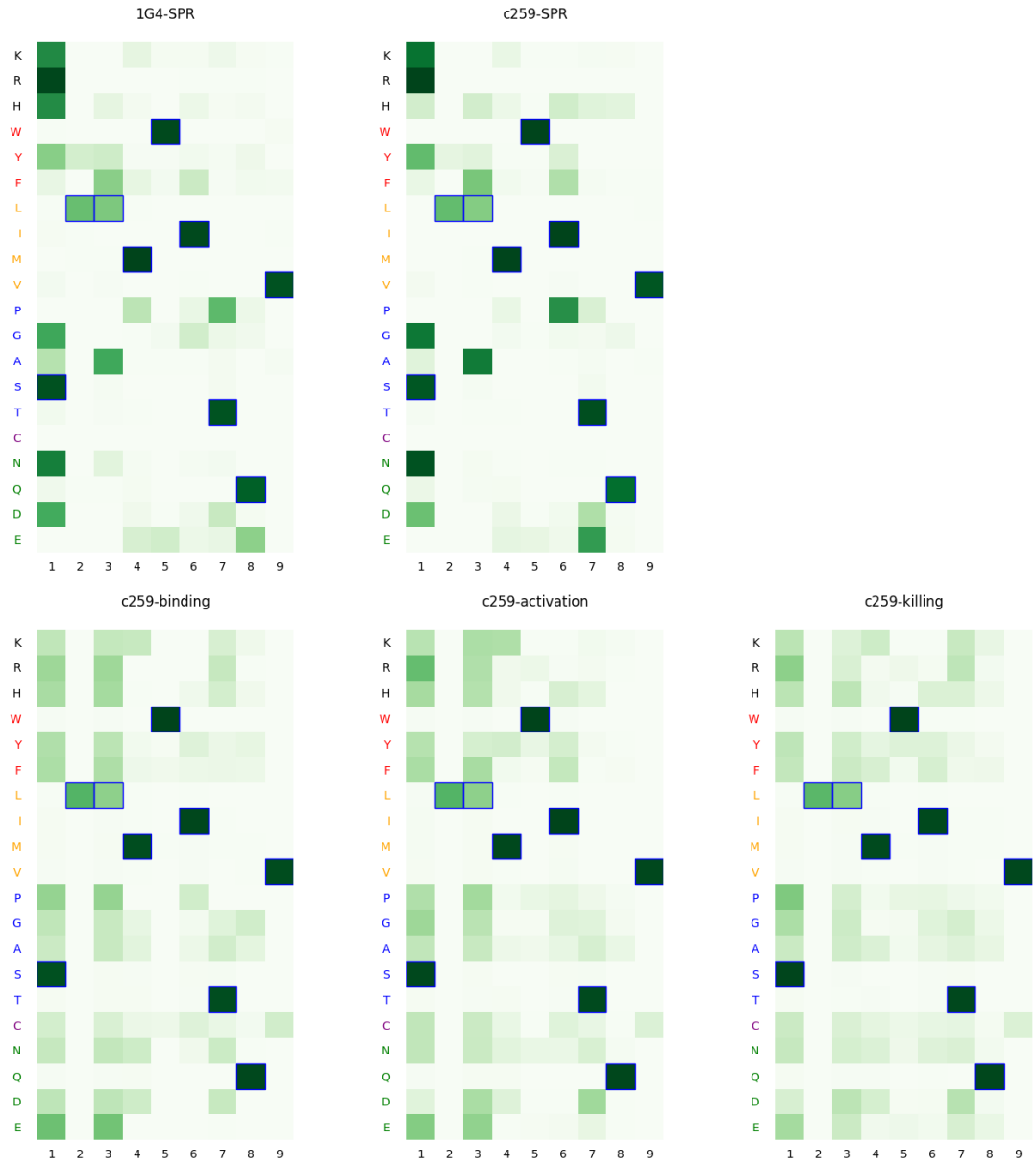


Figure 5-9: TCR footprints for NY-ESO-1 specific TCRs across characterisation methods. Top | SPR data. **Bottom |** Binding and functional data from (211).

The predictive performance of the Orpheus pipeline was next evaluated for each dataset using baseline and optimised models (see **Chapter 3**), with the objective of retrieving all the reported agonists of the 1G4 or c259 TCRs (**Table 5-2** and **5-3**) as high as possible in a ranked list of peptides drawn from the UniProt consensus proteome (114). As for yeast data screens (**Chapter 3**), an optimised model outperformed the baseline PWM across all datasets (**Fig. 5-10**). The Orpheus pipeline was able to retrieve 9V and wildtype NY-ESO-1 at positions 1 and 2 of 1.1 x 10^7 reference peptides, irrespective of the TCR or the characterisation method. However, differences in retrieval performance were observed for other agonists between c259 studies (**Fig. 5-11**, **Table 5-2** and **5-3**).

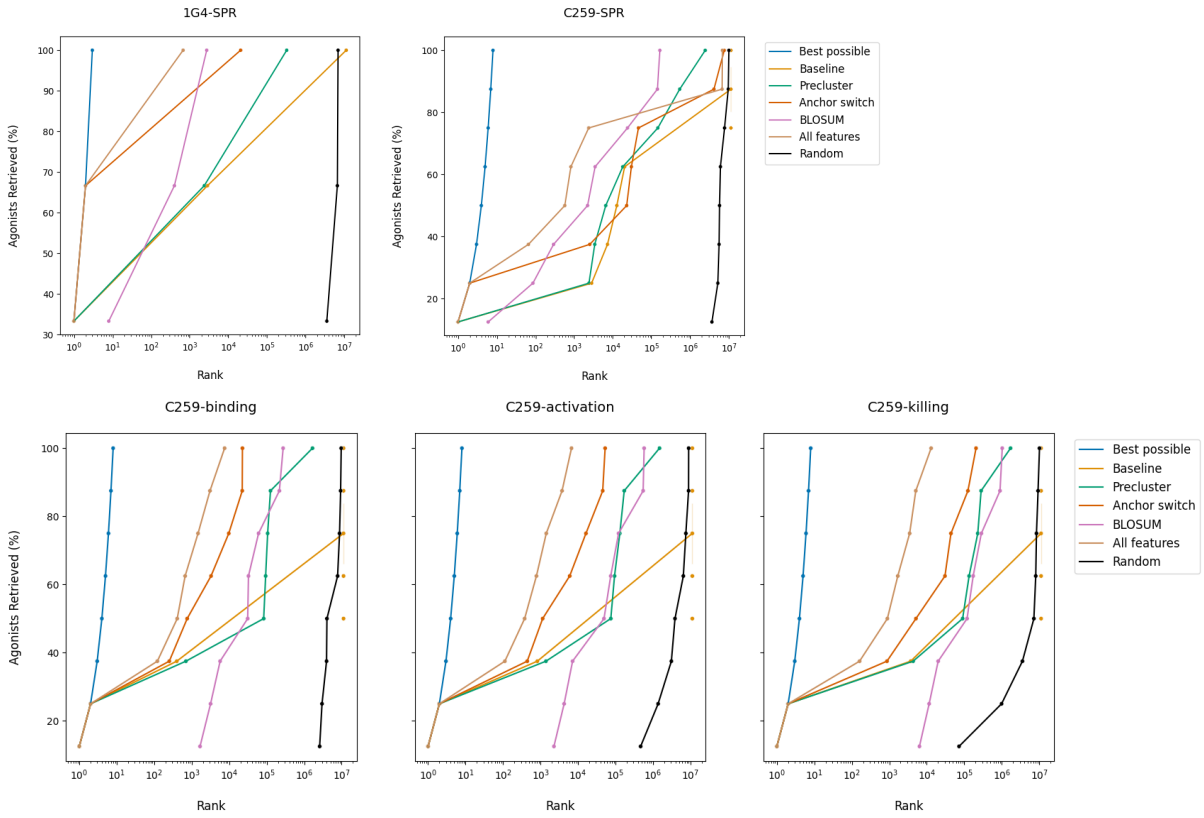


Figure 5-10: Model configurations. Comparing model performance across datasets and feature representations, showing ranked retrieval of known agonists of the 1G4 (69) and c259 (211) TCRs.

Activation and multimer binding data for c259 outperformed SPR and killing datasets in ranked retrieval of known agonists (**Fig. 5-11**), all known agonists being returned by rank 6711 and 7532, respectively (**Table 5-3**). The performance difference between functional binding and SPR screens was most notable for MLAQWCTQA and TLLLWLCQA, where Orpheus predictions made using SPR data failed to identify these agonists. Inspection of amino acid count data (**Fig. 5-2**) revealed that this discrepancy was most likely due to missing values for methionine at position 1 in the case of MLAQWCTQA, and missing cysteine values at position 6 and 7 for MLAQWCTQA and TLLLWLCQA, respectively, in the SPR dataset. Indeed, reducing SPR and multimer binding datasets to a common subset of input peptides resulted in similar retrieval traces, suggesting that data completeness plays an important role in recall for some agonists (**Fig. 5-12**). However, combining SPR and functional datasets negatively impacted retrieval performance when compared with the individual functional datasets alone (**Fig. 5-13**).

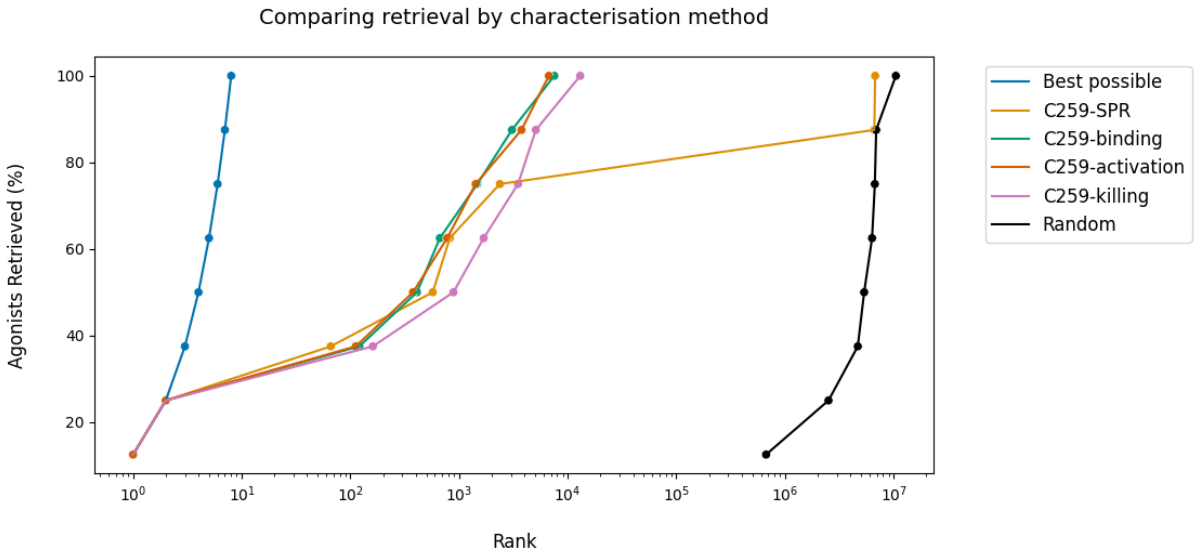


Figure 5-11: Performance by characterisation method. Showing ranked retrieval of known agonists of the c259 TCR with an optimised PWM model.

Table 5-2: Ranked retrieval of known agonists of 1G4. Showing the position of experimentally confirmed functional agonists identified in (69). Predictions made with a combined PWM model.

Agonist	SPR
SLLMWITQV	1
SLLMWITQC	2
FLTLWLTQV	672

Table 5-3: Ranked retrieval of known agonists of c259. Showing the position of experimentally confirmed functional agonists of c259 identified in (211). Predictions made with a combined PWM model.

Agonist	SPR	Binding	Activation	Killing
SLLMWITQV	1	1	1	1
SLLMWITQC	2	2	2	2
NVLLWITAL	66	121	112	161
FLTLWLTQV	573	410	377	888
QVFLWLAQV	822	666	778	1685
MLAQWCTQA	6752981	1473	1424	3489
TLLLWLCQA	6641162	3069	3756	5093
TQIQWATQV	2375	7532	6711	13046

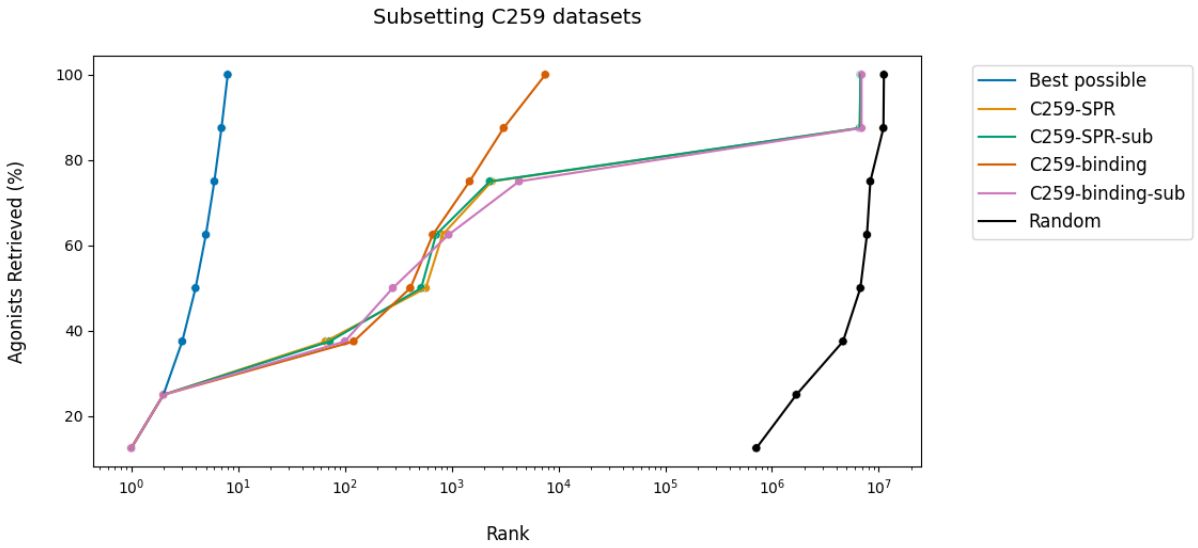


Figure 5-12: Subset analysis. Performance between binding datasets with full and reduced (sub) peptide sets, showing ranked retrieval of known agonists of the c259 TCR with an optimised PWM model.

