Immunoinformatic identification of CD8+ T-cell epitopes



Chloe Hyun-Jung Lee

Kellogg College

University of Oxford

Supervisors:

Prof. Hashem Koohy

Prof. Alison Simmons

A thesis submitted for the degree of

Doctor of Philosophy

October 2022

Acknowledgements

This thesis represents a huge amount of work over a number of years working closely with a number of extremely talented individuals who have provided help, advice and support.

I would like to thank Professor Hashem Koohy and Professor Alison Simmons for supervising this work and providing the opportunity to pursue these projects. I am grateful for their guidance and for providing an environment that was both friendly and scientifically stimulating. I would also like to thank my thesis committee, Professor Graham Ogg, Professor Omer Dushek and Dr. Mikhail Shugay for their feedback as the project developed.

Within the lab, I would like to thank Dr. Agne Antanaviciute, Tarun Gupta, Dr. Ottakandathil Babu and Dr. Anna Aulicino - the single-cell project was a collaborative effort involving clinicians, wet lab experimentalists and computational biologists and it would not have been possible without their advice and contributions. I'd also like to thank Dr. Paul Buckley for working closely on COVID-19 and immunogenicity projects. I am also grateful to Jaesung Huh and Myeongjun Jang for the discussions on deep learning techniques, and value conversations with colleagues at CCB. It has been extremely grateful to work closely with everyone in the group, and was a pleasure to learn and work at such a collaborative environment.

I appreciate the discussions, advice and afternoon chats from all members of Koohy and Simmons group, as well as all co-authors of publications that have resulted from this work. I also acknowledge the help of the National Institute of Health Research for funding my DPhil.

I would like to express my heartfelt gratitude to my family. Thank you for all the encouragement, warm advice and faith in wherever I have been, and thank you for your prayers and support throughout this entire journey. Lastly, I would like to thank God Almighty for the opportunity and strength to undertake this journey.

Attributions

This thesis represents the results of collaborations and contributions of different researchers I had the privilege to work with during my PhD. Here are the chapter-by-chapter contributions in the works included in this thesis.

- Chapter 1-2. All content is contributed exclusively by me.
- Chapter 3. For this multi-modal single-cell study, I worked very closely with Dr. Agne Antanaviciute and Dr. Rosana Ottakandathil Babu on bioinformatics analysis, and Tarun Gupta and Dr. Anna Aulicino on sample collection, processing, and data generation. Within this project for the manuscript in preparation, I conducted single-cell transcriptomics and TCR analysis, which included analysing immune repertoire composition, TCR dynamics, clonotype sharing, differential abundance in various pathologies, tissues, or individuals, and inferring T-cell specificity, and thus included in this chapter. Other analysis, such as those involving spatial transcriptomics data led by Dr. Agne Antanaviciuta, have been excluded from this chapter.
- Chapter 4. All analysis in this chapter, such as predicting CD8+ T-cell epitopes from coronavirus strains, analysing cross-reactivity potential, and evaluating self-reactivity, was done by me, except for the analysis on HLA-dependent variations in SARS-CoV-2 CD8+ T-cell cross-reactivity, which was done in collaboration with Dr. Paul Buckley.
- Chapter 5. All analysis in this chapter was done by me, except for the benchmarking of existing models done in collaboration with Dr. Paul Buckley. Glioblastoma (GBM) data used for benchmarking analysis was generated by Dr. Ruichong Ma.
- Chapter 6. All analysis in this chapter was done by me, except for the analysis on the differences in T-cell recognition features between neoepitopes and pathogenic peptides, which was done in collaboration with Dr. Paul Buckley.
- Chapter 7. The TRAP and RSAT models presented in this chapter were developed exclusively by me. The application of TRAP to study the immune escape of SARS-CoV-2 was done in collaboration with Dr. Paul Buckley.
- Chapter 8. All content is contributed exclusively by me.

Publication statement

Some of the work included in this thesis had been previously published. The works included in this thesis comply with the 're-use of published material' guidance in the following ways: I was the first or co-first author of the publications, I was involved in the creation of all figures, and I have permission from my co-authors and supervisors to re-use the figures and contents of the manuscript.

A list of publications directly linked with the results of this work includes:

- **Lee CH**, Koohy H. In silico identification of vaccine targets for 2019-nCoV. F1000Res 9, 145 (2020)
- **Lee CH**, Pinho MP, Buckley PR, Woodhouse IB, Ogg G, Simmons A, Koohy H. Potential CD8+ T-cell Cross-Reactivity Against SARS-CoV-2 Conferred by Other Coronavirus Strains. Front. Immunol. 11, (2020).
- **Lee CH**, Salio M, Napolitani G, Ogg G, Simmons A, Koohy H. Predicting Cross-Reactivity and Antigen Specificity of T-cell Receptors. Front. Immunol. 11, (2020).
- Buckley, PR*, **Lee, CH***, Ma, R*, Woodhouse I, Woo J, Tsvetkov VO, Shscherbinin BS, Antanaviciute A, Shugay M, Rei M, Simmons A, Koohy H. Evaluating performance of existing computational models in predicting CD8+ T-cell pathogenic epitopes and cancer neoantigens. Briefings in Bioinformatics. *These authors contributed equally.
- **Lee CH**, Antanaviciute A., Buckley PR, Simmons A. & Koohy, H. To what extent does MHC binding translate to immunogenicity in humans? Immunoinformatic (2021)

Details of copyrights are described in the ‘Copyright of figures and tables’ chapter at the end of the thesis.

Abstract

Antigen-specific T-cells play a crucial role in the adaptive immune response by providing a defence mechanism against pathogens and maintaining tolerance against self-antigens. This sparked interest in the development of epitope-based vaccines and immunotherapies that elicit antigen-specific T-cell responses. However, screening the antigens driving the response is currently labour-intensive, low-throughput and costly. Due to the limitations of experimental approaches, computational methods for predicting CD8+ T-cells have started to emerge. However, predicting the T-cell recognition potential of MHC-presented peptides has shown to be more challenging than predicting MHC ligands, and the full spectrum of features underlying peptide immunogenicity remains to be explored. Hence, this thesis presents a systems biology approach to study features of peptide immunogenicity and accurately predict CD8+ T-cell epitopes from HLA-I presented pathogenic or cancer peptides.

The thesis begins with an immunoinformatic analysis of antigen-specific T-cell profiles in the contexts of autoinflammatory and infectious diseases. In autoinflammatory disease, the multi-modal single-cell sequencing of ulcerative colitis and checkpoint treatment-induced colitis revealed pathology-specific differential expressions of cytotoxic T-cells. The current technologies, however, were unable to identify the source antigen, emphasising the importance of predicting T-cell targets to better understand disease pathology. Moreover, in infectious diseases, CD8+ T-cell epitope prediction algorithms facilitated the understanding of disease heterogeneity and vaccine design during the COVID-19 pandemic, but many existing algorithms were found to be ill-suited for predicting epitopes from emerging pathogens.

Therefore, a novel computational workflow was developed for an accurate and robust prediction of source antigens driving the cellular immune response. First, an unbiased evaluation of state-of-the-art algorithms revealed that they perform poorly on both cancer neoepitopes (e.g. glioblastoma) and pathogenic (e.g. SARS-CoV-2) epitopes. After investigating the reasons for low performance, TRAP, a deep learning workflow for context-specific prediction of CD8+ T-cell epitopes, was developed to effectively capture T-cell recognition motifs. One application of TRAP was demonstrated by using it to investigate the immune escape potential of all theoretical SARS-CoV-2 mutants. Thus, this thesis presents a novel computational platform for accurately predicting CD8+ T-cell epitopes to foster a better understanding of TCR:pMHC interaction and the development of effective clinical therapeutics.

Contents

Acknowledgements **2**

Attributions **3**

Publication statement **4**

Abstract **5**

List of abbreviations **14**

Chapter 1. Introduction **18**

Development of T-cell specificity 18
Antigen-specific T-cells for basic, clinical and translational research 19
Experimental strategies of T-cell epitope screening 21
Limitations of experimental approaches 23
Immunoinformatic identification of CD8+ T-cell epitopes 24
Remaining challenges in epitope prediction 24

Chapter 2. Outline of thesis **27**

Chapter 3. Divergent T-cell cytotoxicity programs between ulcerative and checkpoint inhibitor-induced colitis **29**

Introduction 29
Results 31
Unbiased scRNA-Seq reveal analogous and divergent T-cell state perturbation in checkpoint induced and ulcerative colitis 31
Divergent T-cell state perturbations in CC and UC 32
ZNF683+ TRMs enriched in checkpoint-induced compared to ulcerative colitis . 32
Single-cell TCR tracking reveals contrasting T-cell migration dynamics in CC and UC 39
Diverse unknown antigens drive T-cell migration from blood to the site of inflammation 43
Discussion 45

Methods . . . 47

Chapter 4. CD8+ T-cell targets against human coronaviruses 52
Introduction . . . 52
Results . . . 54
SARS-CoV-2 peptides sharing high sequence similarity to epitopes . . . 54
Identify T-cell targets conveying cross-protection . . . 55
Expand public epitopes by sequence similarity . . . 59
Validating predictions by epitopes deposited in IEDB . . . 60
HLA-dependent variation in SARS-CoV-2 CD8+ T-cell cross-reactivity with human coronaviruses . . . 62
Potential risk of self-reactivity by SARS-CoV-2 derived predicted epitopes . . . 64
Discussion . . . 66
Methods . . . 68

Chapter 5. Evaluate performance of existing computational methods in predicting CD8+ T-cell cancer and pathogenic epitopes 70
Introduction . . . 70
Results . . . 72
Computational strategies to identify CD8+ T-cell epitopes . . . 72
Evaluate performance of existing models in predicting pathogenic epitopes . . . 80
Evaluate performance of existing models in predicting cancer neoepitopes . . . 82
The effects of cross-HLA variation in predicting peptide immunogenicity . . . 86
Considerations in modelling immunogenicity . . . 87
Discussion . . . 91
Methods . . . 92

Chapter 6. Sources of low accuracies in identifying epitopes from emerging pathogen 94
Introduction . . . 94
Results . . . 96
Extent of MHC presentation translated to immunogenicity in humans . . . 96
Difference in T-cell recognition features between neoepitopes and pathogenic peptides . . . 97
Dissimilarity to self-proteome have low discriminative power for pathogenic peptides 102
Cross-species variation in T-cell recognition features . . . 102
High performance of immunogenicity models incorporating peptide-HLA pairs may be a reflection of HLA bias from imbalanced dataset . . . 105
Discussion . . . 108
Methods . . . 109

Chapter 7. A deep learning-based platform to predict T-cell recognition potential of MHC-I presented pathogenic and self-peptides 111

Introduction . . . 111
Results . . . 113
TRAP: A robust deep learning workflow to predict CD8+ T-cell recognition of MHC-I presented pathogenic and self-peptides . . . 113
Mitigate HLA bias by employing peptide sequences at TCR contact positions . . . 115
Deep learning architecture captures T-cell recognition motifs . . . 120
Sequence patterns discriminating immunogenicity . . . 123
Abstain low-confidence predictions to improve robustness . . . 125
Relative similarity to autoantigens or tumor-associated antigens (RSAT) as a novel feature of pathogenic peptide immunogenicity . . . 130
Application of TRAP to study immune escape of SARS-CoV-2 variants . . . 132
Discussion . . . 137
Methods . . . 140

Chapter 8. Concluding Discussion **147**
Summary of findings . . . 147
Divergent T-cell profiles between ulcerative and checkpoint treatment-induced colitis, likely driven by diverse unknown antigen . . . 148
Predicting CD8+ T-cell targets enhance understanding of COVID-19 pathology and facilitate vaccine design . . . 149
Existing computational models fail to accurately predict CD8+ T-cell targets in both pathogen and cancer settings . . . 151
Development of a deep learning workflow to discriminate CD8+ T-cell targets from MHC-I ligands . . . 152
Future directions . . . 153
Rules underlying antigen recognition by T-cells . . . 153
Rules underlying cross-reactivity of T-cells . . . 154
Rules underlying shared antigen-specificity of T-cells . . . 155
pMHC:TCR interactome . . . 156

Chapter 9. References **158**

Chapter 10. Copyright of figures and tables **176**

Appendix **179**

Word Count: 44302 (without Bibliography and Appendix)

List of Figures

1.1 Development of functional antigen-specific T-cells. 20
1.2 Cross-reactive and cross-species TCRs. 26

3.1 Multi-modal single-cell atlas of Healthy, Ulcerative colitis and Checkpoint inhibitor-induced colitis. 31
3.2 Convergence of transcriptional profiles in UC and CC. 33
3.3 Immune checkpoint gene expression in scRNA-Seq data. 34
3.4 Single-cell RNA-Seq on T-cell subpopulations. 34
3.5 Differential T-cell responses in colonic tissue and blood in colitis. 35
3.6 T-cell abundance in CC compared to UC. 36
3.7 Condition-specific local neighbourhood abundance differences. 37
3.8 ZNF683+ TRMs enriched in CC compared to UC. 38
3.9 Abundance and proportion of cycling T-cells. 39
3.10 T-cell trafficking and clonal expansion in healthy, UC and CC. 40
3.11 Trafficking and clonal expansions of T-cells in Health and Disease. 41
3.12 Increased expression of tissue egress and lymph-node homing markers in UC. . . 42
3.13 Using public database to reference TCR specificity. 44
3.14 Predicting TCR specificity of UC and CC clones. 45

4.1 SARS-CoV-2 peptides with high sequence similarity to epitopes deposited in IEDB. 56
4.2 Sequence homology between coronavirus strains. 56
4.3 Number of shared and private 9-mer peptides across coronavirus strains. 57
4.4 Number of 9-mer peptides predicted to bind HLA alleles and immunogenic. . . . 58
4.5 Extent of potential CD8+ T-cell cross-reactivity across coronavirus strains. . . . 59
4.6 Number of shared predicted epitopes between SARS-CoV-2 and other coronavirus strains by allowing up to two mismatches. 60
4.7 Landscape of T-cell responses against SARS-CoV-2 epitopes. 63
4.8 Sequence similarity with human proteome. 64

5.1 Features associated with TCR:pMHC interaction. 72
5.2 Models are not reliable in predicting epitopes for an emerging virus and exhibit room for improvement in predicting immunogenic cancer neoantigens. 81

5.3 Majority of the immunogenicity models do not outperform random prediction by bootstrapping. . . . 82
5.4 Models are not reliable in predicting neoepitopes from glioblastoma patients. . . . 84
5.5 Models do not show high performance on TESLA neoantigen dataset. . . . 85
5.6 HLA-variation in the number of epitopes and non-epitopes for model training. . . . 86
5.7 Effects of cross-HLA variation on predicting peptide immunogenicity. . . . 88
5.8 Predictability of immunogenicity models on HLA-A*02:01 binding. . . . 89

6.1 Extent of MHC presentation and T-cell recognition in human and mouse systems. 97
6.2 Difference in magnitude of MHC binding affinity in discriminating epitopes and non-epitopes for pathogenic and cancer peptides. . . . 98
6.3 Difference in magnitude of MHC binding stability in discriminating epitopes and non-epitopes for pathogenic and cancer peptides. . . . 99
6.4 Difference in hydrophobicity in discriminating epitopes and non-epitopes for pathogenic and cancer peptides. . . . 100
6.5 Comparing MHC presentation features on two independent cancer datasets, in-house glioblastoma (GBM) and TESLA datasets. . . . 101
6.6 Dissimilarity to self-proteome cannot distinguish pathogenic epitopes from non-epitopes. . . . 102
6.7 Cross-species variation in T-cell recognition features. . . . 103
6.8 Models trained using cross-species peptides cannot effectively predict immunogenicity of peptides derived from unseen pathogen. . . . 104
6.9 Sequence patterns associated with immunogenicity across different pathogens. . . . 105
6.10 Performance of DeepImmuno. . . . 106
6.11 High performance may be a reflection of HLA bias. . . . 107
6.12 DeepImmuno show lower performance on HLA-balanced data. . . . 107

7.1 Performance of TRAP. . . . 114
7.2 Effect of anchor and contact positions on peptide immunogenicity. . . . 116
7.3 Intra- and inter-HLA variation in positivity score. . . . 119
7.4 Pathogenic and self-antigen datasets. . . . 120
7.5 Performance of pre- and post-padding strategies. . . . 121
7.6 Compare amino acid embedding strategies. . . . 122
7.7 Compare performance of different deep learning architectures. . . . 123
7.8 Adding hallmarks of immunogenicity. . . . 124
7.9 Sequence motifs enriched or depleted in pathogenic and self-epitopes. . . . 126
7.10 Compare enriched motifs across pathogenic and self-peptides. . . . 127
7.11 Clusters of peptides sharing high sequence similarity by pairwise global alignment score at contact positions. . . . 127
7.12 Detect out-of-distribution using autoencoders. . . . 128
7.13 OOD detection using calibration methods. . . . 129
7.14 Relative similarity to autoantigens and tumor-associated antigens (RSAT). . . . 131

7.15 RSAT can effectively discriminate pathogenic epitopes from non-epitopes of varying lengths and species. 131
7.16 Performance of TRAP model trained using SARS-CoV data. 133
7.17 Impact of amino acid substitutions on SARS-CoV-2 immunogenicity. 133
7.18 Impact of position-specific substitutions on epitope immunogenicity. 135
7.19 Neighbour network illustrating deleterious and beneficial mutations from Wuhan Hu-1 epitopes. 136

8.1 Interplay between unique clusters of pMHCs and TCRs. 156

S1 Alignment of spike protein sequences from different coronavirus strains 180
S2 Cross-reactive TCRs recognizing epitopes derived from multiple species. 182
S3 Workflow for predicting antigen specificity of TCRs. 183

List of Tables

1.1 Experimental strategies of T-cell epitope screening . . . 22
1.1 Experimental strategies of T-cell epitope screening . . . 23

4.1 Comparison of normalized sequence alignment score for peptides with exact and non-exact matches. . . . 54
4.2 SARS-CoV-2 peptides having exact matches with SARS-CoV epitopes deposited in IEDB. . . . 54
4.2 SARS-CoV-2 peptides having exact matches with SARS-CoV epitopes deposited in IEDB. . . . 55
4.3 Number of unique peptides by strain that have a matching pattern with immunogenic peptides in the IEDB. . . . 61
4.4 Compare prediction with functional studies from Nelde *et al.* . . . 62
4.5 Predicted epitopes from SARS-CoV-2 that are single amino acid variants of human proteome counterparts. . . . 64
4.5 Predicted epitopes from SARS-CoV-2 that are single amino acid variants of human proteome counterparts. . . . 65

5.1 List of algorithms to predict immunogenicity . . . 75
5.1 List of algorithms to predict immunogenicity . . . 76
5.1 List of algorithms to predict immunogenicity . . . 77
5.1 List of algorithms to predict immunogenicity . . . 78
5.1 List of algorithms to predict immunogenicity . . . 79
5.1 List of algorithms to predict immunogenicity . . . 80
5.2 Examples of reasons for excluding models from benchmarking . . . 80

7.1 Sources of intra-HLA variation. . . . 118

S1 Cluster summary of Tissue CD3+ cells. . . . 184
S1 Cluster summary of Tissue CD3+ cells. . . . 185
S1 Cluster summary of Tissue CD3+ cells. . . . 186
S1 Cluster summary of Tissue CD3+ cells. . . . 187
S1 Cluster summary of Tissue CD3+ cells. . . . 188
S1 Cluster summary of Tissue CD3+ cells. . . . 189

S1 Cluster summary of Tissue CD3+ cells. 190
S1 Cluster summary of Tissue CD3+ cells. 191
S1 Cluster summary of Tissue CD3+ cells. 192
S2 Cluster summary of Blood CD3+ cells 193
S2 Cluster summary of Blood CD3+ cells 194
S2 Cluster summary of Blood CD3+ cells 195
S2 Cluster summary of Blood CD3+ cells 196
S3 Cluster Summary of Epithelial, Stromal and Immune populations 197
S3 Cluster Summary of Epithelial, Stromal and Immune populations 198
S3 Cluster Summary of Epithelial, Stromal and Immune populations 199
S3 Cluster Summary of Epithelial, Stromal and Immune populations 200
S4 List of 9-mer peptides from Nelde et al. compared with predictions. 219
S4 List of 9-mer peptides from Nelde et al. compared with predictions. 220

List of abbreviations

Abbreviation	Descriptions
5HT1B	5-hydroxytryptamine receptor 1B. Distributed throughout the central nervous system
AAT	Autoimmunity, Allergy or Tumor-associated antigens
AE	Autoencoder
AI	Artificial intelligence
AIM	Activation induced marker
AMD	Age-related macular degeneration
ANN	Artificial neural network
ANOVA	Analysis of variance
APC	Antigen-presenting cells
ASA	Accessible surface area
B7	A type of integral membrane protein found on activated antigen-presenting cells (APC) that can produce a costimulatory or a coinhibitory signal.
BA	Binding affinity
BCR	B cell receptor
BERT	Bidirectional Encoder Representations from Transformers
BLOSUM	Blocks substitution matrix. Substitution matrix used for sequence alignment of proteins
CC	Checkpoint inhibitor-induced colitis
CCR	Chemokine receptor
CD103	Cluster of differentiation 103 antigen, also known as integrin αEβ7 (ITGAE). It is a transmembrane heterodimer complex that is involved in cell to cell or cell to matrix interaction
CD107	Cluster of differentiation 107 antigen. Also known as lysosome-associated membrane glycoprotein 1
CD28	Cluster of differentiation 28 antigen. A primary costimulatory molecule for T-cell activation
CD3	Cluster of differentiation 3 antigen. T-cell co-receptor that is involved in activating both the cytotoxic T-cell (CD8+ naive T-cells) and T helper cells (CD4+ naive T-cells)

Abbreviation	Descriptions
CD4	Cluster of differentiation 4 antigen. Co-receptor for the T-cell receptor (TCR) found on the surface of immune cells such as T helper cells, monocytes, macrophages, and dendritic cells
CD44	Cluster of differentiation 44 antigen. Cell-surface glycoprotein involved in cell–cell interactions, cell adhesion and migration.
CD45	Cluster of differentiation 45 antigen. Also known as Protein tyrosine phosphatase, receptor type, C
CD8	Cluster of differentiation 8 antigen. Co-receptor playing a role in T-cell signalling and aiding with cytotoxic T-cell-antigen interactions
CDR	Complementarity-determining region
CITE-Seq	Cellular Indexing of Transcriptomes and Epitopes by sequencing
CMV	Cytomegalovirus
CNN	Convolutional neural network
COVID-19	Coronavirus Disease 2019
CTL	Cytotoxic T lymphocytes
CX3CR1	CX3C motif chemokine receptor 1. Also known as the fractalkine receptor or G-protein coupled receptor 13
CXCL9	Chemokine ligand 9. Plays role to induce chemotaxis, promote differentiation and multiplication of leukocytes, and cause tissue extravasation.
DMF4	Human anti-MART-1 T-cell receptor 4
DMF5	Human anti-MART-1 T-cell receptor 5
DNA	Deoxyribonucleic acid
DNN	Deep neural network
EBV	Epstein–Barr virus
ELISA	Enzyme-linked immunosorbent assay
ELISPOT	Enzyme-linked immunosorbent spot
FN	False negative
FP	False positive
GBM	Glioblastoma
GLIPH	Grouping of lymphocyte interactions by paratope hotspots
GLUT4	Glucose transporter type 4. Insulin-regulated glucose transporter found primarily in adipose tissues and striated muscle (skeletal and cardiac)
GZM	Granzyme
HAVCR2	Hepatitis A virus cellular receptor 2
HCFC1	HosT-cell Factor C1. Involved in control of the cell cycle as well as having regulatory roles in a multitude of processes related to transcription
HCFC1	HosT-cell Factor C1
HCV	Hepatitis C
HIV	Human immunodeficiency virus
HLA	Human Leukocyte Antigen
HSC	Hematopoietic stem cells

Abbreviation	Descriptions
HTR1B	5-Hydroxytryptamine (G-protein coupled receptor for serotonin) Receptor 1B
IAV	Influenza A virus
ICB	Immune checkpoint blockade
ICS	Intracellular cytokine staining
IEDB	Immune Epitope Database
IEL	Intraepithelial lymphocyte
IFN	Interferon
IL7R	Interleukin-7 receptor
ITGAE	Integrin Subunit Alpha E
LSTM	Long short-term memory
MAGE	Melanoma Antigen Gene
MART	Melanoma antigen recognized by T-cells
MERS	Middle East Respiratory Syndrome
MHC	Major histocompatibility complex
MLP	Multilayer Perceptron
MT	Mutant
NLP	Natural language processing
OHE	One hot encoding
OOD	Out of distribution
ORF	Open reading frame
PAM	Protospacer adjacent motif
PBMC	Human peripheral blood mononuclear cells
PD1	Programmed cell death protein 1
PDB	Protein Data Bank
PLM	Pretrained language models
PRF1	Perforin-1
PRXL2A	Redox-regulatory protein FAM213A
PSSM	Position specific scoring matrices
PTPC1	Protein Tyrosine Phosphatase Domain Containing 1
PTPDC1	Protein Tyrosine Phosphatase Domain Containing 1
PWM	Position weight matrix
PXL2A	Peroxiredoxin-like 2 activated in M-CSF stimulated monocyte. Enables antioxidant activity
RAG	Recombination-activating gene
RI	Relative immunogenicity
RMSD	Root mean square deviation
RMSE	Root mean square error
RNBP6	RAN Binding Protein 6
ROC-AUC	Receiver operating characteristic area under the curve
RSAT	Relative Similarity to Autoantigens or Tumor-associated Antigens
S1PR1	Sphingosine-1-Phosphate Receptor 1

Abbreviation	Descriptions
SLC2A4	Glucose transporter type 4
TBX21	T-Box Transcription Factor 21
TCR	T-cell receptor
TNF	Tumor necrosis factor
TOR1B	Torsin Family 1 Member B
TRAP	T-cell recognition potential of HLA-I presented peptides
TRM	Tissue resident memory
TUNAPASTA	Tumor neoantigen prediction, assembly and tabulation algorithm
UC	Ulcerative colitis
UMAP	Uniform Manifold Approximation and Projection for Dimension Reduction
WT	Wild-type
ZAFFI	Zlab affinity enhancement; an algorithm to predict the effect of point mutations on binding affinity of TCRs.
ZEB1	Zinc Finger E-Box Binding Homeobox 1
ZNF683	Zinc Finger Protein 683
ZP4	Zona Pellucida Glycoprotein 4

Chapter 1. Introduction

In part published in

Lee CH, Salio M, Napolitani G, Ogg G, Simmons A, Koohy H. Predicting Cross-Reactivity and Antigen Specificity of T-cell Receptors. Front. Immunol. 11, (2020).

Innate and adaptive immune systems orchestrate to maintain healthy immune functions. The adaptive immunity provides long-lasting humoral and cellular immune responses mediated by B and T-cells, respectively. Among these, T-cells have unique specificity to antigens conferred by their T-cell receptors (TCRs) and play a critical role in various pathologies. In infectious disease, pathogen-specific T-cells coordinate defence mechanisms, whereas in cancer, tumor-specific T-cells are involved in an effective anti-tumor response. In addition to pathological contexts, antigen-specific T-cells also mediate the development of immunological memory (e.g., after vaccination) and the maintenance of tolerance to self-antigens. Because of their importance, identifying the antigens that trigger the T-cell response has been of great interest for developing clinical therapeutics.

Development of T-cell specificity

T-cells develop in the thymus where they mature and differentiate into distinct subsets that play critical roles in controlling and shaping the immune response[1] (Figure 1.1). The earliest T-cells in the thymus lack both CD4 and CD8 co-receptors (termed double-negatives). Thymocytes upregulate recombination genes in order to undergo V(D)J recombination[2], a process of DNA rearrangement and random chain pairing that results in the formation of α- and β-chains for $\alpha\beta$ TCR[3]. Following the rearrangements, positive and negative thymic selection work to train an effective immune system that responds efficiently to pathogens or cancer cells while ignoring or tolerating self[1,4].

The CD4+CD8+ thymocytes migrate into the thymic cortex, where they are presented with self-antigens[5]. Only thymocytes that interact strongly with major histocompatibility complex-I (MHC-I) or MHC-II receive the survival signal, while those that do not interact strongly enough die[6]. The double-positives that interact with MHC-I mature into 'killer' CD8+ cells, which initiate direct cytotoxic responses against infected or cancerous cells[1]. On the other hand, those that interact

with MHC-II molecules become 'helper' CD4+ T-cells, which orchestrate the adaptive immune system by producing pro-inflammatory, chemotactic, as well as immunoprotective cytokines[7] that dampen the immune response[8]. This process ensures that the surviving thymocytes can respond to MHC molecules to serve important immune functions[9].

As this process does not screen for thymocytes that may cause autoimmunity, the following negative selection process ensures that potentially self-reactive cells are eliminated[10]. Thymocytes that bind too strongly to self-antigens receive an apoptotic signal that leads to cell death[11], while some others become regulatory T (Treg) cells[12]. The negative selection is an important component of central tolerance as it prevents the formation of self-reactive T-cells that can cause autoimmune diseases in the host. Thus, typical naive T-cells that leave the thymus are self-restricted, self-tolerant, and single-positive T-cells.

T-cells then use their antigen receptors to recognise peptides and lipids derived from invading organisms presented by the host MHC or lipid-presenting molecules, such as CD1 or MR1. With exceptions of cross-presentation, endogenous or intracellular antigens are presented on MHC-I molecules expressed by all nucleated cells, while exogenous or extracellular antigens are presented on MHC-II molecules expressed by antigen presenting cells (APCs), such as dendritic cells, macrophages and B cells[13]. The peptides presented by MHC molecules are then recognised by T-cells that are specific to pMHCs (i.e. antigen-specific T-cells)[14].

In addition to the TCR:pMHC interaction (i.e. the first signal), antigen-naïve T-cells require appropriate co-stimulation (i.e. the second signal) and immunostimulatory cytokine environment at the time of antigen encounter in order to expand and differentiate into memory and effector T-cells. Because self-peptides are rarely presented with suitable co-stimulation, this mechanism prevents inappropriate responses to self. After activation, T-cells divide rapidly and secrete cytokines that regulate or assist the immune response.

As a result of the sophisticated development and differentiation process, a unique feature of T-cells is their ability to discriminate between healthy and abnormal (e.g. infected or cancerous) cells in the body[15], known as antigen discrimination. Healthy cells typically express a large number of self-derived pMHC on their surface, which are generally ignored by T-cells. However, when the same cells contain even minute amounts of pathogen-derived pMHC, T-cells become activated and initiate an immune response.

Antigen-specific T-cells for basic, clinical and translational research

T-cells' ability to induce antigen-directed cytotoxicity has proven instrumental in fighting diseases, as evidenced by checkpoint blockade immunotherapy, adoptive cellular therapy and cancer vaccinology. For example, epitope-based vaccination can trigger an HLA-restricted T-cell immune response while avoiding potentially toxic or immune-suppressive responses[16–20]. Such approaches have shown to offer the benefits of low cost, greater safety and a broader spectrum of target diseases ranging from virus infections to Alzheimer's disease[22,23]. Given the breadth of clinical applications, it is critical to fully comprehend the rules governing T-cell recognition of their cognate pMHC complexes for safer therapeutic development. However, one key under-explored

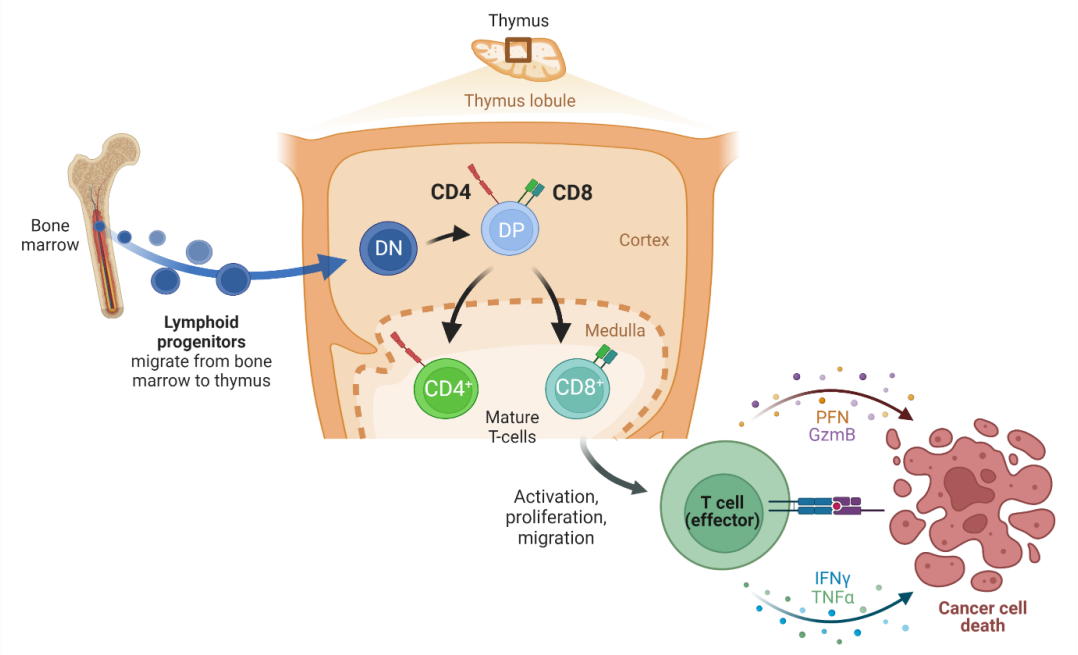


Figure 1.1: Development of functional antigen-specific T-cells. T-cells originated from bone marrow migrate to the thymus where they mature into antigen-specific T-cells. Thymocytes upregulate expression of recombination genes to form α- and β-chains, and $\alpha\beta$TCR+CD4+CD8+ (DP) thymocytes undergo positive and negative selection to become self-restricted, self-tolerant and self-positive T-cells. When naïve T-cells are exposed to an antigen signal via the TCR, a co-stimulatory signal, and a cytokine signal, they differentiate into various effector T cell subsets, such as CD8+ T cells (killer cells) or CD4+ T cells (helper cells) that trigger antiviral, antitumor or autoimmune responses.

area is the ability to identify T-cell targets i.e. whether viral or cancer-associated antigens are likely to be recognised by T-cells and elicit an effective immune response.

The significance of identifying T-cell targets has never been clearer than during the ongoing COVID-19 pandemic. Characterising T-cell epitopes was essential for understanding disease onset, pathogen clearance, heterogeneous response to natural infection or vaccination, as well as developing vaccines or treatments[24–27]. Because identifying T-cell epitopes is at the heart of understanding disease pathology and developing therapeutics, the community has made significant efforts toward creating robust systems for screening or predicting T-cell epitopes.

Despite recent advances, identifying a reliable target for such T-cell mediated therapies remains a challenge[28]. Foremost, finding a target that is both safe and effective remains a difficult task[28,29]. The common targets of T-cell based cancer immunotherapy are tumor-specific antigens (e.g. neoantigens and viral antigens) or tumor associated antigens (e.g. cancer/testis (CT), overexpressed or differentiation antigens). While high specificity is the most important consideration when choosing an antigen, potential off-target effects and dose-limiting toxicity must also be taken into account. Furthermore, the expression of cancer-related antigens within tumour cells may vary, allowing some tumour cells to evade antigen-targeted therapies[30].

Therefore, a comprehensive understanding in the landscape of MHC-presented peptides recognised by T-cells is vital to design safe and effective therapeutics. In recent years, MHC binding predictions have advanced to achieve high performance for common HLA-I alleles[31]. However, because MHC presentation is required but not sufficient to elicit an immune response, identifying T-cell epitopes from a pool of MHC ligands is of great interest. There have been experimental and computational efforts to characterise and predict T-cell epitopes in order to expand the list of potential immune targets[32].

Experimental strategies of T-cell epitope screening

Advances in experimental approaches have aided in the identification and characterisation of T-cell epitopes. Table 1.1 provides examples of various strategies employed for shortlisting T-cell antigens.

Briefly, a robust method for epitope identification is using pools of overlapping peptides or combinatorial libraries to systematically screen for T-cell epitopes. A large pool of short peptides (e.g., 9-10aa for MHC-I) are synthesized[33] and tested against a small number of TCRs (e.g. 4-8 TCRs per experiment on yeast display system). Recent advancement in cell-based platforms, such as signalling and antigen-presenting bifunctional receptors (SABRs)[34] and T-Scan[35], have also allowed the use of T-cell clones or TCR-transduced T-cells without the need to expand soluble TCRs. Moreover, the immunopeptidomics approach based on mass spectrometry enables the identification of naturally occurring candidate epitopes on HLA molecules for subsequent immunogenicity testing.

The systematic epitope screening yields a highly curated and biologically relevant list of peptides for *in vitro* cytokine detection. These assays are useful for reducing large peptide pools to

individual peptides and determining the minimum epitope. Also, the specific peptide-HLA complexes can be synthesized for tetramer assays, allowing for multiparameter phenotyping, functional characterisation, and TCR sequencing. Single-cell approaches have enabled the profiling of antigen-specific T-cells against >100 DNA barcode-labeled MHC multimers[36,37]. These experimental strategies can shed light on the specificity, function, and quality of epitopes that elicit T-cell responses, as well as how these can be harnessed for stronger and longer-lasting therapeutic efficacy.

The peptides characterized by T-cell assays are deposited in Immune Epitope Database and Analysis Resource (IEDB)[38]. The IEDB is a freely available resource that contains a collection of experimentally measured epitopes and a suite of tools for analysing and predicting epitopes. Epitopes with cognate TCRs have also been deposited in databases such as VDJdb [39] and McPAS-TCR [40]. To date, the VDJdb contains > 60,000 TCR specificity records linked with their epitope and MHC, and McPAS-TCR holds > 5000 TCRs associated with pathogenic conditions e.g. pathogen infection, cancer and autoimmunity. As will be described in the following sections, the information provided in the repository serves as the foundation for building computational tools for better prediction of T-cell epitopes from emerging pathogen or tumor tissues.

Table 1.1: Experimental strategies of T-cell epitope screening

Strategy	**Advantages**	**Drawbacks**
Peptide screening using combinatorial peptide libraries E.g. Yeast display system, Signaling and antigen-presenting bifunctional receptors (SABRs) and T-Scan	• Unbiased, rapid and high-throughput screening using pools of overlapping peptides or combinatorial libraries • Able to use T-cell clones or TCR-transduced T-cells to avoid the need for expanding soluble TCR[34,35,41,42].	• Need to prepare peptide libraries for each experiment • Limited # of TCRs can be tested at once • Most unable to screen for epitopes outside scope of peptide libraries e.g. post-translational modifications

Table 1.1: Experimental strategies of T-cell epitope screening

Strategy	Advantages	Drawbacks
Immunopeptidomics by mass spectrometry	• Screen for naturally presented candidate epitopes on HLA molecules • Allow identification of ligands with unpredictable binding modes (e.g., overhanging or extended HLA-I ligands)[43,44], as well as post-translational modifications (e.g., glycosylation, phosphorylation, and deamination) and those from non-canonical sources (e.g., alternative reading frames)	• Previous knowledge required for mass spectrometry analysis of eluted peptides • Search engines for matching MS/MS spectrum may be biased towards tryptic peptides, posing challenge for identification of HLA ligands generated by unspecific protease cleavage[45]. • Potential losses of low-affinity ligands
Epitope characterisation by *in vitro* interrogation E.g. Cytokine detection using ICS, ELISpot and Tetramer assays	• Useful for narrowing down large peptide pools to individual peptides and determine minimal epitope • ICS assays allow for the measurement of multiple molecules (e.g. IFN-γ, TNF and IL-2), and delineation of cell phenotypes and functionality. ELISpot is more sensitive at measuring low-level responses[46]. • Tetramer assays enable profiling of antigen-specific T-cells against >100 DNA barcode-labelled MHC multimers[36,37].	• Relatively low-throughput • Require optimization of each readouts

Limitations of experimental approaches

Although recent experimental approaches provide greater flexibility in screening T-cell epitopes and have made significant contributions to our current understanding of T-cell immunology, there remain limitations. These include a) the small number of TCRs that can be tested against

peptide libraries in a single experiment, b) the need to prepare new peptide libraries for each analysis of pMHC specificity, c) often the requirement to generate individual recombinant TCR, T-cell clones, or reporter cells expressing TCR for screening, and d) the high false positive and negative rates. In fact, because different assays and readouts have different levels of specificity and sensitivity, different thresholds must be used to define positivity, which may introduce an additional layer of noise and bias, resulting in less reproducibility in other contexts and systems. Furthermore, differences in preparative, experimental, analytical and bioinformatics workflows may have an impact on the final readouts, resulting in a different list of immunopeptidome or T-cell epitopes.

Moreover, while experimental methods for identifying MHC-I ligands have matured, presentation does not always imply T-cell recognition and activation. Indeed, the fraction of epitopes (i.e. immunogenic peptides) across different species is currently unknown and appears variable. For example, Croft *et al.* reported that out of ~170 vaccinia virus MHC-I presented peptides eluted from mass spectrometry, ~40% were found immunogenic across more than half of the mice screened[47]. On the other hand, Hensen *et al.* found that only 4/54 influenza A virus and 6/41 influenza B virus peptides were immunogenic[48], indicating that HLA binding is necessary but not sufficient for immunogenicity and that subsequent T-cell assays will be required to identify T-cell epitopes.

Immunoinformatic identification of CD8+ T-cell epitopes

Given the limitations of these functional assays, *in silico* predictions of MHC binding and T-cell recognition have been of great interest to shortlist potential candidates. The increasing wealth of experimental and sequencing data has facilitated advancements of *in silico* approaches to analysing, modelling and predicting TCR:pMHC interaction.

A plethora of computational models has been developed to aid in the prediction of T-cell targets and to foster a better understanding of the characteristics underpinning peptide immunogenicity[49–54]. These models utilized features, such as amino acids at contact position, hydrophobicity, peptide-MHC binding affinity and stability, dissimilarity to self-proteome and foreignness score as correlates to T-cell recognition and reported variable and often poor sensitivity or specificity (Details about existing models are described in Chapter 5). Many existing models performed poorly on emerging pathogens[55,56], and progress in improving the performance appears to be incremental, suggesting that there are still remaining challenges in predicting T-cell epitopes.

Remaining challenges in epitope prediction

An effective antigen-specific CD8+ T-cell response to exogenous pathogens or endogenous threats relies on tightly regulated processing and presentation of antigenic peptides by class I MHCs, followed by recognition of the peptide-MHC (pMHC) by cognate CD8+ T-cells. Therefore, immunogenic peptides encompass attributes associated with features underlying MHC presentation *and* T-cell recognition[52]. Among these, it has been demonstrated that features

attributed to MHC presentation are more prominent than those attributed to TCR recognition, with strongly conserved motifs at anchor positions being one such feature[57,58]. Indeed, recent cutting-edge models[59,60], such as the widely used NetMHCpan[60], have demonstrated impressive performance in predicting MHC presentation on certain alleles. On the other hand, T-cell recognition features are less conserved owing to TCR diversity. In addition, the scarcity of peptides tested by T-cell assays, as well as the lack of true negative datasets (i.e. presented but not recognized by T-cells), hampered our understanding of the properties underlying T-cell recognition.

First, the number of extensively characterised epitopes for cognate TCRs remains limited. To understand the breadth of available epitope and TCR pairs, we retrieved epitope TCR pairs from IEDB, VDJdb, McPAS-TCR, TBAdb[61], 10X dextramer datasets[62], and from epitope screening studies[34,41,63] that employed signalling and antigen-presenting bifunctional receptors (SABR), trogocytosis and yeast display system. After removing duplicate entries, there were 1,375 unique epitopes, 157,535 CDR3βs, and 167,076 epitopes:CDR3β pairs (Figure 1.2A).

Of these epitopes, 594, 299, 38 were presented by HLA-A*02:01, Class-I and DRB1*15:01, respectively, with the remaining HLAs being largely under-represented. In fact, ~40% of the epitopes were presented by HLA-A*02:01 and ~24% were labelled with non-specific HLAs e.g. Class-I. TCRs have been shown to engage with both peptide and MHC concurrently[64], implying that MHC specificity plays a role in T cell antigen recognition. The current dataset, however, is evidently skewed towards highly prevalent HLA alleles, posing a challenge to develop a robust model that encompasses a diverse breadth of HLA alleles.

Out of the 1,375 epitopes, 581 epitopes had one cognate TCR, while 216 and 158 epitopes had $\geq$ 50 and $\geq$ 100 cognate TCRs, respectively (Figure 1.2B). Notably, cross-reactive and cross-species TCRs accounted for only 9.5% and 5.7% of TCRs in the database, respectively (Figure 1.2C), with 44/157,535 (0.027%) TCRs recognising $\geq$ 10 cognate epitopes (Figure 1.2D). With the theoretical estimate that any single T-cell can recognise up to 10^4-10^6 epitopes for an effective immune response[65], this may represent the extent of epitope under-sampling we have on our hands.

Moreover, immunogenicity is influenced by a variety of other factors, including co-stimulation, proliferation, and cytotoxicity[66]. While these factors can influence the magnitude and shape of a T-cell response[67], they are not yet available to provide context in the prediction.

Third, the different sensitivity of experimental assays used to evaluate T-cell responses can result in noisy data, which is often exacerbated by the lack of 'true negative' peptides. Here, the true negatives would be MHC-ligands that have been tested against a pool of T-cells and have been found to be negative in all conditions. However, the diversity of T-cell repertoires and the varying levels of sensitivity of the assays call defining the true negative into question. Taken together, identifying immunogenic peptides has proven to be more challenging than identifying MHC ligands.

Lastly, although existing models reported high accuracies by 10-fold cross-validation, there was

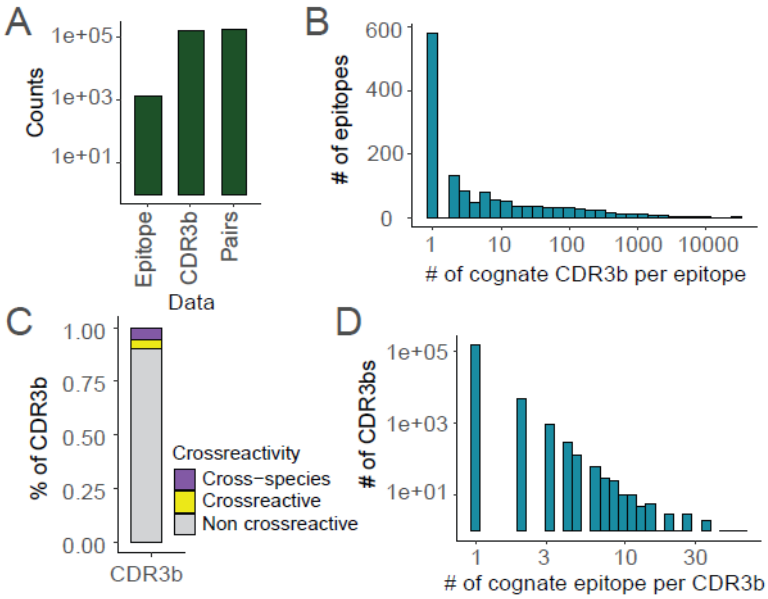


Figure 1.2: Cross-reactive and cross-species TCRs. A. Statistics on the number of unique epitopes, CDR3bs and epitope:CDR3β pairs in the database. B. Histogram on a number of cognate TCRs deposited per epitope. C. Proportion of cross-reactive and cross-species (i.e. TCRs recognizing epitopes derived from multiple species) TCRs. D. Histogram on the number of cognate epitopes for each TCR.

a lack of systematic and unbiased evaluation of their performance in predicting CD8+ T-cell targets from emerging pathogens, such as SARS-CoV-2, and cancer neoepitopes. Therefore, there was limited understanding in the community about whether these models can achieve reliable performance in identifying pathogenic or cancer neoepitopes.

Chapter 2. Outline of thesis

This thesis aims to foster a deeper understanding of the rules that govern antigens' ability to elicit immune responses when they are recognised by T-cells. It covers a comprehensive landscape of features underlying peptide immunogenicity, novel computational strategies for predicting CD8+ T-cell epitopes from HLA-I presented cancer or pathogenic peptides, and demonstrate how their application can provide insights into disease mechanisms or aid in therapeutic development.

The first chapter (Introduction) provided background about the antigen-specific nature of T-cell responses and highlighted the importance of identifying T-cell targets for basic and translational research. It also covered the current methodologies for screening, characterising and predicting CD8+ T-cell epitopes, as well as remaining challenges in epitope prediction.

Chapter 3 and 4 describe case studies of autoinflammatory (Chapter 3) and infectious (Chapter 4) diseases to demonstrate the compelling need to better understand the antigen-specific T-cell response, which requires the ability to identify T-cell targets.

In the context of autoinflammatory disease, Chapter 3 presents a multi-modal single-cell study to gain further insights into the molecular mechanisms of checkpoint inhibitor-induced colitis (CC) in comparison to ulcerative colitis (UC) and healthy individuals. By focusing on immune compartments, especially on T-cells, we observed pathology-specific enrichment of activated and clonally expanded CD8+ T-cells, albeit with different phenotypic signatures. Moreover, distinct subpopulations of cytotoxic T-cells were migrating from circulating blood to the site of inflammation in a pathology-specific manner. However, these technologies were not able to pinpoint the antigen-specificity of these T-cell populations. This illustrated that the current multi-modal technologies have limitations in identifying the source antigen, and thus computational approaches may offer solutions to this unmet need.

Next, Chapter 4 uses COVID-19 as a representation of infectious disease to demonstrate how the ability to predict CD8+ T-cell targets has facilitated vaccine development and understanding of disease pathology. In early 2020, when access to samples from infected individuals was limited, I conducted *in silico* identification of CD8+ T-cell targets to guide toward the best vaccine targets against SARS-CoV-2. Moreover, the study shed light on the heterogeneous outcome of natural infection or vaccination by assessing the extent to which pre-exposure to other coronavirus strains may confer T-cell cross-immunity, as well as the role of patient genetics in driving variations in T-cell response.

The above-mentioned examples highlight the critical impacts of the ability to predict T-cell targets. Recently, a number of computational approaches for predicting CD8+ T-cell epitopes have emerged. These models, however, were not systematically compared. Therefore, Chapter 5 presents an unbiased evaluation of existing algorithms for predicting CD8+ T-cell epitopes against SARS-CoV-2 and cancer neoepitopes, and reports poor performance in both settings. These findings inspired me to develop a robust model capable of accurately predicting CD8+ T-cell epitopes in the context of treating cancer, inflammatory, and infectious diseases.

To build a robust model, I first investigated sources of poor performance, especially in emerging pathogens (Chapter 6). This resulted in findings that a) known hallmarks of immunogenicity, such as MHC binding affinity and hydrophobicity, may not be universally applicable to pathogenic and cancer epitopes, b) peptides derived from different organisms may not share common features predictive of immunogenicity, and c) high performance is almost entirely misattributed to HLA bias inherent in training and validation data.

Based on these findings, Chapter 7 presents TRAP (T-cell recognition potential of HLA-I presented peptides), a deep learning-based model that addresses the limitations and effectively captures T-cell recognition motifs from HLA-I presented pathogenic or self-peptides. Novel strategies were implemented, such as a) using transfer learning to extract amino acid embeddings from pre-trained large-scale protein language models, b) capturing T-cell recognition motifs with a deep learning-based model, c) building separate models for pathogenic and self-peptides to account for divergent immunogenicity-related features, and d) avoiding low confidence predictions. The TRAP was then used to assess the immune escape potential of emerging pathogens as well as the impact of deleterious mutations that may give rise to variants of concern.

Taken together, this thesis presents a systems immunology approach to understanding T-cell dynamics in various pathologies and predicting the cognate antigens driving the response. The studies presented here will enable more accurate identification of CD8+ T-cell epitope, which will aid in basic, clinical and translational research for a wide range of therapeutic applications.

Chapter 3. Divergent T-cell cytotoxicity programs between ulcerative and checkpoint inhibitor-induced colitis

To be published in

Gupta T*, Antanaviciute A*, Aulicino A*, **Lee CH***, Babu RO*, D Fawkner-Corbett, AS Geros, E Bridges, E Fryer, NK Alham, M Jagielowicz, K Parikh, H Chen, H Koohy, A Simmons. Spatially Organised Multicellular Immune Niches Drive Barrier Pathology in Immune Checkpoint Colitis. *Manuscript in preparation.* *These authors contributed equally

Introduction

Immune checkpoint blockade (ICB) therapy has revolutionized cancer treatments with some patients experiencing remarkable clinical efficacy. However, 35% to 50% of ICIs patients experience gastrointestinal (GI) immune-related adverse events (irAEs) such as checkpoint inhibitor-induced colitis (CC)[68,69], but there are no biomarkers or treatments that predict or effectively treat these side effects.

Currently, CC is empirically treated with the same drugs as ulcerative colitis (UC) despite a failure rate of up to 40%[70,71]. However, given that UC is a complex autoinflammatory disorder involving colonic barrier dysfunction, genetic predisposition and immune dysregulation[72] in contrast to CC, it is unclear whether colitis arising from different aetiologies leads to identical or disparate cellular pathology. Elucidating these mechanisms of tissue destruction occurring in either disease is required to develop precision medical approaches and refine therapeutic strategies. This knowledge is essential for the treatment of CC, where it is desirable to inhibit off-target immune activation but preserve anti-tumour responses.

Recent advances in high throughput single-cell RNA sequencing have transformed our understanding of the cellular anatomy of the colon in development, health and UC[73–77]. This approach has revealed differential cellular composition and cross-talk, as well as transcriptional regulatory

networks as hallmarks of barrier breakdown, mesenchymal remodelling and inflammatory immune infiltration in UC[73–77]. On the other hand, the contribution of the epithelial barrier, surrounding niche, circulating immune populations and mucosal lymphoid structures in orchestrating inflammation in CC remains unknown. Single-cell analysis in CC revealed that proliferative cytotoxic tissue resident CD8s are the most over-represented and thus most likely colitigenic T-cell population[78,79]. However, it is unclear how these cells cause or initiate tissue damage in CC.

Here, single-cell transcriptomics was combined with T-cell receptor (TCR) profiling to compare colonic epithelial, mesenchymal and immune cell interactions between CC and UC, with a focus on T-cell compartments. This resource documents the commonalities and differences between these two conditions, emphasising distinct T-cell dynamics that contribute to altered local tissue microenvironment and perpetuation of colitis.

Results

Unbiased scRNA-Seq reveal analogous and divergent T-cell state perturbation in checkpoint induced and ulcerative colitis

To conduct unbiased analysis of colitis, a multi-modal (RNA-Seq, VDJ-Seq, CITE-Seq) single-cell reference atlas of matched colonic and PBMC samples with active UC (UC_I) and CC colitis (CC_I), as well as healthy (HC), UC uninvolved (UC_NI) and immunotherapy-treated patients who did not go on to develop colitis (CC_NI) controls was generated, capturing a total of 155,421 cells from epithelial, stromal and immune compartments from 64 donors (Figure 3.1A). Following data processing, quality control and batch correction, unsupervised clustering identified the majority of the cell types anticipated in the colon (Figure 3.1B, Supplementary Table 1-3), which were annotated using cluster marker genes with reference to previously published single-cell resources[74–78,80,81] (Methods).

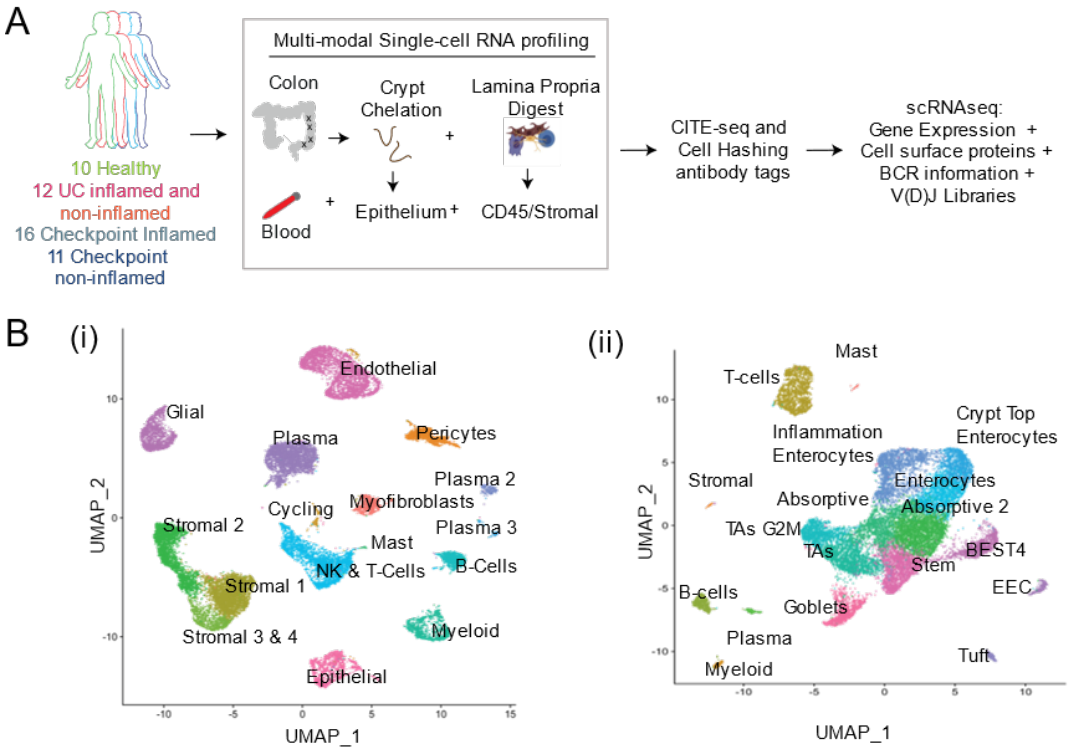


Figure 3.1: Multi-modal single-cell atlas of Healthy, Ulcerative colitis and Checkpoint inhibitor-induced colitis. A. Schematic overview of the experimental design of multi-modal single-cell RNA sequencing. B. UMAP embedding visualising scRNA-Seq clusters of cell populations obtained from stromal/CD45+ (i) cell isolation strategy and epithelial crypt dissociation (ii)

Principal component analysis of sample pseudobulk of epithelial, stromal and immune cell populations showed a clear separation between inflamed and non-inflamed conditions, while differences between CC and UC inflamed samples were a lot less pronounced (Figure 3.2A-C). The differential gene expression and cell abundance analysis also highlighted the global convergence

of CC and UC inflammation towards very similar transcriptomic and cellular responses, such as interferon gamma (IFNg) response and antigen processing and presentation pathways (Figure 3.2D). This analysis emphasized the importance of studying inflammatory states collectively in order to distinguish between convergent, non-specific inflammation responses in tissue and pathology-specific alterations.

To examine the direct targets of checkpoint inhibitor therapy, the expression of *CTLA4* and *PD1* was analysed, and it was found that they were exclusively expressed by T-cell sub-populations in our data [16](Figure 3.3). Moreover, previous scRNA-seq analysis showed accumulation of CD8 tissue resident memory (TRM) T-cells in immune infiltrates of ICB colitis patients that differentiated to a cytotoxic phenotype[78]. Therefore, additional 40,115 $CD3^+$ cells from CC-inflamed, UC-inflamed and HC donor biopsies and paired PBMCs were sequenced to analyse T-cell trafficking between colon and periphery, identify shared and divergent T-cell responses between CC and UC, and characterize protein expression by CITE-Seq analysis.

This data was integrated with unsorted T-cells from prior experiments, for a total of 41,144 $CD3^+$ cells in biopsies and 36,176 from PBMCs (Methods). Following integrated cluster analysis and visualisation (Figure 3.4A-B), 21 distinct T-cell states were identified in biopsy samples and 16 in PBMCs, which were subsequently annotated based on cluster protein and mRNA markers and reconstructed transcription factor regulatory networks (Methods) with reference to previously reported reference datasets (Figure 3.4C)[77,78,80,81]

Aligned with previous observations, one of the major drivers of transcriptomic differences was inflammation status regardless of pathology in colonic tissues, with cells from inflamed and non-inflamed samples forming separate clusters. (Figure 3.5A-D). In both CC and UC inflamed samples, a significant proportional increase in Th17/Tc17 cells, exhausted $CD8^+$ cells ($CD8^+$ $HAVCR2^+$) and actively cycling cells (Figure 3.5E) was observed, with notable shifts in other populations, including Tregs (Figure 3.6).

Divergent T-cell state perturbations in CC and UC

Although cells from inflamed tissues shared many similarities, different T-cell dynamics in CC and UC were also found. First, there was an increase in follicular T-cells (Tfh), $GZMK^+$ effector populations, and $CD4^+$ and $CD8^+$ naïve T-cell clusters in UC but not in CC (Figure 3.6). Also, a relative depletion in TRM populations (TRM1, $ZNF683^+$ cells) was observed in UC when compared to both HC and CC samples[79,83,84]. Lastly, while increased in both CC and UC inflamed samples, *CTLA4*/CTLA4 and PD1/*PDCD1* expressing exhausted and cycling $CD8^+$ cells were more prevalent in CC samples.

ZNF683+ TRMs enriched in checkpoint-induced compared to ulcerative colitis

As many T-cell populations described here exist on a phenotypic continuum (differentiation, activation, exhaustion), this observation was supported by a differential abundance analysis of local k-nearest neighbour (knn) graph neighbourhoods (Figure 3.7, Methods). This further

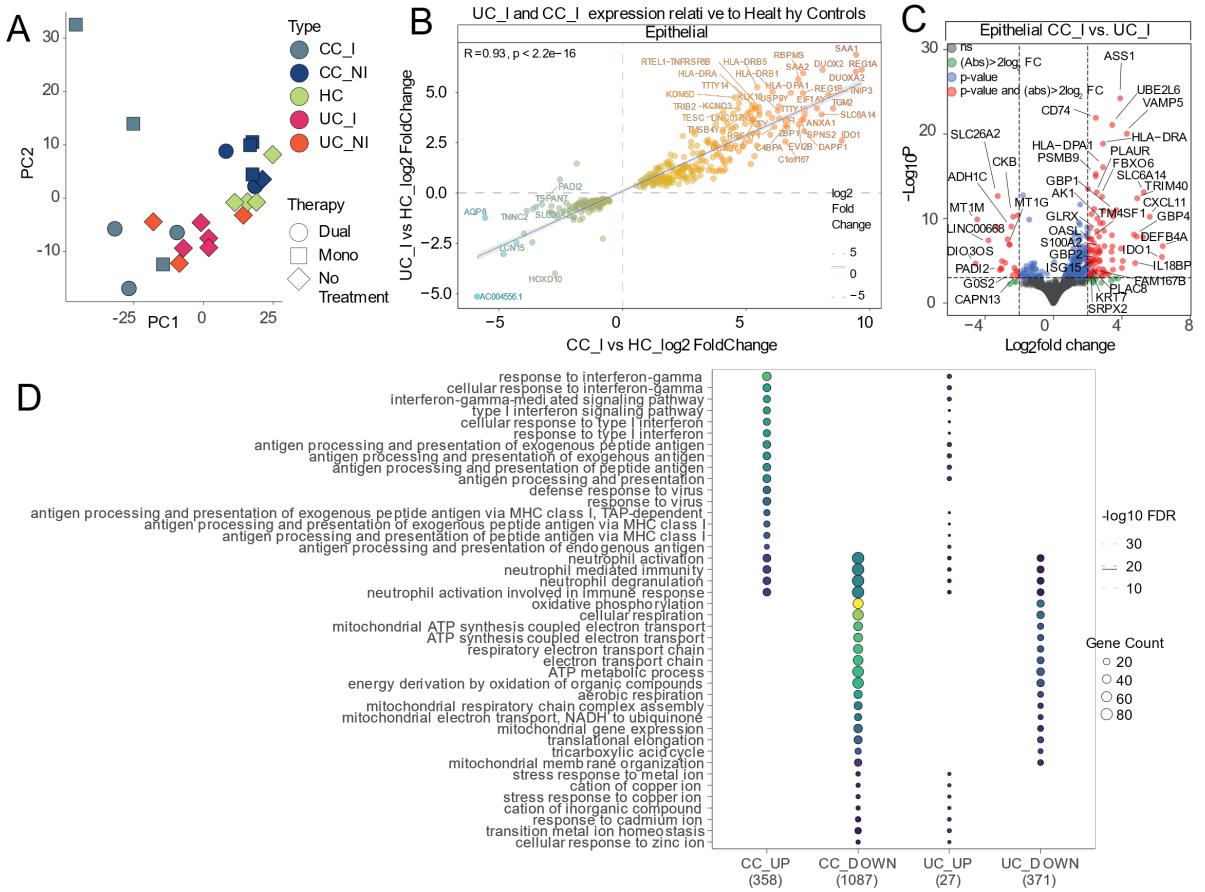


Figure 3.2: Convergence of transcriptional profiles in UC and CC. A. Principal component analysis plot of per sample, normalised scRNA-Seq pseudo-bulk of epithelial cells using the top 1000 most highly variable genes. The first two principal components are shown. B. Correlation between log 2-fold changes across all genes in UC inflamed vs healthy control samples and CC inflamed vs healthy control samples in epithelial compartment scRNA-Seq. Selected genes are labelled. C. Volcano plot showing differentially expressed genes in epithelial cells between UC inflamed and CC inflamed donors. Selected genes are labelled. D. Dotplot visualising top gene ontology terms enriched in upregulated (left, CC_UP) and downregulated (left middle, CC_DOWN) genes in CC epithelium when compared to checkpoint inhibitor non-colitis and upregulated (right middle, UC_UP) and downregulated (right, CC_DOWN) genes in UC epithelium when compared to healthy control samples.

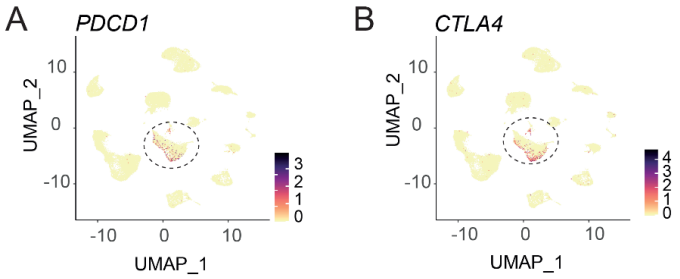


Figure 3.3: Immune checkpoint gene expression in scRNA-Seq data. UMAP overlay visualising expression of *PDCD1* (A) and *CTLA4* (B) transcript in across all stromal/immune populations.

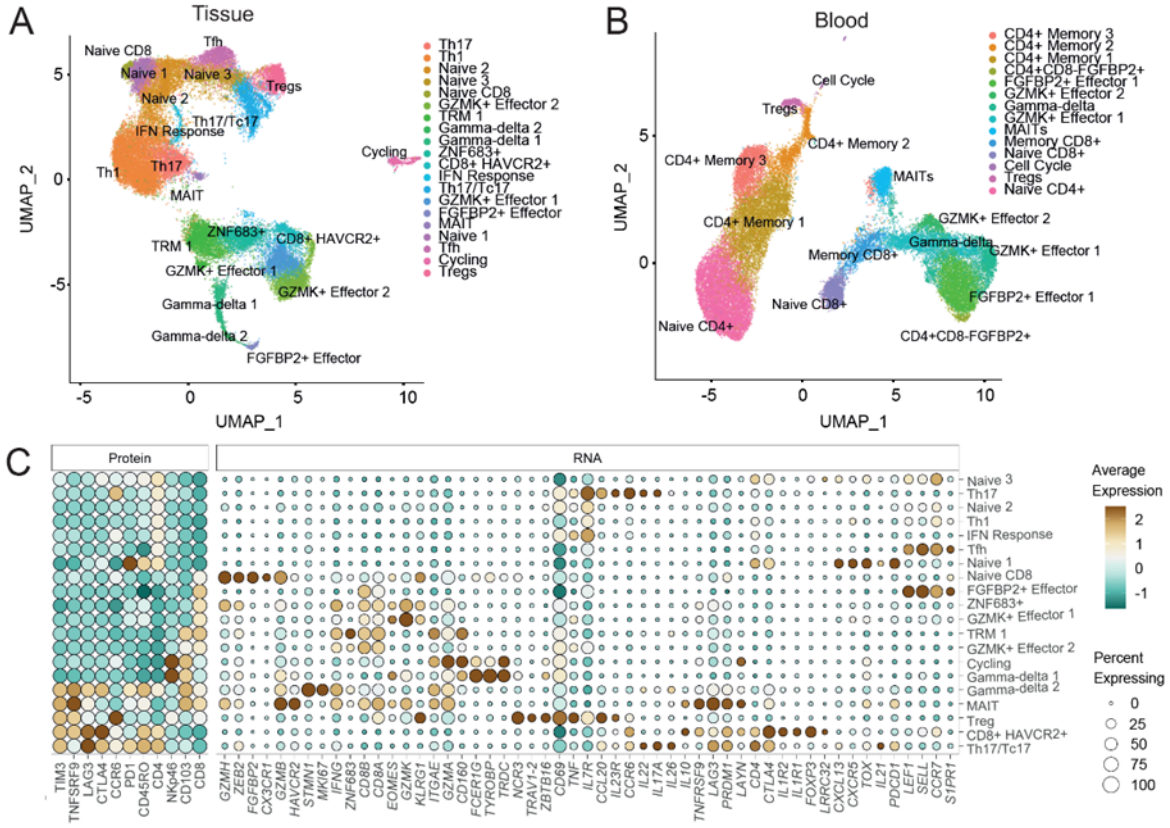


Figure 3.4: Single-cell RNA-Seq on T-cell subpopulations. A-B. UMAP embedding visualising scRNA-Seq clusters of T-cell subpopulations in UC, CC and HC samples from colon tissue biopsies (A) and PBMCs (B). C.Dotplot heatmap showing top T-cell cluster marker transcript (bottom) and CITE-Seq protein panel (top) distribution across all T-cell clusters in tissue biopsy data.

