Deep learning methods for Immunotherapy

Schaap-Johansen, Anna-Lisa

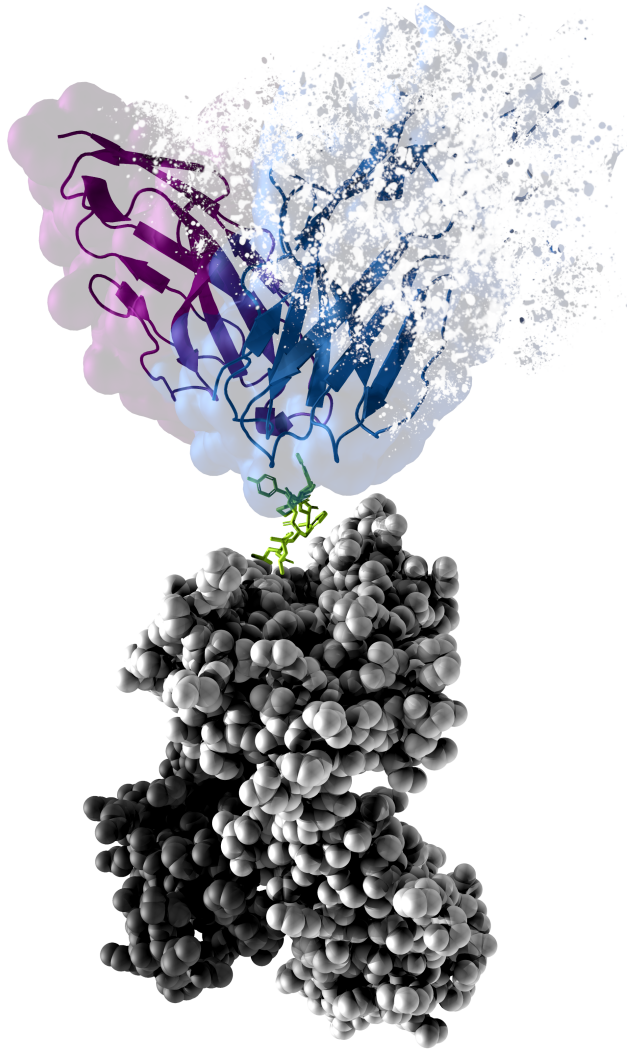
Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

DTU Health Tech
Department of Health Technology

Deep Learning methods for Immunotherapy

PhD Thesis



Anna-Lisa Schaap-Johansen
February 2022

Deep Learning methods for Immunotherapy

Anna-Lisa Schaap-Johansen

February, 2022

Contents

Abbreviations

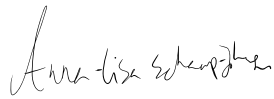
ADAM	Adaptive moment estimation
AUC	Area under the ROC curve
CD	Cluster of differentiation
CDR	Complementarity determining regions
CNN	Convolutional Neural Network
D	Diversity
ERGO-I	Peptide TCR matching prediction
ERGO-II	Peptide TCR matching prediction
FN	False negative
FNN	Feed forward Neural Network
FP	False positive
FPR	False positive rate
HLA	Human leukocyte antigen
ImRex	interaction map recognition
J	Joining
LSTM	Long Short Term Memory
LYRA	Lymphocyte receptor automated modeling
MCC	Matthews Correlation Coefficient
MHC	Major histocompatibility complexes
RNN	Recurrent Neural Network

ROC	Receiver operator characteristic
SGD	Stochastic gradient descent
Tanh	Hyperbolic tangent
TCR	T cell receptor
TITAN	TCR epitope bimodal attention networks
TN	True negative
TP	True positive
TPR	True positive rate
V	Variable

Preface

The work presented in this thesis was carried out at the Department of Health Technology in the AI for Immunological Molecules (AIM) group at the Technical University of Denmark under associate professor Paolo Marcatili's main supervision and co-supervision of professor Xiangdong Fang and associate professor Simon Rasmussen. The work presented was carried out between January 2018 and February 2022.

The thesis consists of a general introduction explaining the essential concepts needed to understand the scope of the thesis, one peer-reviewed publication, one manuscript in preparation with the abstract accepted in a journal, one manuscript in preparation, and an epilogue.



Kongens Lyngby, February 2022
Anna-Lisa Schaap-Johansen

Abstract

The immune system is instrumental in recognizing and defending the body from infections or malfunctioning cells. Although we have improved our understanding of the immune system substantially in the last decades, many questions remain to be answered. Computational tools can play a major role in helping us obtain a better insight into the immune system and be instrumental in improving therapies and diagnostic routines. Immunotherapy is one of the novel fields that computational tools have supported. The core concept of immunotherapy is to exploit the patient's own immune system to fight cancer and other diseases by eliciting or suppressing immune responses targeted at specific molecules. A commonly used strategy in immunotherapy is to identify cancer-specific peptides - named neoantigens/neoepitopes - that can be recognized and targeted by the immune system.

In the first part of this thesis, we present a review paper that provides readers with a general overview of current computational tools for predicting T cell epitopes and neoepitopes, guiding the reader through their potential uses, the data needed, and their advantages and disadvantages. The work also discusses potential future perspectives, uncovering potentially important directions for people to take going forward.

A key element of immunotherapy is the ability to elicit an effective and targeted immune response. T cells carry out different roles in the immune response; some actively find and eliminate infected or pathogenic cells (CD8+ T cells), while others regulate the overall immune response (CD4+ T cells). However, our understanding of the genesis and action modes of such cell types is still limited.

In the second paper, we present a model developed to predict the lineage of a T cell, whether it is a CD8+ or CD4+ T cell, from its T cell receptor. We also discuss the possibility that not all T cell receptors may be specific for a certain lineage but exhibit plasticity in their lineage choice.

The T cell receptor expressed on the surface of T cells interacts with peptides presented by the major histocompatibility complex (MHC) found on the surface of specific cells. Upon recognition of a MHC presented peptide, an immune response will be elicited. We still do not fully comprehend which complexes a T cell receptor will interact with, and what differentiates a binding from a non-binding complex.

In the third project of this thesis, we provide a study showing that energies calculated from modeled structures carry some predictive power in differentiating binding from non-binding complexes. We also show that it is challenging to identify generalizing patterns across peptides due to the absence of clear sequence patterns that can distinguish binders from non-binders.

We hope that the research conducted in this thesis will provide valuable insights regarding T cell receptors and that this research can be used as a stepping stone to improve immunotherapy in the future.

Dansk resumé

Immunsystemet er medvirkende til at genkende og forsvare kroppen mod infektioner eller dårligt fungerende celler. Selvom vi har forbedret vores forståelse af immunsystemet væsentligt i de sidste årtier, er der stadig mange spørgsmål, der mangler at blive besvaret. Beregningsværktøjer kan spille en stor rolle i at hjælpe os med at opnå en bedre indsigt i immunsystemet samt være medvirkende til at forbedre terapier og diagnostiske rutiner. Immunterapi er et af de nye områder, som beregningsværktøjer har understøttet. Kernekonceptet for immunterapi er at udnytte patientens eget immunsystem til at bekæmpe kræft og andre sygdomme ved at fremkalde eller undertrykke immunresponser rettet mod specifikke molekyler. En almindeligt anvendt strategi i immunterapi er at identificere kræftspecifikke peptider - kaldet neoantigener/neoepitoper - som kan genkendes og målrettes af immunsystemet.

I den første del af denne afhandling præsenterer vi et gennemgangspapir, der giver læserne et generelt overblik over aktuelle beregningsværktøjer til at forudsige T-celle epitoper og neoepitoper, som guider læseren gennem deres potentielle anvendelser, de nødvendige data og deres fordele og ulemper. Arbejdet diskuterer også potentielle fremtidsperspektiver og afdækker potentielt vigtige retninger som man bør tage fremadrettet.

Et nøgleelement i immunterapi er evnen til at fremkalde et effektivt og målrettet immunrespons. T-celler udfører forskellige roller i immunresponset; nogle finder og fjerner aktivt inficerede eller patogene celler (CD8+ T-celler), mens andre regulerer det overordnede immunrespons (CD4+ T-celler). Vores forståelse af disse celletypers tilblivelse og handlingsmåder er dog stadig begrænset.

I den anden artikel præsenterer vi en model udviklet til at forudsige afstamningen af en T-celle, hvad enten det er en CD8+ eller CD4+ T-celle, fra dens T-cellereceptor. Vi diskuterer også muligheden for, at ikke alle T-cellereceptorer kan være specifikke for en bestemt afstamning, men udviser plasticitet i deres afstamningsvalg.

T-cellereceptoren udtrykt på overfladen af T-celler interagerer med peptider præsenteret af histokompatibilitetskompleks (MHC), som findes på overfladen af specifikke celler. Ved genkendelse af et MHC præsenteret peptid vil et immunrespons blive fremkaldt. Vi forstår stadig kun til dels, hvilke komplekser

en T-cellereceptor vil interagere med, og hvad der adskiller et bindende fra et ikke-bindende kompleks.

I det tredje projekt i denne afhandling præsenterer vi en undersøgelse, der viser, at energier beregnet ud fra modellerede strukturer har en vis forudsigelsesevne til at differentiere bindende fra ikke-bindende komplekser. Vi viser også, at det er udfordrende at identificere generaliserende mønstre på tværs af peptider på grund af manglende klare sekvensmønstre, der kan skelne bindere fra ikke-bindere.

Vi håber, at forskningen udført i denne afhandling vil give værdifuld indsigt vedrørende T-cellereceptorer, og at denne forskning kan bruges som et springbræt til at forbedre immunterapi i fremtiden.

Acknowledgements

First, I would like to thank my supervisor Paolo Marcatili. I have learned a lot during my time as a PhD student and I would like to express my gratitude for going on this journey with me.

I would like to thank my co-supervisors Fang Xiangdong and Simon Rasmussen for the interesting scientific discussions.

My thanks to Siqi Liu from Beijing Genomics Institute for our interesting scientific discussions and for hosting me in China. It was truly interesting to spend time in your lab.

I would also like to thank Sino-Danish center for making this PhD possible by funding it with the SDC grant.

Thanks to both the previous and current members of AI for Immunological Molecules. My gratitude especially goes to Milena Vujović and Magnus Haraldson Høie for the interesting days and evenings filled with fun and science.

I would like to thank my office mates and especially Marianne, Kristine, Birkir, Sara, Nicola, Allessandro and Mona for all the nice conversations when we needed a break, scientific discussions and support during this time.

Also thanks to Vanessa, Kamilla, Martin and Jose for the collaborations, support and being willing to answer any questions I ask regardless.

The whole Bioinformatics department for the great and insightful discussions, great Friday breakfasts and friendly environment. It has been a pleasure to get to know you all.

My parents, my family and my friends for all the support and affection. Especially my mother who is always there for me when I need her. I am very lucky to have you as my mother.

Finally I would like to thank Antonin, my companion and partner. I am so happy that this PhD allowed me to meet you. Thank you for all your adorable humor, care, love and support during this period. Thank you for helping me with the figures and taking them to the next level. I look forward to the days and experiences that lay ahead of us.

Papers included in the thesis

- **Anna-Lisa Schaap-Johansen**, Milena Vujovic, Annie Borch, Sine Reker Hadrup and Paolo Marcatili.
  T cell Epitope Prediction and Its Application to Immunotherapy. Frontiers in Immunology 12 (2021): 2994.

- **Anna-Lisa Schaap-Johansens**, , Kamilla Kjærgaard Munk,Martin Closter Jespersen, Vanessa Isabell Jurtz, Tina Funck and Paolo Marcatili.
  Can we predict T cell lineage from sequence only? Abstract accepted Frontiers in Immunology, manuscript in preparation.

- **Anna-Lisa Schaap-Johansen** and Paolo maractili.
  Global energy terms for improved TCR-pMHC binding prediction. Manuscript in preparation.

Papers not included in the thesis

- Milena Vujovic, Kristine F. Degn, Frederikke I. Marin, **Anna-Lisa Schaap-Johansen**, Benny Chain, Thomas L. Andresen, Joseph Kaplinsky and Paolo Marcatili.
  T cell receptor sequence clustering and antigen specificity. Computational and Structural Biotechnology Journal 18 (2020): 2166-2173.

- Jon Ashley, **Anna-Lisa Schaap-Johansen**, Mohsen Mohammadniaei, Maryam Naseri, Paolo Marcatili, Marta Prad and Yi sun.
  Terminal deoxynucleotidyl transferase-mediated formation of protein binding polynucleotides. Nucleic acids research 49, no. 2 (2021): 1065-1074.

- Anna Vardi, Andreas Agathangelidis, Sofia Gkagkaridou, **Anna-Lisa Schaap-Johansen**, Maria Karipidou, Anna Boukla, Asimina Fylaktou, Niki Stavroyianni, Michail Iskas, Achilles Anagnostopoulos, Sine Reker Hadrup, Anastasia Chatzidimitriou, Paolo Marcatili, Kostas Stamatopoulos.
  The clonotypic BCR IG of CLL Patients Contain Predicted T-Cell Class I Epitopes with Shared Structural Properties. Blood vol. 138 (2021): 1540-1540.

- Julio Vacacela, **Anna-Lisa Schaap-Johansen**, Patricia Manikova, Paolo Marcatili, Marta Prado, Yi Sun and Jon Ashley.
  The Protein-Templated Synthesis of Enzyme-Generated Aptamers. Angewandte Chemie international Edition (2022), e202201061.

- **Anna-Lisa Schaap-Johansen** and Paolo Marcatili.
  A computational pipeline for predicting cancer neoepitopes. Book chapter in submission.

1 Introduction

1.1 Scope of thesis

This thesis focuses on applying machine learning models to deepen our understanding of immunological bioinformatics by tackling a subset of the current deficiencies in the research field. The immune system is one of the most complex biological systems to study. Even today, all the cellular functions are yet to be fully comprehended. One such area is the adaptive immune system, which is an essential part of the immune system. The primary purpose of the adaptive immune system is to ensure that the host is healthy by detecting and eliminating both malfunctioning cells and pathogenic infections (1). One of the essential elements in the adaptive immune system is the T cell. The T cell identifies abnormal cells by utilizing T cell receptors expressed on their surface, which interact with peptides presented by the major histocompatibility complexes (MHCs) found on the surface of specific cells such as antigen presenting cells.

Two main types of T cells exist, namely CD4+ and CD8+ T cells. These T cells have different functions in the adaptive immune system and interact differently with cells in the body. However, what determines the lineage and defining characteristics of a T cell from a given lineage is a field under active research. Appreciation of what distinguishes a T cell in one lineage from a T cell in another can potentially provide us with information depicting how a T cell engages with cells and change our perception of what the T cell recognizes. Therefore, one of the areas this thesis investigates is the lineage of a T cell, aiming to understand if it can be determined based on the T cell receptor (TCR).

T cells become activated and proliferate upon recognizing presented peptides. Once activated, the T cells can set processes in motion to eliminate malfunctioning or pathogen-infected cells (1). Peptides capable of inducing an immune response are known as T cell epitopes. However, the presentation of a peptide does not necessarily entail a T cell driven immune response. A

better understanding of what drives a T cell immune response can help improve T cell-based immunotherapies with the objective of either activating or suppressing the body's own immune system to help treat disease (2). T cell epitopes are mainly discovered through experimental methods; this is both expensive and time-consuming. Therefore, developing more cost-effective, less time-consuming, and more reliable tools for predicting T cell epitopes is of great interest for both the industry and academic research.

This thesis seeks to investigate T cell-based immunotherapy and the central workings of the current approaches and tools. A range of tools have been made to predict which peptides will elicit an immune response; however, most of these methods focus on MHC presentation. Therefore, in one of the projects included in this thesis, we try to address this area by studying whether it is possible to improve the prediction capabilities using structural energy calculations and the TCR sequence information.

The overall aim of this thesis was to study T cells and their receptors. This was studied to develop methods for deepening our understanding and improving the prediction of T cell recognition, aiming to improve T cell based immunotherapy.

1.2 Structure of the Thesis

The thesis is divided into seven chapters subsequent to this. The first four, chapters 2, 3, 4, and 5 are theoretical explanatory chapters that set the scene for the published papers and explain the underlying concepts. Chapters 6, 7, and 8 present the research conducted during this Ph.D., and chapter 9 summarizes and provides future perspectives.

Chapter 2 covers the background of the immune system, with a focus on the adaptive system's functionality and interactions.

Chapter 3 covers the background of machine learning, expanding on the different methods utilized throughout the PhD.

Chapter 4 covers different ways of representing data and ways of evaluating machine learning methods.

Chapter 5 covers T cell based immunotherapy and the current tools within the field of T cell epitope prediction for T cell-based immunotherapy.

Chapter 6 introduces the first scientific paper included in this thesis. The main aim of this publication was to review T cell based immunotherapy and the tools currently available within this field.

Chapter 7 presents an ongoing project where the abstract has been accepted. This project investigates the possibility of predicting whether a TCR is from the CD4+ or CD8+ T cell lineage based on the TCR sequence using convolutional neural networks (CNNs). Furthermore, the manuscript also aims to create a discussion regarding the potential plasticity of T cells.

Chapter 8 presents the second ongoing project included in the thesis. This project aims to study whether the prediction of T cell recognition can be improved by including structure-based energy terms in the prediction method.

Chapter 9 provides a summary of the thesis, reflects on all three projects, and provides future perspectives.

2 The immune system

The immune system is the essential component of homeostasis. The primary role of the immune system is to protect against outside intruders such as viruses, bacteria, organisms, or other agents causing disease, which can collectively be referred to as pathogens. The immune system can generally be divided into two subgroups, the innate and the adaptive immune systems. The innate immune system is considered fast and the first line of defense. It engages in a nonspecific manner and provides a more general defense against pathogens. Although the innate immune system is regarded as nonspecific, it is still a very powerful system, capable of effectively discriminating between host cells and pathogens. The innate immune system utilizes germline-encoded receptors capable of recognizing features that are common to many pathogens and can therefore recognize broad classes of pathogens. While the innate immune system can target a broad class of pathogens, the germline-encoding of the receptors restricts ther adaptability to recognize more diverse pathogens. Unlike the innate immune system, the adaptive immune system can target the more diverse pathogens not covered by the innate immune system, launching a very precise response against these particular pathogens. However, the response from the adaptive immune system is slow to develop upon first exposure to new pathogens and it is very energy consuming (1).

2.1 The adaptive immune system

The adaptive immune system can additionally be divided into two groups responsible for the humoral immune response and the cell-mediated immune response. Whilst there is some overlap between these two groups, such as both of them being a type of lymphocytic cell interacting with other lymphocytic cells and originating in the bone marrow, they differ in their function and how they further develop. The humoral response is primarily driven by what is known as B cells. This type of cell originates and develops in the bone marrow, thus the name B cells. The main focus of the humoral response is to target extracellular pathogens. On the other hand, the cell mediated response focuses on inspecting and identifying aberrant cells as in the case of cancer cells, or cells

with intracellular pathogens, due to viral or bacterial infections. Cells from the mediated response, unlike B cells, do not develop in the bone marrow, but instead migrate from the bone marrow to the thymus for further maturation (1), thereby naming this type of cells T cells. The main aim of this thesis was to develop prediction methods to improve our understanding of T cells and how they interact. The following sections will therefore give an overview of the important components that play a part in what defines the T cell and its interaction.

2.2 T cell lineage

The cell-mediated response can be roughly divided into two overall lineages, determined by whether a T cell expresses a cluster of differentiation (CD) 8 or 4 co-receptor. The expressed CD8 or CD4 co-receptor, together with a protein complex also present on the surface of a T cell named a T cell receptor (TCR), interacts with cells to determine which cells are abnormal (3). We will elaborate further on cell abnormality in the first article of the thesis in chapter 6. Interestingly, before differentiating into either the CD8 or CD4 lineage, all T cells start out as double-positive T cells, expressing both the CD8 and CD4 co-receptor on their cell surface (4). One could assume that there may be a clear distinction between the two lineages. However, as we will cover in more depth in the manuscript in chapter 7, this distinction may not be that clear-cut. The key interaction to distinguish abnormal from normal cells is to be found between the TCR and peptides bound and presented by the Major Histocompatibility Complex (MHC). It is generally believed that the TCR together with the CD8 or CD4 co-receptor expressed on the surface of the T cell determines whether the T cell will interact with MHC class I or class II bound peptide complexes, respectively as shown in figure 2.1. MHC proteins are expressed on the surface of cells and differ in i) which cells present them dependent on the MHC class, where MHC class I is present on all nucleated cells, whereas MHC class II molecules are only expressed by antigen-presenting cells (5) and ii) how the peptides they present are obtained. The MHC class I molecules display intracellularly derived peptides, whereas peptides presented by MHC class II are mainly derived from extracellular proteins. It is commonly believed that CD8+ T cells interact with peptide-bound MHC class I complexes and, upon engagement, become activated cytotoxic T cells specialized in killing the target cells. CD4+ T cells instead primarily engage with peptide-bound MHC class II complexes. Upon contact, the majority of CD4+ T cells either become activated T helper cells, known for stimulating activated CD8+ T cell

expansion and B cell development, or T regulatory cells, which induce tolerance by suppressing the immune response against self (6).