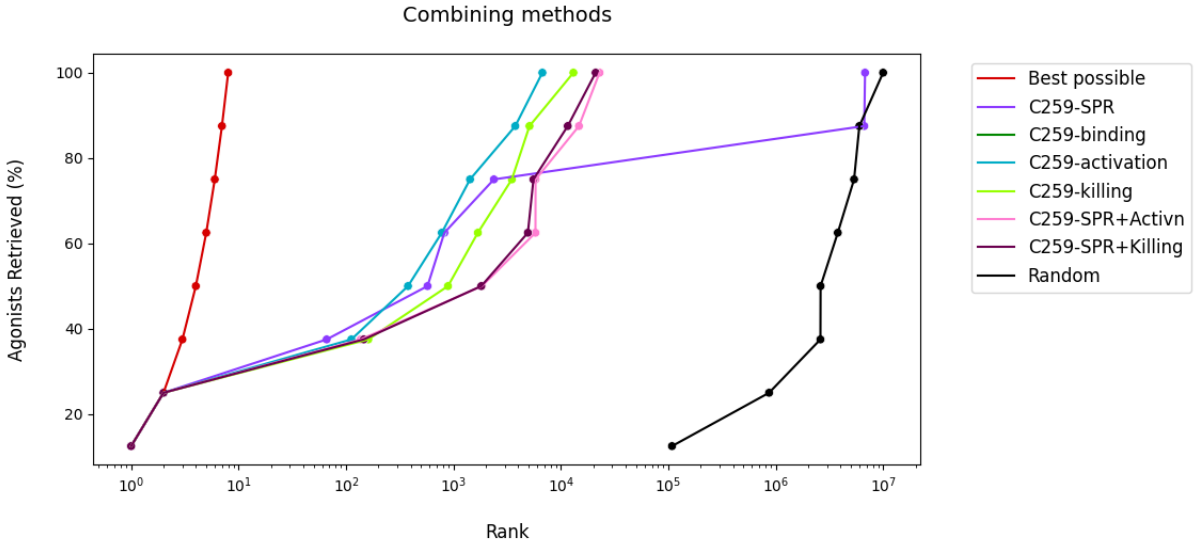


Figure 5-13: Merging characterisation methods. Blending SPR and functional data, showing ranked retrieval of known agonists of the c259 TCR with an optimised PWM model.

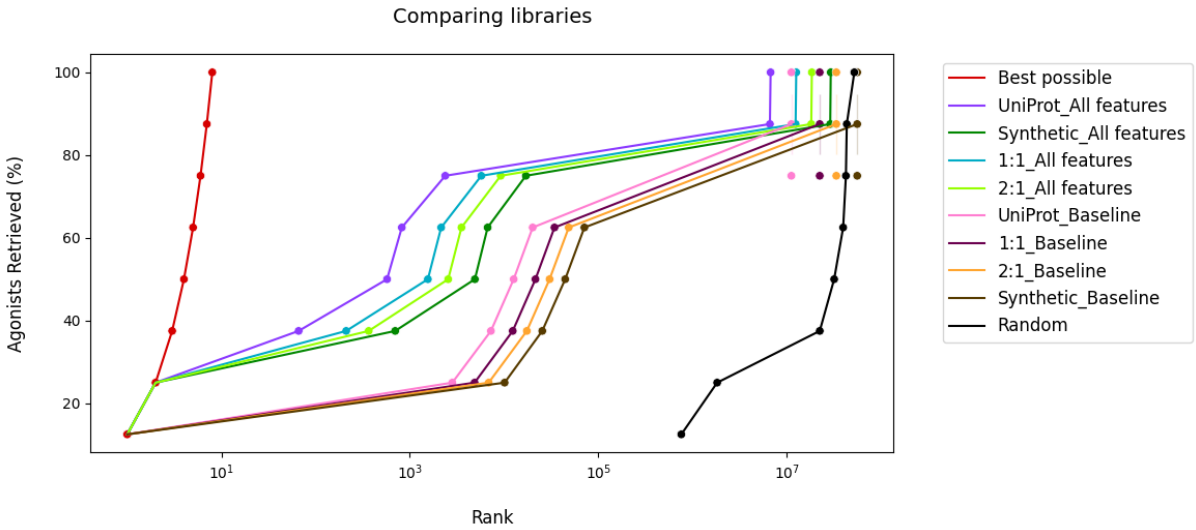


Figure 5-14: Making predictions for non-physiological peptides. Extending the predictive pipeline to synthetic reference libraries, showing ranked retrieval of known agonists of the c259 TCR with a baseline or combined ("All features") PWM model. Models trained using SPR data from the Carterra instrument.

Thus far, the Orpheus pipeline had only been applied to retrieval of agonists present in or added to the consensus human proteome, as a means of screening the likelihood of a given TCR recognising self-peptides. To establish whether the platform could identify known agonists from peptides of any origin, a synthetic library of 5.6×10^7 9-mers was generated through a random draw from each of the 20 natural amino acids at each position. The predictive pipeline was then applied to the fully random library, or to a 1:1 or 1:2 mix of human and synthetic peptides (**Fig. 5-14**). The Orpheus pipeline was able reproducibly to retrieve NY-ESO-1 9V and wildtype peptides at positions 1 and 2, even when drawing from a randomly generated synthetic library of 56 million possible targets, using SPR data gathered with the Carterra instrument (**Table 5-4**). Comparison of the top predicted hits between human and synthetic libraries revealed similar but distinct peptides (**Table 5-5**).

Table 5-4: Predictions for non-physiological peptides. Ranked retrieval of known agonists of c259 using consensus human, fully synthetic or mixed libraries with an optimised PWM model trained on SPR data.

Agonist	Human	1:1	1:2	Synthetic
SLLMWITQV	1	1	1	1
SLLMWITQC	2	2	2	2
NVLLWITAL	66	211	366	699
FLTLWLTQV	573	1554	2553	4901
QVFLWLAQV	822	2151	3532	6696
TQIQWATQV	2375	5736	9211	17048
TLLLWLCQA	6641211	12444466	18243664	28999978
MLAQWCTQA	6753031	12630316	18504097	29370199

Table 5-5: Top predicted hits for an optimised PWM trained on SPR data for c259. Highlighted peptides: NY-ESO-1 (WT and 9V).

Rank	UniProt	1:1	1:2	Synthetic
1	SLLMWITQV	SLLMWITQV	SLLMWITQV	SLLMWITQV
2	SLLMWITQC	SLLMWITQC	SLLMWITQC	SLLMWITQC
3	DLLMEIEQV	DLLMEIEQV	DLLMEIEQV	RQFMWIPQM
4	YLAMYPDQV	GMANWIDQL	GMANWIDQL	HLFMYITQI
5	RALMYIPQL	GMHMWPTQA	KVNMWIEQV	GMANWIDQL
6	KMYMEFPQL	RQYDWITQV	GMHMWPTQA	KVNMWIEQV
7	RMHMWVEDV	YMLMWETQI	HMAMWIHRL	FMWMWIEQL
8	KVADYIPQL	YVYMYITQA	RQYDWITQV	GMHMWPTQA
9	TVAMWIPGV	YLAMYPDQV	YMLMWETQI	GTAMWIERV
10	SLFGWISQL	NTLMWIPEV	YLFTWIPQV	KVFMWLHQV
11	RLFKHIEQL	RALMYIPQL	YVYMYITQA	HMAMWIHRL
12	SLALWVKQV	NLAMEINQV	YLAMYPDQV	RLMMWIDQC
13	DLLEWIEQT	GMHMFIHQI	YMAMYATQL	YQSMWIEQV
14	RLYKWIPQN	GMALFITQV	NTLMWIPEV	RQYDWITQV
15	NLADWYSQV	KMYMEFPQL	HMAMWREQL	YMLMWETQI
16	GLLNWITGA	GMFLWITHV	RCFMWIPQY	YLFTWIPQV
17	KTALWITRI	RMHMWVEDV	RALMYIPQL	YVYMYITQA
18	SCALWIPEV	KVADYIPQL	NYFMWLEQV	SQYMWVTQA
19	RLYDWIPQN	NQYMWISQM	NLAMEINQV	YMAMYATQL
20	YLFDEIDQA	KVADWITRI	GMHMFIHQI	NTLMWIPEV

Considering the significant role of data completeness in dictating model performance (**Fig. 5-12**), the Orpheus platform was next applied to wildtype NY-ESO-1 specific TCRs characterised using phage display of diverse peptide libraries (**Table 5-1**) (107). Briefly, in that study NYES1-3 TCRs were identified from healthy donor PBMCs after exposure to NY-ESO-1 pulsed APCs. The binding preferences of each of these three receptors was then characterised using bacteriophage display with a random peptide-HLA-A*02:01 single chain trimer library as described in (107). The resulting randomised library was considerably deeper than those generated by XSM (**Table 5-1**), with improved coverage over anchor positions (**Fig. 5-3**). Inspection of the TCR heat maps of NYES1-3 (**Fig. 5-15**) confirmed the distinct binding modes reported in (107), wherein NYES1 and S2 prefer peptides with the characteristic p5 tryptophan of the "MW peg", as seen for 1G4 and c259, whilst NYES3 tolerated a greater diversity of amino acids at this position and indeed along the entirety of the SLLMWITQV 9-mer.

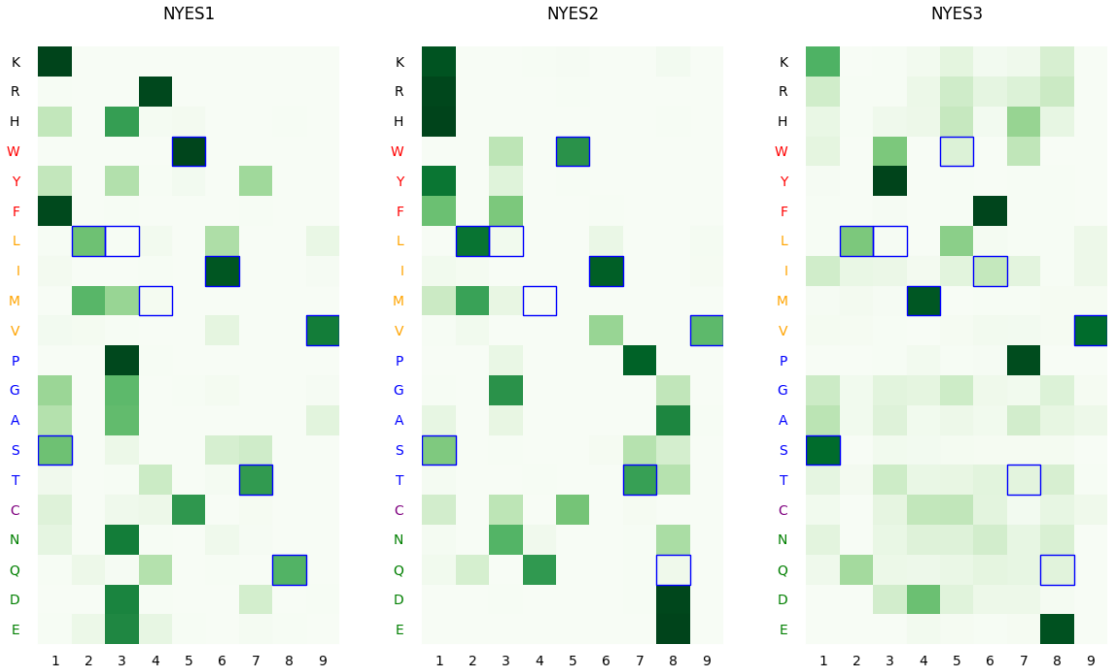


Figure 5-15: TCR heat maps for NYES1-3.

Applying the predictive pipeline to phage display data for NYES1-3, the best model was able to retrieve NY-ESO-1 by positions 6, 30, and 733 for NYES1, 2 and 3 respectively. Unlike in prior yeast display, binding, and functional experiments, a baseline model performed similar to or better than the optimised model (**Fig. 5-16**).

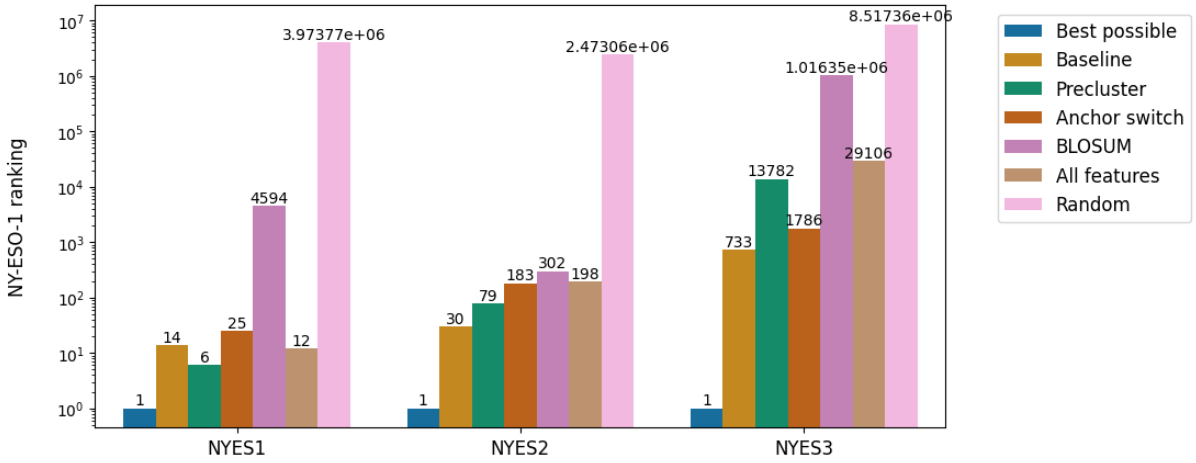


Figure 5-16: Performance by model architecture for NYES1-3 TCRs. Showing the rank at which each model configuration retrieved NY-ESO-1 from a reference proteome.