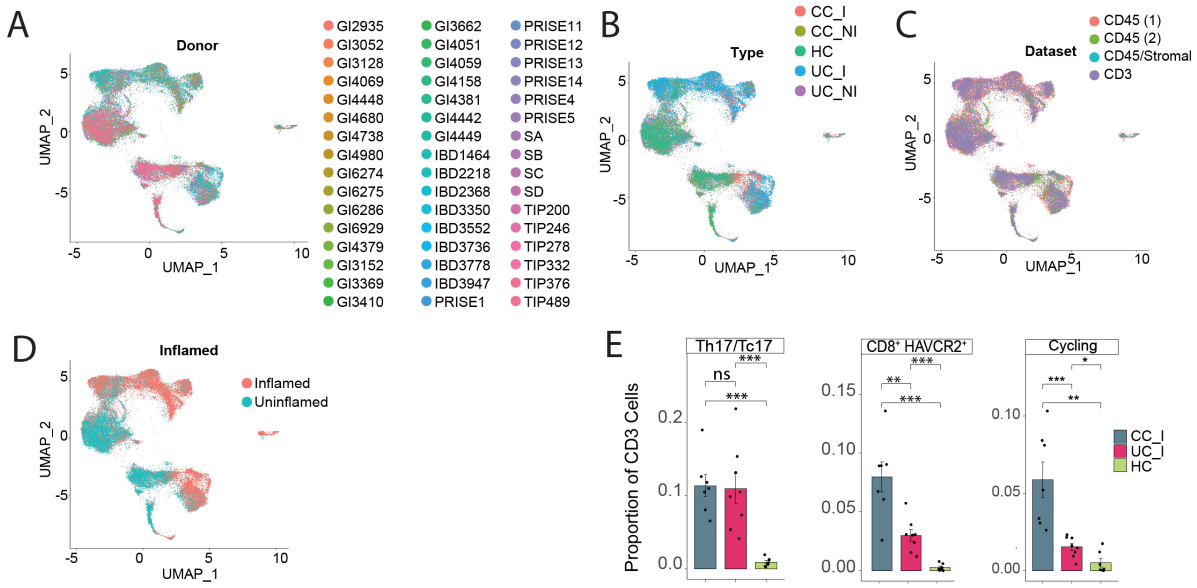


Figure 3.5: Differential T-cell responses in colonic tissue and blood in colitis. A-D. UMAP overlay visualising individual donor distribution (A), sample type (B), dataset isolation strategy (C) and sample inflammation status (D) across scRNA-seq T-cell dataset embedding from PBMC samples. E. Selected T-cell subpopulation abundance changes as a fraction of all T-cells in sample in Checkpoint-inhibitor induced colitis (CC_I), inflamed ulcerative colitis (UC_I), and healthy control (HC) samples, shown as bar plots. Wilcox rank test, p-value < 0.05 $*$; p-value < 0.01 $**$; p-value < 0.001 $***$; ns = not significant. Error bars represent standard error of the mean (SEM).

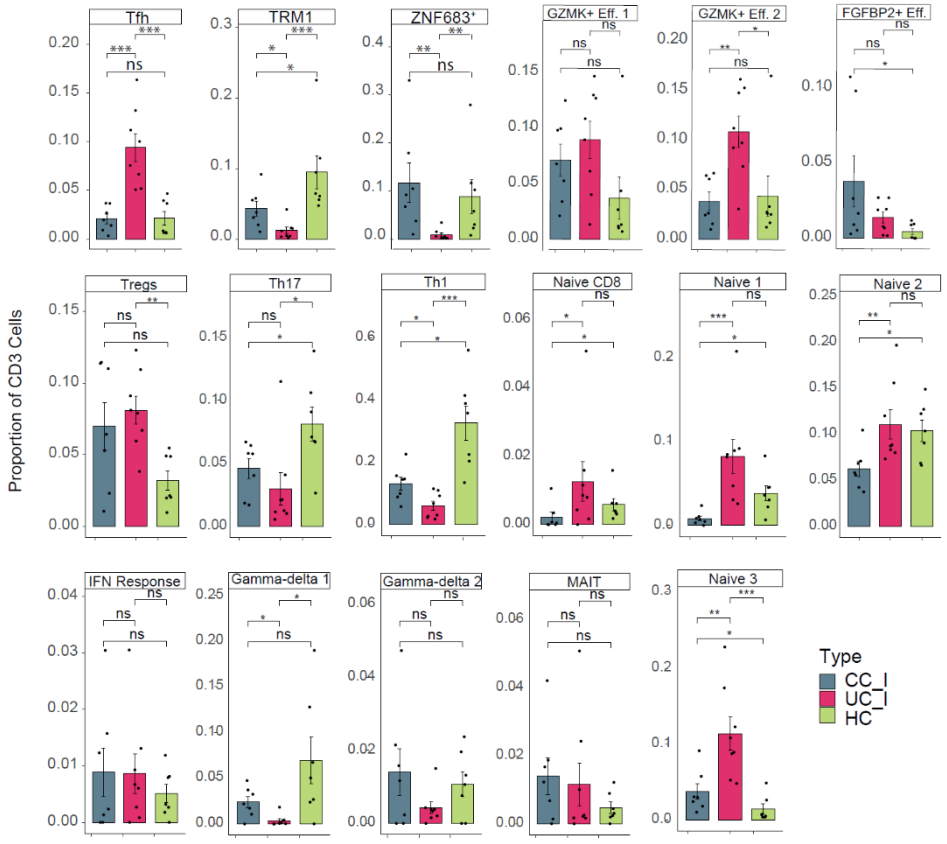


Figure 3.6: T-cell abundance in CC compared to UC. Selected T-cell subpopulation abundance changes as a fraction of all T-cells in sample in Checkpoint-inhibitor induced colitis (CC_I), inflamed ulcerative colitis (UC_I), and healthy control (HC) samples, shown as bar plots. Wilcox rank test, p-value < 0.05 $*$; p-value < 0.01 $**$; p-value < 0.001 $***$; ns = not significant. Error bars represent standard error of the mean (SEM).

highlighted ZNF683+ TRMs as one of the most enriched T-cell populations when comparing CC vs UC samples.

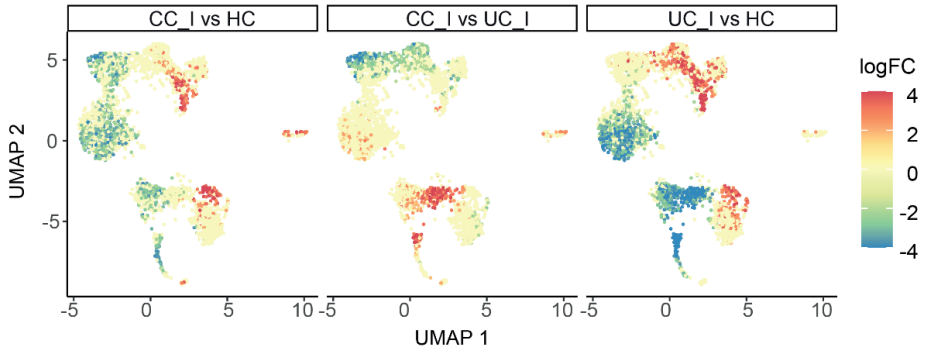


Figure 3.7: UMAP overlays visualising log fold changes of condition-specific local neighbourhood abundance differences. Each point represents a neighbourhood of 10 cells within the single-cell k-nearest neighbour graph visualised as a UMAP embedding.

There were two CD103+ TRM states (denoted as TRM 1 and ZNF683+) found using the pseudotime analysis (Figure 3.8A), which likely exist on a continuum, with cells from CC inflamed samples skewed towards the end of this trajectory. Based on mRNA and protein expression profiles and TF regulatory network analysis, ZNF683^{+} cells represent a reactivated TRM cell state. In addition to expressing CD103/ITGAE and other residency markers[83], ZNF683+ cells expressed transcription factors regulating T-cell terminal differentiation: ZNF683 (Hobit), ZEB2, ZEB1, TBX21 (T-bet), and PRDM1 (Blimp-1), which were also confirmed by TF regulatory network analysis (Figure 3.8B). ZNF683- cells, on the other hand, retained expression of markers associated with a higher potential to persist and seed diverse T-cell memory populations, such as *IL7R*[85] (Figure 3.8B). Reactivated ZNF683^{+} cells further acquired expression of T-cell cytotoxicity genes (*GZMB, GZMA, PRF1, GZMH*)[81]. However, expression of co-inhibitory molecules, including checkpoint inhibitor therapy targets CTLA4 and PD1/PDCD1, was restricted to a few cells at the end of the trajectory (Figure 3.8B-C), implying that the strongest T-cell responses in CC may not be due to direct drug binding effects.

Next, we questioned whether this reactivation is consistent with local T-cell expansion in CC and observed that the majority of actively proliferating cells in both CC and HC were having ZNF683^{+} phenotype. In CC, however, the overall frequency of proliferating T-cells was substantially increased (Figure 3.9A), implying that the normal TRM reactivation pathway is greatly amplified when compared to health. Conversely, in UC, the most proliferative cells were Tc17/Th17 cells, with ZNF683^{+} phenotype being much rarer, in line with previous reports (Figure 3.9B)[73].

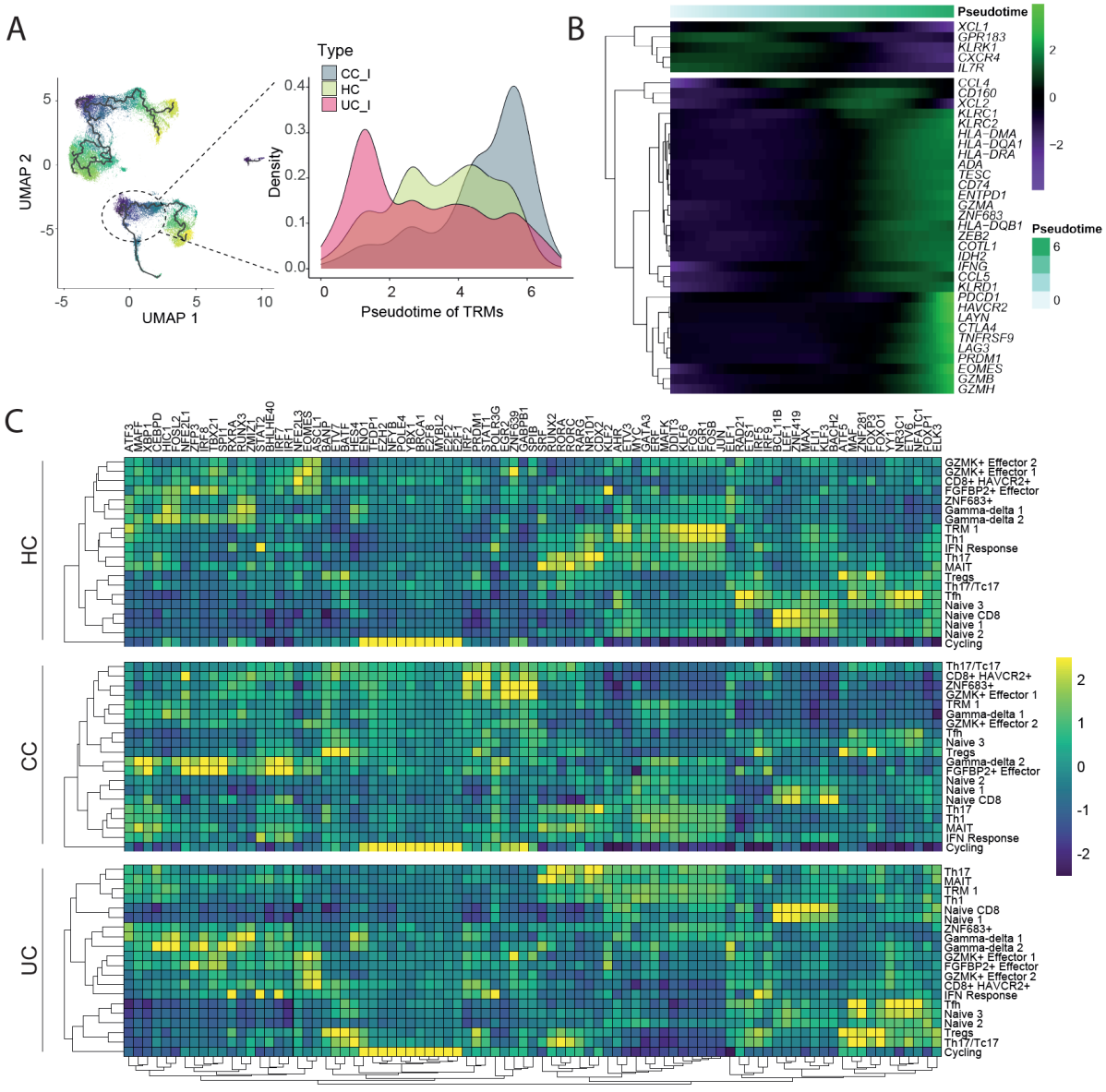


Figure 3.8: ZNF683+ TRMs enriched in CC compared to UC. A. Transcriptomic pseudotime trajectory analysis visualised over the UMAP embedding of colonic tissue T-cells (left). Density plot visualising distribution of all CD103+ TRM cells (cells indicated with dashed circle over UMAP in left) along pseudotime in CC, UC and HC samples (right). B. Pseudotime heatmap of selected TRM pseudotime varying genes. For each gene, loess fit against pseudotime is shown. C. Heatmap visualising scaled, mean transcription factor network activity AUC value per T-cell subpopulation in cells from health (top), CC (middle) and UC (bottom).

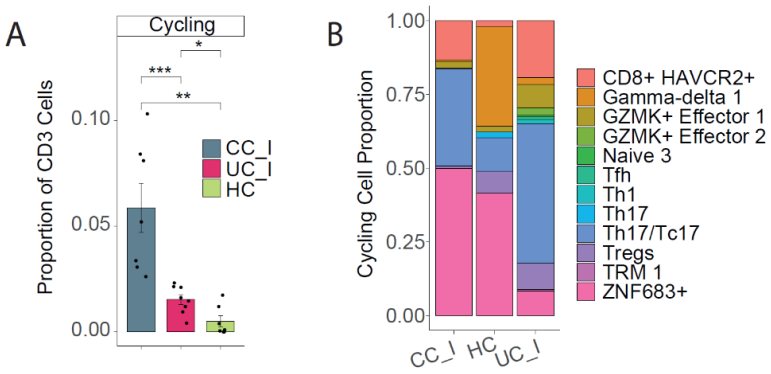


Figure 3.9: Abundance and proportion of cycling T-cells. A. Actively proliferating T-cell abundance changes as a fraction of all T-cells in sample in Checkpoint-inhibitor induced colitis (CC_I), inflamed ulcerative colitis (UC_I), and healthy control (HC) samples, shown as bar plots. Wilcox rank test, p-value < 0.05 $*$; p-value < 0.01 $**$; p-value $< 0.001$$***$; ns = not significant. Error bars represent standard error of the mean (SEM). B. Proportion barplot visualising the distribution of non-cycling T-cell cluster cell phenotypes within cycling T-cells in CC, HC and UC samples.

Single-cell TCR tracking reveals contrasting T-cell migration dynamics in CC and UC

To assess the landscape of tissue clonal dynamics and T-cell trafficking in CC and UC, matched single-cell VDJ sequencing data was used to reconstruct TCR repertoire, yielding 40,745 total clones. As previously reported[77,78], $CD8^+$ T-cells were highly clonal compared to $CD4^+$ populations in both blood and tissue across all conditions (Figure 3.10A), limiting the confidence in interpretation of CD4+ T-cell clonal migration and differentiation dynamics. As such, the analysis was focused on CD8+ compartment, where $CD8^+$ effector and memory cells, exhausted and cycling cells constituted the most expanded populations.

In both blood and tissue, $CD8^+$ effector cells were also the most trafficked population, although very little clonal sharing was found between TRMs and blood T-cells in health. This suggested either recirculation of clones or tissue-retention and acquisition of resident phenotypes by circulating cells (Figure 3.10B-C). Moreover, the relative strength of pairwise clonal repertoire overlaps was calculated among all clusters in blood and tissue within each condition using Morisita's index, which allows to account for differences in the overall size of repertoires between sub-populations. Using this approach, shared clonal networks were reconstructed for both $CD4^+$ and $CD8^+$ populations (Figure 3.11).

In health, in $CD8^+$ populations, clusters from tissue and blood largely formed their own communities, with little blood-to-tissue trafficking (Figure 3.11A(i)). The TRM population in tissue formed the largest clonal overlap, further supporting the evidence of a re-activation continuum between TRM and $ZNF683^+$ cells. In blood, large clonal overlaps were observed between GZMK-expressing effectors and $FGFBP2^+/CX3CR1^+$ effector cells, the latter of which

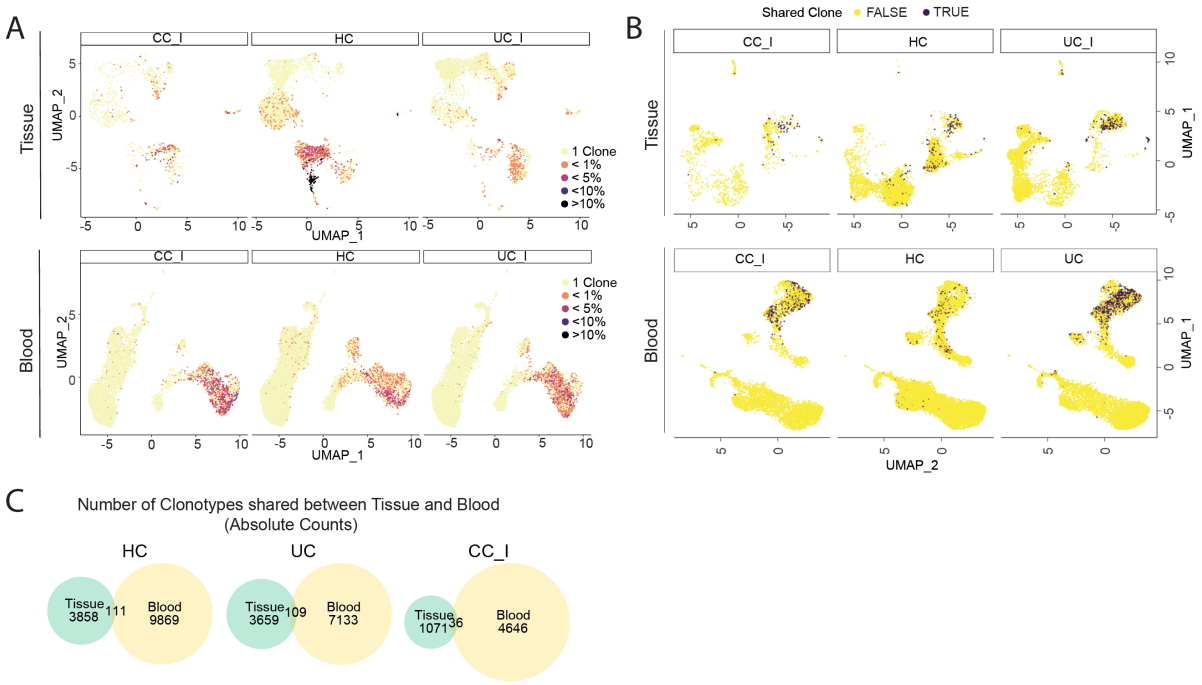


Figure 3.10: T-cell trafficking and clonal expansion in healthy, UC and CC. A. UMAP overlay showing per sample clone size distribution in single T-cells in colonic tissue samples (top) and PBMC samples (bottom). B UMAP overlay visualising the cluster distribution of TCR clones in PBMC samples shared with colonic tissue biopsies (top) and TCR clones in colonic tissue biopsies shared with paired PBMC samples (bottom). C. Venn diagrams visualising the overlap between total clonotypes shared between PBMC and colonic biopsy tissue samples in HC, UC and CC samples.

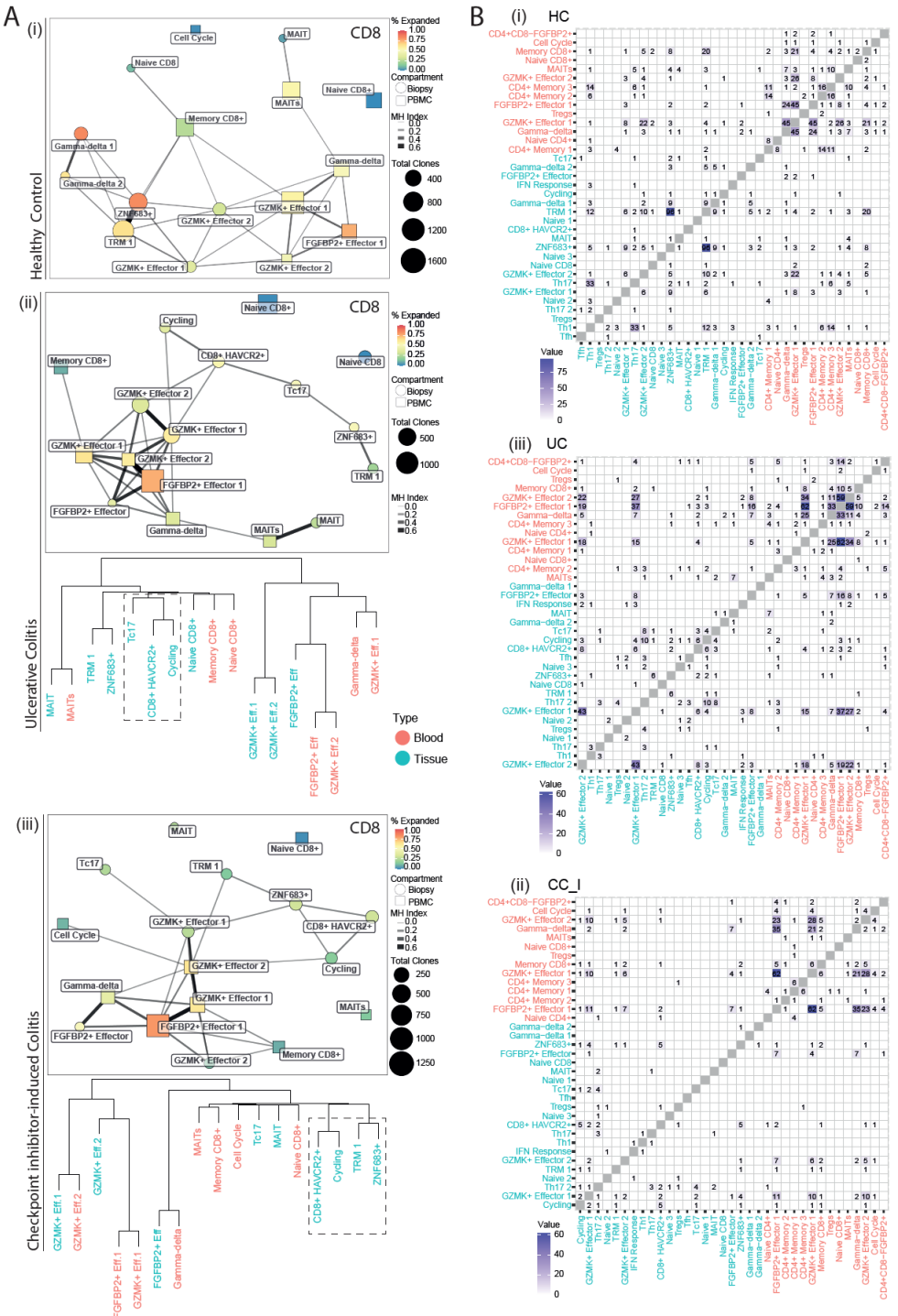


Figure 3.11: Trafficking and clonal expansions of T-cells in Health and Disease. A. Network plot visualising TCR clonal trafficking between PBMC and colonic tissue T-cell populations in samples from health (i), UC (ii) and CC (iii) in CD8+ T-cells. Each node represents a T-cell population as defined by scRNA-Seq transcriptome phenotype while edges represent clonal sharing as defined by Morisita's Index. Hierarchical clustering based on Morisita's index overlap is shown for UC and CC samples below. Resident CD8+ population clusters are highlighted with a dashed box. B. Absolute clonotype overlap count between all T-cell subpopulations in tissue and PBMC paired samples in HC (i), CC (ii) and UC (iii). The gamma-delta clusters are occupied by alpha-beta TCRs with similar transcriptomics profiles of gamma-delta populations.

has been shown to be restricted to vessels[86] but increased in CC (Figure 3.6) Sharing between blood and tissue MAIT-cells and GZMK+ effector populations accounted for the majority of clone trafficking in health.

In UC, completely distinct dynamics were observed, with greatest community formed by the trafficking of all effector populations that were highly expanded (Figure 3.11A(ii), 3.11B(ii)). In addition, there was an increase in MAIT-cell trafficking when compared to health. In line with previous reports in UC[73,83], exhausted, cycling and Tc17 populations were also closely linked, and shared few clones with circulating cells, suggesting the expansion of Tc17 residenT-cells in UC is restricted to tissue (Figure 3.11B(ii)). According to our CITE-Seq data, Tc17 cells expressed the tissue residency cell marker CD103, but not Th17 cells (Figure 3.12A). Additionally, tissue effector cells displayed increased expression of tissue egress (*S1PR1*)[87] and lymph-node homing markers (*SELL, CCR7*)[88] the latter of which were restricted to naïve cells in health and in CC (Figure 3.12B-C).

In CC, there was an increase in trafficking between effector populations in the blood and tissue when compared to health, but to a lesser extent when compared to UC (Figure 3.11A(iii), 3.11B(iii)). Within tissue, exhausted, cycling, TRM and ZNF683^{+} cells formed their own community, with few blood-shared clones, suggesting these cells were forming a local tissue response. However, in contrast to UC, Tc17 cells were not closely clonally linked to this group and ZNF683+ population was central instead (Figure 3.11A(iii)).

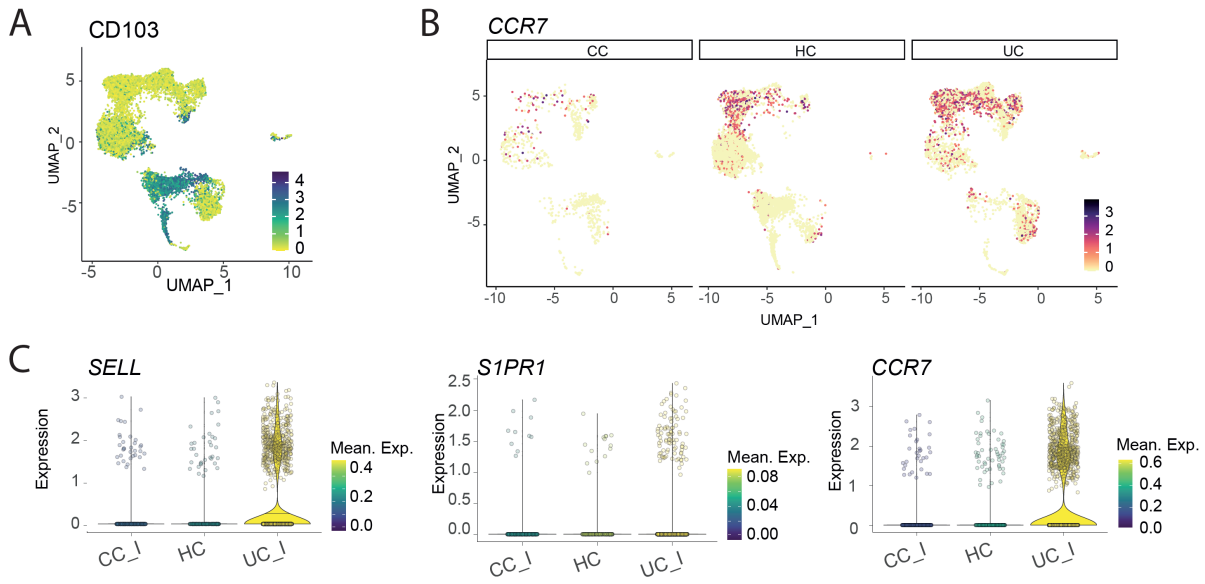


Figure 3.12: Increased expression of tissue egress and lymph-node homing markers in UC. A. UMAP overlay visualising expression of CD103 protein as measured by CITE-Seq in colonic biopsy tissue T-cell subpopulations. B. UMAP overlay visualising *CCR7* expression in all ST slides in CC, HC and UC sections. C. Violin plots showing expression distribution of CCR7, SELL and S1PR1 in GZMK+ Effector T-cell populations in colonic biopsies in UC, CC and HC samples.

Diverse unknown antigens drive T-cell migration from blood to the site of inflammation

To investigate the potential antigens driving T-cell migration, TCRs were first cross-referenced to public databases (VDJdb, McPAS, TBAdb)[39,40,61], and identified 1,263 unique clonotypes (exact TRB matches) linked to 47 known antigens or pathology in our datasets. However, these public TCRs were largely restricted to non-expanded and non-trafficked T-cells, suggesting that the majority of T-cell responses in all conditions are likely due to unknown antigens (Figure 3.13). Notably, we did not observe expansion of T-cell clones previously associated with cancer neo-antigens in CC, although this mechanism cannot be ruled out due limitations in available data and privacy or HLA-specificity of TCR repertoires.

To further investigate whether trafficking is likely to be a response to a particular antigen or a diverse set of antigens, all trafficked CDR3 sequences were clustered using GLIPH2[89] algorithm. The majority of the trafficked TCR clones belonged to distinct antigen-specificity groups in both UC and CC, suggesting a response to diverse antigens (Figure 3.14).

Taken together, this suggests that trafficked and locally expanded T-cells contribute to mucosal inflammation in both CC and UC. However, in UC T-cell trafficking is substantially increased with relative depletion of TRM and effector populations in tissue, with Tc17 and exhausted cells being central to the network of locally proliferating resident population. In CC, on the other hand, cycling, reactivated TRM and exhausted cell constituted their own cluster, forming a local tissue response, with ZNF683+ at central of the network. In search for cognate antigen, we found that T-cell expansion or trafficking was a response to diverse yet unknown antigen. Further characterization of their targets would be essential to understanding the complete picture of inflammation in colitis and identifying therapeutic targets.

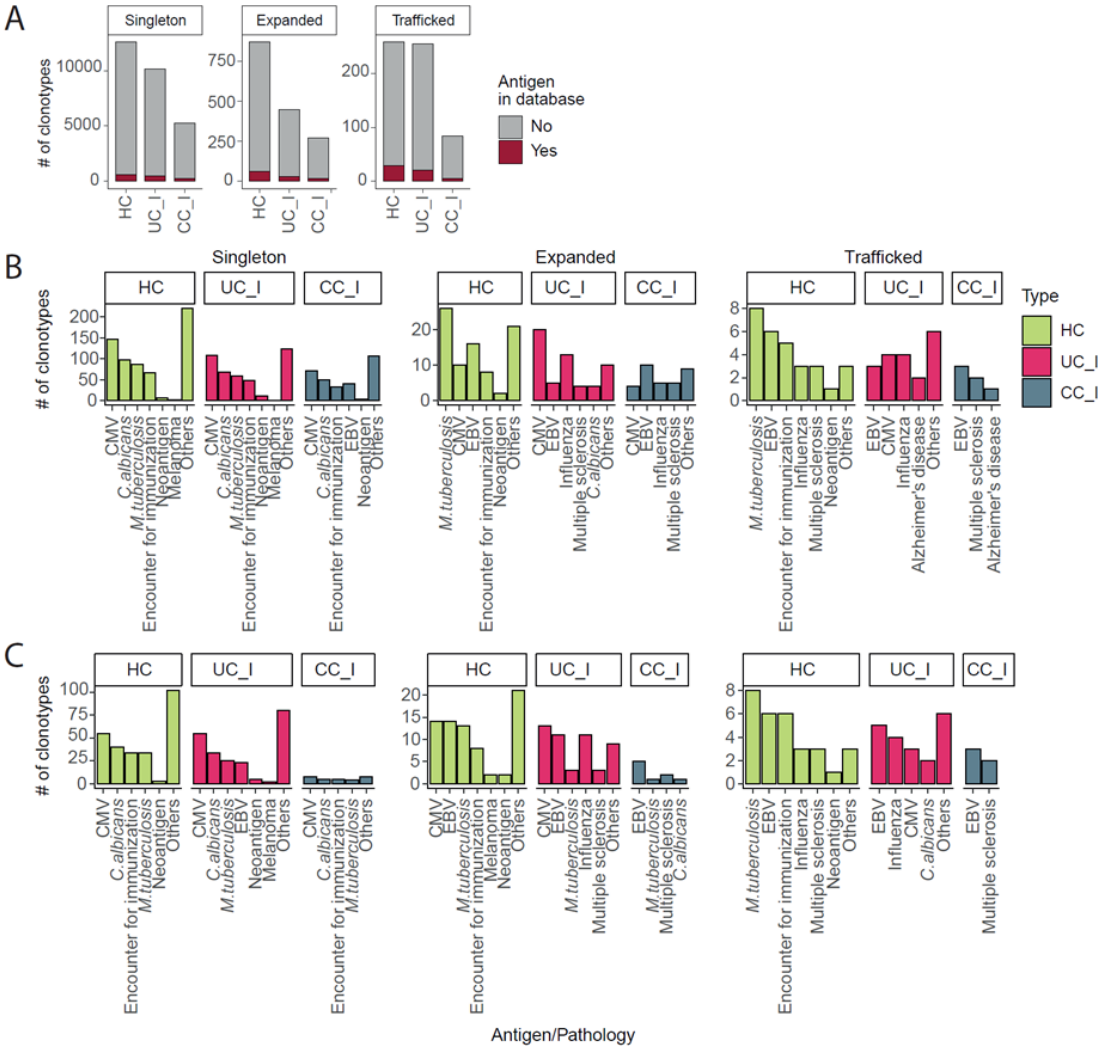


Figure 3.13: Using public database to reference TCR specificity. A. Number of TCR clonotypes with known antigen specificities in public TCR-antigen databases. Non-expanded (left), expanded (middle) and trafficked between blood and tissue (right) clonotypes are shown. B. Barplot visualising top known antigens for public TCR sequences from VDJdb, McPAS and TBAdb databases in PBMC samples (right) and colonic tissue biopsy samples (left) in unexpanded/singleton clonotypes (left), expanded clonotypes (middle) and trafficked clonotypes (bottom) in HC, UC and CC samples.

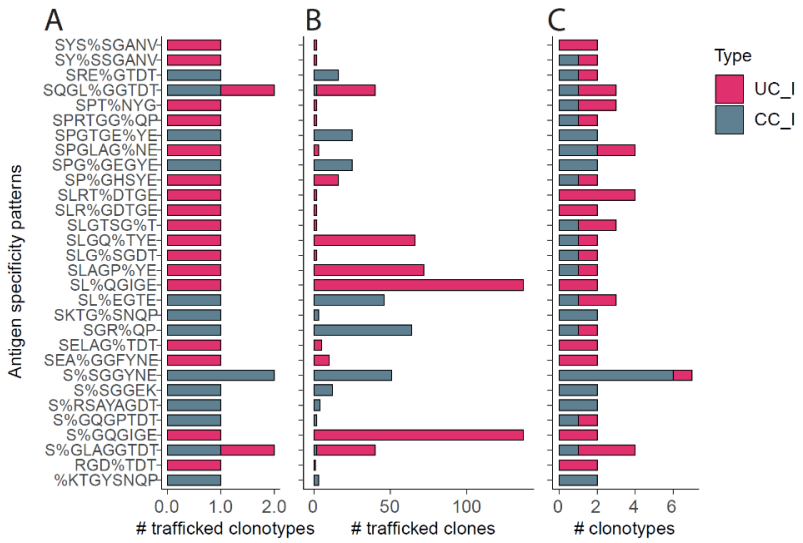


Figure 3.14: Predicting TCR specificity of UC and CC clones. Largest recovered GLIPH2 TCR antigen specificity groups, visualising the counts of trafficked clonotypes per group (A), trafficked clones/expansion per group (B) and total number of trafficked and non-trafficked clonotypes per specificity group (C) in UC and CC samples.

Discussion

Using integrative analysis of multi-modal single-cell profiling in a cohort of checkpoint inhibitor therapy induced and ulcerative colitis patients, key transcriptomic perturbations that delineate these conditions and how these converge into discrete multi-cellular niches within the colonic tissue architecture were identified.

The fact that two conditions induce similar transcriptional programmes suggests that many downstream inflammatory responses in the gut converge regardless of disease aetiology. On the other hand, divergent inflammatory tissue microenvironment domains were found in both CC and UC by mapping disease-specific alterations identified at the single-cell level in tissue.

While increased T-cell infiltration of the mucosa was observed in both conditions, CC was characterised by a significant increase in highly proliferative CD8+ populations, whereas CD4+ T-cells were more prevalent in UC. In CC, there was an increase in both IFNγ and T-cell target chemokine signalling near the crypt tops, indicating an acute inflammatory phase that persisted even within the mucosa. This suggests that a component of the CC pathology may be due to an inability to dampen immune responses once they have been initiated as part of the normal inflammation and damage resolution cycle.

Clonal tracking analysis of TCR repertoires in blood and tissue revealed that, while T-cell response in both conditions consists of a systemic, (re-)circulating component as well as a local resident T-cell expansion, significant differences exist. In UC, higher levels of T-cell trafficking and retention of lymph node homing markers were observed even within effector populations.

According to previous findings[73,77], CD8+ resident memory expansions in UC were skewed towards Tc17 polarised cells that did not recirculate, whereas in CC, a reactivated intra-epithelial ZNF683+/Hobit population that formed the majority of locally-proliferating, non-trafficked T-cells.

ZNF683+ cells with a similar phenotype were recently identified as the intra-tumor lymphocytes responsible for anti-tumor responses in lung cancer after checkpoint inhibitor treatment[90], implying that activation of analogous T-cell populations may be responsible for both anti-tumor responses and adverse immune effects in the gut. While it is likely that these cells are targeting cancer neo-antigens within the tumour micro-environment, the trigger of responses within the gut is less clear. Our TCR repertoire analysis revealed that clonal expansions in CC are not consistent with known public TCRs/antigens and are unlikely to have developed in response to a single antigen. Although there is some evidence that the local microbiota plays a role in driving inflammation in some CC patients, it is unclear whether these cells target luminal bacterial or viral antigens, or altered self-peptides from benign somatic mutations. Because epithelium/barrier sites typically have a higher somatic mutation burden, the latter hypothesis is especially intriguing in light of the increased presence of reactivated TRM in the intra-epithelial space.

Taken together, our findings suggest that checkpoint inhibitor-induced colitis may result from disruption of several populations and T-cell functions. Consequently, the downstream alterations within the local tissue microenvironment enables the recruitment and activation of bystander T-cells (e.g., non-resident cytotoxic CD8s, non-exhausted TRMs) and other immune populations, and this may contribute to the perpetuation of colitis.

Checkpoint-inhibitor induced colitis occurs in up to 60% of patients undergoing cancer immunotherapy. However, as a relatively new entity, little is known about the molecular mechanisms underpinning its pathology. Our work showed that T-cell trafficking occurs at low levels in CC, implying that vedolizumab, a key biologic used to treat UC, would have limited efficacy. Instead, our findings suggest that disrupting microdomain anatomy by targeting specific tissue-localized trafficking pathways may be an alternative approach to non-specific biologics, potentially opening up new avenues for personalised medicine.

This work demonstrated the power of single-cell technologies to uncover T-cell compositions and cross-talks in autoinflammatory diseases, as well as identify pro-inflammatory T-cells. However, we were unable to identify the target antigens driving the divergent T-cell dynamics. In this regard, computational approaches to understand the features underlying T-cell recognition of cognate antigens and identifying CD8+ T-cell targets may provide solutions to this unmet need.

Methods

This is a selection of methods relevant to the content of this thesis. Detailed methods for sample handling, processing, cell isolation, ST data analysis should be referenced to the manuscript in preparation.

Full ST and scRNA-Seq datasets have been made available via an interactive data portal at https://simmonslab.shinyapps.io/CheckpointUCDataPortal/.[reviewer access – username: reviewer password: wncvm0sz7fsd2588].

Patient Sample Collection

Samples were collected with informed consent during clinical mandated visits in accordance with local and national patient safety directives. At the same clinical visit, 3-4 biopsy pairs were taken at endoscopy and 10mls of blood were collected. The operating surgeon collected surgical samples from tissue that exceeded clinical histopathological requirements. Inflammation status was confirmed through endoscopic examination and review of the clinical histopathology report issued by experienced consultant pathologists employed for NHS patient care. Individual patient details (including gender, medication and disease severity) have been anonymized and deposited at Mendeley Data

(accession:https://data.mendeley.com/datasets/7z8yx644hb/draft?a=e8b9e179-8fb7-448e-a481-7cd4c090f71a).

Single-Cell RNA sequencing

***Epithelial/Lamina Propria/Matched PBMC cells*:** Tissue dissociation was performed as previously described[76]. Briefly, 0.5 million cells from each fraction (Epithelial, CD45/Stromal and PBMC) were incubated for 30 minutes at 4°C with a mix of optimised concentrations of hashing and CITE-seq antibodies after 10 minutes of blocking with Trustain FcX (Biolegend). To remove unbound antibodies, cells were washed in staining buffer at each step. Cells from up to five different samples were mixed together in a 1:1 ratio. Each pool was loaded into a different channel of the 10X Chromium single-cell platform, one for the epithelial fraction and one for the lamina propria fraction. Each channel received an input of 30,000 single-cells per pool, with a recovery rate of approximately 10,000 cells. 10x Genomics Library Kits were used to create the libraries. From matched blood samples, cells from four different samples were pooled at 1:1 ratio. Similarly, each channel got 30,000 cells from each pool. Gene expression (GEX) and Antibody-derived Tag (ADT) Libraries were prepared using 10x Genomics Library Kits (10X Genomics, CG000186,Rev A). In addition, from the lamina propria and PBMC pools, single-cell TCR and BCR libraries were generated. ***Tissue CD45+ cells*:** Live CD45+ cells were sorted as described above, counted and resuspended at a final concentration of 1million/mL. 10,000 cells were loaded into the 10X scRNA platform for each sample. GEX Libraries were prepared using 10x Genomics Library Kits (10x Genomics, CG000183,Rev A). ***CD3+ cells*:** Live CD3+ cells from tissue and blood were sorted as described above. Cells were counted from three samples, pooled in a 1:1 ratio, and 20,000 cells from each pool were loaded into the 10X scRNA platform. GEX, ADT and TCR libraries were prepared using 10x Genomics Library Kits (10x Genomics,

CG000208,Rev F). Sequencing of these libraries was carried out on an Illumina NovaSeq 6000 platform with 1% PhiX (Novogene).

Raw sequencing data processing

All raw sequencing data was converted from bcl to fastq format using the Illumina bcl2fastq software, version 2.20.0.422, with each sample index barcode allowing for up to one mismatch. FastQC software was used to check the quality of raw sequence reads[91].

Raw 10X scRNA-Seq and CITE-Seq data processing

Cellranger software from 10X Genomics was used to process, align, and summarise unique molecular identifier (UMI) counts against the hg38 (10x reference: refdata-gex-GRCh38-2020-A) human reference genome for each sequenced scRNA-Seq pool. Matched protein CITE-Seq and hashing antibody panel data were processed together with scRNA-Seq as matched feature barcoding libraries. Antibody tag UMI counts were calculated using a joint feature barcoding sequence reference of TotalSeq antibody sequences of hashing antibodies and protein expression target panel.

Raw UMI count matrices were imported into R for further processing. For each scRNA-Seq sample, cell calling was performed using 'emptyDrops'[92] function from DropletUtils on the full raw count matrices of all barcodes in the 10x barcode whitelist to distinguish cells from empty droplets containing only ambient RNA. Raw count matrices were corrected for Illumina index swapping using 'swappedDrops'[93].

10x scRNA-Seq data normalization and dimension reduction

For each 10x reaction, Seurat R package[94] was used to normalise expression values for total UMI counts per cell. The mean-variance relationship was used to identify highly variable genes, and principal-component analysis was used to reduce dimensionality. Scree plots were used to determine the number of principal components to use for clustering analyses for each pool. Cells were then clustered using Louvain algorithm for modularity optimization using kNN graph as input. Cell clusters were visualized using UMAP algorithm[95] with principal components as input and n.neighbors = 30, spread = 1 and min.dist = 0.1.

Cell type annotation

scRNA-Seq cell populations were annotated using a combination of known marker genes and protein expression profiles (summarised in Supplementary Table 1-3 for each population/cluster) and using previously published scRNA-Seq reference atlas datasets in the colon and PBMCs[75–77,80,81,83]. In the case of the latter, we used Seurat to perform label transfer between datasets for each cell type in reference 10x datasets for each cell in our data and assigned putative cluster identities to populations based on the most frequently predicted cell label for each cluster. The classification results were then compared to known marker gene expression. The most specific or informative marker was used to identify finer or higher resolution clusters or condition-specific populations.

Trajectory analysis

The Monocle 3 algorithm was used to calculate T-cell embedding trajectory using integrated, batch corrected, dimension-reduced data[96]. Seurat objects were converted to Monocle 3 cell_data_set objects and cells were re-clustered to enable learning of disjoint graphs in multiple partitions for CD4+ and CD8+ populations separately. Trajectory reconstruction was performed, with closed loops and multiple partitions. The start of the trajectory was denoted with the node within the naïve cell cluster, and pseudotime was computed along the trajectory. To calculate cell densities over pseudotime from the TRMs to reactivated TRMs, cells along the branch of trajectory were selected, and density distribution was calculated and visualised using gaussian density kernel estimate as implemented in ggplot2 function stat_density(), with default parameters.

Genes varying with pseudotime were detected by fitting generalised linear negative binomial models to each gene in cells within the trajectory from TRMs to reactivated TRM. To test for non-linear relationship of gene expression on pseudotime, natural splines (3 degrees of freedom) were constructed for the pseudotime variable and a likelihood ratio tests were used to detect pseudotime varying genes between full and reduced (no pseudotime) models. p-values were further corrected for multiple testing using Benjamini-Hochberg multiple testing correction. In order to visualise selected pseudotime varying gene expression patterns over pseudotime, a loess regression fit of expression over pseudotime was generated for each gene across all cells within the TRM branch of the trajectory. Smoothed expression patterns obtained from the fits were visualised as heatmaps using R package 'pheatmap'. Hierarchical clustering with complete linkage were used to cluster genes.

Condition-specific clusters

To identify condition specific clusters, for each cluster of each sample within each cellular compartment, cell counts were normalised to the total number of cells detected within that compartment in a given sample and the proportions of cells were compared using a two-sided Wilcoxon test, with p-values < 0.05 considered as significantly different. The R package ggpubr was used to visualise proportion distributions, cluster means and standard errors. The graph-based differential abundance analysis was performed using R package miloR[97] for cell type populations that exist on a phenotypic continuum rather than discrete clusters (e.g. T-cell subtypes). The integrated, batch-corrected harmony components were used for nearest neighbour graph reconstruction, with k=10 nearest neighbours used for neighbourhood definitions throughout.

Transcription factor analysis

The Rpackage SCENIC workflow was used to detect active transcription factor modules in all datasets. The normalised single-cell gene expression matrix was first filtered to exclude any genes detected in fewer than 20cells. The RcisTarget database, containing transcription factor motif scores for gene promoters and transcription start sites for the hg38 human reference genome, was downloaded from https://resources.aertslab.org/cistarget/databases/homo_sapiens/hg38/refseq_r80/mc9nr/gene_based/ and the expression matrix was further filtered to include only genes available in the RcisTarget database.

The remaining genes were used to compute a gene–gene correlation matrix for co-expression module detection using the random forest-based GENIE3 algorithm[98]. The Rpackage SCENIC[99] was used to perform transcription factor network analysis to detect co-expression modules enriched for target genes of each candidate TF from the RcisTarget database. The AUCell package[99] was used to compute a score for each TF module in each individual cell. The generalised linear models were used to test for condition or cluster dependence of TF AUC values to identify condition- or cluster-specific TF modules. The batch and gene detection rates were also considered as AUC values were highly dependent on gene detection rate of individual cells, and they were used to block co-variates in the model formula. The resulting p-values were further adjusted for multiple testing using Benjamini-Hochberg multiple testing correction.

Receptor-ligand interaction analysis

The R package 'CellChat'[100] was used to infer receptor-ligand interactions between all single-cell clusters identified from scRNA-Seq data, as well as condition-specific interactions. Cell-to-cell crosstalk was performed separately on each condition, and then differenT-cellChat objects for conditions were merged together to model communication probability and significant communications between conditions. The normalised data matrix and cluster label metadata were provided as input to compute probability and infer cellular communication networks using CellChat functions, "computeCommunProbPathway", "aggregateNet", "computeCommunProb", "identifyCommunicationPatterns", and "compareInteractions" with default parameters. Specific interactions were visualised using Circos plots.

scTCR-Seq data processing

Single-cell TCR clonotypes were assembled using Cellranger VDJ software. Because paired VDJ enrichment libraries were not available for intra epithelial T-cells obtained from epithelial crypt dissociation isolation (processed using 5' scRNA-Seq chemistry), TCR sequences were inferred from transcriptome data using TRUST4 software[101] with the same VDJ reference. Single-cell barcodes were then used in both cases to link corresponding VDJ clonotypes and gene expression data.

TCR clonality analysis

VDJTools[102] software was used to compute TCR repertoire statistics for individual samples, as well as cluster-level statistics. Clonotypes were defined by paired CDR3$\alpha\beta$ nucleotide sequence for analysing clonal expansion and trafficking. T-cell clonality was defined by computing Shannon Entropy and the fraction of total clonal space occupied by each clonotype in a sample. A clonotype was considered expanded if more than one clone was detected within the same sample and sample type (blood and biopsy samples from the same donor were considered separately and addressed by trafficking analyses).

Reference TCR specificity using public database

TCR CDR3 amino acid sequences were assembled and compared to known and/or public TCR sequences in VDJdb (vdjdb-2021-02-02)[39], McPAS (downloaded 2021-07-21)[40] and TBAdb

(downloaded 2021-07-21). GLIPH2[89] was used to predict antigen specificity of TCR clonotypes and specificity clusters with Fisher's p-values ≤ 0.05 was analysed.

TCR dynamic analysis

To examine TCR repertoire overlaps and assess phenotypic conversion and trafficking, Morisita's index was computed using R divo package between all pairwise blood and tissue T-cell clusters. The CD4+ and CD8+ cell populations were considered separately to aid clarity in the analysis and visualisation of shared clonal networks between and within blood and tissue T-cell populations. Little clonal sharing was observed between two lineages as expected. For CD8+ T-cell networks, we further filtered out edges with very low clonal overlap for clarify. The filter thresholds were defined based on the distribution of Morisita's index across all CD4 and CD8 cell populations, as these are likely to be spurious connections or noise, with all edges below 95^{th} percentile being considered low confidence. Because CD4+ cell populations are much less clonal than CD8+ cell populations, all edges were kept for visualisation. However, a much larger sampling of CD4+ cells from each individual would be required to confidently infer clonal dynamics of CD4+ populations, as most of the TCR sharing between populations was observed as only one or two clones, which may be spurious in the cases of closely phenotypically related cell clusters (but not blood-tissue sharing) due to sources of errors in the dataset such as barcode multiplets[103]. Heatmaps were made using raw, unfiltered counts of all overlaps in CD4 and CD8 populations in CC, UC and HC. The R package ggraph was used to create weighted, undirected networks, with edge weights defined as Morisita's clone overlap index. Graphs were visualised using a force-directed Fruchterman-Reingold layout. Edge width ranges were standardised between conditions, as before, for comparable visualisation. Hierarchical clustering of CD4 and CD8 T-cell populations was also performed using the pairwise Morisita's overlap index matrix to define population similarity in CC, HC, and UC samples. Clustering was performed in R using the hclust() function in a complete linkage method.

Chapter 4. CD8+ T-cell targets against human coronaviruses

Original publications:

- **Lee CH**, Koohy H. In silico identification of vaccine targets for 2019-nCoV. F1000Res 9, 145 (2020)
- **Lee CH**, Pinho MP, Buckley PR, Woodhouse IB, Ogg G, Simmons A, Koohy H. Potential CD8+ T-cell Cross-Reactivity Against SARS-CoV-2 Conferred by Other Coronavirus Strains. Front. Immunol. 11, (2020).
- Buckley, PR, **Lee, CH**, Pinho MP., Babu, RO, Woo, J, Antanaviciute, A, Simmons, A, Ogg, G and Koohy, H., HLA-dependent variation in SARS-CoV-2 CD8+ T-cell cross-reactivity with human coronaviruses. Immunology (2022).

Introduction

In early 2020, faced with a rapid epidemical outbreak of a novel and unknown pathogen, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a key bottleneck for vaccine development was limited access to samples of infected individuals. We therefore set out to computationally identify CD8+ T-cell targets from SARS-CoV-2 proteome to provide a shortlist of vaccine candidates for downstream experimental validation, thereby accelerating the development pipeline[104].

Moreover, while the severity of symptoms had been linked to age, gender and comorbidities such as cardiovascular disease and chronic respiratory disease[105], the underlying mechanism of the broad variation in susceptibility and severity to COVID-19 was not fully understood[106]. It was however accepted that an altered immune response is a key contributor to pathology[107,108], and the balance between the host's generation of protective and pathological immune responses may be a vital factor governing the disease outcome.

To explore factors underlying heterogeneous immune response, we investigated the immune memory from related pathogens that have previously been shown to help reduce severity and spread of the diseases[109–111]. The pre-existing T-cell immunity through cross-reactivity to familial

coronavirus strains may provide individuals with protection or enhanced susceptibility against the SARS-CoV-2[112–114]. Therefore, we sought to characterize the potential for T-cell cross-protection and identify targets for developing an universal coronavirus vaccine.

The coronavirus strains that infect humans belong to either alpha and beta genera. The alphacoronavirus contains HCoV-229E and HCoV-NL63, while the betacoronavirus contains HCoV-OC43, and HCoV-HKU1, as well as middle east respiratory syndrome coronavirus (MERS-CoV), SARS-CoV-1 and SARS-CoV-2[115]. MERS-CoV and SARS-CoV cause more severe symptoms including respiratory tract disease, whereas NL63, 229E, OC43, and HKU1 usually only cause mild to moderate symptoms such as cough, runny nose, fever and sore throat[116].

This study uses immunoinformatic methods to shortlist vaccine candidates and understand disease heterogeneity. To shortlist vaccine targets, SARS-CoV-2 peptides with high sequence similarity to previously characterised epitopes were identified[104], and CD8+ T-cell targets were predicted using immunogenicity algorithms[55]. To better understand pathology of heterogeneous disease outcome, predicted CD8+ T-cell epitopes were compared across seven coronavirus strains and potential cross-protection conferred by pre-exposure to other coronaviruses was evaluated. The role of patient genetics in the presentation of cross-species epitopes was also investigated, and it was found that patient with certain HLA alleles (e.g. HLA-B∗07:02, C∗07:02 and A∗03:01) are more likely to have cross-reactive CD8+ T-cells that recognise SARS-CoV-2 epitopes homologous to HCoV[117]. Lastly, the risk of self-reactivity of predicted CD8+ T-cell epitopes was assessed in order to select candidates with low toxicity and high efficacy. These efforts aimed to improve understanding of the heterogeneous response to disease and vaccination, as well as to foster the development of effective therapeutics.

Results

SARS-CoV-2 peptides sharing high sequence similarity to epitopes

All peptides in IEDB (as of 13-02-2020) that have been reported immunogenic using T-cell assays and have humans as the host organism were retrieved. A local sequence alignment of SARS-CoV-2 proteome was conducted against 35,225 IEDB peptides, and 28 epitopes (i.e. peptides characterised positive by T-cell assays) were found to have identical sequence matches (Table 4.1). Notably, all matching epitopes were derived from SARS-CoV-1. These peptides have been shown to bind various HLA alleles, with a higher number binding to HLA-A:02:01, from both class I and class II that can be target for CD8+ and CD4+ T-cells respectively.

In addition to 28 identical matches, we extracted SARS-CoV-2 peptides having high sequence similarity to known epitopes in IEDB. Based on the normalized alignment score of exact matches, the threshold was set at alignment score of 4 to extract peptides of high sequence similarity (Table 4.2), and additional 48 peptides were identified, of which 36 were derived from SARS-CoV-1 (Figure 4).

Table 4.1: Comparison of normalized sequence alignment score for peptides with exact and non-exact matches.

Species	**# of epitopes**
Severe acute respiratory syndrome-related coronavirus	28
Homo sapiens	7
SARS coronavirus Tor2	5
Yellow fever virus 17D	2
(Unlabelled)	1
Hepatitis C virus genotype 1	1
Human papillomavirus type 16	1
SARS coronavirus BJ01	1
SARS coronavirus TJF	1
SARS coronavirus Urbani	1
Total	**48**

Table 4.2: SARS-CoV-2 peptides having exact matches with SARS-CoV epitopes deposited in IEDB. Antigen and HLA information are retrieved from IEDB for corresponding epitopes.

IEDB peptide	**SARS-CoV-2 pattern**	**Antigen**	**HLA allele**
TLACFVLAAV	TLACFVLAAV	Membrane glycoprotein	HLA-A∗02:01
AFFGMSRIGMEVTPSGTW	AFFGMSRIGMEVTPSGTW	N protein	
ALNTPKDHI	ALNTPKDHI	Nucleoprotein	HLA-A∗02:01
AQFAPSASAFFGMSR	AQFAPSASAFFGMSR	nucleocapsid protein	HLA class II
AQFAPSASAFFGMSRIGM	AQFAPSASAFFGMSRIGM	N protein	
GMSRIGMEV	GMSRIGMEV	Nucleoprotein	HLA-A∗02:01

Table 4.2: SARS-CoV-2 peptides having exact matches with SARS-CoV epitopes deposited in IEDB. Antigen and HLA information are retrieved from IEDB for corresponding epitopes.

ILLNKHIDA	ILLNKHIDA	Nucleoprotein	HLA-A∗02:01
IRQGTDYKHWPQIAQFA	IRQGTDYKHWPQIAQFA	N protein	
KHWPQIAQFAPSASAFF	KHWPQIAQFAPSASAFF	N protein	
LALLLLDRL	LALLLLDRL	Nucleoprotein	HLA-A∗02:01
LLLDRLNQL	LLLDRLNQL	Nucleoprotein	HLA-A∗02:01
LLNKHIDAYKTFPPTEPK	LLNKHIDAYKTFPPTEPK	N protein	
LQLPQGTTL	LQLPQGTTL	Nucleoprotein	HLA-A∗02:01
RRPQGLPNNTASWFT	RRPQGLPNNTASWFT	nucleocapsid protein	HLA class I
YKTFPPTEPKKDKKKK	YKTFPPTEPKKDKKKK	N protein	
ILLNKHID	ILLNKHID	Nucleoprotein	HLA-A∗02:01
MEVTPSGTWL	MEVTPSGTWL	nucleocapsid protein	HLA-B∗40:01
ALNTLVKQL	ALNTLVKQL	S protein	HLA-A∗02:01
FIAGLIAIV	FIAGLIAIV	Spike glycoprotein precursor	HLA-A2
LITGRLQSL	LITGRLQSL	Spike glycoprotein precursor	HLA-A2
NLNESLIDL	NLNESLIDL	S protein	HLA-A∗02:01
QALNTLVKQLSSNFGAI	QALNTLVKQLSSNFGAI	S protein	HLA-DRB1∗04:01
RLNEVAKNL	RLNEVAKNL	Spike glycoprotein precursor	HLA-A∗02:01
VLNDILSRL	VLNDILSRL	S protein	HLA-A∗02:01
VVFLHVTYV	VVFLHVTYV	Spike glycoprotein precursor	HLA-A∗02:01
GAALQIPFAMQMAYRF	GAALQIPFAMQMAYRF	S protein	HLA-DRA∗01:01/ DRB1∗07:01
MAYRFNGIGVTQNVLY	MAYRFNGIGVTQNVLY	S protein	HLA-DRB1∗04:01
QLIRAAEIRASANLAATK	QLIRAAEIRASANLAATK	S protein	HLA-DRB1∗04:01

Identify T-cell targets conveying cross-protection

As a complementary reciprocal approach, a *de novo* search for immunogenic peptides in the SARS-CoV-2 proteome was performed. Along with SARS-CoV-2, epitopes were predicted from six other coronaviruses and the level of potential cross-reactivity across seven coronavirus strains was analysed to assess potential cross-protection conferred by pre-exposure to other coronaviruses, and identify vaccine targets applicable to multiple human coronavirus strains.

To first assess the homology of proteome sequences among alpha and betacoronavirus strains, we conducted sequence alignment and generated a phylogenetic tree of encoded proteins between NL63, 229E, OC43, HKU1, MERS-CoV, SARS-CoV-1 and SARS-CoV-2 (see Methods). According to sequence alignments (Supplementary Figure 1) and phylogenetic trees of encoded proteins

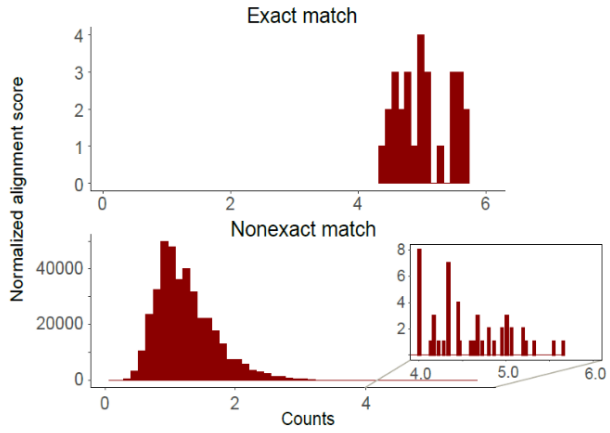


Figure 4.1: SARS-CoV-2 peptides with high sequence similarity to epitopes deposited in IEDB. Number of epitopes having high sequence similarity to SARS-CoV-2 grouped by their source organism.

(Figure 4.2), the alpha and betacoronavirus strains shared a high sequence similarity within their own genera, with NL63 and 229E sharing a higher similarity than betacoronaviruses. Phylogenetic analysis of betacoronavirus strains revealed greater similarity between OC43 and HKU1, and between SARS-CoV-1 and SARS-CoV-2. Notably, MERS-CoV had proteome sequences that were relatively distinct from all other coronavirus strains studied (Figure 4.2).

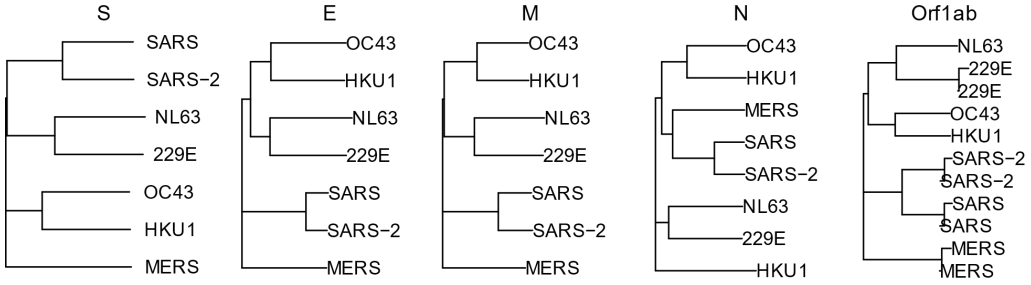


Figure 4.2: Sequence homology between coronavirus strains. Phylogenetic tree of encoded proteins, spike protein (S), envelope protein (E), and replicase polyprotein (Orf1ab) for seven coronavirus strains. The recurrent strain annotations are due to presence of two open reading frames annotated with the same functional protein.

For *de novo* search of immunogenic peptides, 9-mers were generated from the SARS-CoV-2 proteome with a window of 9 amino acids and a step length of 1 amino acid (9613 in total) (Figure 4.3). Then, the antigen presentation potential of 9-mer peptides was examined for the ten most prevalent class I HLA alleles. These include HLA-A∗0101, 0201, 0301, 2402, HLA-B∗0702, 4001, 0801 and HLA-C∗0702, 0401, 0701[118]. The MHC presentation was predicted by NetMHCpan v4.0[119]. On average, 2559 (SD = 120) peptides were predicted to bind at least one HLA allele (Figure 4.4A), accounting for ~21% of the total number of 9-mer peptides found across different strains. There were 66 peptides from seven strains that were predicted to bind at least 7 different

HLA alleles, 9 of which were derived from SARS-CoV-2.

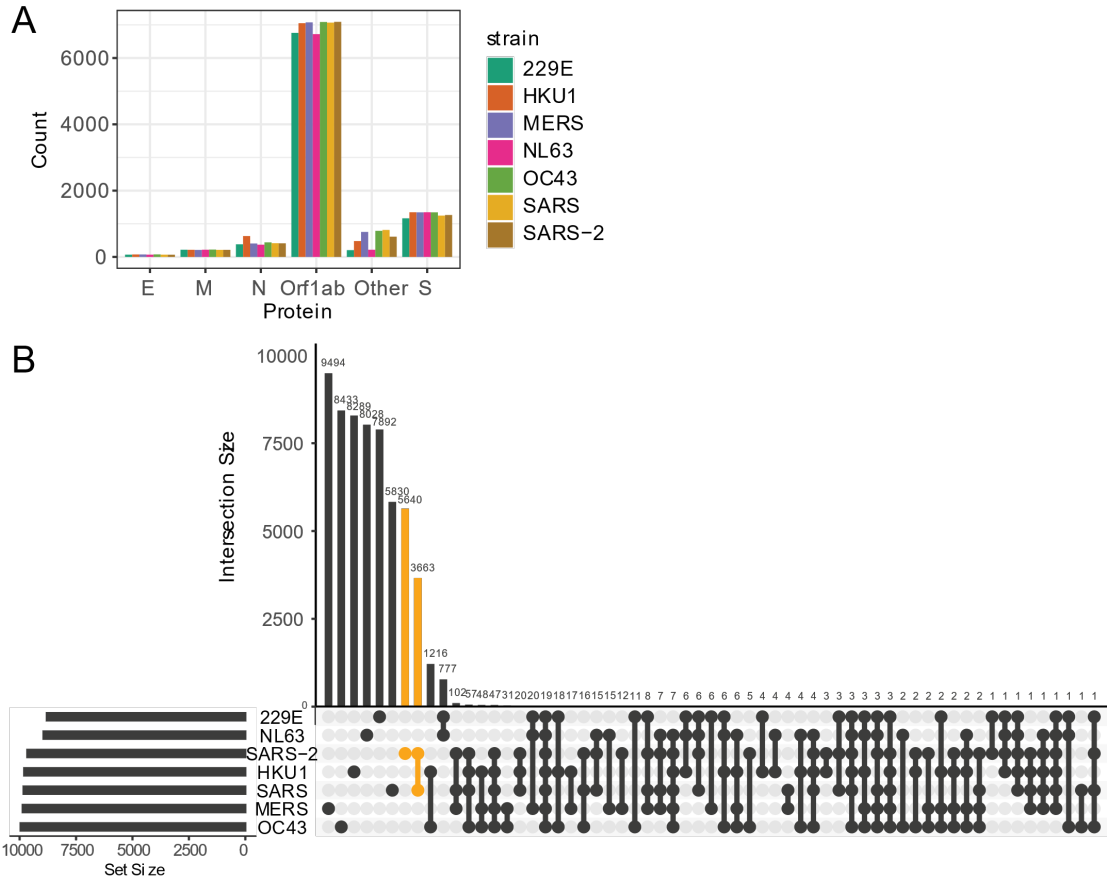


Figure 4.3: Number of shared and private 9-mer peptides across coronavirus strains. A. Number of 9-mer peptides generated from each coronavirus strains grouped by functional proteins, envelope protein (E), membrane protein (M), nucleocapsid protein (N) and replicase polyprotein (Orf1ab), spike protein (S), and other encoded proteins (Other). B. Number of shared and private 9-mer peptides between coronavirus strains. The UpSet plot illustrates interaction of 9-mer peptides between seven coronavirus strains. The number of 9-mer peptides unique from SARS-CoV-2 or shared with SARS-CoV are coloured in orange.

Although viral antigen presentation is a vital step in triggering immune responses, not all MHC binding peptides are immunogenic. We therefore set out to predict T-cell immunogenicity (i.e. the ability of a peptide presented by an MHC molecule to interact productively with a T-cell receptor) of all peptides predicted to bind at least one HLA allele. The existing immunogenicity predicting models were benchmarked and a recently published model called *Repitope*[120] was found to be the best performing existing model to predict viral epitope immunogenicity.

We therefore utilized Repitope (see Methods) and identified 4894 unique predicted binders (~30%, out of 16,096) to be immunogenic (i.e. predicted epitopes), and the proportion of such predicted epitopes was comparable across different strains (Figure 4.4B). On average, 429 (SD=26) epitopes were predicted to bind at least one HLA type and be immunogenic.

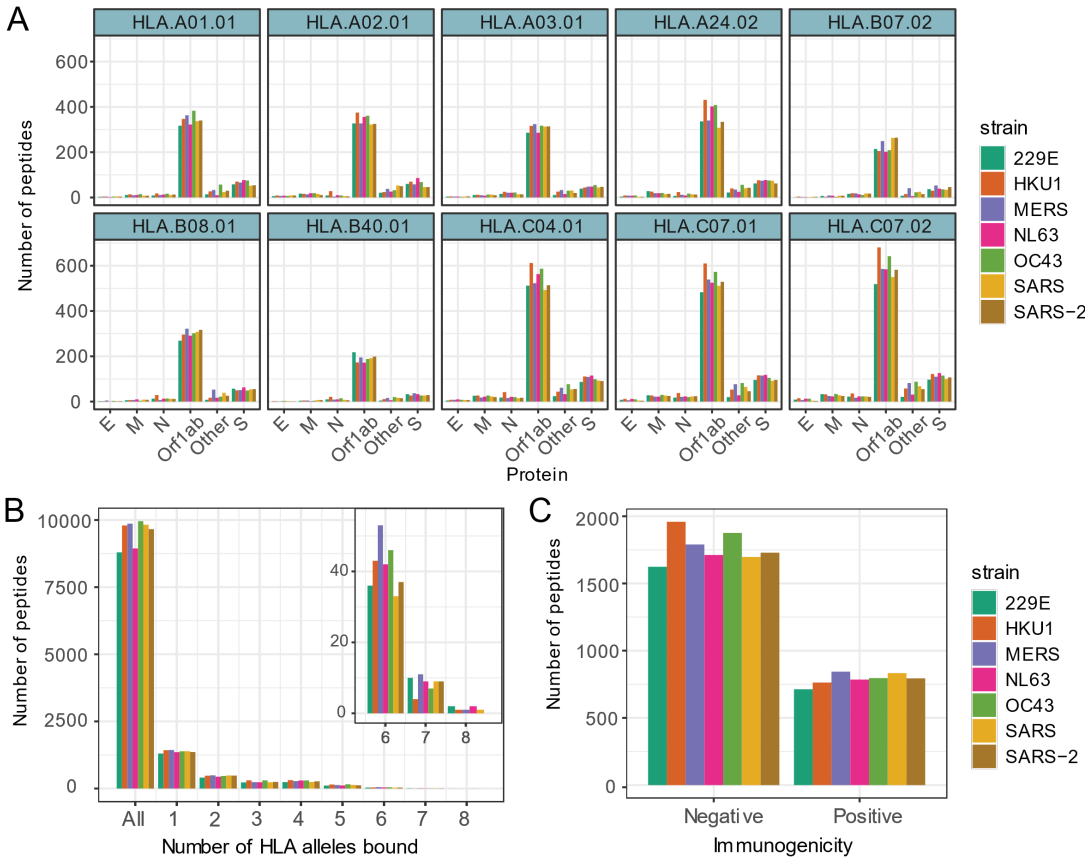


Figure 4.4: Number of 9-mer peptides predicted to bind HLA alleles and immunogenic. A. Number of 9-mer peptides from each coronavirus strains predicted to bind annotated HLA alleles. B. Number of peptides predicted to bind equal to specified number of HLA alleles. C. Number of peptides predicted to productively interact with TCRs by Repitope prediction.