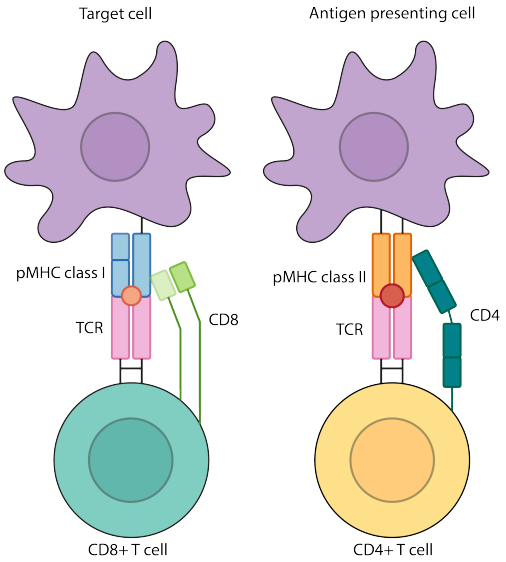


Figure 2.1. An illustration showing on the left side a CD8+ T cell binding to a MHC class I presented peptide and on the right side a CD4+ T cell binding to MHC class II presented peptide.

2.3 T cell receptor structure

The TCR is a heterodimeric protein expressed at the cell membrane. This heterodimeric complex consists of two transmembrane chains capable of recognizing peptides presented by an MHC complex with the interaction stabilized by the CD8 or CD4 co-receptor. Two kinds of TCRs can be expressed at the cell membrane, and these are defined by the component chains, which can be either α/β or γ/δ. Most T cells express TCR composed of α/β chains, with a minority of only about 5% of T cells expressing TCRs with γ/δ chains (7). The scope of this thesis is focused on T cells expressing α/β TCRs; therefore, the following section will be exclusively dedicated to these. The two chains

in the TCR each contain a constant region and an N-terminal variable region. There are three complementarity determining regions (CDRs) located within the variable region, namely CDR1, CDR2, and CDR3. These three regions have a loop structure and are the part of the TCR mainly responsible for the recognition of specific peptide-MHC complexes, shown in figure 2.2.

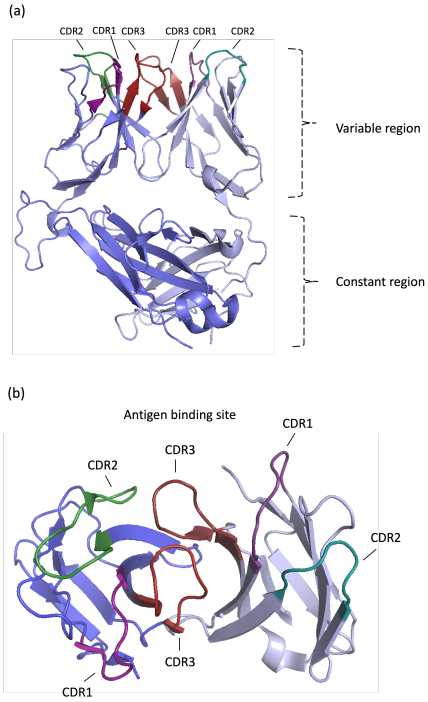


Figure 2.2. An illustration showing the structure of a T cell receptor, (a) provides a side view of the T cell receptor, showing the constant region and variable regions and the highlighted complementarity determining regions CDRs. (b) gives a top view of the CDRs, showing the location which impacts antigen specificity the most. The images were made with PyMOL using the 1OGA PDB structure.

2.4 T cell diversity

In order to recognize a wide variety of pathogens and abnormal cells, an extensive diverse repertoire of T cell receptors (TCR) is necessary. The main driver of this diversity found in TCRs is a mechanism known as V(D)J recombination, which occurs during early T cell development (1). Overall, this process works by recombining gene segments called variable (V), diversity (D), and joining (J) segments, forming the variable domain of a TCR chain. The main drivers of MHC presented peptide recognition are the CDRs, which coincide with being the most variable part of the TCR. As shown in figure 2.3, the CDR regions are encoded differently. The CDR1 and CDR2 regions of both the α and the β chain are both only encoded by V gene segments, whereas the CDR3 regions span more gene segments. In detail, CDR3 on the β chain is encoded with segments from the V, J, and D genes with the addition of nucleotides at the junction between gene segments. The CDR3 of the α chain also spans V and J gene segments with the addition of nucleotides, but unlike the β chain, it does not include a D gene segment. The diversity found in these variable domains is a product of an almost random combination of the multiple different variations of the V, D, and J genes, consolidated by the addition of nucleotides added to the junction between the gene segments (8).

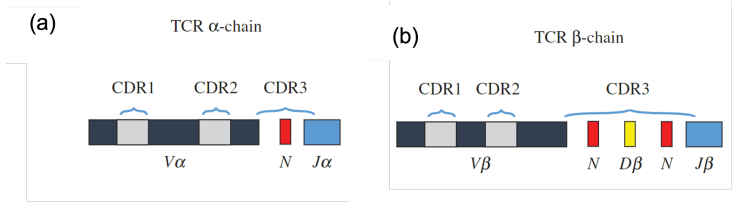


Figure 2.3. This illustrates V(D)J recombination, which rearranges the Variable (V), Joining (J) and Diversity (D) gene segments to create variability in TCR receptors. This figure shows the process for the CDRs which impacts TCR specificity the most. Figure adapted from Laydon et al. 2015 (8)

2.5 TCR recognition of the peptide-MHC complex

T cells expressing the CD8 co-receptor are often associated with MHC class I presented peptide recognition. Contrarily to MHC class II bound peptides,

which are longer and often more variable, peptides presented by the MHC class I molecule are typically between 8 to 14 amino acid residues long, with 9 residue peptides being one of the most abundant lengths (9). The MHC class I molecule is heterodimeric and composed of two polypeptide chains: the $\beta 2$ microglobulin and the membrane-spanning α chain. Unlike the $\beta 2$ microglobulin encoded by the B2m gene, the α chain is very polymorphic and is encoded in humans by the human leukocyte antigen (HLA) locus (10). The alpha chain folds into three domains, namely $\alpha 1$, $\alpha 2$, and $\alpha 3$. The region between the $\alpha 1$ and $\alpha 2$ domains is known as the peptide-binding groove, and it is this binding groove that restricts peptide length, as shown in figure 2.4. In MHC class I molecules the binding groove has narrow ends, which forces residues of the peptide up to accommodate the length of the peptide as it increases. The bulged conformation generally assumed by the peptides can then be recognized by TCRs (11).

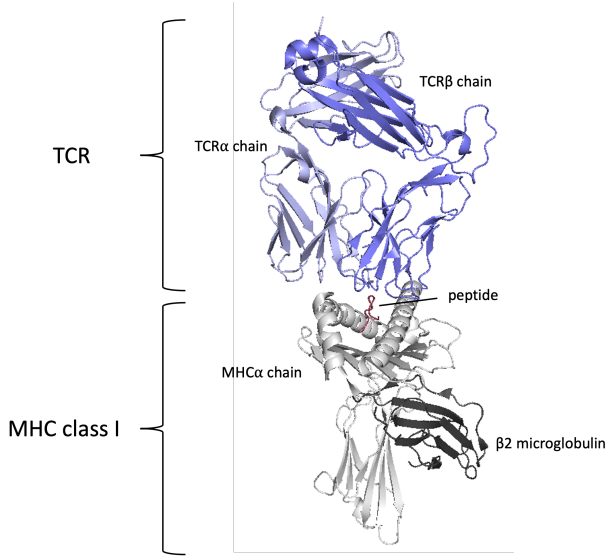


Figure 2.4. A structural representation of a T cell receptor (TCR) binding to a Major histocompatibility complex (MHC) class I presented peptide. This image was made with PyMOL using he 1OGA PDB structure.

The CDR1, CDR2, and CDR3 are the most variable regions of the TCR and mainly drive the interaction with peptide-MHC complexes, as stated previously. The CDR3 loops, and especially the CDR3 β loops, are very diverse, and the part of the TCR accounting for most of the peptide recognition and specificity. The CDR1 and CDR2 loops have less variability and are the regions of the TCR, which primarily interact with the MHC (12) as is shown in figure 2.4. The impact of the CDR loops regarding TCR peptide-MHC recognition makes them an important element to include when predicting TCR peptide-MHC interaction.

The structure of TCR peptide-MHC (TCRpMHC) complexes can provide an additional level of information to models predicting TCR peptide-MHC interaction that can potentially improve the prediction capabilities of these models. However, the hypervariability of the CDRs makes it challenging to model TCRs, and methods modeling TCR structures are scarce. One such tool is Lymphocyte receptor automated modeling (LYRA), developed by Klausen et al. in 2015 (13), which uses templates from a TCR database to model TCRs. A tool developed by Jensen et al. in 2019, named TCRpMHCmodels (14), models TCRpMHC complexes by combining modeled TCRs generated using LYRA with modeled peptide-MHC complexes from MODELLER (15), generated using peptide-MHC templates from a peptide-MHC database.

Naturally, not all peptides trigger a cellular immune response mediated by T cells. In most cases, it is more common for parts of an antigen known as an antigenic determinant or epitope to trigger a T cell response. However, in certain instances, such as with autoimmune disease, peptides categorized as self-peptides presented by healthy cells can provoke an unwanted T cell response as well (16). Therefore what defines an epitope and what triggers a T cell-mediated immune response has shown to be a challenging task to predict. In addition to being a difficult task to predict, it is also a field where significant progress has been slowed down by the limited data linking TCR sequences to their target peptide-MHC complexes.

3 machine learning

Machine learning is a field of study which develops computational algorithms capable of learning the patterns in a dataset without following explicit instructions. The field of machine learning has seen rapid development in the complexity of the models over the years. Earlier models such as decision trees and simple neural networks have developed into more complex models such as random forests and neural network-based Deep learning models (17). The development of newer and more complex models is mainly due to i) improved hardware technologies and ii) the increase in data availability. The new computational hardware is continuously improving and has already reached a state where it is possible to train networks on millions of data points in just a couple of hours or days. The exponential increase in data is also a critical factor in the development of models regarding their applicabilities and the possibilities (18). The field of machine learning can overall be divided into two main groups; supervised and unsupervised learning. Supervised learning utilizes labeled data, meaning knowledge of the true target value of given data input. The algorithm approximates the most optimal function to describe the data by comparing and optimizing the model towards the true labels. Unsupervised learning instead learns how to group the data purely based on the data input itself, without the assistance of labels. One of the more widely used methods within unsupervised learning is clustering, which clusters data based on similarities between the data points.

3.1 Random forest

Bioinformaticians use machine learning for their predictive power and to better understand the biology driving the prediction. Models such as neural networks are often considered a black box since it can be complicated to decipher how the model arrived at its predictions, where models such as Random Forest allow for a more straightforward attainable insight into the model's decisions, making it an important and often used method within bioinformatics.

The random forest is a supervised learning algorithm, which operates as an ensemble, consisting of an often large number of individual diverse decision

trees. Decision Trees work by splitting the data conditioned on one of the input features. The split generates two or more branches as output; each branch will continue splitting in an iterative manner until the data is exhausted, generating a tree-like structure. A visualization showing the decision path and the tree like structure of a decision tree can be seen in figure 3.1. Decision trees split the data based on the variables which can most efficiently split the data to match the true target (19). A split is decided based on an estimated change of impurity between classes, where the split leading to the lowest reduction in impurity is chosen using the Gini index (20). This way, the tree can discover which features are more informative regarding the different classes and thereby decide which features can most cleanly divide the data into the different classes. This setup makes decision trees very good at predicting on the training data, but they often fail to generalize well to new data (21).

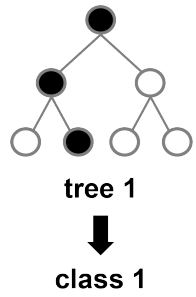


Figure 3.1. An illustration of a decision tree. The colored circles indicate the decision path taken to make the final prediction, where in this case the tree predicts class 1 for the given input.

The random forest, developed in 1995 by Tin Kam Ho (22) and further optimized in 2001 by Leo Breiman (23), aims to rectify the inability of the individual tree to generalize by using the concept of - *the wisdom of the crowds*. During training of the random forest, subsets of features and samples are randomly selected to build multiple decision trees as can be seen in figure 3.2. Each decision tree makes a prediction of the class for a given sample, and the class predicted by a majority of trees, becomes the final random forest prediction for the sample. The random forest is a method that is often used as a baseline and, in many instances it is able to perform comparably to other more advanced methods such as neural networks as well as being computationally fast to train.

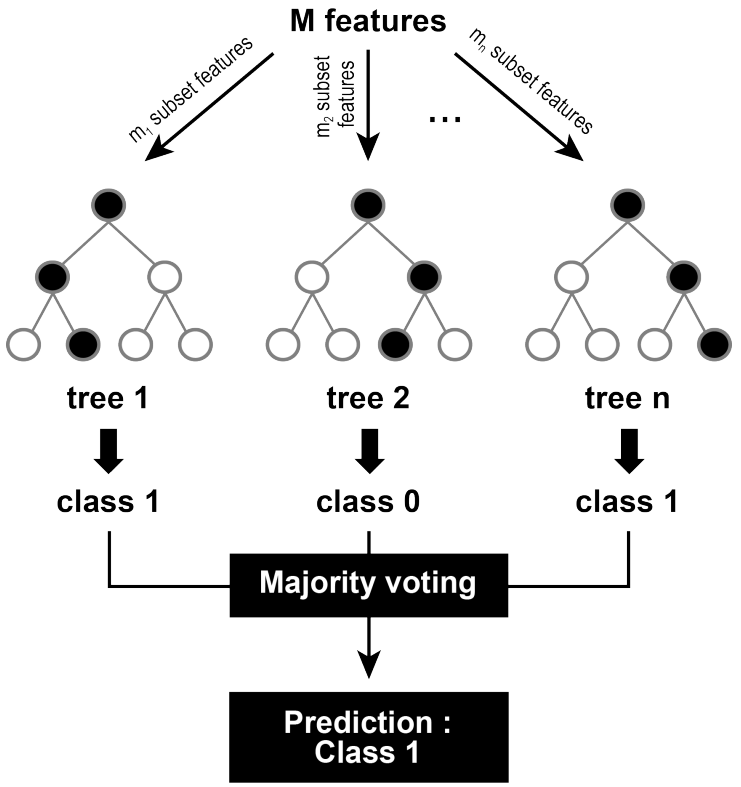


Figure 3.2. This illustrates a random forest consisting of n trees. The colored circles indicate the decision path taken to make the final prediction. The final class determined by the random forest is based on majority voting. In this case most trees voted class 1, making that the prediction for the given input.

3.2 Neural Networks

The idea behind a neural network is to mimic how a human brain utilizes neurons to process information. In an artificial neural network, these neurons attempt to discover any underlying relationships present in the data, aiming to utilize these to gain further understanding or to answer a given question (24).

Neural networks have received increasing interest and development in recent years. Two elements, in particular, have contributed to the growth of this field, with the first element being the advancement and availability of greater computational power and the second element being the development of novel neural network architectures. The increase in computational power has allowed the user to expand the parameter space to contain millions of parameters on substantial datasets within a reasonable time. This enables the user to overcome a major limitation that can greatly impact the type of questions being inquired and answers obtained. In addition to allowing for the parameter space to expand, the greater computational power availability has enabled the development of more complex and computationally intensive neural networks. The evolution and combination of these two elements have helped promote methods achieving increased predictive performances on previously difficult tasks and impacted the shape of fields such as language processing, image processing, bioinformatics, and more (25).

Although deep learning has attracted a lot of attention and development in recent years and is still developing at rapid speed, it is not a new discovery. In 1943 McCulloch and Pitts (26) created a mathematical model mimicking the functionality of neurons found in the brain. They designed an artificial neuron, which fundamentally works by aggregating boolean inputs presented to the model and basing its decision on whether the aggregated value is below or above a given threshold. In 1958 Rosenblatt (27) invented the perceptron, which further evolved the artificial neuron model by adding weights to the inputs, thus allowing for some inputs to be assigned greater importance and introducing the possibility of using non-boolean values as well. These and other discoveries throughout the years have laid the foundation for deep learning and artificial neural networks in general, and have paved the way for the development of newer and more complex types of neural networks (25).

Fundamentally, an artificial neural network is created by combining multiple artificial neurons. An artificial neural network can contain few or multiple layers of artificial neurons. The most basic architecture of an artificial neural network consists of an input layer, a hidden layer, and an output layer. Each layer in a network is comprised of neurons connected by weights. An activation function can be applied to the output of the calculated weighted sum of the input values and an added bias term of a neuron to perform a non-linear transformation to help the network learn complex patterns in the data as shown in figure 3.3.

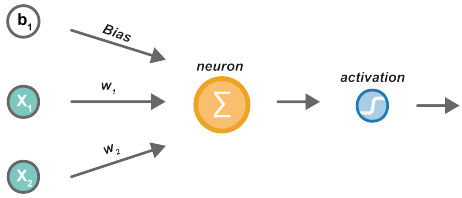


Figure 3.3. An illustration which shows the function of a single neuron. The inputs are weighted with their respective weights, which is then summed together with a bias, and is then fed through an activation function.

Some of the more well used activation functions are hyperbolic tangent (tanh), sigmoid (σ) and rectified linear unit (ReLU) which are illustrated in figure 3.4. The bias value is added to allow for the activation function to be shifted, which adds flexibility with regards to fitting the data better. Depending on the number of hidden layers, a network can either be categorized as a shallow network or a deep network. Often networks with no more than two hidden layers are classified as a shallow network, whereas networks containing more hidden layers than two are considered deep networks.

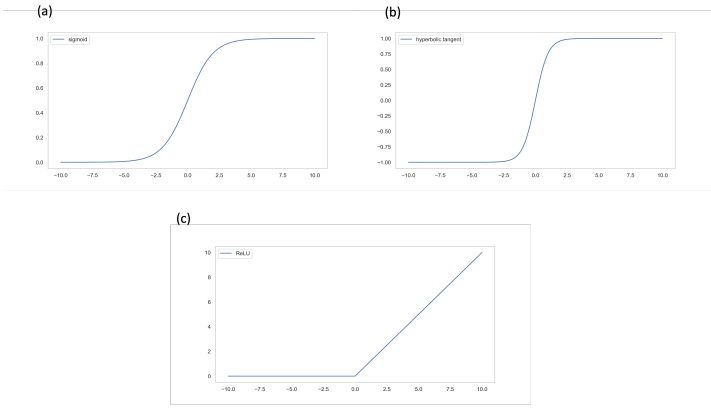


Figure 3.4. This illustrates three common activation functions, with (a) being the Sigmoid function, (b) the tanh function and (c) being the RelU function.

The training of a neural network can be divided into two parts, namely forward and backward propagation (24). The series of calculations performed from when the input is fed into the network, passing through the hidden layers and finally produces a prediction at the output layer is called forward propagation. Backward propagation on the other hand is fundamentally the chain rule of calculus. Essentially it is a method to compare the predicted values with a given label, which is used to calculate gradients for all the weights in the network, with the objective of minimizing error estimated by a loss function. Unlike the forward propagation starting from the input layer and going forward in the network, the backward propagation begins from the output layer and goes backwards in the network. The gradients calculated are used to iteratively optimize the weights in the network by adjusting them in a direction that reduces the error. This process is done by using the gradient descent algorithm, where one of the more well-known methods is the stochastic gradient descent (SGD). Newer and more advanced optimization techniques have been developed, with one of the most widespread being the Adaptive Moment Estimation (ADAM) optimizer (28). Where SGD uses a single never changing learning rate for the weights during training of the network, ADAM instead optimizes a learning rate individually for the different parameters in the network. One of the main drivers behind the popularity of ADAM is that it is easy to use and often leads to fast algorithm convergence. The amount of weight adjustment is controlled by multiplying the gradients with a learning rate. This hyper-parameter determines how big a ”step” the gradient will take towards the minimum determined by a loss function. Lower learning rates lead to a slower travel towards the minimum and vice versa. Although a large learning rate would lead to faster training of a model due to the large step, it will also lead to divergence of the algorithm as the algorithm is constantly “jumping” over the minimum of the function and thus not converging. Contrarily, although using a small learning rate, in general should ensure convergence, a too small learning rate can also take a very long time to converge as well as get stuck on a plateau region.