Discussion

We demonstrate that the Orpheus pipeline can retrieve the cognate agonists of varied NY-ESO-1 specific TCRs, leveraging data determined using a variety of experimental methods and readouts and spanning both engineered and wildtype TCRs. Comparing pipeline performance for the c259 TCR across methods provides a lens into the comparative performance of binding and functional screens. Notably, multimer binding and activation data appear to outperform datasets produced with the Carterra instrument or killing assays. However, in the context of the relatively shallow peptide libraries generated using XSM, we find that retrieval performance is driven by dataset completeness more than the choice of characterisation method. Simply blending XSM datasets does not yield the best performance (**Fig. 5-13**), suggesting a more nuanced relationship between signal-to-noise ratio, coverage of amino acids at each position in a target peptide, binding vs. functional data, and recall of weaker agonists.

The relative ranking of the NY-ESO-1 9V peptide correlates with TCR affinity, such that the best ranking for the c259 (820 nM), 1G4 (3.7 μM), NYES1 (7.0 μM), NYES2 (7.1 μM), and NYES3 (≥ 82 μM) TCRs was position 1, 1, 14, 30 and 733 of 1.1 x 10^7 respectively. Comparison of retrieval performance using different model architectures reveals that features such as anchor switching, developed to compensate for hard-coded anchor assumptions in the input library, do not translate to improved performance with fully random phage display libraries. Caution is therefore needed when applying the model to libraries produced using different methods.

Limitations and further work

We demonstrate the capacity of the Orpheus pipeline to retrieve the cognate agonists of the c259 TCR with SPR, multimer and functional data, and for the NYES1-3 TCRs using phage display. Unfortunately, no datasets were available either for phage display of the c259 TCR, or for binding and/ or functional characterisation of NYES1-3. Furthermore, the NY-ESO-1 SLLMWITQV peptide negatively impacts the recognition of the Ag2 signal peptides and precludes yeast pMHC display screening of NY-ESO-1 specific TCRs. Together, these factors prevented a like-for-like analysis of yeast, phage and XSM display methods across c259 or NYES1-3 TCRs. Functional validation of top predicted binders (**Table 5-5**) is ongoing, which may provide further insight into both subdominant recognition motifs and enable a more detailed evaluation of the factors driving differential model performance (**Fig. 5-16**).

Chapter acknowledgements

The Dushek Group and Immunocore Ltd. provided NY-ESO-1-specific TCR datasets. Thanks are due to both for the kind provision of these data and for feedback on the analysis performed.

Chapter 6: Discussion

In this DPhil research project, we have investigated the capacity of supervised and unsupervised models to "de-orphanise" TCRs of clinical interest, in the context of both TCR- and antigen-directed discovery. We have identified major limitations in computational specificity inference with both SPMs and UCMs, turning instead to synthetic peptide display libraries as a rich source of binding data for individual receptors. Here, we summarise the impact and limitations of the project and explore promising avenues for future research.

Generalisable inference of TCR specificity remains unsolved

Despite the exponential growth in immune repertoire data and advances in data science and artificial intelligence, quantitative immunology still lacks a systematic and generalisable framework for inference of antigen specificity for orphan TCRs. We (65) and many others (93, 116) have noted improvements in predictions made for "seen" epitopes. However, notwithstanding the release of many new models in the interim (243-249), predictive accuracy for unseen epitopes remains stubbornly close to a random guess (93). Which then are the most promising potential paths forward? Significant challenges include the complexity of the underlying immunology, restricted availability of TCR-epitope label pairs, and limitations in modelling approaches. Overcoming these challenges may bring us closer to realising the "holy grail" of specificity inference.

Improving modelling of the underlying immunology

The underlying immunological correlates of T cell interaction with their cognate ligands are highly variable and only partially understood, with critical consequences for model design. Importantly, TCR antigen specificity inference is just one part of the larger puzzle of antigen immunogenicity prediction (123) which can be condensed into three key phases: antigen processing and presentation by MHC; TCR recognition; and functional T cell activation.

Antigen processing and presentation pathways have been extensively studied, and computational predictors of pMHC binding affinity have achieved near-perfect ROC-AUC for some alleles (122, 215). However, this problem is far from solved, particularly for MHC class II and for less frequent class I alleles (101). Mechanistic modelling of processing and presentation pathways (250), more sensitive peptide characterisation methods (251, 252), and class-specific modelling advances (253) could help to close this gap. Likewise, recent additions to our knowledge of the binding profiles of non-classical MHC (112, 254) provide hope for

extending our understanding of TCR antigen specificity to unconventional T cells and non-peptide antigens.

The inherent ability of TCRs to recognise multiple targets poses a fundamental challenge to specificity inference. The findings of **Chapters 3-5** add to previous studies (105, 107, 109, 112) demonstrating how synthetic peptide display methodologies can be leveraged to characterise receptor promiscuity. However, these TCR-directed methods focus on individual receptors, limiting their utility for more generalisable computational solutions. Although display data offer a valuable source of positive and negative TCR-pMHC training examples, PWMs trained on TCR-specific datasets cannot generalise to other receptor-antigen combinations. Library-on-library methods such as receptor-antigen pairing by targeted retroviruses (RAPTR) (94) may bridge the gap between antigen- and TCR-directed methods, however the scalability of these systems remains uncertain.

Many groups have attempted to bypass the complexity of the TCR-pMHC interface by predicting antigen immunogenicity independent of the TCR, as a direct mapping from peptide sequence to some functional measure of T cell activation (reviewed in (121)). However, similar limitations have been encountered for those models as have been described for specificity inference. Many predictors are trained using epitopes from the Immune Epitope Database labelled with readouts from single time points (101). The definition of functional activation is also both variable and highly sensitive to study conditions. Achar and colleagues have demonstrated that integrating cytokine responses over time can improve prediction of response for a limited set of peptides, however these studies do not link peptide response to individual or related TCRs (255). Antigen load and affinity may also play important roles (256, 257).

Thus, models capable of predicting functional T cell response will likely need to bridge from antigen presentation to TCR antigen recognition, T cell activation and effector differentiation, and to integrate complex tissue-specific cytokine, cell phenotype, and spatiotemporal datasets. Whilst T cell-independent predictors of immunogenicity have clear translational benefits, only after we can dissect the relative contribution of the three stages described above will we truly understand what determines antigen immunogenicity.

Harnessing new data sources

Compared to the enormous diversity of the TCR-pMHC interface, publicly available repositories of TCR-epitope pairs remain severely limited (**Fig. 1-3**). However, recent consolidation efforts have produced carefully curated datasets of 5.3 million redundant and 1.6 million non-redundant TCR sequences, annotated with valuable biological metadata (15). This provides an opportunity to boost model performance through *transfer learning*: a technique in which models are trained on a large volume of unlabelled or partially labelled data, and the

patterns learned from those datasets are used to inform a second predictive task. This technique has been widely adopted in computational biology, including in predictive tasks for T and B cell receptors (148, 151, 188). Indeed, the best-performing configuration of TITAN used a TCR module that had been pretrained on a BindingDB database of 471,017 protein-ligand pairs (119), and several subsequent releases have exploited pre-training as a means of boosting performance (248, 258, 259). We estimate that there are 7.1 x 10^5 unique TCR sequences having known antigen specificity, the majority of which from the MIRA study **(Fig. 1-3)**, however the true magnitude of labelled and unlabelled data required for improved prediction of unseen epitopes is unknown.

Exploiting advances in structure prediction

Recent breakthroughs in protein structure prediction, for example with AlphaFold-3 (AF3) (260) and RFdiffusion (261), have transformed protein structural biology. These models have now extended from high-fidelity prediction of monomeric protein structures (149), to reconstruction of complex protein-protein and protein-ligand assemblies (187). There are clear opportunities to leverage these advantages for inference of TCR specificity, for example through use of predicted structure features as model inputs (148, 243). However, the reliability of structure-driven inference is limited by the scale of available training data: just 60 non-redundant TCR structures exist at the time of writing (169). Predictive performance of AF3 dips on hypervariable CDR loops, and domain-specific models have been shown to achieve similar or improved performance at faster runtimes (262). Whilst AF3 reported an improvement in antibody-specific structure prediction over its predecessor AF2-multimer, only 60% of predictions passed the CAPRI threshold score for protein-antibody docking (260, 263). These performance limitations may be attributed to the inherent conformational flexibility of the CDRs, variability between apo- and ligand-bound forms of the TCR, and/or a lack of TCR-specific co-evolutionary information: multiple sequence alignments being a key input to top-performing structure prediction models. Coordinated efforts to solve and release TCR-pMHC structures at high throughput may improve the fidelity of CDR loop modelling, and so of the predicted structures. The need is greatest for TCRs whose CDR loops adopt structures that do not fall within the “canonical forms” observed in many productive TCR-pMHC pairs, for weaker interactions, and for wildtype TCRs - all of which are underrepresented in current structural datasets.

Simple sequence-driven clustering outperforms more complex models

The performance of UCMs as predictors of TCR specificity was also found to be highly variable (**Chapter 2**). The five UCMs studied were able to group TCRs into antigen-specific clusters, even when provided with as few as ten representatives per epitope (**Fig. 2-6a**).

Performance was highly sensitive to the dataset, the amount of information provided, and the epitope being studied. Thus, whether TCR sequence clusters reflect recognition of unseen or known seen antigens, will depend on the condition under investigation, the available input and reference data, and the experimental methodology. Notably, a simple sequence distance model performed comparably to more complex algorithms (**Fig. 2-5**), yielded similar sequence logos for large epitope-specific clusters (**Fig. 2-8**), and was considerably more scalable (**Fig. 2-9**). A plausible strategy is, therefore, to combine two or more of the models - for example, a Hamming distance model with GIANA, ClusTCR, or tcrdist3 - to achieve robust cluster assignments. Practical considerations including ease of use must also be considered, for instance only GLIPH2 is currently available as both a web tool and a command-line executable.

The Orpheus pipeline performs well in varied contexts

Turning to TCR-directed methods, yeast pMHC screens yield a quantitative measure of receptor binding to over one billion synthetic peptides. By training a context-specific, augmented PWM on late-stage screening rounds, we retrieved many of the known human agonists of the a3a, DMF5, ngk2c, as4.3, Ob.A12, ObF3, c259, 1G4, and NYES1-3 TCRs within a range that could feasibly be functionally validated in a research setting, referencing a computational library of over eleven million self-peptides. These receptors exhibit binding affinities of 0.8-500μM, to peptides of length 8-14 residues, characterised using different methods and input data types, presented by class I and class II MHC molecules, and spanning varied disease contexts. This is, to our knowledge, the first TCR specificity predictor capable of deployment across such diverse conditions.

Balancing model robustness with adaptability for a given receptor is an important consideration in computational design. The Orpheus pipeline provides an off-the-shelf solution for prediction of TCR specificity given a pMHC screening campaign, and our best model retrieved the agonists of TCR-Ts, natural TCRs and NK receptors without the need for parameter tuning. This feature was designed to enable “blind” characterisation of receptors for which there are no known cognate pMHC, but the final configuration may not be optimal for all receptor-antigen pairs or methods of characterisation. For example, a baseline PWM outperformed the combined model in retrieval of NY-ESO-1 for NYES1-3 (**Fig. 5-16**), and differences emerged between model setups with class II restricted TCRs, particularly for weaker agonists (**Table 3-7**). Investigators may therefore wish to explore different configurations, or a combination of Orpheus predictions with more customisable techniques such as those deployed in (109).

We found that simple models outperformed more sophisticated machine learning approaches in retrieval of cognate agonists (**Fig. 3-5** and **3-6**), even when provided with hundreds of thousands of training instances (**Table 3-1**). This revealed a surprising limitation of machine learning models not only in the context of supervised prediction of TCR epitope specificity for

unseen epitopes (189), and in unsupervised clustering (189), but also in a narrower context with a bounty of positive and negative input data. This finding suggests that until the diversity, quality and volume of available TCR-epitope binding data is dramatically increased, specifically aiming to fill the gaps in chain representation, epitope source and HLA restriction identified in this study, incremental data-driven knowledge is more likely to yield benefits than more sophisticated modelling approaches.

Multi-modal data approaches boost performance of the predictor

Integrating affinity-driven selection data and functional activity for the a3a TCR improved the retrieval of agonists over selection data alone (**Fig. 3-8**). This agrees with a broader trend in the field of computational immunology, that emphasises systematic integration of sequence, phenotypic, spatial and functional information in characterising immune responses (255, 264). Combining receptor sequence with single-cell transcriptomic data can also enhance inference of T cell antigen specificity (141). By taking a graph theoretical approach, Schattgen and colleagues (194) reported an association between clonotype clustering with the cellular phenotypes derived from gene expression and surface marker expression.

Multi-modal approaches such as this pave the way for iterative cycles of clonotypically-expanded effector T cell discovery; cross-reactivity testing with yeast pMHC display; evaluation of synthetic peptides for functional agonism; prediction of specificity against wildtype and/or non-physiological reference libraries; functional validation of the predicted hits; and integration of binding and co-culture data. However, we find that not all functional data is created equally: merging of binding and activation or killing data did not improve agonist retrieval for c259 (**Fig. 5-13**). Furthermore, whilst combining diverse data sources may enable the discovery of strong binders or functional agonists, it risks biasing the search to a local minimum of the peptide sequence space, and so overlooking subdominant agonists. The importance of sequence coverage in predictive performance is most clearly exemplified in XSM experiments for NY-ESO-1 specific TCRs (**Chapter 5**), wherein dataset completeness appears to be more important than the method of characterisation for small libraries (**Fig. 5-11** and **5-12**).

This principle also applies to deeper yeast and phage display screens. For example, the strongest functional agonists of the a3a TCR are the MAGE-like peptides MAGE-A3 (EVDPIGHLY) and MAGE-A6 (EVDPIGHVY) (**Fig. 3-7**). Were an initial shallow library screen to return only MAGE-like peptides, and the search further directed towards similar peptides using stimulation data from co-culture experiments, one might overlook weaker agonists such as titin that share only 5/9 residues with MAGE-A3. It is therefore important that the initial screen be of sufficient depth to capture subdominant peptide clusters, and that as much of the sequence space of enriched mimotopes or predicted hits be functionally tested as

possible. Despite this risk of narrow search bias, the cyclic approach to discovery described above has now revealed over fifty agonists of the a3a TCR.

The clustering and visualisation approaches developed in **Chapter 4** (**Fig. 4-9**) may aid in this discovery process, and are now being actively deployed in the de-orphanisation of natural cancer-associated TCRs. Using the tools and techniques developed in this project, the Fernandes group have now successfully identified wildtype targets of an effector T cell clonally expanded in a patient with lung cancer: a manuscript that is currently under preparation together with the Dong group at the CAMS Oxford Institute. It should be noted that the network graph clusters produced are sensitive to the hyperparameters chosen when grouping peptides on the basis of embedding and sequence distance (**Fig. 4-10** to **4-12**) and can therefore be adapted to search for peptide relationships at different scales.