With the pool of predicted epitopes, Venn diagrams were generated to illustrate private and public epitopes across different coronavirus strains (Figure 4.5). From a total of 794 predicted epitopes from SARS-CoV-2, 411 (/794, ~52%) were private, while the remaining 383 (/794, ~48%) were public. Among these, 379 (/383, ~99%) were shared with SARS-CoV and 21 (/383, ~5.4%) shared with either one of four common coronaviruses, 229E, HKU1, NL63 and OC43. Only one predicted epitope – SLAIDAYPL derived from Orf1ab – was found in all strains. Many of the public epitopes were derived from the Orf1ab protein due to its long sequence. This analysis suggested a high level of CD8+ T-cell cross-reactivity between SARS-CoV-2 and SARS-CoV-1, but limited extent with other common flu-like coronaviruses.

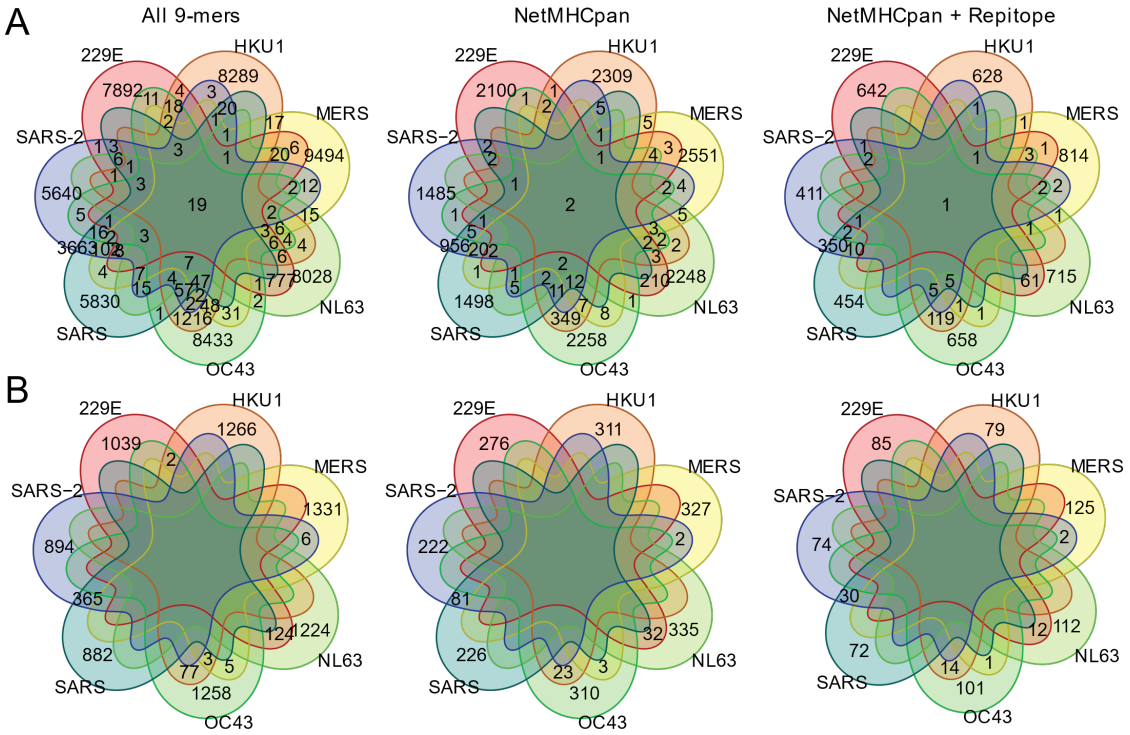


Figure 4.5: Extent of potential CD8+ T-cell cross-reactivity across coronavirus strains. A.B Venn diagram of private and public 9-mer peptides shared across coronaviruses from a complete set of peptides to after MHC presentation prediction by NetMHCpan and immunogenicity prediction by Repitope, for all encoded proteins (A) and peptides derived from spike protein only (B).

Expand public epitopes by sequence similarity

In addition to public epitopes having exact homologies, predicted epitopes with a high sequence similarity may also trigger cross-reactivity across coronavirus strains. We thereby expanded the previous set of public predicted SARS-CoV-2 epitopes to include those with up to two amino acid mismatches (Figure 4.6A). In addition to 383 predicted epitopes from SARS-CoV-2 that were shared with other coronavirus strains, allowing one or two amino acid mismatches resulted in an increase of 173 and 159 unique public epitopes, respectively.

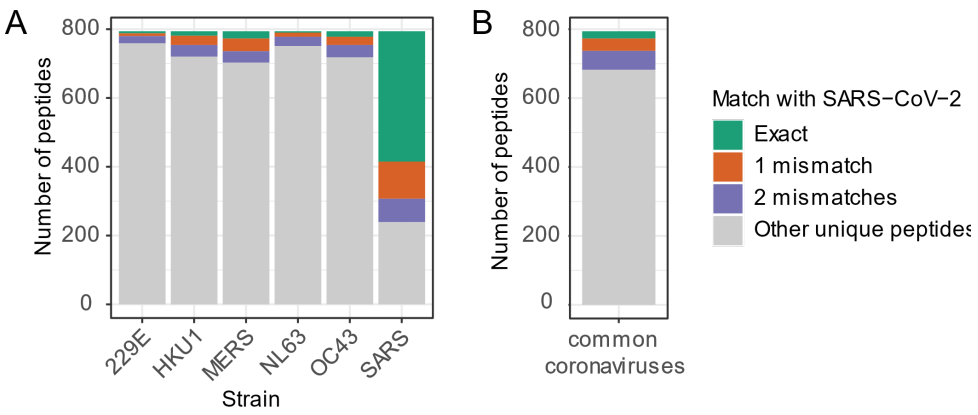


Figure 4.6: Number of shared predicted epitopes between SARS-CoV-2 and other coronavirus strains by allowing up to two mismatches. A, Map of public peptides our of 794 SARS-CoV-2 predicted epitopes. Note that it includes duplicated peptides that may be shared across coronavirus strains, i.e. peptides shared across SARS-CoV-2, SARS-CoV and MERS-CoV are counted in both SARS and MERS. B. Unique predicted epitopes from four common coronaviruses, 229E, HKU1, NL63 and OC43, shared with 794 SARS-CoV-2 predicted epitopes.

While SARS-CoV-2 shared the most with SARS-CoV-1 and MERS-CoV, individuals have a greater exposure to common coronaviruses, 229E, HKU1, NL63 and OC43. To assess the extent of cross-reactive responses that can be conferred by common coronaviruses, we investigated the number of unique SARS-CoV-2 predicted epitopes shared with either one of the four coronaviruses, namely 229E, HKU1, NL63 and OC43. There were 112 unique peptides out of 794 (14%) SARS-CoV-2 predicted epitopes shared with four common coronaviruses with a high sequence similarity (Figure 4.6B). Among the 112 peptides, 21 (/794, 2.6%) had exact matches with SARS-CoV-2 peptides, while 36 (/794, 4.5%) and 55 (/794, 6.9%) peptides had one and two mismatches, respectively.

Validating predictions by epitopes deposited in IEDB

For validation of predicted epitopes, the list of predicted epitopes was compared with epitopes characterized by *in vitro* T-cell assays deposited in IEDB. All 7,869 linear epitopes were retrieved from IEDB (as of 05-05-2020), which were presented on MHC-I alleles and have human as the host organism, then were compared with the list of 4,894 unique predicted epitopes from SARS-CoV-2.

From the 7,869 IEDB immunogenic peptides, there were 34 unique peptides derived from coronavirus strains, one from OC43, two from 229E and the remaining from SARS-CoV-1. The 25 unique 9-mer predicted epitopes from four coronavirus strains, 229E, NL63, SARS-CoV-1 and SARS-CoV-2 had matching pattern with 24 IEDB peptides (Table 4.3). The matching epitopes from NL63 and SARS-CoV-2 found in our analysis were shared with 229E and SARS-CoV-1 respectively. Consequently, all 229E and NL63 predicted epitopes were matching with 229E IEDB peptides, whereas all SARS-CoV-1 and SARS-CoV-2 predicted epitopes were matching with

SARS-CoV IEDB peptides. None of our predicted epitopes matched with peptides derived from non-coronavirus microorganisms, suggesting a restricted potential for cross-reactivity between coronavirus and other virus strains.

Notably, 22 of 31(71%) SARS-derived immunogenic peptides in IEDB corresponded to our 23 SARS-derived 9-mer predicted epitopes (Table 4.2). Among the 9 false negative SARS-derived IEDB peptides, 7 were predicted to be MHC nonbinders and 2 non-immunogenic. Despite the false negatives, shortlisted predicted epitopes yielded 71% of true positives, indicating a higher probability of identifying immunogenic peptides to facilitate targeted identification and validation of vaccine candidates.

Table 4.3: Number of unique peptides by strain that have a matching pattern with immunogenic peptides in the IEDB. *Multiple 9-mer peptides may have matching pattern with a single IEDB epitope and vice versa.* *Match with 229E IEDB peptide, ˆMatch with SARS peptides

Strain	**Number of predicted 9-mer epitopes with matches in IEDB**	**Number of IEDB peptide matches (total for each strain)**
229E	2	2 (2)
NL63	1*	0 (0)
OC43	0	0 (1)
SARS	23	22 (31)
SARS-2	16ˆ	0 (0)

In a separate study, Nelde *et al.*[121] predicted immunogenicity of 1,739 SARS-CoV-2 derived HLA-I peptides using the integration of two algorithms, SYFPEITHI and NetMHCpan. They then chose 100 of 1,739 peptides for functional validation and reported 29/100 (29%) peptides as naturally occurring CD4+ or CD8+ T-cell epitopes. We compared 9-mer peptides (59 peptides) with our predictions, and found that 39/59 (66%) peptides to agree with our predictions (Table 4.4, Supplementary Table 4).

Table 4.4: Compare prediction with functional studies from Nelde *et al.* Confusion matrix illustrating predictive accuracy of our prediction by comparing $CD4^+$ or $CD8^+$ T-cell immunogenicity characterized by IFNγ ELISPOT screening. Predicted immunogenicity refers to qualitative classification of immunogenicity based on predicted score of Repitope. Predicted HLA alleles lists all HLA alleles out of 10 HLA alleles (HLA-A*0101, 0201, 0301, 2402, HLA-B*0702, 4001, 0801 and HLA-C*0702, 0401, 0701 alleles) predicted to bind the particular peptide by NetMHCpan 4.0 prediction.

	Functional validation	
Predictions	Negative	Positive
Negative	34	9
Positive	11	5
Accuracy : 0.661		
95% CI : (0.5261, 0.7792)		
Sensitivity : 0.7556		
Specificity : 0.3571		
Prevalence : 0.7627		
Detection Rate : 0.5763		
Detection Prevalence : 0.7288		
Balanced Accuracy : 0.5563		

HLA-dependent variation in SARS-CoV-2 CD8+ T-cell cross-reactivity with human coronaviruses

To further explore the correlates of cross-reactivity, the landscape of T-cells recognizing SARS-CoV-2 epitopes homologous to other human coronaviruses (HCoV) was analysed using MIRA dataset[122]. The MIRA dataset is comprised of over 160,000 SARS-CoV-2-specific T-cell receptors (TCRs) from 239 healthy individuals or 90 COVID-19 convalescent patients, and 792 unique SARS-CoV-2 peptides, 54 of which were SARS-CoV-2-HCoV homologous peptides.

To depict the landscape of public TCRs in HC and COVID-19 convalescent patients, the public TCRs that were characterised to bind epitopes from both SARS-CoV-2 and other coronavirus strains were illustrated by a bipartite graph (Figure 4.7A). The graph was annotated to indicate whether the SARS-CoV-2 epitopes were homologous or non-homologous to other human coronaviruses. This revealed two hubs, the first of which was dominated by cross-reactive TCRs that recognised both HCoV homologous and non-homologous SARS-CoV-2 peptides (green nodes), and second of which was dominated by TCRs that primarily bound non-homologous peptides. Indeed, while healthy individuals responded to a mix of HCoV homologous and non-homologous SARS-CoV-2 peptides, the majority of COVID-19 convalescent patients (51/86) were almost entirely (>99%) occupied by TCRs that recognised non-homologous peptides.

To further elucidate correlates of T-cells binding HCoV homologous epitopes from COVID-19 convalescent patients, public TCRs from 12 patients were retrieved, which predominately (>50%)

recognized homologous peptides (i.e. PubTCR-Homologous). We then sampled public TCRs from 51 patients whose majority of TCRs recognized non-homologous peptides (i.e. PubTCR-Non-homologous). By comparing HLAs between two groups, individuals in PubTCR-Homologous group were found to have statistical enrichment in HLA-B*07:02, C*07:02 and A*03:01 usage (Figure 4.7B). Notably, the systematic framework for examining specificity across all SARS-CoV-2 epitopes in the database against a large pool of HLAs aided experimental studies, which were previously limited to a small number of epitopes and HLAs. This finding was consistent with observation from Francis *et al.* where a subset of COVID-19 patients carrying specific HLA alleles (e.g. HLA-B*07:02) had pre-existing CD8+ T-cell memory to SARS-CoV-2[123].

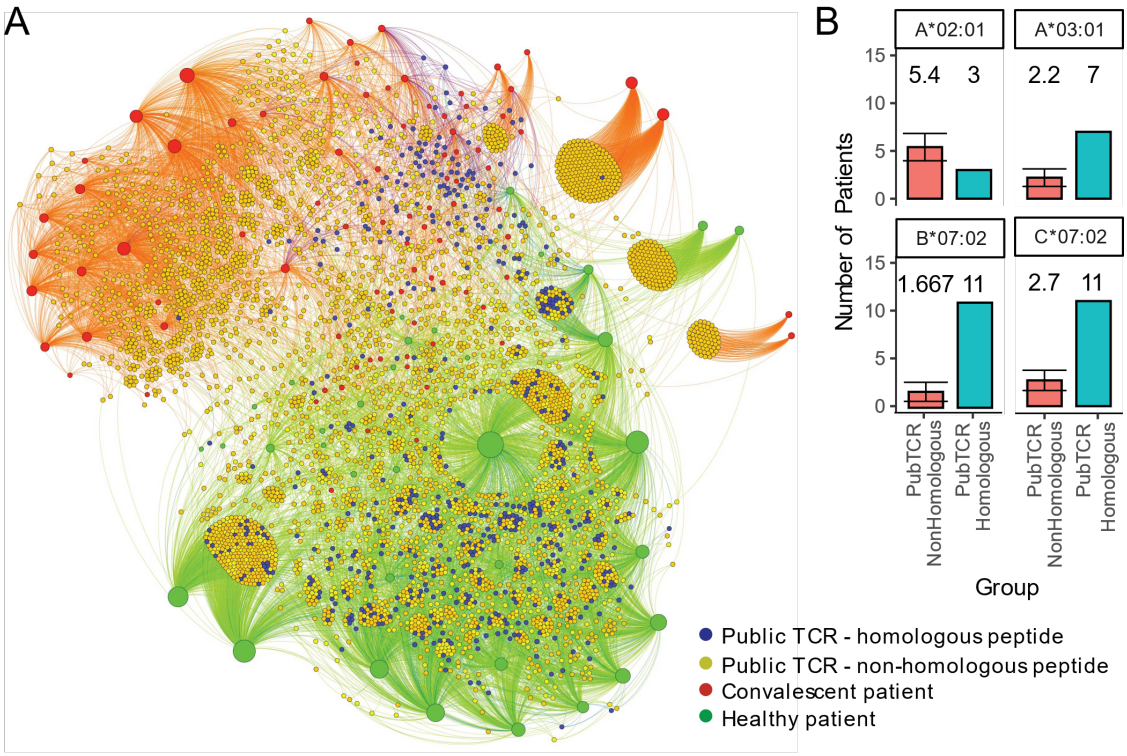


Figure 4.7: Landscape of T-cell responses against SARS-CoV-2 epitopes. A. Bipartite network graph showing public TCRs binding SARS-CoV-2 epitopes homologous or non-homologous to HCoV peptides in healthy or COVID-19 convalescent individuals. TCRs are color-coded by whether they recognize only homologous peptide (blue), only non-homologous peptides (orange) or both (yellow) from COVID-19 convalescent patients (red) or healthy individuals (green). Node size reflects degree of connectivity i.e. the quantity of an individual's TCRs that are shared with other patients. B. Barplot showing the number of COVID-19 patients carrying specified HLA-I alleles. Patients are grouped by whether their cross-reactive TCRs predominately recognize 'non-homologous' (n=12, sampled 10 times) or 'homologous' (n=12) HCoV peptides.

Potential risk of self-reactivity by SARS-CoV-2 derived predicted epitopes

Viral peptides that are identical or highly similar to those of the host organism would not elicit high-affinity T-cell responses, because most T-cells that bind with high affinity to self-peptides are eliminated during thymic negative selection. However, some of these potentially self-reactive T-cells escape thymic selection, and when primed in the context of an infection they may cause autoimmune or autoinflammatory responses[124,125]. Thus, when looking for potential candidates for vaccine design, it is important to assess the sequence similarity of predicted epitopes to their best matching counterparts in human proteome.

By comparing predicted epitopes from SARS-CoV-2 with the human proteome, none of the predicted epitopes had identical match but 10 and 184 epitopes had one and two amino acid mismatches respectively with their best matching human proteome counterparts (Figure 4.8). The predicted epitopes differing by only one amino acid are listed in Table 4.5. A number of genes in the list were associated with glycoprotein, including HTR1B, HAVCR2, HCFC1, PAM, SLC2A4, TOR1B and ZP4. With glycoprotein being one of the key molecules involved in innate and adaptive immune response, hyperactivation or malfunction of glycoproteins may trigger immunopathology, as exemplified by an autoimmune disease, rheumatoid arthritis[126].

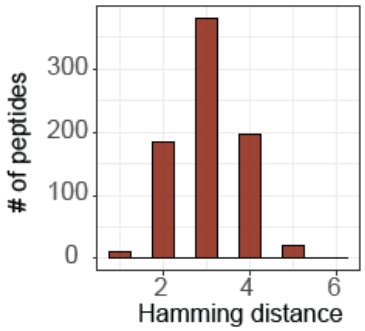


Figure 4.8: Sequence similarity with human proteome. Distribution of hamming distance between SARS-CoV-2 derived peptides and human proteome counterparts (the region most similar to corresponding virus peptides).

Table 4.5: Predicted epitopes from SARS-CoV-2 that are single amino acid variants of human proteome counterparts. SARS-CoV-2 protein and human protein denote the proteins that SARS-CoV-2 peptide and human proteome patterns are derived, respectively. Note that only the best matching counterparts for each predicted SARS-CoV-2 epitopes are listed.

SARS-CoV-2 Peptide	Human proteome pattern	SARS-CoV-2 protein	Human protein	Gene
FLALITLAT	LLALITLAT	ORF7a protein (ORF7a)	P28222|5HT1B	HTR1B
GDAALALLL	GAAALALLL	nucleocapsid phosphoprotein (N)	O14657|TOR1B, Q9BRX8|PXL2A	TOR1B, PRXL2A
GLPGTILRT	GQPGTILRT	orf1ab polyprotein (Orf1ab)	P51610|HCFC1	HCFC1

Table 4.5: Predicted epitopes from SARS-CoV-2 that are single amino acid variants of human proteome counterparts. SARS-CoV-2 protein and human protein denote the proteins that SARS-CoV-2 peptide and human proteome patterns are derived, respectively. Note that only the best matching counterparts for each predicted SARS-CoV-2 epitopes are listed.

SARS-CoV-2 Peptide	**Human proteome pattern**	**SARS-CoV-2 protein**	**Human protein**	**Gene**
GLTVLPPLL	GLTVLPALL	surface glycoprotein (S)	P14672|GLUT4	SLC2A4
IPIGAGICA	IYIGAGICA	surface glycoprotein (S)	Q8TDQ0|HAVR2	HAVCR2
IVNSVLLFL	TVNSVLLFL	envelope protein (E)	O60518|RNBP6	RANBP6
QLSLPVLQV	QLLLPVLQV	orf1ab polyprotein (Orf1ab)	Q8IWE2|NXP20	FAM114A1
SLPINVIVF	SLPINVQVF	orf1ab polyprotein (Orf1ab)	Q12836|ZP4	ZP4
TPGSGVPVV	EPGSGVPVV	orf1ab polyprotein (Orf1ab)	P19021|AMD	PAM
VLPQLEQPY	VLPQNEQPY	orf1ab polyprotein (Orf1ab)	A2A3K4|PTPC1	PTPDC1

Discussion

This study presents our initial efforts to produce a comprehensive *in silico* map of immune targets against SARS-CoV-2 at the beginning of the pandemic. In a series of studies, we i) shortlisted SARS-CoV-2 peptides based on their similarity to previously characterized epitopes, ii) conducted *de novo* epitope searches using the existing immunogenicity algorithms, and iii) evaluated the potential for cross-reactivity with other coronaviruses.

Furthermore, while COVID-19 appears to be a response to SARS-CoV-2 epitopes that are not homologous to HCoV peptides, patients with certain HLA alleles (e.g. HLA-B*07:02, -C*07:02, -A*03:01) were found more likely to possess cross-reactive TCRs capable of binding both SARS-CoV-2 and other human coronaviruses. It is therefore possible that SARS-CoV-2 naïve individuals with certain HLAs be at a lower risk of severe disease or experience enhanced vaccine response if previously been exposed to endemic coronaviruses. However, a direct link to pre-existing immunity requires further investigations.

Along with the *in silico* profile of public and private epitopes, the potential risk of self-reactivity imposed by high sequence similarity to the human proteome was evaluated. Considering the reports of immunopathology causing lung, heart, liver, intestine, genital and kidney failures in COVID-19 patients[127], peptides from SARS-CoV-2 may carry high risk of immunopathology and should be carefully selected before proceeding with vaccination.

In this study, the potential immunogenicity was determined by peptide sequence alone but other factors, such as turnover rate and expression level of parent proteins, may contribute to the selection of immunodominant epitopes. For example, Bojkova *et al.* characterized changes in the protein level of SARS-CoV-2 infected Caco-2 cells, and showed 3-6 folds (in log2) increase in protein expressions after 20h compared to 2h post infection, with nucleocapsid and membrane proteins having the highest and the replicase polyprotein 1ab having the lowest fold change, respectively[128]. Therefore, although the membrane, nucleoprotein, and spike proteins had fewer predicted epitopes than the replicase polyprotein 1ab, their higher abundance may have a significant impact on determining their immunodominance and skewing the immune response against their epitopes.

In regards to the accuracy of predictive models used in this study, the algorithms for predicting MHC presentation have matured significantly over the last decade thanks to extensive datasets, especially for the most prevalent HLA types. On the other hand, predicting immunogenicity is a challenging and unresolved problem. These algorithms are trained on a small dataset and do not consider other factors influencing the immunogenicity of a given epitope, such as abundance, expression patter and localization. As a result, even though the best model was used for classifying immunogenicity, the predictions were suboptimal and should be interpreted with caution.

Moreover, the classification thresholds of these algorithms have intrinsic trade-off between sensitivity and specificity. In the case of SARS-CoV-1 derived peptides, strong binding of HLA-A*02:01 has been shown to correlate with stimulation of IFNγ secretion in the PBMC

from HLA-A2 SARS-recovered donors *ex vivo*[129,130], but we cannot rule out the possibility of immunogenicity from low-binding SARS-CoV-2 peptides. Furthermore, adding a filter to predict T-cell recognition increases the risk of false negatives from HLA ligands. While *in silico* map was generated based on NetMHCpan and Repitope's recommended thresholds, more stringent or relaxed thresholds can be applied depending on the availability of resources for functional validations.

To validate *in silico* predictions, the predicted epitopes were compared with those characterized by *in vitro* T-cell assays deposited in IEDB, and 22 out of 31 SARS-CoV-1 IEDB epitopes were found to match our predictions. Because coronavirus-related peptides in IEDB have varying lengths of up to 25aa, multiple 9-mer epitopes may match with a single peptide in IEDB.

To estimate the false negative rate, we identified epitopes that did not pass our filters. 23 out of 43732 predicted non-binders and 2 out of 11202 predicted binder-non-immunogenic peptides matched with 10 coronavirus-derived epitopes in IEDB (9 from SARS-CoV and 1 from OC43). While the presence of multiple 9-mers matching with the same IEDB epitope is not surprising, exclusion of some matches by the MHC binding and immunogenicity filtering may facilitate determining the regions of peptide that are most likely to elicit the T-cell response.

The accuracy was evaluated by comparing the predictions to functionally validated SARS-CoV-2 epitopes from Nelde *et al.*, and showed that 66% of the peptides agreed with the predictions. Notably, Nelde *et al.* also demonstrated presence of T-cell responses in unexposed individuals to 31% of their validated HLA class I epitopes, indicating the presence of CD8+ T-cell cross-reactivity with common coronaviruses. In agreement with our prediction, Orf1ab was the protein most recognized by the class I epitopes they found in healthy donors, while membrane (M) and nucleocapsid (N) proteins were not recognized. This points to the potential cross-reactivity with other coronavirus strains and suggests the significance of conceptualizing the map of cross-reactive potential among coronavirus strains.

We believe that the comprehensive profile of predicted SARS-CoV-2 epitopes and *in silico* assessment of their self-reactivity and cross-reactivity assisted biologists with targeted function validation, and facilitated the design of vaccines capable of protecting against multiple prevalent virus strains. These studies were followed up by other studies[131–133] and contributed to better understanding of heterogeneity in the susceptibility and severity to the disease.

This work highlights the importance of the ability to predict CD8+ T-cell targets not only for shortlisting vaccine candidates but also for better understanding disease pathology. However, there is currently a knowledge gap in understanding the characteristics that underpin peptide immunogenicity for accurate CD8+ T-cell target prediction. With numerous efforts made to address this challenge, a systematic and unbiased framework to evaluate the success of these studies are discussed in the following chapters.

Methods

Retrieval of coronavirus proteome sequences

The proteome sequences of coronavirus strains were obtained from NCBI. The reference numbers for these sequences are NC_002645.1 (229E), NC_006577.2 (HKU1), NC_019843.3 (MERS-CoV), NC_005831.2 (NL63), NC_006213.1 (OC43), NC_004718.3 (SARS-CoV) and NC_045512.2 (SARS-CoV-2).

Sequence alignment and phylogenetic tree of encoded proteins in coronaviruses

For multiple sequence alignment and phylogenetic tree generation, the open reading frames were grouped by their encoded proteins – spike protein (S), envelope protein (E), membrane protein (M), nucleocapsid protein (N), replicase polyprotein (Orf1ab) and other encoded regions (Other) – prior to analysis. The multiple sequence alignment was conducted using 'msa' function from R msa v1.4.3 package and visualized by 'msaPrettyPrint' function. The phylogenetic tree was produced by plotting 'identity' distance generated by 'dist.alignment' function from R seqinr v3.6.1 package.

Sequence similarity with human proteome and epitopes deposited in IEDB

The sequence similarity of peptides derived from SARS-CoV-2 with human proteome counterparts and IEDB peptides was computed by first identifying best global-local alignment by 'pairwiseAlignment' function from Biostrings v2.40.2 package with a high gap penalty, gapOpening of 100 and gapExtension of 100. The number of mismatches between best aligned pair was computed by hamming distance using 'stringdist' function from R stringdist v0.9.5.5 package.

Peptide generation from proteome sequences

Each encoded proteins were fragmented into 9-mer peptides by scanning the proteome with a window of 9 amino acids and step length of 1 amino acid. For strains containing two open reading frames annotated with the same functional protein e.g. 229E, MERS-CoV, SARS-CoV and SARS-CoV-2 having two open reading frames annotated for Orf1ab, 9-mer peptides were generated from both encoded proteins and unique set of peptides were selected for subsequent analysis.

MHC presentation prediction

The antigen presentation of MHC was predicted using NetMHCpan v4.0[119], a model trained on binding affinity and eluted ligand data, against HLA-A*0101, 0201, 0301, 2402, HLA-B*0702, 4001, 0801 and HLA-C*0702, 0401, 0701 alleles. Peptides with rank score <= 2.0 were categorized as positive HLA-binder.

Immunogenicity Prediction

The Repitope package was retrieved from GitHub repository (https://github.com/masato-ogishi/Repitope.git). After feature computation and feature selection, the published Repitope

'MHCI_Human' model was used to extrapolate probabilistic immunogenicity scores for our dataset and made binary classification of immunogenicity.

SARS-CoV-2 specific TCR repertoire data (MIRA dataset)

The COVID-19 MIRA dataset (>160k high-confidence SARS-CoV-2-specific TCRs) was downloaded from https://clients.adaptivebiotech.com/pub/covid-2020. The data contains TCR repertoire against SARS-CoV-2 epitopes from COVID-19 convalescent patients and healthy individuals.

Network analysis of cross-reactive TCRs recognizing SARS-CoV-2 homologous and/or non-homologous to HCoV

A public TCR is defined as TCRs having same CDR3 sequence and V and J gene observed in more than one patient in the MIRA dataset. All graphs were first generated using iGraph in R, exported to Cytoscape using the createNetworkFromIgraph function in the RCy3 package. From cytoscape, all graphs were exported as .graphml files and read into Gephi. In Gephi, either 'ForceAtlas' and 'Fruchterman-Reingold' templates were used. In almost all cases, gravity and repulsion parameters were adjusted to improve visual aesthetics.

Chapter 5. Evaluate performance of existing computational methods in predicting CD8+ T-cell cancer and pathogenic epitopes

Original publications:

- **Lee CH**, Salio M, Napolitani G, Ogg G, Simmons A, Koohy H. Predicting Cross-Reactivity and Antigen Specificity of T-cell Receptors. Front. Immunol. 11, (2020).
- Buckley, PR*, **Lee, CH***, Ma, R*, Woodhouse I, Woo J, Tsvetkov VO, Shscherbinin BS, Antanaviciute A, Shugay M, Rei M, Simmons A, Koohy H. Evaluating performance of existing computational models in predicting CD8+ T-cell pathogenic epitopes and cancer neoantigens. Briefings in Bioinformatics. *These authors contributed equally.

Introduction

The dual nature of the peptide specific TCR recognition interface, comprised of both peptide and MHC, makes predicting interaction between TCR and pMHC uniquely challenging. The T-cell specificity is largely determined by the promiscuous peptides due to a relatively invariant interaction with MHC molecule[134,135]. Additionally, TCR:pMHC recognition is influenced by peptide length, physicochemical properties, peptide sequence especially at contact positions (i.e. P3-P9 of 9aa peptide), MHC haplotype and structural landscape [136]. Over the years, TCR:pMHC interactions have been extensively studied, providing a wealth of data for modelling from various perspectives.

In recent years, a number of computational algorithms have been developed for predicting peptide immunogenicity. These models utilized sequence-based, biophysical and structural features underlying TCR:pMHC interaction, and reported ROC-AUC values as high as 0.85. With a wide range of applications ranging from identifying targets for cancer immunotherapies to vaccine development, there has been a growing interest in using such algorithms to shortlist candidates for functional validations.

Despite the greater interest, we previously observed that many existing immunogenicity algorithms show suboptimal performance. Moreover, no systematic and unbiased evaluation of the performance of these models has been performed in the context of emerging pathogens and cancer neoepitopes. With previous work demonstrating the limitations of some models in predicting SARS-CoV-2 epitopes, we aim to benchmark and guide the use of existing algorithms in various immunological settings.

This chapter begins with a literature review of computational models for predicting CD8+ T-cell epitopes. The performance of widely used and publicly available immunogenicity algorithms is evaluated for predicting cancer and pathogenic epitopes. The existing models performed poorly in both settings, indicating significant room for improvement.

Results

Computational strategies to identify CD8+ T-cell epitopes

A number of algorithms have been developed to predict immunogenicity of peptides based on features, such as amino acid compositions, positional information, and physicochemical, biophysical and structural properties (Table 5.1, Figure 5.1). We describe the existing models that have been developed based on sequence, biophysical, kinetic and structural characteristics of TCR:pMHC interactions.

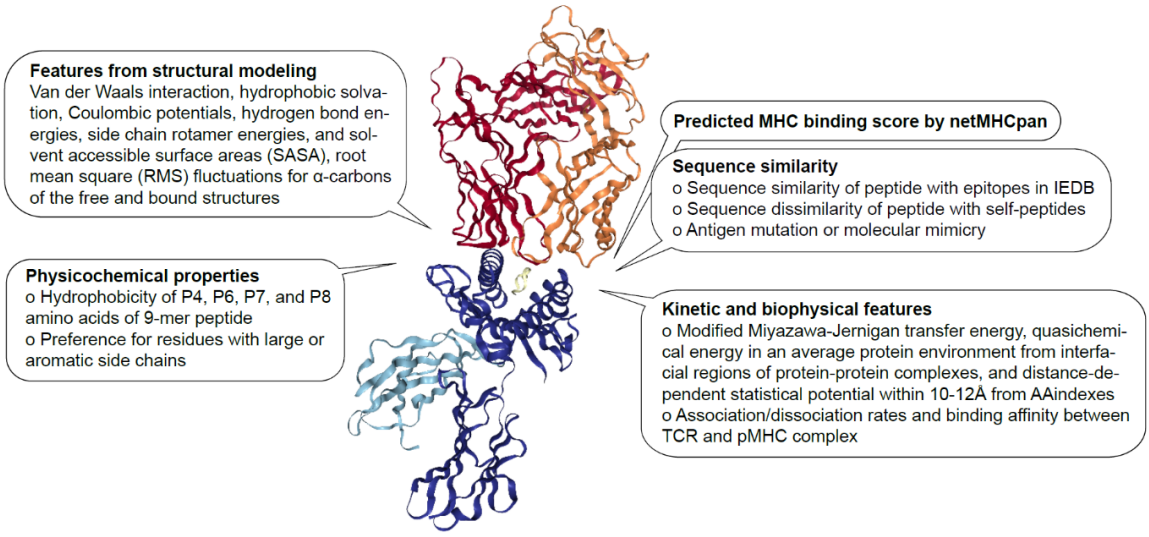


Figure 5.1: Features associated with TCR:pMHC interaction. Description of sequence-based, structural, kinetic and biophysical features previously found to be associated with pMHC recognition by TCR The diagram is 1G4 TCR bound to NY-ESO-1/HLA-A∗02:01 (PDB 2BNR) where TCRα, TCRβ, MHC, b2-microglobulin and peptide are coloured in orange, red, blue, light blue and yellow, respectively

Using sequence-based features

Most widely used immunogenicity models employ sequence-based features. Chronologically, Calis *et al.* developed IEDB model based on the observation that contact positions (P4-P6 in 9aa peptide) and amino acids with large and aromatic side chains play important roles in peptide immunogenicity[49]. Then, Trolle *et al.* integrated predictions from pMHC binding affinity, pMHC stability and T-cell propensity to improve immunogenicity prediction beyond MHC binding (named NetTepi)[53]. In subsequent studies, Pogorelyy *et al.* utilized physicochemical properties of amino acids by transforming peptide sequence into a vector of Kidera factors (named iPred)[54]. Similarly, Ogishi *et al.* incorporated amino acid index database (AAindex)[137] and other physicochemical properties, together with TCR-peptide contact potential profiling (CPP)-based features into a linear coordinate system through a machine-learning technique known as Extremely Randomized Tree (ERT) algorithm[50] (named Repitope).

Recently, Wells *et al.* and Gao *et al.* incorporated low agretopicity (ratio in MHC binding affinity between cancer neoepitope and human proteome), high foreignness score (similarity to pathogenic epitopes) and low similarity to human proteome as correlates of immunogenicity (named TESLA and GAO respectively)[52,138]. In the latest study, Li *et al.* encoded peptide and HLA sequences using BLOSUM50 substitution matrix, and adopted convolutional neural network (CNN) to predict immunogenicity potential of peptide-MHC pairs (named DeepImmuno)[51]. This model has delivered the best results to date, with ROC-AUC of 0.83-0.86.

Because of their ease of use and greater breadth of peptides applicable for prediction, sequence-based models are now more widely used in a variety of applications. While models using biophysical, kinetic and structural properties have started to emerge, they remain limited to restricted number of peptide-MHC-TCR interactions. As such, our work focuses on sequence-based models.

Using biophysical and kinetic features

In addition to models incorporating sequence information, biophysical studies revealed that some interactions at the pMHC surface appear to be more important in triggering a T-cell responses[134,139]. This raised a hypothesis that even TCRs sharing a similar TCR footprint may have their own "interaction profile" [140], and claimed that while conventional hotspots were attributed to amino acid sequences, the concept of hotspot should be expanded to account for interaction features, such as hydrogen bonds, van der Waals forces, pockets and coordination of water molecules [139,141].

A collective effort has identified biological and physical parameters that modulate TCR:pMHC engagement and T-cell response (reviewed in Spear *et al.*[142]), which include binding affinity, association and dissociation rates, half-life of interaction, changes in heat capacity, competition for CD3 molecules and conformational adaptability of TCRs[142–157]. Taking these biophysical and kinetic features into account may help to effectively reflect the two-dimensional (2D) and dynamic *in vivo* system whilst modelling TCR:pMHC interactions.

Although we are not currently in the position to perform an *ab initio* prediction based on biophysical and/or structural data, recent studies have started to incorporate physical and chemical properties in developing a generalized TCR scoring function. Riley *et al.* used a multi-linear regression model with 16 full-atom Rosetta terms [158,159] and six flexibility terms as predictor variables and experimental binding energies as response variables to estimate the effect of point mutations on thermodynamic stability of the TCR. They identified six significant features, four structural terms (van der Waals attractive and repulsive forces, solvation energies and sidechain hydrogen bonding) and two flexibility terms (root mean square (RMS) fluctuations for α-carbons of the free and bound structures), all of which contributed to the improved performance of the scoring function.

In another application, Haider *et al.* aimed to engineer an affinity enhanced A6 TCR targeting Tax peptide/HLA-A2 complex[16]. They created a set of 219 fitted scoring functions using kinetic and potential energy terms and identified a function (named ZAFFI score) that was best capable

of reproducing affinity changes upon 648 mutations on ovomucoid turkey inhibitor molecule. Pierce *et al.* then developed an improved scoring function (ZAFFI 1.1) having a higher correlation for a set of DF5 point mutations[19]. ZAFFI 1.1 includes six terms: van der Waals attractive and repulsive components, desolvation, intra-residue clash, hydrogen bonding and Coulombic electrostatic force.

Using structural features

The 3D crystal structures of T-cell receptor and their cognate peptide-MHC (pMHC) complexes have been resolved and deposited in Protein Data Bank (PDB) [160]. The structural T-cell receptor database (STCRDab) contains >340 PDB entries and >480 $\alpha\beta$TCR structures and facilitates analysis and visualization of TCR structures[161]. This database includes information about MHC type, antigen specificity, Vα-Vβ pairing, orientation, dissociation rate (Kd) and CDR type. Additionally, TCR3d provides information on germline gene usage, antigen binding mode and interface features[162].

Based on the cognate peptide, MHC and TCR structures in the aforementioned database, there have been a number of attempts to predict peptide-MHC conformations, including docking algorithms[163,164], protein threading[165], all-atom molecular dynamics (MD) simulations [166–168], energy minimization[169] and hybrid of these approaches[170]. Similarly, MD or Monte Carlo simulations, TCR:pMHC hydrogen bond network analysis[171,172], binding free energy simulation[173] and CDR loop characterization[155] have been used to model pMHC-TCR.

Furthermore, the features retrieved from structural modelling were utilized to predict TCR:pMHC complex formation[174,175]. Mendes *et al.* performed structure-based prediction on viral epitopes. They demonstrated that accessible surface area (ASA) is predictive of immunogenicity by using structural features as input for a multivariate statistical model. Similarly, Riley *et al.* demonstrated that hydrophobic SASA and hydrophobic solvation energy values at peptide positions 5, 7 and 8 were in the top 10% of all weights in the neural network for predicting immunogenicity [176].

The models using biophysical, kinetic and structural properties remain limited to restricted number of peptide-MHC-TCR pairs. For example, with <500 pMHC:TCR 3D structures deposited in the database skewed towards prevalent HLAs and viral or cancer epitopes, it is difficult to train model that is robust to a wide range of peptides and HLAs. Further biophysical, kinetic and structural characterisation of TCR:pMHC complexes will allow advances in computational methods for identifying T-cell epitopes.

Table 5.1: List of algorithms to predict immunogenicity

Reference	Training Data	Algorithm	Discriminative Features (Immunogenicity)
Tung, C.-W. *et al.* [177] (POPISK)	Trained on 9-mer HLA-A2 restricted peptides. From MHCPEP, SYFPEITHI and IEDB, consist of 558 immunogenic, 527 non-immunogenic peptides	Decision tree learning methods to identify informative physicochemical properties from 531 physicochemical properties retrieved from version 9.0 of amino acid index (AAindex) database. Support vector machine with a weighted string kernels for immunogenicity prediction.	Top AAindex contributors: i) Retention coefficient in HPLC, pH2.1, ii) Principal property value z2, iii) Hydrophobicity scale from native proteins, iv) Normalized composition of membrane proteins, and v) pK-C. Found positions 4, 6, 8 and 9 critical for 9-mer peptide
Calis, J. J. A. *et al.* [49] (IEDB)	Trained on 9-mer from MHC-I associated peptides. From IEDB and three immunogenicity studies in mice (Assarsson *et al.*[178], Kotturi *et al.*[179]. and unpublished data on Coxiella Burnetti-derived peptides), consist of 600 immunogenic and 181 non-immunogenic peptides	Per non-anchor residue of the presented peptide, log enrichment score calculated as ratio between the fraction of specific amino acid in immunogenic vs. non-immunogenic data, then score weighted to the importance of that position measured as Kullback-Leibler divergence. The weighted log enrichment scores of all (non-anchor) residues summed as immunogenicity score.	Preference for residues with larger or aromatic side chains Positions 4-6 critical for 9-mer peptide
Trolle, T. & Nielsen, M. [53] (NetTepi)	Trained on 9-mer peptides covering 9 HLA alleles. From 295 T-cell epitopes from SYFPEITHI and 1216 T ell epitopes from IEDB, allele-balanced training data created by randomly selecting 50 epitopes from each of 9 HLA alleles except 2 alleles having 14 epitopes each, Total 378 epitopes.	Weighted sum of pMHC binding affinity (NetMHCcons[180]), pMHC stability (NetMHCstab[181]) and T-cell propensity prediction[49]. Optimal relative weights obtained.	Performance gain obtained by summing pMHC binding affinity, pMHC stability predictions and T-cell propensity than individual predictions.

Table 5.1: List of algorithms to predict immunogenicity

Reference	Training Data	Algorithm	Discriminative Features (Immunogenicity)
Chowell, D. *et al.* [182]	Trained on 9-mer H-2D^b and HLA-A2 restricted peptides (separately for two ANN-Hydro models). From IEDB, 204 immunogenic and 232 nonimmunogenic (self-peptides from MHC ligand elution experiment with no known immunogenicity) for H-2D^b, and 372 immunogenic and 201 nonimmunogenic peptides for HLA-A2.	Hydrophobicity-based artificial neural network (ANN-Hydro) based on numeric sequence of amino acid hydrophobicity.	Strong bias towards hydrophobic amino acids at TCR contact residues (P4, P6, P7 and P8 for 9-mers) within immunogenic epitopes. Negative correlation between polarity of amino acids and immunogenicity.
Łuksza, M. *et al.* [183]	Trained on 2552 MHC-I immunogenic peptides from IEDB. Neoantigens with mutations generated from non-hydrophobic, wild-type residues at positions 2 and 9 excluded (as prediction of MHC affinities for wild-type peptides with non-hydrophobic anchor residues led to non-informative amplitudes).	Recognition potential of a neoantigen = A x R, where amplitude (A) is relative probability that a neoantigen is presented on MHC-I whereas its wild-type counterpart is not, and R is probability that neoantigen will be recognized by TCR repertoire. R defined by a multistate thermodynamic model, treating sequence similarity as proxy for binding energies.	High sequence similarity of a given neoantigen with epitopes in IEDB by gapless alignment with BLOSUM62 amino acid similarity matrix
Bjerregaard, A.-M. *et al.* [184]	From 13 publications, analysed total 1948 peptide-HLA complexes, of which 53 reported immunogenic.	HLA binding prediction by NetMHCpan-4.0. Similarity between each neo- and normal peptide using kernel similarity measure proposed by Shen *et al.*[185].	High predicted binding score (HLA binding strength). Peptide sequence dissimilarity to self (wild-type counterpart of the neopeptide), especially for those with comparable HLA binding.
Pogorelyy, M. V. *et al.* [54] (iPred)	Trained on 9-mer peptides. From Chowell *et al.*[182], 3671 immunogenic and 3911 non-immunogenic peptides	Principal component analysis and dimensionality reduction on 10-dimensional vectors of Kidera factor sums for each epitope. Fit multinomial Gaussian model using expectation maximization to estimate probability of being immunogenic.	Distinct physicochemical properties in Kidera space

Table 5.1: List of algorithms to predict immunogenicity

Reference	Training Data	Algorithm	Discriminative Features (Immunogenicity)
Jurtz, V. I. *et al.* [186] (NetTCR)	Trained on 8,920 TCRβ CDR3 sequences and 91 HLA-A2 cognate peptides obtained from IEDB. 379 TCR and 16 peptides from the MIRA assay in Klinger *et al.*[187] Negative data from eluted peptide ligands from self (i.e. human) proteins, a set of 200,000 TCR CDR3 sequences from 20 healthy donors and creating internal incorrect combinations of TCRs and peptides.	Convolutional neural networks (CNN) to predict whether a given TCR is able to recognize a specific peptide, with amino acid sequences of peptide and CDR3 region of TCRβ chain as input. CNNs scans the input and detects pattern to be integrated into network.	Conserved sequence patterns of peptide-TCR pairs encoded by BLOSUM50 matrix
Smith, C. C. *et al.*[188]	Trained on 8-11mer 141 epitopes from MHC-I $H2^b$ and $H2^d$ haplotypes	Using amino acid features (tiny, small, aliphatic, aromatic, non-polar, polar, charged, basic and acidic), variables derived by presence/absence of each feature at each absolute and relative position, at site of SNV mutation, at being/middle/end residues, difference of each feature in mutated vs. reference antigen. Most predictive features into gradient boosting algorithm and trained by 10,000-fold cross-validation.	Peptide biochemical features: valine at position 1, valine at last position, small amino acids at the last position, basic amino acids of the reference at the mutated position, changes in the mutated position to a small amino acid, lysine at relative site 1, and presence of valine within the first 3 positions.

Table 5.1: List of algorithms to predict immunogenicity

Reference	Training Data	Algorithm	Discriminative Features (Immunogenicity)
Ogishi, M. & Yotsuyanagi, H. [50] (Repitope)	Trained on 8-11mer MHC-I and 11-30mer MHC-II peptides. From IEDB, LANL HIV and HCV database and TANTIGEN database, 6,957 HLA-I and 16,642 HLA-II immunogenic peptides. 191,326 TCR CDR3β sequences obtained from MiXCR.	TCR-peptide contact potential profiling (CPP) by optimal alignment between CDR3β (randomly down-sampled to 10,000 sequences) and peptides and using pairwise contact potential scales from AAindex. Peptide sequence-based estimates of physicochemical properties (= peptide descriptors) using: *aIndex, blosumIndices, boman, charge, crucianProperties, fasgaiVectors, hmoment, hydrophobicity, instaIndex, kideraFactors, mswhimScores, pI, protFP, chseScales and zScales*. Most predictive peptide descriptors and CPP features compressed into a linear coordinate system through extremely randomized tree (ERT) algorithm.	Physicochemical and CPP features: features from short (3-aa and 4-aa) and longest (8-aa and 11-aa for MHC-I and MHC-II respectively) fragments, skewness- and kurtosis-derived features and AAindexes, including inverse of modified Miyazawa-Jernigan transfer energy, inverse of quasichemical energy in an average protein environment from interfacial regions of protein-protein complexes, and distance-dependent statistical potential within 10-12Å
Riley, T. P. *et al.*[176]	Trained on 9-mer HLA-A2 restricted peptides. 155 immunogenic from IEDB, 2756 HeLa HLA-A2 binding self-peptides and 1044 HLA-A2 nonbinders.	A feed-forward neural network with inputs describing structural and structure-based energetic features of 9aa in peptide sequence and peptide-HLA complex. Structural and energy features are those comprising Talaris2014 energy function.	Structural and energic features: van der Waals interaction, hydrophobic solvation, Coulombic potentials, hydrogen bond energies, side chain rotamer energies, and solvent accessible surface areas (SASA).

Table 5.1: List of algorithms to predict immunogenicity

Reference	Training Data	Algorithm	Discriminative Features (Immunogenicity)
		$A(s,M) = \frac{1}{binding(s,M)} \times R(s,\text{pathogen}) \times R(s,\text{self})$	
Gao *et al.* [138] (GAO)	Foreign peptides retrieved from IEDB by the following search criteria: "Epitope: Linear; Assay: T Positive; MHC Restriction: MHC I; Disease: Infectious disease; Host: Human", excluded HIV peptides to avoid bias in training and testing the model. Self-peptides retrieved by criteria: "Epitope: Linear; Assay: MHC Ligand Assay Positive; Organism: Homo Sapiens"	classifier based on 'CTL response metric' A(s, M), where s is the sequence of the test peptide and M respective HLA. R describes sequence similarity of peptide s to either pathogen-derived or self-peptides. Peptide having A(s, M) score greater than threshold classified as immunodominant peptide (e.g. threshold for immunodominant SARS-CoV-2 peptide as 7024).	Incorporate HLA Binding probability, similarity to pathogenic peptides and similarity to human proteome.
Schmidt *et al.* [189] (PRIME)	Immunogenic (n = 1,282) and non-immunogenic (n = 3,676) peptides collected from recent studies, and comprise both pathogen and cancer testis antigens as well as cancer neo-epitopes.	Logistic regression encoding each pHLA pair as 21-dimensional vector consisting of predicting HLA-I affinity and frequencies of 20 amino acids at MIA positions (i.e. minimal impact on HLA-I alleles).	Amino acid frequency at contact positions
Li *et al.* [51] (DeepImmuno)	8971 peptides from IEDB for training (linear, T-cell assays, MHC-I, human and any disease, with 4-digit HLA alleles, discarded redundant peptides), 9-10 amino acids. Validated using 608 neoepitopes from TESLA study and 100 SARS-CoV-2 peptides.	Convolution-neural network on BLOSUM50-encoded peptide and HLA sequences, followed by two fully connected dense layers to output predictive value for each pair.	Peptide and HLA sequence encoded by BLOSUM50

Table 5.1: List of algorithms to predict immunogenicity

Reference	Training Data	Algorithm	Discriminative Features (Immunogenicity)
Wells *et al.* [52] (TESLA)	pMHC set for recognition features comprised of 12 immunogenic and 17 non-immunogenic peptides.	Fisher's exact test and ranking of peptides by MHC binding and T-cell recognition features.	High MHC binding affinity and long half-life, low agretopicity (ratio of mutant binding to wild-type binding affinity) or high foreignness score (TCR recognition probability derived from homology to known pathogenic peptides in IEDB).

Table 5.2: Examples of reasons for excluding models from benchmarking

Model	Reason for exclusion
Riley, T. P. *et al.* [176]	• To our knowledge, the model is not publicly available. • Limited breadth of HLAs available to predict epitopes, with greater focus on HLA-A*02:01.
Jurtz, V. I. *et al.* [186]	• Predicts interaction of HLA-A*02:01-bound peptides with specific CDR3 sequences. Out of scope for the present study.
Tung, C. W. *et al.* [177]	• Unable to access the webserver or locate the model

Evaluate performance of existing models in predicting pathogenic epitopes

To conduct an unbiased evaluation of existing immunogenicity models, we gathered all publicly available models and excluded those that were unable to perform a systematic assessment (Table 5.2). The following models were tested for their ability to predict whether MHC-presented peptides elicit T-cell response (i.e. are immunogenic): IEDB model[49], NetTepi[53], iPred[54], Repitope[50], PRIME[189], DeepImmuno[51] and Gao[138]. In addition to T-cell response models, as Paul *et al.* found that a large proportion of HLA-presented vaccinia virus peptides were immunogenic, NetMHCpan 4.0[60] was included as a highly accurate MHC binding classifier to assess its predictability in human setting.

To benchmark the performance of these models, SARS-CoV-2 peptides that had been functionally characterised by T-cell assays were collected. The SARS-CoV-2 dataset has distinct advantages. First, ongoing efforts to understand T-cell response against COVID-19 allowed characterization of comprehensive list of peptides from SARS-CoV-2 proteome, providing mega-pool of peptides. Second, with an exception of peptides sharing high sequence homology with other coronavirus e.g. MERS-CoV or SARS-CoV, majority of peptides have not been included in the training dataset

due to their recent emergence. These peptides are therefore ideal test data for unbiased evaluation of models trained on partially different datasets in predicting epitopes from an emerging pathogen, enabling a benchmark of models in a realistic setting. Thus, the capability of these models to accurately predict CD8+ T-cell epitopes was assessed upon the emergence of a novel pathogen.

The HLA-I presented SARS-CoV-2 peptides were retrieved from IEDB and VIPR dataset[190] (both accessed on 07-10-2021). Two filters were then applied. First, we limited ourselves on peptides having 9 or 10 amino acids in length, due to low number of peptides of other lengths. Second, only peptides presented on 13 HLA alleles were included for which all models were applicable. The MERS-CoV or SARS-CoV-1 identical homologs were retained in the dataset despite the possibility of being included in the training dataset to reflect real-life scenario. Application of these filters left 858 SARS-CoV-2 peptides, of which ~63% were immunogenic and ~37% non-immunogenic.

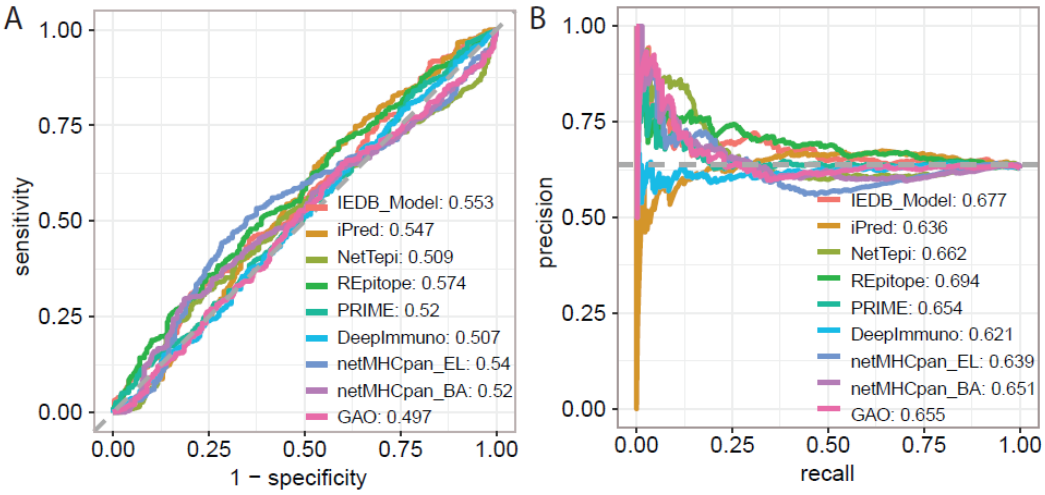


Figure 5.2: Models are not reliable in predicting epitopes for an emerging virus and exhibit room for improvement in predicting immunogenic cancer neoantigens. A. ROC-curves of models tested 'as published' against 858 SARS-CoV-2 peptides of lengths 9 and 10. B. PR-curves of models tested 'as published' against the SARS-CoV-2 peptides.

The performance of aforementioned 8 models was evaluated on the SARS-CoV-2 dataset. The model performance was assessed using both receiver operating characteristic area-under-the-curve (ROC-AUC) and precision-recall area-under-the-curve (PR-AUC). The PR-AUC metric was included to reflect unbalanced classes in the test data by tailoring detection of minority class. The ROC-AUC in predicting epitopes from SARS-CoV-2 data ranged from 0.497 from Gao to 0.574 from Repitope (Figure 5.2A), suggesting suboptimal performance of all models in predicting CD8 T-cell targets from emerging viruses. Similarly, PR-AUC ranged from 0.621 to 0.694 from DeepImmuno and Repitope, respectively (Figure 5.2B). Given that ~63% of data is immunogenic, the models performed marginally better than random.

In a subsequent bootstrapping analysis, the distribution of PR-AUC scores from a dataset with the original label (i.e. benchmarked PR-AUCs) was compared to that with a randomly shuffled

label (i.e. shuffled PR-AUCs) to assess the models' predictive capacity against the test dataset. The z-scores illustrate the deviations in the distribution between true and random predictions. Indeed, bootstrap analysis revealed that the majority of models did not outperform the baseline, with only Repitope (ROC-AUC z-score = 3.69, PR=AUC z-score = 3.87) and the IEDB model (ROC-AUC z-score = 2.54, PR-AUC z-score = 2.88) outperforming random predictions (Figure 5.3).

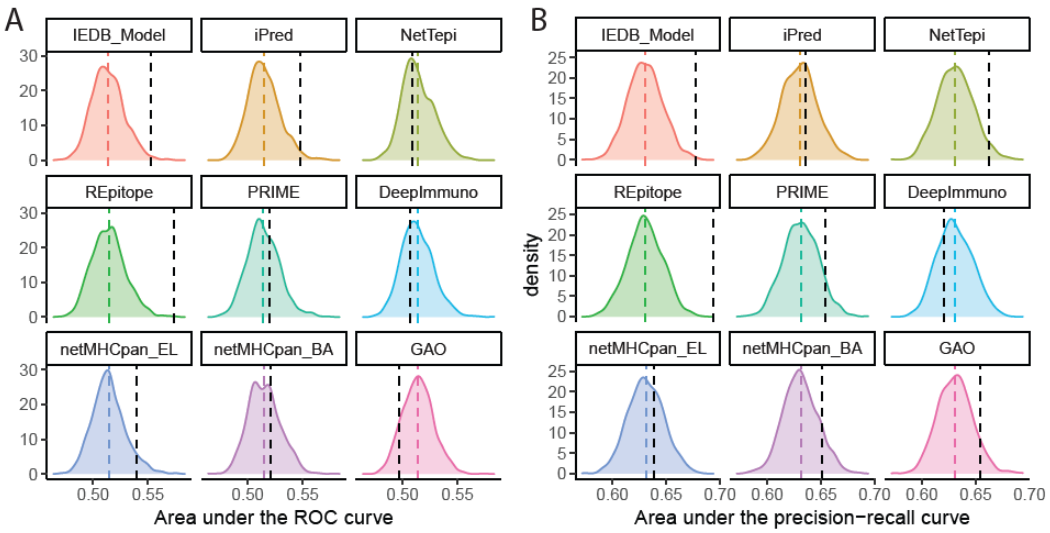


Figure 5.3: Majority of the immunogenicity models do not outperform random prediction by bootstrapping. Density plots showing model performance measured by ROC-AUC (A) and PR-AUC (B), relative to a bootstrapped random distribution. The scores are sampled 1000 times from the datasets with the original label or randomly shuffled label. Black line shows the 'true' ROC-AUC (A) or PR-AUC (B) measured from the performance evaluation. Coloured lines show the mean of the randomly generated distribution reflecting the background.

Taken together, neither HLA ligand predictors nor models specifically designed to predict T-cell recognition of MHC-presented peptides could reliably identify epitopes from an emerging pathogen. This analysis pinpoints that HLA binding classifiers may not be sufficient to predict immunogenicity, especially in the case of an emerging pathogen in human setting. This indicates that while presentation is necessary, it is not sufficient for T-cell immunogenicity. Furthermore, this analysis illustrates that the current models do not appear to outperform MHC presentation predictors, and suggest a need to develop an accurate and robust classifier able to shortlist CD8+ T-cell epitopes beyond HLA presentation.

Evaluate performance of existing models in predicting cancer neoepitopes

One key application of immunogenicity classifier is identifying cancer neoepitopes for personalized cancer immunotherapy or vaccine developments. Shortlisting cancer neoepitopes is a 'needle in a haystack' problem in which one aims to find an extremely small number of epitopes from substantially imbalanced dataset. In fact, previous studies reported that neoantigens comprised

only ~6% of candidate cancer peptide list[52,191], implying that the current prediction pipeline has a high false positive rates. Nevertheless, the ability of models to accurately identify a small number of epitopes is critical.

To benchmark the algorithms, two independent cancer neoantigen datasets were employed, both of which were intrinsically imbalanced. They are i) our in-house glioblastoma (GBM) dataset and ii) publicly available dataset from Tumor Neoantigen Selection Alliance (TESLA) consortium[52]. The GBM dataset comprises HLA-A*02:01 presented peptides from glioblastoma cancer patients. Any peptides found in any model's training data were excluded, yielding 123 peptides with 9-10 amino acids in lengths of which 25 (20%) were Positives. Similarly, any peptides observed in training data from the 'TESLA' dataset were excluded, and only retained peptides that were presented on HLA alleles for which all models are applicable, leaving peptides from 7 of the 13 original HLA-I alleles. This resulted 27 (~6.7%) immunogenic and 372 (~93%) non-immunogenic peptides of lengths 9 and 10.

The performance of the models was evaluated on GBM and TESLA datasets using PR-AUC metric. As cancer neoepitope dataset is highly imbalanced, ROC-AUC can be misleading as this metric may underrepresent the minority class[192]. In addition to the 8 previously discussed models, 'DeepHLApan', which was specifically designed for identifying cancer neoepitopes, was included[193].

For GBM dataset, the PR-AUC ranged from 0.20 for DeepImmuno to 0.437 for iPred (Figure 5.4A). With the exception of DeepHLApan which had high number of false negatives, most of algorithms had low accuracy due to a high number of false positives (Figure 5.4B). Notably, despite having one of the highest ROC-AUC from 10-fold cross-validations in the original publication, DeepImmuno had the highest number of false positives. Interestingly, netMHCpan (BA and EL) and PRIME identified the highest number of cancer neoepitopes, 19, 18, and 18 respectively, for a total of 25 confirmed neoantigens (Figure 5.4B). While the models showed suboptimal PR-AUC scores, the majority of models outperformed random in bootstrap analysis, with netMHCpan_BA (z=5.2) having the most predictive power against the GBM dataset, followed by iPred (z=5.03) and NetTepi (z=4.38) (Figure 5.4C).

Similarly, suboptimal PR-AUC scores were observed on the TESLA dataset, ranging from 0.051 for Repitope to 0.211 for *Gao* (Figure 5.5A). By bootstrapping analysis, Gao (z=6.65), netMHCpan_EL (z=6.2), NetTepi (z=4.88), PRIME (z=4.55) and netMHCpan_BA (z=4.53) were found to perform better than random (Figure 5.5C). Gao computes immunogenicity of a peptide using dissimilarity to self and similarity to viral peptides. These are features identified by Wells *et al.* in the context of the TESLA dataset, which may explain Gao's superior performance. PRIME identified the most TESLA neoantigens (26/27) followed by netMHCpan_EL which found 22/27. However, all models, including DeepHLApan, produced high numbers of false positives (Figure 5.5B).

In summary, on the TESLA dataset, the only model that outperformed NetMHCpan was Gao predictor, although this was not true on GBM dataset. The performance of these models in predicting cancer neoantigens showed a high degree of cross-data inconsistency. Model

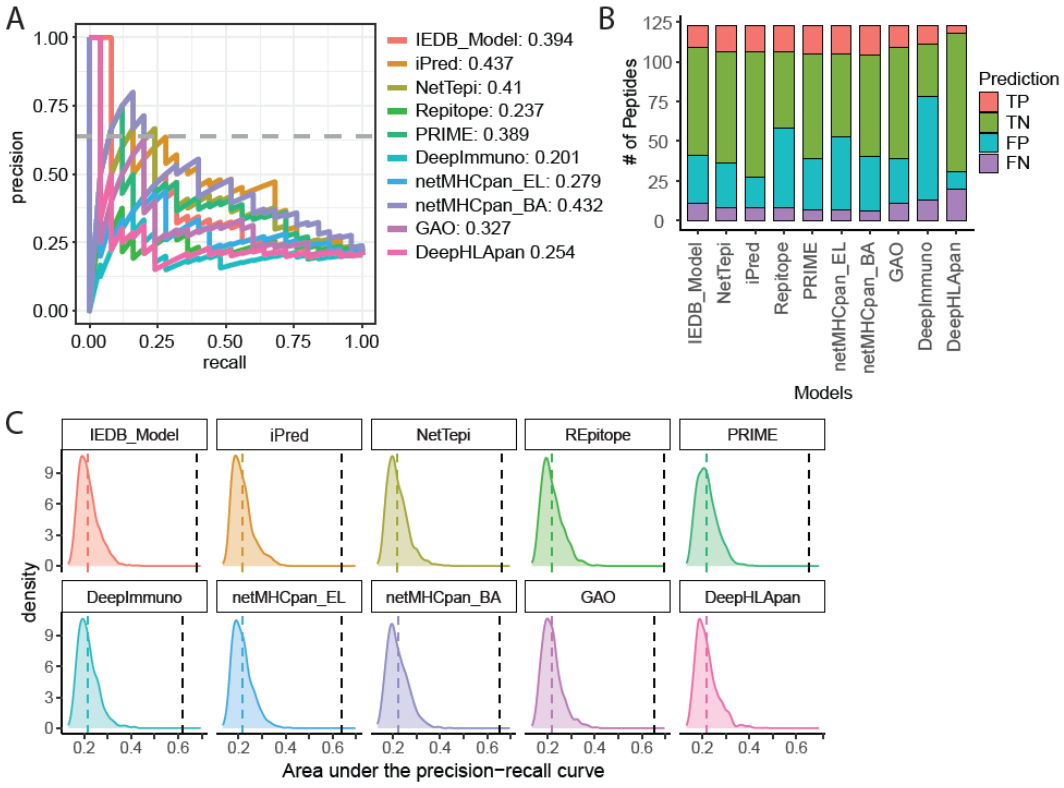


Figure 5.4: Models are not reliable in predicting neoepitopes from glioblastoma patients. A. PR curves showing model performance against the GBM neoantigen dataset. B. Number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) predicted by models against neoantigens from glioblastoma patients. C. PR-AUCs of each model against the GBM dataset (black dashed line) contrasted with a random 'bootstrap' distribution.

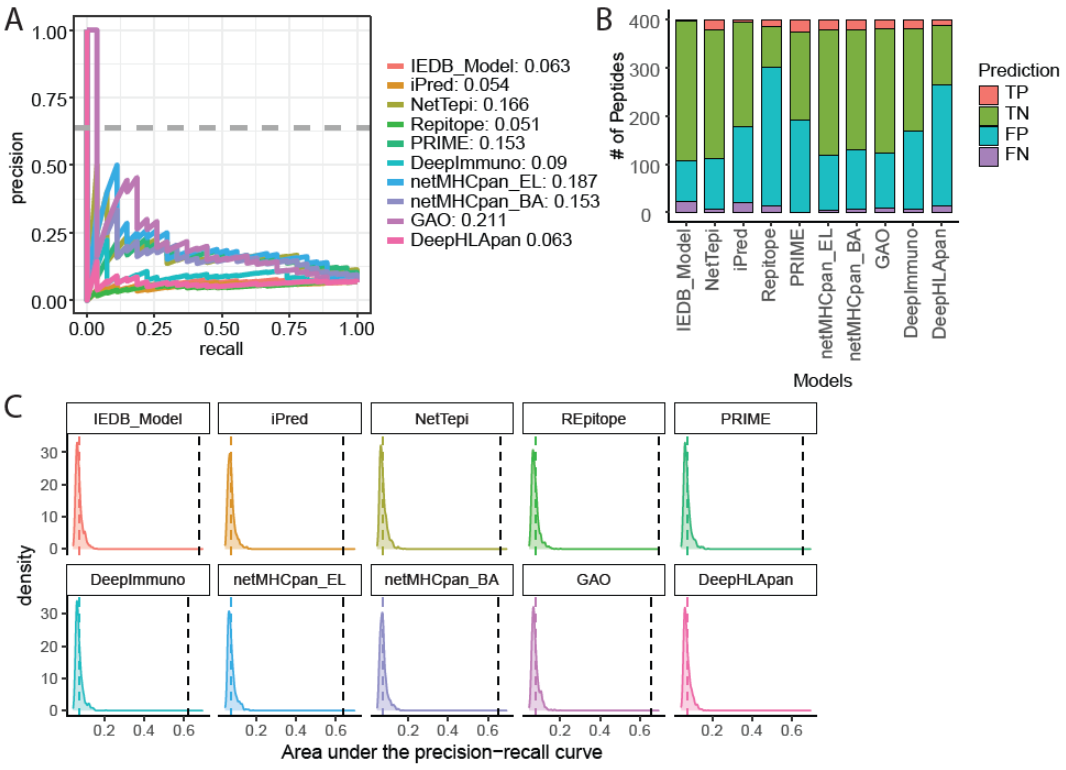


Figure 5.5: Models do not show high performance on TESLA neoantigen dataset. A. PR curves showing model performance against the TESLA neoantigen dataset. B. Number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) predicted by models against TESLA neoantigen dataset. C. PR-AUCs of each model against the TESLA dataset (black dashed line) contrasted with a random 'bootstrap' distribution

performance was generally hampered by a high false positive rate that contributed to low precision. It is worth noting that although high false positive and low false negative rates may be preferred in some medical settings, immunogenicity predictions would favour achieving optimal performance by balancing false positive and false negative rates. Taken together, these findings suggest that improving precision to robustly shortlist T-cell targets may be an avenue for improving neoepitope prediction beyond HLA ligand predictors.