3.2.1 Feed Forward Neural Networks

Different types of networks exist, but one of the most well-known and widely used networks is the feed-forward neural network (FNN), which is a fully connected network (24). In a fully connected network, every neuron in each layer will be connected to all the neurons in the previous layer as well as all the

neurons in the following layer. FNNs work in a hierarchical fashion, where information flows from the input layer through the hidden layer if present and then ultimately through the output layer. The structure of a simple FNN with one hidden layer is illustrated in figure 3.5.

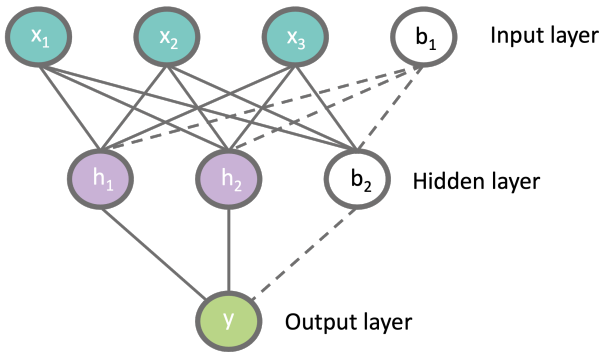


Figure 3.5. A feed forward neural network architecture with an input layer, one hidden layer, bias terms and an output layer.

FNNs are primarily used for supervised learning tasks and although it is a widely used model which is fairly straightforward to understand and quick to train, it has some limitations. FNNs are not designed to retain any spatial information, and thus loses any contextual information. This can be rather problematic when working with biological sequences as the residue placement in the sequence and in regards to each other can have a big impact.

3.2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs), inspired by the visual cortex (29), are a widely used and well-performing type of neural network which is especially ideal for image processing tasks. Unlike FNNs, where each hidden neuron is connected to all input neurons, the CNN breaks with this full connectivity by being sparsely connected via each hidden unit connecting to a subset of adjacent units instead. This restricted region of connectivity is also known as the receptive field, filter, or kernel (30). The size of the kernel, also called the kernel size, is determined by the number of adjacent units chosen to be processed together, thus deciding the number of neighboring spatial information that a receptive field will cover. What the receptive field processes is also determined

by what is known as the stride; this number decides how many units at a time the receptive field is shifted, ultimately deciding how much the receptive fields should overlap. When sliding the kernel over the positions in the input, an array of numbers is generated, commonly termed an activation map or a feature map as illustrated in figure 3.6. It should be mentioned that the set of weights of a specific kernel remains the same regardless of the position in the input, functioning as a type of feature identifier. This way, the kernel can detect the same features or patterns located at different positions in the input data. Conceptually convolutional filter weights get updated during the CNN training, making them specific to the input data enabling them to detect useful patterns for prediction.

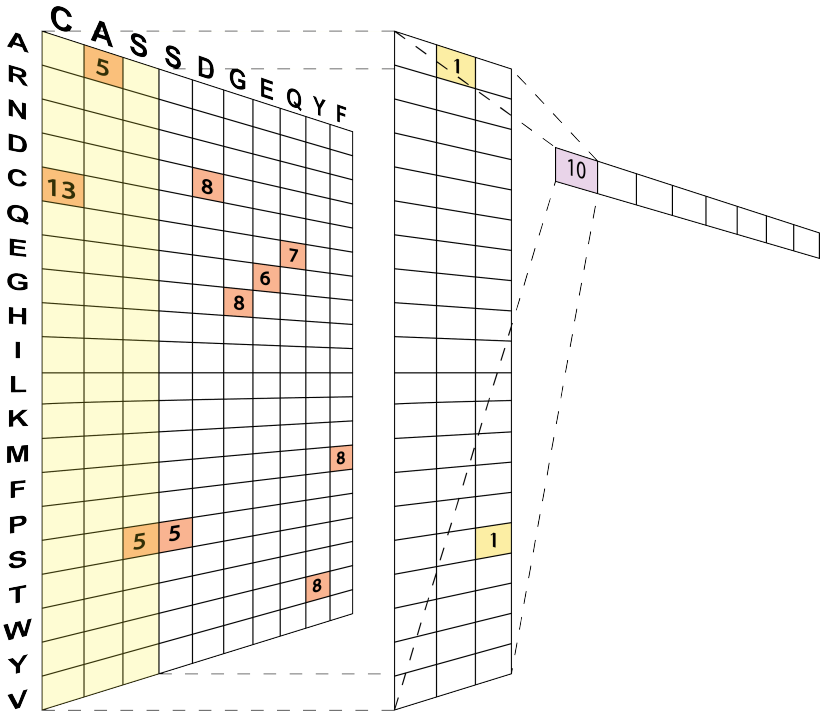


Figure 3.6. An illustration of a BLOSUM encoded CDR3β sequence with a convolutional filter/kernel of size 3 and a stride of 1 moving over it producing a feature map.

Feature maps, however, can have a large spatial dimension, meaning a lot of parameters, which can be computationally heavy and lead to the model learning non-generalizable details and noise in the data, causing the model to overfit. One way to solve this problem and make the feature maps less sensitive to the location of the features is by using pooling layers after the convolutional layers. Different pooling operations exist, where max-pooling is one of the more widely used methods. Max-pooling works by sliding a filter of a specified window width across the feature map selecting the max element from the region covered by the filter. This operation should result in a new feature map only containing the most prominent features from the previous feature map. Ideally, this would produce feature maps with a condensed resolution, eliminating irrelevant details and extracting only the most important features. In addition to creating a feature map that should be more robust to any potential changes in the position of features in an image, they can also enable CNNs to process inputs of different lengths. For networks such as FNNs and CNNs, using inputs with different spatial sizes will output a different number of features after being processed. Unlike an FNN, a CNN can have a pooling implemented to reduce the number of features to a specific size when extracting relevant features, thereby equalizing the number of features across samples with variable lengths.

CNNs are also rapid to train and produce feature maps with distilled features that ideally contain short-range context-dependent input data representations. The speed lies in the fact that the computation of each filter across the input is parallelizable. Although fast to train and capable of identifying short-range dependencies, one major limitation of CNNs is their inability to model long-range context. When working with sequencing data, residues at positions far away from each other can hold important information, but any signal spanning more positions than what a filter covers will not be detected. This can be dealt with by using neural networks capable of detecting long-range dependencies, such as networks implementing recurrent connections. It can still be of interest to cover short-range dependencies in a sequence, and it can therefore be an advantage to take the output from the CNNs and feed it into a neural network with recurrent connections.

3.2.3 Long Short-Term Memory Neural Networks

Long Short-Term Memory neural networks (LSTMs) (31) are a type of network with recurrent connections, making them better at processing temporal infor-

mation than CNNs. LSTMs are an advanced version of Recurrent Neural Networks (RNNs) (32), and thus an elaboration regarding RNNs is needed in order to understand the LSTM. RNNs like a FNN also utilize the backward propagation algorithm to calculate the gradients in the network. However, unlike the backward propagation algorithm implemented in regular FNNs, backward propagation in RNNs has time dependency in the algorithm. This backward propagation algorithm is commonly known as backward propagation through time. Conceptually, the main difference between the two algorithms is the fact that computing the gradient for a given state requires the computation of all the previous states as well. This works by "unfolding" the RNN loop in time, where the three-layer structure in the RNN, the input layer, hidden state, and output layer exists in an amount equal to the number of positions in a given sequence. The network can in simple terms be considered as multiple copies of the same structure, where each passes a message to a successor in the network. Thereby the current output predicted depends on the current state as well as the previous states. Backward propagation is applied across the unfolded RNN, where the errors are accumulated for each timestep, and weights are updated through the network once the network has been "rolled back up". An illustration of this can be found in figure 3.7.

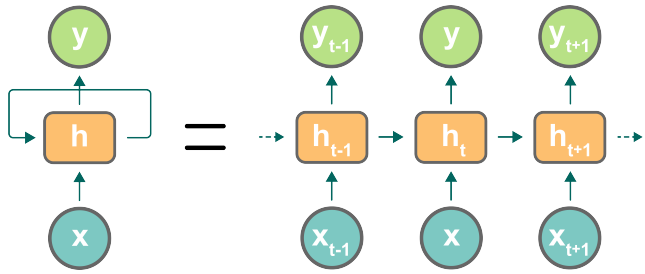


Figure 3.7. The concept of the recurrent neural network when "rolled up" as can be seen on the left and how it looks "rolled out", as can be seen on the right of the figure.

In a standard RNN, the repeating modules will have a single layer, such as a tanh layer adding nonlinearity to the input. A single RNN cell is visualized in figure 3.8, showing the input being multiplied with the previous output and thereafter being fed through a tanh activation function, which is then passed on to the next state.

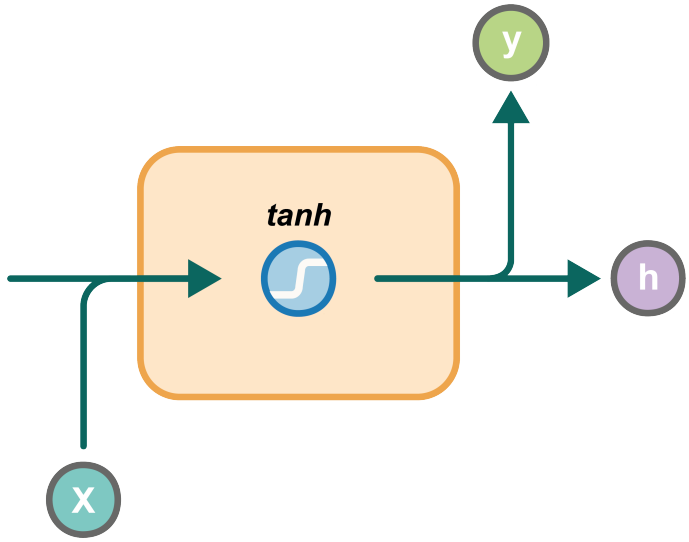


Figure 3.8. The recurrent neural network cell state, where output is fed into the next step in the network. Here x is the input, h is the hidden state and y is the prediction.

The implementation of the chain rule in regular RNNs allows for undesired events such as vanishing or exploding gradients to occur. The length of a sequence determines the number of matrix multiplications a gradient must go through, where the longer a sequence is the more calculations are necessary. In cases where the gradients have a value smaller than one, the gradient will shrink exponentially, which will result in the gradients having values nearing zero, also known as vanishing gradients. Contrarily, in the event where gradients have a value larger than one, an exponential increase of the number occurs and increases until it is not possible to compute.This is also known as exploding gradients. The issues with the gradients indicate that the network's sensitivity to past inputs will decay to a certain extent with every new input introduced until, at some point, the new inputs will have made the network forget about the initial inputs. This unfortunately means that as the length of the sequence grows, the ability of the RNN to connect the information decreases. Multiple approaches have been suggested to address this problem with one of the more well-known being LSTMs.

The LSTM, unlike the RNN, has been designed to prevent any problems arising due to long-term dependencies. Like a typical RNN, the LSTM consists of a chain of connected repeating modules of neural networks. However, instead of having only a single operation to process the data, the LSTM has multiple operations to process the data as can be gathered from figure 3.9. The key concept of the LSTM is the cell state and the various gates implemented in the architecture. The cell state works as the "memory" in the network and the gates decide what information should be added or removed from the cell state. During training of the LSTM the gates learn what information is relevant to remember or to forget.

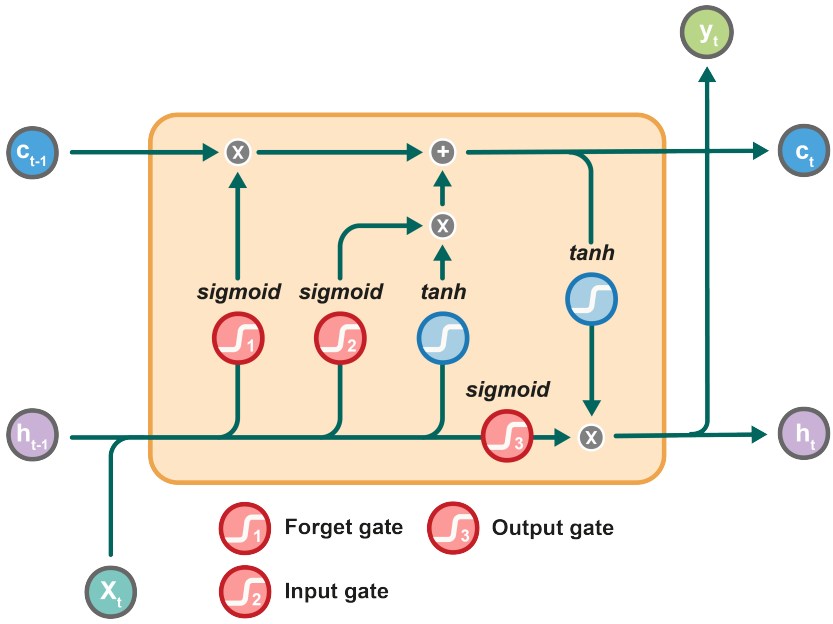


Figure 3.9. The long-short term memory cell state. The sigmoid functions denote the gates, where sigmoid1 is the forget gate, sigmoid2 is the input gate and sigmoid3 is the output gate. Here c indicates the cell state, h is the hidden state, X in the blue green circle is the input, y is the prediction and the grey circles denote element-wise addition (+) or multiplication (x).

The first gate data will go through in a LSTM cell is the forget gate. This gate determines what information should be kept or eliminated, which is done by passing information from the current state together with information from the previous hidden state through a sigmoid function. The sigmoid function outputs values between 0 and 1, where 1 means to keep and 0 means to forget the information. The idea of this gate is to decide what information from the prior steps is relevant to keep.

The next step in the network is the input gate. This gate is used to update the cell state. This is done by first passing the previous hidden state together with the current input onto a sigmoid function. In this step the sigmoid transformed values can instead be understood as deciding which information is important, where 1 denotes the information as being important and 0 as non important. To regulate the network, the hidden state and current input is also passed through a tanh function, which transforms the values to be between -1 and 1. The output from the sigmoid and tanh functions are then multiplied together, such that the sigmoid function determines what information from the tanh is important to pass on to the cell state. This step helps the network determine which information from the current step should be added to the cell state.

Thereafter, the next action is to calculate the cell step. The first step is to pointwise multiply the forget vector generated by the forget gate, where a value in the cell state will be forgotten if multiplied with values near 0 in the forget vector. Thereafter a pointwise addition is done with the outputs from the input gate, which updates the cell state to a new cell state, containing values relevant to the neural network.

The last step is the output gate, which decides the values of the next hidden state. The hidden state is used to contain information about the previous inputs as well as for making predictions. This step is carried out by passing the previous hidden state and current input into a sigmoid function and passing the new cell state into a tanh function. The output from the sigmoid function and tanh function are multiplied, which decides what information should be outputted to the hidden state. If there are more timesteps, the new hidden state and new cell state will be transferred over to the next timestep.

Although the LSTM in general solves any problems due to long-term dependency it has a major limitation, namely that they cannot be parallelized since the positions in an input need to be processed in a sequential manner.

This results in the training process requiring a lot of time which in certain experimental setups can be problematic.

3.2.4 Regularization

Although improved computational power allows for an expansion of the parameter space, which can expand the complexity of a model, thereby increasing the potential of learning a problem, too many parameters can have the opposite effect on the prediction capability of a model. A lot of parameters can also increase the chances of a model learning noise and trivial patterns present in the data, also known as overfitting, which can make the model incapable of generalizing well to never before seen data (33).

Early stopping is a regularization method where the training process is interrupted if a model has not improved or stops improving its performance on a hold out set, known as a validation set, over an arbitrarily decided number of training epochs (34). If a model is not trained long enough (too few epochs), it might not be able to identify the underlying patterns in a dataset. However, neural networks with enough parameters have the ability to fit training data perfectly if trained for enough epochs. Early stopping can aid in letting the network train for enough epochs to learn underlying dataset patterns, while still avoiding possible increased generalization error due to overfitting of the training dataset.

Dropout is a very commonly used trick to reduce the ability of a neural network to overfit (35). It works by masking units from a neural network layer randomly with a certain probability when training the network. The masking works by setting the activation value of the randomly selected units to zero, which ensures no interaction between the "dropped" unit and the previous or the following layer. This method reduces the risk of overfitting since a different subset of units is trained at each iteration, thereby decreasing the possibility that the network becomes dependent on only specific units in a network assigning large weights to them.

Batch normalization is a technique which helps standardize the inputs to a layer. This can help stabilize the learning process as well as increase the speed of training by reducing the number of training epochs required to learn and predict well (36).

4 Data representation and evaluation

4.1 Encoding

Computational models can only compute data input represented in a mathematical format. However, when working with biological data such as in the case of gene sequences or peptide sequences, these sequences are represented by nucleotides and amino acids respectively. Thus these different biological alphabets are required to be translated into an alphabet the computer will understand. There exists multiple ways of translating these sequences, with one of the most simple and well known approaches being one-hot-encoding.

One-hot-encoding

This method works by representing letters in an alphabet as a binary vector with the size of the alphabet in question. Each letter is assigned a unique position in the vector, which for a given letter will be represented with a 1 at the letters unique position in the vector and 0's for the remaining letters in the vector. This way each letter in the alphabet has a unique representation vector.

BLOSUM encoding

One major caveat regarding one-hot-encoding is that all pairwise distances are assumed to be identical. However, this is not the case for amino acids. It is well known that some amino acids can have similar properties, and thus more easily be interchanged with each other without having any noteworthy impact on protein function or structure, whereas other amino acids cannot. Substitution matrices such as the BLOSUM matrix (37) on the other hand takes these similarities and dissimilarities into account. The BLOSUM matrix represents each amino acid as a vector with a calculated log-odds score approximating to the closest integer, indicating how likely a given amino acid is to be substituted by another given amino acid. This way, similar amino acids with similar residues will have a higher probability of substitution and vice versa, which

can be more informative for the network than assuming all amino acids to be equally different.

Energy encoding

In certain computational experimental setups, such as predicting TCR recognition of MHC presented peptides, information regarding the structure of the complex can add valuable information to the prediction capabilities of a model, such as the potential stability of a protein complex. One way of representing the structural information is by calculating the potential global energies of the modeled complex structures. Two widely used methods for calculating the potential energy of protein structures are FoldX (38; 39) and the Rosetta Energy Function 2015 (REF15) (40; 41). Both methods use a chemical force field, which utilizes a defined set of functions and parameters to calculate the potential energy of a given chemical structure.

The Rosetta energy function is a model which utilizes physical and mathematical assumptions parametrized from small molecule and X-ray crystal structure data. The Rosetta energy function calculates the potential energy by approximating the energy of a biomolecule conformation. This is done by performing a weighted sum of individual energy terms. One of the most important energy term is a so-called statistical pair potential, derived from the underlying statistics of experimentally derived structures of observing different amino acids at any given distance.

FoldX is an empirically derived force field. This algorithm was calibrated utilizing experimentally obtained mutational free energy changes from more than 1000 point mutants. The main functionality of FoldX is to calculate the free energy of macromolecules, allowing for calculating among other things, the stability and interaction energy of a protein complex.

4.2 Performance metrics

Measuring the performance of a machine learning model makes it possible to evaluate the model's performance and also enables the possibility of comparison. However, selecting the metric for measuring a model's performance is often not a trivial task, especially for imbalanced datasets, which is a recurring phenomenon in the field of biology. Imbalanced data typically refers to a classification problem where the classes in the dataset are not equally distributed. In the case of a binary classification problem, there may be more

negative than positive samples in the dataset and thus create bias towards the negative samples in the model.

The majority of performance measurements are calculated utilizing the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). One of the most commonly used metrics in machine learning is accuracy, which based on the before mentioned categories estimates the number of correct predictions made.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Accuracy, although having a major advantage of easy interpretation, suffers from the disadvantage of not accounting for class imbalance. A high accuracy in an imbalanced data setup can therefore simply be due to the model purely predicting the majority class.