We propose novel metrics of cross-reactivity for therapeutic candidate prioritisation

Faced with a selection of clonally expanded, functionally relevant T cells implicated in a given disease, how can one prioritise candidate receptors for further development? Despite extensive study of TCR promiscuity, and the translational benefits of engineered TCRs having improved selectivity, to date there have been few attempts to define consistent metrics of cross-reactivity. In **Chapter 4**, we propose CR50, local and global positional entropy as measures of the diversity of enriched peptides bound by a TCR of interest. These formulae allow us to rank variants of the a3a TCR engineered to have improved selectivity towards MAGE-A3, and to prioritise v9 on the basis of validatory functional experiments. Further structural, functional and translational studies of this promising TCR-T candidate are ongoing.

Limitations and further work

This research has illustrated the potential of simple statistically informed predictors to retrieve the cognate agonists of varied receptors in diverse contexts. However, this project carries important limitations in both experimental and computational design.

Limitations of yeast display for inference of specificity

Yeast pMHC display is non-physiological

Single-chain pMHC trimers are non-physiological constructs that have been carefully engineered to permit their expression on yeast. Whilst electroporated yeast libraries are tested

for pMHC presentation using flow cytometry prior to screening (214), the extent to which yeast display of single-chain trimers truly mimics native presentation by human APCs is not fully understood. For example, it is likely that differences in the post-translational modification (PTM) of peptides shifts the suite of presented immunogenic peptides (the *immunopeptidome*) between yeast libraries and the native context (265). The ranked list of peptides returned may therefore miss clinically relevant modified self-peptides. Indeed, PTMs such as citrullination have been shown to play a role in influencing T cell responses in rheumatoid arthritis (266). Yeast display is also currently limited to the study of peptide antigens, and therefore overlooks lipid or metabolite antigen recognition by unconventional T cells via CD1 and MR1. The post-translationally modified peptide and non-peptide immunopeptidome may be probed using techniques such as mass spectrometry (254, 265). However, predictive modelling of modified peptides and non-peptide antigens would require a more sophisticated representation of the input than the amino acid sequences captured in the present study. Here, lessons from the world of machine learning-enabled small molecule discovery, such as the use of SMILES molecular representations (119, 267), could present a possible solution. The amino acid preferences of a given MHC allele at anchor positions are among the most significant assumptions made during yeast library design, and we find that substituting experimentally determined frequencies into the PWM can yield improved predictive performance **(Fig. 3-3**). Whilst mass spectrometry experiments are deepening our knowledge of the anchor preferences of many prevalent HLA types (20), coverage remains incomplete for less common class I and for class II alleles. Expansion of these datasets will likely translate into improved modelling of anchors and so to improved model performance.

A fundamental experimental limitation of TCR-directed screens such as yeast display is that they are currently restricted to studies of single, soluble receptors. Whilst this enables deep characterisation of TCRs that are more likely to be tractable for therapeutic use, it overlooks receptors that cannot be expressed in a soluble form. The studied TCRs are also characterised in a non-native context, being isolated on the surface of magnetic beads (**Chapter 3**). This is pertinent as recent studies comparing the format of natural TCRs and chimaeric antigen receptor (CAR)-T cells suggest an important impact of receptor format on antigen sensitivity (268, 269). Methodologies extending the breadth of pMHC complexes available in synthetic display screens, with display of both pMHC and TCR in a more native-like configuration, would likely return higher fidelity binding profiles.

Synthetic pMHC display is not appropriate for characterising weak interactions

Furthermore, there exists an upper affinity threshold for effective screening using yeast-pMHC libraries. Weak TCR-pMHC interactions such as those for the wildtype A3 TCR, which recognises MAGE-A3 with a KD of 500 μM (213), were poorly characterised: producing indistinct heat maps (**Fig. 4-9**, top row), logos similar to pre-selection analyses (**Fig. 4-9**, second row), and datasets that were weakly predictive of known antigen specificity (**Table 4-**

4). Accurate retrieval of the cognate agonists of the wildtype as4.3, nkg2c and NYES1-3 receptors suggests that synthetic peptide display may nonetheless be effectively applied to receptors having an affinity of ≤100 μM.

At the other end of the spectrum, the Orpheus pipeline retrieved the primary targets of TCRs such as c259 that engage pMHC with nanomolar affinities (**Table 5-1**). However, TCR-Ts may be engineered to bind their targets in the picomolar (270) or femtomolar (271) range. The extent to which binding at this level of affinity translates into an interpretable specificity profile with yeast pMHC display, is under active investigation. Early unpublished work suggests that tebentafusp, the first approved TCR-T engineered to recognise gp100 with picomolar affinity (272), may recognise not only both 9-mer and 10-mer peptides, but enriches distinct groups of peptides distal in sequence from its known target. Use of the Orpheus pipeline has also enabled both retrieval of gp100 at rank 1 (manuscript under preparation), and identification of other hitherto unreported wildtype agonists of tebentafusp, from the self-proteome. For stronger interactions such as the K-RAS specific TCR JDIa96b35 (212), the results of phage display analyses suggest that deeper sequencing will likely be required to uncover subdominant but functional agonists.

There is a need for more experimental repeats and more varied conditions

Whilst the scale of the peptides screened in a yeast pMHC screen is of order 10^9 (**Table 1-1**), limitations in sequencing depth mean that only a subset of these can be characterised in a single selection- typically in the range 10^6-10^7 unique peptides (**Table 3-1**). Whilst the predictive accuracy of the pipeline suggests that this coverage is sufficient to identify the strongest binders, it is possible that weaker clinically relevant agonists are being overlooked. Practical limitations in time and funding prevent the repetition of library preparation by an individual researcher, and thus, the extent to which a given library recapitulates the true binding profile of a TCR of interest under common conditions remains unknown. Indeed, variance is observed in pipeline performance for the DMF5 TCR when the model is provided with screens conducted by different groups (**Tables 3-4** and **3-5**). These differences might be due to the random nature of library generation, the relative depth of sequencing (**Table 3-1**), or differences in user technique and/or experimental procedure. For example, it is possible that different buffer conditions might alter the effective pH of the solution during selection, impacting the ionisation state of key residues involved in TCR-pMHC recognition, and so shifting the antigenic landscape of enriched peptides. This is particularly relevant in the design of TCR-Ts whose intended targets are presented in a hypoxic tumour microenvironment. Deeper sequencing might reveal novel clusters of agonist peptides, however the number of peptides that must be tested in a screen for complete characterisation of a given receptor remains unknown. Whilst inexpensive tools and techniques - including those explored for fastq2pep in **Chapter 4** - can be exploited to understand and reduce the impact of contaminants, sophisticated and cost-intensive measures will be required to eliminate them entirely. The problem of contamination

is compounded with the characterisation of more and more TCR-HLA pairs in a facility. A collaborative approach in which multiple repeats of the same experiment are conducted under varied conditions by different groups, including via yeast and phage display and XSM, could help to characterise the contribution of each of these factors to predictive performance.

Limitations in modelling approach

Consensus proteomes may not be representative

Gaps in the antigenic landscape that come as a result of the random nature of the yeast pMHC screen are exacerbated by the reliance of our pipeline on consensus human proteomes, which are unlikely to be fully representative of the spectrum of peptides exhibited in an individual. Specifically, they exclude the broad spectrum of sequences contributed by commensal and/or pathogenic organisms present in the body, which may provoke T cell responses either towards infected tissues, or through cross-recognition of similar self-peptides. For example, a conserved microbial peptide YEIH was among the strongest reported agonists of the as4.3 TCR reported in (110). Reference proteomes also fail to account for antigenic variation driven by hereditary differences and environmental exposure, and are limited to *canonical* peptides encoded by exons. However, recent studies have revealed a hidden world of non-canonical intronic but translated peptides, including those derived from endogenous retroviruses or alternative splicing, which alter the immunopeptidome between individuals and disease settings (251, 273). Possible solutions to this problem include broadening the reference proteome to include non-canonical peptides, and merging reference proteomes from multiple individuals and across prevalent microbes. However, the selection of which samples and organisms to include may be biased towards a specific research objective, and may consequently overlook functional agonists from other organisms. A screen of all possible peptides of a given length is likely a more comprehensive solution, albeit one that requires greater computational resource than available in a typical research setting, and that would result in dramatically increased model runtimes. A first step towards this approach is included in the small subsample of random peptides incorporated into c259 screens in **Chapter 5**, however investigation of the extent to which the top-ranked non-physiological peptides are capable of functional agonism is ongoing.

TCR-pMHC binding is not sufficient for functional activation

The degree to which antigen binding to MHC is sufficient to recapitulate functional immunogenicity remains contested (120). The potential impact of this omission is to overlook features that are important for the separation of binders from functional agonists. As observed for the retrieval of titin as a predicted binder for a3v9 (**Table 4-4**), one limitation of the Orpheus pipeline might be that - whilst amino acid sequence representations enable it to discern broad patterns inherent in complementary peptides - they may not define the granular rules within

clusters which separate agonists from non-agonists. Even if the sequence space is extensively sampled in the library, the model is only learning from the positive class of enriched peptides. There is, therefore, the potential for false positives, which resemble true agonists but may lack a key interaction influencing the kinetics of TCR engagement. The manner in which peptides are scored using a PWM (**Eq. 3-1b**) also assumes that each residue is independent, whereas complementary motifs are a product of the wider sequence context that is not picked up within this framework.

Whilst reasonable performance has been observed for the Orpheus pipeline in diverse contexts, the severe consequences of off-target binding imply that display screens be used as a complementary approach to therapeutic de-risking, alongside functional assays, rather than as a standalone solution. Sub-selecting the reference proteome on the basis of predicted MHC binding improves agonist ranking (**Table 3-9**) but incurs an additional risk of overlooking peptides that are presented when expressed at high levels in a given tissue. A qualitative review of model performance between display and XSM methods suggests that Orpheus predictions from datasets with fewer than one thousand unique hits should be interpreted with caution, although this threshold will be necessarily TCR-dependent. Opportunities to improve model performance whilst preserving recall include re-ranking predicted hits on both the PWM score, the predicted probability of MHC binding, TCR-independent immunogenicity (274-277), peptide abundance (278), or some combination of the above.

Further work

In the present study, we have established proof-of-concept of the capacity of the Orpheus pipeline to de-orphanise therapeutically relevant natural and engineered TCRs, when provided with deep yeast pMHC screens. This software will be made freely available at www.github.com/hudsondan/Orpheus. Application of the pipeline to de-orphanise therapeutic TCRs, including Tebentafusp, natural TCRs implicated in cancer and autoimmunity, and killer cell immunoglobulin-like receptors (KIRs), is ongoing between the Fernandes laboratory and its collaborators. Screening of other therapeutic modalities, including TCR mimetic antibodies, pMHC minibinders, CAR-T cells and bispecifics, could provide novel insights into cross-reactivity beyond the TCR.

Finally, rich datasets such as those provided by synthetic yeast pMHC display may serve as a bridge between TCR-directed and antigen-directed methods, and so provide a possible solution to more generalisable inference. Statistical analyses of the TCR repertoires of large HLA-typed cohorts, and structural studies of TCRs having known class I MHC restriction, support the tractability of predicting MHC restriction from TCR sequence alone (279, 280). This suggests a possible architecture wherein the MHC restriction of a given TCR is inferred from its

sequence, and linked to the most likely candidates from the cognate immunopeptidome via MHC-specific predictors such as those developed in this work.

Bibliography

1. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell*. 2020; **181**(2): 281-92.e6.
2. Glanzmann E, Riniker P. Essential lymphocytophthisis; new clinical aspect of infant pathology. *Ann Paediatr*. 1950; **175**(1-2):1-32.
3. Meuer SC, Fitzgerald KA, Hussey RE, Hodgdon JC, Schlossman SF, Reinherz EL. Clonotypic structures involved in antigen-specific human T cell function. Relationship to the T3 molecular complex. *J Exp Med*. 1983; **157**(2): 705-19.
4. Pardoll DM, Fowlkes BJ, Bluestone JA, Kruisbeek A, Maloy WL, Coligan JE, et al. Differential expression of two distinct T-cell receptors during thymocyte development. *Nature*. 1987; **326**(6108): 79-81.
5. Sušac L, Vuong MT, Thomas C, von Bülow S, O'Brien-Ball C, Santos AM, et al. Structure of a fully assembled tumor-specific T cell receptor ligated by pMHC. *Cell*. 2022; **185**(17): 3201-13.e19.
6. Punt JA, Roberts JL, Kearse KP, Singer A. Stoichiometry of the T cell antigen receptor (TCR) complex: each TCR/CD3 complex contains one TCR alpha, one TCR beta, and two CD3 epsilon chains. *J Exp Med*. 1994; **180**(2): 587-93.
7. Schodin BA, Tsomides TJ, Kranz DM. Correlation Between the Number of T Cell Receptors Required for T Cell Activation and TCR–Ligand Affinity. *Immunity*. 1996; **5**(2): 137-46.
8. Groh V, Porcelli S, Fabbi M, Lanier LL, Picker LJ, Anderson T, et al. Human lymphocytes bearing T cell receptor gamma/delta are phenotypically diverse and evenly distributed throughout the lymphoid system. *J Exp Med*. 1989; **169**(4): 1277-94.
9. Hviid L, Akanmori BD, Loizon S, Kurtzhals JAL, Ricke CH, Lim A, et al. High frequency of circulating γδ T cells with dominance of the Vδ1 subset in a healthy population. *Int Immunol.* 2000; **12**(6): 797-805.
10. De Libero G. Tissue distribution, antigen specificity and effector functions of γδ T cells in human diseases. *Springer Seminars in Immunopathology*. 2000; **22**(3): 219-38.
11. Hedrick SM, Cohen DI, Nielsen EA, Davis MM. Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. *Nature*. 1984; **308**(5955): 149-53.
12. Yanagi Y, Yoshikai Y, Leggett K, Clark SP, Aleksander I, Mak TW. A human T cell-specific cDNA clone encodes a protein having extensive homology to immunoglobulin chains. *Nature*. 1984; **308**(5955): 145-9.
13. Sewell AK. Why must T cells be cross-reactive? *Nat Rev Immunol*. 2012; **12**: 669-77.
14. Schatz DG, Oettinger MA, Blatimore D. The V(D)J recombination activating gene, RAG-1. *Cell*. 1989; **59**(6): 1035-48.
15. Raybould MIJ, Greenshields-Watson A, Agarwal P, Aguilar-Sanjuan B, Olsen TH, Turnbull OM, et al. The Observed T Cell Receptor Space database enables paired-chain repertoire mining, coherence analysis, and language modeling. *Cell Rep.* 2024; **43**(9).
16. Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P, et al. A direct estimate of the human αβ T cell receptor diversity. Science. 1999; **286**: 958-61.
17. Nguyen AT, Szeto C, Gras S. The pockets guide to HLA class I molecules. *Biochem Soc Trans*. 2021; **49**(5): 2319-31.
18. Schrödinger LLC, The PyMOL Molecular Graphics System, Version 2.0. 2022.
19. Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT, the international ImMunoGeneTics information system 25 years on. *Nucleic Acids Res*. 2015; **43**(D1): D413-D22.
20. Tadros DM, Eggenschwiler S, Racle J, Gfeller D. The MHC Motif Atlas: a database of MHC binding specificities and ligands. *Nucleic Acids Res*. 2023; **51**(D1): D428-D37.
21. Dausset J. The Major Histocompatibility Complex in Man. *Science*. 1981; **213**(4515): 1469-74.