The effects of cross-HLA variation in predicting peptide immunogenicity

By exploring variability in the number of epitopes and non-epitopes across different HLA alleles, there was a substantial difference in the datasets used to train different models (Figure 5.6A). Indeed, in the training data for iPred, we observed that epitopes dominate HLA-A supertypes while non-epitopes dominate HLA-B (Figure 5.6B). As a result, HLA-A presented peptides are more prone to receive higher immunogenicity score from this model regardless of their true immunogenicity status.

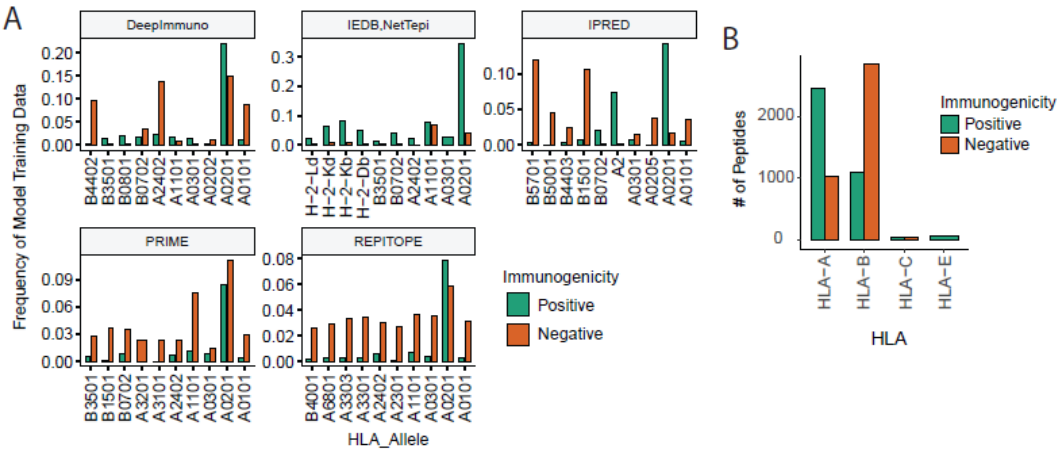


Figure 5.6: HLA-variation in the number of epitopes and non-epitopes for model training. A. Barplots showing the distribution of immunogenic and non-immunogenic peptides for the top HLA alleles in each model's training dataset. Gao's training data is not included as the immunogenicity status are continuous. B. Barplot showing the HLA supertype distribution for immunogenic and non-immunogenic peptides amongst the Chowell dataset used to train iPred.

We therefore sought to evaluate whether imbalances in the number of epitopes and non-epitopes across different HLAs have contributed to suboptimal performance of these models. We hypothesized that the imbalanced dataset may lead to skewed predictions associated to HLA, in case that TCR recognition potential is not explicitly deconvoluted from antigen presentation features. This may lead to predictions reflecting antigen presentation potential due to more prominent HLA binding features[58,59] instead of both HLA presentation and TCR recognition propensities. We therefore aimed to explore if the immunogenicity models were biased in their prediction by HLA alleles.

We first examined the composition of epitopes and non-epitopes on the most prominent ten HLA

alleles in each model's training datasets. In all training datasets, HLA-A*02:01 had the most peptides, and the number of epitopes and non-epitopes was highly imbalanced on all HLA alleles, including HLA-A*02:01 (Figure 5.7A). Even when peptides in each models' training data were separated based on whether they bind HLA-A*02:01 (hereby A*02:01+) or not (A*02:01-), a significant difference in the prevalence of epitopes and non-epitopes was found (Figure 5.7B). With the exception of IEDB and NetTepi, epitopes dominated A*02:01- group, implying that non-epitopes in the training data were primarily presented by HLAs other than HLA-A*02:01.

We therefore hypothesized that these models may capture HLA binding features, resulting in biased predictions by HLA alleles. To test this hypothesis, the predicted immunogenicity scores generated by each model from SARS-CoV-2 peptides were gathered and classified according to whether the peptide is immunogenic or not and whether it binds HLA-A*02:01 or not. There was no significant difference in immunogenicity scores between epitopes and non-epitopes that bound on HLA-A*02:01 (i.e. A0201+ Pve and Nve) in some of the models (Figure 3C). However, A02:01+ Nve in Repitope (Cohen's d = 0.454), DeepImmuno (Cohen's d = 0.347), NetTepi (Cohen's d = 0.338) and iPred (Cohen's d = 0.206) resulted in a higher immunogenicity score than A02:01- Pve. This suggested that the model favours A02:01 binding over immunogenicity potential and skews prediction scores regardless of true immunogenicity status.

We next investigated how well these models detect HLA presentation features as opposed to T-cell recognition features. The predicted immunogenicity scores from previous SARS-CoV-2 benchmarking study were again used to test if the models could predict whether or not peptides are bound on HLA*02:01. Remarkably, all models performed better (Figure 5.8) than when the scores were used to predict immunogenicity potential (Figure 5.2A), implying that immunogenicity predictors may include more dominant HLA features than T-cell recognition features.

Our analysis suggests that the models may have learned prominent HLA binding features over the T-cell recognition features, given the imbalanced number of epitopes and non-epitopes across different HLA alleles in the training datasets. This suggests that the suboptimal performance of these models may partly be due to data bias, and indicate the need to balance the number of epitopes and non-epitopes at least for the prevalent HLA alleles and/or develop strategies to regress out the effect of HLA imbalances. Therefore, in order to accurately predict T-cell recognition potential of HLA-presented peptides, peptide compositions and modelling strategies should be carefully considered to minimize the potential HLA bias.

Considerations in modelling immunogenicity

Given the limitations in the current methods to predict TCR:pMHC recognition, this section describes a few factors to consider when building immunogenicity classifiers. First, a key challenge is the lack of true negative datasets for TCR-epitope interaction as well as cross-reactivity information. Several groups tackled this limitation by simulating a background or negative data [50,54,186,194]. For example, Jurtz *et al.* created incorrect combinations of TCRs and peptides i.e. linking TCR sequences with a random peptide different from the cognate target, and produced a balanced set of positive and negative data.

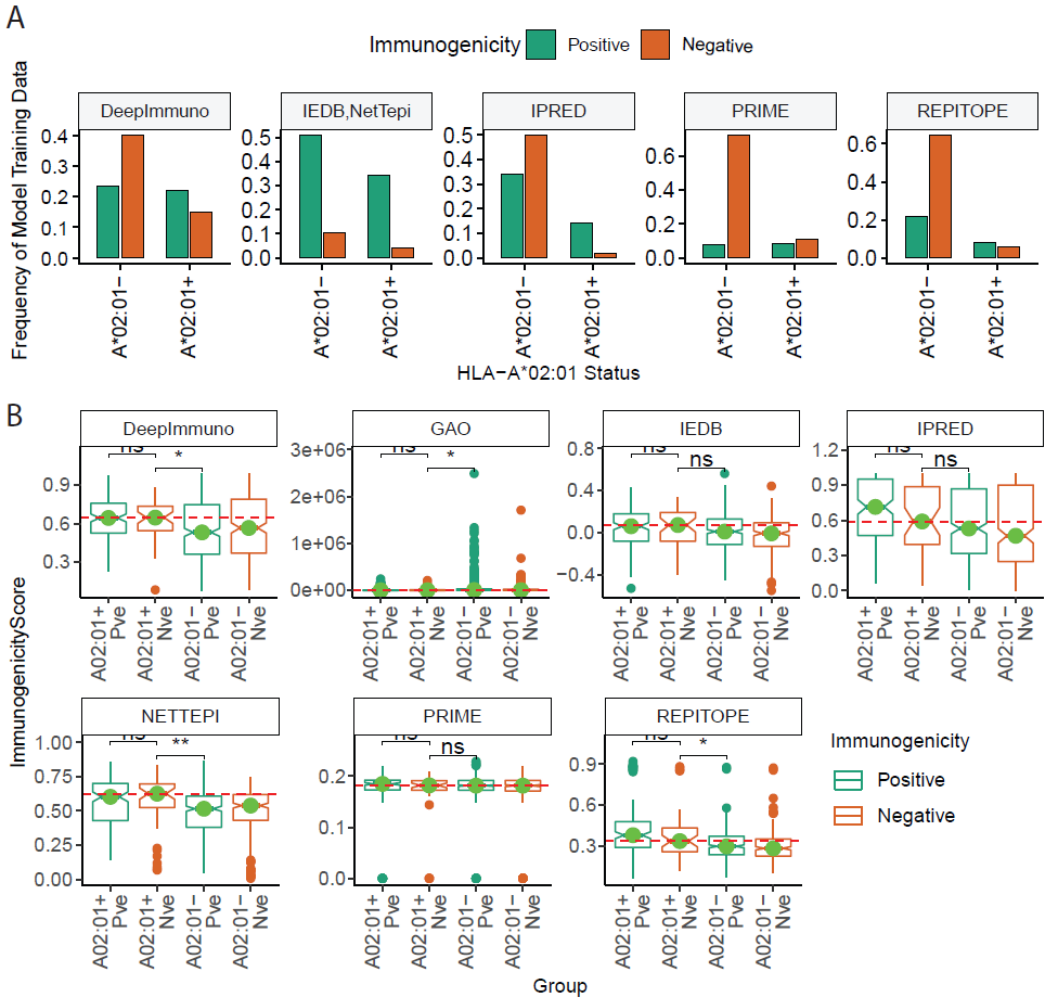


Figure 5.7: Effects of cross-HLA variation on predicting peptide immunogenicity. A. Barplots showing the distribution of immunogenic and non-immunogenic peptides for whether they bind HLA-A02:01 (labelled A*02:01+) or another allele (A*02:01-). B. Notched boxplots showing subtle differences between the immunogenicity scores created by each model, based on whether the peptide binds. HLA-A*02:01 (A02:01+) or another allele (A02:01-), and whether the peptide is immunogenic (Pve, labelled blue) or non-immunogenic (Nve, labelled red). For 'Gao', the y-axis is log10 scaled.

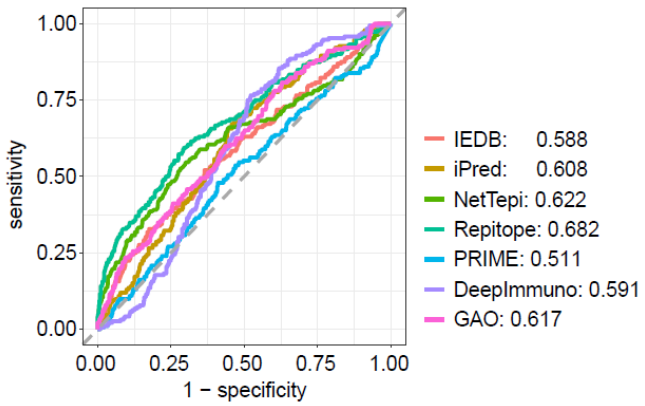


Figure 5.8: Predictability of immunogenicity models on HLA-A*02:01 binding. ROC-AUC of models when prediction was made for SARS-CoV-2 peptides and used to predict HLA-A*02:01 binding status. Significance was assessed using Wilcoxon tests.

Additionally, the existing datasets describe immunogenicity in a binary format, whereas it is evident that the T-cell response comes in a continuum from a mild to a strong response and varies in functional outcomes such as differential cytokine production. Quantitative T-cell response measures associated with each epitope will pave the way for more rigorous modelling.

Third, there is a considerable heterogeneity in the experimental methodologies employed to assess T-cell responses. Although standardizing T-cell assays into a single readout is practically difficult, accuracy of predictive algorithms may be enhanced by reflecting the sensitivity and specificity of each assays used for characterizing epitope.

Fourth, exhaustive screenings have been performed based on an assumption of invariant MHC interaction. Previous research, however, suggested that a TCR could recognise peptides bound by non-canonical HLA molecules[195,196]. van der Zwan *et al.* reported cross-reactivity of virus-specific T-cells to HLA-A and -B molecules, as well as HLA-B*08:01-restricted EBV-specific T-cell against HLA-C*01:02[197]. From a clinical standpoint, a severe off-target toxicity was reported by adoptive cell transfer of T-cells targeting HLA-A*01:01 MAGE-A3 complexes by binding to a Titin-derived peptide displayed on HLA-A*01:01 [198]. Thus, it may be necessary to screen for peptides bound on non-predicted HLA alleles to project the full scope of cross-reactivity, and we must keep in mind that the current sets of data may only represent the tip of an iceberg.

Lastly, while TCR:pMHC interactions exhibit a remarkable capacity for discrimination, they are often sloppy and cross-reactive. Nevertheless, as exemplified by thymic selection, weaker affinities play an essential role in supporting the sensitive detection of a wide range of cognate antigens while keeping it well-balanced from self-reactivity[199,200]. Moreover, low and high-affinity T-cells may involve in biological processes differently in specific conditions, such as effector and memory differentiation, metabolic reprogramming, and immune response[201,202]. Given the dynamic nature of weak interactions and their potential functional implications, we may need to deviate from lessons learned from well optimized interfaces. We should note that, due to the

difficulties in measuring low affinity interactions, existing data may be biased in favour of high affinity interactions[203], and that scoring functions and parameters may need to be re-examined to reflect the dynamic interplay of low and high-affinity T-cells for an efficient immune response[204].

Discussion

In this chapter, existing immunogenicity models were benchmarked for identifying CD8+ T-cell epitopes from emerging pathogens or cancer. It began by reviewing the current list of immunogenicity algorithms that have been developed to date. Then, an unbiased evaluation of MHC binding and immunogenicity models was performed on SARS-CoV-2, in-house glioblastoma (GBM) and TESLA datasets, and current methods showed suboptimal performance in both contexts. Lastly, the extent to which HLA bias contributes to immunogenicity predictions was assessed, and it was found that HLA binding features appear to dominate immunogenicity prediction in current modelling strategies.

Most models did not outperform MHC binding predictions when predicting CD8+ T-cell targets from emerging viruses or cancer, with all models performing poorly. Only for predicting cancer neoepitopes the tested models marginally outperform NetMHCpan. PRIME and NetTepi achieved high recall (i.e. high true positives), albeit at the expense of many false positives (i.e. low precision). Indeed, many models had a high false positive rate.

While we sought to provide an unbiased evaluation of existing algorithms, there exist limitations. First, the analysis had to be restricted to certain HLAs and peptide lengths for which all models were applicable. Comprehensive list of peptides from other HLAs and peptide lengths would be required to generalize our observations. Second, despite the potential benefits of re-training models, pre-trained models from original publications were used. This aimed to report expected accuracies when end users employ the models as published and provide fair comparisons across models without re-calibrating parameters on new training datasets.

Third, SARS-CoV-2 peptides were used as the test dataset because they have the unique advantage of having never been seen by any of the models. However, depending on the training dataset employed for each models, there may be greater divergence between training and test datasets (i.e. out-of-distribution generalization problem). This issue will be addressed in greater detail in the following chapters.

Forth, HLA bias was assessed with an emphasis on HLA-A*02:01 binding due to limited number of peptides bound on other HLA alleles and characterized by T-cell assays. More information from other HLAs will permit HLA-specific analysis.

In the last section, we discussed key challenges in the current datasets and factors to consider when modelling peptide immunogenicity, such as lack of true negatives, heterogeneity in methodologies and readouts, assumption of invariant HLA interactions and limited sensitivity to capture low-affinity TCR:pMHC interactions. We believe taking these factors into account will facilitate making more accurate and robust predictions of highly complex and multifactorial phenomenon.

Methods

SARS-CoV-2 data for benchmarking

The SARS-CoV-2 datasets were retrieved from IEDB and VIPR[190] (both accessed 07-10-2021). Peptides were restricted to those having 9-10 amino acids in length due to a low number of peptides with other lengths, such as 8-aa peptide having 32 epitopes and 8 non-epitopes, and 11-aa peptides having 69 epitopes or 28 non-epitopes. Peptides bound on HLA-A, HLA-B, or HLA-C alleles were included. Those having single Positive readout but ≥2 Negative readouts were excluded, to only accommodate high-confidence Positives. Otherwise, in case of contradictory readouts, the designated peptide-HLA was considered Positive. Because NetTepi was only applicable to peptides bound on one of 13 HLA alleles, peptides bound on the following HLA alleles were retained: HLA-A*02:01, HLA-B*58:01, HLA-B*15:01, HLA-B*35:01, HLA-B*07:02, HLA-A*01:01, HLA-A*03:01, HLA-A*11:01, HLA-A*24:02, HLA-A*26:01, HLA-B*27:05, HLA-B*39:01, HLA-B*40:01.

GBM data for benchmarking

The cancer neoantigens from glioblastoma patients were first predicted by calling somatic variants from whole exome sequencing data using the in-house tool 'TUNAPASTA v0.5', which is based on MuPeXI workflow[205]. The somatic variants underwent MHC binding prediction by NetMHCpan and retained those having priority score > 50. Selected peptides were synthesized by Pepscan (Netherlands) and used for screening antigen-specific T-cells from HLA-A02:01+ GBM patients' and healthy donors' peripheral blood mononuclear cells (PBMC). PBMC was primed with cancer neoepitope candidates *in vitro* as previously described[206,207]. The pMHC(HLA-A02:01)-tetramers with peptides of interest[208] were assembled, then incubated with $1x10^5$ primed T-cells. Peptide-specific T-cells were sorted by flow cytometry. Isolated T-cells were then expanded and functionally validated by co-culturing them with corresponding peptides at various concentrations, followed by immunostaining of TNF-a (clone MAb11, Biolegend) and CD107a (clone H4A3, Biolegend).

TESLA data for benchmarking

The TESLA dataset was retrieved from Wells et al[52]. In this study, authors predicted cancer neoantigens from six patients, yielding 608 predicted neoepitopes, of which 37 were characterized immunogenic. For benchmarking, peptides having 9-10 amino acids in length and were presented on one of 13 HLA alleles were included where all models could be applied.

Immunogenicity models for benchmarking

All publicly available models were gathered, and excluded those that were unable to perform comparative assessment (Example of reasons described in Table 5.1 and 5.2). This left 7 models: IEDB model, NetTepi, iPred, Repitope, PRIME, DeepImmuno, Gao and DeepHLApan. NetMHCpan 4.0 eluted ligand and binding affinity predictions were added to compare with MHC binding predictions. All models, with the exception of DeepHLApan, were tested on SARS-CoV-2 and cancer neoepitope data, whereas DeepHLApan was only tested on cancer neoepitope data.

All models were downloaded and ran locally, except DeepImmuno that could not be run locally and instead used webserver at: https://deepimmuno.research.cchmc.org/

Training datasets of each models

Training data for IEDB model was retrieved from Calis *et al.*[49]. The training data for iPred ('Chowell' dataset) was retrieved from https://github.com/antigenomics/ipred/tree/master/classifier. The MHC-I human training data for Repitope model retrieved from https://github.com/masato-ogishi/Repitope. Training data for PRIME was downloaded from Schmidt *et al.*[189]. Training data for DeepImmuno was downloaded from https://github.com/frankligy/DeepImmuno#deepimmuno-cnn

Performance evaluation

All models were benchmarked using default settings. For each model, ROC curves were generated using the pROC package and PR curves using yardstick. To generate confusion matrices, optimal threshold values were determined from ROC curves using youden index, conducted binary classifications and generated matrices using confusionMatrix function in the R caret package. The youden index uses ROC curves to determine a threshold that maximises [1-sensitivity+specificity].

Bootstrapping analysis

Different models predicted immunogenicity scores on test datasets. On this dataset, the immunogenicity label of each peptide was randomly shuffled. The PR-AUCs were computed from the shuffled dataset and repeated the process 1000 times, yielding a distribution of 'shuffled PR-AUCs'. This reflects the performance of random, unskilled models. The true 'benchmarked PR-AUCs' were then compared to the 'shuffled PR-AUCs' to determine the extent to which the model outperforms unskilled models. Z-scores were calculated by subtracting the mean of the 'shuffled PR-AUC' distribution from the 'benchmarked PR-AUC' distribution and dividing by the standard deviation of the 'shuffled PR-AUC' distribut

Chapter 6. Sources of low accuracies in identifying epitopes from emerging pathogen

Original publications:

- **Lee CH**, Antanaviciute A., Buckley PR, Simmons A. & Koohy, H. To what extent does MHC binding translate to immunogenicity in humans? Immunoinformatic (2021)
- Buckley, PR*, **Lee, CH***, Ma, R*. Woodhouse I, Woo J, Tsvetkov VO, Shscherbinin BS, Antanaviciute A, Shugay M, Rei M, Simmons A, Koohy H.. Evaluating performance of existing computational models in predicting CD8+ T-cell pathogenic epitopes and cancer neoantigens. Briefings in Bioinformatics. *These authors contributed equally.

Introduction

Previous efforts to decipher properties underlying T-cell recognition of MHC-presented peptides have contributed to the development of immunogenicity classifiers.

Early research linked immunogenicity to large and aromatic side chains[49] and hydrophobicity[54,177,182,209]. Specifically for cancer neoepitopes, Wells *et al.* reported agretopicity (i.e. the ratio of binding affinity between neoepitope and wild-type counterpart) and foreignness score (i.e. similarity of neoepitope to previously characterized epitopes) as correlates of T-cell recognition[52]. Furthermore, Richman *et al.*[210] and Devlin *et al.*[211] demonstrated that structural dissimilarity to the non-mutated human proteome is another recognition feature that allows neoepitopes to escape immune tolerance. While many algorithms utilize features learnt from cancer neoepitope studies, the extent to which self- and foreign peptides share common immunogenicity features has been called into question.

Moreover, we previously reported that existing immunogenicity algorithms perform poorly in predicting epitopes from cancer and emerging viral pathogen[56]. The limitations of the current methodologies urged to investigate the sources of poor performance, especially on emerging pathogen.

Here, with many existing algorithms performing poorly in predicting pathogenic and cancer epitopes, the sources of poor accuracy were investigated to develop accurate, robust and biologically meaningful classifiers. First, the extent to which MHC binding correlates to immunogenicity in humans was examined to emphasise the importance of deciphering T-cell recognition features in addition to MHC binding properties. The discriminative power of known immunogenicity hallmarks was then compared across cancer neoepitopes and peptides from various pathogenic species. Finally, the reason why models incorporating peptide and HLA alleles performed poorly on cancer and emerging pathogens was investigated, and potential HLA bias from an imbalanced training dataset was found. This chapter thereby uncovered potential causes of poor performance to lay the foundation for building robust and accurate classifiers for self- and pathogenic peptides.

Results

Extent of MHC presentation translated to immunogenicity in humans

High-throughput peptide-HLA binding assays have enabled comprehensive characterization of MHC-presented peptides, which in turn led to the development of highly accurate models to predict peptide-MHC binding, especially for class I molecules[31]. Moreover, large-scale screening of pMHC-specific T-cells and candidate T-cell epitopes have started to shed light on pMHC-TCR interactions[35,212] in the context of T-cell cross-reactivity and antigen specificity[62,213,214]. With expanding data from both MHC binding and T-cell recognition studies, the impact of these processes on peptide immunogenicity, or a peptides' ability to trigger an immune response, has also been investigated.

To determine what fraction of MHC-presented viral peptides are recognized by T-cells in mouse, Croft *et al.* screened 172 vaccinia virus (VACV) peptides presented on H-2D^{b} and H-2K^{b} by high-resolution mass spectrometry and then tested for IFNγ response on 8 virus-infected mice[215]. They observed that 83% and 39% of peptides had T-cell response in at least one and five mice respectively, and concluded that in a mouse model, majority of MHC-bound vaccinia virus peptides are immunogenic.

Notably, 99 immunogenic and 25 non-immunogenic vaccinia virus peptides from this study were deposited in Immune Epitope Database and Analysis Resource (IEDB), 64 of which also had entries from other studies[216–219] that predominantly used *ex vivo* splenocytes or CD8^{+} T-cells co-cultured with VACV-incubated mouse cell lines. Interestingly, of these 64 peptides, 47 had contradictory readouts across different studies, mostly attributed to the use of splenocytes as effector population opposed to CD8^{+} T-cells as in Croft *et al.*, suggesting potential disparities in immunogenicity due to different experimental setups.

In a follow-up study, Paul *et al.*[31] leveraged the data from Croft *et al.* to evaluate the performance of 17 models in predicting T-cell epitopes. They reported that all evaluated models, which were originally developed for predicting peptide presentation and processing, can predict T-cell epitopes better than random, with 12 achieving AUCs greater than 0.9. Studies such as Paul *et al.* and Croft *et al.* raise the question of the extent to which these models can be used to predict peptide immunogenicity, or more fundamentally, the extent to which presentation features can be translated into T-cell response in humans.

Considering the difficulties of testing peptide immunogenicity in humans *in vivo*, the contribution of MHC binding on immunogenicity was inferred by analysing the proportion of MHC-bound peptides characterized immunogenic by *in vivo* or *in vitro* T-cell assay. HLA-A*02:01-restricted VACV peptides that were tested by T-cell assays were retrieved from the IEDB. To ensure MHC binding, peptides were subjected for NetMHCpan 4.1 prediction and T-cell response were only analysed on predicted MHC binders. While pathogenic and self-peptides displayed similar levels of MHC presentation in human and mouse (Figure 6.1A, 6.1C), 48% and 33% of HLA-A*02:01- and H2-Db-bound peptides were non-immunogenic, respectively (Figure 6.1B, 6.1D). Also, the proportion of immunogenic VACV peptides ranged from 2.4% to 69% across studies[220–223]. This

suggests that the fraction of presented peptides that are immunogenic may be more variable, and both context and organism-dependent.

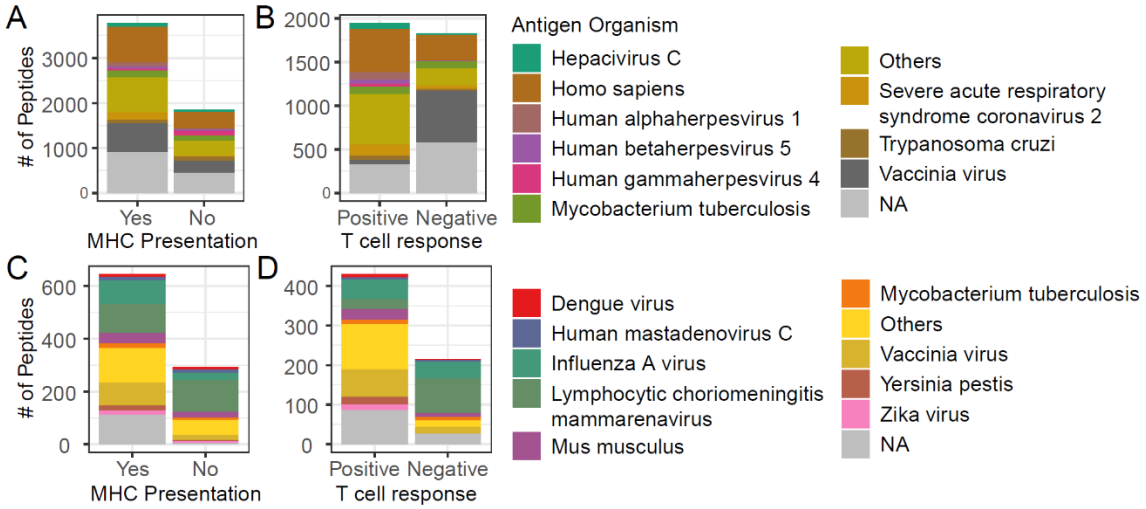


Figure 6.1: Extent of MHC presentation and T-cell recognition in human and mouse systems. Retrieved pathogenic and self-peptides characterized by T-cell assays from IEDB. A, C. NetMHCpan 4.1 prediction of HLA-A*02:01(A)- and H2-D^{b} (C)-restricted peptides for presentation of respective MHC molecules. B, D. Response of T-cells, tumor infiltrating lymphocytes, Jurkat T-cells, T-cell hybridoma or PBMC for HLA-A*02:01- (B) or H2-D^{b}-(D) presented peptides. Peptides characterized using splenocytes as effector population were removed from analysis due to high variability.

Difference in T-cell recognition features between neoepitopes and pathogenic peptides

Benchmarking studies revealed that many algorithms employing features learned from cancer studies did not perform well on pathogenic peptides. We therefore sought to investigate whether the observed variabilities in performances between pathogenic and cancer settings stem from differences in immunogenicity features. Wells *et al.* previously reported a comprehensive list of features associated with T-cell recognition for cancer neoepitopes, such as binding affinity, binding stability and fraction of hydrophobic amino[52]. Therefore, the capacity of these features to distinguish pathogenic epitopes from non-epitopes was examined.

The HLA-I bound pathogenic peptides were retrieved from the IEDB (accessed 17-02-22, 'pathogenic' dataset) and cancer neoepitopes from Wells *et al.*[52] (TESLA 'cancer' dataset). The entire TESLA dataset of 608 peptides was used, with ~6% being immunogenic and ~94% not. The 'pathogenic' dataset contained 23,958 peptides where ~26% were immunogenic and ~74% were not. The differences in MHC binding affinity, binding stability and fraction of hydrophobic amino acids were then compared on pathogenic peptides. The agretopicity, foreignness score and tumor abundance were not included in the analysis because they are more cancer-specific.

First, while both cancer and pathogenic peptides have stronger MHC binding affinities for epitopes compared to non-epitopes, cancer peptides in general exhibit stronger affinities than their pathogenic counterparts (Figure 6.2A-C). Pathogenic epitopes had stronger affinities than non-epitopes, consistent with cancer peptides (Figure 6.2B). However, the median binding affinity of cancer non-epitopes was comparable to that of pathogenic epitopes especially above 100nM (Figure 6.2B-C), implying that pathogenic and cancer peptides may be subjected to different thresholds.

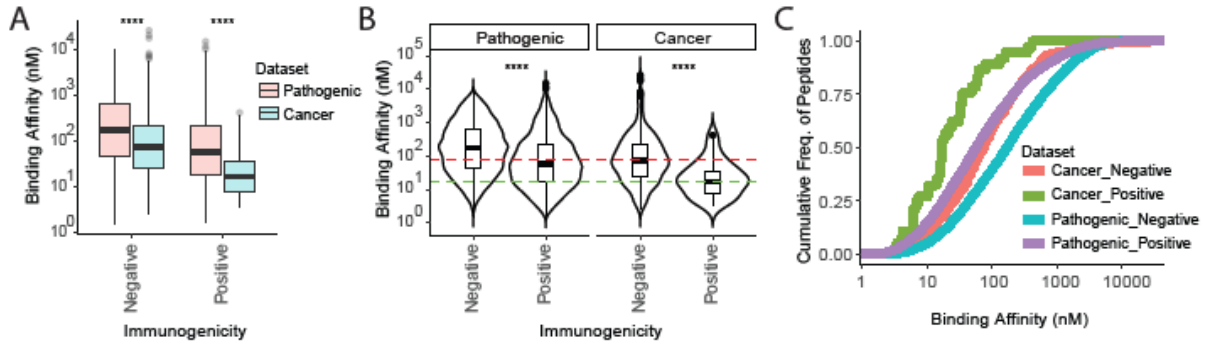


Figure 6.2: Difference in magnitude of MHC binding affinity in discriminating epitopes and non-epitopes for pathogenic and cancer peptides. A-B. Boxplots (A) and violin (B) plots comparing epitopes and non-epitopes between pathogenic and cancer peptides. Green and red dashed lines show the median of binding affinities for cancer (TESLA) epitopes and non-epitopes, respectively. C. Line plot showing the empirical cumulative distribution of binding affinities.

Second, a similar trend was observed in MHC binding stability. Consistent with MHC binding affinity, cancer peptides generally showed greater stability compared to pathogenic counterparts (Figure 6.3A-C). Pathogenic epitopes had higher stability than non-epitopes, and cancer non-epitopes had comparable stability to those of pathogenic epitopes (Figure 6.3B-C). To ensure the trend is consistent across different HLA alleles, pathogenic peptides bound on five common HLAs were examined. Pathogenic epitopes had stronger binding affinities (Figure 6.3D) and more stable binding (Figure 6.3E) than non-epitopes across HLAs, with the exception of HLA-C*07:02, which may be attributed to the low number of peptides.

Third, pathogenic epitopes were more hydrophobic than their cancer counterparts (Figure 6.4A-C). Notably, pathogenic and cancer neoepitopes showed opposite directionalities in hydrophobicity (Figure 4B). Moreover, the pattern was not consistent across HLA alleles, with only pathogenic epitopes presented by HLA-A*01:01 showing greater hydrophobicity (Figure 6.4D). With the high variability across cancer and pathogenic peptides, as well as across different HLA alleles, further work is required to determine whether the trend is a biological phenomenon or a technical artefact.

The trends in T-cell recognition features were compared across in-house glioblastoma (GBM) and TESLA datasets to further investigate variability across different cancer datasets. The binding affinity (Figure 6.5A), binding stability (Figure 6.5B) as well as fraction of hydrophobicity

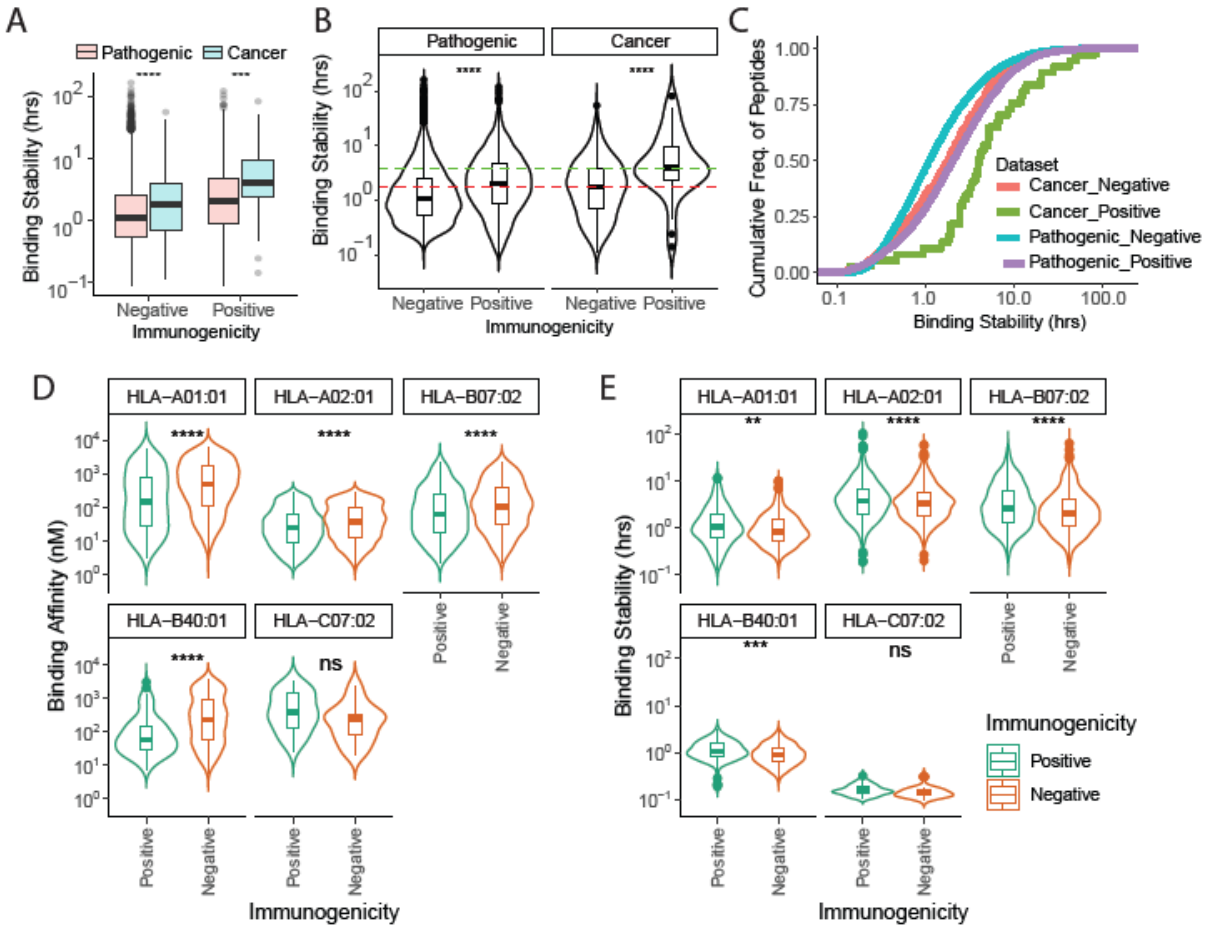


Figure 6.3: Difference in magnitude of MHC binding stability in discriminating epitopes and non-epitopes for pathogenic and cancer peptides. A-B. Boxplots (A) and violin (B) plots comparing epitopes and non-epitopes between pathogenic and cancer peptides. Green and red dashed lines show the median of binding stabilities for cancer (TESLA) epitopes and non-epitopes, respectively. C. Line plot showing the empirical cumulative distribution of binding affinities. D-E Violin plots showing binding affinity (D) and stability (E) for five most prevalent HLAs.

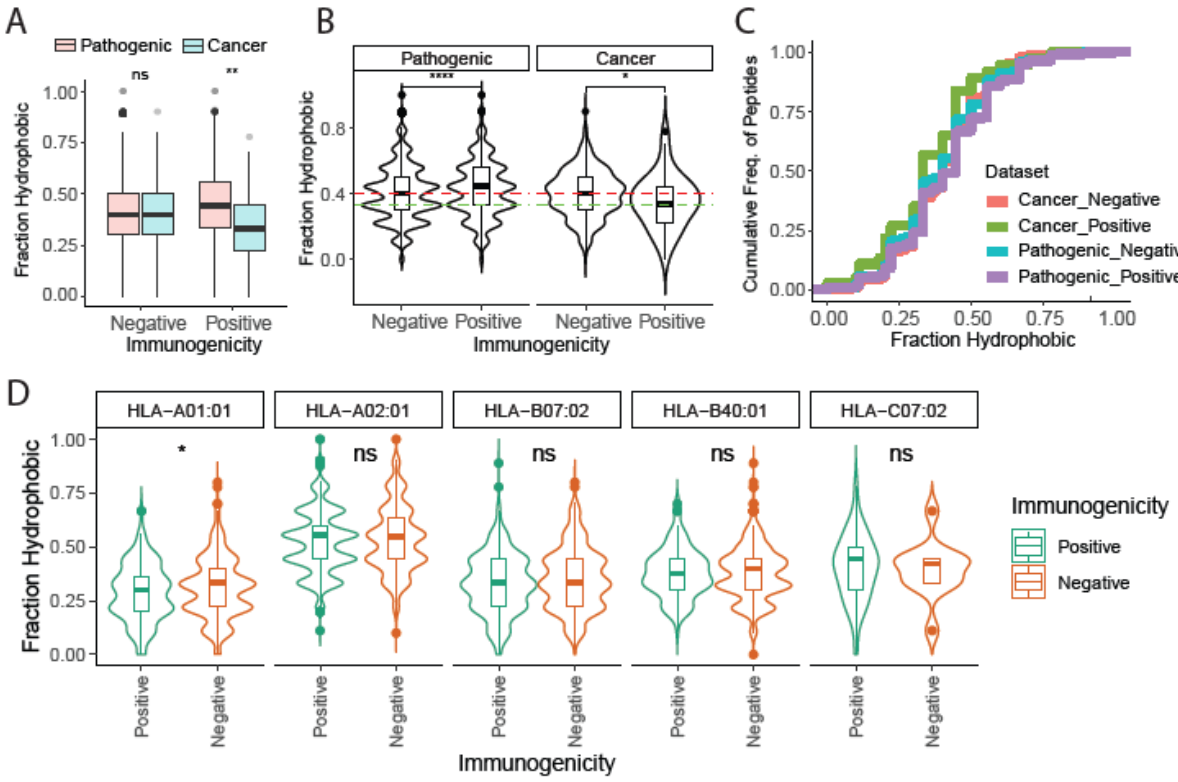


Figure 6.4: Difference in hydrophobicity in discriminating epitopes and non-epitopes for pathogenic and cancer peptides. A-B. Boxplots (A) and violin (B) plots comparing epitopes and non-epitopes between pathogenic and cancer peptides. Green and red dashed lines show the median of binding stabilities for cancer (TESLA) epitopes and non-epitopes, respectively. The hydrophobicity computed by fraction of hydrophobic amino acids in the peptide sequence. C. Line plot showing the empirical cumulative distribution of binding affinities. D. Violin plots for five common HLAs.

(Figure 6.5C) could effectively discriminate epitopes on both datasets. However, as suggested by previous pathogenic studies, the analysis indicated that the thresholds required to differentiate immunogenicity may be tailored specific to the context of test peptides.

In summary, these findings hinted that cancer and pathogenic epitopes may have different T-cell recognition features in magnitudes or directionality, and that context-specific training datasets may be required to tailor thresholds for accurate predictions. Of note, the analysis should be interpreted with caution because it is yet unclear whether the differences observed from the existing data repository are biological or the result of data-related artefacts.

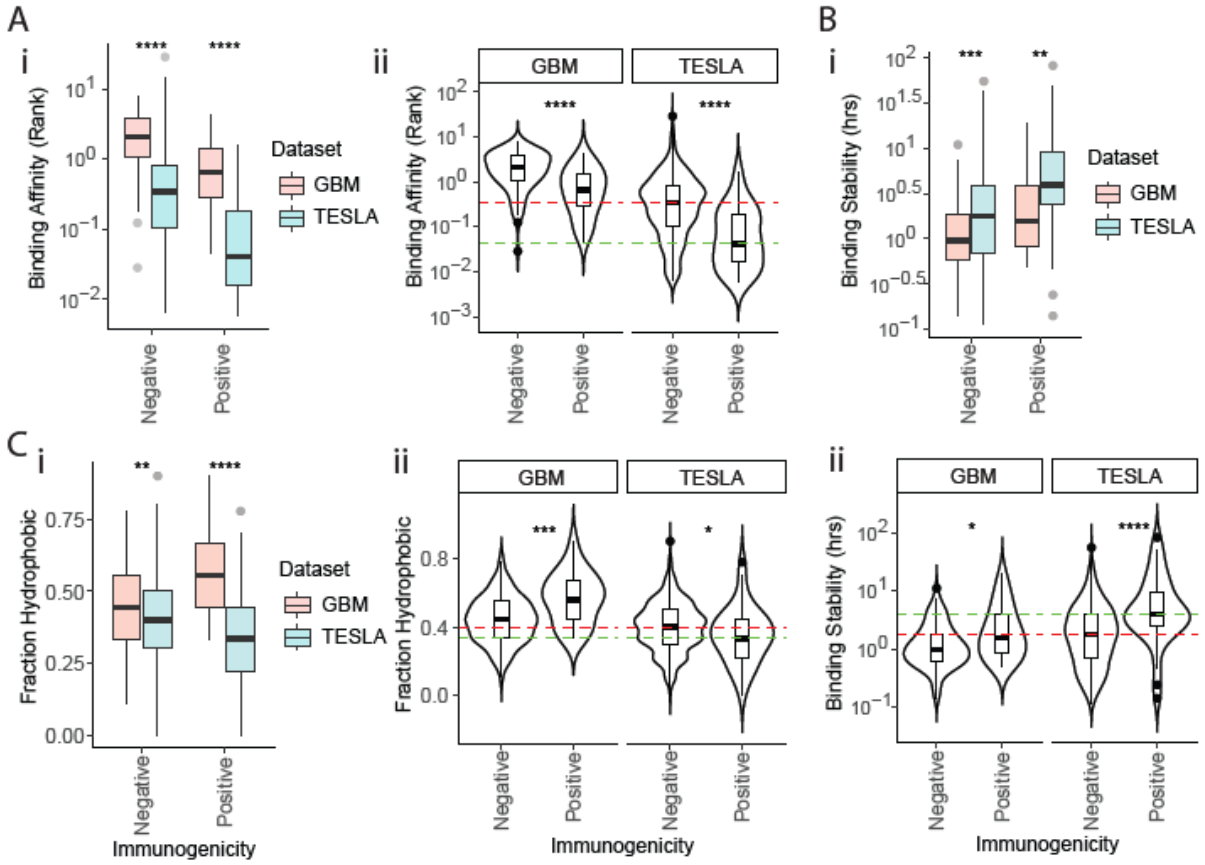


Figure 6.5: Comparing MHC presentation features on two independent cancer datasets, in-house glioblastoma (GBM) and TESLA datasets. A-C. Boxplots and violin plots comparing epitopes and non-epitopes between GBM and TESLA datasets for binding affinity (A), binding stability (B) and hydrophobicity (C). Green and red dashed lines show the median of binding stabilities for cancer (TESLA) epitopes and non-epitopes, respectively. Significance was assessed using Wilcoxon tests.

Dissimilarity to self-proteome have low discriminative power for pathogenic peptides

With previous studies reporting dissimilarity to self-proteome as a T-cell recognition feature for cancer neoepitopes[210,211], we investigated whether the same feature can discriminate pathogenic epitopes. The sequence similarity between pathogenic peptide and human proteome was calculated using the BLOSUM62 substitution matrix, and there observed no statistical difference in dissimilarity to self between pathogenic epitope and non-epitopes (Figure 6.6A).

One reason might be contributed to highly heterologous sequence of pathogenic peptides to human proteome when compared to cancer neoepitopes. Looking at Hamming distance, the majority of pathogenic peptides had 2-5 amino acid differences (Figure 6.6B) whereas neoepitopes had single nucleotide variations (SNVs) or indels. Notably, the dominance of SNVs and indels in cancer neoepitopes is attributed to the inability of current methodologies to call CD8+ T-cell targets from peptides with frameshift or structural variations. Given the inherent differences between neoepitope and pathogenic peptides, different features would be required to estimate T-cell recognition potential.

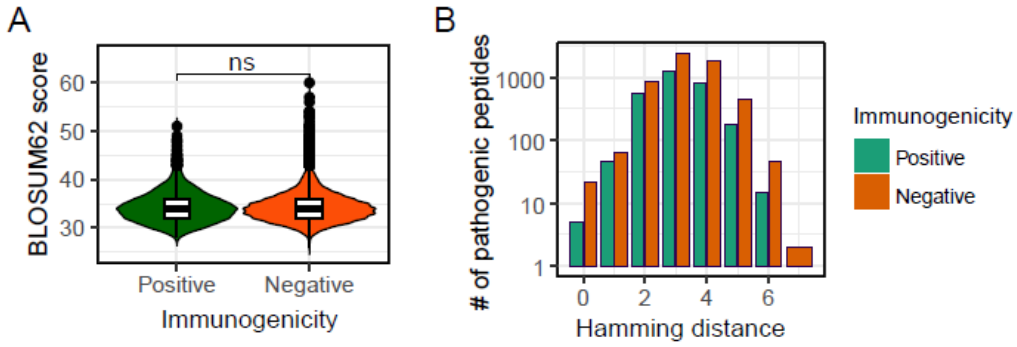


Figure 6.6: Dissimilarity to self-proteome cannot distinguish pathogenic epitopes from non-epitopes. A. Retrieved IEDB pathogenic peptides tested by T-cell assays, and identified those predicted to bind respective HLAs by NetMHCpan 4.1. Conducted optimal global-local alignment between 2,875 pathogenic epitopes and 5,683 non-epitopes to human proteome. Statistics shown is p-values from Student's t-test (p value = 0.151). B. Hamming distance between pathogenic peptides and best counterparts in human proteome in log10 scale.

Cross-species variation in T-cell recognition features

In addition to comparing pathogenic and cancer neoepitopes, we investigated whether the same recognition features can be applied to peptides derived from different pathogenic species (Figure 6.7A). First, it was evaluated to what extent known markers of immunogenicity, such as MHC binding affinity[53,60] and the percentage of hydrophobic amino acids in peptide sequence[209], can be used to identify epitopes from different species. Notably, these features have been used as predictors in models such as TESLA[52], iPred[54], NetTepi[53] and GAO[138]. Although two features could effectively distinguish epitopes from non-epitopes at macroscopic level, where epitopes had

higher MHC binding rank (-log transformed) and hydrophobicity with medium effect sizes (Figure 6.7B-C), they no longer showed statistical significance on many peptides at the species-level (Figure 6.7D-E). For example, SARS-CoV-2 epitopes did not differ significantly from non-epitopes on hydrophobicity.

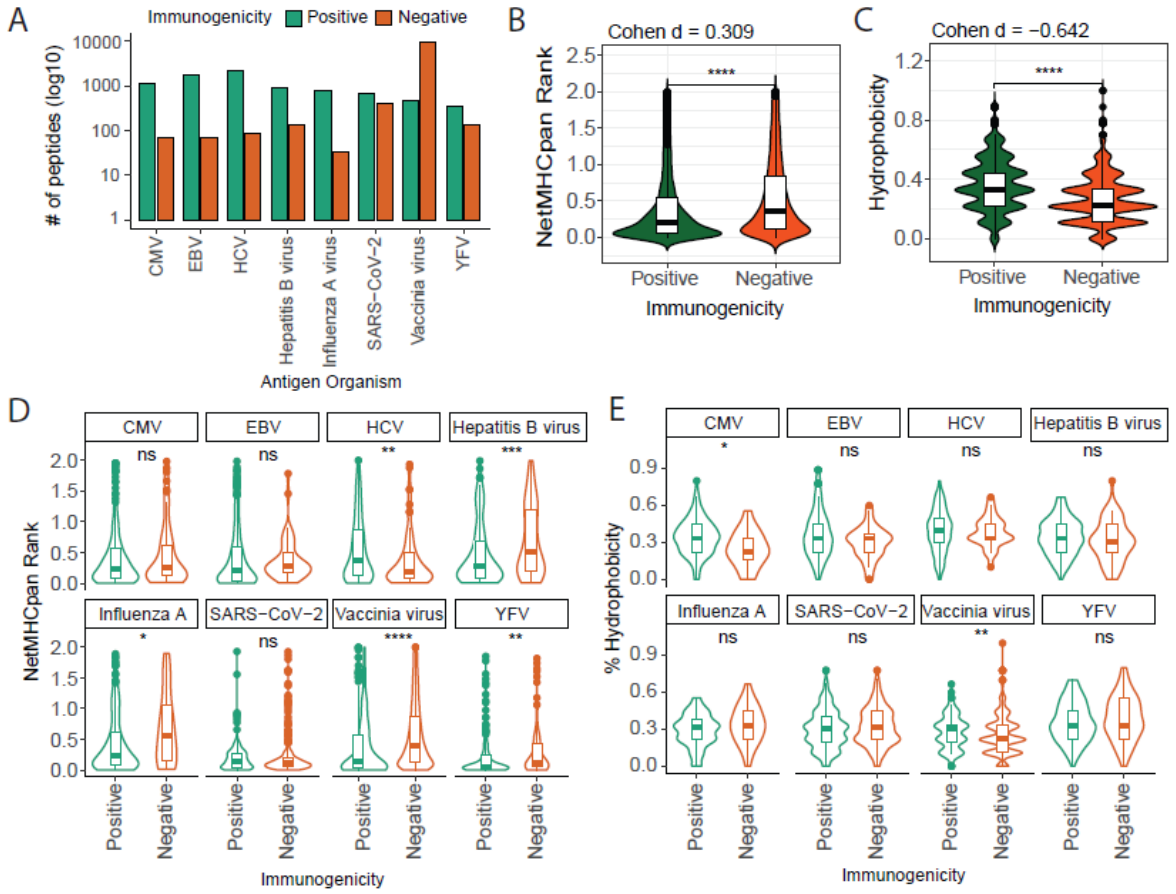


Figure 6.7: Cross-species variation in T-cell recognition features. A. Statistics of peptides in pathogenic data by species of origin. B. Distribution of MHC binding rank (-log2 transformed) predicted by NetMHCpan 4.0. C. Proportion of hydrophobic amino acids, A, V, L, M and W, in peptide sequence. Contains the Cohen's d value showing the effect sizes differentiating positives and negatives as small (d = 0.2), medium (d = 0.5) and large (0.8). D. Distribution of hydrophobicity facet by species of origin. E. Distribution of MHC binding rank facet by species of origin.

Second, the ability of cross-species peptides (i.e. peptides derived from other species) to predict epitopes of unseen pathogens was assessed. The peptideTcell data was divided randomly or in a cross-species manner: i) random split to 90% train vs. 10% test (i.e. representation of 10-fold cross-validation), or ii) cross-species split to Non-SARS-CoV-2 (SARS-2) train vs. SARS-2 test, and Non-vaccinia virus (VACV) train vs. VACV test (Figure 6.8A). The performance of XGBOOST classifiers was then compared on these datasets (Method). While no substantial difference was observed during training (Figure 6.8B), cross-species model showed lower accuracy

and higher root mean squared error (RMSE) (Figure 6.8C). This implied that cross-species peptides may not share common features associated to predicting immunogenicity (i.e. predictive features in training and test data are likely to be different which is known as out-of-distribution generalization problem), resulting limited accuracy on unseen pathogen.

To support this, differential position specific scoring matrices (dPSSMs) were generated to compare immunogenicity-related sequence patterns between train and test datasets (Methods). While random split data shared similar patterns between train and test datasets, such as enrichment of L, G/E, G, hydrophobic resides and T on P3-P8 (Figure 6.9A), cross-species data showed low homology (Figure 6.9B-C), and dPSSM scores failed to predict immunogenicity (Figure 6.9D, Method).

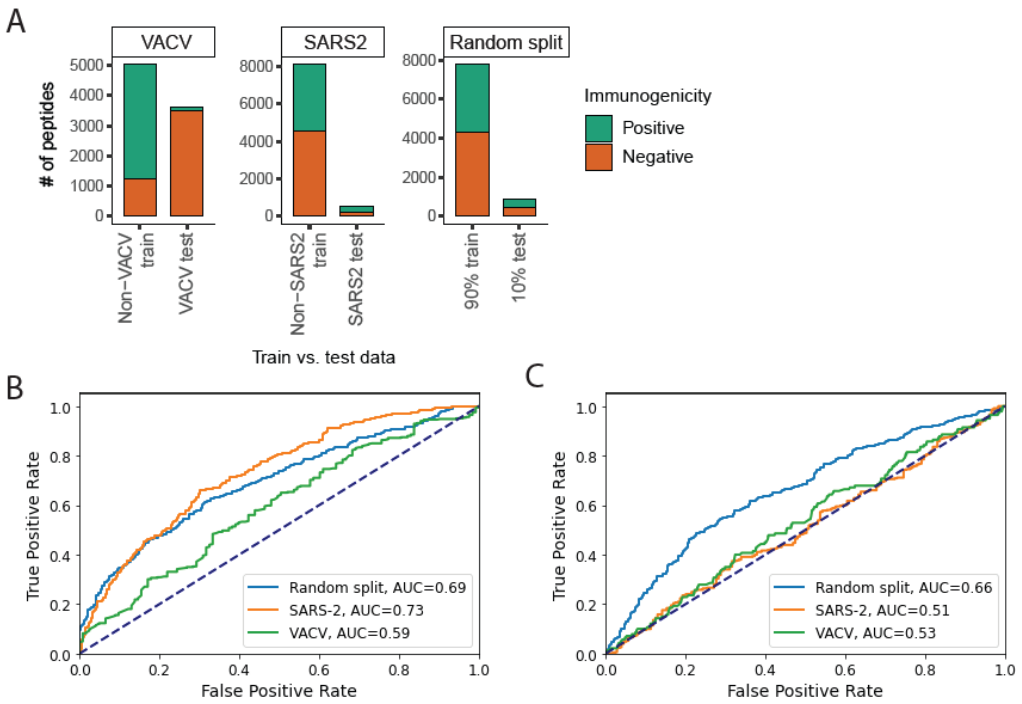


Figure 6.8: Models trained using cross-species peptides cannot effectively predict immunogenicity of peptides derived from unseen pathogen. A. Statistics of peptides in cross-species dataset i.e. split non-vaccinia virus (non-VACV) peptides for training and VACV peptides for testing, non-SARS-CoV-2 (non-SARS2) peptides for training and SARS-2 peptides for testing, and data randomly split into 90% train and 10% test, as resemblance of 10-fold cross-validation. B-C. ROC-AUC curves of XGBOOST classifiers trained by 90% random data, non-SARS2, non-VACV peptides during 10-fold validation (B) and on test datasets - 10% random data, SARS2 and VACV peptides, respectively (C).

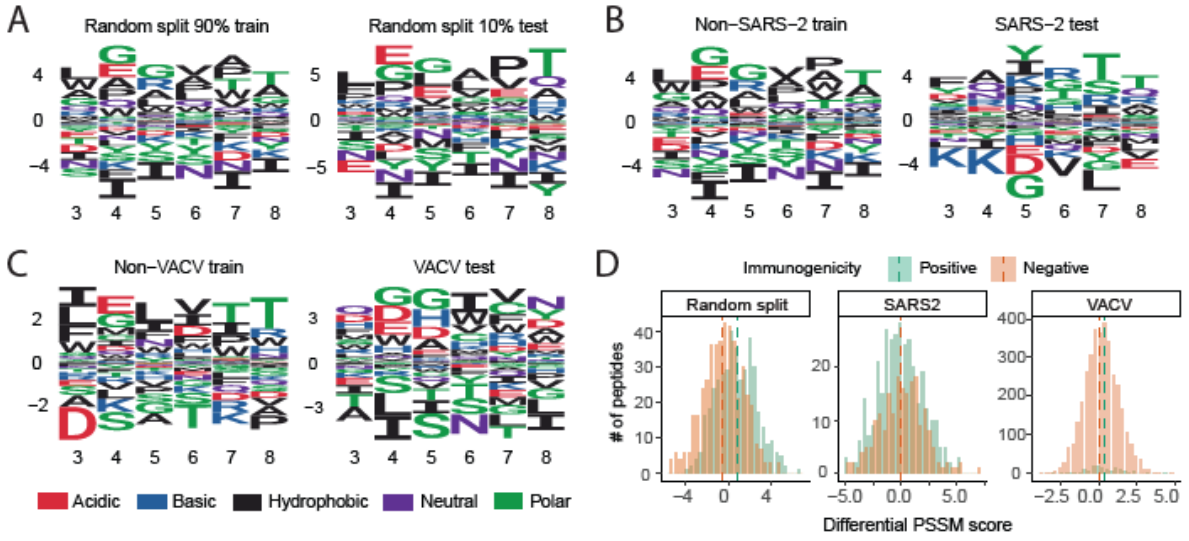


Figure 6.9: Sequence patterns associated with immunogenicity across different pathogens. A-C. Sequence logo of amino acids enriched in epitopes (Positive) compared to non-epitopes (Negative), in each position of contact residues for randomly split data (A), SARS-2 split data (B) and vaccinia virus split data (C). D. Distribution of differential position specific scoring matrix (dPSSM) score, computed by using dPSSMs trained using 90% random data, non-SARS2, non-VACV peptides on 10% random data, SARS2 and VACV peptides, respectively.

High performance of immunogenicity models incorporating peptide-HLA pairs may be a reflection of HLA bias from imbalanced dataset

In addition to the out-of-distribution generalization problem, it was found that high reported performance from latest models incorporating peptide-HLA pairs was driven by HLA bias introduced by imbalanced dataset. Recently, deep learning models incorporating peptide-HLA pairs reported ROC-AUC of ~0.85 by 10-fold cross-validation[51] (Figure 6.10). While they reported the highest accuracy to date, we observed poor performance on SARS-CoV-2 peptides in a benchmarking study[56]. Here, we hypothesized that this poor agreement may be due to differences in datasets and investigated the cause of disagreement.

First, to evaluate the ability of the model to discriminate Positive vs. Negative bound on the same HLA allele, the model was trained and tested in single HLA-level on HLA-A*02:01 and HLA-A*24:02, which had the highest numbers of characterised peptides. This was then compared to a model trained with the same number of randomly sampled pMHCs (Figure 6.11C). Here, the model trained on HLA-A*02:01 peptides performed marginally better than random (Figure 6.11A), and the model trained on HLA-A*24:02 peptides showed high variability (Figure 6.11B). In addition, a model was trained using HLA-balanced data (i.e. the same number of epitopes and non-epitopes for each HLA), and HLA-balanced training data substantially reduced classifier performance (Figure 6.12). This suggested that current models are biased toward classifying for certain over- or under- represented HLA alleles, and that their reported performance does not

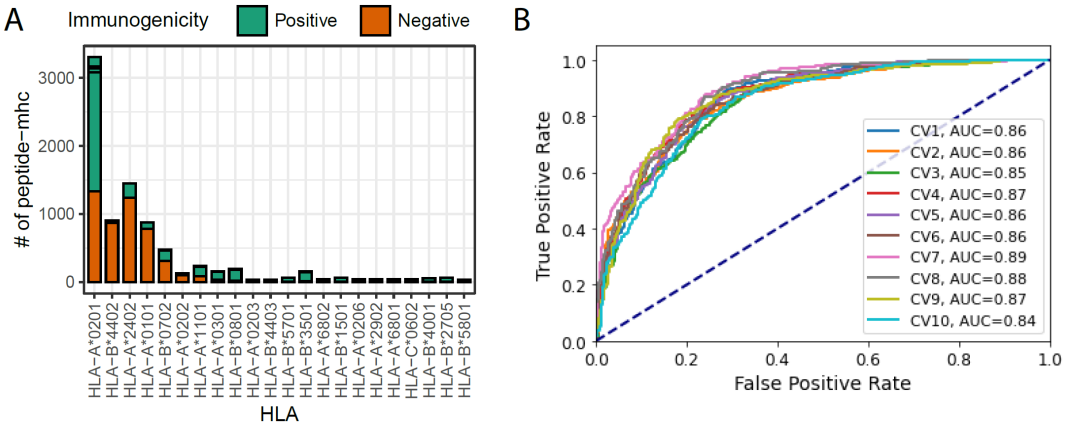


Figure 6.10: Performance of DeepImmuno. A. Statistics of peptides for each HLA allele from DeepImmuno training dataset. B. ROC curve reproducing DeepImmuno performance by 10-fold cross-validations.

reflect their true-world accuracy.

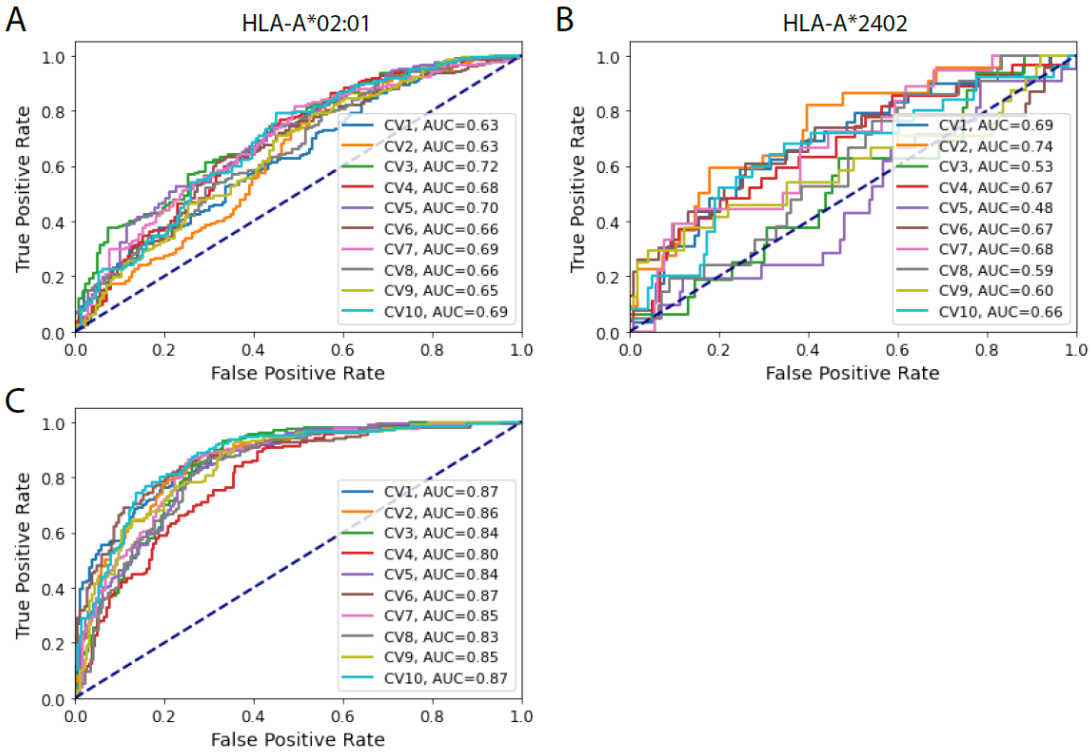


Figure 6.11: High performance may be a reflection of HLA bias. A-B. ROC curve showing performance on single HLA allele, on HLA-A*02:01 only (A) or HLA-A*24:02 only (B). C. ROC curve showing performance on randomly sampled 3303 peptide-HLA pairs (i.e. equivalent number of pairs for HLA*02:01) to demonstrate that reduced accuracies is not due to lower number of training data.

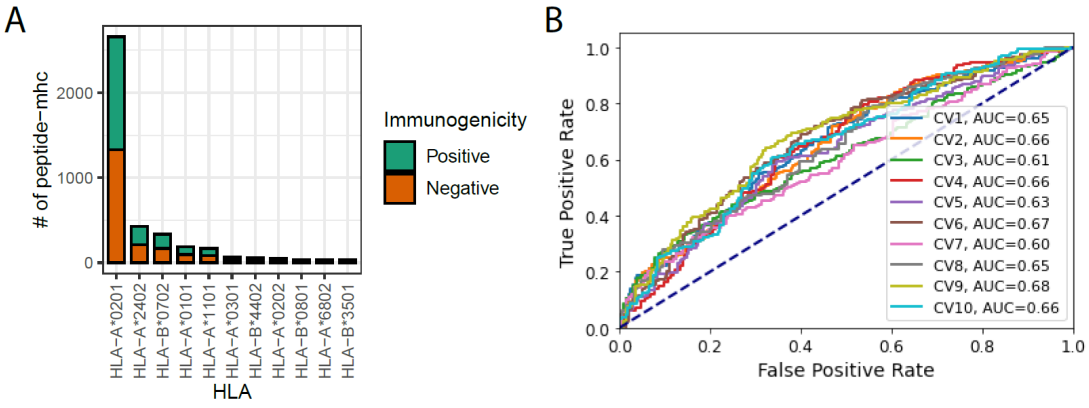


Figure 6.12: DeepImmuno show lower performance on HLA-balanced data. A. Statistics on number of peptides per HLA after down-sampling to balance number of epitopes and non-epitopes. B. Performance of classifier using down-sampled dataset to balance number of epitopes and non-epitopes per HLA allele.

Discussion

In this chapter, potential sources of poor performance were investigated to help guide the development of future models. We first reported that MHC binding may not translate to immunogenicity, especially in human context, and that T-cell recognition features are required to accurately predict immune potential. In fact, there is a growing consensus that MHC presentation alone is insufficient to elicit an immune response for cancer neoepitopes[52,191,224]. Our finding supports this claim and suggest that additional T-cell recognition features would be required to decipher T-cell targets from HLA ligands.

The out-of-generalization problems were showcased across cancer neoepitopes and pathogenic peptides, as well as peptides derived from various pathogenic species. The dissimilarity to the self-proteome used as an example hallmark from cancer neoepitope studies that may not apply to pathogenic peptides. Indeed, Koncz *et al.* recently reported that while dissimilarity to the self-proteome is essential to discriminate self versus nonself, highly dissimilar peptides may be less immunogenic. This is because T-cells reactive to peptides that are highly dissimilar to the self-proteome will not have survived from self-mediated positive selection[225]. Therefore, while dissimilarity to self-proteome at a moderate level may indicate immunogenicity e.g. cancer neoepitopes with single nucleotide variation from self, this may not be a linear relationship, especially at a high level of dissimilarity, such as pathogenic peptides.

The low performance was also attributed to divergent discriminative features between cancer neoepitopes and pathogenic epitopes in directionality or magnitudes[56,226]. Our work suggests that the presentation features, such as binding affinity, peptide stability, and fraction of hydrophobicity, are associated with peptide immunogenicity for both cancer neoantigens and viral peptides, albeit at different thresholds. Finally, the high performance of models incorporating peptide-HLA pairs was shown to be a reflection of HLA bias from an imbalanced dataset.

In addition to model-related problems, data-related issues were observed, which may impede the development of an accurate classifier. These include: i) small sample size especially for less prevalent HLA alleles, ii) an imbalanced number of epitopes and non-epitopes across many HLA alleles, iii) variation in functional assays employed for characterization with varying degrees of sensitivity and specificity, and iii) lack of true 'negatives' as they can be recognized by at least one TCR in the repertoire.

Taken together, these findings suggest that models trained by pathogenic peptides may not be directly applicable to cancer neoepitopes and vice versa, and that different features may be required to shortlist epitopes in pathogen vs. cancer settings, with a greater emphasis on T-cell recognition residues. We hope that the findings from this study will assist end users in making evidence-based decisions about which model and parameters to use with their data and research questions.

Methods

Data for pathogenic vs. cancer comparison

For cancer neoepitopes, the 'TESLA' dataset was downloaded from Wells et al[52]. In this study, 608 neoepitopes from 6 patients were predicted, of which 37 were characterized to be immunogenic. All 608 peptides of lengths 8-14 amino acids were employed for this analysis. To curate the 'non-cancer pathogenic' dataset, MHC class I peptides was retreived from the IEDB (accessed 17-02-2022) and combined with peptides reposited in *Repitope* package. Peptides of lengths 8-14 and those having 4-digit HLA resolution were retained and peptides derived from "homo sapien" or "cancer" were excluded. Any peptides having contradictory immunogenicity observations, i.e. pMHCs characterized positive and negative by T-cell assays, were annotated as positive adopting the approach described previously by Ogishi et al[50]. Lastly, only peptides predicted to bind their corresponding HLA allele (NetMHCpan 4.0 with a cutoff Rank threshold of 2.0) were retained for analysis.