Another commonly used metric for model evaluation is Receiver Operator Characteristic (ROC) Area Under the ROC Curve (AUC). The ROC curve is estimated based on the true positive rate (TPR) and the false positive rate (FPR). Here TPR denotes the proportion of positives correctly predicted as positives, and FPR as the proportion of negatives incorrectly called as positives. The different points on the ROC curve correspond to all the possible decision thresholds to determine whether the results are positive or negative. AUC based on the ROC curve then summarizes the overall diagnostic accuracy of the model. However, ROC AUC is not built to reflect the minority class in a highly imbalanced dataset well, since this metric does not place more emphasis on one class over the other. Although less sensitive to imbalanced data than accuracy, a highly imbalanced dataset can still produce misleading results when using ROC AUC.

$$TPR = \frac{TP}{TP + FN} \tag{4.2}$$

$$FPR = \frac{FP}{FP + TN} \tag{4.3}$$

A metric more resilient to imbalanced data is Matthews Correlation Coefficient (MCC). This metric takes the number of examples into consideration, making this metric much more robust to class imbalance compared to the other metrics mentioned.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.4}$$

4.3 Homology partitioning

It is not uncommon for datasets to contain similar sequences, a concept known as data redundancy. However, the presence of data redundancy is sometimes an overlooked issue, which can cause major problems downstream in regards to model training and validation, which we will expand on in this section.

Biological homology is a result of a shared evolutionary history. This concept complicates the analysis of DNA, RNA, or protein sequences due to similarities, rendering the analysis of samples difficult (42). Whether samples are homologous or not is typically inferred from the nucleotide or amino acid sequence similarity among samples, where significant sequence similarity over a certain threshold strongly indicates samples being homologous. This sequence similarity can lead to the occurrence of redundancy in biological datasets, meaning that multiple very similar data points may be present in the data at the same time (43). In addition to homology, another cause of data redundancy can be that some types of sequences may be more prevalent than others in a dataset due to specific research interests, or as in the case when working with TCRs, where some clones are often more expanded than other clones (44), which can generate a biased representation of those specific sequences. The redundancy present in biological data, or overrepresentation of certain sequences, complicates the procedure of optimal data partitioning. Currently a very common practice in machine learning is to randomly separate data into a train, validation and test set. However, if sequence homology is not accounted for when building a machine learning model based on a biological sequence dataset, the model may seemingly predict well, but is in fact just an overestimation of the predictive performance. Hence, instead of reflecting the models ability to interpolate or extrapolate, the presence of similar sequences in both the training and test set is rather showcasing the models ability to reproduce its own input (42). This puts the model at a disadvantage since the model

would not have learned the general patterns of the overall data in these instances. The inability to capture general patterns makes the model incapable of generalizing to novel data that has not already been presented to the model at any given point during the construction of the model.

Solving the issue of redundancy due to sequence similarity can be approached from different angles. Many methods exist, but the approach used in this thesis is homology partitioning. Sequences with a similarity equal to or above a specified threshold are clustered together (45). The clustering helps identify similar data points, which are then partitioned together, ensuring that no overlap exists between the train and test set. Furthermore, this also prevents a potentially already scarce dataset from becoming even more scarce by avoiding size reduction. A potentially major disadvantage of this method is that certain types of sequences may be overrepresented in the dataset. This poses a risk of biasing the model due to overrepresented sequences being presented more frequently to the model. The bias which arises can be dealt with in different ways, such as giving the data points weights, increasing the possibility of underrepresented sequences being presented to the model. Another method could be only to present one sequence per cluster, where the sequence is chosen randomly in each training iteration, thereby removing any potential overrepresentation present in a large cluster. A third method could be to penalize the model harder for getting an underrepresented sample wrong compared to an overrepresented sample.

4.4 Cross-validation

Cross-validation is a technique in machine learning that enables the use of the same dataset both for training and testing a model by cycling partitions of the data. In the commonly used k-fold cross-validation method, the data is partitioned into k equal parts, where each fold is used as a test, and the remaining k-1 partitions are used to train the model repeated k times. This process is illustrated in figure 4.1, where k is set to 5. It is a technique that allows robust estimation and evaluation of model performance as well as an effective procedure to evaluate models trained on limited data. The improved estimation of model performance lies in the fact that if a random subset of the data is used as a test set, we might be underestimating or overestimating model performance since this subset may hold a bias.

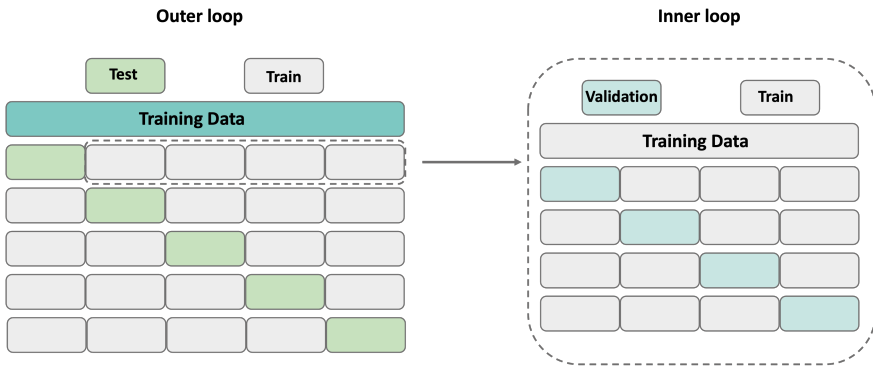


Figure 4.1. 5-fold cross-validation scheme on the left showing how a different partition will be a test set in each cross-validation cycle. On the right a nested 5-fold cross-validation scheme of a single test fold.

If one wants to train a neural network using a training, validation, and test set, k-fold cross-validation is not applicable. In this instance, the network is trained using the training set with the validation set as a guide to help select the best hyper-parameters and stop the model from overfitting by performing early-stopping with the performance of the trained model being estimated using the test set. Here, a nested k-fold cross-validation procedure can be a potential solution to this issue. As in the regular k-fold cross-validation setup, the data is still divided into k equal folds but are now divided into two levels. In the first level, a partition is chosen as a test set as done in a regular cross-validation setup. However, instead of training the model on the remaining k-1 partitions, these are instead moved to the second level, where one partition will be used as the validation, and the remaining k-2 partitions will be used as the training set. In the case of dividing the data into five equal folds, one fold would be used as the test, one fold as the validation, and the remaining three folds as the training set as shown in figure 4.1. The first level will be run five times so that each fold will be used as a test once. The second will be run four times each, ensuring that each fold in the remaining four partitions will be used as a validation set at least once. This amounts to a total of 20 independent models trained on the same dataset to estimate the overall performance of the ensembled models.

5 Immunotherapy and current tools

The primary purpose of this thesis was to develop deep learning methods for immunotherapy. Before expanding on which tools are currently available within the field, we will first cover what immunotherapy is and how it can be used. Immunotherapy is a type of treatment that aims to suppress or activate the body's own immune system to treat disease. Different kinds of immunotherapies exist, such as Adoptive cellular therapy, Immune checkpoint therapy, cytokine therapy, monoclonal antibodies, and cancer vaccines. Immunotherapy is generally used to treat cancer and can be used by itself when more traditional anti-tumor therapies are not effective or in combination to enhance their effect (46). Genome aberrations are often a typical feature of most cancer types (47). Although these aberrations often play an important role in cancer development, they can be exploited for immune system recognition by recognizing cancer-specific peptides known as neoepitopes. Having computational tools for predicting epitopes and neoepitopes has already been recognized as being important for the successful development of many cancer immunotherapies (48). Different computational tools exist to help discover potential immunotherapy targets. These tools can generally be split into two separate groups, which are sequence based and structural based prediction methods. The majority of these methods primarily identify epitopes and neoepitopes presented by the MHC molecule without taking TCR recognition into account; a more in-depth explanation and overview can be found in the first article, in chapter 6.

Newer tools such as pEptide tcR matchinG predictiOn (ERGO)-I (49), ERGO-II (50), interaction map recognition (ImRex) (51), Tcr epITope bimodal Attention Networks (TITAN) (52) and NetTCR2 (53) are developed with TCR recognition as the main focus point. These tools all utilize deep learning frameworks for predicting peptide recognition. The frameworks used are CNNs (ImRex and NetTCR2), CNNs with attention (TITAN) and LSTMs or Autoencoders (ERGO-I, ERGO-II). The majority of these frameworks were generally chosen due to their ability to catch certain contextual information. Models such as ERGO-I, ImRex, and TITAN use the CDR3 β region to predict potential epitopes and neoepitopes, whereas tools such as ERGO-II and NetTCR2 use paired CDR3 sequences for this task. These tools are all sequence

based prediction methods; therefore, in the third article, chapter 8, we study whether global energies from modeled structures can improve T cell peptide recognition prediction.

6 T cell Epitope Prediction and Its Application to Immunotherapy

This chapter presents a review that has been peer-reviews and published in Frontiers in Immunology. In this review, we present current tools available for epitope and neoepitope prediction at the time of writing. We create an overview of the type of data the different tools can use as input, what kind of analysis they are able to perform, and potential pros and cons with the different methods. This review also provides a discussion on areas that could be of potential interest in the future for improving the accuracy of epitope and neoepitope prediction tools.

The main goal of this project was to create an overview to enable the reader to make informed choices in which tools may be applicable for them to use. This review was intended to function as a tool to help those who are in the midst of or about to design new experiments making the reader aware of what type of data is necessary for the question they are trying to solve with a specific tool. This review was also meant as a guide to help the reader understand what tools they can use with the data they have already collected.

frontiers in Immunology

REVIEW
published: 15 September 2021
doi: 10.3389/fimmu.2021.712488

T Cell Epitope Prediction and Its Application to Immunotherapy

*Anna-Lisa Schaap-Johansen, Milena Vujović, Annie Borch, Sine Reker Hadrup and Paolo Marcatili**

Department of Health Technology, Technical University of Denmark, Lyngby, Denmark

T cells play a crucial role in controlling and driving the immune response with their ability to discriminate peptides derived from healthy as well as pathogenic proteins. In this review, we focus on the currently available computational tools for epitope prediction, with a particular focus on tools aimed at identifying neoepitopes, i.e. cancer-specific peptides and their potential for use in immunotherapy for cancer treatment. This review will cover how these tools work, what kind of data they use, as well as pros and cons in their respective applications.

Keywords: epitope prediction, neoantigens, neoepitope prediction, T cell, TCR, T cell receptor

OPEN ACCESS

Edited by:
Sandra Tuyaerts, University Hospital Brussels, Belgium

Reviewed by:
Scott Christley, University of Texas Southwestern Medical Center, United States
Cristina Maccalli, Sidra Medicine, Qatar

***Correspondence:**
Paolo Marcatili
pamar@dtu.dk

Specialty section:
This article was submitted to Cancer Immunity and Immunotherapy, a section of the journal Frontiers in Immunology

Received: *20 May 2021*
Accepted: *12 July 2021*
Published: *15 September 2021*

Citation:
Schaap-Johansen A-L, Vujović M, Borch A, Hadrup SR and Marcatili P (2021) T Cell Epitope Prediction and Its Application to Immunotherapy. Front. Immunol. 12:712488. doi: 10.3389/fimmu.2021.712488

INTRODUCTION

T cells recognize and survey peptides (epitopes) presented by major histocompatibility complex (MHC) molecules on the surface of nucleated cells. To be able to perform this task, T cells must be able to differentiate between native "self" peptides versus peptides deriving from pathogens, infections or genomic mutations. In order to effectively mount and initiate an immune response, T cells must undergo activation. The main requirement of T cell activation is the molecular recognition between the T cell receptor (TCR) expressed on the T cell surface and peptide-MHC complexes (pMHC) presented on the surface of other cells. This precise recognition process is of paramount importance for a well-functioning immune system, and is shaped by a mechanism named central tolerance. In order to ensure that T cells do not react against ubiquitous peptides found in an individual, T cells undergo the process of negative selection. Early in their development, T cells are presented with a plethora of self-peptides, where any T cell that recognizes self-peptides is eliminated, leaving only T cells with little or no specificity for self. Cases in which this mechanism fails and T cells recognize self-epitopes are typically associated with harmful effects on the organism and might result in autoimmune disorders.

As mentioned earlier, T cells recognize epitopes only when they are presented by MHC molecules. Early in the thymic development of T cells, they undergo the process of positive selection ensuring that they bind to host MHC molecules. There exist two classes of MHC molecules: class I expressed on surfaces of all nucleated cells and class II found on surfaces of specialized antigen-presenting cells (APCs). As two classes of MHC molecules occur, two types of T cells are specially equipped for binding to the MHC I and II, the CD8+ and CD4+ T cells, respectively. The general focus of this review will be on cytotoxic CD8+ T cell binding to MHC I presented epitopes.

The immune system in general is very good at identifying "foreign" peptides stemming from bacterial or viral infections. On the other hand, as initially proposed by Burnet and Thomas through the idea of immunosurveillance (1, 2), the same process can also protect our organism from cancer,

by recognizing cancer-specific peptides (neoepitopes) generated by somatic mutations or genomic aberrations (**Figure 1**). The ability of the immune system to target cancer cells has been exploited by a novel class of therapies, such as adoptive T cell therapy and cancer vaccines, named immunotherapies. These approaches, by exploiting the high selectivity of the immune system, have the advantage to be more specific and less invasive than traditional cancer therapies, and potentially effective even at later stages by providing immunological memory.

Broadly, immunotherapy can be divided into two categories: "active" and "passive". The "active" works to stimulate T cells of the individual's immune system into attacking tumor cells i.e. effectively training the immune system *in vivo*. The "passive", focuses on *in-vitro* training and subsequent injection of immune agents that will help battle the disease *in vivo* (3). Passive immunotherapy includes therapies such as adoptive cell therapy, cytokine injection, monoclonal antibodies and lymphocytes (4, 5). Active immunotherapies encompass therapies such as non-specific immunomodulation and vaccination (6, 7).

Computational tools for epitope prediction have been recognized as being crucial for successful development of various cancer immunotherapies (8). This review will therefore give an overview of both general and cancer specific epitope prediction tools and discuss the pros and cons of the different tools and future perspectives in the field.

EPITOPE PREDICTION METHODS

As mentioned before, a peptide needs to be presented by an MHC I molecule for it to be able to elicit effector T cell responses. Contrarily to MHC II molecules, which can bind to peptides that are longer and more variable, MHC I binding is restricted to peptides typically 8-14 amino acid long in sequence and that some of the residues in the peptide, denoted anchor residues, are important for peptide-MHC binding (9) (**Figure 2**). In most human alleles the anchors are the second and the last residues in the peptide (10), but this depends on the allele and species. The binding of peptides to MHC molecules is therefore a very selective step, which has been a major focus in many epitope prediction models. However, most peptides presented by MHC molecules will not elicit an immune response as they do not evoke TCR specific recognition by the T cell. In order to shed light on this interaction, computational models are being constructed with the goal of predicting T cell recognition of the presented peptide and its connection to an overall immune response. Epitope prediction can thus currently be divided into

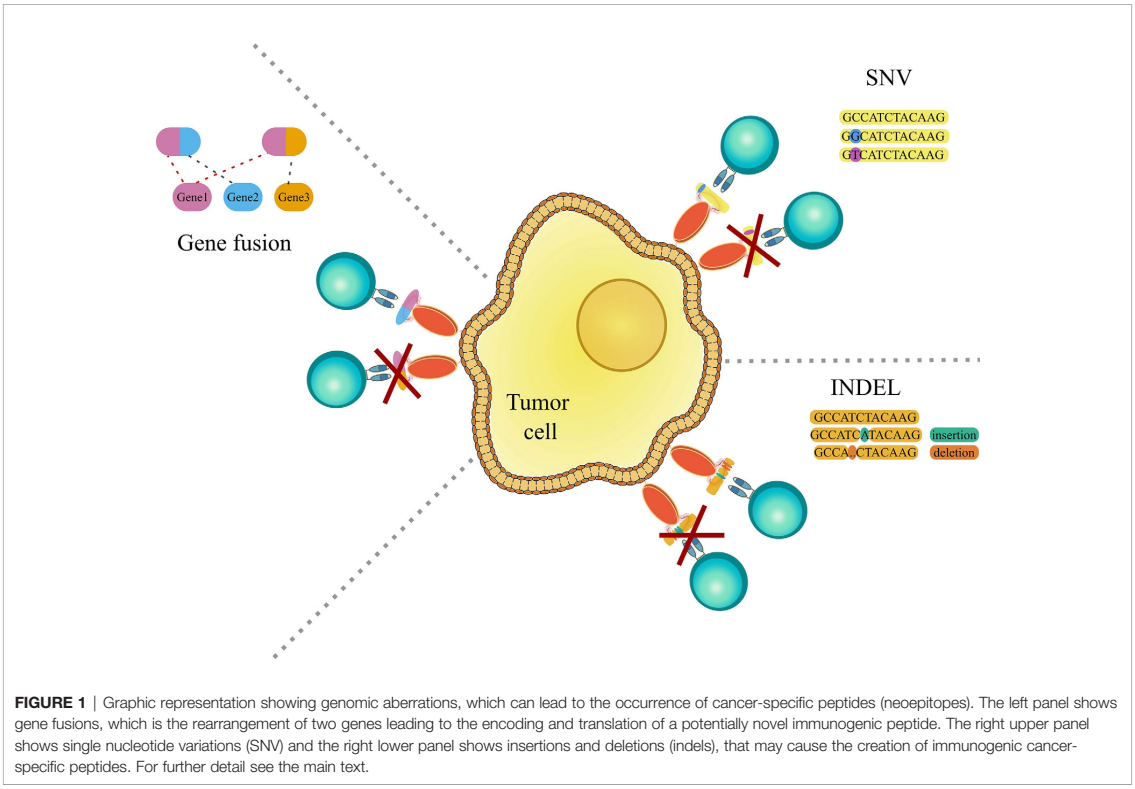


FIGURE 1 | Graphic representation showing genomic aberrations, which can lead to the occurrence of cancer-specific peptides (neoepitopes). The left panel shows gene fusions, which is the rearrangement of two genes leading to the encoding and translation of a potentially novel immunogenic peptide. The right upper panel shows single nucleotide variations (SNV) and the right lower panel shows insertions and deletions (indels), that may cause the creation of immunogenic cancer-specific peptides. For further detail see the main text.

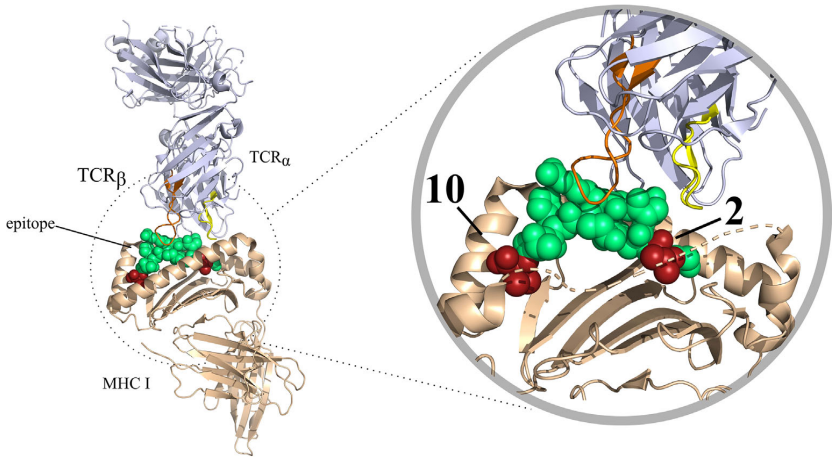


FIGURE 2 | T-cell interaction with a pMHC complex rendered in PyMOL (PDB code: 6TRO). Here MHC I is shown as colored in beige. The TCR is colored blue white. The CDR3 variable regions of the T-cell have been colored in different colors, these are as follows: CDR3 α colored in yellow, CDR3β colored in orange. The bound peptide is colored in green, with the anchor residues are colored in red.

two main focus areas. The first addresses the presentation of peptides by MHC molecules. Extensive reviews on this subject have been published recently, and we single out the in depth work by Peters et al. (11). In this review, we mainly focus on the second part of the interaction: predicting T cell recognition of pMHC complexes.