22. De Libero G, Chancellor A, Mori L. Antigen specificities and functional properties of MR1-restricted T cells. *Mol Immunol.* 2021; **130**:148-53.
23. de Jong A, Ogg G. CD1a function in human skin disease. *Mol Immunol.* 2021; **130**: 14-9.
24. Liu X, Dai S, Crawford F, Frugé R, Marrack P, Kappler J. Alternate interactions define the binding of peptides to the MHC molecule IAb. *PNAS*. 2002; **99**(13): 8820-5.
25. Zareie P, Szeto C, Farenc C, Gunasinghe SD, Kolawole EM, Nguyen A, et al. Canonical T cell receptor docking on peptide–MHC is essential for T cell signaling. *Science*. 2021; **372**(6546): eabe9124.
26. Raybould MIJ, Nissley DA, Kumar S, Deane CM. Computationally profiling peptide:MHC recognition by T-cell receptors and T-cell receptor-mimetic antibodies. *Front Immunol.* 2023; **13**.
27. Gras S, Chadderton J, Del Campo CM, Farenc C, Wiede F, Josephs TM, et al. Reversed T Cell Receptor Docking on a Major Histocompatibility Class I Complex Limits Involvement in the Immune Response. *Immunity*. 2016; **45**(4): 749-60.
28. Shinzawa M, Moseman EA, Gossa S, Mano Y, Bhattacharya A, Guinter T, et al. Reversal of the T cell immune system reveals the molecular basis for T cell lineage fate determination in the thymus. *Nat Immunol.* 2022; **23**(5): 731-42.
29. Garcia KC, Scott CA, Brunmark A, Carbone FR, Peterson PA, Wilson IA, et al. CD8 enhances formation of stable T-cell receptor/MHC class I molecule complexes. *Nature*. 1996; **384**(6609): 577-81.
30. Irvine DJ, Purbhoo MA, Krogsgaard M, Davis MM. Direct observation of ligand recognition by T cells. *Nature*. 2002; **419**(6909): 845-9.
31. Laugel B, van den Berg HA, Gostick E, Cole DK, Wooldridge L, Boulter J, et al. Different T Cell Receptor Affinity Thresholds and CD8 Coreceptor Dependence Govern Cytotoxic T Lymphocyte Activation and Tetramer Binding Properties. *J Biol Chem*. 2007; **282**(33): 23799-810.
32. Kern PS, Teng M-k, Smolyar A, Liu J-h, Liu J, Hussey RE, et al. Structural Basis of CD8 Coreceptor Function Revealed by Crystallographic Analysis of a Murine CD8αα Ectodomain Fragment in Complex with H-2Kb. *Immunity*. 1998; **9**(4): 519-30.
33. Lichtenheld MG, Olsen KJ, Lu P, Lowrey DM, Hameed A, Hengartner H, et al. Structure and function of human perforin. *Nature*. 1988; **335**(6189): 448-51.
34. Rouvier E, Luciani MF, Golstein P. Fas involvement in Ca(2+)-independent T cell-mediated cytotoxicity. *J Exp Med*. 1993; **177**(1): 195-200.
35. Brehm MA, Daniels KA, Welsh RM. Rapid Production of TNF-α following TCR Engagement of Naive CD8 T Cells1. *J Immunol*. 2005; **175**(8): 5043-9.
36. Nathan CF, Murray HW, Wiebe ME, Rubin BY. Identification of interferon-gamma as the lymphokine that activates human macrophage oxidative metabolism and antimicrobial activity. *J Exp Med*. 1983; **158**(3): 670-89.
37. Addison EG, North J, Bakhsh I, Marden C, Haq S, Al-Sarraj S, et al. Ligation of CD8α on human natural killer cells prevents activation-induced apoptosis and enhances cytolytic activity. *Immunology*. 2005; **116**(3): 354-61.
38. Walker LJ, Kang Y-H, Smith MO, Tharmalingham H, Ramamurthy N, Fleming VM, et al. Human MAIT and CD8αα cells develop from a pool of type-17 precommitted CD8+ T cells. *Blood*. 2012; **119**(2): 422-33.
39. Kjer-Nielsen L, Patel O, Corbett AJ, Le Nours J, Meehan B, Liu L, et al. MR1 presents microbial vitamin B metabolites to MAIT cells. *Nature*. 2012; **491**(7426): 717-23.
40. Jönsson P, Southcombe JH, Santos AM, Huo J, Fernandes RA, McColl J, et al. Remarkably low affinity of CD4/peptide-major histocompatibility complex class II protein interactions. *PNAS*. 2016; **113**(20): 5682-7.
41. Tuzlak S, Dejean AS, Iannacone M, Quintana FJ, Waisman A, Ginhoux F, et al. Repositioning TH cell polarization from single cytokines to complex help. *Nat Immunol*. 2021; **22**(10): 1210-7.

42. Kondo M, Weissman IL, Akashi K. Identification of Clonogenic Common Lymphoid Progenitors in Mouse Bone Marrow. *Cell*. 1997; **91**(5):661-72.
43. Pui JC, Allman D, Xu L, DeRocco S, Karnell FG, Bakkour S, et al. Notch1 expression in early lymphopoiesis influences B versus T lineage determination. *Immunity*. 1999; **11**(3): 299-308.
44. Lind EF, Prockop SE, Porritt HE, Petrie HT. Mapping Precursor Movement through the Postnatal Thymus Reveals Specific Microenvironments Supporting Defined Stages of Early Lymphoid Development. *J Exp Med*. 2001; **194**(2): 127-34.
45. Raulet DH, Garman RD, Saito H, Tonegawa S. Developmental regulation of T-cell receptor gene expression. *Nature*. 1985; **314**(6006):103-7.
46. Miller JFAP. Analysis of the Thymus Influence in Leukaemogenesis. *Nature*. 1961; **191**(4785):248-9.
47. Anderson G, Jenkinson EJ, Moore NC, Owen JJT. MHC class II-positive epithelium and mesenchyme cells are both required for T-cell development in the thymus. *Nature*. 1993; **362**(6415): 70-3.
48. Starr TK, Jameson SC, Hogquist KA. Positive and Negative Selection of T Cells. *Ann Rev Immunol.* 2003;**21**:139-76.
49. Medawar PB. The behaviour and fate of skin autografts and skin homografts in rabbits: A report to the War Wounds Committee of the Medical Research Council. *J Anat*. 1944; **78**(Pt 5): 176-99.
50. Vos O, Davids JA, Weyzen WW, Van Bekkum DW. Evidence for the cellular hypothesis in radiation protection by bone marrow cells. Acta physiologica et pharmacologica Neerlandica. 1956; **4**(4): 482-6.
51. Anderson MS, Venanzi ES, Klein L, Chen Z, Berzins SP, Turley SJ, et al. Projection of an Immunological Self Shadow Within the Thymus by the Aire Protein. *Science*. 2002; **298**(5597): 1395-401.
52. Kappler JW, Roehm N, Marrack P. T cell tolerance by clonal elimination in the thymus. *Cell*. 1987; **49**(2): 273-80.
53. Redmond WL, Sherman LA. Peripheral tolerance of CD8 T lymphocytes. *Immunity*. 2005; **22**(3): 275-84.
54. Uribe-Querol E, Rosales C. Phagocytosis: Our Current Understanding of a Universal Biological Process. *Front Immunol.* 2020;**11**:1066.
55. Clark GJ, Silveira PA, Hogarth PM, Hart DNJ. The cell surface phenotype of human dendritic cells. *Sem Cell Dev Bio*. 2019; **86**:3-14.
56. Sckisel GD, Bouchlaka MN, Monjazeb AM, Crittenden M, Curti BD, Wilkins DEC, et al. Out-of-Sequence Signal 3 Paralyzes Primary CD4(+) T-Cell-Dependent Immunity. *Immunity*. 2015; **43**(2): 240-50.
57. Straus DB, Weiss A. Genetic evidence for the involvement of the lck tyrosine kinase in signal transduction through the T cell antigen receptor. *Cell*. 1992; **70**(4): 585-93.
58. Brameshuber M, Kellner F, Rossboth BK, Ta H, Alge K, Sevcsik E, et al. Monomeric TCRs drive T cell antigen recognition. *Nat Immunol*. 2018; **19**(5): 487-96.
59. Chan AC, Irving BA, Fraser JD, Weiss A. The zeta chain is associated with a tyrosine kinase and upon T-cell antigen receptor stimulation associates with ZAP-70, a 70-kDa tyrosine phosphoprotein. *PNAS*. 1991; **88**(20): 9166-70.
60. Mariuzza RA, Agnihotri P, Orban J. The structural basis of T-cell receptor (TCR) activation: An enduring enigma. *J Biol Chem*. 2020; **295**(4): 914-25.
61. Davis SJ, van der Merwe PA. The kinetic-segregation model: TCR triggering and beyond. *Nat Immunol.* 2006; **7**(8): 803-9.
62. Jenkins MK, Ashwell JD, Schwartz RH. Allogeneic non-T spleen cells restore the responsiveness of normal T cell clones stimulated with antigen and chemically modified antigen-presenting cells. *J Immunol.* 1988; **140**(10): 3324-30.

63. Cossarizza A, Ortolani C, Paganelli R, Barbieri D, Monti D, Sansoni P, et al. CD45 isoforms expression on CD4+ and CD8+ T cells throughout life, from newborns to centenarians: implications for T cell memory. *Mech Ageing Dev.* 1996; **86**(3):173-95.
64. Hommel M, Kyewski B. Dynamic changes during the immune response in T cell-antigen-presenting cell clusters isolated from lymph nodes. *J Exp Med.* 2003; **197**(3): 269-80.
65. Hudson D, Fernandes RA, Basham M, Ogg G, Koohy H. Can we predict T cell specificity with digital biology and machine learning? *Nat Rev Immunol*. 2023.
66. Schaap-Johansen AL, Vujović M, Borch A, Hadrup SR, Marcatili P, . T Cell Epitope Prediction and Its Application to Immunotherapy. *Frontiers Media SA*; 2021. p. 712488.
67. Waldman AD, Fritz JM, Lenardo MJ. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nat Rev Immunol.* 2020; **20**(11): 651-68.
68. Linette GP, Stadtmauer EA, Maus MV, Rapoport AP, Levine BL, Emery L, et al. Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood*. 2013; **122**(6): 863-71.
69. Foldvari Z, Knetter C, Yang W, Gjerdingen TJ, Bollineni RC, Tran TT, et al. A systematic safety pipeline for selection of T-cell receptors to enter clinical use. *NPJ Vaccines*. 2023; **8**(1):126.
70. Hernando B, Dietzen M, Parra G, Gil-Barrachina M, Pitarch G, Mahiques L, et al. The effect of age on the acquisition and selection of cancer driver mutations in sun-exposed normal skin. *Ann Oncol.* 2021; **32**(3): 412-21.
71. Sesma A, Pardo J, Cruellas M, Gálvez EM, Gascón M, Isla D, et al. From Tumor Mutational Burden to Blood T Cell Receptor: Looking for the Best Predictive Biomarker in Lung Cancer Treated with Immunotherapy. *Cancers* (Basel). 2020; **12**(10).
72. Scott AC, Dündar F, Zumbo P, Chandran SS, Klebanoff CA, Shakiba M, et al. TOX is a critical regulator of tumour-specific T cell differentiation. *Nature*. 2019; **571**(7764): 270-4.
73. Yost KE, Satpathy AT, Wells DK, Qi Y, Wang C, Kageyama R, et al. Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nat Med.* 2019; **25**(8):1251-9.
74. Wherry EJ, Kurachi M. Molecular and cellular insights into T cell exhaustion. *Nat Rev Immunol*. 2015; **15**(8): 486-99.
75. Daniel B, Yost KE, Hsiung S, Sandor K, Xia Y, Qi Y, et al. Divergent clonal differentiation trajectories of T cell exhaustion. *Nat Immunol*. 2022; **23**(11): 1614-27.
76. Shakiba M, Zumbo P, Espinosa-Carrasco G, Menocal L, Dündar F, Carson SE, et al. TCR signal strength defines distinct mechanisms of T cell dysfunction and cancer evasion. *J Exp Med*. 2021; **219**(2): e20201966.
77. Dan JM, Mateus J, Kato Y, Hastie KM, Yu ED, Faliti CE, et al. Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. *Science*. 2021; **371**(6529): eabf4063.
78. Swanson PA, Padilla M, Hoyland W, McGlinchey K, Fields PA, Bibi S, et al. AZD1222/ChAdOx1 nCoV-19 vaccination induces a polyfunctional spike protein–specific TH1 response with a diverse TCR repertoire. *Sci Trans Med*. **13**(620): eabj7211.
79. Bjornevik K, Cortese M, Healy BC, Kuhle J, Mina MJ, Leng Y, et al. Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. *Science*. 2022; **375**(6578): 296-301.
80. Joglekar AV, Li G. T cell antigen discovery. *Nat Methods*. 2021; **18**(8): 873-80.
81. Hadrup SR, Bakker AH, Shu CJ, Andersen RS, van Veluw J, Hombrink P, et al. Parallel detection of antigen-specific T-cell responses by multidimensional encoding of MHC multimers. *Nat Methods*. 2009; **6**(7): 520-6.
82. Newell Evan W, Sigal N, Bendall Sean C, Nolan GP, Davis MM. Cytometry by Time-of-Flight Shows Combinatorial Cytokine Expression and Virus-Specific Cell Niches within a Continuum of CD8+ T Cell Phenotypes. *Immunity*. 2012; **36**(1): 142-52.
83. Bentzen AK, Marquard AM, Lyngaa R, Saini SK, Ramskov S, Donia M, et al. Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes. *Nat Biotechnol*. 2016; **34**(10): 1037-45.