Data for pathogenic comparison

Similar strategy was used to prepare data for comparing immunogenicity-related features across peptides derived from different pathogenic species. The peptides of lengths 9-10 amino acids were retrieved from IEDB and underwent similar filtering strategies. To tackle the problem of no true negatives, immunogenicity potential was considered as a continuum with some peptides likely to trigger greater T-cell response and/or able to bind numerous TCRs. Therefore, 'Negatives' were defined as peptides having very low immunogenicity potential that have not yet triggered T-cell response in $\geq$3 tests. Peptides having contradictory immunogenicity annotations were categorized as 'Positive' and only retained Negatives $\geq$3 tests. This resulted 5,093 immunogenic ('Positive') and 6,628 non-immunogenic ('Negative') peptides. For pathogenic datasets, we subset peptides originating from non-human species, resulting in 4,000 Positive and 6,097 Negative pathogenic peptides. The dataset include information about peptide sequence, binary immunogenicity, HLA allele, source antigen and organism.

DeepImmuno training data

DeepImmuno training data retrieved from GitHub repository, including peptide and HLA information. https://github.com/frankligy/DeepImmuno

MHC binding and stability predictions

For analysis, NetMHCpan4.0 was used to predict binding affinities of peptides to corresponding HLAs. NetMHCstabpan[181] was used to predict binding stabilities.

Hydrophobicity

The 'fraction of hydrophobicity' was calculated as the fraction of a peptide's residues that were hydrophobic. Hydrophobic residues were considered as "V", "I", "L", "F", "M", "W" and "C".

Dissimilarity to self

The sequence similarity between pathogenic peptides and human proteome counterparts was computed by first identifying best global-local alignment by "pairwiseAlignment" function from Biostrings v2.40.2 package with a high gap penalty, gapOpening of 100 and gapExtension of 100. The number of mismatches between best aligned pair was computed by hamming distance using "stringdist" function from R stringdist v0.9.5.5 package.

Differential amino acid frequencies

The probability frequency of each amino acids in each position was represented by position specific scoring matrices using "consensusMatrix" function from R Biostrings v2.56.0 package. The dPSSMs from Positive and Negative peptides were standardised by centre and scaling, and differential dPSSM was generated by computing the difference between the two. The differential PSSMs were generated using random split, VACV and SARS-2 datasets. The predictability of dPPSM scores was evaluated by leave-one-out cross-validations.

Amino acid usage classifier

Each amino acids in peptide sequence at contact positions were represented by their physico-chemical properties using "aaDescriptors" function in R Peptides v2.4.4 package. The amino acid descriptors include kideraFactor, zScales, tScales, vhseScales, protFP, stScales, blosumIndices, mswhimScores, crucianiProperties and fasgaiVectors, which describe properties such as polarity, electronic properties, hydrophobicity, α-helix/bend preference, β-sheet, bulkiness/size of side-chains, hydrogen-bonding, isoelectric point and structural topology[227]. In addition to amino acid-level embedding, the peptide-wide property was added by averaging across all positions for each peptide sequence. The embeddings from random split and cross-species datasets were used to generate XGBoost classifier using "XGBClassifier" function in python xgboost v0.90 package. The set of hyperparameters, such as alpha, gamma, max_depth and colsample_bytree, were optimized by grid search for each dataset.

Chapter 7. A deep learning-based platform to predict T-cell recognition potential of MHC-I presented pathogenic and self-peptides

To be published in:

Lee CH *et al.*, TRAP: A robust deep learning workflow to predict CD8+ T-cell epitopes. *Manuscript in preparation*

Introduction

T-cell recognition of pathogenic peptides, autoantigens and cancer neoepitopes presented by MHC molecules has been extensively studied, especially for the applications in cancer immunotherapies or vaccine development [228–230], as well as to better understand T-cell dynamics in maintaining health and fighting diseases. However, current experimental approaches for the identification of T-cell targets are low throughput and laborious[32,63,231].

While there have been great efforts to foster the process, computational prediction of T-cell epitopes i.e., T-cell recognition potential of MHC-presented peptides, also remains challenging. Previous efforts to decipher properties underlying T-cell recognition of peptide-MHC complexes (pMHCs) identified large and aromatic side chains[49] and hydrophobicity[54,177,182,209] to be associated with immunogenicity. Specifically for cancer neoepitopes, Wells *et al.* reported agretopicity (i.e. the ratio of binding affinity between neoepitope and wild-type counterpart) and foreignness score (i.e. similarity of neoepitope to previously characterized epitopes) as correlates of T-cell recognition of cancer neoepitopes[52]. Furthermore, Richman *et al.*[210] and Devlin *et al.*[211] demonstrated that dissimilarity to the non-mutated human proteome is another recognition feature that allows neoepitopes to escape immune tolerance.

However, we have previously described distinct discriminative features between cancer and pathogenic epitopes[226] (Chapter 6). Along with the context-specific properties, other factors, such as limited training data and highly diverse T-cell receptor (TCR) repertoire, which leads

to less conserved recognition motifs, contribute to the challenge of predicting T-cell recognition potential. As a result, many existing models perform poorly against both cancer and emerging viral pathogens[56].

Over the last decade, deep learning and natural language processing (NLP) have transformed biomedical research and have offered breakthrough discoveries. Because of their ability to extract complex sequential or non-sequential patterns from large amounts of data, deep learning architectures have been used in the prediction of peptide-HLA binding[60,232,233] and TCR specificity[214,234,235]. Moreover, transformer-based pretrained language models (PLMs), which revolutionised modern NLP by combining the power of transformers, self-supervised learning and transfer learning[236–238], have advanced the prediction of protein structure and function[239–242]. In particular, the protein transformer based PLMs contain background knowledge learnt form large volume of protein sequences that can be transferred to downstream tasks.

However, even though such models perform well with data on which they were developed for, they often lack the ability to generalized to out-of-distribution data due to lack of domain-specific knowledge.

This study presents TRAP (T-cell recognition potential of HLA-I presented peptides), a robust deep learning workflow to predict CD8+ T-cell epitopes from MHC-I presented pathogenic and self-peptides. To circumvent the causes of poor performance discussed in previous chapters, TRAP provides a step-by-step workflow that addresses the issues and achieves superior and reliable performance. Finally, we show how TRAP can reliably predict CD8+ epitopes, and can be used to study the potential of T-cell immune escape of SARS-CoV-2 mutants. While many existing models predicted immunogenicity based on MHC binding, the TRAP workflow takes one step ahead by predicting immunogenicity from HLA-I presented pathogenic and self-peptides. The accurate prediction of CD8+ epitopes by the help of TRAP will greatly accelerate shortlisting targets for therapeutic applications.

Results

TRAP: A robust deep learning workflow to predict CD8+ T-cell recognition of MHC-I presented pathogenic and self-peptides

The TRAP (T-cell recognition potential of HLA-I presented peptides) is presented as a comprehensive platform for predicting CD8+ T-cell immunogenicity of HLA-I presented pathogenic and self-peptides (Figure 7.1A).

In this workflow, peptide sequences of 9-10 amino acids in length are encoded using amino acid embeddings derived from protein transformer-based pretrained language models (PLMs) by transfer learning. TRAP then uses a 1D convolutional neural network to extract T-cell recognition motifs of varying lengths, as well as a decision tree classifier to eliminate low-confidence predictions to improve accuracy. The immunogenicity of pathogenic peptides with low-confidence predictions is predicted using a novel metric called Relative Similarity to Autoantigens or Tumor-associated antigens (RSAT), which was developed as a complement to "dissimilarity to self" from cancer studies.

The performance of TRAP was benchmarked against existing immunogenicity models, namely NetTepi, IEDB, PRIME, DeepImmuno and TESLA by 10-fold cross-validation on the same datasets (Method). Due to the nature of the existing models, HLA-agnostic predictions were performed for IEDB, iPred, NetTepi PRIME and Repitope, and HLA-restrictive prediction for DeepImmuno. The HLA-restrictive prediction was done using HLA-balanced data where the number of epitopes and non-epitopes were balanced per HLA to validate their prediction irrespective of HLA-associated bias. TRAP was one of the best pathogenic models, achieving ROC-AUC of 0.751 for HLA-agnostic and 0.709 for HLA-restrictive predictions (Figure 7.1B). It was also the best self-antigen model, with ROC-AUC of 0.931 for HLA-agnostic and 0.703 for HLA-restrictive predictions (Figure 7.1C). It is worth noting that the higher performance of self-antigen model is due to the inclusion of a broader range of MHC binding scores to overcome with data limitation, and thus a higher association between binding score and CD8+ T-cell epitope prediction.

While Repitope slightly outperformed TRAP in pathogenic datasets, the robustness of its prediction on emerging pathogens was assessed. The coronavirus TRAP and Repitope models were trained using 1511 Non-Wuhan SARS-CoV-2 peptides (Figure 7.1D), which were then used to predict immunogenicity of 66 Wuhan SARS-CoV-2 peptides (Figure 7.1E). Here, TRAP was better at extracting immunogenicity-related features from limited data of emerging pathogens (Figure 7.1D), and translating them onto related species (Figure 7.1E).

Therefore, TRAP can predict CD8+ T-cell epitopes in both HLA-agnostic and restrictive contexts, and performed better in both. The TRAP can also make accurate predictions about emerging pathogens and suitable for shortlisting therapeutic candidates when data is limited. The following sections will discuss the strategies used to overcome constraints and improve performance.

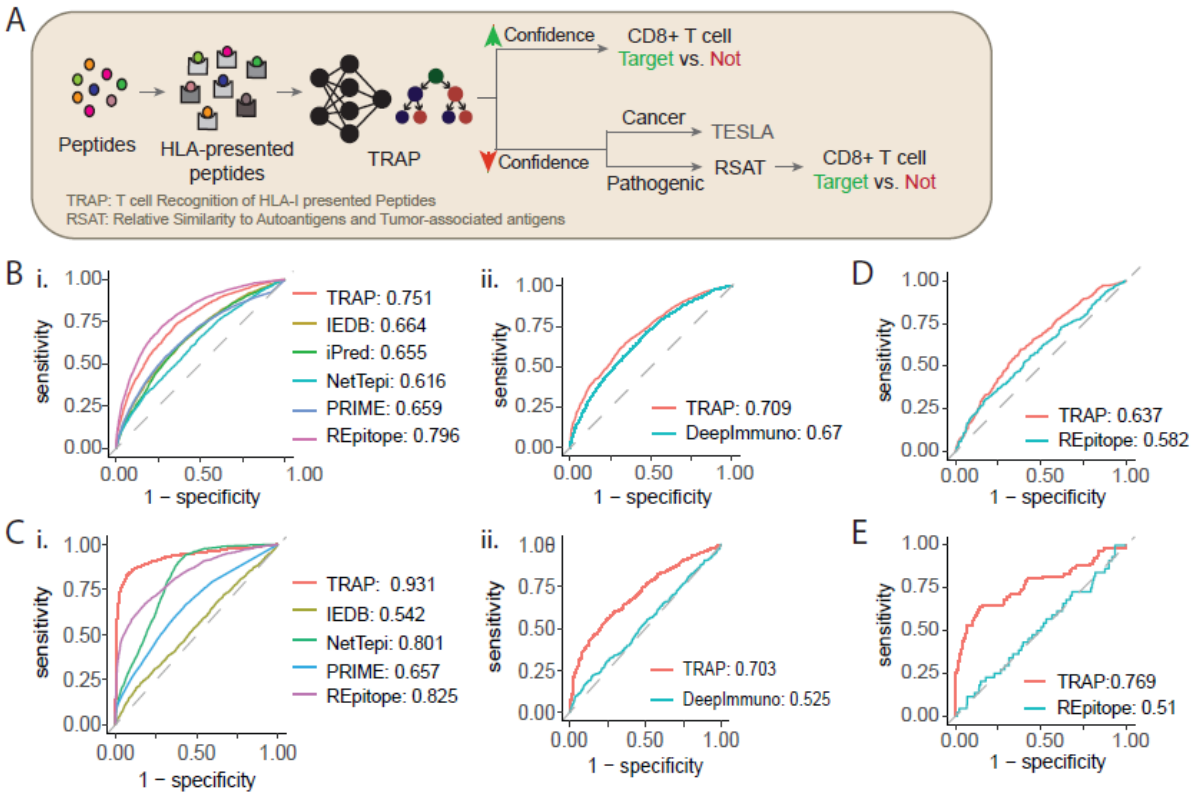


Figure 7.1: Performance of TRAP. A. Schematic diagram of TRAP (T-cell recognition potential of HLA-I presented peptides), a robust deep learning based workflow to predict CD8+ T-cell epitopes from MHC-I presented pathogenic or self-peptides. Once peptides have been predicted by NetMHCpan to bind HLA alleles, the TRAP uses the peptide sequence and NetMHCpan rank scores as inputs to predict immunogenicity of the peptide with the respective HLA binding affinity. The TRAP predicts immunogenicity and provides confidence in the prediction. If the prediction is abstained due to low confidence, we recommend predicting cancer neoepitopes using TESLA[52], which is known to use more general features such as agretopicity and dissimilarity to self-proteome, and pathogenic peptides with RSAT (Relative Similarity to Autoantigens or Tumor-associated antigens). B-C. ROC curve comparing performance of TRAP with existing models, such as IEDB, iPred, NetTepi, PRIME, Repitope and DeepImmuno for pathogenic (B) and self-peptides (C). Due to nature of these models, IEDB, iPred, NetTepi PRIME and Repitope were predicted in HLA-agnostic manner (i.e.by peptide sequence only) (i), and DeepImmuno predicted in HLA-restrictive manner (i.e. by peptide-HLA) using HLA-balanced data (ii). D. ROC curve comparing performance of TRAP and Repitope that have been trained using 1511 Non-Wuhan SARS-CoV-2 peptides by 10-fold cross-validations. E. ROC-curve comparing performance of Non-Wuhan SARS-CoV-2 trained TRAP and Repitope models in predicting 66 Wuhan SARS-CoV-2 peptides.

Mitigate HLA bias by employing peptide sequences at TCR contact positions

Previous studies reported contribution of anchor positions (position 2 (P2) and P9 of 9aa peptide) for MHC binding[243] and contact positions (P3-P8) in T-cell recognition[49]. Correspondingly, HLA supertypes drove clustering of peptides at anchor positions (Figure 7.2A) and TCR specificities at contact positions (Figure 7.2C-D) on our data, with peptides bound by the same TCR showing conserved motifs (Figure 7.2B). However, because of the strong conserved pattern in anchor positions, HLA supertype not only drove clustering peptides at anchor positions but also acted as the strongest covariate driving clustering of peptides in full sequence (Figure 7.2E). With such a strong conserved pattern, MHC binding features may dominate immunogenicity predictions when the full peptide sequence is used for model training. In fact, we observed that some of existing immunogenicity classifiers were more capable of predicting dominant HLA type (i.e. whether peptides were bound to HLA-A*02:01 or not) than peptide immunogenicity[56]. As we aimed to predict T-cell recognition potential once peptides are bound to HLA alleles, we incorporated contact positions only i.e. positions 3-8 (P3-P8) of 9aa peptide and P3-P9 of 10aa peptides, not only to focus on T-cell recognition patterns, but also to rule out the need to balance number of epitopes vs. non-epitopes by HLA supertypes for model development.

Moreover, we investigated whether neglecting HLA information when predicting T-cell recognition potential of MHC-presented peptides from the current dataset would result in a significant loss of information. First, a novel metric was developed for converting qualitative measurements (i.e. Positive or Negative) to a quantitative 'positivity score' that reflects confidence in the immunogenicity (Methods). Briefly, the metric took into account i) the number of experiments conducted (Figure 7.3A), ii) percentage tested positive (Figure 7.3B) and iii) number of cognate TCRs where available (Figure 7.3C) to compute the positivity score. Due to limited pool of peptides with cognate TCRs, the majority of peptides' positivity score was computed by translating the number of experiments conducted and % tested positive, and the number of cognate TCRs was incorporated to add greater weights to positivity (all score with cognate TCR > 1). The positivity scores were computed for each entry for each peptide-HLA pair, and ranged from 0 to 2.3, with Negatives ranging from 0-1 and Positives ranging from 1-2.3 (Figure 7.3D).

Using the positivity scores of each peptide-HLA entry, the extent of inter-HLA variation (i.e. variation across different HLA alleles) was compared to intra-HLA variation (i.e. variation within the same HLA allele) for peptides bound on multiple HLA alleles. The intra-HLA variation accounted for the majority of data variability, while inter-HLA variation contributes less (Figure 7.3E). This is due to other biological (e.g. different effector cell, antigen presenting cells etc.) or technical (e.g. assay type, culture conditions, research groups etc.) factors causing greater discrepancies in positivity score than HLA alleles (Figure 7.3F, Table 1). This implied that when it came to each peptide, there wasn't much difference in immunogenicity based on the HLAs to which it was bound, but when it came to HLAs, each HLA This implied that there wasn't much difference in immunogenicity between peptides based on the HLAs to which they were bound for this dataset. When it came to HLAs, however, each HLA had different pool of peptides, resulting in a different distribution of positive and negative peptides, resulting in HLA-bias. As a result, an HLA-generalised approach was adopted in which peptide-HLA binding information

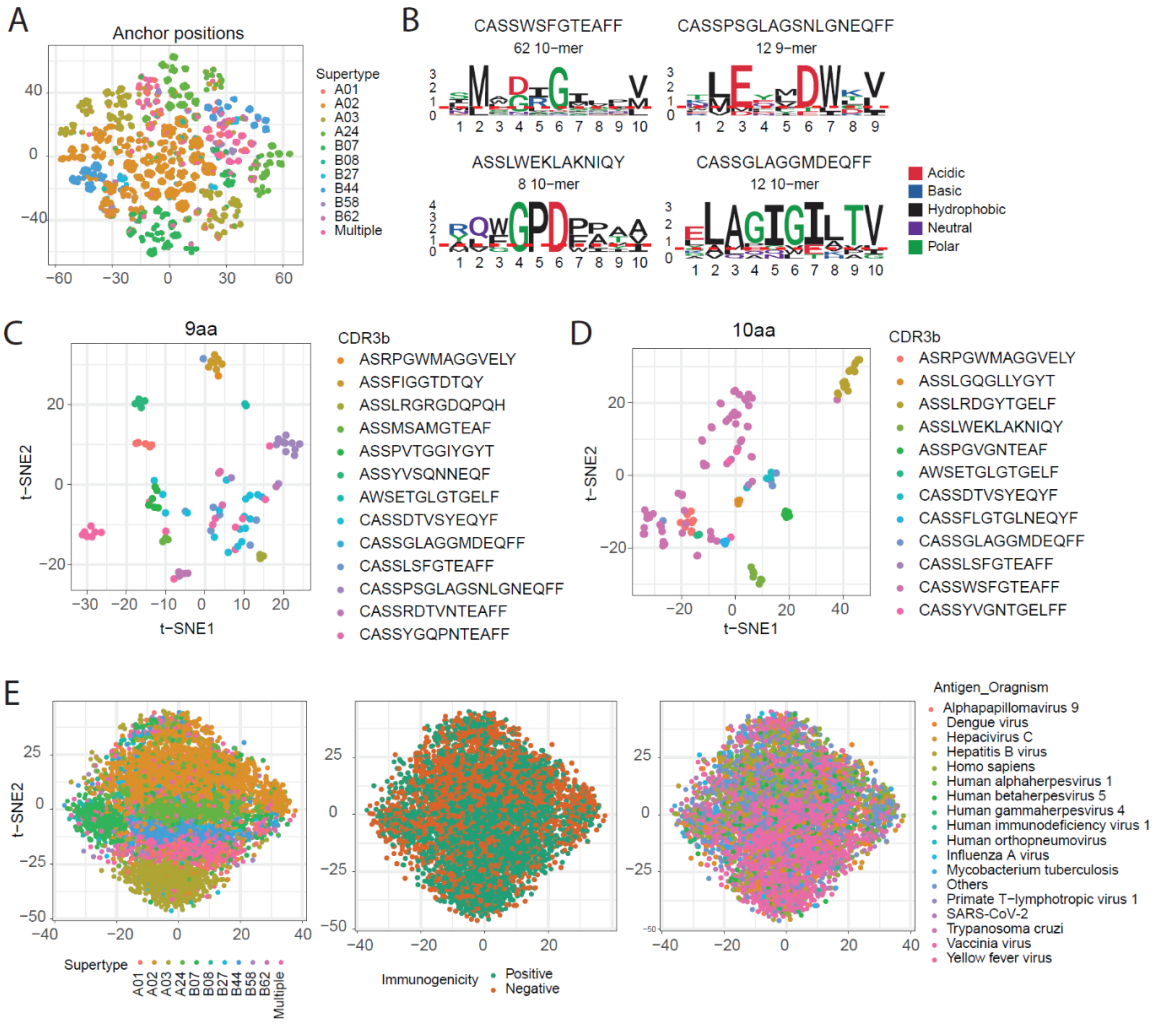


Figure 7.2: Effect of anchor and contact positions on peptide immunogenicity. A. t-SNE embedding of peptides described by amino acid descriptors (aaDescriptors) for amino acids at anchor positions (i.e. P2 and P9 of 9aa peptides) and peptide-wide aaDescriptors (i.e. averaged across full peptide sequence), coloured by HLA supertypes. B. Sequence logos of peptides recognized by the same TCR, coloured by amino acid properties. Notably, these are examples of TCRs having the most number of cognate peptides deposited in the database. However, except for the first TCR, number of peptides may be low to represent sequence conservation with statistical confidence. C-D. t-SNE embedding of peptide-wide descriptors and amino acids at contact positions, coloured by cognate TCRs for 9aa peptides (C) and 10aa peptides (D). E. t-SNE embeddings of amino acids at all positions and peptide-wide descriptors, coloured by HLA supertype (left), immunogenicity (middle) and species of origin (right).

was used instead of the HLA allele itself.

Table 7.1: Sources of intra-HLA variation. The cases of intra-HLA variation include biological and technical factors, such as different assay, effector population and antigen presenting cells in characterizing immunogenicity of peptides bound on the same HLA. '-' means information is not available. For peptide-HLAs without supporting information, positivity scores were denoted as binary values.

Peptide	**Immunogenicity**	**MHC Restriction**	**Antigen Organism**	**# Tested**	**# Responded**	**Response frequency**	**In vitro stimulation**	**Assay**	**Assay group**	**Effector cell culture condition**	**Antigen presenting cells**
GILGFVFTL	Positive	HLA-A*02:01	Influenza A virus	-	-	-	Restimulation in vitro	51 chromium	cytotoxicity	Cell Line / Clone	B cell
GILGFVFTL	Positive	HLA-A*02:01	Influenza A virus	-	-	-	Restimulation in vitro	ELISPOT	IFNg release	Direct Ex Vivo	PBMC
GILGFVFTL	Positive	HLA-A*02:01	Influenza A virus	-	-	-	Restimulation in vitro	multimer/ Tetramer	qualitative binding	Short Term Restimulated	
GILGFVFTL	Positive	HLA-A*02:01	Influenza A virus	6	6	100		ICS	CXCL9/MIG release	Direct Ex Vivo	PBMC
GILGFVFTL	Positive	HLA-A*02:01	Influenza A virus	5	2	40		ELISPOT	IFNg release	Direct Ex Vivo	PBMC
GILGFVFTL	Positive	HLA-A*02:01	Influenza A virus	1	1	100	Restimulation in vitro	51 chromium	cytotoxicity	Cell Line / Clone	B cell

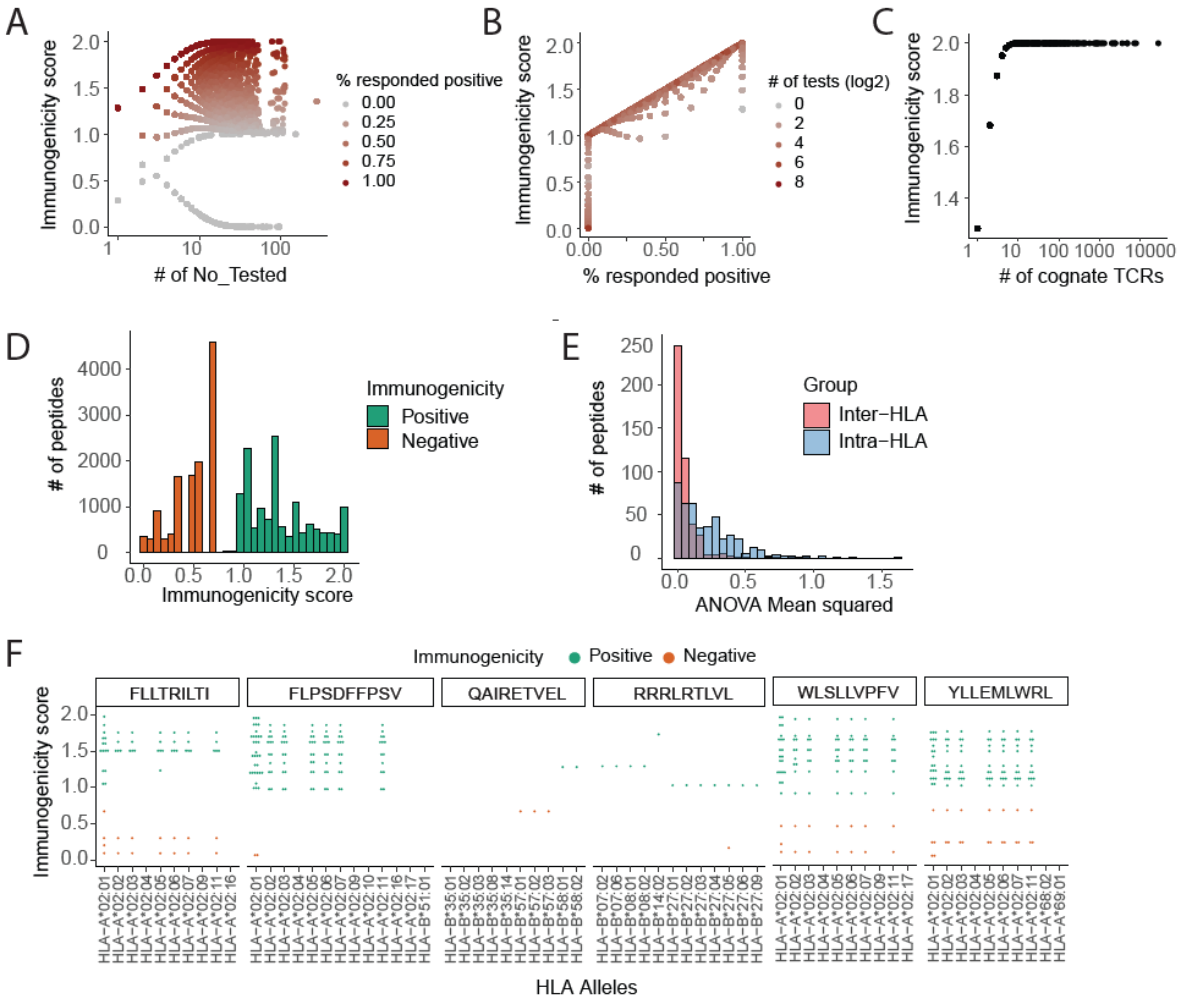


Figure 7.3: Intra- and inter-HLA variation in positivity score. A-C. Convert binary to quantitative positivity score. The positivity score computed using number of test conducted (A), percentage responded positive (B) and number of cognate TCRs (C). D. Distribution of positivity score.. In essence, for positive peptides, higher the positivity score, more likely to be immunogenic. For negative peptides, lower the score, more likely to be non-immunogenic. E. Distribution of intra-HLA and inter-HLA variation for peptides having different peptide-HLA entries. The variation described by ANOVA mean squared. F. Example of peptides having higher intra-HLA variation than inter-HLA variation.

Deep learning architecture captures T-cell recognition motifs

As potential causes of poor performance, previous sections discussed HLA bias, out-of-distribution (OOD) generalisation problem, and a limited breadth of training data. These issues were addressed in TRAP by i) using peptide sequence at contact positions, ii) building separate models for pathogenic and self-peptides, iii) encoding peptide sequences using amino acid properties derived from protein transformer-based pretrained language models (PLMs), iv) devising one-dimensional convolutional neural network (1D CNN) architecture designed to capture T-cell recognition motifs and v) employing OOD detection module, which outperformed other machine learning algorithms.

First, pathogenic and self-antigen datasets were prepared to build context-specific models (Methods). Briefly, pathogenic dataset is a subset of the peptideTcell data, comprising only pathogen-derived peptides (Figure 7.4A). For self-antigen dataset, autoantigens, tumor-associated antigens and cancer neoepitopes were retrieved from Cancer peptide database, dbPepNeo, IEDB, McPAS-TCR and NEPdb databases as epitopes, and benign HLA-I ligands expressed in thymus as non-epitopes (Figure 7.4B-C). The HLA-I ligands (exclusively) expressed in the thymus were assumed to be involved in negative selection of T-cells, and thus no repertoire to recognise these peptides should be present. To the best of our knowledge, this is the first study to use the concept of negative thymic selection to distinguish self-epitopes from non-epitopes.

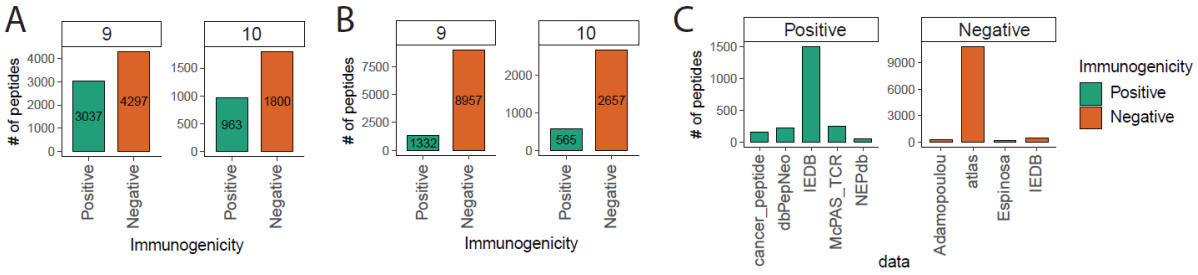


Figure 7.4: Pathogenic and self-antigen datasets. A. Statistics on number of peptides in pathogenic dataset. B. Statistics on number of peptides in self-antigen dataset. C. Statistics on number of self-epitopes and HLA-I ligands expressed in thymus, retrieved from publications and databases.

Second, because peptides vary in lengths, the optimal padding strategy for aligning 6-mer with 7-mer contact position peptides was investigated. The predictive accuracies of simple dense layer classification, 1D CNN, 2D CNN and bidirectional recurrent neural network (biRNN) models were evaluated using pre- and post-paddings. No significant difference was observed across pathogenic and self-antigen data, with pre-padding achieving slightly better performance in self-antigen data on the BiRNN model (Figure 7.5A). Moreover, the relative location of k-mer motifs was compared on 6-mer and 7-mer peptides, and many 3-mer motifs found at P3 of the 9aa peptide were found at P4 of the 10aa peptide (Figure 7.5B). Therefore, to align T-cell recognition hotspots with biological context, peptides of shorter length were pre-padded.

Third, to address data limitation, peptides were encoded using amino acid properties derived

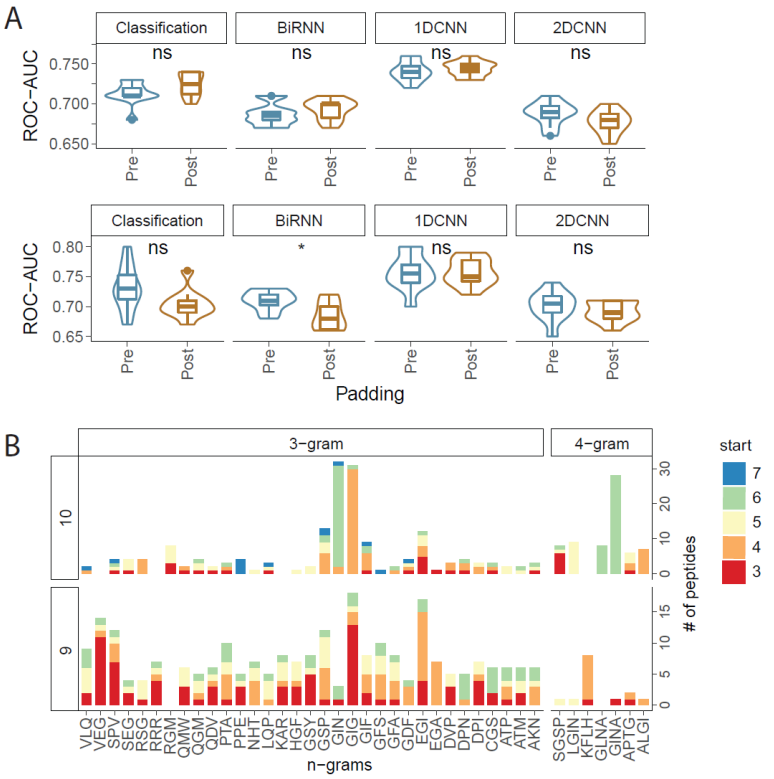


Figure 7.5: Performance of pre- and post-padding strategies. A. ROC-AUC values of pre- vs. post-padding strategies on respective deep learning architectures for pathogenic (top) and self-antigens (bottom). Each point represents ROC-AUC value from one round of 10-fold cross-validations. B. Example of n-grams found in both 9aa and 10aa peptides, coloured by the start position of respective n-gram in the peptides.

from protein transformer-based pretrained language models (PLMs) by transfer learning. The protein transformer-based PLMs were trained using millions to billions of protein sequences and carry the most representative 1024 embeddings, which describe the physicochemical, structural or electrostatic properties of amino acids in protein space[242]. The performance of classification model was compared when peptides were encoded by one-hot-encoding (OHE), amino acid descriptors and embeddings from five protein transformer-based PLMs (Methods). ProtT5-XL-UniRef showed the highest performance, with average ROC-AUC of 0.72 for pathogenic (Figure 7.6A) and 0.724 for self-antigen datasets (Figure 7.6B).

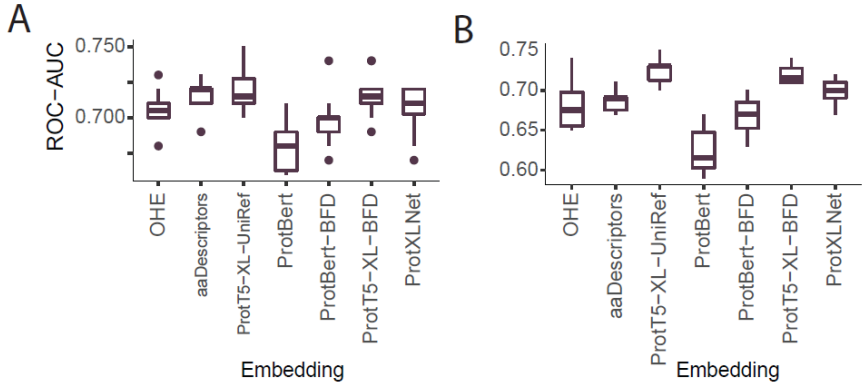


Figure 7.6: Compare amino acid embedding strategies. A-B. ROC-AUC values representing single dense classification prediction using different encoding strategies, including one-hot encoding (OHE), amino acid descriptors (aaDescriptors) and embeddings from protein transformer-based PLMs on pathogenic peptides (A) and self-antigens (B). Each point represents ROC-AUC value from one round of 10-fold cross-validations.

Forth, the capability of different machine learning and deep learning architectures to classify immunogenicity potential was assessed. The performances of XGBOOST, BiRNN, bidirectional long short-term memory (BiLSTM), 1D CNN, 2D CNN and transformer models were compared. Briefly, RNNs or LSTMs are popular sequence models efficient in predicting sequential text data[244], whereas CNNs are suitable for scanning across text or image and detect local patterns by using 'kernels'. 1D CNNs have widely been used for text and 2Ds for image classifications[245]. The 1D CNN model using ProtT5-XL-UniRef embeddings achieved the highest average ROC-AUC of 0.74 for pathogenic (Figure 7.7A) and 0.75 for self-antigens (Figure 7.7B). We surmised that T-cell recognition motifs captured by 1D CNN model have been more predictive of immunogenicity than sequential relationship captured by RNN models.

In addition to peptide sequences, we examined whether incorporating other hallmarks of immunogenicity, such as predicted MHC binding rank[53,60] and hydrophobicity[209] can further improve classifier performance. Particularly, given that the current model input sequences at contact positions only, adding MHC binding rank would allow the model to accommodate peptide-MHC binding information. This strategy has several advantages. First, it feeds the binding information without biasing HLAs, avoiding the need to balance training data by HLAs. Second, using

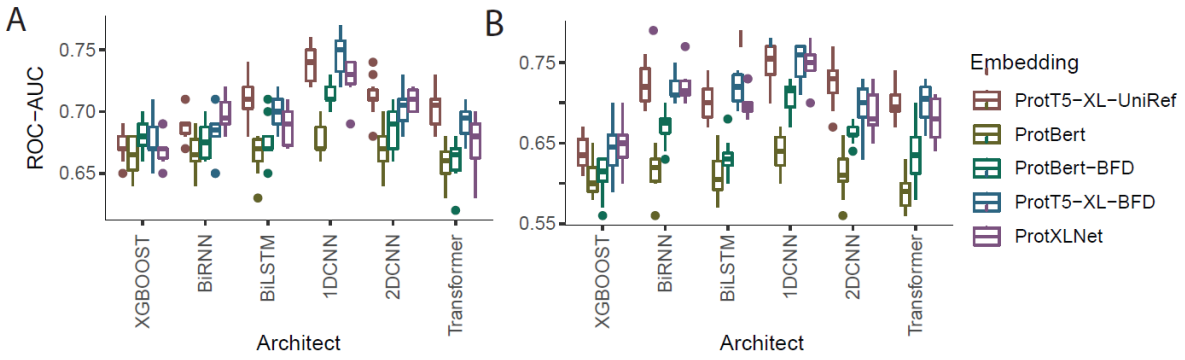


Figure 7.7: Compare performance of different deep learning architectures. A-B. ROC-AUC values comparing performance of different machine learning and deep learning architectures using embeddings from 5 protein transformer-based PLMs for pathogenic (A) and self-peptides (B). XGBOOST: extreme gradient boosting. BiRNN: bidirectional recurrent neural network. BiLSTM: bidirectional Long short-term memory. 1DCNN: 1-dimensiona convolutional neural network. 2DCNN: 2-dimensional CNN. Each point represents ROC-AUC value from one round of 10-fold cross-validations.

NetMHCpan predictions provides the most accurate prediction without requiring the model to be re-trained using massive peptide-HLA binding data. By integrating MHC binding rank and hydrophobicity as fully connected layers (i.e. Multilayer perceptrons, MLPs) and optimizing hyperparameters (Figure 7.8A, Methods), performance was improved to an average ROC-AUC of 0.76 for pathogenic (Figure 7.8B) and 0.90 for self-antigens (Figure 7.8C). Of note, large improvement in self-antigen model was due to greater difference in MHC binding rank between self-epitopes vs. non-epitopes, as a consequence of relaxed MHC binding rank threshold ($\leq$ 10 MHC binding rank) to account for as many self-epitopes.

Therefore, representing peptide sequences at contact positions by ProtT5-XL-UniRef amino acid embeddings, extracting T recognition motifs by 1D CNN kernels, and adding MHC binding rank and hydrophobicity as MLPs could effectively achieve superior performance in classifying immunogenicity.

Sequence patterns discriminating immunogenicity

Among different architectures, 1D CNN model that captures local motifs achieved the best performance. Moreover, extensive screening of pMHC library against a single TCR revealed dominant hotspots or motifs in the cognate peptides[246,247]. We therefore set out to expand this observation and explore enrichment of n-grams (i.e. contiguous sequence of n-amino acids)[248] and position-specific k-mer motifs (i.e. contiguous or non-contiguous sequence of k amino acids at specific positions) in contact residues of pathogenic and self-peptides.

The ratio (in normalized frequency) of n-gram or positions-specific k-mer motifs between epitopes and non-epitopes was computed (Methods). The top-ranking n-grams were GIG, GINA, GIF,

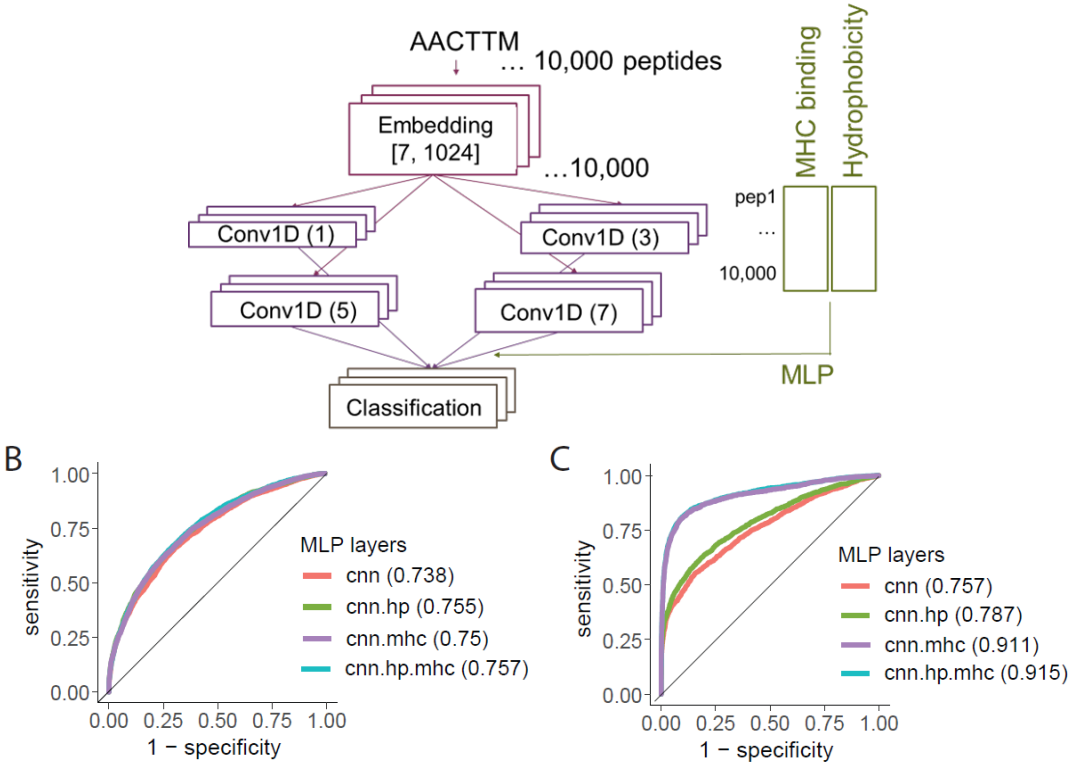


Figure 7.8: Adding hallmarks of immunogenicity. A. Schematic diagram of model incorporating MHC binding potential and hydrophobicity. B-C. ROC curve comparing performance of models after adding MHC binding rank and/or hydrophobicity to peptide sequence-based 1D CNN model for pathogenic (B) and self-peptides (C).

LGIN, VEG and SGSP for pathogenic epitopes and SC, GIGI, IC, QC and CA for self-epitopes (Figure 7.9A). In addition, top position-specific k-mer motifs were .EG.L., .E.IL. and .GIG... for pathogenic epitopes (Figure 7.9B) and GIG.., ..M.P., .G.GI.., and AGI.... for self-epitopes (Figure 7.9C). Notably, GIG, which was previously associated with DMF5 TCR, was one of the top n-grams and position-specific k-mer motifs.

To investigate prevalence of sequence patterns across pathogenic and self-epitopes, we compared n-grams and position-specific k-mer motifs enriched in individual analysis. The 48 n-grams and total 298 position-specific k-mer motifs from 9 and 10aa peptides (Figure 7.10A) were shared between pathogenic and self-epitopes. These include GIG, GINA, GMP, ALGI and APTG n-grams (Figure 7.10B), and ....VP, W..P.., .G.GI.., and .GIG... position-specific k-mer motifs (Figure 7.10C).

Furthermore, pairwise global-alignment scores across pathogenic peptides were computed to identify clusters of epitopes sharing high sequence homology (Method). Plotting highly similar peptides revealed clusters dominated by Positive peptides (Figure 7.11A). These clusters contained peptides having patterns, such as YNTI... FG(Y/F)PV(F/Y). F(E/D)(K/R)S.., G.GW.. .DR.WC, GPG.R.P, P.DFFP., .GIGI.., .DRGM., VALG.NA. and .LGLNA. (Figure 7.11B).

Our analyses therefore suggest that TCR contact positions of immunogenic peptides exhibit preferences towards presenting certain motifs while disfavouring others. We also observed similarities and differences between pathogenic and self-peptides. We believe these shortlisted sequence patterns can be target for functional validations in identifying immune hotspots.

Abstain low-confidence predictions to improve robustness

While the DNN models could achieve higher accuracies than existing models, they are nonetheless far from perfect predictions. Partially due to the limited number of peptides characterized to fill the full combinatorial peptide space, the performance of sequence-based model drops substantially in predicting peptides nonhomologous to those of training data, called the out-of-distribution (OOD) generalization problem[249]. In contexts where incorrect predictions can have severe consequences e.g. in healthcare or security, using model uncertainty to decide when to trust or abstain the prediction can facilitate rejecting false predictions and improving model robustness.

There are generally two types of OODs: background shift (i.e. shift in population-level features that do not depend on classification labels) and semantic shift (i.e. shift in features that are correlated with the context and label)[250]. Density or proximity-based models are typically better at detecting background shift, while calibration methods (i.e. using model's prediction confidence) perform better when detecting semantic shifts. In our study, the greatest OOD came from peptides originating from different species that had moderately different features associated with immunogenicity (i.e. semantic features). This aligns with previous findings that the majority of OOD stems from semantic shift and thus calibration methods generally outperform proximity-based methods (i.e. autoencoder reconstruction)[250].

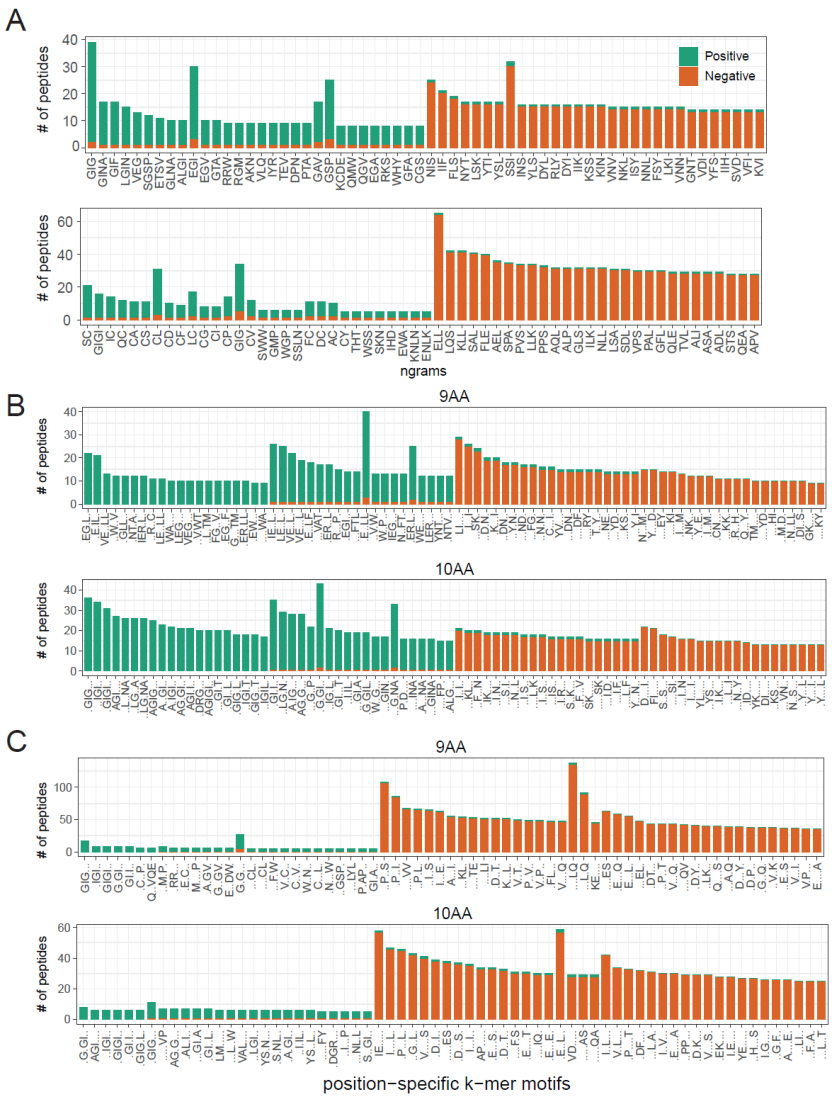


Figure 7.9: Sequence motifs enriched or depleted in pathogenic and self-epitopes. A. Top n-grams enriched or depleted in epitopes for pathogenic (top) or self-antigens (bottom). B-C. Top position-specific k-mer motifs enriched or depleted in 9aa (B) or 10aa (C) epitopes for pathogenic (top) or self-antigens (bottom).

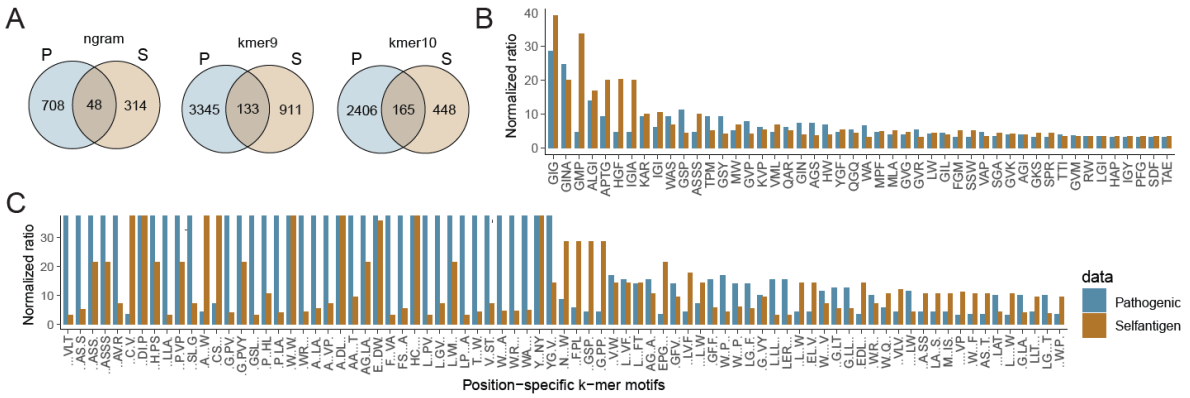


Figure 7.10: Compare enriched motifs across pathogenic and self-peptides. A. Venn diagram comparing enriched n-grams (left) or position-specific k-mer motifs from 9aa (middle) or 10aa peptides (right) between pathogenic and self-peptides. B. Normalized enrichment ratio of shared n-grams enriched in both pathogenic and self-epitopes. C. Normalized ratio of shared position-specific k-mer motifs found from 9aa peptides enriched in both pathogenic and self-epitopes.

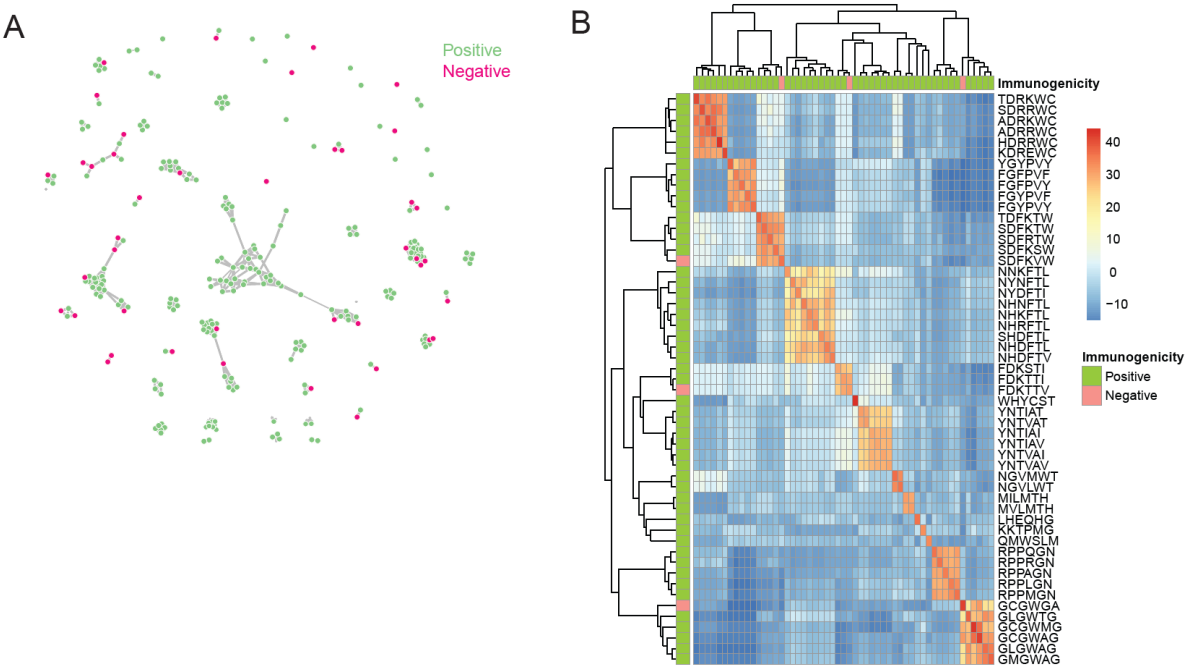


Figure 7.11: Clusters of peptides sharing high sequence similarity by pairwise global alignment score at contact positions. A. Network graph illustrating pairwise global alignment score between 9aa pathogenic peptides. Shown are peptides having ≥22 alignment score with ≥3 other peptides. B Heatmap showing pathogenic peptides having global alignment score ≥ 27 with ≥ 3 other peptides for 9aa peptides.

Here, several OOD detection methods were tested to avoid low-confidence predictions. Different autoencoder architectures, which are often used for anomaly detection, were first investigated[251]. We assessed autoencoder, variational autoencoder and denoising autoencoder architectures, and compared reconstruction loss between correctly vs. incorrected predicted test peptides (Figure 7.12, Methods). However, no significant difference in reconstruction loss between correctly and incorrectly predicted peptides was observed, implying that the autoencoder-based methods cannot effectively identify OOD inputs. We speculated that unsupervised anomaly detection methods, such as reconstruction loss from autoencoders, were ineffective because 'species' is not a primary driver of peptide clustering at contact positions, which is often driven by the TCRs that they are bound by (Figure 7.2E). Likewise, density-based or distance-based methods that use sequence similarity would not be able to detect OODs.

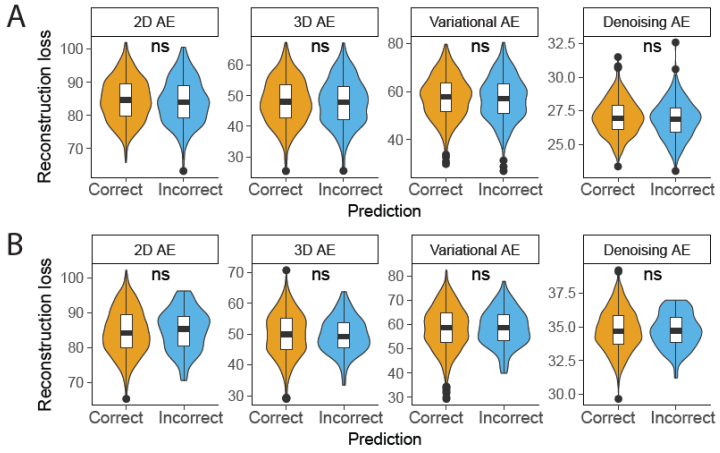


Figure 7.12: Detect out-of-distribution using autoencoders. A. Anomaly detection using autoencoders, namely 2-dimension autoencoder (2D AE), 2-dimension autoencoder (3D AE), variational autoencoder and denoising autoencoders on pathogenic (A) and self-peptides (B). Statistical significance by p-values from Student's t-test. ns: non-significant.

Therefore, calibration methods such as maximum softmax probability (MaxProb), temperature scaling, average MaxProb from ensemble of 10 models (called Ensemble) and Monte Carlo dropout were tested, because they were reported to effectively detect semantic shifts (Methods). We observed that all four methods could significantly differentiate correct vs. incorrect predictions for both pathogenic and self-antigens (Figure 7.13A-D). By evaluating predictive power of these metrics on decision tree classifier, MaxProb and Ensemble had the highest feature importance (Figure 7.13E), and could achieve comparable performance to the model incorporating all four metrics with average ROC-AUC of 0.923 for pathogenic and 0.806 for self-antigens (Figure 7.13F). With these, a MaxProb and Ensemble-based OOD detection module was introduced downstream of 1D CNN model prediction to only report peptides that are likely to have a correction prediction.

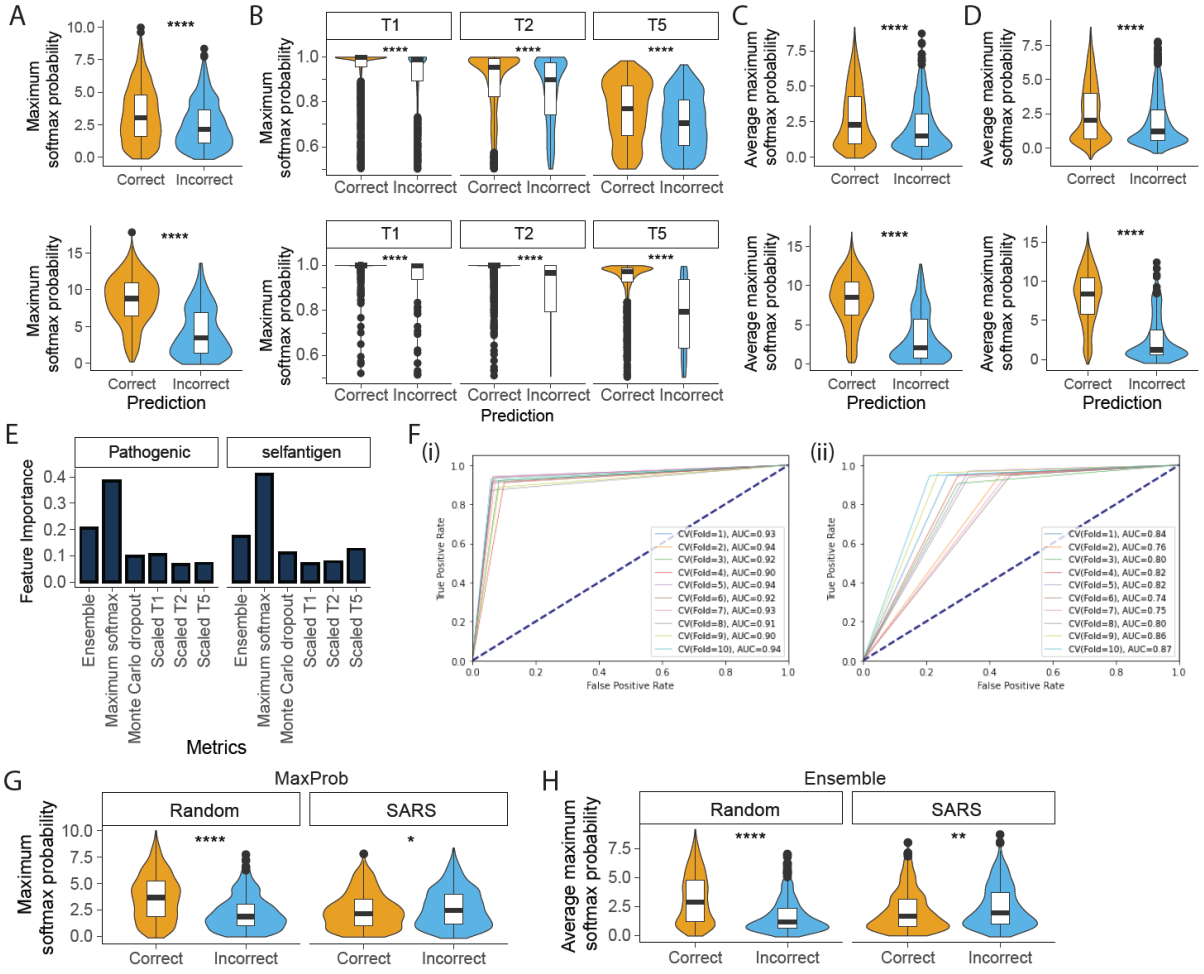


Figure 7.13: OOD detection using calibration methods. A-D. Distribution of different calibration-based metrics on pathogenic (top) and self-peptides (bottom) discriminating correctly vs. incorrectly predicted peptides, using maximum softmax probability (A), temperature scaling with varying temperatures that scale logit values (B), average maximum softmax probability from ensemble of 10 models (C) and average maximum softmax probability from Monte Carlo dropouts having 100 reiterations with stochastic dropouts of 0.6 (D). E. Contribution of calibration metrics on decision tree classifier. F. ROC curve illustrating performance of DecisionTree OOD classifiers using maximum softmax probability and ensemble on pathogenic (i) and self-peptides (ii). G-H. Distribution of maximum softmax probability (G) and average maximum softmax probability from ensemble of 10 models (H) resulted from randomly split data (resemblance of 10-fold cross-validation) and cross-species data (i.e. model trained using non-SARS-CoV-2 peptides to predict SARS-CoV-2 peptides).

Relative similarity to autoantigens or tumor-associated antigens (RSAT) as a novel feature of pathogenic peptide immunogenicity

We and others[225,226] have shown that some of highly predictive metrics found from cancer neoepitope studies, such as 'dissimilarity to self', may not be applicable to pathogenic peptides. To address this, we present an alternative solution to this problem by a novel metric, a relative similarity to autoantigens or tumor-associated antigens (RSAT) to estimate immunogenicity potential on pathogenic peptides. For pathogenic peptides that were abstained due to low confidence prediction, RSAT will provide an additional estimate of immunogenicity potential.

The 'dissimilarity to self' stems from the paradigm of negative selection where T-cells are negatively selected against strongly binding self-peptides and thus should lack T-cell repertoire binding self-proteome[210,211]. However, there is another side of the story where T-cells are positively selected by binding to low or moderate binding self-peptides. In fact, recent studies attributed the inability of immune system to recognize a large number of pathogenic peptides, most of which are highly dissimilar to human proteome, to the mechanism of positive selection where only T-cells bound by low or moderately binding self-peptides survive to trigger immune response[225]. Therefore, we hypothesized that pathogenic peptides homologous to immunogenic self-peptides, such as autoantigens or tumor-associated antigens, may be more likely to trigger immune response, and assessed whether relative similarity to autoantigens or tumor-associated antigens compared to reference human proteome can be a predictor of immunogenicity for pathogenic peptides.

To test our hypothesis, 5023 unique Autoimmunity, Allergy or Tumor-associated antigens (AATs) were retrieved from NEPdb[252], dbPepNeo[253], McPAS-TCR[40], IEDB[38] and cancer antigenic peptide database (https://caped.icp.ucl.ac.be) (Figure 7.14A). We then computed similarity between pathogenic peptides and AATs, and only retained pathogenic peptides having comparable AAT counterparts (Methods). On these pathogenic peptides, we then computed the similarity between pathogenic peptide and closest human proteome counterpart to AAT, and ratio between two i.e. relative similarity of pathogenic peptide to autoantigens or tumor-associated antigens (named RSAT).

From pathogenic data, 263/3082 pathogenic epitopes and 357/6756 non-epitopes had comparable counterparts and AATs by threshold match score of 0.6 (Method, Figure 7.14B). The RSAT was computed on these peptides and found a significant difference between epitopes vs. non-epitopes across peptide lengths 9 and 10aa (Figure 7.15A), and across different pathogenic species (Figure 7.15B), indicating that RSAT could effectively classify peptides of varying lengths and pathogen species. Despite the fact that each species has a limited number of peptides, positives had higher RSAT values with low (SARS-CoV-2 and M.tuberculosis), medium (EBV) and high (HCV and vaccinia virus) effect sizes.

Here, we developed a novel metric, RSAT, to estimate immune potential of pathogenic peptides based on relative similarity to known auto- or tumor-associated epitopes. We demonstrated that RSAT can effectively discriminate epitopes of different lengths and pathogens. We appreciate that the low number of Autoimmunity, Allergy or Tumor-associated antigens (AATs) limits

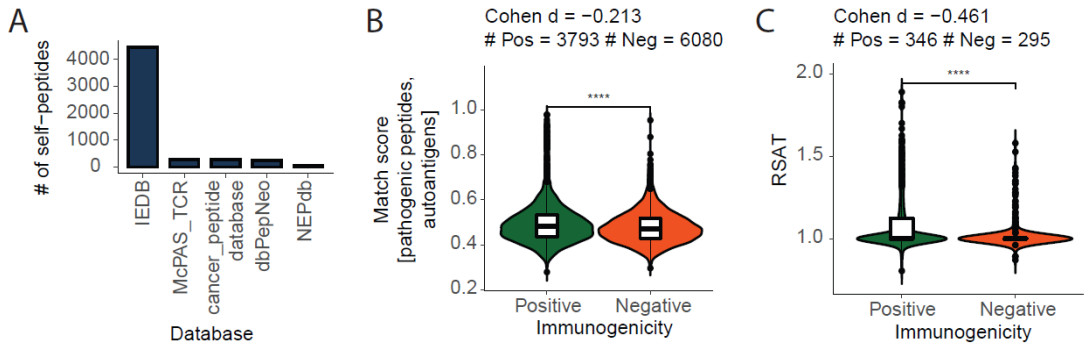


Figure 7.14: Relative similarity to autoantigens and tumor-associated antigens (RSAT). A. Bar plot illustrating the number of self-peptides (i.e. cancer neoepitopes, autoantigens, tumor-associated antigens) retrieved from different databases. B. Distribution of match score between pathogenic peptides and best self-peptide counterparts. Contains the Cohen's d value showing the effect sizes differentiating positives and negatives as small (d = 0.2), medium (d = 0.5) and large (0.8), as well as the number of positive and negative peptides in each species. C. Distribution of RSAT computed for pathogenic peptides having comparable self-antigen counterparts (by match score $\geq$ 0.6).

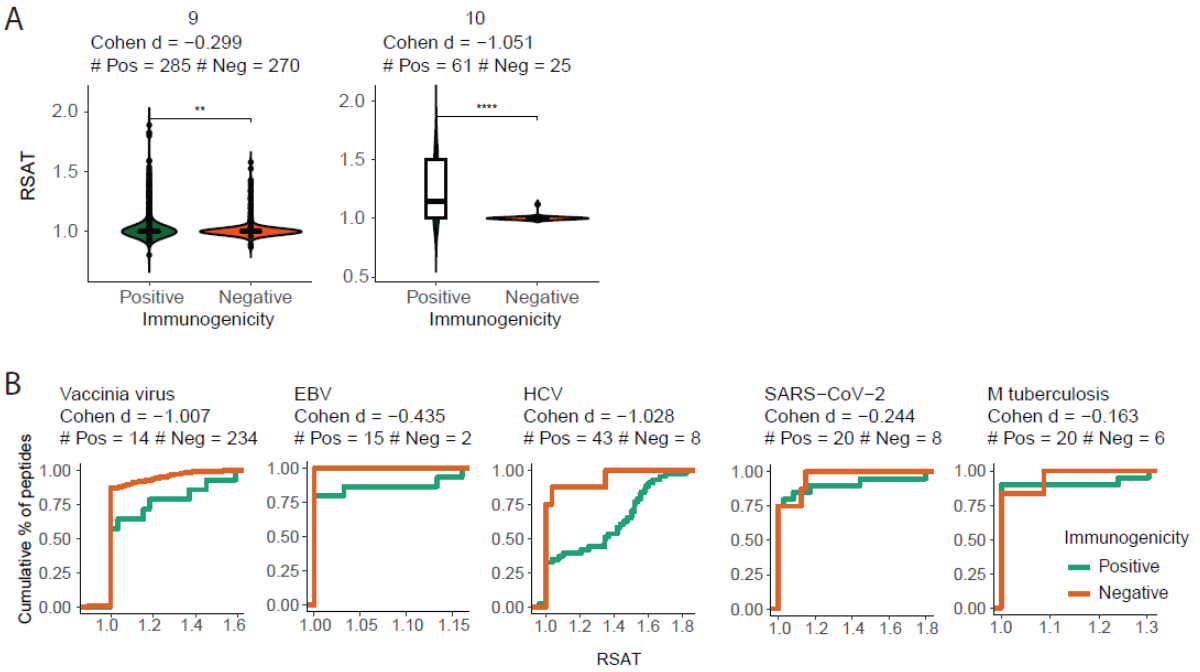


Figure 7.15: RSAT can effectively discriminate pathogenic epitopes from non-epitopes of varying lengths and species. A. Distribution of RSAT for pathogenic peptides having 9aa and 10aa in length. B. Distribution RSAT by species of origin.

the use of RSAT on broader range peptides. However, we envision that as data becomes more abundant, RSTA will become more popular.

Application of TRAP to study immune escape of SARS-CoV-2 variants

Despite widespread vaccination, the emergence of numerous Variants of Concern (VOC), including Omicron BA1 and its derivative subvariants BA2, 4, 5[254,255], poses a potential threat. Although Omicron has milder symptoms than previous strains[256], it has higher infectivity or transmissibility rate, resulting in unprecedented healthcare turbulence and social disruptions, in addition to the massive economic burden.

Effective surveillance and monitoring systems with help of predictive models can rapidly detect concerning variants that may have a fitness advantage of the ability to evade immune response[257]. While the escape potential of humoral immunity has been studied, T-cell immunity has received little attention. T-cell responses are an important component of the so-called "wall of immunity" against COVID-19. Moreover, when a VOC emerges, there is considerable uncertainty about how novel mutations will impact T-cell immunogenicity. In such cases, the ability to quickly estimate the effects of newly emerged mutations on T-cell immunogenicity would aid in reducing uncertainty for SARS-CoV-2 and other diseases.

Indeed, a comprehensive work by Dolton *et al.* focused on HLA-A*02 epitope YPG, which was found to evade >120 TCRs as a result of the P→L mutation, demonstrating T-cell escape in the wild[258]. Furthermore, a non-Hodgkin's lymphoma patient infected with SARS-CoV-2, Stavenich *et al.*, discovered two mutations that were experimentally confirmed to allow the virus to evade the patient's CD8+ T-cell response[259].