One of the first attempts at defining the immunogenic potential of peptides was based on their local and global physico-chemical characteristics, regardless to the specific T cell interaction. One of such tools is POPI (12), which is a support vector machine (SVM) based method. SVMs are machine learning tools that can identify complex non-linear relationships between the input data and the predicted variable. In this case, a feature set of physico-chemical properties derived from MHC I binding peptides is used to predict the peptide's immunogenicity. POPI uses averaged values of the physico-chemical properties independent of the amino acid positions in the peptides, therefore being unable to take local information into consideration in the predictions.

Another model named POPISK (13), by the same group, tries to improve on this by utilizing a SVM in conjunction with a weighted degree string kernel. The model is seemingly only capable of predicting immunogenicity for HLA-A2-binding peptides. Where predictions reached an overall accuracy (ACC) of 0.68 and 0.74 for area under the curve (AUC). The ACC and AUC are calculations based on a confusion matrix, which in different ways essentially estimates how often an algorithm predicts correctly. In both cases, a perfect prediction would have both ACC and AUC equal to 1, and lower values for worse predictions. A more exhaustive introduction to accuracy metrics for prediction tools can be found in Peters et al. (11). It should be mentioned that the dataset was not pre-processed to remove or reduce the redundancy - i.e. very similar peptides might be present. This has been observed to have a negative impact on the methods' ability to generalize, that is the ability of an algorithm to achieve good results on data that is different from the data used to train. A typical strategy to deal with this issue is to perform some form of homology reduction to reduce redundancy. In the discussion we will discuss more about the importance of such procedure when assessing the actual accuracy of prediction tools. Furthermore, it should be noted that both POPI and POPISK are not available for general use anymore.

Calis et al. created the immunogenicity model (14) based on experimental indications. The authors discovered that T cells show a preference for binding peptides containing aromatic and large amino acids. They also showed that positions 4-6 were important in regards to immunogenicity. Based on this information, a scoring model was created which scores peptides based on the ratio of an amino acid between a non-immunogenic and immunogenic dataset. Furthermore, it weights the amino acid based on its position in the ligand. The authors estimated the accuracy of the model on new MHC I binding peptides, and obtained an AUC of about 0.65, thus the model is only to some extent predictable. It should be noted, that where models such as POPISK only is capable of predicting TCR propensity for HLA-A *02:01, the Calis et al. immunogenicity model can make predictions for any MHC I molecule.

PAAQD (15) is a model which focuses on predicting T cell reactivity. It works by encoding nine-mer peptides which are processed in a random forest algorithm, in order to predict the immunogenicity of a peptide binding to MHC I. The peptides are

numerically encoded by combining information regarding quantum topological molecular similarity (QTMS) descriptors and amino acid pairwise contact potentials (AAPPs). In the article it was mentioned that an ACC of 0.72 and a AUC of 0.75 was obtained for immunogenicity prediction. It obtained a higher AUC and ACC than POPISK and a higher AUC than the immunogenicity model by Calis et al., however, like POPISK, no homology reduction was done to reduce redundancy. Furthermore the model had a focus on HLA-A2 and will have limited success in predicting immunogenic peptides for other HLA molecules.

Jørgensen and Ramussen, who developed NetMHCstab (16) and NetMHCstabpan (17) respectively, theorized that instead of entirely focusing on the HLA binding affinity one should also take pMHC stability into account to predict immunogenic MHC I ligands. They based this hypothesis on the assumption that a more stable presentation of an epitope bound to an MHC will increase the likelihood of a T cell recognizing the epitope. However, as the authors have also indicated in the papers themselves, stability alone did not give as good results as combining a stability predictor with a pMHC I binding predictor.

Experimental investigation of peptide presentation and binding by Schmidt et al. (18) showed poor correlation with predictions for the same peptides by NEtMHCstab and NetMMHCpan in combination with a binding affinity predictor. These models were outperformed by another epitope prediction model: NetTepi (19). This model has been built on top of previous efforts and combines: peptide-MHC stability using NetMHCstab, T cell propensity predictions using the immunogenicity model by Calis et al. and peptide-MHC binding affinity using NetMHCcons (20). The model has been stated to be capable of predicting T cell epitope for multiple HLA molecules with a sensitivity of 90% and a false positive rate of 1.5%.

One of the newer models for predicting which epitopes will be recognized by T cells is NetTCR (21). NetTCR implements a convolutional neural network (CNN) model to predict TCR recognition of a peptide. CNNs are a type of neural network which are very popular for different tasks (e.g. image recognition) and capable of identifying local patterns in the input data. The model takes as input a HLA-A *02:01 binding MHC I peptides and the CDR3 protein sequence of a T cell receptor. The model obtained a somewhat high AUC of 0.727. The AUC is lower than the AUC for POPISK (0.74) and PAAQD (0.75). However, it should be noted that unlike POPISK and PAAQD, NetTCR performed homology reduction to reduce any redundancy in the data.

A major bottleneck in improving the accuracy of models is in the limited amount of available training data. However, several databases collecting experimental immunogenicity data are now available, with one of the first to pioneer this area being SYFPEITHI from Rammensee et al. in 1999 (22). Newer databases have since been created such as IEDB (23), VDJdb (24), McPAS-TCR (25), ATLAS (26) and STCRDab (27). The steadily increasing amount of experimental data will support the generation of models with greater prediction power.

STRUCTURAL EPITOPE PREDICTION

The energetic balance of the TCR-pMHC interaction is one of the main drivers in dictating the initiation of an immune response. As evident from structural (28) and mutagenesis studies (29), this balance is very delicate. All circulating T cells have undergone the so-called positive selection process, meaning that they must bind with low affinity to MHC molecules, regardless of the specific epitope. Additionally, TCR interaction is highly cross-reactive, meaning that a single TCR will potentially be able to bind to thousands of peptides. This poses a serious hurdle to develop computational tools to predict immunogenicity based on structural calculations. In recent years, it has been shown that, when using fine-grained molecular dynamics (MD) simulations, one can to some extent predict TCR-pMHC interactions (30). Unfortunately, this approach is neither very precise nor feasible. For such calculations, high quality structures of the interacting molecules are needed, and the current available amount of solved structures for TCRs is very limited - less than three hundred at the time of writing. In contrast, the number of different TCRs that circulate at any time in humans is 10^6 to 10^8 (31), and the theoretical numbers of different TCRs is at least 4×10^{11} (32). This stark difference greatly reduced the usefulness of such methods to a tiny minority of the available cases. Even when solved structures are available, MD simulations are very demanding in terms of computing time. The dynamics of the TCR-pMHC interaction, especially regarding their dissociation rate, have time scales that are currently at the very limit of what one can achieve with full-grain MD Simulations.

Some works have focused on solving these 2 problems - the lack of structural information and the need for more efficient structure-based algorithms. It is now possible to model to a very good accuracy TCRs, pMHCs, and their complexes. Without delving in too much detail, most currently available methods (33–35) can model pMHC complexes to a very good accuracy - often less than 1Å Root Mean Square Deviation (RMSD) - from the native structure, and almost as good as the experimentally resolved structures. TCRs can also be modeled with good accuracy (in general less than 2Å RMSD), with some minor exception for the CDR3 regions of both TCR chains. The real culprit of all modeling tools is in predicting the correct mutual orientation of the TCR with respect to the pMHC, for which only a decent accuracy can be achieved: approximately, only 50% of the molecular contacts between TCRs and pMHC are recovered in the model. Given the current accuracy of the modeling tools for TCR-pMHC complexes, together with the computational cost of running detailed atomistic simulation, underline the need of more coarse-grained models, that can ease both the aforementioned problems. In recent years, Lanzarotti and co-workers (36, 37) used TCR-pMHC models to refine existing computational force fields [Rosetta (38) and FoldX (39)], and combined such refined energy calculations in a simple statistical framework to improve the prediction of existing TCR-pMHC complexes. The authors show that, even in such a simple approach, it is possible to exploit structural models to identify, among a pool of TCRs and pMHCs, the actual interacting partners.

The same results have recently been confirmed using a similar approach (40). The authors show that, by investigating the energy and the structural variability in TCR-pMHC models, it is possible to improve the prediction of TCR-pMHC pairs. At the current stage, structure-based methods can greatly reduce the number of false positive predictions obtained by sequence-only methods, at the cost of reduced sensitivity.

NEOANTIGEN PREDICTION

Genome aberrations are a typical feature of many cancer types (41). On the one hand such aberrations are linked to the cancer occurrence and growth, i.e. by disrupting normal cell cycle and apoptosis control. On the other hand, they can be exploited by the immune system to recognize and eliminate cancer cells. As mentioned previously, neoepitopes have been a major target of immunotherapy approaches such as adoptive T cell therapy or cancer vaccination. Several computational tools have been developed to assist and improve immunotherapy. The main rationale of these tools is to first identify aberrations in the cancer genome, and then, to a different extent and with individual approaches, to predict the ones that are more likely to trigger an effective immune response. Besides genomic aberrations, events such as post-translational modifications (PTMs) (42) and peptides derived from non-coding regions (43) can also cause neoepitopes to arise. However, due to the limited availability of data and of the biological basis of these, there are currently only very few computational tools for their analysis and prediction (44). Broadly speaking, the available tools can be categorized by the type of input data they accept, by the type of variants they can call, and by the strategy used to filter or prioritize the most immunogenic variants. Regarding the first point, neoepitopes can arise due to events such as single nucleotide variations (SNV), insertions and deletions (indels), intron retention, and chromosomal aberrations (45–48). While most of the tools can predict neoepitopes from SNVs [Epi-Seq, TIminer, Neopepsee, DeepAntigen], some also incorporate indels [pVACseq, MuPeXI, Epidisco, OpenVax, NeoEpiScope, CloudNeo, pTuneos, antigen.garnish, NeoPredPipe, TSNAD], and others only focus on indels [ScanNeo], gene fusions [NeoFuse, INTEGRATE-neo], or they let the users input the variants as peptides [EDGE, DeepHLApan], for an overview see **Table 1**. Another difference between the tools is the types of data that these models rely on. In most cases the tools use whole genome sequencing (WGS), whole exome sequencing (WES), transcriptome sequencing (RNA-seq), peptide sequencing, or a combination of those. Finally, in order to filter and prioritize neoepitopes, many tools incorporate predictions from NetMHC (68) and NetMHCpan (69), alongside some other tools for predicting MHC binding. In the following, we will briefly present the available tools based on the characteristic that we have just discussed.

Single Data–Based Models

Both RNA-seq and DNA-seq data can be exploited to identify variants in the cancer genome, and several tools make use of these data to predict neoantigens. It is important to notice that these two experimental methods provide complementary information. DNA-seq data is in general more sensitive, i.e. it can identify more variants. RNA-seq experiments can be used to generate expression levels at the gene or, as at the transcript level, thus helping to prioritise variants that are present in highly abundant genes over those that have low or no expression. It should be noted that the transcript level is often recommended, since this can further give information regarding events important for neoepitope prediction, such as isoform selection

TABLE 1 | Overview of the different neoantigen prediction tools.

	Bioinformatic tools for neoantigen prediction						
Tool	**DNA**	**RNA**	**Peptide**	**SNV**	**indels**	**Gene fusion**	**Reference**
Epi-seq		X		X			(49)
TIminer	X	X		X			(50)
Neopepsee	X	X		X			(51)
DeepAntigen	X	X		X			(52)
PVACseq	X	X		X	X		(53)
Mupexi	X	X		X	X		(54)
Epidisco	X	X		X	X		(55)
OpenVax	X	X		X	X		(56)
Neoepiscope	X	X		X	X		(57)
CloudNeo	X	X		X	X		(58)
pTuneous	X	X		X	X		(59)
antigen.garnish	X	X		X	X		(60)
NeoPredPipee	X	X		X	X		(61)
TSNAD	X	X		X	X		(62)
ScanNeo		X			X		(63)
NeoFuse		X				X	(64)
INTEGRATE-neo	X	X				X	(65)
EDGE		X	X	X	X	X	(66)
DeepHLApan			X	X	X	X	(67)

and alternative splicing (70–72). Peptide sequencing can also be used for neoantigen prediction. This holds information regarding whether a gene is actually expressed or not at the protein level. This is very important information; identified variants at DNA or RNA level are not always expressed at protein level. The reader should take this into account when deciding which tools they want to use.

Epi-Seq (49) is a tool which only uses tumor RNA-seq data. Epi-Seq works as a wrapper tool, i.e. it combines the output of other tools to perform an integrated prediction. It only supports SNV variant calling and neoantigen prediction from those calls. The Epi-Seq pipeline is very useful when only RNA-seq data is available. However, since the pipeline only focuses on SNV variants other potentially important variants are not predicted on.

ScanNeo (63) is a tool capable of predicting neoepitopes from small to large-sized indels. ScanNeo is a wrapper tool, which takes as input RNA-seq data. The three major steps in its pipeline are i) indels discovery, ii) annotation and filtering and iii) neoantigen prediction. ScanNeo uses NetMHC in its pipeline. Besides NetMHC, the tool also employs NetMHCpan in its pipeline to predict peptides that bind to HLA class I with high affinity.

NeoFuse (64) is a computational pipeline predicting neoantigens from gene fusions. It is a wrapper tool which uses raw RNA-seq data from patient tumors as input to do HLA class 1 typing, predict fusion peptides and quantification of gene expression. MHCflurry (73) to predict pMHC binding and the gene expression levels are utilized to filter out candidate fusion neoantigens. Like Epi-seq this is convenient when only tumor RNA-seq data is available.

DeepHLAPan (67) is a recurrent neural network-based approach, which takes both peptide-HLA binding and potential peptide-HLA immunogenicity into account. The tool predicts neoepitopes utilizing HLA class I typing provided by the user and peptides. The tool further filters the candidate neoantigens based on a score generated by an immunogenicity model based on immunogenicity data from IEDB.

Data Integration–Based Models

Next generation sequencing (NGS) has made it easier to sequence in parallel the DNA and RNA of a patient. By integrating the use of both DNA and RNA data, the researcher can call somatic mutations from the DNA and quantify gene and transcript expression from the RNA data, which can help in identifying which variants are more likely to be expressed. Also in this case, most of the computational tools are in fact wrappers of multiple different methods which are integrated in multi-step workflows to perform the neoepitope prediction. Besides integrating DNA and RNA data, it is also possible to predict neoepitopes from peptide and RNA sequencing data. The peptide data enables us to know which genes are actually expressed at protein level and the RNA data helps with identifying which of the peptides will be presented by the HLA alleles, since expression of messenger RNA is strongly correlated with HLA peptide presentation (74). In general integrating data can often help in generating more accurate predictions, as many of the tools which will be mentioned in this section also have shown in their studies. When choosing tools, the reader should keep in mind the somatic variations they want to account for and what kind of data they possess.

pVACseq (53) is a neoantigen prediction tool, which can work with either WES or WGS data together with RNA data. This tool can predict neoantigens from small indels and SNVs. pVACseq utilizes HLAminer (75) to infer the patients HLA class I typing and NetMHC to predict HLA class I restricted epitopes. The tool prioritizes neoepitopes based on sequencing depth and fraction of reads containing the variant allele.

INTEGRATE-neo (65) is another tool which also uses NetMHC in its pipeline. This tool is based on INTEGRATE (76), which uses DNA sequencing data to predict peptides generated by gene fusion events, and thereafter uses HLAminer to perform in silico HLA typing, and lastly uses NetMHC to predict neoantigens based on the gene fusions. Where the other tools can work just with the DNA data, optionally also integrating RNA data into their pipelines, INTEGRATE-neo requires the use of both DNA and RNA. A tool suite named pVACtools which includes pVACseq and INTEGRATE-Neo among other tools to not only account for SNVs and small indels but also include support for structural variants.

MuPeXI (54) like pVACseq requires the user to provide HLA types, somatic variants and optionally gene expression estimates. The tool predicts neoantigens from SNVs and indels. The tool can use either WES or WGS data and optionally also RNA data and have similar features to pVACseq. However, unlike pVACseq, MuPeXI also offers i. a priority score to rank peptides ii. a comprehensive search for self-similarity peptides and lastly iii. besides being a downloadable command-line tool it is also available as a webserver. Furthermore, this model incorporates the use of NetMHCpan (69) in its pipeline instead of NetMHC.

Epidisco (55) takes as input wild type DNA, tumor DNA and tumor RNA sequencing data. The tool maps the normal and tumor DNA samples to the human GRCh37 reference genome. Epidisco, like many of the other tools mentioned works as a wrapper around other existing tools, and also like many of the other tools, Epidisco uses NetMHCpan in its pipeline. The tool supports SNV and indel based neoantigen prediction. Epidisco focuses on vaccine peptide selection, and generates a ranked list of peptide candidates.

TIminer (50), like many of the other tools, is a tool which as input requires a pre-existing set of variants derived from DNA. The tool also incorporates NetMHCpan in its pipeline and unlike other tools it is able to process raw RNA-seq data which may obtain more information relevant for neoantigen prediction. This tool, however, only supports neoantigen prediction from SNVs.

OpenVax (56) is another pipeline which integrates the use of NetMHCpan into its pipeline, however, it is also possible to choose other MHC binding peptide predictors. The OpenVax pipeline, unlike many of the other tools takes as input raw DNA and RNA sequencing files. The OpenVax pipeline has also included somatic variant calling tools in its pipeline which are

capable of calling SNVs and indel variants. It has a ranking function similar to MuPeXI, but with less features, namely MHC class I affinity scores and RNA-seq read count based variant expression.

NeoEpiScope (57) is another tool which can use NetMHCpan in its pipeline. The tool in general uses MHCflurry or MHCnuggets, however, NetMHCpan can also be used if installed individually. Like many of the other tools, NeoEpiScope requires as input a set of somatic variants and supports SNV and indel based neoantigen prediction. The main focus of this tools is to prioritize handling phased variants. To use the phasing function, the user must submit patient haplotypes.

CloudNeo (58) is a tool developed for cloud computing, created to eliminate the need for local infrastructure investment in computation, data storage and transfer, while also providing scalable computational capabilities for neoantigen identification. CloudNeo is a wrapper like many of the other tools which also utilizes NetMHCpan in its pipeline. CloudNeo supports SNVs and indels for neoantigen prediction. Although CloudNeo uses RNA data in its pipeline, it seemingly only utilizes the RNA data for HLA typing, however, DNA data can also be used for this purpose.

Neopepsee (51) is a tool which takes as input a list of somatic mutations and raw RNA seq data. The tool focuses on non-synonymous somatic mutations and works as a wrapper tool, which uses tools such as NetMHCpan to predict MHC binding affinity. For peptides with the highest binding affinity, immunogenicity features are then calculated and fed into a locally weighted naïve Bayes classifier. The idea with Neopepsee is to use a classifier to decrease the amount of false-positives that using only binding affinity would provide.

pTuneos (59) predicts and prioritizes candidate neoantigens from SNVs and indels. The tool is a wrapper tool, which takes as input raw WGS/WES tumor normal matched sequencing data and optionally also tumor RNA-seq. The tool utilizes HLA class I typing and NetMHCpan to predict binding affinity of normal and mutant peptides, which is then run through a random forest model to predict a T cell recognition probability. Finally they use a scoring schema to evaluate whether a candidate neoepitope that can be recognized by a T cell will be naturally processed and presented. This can be used to prioritize the peptides based on *in vivo* immunogenicity.

The package antigen.garnish (60) is an wrapper tool in R, utilizing NetMHCpan among others for peptide MHC binding in its pipeline. It predicts neoantigens from SNVs and indels. Besides MHC binding it also takes hydrophobicity, comparison of MHC binding affinity between mutated and non-mutated counterpart, and dissimilarity into account. Furthermore, the tool also calculates a TCR recognition probability based on the dissimilarity.

NeoPredPipe (61) is another tool which incorporates NetMHCpan into its pipeline. Like many of the other tools the user has to submit files regarding patient haplotypes and SNVs and indels. NeoPredPipe unlike the other tools provides the opportunity of neoantigen prediction on multi-region sequencing data and also asses the intra-tumor heterogeneity, which is done based on multi-region samples, where the neoantigen burden is reported for clonal, subclonal and shared variants. NeoPredPipe furthermore also predicts the likelihood of TCR recognition. This based on the probability of the mutant epitopes ability to bind to MHC I molecules and the epitopes similarity to pathogenic peptides.