84. Altman JD, Moss PAH, Goulder PJR, Barouch DH, McHeyzer-Williams MG, Bell JI, et al. Phenotypic Analysis of Antigen-Specific T Lymphocytes. *Science*. 1996; **274**(5284): 94-6.
85. Yao Y, Wyrozżemski Ł, Lundin KEA, Sandve GK, Qiao S-W. Differential expression profile of gluten-specific T cells identified by single-cell RNA-seq. *PLOS ONE*. 2021; **16**(10): e0258029.
86. Kurtulus S, Hildeman D. Assessment of CD4(+) and CD8 (+) T cell responses using MHC class I and II tetramers. *Methods Mol Biol*. 2013; **979**: 71-9.
87. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, et al. Identifying specificity groups in the T cell receptor repertoire. *Nature*. 2017; **547**(7661): 94-8.
88. Bosselut R. T cell antigen recognition: Evolution-driven affinities. *PNAS*. 2019; **116**(44): 21969-71.
89. Emerson RO, Dewitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet*. 2017; **49**(5): 659-65.
90. Wang X, He Y, Zhang Q, Ren X, Zhang Z. Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2. *Gen Prot Bio.* 2021; **19**(2): 253-66.
91. Zhang W, Hawkins PG, He J, Gupta NT, Liu J, Choonoo G, et al. A framework for highly multiplexed dextramer mapping and prediction of T cell receptor sequences to antigen specificity. *Sci Adv.* 2021; **7**(20): eabf5835.
92. Dolton G, Zervoudi E, Rius C, Wall A, Thomas HL, Fuller A, et al. Optimized Peptide-MHC Multimer Protocols for Detection and Isolation of Autoimmune T-Cells. *Front Immunol.* 2018; **9**: 1378.
93. Meysman P, Barton J, Bravi B, Cohen-Lavi L, Karnaukhov V, Lilleskov E, et al. Benchmarking solutions to the T-cell receptor epitope prediction problem: IMMREP22 workshop report. *ImmunoInformatics*. 2023; **9**: 100024.
94. Dobson CS, Reich AN, Gaglione S, Smith BE, Kim EJ, Dong J, et al. Antigen identification and high-throughput interaction mapping by reprogramming viral entry. *Nat Methods.* 2022; **19**(4): 449-60.
95. Guo XJ, Elledge SJ. V-CARMA: A tool for the detection and modification of antigen-specific T cells. *PNAS*. 2022; **119**(4).
96. Joglekar AV, Leonard MT, Jeppson JD, Swift M, Li G, Wong S, et al. T cell antigen discovery via signaling and antigen-presenting bifunctional receptors. *Nat Methods*. 2019; **16**(2): 191-8.
97. Tomasz K, Mohammad HD, Charlotte IW, Nouran SA, Zachary CH, Kai WW, et al. T-Scan: A Genome-wide Method for the Systematic Discovery of T Cell Epitopes. *Cell*. 2019; **178**(4): 1016-28.e13.
98. Pan X, Huang L-C, Dong T, Peng Y, Cerundolo V, McGowan S, et al. Combinatorial HLA-peptide bead libraries for high throughput identification of CD8+ T cell specificity. *J Immunol Methods*. 2014; **403**(1): 72-8.
99. Li G, Bethune MT, Wong S, Joglekar AV, Leonard MT, Wang JK, et al. T cell antigen discovery via trogocytosis. *Nat Methods*. 2019; **16**(2): 183-90.
100. Bagaev DV, Vroomans RMA, Samir J, Stervbo U, Rius C, Dolton G, et al. VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res*. 2020; **48**(D1): D1057-D62.
101. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res*. 2019; **47**: D339-D43.
102. Nolan S, Vignali M, Kaplan IM, Biotechnologies A, Svejnoha E, Craft T, et al. A large-scale database of T-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. 2020. Preprint on ArXiv.
103. Tickotsky N, Sagiv T, Prilusky J, Shifrut E, Friedman N, . McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics*. 2017; **33**(18): 2924-9.
104. Brophy SE, Holler PD, Kranz DM, . A yeast display system for engineering functional peptide-MHC complexes. *J Immunol Methods*. 2003; **272**: 235-46.

105. Birnbaum ME, Mendoza JL, Sethi DK, Dong S, Glanville J, Dobbins J, et al. Deconstructing the peptide-MHC specificity of T cell recognition. *Cell*. 2014; **157**: 1073-87.
106. Crawford F, Jordan KR, Stadinski B, Wang Y, Huseby E, Marrack P, et al. Use of baculovirus MHC/peptide display libraries to characterize T-cell receptor ligands. *Immunol Rev.* 2006; **210**: 156-70.
107. Coles CH, Mulvaney RM, Malla S, Walker A, Smith KJ, Lloyd A, et al. TCRs with Distinct Specificity Profiles Use Different Binding Modes to Engage an Identical Peptide-HLA Complex. *J Immunol*. 2020; **204**(7): 1943-53.
108. Zhao X, Kolawole EM, Chan W, Feng Y, Yang X, Gee MH, et al. Tuning T cell receptor sensitivity through catch bond engineering. *Science*. 2022; **376**.
109. Gee MH, Han A, Lofgren SM, Beausang JF, Mendoza JL, Birnbaum ME, et al. Antigen Identification for Orphan T Cell Receptors Expressed on Tumor-Infiltrating Lymphocytes. *Cell*. 2018; **172**(3): 549-63.e16.
110. Yang X, Garner LI, Zvyagin IV, Paley MA, Komech EA, Jude KM, et al. Autoimmunity-associated T cell receptors recognize HLA-B*27-bound peptides. *Nature*. 2022; **612**(7941): 771-7.
111. Rosenberg AM, Ayres CM, Medina-Cucurella AV, Whitehead TA, Baker BM. Enhanced T cell receptor specificity through framework engineering. *Front Immunol*. 2024; **15**.
112. Huisman BD, Guan N, Rückert T, Garner L, Singh NK, McMichael AJ, et al. High-throughput characterization of HLA-E-presented CD94/NKG2x ligands reveals peptides which modulate NK cell activation. *Nat Comms*. 2023; **14**(1).
113. Yang A, Jude KM, Lai B, Minot M, Kocyla AM, Glassman CR, et al. Deploying synthetic coevolution and machine learning to engineer protein-protein interactions. *Science*. 2023; **381**(6656): eadh1720.
114. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res*. 2023; **51**(D1): D523-D31.
115. Kula T, Dezfulian MH, Wang CI, Abdelfattah NS, Hartman ZC, Wucherpfennig KW, et al. T-Scan: A Genome-wide Method for the Systematic Discovery of T Cell Epitopes. *Cell*. 2019; **178**(4): 1016-28.e13.
116. Moris P, Postovskaya A, Gielis S, De Neuter N, Bittremieux W, Ogunjimi B, et al. Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. *Brief Bioinform.* 2021; **22**(4): 1-12.
117. Huang H, Wang C, Rubelt F, Scriba TJ, Davis MM, . Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nat Biotechnol*. 2020; **38**: 1194-202.
118. Mayer-Blackwell K, Schattgen S, Cohen-Lavi L, Crawford JC, Souquette A, Gaevert JA, et al. TCR meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs. *eLife*. 2021; **10**: e68605.
119. Weber A, Born J, Rodriguez Martínez M, et al. TITAN: T Cell Receptor Specificity Prediction with Bimodal Attention Networks. *Bioinformatics*. 2021; **37**(S1) :I237-I44.
120. Lee CH, Antanaviciute A, Buckley PR, Simmons A, Koohy H. To what extent does MHC binding translate to immunogenicity in humans? *Immunoinformatics*. 2021; **3-4**:100006.
121. Buckley PR, Lee CH, Ma R, Woodhouse I, Woo J, Tsvetkov VO, et al. Evaluating performance of existing computational models in predicting CD8+ T cell pathogenic epitopes and cancer neoantigens. *Brief Bioinform*. 2022; **23**(3).
122. Mösch A, Raffegerst S, Weis M, Schendel DJ, Frishman D, et al. Machine Learning for Cancer Immunotherapies Based on Epitope Recognition by T Cell Receptors. *Front Genet.* 2019; **10**: p. 1141.
123. Wells DK, van Buuren MM, Dang KK, Hubbard-Lucey VM, Sheehan KCF, Campbell KM, et al. Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. *Cell*. 2020; **183**(3): 818-34.e13.

124. Valkiers S, de Vrij N, Gielis S, Verbandt S, Ogunjimi B, Laukens K, et al. Recent advances in T-cell receptor repertoire analysis: Bridging the gap with multimodal single-cell RNA sequencing. *Immunoinformatics*. 2022; **5**:100009.
125. Lee CH, Salio M, Napolitani G, Ogg G, Simmons A, Koohy H, et al. Predicting Cross-Reactivity and Antigen Specificity of T Cell Receptors. *Front Immunol*. 2020; **11**: 2498.
126. Vujovic M, Degn KF, Marin FI, Schaap-Johansen AL, Chain B, Andresen TL, et al. T cell receptor sequence clustering and antigen specificity. *Comp Struct Biol J*. 2020; **18**: 2166-73.
127. Katayama Y, Yokota R, Akiyama T, Kobayashi TJ. Machine Learning Approaches to TCR Repertoire Analysis. *Front Immunol*. 2022; **13**: 858057.
128. Pearson K, . On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science.* 1901; **2**: 559-70.
129. McInnes LaH, John and Saul, Nathaniel and Grossberger, Lukas. UMAP: Uniform Manifold Approximation and Projection. *J Open Source Software*. 2018; **3**(29): 861.
130. Cai M, Bang S, Zhang P, Lee H. ATM-TCR: TCR-Epitope Binding Affinity Prediction Using a Multi-Head Self-Attention Model. *Front Immunol*. 2022; **13**: 893247.
131. Pavlović M, Scheffer L, Motwani K, Kanduri C, Kompova R, Vazov N, et al. The immuneML ecosystem for machine learning analysis of adaptive immune receptor repertoires. *Nat Mach Intell.* 2021; **3**(11): 936-44.
132. Heikkilä N, Vanhanen R, Yohannes DA, Kleino I, Mattila IP, Saramäki J, et al. Human thymic T cell repertoire is imprinted with strong convergence to shared sequences. *Mol Immunol.* 2020; **127**: 112-23.
133. Montemurro A, Schuster V, Povlsen HR, Bentzen AK, Jurtz V, Chronister WD, et al. NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. *Nat Comms Bio.* 2021; **4**(1).
134. 10X Genomics. A new way of exploring immunity - linking highly multiplexed antigen recognition to immune repertoire and phenotype 2019 [Available from: https://pages.10xgenomics.com/rs/446-PBO-704/images/10x_AN047_IP_A_New_Way_of_Exploring_Immunity_Digital.pdf.]
135. Ehrlich R, Kamga L, Gil A, Luzuriaga K, Selin LK, Ghersi D. SwarmTCR: a computational approach to predict the specificity of T cell receptors. *BMC Bioinformatics*. 2021; **22**(1): 422.
136. Dean J, Emerson RO, Vignali M, Sherwood AM, Rieder MJ, Carlson CS, et al. Annotation of pseudogenic gene segments by massively parallel sequencing of rearranged lymphocyte receptor loci. *Gen Med*. 2015; **7**(1): 123.
137. Luu AM, Leistico JR, Miller T, Kim S, Song JS. Predicting TCR-Epitope Binding Specificity Using Deep Metric Learning and Multimodal Learning. *Genes (Basel).* 2021; **12**(4).
138. Zhang W, Wang L, Liu K, Wei X, Yang K, Du W, et al. PIRD: Pan Immune Repertoire Database. *Bioinformatics*. 2020; **36**(3): 897-903.
139. Jokinen E, Huuhtanen J, Mustjoki S, Heinonen M, Lähdesmäki H. Predicting recognition between T cell receptors and epitopes with TCRGP. *PLoS Comp Bio*. 2021; **17**(3): e1008814.
140. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*. 2017; **547**(7661): 89-93.
141. Fischer DS, Wu Y, Schubert B, Theis FJ. Predicting antigen specificity of single T cells based on TCR CDR3 regions. *Mol Syst Biol*. 2020; **16**(8): e9416.
142. Tong Y, Wang J, Zheng T, Zhang X, Xiao X, Zhu X, et al. SETE: Sequence-based Ensemble learning approach for TCR Epitope binding prediction. *Comput Biol Chem*. 2020; **87**:107281.
143. Valkiers S, Van Houcke M, Laukens K, Meysman P. ClusTCR: a python interface for rapid clustering of large sets of CDR3 sequences with unknown antigen specificity. *Bioinformatics*. 2021; **37**(24): 4865-7.
144. Snyder TM, Gittelman RM, Klinger M, May DH, Osborne EJ, Taniguchi R, et al. Magnitude and dynamics of the T-cell response to SARS-CoV-2 infection at both individual and population levels. *Preprint on medRxiv*; 2020.