Therefore, with increasing evidence indicating the impact of VOC, we combined TRAP with *in silico* mutagenesis to not only dissect the impact of *existing* mutations on T-cell responses, but also to predict *theoretical* mutations that may be most detrimental to T-cell immunity.

Accuracy of TRAP trained using SARS-CoV-2 data

First, a coronavirus TRAP model was built using functionally characterised SARS-CoV-1 and SARS-CoV-2 epitopes and non-epitopes (Methods). To validate its accuracy, the coronavirus TRAP model was trained using all SARS-CoV peptides except the Wuhan Hu-1 test peptides, yielding ROC-AUC of 0.77 by 10-fold cross-validation (Figure 7.16). The performance was higher than other T-cell recognition models trained using the same dataset.

In silico mutagenesis of SARS-CoV-2

Using the coronavirus TRAP model, we investigated how single point mutations in 9 or 10 amino acid Wuhan Hu-1 epitopes affect immunogenicity. We employed *in silico* mutagenesis and subjected the mutants to NetMHCpan and TRAP predictions to generate immunogenicity scores (Methods). For each mutant peptide MT_k, the immunogenicity scores were averaged per MHC to assess the effect of the mutation in pan-HLA manner, and the log ratios between MT_k and its corresponding WT_i at TCR contact positions were examined (Methods).

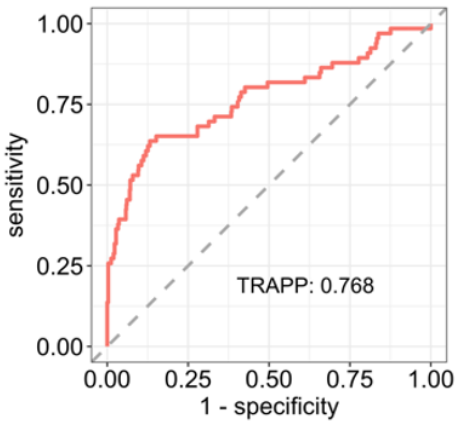


Figure 7.16: Performance of TRAP model trained using SARS-CoV data. ROC-AUC curve illustrating the performance of TRAP model trained using SARS-CoV peptides functionally characterised by T-cell assays.

First, we investigated the impact of each amino acids irrespective of their positions. We found that displacement of basic lysine (K) and polar cysteine (C) and glycine (G) increased the immunogenicity potential (log ratio > 0) on 9-mer epitopes (Figure 7.17A). Similarly, replacement with these amino acids reduced the immunogenicity. On the other hand, displacement of hydrophobic amino acids, such as tryptophan (W), phenylalanine (F) and alanine (A), in TCR contact positions reduced immunogenicity. Similarly on 10-mer epitopes, displacement of arginine (R) , glycine (G) and lysine (K) increased immunogenicity, whereas displacement of hydrophobic amino acids were predominantly detrimental (Figure 7.17B).

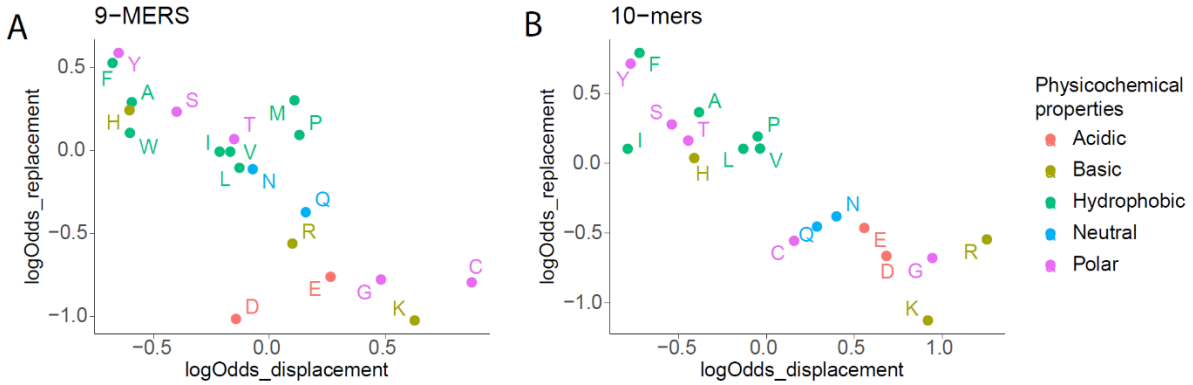


Figure 7.17: Impact of amino acid substitutions on SARS-CoV-2 immunogenicity. A-B. Dot plot illustrating log odds ratios of displacing or replacing with the designated amino acid on 9-mer (A) or 10-mer (B) epitopes.

Next, we examined the impact of positional mutations (i.e. A_P_B: A = original amino acid, P = position of the mutation, B = new amino acid) at TCR contact positions of 9-mer peptides (Figure 7.17A). We again observed the displacement of hydrophobic amino acids reducing immunogenicity. Furthermore, it was commonly observed that displacement of F or Y in P3 was detrimental to

immunogenicity. The pattern was more pronounced on 10-mer epitopes, where hydrophobic and polar residues dominated the most 'positive' replacement for immunogenicity, while acidic, basic and neutral residues populated the most detrimental replacements (Figure 7.17B-C). This finding not only sheds light on the molecular biology underpinning immunogenicity, but also indicates that the physicochemical properties of the replacing amino acid may be critical for estimating the mutation's impact on T-cell immunogenicity.

To estimate the overall impact of mutations on each WT Wuhan Hu-1 epitopes, we analysed average log ratio scores per epitope and found that substitutions negatively impact peptide immunogenicity to varying degrees (Figure 7.18D). For example, LLW* and APG* were predicted to be escape-prone (i.e. mutations leading reduced T-cell response), whereas NGV*, VAK*, EEL* and SPR* were predicted to exhibit no substantial difference.

Neighbour-network analysis to visualise immune escape

To visualised the impact of mutations on each WT Wuhan Hu-1 epitope, the 'neighbour network' approach was adopted from Ogishi *et al.*[50]. Briefly, Ogishi *et al.* showed the relationship between wild-type epitope and its single amino acid mutants in a network representation (Methods). We used their approach to build neighbour networks for epitopes of interest (Figure 7.19).

Here, the epitope APG* again showed considerable routes for T-cell escape (Figure 7.19A), with mutations in P2 (cluster 3), P7 (cluster 8) and P9 (cluster 2) leading to reduced immunogenicity. In particular, substitution in P7 (cluster 8) of APG* results in K417 mutation in Omicron. Interestingly, while this cluster is dominated by gain-of-immunogenicity mutants, the Omicron mutation is one of the only two P7 mutants predicted to lose immunogenicity. This analysis indicate that APG* is not robustly immunogenic and that the mutation in Omicron is one of the deleterious substitutions that could be made in the position.

SPR*, on the other hand, would remain immunogenic after mutations at all positions, except P2 (Figure 7.19B). Particularly, the P→H mutation in P2 led to Omicron mutation, which compromised binding to HLA-B*07 and thus loss of immunogenicity. This analysis shows that Omicron used one of the few SPR* escape routes to gain immunogenicity in the real world. This finding is especially intriguing because this SPR* epitope contains the putative superantigen core 'PRARR' which has been linked to post-COVID-19 inflammatory disorders. Thus, our analysis revealed that BA1-Omicron mutation SPR* in P2 may result in a narrower MHC binding repertoire, but that mutations in other positions would increase immunogenicity.

Taken together, this study highlights the application of *in silico* mutagenesis, TRAP and neighbour-network analysis to investigate the effect of theoretical single point mutations in T-cell response. In fact, we observed real-world scenarios that resulted in Omicron mutations. This analysis shows that in addition to identifying therapeutic targets, TRAP is a powerful tool for rapidly understanding theoretical landscape of immune potential of pathogenic or self-antigens and identify emerging variants of concern.

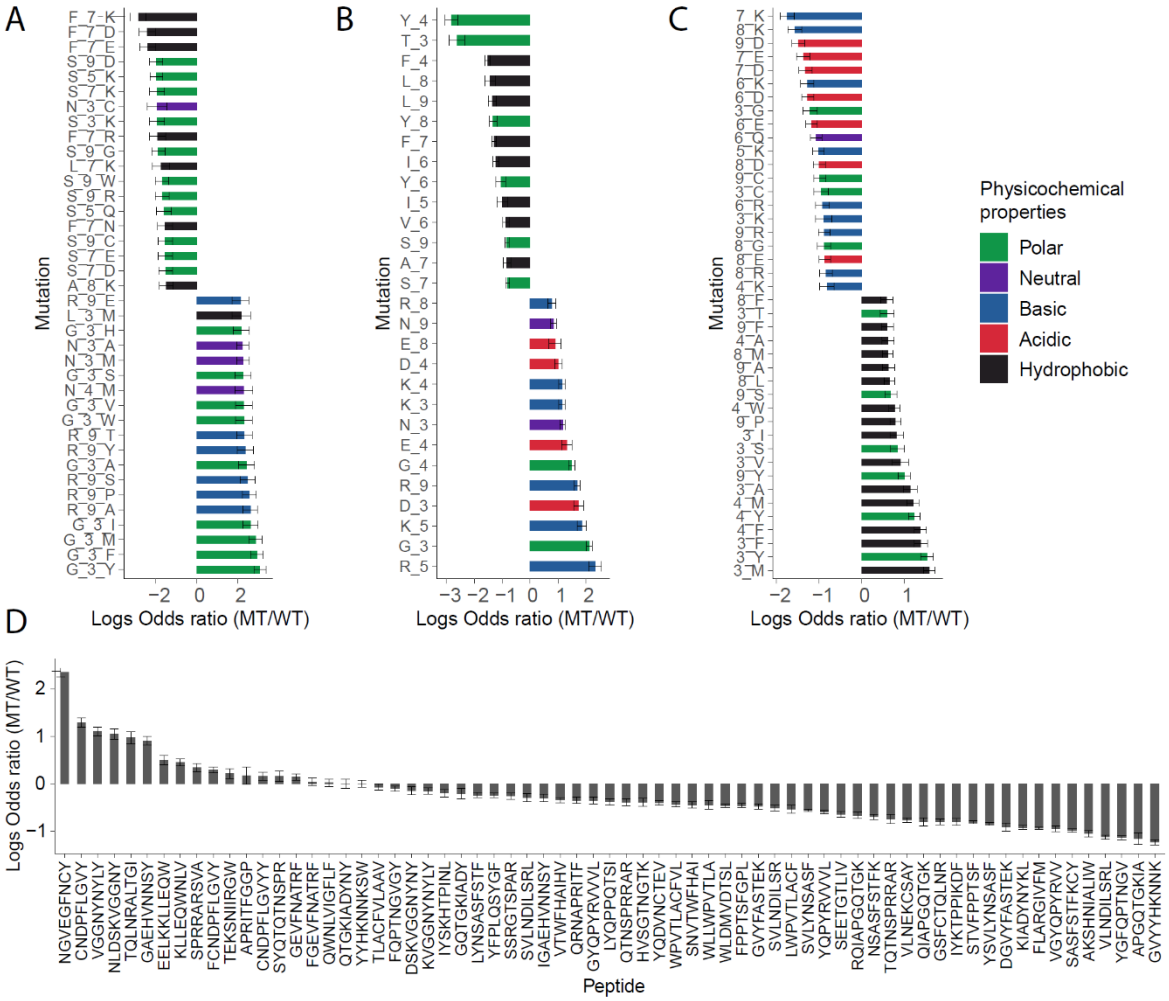


Figure 7.18: Impact of position-specific substitutions on epitope immunogenicity. A. Log odds ratio showing the impact of positional mutations (i.e. A_P_B: A = original amino acid, P = position of the mutation, B = new amino acid) on peptide immunogenicity for 9-mer. Showing positional mutations with 5% highest and lowest of logs ratios. B. Log ratios for replacing amino acids into designated position (i.e. A_P: A = replacing amino acid, P = position of the mutation). C. Log ratios for displacing amino acid from designated position (i.e. P_B; P = position of the mutation, B = displacing amino acid). D. Average log odds ratio scores for each epitope. The log ratio scores from mutations at different positions were aggregated and averaged to compute the averaged log ratio score per epitope.

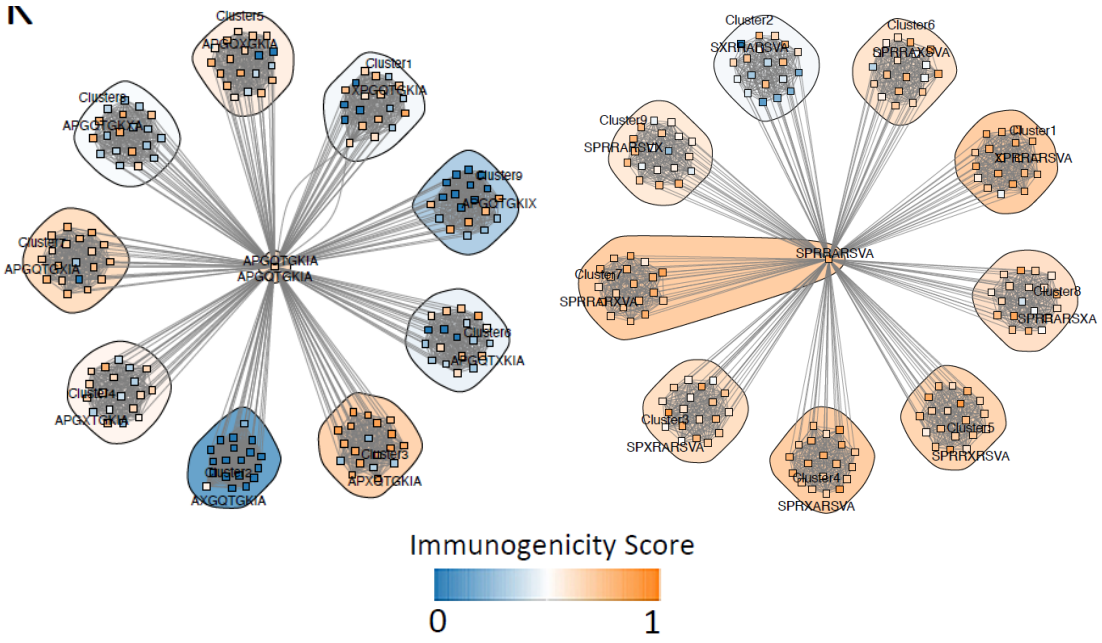


Figure 7.19: Neighbour network illustrating deleterious and beneficial mutations from Wuhan Hu-1 epitopes. Neighbour network illustrating the impact of single point mutations on immunogenicity for APG*(left) and SPR*(right) epitopes. Each cluster shows point mutations in each positions. Peptides having gain-of-immunogenicity coloured in orange, and peptides with loss-of-immunogenicity coloured in blue.

Discussion

This study presents TRAP, a deep learning platform that allows context-specific identification of CD8+ T-cell pathogenic and self-epitopes. The workflow aims to address sources underlying poor performance of existing models, such as HLA bias, limited training data and out-of-distribution generalisation problem. In this workflow, we adopted HLA-generalised approach, utilised amino acid embeddings from transformer-based PLMs, employed 1D CNN model to extract T-cell recognition motifs, added layers to account for MHC binding, devised OOD detection module and developed a novel metric for estimating immunogenicity potential of pathogenic peptides. We further demonstrated an application of TRAP by using it to assess the impact of random mutations in giving rise to variants of concern.

For self-antigen model, the autoantigens, tumor-associated antigens and cancer neoepitopes were collected as epitopes, and benign HLA-I ligands expressed in thymus as non-epitopes (Figure 7.4B-C). We assumed that HLA-I ligands (exclusively) expressed in the thymus are involved in negative selection of T-cells, and thus there is no repertoire recognising these peptides. Notably, our novel approach of incorporating the HLA-I ligands allowed for a clearer separation of self-epitopes versus non-epitopes. When the data is limited, biological knowledge, such as negative thymic selection, can be used to bridge the data gap, and aid in the development of a more accurate classifier.

In benchmarking TRAP performance, we found that while Repitope performed well on pathogenic data, it performed poorly on emerging pathogens. Our previous analysis (Chapter 6) revealed that Repitope was one of the models with skewed prediction for prevalent HLAs, with HLA-A*02:01-bound peptides having a higher immunogenicity score than non-HLA-A*02:01 bound peptides[56]. This suggested that the model considered A02:01 binding as a superior feature to T-cell recognition potential, skewing the prediction. While TRAP performed slightly lower on the pathogenic dataset, the HLA-generalised approach of TRAP eliminates the possibility of HLA bias. Therefore, end users should choose which model to pursue based on the peptides and HLA alleles of interest.

While these approaches improved accuracy and robustness of the prediction, there still remains challenges. Foremost, while calibration methods could effectively detect incorrect prediction in 10-fold cross-validation, they failed on species-specific dataset e.g. model trained on non-SARS-CoV-2 peptides to predict immunogenicity of SARS-2 peptides (Figure 7.13G-H). We surmised this is because of "spurious semantic features" (i.e. features having discriminative power in training data but no longer predictive in test data) that were driving overly confident incorrect predictions. In a recent case study, Arora *et al.* benchmarked calibration methods on real-world challenge data and reported that models were over-confident on the OOD examples because while spurious semantic features were still prominent, they no longer were predictive on test peptides. Therefore, calibration methods, which use model confidence to detect OOD, tend to struggle on challenge data, producing accuracies close to random. With spurious semantic shift being one of the remaining challenges in the NLP, further advances in deep learning and NLP will also facilitate improving robustness of predictions on complex biological data.

Second, following our workflow, peptides may be abstained due to low-confidence predictions. For cancer peptides, we advise using TESLA[52] that incorporates context-specific generic features for neoepitope prediction. The pathogenic peptides having low-confidence predictions will be directed to RSAT in TRAP workflow. However, because RSAT is only applicable to peptides with self-peptide counterparts, its coverage remains limited. Further characterisation of self-epitopes will expand RSAT coverage to embrace a broader range of pathogenic peptides.

Third, TRAP takes an HLA-generalised approach to avoid HLA bias, as the current dataset contains little intra-HLA variation in peptide immunogenicity. However, characterizing peptide immunogenicity on broader range of HLAs may provide insight into the effect of HLA on overall immunogenicity. Moreover, because the specific interaction between peptide and T-cell receptor is MHC-restricted[260,261], MHC-focused studies would be required to model specific TCR:pMHC recognition.

Forth, T-cell recognition depends not only on peptide-MHC complex, but also on multiple other factors, such as cytokine microenvironment, co-stimulatory molecules *in vivo* and availability of TCR repertoires that are often highly stochastic and individualized[262], adding complication to T-cell activation and function[263]. While there are almost no record of these determinants[264], studies highlighting the influence of microenvironment on T-cell response will be pivotal to recapitulating T-cell function for therapeutic applications.

In addition, we incorporated peptide sequence at the contact position to gauge on T-cell recognition motifs. However, some studies have shown that position 1 (P1) of 9-mer peptide plays an important role in T-cell binding, as demonstrated by P1, 3, 4, 5, 9 being critical for MAGE A3 binding by a3a-engineered T-cells[265]. Although it is difficult to deconvolute the roles of different positions in MHC and T-cell binding, more studies may shed light on the role of P1.

Fifth, due to the lack of true Negative self-antigens, we retrieved benign HLA-I ligands exclusively expressed in thymus for their relevance to sharing T-cell repertoire and central tolerance. However, not all peptides expressed in thymus will be involved in negative selection of T-cells with the notion that positive selection-inducing self-peptides are displayed by cortical thymic epithelial cells (cTEC)[266]. While there have been studies describing Tregs reacting to self-peptides present in the thymus[267,268], given that the models are targeted for CD8+ T-cells, we rationed that majority of benign HLA-I ligands will be non-immunogenic. However, future studies characterizing HLA-I ligand expressed in medulla thymic epithelial cells (mTEC) will improve confidence as 'non-immunogenic self-antigens'.

Lastly, current breadth of peptides characterized by T-cell assays are far from filling the full combinatorial peptide space, especially for CD4+ T-cell targets. Also, while this model is limited to peptide specific binding of CD8+ T-cells, other T-cells are specific for lipid or small molecules like metabolites. As such, high-throughput screening of immune targets and antigen-specific TCRs by the help of recent technological advancement will greatly foster the process of model development.

Despite current data and model constraints, the TRAP outperforms existing models in predicting

CD8+ T-cell epitopes and will provide a more robust, accurate, and biologically meaningful list of peptides for functional validations, thereby accelerating therapeutic development.

Methods

Data preparation

PeptideTcell data

Peptides that bind MHC-I molecules are typically restricted to 8-10 amino acids (aa) in lengths due to closed structure of peptide-binding groove. Given limited number of 8aa peptides in databases, peptides of lengths 9-10 amino acids (aa) characterized by T-cells were retrieved from IEDB. These included cancer neoepitopes, autoantigens or pathogenic peptides characterized with or without nutritional supplements. The peptides without HLA allele or serotype annotations were removed. To ensure MHC binding, peptides were subjected to NetMHCpan 4.0 prediction and retained predicted MHC binder having rank ≤ 2.0. Peptides having contradictory immunogenicity annotations were categorized as 'Positive' and only retained Negatives $\geq$3 tests. This resulted 5,093 immunogenic ('Positive') and 6,628 non-immunogenic ('Negative') peptides in PeptideTcell data. The dataset include information about peptide sequence, binary immunogenicity, HLA allele, source antigen and organism. This data is the foundation for sequence pattern analysis, model development, RSAT validation (for only pathogenic peptides).

Pathogenic data

For analysis of enriched sequence patterns and development of deep learning-based models, datasets for pathogenic and self-antigens were prepared separately. For pathogenic datasets, we subset peptides originating from non-human species, resulting in 4,000 Positive and 6,097 Negative pathogenic peptides.

Self-antigen data

For immunogenic self, autoantigens, tumor-associated antigens and cancer neoepitopes having 9-10 amino acids were collected from different databases. 162 epitopes were gathered from Cancer peptide database, 228 from dbPepNeo, 1506 from IEDB, 256 from McPAS-TCR, and 256 from NEPdb. For Negative self-antigens, we gathered HLA-I ligands expressed in thymus suggested to be involved negative selection of T-cells. 240 HLA-I ligands were collected from Adamopoulou *et al.* expressed in negatively selecting dendritic cells, 187 HLA-I ligands from Espinosa *et al.* expressed in thymus and 10,840 benign HLA-I presented peptides from HLA Ligand Atlas, exclusively expressed in thymus. Of note, we did not include 'Negatives' from IEDB for self-antigen data as many were tested due to their association with tumor-associated antigens e.g. cancer/testis antigen 1, melanoma-associated antigen 9. Yet, there was limited evidence in non-immunogenicity of these peptides. Therefore, only MHC-I peptides expressed in thymus were included. After pre-processing, removing duplicates and filtering for peptides with NetMHCpan rank ≤ 10, the self-antigen dataset was composed of 1,606 Positive and 10,915 Negative peptides.

Benchmarking data

The benchmarking analysis has been done in HLA-agnostic (i.e. peptide-based) or restricted (i.e. peptide-HLA based) manner depending on the inputs for different models. The TRAP makes

prediction on peptide sequence taking HLA binding rank into account, which is the combination of both agnostic and restricted approaches. The HLA agnostic approach was done on all peptides in aforementioned pathogenic and self-antigen datasets. The HLA-restricted prediction was done against epitopes that were bound to 13 HLAs for which NetTepi could be performed, which are HLA-A*02:01, HLA-B*58:01, HLA-B*15:01, HLA-B*35:01, HLA-B*07:02, HLA-A*01:01, HLA-A*03:01, HLA-A*11:01, HLA-A*24:02, HLA-A*26:01, HLA-B*27:05, HLA-B*39:01, HLA-B*40:01. The list of models conducted in HLA-restricted manner are iPred, PRIME, NetTepi, IEDB and DeepImmuno. Specifically for DeepImmuno, additional filter was applied to exclude peptides that were bound to HLAs that DeepImmuno could not process.

DeepImmuno data

DeepImmuno training data retrieved from GitHub repository, including peptide and HLA information. https://github.com/frankligy/DeepImmuno

10-fold CV and cross-species comparison

The peptideTcell data was divided randomly or in cross-species manner: i) 90% train vs. 10% test random split (i.e. representation of 10-fold cross-validation), ii) Non-SARS-CoV-2 (SARS-2) train vs. SARS-2 test, and iii) Non-vaccinia virus (VACV) train vs. VACV test.

To compare XGBOOST classifiers trained and tested on 10-fold vs. cross-species manner, each amino acids in peptide sequence at contact positions were first represented by their physicochemical properties using "aaDescriptors" function in R Peptides v2.4.4 package. The amino acid descriptors include kideraFactor, zScales, tScales, vhseScales, protFP, stScales, blosumIndices, mswhimScores, crucianiProperties and fasgaiVectors, which describe properties such as polarity, electronic properties, hydrophobicity, α-helix/bend preference, β-sheet, bulkiness/size of side-chains, hydrogen-bonding, isoelectric point and structural topology[227]. In addition to amino acid-level embedding, the peptide-wide property was added by averaging across all positions for each peptide sequence. The embeddings from random split and cross-species datasets were used to generate XGBoost classifier using "XGBClassifier" function in python xgboost v0.90 package. The set of hyperparameters, such as alpha, gamma, max_depth and colsample_bytree, were optimized by grid search for each dataset. The model was trained using train datasets: 90% train, Non-SARS-2 and Non-VACV peptides by 10-fold cross-validations. The trained model were tested on 10% test, SARS-2 and VACV peptides, respectively as representatives of 10-fold CV and cross-species comparisons.

To analysis sequence homology between train vs. test datasets, differential position specific scoring matrices (dPSSMs) were generated for each train and test datasets. The probability frequency of each amino acids in each position was represented by position specific scoring matrices using "consensusMatrix" function from R Biostrings v2.56.0 package. The dPSSMs from Positive and Negative peptides were standardised by centre and scaling, and computed difference between two to generate differential dSSM. To estimate discriminative power of dPSSM score, we generated dPSSMs using training data and used the matrices to score respective test peptides for their immunogenicity potential.

Immunogenicity score

The immunogenicity score was computed taking three factors into account: 1) number of experiments conducted, 2) percentage tested positive and 3) number of cognate TCRs if available, using the following equations.

$$(1)\ IS = pr + \left(1 - e^{-\frac{TN}{Qp}}\right)$$

$$(2)\ IS = 1 - \left[pr + \left(1 - e^{-\frac{TN}{Qn}}\right)\right]$$

$$(3)\ x = \frac{1}{1 + e^{-nTCR}}$$

$$(4)\ IS = \min' + \frac{\max' - \min'}{\max(x) - \min(x)} \cdot (x - \min(x))$$

(1) computes immunogenicity score (IS) for Positives peptides that do not have TCR information available, where pr represents % responded positive and TN total number of tests conducted and Qp number of tests designated as a threshold for a sufficient number of tests for Positive peptides (Qp = 3). (2) computes IS for Negative peptides that do not have TCR information available, where Qn represents the designated threshold for sufficient number of tests for Negative peptides (Qn = 5). (3) and (4) compute for Positives having cognate TCR. (3) is a sigmoid function that translate number of cognate TCR (nTCR) to values in logistic growth curve and (4) scales the distribution to align with scores from (1), where min' equals mean of the distribution (~1.28 equivalent to 1 test and 1 responded positive) and max' = 2.0.

Intra- and inter-HLA variability

The HLA effect was computed on 2,349 peptides having entries from multiple HLA alleles. One-way ANOVA was conducted for the effect of HLA on immunogenicity score for each peptide-HLA pair using 'aov' function from R stats v4.0.5 package. The distributions show mean squared from summary output.

Differential sequence patterns

The n-grams, contiguous sequence of n-amino acids, were generated using 'ngram' function from R ngram v3.1.0 package. The number of peptides containing the n-grams was counted for Positive and Negative respectively and normalized by the total number of Positive and Negative peptides. Then, compute $\text{ratio}_{\text{n-gram}}$ = normalized # of Positive peptides containing the n-gram / normalized # of Negative peptides containing the n-gram, and shortlist top differential n-grams by the ratio score. Similarly, we generated all possible combinations of position-specific k-mer motifs (i.e. contiguous or non-contiguous sequence of k amino acids restricted to peptides of same lengths) , where e.g. .M.W. denotes MW pattern at P2 and P4 of 5 amino acid peptide. We computed ratio $_{\text{positional k-mer}}$ = normalized # of Positive peptides containing the positional k-mer / normalized # of Negative peptides containing the positional k-mer to shortlist top differential position-specific k-mer motifs. For visualization, patterns were categorized by their

normalized ratios, where 'lows' have ¼ < ratio < 4, 'high in pos' have ratio ≥ 4, 'high in neg' have ratio ≤ ¼, and 'onlys' have motifs in either positive or negative sets, and top differential patterns from 'high' or 'only' categories were visualized by barplot. To identify shared enriched patterns between pathogenic and self-antigens, only n-grams or positional k-mer motifs having ratio ≥ 3 or ≥ 3 positive peptides (for onlys) were pre-selected for comparison.

The pairwise sequence similarity between all peptides was computed using global alignment using 'pairwiseAlignment' function in Biostrings v2.56.0 package[269], using BLOSUM62 substitution matrix and default parameters. For clusters of highly similar peptides, the peptides having alignment scores >=22 with >= 3 other peptides were visualized into network graph using R igraph v1.2.6 package.

Deep learning models for pathogenic and self-epitope prediction

The peptide sequences at contact positions, i.e. P3-P8 of 9aa and P3-P9 of 10aa peptides, were encoded either by one-hot-encoding having dimension [m, 7, 21], amino acid descriptors [m, 7, 49] or amino acid embeddings from protein transformer-based pretrained language models (PLMs) models [m, 7, 1024], where m represents number of peptides. The amino acid descriptors include kideraFactors, tScales, protFP, BLOSUM, stScales and MSWHIM captured by "aaDescriptors" function from R Peptides v2.4.4 package, and Atchley factors from "AAMetric.Atchley" function in R HDMD v1.2 package. In addition, amino acids were embedded using five protein transformer-based PLMs, including prot_t5_xl_uniref50, prot_bert, prot_bert_bfd, prot_t5_xl_bfd, prot_xlnet from Rostlab using Tokenizer and EncoderModel functions from python transformers v4.19.0 package. These models are based on T5 or BERT and were pretrained on a large corpus of protein sentences, e.g. UniRef50, a dataset consisting of 45 million protein sequences, in a semi-supervised fashion. Further details about transformer-based PLMs can be found in https://huggingface.co/Rostlab.

To account for peptides of varying lengths, the peptides coming from 9aa peptides were padded (i.e. adding non-relevant number to the borders of the matrix) either in the front (i.e. pre-padding) or at the back (i.e. post-padding). First, we computed all possible n-grams from 6- and 7-mer peptides, and analysed if the same n-grams were present in both 6- and 7-mers. We found that many n-grams were present in both 6- and 7-mers, i.e. n-grams having ratio non-NA or non-zero (Figure 4Bi-ii). The ratio of 1 means the respective motif was equally prevalent in 9- and 10-mer peptides. We then aligned by their respective positions, and observed that many 3-grams located in P3-P5 of 9aa peptides were positioned at P4-P6 of 10aa peptides, indicating padding in front of 6-mer peptide (i.e. pre-padding) would align with hotspots in 7-mers (Figure 4Biii).

The classification accuracies of different deep learning architectures were compared between simple dense layer (i.e. classification), biRNN, biLSTM, 1D CNN, 2D CNN and transformer models. The biRNN model contains two biRNN layers each having 512 units, followed by dense layer of 256 units and classification layer with dropouts in-between. The biLSTM model has the same structure as biRNN model. 2D CNN model has two Conv2D layers, with filters = 16 and 32 respectively, followed by MaxPool2D, Flatten, dense layer of 256 units and classification layer with dropouts in-between. The transformer model for pathogenic and self-antigen data had 10

and 2 attention heads respectively and 32 hidden layers in feed forward network, followed by GlobalAveragePooling1D, dense layer with 128 units and classification, with dropouts in-between. The final model is 1D CNN with kernel sizes 1, 3, 5, 7, each max pooled and concatenated to a layer. In parallel, -log transformed NetMHCpan rank and hydrophobicity (i.e. proportion of A, V, L, M, W) have been added as a MLP layer. 1D CNN and MLP layers are concatenated, put into dense layer of 256 units followed by classification. The hyperparameters of the final 1D CNN models have been optimized by grid search. The final hyperparameters for pathogenic model is: learning rate = 1e-05, weight decay = 1e-06, dropout rate = 0.1, batch size = 50, MLP dense = 2000 and dense layer node = 256, giving ROC-AUC of 0.764 by 10-fold cross-validation. The final hyperparameters for self-antigen model is: learning rate = 0.001, weight decay = 0.01, dropout rate = 0.2, batch size = 100, MLP dense = 1500 and dense layer node = 512, giving ROC-AUC of 0.943 by 10-fold cross-validation. All deep learning architectures are implemented using python TensorFlow v2.8.0 package.

Out of distribution (OOD) detection

Autoencoder is a type of unsupervised neural network that have smaller number of neurons in the hidden layers than input layer. This allows the architecture to extract essential information from the input layer, preserve in lower-dimension, and employ to reconstruct output. We use the difference between the input and reconstructed output (i.e. reconstruction loss) as a metric for anomaly detection, as autoencoders cannot effectively reconstruct patterns not learnt from training data[251].

For anomaly detection, peptides encoded by prot_t5_xl_uniref50 transformer-based PLM were applied onto 2D and 3D autoencoders, variational autoencoder and denoising autoencoder models, built using 'AutoEncoder' function from python pyod v1.0.1 package. The autoencoder models trained using 90% random train and non-SARS-CoV-2 data were used to predict 10% test and SARS-CoV-2 data respectively, and computed reconstruction loss between original test and predicted test values.

For calibration methods, the 1D CNN model was trained using softmax activation function, and computed maximum softmax probability and temperature-scaled softmax probability (T=2) for each test peptide. The maximum softmax probability is the maximum probability between class 0 (Negative) and 1 (Positive). The temperature scaling softens the softmax probability with $T > 1$, making the network slightly less confidence and make the confidence scores reflect the true probabilities[270]. For Ensemble, 10 different models each generated by 10-fold training dataset were used to produce prediction scores (i.e. softmax probabilities) on test data, merged maximum softmax probabilities and computed the average of 10 models. For Monte Carlo dropout, predictions from 100 different reiterations with stochastic Dropout(0.6) were merged and averaged per test peptide. The final OOD classifier uses maximum softmax probability and average softmax probability from Ensemble of 10 models as inputs to decision tree classifier to predict correct vs. incorrect predictions. The classifier is built using 'DecisionTreeClassifier' function from python sklearn v1.0.2 package.

Relative similarity to autoantigens or tumor-associated antigens (RSAT)

To compute RSAT, total 5023 unique cancer neoepitopes, autoantigens, tumor-associated antigens and other self-epitopes were retrieved from IEDB, dbPepNeo, NEPdb, McPAS-TCR and tumor antigenic peptide database. First, only pathogenic peptides having comparable self-epitope counterparts are retrained by computing Match score[271] (5) between pathogenic peptide and self-epitopes, where BL represents global-local alignment score using BLOSUM62 matrix, p represents pathogenic peptide and se self-epitopes. Only pathogenic peptides having match score ≥ 0.6 are retained to compute RSAT.

$$(5)\ Match\ score = \frac{BL(p,\ se)}{\sqrt{BL(p,\ se) * BL(p,\ se)}}$$

For pathogenic peptides having comparable self-epitope counterpart, compute relative similarity to self-epitope. First, compute maximum global-local alignment score between pathogenic peptide and self-epitope using BLOSUM62 substitution matrix. Second, compute maximum global alignment score between pathogenic peptide and best self-epitope counterpart in human proteome. Third, compute ratio between alignment score from self-epitope and human proteome counterpart.

Training TRAP to predict immunogenic SARS-CoV-2 epitopes

To train coronavirus TRAP model, 2,295 unique peptide-HLA coronavirus epitopes (including all SARS-CoV-1 and SARS-CoV-2 variants) were collected from IEDB and MIRA in July 2022. For peptides having contradictory immunogenicity response, i) 37 peptides having 1 Positive and $\geq$2 Negative observations, ii) Negative peptide-HLAs having greater number of Positive than Negative observations, and iii) 55 Wuhan Hu-1 epitopes having mutations in Omicron to avoid overlap with test dataset were excluded. Pre-processing and filtering steps left 1,641 unique peptide-HLAs, of which 566 (34.5%) are Negative and 1075 (65.5%) were Positives.

Of these, 1,511 coronavirus epitopes were used to train the coronavirus TRAP model, excluding SARS-CoV-2 Wuhan epitope that were reserved for testing. The embeddings from ProtT5-XL-UniRef model was used to build TRAP comprised of 1D CNN architecture with MLPs incorporating MHC binding rank score and % of hydrophobic amino acids in the peptide.

Validate TRAP's performance in identifying Omicron epitopes

We first validated the performance of coronavirus TRAP model in identifying Wuhan Hu-1 epitopes with Omicron variants. 66 Positive pMHCs were kept to be used for test dataset. The prediction scores of Negatives were estimated by randomly sampling 66 validation predictions 10 times from 10-fold cross-validations. Then, ROC-AUC, PR-AUC, precision, recall, sensitivity, specificity, balanced accuracy, detection rate, prevalence and F1 score were computed for each random sampling.

In silico mutagenesis

From the HLA-I presented SARS-CoV-2 epitopes retrieved from IEDB and MIRA, theoretical mutants were generated from 171 9-mer and 190 10-mer Wuhan Hu-1 originated epitopes by

single point substitutions. We subjected the wild-type (Wuhan Hu-1 epitope, WT) and mutants (MT) to NetMHCpan predictions on 64 HLAs and excluded peptide-HLA pairs where both WT and MT were predicted as non-binder for respective HLA. The WT-MT pairs where both WT and MT were predicted as MHC binders were subsequently subjected TRAP to generate T-cell recognition scores. As TRAP aims to predict T-cell recognition potential on MHC binders only, we imputed the pseudo-zero T-cell recognition scores of 0.01 for predicted non-MHC binders.

Evaluate gain or loss of immunogenicity by mutations

To compute the overall immunogenicity potential, we combined the metrics for antigen presentation (i.e. NetMHCpan prediction) and T-cell recognition (i.e. TRAP prediction).

For antigen presentation metric, we transformed percentile rank from NetMHCpan prediction to range 0-1 using negative logistic function (below). As all the binders range from 0-2, the function evaluates 0.5 (for rank (x) = 2) to 1 (as rank approaches 0). K is a parameter that defines the slope of the curve. We selected k=1.5 to avoid strong discrepancies between two tails of the percentile ranks.

$$MHCScore(x) = \frac{1}{1 + e^k(x - 2)}$$

To generate the overall immunogenicity score, we multiplied the transformed MHCScore with TRAPscore (below). An ImmunogenicityScore was computed for each wild-type and mutant pMHCs.

$$ImmunogenicityScore = MHCScore \times TRAPscore$$

Neighbour Networks

To visualize the immune escape potential, neighbor networks were generated for each epitopes of interest. The approach was adopted from Ogishi *et al.* and their Repitope package https://github.com/masato-ogishi/Repitope. We generated network representations of the single amino acid substitution trajectories, with the mutants with a single substitution are defined as neighbours, connected by edges. Clustering was performed using walktrap algorithm from R igraph package.

Chapter 8. Concluding Discussion

Summary of findings

Antigen-specific T-cells are essential for eliminating intracellular infections, triggering anti-tumor response as well as activating innate immune system. With the growing interest in identifying the cognate antigens of antigen-specific T-cells, many efforts have been made to improve experimental and computational methods for screening, predicting or characterising T-cell epitopes. However, when compared to MHC binding algorithms, models predicting the T-cell recognition potential of MHC-presented peptides performed poorly in both pathogenic and cancer settings[56].

The focus of this work has been to use a systems immunology approach to study divergent antigen-specific T-cell response in different pathologies, to understand features governing peptide immunogenicity and to develop an accurate and robust classifier capable of differentiating CD8+ T-cell targets from MHC-I ligands.

Immunoinformatic approaches are well-suited for gaining insights into antigen-specific T-cell responses. First, breakthroughs in single-cell sequencing technologies have greatly improved our understanding of the cellular interactions as well as clonal dynamics that drive immune responses during inflammation. In chapter 3 of this thesis, the immunoinformatic analysis of single-cell data provided detailed insights into divergent T-cell dynamics between two aetiologies of colitis. This study highlighted that despite these technological advances, the inability to infer T-cell epitopes remains a significant challenge that needs to be addressed.

Second, one major obstacle in designing a high-throughput multiplexed dextramer immune profiling experiment is that the target antigens are not known as a priori. Additionally, testing all possible antigens against antigen-specific T-cells is time-consuming, labour-intensive and expensive. Therefore, an efficient computational pipeline would greatly facilitate shortlisting target antigens for functional validations. In COVID-19 studies presented in chapter 4 of this thesis, we predicted CD8+ T-cell targets from the SARS-CoV-2 proteome, investigated the genetic basis of heterogeneous disease outcomes, examined the impact of mutations to detect mutations of concern and evaluated potential self-reactivity for safer vaccine designs.

Lastly, advances in artificial intelligence have enabled the extraction of complex patterns from non-linear relationships, as well as the development of large-scale pre-trained protein language models that can provide solutions to data constraints. By transfer learning, information learned

from the large-scale models could be transferred to supplement the limited data of our problem. With this, a deep learning based workflow enabled an accurate and robust prediction of CD8+ T-cell epitopes from HLA-I presented pathogenic or cancer peptides.

In this thesis, pathology-specific profiles of antigen-specific T-cells were investigated in the contexts of autoinflammatory and infectious diseases. There was an unmet need in finding the antigens driving the cellular immune response, requiring the development of new strategies for identifying CD8+ T-cell epitopes. To begin, a comprehensive landscape of features governing the peptide immunogenicity was explored. Following that, the performance of existing models was evaluated, and novel deep learning-based strategies for more accurate prediction of CD8+ T-cell targets were developed. The works presented in this thesis will serve as the basis for future research into emerging pathogens, cancer neoantigens and targets for effective vaccines and T-cell mediated therapies.

Divergent T-cell profiles between ulcerative and checkpoint treatment-induced colitis, likely driven by diverse unknown antigen

Checkpoint inhibitor-induced colitis affects up to 60% of cancer immunotherapy patients. However, because it is a new entity, little was known about the molecular mechanisms underlying its pathology.

The multi-modal (RNA-Seq, VDJ-Seq and CITE-Seq) single-cell technologies have become a powerful system to generate expression profiles of highly heterogeneous cell populations. The single-cell profiling of matched colonic and PBMC samples was performed in a cohort of checkpoint inhibitor therapy-induced and ulcerative colitis patients. To the best of our knowledge, this was the first study to investigate CC by a) comparing cellular heterogeneity with UC, b) integrating spatial context with multiplexed transcriptional measurements, and c) tracking T-cell dynamics from blood to the site of inflammation via immune profiling of both blood and tissue. Although not discussed in this thesis, non-immune compartments such as epithelium and stromal cells, which had not been studied previously, were also investigated.

A number of interesting aspects of T-cell activity have been uncovered. First, disease-specific changes were mapped at the single-cell level. When CC was compared to UC, CC had a significant increase in a highly proliferative CD8+ population, whereas UC had more CD4+ T-cells. In CC, we observed an overall increase in TRM reactivation pathways marked by the ZNF683+/Hobit TRM population. In contrast, Tc17/Th17 cells were the most proliferative cells in UC, with ZNF683+ cells being much rare. This finding was supported by spatial transcriptomics data, which revealed that exhausted CD8+ HAVCR2+ and ZNF683+ cells had a higher signal in CC slides across epithelial crypts. ZNF683$^+$ cells with a similar phenotype were recently implicated as the intra-tumor lymphocytes responsible for anti-tumor responses in lung cancer following checkpoint inhibitor treatment[90]. This suggests that activation of analogous T-cell populations may be responsible for both anti-tumor responses and adverse immune effects in the gut.

Second, using single-cell VDJ data, the landscape of T-cell migration from circulating blood to the site of inflammation was investigated. While T-cell trafficking of effector populations was

observed in both UC and CC, it was less pronounced in CC, with even the effector populations in UC retaining the lymph node homing markers. In terms of locally proliferating T-cells, CD8+ resident memory expansions in UC were skewed towards Tc17 polarised cells, as previously reported[73,77]. These Tc17 populations had a high degree of clonal overlap with the exhausted and cycling populations. In CC, on the other hand, the majority of locally-proliferating non-trafficked T-cells were formed by a reactivated ZNF683^{+} population, which clonally overlapped with exhausted and cycling populations. While there is currently no biomarker or treatment for gastrointestinal immune related adverse effects (irAEs) such as CC, our work shows that T-cell trafficking occurs at lower levels in CC compared to UC, implying that vedolizumab, a key biologic used to treat UC, may have limited efficacy in CC. Further characterisation of the targets will be required in order to understand the full picture of the inflammation in colitis and identify therapeutic targets.

Third, clustering analysis of clonally expanded and/or migrated T-cells revealed that T-cell responses were likely triggered by a variety of antigens, but the source antigen could not be identified. The source antigen of these T-cells were investigated by cross-referencing to public databases, but mapping T-cell specificity proved difficult due to the small pool of known antigen-cognate TCR pairs. Because ZNF683+ cells have been linked to anti-tumor responses, analogous cells with this phenotype are likely to be targeting cancer neoantigens in the tumour microenvironment. However, the trigger of the gut responses is less clear. Furthermore, it remains unclear whether these cells target luminal bacterial or viral antigens, or altered self-peptides from benign somatic mutations. As barrier sites typically have a higher somatic mutation burden, the latter hypothesis is especially interesting in light of the increased presence of reactivated TRMs in the intra-epithelial space.

To investigate this hypothesis, preliminary experiments are in progress using a yeast display system to screen cognate antigens of key selected TCRs against combinatorial peptide libraries. Clones of interest were selected from checkpoint inhibitor-treated samples and controls based on a) evidence of clonal expansion (minimum of 3 cells), b) evidence of Nivolumab binding, and c) no known antigen specificity in public databases. CD8+ T cell clones that demonstrated progression from normal/re-activated to cycling to PD1+ expressing exhausted/highly activated phenotypes were of particular interest. Uncovering the antigen specificity of these orphan TCRs in future research will be a step towards a better understanding of the antigens likely driving divergent T-cell responses in colitis of different aetiology.

Predicting CD8+ T-cell targets enhance understanding of COVID-19 pathology and facilitate vaccine design

In early 2020, there was an urgent need to understand the pathophysiology of a novel and unknown pathogen, SARS-CoV-2, in order to prevent its spread and develop effective vaccines and treatments. Chapters 4-7 in this thesis described the initiatives to shortlist vaccine candidates, understand heterogeneity in COVID-19 pathology, and develop a computational workflow to identify COVID-19 vaccine targets. Because emerging variants continue to pose a global threat, the novel workflow was adopted to assess their impact.

Optimal selection of T-cell epitopes was crucial to accelerate vaccine development. We conducted *in silico* search and were among the first to report a comprehensive list of vaccine candidates[104]. First, 76 SARS-CoV-2 peptides were found to have high sequence similarity to previously identified coronavirus epitopes. We then conducted a systematic immunogenicity prediction on all possible SARS-CoV-2 peptides as a quick and resource-efficient solution to assist biologists and clinicians in preventing the spread of infection. While working on this manuscript, we became aware of a similar study comparing SARS-CoV-2 with SARS-CoV-1 epitopes[272]. The fact that these two studies were carried out independently using different approaches but yielded consistent results demonstrates a high level of confidence in reproducing the results.

In addition, with the growing concern about the emergence of new variants, we aimed to investigate how pre-exposure to other coronavirus strains affected the heterogeneous response to natural infection or vaccination. On seven coronavirus strains, a *de novo* search for CD8+ T-cell epitopes was performed and potential CD8+ T-cell cross-reactivity was investigated. 794 predicted SARS-CoV-2 epitopes were highly similar to other coronavirus strains, 21 of which were identical to one of four common flu-like coronaviruses. Because little was known about T-cell cross-reactivity or the presence of immune memory at the time of the study, this systematic analysis provided a comprehensive picture of the potential cross-protection that can be conferred by other coronavirus strains. As coronavirus variants continue to pose a global threat, similar research evaluating the impact of emerging variants based on pre-existing immunity will be extremely valuable.

Moreover, the work presented in chapter 4 was greatly useful in resolving a conflicting body of evidence regarding T-cell cross-reactivity and how it influences heterogeneity in disease pathology and clinical outcome. At the time of the study, Nguyen *et al.* found little evidence of cross-reactivity of TCRs recognising the immunodominant SARS-CoV-2 SPR* epitope in association to HLA-B*07:02 allele[273]. In contrast, Francis *et al.* demonstrated the presence of pre-existing CD8+ T-cell memory to SARS-CoV-2 in naïve samples, conditioned by genotype, and suggested that unexposed individuals with specific HLA alleles (e.g., B*07:02) may be more likely to possess cross-reactive memory T-cells specific for SPR* SARS-CoV-2 epitopes[123]. Therefore, the role of genetics in the immune response to vaccination or natural infection was investigated. Patients with specific HLA alleles (e.g. HLA-B*07:02, C*07:02 and A*03:01) were found more likely to have cross-reactive TCRs capable of binding both SARS-CoV-2 and other human coronaviruses. This observation was confirmed by a separate functional study that looked into the HLA association of T-cell response[274]. Although this study was unable to establish a direct link to pre-existing immunity, it demonstrated that cross-immunity is not binary and is likely to be influenced by a number of other factors including HLAs. Moreover, patients with cross-reactive TCRs may have more robust protection against SARS-CoV-2 variants, other human coronaviruses or even future emerging coronavirus strains.

Lastly, the risk of self-reactivity for SARS-CoV-2 vaccine candidates was evaluated by comparing them to the reference human proteome. While negative selection has been proposed to eliminate self-reactive T-cell precursors, viral antigens that are structurally similar to self-antigens have been implicated to induce autoimmunity via molecular mimicry[275]. It was also hypothesised

that a non-specific antiviral immune response may cause damaged cells to release self-antigens, resulting in bystander activation of autoreactive T-cell[275,276]. Indeed, Venkatakrishnan *et al.*[277] identified peptides that are identical between SARS-CoV-2 and the human proteome, and reported that the genes giving rise to these peptides are expressed in tissues implicated in COVID-19 pathogenesis. As such, SARS-CoV-2 peptides that are highly similar to the self-proteome may have immunopathological or autoimmune potential. To mitigate the potential toxicity, we assessed the risk of self-reactivity of predicted CD8+ T-cell epitopes. There were no predicted epitopes identical to the human proteome, but there were 10 and 184 predicted epitopes with one and two amino acid mismatches respectively, that were highly similar to the human proteome. Our work aimed to guide toward the safe selection of epitopes for therapeutic development.

These studies presented collective efforts to understand the mechanism underlying COVID-19 pathology and propose therapeutic targets for preventing or treating the disease. The comprehensive profile of predicted SARS-CoV-2 epitopes, as well as analyses of their self-reactivity and cross-reactivity provide a valuable resource for the development of vaccines that can protect against multiple prevalent coronavirus strains. This work also laid the foundation for assessing the performance of existing immunogenicity classifiers and motivated the development of a more accurate and robust algorithm.

Existing computational models fail to accurately predict CD8+ T-cell targets in both pathogen and cancer settings

The colitis and COVID-19 studies demonstrated the importance of identifying CD8+ T-cell targets that drive antigen-specific T-cell responses in understanding divergent pathologies and developing therapeutics. Several methods for identifying CD8+ T-cell epitopes have previously been developed, utilising features such as positional information, physicochemical properties and similarity to known epitopes. However, with the recent emergence of computational models, the accuracy and variability of model performance in different contexts, such as in predicting cancer neoepitopes or epitopes from emerging pathogens, have not yet been determined. This posed a particular challenge in estimating the accuracy of vaccine target prediction against SARS-CoV-2, which was later found to have 66% accuracy when compared to functional study by Nelde *et al.*[278].

Therefore, to address this limitation, eight immunogenicity and one MHC binding algorithms were systematically benchmarked in predicting CD8+ T-cell epitopes in emerging pathogen and cancer settings, and all models performed poorly in both settings. The immunogenicity classifiers did not outperform MHC binding predictors in predicting pathogenic epitopes, and performed only marginally better in predicting cancer neoepitopes, with many algorithms producing a high number of false positives (low precision). This was a significant step forward in providing insights about existing models' poor performance and assisting end users in making evidence-based decisions about which model and parameters to use for their data and research questions.

The next big question was finding out the causes of the poor performance, as well as the reasons for discrepancies between reported performance in the original publications and benchmarking

studies or other applications. We found that, first, an imbalanced number of epitopes and non-epitopes across different HLAs in the training datasets led models to erroneously capture dominant HLA binding features. Second, the T-cell recognition features may differ between neoepitopes and pathogenic peptides, as well as between peptides derived from different pathogenic species (i.e. named out-of-distribution generalisation problem). Third, the high performance of immunogenicity models incorporating peptide-HLA pairs are likely due to HLA bias from imbalanced datasets. These findings were critical in laying the foundation for the development of a more robust classifier.

Development of a deep learning workflow to discriminate CD8+ T-cell targets from MHC-I ligands

Most immunogenicity algorithms face difficulties in learning classification features from a limited number of epitopes and non-epitopes, as well as overcoming the out-of-distribution generalisation problem from peptides derived from different hosts, organisms or diseases. To address these challenges, a novel immunoinformatic workflow was developed for identifying CD8+ T-cell epitopes that leverage recent advances in protein transformer-based PLMs and deep learning architectures.

The TRAP (T-cell recognition potential of HLA-I presented peptides) is a robust deep learning workflow for predicting CD8+ T-cell epitopes from MHC-I presented pathogenic and self-peptides. To address the aforementioned challenges, TRAP used peptide sequences only at contact positions to avoid HLA bias caused by capturing dominant HLA binding features at anchor positions. Second, separate models for pathogenic and self-peptides were built to address the out-of-distribution generalisation problem. Third, to overcome data limitations, peptide sequences were encoded using amino acid properties derived from protein transformer-based pre-trained language models (PLMs). This enabled more information about the physicochemical, electrostatic or biophysical properties of amino acids to be fed into the system. Fourth, T-cell recognition motifs were captured using a one-dimensional convolutional neural network (1D CNN) architecture. Fifth, the MHC binding rank score predicted by the most accurate algorithm (NetMHCpan) was added to provide the information about MHC binding. Lastly, an out-of-distribution detection module was employed to abstain from low-confidence predictions for peptides that are very different from those in the training dataset. By these approaches, TRAP outperformed other machine learning algorithms.

Another metric proposed to estimate immunogenicity from cancer neoepitope studies was dissimilarity to self. However, the dissimilarity to self could not be applied to pathogenic peptides because they have highly heterologous sequences compared to the human proteome. Therefore, a novel metric termed RSAT (Relative similarity to autoantigens or tumor-associated antigens) was developed to estimate the immunogenicity of peptides from emerging pathogens. To compensate for the high dissimilarity between pathogenic and human proteomes, this metric compares the similarity of pathogenic peptides to the reference human proteome ('healthy') with respect to 'immunogenic' self-peptides such as autoantigens or tumor-associated antigens. This allowed for the identification of pathogenic epitopes despite their sequence dissimilarity to the human

proteome.

The TRAP model was used to study the immune escape potential of SARS-CoV-2 variants as well as the impact of mutagenesis in giving rise to deleterious variants of concern. We combined *in silico* mutagenesis, MHC binding (i.e. NetMHCpan) and T-cell recognition (i.e. TRAP) predictions to draw the trajectory of the deleterious or beneficial mutations. The development of a robust immunogenicity classifier enabled surveillance of SARS-CoV-2 variants, as well as the detection of variants that may result in increased transmissibility, disease severity or other adverse effects on public health. Although not discussed in this thesis, the TRAP self-antigen model was used to shortlist cancer neoepitopes from glioblastoma patients as well. Such applications demonstrate how TRAP can be used in practice.

In conclusion, chapter 7 of this thesis presented metrics and algorithms developed for improving the prediction of CD8+ T-cell epitopes from MHC-I ligands. We demonstrated how these models can be used to monitor and assess variants of concern to mitigate the impact of future pandemics. They will serve as a toolkit for identifying CD8+ T-cell targets for a wide range of applications in basic and translational research.

Future directions

In this work, systems immunology and artificial intelligence approach was used to elucidate antigen-specific T-cell responses, with the goal of determining whether viral or cancer-associated antigens are likely to be recognised by T-cells and elicit an immune response. In the context of pMHC:TCR interaction, this thesis focused on the first component, namely the features underlying peptide immunogenicity. However, TCR-associated features, including T-cell cross-reactivity and shared antigen specificity, have also been the subject of extensive research with significant achievements, but challenges remain.

Although recent biological and computational advances to screen antigenic peptides and profile TCR repertoires have greatly improved our understanding of the pMHC:TCR interaction, the picture is far from complete. The following sections will describe future directions to extend the study of peptide immunogenicity to better understand the rules underlying antigen recognition by *specific* T-cells, as well as the application of immunogenicity models to systematically study the impact of emerging pathogens. Next, it will describe future directions to study the other side of TCR:pMHC interaction - the rules governing cross-reactivity and shared antigen-specificity of T-cells, and how these two pillars should be studied in the light of one another to depict the entire pMHC:TCR interactome.

Rules underlying antigen recognition by T-cells

On the first component of pMHC:TCR interaction, it will be valuable to extend the work of peptide immunogenicity to investigate the ability of peptides to be recognised by *specific* T-cells. Here, the peptide immunogenicity was investigated at the organismal level i.e. whether the peptide can elicit a response from *any* T-cell. The limitation was largely due to the limited

pool of available peptides characterised for their cognate TCRs, however, advances in screening methods will lead to availability of more comprehensive datasets in the future, thus enabling development of more tailored immunogenicity models. It is becoming apparent that predicting specific interactions between TCR and cognate pMHCs is crucial for developing personalised therapies and tailoring vaccines or treatments to individuals' TCR repertoire. Therefore, screening antigen-specific TCR against a larger pool of epitopes from various origins and pathologies will greatly aid in learning peptide features that allow interaction with specific TCRs.

Moreover, based on the groundwork of this work, using models like TRAP to systematically evaluate the impact of mutations on the emergence of deleterious pathogens will be of great interest. Emerging pathogens have posed a significant threat in recent years, and new variants and pathogens are expected to rise in the coming years[279]. It is therefore critical to surveil variants of concern and assess the immune escape potential of these variants. Furthermore, in order to accurately determine deleterious variants, it is essential to assess the extent to which models trained by other pathogens can be generalised and thus used for emerging pathogens. Thus, an iterative process of refining the training data, model architecture and validating predictions should be followed to mitigate the impact of another pandemic.

Rules underlying cross-reactivity of T-cells

It is now well accepted that peptide-specific TCRs exhibit high levels of cross-reactivity[65,280,281] i.e. a TCR may be capable of recognising more than one peptide-MHC molecule. TCR cross-reactivity is a double-edged sword: while a highly cross-reactive T-cell can effectively screen for a broad spectrum of epitopes, this may also lead to dysregulated T-cell responses, which may contribute to allergy, immunopathology, autoimmunity and chronic infection[282–285]. Furthermore, prior exposure to degenerate T-cells can lead to a breakdown of tolerance[286–291], inducing a polarised response to a pathogen or vaccination [292,293]. As a result, there has been a great interest in deciphering TCR cross-reactivity to better understand disease pathology and develop safer therapeutics.

However, while any single TCR has been proposed have the potential to recognise as many as 10^4-10^7 different MHC-bound epitopes[65], cross-reactive and cross-species TCRs account for only 9.5% and 5.7% of TCRs in the current database, respectively (Figure 1.1C, Supplementary Figure 2), implying substantial under-sampling in TCR cross-reactivity.

Several experimental methods exist for determining the polyspecificity of TCRs[34–37,41,42,212,294–299]; however, these are labour-intensive and low-throughput. Therefore, learning the general features governing the breadth or constituents of cross-reactivity has been challenging due to limited data, and algorithms for predicting the TCR cross-reactivity are still in their infancy.

Therefore, the development of experimental methodologies that allow for the simultaneous screening of peptide libraries against *numerous* TCRs will greatly aid in learning general characteristics that determine the breadth and targets of cross-reactive TCRs. Moreover, a repository of both positive and negative epitope TCR binding data will be highly valuable in narrowing down true cross-reactive space, rather than naively assuming non-binding by the

background. Such approaches will lead to better understanding of T-cell cross-reactivity, which has a wide range of applications such as identifying antigens with a high risk of cross-reactivity or off-target toxicity, and ultimately selecting a safer pool of therapeutic targets for personalised therapies[300,301].

Rules underlying shared antigen-specificity of T-cells

Recent advances in bulk and single-cell sequencing technologies allowed high-throughput generation of antigen-specific TCRs. These datasets served as the foundation for several recent studies attempting to predict TCR specificity groups based on CDR sequences[62,213,235,302–313] (Supplementary Figure 3).

However, only a few viral or cancer-associated antigens have been extensively screened. As a result, knowledge is limited to the few selected epitopes, and computational methods are biased toward the well-studied epitopes. To compensate for the limited data, unsupervised clustering methods have been the dominant approach for predicting shared antigen-specificity i.e. grouping unlabelled TCR sequences based on their similarity, which reflects the likelihood of recognising the same or related antigens. This approach, however, cannot address the fundamental question of what the cognate antigen is.

Hence, the foremost should be screening antigen-specific T-cells against a broader range of antigens, in the context of different hosts, pathogens, diseases and body systems. Second, with an increasing volume of unlabelled data from high-throughput sequencing, semi-supervised clustering with the help of labelled data may help understanding the specificity of the unlabelled data, though we continue to struggle with labelling orphan TCRs in an unbiased manner. In this regard, screening against a larger pool of antigens will fundamentally help in the provision of labelled data for semi-supervised clustering. Third, the majority of models only incorporate TCRβ sequence and VJ gene usage due to a lack of paired alpha/beta or gamma/delta TCRs. Furthermore, there is little use of expression data from transcriptomics, proteomics or spatial data due to limited understanding of the relationship between expression profiles and T-cell specificity. Combining information on paired TCR, cellular composition, cross-talk, and pathology-specific spatial distribution with TCR specificity would greatly improve inference and interpretation of T-cell antigen specificity.

Finally, experiments resembling biological conditions, such as a pool of intestinal antigens versus a pool of gut TCRs, or a pool of cancer-associated antigens / self-antigens versus tumor microenvironment TCRs, would shed more light on disease mechanisms. While combinatorial display systems are useful for unbiased screening, presented peptides are artificial and even if they do elicit an *in vitro* response, some may never be encountered by T-cell *in vivo*. Furthermore, each experiment screens a limited number of TCRs at a time, which restricts understanding the full picture of the interaction of diverse antigens and TCRs forming the resultant immune response. A higher throughput system would provide greater insights into antigens that are likely to be presented and recognised, the presence of agnostic or antagonistic binding, likely competition between antigens or TCRs, differences in binding affinity or the presence of molecular

mimicry affecting the final immune response.

pMHC:TCR interactome

Cross-reactivity and shared antigen-specificity are often studied separately; thus, understanding the relationship between the two closely interconnected notions seems to be underrepresented in the research community. In order to set the foundation for a better understanding of T-cell response, these two pillars of adaptive immunity should be investigated concurrently.

Depicting the cross-recognition of TCRs and pMHCs in $>10^6$ space, modelling pMHC:TCR landscape should be taken as a cross-talk between unique and representative clusters of pMHCs and TCRs rather than individual entities, namely 'pMHC:TCR interactome' (Figure 8.1).

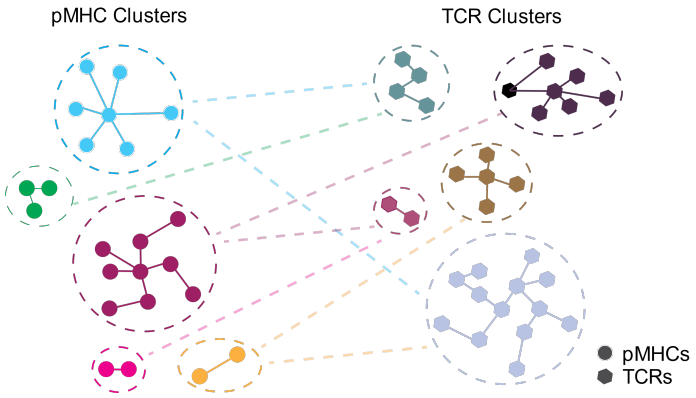


Figure 8.1: Interplay between unique clusters of pMHCs and TCRs.

In an ideal world with an accurate distance measure, pMHCs in the same cluster should share the common specificity towards TCRs and vice versa. Each node denotes pMHC (circle) or TCR (polygon) entities and the edge denote the distance with the closest pMHC or TCR, respectively.

However, there are remaining challenges in constructing the pMHC:TCR interactome. First, the number of available TCRs and surveillance pMHCs falls far short of covering the entire combinatorial interactome space. The development of methodologies for screening numerous epitopes against a pool of TCRs (i.e. many pMHCs vs. many TCRs) will be a huge step forward.

Furthermore, as discussed by Bradley *et al.*, we currently do not know whether TCRs that are predicted to be similar by antigen specificity algorithms e.g. GLIPH or TCRDist are expected to have similar cross-reactome[314]. An intuitive answer would be that TCRs with a shorter distance, especially those within the same cluster, will have a greater overlap of cross-reactive peptides. However, structural studies have revealed that sequence similarity cannot adequately represent cross-recognition potential as TCRs can have multiple configurations for different classes of epitopes[315].

Therefore, to model this dynamic interplay, an accurate distance measure to group TCRs informative of their antigen specificity and cross-reactivity may be required. This will require assessment of all identified features, such as paired TCR$\alpha\beta$ sequences, k-mer motifs, physicochemical prop-

erties as well as structural, physical and kinetic parameters, in order to derive a minimum set of features that maximize association to modelling the interactome. These features will become a toolkit for developing TCR and pMHC distance measures that partition $>10^6$ TCRs and $>10^{18}$ peptides into designated clusters. The relationship between TCR and pMHC clusters may segregate into a linear function or yield an indistinct pattern in which even repertoires for closely related epitopes have divergent landscapes with very limited overlap.

Ultimately, building the complete map portraying the pMHC:TCR interaction will provide opportunities to describe the immune system's dynamic response to disease or therapeutics. This will enable the development of personalised treatments that take into account the individual's endogenous ability to target a given disease-specific antigen as well as the individual's risk of autoimmunity.

In summary, future immunoinformatic analysis of T-cell epitopes will likely focus on scaling up the breadth of antigens, personalising targets to individual's TCR repertoire and improving the robustness of T-cell epitope and TCR specificity predictions. This requires the development of high-throughput screening assays, comprehensive antigen TCR datasets, and tailored modelling strategies. Furthermore, data analysis strategies for integrating multiple layers of information, such as TCR specificity with expression and spatial information, will enable to make the best use of the data.

Chapter 9. References

1. Kumar, B. V., Connors, T. J. & Farber, D. L. Human T Cell Development, Localization, and Function throughout Life. *Immunity* **48**, 202–213 (2018).

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

3. von Boehmer, H. & Fehling, H. J. Structure and Function of the Pre-T Cell Receptor. *Annual Review of Immunology* **15**, 433–452 (1997).

4. Nikolich-Žugich, J., Slifka, M. K. & Messaoudi, I. The many important facets of T-cell repertoire diversity. *Nat Rev Immunol* **4**, 123–132 (2004).

5. Robey, E. & Fowlkes, B. J. Selective Events in T Cell Development. *Annual Review of Immunology* **12**, 675–705 (1994).

6. Deftos, M. L., Huang, E., Ojala, E. W., Forbush, K. A. & Bevan, M. J. Notch1 Signaling Promotes the Maturation of CD4 and CD8 SP Thymocytes. *Immunity* **13**, 73–84 (2000).

7. van den Broek, T., Borghans, J. A. M. & van Wijk, F. The full spectrum of human naive T cells. *Nat Rev Immunol* **18**, 363–373 (2018).

8. How regulatory T cells work | Nature Reviews Immunology. https://www.nature.com/articles/nri2343.

9. Merkenschlager, M. *et al.* How Many Thymocytes Audition for Selection? *Journal of Experimental Medicine* **186**, 1149–1158 (1997).

10. Zerrahn, J., Held, W. & Raulet, D. H. The MHC Reactivity of the T Cell Repertoire Prior to Positive and Negative Selection. *Cell* **88**, 627–636 (1997).

11. von Boehmer, H., Teh, H. S. & Kisielow, P. The thymus selects the useful, neglects the useless and destroys the harmful. *Immunology Today* **10**, 57–61 (1989).

12. Li, M. O. & Rudensky, A. Y. T cell receptor signalling in the control of regulatory T cell differentiation and function. *Nat Rev Immunol* **16**, 220–233 (2016).

13. Germain, R. N. MHC-dependent antigen processing and peptide presentation: Providing ligands for T lymphocyte activation. *Cell* **76**, 287–299 (1994).

14. Swain, S. L. T Cell Subsets and the Recognition of MHC Class. *Immunological Reviews* **74**, 129–142 (1983).

15. Feinerman, O., Germain, R. N. & Altan-Bonnet, G. Quantitative challenges in understanding ligand discrimination by $\alpha\beta$ T cells. *Molecular Immunology* **45**, 619–631 (2008).

16. Haidar, J. N. *et al.* Structure-Based Design of a T Cell Receptor Leads to Nearly 100-Fold Improvement in Binding Affinity for pepMHC. *Proteins* **74**, 948–960 (2009).

17. Hellman, L. M. *et al.* Improving T Cell Receptor On-Target Specificity via Structure-

Guided Design. *Molecular Therapy* **27**, 300–313 (2019).

18. Malecek, K. *et al.* Specific Increase in Potency via Structure-Based Design of a TCR. *The Journal of Immunology* **193**, 2587–2599 (2014).

19. Pierce, B. G. *et al.* Computational Design of the Affinity and Specificity of a Therapeutic T Cell Receptor. *PLOS Computational Biology* **10**, e1003478 (2014).

20. Zoete, V., Irving, M., Ferber, M., Cuendet, M. & Michielin, O. Structure-Based, Rational Design of T Cell Receptors. *Front. Immunol.* **4**, (2013).