TSNAD (62) is a tool which earlier had netmhcpan integrated in its pipeline, however, in their version 2.0, which was updated in 2019, they replaced NetMHCpan with the earlier mentioned DeepHLAPan to predict binding of the mutant epitopes to MHC I molecules. TSNAD works by, like many of the other tools by integrating multiple tools into its pipeline. The tool takes as input raw read of tumor normal DNA pairs. The sequences can either be mapped to GRCh37 or GRCh38. In the updated version, raw RNA-seq data can optionally be added to help filter neoantigens. The tool supports neoantigen prediction from SNVs and indels.

DeepAntigen (52) is a deep sparse neural network model based on group feature selection (DNN-GFS). Uniquely this model bases its predictions on the DNA loci of the neoantigens in a 3D genome perspective. The authors discovered that the DNA loci of the immunonegative and immunopositive MHC class I neoantigens have distinct spatial distributions. The model uses preprocessed WES and messenger RNA-seq for calling somatic mutations and estimating gene expression. The model also takes as input Hi-C (77) data (captures chromosome conformation) for 3D genome information. However, this method can only predict neoepitopes from non-synonymous point mutations and 9 mer peptides.

EDGE (66) is a commercial platform for neoantigen identification. The EDGE model is a neural network trained on HLA peptide mass spectrometry data and RNA-seq data from various human tumors. The model uses HLA class I type and sequence, RNA and peptide sequencing data or peptides generated from somatic variant calling data to predict neoantigens. Although the model does not incorporate TCR binding, it is still to a certain extent able to capture T cell recognition with the addition of RNA expression.

DISCUSSION

In recent years, the number of computational tools for epitope and neoepitope prediction has exploded. In many cases, these tools combine the results of other methods, using different heuristic approaches, to perform their predictions. Unfortunately, the amount and quality of available data make it difficult to decide which of these approaches are sound, and which are not. As an example, many of the currently existing epitope and neoepitope prediction methods are mainly focusing on MHC presentation. This is because, from a quantitative point of view, MHC binding is the most selective step. According to Yewdell et al. around 1 in 200 peptides bind to MHC class I with an affinity strong enough (500 nM or lower) to induce a immune response (78). Other studies, such as Sette et al. (79), also indicated an MHC affinity threshold of 500 nM to be associated with T cell recognition of HLA class I

bound peptides. Moreover, MHC binding is considered necessary but not sufficient for a molecule to be immunogenic: in general only the minority of epitopes predicted are immunogenic (80–82). However, this paradigm has been challenged on many occasions. In particular for neoepitopes, there is not a general consensus on the fact that a strong MHC binding is connected to immunogenicity. A recent study by Bjerregaard et al. (83), supports the theory that strong binders are immunogenic. Their study indicated that immunogenic neopeptides bind significantly stronger compared to non-immunogenic peptides and that they in general bind with a strong affinity. However, Duan et al. (49) deemed binding affinity scores alone, especially from NetMHC, as not being an effective predictor of tumor rejection and immunogenicity. In fact, in their study they noticed that the epitopes that did elicit tumor protection were in general not strong MHC class I binders. They therefore created an algorithm which subtracts the predicted NetMHC scores of unmutated counterpart peptides from the NetMHC scores of the mutated peptides. This setup is referred to as the differential agretopicity index (DAI). The idea is that this can reflect to which degree the binding of mutated peptides differ from their unmutated counterparts (49). Even this score, however, performed poorly for identifying effective neoepitopes (84). Similar indications have also been made by (85) and (86), where it was shown that not only peptides predicted as strong binders but also peptides predicted as weak binders or non-binders are capable of initiating a T cell response. At the current stage, there's no clear consensus on the importance of MHC binding for identifying dominant epitopes and neoepitopes. Further studies will be needed to decide if and how the threshold of 500 nM routinely being used as a threshold for peptide selection should be reconsidered.

The lack of experimental data is also among the causes of another potential problem. The datasets that are used to train these models are often very redundant: they contain many epitopes that are either identical or very similar. If not properly managed, redundancy can cause the tools to overfit: this means that their actual prediction accuracy on new data will be worse than the one reported in the publications. As a general suggestion, we encourage the users to check that the tools they are using take redundancy into account, for example by performing homology reduction procedures (87), rather than basing their choice on a purely numerical comparison of the accuracies reported in the papers.

A potentially very important but much less studied area is PTMs. Different PTMs exist such as phosphorylation, ubuiquitinylation, glycosylation, methylation, citrullination, to name a few. PTMs have been thought to be potential neoepitope candidates. This is based on the theory that peptides with aberrant PTMs have not been exposed to the immune system and thus potentially not subject to central tolerance. It has been shown that PTM self-antigens are capable of escaping central tolerance and being recognized by the immune system (88). Aberrant PTMs have been discovered in multiple cancers. Increased levels of glycans have for example been observed in cancers such as breast cancer (89, 90). However, identifying glycosylation sites as well as other PTM sites is not an easy task. In general mass spectrometry is often not capable of identifying less abundant proteins, due to its low sensitivity, thus capturing PTM information can be difficult due to the general low abundance.

Another lesser explored avenue are neoantigens derived from generally considered non-coding regions of the genome. Since they are less explored and studied, they are less utilized for analysis. Despite this, Laumont et al. (43) showed in their recent study that non-coding regions were possibly a considerable source of neoantigens.

There are still many events which are partially or completely disregarded by the current prediction models but can affect peptide binding and T cell recognition. Some examples include PTMs, local environment, self-similarity, clonality, and non-coding derived peptides. Moving forward, a tool which covers as many different neoepitope causing events as possible would be ideal. Another open question is whether some genomic aberrations are more effective than others for attacking the cancer cells. This begs the question of whether this is a generalized property or inherently specific for individual cancers, thus impairing the effectiveness of one-fits-all models.

Some of the tools presented in this review have been used in developing therapies that are being tested in ongoing clinical and pre-clinical trials. To mention a few, the development of neoantigen targeted personalized cancer treatments for cancers such as melanoma (91), glioblastoma (92) and non-small cell lung cancer (93) have been showing promising results. In particular, the use of tools that rely heavily on mhc binding prediction has propelled the discovery of candidates for test and use in targeted personalized immunotherapy in these studies. Even though these trials had encouraging results, they have also met some limitations in regards to the efficiency of the targeted immunotherapy, indicating that we are still in the early stages of development for neoepitope prediction tools. We envision that a growing amount of evidence on neoepitopes and on the ability of different tools to predict them will have a major impact on the development of better epitope and neoepitope prediction tools, and in turn help guide future immunotherapies.

AUTHOR CONTRIBUTIONS

A-LS-J and PM conceived and wrote the paper. MV created the figures together with A-LS-J and corrected and commented the paper. AB and SH corrected and commented the paper. All authors contributed to the article and approved the submitted version.

FUNDING

A-LS-J is funded by the 2018 SDC grants.

REFERENCES

1. Burnet FM. Immunological Aspects of Neoplasia. In: Schwartz RS, editor. *Prog Tumor Res.* Basel, Karger (1970). vol 13, pp. 1-27. doi: 10.1159/000386035
2. Thomas L. On Immunosurveillance in Human Cancer. *Yale J Biol Med* (1982) 55:329–33. doi: 10.18632/oncotarget.2998
3. Galluzzi L, Vacchelli E, Bravo-San Pedro JM, Buqué A, Senovilla L, Baracco EE, et al. Classification of Current Anticancer Immunotherapies. *Oncotarget* (2014) 5:12472–508. doi: 10.18632/oncotarget.2998
4. Humphries C. Adoptive Cell Therapy: Honing That Killer Instinct. *Nature* (2013) 504. doi: 10.1038/504S13a
5. Nagasawa DT, Fong C, Yew A, Spasic M, Garcia HM, Kruse CA, et al. Passive Immunotherapeutic Strategies for the Treatment of Malignant Gliomas. *Neurosurgery Clinics of North America* (2012) 23(3):481-95. doi: 10.1016/j.nec.2012.04.008
6. Satoh Y, Esche C, Gambotto A, Shurin GV, Yurkovetsky ZR, Robbins PD, et al. Local Administration of IL-12-Transfected Dendritic Cells Induces Antitumor Immune Responses to Colon Adenocarcinoma in the Liver in Mice. *J Exp Ther Oncol* (2002) 2:337–49. doi: 10.1046/j.1359-4117.2002.01050.x
7. Rice J, Ottensmeier CH, Stevenson FK. DNA Vaccines: Precision Tools for Activating Effective Immunity Against Cancer. *Nat Rev Cancer* (2008) 108-20. doi: 10.1038/nrc2326
8. Singh-Jasuja H, Emmerich NP, Rammensee HG. The Tübingen Approach: Identification, Selection, and Validation of Tumor-Associated Hla Peptides for Cancer Therapy. *Cancer Immunol Immunother* (2004) 53:187–95. doi: 10.1007/s00262-003-0480-x
9. Mommen GP, Frese CK, Meiring HD, Gaans-van Den Brink J, De Jong AP, Van Els CA, et al. Expanding the Detectable HLA Peptide Repertoire Using Electron-Transfer/Higher-Energy Collision Dissociation (EThcD). *Proc Natl Acad Sci USA* (2014) 111:4507–12. doi: 10.1073/pnas.1321458111
10. Falk K, Rotzschke O, Stevanovic S, Jung G, Rammenseet HG. Allele-Specific Motifs Revealed by Sequencing of Self-Peptides Eluted From MHC Molecules. *Tech Rep* (1991) 351:290–6. doi: 10.1038/351290a0
11. Peters B, Nielsen M. Sette A. T Cell Epitope Predictions. *Annu Rev Immunol* (2020) 38:123–45. doi: 10.1146/annurev-immunol-082119-124838
12. Tung CW, Ho SY. POPI: Predicting Immunogenicity of MHC Class I Binding Peptides by Mining Informative Physicochemical Properties. *Bioinformatics* (2007) 23:942–9. doi: 10.1093/bioinformatics/btm061
13. Tung CW, Ziehm M, Kämper A, Kohlbacher O, Ho SY. POPISK: T-Cell Reactivity Prediction Using Support Vector Machines and String Kernels. *BMC Bioinf* (2011) 12:446. doi: 10.1186/1471-2105-12-446
14. Calis JJ, Maybeno M, Greenbaum JA, Weiskopf D, De Silva AD, Sette A, et al. Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. *PloS Comput Biol* (2013) 9:1003266. doi: 10.1371/journal.pcbi.1003266
15. Saethang T, Hirose O, Kimkong I, Tran VA, Dang XT, Nguyen LAT, et al. PAAQD: Predicting Immunogenicity of MHC Class I Binding Peptides Using Amino Acid Pairwise Contact Potentials and Quantum Topological Molecular Similarity Descriptors. *J Immunol Methods* (2013) 387:293–302. doi: 10.1016/j.jim.2012.09.016
16. Jørgensen KW, Rasmussen M, Buus S, Nielsen M. NetMHCstab - Predicting Stability of Peptide-MHC-I Complexes; Impacts for Cytotoxic T Lymphocyte Epitope Discovery. *Immunology* (2014) 141:18–26. doi: 10.1111/imm.12160
17. Rasmussen M, Fenoy E, Harndahl M, Kristensen AB, Nielsen IK, Nielsen M, et al. Pan-Specific Prediction of Peptide–MHC Class I Complex Stability, a Correlate of T Cell Immunogenicity. *J Immunol* (2016) 197:1517–24. doi: 10.4049/jimmunol.1600582
18. Schmidt J, Guillaume P, Dojcinovic D, Karbach J, Coukos G, Luescher I. In Silico and Cell-Based Analyses Reveal Strong Divergence Between Prediction and Observation of T-Cell-Recognized Tumor Antigen T-Cell Epitopes. *J Biol Chem* (2017) 292:11840–9. doi: 10.1074/jbc.M117.789511
19. Trolle T, Nielsen M. NetTepi: An Integrated Method for the Prediction of T Cell Epitopes. *Immunogenetics* (2014) 66:449–56. doi: 10.1007/s00251-014-0779-0
20. Karosiene E, Lundegaard C, Lund O, Nielsen M. NetMHCcons: A Consensus Method for the Major Histocompatibility Complex Class I Predictions. *Immunogenetics* (2012) 64:177–86. doi: 10.1007/s00251-011-0579-8
21. Jurtz VI, Jessen LE, Bentzen AK, Jespersen MC, Mahajan S, Vita R, et al. NetTCR: Sequence-Based Prediction of TCR Binding to Peptide-MHC Complexes Using Convolutional Neural Networks. *bioRxiv* (2018). doi: 10.1101/433706
22. Rammensee HG, Bachmann J, Emmerich NPN, Bachor OA, Stevanović S. SYFPEITHI: Database for MHC Ligands and Peptide Motifs. *Immunogenetics* (1999) 50:213-9. doi: 10.1007/s002510050595
23. Fleri W, Paul S, Dhanda SK, Mahajan S, Xu X, Peters B, et al. The Immune Epitope Database and Analysis Resource in Epitope Discovery and Synthetic Vaccine Design. *Front Immunol* (2017). doi: 10.3389/fimmu.2017.00278
24. Bagaev DV, Vroomans RM, Samir J, Stervbo U, Rius C, Dolton G, et al. VDJdb in 2019: Database Extension, New Analysis Infrastructure and a T-Cell Receptor Motif Compendium. *Nucleic Acids Res* (2020) 48:D1057–62. doi: 10.1093/nar/gkz874
25. Tickotsky N, Sagiv T, Prilusky J, Shifrut E, Friedman N. McPAS-TCR: A Manually Curated Catalogue of Pathology-Associated T Cell Receptor Sequences. *Bioinformatics* (2017) 33:2924–9. doi: 10.1093/bioinformatics/btx286
26. Borrman T, Cimons J, Cosiano M, Purcaro M, Pierce BG, Baker BM, et al. ATLAS: A Database Linking Binding Affinities With Structures for Wild-Type and Mutant TCR-pMHC Complexes. *Proteins: Struct Funct Bioinf* (2017) 85:908–16. doi: 10.1002/prot.25260
27. Leem J, De Oliveira SH, Krawczyk K, Deane CM. STCRDab: The Structural T-Cell Receptor Database. *Nucleic Acids Res* (2018) 46:D406–12. doi: 10.1093/nar/gkx971
28. Rudolph MG, Wilson IA. The Specificity of TCR/pMHC Interaction. *Current Opinion Immun* (2002) 14(1):52-65. doi: 10.1016/S0952-7915(01)00298-9
29. Bentzen AK, Such L, Jensen KK, Marquard AM, Jessen LE, Miller NJ, et al. T Cell Receptor Fingerprinting Enables in-Depth Characterization of the Interactions Governing Recognition of Peptide–MHC Complexes. *Nat Biotechnol* (2018) 36:1191–6. doi: 10.1038/nbt.4303
30. Knapp B. Deane CM. T-Cell Receptor Binding Affects the Dynamics of the Peptide/MHC-I Complex. *J Chem Inf Model* (2016) 56:46–53. doi: 10.1021/acs.jcim.5b00511
31. Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee JY, et al. Diversity and Clonal Selection in the Human T-Cell Repertoire. *Proc Natl Acad Sci USA* (2014) 111:13139–44. doi: 10.1073/pnas.1409155111
32. Jenkins MK, Chu HH, McLachlan JB, Moon JJ. On the Composition of the Preimmune Repertoire of T Cells Specific for Peptide–Major Histocompatibility Complex Ligands. *Annu Rev Immunol* (2010) 28:275–94. doi: 10.1146/annurev-immunol-030409-101253
33. Jensen KK, Rantos V, Jappe EC, Olsen TH, Jespersen MC, Jurtz V, et al. TCRpMHCmodels: Structural Modelling of TCR-pMHC Class I Complexes. *Sci Rep* (2019) 9. doi: 10.1038/s41598-019-50932-4
34. Gielis S, Moris P, Bittremieux W, De Neuter N, Ogunjimi B, Laukens K, et al. Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires. *Front Immunol* (2019) 10:2820. doi: 10.3389/fimmu.2019.02820
35. Li S, Wilamowski J, Teraguchi S, van Eerden FJ, Rozewicki J, Davila A, et al. Structural Modeling of Lymphocyte Receptors and Their Antigens. *Methods Mol Biol (Humana Press Inc)* (2019) 2048:207–29. doi: 10.1007/978-1-4939-9728-2_17
36. Lanzarotti E, Marcatili P, Nielsen M. Identification of the Cognate Peptide-MHC Target of T Cell Receptors Using Molecular Modeling and Force Field Scoring. *Mol Immunol* (2018) 94:91–7. doi: 10.1016/j.molimm.2017.12.019
37. Lanzarotti E, Marcatili P, Nielsen M. T-Cell Receptor Cognate Target Prediction Based on Paired α and β Chain Sequence and Structural CDR Loop Similarities. *Front Immunol* (2019) 10:2080. doi: 10.3389/fimmu.2019.02080
38. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab Initio Protein Structure Prediction of CASP III Targets Using ROSETTA. *Proteins* (1999) Suppl 3. doi: 10.1002/(SICI)1097-0134(1999)37:3+<171::AID-PROT21>3.3.CO;2-Q
39. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX Web Server: An Online Force Field. *Nucleic Acids Res* (2005) 33:W382–8. doi: 10.1093/nar/gki387
40. Aranha MP, Jewel YSM, Beckman RA, Weiner LM, Mitchell JC, Parks JM, et al. Combining Three-Dimensional Modeling With Artificial Intelligence to