145. Zhang H, Zhan X, Li B, . GIANA allows computationally-efficient TCR clustering and multi-disease repertoire classification by isometric transformation. *Nat Comms*. 2021; **12**(4699): 1-11.
146. Zhang SQ, Ma KY, Schonnesen AA, Zhang M, He C, Sun E, et al. High-throughput determination of the antigen specificities of T cell receptors in single cells. *Nat Biotechnol.* 2018; **36**: 1156-1159.
147. Zhang H, Liu L, Zhang J, Chen J, Ye J, Shukla S, et al. Investigation of Antigen-Specific T-Cell Receptor Clusters in Human Cancers. *Clin Canc Res*. 2020; **26**(6): 1359-71.
148. Bradley P. Structure-based prediction of T cell receptor:peptide-MHC interactions. *eLife*. 2023;**12**.
149. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021; **596**(7873): 583-9.
150. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000; **28**(1): 235-42.
151. Wu K, Yost KE, Daniel B, Belk JA, Xia Y, Egawa T, et al. TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-xbinding analyses. Preprint at *bioRxiv.* 2021.
152. Chen SY, Yue T, Lei Q, Guo AY. TCRdb: a comprehensive database for T-cell receptor sequences with powerful search function. *Nucleic Acids Res*. 2021; **49**(D1): D468-d74.
153. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res*. 2016; **44**(D1): D1045-53.
154. Dines JN, Manley TJ, Svejnoha E, Simmons HM, Taniguchi R, Klinger M, et al. The ImmuneRACE Study: A Prospective Multicohort Study of Immune Response Action to COVID-19 Events with the ImmuneCODE™ Open Access Database. *Preprint at medRxiv*. 2020.
155. Lu T, Zhang Z, Zhu J, Wang Y, Jiang P, Xiao X, et al. Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nat Mach Intell*. 2021; **3**(10): 864-75.
156. Chen G, Yang X, Ko A, Sun X, Gao M, Zhang Y, et al. Sequence and Structural Analyses Reveal Distinct and Highly Diverse Human CD8(+) TCR Repertoires to Immunodominant Viral Antigens. *Cell Rep*. 2017; **19**(3): 569-83.
157. Huth A, Liang X, Krebs S, Blum H, Moosmann A. Antigen-Specific TCR Signatures of Cytomegalovirus Infection. *J Immunol*. 2019; **202**(3): 979-90.
158. Springer I, Besser H, Tickotsky-Moskovitz N, Dvorkin S, Louzoun Y. Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs. *Front Immunol.* 2020; **11**: 1803.
159. Kanakry CG, Coffey DG, Towlerton AM, Vulic A, Storer BE, Chou J, et al. Origin and evolution of the T cell repertoire after posttransplantation cyclophosphamide. *JCI Insight.* 2016; **1**(5).
160. Chronister WD, Crinklaw A, Mahajan S, Vita R, Koşaloğlu-Yalçın Z, Yan Z, et al. TCRMatch: Predicting T-Cell Receptor Specificity Based on Sequence Similarity to Previously Characterized Receptors. *Front Immunol.* 2021; **12**: 640725.
161. Sidhom JW, Larman HB, Pardoll DM, Baras AS. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nat Commun*. 2021; **12**(1): 1605.
162. Dens C, Bittremieux W, Affaticati F, Laukens K, Meysman P, . Interpretable deep learning to uncover the molecular binding patterns determining TCR-epitope interactions. *Immuninformatics*. 2023; **11**: 100027.
163. Springer I, Tickotsky N, Louzoun Y, . Contribution of T Cell Receptor alpha and beta CDR3, MHC typing, V and J genes to peptide binding prediction. *Front Immunol*. 2021; **12**: 1436.
164. Grazioli F, Mösch A, Machart P, Li K, Alqassem I, O'Donnell TJ, et al. On TCR binding predictors failing to generalize to unseen peptides. *Front Immunol*. 2022; **13**: 1014256.
165. Soto C, Bombardi RG, Kozhevnikov M, Sinkovits RS, Chen EC, Branchizio A, et al. High Frequency of Shared Clonotypes in Human T Cell Receptor Repertoires. *Cell Rep.* 2020; **32**(2): 107882.

166. Woolhouse ME, Gowtage-Sequeria S. Host range and emerging and reemerging pathogens. *Emerg Infect Dis.* 2005; **11**(12): 1842-7.
167. Mason D. A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunology Today*. 1998; **19**(9): 395-404.
168. Singh NK, Riley TP, Baker SCB, Borrman T, Weng Z, Baker BM. Emerging Concepts in TCR Specificity: Rationalizing and (Maybe) Predicting Outcomes. J Immunol. 2017; **199**(7): 2203-13.
169. Leem J, de Oliveira SH P, Krawczyk K, Deane CM. STCRDab: the structural T-cell receptor database. *Nucleic Acids Res*. 2017; **46**(D1): D406-D12.
170. Degauque N, Brouard S, Soulillou J-P. Cross-Reactivity of TCR Repertoire: Current Concepts, Challenges, and Implication for Allotransplantation. *Front Immunol*. 2016; **7**.
171. Jerne NK. The Natural Selection theory of antibody formation. *PNAS*. 1955; **41**:849-57.
172. Welsh RM, Che JW, Brehm MA, Selin LK. Heterologous immunity between viruses. *Immunol Rev*. 2010; **235**(1): 244-66.
173. Iheagwara UK, Beatty PL, Van PT, Ross TM, Minden JS, Finn OJ. Influenza Virus Infection Elicits Protective Antibodies and T Cells Specific for Host Cell Antigens Also Expressed as Tumor-Associated Antigens: A New View of Cancer Immunosurveillance. *Can Immunol Res.* 2014; **2**(3): 263-73.
174. Johnson LA, Heemskerk B, Powell DJ, Jr., Cohen CJ, Morgan RA, Dudley ME, et al. Gene Transfer of Tumor-Reactive TCR Confers Both High Avidity and Tumor Reactivity to Nonreactive Peripheral Blood Mononuclear Cells and Tumor-Infiltrating Lymphocytes1. *J Immunol*. 2006; **177**(9): 6548-59.
175. Morgan RA, Chinnasamy N, Abate-Daga D, Gros A, Robbins PF, Zheng Z, et al. Cancer Regression and Neurological Toxicity Following Anti-MAGE-A3 TCR Gene Therapy. *J Immunother*. 2013; **36**(2): 133-51.
176. Parkhurst MR, Yang JC, Langan RC, Dudley ME, Nathan D-AN, Feldman SA, et al. T Cells Targeting Carcinoembryonic Antigen Can Mediate Regression of Metastatic Colorectal Cancer but Induce Severe Transient Colitis. *Mol Ther*. 2011; **19**(3): 620-6.
177. Fleischhauer K, Kernan NA, O'Reilly RJ, Dupont B, Yang SY. Bone marrow-allograft rejection by T lymphocytes recognizing a single amino acid difference in HLA-B44. *NEJM*. 1990; **323**(26): 1818-22.
178. Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. *Nature*. 1988; **334**(6181): 395-402.
179. Nathan P, Hassel JC, Rutkowski P, Baurain J-F, Butler MO, Schlaak M, et al. Overall Survival Benefit with Tebentafusp in Metastatic Uveal Melanoma. *NEJM*. 2021; **385**(13): 1196-206.
180. Corrie BD, Marthandan N, Zimonja B, Jaglale J, Zhou Y, Barr E, et al. iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunol Rev*. 2018; **284**(1): 24-41.
181. Andreatta M, Corria-Osorio J, Müller S, Cubas R, Coukos G, Carmona SJ. Interpretation of T cell states from single-cell transcriptomics data using reference atlases. *Nat Comms.* 2021; **12**(1): 2965.
182. Mayer A, Callan CG. Measures of epitope binding degeneracy from T cell receptor repertoires. *PNAS*. 2023; **120**(4): e2213264120.
183. Quaratino S, Thorpe CJ, Travers PJ, Londei M. Similar antigenic surfaces, rather than sequence homology, dictate T-cell epitope molecular mimicry. *PNAS*. 1995; **92**(22): 10398-402.
184. Lanzarotti E, Marcatili P, Nielsen M. T-Cell Receptor Cognate Target Prediction Based on Paired α and β Chain Sequence and Structural CDR Loop Similarities. *Front Immunol*. 2019; **10**:2080.
185. Leman JK, Weitzner BD, Lewis SM, Adolf-Bryfogle J, Alam N, Alford RF, et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods*. 2020; **17**(7): 665-80.

186. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*. 2021; **118**(15).
187. Evans R, O'neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer. *Preprint at bioRxiv*. 2022.
188. Jiang Y, Huo M, Li SC. TEINet: a deep learning framework for prediction of TCR-epitope binding specificity. *Briefings in Bioinformatics*. 2023; **24**(2).
189. Hudson D, Lubbock A, Basham M, Koohy H. A comparison of clustering models for inference of T cell receptor antigen specificity. *ImmunoInformatics*. 2024; **13**:100033.
190. Hayashi F, Isobe N, Glanville J, Matsushita T, Maimaitijiang G, Fukumoto S, et al. A new clustering method identifies multiple sclerosis-specific T-cell receptors. *Ann Clin Transl Neurol*. 2021; **8**(1): 163-76.
191. Pogorelyy MV, Rosati E, Minervina AA, Mettelman RC, Scheffold A, Franke A, et al. Resolving SARS-CoV-2 CD4+ T cell specificity via reverse epitope discovery. *Cell Rep Med*. 2022; **3**(8): 100697.
192. Wang Y, Duan F, Zhu Z, Yu M, Jia X, Dai H, et al. Analysis of TCR Repertoire by High-Throughput Sequencing Indicates the Feature of T Cell Immune Response after SARS-CoV-2 Infection. *Cells*. 2021; **11**(1): 68.
193. Drost F, An Y, Dratva LM, Lindeboom RG, Haniffa M, Teichmann SA, et al. Integrating T-cell receptor and transcriptome for large-scale single-cell immune profiling analysis. *Preprint at bioRxiv*. 2021.
194. Schattgen SA, Guion K, Crawford JC, Souquette A, Barrio AM, Stubbington MJT, et al. Integrating T cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (CoNGA). *Nat Biotechnol*. 2022; **40**(1): 54-63.
195. Sethna Z, Elhanati Y, Callan CG, Walczak AM, Mora T, . OLGA: Fast computation of generation probabilities of B-and T-cell receptor amino acid sequences and motifs. *Bioinformatics*. 2019; **35**(17): 2974-81.
196. Pogorelyy MV, Minervina AA, Shugay M, Chudakov DM, Lebedev YB, Mora T, et al. Detecting T cell receptors involved in immune responses from single repertoire snapshots. *PLoS Bio*. 2019; **17**(6): e3000314.
197. Pedregosa F, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, et al. Scikit-learn: Machine Learning in Python. *JMLR*. 2011; **12**: 2825-30.
198. Meysman P, De Neuter N, Gielis S, Bui Thi D, Ogunjimi B, Laukens K. On the viability of unsupervised T-cell receptor sequence clustering for epitope preference. *Bioinformatics*. 2018; **35**(9): 1461-8.
199. Postovskaya A, Vujkovic A, deBlock T, van Petersen L, van Frankenhuijsen M, Brosius I, et al. Leveraging T-cell receptor – epitope recognition models to disentangle unique and cross-reactive T-cell response to SARS-CoV-2 during COVID-19 progression/resolution. *Front Immunol*. 2023; **14**.
200. Graphpad v10.0.3 for mac gs, Boston, Massachusetts, USA.
201. Hunter JD. Matplotlib: A 2D graphics environment. *Comp Sci Eng*. 2007; **9**(3): 90-5.
202. Waskom M. seaborn: statistical data visualization. *J Open Source Software*. 2021; **6**(60): 3021.
203. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004; **32**(5): 1792-7.
204. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: A Sequence Logo Generator. *Gen Res*. 2004; **14**(6): 1188-90.
205. Quast NP, Abanades B, Guloglu B, Karuppiah V, Harper S, Raybould MIJ, et al. T-cell receptor structures and predictive models reveal comparable alpha and beta chain structural diversity despite differing genetic complexity. *Preprint at bioRxiv*. 2024.
206. McMaster B, Thorpe C, Ogg G, Deane CM, Koohy H. Can AlphaFold's breakthrough in protein structure help decode the fundamental principles of adaptive cellular immunity? *Nat Methods.* 2024; **21**: 766-776.

207. Klebanoff CA, Chandran SS, Baker BM, Quezada SA, Ribas A. T cell receptor therapeutics: immunological targeting of the intracellular cancer proteome. *Nat Rev Drug Disco*. 2023; **22**: 996-1017.
208. Yin Y, Mariuzza RA. The Multiple Mechanisms of T Cell Receptor Cross-reactivity. *Immunity*. 2009; **31**(6): 849-51.
209. Laura AJ, Richard AM, Mark ED, Lydie C, James CY, Marybeth SH, et al. Gene therapy with human and mouse T-cell receptors mediates cancer regression and targets normal tissues expressing cognate antigen. *Blood*. 2009; **114**(3): 535-46.
210. Saligrama N, Zhao F, Sikora MJ, Serratelli WS, Fernandes RA, Louis DM, et al. Opposing T cell responses in experimental autoimmune encephalomyelitis. *Nature*. 2019; **572**(7770): 481-7.
211. Karapetyan AR, Chaipan C, Winkelbach K, Wimberger S, Jeong JS, Joshi B, et al. TCR fingerprinting and off-target peptide identification. *Front Immunol*. 2019; **10**: 2501.
212. Poole A, Karuppiah V, Hartt A, Haidar JN, Moureau S, Dobrzycki T, et al. Therapeutic high affinity T cell receptor targeting a KRASG12D cancer neoantigen. *Nat Comms.* 2022; **13**(1): 5333.
213. Cameron BJ, Gerry AB, Dukes J, Harper JV, Kannan V, Bianchi FC, et al. Identification of a Titin-Derived HLA-A1-Presented Peptide as a Cross-Reactive Target for Engineered MAGE A3-Directed T Cells HHS Public Access. *Sci Trans Med*. 2013; **5**: 197-03.
214. McCarthy JEV. Functional and structural correlates of T cell receptor specificity and cross-reactivity. *DPhil thesis, University of Oxford*; 2024.
215. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res*. 2020; **48**(W1): W449-W54.
216. Vazquez-Lombardi R, Jung JS, Schlatter FS, Mei A, Mantuano NR, Bieberich F, et al. High-throughput T cell receptor engineering by functional screening identifies candidates with enhanced potency and specificity. *Immunity*. 2022; **55**(10): 1953-66.e10.
217. Riley TP, Hellman LM, Gee MH, Mendoza JL, Alonso JA, Foley KC, et al. T cell receptor cross-reactivity expanded by dramatic peptide–MHC adaptability. *Nat Chem Bio*. 2018; **14**(10): 934-42.
218. Hellman LM, Foley KC, Singh NK, Alonso JA, Riley TP, Devlin JR, et al. Improving T Cell Receptor On-Target Specificity via Structure-Guided Design. *Mol Ther*. 2019; **27**(2): 300-13.
219. Ito Y, Nakamura Y, Takahashi N, Saito S, Aleksic B, Iwata N, et al. A genetic association study of the FXYD domain containing ion transport regulator 6 (FXYD6) gene, encoding phosphohippolin, in susceptibility to schizophrenia in a Japanese population. *Neurosci Lett.* 2008; **438**(1): 70-5.
220. Lee JK, Stewart-Jones G, Dong T, Harlos K, Di Gleria K, Dorrell L, et al. T Cell Cross-Reactivity and Conformational Changes during TCR Engagement. *J Exp Med*. 2004; **200**(11): 1455-66.
221. Wooldridge L, Ekeruche-Makinde J, van den Berg HA, Skowera A, Miles JJ, Tan MP, et al. A Single Autoimmune T Cell Receptor Recognizes More Than a Million Different Peptides. *J Biol Chem*. 2012; **287**(2): 1168-77.
222. Dolton G, Rius C, Wall A, Szomolay B, Bianchi V, Galloway SAE, et al. Targeting of multiple tumor-associated antigens by individual T cell receptors during successful cancer immunotherapy. *Cell*. 2023; **186**(16): 3333-49.e27.
223. Raman MCC, Rizkallah PJ, Simmons R, Donnellan Z, Dukes J, Bossi G, et al. Direct molecular mimicry enables off-target cardiovascular toxicity by an enhanced affinity TCR designed for cancer immunotherapy. *Sci Rep*. 2016; **6**(1): 18851.
224. Gaugler B, Van den Eynde B, van der Bruggen P, Romero P, Gaforio JJ, De Plaen E, et al. Human gene MAGE-3 codes for an antigen recognized on a melanoma by autologous cytolytic T lymphocytes. *J Exp Med*. 1994; **179**(3): 921-30.
225. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*. 2014; **30**(5): 614-20.

226. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009; **25**(11): 1422-3.
227. McInnes L, Healy J, Astels S. hdbscan: Hierarchical density based clustering. *J Open Source Software*. 2017; **2**(11): 205.
228. Bonald T, De Lara N, Lutz Q, Charpentier B. Scikit-network: graph analysis in Python. *J Mach Learn Res*. 2020; **21**(1).
229. Dotmatics. Geneious.
230. Van Der Maaten L, Hinton G, . Visualizing Data using t-SNE. *JMLR*. 2008; **9**: 2579-605.
231. Lecerf M, Lacombe RV, Dimitrov JD. Polyreactivity of antibodies from different B-cell subpopulations is determined by distinct sequence patterns of variable region. *Front Immunol.* 2023; **14**.
232. Chen Y-T, Scanlan MJ, Sahin U, Türeci Ö, Gure AO, Tsang S, et al. A testicular antigen aberrantly expressed in human cancers detected by autologous antibody screening. *PNAS*. 1997; **94**(5): 1914-8.
233. Thomas R, Al-Khadairi G, Roelands J, Hendrickx W, Dermime S, Bedognetti D, et al. NY-ESO-1 Based Immunotherapy of Cancer: Current Perspectives. *Front Immunol.* 2018; **9**:fimmu.2018.00947.
234. Chen J-L, Dunbar PR, Gileadi U, Jäger E, Gnjatic S, Nagata Y, et al. Identification of NY-ESO-1 Peptide Analogues Capable of Improved Stimulation of Tumor-Reactive CTL1. *J Immunol.* 2000; **165**(2): 948-55.
235. Lin MJ, Svensson-Arvelund J, Lubitz GS, Marabelle A, Melero I, Brown BD, et al. Cancer vaccines: the next immunotherapy frontier. *Nat Canc*. 2022; **3**(8): 911-26.
236. Lopez J, Shinde R, Burgess MA, Sato T, Thistlethwaite F, Van Tine B, et al. 1309TiP - Trial in progress: First-in-human study of a novel anti-NY-ESO-1–anti-CD3, TCR-based bispecific (IMCnyeso) as monotherapy in NY-ESO-1/LAGE-1A-positive advanced solid tumours (IMCnyeso-101). *Ann Oncol*. 2019; **30**: v532.
237. Liu X, Xu Y, Xiong W, Yin B, Huang Y, Chu J, et al. Development of a TCR-like antibody and chimeric antigen receptor against NY-ESO-1/HLA-A2 for cancer immunotherapy. *J Immunother Canc*. 2022; **10**(3): e004035.
238. Pan Q, Weng D, Liu J, Han Z, Ou Y, Xu B, et al. Phase 1 clinical trial to assess safety and efficacy of NY-ESO-1-specific TCR T cells in HLA-A∗02:01 patients with advanced soft tissue sarcoma. *Cell Rep Med.* 2023; **4**(8): 101133.
239. Chen J-L, Stewart-Jones G, Bossi G, Lissin NM, Wooldridge L, Choi EML, et al. Structural and kinetic basis for heightened immunogenicity of T cell vaccines. *J Exp Med.* 2005; **201**(8): 1243-55.
240. Robbins PF, Li YF, El-Gamil M, Zhao Y, Wargo JA, Zheng Z, et al. Single and dual amino acid substitutions in TCR CDRs can enhance antigen-specific T cell functions. *J Immunol.* 2008; **180**(9): 6116-31.
241. Robbins PF, Morgan RA, Feldman SA, Yang JC, Sherry RM, Dudley ME, et al. Tumor Regression in Patients With Metastatic Synovial Cell Sarcoma and Melanoma Using Genetically Engineered Lymphocytes Reactive With NY-ESO-1. *J Clin Oncol*. 2011; **29**(7): 917-24.
242. D'Angelo SP, Melchiori L, Merchant MS, Bernstein D, Glod J, Kaplan R, et al. Antitumor Activity Associated with Prolonged Persistence of Adoptively Transferred NY-ESO-1 c259T Cells in Synovial Sarcoma. *Cancer Discovery*. 2018; **8**(8): 944-57.
243. Zhao Y, He B, Xu F, Li C, Xu Z, Su X, et al. DeepAIR: A deep learning framework for effective integration of sequence and 3D structure to enable adaptive immune receptor analysis. *Sci Adv.* 2023; **9**(32): eabo5128.
244. Fast E, Dhar M, Chen B. TAPIR: a T-cell receptor language model for predicting rare and novel targets. *Preprint on bioRxiv*. 2023.

245. Croce G, Bobisse S, Moreno DL, Schmidt J, Guillame P, Harari A, et al. Deep learning predictions of TCR-epitope interactions reveal epitope-specific chains in dual alpha T cells. *Nat Comms*. 2024; **15**(1): 3211.
246. Tickotsky N. POP-UP TCR: Prediction of Previously Unseen Paired TCR-pMHC. *Preprint on bioRxiv.* 2023.
247. Kwee BPY, Messemaker M, Marcus E, Oliveira G, Scheper W, Wu CJ, et al. STAPLER: Efficient learning of TCR-peptide specificity prediction from full-length TCR-peptide data. *Preprint on bioRxiv.* 2023.
248. Meynard-Piganeau B, Feinauer C, Weigt M, Walczak AM, Mora T. TULIP — a Transformer based Unsupervised Language model for Interacting Peptides and T-cell receptors that generalizes to unseen epitopes. *Preprint on bioRxiv*. 2024.
249. Chen J, Zhao B, Lin S, Sun H, Mao X, Wang M, et al. TEPCAM: Prediction of T-cell receptor–epitope binding specificity via interpretable deep learning. *Protein Science*. 2024; **33**(1): e4841.
250. Boulanger DSM, Eccleston RC, Phillips A, Coveney PV, Elliott T, Dalchau N. A Mechanistic Model for Predicting Cell Surface Presentation of Competing Peptides by MHC Class I Molecules. *Front Immunol*. 2018; **9**: 1538.
251. Liao H, Barra C, Zhou Z, Peng X, Woodhouse I, Tailor A, et al. MARS an improved de novo peptide candidate selection method for non-canonical antigen target discovery in cancer. *Nat Comms*. 2024; **15**(1): 661.
252. Chong C, Müller M, Pak H, Harnett D, Huber F, Grun D, et al. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat Comms*. 2020; **11**(1): 1293.
253. Racle J, Guillaume P, Schmidt J, Michaux J, Larabi A, Lau K, et al. Machine learning predictions of MHC-II specificities reveal alternative binding mode of class II epitopes. *Immunity*. 2023; **56**(6): 1359-75.e13.
254. Huang S, Shahine A, Cheng TY, Chen YL, Ng SW, Balaji GR, et al. CD1 lipidomes reveal lipid-binding motifs and size-based antigen-display mechanisms. *Cell*. 2023; **186**(21): 4583-96.e13.
255. Achar SR, Bourassa FXP, Rademaker TJ, Lee A, Kondo T, Salazar-Cavazos E, et al. Universal antigen encoding of T cell activation from high-dimensional cytokine dynamics. *Science*. 2022; **376**(6595): 880-4.
256. Keck S, Schmaler M, Ganter S, Wyss L, Oberle S, Huseby ES, et al. Antigen affinity and antigen dose exert distinct influences on CD4 T-cell differentiation. *PNAS*. 2014; **111**(41): 14852-7.
257. van Panhuys N, Klauschen F, Germain RN. T-cell-receptor-dependent signal intensity dominantly controls CD4(+) T cell polarization In Vivo. *Immunity*. 2014; **41**(1): 63-74.
258. Han Y, Yang Y, Tian Y, Fattah FJ, Itzstein MSv, Hu Y, et al. pan-MHC and cross-Species Prediction of T Cell Receptor-Antigen Binding. *Preprint at bioRxiv*. 2023.
259. Nagano Y, Pyo A, Milighetti M, Henderson J, Shawe-Taylor J, Chain B, et al. Contrastive learning of T cell receptor representations. *Preprint at aRxiv*. 2024.
260. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*. 2024; **630**(8016): 493-500.
261. Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, et al. De novo design of protein structure and function with RFdiffusion. *Nature*. 2023; **620**(7976): 1089-100.
262. Abanades B, Wong WK, Boyles F, Georges G, Bujotzek A, Deane CM. ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. *Nat Comms Bio.* 2023; **6**(1): 575.
263. Richard E, Michael ON, Alexander P, Natasha A, Andrew S, Tim G, et al. Protein complex prediction with AlphaFold-Multimer. *Preprint at bioRxiv*. 2022.
264. Liu S, Iorgulescu JB, Li S, Borji M, Barrera-Lopez IA, Shanmugam V, et al. Spatial maps of T cell receptors and transcriptomes reveal distinct immune niches and interactions in the adaptive immune response. *Immunity*. 2022; **55**(10): 1940-52.e5.

265. Kacen A, Javitt A, Kramer MP, Morgenstern D, Tsaban T, Shmueli MD, et al. Post-translational modifications reshape the antigenic landscape of the MHC I immunopeptidome in tumors. *Nat Biotechnol*. 2023; **41**(2): 239-51.
266. Curran AM, Girgis AA, Jang Y, Crawford JD, Thomas MA, Kawalerski R, et al. Citrullination modulates antigen processing and presentation by revealing cryptic epitopes in rheumatoid arthritis. *Nat Comms*. 2023; **14**(1): 1061.
267. Prasasty VD, Istyastono EP. Data of small peptides in SMILES and three-dimensional formats for virtual screening campaigns. *Data in Brief*. 2019; **27**: 104607.
268. Harris DT, Hager MV, Smith SN, Cai Q, Stone JD, Kruger P, et al. Comparison of T Cell Activities Mediated by Human TCRs and CARs That Use the Same Recognition Domains. *J Immunol.* 2018; **200**(3): 1088-100.
269. Burton J, Siller-Farfán JA, Pettmann J, Salzer B, Kutuzov M, van der Merwe PA, et al. Inefficient exploitation of accessory receptors reduces the sensitivity of chimeric antigen receptors. *PNAS*. 2023; **120**(2): e2216352120.
270. Li Y, Moysey R, Molloy PE, Vuidepot A-L, Mahon T, Baston E, et al. Directed evolution of human T-cell receptors with picomolar affinities by phage display. *Nat Biotechnol.* 2005; 23(3): 349-54.
271. Liddy N, Bossi G, Adams KJ, Lissina A, Mahon TM, Hassan NJ, et al. Monoclonal TCR-redirected tumor cell killing. *Nat Med*. 2012; **18**(6): 980-7.
272. Middleton MR, McAlpine C, Woodcock VK, Corrie P, Infante JR, Steven NM, et al. Tebentafusp, A TCR/Anti-CD3 Bispecific Fusion Protein Targeting gp100, Potently Activated Antitumor Immune Responses in Patients with Metastatic Melanoma. *Clin Canc Res*. 2020; **26**(22): 5869-78.
273. Lodha M, Erhard F, Dölken L, Prusty BK. The Hidden Enemy Within: Non-canonical Peptides in Virus-Induced Autoimmunity. *Front Microbiol*. 2022; **13**:840911.
274. Mattei AE, Gutierrez AH, Seshadri S, Tivin J, Ardito M, Rosenberg AS, et al. In silico methods for immunogenicity risk assessment and human homology screening for therapeutic antibodies. mAbs. 2024; **16**(1): 2333729.
275. Lee CH, Huh J, Buckley PR, Jang M, Pinho MP, Fernandes RA, et al. A robust deep learning workflow to predict CD8 + T-cell epitopes. *Gen Med.* 2023; **15**(1).
276. Calis JJ, Maybeno M, Greenbaum JA, Weiskopf D, De Silva AD, Sette A, et al. Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comp Bio*. 2013; **9**(10): e1003266.
277. Paul S, Lindestam Arlehamn CS, Scriba TJ, Dillon MBC, Oseroff C, Hinz D, et al. Development and validation of a broad scheme for prediction of HLA class II restricted T cell epitopes. *J Immunol Methods*. 2015; **422**:28-34.
278. Muntel J, Boswell SA, Tang S, Ahmed S, Wapinski I, Foley G, et al. Abundance-based classifier for the prediction of mass spectrometric peptide detectability upon enrichment (PPA). *Mol Cell Proteomics*. 2015; **14**(2): 430-40.
279. DeWitt WS, III, Smith A, Schoch G, Hansen JA, Matsen FAIV, Bradley P. Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *eLife*. 2018; **7**: e38358.
280. Gupta S, Sgourakis NG. A structure-guided approach to predict MHC-I restriction of T cell receptors for public antigens. *Preprint at bioRxiv*. 2024.