21. Zhao, Y. *et al.* High-Affinity TCRs Generated by Phage Display Provide CD4+ T Cells with the Ability to Recognize and Kill Tumor Cell Lines. *The Journal of Immunology* **179**, 5845–5854 (2007).

22. Gate, D. *et al.* Clonally expanded CD8 T cells patrol the cerebrospinal fluid in Alzheimer's disease. *Nature* **577**, 399–404 (2020).

23. Jiang, W. *et al.* In vivo clonal expansion and phenotypes of hypocretin-specific CD4 + T cells in narcolepsy patients and controls. *Nature Communications* **10**, 1–17 (2019).

24. Grifoni, A. *et al.* SARS-CoV-2 human T cell epitopes: Adaptive immune response against COVID-19. *Cell Host & Microbe* **29**, 1076–1092 (2021).

25. Lang-Meli, J. *et al.* SARS-CoV-2-specific T-cell epitope repertoire in convalescent and mRNA-vaccinated individuals. *Nat Microbiol* **7**, 675–679 (2022).

26. Mateus, J. *et al.* Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. *Science* **370**, 89–94 (2020).

27. Haynes, W. A. *et al.* High-resolution epitope mapping and characterization of SARS-CoV-2 antibodies in large cohorts of subjects with COVID-19. *Commun Biol* **4**, 1–14 (2021).

28. Stromnes, I. M., Schmitt, T. M., Chapuis, A. G., Hingorani, S. R. & Greenberg, P. D. Readapting T cells for cancer therapy: from mouse models to clinical trials. *Immunological Reviews* **257**, 145–164 (2014).

29. Zhao, Q. *et al.* Engineered TCR-T Cell Immunotherapy in Anticancer Precision Medicine: Pros and Cons. *Front Immunol* **12**, 658753 (2021).

30. Whittington, M. D., Ollendorf, D. A. & Campbell, J. D. Accounting for All Costs in the Total Cost of Chimeric Antigen Receptor T-Cell Immunotherapy. *JAMA Oncology* **4**, 1784–1785 (2018).

31. Paul, S. *et al.* Benchmarking predictions of MHC class I restricted T cell epitopes in a comprehensively studied model system. *PLoS Comput Biol* **16**, e1007757 (2020).

32. Joglekar, A. V. & Li, G. T cell antigen discovery. *Nature Methods* 1–8 (2020) doi:10.1038/s41592-020-0867-z.

33. Singh, S. K. *et al.* The simultaneous ex vivo detection of low-frequency antigen-specific CD4+ and CD8+ T-cell responses using overlapping peptide pools. *Cancer Immunol Immunother* **61**, 1953–1963 (2012).

34. Joglekar, A. V. *et al.* T cell antigen discovery via signaling and antigen-presenting bifunctional receptors. *Nature Methods* **16**, 191–198 (2019).

35. Kula, T. *et al.* T-Scan: A Genome-wide Method for the Systematic Discovery of T Cell Epitopes. *Cell* **178**, 1016-1028.e13 (2019).

36. Bentzen, A. K. *et al.* T cell receptor finger-

printing enables in-depth characterization of the interactions governing recognition of peptide–MHC complexes. *Nature Biotechnology* **36**, 1191–1196 (2018).

37. Zhang, S.-Q. *et al.* High-throughput determination of the antigen specificities of T cell receptors in single cells. *Nature Biotechnology* **36**, 1156–1159 (2018).

38. Fleri, W. *et al.* The Immune Epitope Database and Analysis Resource in Epitope Discovery and Synthetic Vaccine Design. *Frontiers in Immunology* **8**, (2017).

39. Bagaev, D. V. *et al.* VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res* **48**, D1057–D1062 (2020).

40. Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E. & Friedman, N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* **33**, 2924–2929 (2017).

41. Li, G. *et al.* T cell antigen discovery via trogocytosis. *Nature Methods* **16**, 183–190 (2019).

42. Pan, X. *et al.* Combinatorial HLA-peptide bead libraries for high throughput identification of CD8+ T cell specificity. *Journal of Immunological Methods* **403**, 72–78 (2014).

43. Crooke, S. N., Ovsyannikova, I. G., Kennedy, R. B. & Poland, G. A. Identification of naturally processed Zika virus peptides by mass spectrometry and validation of memory T cell recall responses in Zika convalescent subjects. *PLoS One* **16**, e0252198 (2021).

44. Johnson, K. L., Ovsyannikova, I. G., Mason, C. J., Bergen, H. R. & Poland, G. A. Discovery of naturally processed and HLA-presented class I peptides from vaccinia virus infection using mass spectrometry for vaccine development. *Vaccine* **28**, 38–47 (2009).

45. Wilhelm, M. *et al.* Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nat Commun* **12**, 3346 (2021).

46. Karlsson, A. C. *et al.* Comparison of the ELISPOT and cytokine flow cytometry assays for the enumeration of antigen-specific T cells. *Journal of Immunological Methods* **283**, 141–153 (2003).

47. Croft, N. P. *et al.* Most viral peptides displayed by class I MHC on infected cells are immunogenic. *PNAS* **116**, 3112–3117 (2019).

48. Hensen, L. *et al.* CD8+ T cell landscape in Indigenous and non-Indigenous people restricted by influenza mortality-associated HLA-A*24:02 allomorph. *Nat Commun* **12**, 2931 (2021).

49. Calis, J. J. A. *et al.* Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. *PLOS Computational Biology* **9**, e1003266 (2013).

50. Ogishi, M. & Yotsuyanagi, H. Quantitative Prediction of the Landscape of T Cell Epitope Immunogenicity in Sequence Space. *Front. Immunol.* **10**, (2019).

51. Li, G., Iyer, B., Prasath, V. B. S., Ni, Y. & Salomonis, N. DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity. *Briefings in Bioinformatics* (2021) doi:10.1093/bib/bbab160.

52. Wells, D. K. *et al.* Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. *Cell* **183**, 818-834.e13 (2020).

53. Trolle, T. & Nielsen, M. NetTepi: an integrated method for the prediction of T cell epitopes. *Immunogenetics* **66**, 449–456 (2014).

54. Pogorelyy, M. V. *et al.* Exploring the pre-immune landscape of antigen-specific T cells. *Genome Medicine* **10**, 68 (2018).

55. Lee, C. H. *et al.* Potential CD8+ T Cell Cross-Reactivity Against SARS-CoV-2 Conferred by Other Coronavirus Strains. *Front. Immunol.* **11**, (2020).

56. Buckley, P. R. *et al.* Evaluating performance of existing computational models in predicting CD8+ T cell pathogenic epitopes and cancer neoantigens. *Briefings in Bioinformatics* **23**, bbac141 (2022).

57. Karnaukhov, V. *et al.* HLA binding of self-peptides is biased towards proteins with specific molecular functions. *bioRxiv* 2021.02.16.431395 (2021) doi:10.1101/2021.02.16.431395.

58. Bassani-Sternberg, M. & Gfeller, D. Unsupervised HLA Peptidome Deconvolution Improves Ligand Prediction Accuracy and Predicts Cooperative Effects in Peptide–HLA Interactions. *The Journal of Immunology* **197**, 2492–2499 (2016).

59. Bassani-Sternberg, M. *et al.* Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLOS Computational Biology* **13**, e1005725 (2017).

60. Jurtz, V. *et al.* NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *The Journal of Immunology* **199**, 3360–3368 (2017).

61. Zhang, W. *et al.* PIRD: Pan Immune Repertoire Database. *Bioinformatics* **36**, 897–903 (2020).

62. Zhang, W. *et al.* A framework for highly multiplexed dextramer mapping and prediction of T cell receptor sequences to antigen specificity. *Science Advances* **7**, eabf5835 (2021).

63. Birnbaum, M. E. *et al.* Deconstructing the Peptide-MHC Specificity of T Cell Recognition. *Cell* **157**, 1073–1087 (2014).

64. Kedzierska, K. & Koutsakos, M. The ABC of Major Histocompatibility Complexes and T Cell Receptors in Health and Disease. *Viral Immunology* **33**, 160–178 (2020).

65. Mason, D. A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunology Today* **19**, 395–404 (1998).

66. Zhang, N. & Bevan, M. J. CD8+ T Cells: Foot Soldiers of the Immune System. *Immunity* **35**, 161–168 (2011).

67. Paludan, S. R., Pradeu, T., Masters, S. L. & Mogensen, T. H. Constitutive immune mechanisms: mediators of host defence and immune regulation. *Nat Rev Immunol* **21**, 137–150 (2021).

68. Dougan, M., Luoma, A. M., Dougan, S. K. & Wucherpfennig, K. W. Understanding and treating the inflammatory adverse events of cancer immunotherapy. *Cell* **184**, 1575–1588 (2021).

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

70. Bishu, S. *et al.* Efficacy and Outcome of Tofacitinib in Immune checkpoint Inhibitor Colitis. *Gastroenterology* **160**, 932-934.e3 (2021).

71. Hashash, J. G., Francis, F. F. & Farraye, F. A. Diagnosis and Management of Immune Checkpoint Inhibitor Colitis. *Gastroenterol Hepatol (N Y)* **17**, 358–366 (2021).

72. Chang, J. T. Pathophysiology of Inflammatory Bowel Diseases. *New England Journal of Medicine* **383**, 2652–2664 (2020).

73. Corridoni, D. *et al.* Single-cell atlas of colonic CD8 + T cells in ulcerative colitis. *Nature Medicine* **26**, 1480–1490 (2020).

74. Fawkner-Corbett, D. *et al.* Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell* **0**, (2021).

75. Kinchen, J. *et al.* Structural Remodeling of the Human Colonic Mesenchyme in Inflammatory Bowel Disease. *Cell* **175**, 372-386.e17 (2018).

76. Parikh, K. *et al.* Colonic epithelial cell diversity in health and inflammatory bowel disease. *Nature* **567**, 49–55 (2019).

77. Smillie, C. S. *et al.* Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **178**, 714-730.e22 (2019).

78. Luoma, A. M. *et al.* Molecular Pathways of Colon Inflammation Induced by Cancer Immunotherapy. *Cell* (2020) doi:10.1016/j.cell.2020.06.001.

79. Sasson, S. C. *et al.* Interferon-Gamma–Producing CD8+ Tissue Resident Memory T Cells Are a Targetable Hallmark of Immune Checkpoint Inhibitor–Colitis. *Gastroenterology* **0**, (2021).

80. Martin, J. C. *et al.* Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. *Cell* **178**, 1493-1508.e20 (2019).

81. Szabo, P. A. *et al.* Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nat Commun* **10**, 4706 (2019).

82. Keir, M. E., Butte, M. J., Freeman, G. J. & Sharpe, A. H. PD-1 and Its Ligands in Tolerance and Immunity. *Annual Review of Immunology* **26**, 677–704 (2008).

83. Boland, B. S. *et al.* Heterogeneity and clonal relationships of adaptive immune cells in ulcerative colitis revealed by single-cell analyses. *Science Immunology* (2020).

84. Roosenboom, B. *et al.* Intestinal CD103+CD4+ and CD103+CD8+ T-Cell Subsets in the Gut of Inflammatory Bowel Disease Patients at Diagnosis and During Follow-up. *Inflammatory Bowel Diseases* **25**, 1497–1509 (2019).

85. Kaech, S. M. *et al.* Selective expression of the interleukin 7 receptor identifies effector CD8 T cells that give rise to long-lived memory cells. *Nat Immunol* **4**, 1191–1198 (2003).

86. Yamauchi, T. *et al.* T-cell CX3CR1 expression as a dynamic blood-based biomarker of response to immune checkpoint inhibitors. *Nat Commun* **12**, 1402 (2021).

87. Baeyens, A., Fang, V., Chen, C. & Schwab, S. R. Exit Strategies: S1P Signaling and T Cell Migration. *Trends in Immunology* **36**, 778–787 (2015).

88. Martin, M. D. & Badovinac, V. P. Defining Memory CD8 T Cell. *Frontiers in Immunology* **9**, (2018).

89. Huang, H., Wang, C., Rubelt, F., Scriba, T. J. & Davis, M. M. Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nature Biotechnology* **38**, 1194–1202 (2020).

90. Caushi, J. X. *et al.* Transcriptional programs of neoantigen-specific TIL in anti-PD-1-treated lung cancers. *Nature* **596**, 126–132 (2021).

91. Andrews, T. S., Kiselev, V. Y., McCarthy, D. & Hemberg, M. Tutorial: guidelines for the computational analysis of single-cell RNA se-

quencing data. *Nature Protocols* 1–9 (2020) doi:10.1038/s41596-020-00409-w.

92. Lun, A. T. L. *et al.* EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biology* **20**, 63 (2019).

93. Griffiths, J. A., Richard, A. C., Bach, K., Lun, A. T. L. & Marioni, J. C. Detection and removal of barcode swapping in single-cell RNA-seq data. *Nat Commun* **9**, 2667 (2018).

94. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411–420 (2018).

95. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2020).

96. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**, 979–982 (2017).

97. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat Biotechnol* **40**, 245–253 (2022).

98. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLOS ONE* **5**, e12776 (2010).

99. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083–1086 (2017).

100. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat Commun* **12**, 1088 (2021).

101. Song, L. *et al.* TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. *Nat Methods* **18**, 627–630 (2021).

102. Shugay, M. *et al.* VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLOS Computational Biology* **11**, e1004503 (2015).

103. Lareau, C. A., Ma, S., Duarte, F. M. & Buenrostro, J. D. Inference and effects of barcode multiplets in droplet-based single-cell assays. *Nat Commun* **11**, 866 (2020).

104. Lee, C. H. & Koohy, H. In silico identification of vaccine targets for 2019-nCoV. *F1000Res* **9**, 145 (2020).

105. Caramelo, F., Ferreira, N. & Oliveiros, B. Estimation of risk factors for COVID-19 mortality - preliminary results. *medRxiv* 2020.02.24.20027268 (2020) doi:10.1101/2020.02.24.20027268.

106. Dowd, J. B. *et al.* Demographic science aids in understanding the spread and fatality rates of COVID-19. 18.

107. Tay, M. Z., Poh, C. M., Rénia, L., MacAry, P. A. & Ng, L. F. P. The trinity of COVID-19: immunity, inflammation and intervention. *Nat Rev Immunol* (2020) doi:10.1038/s41577-020-0311-8.

108. Verity, R. *et al.* Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases* S1473309920302437 (2020) doi:10.1016/S1473-3099(20)30243-7.

109. Epstein, S. L. & Price, G. E. Cross-protective immunity to influenza A viruses. *Expert Rev Vaccines* **9**, 1325–1341 (2010).

110. Frank, S. A. *Immunological Variability of Hosts. Immunology and Evolution of Infectious Disease* (Princeton University Press, 2002).

111. Auladell, M. *et al.* Recalling the Future: Immunological Memory Toward Unpredictable Influenza Viruses. *Front. Immunol.* **10**, (2019).

112. Nickbakhsh, S. *et al. Epidemiology of seasonal coronaviruses: Establishing the context for COVID-19 emergence.* http://medrxiv.org/lookup/doi/10.1101/2020.03.18.20037101 (2020) doi:10.1101/2020.03.18.20037101.

113. Khan, S. *et al. Analysis of Serologic Cross-Reactivity Between Common Human Coronaviruses and SARS-CoV-2 Using Coronavirus Antigen Microarray.* http://biorxiv.org/lookup/doi/10.1101/2020.03.24.006544 (2020) doi:10.1101/2020.03.24.006544.

114. Cheng, M. *et al.* Cross-reactivity of antibody against SARS-coronavirus nucleocapsid protein with IL-11. *Biochemical and Biophysical Research Communications* **338**, 1654–1660 (2005).

115. Lim, Y. X., Ng, Y. L., Tam, J. P. & Liu, D. X. Human Coronaviruses: A Review of Virus–Host Interactions. *Diseases* **4**, (2016).

116. Common Human Coronaviruses | CDC. https://www.cdc.gov/coronavirus/general-information.html (2020).

117. Buckley, P. R. *et al.* HLA-dependent variation in SARS-CoV-2 CD8 + T cell cross-reactivity with human coronaviruses. *Immunology* **166**, 78–103 (2022).

118. Maiers, M., Gragert, L. & Klitz, W. High-resolution HLA alleles and haplotypes in the United States population. *Human Immunology* **68**, 779–788 (2007).

119. Jurtz, V. *et al.* NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *The Journal of Immunology* **199**, 3360–3368 (2017).

120. Ogishi, M. & Yotsuyanagi, H. Quantitative Prediction of the Landscape of T Cell Epitope Immunogenicity in Sequence Space. *Front. Immunol.* **10**, 827 (2019).

121. SARS-CoV-2 T-cell epitopes define heterologous and COVID-19-induced T-cell recog- nition (2020) doi:10.21203/rs.3.rs-35331/v1.

122. Nolan, S. *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. *Res Sq* rs.3.rs-51964 (2020) doi:10.21203/rs.3.rs-51964/v1.

123. Francis, J. M. *et al.* Allelic variation in class I HLA determines CD8+ T cell repertoire shape and cross-reactive memory responses to SARS-CoV-2. *Science Immunology* **7**, eabk3070 (2021).

124. Frankild, S., Boer, R. J. de, Lund, O., Nielsen, M. & Kesmir, C. Amino Acid Similarity Accounts for T Cell Cross-Reactivity and for "Holes" in the T Cell Repertoire. *PLOS ONE* **3**, e1831 (2008).

125. Petrova, G., Ferrante, A. & Gorski, J. Cross-Reactivity of T Cells and Its Role in the Immune System. *Crit Rev Immunol* **32**, 349–372 (2012).

126. Rudd, P. M., Elliott, T., Cresswell, P., Wilson, I. A. & Dwek, R. A. Glycosylation and the Immune System. *Science* **291**, 2370–2376 (2001).

127. How COVID-19 kills: Explaining the mechanisms, symptoms, and diagnosis of the new coronavirus. *ScienceDaily* https://www.sciencedaily.com/releases/2020/05/200513081810.h

128. Bojkova, D. *et al.* Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature* **583**, 469–472 (2020).

129. Tsao, Y.-P. *et al.* HLA-A∗0201 T-cell

epitopes in severe acute respiratory syndrome (SARS) coronavirus nucleocapsid and spike proteins. *Biochem Biophys Res Commun* **344**, 63–71 (2006).

130. Zhou, M. *et al.* Screening and Identification of Severe Acute Respiratory Syndrome-Associated Coronavirus-Specific CTL Epitopes. *The Journal of Immunology* **177**, 2138–2145 (2006).

131. Kundu, R. *et al.* Cross-reactive memory T cells associate with protection against SARS-CoV-2 infection in COVID-19 contacts. *Nat Commun* **13**, 80 (2022).

132. Jia, L. *et al.* Pre-existing antibodies targeting a linear epitope on SARS-CoV-2 S2 cross-reacted with commensal gut bacteria and shaped vaccine induced immunity. 2021.07.13.21260404 Preprint at https://doi.org/10.1101/2021.07.13.21260404 (2022).

133. Su, W. *et al.* The Association Between Previous Influenza Vaccination and COVID-19 Infection Risk and Severity: A Systematic Review and Meta-analysis. *American Journal of Preventive Medicine* **0**, (2022).

134. Adams, J. J. *et al.* Structural interplay between germline interactions and adaptive recognition determines the bandwidth of TCR-peptide-MHC cross-reactivity. *Nature Immunology* **17**, 87–94 (2016).

135. Riley, T. P. & Baker, B. M. The intersection of affinity and specificity in the development and optimization of T cell receptor based therapeutics. *Seminars in Cell & Developmental Biology* **84**, 30–41 (2018).

136. Ekeruche-Makinde, J. *et al.* Peptide length determines the outcome of TCR/peptide-MHCI engagement. *Blood* **121**, 1112–1123 (2013).

137. Kawashima, S. & Kanehisa, M. AAindex: Amino Acid index database. *Nucleic Acids Res* **28**, 374–374 (2000).

138. Gao, A. *et al.* Predicting the Immunogenicity of T cell epitopes: From HIV to SARS-CoV-2. 2020.05.14.095885 Preprint at https://doi.org/10.1101/2020.05.14.095885 (2020).

139. Song, I. *et al.* Broad TCR repertoire and diverse structural solutions for recognition of an immunodominant CD8 + T cell epitope. *Nature Structural & Molecular Biology* **24**, 395–406 (2017).

140. Antunes, D. A. *et al.* Interpreting T-Cell Cross-reactivity through Structure: Implications for TCR-Based Cancer Immunotherapy. *Front. Immunol.* **8**, (2017).

141. Cuendet, M. A., Zoete, V. & Michielin, O. How T cell receptors interact with peptide-MHCs: A multiple steered molecular dynamics study. *Proteins: Structure, Function, and Bioinformatics* **79**, 3007–3024 (2011).

142. Spear, T. T., Evavold, B. D., Baker, B. M. & Nishimura, M. I. Understanding TCR affinity, antigen specificity, and cross-reactivity to improve TCR gene-modified T cells for cancer immunotherapy. *Cancer Immunol Immunother* **68**, 1881–1889 (2019).

143. Stone, J. D., Chervin, A. S. & Kranz, D. M. T-cell receptor binding affinities and kinetics: impact on T-cell activity and specificity. *Immunology* **126**, 165–176 (2009).

144. Krogsgaard, M. *et al.* Evidence that Structural Rearrangements and/or Flexibility during TCR Binding Can Contribute to T Cell Activation. *Molecular Cell* **12**, 1367–1378 (2003).

145. Degano, M. *et al.* A Functional Hot Spot for Antigen Recognition in a Superagonist

TCR/MHC Complex. *Immunity* **12**, 251–261 (2000).

146. Kalergis, A. M. & Nathenson, S. G. Altered Peptide Ligand-Mediated TCR Antagonism Can Be Modulated by a Change in a Single Amino Acid Residue Within the CDR3β of an MHC Class I-Restricted TCR. *The Journal of Immunology* **165**, 280–285 (2000).

147. Liu, B., Chen, W., Evavold, B. D. & Zhu, C. Accumulation of Dynamic Catch Bonds between TCR and Agonist Peptide-MHC Triggers T Cell Signaling. *Cell* **157**, 357–368 (2014).

148. Lyons, G. E. *et al.* Influence of Human CD8 on Antigen Recognition by T-Cell Receptor–Transduced Cells. *Cancer Res* **66**, 11455–11461 (2006).

149. Moore, T. V. *et al.* Relationship between CD8-dependent antigen recognition, T cell functional avidity, and tumor cell recognition. *Cancer Immunol Immunother* **58**, 719 (2008).

150. Pryshchep, S., Zarnitsyna, V. I., Hong, J., Evavold, B. D. & Zhu, C. Accumulation of Serial Forces on TCR and CD8 Frequently Applied by Agonist Antigenic Peptides Embedded in MHC Molecules Triggers Calcium in T Cells. *The Journal of Immunology* **193**, 68–76 (2014).

151. Spear, T. T., Nagato, K. & Nishimura, M. I. Strategies to genetically engineer T cells for cancer immunotherapy. *Cancer Immunol Immunother* **65**, 631–649 (2016).

152. Spear, T. T. *et al.* Critical biological parameters modulate affinity as a determinant of function in T-cell receptor gene-modified T-cells. *Cancer Immunol Immunother* **66**, 1411–1424 (2017).

153. Spear, T. T. *et al.* Altered Peptide Ligands Impact the Diversity of Polyfunctional Phenotypes in T Cell Receptor Gene-Modified T Cells. *Molecular Therapy* **26**, 996–1007 (2018).

154. Thomson, C. T., Kalergis, A. M., Sacchettini, J. C. & Nathenson, S. G. A Structural Difference Limited to One Residue of the Antigenic Peptide Can Profoundly Alter the Biological Outcome of the TCR-Peptide/MHC Class I Interaction. *The Journal of Immunology* **166**, 3994–3997 (2001).

155. Tsuchiya, Y., Namiuchi, Y., Wako, H. & Tsurui, H. A study of CDR3 loop dynamics reveals distinct mechanisms of peptide recognition by T-cell receptors exhibiting different levels of cross-reactivity. *Immunology* **153**, 466–478 (2018).

156. Valitutti, S., Müller, S., Cella, M., Padovan, E. & Lanzavecchia, A. Serial triggering of many T-cell receptors by a few peptide–MHC complexes. *Nature* **375**, 148–151 (1995).

157. Aleksic, M. *et al.* Dependence of T Cell Antigen Recognition on T Cell Receptor-Peptide MHC Confinement Time. *Immunity* **32**, 163–174 (2010).

158. Leaver-Fay, A. *et al.* Chapter Six - Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement. in *Methods in Enzymology* (ed. Keating, A. E.) vol. 523 109–143 (Academic Press, 2013).

159. Riley, T. P. *et al.* A generalized framework for computational design and mutational scanning of T-cell receptor binding interfaces. *Protein Eng Des Sel* **29**, 595–606 (2016).

160. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).

161. Leem, J., de Oliveira, S. H. P., Krawczyk, K. & Deane, C. M. STCRDab: the structural T-cell receptor database. *Nucleic Acids Res* **46**, D406–D412 (2018).

162. Gowthaman, R. & Pierce, B. G.

TCR3d: The T cell receptor structural repertoire database. *Bioinformatics* **35**, 5323–5325 (2019).

163. Desmet, J., Wilson, I. A., Joniau, M., De Maeyer, M. & Lasters, I. Computation of the binding of fully flexible peptides to proteins with flexible side chains. *The FASEB Journal* **11**, 164–172 (1997).

164. Leach, A. R. Ligand docking to proteins with discrete side-chain flexibility. *Journal of Molecular Biology* **235**, 345–356 (1994).

165. Bui, H.-H., Schiewe, A. J., Grafenstein, H. von & Haworth, I. S. Structural prediction of peptides binding to MHC class I molecules. *Proteins: Structure, Function, and Bioinformatics* **63**, 43–52 (2006).

166. Fagerberg, T., Cerottini, J.-C. & Michielin, O. Structural Prediction of Peptides Bound to MHC Class I. *Journal of Molecular Biology* **356**, 521–546 (2006).

167. Monje-Galvan, V., Warburton, L. & Klauda, J. B. Setting Up All-Atom Molecular Dynamics Simulations to Study the Interactions of Peripheral Membrane Proteins with Model Lipid Bilayers. in *Intracellular Lipid Transport: Methods and Protocols* (ed. Drin, G.) 325–339 (Springer, 2019). doi:10.1007/978-1-4939-9136-5_22.

168. Perilla, J. R., Hadden, J. A., Goh, B. C., Mayne, C. G. & Schulten, K. All-Atom Molecular Dynamics of Virus Capsids as Drug Targets. *J. Phys. Chem. Lett.* **7**, 1836–1844 (2016).

169. Park, M.-S., Park, S. Y., Miller, K. R., Collins, E. J. & Lee, H. Y. Accurate structure prediction of peptide–MHC complexes for identifying highly immunogenic antigens. *Molecular Immunology* **56**, 81–90 (2013).

170. Kish-Catalone, T. M., Lu, W., Gallo, R. C. & DeVico, A. L. Preclinical Evaluation of Synthetic 2 RANTES as a Candidate Vaginal Microbicide To Target CCR5. *Antimicrobial Agents and Chemotherapy* **50**, 1497–1509 (2006).

171. Knapp, B. *et al.* pyHVis3D: visualising molecular simulation deduced H-bond networks in 3D: application to T-cell receptor interactions. *Bioinformatics* **34**, 1941–1943 (2018).

172. Zhang, H. *et al.* The contribution of major histocompatibility complex contacts to the affinity and kinetics of T cell receptor binding. *Scientific Reports* **6**, 1–11 (2016).

173. Wan, S., Knapp, B., Wright, D. W., Deane, C. M. & Coveney, P. V. Rapid, Precise, and Reproducible Prediction of Peptide–MHC Binding Affinities from Molecular Dynamics That Correlate Well with Experiment. *J. Chem. Theory Comput.* **11**, 3346–3356 (2015).

174. Mendes, M. F. A., Antunes, D. A., Rigo, M. M., Sinigaglia, M. & Vieira, G. F. Improved structural method for T-cell cross-reactivity prediction. *Molecular Immunology* **67**, 303–310 (2015).

175. Pierce, B. G. & Weng, Z. A flexible docking approach for prediction of T cell receptor–peptide–MHC complexes. *Protein Science* **22**, 35–46 (2013).

176. Riley, T. P. *et al.* Structure Based Prediction of Neoantigen Immunogenicity. *Front. Immunol.* **10**, (2019).

177. Tung, C.-W., Ziehm, M., Kämper, A., Kohlbacher, O. & Ho, S.-Y. POPISK: T-cell reactivity prediction using support vector machines and string kernels. *BMC Bioinformatics* **12**, 446 (2011).

178. Assarsson, E. *et al.* A quantitative analysis of the variables affecting the repertoire of T cell specificities recognized after vaccinia virus

infection. *J. Immunol.* **178**, 7890–7901 (2007).

179. Kotturi, M. F. *et al.* A Multivalent and Cross-Protective Vaccine Strategy against Arenaviruses Associated with Human Disease. *PLOS Pathogens* **5**, e1000695 (2009).

180. Karosiene, E., Lundegaard, C., Lund, O. & Nielsen, M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* **64**, 177–186 (2012).

181. Jørgensen, K. W., Rasmussen, M., Buus, S. & Nielsen, M. NetMHCstab – predicting stability of peptide–MHC-I complexes; impacts for cytotoxic T lymphocyte epitope discovery. *Immunology* **141**, 18–26 (2014).

182. Chowell, D. *et al.* TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *PNAS* **112**, E1754–E1762 (2015).

183. Łuksza, M. *et al.* A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* **551**, 517–520 (2017).

184. Bjerregaard, A.-M. *et al.* An Analysis of Natural T Cell Responses to Predicted Tumor Neoepitopes. *Front. Immunol.* **8**, (2017).

185. Shen, W.-J., Wong, H.-S., Xiao, Q.-W., Guo, X. & Smale, S. Towards a Mathematical Foundation of Immunology and Amino Acid Chains. *arXiv:1205.6031 [cs, q-bio, stat]* (2012).

186. Jurtz, V. I. *et al.* NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. *bioRxiv* 433706 (2018) doi:10.1101/433706.

187. Klinger, M. *et al.* Multiplex Identification of Antigen-Specific T Cell Receptors Using a Combination of Immune Assays and Immune Receptor Sequencing. *PLOS ONE* **10**, 1–21 (2015).

188. Smith, C. C. *et al.* Machine-Learning Prediction of Tumor Antigen Immunogenicity in the Selection of Therapeutic Epitopes. *Cancer Immunol Res* **7**, 1591–1604 (2019).

189. Schmidt, J. *et al.* Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. *Cell Reports Medicine* **2**, 100194 (2021).

190. Pickett, B. E. *et al.* ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Research* **40**, D593–D598 (2012).

191. Yadav, M. *et al.* Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515**, 572–576 (2014).

192. Saito, T. & Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE* **10**, e0118432 (2015).

193. Wu, J. *et al.* DeepHLApan: A Deep Learning Approach for Neoantigen Prediction Considering Both HLA-Peptide Binding and Immunogenicity. *Front. Immunol.* **0**, (2019).

194. Springer, I., Besser, H., Tickotsky-Moskovitz, N., Dvorkin, S. & Louzoun, Y. Prediction of specific TCR-peptide binding from large dictionaries of TCR-peptide pairs. *bioRxiv* 650861 (2020) doi:10.1101/650861.

195. Gaston, J. S., Rickinson, A. B. & Epstein, M. A. Cross-reactivity of self-HLA-restricted Epstein-Barr virus-specific cytotoxic T lymphocytes for allo-HLA determinants. *J Exp Med* **158**, 1804–1821 (1983).

196. Amir, A. L. *et al.* Allo-HLA reactivity of

virus-specific memory T cells is common. *Blood* **115**, 3146–3157 (2010).

197. van der Zwan, A. *et al.* Cross-Reactivity of Virus-Specific CD8+ T Cells Against Allogeneic HLA-C: Possible Implications for Pregnancy Outcome. *Front. Immunol.* **9**, (2018).

198. Raman, M. C. C. *et al.* Direct molecular mimicry enables off-target cardiovascular toxicity by an enhanced affinity TCR designed for cancer immunotherapy. *Scientific Reports* **6**, 1–10 (2016).

199. Ziegler, A., Müller, C. A., Böckmann, R. A. & Uchanska-Ziegler, B. Low-affinity peptides and T-cell selection. *Trends in Immunology* **30**, 53–60 (2009).

200. van der Merwe, P. A. & Dushek, O. Mechanisms for T cell receptor triggering. *Nature Reviews Immunology* **11**, 47–55 (2011).

201. Martinez, R. J. & Evavold, B. D. Lower Affinity T Cells are Critical Components and Active Participants of the Immune Response. *Front. Immunol.* **6**, (2015).

202. Martinez, R. J., Andargachew, R., Martinez, H. A. & Evavold, B. D. Low-affinity CD4+ T cells are major responders in the primary immune response. *Nature Communications* **7**, 13848 (2016).

203. Gee, M. H. *et al.* Stress-testing the relationship between T cell receptor/peptide-MHC affinity and cross-reactivity using peptide velcro. *PNAS* **115**, E7369–E7378 (2018).

204. Rudolph, M. G., Stanfield, R. L. & Wilson, I. A. How tcrs bind mhcs, peptides, and coreceptors. *Annu. Rev. Immunol.* **24**, 419–466 (2006).

205. Bjerregaard, A.-M., Nielsen, M., Hadrup, S. R., Szallasi, Z. & Eklund, A. C. MuPeXI: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunol Immunother* **66**, 1123–1130 (2017).

206. Chen, J.-L. *et al.* NY-ESO-1 specific antibody and cellular responses in melanoma patients primed with NY-ESO-1 protein in ISCOMATRIX and boosted with recombinant NY-ESO-1 fowlpox virus. *International Journal of Cancer* **136**, E590–E601 (2015).

207. Ali, M. *et al.* Induction of neoantigen-reactive T cells from healthy donors. *Nat Protoc* **14**, 1926–1943 (2019).

208. Rodenko, B. *et al.* Generation of peptide–MHC class I complexes through UV-mediated ligand exchange. *Nat Protoc* **1**, 1120–1132 (2006).

209. Chowell, D., Krishna, S., Blattman, J. & Anderson, K. Hydrophobicity is a hallmark of immunogenic MHC class I T cell epitopes (APP2P.111). *The Journal of Immunology* **192**, 43.12-43.12 (2014).

210. Richman, L. P., Vonderheide, R. H. & Rech, A. J. Neoantigen Dissimilarity to the Self-Proteome Predicts Immunogenicity and Response to Immune Checkpoint Blockade. *Cell Systems* **9**, 375-382.e4 (2019).

211. Devlin, J. R. *et al.* Structural dissimilarity from self drives neoepitope escape from immune tolerance. *Nature Chemical Biology* **16**, 1269–1276 (2020).

212. Gee, M. H. *et al.* Antigen Identification for Orphan T Cell Receptors Expressed on Tumor-Infiltrating Lymphocytes. *Cell* **172**, 549-563.e16 (2018).

213. Glanville, J. *et al.* Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).

214. Sidhom, J.-W., Larman, H. B., Pardoll, D. M. & Baras, A. S. DeepTCR is a deep learn-

ing framework for revealing sequence concepts within T-cell repertoires. *Nature Communications* **12**, 1605 (2021).

215. Croft, N. P. *et al.* Most viral peptides displayed by class I MHC on infected cells are immunogenic. *Proc Natl Acad Sci USA* **116**, 3112–3117 (2019).

216. Moutaftsi, M. *et al.* A consensus epitope prediction approach identifies the breadth of murine T CD8+ -cell responses to vaccinia virus. *Nat Biotechnol* **24**, 817–819 (2006).

217. Ishizuka, J. *et al.* Quantitating T Cell Cross-Reactivity for Unrelated Peptide Antigens. *The Journal of Immunology* **183**, 4337–4345 (2009).

218. Moutaftsi, M. *et al.* Correlates of protection efficacy induced by vaccinia virus-specific CD8+ T cell epitopes in the murine intranasal challenge model. *Eur J Immunol* **39**, 717–722 (2009).

219. Tscharke, D. C. *et al.* Identification of poxvirus CD8+ T cell determinants to enable rational design and characterization of smallpox vaccines. *Journal of Experimental Medicine* **201**, 95–104 (2004).

220. Oseroff, C. *et al.* HLA class I-restricted responses to vaccinia recognize a broad array of proteins mainly involved in virulence and viral gene regulation. *PNAS* **102**, 13980–13985 (2005).

221. Gilchuk, P. *et al.* Discovering naturally processed antigenic determinants that confer protective T cell immunity. *J Clin Invest* **123**, 1976–1987 (2013).

222. Meyer, V. S. *et al.* Long-Term Immunity against Actual Poxviral HLA Ligands as Identified by Differential Stable Isotope Labeling. *The Journal of Immunology* **181**, 6371–6383 (2008).

223. Kotturi, M. F. *et al.* Of mice and humans: how good are HLA transgenic mice as a model of human immune responses? *Immunome Res* **5**, 3 (2009).

224. Capietto, A.-H. *et al.* Mutation position is an important determinant for predicting cancer neoantigens. *J Exp Med* **217**, (2020).

225. Koncz, B. *et al.* Self-mediated positive selection of T cells sets an obstacle to the recognition of nonself. *PNAS* **118**, (2021).

226. Lee, C. H., Antanaviciute, A., Buckley, P. R., Simmons, A. & Koohy, H. To what extent does MHC binding translate to immunogenicity in humans? *ImmunoInformatics* **3–4**, 100006 (2021).

227. Osorio, D., Rondón-Villarreal, P. & Torres, R. Peptides: A Package for Data Mining of Antimicrobial Peptides. *The R Journal* **7**, 4 (2015).

228. Krammer, F. SARS-CoV-2 vaccines in development. *Nature* **586**, 516–527 (2020).

229. Ott, P. A. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217–221 (2017).

230. Zacharakis, N. *et al.* Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer. *Nature Medicine* **24**, 724–730 (2018).

231. Rubelt, F. *et al.* Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nat Immunol* **18**, 1274–1278 (2017).

232. Jensen, K. K. *et al.* Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* **154**, 394–406 (2018).

233. Racle, J. *et al.* Robust prediction of HLA

class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat Biotechnol* **37**, 1283–1286 (2019).

234. Lu, T. *et al.* Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nat Mach Intell* 1–12 (2021) doi:10.1038/s42256-021-00383-2.

235. Weber, A., Born, J. & Rodriguez Martínez, M. TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics* **37**, i237–i244 (2021).

236. Kalyan, K. S., Rajasekharan, A. & Sangeetha, S. AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing. Preprint at https://doi.org/10.48550/arXiv.2108.05542 (2021).

237. Roshanzamir, A., Aghajan, H. & Soleymani Baghshah, M. Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech. *BMC Medical Informatics and Decision Making* **21**, 92 (2021).

238. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* http://arxiv.org/abs/1810.04805 (2019) doi:10.48550/arXiv.1810.04805.

239. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS* **118**, (2021).

240. Rao, R. *et al.* Evaluating Protein Transfer Learning with TAPE. *Adv Neural Inf Process Syst* **32**, 9689–9701 (2019).

241. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* **16**, 1315–1322 (2019).

242. Elnaggar, A. *et al. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning.* http://biorxiv.org/lookup/doi/10.1101/2020.07.12.199554 (2020) doi:10.1101/2020.07.12.199554.

243. Ruppert, J. *et al.* Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. *Cell* **74**, 929–937 (1993).

244. Cho, K. *et al.* Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Preprint at https://doi.org/10.48550/arXiv.1406.1078 (2014).

245. Kim, Y. Convolutional Neural Networks for Sentence Classification. Preprint at https://doi.org/10.48550/arXiv.1408.5882 (2014).

246. Wooldridge, L. *et al.* CD8 Controls T Cell Cross-Reactivity. *The Journal of Immunology* **185**, 4625–4632 (2010).

247. Bentzen, A. K. & Hadrup, S. R. T-cell-receptor cross-recognition and strategies to select safe T-cell receptors for clinical translation. *Immuno-Oncology Technology* **2**, 1–10 (2019).

248. Schmidt, D. & Heckendorf, C. ngram: Fast n-Gram 'Tokenization'. (2021).

249. Hendrycks, D. & Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. Preprint at https://doi.org/10.48550/arXiv.1610.02136 (2018).

250. Arora, U., Huang, W. & He, H. Types of Out-of-Distribution Texts and How to Detect Them. in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* 10687–10701 (Asso-

ciation for Computational Linguistics, 2021). doi:10.18653/v1/2021.emnlp-main.835.

251. Zhou, C. & Paffenroth, R. C. Anomaly Detection with Robust Deep Autoencoders. in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 665–674 (Association for Computing Machinery, 2017). doi:10.1145/3097983.3098052.

252. Xia, J. *et al.* NEPdb: A Database of T-Cell Experimentally-Validated Neoantigens and Pan-Cancer Predicted Neoepitopes for Cancer Immunotherapy. *Frontiers in Immunology* **12**, (2021).

253. Tan, X. *et al.* dbPepNeo: a manually curated database for human tumor neoantigen peptides. *Database* **2020**, (2020).

254. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* **19**, 409–424 (2021).

255. Chen, K.-W. K., Tsung-Ning Huang, D. & Huang, L.-M. SARS-CoV-2 variants – Evolution, spike protein, and vaccines. *Biomedical Journal* **45**, 573–579 (2022).

256. Sigal, A. Milder disease with Omicron: is it the virus or the pre-existing immunity? *Nat Rev Immunol* **22**, 69–71 (2022).

257. Subissi, L. *et al.* An early warning system for emerging SARS-CoV-2 variants. *Nat Med* **28**, 1110–1115 (2022).

258. Dolton, G. *et al.* Emergence of immune escape at dominant SARS-CoV-2 killer T cell epitope. *Cell* **185**, 2936-2951.e19 (2022).

259. Stanevich, O. *et al. SARS-CoV-2 escape from cytotoxic T cells during long-term COVID-19.* https://www.researchsquare.com/article/rs-750741/v1 (2021) doi:10.21203/rs.3.rs-750741/v1.

260. Zinkernagel, R. M. & Doherty, P. C. Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. *Nature* **248**, 701–702 (1974).

261. La Gruta, N. L., Gras, S., Daley, S. R., Thomas, P. G. & Rossjohn, J. Understanding the drivers of MHC restriction of T cell receptors. *Nat Rev Immunol* **18**, 467–478 (2018).

262. Duan, F. *et al.* Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. *J Exp Med* **211**, 2231–2248 (2014).

263. Chen, L. & Flies, D. B. Molecular mechanisms of T cell co-stimulation and co-inhibition. *Nat Rev Immunol* **13**, 227–242 (2013).

264. Achar, S. R. *et al.* Universal antigen encoding of T cell activation from high-dimensional cytokine dynamics. *Science* **376**, 880–884 (2022).

265. Cameron, B. J. *et al.* Identification of a Titin-Derived HLA-A1–Presented Peptide as a Cross-Reactive Target for Engineered MAGE A3–Directed T Cells. *Science Translational Medicine* **5**, 197ra103-197ra103 (2013).

266. Takada, K. & Takahama, Y. Positive-selection-inducing self-peptides displayed by cortical thymic epithelial cells. *Adv Immunol* **125**, 87–110 (2015).

267. Leonard, J. D. *et al.* Identification of Natural Regulatory T Cell Epitopes Reveals Convergence on a Dominant Autoantigen. *Immunity* **47**, 107-117.e8 (2017).

268. Davis, M. M. Not-So-Negative Selection. *Immunity* **43**, 833–835 (2015).

269. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: Efficient

manipulation of biological strings. (2022) doi:10.18129/B9.bioc.Biostrings.

270. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On Calibration of Modern Neural Networks. Preprint at https://doi.org/10.48550/arXiv.1706.04599 (2017).

271. Bresciani, A. *et al.* T-cell recognition is shaped by epitope sequence conservation in the host proteome and microbiome. *Immunology* **148**, 34–39 (2016).

272. Ahmed, S. F., Quadeer, A. A. & McKay, M. R. Preliminary Identification of Potential Vaccine Targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. *Viruses* **12**, 254 (2020).

273. Nguyen, T. H. O. *et al.* CD8+ T cells specific for an immunodominant SARS-CoV-2 nucleocapsid epitope display high naive precursor frequency and TCR promiscuity. *Immunity* **54**, 1066-1082.e5 (2021).

274. Olafsdottir, T. A. *et al.* HLA alleles, disease severity, and age associate with T-cell responses following infection with SARS-CoV-2. *Commun Biol* **5**, 1–14 (2022).

275. Smatti, M. K. *et al.* Viruses and Autoimmunity: A Review on the Potential Interaction and Molecular Mechanisms. *Viruses* **11**, 762 (2019).

276. Fujinami, R. S., von Herrath, M. G., Christen, U. & Whitton, J. L. Molecular Mimicry, Bystander Activation, or Viral Persistence: Infections and Autoimmune Disease. *Clinical Microbiology Reviews* **19**, 80–94 (2006).

277. Venkatakrishnan, A. J. *et al.* Benchmarking evolutionary tinkering underlying human–viral molecular mimicry shows multiple host pulmonary–arterial peptides mimicked by SARS-CoV-2. *Cell Death Discov.* **6**, 1–14 (2020).

278. Nelde, A. *et al.* SARS-CoV-2-derived peptides define heterologous and COVID-19-induced T cell recognition. *Nature Immunology* 1–12 (2020) doi:10.1038/s41590-020-00808-x.

279. Baker, R. E. *et al.* Infectious disease in an era of global change. *Nat Rev Microbiol* **20**, 193–205 (2022).

280. Matzinger, P. & Bevan, M. J. Why do so many lymphocytes respond to major histocompatibility antigens? *Cellular Immunology* **29**, 1–5 (1977).

281. Sewell, A. K. Why must T cells be cross-reactive? *Nature Reviews Immunology* **12**, 669–677 (2012).

282. Cornberg, M. *et al.* Clonal Exhaustion as a Mechanism to Protect Against Severe Immunopathology and Death from an Overwhelming CD8 T Cell Response. *Front Immunol* **4**, (2013).

283. Pomés, A. & Schulten, V. Cross-reactivity in allergy: A double-edged sword. *Allergy* **75**, 9–11 (2020).

284. Selin, L. K. *et al.* CD8 memory T cells: cross-reactivity and heterologous immunity. *Seminars in Immunology* **16**, 335–347 (2004).

285. Welsh, R. M. & Fujinami, R. S. Pathogenic epitopes, heterologous immunity and vaccine design. *Nature Reviews Microbiology* **5**, 555–563 (2007).

286. Albert, L. J. & Inman, R. D. Molecular Mimicry and Autoimmunity. *New England Journal of Medicine* **341**, 2068–2074 (1999).

287. Oldstone, M. B. A. Molecular mimicry and immune-mediated diseases. *The FASEB*

Journal **12**, 1255–1265 (1998).

288. Pender, M. P., Csurhes, P. A., Burrows, J. M. & Burrows, S. R. Defective T-cell control of Epstein–Barr virus infection in multiple sclerosis. *Clin Transl Immunology* **6**, e126 (2017).

289. Wim Ang, C., Jacobs, B. C. & Laman, J. D. The Guillain–Barré syndrome: a true case of molecular mimicry. *Trends in Immunology* **25**, 61–66 (2004).

290. Wucherpfennig, K. W. & Strominger, J. L. Molecular mimicry in T cell-mediated autoimmunity: Viral peptides activate human T cell clones specific for myelin basic protein. *Cell* **80**, 695–705 (1995).

291. Zhao, Z.-S., Granucci, F., Yeh, L., Schaffer, P. A. & Cantor, H. Molecular Mimicry by Herpes Simplex Virus-Type 1: Autoimmune Disease After Viral Infection. *Science* **279**, 1344–1347 (1998).

292. Chen, A. T. *et al.* Loss of Anti-Viral Immunity by Infection with a Virus Encoding a Cross-Reactive Pathogenic Epitope. *PLOS Pathogens* **8**, e1002633 (2012).

293. Wlodarczyk, M. F., Kraft, A. R., Chen, H. D., Kenney, L. L. & Selin, L. K. Anti–IFN-γ and Peptide-Tolerization Therapies Inhibit Acute Lung Injury Induced by Cross-Reactive Influenza A–Specific Memory T Cells. *The Journal of Immunology* **190**, 2736–2746 (2013).

294. Bijen, H. M. *et al.* Preclinical Strategies to Identify Off-Target Toxicity of High-Affinity TCRs. *Molecular Therapy* **26**, 1206–1214 (2018).

295. Ekeruche-Makinde, J. *et al.* T-cell Receptor-optimized Peptide Skewing of the T-cell Repertoire Can Enhance Antigen Targeting. *J. Biol. Chem.* **287**, 37269–37281 (2012).

296. Kisielow, J., Obermair, F.-J. & Kopf, M. Deciphering CD4 + T cell specificity using novel MHC–TCR chimeric receptors. *Nature Immunology* **20**, 652–662 (2019).

297. Gotch, F., McMichael, A. & Rothbard, J. Recognition of influenza A matrix protein by HLA-A2-restricted cytotoxic T lymphocytes. Use of analogues to orientate the matrix peptide in the HLA-A2 binding site. *J Exp Med* **168**, 2045–2057 (1988).

298. Hu, N. *et al.* Highly conserved pattern of recognition of influenza A wild-type and variant CD8+ CTL epitopes in HLA-A2+ humans and transgenic HLA-A2+/H2 class I-deficient mice. *Vaccine* **23**, 5231–5244 (2005).

299. Petrova, G. V., Naumova, E. N. & Gorski, J. The Polyclonal CD8 T Cell Response to Influenza M158–66 Generates a Fully Connected Network of Cross-Reactive Clonotypes to Structurally Related Peptides: A Paradigm for Memory Repertoire Coverage of Novel Epitopes or Escape Mutants. *The Journal of Immunology* **186**, 6390–6397 (2011).

300. Haase, K., Raffegerst, S., Schendel, D. J. & Frishman, D. Expitope: a web server for epitope expression. *Bioinformatics* **31**, 1854–1856 (2015).

301. Jaravine, V., Mösch, A., Raffegerst, S., Schendel, D. J. & Frishman, D. Expitope 2.0: a tool to assess immunotherapeutic antigens for their potential cross-reactivity against naturally expressed proteins in human tissues. *BMC Cancer* **17**, (2017).

302. Dash, P. *et al.* Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).

303. Zhang, H. *et al.* Investigation of Antigen-Specific T-Cell Receptor Clusters in Human Cancers. *Clin Cancer Res* 1078-0432.CCR-19–3249 (2019) doi:10.1158/1078-0432.CCR-19-

3249.

304. Pogorelyy, M. V. & Shugay, M. A Framework for Annotation of Antigen Specificities in High-Throughput T-Cell Repertoire Sequencing Studies. *Front. Immunol.* **10**, (2019).

305. Thomas, N. *et al.* Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics* **30**, 3181–3188 (2014).

306. Cinelli, M. *et al.* Feature selection using a one dimensional naïve Bayes' classifier increases the accuracy of support vector machine classification of CDR3 repertoires. *Bioinformatics* **33**, 951–955 (2017).

307. Priel, A., Gordin, M., Philip, H., Zilberberg, A. & Efroni, S. Network Representation of T-Cell Repertoire— A Novel Tool to Analyze Immune Response to Cancer Formation. *Front. Immunol.* **9**, (2018).

308. DeWitt, W. S., III *et al.* Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *eLife* **7**, e38358 (2018).

309. Meysman, P. *et al.* On the viability of unsupervised T-cell receptor sequence clustering for epitope preference. *Bioinformatics* **35**, 1461–1468 (2019).

310. Thakkar, N. & Bailey-Kellogg, C. Balancing sensitivity and specificity in distinguishing TCR groups by CDR sequence similarity. *BMC Bioinformatics* **20**, 241 (2019).

311. Montemurro, A. *et al.* NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. *Commun Biol* **4**, 1–13 (2021).

312. Gielis, S. *et al.* TCRex: a webtool for the prediction of T-cell receptor sequence epitope specificity. *bioRxiv* 373472 (2018) doi:10.1101/373472.

313. 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. *Frontiers in Immunology* **12**, (2021).

314. Bradley, P. & Thomas, P. G. Using T Cell Receptor Repertoires to Understand the Principles of Adaptive Immune Recognition. *Annual Review of Immunology* **37**, 547–570 (2019).

315. Riley, T. P. *et al.* T cell receptor cross-reactivity expanded by dramatic peptide–MHC adaptability. *Nat Chem Biol* **14**, 934–942 (2018).

Chapter 10. Copyright of figures and tables

The copyrights for the figures and tables in the thesis are as follows. All other figures have yet to be published.

The figures and tables were published in F1000Research, Immunology, Immunoinformatics, and Briefings in Bioinformatics. F1000Research and all Frontiers content are licenced under a CC-BY. Immunoninformatics is licenced under the CC BY-NC-ND 4.0 licence. The editor has granted permission for Briefings in Bioinformatics. The Copyright Clearance Centre has approved the request to reuse figures from Immunology under John Wiley & Sons – Books platform. Therefore, all figures have been redistributed with acknowledgements.

Chapter 1.

- Figure 1 has been created with BioRender.com using template created by Dr. Nabil Zeidan, PhD (Codesigner).

Chapter 4.

- Figure 1-2. **Lee CH**, Koohy H. In silico identification of vaccine targets for 2019-nCoV. F1000Res 9, 145 (2020)
- Figure 3-6. **Lee CH**, Pinho MP, Buckley PR, Woodhouse IB, Ogg G, Simmons A, Koohy H. Potential CD8+ T-cell Cross-Reactivity Against SARS-CoV-2 Conferred by Other Coronavirus Strains. Front. Immunol. 11, (2020).
- Figure 7. Buckley, PR, **Lee, CH**, Pinho MP., Babu, RO, Woo, J, Antanaviciute, A, Simmons, A, Ogg, G and Koohy, H., HLA-dependent variation in SARS-CoV-2 CD8+ T-cell cross-reactivity with human coronaviruses. Immunology (2022).
- Figure 8. **Lee CH**, Pinho MP, Buckley PR, Woodhouse IB, Ogg G, Simmons A, Koohy H. Potential CD8+ T-cell Cross-Reactivity Against SARS-CoV-2 Conferred by Other Coronavirus Strains. Front. Immunol. 11, (2020).
- Table 1. **Lee CH**, Koohy H. In silico identification of vaccine targets for 2019-nCoV. F1000Res 9, 145 (2020).

- Table 2. **Lee CH**, Pinho MP, Buckley PR, Woodhouse IB, Ogg G, Simmons A, Koohy H. Potential CD8+ T-cell Cross-Reactivity Against SARS-CoV-2 Conferred by Other Coronavirus Strains. Front. Immunol. 11, (2020).
- Table 4. **Lee CH**, Pinho MP, Buckley PR, Woodhouse IB, Ogg G, Simmons A, Koohy H. Potential CD8+ T-cell Cross-Reactivity Against SARS-CoV-2 Conferred by Other Coronavirus Strains. Front. Immunol. 11, (2020).

Chapter 5.

- Figure 1. **Lee CH**, Salio M, Napolitani G, Ogg G, Simmons A, Koohy H. Predicting Cross-Reactivity and Antigen Specificity of T-cell Receptors. Front. Immunol. 11, (2020).
- Figure 2-8. Buckley, PR*, **Lee, CH***, Ma, R*, Woodhouse I, Woo J, Tsvetkov VO, Shscherbinin BS, Antanaviciute A, Shugay M, Rei M, Simmons A, Koohy H. Evaluating performance of existing computational models in predicting CD8+ T-cell pathogenic epitopes and cancer neoantigens. Briefings in Bioinformatics. *These authors contributed equally.
- Table 1. **Lee CH**, Salio M, Napolitani G, Ogg G, Simmons A, Koohy H. Predicting Cross-Reactivity and Antigen Specificity of T-cell Receptors. Front. Immunol. 11, (2020).
- Table 2. Buckley, PR*, **Lee, CH***, Ma, R*, Woodhouse I, Woo J, Tsvetkov VO, Shscherbinin BS, Antanaviciute A, Shugay M, Rei M, Simmons A, Koohy H. Evaluating performance of existing computational models in predicting CD8+ T-cell pathogenic epitopes and cancer neoantigens. Briefings in Bioinformatics. *These authors contributed equally.

Chapter 6.

- Figure 1. **Lee CH**, Antanaviciute A., Buckley PR, Simmons A. & Koohy, H. To what extent does MHC binding translate to immunogenicity in humans? Immunoinformatic (2021)
- Figure 2-5. Buckley, PR*, **Lee, CH***, Ma, R*. Woodhouse I, Woo J, Tsvetkov VO, Shscherbinin BS, Antanaviciute A, Shugay M, Rei M, Simmons A, Koohy H.. Evaluating performance of existing computational models in predicting CD8+ T-cell pathogenic epitopes and cancer neoantigens. Briefings in Bioinformatics. *These authors contributed equally.
- Figure 6. **Lee CH**, Antanaviciute A., Buckley PR, Simmons A. & Koohy, H. To what extent does MHC binding translate to immunogenicity in humans? Immunoinformatic (2021)

Chapter 8.

- Figure 1. **Lee CH**, Salio M, Napolitani G, Ogg G, Simmons A, Koohy H. Predicting Cross-Reactivity and Antigen Specificity of T-cell Receptors. Front. Immunol. 11, (2020).

- Figure 4. **Lee CH**, Salio M, Napolitani G, Ogg G, Simmons A, Koohy H. Predicting Cross-Reactivity and Antigen Specificity of T-cell Receptors. Front. Immunol. 11, (2020).

Appendix

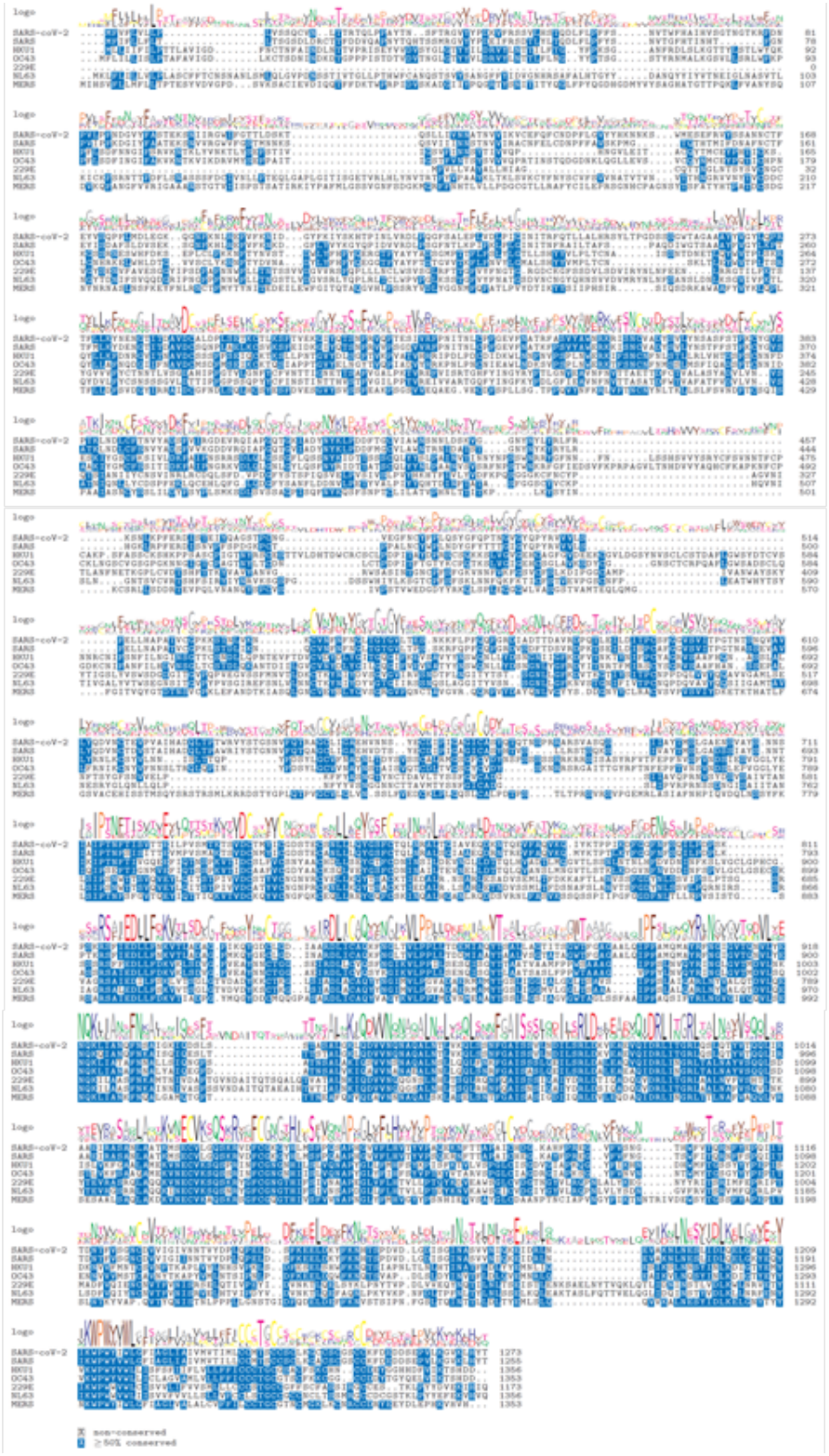
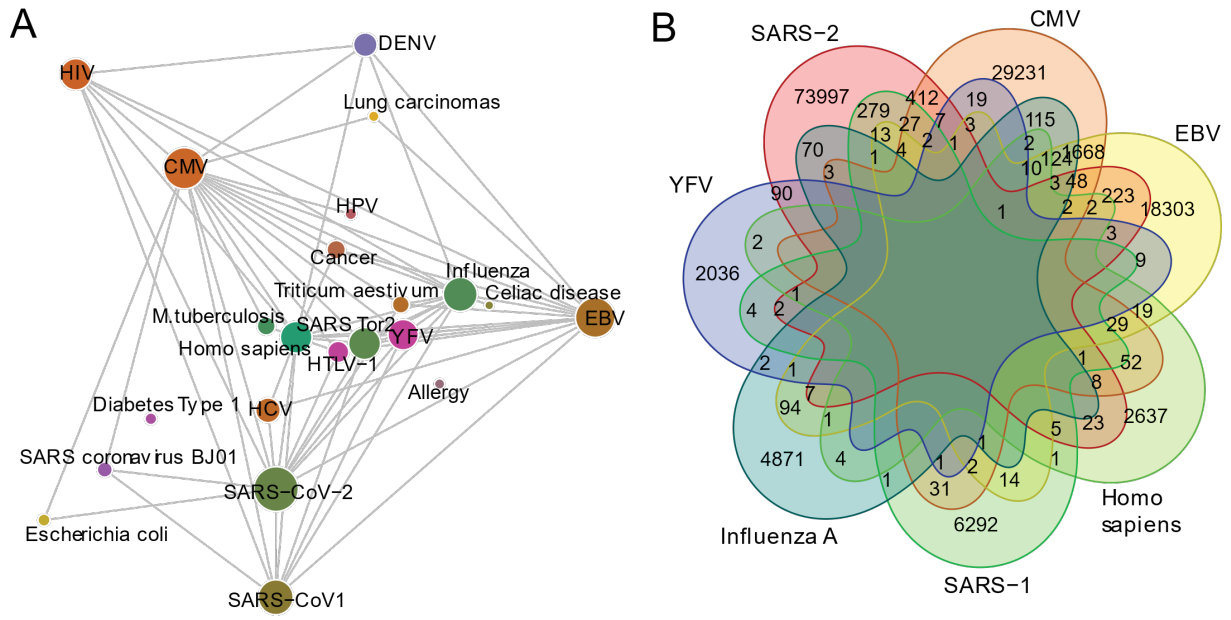


Figure S1: Alignment of spike protein sequences from different coronavirus strains

Supplementary Figure 2. Antigens recognised by cross-reactive TCRs.
Despite the substantial under-sampling of T-cell epitopes, we mapped the overlap of TCR repertoires recognizing epitopes from different species, and saw TCRs that recognized epitopes from EBV, CMV, influenza, SARS-CoV-2 and Homo sapiens were at the hubs to recognize epitope from other species (Supplementary Figure 2A). Among these, a few TCRs could recognize epitopes from 4-6 different species (Supplementary Figure 2B), such as ASSFSGNTGELF, ASSLGNTEAF, ASSLSGYEQY, ASSQDPSRTYEQY, ASSSGLAGGNEQF, ASSSTGGGEKDQPQH, CASSLGVAGTNTGELFF, CASSLMRGGTYNSPLHF, CASSPRDRERGEQYF, CASSQDPSRTYEQYF, CASSSTGGGEKDQPQHF, CSVDLEANYGYTF, SASPGDYEQY, SVDLEANYGYT, ASSPGQGYEQY, ASSLGGTGELF and ASSLAGGSYEQY, where ASSIRSSYEQY could recognize epitopes from all 6 species. Of note, the analysis should be interpreted with caution, as the dataset is heavily biased toward pathogen that have been extensively studied, such as CMV, EBV and influenza virus, and has limited pool of cross-reactive TCRs.
To examine sequence conservation of epitopes bound by the same cross-reactive TCRs, we selected the most cross-reactive and cross-species TCRs in the database (Supplementary Figure 2C). We observed that some TCRs had prominent recognition motifs (Supplementary Figure 2Di, 2Diii), while others were more promiscuous (Supplementary Figure 2Div). Notably, 12-mer and 14-mer epitopes recognized by the same TCR shared common patterns, implying that the same pattern was preserved across peptides of different lengths (Supplementary Figure 2Di-ii).



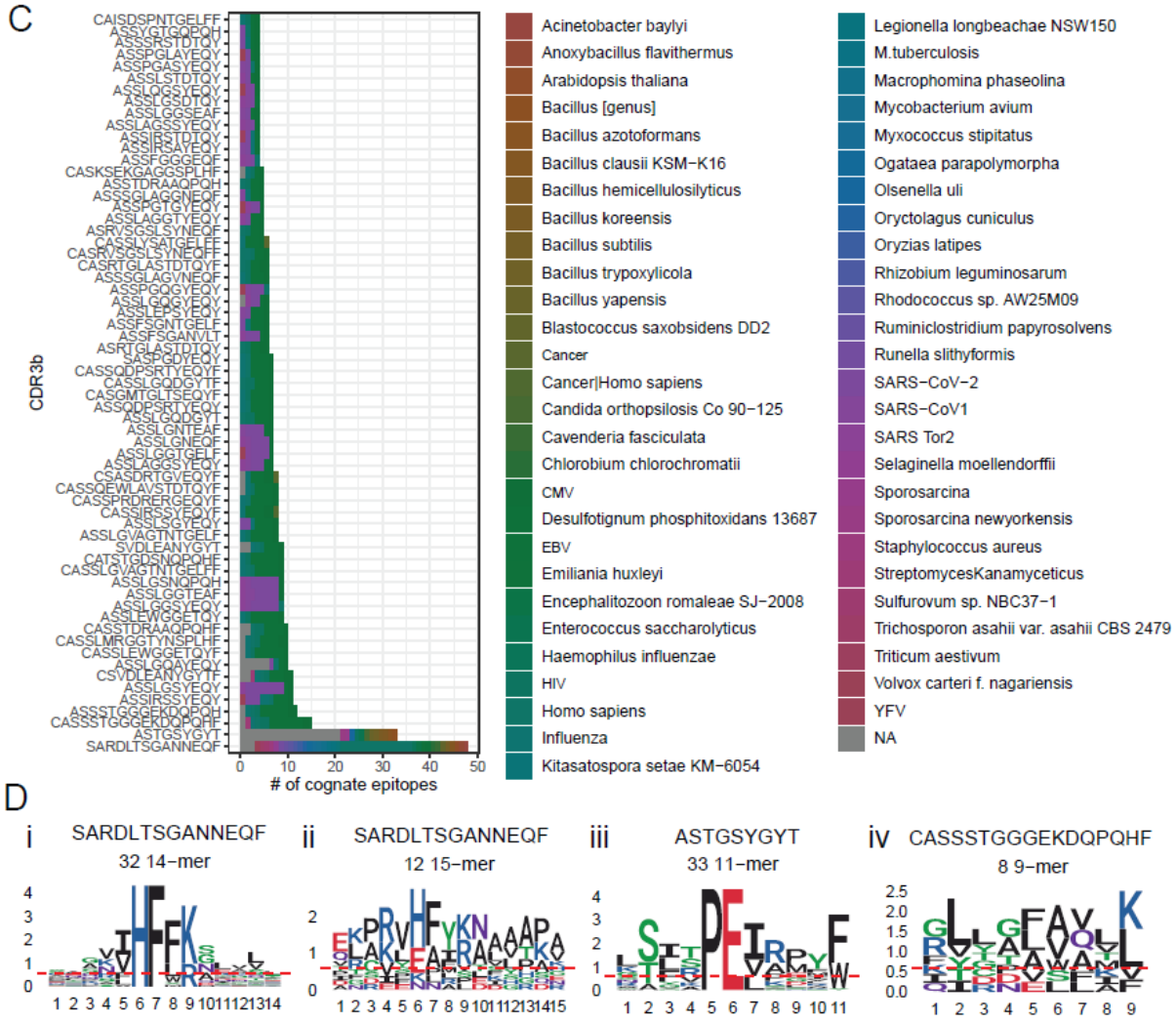


Figure S2: Cross-reactive TCRs recognizing epitopes derived from multiple species. A. Network graph illustrating TCR repertoire overlap that recognized epitopes across different species, where thickness of edges represent Morisitae's index and node represents number of unique epitopes for each species. B. Venn Diagram illustrating TCRs recognizing epitopes originated from 7 species, CMV, EBV, Human, SARS-CoV-1, SARS-CoV-2, influenza A and yellow fever virus. C. List of most cross-species TCR coloured by the antigen organism. D. Sequence logo of epitopes bound by the cross-species TCRs, where titles show CDR3b sequence, number of epitopes and length of peptides.