Increase Specificity and Precision in Peptide–MHC Binding Predictions. *J Immunol* (2020) 205:1962–77. doi: 10.4049/jimmunol.1900918
41. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of Mutational Processes in Human Cancer. *Nature* (2013) 500:415–21. doi: 10.1038/nature12477
42. Malaker SA, Penny SA, Steadman LG, Myers PT, Loke JC, Raghavan M, et al. Identification of Glycopeptides as Posttranslationally Modified Neoantigens in Leukemia. *Cancer Immun Res* (2017) 5(5):376–84. doi: 10.1158/2326-6066.CIR-16-0280
43. Laumont CM, Vincent K, Hesnard L, Audemard R, Bonneil R, Laverdure JP, et al. Noncoding Regions Are the Main Source of Targetable Tumor-Specific Antigens. *Sci Trans Med* (2018) 10. doi: 10.1126/scitranslmed.aau5516
44. Solleder M, Guillaume P, Racle J, Michaux J, Pak HS, Müller M, et al. Mass Spectrometry Based Immunopeptidomics Leads to Robust Predictions of Phosphorylated Hla Class I Ligands. *Mol Cell Proteomics* (2020) 19:390–404. doi: 10.1074/mcp.TIR119.001641
45. Wickström SL, Lövgren T, Volkmar M, Reinhold B, Duke-Cohan JS, Hartmann L, et al. Cancer Neoepitopes for Immunotherapy: Discordance Between Tumor-Infiltrating T Cell Reactivity and Tumor MHC Peptidome Display. *Front Immunol* (2019) 10:2766. doi: 10.3389/fimmu.2019.02766
46. Smart AC, Margolis CA, Pimentel H, He MX, Miao D, Adeegbe D, et al. Intron Retention Is a Source of Neoepitopes in Cancer. *Nat Biotechnol* (2018) 36:1056–63. doi: 10.1038/nbt.4239
47. Grade M, Difilippantonio MJ, Camps J. Patterns of Chromosomal Aberrations in Solid Tumors. *Chromosomal Instability Cancer Cells (Springer Int Publ)* (2015) 200:115–42. doi: 10.1007/978-3-319-20291-4_6
48. Wei Z, Zhou C, Zhang Z, Guan M, Zhang C, Liu Z, et al. The Landscape of Tumor Fusion Neoantigens: A Pan-Cancer Analysis. *iScience* (2019) 21:249–60. doi: 10.1016/j.isci.2019.10.028
49. Duan F, Duitama J, Al Seesi S, Ayres CM, Corcelli SA, Pawashe AP, et al. Genomic and Bioinformatic Profiling of Mutational Neoepitopes Reveals New Rules to Predict Anticancer Immunogenicity. *J Exp Med* (2014) 211:2231–48. doi: 10.1084/jem.20141308
50. Tappeiner E, Finotello F, Charoentong P, Mayer C, Rieder D, Trajanoski Z. TIminer: NGS Data Mining Pipeline for Cancer Immunology and Immunotherapy. *Bioinformatics* (2017) 33:3140–1. doi: 10.1093/bioinformatics/btx377
51. Kim S, Kim HS, Kim E, Lee MG, Shin EC, Paik S, et al. Neopepsee: Accurate Genome-Level Prediction of Neoantigens by Harnessing Sequence and Amino Acid Immunogenicity Information. *Ann Oncol* (2018) 29:1030–6. doi: 10.1093/annonc/mdy022
52. Shi Y, Guo Z, Su X, Meng L, Zhang M, Sun J, et al. DeepAntigen: A Novel Method for Neoantigen Prioritization *via* 3D Genome and Deep Sparse Learning. *Bioinformatics* (2020) 36(19):4894–901. doi: 10.1093/bioinformatics/btaa596
53. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, et al. pVAC-Seq: A Genome-Guided *In Silico* Approach to Identifying Tumor Neoantigens. *Genome Med* (2016) 8:11. doi: 10.1186/s13073-016-0264-5
54. Bjerregaard AM, Nielsen M, Hadrup SR, Szallasi Z, Eklund AC. MuPeXI: Prediction of Neo-Epitopes From Tumor Sequencing Data. *Cancer Immunol Immunother* (2017) 66:1123–30. doi: 10.1007/s00262-017-2001-3
55. Mondet S, Aksoy BA, Rozenberg L, Hodes I, Hammerbacher J. Bioinformatics Workflow Management With the Wobidisco Ecosystem. *bioRxiv* (2017). doi: 10.1101/213884
56. Kodysh J, Rubinsteyn A. OpenVax: An Open-Source Computational Pipeline for Cancer Neoantigen Prediction. *Methods Mol Biol (Humana Press Inc)* (2020) 2120:147–60. doi: 10.1007/978-1-0716-0327-7_10
57. Wood MA, Nguyen A, Struck AJ, Ellrott K, Nellore A, Thompson RF. Neoepiscope Improves Neoepitope Prediction With Multivariant Phasing. *Bioinformatics* (2020) 36:713–20. doi: 10.1093/bioinformatics/btz653
58. Bais P, Namburi S, Gatti DM, Zhang X, Chuang JH. CloudNeo: A Cloud Pipeline for Identifying Patient-Specific Tumor Neoantigens. *Bioinformatics* (2017) 33:3110–2. doi: 10.1093/bioinformatics/btx375
59. Zhou C, Wei Z, Zhang Z, Zhang B, Zhu C, Chen K, et al. PTuneos: Prioritizing Tumor Neoantigens From Next-Generation Sequencing Data. *Genome Med* (2019) 11:67. doi: 10.1186/s13073-019-0679-x
60. Richman LP, Vonderheide RH, Rech AJ. Neoantigen Dissimilarity to the Self-Proteome Predicts Immunogenicity and Response to Immune Checkpoint Blockade. *Cell Syst* (2019) 9:375–82. doi: 10.1016/j.cels.2019.08.009
61. Schenck RO, Lakatos E, Gatenbee C, Graham TA, Anderson AR. NeoPredPipe: High-Throughput Neoantigen Prediction and Recognition Potential Pipeline. *BMC Bioinf* (2019) 20:264. doi: 10.1186/s12859-019-2876-4
62. Zhou Z, Lyu X, Wu J, Yang X, Wu S, Zhou J, et al. TSNAD: An Integrated Software for Cancer Somatic Mutation and Tumour-Specific Neoantigen Detection. *R Soc Open Sci* (2017) 4(4):170050. doi: 10.1098/rsos.170050
63. Wang TY, Wang L, Alam SK, Hoeppner LH, Yang R. ScanNeo: Identifying Indel-Derived Neoantigens Using RNA-Seq Data. *Bioinformatics* (2019) 35:4159–61. doi: 10.1093/bioinformatics/btz193
64. Fotakis G, Rieder D, Haider M, Trajanoski Z, Finotello F. NeoFuse: Predicting Fusion Neoantigens From RNA Sequencing Data. *Bioinformatics* (2020) 36:2260–1. doi: 10.1093/bioinformatics/btz879
65. Zhang J, Mardis ER, Maher CA. INTEGRATE-Neo: A Pipeline for Personalized Gene Fusion Neoantigen Discovery. *Bioinformatics* (2017) 33:555–7. doi: 10.1093/bioinformatics/btw674
66. Bulik-Sullivan B, Busby J, Palmer CD, Davis MJ, Murphy T, Clark A, et al. Deep Learning Using Tumor HLA Peptide Mass Spectrometry Datasets Improves Neoantigen Identification. *Nat Biotechnol* (2019) 37:55–71. doi: 10.1038/nbt.4313
67. Wu J, Wang W, Zhang J, Zhou B, Zhao W, Su Z, et al. DeepHLApan: A Deep Learning Approach for Neoantigen Prediction Considering Both HLA-Peptide Binding and Immunogenicity. *Front Immunol* (2019) 10:2559. doi: 10.3389/fimmu.2019.02559
68. Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, et al. Reliable Prediction of T-Cell Epitopes Using Neural Networks With Novel Sequence Representations. *Protein Sci* (2003) 12:1007–17. doi: 10.1110/ps.0239403
69. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, et al. NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence. *PloS One* (2007) 2: e796. doi: 10.1371/journal.pone.0000796
70. Nicolae M, Mangul S, Măndoiu II, Zelikovsky A. Estimation of Alternative Splicing Isoform Frequencies From RNA-Seq Data. *Algorithms Mol Biol* (2011) 6:9. doi: 10.1186/1748-7188-6-9
71. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing. *Nat Genet* (2008) 40:1413–5. doi: 10.1038/ng.259
72. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding Mechanisms Underlying Human Gene Expression Variation With RNA Sequencing. *Nature* (2010) 464:768–72. doi: 10.1038/nature08872
73. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Syst* (2018) 7:129–32. doi: 10.1016/j.cels.2018.05.014
74. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, et al. Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-Allelic Cells Enables More Accurate Epitope Prediction. *Immunity* (2017) 46:315–26. doi: 10.1016/j.immuni.2017.02.007
75. Warren RL, Choe G, Freeman DJ, Castellarin M, Munro S, Moore R, et al. Derivation of HLA Types From Shotgun Sequence Datasets. *Genome Med* (2012) 4:95. doi: 10.1186/gm396
76. Zhang J, White NM, Schmidt HK, Fulton RS, Tomlinson C, Warren WC, et al. INTEGRATE: Gene Fusion Discovery Using Whole Genome and Transcriptome Data. *Genome Res* (2016) 26:108–18. doi: 10.1101/gr.186114.114
77. van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, et al. Hi-C: A Method to Study the Three-Dimensional Architecture of Genomes. *J Vis Exp* (2010) (39):e1869. doi: 10.3791/1869
78. Yewdell JW, Bennink JR. Immunodominance in Major Histocompatibility Complex Class I-Restricted T Lymphocyte Responses. *Annu Rev Immunol* (1999) 17:51–88. doi: 10.1146/annurev.immunol.17.1.51
79. Sette A, Vitiello A, Reherman B, Fowler P, Nayersina R, Kast WM, et al. The Relationship Between Class I Binding Affinity and Immunogenicity of Potential Cytotoxic T Cell Epitopes. *J Immunol* (1994) 153:5586–92.
80. Croft NP, Smith SA, Pickering J, Sidney J, Peters B, Faridi P, et al. Most Viral Peptides Displayed by Class I MHC on Infected Cells Are Immunogenic. *Proc Natl Acad Sci USA* (2019) 116:3112–7. doi: 10.1073/pnas.1815239116

81. Zhong W, Reche PA, Lai CC, Reinhold B, Reinherz EL. Genome-Wide Characterization of a Viral Cytotoxic T Lymphocyte Epitope Repertoire. *J Biol Chem* (2003) 278:45135–44. doi: 10.1074/jbc.M307417200
82. Dönnes P, Kohlbacher O. Integrated Modeling of the Major Events in the MHC Class I Antigen Processing Pathway. *Protein Sci* (2005) 14:2132–40. doi: 10.1110/ps.051352405
83. Bjerregaard AM, Nielsen M, Jurtz V, Barra CM, Hadrup SR, Szallasi Z, et al. An Analysis of Natural T Cell Responses to Predicted Tumor Neoepitopes. *Front Immunol* (2017) 8:1566. doi: 10.3389/fimmu.2017.01566
84. Koşaloğlu-Yalçın Z, Lanka M, Frentzen A, Logandha Ramamoorthy Premlal A, Sidney J, Vaughan K, et al. Predicting T Cell Recognition of MHC Class I Restricted Neoepitopes. *OncoImmunology* (2018) 7:e1492508. doi: 10.1080/2162402X.2018.1492508
85. Fritsch EF, Rajasagi M, Ott PA, Brusic V, Hacohen N, Wu CJ. HLA-Binding Properties of Tumor Neoepitopes in Humans. *Cancer Immun Res* (2014) 2(6). doi: 10.1158/2326-6066.CIR-13-0227
86. Ghorani E, Rosenthal R, McGranahan N, Reading JL, Lynch M, Peggs KS, et al. Differential Binding Affinity of Mutated Peptides for MHC Class I Is a Predictor of Survival in Advanced Lung Cancer and Melanoma. *Ann Oncol* (2018) 29:271–9. doi: 10.1093/annonc/mdx687
87. Lund O, Frimand K, Gorodkin J, Bohr H, Bohr J, Hansen J, et al. Protein Distance Constraints Predicted by Neural Networks and Probability Density Functions. *Protein Eng* (1997) 10:1242–8. doi: 10.1093/protein/10.11.1241
88. Raposo B, Merky P, Lundqvist C, Yamada H, Urbonaviciute V, Niaudet C, et al. T Cells Specific for Post-Translational Modifications Escape Intrathymic Tolerance Induction. *Nat Commun* (2018) 9(1):353. doi: 10.1038/s41467-017-02763-y
89. de Leoz MLA, Young LJT, An HJ, Kronewitter SR, Kim J, Miyamoto S, et al. High-Mannose Glycans Are Elevated During Breast Cancer Progression. *Mol Cell Proteomics* (2011) 10.(1):M110.002717 doi: 10.1074/mcp.m110.002717
90. Tesařová P, Kalousová M, Trnková B, Soukupová J, Argalášová S, Mestek O, et al. Carbonyl and Oxidative Stress in Patients With Breast Cancer-Is There a Relation to the Stage of the Disease? *Tech Rep* (2007) 54:219–24.
91. Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, et al. An Immunogenic Personal Neoantigen Vaccine for Patients With Melanoma. *Nature* (2017) 547:217–21. doi: 10.1038/nature22991
92. Keskin DB, Anandappa AJ, Sun J, Tirosh I, Mathewson ND, Li S, et al. Neoantigen Vaccine Generates Intratumoral T Cell Responses in Phase Ib Glioblastoma Trial. *Nature* (2019) 565:234–9. doi: 10.1038/s41586-018-0792-9
93. Zhang W, Yin Q, Huang H, Lu J, Qin H, Chen S, et al. Personal Neoantigens From Patients With NSCLC Induce Efficient Antitumor Responses. *Front Oncol* (2021) 11:628456. doi: 10.3389/fonc.2021.628456

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

7 Can we predict T cell lineage form sequence only?

This chapter presents a manuscript in preparation where the abstract has been accepted in the journal; Frontiers in Immunology. T cells can, in general, be divided into two groups, which are CD8+ and CD4+ T cells. Both lineages are important for the adaptive immune system, but determining the lineage choice is still a topic that is garnering much academic interest. Therefore, in this work, we analyze if it is possible to predict the lineage of T cell, whether it is a CD8+ or CD4+ T cell, from its paired α and β TCR sequences. We studied this to investigate whether there are any clear patterns present in the data that may push the choice of lineage in a specific direction. We show that there is a small signal in the data that, to a certain extent, can help classify a lineage. However, in this work, we also show, with the help of logo plots and two sample logo plots, that TCR sequences overall are very similar across T cell lineages and that this is the case for multiple different datasets. We also discovered paired TCR sequences that were present on both CD8+ T cells and CD4+ T cells. All this information combined leads us to question how static the T cell lineage choice is and whether it is possible that T cells may be cross-reactive across MHC classes.

frontiers

Can we Predict T cell lineage from Sequence only?

Anna-Lisa Schaap-Johansen, Kamilla Kjærgaard Munk, Martin Closter Jespersen, Vanessa Isabell Jurtz, Tina Funck and Paolo Marcatili [1]

[1] *Department of Health Technology, Technical University of Denmark, Lyngby, Denmark*

Correspondence*:
Paolo Marcatili
pamar@dtu.dk

ABSTRACT

Cross-reactivity is a well-established property of T cells: a single T cell receptor (TCR) can bind up to one million different peptides presented by MHC molecules. This plasticity has been described extensively for peptides bound by a single MHC molecule and to a minor extent for peptides bound to different alleles of the same type. Here, we investigate if the TCR sequence determines, completely or in part, the type of MHC molecules it interacts with. T cells can be divided into two major groups, namely CD4+ or CD8+ T cells, with TCRs of the former group interacting with peptides presented by class II MHCs, while the latter group interacts with peptides presented by class I MHC. The two T cell groups use the same mechanism and machinery to produce functional T cell receptors; thus, identifying the T cell lineage from the TCR sequence alone is not a trivial task. Multiple theories have been formulated to explain lineage choice, and methods have been developed to try and predict it. In this paper, we present a tool for predicting lineage choice based on T cell receptor sequence only and explore the possibility of T cell cross-reactivity across MHC classes and how this may affect lineage choice prediction.

Keywords: Cross reactivity, T cell, TCR, T cell receptor

1 INTRODUCTION

T cells are cross-reactive: it is known that a T cell can interact with up to one million different peptides (1, 2). In this work, we try to establish if cross-reactivity is also possible for peptides presented by different major histocompatibility complex (MHC) classes. T cells interact with other cells via a T cell receptor (TCR) and a co-receptor, the most common being the CD4 and CD8 co-receptors. The majority of T cells express TCRs with an α- and β-chain, and each of these chains has three complementarity determining regions (CDRs), named CDR1, CDR2, and CDR3, which interacts with the peptide-MHC molecule. The CDR3 loop is the most variable part of the TCR. It is found in the center of the TCR binding site, where it interacts with the peptide, thus accounting for majority of the TCR specificity. The CDR1 and CDR2 loops are less variable and interact mostly with the MHC.

Most mature T cells are characterized by the mutually exclusive expression of either the CD4 or CD8 co-receptor molecule on the surface of the cell. These two T cell populations differ in their function and which MHC they bind to: CD4+ T cells are in general believed to bind to MHC class II, whereas CD8+ T cells bind to MHC class I. TCRs are recombined into their complete sequence before lineage choice occurs.

This raises the question of whether or not the choice of TCR sequence predetermines its intended MHC class.

Multiple theories have been suggested to solve the question of how the T cell lineage choice transpires (3). However, how bipotential thymocyte precursors decide whether to differentiate into a CD8+ cytotoxic T cell or a CD4+ helper T cell is a question that still has to be answered within the field of developmental immunology. A fundamental understanding of what drives lineage choice and what defines a lineage can improve comprehension of T cell receptor repertoires and potentially also advance prediction models based on TCRs and their ligands. A potential way to investigate the connection between a T cell lineage and its TCR sequence is to try and predict the former from the latter. If possible, this would indicate that TCR sequences are not randomly distributed among classes and that some form of selection is present.

Previous works have investigated if the lineage could be predicted from the complete TCR sequence or part of it. As of now, the most well-known models for predicting CD4/CD8 lineage is the support vector machine (SVM) algorithm from Li *et al.* (4) and the Extreme Gradient Boosted decision tree classifier using the XGBoost implementation from Carter *et al.* (5), which is the current state of the art. Li *et al.* used TCR β CDR3 sequences from CD4+ and CD8+ T cells as input for their SVM model. The CDR3 amino acid sequences were converted to numerical arrays consisting of Atchley factors (6). They did not introduce any gaps or pad the sequences; hence they created a SVM model for each length present in the dataset. Although this allows for potential amino acid preferences to be discovered, the length dependency reduces the amount of data available to be trained on and can introduce biases that are unaccounted for. Carter *et al.* showed that another downside to this setup is that the same β CDR3 can be present on both CD4+ and CD8+ T cells, which is not accounted for in the SVM model. Furthermore, Carter *et al.* showed that paired α/β TCR sequences hold more information due to, as they suggested, the presence of synergistic information within the pairing of the α and β chain. As mentioned previously, Carter *et al.* utilized an XGBoost implementation, which takes as inputs paired α and β sequences represented by their V and J genes categorically encoded, together with the length of the CDR3, the CDR3 charge as well as the amino acid frequencies found in the CDR3. One downside to this strategy is that this sequence representation removes any detailed pattern that may be present in the complete α and β sequences.

In this paper, we develop a machine learning approach to study whether we can identify the T cell lineage from its complete TCR sequence only. Interestingly, even though the sequences from CD4+ and CD8+ T cells showed very similar composition and profiles, such model shows a moderate predictive power, supporting the hypothesis of a non-random selection of T cell lineage. Surprisingly, patterns and data in the datasets suggest that cross-reactivity may exist across MHC classes.

2 METHODS

2.1 Data collection

This paper uses four different datasets. The first dataset, which the models were trained on, comes from a single cell sequencing (SCS) experiment by Carter *et al.* (5). This dataset was downloaded from the github repository `https://github.com/JasonACarter/CD4_CD8-Manuscript`. In total seven samples were collected, consisting of α and β paired CD4/CD8 T cells from the peripheral blood of healthy human individuals.

The second dataset was data collected from the VDJdb database (7, 8), which was downloaded from VDJdb.cdr3.net on 22/10/2021. The VDJdb database consists of data from published studies which has been manually parsed into a VDJdb format following VDJdbs own guidelines.

The third dataset was obtained from the McPAS-TCR database (9), which was downloaded from `http://friedmanlab.weizmann.ac.il/McPAS-TCR/` on 28/10/2021. The database consists of manually curated pathology associated T cell receptor sequences gathered from published experimental data. The VDJdb and McPAS datasets were combined to increase the dataset size for downstream experiments.

The fourth and last dataset used in this study originates from samples collected in a previous study (10). The data consists of CD4+/CD8+ T cells isolated from peripheral blood collected from five healthy monozygotic human twin pairs. Unlike the other three datasets, the twin dataset does not contain information about the pairing. However, the large dataset could still deem it useful in regards to analyzing potential patterns present in the dataset.

2.2 Data processing

The TCR sequences of all the datasets were reconstructed using in-house scripts (11). The TCR sequences were reconstructed by using the CDR3 and the V and J gene information. Each reconstructed TCR sequence was then aligned according to the IMGT numbering scheme (12) and saved as an aligned sequence with gaps, with a final length of 138 amino acids.

The original single cell sequencing dataset consists of a total of 97,504 sequences. After processing, a total of 89,428 sequences are left, where 64,500 are CD4+ T cells and 24,928 are CD8+ T cells. The discarded sequences were due to these sequences not complying with the rules set by the in-house scripts, such as the absence of phenylalanine (F) or tryptophan (W) followed by a glycine (G) at the end of the CDR3 sequence.

The VDJdb and McPAS-TCR databases host TCR sequences from other species beyond humans. For consistency with the single cell dataset, we removed all non-human sequences, as well as any unpaired TCR sequences. We also removed sequences with missing V and J genes and TCR sequences having characters instead of letters in their amino acid sequences. The in-house scripts were then used to get the reconstructed α and β TCR sequences. After processing the VDJdb and McPAS-TCR dataset with our in-house scripts, a total of 21,963 sequences remain, where 20,962 of these sequences are from VDJdb and 1,001 from the McPAS dataset. Of the 21,963 sequences, 170 are CD4+ T cell and the remaining 21,793 CD8+ T cell sequences.

The twin dataset originally consisted of 181,285,548 raw sequencing reads. The sequencing data was cleaned, merged, TCR sequences were reconstructed using the in-house scripts, and any sequences that had rearranged loci that were not productive were removed. This led to the final twin dataset consisting of 634,024 α chains and 931,076 beta chains. The V-QUEST tool from IMGT was used to find sequences that were productive.

2.3 Machine learning

To reduce the possible effect of overfitting, we adopted the homology partitioning approach (13). The single cell sequencing data was clustered using MMSeqs2 (14) at 80% identity. The clusters were partitioned in a train, validation, and test set, covering 70%, 10%, and 20% of the processed data, respectively, and contained similar CD4/CD8 ratios. These datasets will be referred to as the internal single cell sequencing (SCS) train, validation, and test set as they were used to train and test the two different machine learning models presented in this paper.