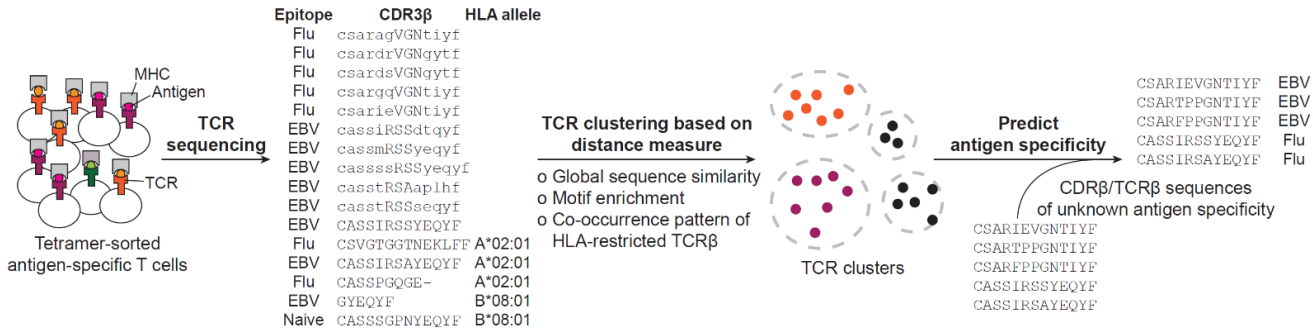


Figure S3: Workflow for predicting antigen specificity of TCRs. The tetramer-sorted antigen specific CDR3β or TCRβ are clustered by distance measure defined by either global sequence similarity, motif enrichment or sequence co-occurrence pattern. Then, each specificity clusters are investigated for their descriptive features, such as enrichment of common V-genes, CDR3 length, clonal expansions, and motif significance, to be considered in making the prediction of antigen specificity. Based on the collection of identified features, previously uncharacterized CDR3βs or TCRβs are predicted for their antigen specificity. The example sequences have been retrieved from Glanville *et al.*[213] and DeWitt *et al.*[308].

Table S1: Cluster summary of Tissue CD3+ cells. Legend: Tr: Trend, p<0.05 (*), p<0.01(**), p<0.001(***)

Cluster	Key Marker Genes	Key Antibodies	Notes	Condition Enrichment	TCR Analysis	Key TFs	References
Th17	*IL26*, *IL17A*, *IL22*, *KIT*, *IL23R*, *CCR6*, *CCL20*	CD4+		Enriched in HC, baseline Th17 cluster	BloodShare: 0.0197 % singleton: 0.86	*JUNB*, *ATF7*, *BATF3*, *FOXP3*, *RORC*, *STAT3*, *RORA*, *RARA*, *RUNX2*	(Capone and Volpe, 2020; Christie and Zhu, 2014; Jaeger et al., 2021; Martin et al., 2019; Smillie et al., 2019; Szabo et al., 2019; Yamazaki et al., 2017)
Th1	*IFNG*	CD4+			BloodShare: 0.0153 % singleton: 0.91	*STAT1*, *STAT4*, *TBX19*, *TBX21*, *GATA3*	(Christie and Zhu, 2014; Evans and Jenner, 2013; Oestreich and Weinmann, 2012; Zhou, 2015)
Naïve 2	*LEF1*, *CCR7*, *SELL*	CD4+			BloodShare: 0.00531 % singleton: 0.98	*RUNX2*, *TCF7*, *KLF2*, *FOXP1*, *LEF1*, *BACH2*, *IKZF1*	(Corridoni et al., 2020; Egawa et al., 2007; Feng et al., 2011; Willinger et al., 2006)
Tfh (2)	*LEF1*, *CCR7*, *SELL*, *CXCR5*, *TOX*, *MAF*, *PDCD1*	CD4+	CCR7+ Tfh subset		BloodShare: 0.00565 % singleton: 0.97	*TCF7*, *FOXP1*, *RUNX2*, *KLF3*, *LEF1*, *BACH2*, *IKZF1*, *ARID5A*, *MAF*	(Corridoni et al., 2020; Egawa et al., 2007; Feng et al., 2011; Willinger et al., 2006)

Table S1: Cluster summary of Tissue CD3+ cells. Legend: Tr: Trend, p<0.05 (*), p<0.01(**), p<0.001(***)

Cluster	Key Marker Genes	Key Antibodies	Notes	Condition Enrichment	TCR Analysis	Key TFs	References
Naïve CD8	*LEF1*, *CCR7*, *SELL*	CD8+	Naïve CD8+ counterpart of naïve CD4+ cells		BloodShare: 0.0196 % singleton: 0.98	*TBX19*, *TBX21*, *LEF1*, *BACH2*, *IKZF1*, *ID1*, *ZEB1*, *STAT3*	(Chen et al., 2018; Kaech and Cui, 2012)
GZMK+ Effector 2	*GZMK*, *IFNG*, *KLRG1*, *EOMES*, *GZMB*-low, *GZMH*-low	CD103- CD8+		Enriched in UC over CC	BloodShare: 0.162 singleton: 0.73	*EOMES*, *STAT3*, *BATF*, *TBX21*	(Corridoni et al., 2020)
TRM	*ITGAE*, *IL7R*	CD103+ CD8+	Tissue-resident memory cells	Depleted in UC but not CC	BloodShare: 0.0733 singleton: 0.65	*EGR1*, *EGR2*, *EGR3*, *STAT4*, *RUNX3*	(Behr et al., 2018; Corridoni et al., 2020)

Table S1: Cluster summary of Tissue CD3+ cells. Legend: Tr: Trend, p<0.05 (*), p<0.01(**), p<0.001(***)

Cluster	Key Marker Genes	Key Antibodies	Notes	Condition Enrichment	TCR Analysis	Key TFs	References
Gamma-delta 2	*TRDC*, *TYROBP*, *NCR3*, *CD160*, *FCER1G*	CD103+	Intra-epithelial, TYROBP+, natural IEL Phenotype. Enriched for gamma-delta T-cells, but also contain alpha-beta T-cells		BloodShare: 0.148 singleton: 0.52	*RORC*, *MAF*, *SOX4*, *TCF7*, *LEF1*, *TBX21*, *EOMES*	(Barros-Martins et al., 2016; Corridoni et al., 2020; Parker and Ciofani, 2020)
Gamma-delta 1	*TRDC*, *NCR3*, *NCR1*, *NCR2*, *CD160*, *FCER1G*	CD103+	Intra-epithelial, TYROBP-, induced IEL phenotype. Enriched for gamma-delta T-cells, but also contain alpha-beta T-cells		BloodShare: 0.0417 singleton: 0.67	*RORA*, *RORC*, *RUNX3*, *BATF3*, *TBX21*, *EOMES*, *SOX5*, *STAT3*, *STAT1*, *STAT5*	(Barros-Martins et al., 2016; Corridoni et al., 2020; Parker and Ciofani, 2020)

Table S1: Cluster summary of Tissue CD3+ cells. Legend: Tr: Trend, p<0.05 (*), p<0.01(**), p<0.001(***)

Cluster	Key Marker Genes	Key Antibodies	Notes	Condition Enrichment	TCR Analysis	Key TFs	References
ZNF683+	*ZNF683*, *ITGAE*, *CD160*, *ZEB2*, *GZMB*	CD103+	Re-activated tissueresident memory cells	Depleted in UC compared to HC Increased in CC	BloodShare: 0.0818 singleton: 0.51	*PRDM1*, *TBX21*, *EOMES*, *ZNF683*, *RUNX3* TF regulatory network analysis showed higher activity of T-cell terminal differentiation TFs for CC in *ZNF683* cells *ZEB1*, *GABPA*, *GABPB1*, *SPIB*, *PBX3*, *TAF7*, *PAX5*, *YBX1*, *ZNF639*	(Behr et al., 2018; Bhullar and Sollars, 2011; Caushi et al., 2021; Cobaleda et al., 2007; David-Fung et al., 2009; Gegonne et al., 2012; Guo et al., 2020; Kanaya et al., 2012; Kragten et al., 2018; Kumar et al., 2017; Luo et al., 2017; Parish and Kaech, 2009; Scott and Omilusik, 2019; Su et al., 1996; Vandereyken et al., 2020; Viganò et al., 2014; Willis et al.,

Table S1: Cluster summary of Tissue CD3+ cells. Legend: Tr: Trend, p<0.05 (*), p<0.01(**), p<0.001(***)

Cluster	Key Marker Genes	Key Antibodies	Notes	Condition Enrichment	TCR Analysis	Key TFs	References
CD8+ HAVCR2+	*HAVCR2*, *LAYN*, *EOMES*, *PDCD1*, *CTLA4*, *TNFRSF9*, *GZMH*, *GZMB*	CD8+ PD1+	Exhausted phenotype CD8+ cells PMID: 31624246	Enriched in UC and CC over HC Enriched in CC over UC	BloodShare: 0.0663 singleton: 0.70	*EOMES*, *TBX21*, *PRDM1*, *BATF*, *STAT1*, *RUNX3*, *TCF3*, *TCF4*, *TCF7*, *TCF21*, *FOXB1*, *BCL3*	(Chen et al., 2019; Li et al., 2018; Seo et al., 2021)
IFN Response	*IFI44L*, *ISG15*, *IFI6*, *MX1*, *ISG20*, *OAS1*		Strong interferon response signature, some sample-specificity		BloodShare: 0.0313 singleton: 0.81	*IRF1*, *IRF2*, *IRF3*, *IRF4*, *IRF5*, *IRF7*, *IRF8*, *IRF9*, *STAT2*, *STAT1*	

Table S1: Cluster summary of Tissue CD3+ cells. Legend: Tr: Trend, p<0.05 (*), p<0.01(**), p<0.001(***)

Cluster	Key Marker Genes	Key Antibodies	Notes	Condition Enrichment	TCR Analysis	Key TFs	References
Th17 2	*IL22*, *IL17A*, *IL26*, *IL21*, *IL23R*, *IFNG*, *PDCD1*, *CTLA4*, *NCR3*, *CCR6*, *CCL20*, *BATF*	CD4+ PD1+	Inflammation-associated counterpart to Th17 cluster	Increased in both CC and UC over HC, no CC-UC difference	BloodShare: 0.0142 singleton: 0.87	*JUNB*, *ATF7*, *BATF3*, *FOXP3*, *RORC*, *STAT3*, *RORA*, *RARA*, *RUNX2*	(Capone and Volpe, 2020; Christie and Zhu, 2014; Yamazaki et al., 2017)
Tc17	*IL22*, *IL17A*, *IL26*, *IL21*, *IL23R*, *IFNG*, *PDCD1*, *CTLA4*, *NCR3*, *CCR6*, *CCL20*	CD8+ PD1+	Inflammation-associated counterpart to Th17 cluster, CD8+, transcriptionally very similar to Th17 2		BloodShare: 0.0361 singleton: 0.69	*STAT3*, *RORC*, *BATF3*, *MAF*, *PRDM1*, *IRF4*, *AHR*, *STAT5*, *GATA3*	(Christie and Zhu, 2014; Mielke et al., 2019)

Table S1: Cluster summary of Tissue CD3+ cells. Legend: Tr: Trend, p<0.05 (*), p<0.01(**), p<0.001(***)

Cluster	Key Marker Genes	Key Antibodies	Notes	Condition Enrichment	TCR Analysis	Key TFs	References
GZMK+ Effector 1	*GZMK*, *IFNG*, *KLRG1*, *EOMES*, *GZMH*,	CD103- CD8+		Increased in both UC and CC over HC, but not significant	BloodShare: 0.259 singleton: 0.56	*EOMES*, *EGR1*, *EGR2*, *EGR3*, *EGR4*, *IRF7*, *IRF8*, *IRF2*, *TBX21*	(Corridoni et al., 2020)
FGFBP2+ Effector	*FGFBP2*, *CX3CR1*, *IFNG*, *KLRG1*, *EOMES*, *GZMH*, *GZMB*	CD103- CD8+	Circulating, vessel confined CD8+ effector cells.	Enriched in CC over HC (and over UC, but not significant)	BloodShare: 0.592 singleton: 0.35	*KLF2*, *KLF3*	(Corridoni et al., 2020; Yamauchi et al., 2021; Zheng et al., 2017)

Table S1: Cluster summary of Tissue CD3+ cells. Legend: Tr: Trend, p<0.05 (*), p<0.01(**), p<0.001(***)

Cluster	Key Marker Genes	Key Antibodies	Notes	Condition Enrichment	TCR Analysis	Key TFs	References
MAIT	*TRAV1-2*, *TRAJ33*, *ZBTB16*, *KLRG1*, *SLC4A10*, *NCR3*	CD8+	Mucosal-associated invariant T-cells (Semi-invariant TCR Alpha Chain)		BloodShare: 0.177 singleton: 0.76	*RORC*, *RORA*, *TBX21*, *EOMES* (Eomes), *IKZF1* (Helios), *EGR1*, *ELK3*, *ZBTB16*, *ZBTB7A*, *ZBTB7B*, *ZBTB17*, *ZBTB14*, *ZBTB33*	(Gérart et al., 2013; Leeansyah et al., 2015)
Naïve 1	*LEF1*, *CCR7*, *SELL*	CD4+			BloodShare: 0 singleton: 0.99	*KLF2*, *KLF3*, *KLF5*, *FOXP1*, *TCF7*, *LEF1*, *BACH2*, *IKZF1*	(Corridoni et al., 2020)
Tfh	*TOX2*, *PDCD1*, *CTLA4*, *CXCR5*, *CXCL13*	CD4+ PD1+	Follicular helper cells	Enriched in UC over HC and CC	BloodShare: 0.00204 singleton: 0.95	*FOXP1*, *STAT3*, *BCL6*, *MAF*	(Christie and Zhu, 2014)
Cycling	*MKI67*, *CTLA4*, *STMN1*	CD8+ and CD4+	A mix of CD4+ and CD8+ (mostly CD8+) actively proliferating cells in G2M and S-phase	Increased in CC over UC and HC, increased in UC over HC	BloodShare: 0.0429 singleton: 0.75	*E2F8*, *E2F2*, *E2F1*, *E2F7*, *E2F4*, *E2F5*	(Arbogast et al., 1999; Shi et al., 2009)

Table S1: Cluster summary of Tissue CD3+ cells. Legend: Tr: Trend, p<0.05 (*), p<0.01(**), p<0.001(***)

Cluster	Key Marker Genes	Key Antibodies	Notes	Condition Enrichment	TCR Analysis	Key TFs	References
Tregs	*FOXP3*, *IL10*, *CTLA4*, *TNFRSF9*, *CCR6*, *IL1R1*, *IL1R2*, *LRRC32*	CD4+	Regulatory T-cells		BloodShare: 0.00619 singleton: 0.93	*STAT5*, *BATF*, *FOXP3*, *IRF4*, *JUNB*, *FOXO1*, *TCF3*, *LEF1*, *PRDM1*, *RUNX1*	(Christie and Zhu, 2014; Koizumi and Ishikawa, 2019; Shevyrev and Tereshchenko, 2019)

Table S2: Cluster summary of Blood CD3+ cells

Cluster	Key Marker Genes	Key Antibodies	Notes	TCR Analysis	DEGs	References
CD4+ Memory 3	*CCR6* *RORa* *CTSH* *TIMP1* *ANXA1*	CD4+ CD45RO	Potentially Th17 phenotype effector/memory cells	TissueShare: 0.0120 Singleton: 0.94		(Boland et al., 2020; Castro et al., 2017; Gavins and Hickey, 2012; Ramesh et al., 2014; Singh et al., 2008)
CD4+ Memory 2	*CCR10* *CCR4*	CD4+ CD45RO	Possible skin-homing CD4 Th1-Th2 pleiotropic effector memory phenotype	TissueShare: 0.0150 Singleton : 0.95	HLA-C in UC (vs CC)	(Andrew et al., 2001; Boland et al., 2020; Hudak et al., 2002; Nakatani et al., 2001; Soler et al., 2003)
CD4+ Memory 1	*KRT1* *TCF7* *SELL*	CD4+ CD45RO	T central memory cells	TissueShare: 0.00190 Singleton : 0.97	FOS/JUN (AP-1) pathway in CC (vs UC) – induces proliferation but also PD-1 HLA-C, HLA-A and STAT1 in UC (vs CC)	(Szabo et al., 2019)

Table S2: Cluster summary of Blood CD3+ cells

Cluster	Key Marker Genes	Key Antibodies	Notes	TCR Analysis	DEGs	References
CD4+ CD8- FGFBP2+	*FGFBP2* *CX3CR1* *HOPX* *GZMB* *GNLY* *TBX21* *KLRG1*	CD4+ CD45RO	Terminally differentiated CD4 effector memory cells. Per literature, in UC, become IELs on trafficking to tissue. However there is considerable donor variation in the population	TissueShare: 0.197 Singleton : 0.43		(Batista et al., 2020; Gerlach et al., 2016; Kobayashi et al., 2007; Nishimura et al., 2002)
FGFBP2+ Effector 1	*FGFBP2* *CX3CR1* *GZMH* *ADGRG1* *SPON2* *ZEB2*	CD8+ CD45RO	Cytotoxic cells, with adhesion and migration markers, Found to correlate with clinical response in Checkpoint therapy	TissueShare: 0.115 Singleton : 0.54	IFITM3 (activation marker) in CC (vs UC) HLA-C, KLRG1, IRF1, CCL5 (markers of terminal differentiation) in UC (vs CC). Also however express RAP1B which is associated with reduced migration into colon.	(Li et al., 2019; Omilusik et al., 2015; Yamauchi et al., 2021)

Table S2: Cluster summary of Blood CD3+ cells

Cluster	Key Marker Genes	Key Antibodies	Notes	TCR Analysis	DEGs	References
GZMK+ Effector 2	*GZMK HLA-DRA FABP5 CD160*	CD8+ CD45RO	Highly active Effector T-cell population – because expressing high amounts of HLA type 1, correlated with autoimmune disease in SLE	TissueShare: 0.194 Singleton : 0.54		(Boland et al., 2020; Szabo et al., 2019; Viallard et al., 2001)
Gamma-delta	*TRDC FCGR3A TYROBP*			TissueShare: 0.0636 Singleton : 0.74		
GZMK+ Effector 1	*GZMK XCL2 CCL4L2 CD160*	CD8+ CD45RO	Effector population Expressing some exhaustion markers	TissueShare: 0.0821 Singleton : 0.63	CTSW in CC (vs UC) – marker of cytotoxicity.	(Szabo et al., 2019)
MAITs	*TRAV1-2, TRAJ33, ZBTB16 SLC4A10*	CD8+ CD45RO	Represent 5-10% of the circulating population	TissueShare: 0.0368 Singleton : 0.76	HLA-C in UC (vs CC) TXNIP (Activation marker) in CC (vs UC)	(Chiba et al., 2018)

Table S2: Cluster summary of Blood CD3+ cells

Cluster	Key Marker Genes	Key Antibodies	Notes	TCR Analysis	DEGs	References
Memory CD8+	*TCF7* *IL7R* *FXYD2* *CCR7*	CD8+ CD45RO	Potentially Central Memory / Resting CD8 T-cells / Gradient from Naïve population	TissueShare: 0.0458 Singleton : 0.87		(Belarif et al., 2018; Boland et al., 2020; Szabo et al., 2019)
Naïve CD8+	*CCR7* *LEF1*	CD8+		TissueShare: 0 Singleton : 0.99	HLA-C in UC, CD69 in CC vs UC	
Cell Cycle	*MKI67* *STMN1*	CD4+ CD8+ CD45RO		TissueShare: 0.0223 Singleton : 0.95		
Tregs	*FOXP3*	CD4+ CD45RO		TissueShare: 0.00558 Singleton : 0.98		
Naïve CD4+	*CCR7* *LEF1*	CD4+		TissueShare: 0.000416 Singleton : 0.99	CD69 and JUNB in CC (vs UC) HLA-C in UC	

Epithelial Compartment

Table S3: Cluster Summary of Epithelial, Stromal and Immune populations

Cluster	Sub-Cluster(s) if present	Key Marker Genes	Abundance	Notes	Key Literature References
Absorptive	Absorptive (1) and (2)	*TIMP3, TAF4B, BAMBI, HIST1H3C, ARC*	Fewer in CC_I vs CC_NI (*)	Two clusters differ because of gradient of gene expression	For all Epithelial subclusters below: (Parikh et al., 2019; Smillie et al., 2019)
BEST4		*BEST4, OTOP2, SPIB, HES4*	Fewer in CC_I vs CC_NI (**)		
Crypt Top Enterocytes		*AQP8, CEACAM1, CEACAM7*	Fewer in CC_I vs CC_NI (**)		
EEC (Enteroendocrine)		*CHGA, PYY, CRYBA2, SCGN, FEV*	No changes in Inflammation		
Enterocytes		*CA1, SLC26A3, SELENBP1*	Fewer in CC_I vs CC_NI (**)		

Table S3: Cluster Summary of Epithelial, Stromal and Immune populations

Cluster	Sub-Cluster(s) if present	Key Marker Genes	Abundance	Notes	Key Literature References
Goblets		*MUC2, TFF3, ITLN1, SPINK4, WFDC2, KLK1, REP15, TPSG1,*	CC_I > CC_NI (Tr)		
Inflammation Enterocytes		*NOS2, WARS, IDO1, DUOX2A, DUOX, CXCL11*	More in CC_I vs CC_NI (**), Same trend in UC		
Stem Cells		*LGR5, ASCL2, SMOC2, OLFM4*	No changes in inflammation		
TA (Transit Amplifying) Cells	TA and TA (G2M) clusters	*MKI67, ZWINT, TYMS. GINS2, CENPM, CDC45, MCM10*	No changes in inflammation	TA and TA G2M differ in degree of replication signal expression	

Table S3: Cluster Summary of Epithelial, Stromal and Immune populations

Cluster	Sub-Cluster(s) if present	Key Marker Genes	Abundance	Notes	Key Literature References
Tuft		*HTR3E, GNG13, SH2D7, SH2D6,TRPM5,AZGP1,*	No changes in inflammation		
Stromal Cells		*CAV1, SOX8, NTRK2, CCL23, MIA, SERPINF2,*	No changes in inflammation	Small residual stromal population in epithelial enriched fraction	
T-Cells		*CD3E, CD69, CD96, CD7, ZNF331, EVL*	CC_I > CC_NI (Tr)	Intra-epithelial T lymphocytes	
B-Cells		*CD19,CD79A, CD79B, MS4A1, CD69, EVL*	No changes in Inflammation		
Mast T-cells		*TPSAB1, TPSB2C*	No changes in Inflammation		
Myeloid Cells		*CD14, C1QA, C1QB, C1QC, CTSB,*	CC_I > CC_NI (*), > UC_I (Tr) and HC (Tr)		

Table S3: Cluster Summary of Epithelial, Stromal and Immune populations

Cluster	Sub-Cluster(s) if present	Key Marker Genes	Abundance	Notes	Key Literature References
Plasma Cells		*IGLC2, IGLC3*	Trend towards more in CC_I vs all other conditions		

Stromal (Enriched Fraction) Compartment

Cluster	Sub-Cluster(s) if present	Key Marker Genes	Abundance	Notes	Key Literature References
Myofibroblasts		*ACTG2, MYH11, LUM, HHIP, SOSTDC1, MFAP5, NPNT, FOXL1*	Depleted in CC_I vs CC_NI(**) and HC(**). No differences between UC and HC.	Production of Lumican higher in inflammation – may be secondary to TNFa stimulation. *F3, POSTN, EDNRB, MRVI1, MMP11, FOXL1, PDGFRA* shared with S2	(Fawkner-Corbett et al., 2021; Kinchen et al., 2018; Martin et al., 2019; Pilling et al., 2015; ShoshkesCarmel et al., 2018)

Cluster	Sub-Cluster(s) if present	Key Marker Genes	Abundance	Notes	Key Literature References
Myofibroblasts 2		*MYH11, GREM1/2, RSPO3, RSPO2, CHRDL1*	No differences in inflammation	*GREM1/2* as BMP antagonists help maintain a stem-cell niche. Abundant in Submucosa. Deeper Myofibroblast cluster (Seen most in resection ST sections B8 and B9) *GREM1* and *GREM2* are a marker of fibroblast activation in lung *RSPO2* shared with S2 *GREM1,GREM2, CHRDL1, RSPO3* shared with S3	(Brügger et al., 2020; Huan et al., 2021; Kosinski et al., 2007a)
Pericytes		*COX4I2, RGS5, NOTCH3, C1QTNF1, FAM162B, STEAP4, KCNJ8*	CC_I>CC_NI(****) and HC(**). Trend towards UC_I>UC_NI and HC.		(Fawkner-Corbett et al., 2021; Kinchen et al., 2018; Martin et al., 2019; Travaglini et al., 2020; Uhlitz et al., 2021)
Endothelial		*PECAM1, CD34, VWF, EGFL7, PLVAP, RAMP2 &3, CLDN5,*	CC_I > CC_NI(***), CC_II > HC(**),	Endothelial cell markers are shared with subcluster 2 of Plasma cells	(Fawkner-Corbett et al., 2021)
Glial		*S100B, NRXN1, GPM6B, CDH19,PLP1, MPZ, SPP1*	CC_I>UC_I(*), but the same as HC		

Cluster	Sub-Cluster(s) if present	Key Marker Genes	Abundance	Notes	Key Literature References
Stromal 1 (Wnt2B+ FosHi and FosLo)	S1 (1) and S1(2) sub-clus-ters.	*ADAMDEC1, FABP4/5, CCL8/7 MCTP1*	Depleted in CC_I vs CC_NI(***) and HC(*). Depleted in UC_I vs UC_NI(*) and HC(Tr)	Gradient of *CCL13, SCT, GSTM1,PROM1, POSTN, DMKN, OASL, GCH1, MX2* expression between two subclusters *WNT2B* and *RSPO3* producer, along with S3 *MCTP1* shared production with Activated S1 cells (*CCL2, ABCA8, COL14A1, APOE* that were Historically used were not specific S1 markers in this dataset)	(Fawkner-Corbett et al., 2021; Kinchen et al., 2018)
Stromal 2 (Wnt5B)		*BMP5, Wnt5B, COL4A5, COL4A6 PDGFRA, NSG1, AGT, ENHO, VSTM2A, HAS1,TRPA1, SOX6, NRG1, GLP2R, ALKAL2* Global S2 markers	Depleted in CC_I vs CC_NI(**) and HC(*)	Supports epithelial stem-cell niche through production of Wnt, in conjunction with *GREM1/2* produced by myofibroblast layer that inhibits BMP that would otherwise inhibit the Wnt pathway. *F3, POSTN, EDNRB, MRVI1, MMP11, FOXL1, PDGFRA* shared with Myofibroblasts *FRZB* shared with Pericytes *APOD* shared with S3 *LIMCH1* shared with Endothelial *HSD17B2* shared with Epithelium	(Halpern et al., 2020; Kinchen et al., 2018; Kosinski et al., 2007b; Melissari et al., 2021; Popescu and Faussone-Pellegrini, 2010; Qi et al., 2017; Smillie et al., 2019)

Cluster	Sub-Cluster(s) if present	Key Marker Genes	Abundance	Notes	Key Literature References
	S2(1) and S2(2)	*Gradient of SCUBE2, MMP11, FGF9, HAS1,GFOD1 DUSP2 between clusters*	As above	S2(1) and S2(2) likely represent fibroblast expression gradient differences across cluster	
	S2 TLL2+	*RSPO2, SMOC2, CPM, CYP1B1, PTGIS, PDGFRL, TLL2, BLNK, TSPAN7, ADAMTSL3, ASB13*	No changes in inflammation	Function of subset unknown. Primary RSPO2 producing S2 cluster *SMOC* is a intestinal stem cell marker and acts to drive fibrosis through MF -> Fibroblast transition in Kidney & Lung (possibly via TGFB1 pathway) Expresses global S2 markers	

Cluster	Sub-Cluster(s) if present	Key Marker Genes	Abundance	Notes	Key Literature References
	S2 NPY+	*NPY, LTBP1, AREG,FST, SLITRK6,SLC4A4, CYTL1,FBN2,F2RL3, SEZ6L2, BMP7,PTX3, GJA1*	CC_I < CC_NI (**) and UC_I(*) and HC (Tr).	Function of subset unknown. Localised around Vascular structures in the deeper stroma by ST Expresses PID pathways for ERBB (Epithelial Growth factor receptor), LIS1 (Neuronal migration), DELTA-NP63 (Hallmark of stemness),A6B1-A6B4-INTEGRIN (Hallmark of stemness) NPY only produced by this cluster, does not also produce PYY or PPY which are otherwise produced by EEC *ITGA8, PTGS1, LEF1, PAPLN, PTX3, GJA1, LTBP1, AKR1C1, TBXAS1 , MMP3, SMOX, BMP7, STC1, NRG1, LOX* shared with activated fibroblasts Expresses global S2 markers but does not produce *RSPO1/2* or *NDUFA4L2* like the other S2 subsets	(Chen et al., 2020; Cox, 2007; El-Salhy and Hausken, 2016)
Stromal 3		*GREM1, OGN, PRELP, IL33, C1QTNF3, EBF1, FGL2, CD34 Global S3 markers*	CC_I>CC_NI(***) and HC(**). UC_I>UC_NI (Tr) and HC(Tr)	Clustered around vascular structures by ST	(Fawkner-Corbett et al., 2021; Kinchen et al., 2018)

Cluster	Sub-Cluster(s) if present	Key Marker Genes	Abundance	Notes	Key Literature References
	S3	*SFRP2, THBS2, AKR1C1,PLA2G2A, OSR2,CHRDL2,AOX1,WISP2, SL, PI, CHRDL1*		*MGST1, MEDAG, CFB, FIBIN* common with activated fibroblast cluster *RSPO3, SPRP1* shared with S1 *MGST1* common with epithelium	
	S3 PI16+	*PI16,OGN,MFAP5*	CC_NI>CC_I(*)	*CPM, MFAP5, VIT* shared with S3 Belongs to deeper S3 Subset (B8) More RSPO3/less immune focused	(Harnack et al., 2019)
	S3 C7+	*C7,C1QTNF3,CCL19,CYP7B1*	No changes in inflammation	*CHRDL1, GPC3* shared with S3 More superficial S3 Subset (B8) More immune focused	
Stromal 4 (S4)		*CCL19, CCL21,CTSH,FDCSP*		Localised around lymphoid aggregates/follicles *CTSH,FDCSP* in common with activated fibroblast cluster	(Kinchen et al., 2018)
Activated Fibroblast		*IL13RA2, IL24, FAP TWIST1, CXCL8, CXCL10, WNT5A* (but not *WNT5B)* Global Activated Fibroblast markers		Formed predominantly of activated S1, S2, CSF3+ clusters *CD82, IL13RA2, IL24, FAP ,TWIST1 , CHI3L1, NQO1, THRC1, CLU, PDPN* also shared with S3 and S4	(Martin et al., 2019)

Cluster	Sub-Cluster(s) if present	Key Marker Genes	Abundance	Notes	Key Literature References
	Activated S1	*CP RAMP1 CHI3L2*	CC_I>CC_NI(*), UC_I>HC(Tr)	*RAMP1, CHI3L2* also expressed by TNFSFB11 (S1) *CSF3* Shared with Activated TNFSFB11 and activated CSF3	
	Activated S2	*DEFB1, FBH1, KIAA0930, GPX7, IL7R, ITGA2,MMP10, MYH11, IL1RL1, CXCL11*	CC_I>CC_NI (**) and HC(*). UC_I>UC_NI(*)	*SMOC1* shared with activated TNFRSF11 (S1) *ITGA2* shared with S2-NPY *CXCL11* shared with Endothelial cells	
	Activated CSF3+	*LIF,IL6,CXCL8,CSF3, RARRES1, RAF1*	UC_I>UC_NI (*), HC(Tr), CC_I>CC_NI(Tr) and HC(Tr)	*CSF3, CXCL6, CXCL8 RARRES1, STEAP1, TWIST1 PRRX1, IL24, CHI3L2* shared with other activated S1 clusters *FDCSP* shared with other activated FB clusters and S4 *CCL19* shared with S4 *GREM1* shared with S3/S4 *TRAF1* shared with Activated C3+	

Cluster	Sub-Cluster(s) if present	Key Marker Genes	Abundance	Notes	Key Literature References
	Activated TN-FSFB11 S1	*IL13RA2,RGS5,MGST1, CXCL6, CXCL8,IL11, RAMP1, STEAP1,RA, RRES1, ASS1,ASL, VIPR2,GK,CE MIP,FIBIN,CHI3L2 PF4V1,CHST15,FAM20A PI15,S, LC16A4, FDCSP,CP*	CC_I>CC_NI (***) and HC(*). UC_I>UC_NI (Tr) and HC (Tr)	*CP* also expressed by S1 close to S3/4 *ASS1* also expressed by S3 *FDCSP* also expressed by S4 *CXCL8* aslo expressed by CSF3+ *SULF2, SNX10, CTHRC1, WISP1* Expressed with other activated fibroblasts RGS5 expressed with pericytes	
	S1 Interferon response	*IFIT1/2, OAS1/2/3/, OASL, CD69*	No differences in inflammation	*CXCL10* (shared with activated Fibroblast cluster) Shares *CCL7/8, ADAMDEC1* with parent S1 cluster	
	Activated C3	*C3, CFB, CD24,TAC3,TNFSF11, TRAF1, CST1*	UC_I> CC_I(**), UC_NI(*) and HC(Tr)	*C3* shared with S3 and S4 *CFB* shared with S1 activated *CD24, TNFSF11* shared with S4. *TRAF1* shared with activated CSF3+. Produces more BMP1 than other activated FB	
Epithelial		*ELF3, EPCAM, KRT19, LGALS4, CLDN3*		Small number of residual epithelial cells in stromal/CD45 enriched population	

Immune Comparment

Cluster	Sub-Cluster(s) if present	Key Marker Genes	Abundance	Notes	Key Literature References
Cycling		*MKI67, ZWINT, CD3D, CD19, PTPRC*	Increased in CC_I and UC_I vs NI and Health.	Conglomerate cluster comprising cycling immune cells, largely T-cells, B cells and plasma cells	
NK&T-cells		*CD3D, CD7, KLRB1*	Analysed in more detail utilizing CD3 dataset		
Mast		*CPA3, TPSAB1, TPSB2, MS4A2, TPSD1, SLC18A2*	No differences in inflammation		
Myeloid		*CD14, AIF1, SPI1, CD68, C1QA, C1QB, C1QC, IGSF6, LST1*	CC_I>CC_NI (***), CC_I>HC(*), Trend towards CC_I>UC_I		
B Cells		*C19,CD22, DMS4A1, BANK1, CXCR5,FCRLA, PIKFYVE, TNFRSF13C, FAM129C*	No changes in inflammation		
cDC		*CD1C, CD1E, FCER1A, CLEC9A, FLT3*	Increased in HC vs CC_I(**) and UC_I(*)		For all Myeloid cell subsets reported below: (Gautier et al., 2012; Martin et al., 2019; Villar and Segura, 2020)

Cluster	Sub-Cluster(s) if present	Key Marker Genes	Abundance	Notes	Key Literature References
pDC		*CLEC4C, IL3RA, BCL11A,TCF4*	Trends towards increase in both UC_I and CC_I		
ToIDC (mregDC)		*CCL19,CCR7,LAMP3, FSCN1*	No differences in Inflammation	Tolerogenic Dendritic cells	
Inflammatory Macrophages	Inflammatory Macrophages (1)	*CCL18, CXCL1, S100A13*	Increased in CC_I vs HC (**) and UC (*). No differences between UC_I and HC		
	Inflammatory Macrophages (2)	*NFKBIA, S100A6, CXCL11, SIRPB1, ETV7,*	Increased in CC_I (*), trend of increase in UC_I vs HC		
	Inflammatory Macrophages (3)	*S100A12, SERPINB2, APOBEC3A,NFKBIA, S100A6, IL1B,*	Increased in UC_I (*) and CC_I(**) compared to HC		
	Inflammatory Macrophages (4)	*INHBA, IL1A, SLAMF1*	Trends towards increase in both UC_I and CC_I		

Cluster	Sub-Cluster(s) if present	Key Marker Genes	Abundance	Notes	Key Literature References
Macrophages (without proinflammatory signals)	Macrophages (1) and (2)	*CD14 CD68*	No differences in inflammation	No pro-inflammatory signal	
	Macrophages (3)	*RNASE1, IGF1, ETV5, MRC1*	Reduced in both UC_I(*) and CC_I(*) compared to HC	Expresses markers of residence (MRC1)	
Cycling		*MKI67*	Increased in HC vs CC_I(**) and UC_I(*)	Cycling Myeloid cells.	
Atypical B Cells		*CD19, ITGAX, CD86, EBI3, GPR137B, BATF, SIGLEC6, FCRL4, SOX5, LIMK1, TESC, MPP6, CD58, HMOX1, WEE1, ZBTB32,*	No differences in inflammation		For all B cell subsets described below: (Boland et al., 2020; Sanz et al., 2019; da Silva et al., 2020; Stewart et al., 2021; Villar and Segura, 2020)
CD27-		*CD19, TCL1A, IL4R, IGHD, BCL7A, CD200*	No differences in inflammation		

Cluster	Sub-Cluster(s) if present	Key Marker Genes	Abundance	Notes	Key Literature References
CD27+		*CD19, CD27, CD70, KLK1, TMEM273, PVT1, TNF*			
Germinal Centre		*CD19, RGS13, SUGCT, SERPINA9, MYO1E, ELL3, HOPX, GCSAM,*	UC_I> CC_I (Tr) and HC (Tr)		
Germinal Centre Cycling		*CD19, RGS13, SUGCT, SERPINA9, MYO1E, ELL3, HOPX, GCSAM, MKI67, ZWINT*	UC_I> CC_I (*) and HC (*)		
Plasma		*ANKRD28, XBP1, PRDM1,TNFRSF17, TRIB1, RRBP1, DUSP5, , SPAG4, SDC1, AQP3.*		CD19 negative	
Interferon Response		*IFI6, IFIT3, WARS*	UC_I > HC(*), CC_I > HC (Tr)		

References

Andrew, D.P., Ruffing, N., Kim, C.H., Miao, W., Heath, H., Li, Y., Murphy, K., Campbell, J.J., Butcher, E.C., and Wu, L. (2001). C-C Chemokine Receptor 4 Expression Defines a Major Subset of Circulating Nonintestinal Memory T-cells of Both Th1 and Th2 Potential. The Journal of Immunology *166*, 103–111.

Arbogast, A., Boutet, S., Phelouzat, M.A., Plastre, O., Quadri, R., and Proust, J.J. (1999). Failure of T Lymphocytes from Elderly Humans to Enter the Cell Cycle Is Associated with Low Cdk6 Activity and Impaired Phosphorylation of Rb Protein. Cellular Immunology *197*, 46–54.

Barros-Martins, J., Schmolka, N., Fontinha, D., Miranda, M.P. de, Simas, J.P., Brok, I., Ferreira, C., Veldhoen, M., Silva-Santos, B., and Serre, K. (2016). Effector $\gamma\delta$ T-cell Differentiation Relies on Master but Not Auxiliary Th Cell Transcription Factors. The Journal of Immunology *196*, 3642–3652.

Batista, N. v., Chang, Y.-H., Chu, K.-L., Wang, K.C., Girard, M., and Watts, T.H. (2020). T-cell–Intrinsic CX3CR1 Marks the Most Differentiated Effector CD4+ T-cells, but Is Largely Dispensable for CD4+ T-cell Responses during Chronic Viral Infection. ImmunoHorizons *4*, 701–712.

Behr, F.M., Chuwonpad, A., Stark, R., and Gisbergen, K.P.J.M. van (2018). Armed and Ready: Transcriptional Regulation of Tissue-Resident Memory CD8 T-cells. Frontiers in Immunology *9*, 1770.

Belarif, L., Mary, C., Jacquemont, L., Mai, H. le, Danger, R., Hervouet, J., Minault, D., Thepenier, V., Nerrière-Daguin, V., Nguyen, E., et al. (2018). IL-7 receptor blockade blunts antigen-specific memory T-cell responses and chronic inflammation in primates. Nature Communications 2018 9:1 *9*, 1–13.

Bhullar, J., and Sollars, V.E. (2011). YBX1 expression and function in early hematopoiesis and leukemic cells. Immunogenetics 2011 63:6 *63*, 337–350.

Boland, B.S., He, Z., Tsai, M.S., Olvera, J.G., Omilusik, K.D., Duong, H.G., Kim, E.S., Limary, A.E., Jin, W., Milner, J.J., et al. (2020). Heterogeneity and clonal relationships of adaptive immune cells in ulcerative colitis revealed by single-cell analyses. Science Immunology *5*, 4432.

Brügger, M.D., Valenta, T., Fazilaty, H., Hausmann, G., and Basler, K. (2020). Distinct populations of crypt-associated fibroblasts act as signaling hubs to control colon homeostasis. PLoS Biology *18*.

Capone, A., and Volpe, E. (2020). Transcriptional Regulators of T Helper 17 Cell Differentiation in Health and Autoimmune Diseases. Frontiers in Immunology *11*.

Castro, G., Liu, X., Ngo, K., Leon-Tabaldo, A. de, Zhao, S., Luna-Roman, R., Yu, J., Cao, T., Kuhn, R., Wilkinson, P., et al. (2017). RORγt and RORα signature genes in human Th17 cells. PLoS ONE *12*.

Caushi, J.X., Zhang, J., Ji, Z., Vaghasia, A., Zhang, B., Hsiue, E.H.-C., Mog, B.J., Hou, W., Justesen, S., Blosser, R., et al. (2021). Transcriptional programs of neoantigen-specific TIL in anti-PD-1-treated lung cancers. Nature 2021 596:7870 *596*, 126–132.

Chen, W., Liu, Y., Liu, W., Zhou, Y., He, H., and Lin, S. (2020). Neuropeptide Y Is an Immunomodulatory Factor: Direct and Indirect. Frontiers in Immunology *11*.

Chen, Y., Zander, R., Khatun, A., Schauder, D.M., and Cui, W. (2018). Transcriptional and Epigenetic Regulation of Effector and Memory CD8 T-cell Differentiation. Frontiers in Immunology *9*, 2826.

Chen, Z., Ji, Z., Ngiow, S.F., Manne, S., Cai, Z., Huang, A.C., Johnson, J., Staupe, R.P., Bengsch, B., Xu, C., et al. (2019). TCF-1-centered transcriptional network drives an effector versus exhausted CD8 T-cell fate decision. Immunity *51*, 840.

Chiba, A., Murayama, G., and Miyake, S. (2018). Mucosal-Associated Invariant T-cells in Autoimmune Diseases. Frontiers in Immunology *0*, 1333.

Christie, D., and Zhu, J. (2014). Transcriptional regulatory networks for CD4 T-cell differentiation. Current Topics in Microbiology and Immunology *381*, 125– 172.

Cobaleda, C., Jochum, W., and Busslinger, M. (2007). Conversion of mature B cells into T-cells by dedifferentiation to uncommitted progenitors. Nature 2007 449:7161 *449*, 473–477.

Corridoni, D., Antanaviciute, A., Gupta, T., D, F.-C., A, A., M, J., K, P., E, R., S, T., D, I., et al. (2020). Single-cell atlas of colonic CD8 + T-cells in ulcerative colitis. Nature Medicine *26*, 1480–1490.

Cox, H.M. (2007). Neuropeptide Y receptors; antisecretory control of intestinal epithelial function. Autonomic Neuroscience: Basic and Clinical *133*, 76–85.

David-Fung, E.-S., Butler, R., Buzi, G., Yui, M.A., Diamond, R.A., Anderson, M.K., Rowen, L., and Rothenberg, E. v. (2009). Transcription factor expression dynamics of early T-lymphocyte specification and commitment. Developmental Biology *325*, 444.

Egawa, T., Tillman, R.E., Naoe, Y., Taniuchi, I., and Littman, D.R. (2007). The role of the Runx transcription factors in thymocyte differentiation and in homeostasis of naive T-cells. The Journal of Experimental Medicine *204*, 1945.

El-Salhy, M., and Hausken, T. (2016). The role of the neuropeptide Y (NPY) family in the pathophysiology of inflammatory bowel disease (IBD). Neuropeptides *55*, 137–144.

Evans, C., and Jenner, R. (2013). Transcription factor interplay in T helper cell differentiation. Briefings in Functional Genomics *12*, 499–511.

Fawkner-Corbett, D., Antanaviciute, A., Parikh, K., M, J., AS, G., T, G., N, A., D, K., D, F., E, M., et al. (2021). Spatiotemporal analysis of human intestinal development at single-cell resolution. Cell *184*, 810-826.e23.

Feng, X., Wang, H., Takata, H., Day, T.J., Willen, J., and Hu, H. (2011). Transcription factor Foxp1 exerts essential cell-intrinsic regulation of the quiescence of naive T-cells. Nature Immunology *12*, 544.

Gautier, E.L., Shay, T., Miller, J., Greter, M., Jakubzick, C., Ivanov, S., Helft, J., Chow, A., Elpek, K.G., Gordonov, S., et al. (2012). Gene expression profiles and transcriptional regulatory pathways underlying mouse tissue macrophage identity and diversity. Nature Immunology *13*, 1118.

Gavins, F.N.E., and Hickey, M.J. (2012). Annexin A1 and the regulation of innate and adaptive immunity. Frontiers in Immunology *3*.

Gegonne, A., Tai, X., Zhang, J., Wu, G., Zhu, J., Yoshimoto, A., Hanson, J., Cultraro, C., Chen, Q.-R., Guinter, T., et al. (2012). The General Transcription Factor TAF7 Is Essential for Embryonic Development but Not Essential for the Survival or Differentiation of Mature T-cells. Molecular and Cellular Biology *32*, 1984.

Gérart, S., Sibéril, S., Martin, E., Lenoir, C., Aguilar, C., Picard, C., Lantz, O., Fischer, A., and Latour, S. (2013). Human iNKT and MAIT-cells exhibit a PLZFdependent proapoptotic propensity that is counterbalanced by XIAP. Blood *121*, 614–623.

Gerlach, C., Moseman, E.A., Loughhead, S.M., Alvarez, D., Zwijnenburg, A.J., Waanders, L., Garg, R., Torre, J.C. de la, and Andrian, U.H. von (2016). The chemokine receptor CX3CR1 defines three antigen-experienced CD8 T-cell subsets with distinct roles in immune surveillance and homeostasis. Immunity *45*, 1270.

Guo, Y., Yuan, X., Li, K., Dai, M., Zhang, L., Wu, Y., Sun, C., Chen, Y., Cheng, G., Liu, C., et al. (2020). GABPA is a master regulator of luminal identity and restrains aggressive diseases in bladder cancer. Cell Death and Differentiation *27*, 1862.

Halpern, K.B., Massalha, H., Zwick, R.K., Moor, A.E., Castillo-Azofeifa, D., Rozenberg, M., Farack, L., Egozi, A., Miller, D.R., Averbukh, I., et al. (2020). Lgr5+ telocytes are a signaling source at the intestinal villus tip. Nature Communications *11*.

Harnack, C., Berger, H., Antanaviciute, A., Vidal, R., Sauer, S., Simmons, A., Meyer, T.F., and Sigal, M. (2019). R-spondin 3 promotes stem cell recovery and epithelial regeneration in the colon. Nature Communications 2019 10:1 *10*, 1–15.

Huan, C., Xu, W., Liu, Y., Ruan, K., Shi, Y., Cheng, H., Zhang, X., Ke, Y., and Zhou, J. (2021). Gremlin2 Activates Fibroblasts to Promote Pulmonary Fibrosis Through the Bone Morphogenic Protein Pathway. Frontiers in Molecular Biosciences *0*, 619.

Hudak, S., Hagen, M., Liu, Y., Catron, D., Oldham, E., McEvoy, L.M., and Bowman, E.P. (2002). Immune Surveillance and Effector Functions of CCR10+ Skin Homing T-cells. The Journal of Immunology *169*, 1189–1196.

Jaeger, N., Gamini, R., Cella, M., Schettini, J.L., Bugatti, M., Zhao, S., Rosadini, C. v., Esaulova, E., Luccia, B. di, Kinnett, B., et al. (2021). Single-cell analyses of Crohn's disease tissues reveal intestinal intraepithelial T-cells heterogeneity and altered subset distributions. Nature Communications *12*.

Kaech, S.M., and Cui, W. (2012). Transcriptional control of effector and memory CD8+ T-cell differentiation. Nature Reviews. Immunology *12*, 749.

Kanaya, T., Hase, K., Takahashi, D., Fukuda, S., Hoshino, K., Sasaki, I., Hemmi, H., Knoop, K.A., Kumar, N., Sato, M., et al. (2012). The Ets Transcription Factor Spi-B Is Essential for the Differentiation of Intestinal Microfold (M) Cells. Nature Immunology *13*, 729.

Kinchen, J., HH, C., K, P., A, A., M, J., D, F.-C., N, A., L, C., E, M.-G., M, A., et al. (2018). Structural Remodeling of the Human Colonic Mesenchyme in Inflammatory Bowel Disease. Cell *175*, 372-386.e17.

Kobayashi, T., Okamoto, S., Iwakami, Y., Nakazawa, A., Hisamatsu, T., Chinen, H., Kamada, N., Imai, T., Goto, H., and Hibi, T. (2007). Exclusive increase of CX3CR1+CD28-CD4+ T-cells in inflammatory bowel disease and their recruitment as intraepithelial lymphocytes. Inflammatory Bowel Diseases *13*, 837–846. Koizumi, S., and Ishikawa, H. (2019). Transcriptional Regulation of Differentiation and Functions of Effector T Regulatory Cells. Cells *8*, 939.

Kosinski, C., Li, V.S.W., Chan, A.S.Y., Zhang, J., Ho, C., Tsui, W.Y., Chan, T.L., Mifflin, R.C., Powell, D.W., Yuen, S.T., et al. (2007a). Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. Proceedings of the National

Academy of Sciences of the United States of America *104*, 15418.

Kosinski, C., Li, V.S.W., Chan, A.S.Y., Zhang, J., Ho, C., Tsui, W.Y., Chan, T.L., Mifflin, R.C., Powell, D.W., Yuen, S.T., et al. (2007b). Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. Proceedings of the National Academy of Sciences of the United States of America *104*, 15418.

Kragten, N.A.M., Behr, F.M., Braga, F.A.V., Remmerswaal, E.B.M., Wesselink, T.H., Oja, A.E., Hombrink, P., Kallies, A., Lier, R.A.W. van, Stark, R., et al. (2018).

Blimp-1 induces and Hobit maintains the cytotoxic mediator granzyme B in CD8 T-cells. European Journal of Immunology *48*, 1644–1662.

Kumar, B. v., Ma, W., Miron, M., Granot, T., Guyer, R.S., Carpenter, D.J., Senda, T., Sun, X., Ho, S.-H., Lerner, H., et al. (2017). Human tissue-resident memory T-cells are defined by core transcriptional and functional signatures in lymphoid and mucosal sites. Cell Reports *20*, 2921.

Leeansyah, E., Svärd, J., Dias, J., Buggert, M., Nyström, J., Quigley, M.F., Moll, M., Sönnerborg, A., Nowak, P., and Sandberg, J.K. (2015). Arming of MAIT-cell Cytolytic Antimicrobial Activity Is Induced by IL-7 and Defective in HIV-1 Infection. PLoS Pathogens *11*.

Li, H., van der Leun, A.M., Yofe, I., Lubling, Y., Gelbard-Solodkin, D., van Akkooi, A.C.J., van den Braber, M., Rozeman, E.A., Haanen, J.B.A.G., Blank, C.U., et al. (2019). Dysfunctional CD8 T-cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma. Cell *176*, 775-789.e18.

Li, J., He, Y., Hao, J., Ni, L., and Dong, C. (2018). High Levels of Eomes Promote Exhaustion of Anti-tumor CD8+ T-cells. Frontiers in Immunology *9*, 2981.

Luo, C.T., Osmanbeyoglu, H.U., Do, M.H., Bivona, M.R., Toure, A., Kang, D., Xie, Y., Leslie, C.S., and Li, M.O. (2017). Ets transcription factor GABP controls T-cell homeostasis and immunity. Nature Communications *8*.

Martin, J.C., Chang, C., Boschetti, G., R, U., M, G., JA, G., K, G., LS, C., S, N., AJ, G., et al. (2019). Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. Cell *178*, 1493-1508.e20.

Melissari, M.-T., Henriques, A., Tzaferis, C., Prados, A., Sarris, M.E., Chouvardas, P., Grammenoudi, S., Kollias, G., and Koliaraki, V. (2021). Col6a1+/CD201+ mesenchymal cells regulate intestinal morphogenesis and homeostasis. BioRxiv 2021.02.16.431453.

Mielke, L.A., Liao, Y., Clemens, E.B., Firth, M.A., Duckworth, B., Huang, Q., Almeida, F.F., Chopin, M., Koay, H.-F., Bell, C.A., et al. (2019). TCF-1 limits the formation of Tc17 cells via repression of the MAF–RORγt axis. The Journal of Experimental Medicine *216*, 1682.

Nakatani, T., Kaburagi, Y., Shimada, Y., Inaoki, M., Takehara, K., Mukaida, N., and Sato, S. (2001). CCR4+ memory CD4+ T lymphocytes are increased in peripheral blood and lesional skin from patients with atopic dermatitis. Journal of Allergy and Clinical Immunology *107*, 353–358.

Nishimura, M., Umehara, H., Nakayama, T., Yoneda, O., Hieshima, K., Kakizaki, M., Dohmae, N., Yoshie, O., and Imai, T. (2002). Dual Functions of

Fractalkine/CX3C Ligand 1 in Trafficking of Perforin+/Granzyme B+ Cytotoxic Effector Lymphocytes That Are Defined by CX3CR1 Expression. The Journal of Immunology *168*, 6173–6180.

Oestreich, K., and Weinmann, A. (2012). Transcriptional mechanisms that regulate T helper 1 cell differentiation. Current Opinion in Immunology *24*, 191– 195.

Omilusik, K.D., Best, J.A., Yu, B., Goossens, S., Weidemann, A., Nguyen, J. v., Seuntjens, E., Stryjewska, A., Zweier, C., Roychoudhuri, R., et al. (2015). Transcriptional repressor ZEB2 promotes terminal differentiation of CD8+ effector and memory T-cell populations during infection. The Journal of Experimental Medicine *212*, 2027.

Parikh, K., Antanaviciute, A., Fawkner-Corbett, D., Jagielowicz, M., Aulicino, A., Lagerholm, C., Davis, S., Kinchen, J., Chen, H.H., Alham, N.K., et al. (2019). Colonic epithelial cell diversity in health and inflammatory bowel disease. Nature 2019 567:7746 *567*, 49–55.

Parish, I.A., and Kaech, S.M. (2009). Diversity in CD8+ T-cell differentiation. Current Opinion in Immunology *21*, 291.

Parker, M.E., and Ciofani, M. (2020). Regulation of $\gamma\delta$ T-cell Effector Diversification in the Thymus. Frontiers in Immunology *11*, 42.

Pilling, D., Vakil, V., Cox, N., and Gomer, R.H. (2015). TNF-α–stimulated fibroblasts secrete lumican to promote fibrocyte differentiation. Proceedings of the National Academy of Sciences of the United States of America *112*, 11929.

Popescu, L.M., and Faussone-Pellegrini, M.-S. (2010). TELOCYTES – a case of serendipity: the winding way from Interstitial Cells of Cajal (ICC), via Interstitial Cajal-Like Cells (ICLC) to TELOCYTES. Journal of Cellular and Molecular Medicine *14*, 729.

Qi, Z., Li, Y., Zhao, B., Xu, C., Liu, Y., Li, H., Zhang, B., Wang, X., Yang, X., Xie, W., et al. (2017). BMP restricts stemness of intestinal Lgr5+ stem cells by directly suppressing their signature genes. Nature Communications *8*.

Ramesh, R., Kozhaya, L., McKevitt, K., Djuretic, I.M., Carlson, T.J., Quintero, M.A., McCauley, J.L., Abreu, M.T., Unutmaz, D., and Sundrud, M.S. (2014). Proinflammatory human Th17 cells selectively express P-glycoprotein and are refractory to glucocorticoids. The Journal of Experimental Medicine *211*, 89.

Sanz, I., Wei, C., Jenks, S.A., Cashman, K.S., Tipton, C., Woodruff, M.C., Hom, J., and Lee, F.E.-H. (2019). Challenges and Opportunities for Consistent Classification of Human B Cell and Plasma Cell Populations. Frontiers in Immunology *10*.

Scott, C.L., and Omilusik, K.D. (2019). ZEBs: Novel Players in Immune Cell Development and Function. Trends in Immunology *40*, 431–446.

Seo, W., Jerin, C., and Nishikawa, H. (2021). Transcriptional regulatory network for the establishment of CD8+ T-cell exhaustion. Experimental & Molecular Medicine *53*, 202.

Shevyrev, D., and Tereshchenko, V. (2019). Treg Heterogeneity, Function, and Homeostasis. Frontiers in Immunology *10*, 3100.

Shi, M., Lin, T.H., Appell, K.C., and Berg, L.J. (2009). Cell cycle progression following naïve T-cell activation is independent of Jak3/γc cytokine signals. Journal of Immunology (Baltimore, Md.: 1950) *183*, 4493.

Shoshkes-Carmel, M., Wang, Y.J., Wangensteen, K.J., Tóth, B., Kondo, A., Massassa, E.E., Itzkovitz, S., and Kaestner, K.H. (2018). Subepithelial telocytes are an important source of Wnts that supports intestinal crypts. Nature *557*, 242.

da Silva, F.A.R., Pascoal, L.B., Dotti, I., Setsuko Ayrizono, M. de L., Aguilar, D., Rodrigues, B.L., Arroyes, M., Ferrer-Picon, E., Milanski, M., Velloso, L.A., et al. (2020). Whole transcriptional analysis identifies markers of B, T and plasma cell signaling pathways in the mesenteric adipose tissue associated with Crohn's disease. Journal of Translational Medicine 2020 18:1 *18*, 1–14.

Singh, S.P., Zhang, H.H., Foley, J.F., Hedrick, M.N., and Farber, J.M. (2008). Human T-cells That Are Able to Produce IL-17 Express the Chemokine Receptor CCR6. The Journal of Immunology *180*, 214–221.

Smillie, C.S., Biton, M., Ordovas-Montanes, J., Sullivan, K.M., Burgin, G., Graham, D.B., Herbst, R.H., Rogel, N., Slyper, M., Waldman, J., et al. (2019). Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. Cell *178*, 714-730.e22.

Soler, D., Humphreys, T.L., Spinola, S.M., and Campbell, J.J. (2003). CCR4 versus CCR10 in human cutaneous TH lymphocyte trafficking. Blood *101*, 1677– 1683.

Stewart, A., Ng, J.C.-F., Wallis, G., Tsioligka, V., Fraternali, F., and Dunn-Walters, D.K. (2021). Single-Cell Transcriptomic Analyses Define Distinct Peripheral B Cell Subsets and Discrete Development Pathways. Frontiers in Immunology *12*.

Su, G.H., Ip, H.S., Cobb, B.S., Lu, M.M., Chen, H.M., and Simon, M.C. (1996). The Ets protein Spi-B is expressed exclusively in B cells and T-cells during development. The Journal of Experimental Medicine *184*, 203.

Szabo, P.A., Levitin, H.M., Miron, M., Snyder, M.E., Senda, T., Yuan, J., Cheng, Y.L., Bush, E.C., Dogra, P., Thapa, P., et al. (2019). Single-cell transcriptomics of human T-cells reveals tissue and activation signatures in health and disease. Nature Communications 2019 10:1 *10*, 1–16.

Travaglini, K.J., Nabhan, A.N., Penland, L., Sinha, R., Gillich, A., Sit, R. v., Chang, S., Conley, S.D., Mori, Y., Seita, J., et al. (2020). A molecular cell atlas of the human lung from single-cell RNA sequencing. Nature 2020 587:7835 *587*, 619–625.

Uhlitz, F., Bischoff, P., Peidli, S., Sieber, A., Obermayer, B., Blanc, E., Trinks, A., Lüthen, M., Ruchiy, Y., Sell, T., et al. (2021). Mitogen-activated protein kinase activity drives cell trajectories in colorectal cancer. BioRxiv 2020.01.10.901579.

Vandereyken, M., James, O.J., and Swamy, M. (2020). Mechanisms of activation of innate-like intraepithelial T lymphocytes. Mucosal Immunology 2020 13:5 *13*, 721–731.

Viallard, J.F., Bloch-Michel, C., Neau-Cransac, M., Taupin, J.L., Garrigue, S., Miossec, V., Mercie, P., Pellegrin, J.L., and Moreau, J.F. (2001). HLA-DR expression on lymphocyte subsets as a marker of disease activity in patients with systemic lupus erythematosus. Clinical and Experimental Immunology *125*, 485.

Viganò, S., Banga, R., Bellanger, F., Pellaton, C., Farina, A., Comte, D., Harari, A., and Perreau, M. (2014). CD160-Associated CD8 T-Cell Functional Impairment Is Independent of PD-1 Expression. PLoS Pathogens *10*.

Villar, J., and Segura, E. (2020). Decoding the Heterogeneity of Human Dendritic Cell Subsets. Trends in Immunology *41*, 1062–1071.

Willinger, T., Freeman, T., Herbert, M., Hasegawa, H., McMichael, A., and Callan, M. (2006). Human naive CD8 T-cells down-regulate expression of the WNT pathway transcription factors lymphoid enhancer binding factor 1 and transcription factor 7 (T-cell factor-1) following antigen encounter in vitro and in vivo. Journal of Immunology (Baltimore, Md.: 1950) *176*, 1439–1446.

Willis, S.N., Tellier, J., Liao, Y., Trezise, S., Light, A., O'Donnell, K., Garrett-Sinha, L.A., Shi, W., Tarlinton, D.M., and Nutt, S.L. (2017). Environmental sensing by mature B cells is controlled by the transcription factors PU.1 and SpiB. Nature Communications *8*.

Xue, H.-H., Jing, X., Bollenbacher-Reilley, J., Zhao, D.-M., Haring, J.S., Yang, B., Liu, C., Bishop, G.A., Harty, J.T., and Leonard, W.J. (2008). Targeting the GA Binding Protein β1L Isoform Does Not Perturb Lymphocyte Development and Function. Molecular and Cellular Biology *28*, 4300.

Yamauchi, T., Hoki, T., Oba, T., Jain, V., Chen, H., Attwood, K., Battaglia, S., George, S., Chatta, G., Puzanov, I., et al. (2021). T-cell CX3CR1 expression as a dynamic blood-based biomarker of response to immune checkpoint inhibitors. Nature Communications 2021 12:1 *12*, 1–14.

Yamazaki, S., Tanaka, Y., Araki, H., Kohda, A., Sanematsu, F., Arasaki, T., Duan, X., Miura, F., Katagiri, T., Shindo, R., et al. (2017). The AP-1 transcription factor JunB is required for Th17 cell differentiation. Scientific Reports *7*.

Zheng, C., Zheng, L., Yoo, J.-K., Guo, H., Zhang, Y., Guo, X., Kang, B., Hu, R., Huang, J.Y., Zhang, Q., et al. (2017). Landscape of Infiltrating T-cells in Liver Cancer Revealed by Single-Cell Sequencing. Cell *169*, 1342-1356.e16.

Zhou, L. (2015). How smart can it be: transcriptional regulation of T helper cells by SMAR1. Mucosal Immunology *8*, 1181–1183.

Table S4: List of 9-mer peptides from Nelde et al. compared with predictions.

Peptide	Immunogenicity	Predicted Immunogenicity	Immunogenicity Predicted Score	Predicted HLA allele
KLFAAETLK	Positive	Positive	0.38296	HLA.A03.01
VYIGDPAQL	Positive	Positive	0.4416	HLA.A24.02|HLA.C04.01|HLA.C07.01|HLA.C07.02
QYIKWPWYI	Positive	Positive	0.43888	HLA.A24.02|HLA.C04.01|HLA.C07.01|HLA.C07.02
FPRGQGVPI	Positive	Positive	0.48152	HLA.B07.02|HLA.B08.01
QRNAPRITF	Positive	Positive	0.4076	HLA.B07.02|HLA.C04.01|HLA.C07.01|HLA.C07.02
AGDSGFAAY	Negative	Positive	0.39248	HLA.A01.01|HLA.C04.01|HLA.C07.01|HLA.C07.02
FLLPSLATV	Negative	Positive	0.52576	HLA.A02.01|HLA.C04.01|HLA.C07.01|HLA.C07.02
FIAGLIAIV	Negative	Positive	0.90184	HLA.A02.01
FLAFVVFLL	Negative	Positive	0.39192	HLA.A02.01
KLLEQWNLV	Negative	Positive	0.38912	HLA.A02.01
NIVNVSLVK	Negative	Positive	0.38504	HLA.A03.01
ASMPTTIAK	Negative	Positive	0.39936	HLA.A03.01
SVLNDILSR	Negative	Positive	0.44368	HLA.A03.01
ASAFFGMSR	Negative	Positive	0.38808	HLA.A03.01
TPINLVRDL	Negative	Positive	0.44528	HLA.B07.02|HLA.B08.01
VRFPNITNL	Negative	Positive	0.39592	HLA.B08.01|HLA.B40.01|HLA.C04.01|HLA.C07.01|HLA.C07.02
LTDEMIAQY	Positive	Negative	0.22984	HLA.A01.01|HLA.C04.01|HLA.C07.01|HLA.C07.02
ALSKGVHFV	Positive	Negative	0.3552	HLA.A02.01|HLA.C04.01
VYFLQSINF	Positive	Negative	0.22664	HLA.A24.02|HLA.C04.01|HLA.C07.01|HLA.C07.02
TPKYKFVRI	Positive	Negative	0.23216	HLA.B07.02|HLA.B08.01
FVKHKHAFL	Positive	Negative	0.2844	HLA.B08.01|HLA.C07.01|HLA.C07.02
DLKGKYVQI	Positive	Negative	0.20912	HLA.B08.01
EAFEKMVSL	Positive	Negative	0.24656	HLA.B07.02|HLA.B08.01|HLA.C04.01|HLA.C07.01|HLA.C07.02
YYQLYSTQL	Positive	Negative	0.34944	HLA.A24.02|HLA.B08.01|HLA.C04.01|HLA.C07.01|HLA.C07.02
NRFLYIIKL	Positive	Negative	0.1784	HLA.B08.01|HLA.B40.01|HLA.C04.01|HLA.C07.01|HLA.C07.02
RQEEVQELY	Negative	Negative	0.1208	HLA.A01.01|HLA.C04.01
SPDDQIGYY	Negative	Negative	0.20296	HLA.A01.01|HLA.B07.02|HLA.C04.01|HLA.C07.01|HLA.C07.02
YIDIGNYTV	Negative	Negative	0.23104	HLA.A01.01|HLA.A02.01|HLA.C04.01|HLA.C07.01|HLA.C07.02
RLFRKSNLK	Negative	Negative	0.27352	HLA.A03.01
RIAGHHLGR	Negative	Negative	0.36184	HLA.A03.01
NLIIKNLSK	Negative	Negative	0.30688	HLA.A03.01

Table S4: List of 9-mer peptides from Nelde et al. compared with predictions.

Peptide	Immunogenicity	Predicted Immunogenicity	Immunogenicity Predicted Score	Predicted HLA allele
VTYVPAQEK	Negative	Negative	0.17376	HLA.A03.01
ASKIITLKK	Negative	Negative	0.25528	HLA.A03.01
NLIIKNLSK	Negative	Negative	0.30688	HLA.A03.01
GVKHVYQLR	Negative	Negative	0.21416	HLA.A03.01
SSTASALGK	Negative	Negative	0.2852	HLA.A03.01
APFLYLYAL	Negative	Negative	0.34688	HLA.B07.02|HLA.B08.01
EPKLGSLVV	Negative	Negative	0.30968	HLA.B07.02|HLA.B08.01
YLKLRSDVL	Negative	Negative	0.21976	HLA.B08.01|HLA.C04.01|HLA.C07.01|HLA.C07.02
EPVLKGVKL	Negative	Negative	0.22752	HLA.B07.02|HLA.B08.01
IIKNLSKSL	Negative	Negative	0.22776	HLA.B07.02|HLA.B08.01
TLDSKTQSL	Negative	Negative	0.16552	HLA.A01.01|HLA.A02.01|HLA.B07.02|HLA.B08.01|HLA.C04.01|HLA.C07.01|HLA.C07.02
VPMEKLKTL	Negative	Negative	0.31872	HLA.B07.02|HLA.B08.01|HLA.C04.01|HLA.C07.01|HLA.C07.02
GAKLKALNL	Negative	Negative	0.34312	HLA.B08.01
YQKVGMQKY	Negative	Negative	0.24856	HLA.A01.01|HLA.C07.01|HLA.C07.02
VLKGVKLHY	Negative	Negative	0.24	HLA.A01.01|HLA.A03.01
FLYLYALVY	Negative	Negative	0.3372	HLA.A01.01|HLA.A03.01
LVKPSFYVY	Negative	Negative	0.19944	HLA.A01.01|HLA.A03.01|HLA.C07.01|HLA.C07.02
WLSYFIASF	Negative	Negative		Predicted nonbinder
KVSIWNLDY	Negative	Negative	0.18952	HLA.A01.01|HLA.A03.01
RQEEVQELY	Negative	Negative	0.1208	HLA.A01.01|HLA.C04.01
IQYIDIGNY	Negative	Negative	0.2272	HLA.A03.01
LLNKHIDAY	Negative	Negative	0.21096	HLA.A01.01|HLA.A03.01
NYMPYFFTL	Negative	Negative	0.2984	HLA.A24.02|HLA.B08.01|HLA.C04.01|HLA.C07.01|HLA.C07.02
IRQEEVQEL	Negative	Negative	0.2604	HLA.C04.01|HLA.C07.01|HLA.C07.02
EYHDVRVVL	Negative	Negative	0.20432	HLA.A24.02|HLA.B08.01|HLA.C04.01|HLA.C07.01|HLA.C07.02
KKADETQAL	Negative	Negative	0.26312	HLA.C07.01|HLA.C07.02
VYDPLQPEL	Negative	Negative	0.32408	HLA.A24.02|HLA.C04.01|HLA.C07.01|HLA.C07.02
IYNDKVAGF	Negative	Negative	0.19256	HLA.A24.02|HLA.C04.01|HLA.C07.01|HLA.C07.02

Immunogenicity refers to immunogenicity by either CD4+ or CD8+ T-cells characterized by IFNγ ELISPOT screening. Predicted immunogenicity refers to qualitative classification of immunogenicity based on predicted score of Repitope. Predicted HLA alleles lists all HLA alleles out of 10 HLA alleles (HLA-A*0101, 0201, 0301, 2402, HLA-B*0702, 4001, 0801 and HLA-C*0702, 0401, 0701

alleles) predicted to bind the particular peptide by NetMHCpan 4.0 prediction. List of 9-mer peptides from Nelde et al. compared with predictions.