The VDJdb-McPAS dataset and a subset of the VDJdb-McPAS dataset were used as external test sets to evaluate the performance of the models. The first test set consisted of the full VDJdb-McPAS combined dataset. The second test was created by clustering the VDJdb-McPAS sequences with the sequences from

the single cell sequencing data with an 80% similarity threshold. Any sequences from VDJdb-McPAS clustering together with sequences used to train the model were discarded. These datasets will be referred to as the full VDJdb-McPAS dataset and the clustered VDJdb-McPAS dataset.

Using the internal SCS train and validating set we trained a convolutional neural network (CNN) to classify CD4+ T cell sequences from CD8+ T cell receptor sequences. The network consists of two 1D convolutional layers, with batch normalization before input, ReLU as the activation function and max-pooling after each convolution. Batchnormalization is also performed after maxpooling for the second convolutional layer. Both convolutional layers have a kernel size of 3 and a stride of 2, the first convolutional layers outputs 50 filters and the second 25 filters. After the batch normalization of the second pooled convolutional layer, we place a dropout layer. Finally, a feed-forward linear layer with 425 hidden unit is present after the dropout and before the output layers. We use BCEWithLogitsoss as the loss function and Adaptive Moment Estimation (Adam) as the optimization algorithm and a batch size of 128. The model structure is a binary classification problem, where 0 denotes CD4+ T cell inputs and 1 denotes CD8+ T cell inputs. Input sequences are encoded using a BLOSUM62 (15) encoding scheme and gaps are encode with zeroes. We train the CNN for 146 epochs, with early stopping set to stop training if validation loss had not decreased for 50 epochs. We test the model on the internal SCS test set and the two versions of the external VDJdb-McPAS test datasets. An illustration of the CNN model can be seen in figure 1.

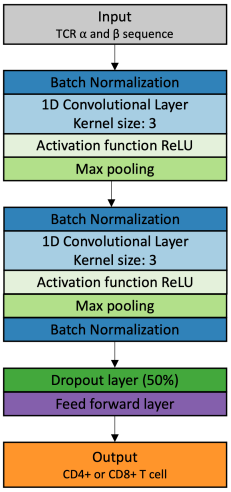


Figure 1. Network architecture of the CNN.

We generated an Extreme Gradient Boosted decision tree classifier using the Python XGBoost implementation. This XGBoost was trained using the internal SCS train set with default parameters as descried by Carter *et al.* in their original model setup. The α and β TCR chain sequences were separately represented by their V and J regions categorically encoded individually, CDR3 length, CDR3 charge, and amino acid frequencies in the CDR3 in that order. The XGBoost was then trained on the encoded paired α and β TCR sequences. The encoding space contained unique V and J genes found in both the single cell sequencing and VDJdb-McPAS combined dataset. The model was tested on an internal single cell sequencing test set and the two versions of the external VDJdb-McPAS test datasets. The train and validation sets were combined into one dataset when training the XGBoost model, with a similar CD4/CD8 ratio as before combining the two datasets.

2.4 Performance measures

The predictive performance of the different machine learning models was measured using the area under the receiver operator characteristic curve (AUC). The receiver operator characteristic (RUC) curve is an evaluation metric for binary classification problems, and the AUC is a measure of the model's ability to distinguish between classes - in this case, TCR sequences belonging to either CD4+ or CD8+ T cells. The higher the AUC, the better the performance of the model.

2.5 Logo plots

Logo plots were created utilizing the Logomaker software from (16). Logomaker requires the sequences to be of the same length; Logomaker was therefore used to create logo plots for the reconstructed and aligned full length sequences of both the α and β chain. Logomaker was also used to create logo plots for CDR3s of length ten and fifteen to ensure that no potential bias had been created after using in-house scripts to reconstruct and align the sequences.

2.6 Two sample logo plots

Two sample logo plots were produced for both the full α and β sequences as well as the CDR3 section of the α and β sequences after being generated using the in-house scripts. The two sample logo plots were created using the software from (17), where a two sample t-test was used and with everything at default except for correcting the p-value using the Bonferroni correction. The two sample logo plot software requires a "positive sample" and "negative sample" in the setup used in this study; CD8 sequences were regarded as "positive sample" and CD4 sequences as "negative sample".

3 RESULTS

We first analyze the SCS dataset for the presence of T cells of different lineage expressing identical TCR sequences. We then proceed to analyze the logo plots of TCR sequences from the different lineages, and eventually, we train a deep neural network to predict the lineage from the paired TCR sequences and analyze the results on different datasets.

3.1 Dataset analysis

It is known that T cells can behave in a cross-reactive manner recognizing multiple peptides. In this paper, we study whether cross-reactivity also can be observed across MHC classes.

In the original SCS dataset, we observe that 632 paired TCR sequences are reported as originating both from a CD4+ and a CD8+ T cell, leading to a total of 1271 samples in the dataset. These samples shared identical CDR3 α and β and the same V and J gene for both α and β. This means that out of a total of 97,504 sequences, 1271 of them had double labels, amounting to around 1,3% of the data having double labels, and 0.6% of the total data being uniquely double labeled, meaning each of the double labels counted only once. Although not a substantial amount, this could still potentially be of interest.

We, therefore, analyzed the logo plots derived from single and double labeled CDR3s from the TCRs (figure 2), we do not observe any significant dissimilarity. This is confirmed by the statistical analysis performed using the two sample logo webserver: we did not discover any individual position which hosted a difference of more than 6.3% between the double and single label T cell receptors (results in supplemental Figure 1). We also analyze if any gene is overrepresented in the setup of the double label vs. single label setup, however, we did not find any. This indicates that the germline does not seem to harbor any information regarding double lineage.

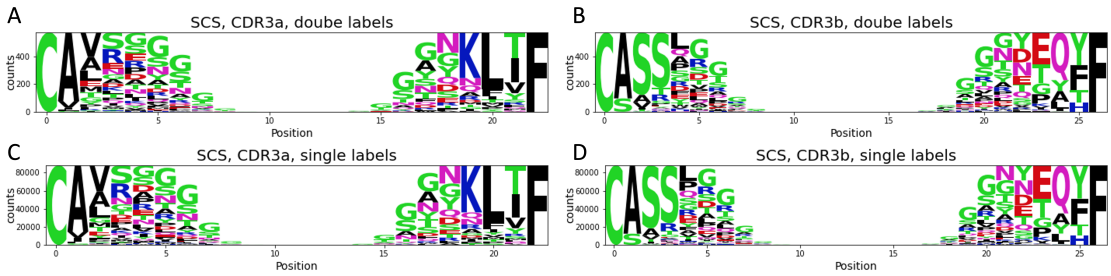


Figure 2. Differences between single and double labeled TCRs within the single cell dataset. Double labeled TCRs have paired TCR sequences that are labeled as both a CD4+ and a CD8+ T cell whereas single labeled TCRs are denoted as either CD4+ or CD8+ T cells. Logo plots showing the difference in the CDR3 region of the α (A) and β chain (B) in the double labeled TCRs and the α (C) and β chain (D) in the single labeled TCRs.

In the logo plots, we see a CAV motif at the beginning positions in the α sequences and a CASS motif in the β sequences, and phenylalanine (F) at the last position for both the α and β sequences shown with tall letters in the logo plots. This is because these amino acids on these positions are generally very conserved and therefore occur in the majority of the TCR sequences. Amino acids, which are present but less frequent at a given position, are shown with a smaller heights to indicate this information.

We then create logo plots to compare CD4+ and CD8+ TCR sequences. Figure 3 illustrates that CD4 CDR3 α and β and CD8 CDR3 α and β have no major difference at any position in the logo plots. The full sequence logo plots for this comparison can be found in supplemental Figure 2.

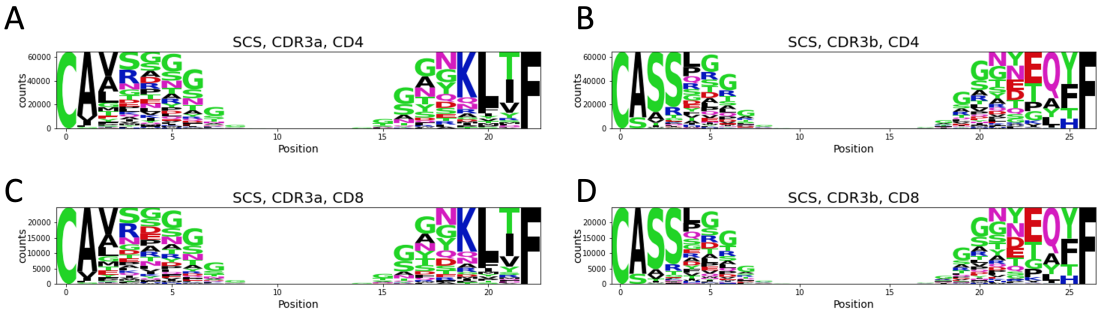


Figure 3. Differences between CD4+ and CD8+ TCRs in the single cell dataset. Logo plots showing the difference in the CDR3 region of the CD4+ TCRs α chain (A) and β chain (B) and the CD8+ TCRs α chain (C) and β chain (D).

To test whether this is an artifact of the alignment protocol used to process the sequences, we gather sequences of the same lengths from the raw data and analyze the corresponding logo plots obtained. As illustrated in figure 4 for CDR3 β with sequences of length 15 and 10, the raw single cell sequencing data prior to processing show the same tendency and no major distinction is present between CD4 CDR3 β and CD8 CDR3 β regardless of length. The same tendency was also found for CDR3 α sequences, as seen in supplemental Figure 3.

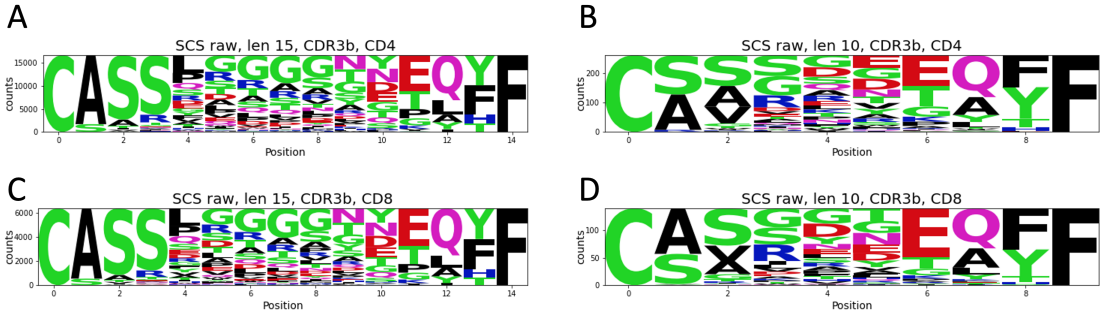


Figure 4. Logo plots showing similarities and differences in the CDR3 β chain of TCR sequences of length 15 and 10. SCS CDR3 β sequences for CD4 of length 15 (A), SCS CDR3 β sequences for CD4 of length 10 (B), SCS CDR3 β sequences for CD8 of length 15 (C), SCS CDR3 β sequences for CD8 of length 10 (D).

Similar tendencies may be present in other datasets. To study whether this is the case, we create logo plots for a twin dataset and VDJdb combined McPAS dataset, namely the VDJdb-McPAS dataset, gathered from the VDJdb and McPAS databases. As can be gathered from figure 5, the logo plots share the same characteristic of there being no clear discrepancy between the CD4 β and CD8 β logo plots. Furthermore, it is also evident from the different logo plots in figure 5 that the logo plots are very comparable between datasets. This is also the case for the CDR3 α sequences in supplemental Figure 4 and the TCRα (supplemental Figure 5) and TCRβ (supplemental Figure 6) sequences.

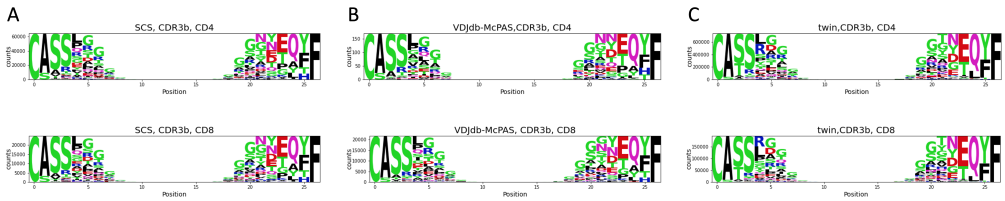


Figure 5. Differences and similarities between the CDR3β chain within different datasets. Comparing logo plots between the single cell dataset (A), the VDJdb-McPAS dataset (B) and the twin dataset (C).

The similarities between the logo plots of the different datasets are quite intriguing. Therefore, we investigate this with a two-sample logo plot, which will indicate whether any statistical differences between CD4 and CD8 per position are present and whether any statistical differences are comparable across the different datasets. As displayed in figure 6, which showcases the two-sample logo plots for the TCR β sequences, there are some statistical differences per position between CD4 and CD8. However, these can be considered relatively minor. Furthermore, the statistical differences are inconsistent throughout the different datasets and even somewhat contradict each other in certain instances. Similar tendencies hold true for the TCRα sequences, present in supplemental Figure 7.

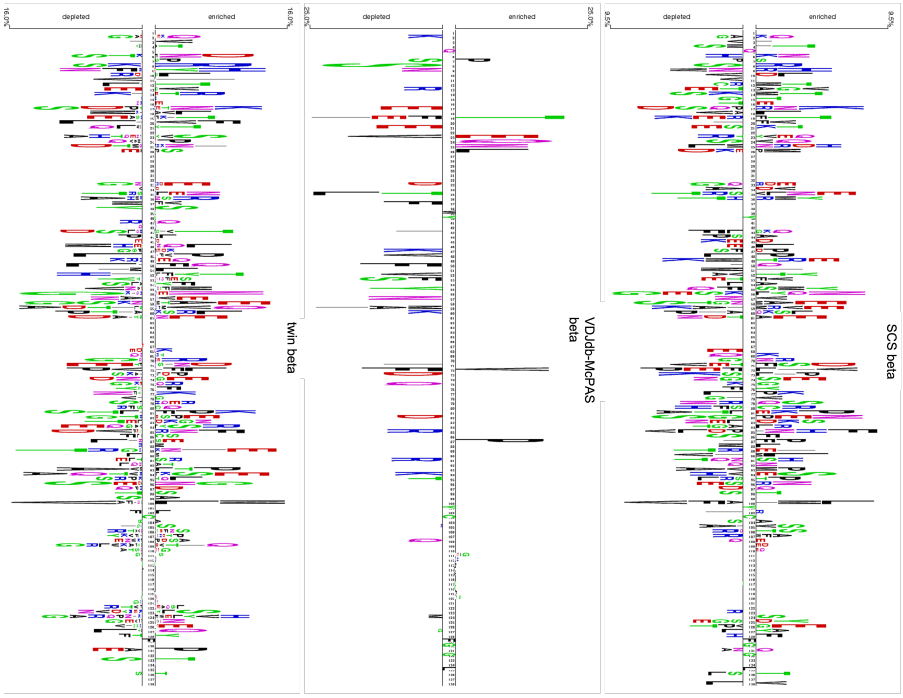


Figure 6. Two-sample logo plot showing the differences in the TCRβ sequences from CD4+ and CD8+ T cells within the different datasets. Comparing two-sample logo plots between the single cell dataset (A), the VDJdb-McPAS dataset (B) and the twin dataset (C). Here an enrichment indicates that a given amino acid at a given position is upregulated in CD8+ T cell TCRβ sequences and vice versa.

3.2 Machine Learning analysis of CD4+/CD8+ TCRs

The logo plots and two sample logo plots were not able to detect any clear patterns. However, it is possible that if the patterns are very complicated, more complex models are needed to discover those patterns.

In the original paper by Carter *et al.*, an Extreme Gradient Boosted decision tree classifier as a model was utilized, which obtained an AUC of 0.64 as their highest AUC. In their approach, the V and J genes were represented using a one-hot encoding, whereas the CDR3 were represented by their length, amino acid composition, and overall charge. We first investigate whether a more complicated model combined with a more informative encoding scheme and a different splitting setup could improve the predictions. Given its ability to discover local patterns in sequence data, we train a CNN. For comparison with the original model, we also train an Extreme Gradient Boosted decision tree classifier. However, as can be seen in figure 7a, we achieve comparable AUC values between our own model and the newly trained Extreme Gradient Boosted decision tree classifier at an AUC of 0.66.

We then check the ability of the models to perform on a different dataset, namely the VDJdb-McPAS dataset, and as can be seen in figure 7b, the CNN outperforms the XGBoost model. The CNN model achieves an AUC of 0.75 and the XGBoost an AUC of 0.65.

The VDJdb-McPAS dataset may contain sequences that have an 80% similarity or higher to the data the models have been trained on. We, therefore, wanted to further test the model on how well it performs on data with less than 80% similarity to the data the model is trained on. As can be observed in figure 7c,

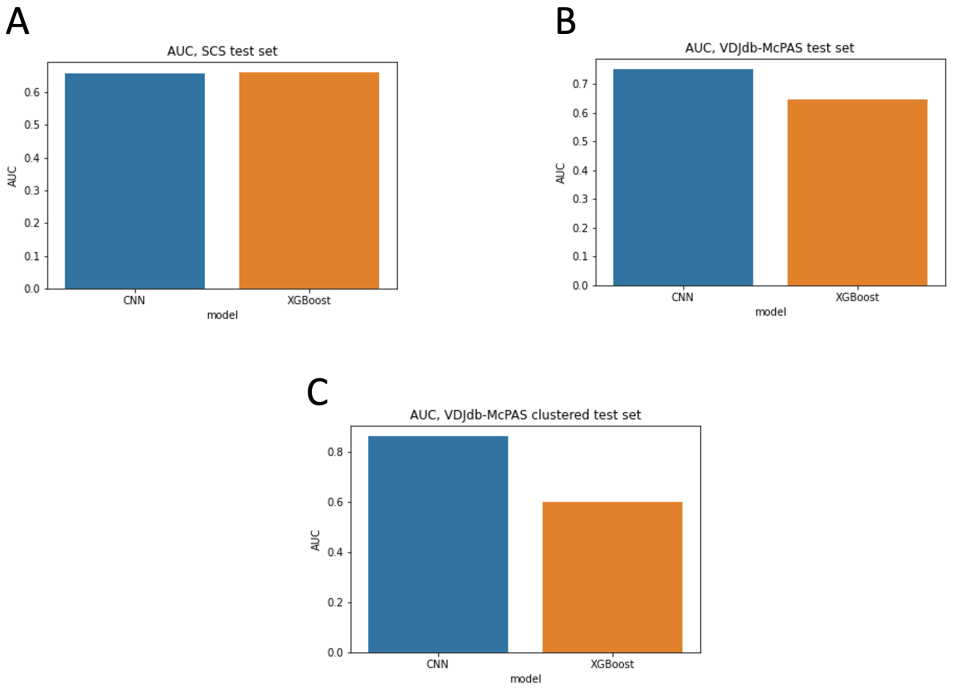


Figure 7. Performance estimation of the CNN and XGBoost models using AUC for the different test sets; SCS test set (A), VDJdb-McPAS test set (B) and VDJdb-McPAS clustered test set (C).

the CNN model obtains an AUC of 0.86, whereas the XGBoost model had an AUC of 0.60. The results gathered from testing the model on the VDJdb-McPAS and the VDJdb-McPAS subset with sequences of less than 80% similarity to the data used to train the models indicate that the CNN model is better at generalizing compared to the XGBoost model.

It has been mentioned that the frequency of amino acids and charge of a T cell receptor can have an impact on the lineage a T cell belongs to (5). However, we did not observe this to improve our model when predicting T cell lineage (results not included).

As mentioned earlier, we do not observe any clearly conserved patterns between the two sample logo plots constructed from the SCS, twin, and VDJdb-McPAS dataset to uncover any potential statistical differences between CD4+ and CD8+ T cells using TCRα and TCRβ sequences. We also observe that some positions show contradicting enrichments of amino acids between the different two sample logo plots. We were unable to discern a considerable difference between logo plots created for CD4+ and CD8+ T cell sequences, whether looking at a subset of the TCR sequence, the CDR3, or the full TCR sequence, both for α and β. These observations can, to a certain degree, explain why this is such a difficult task to predict. However, although we do not notice a clear cut distinction between CD4+ and CD8+ T cell sequences from the logo plot and two sample logo plots, we still obtain a signal when using more complex methods to predict this task.

4 DISCUSSION

In this study, we show that there is a signal - albeit not strong - in regards to predicting T cell lineage from sequence only. The signal is not clearly identified in the amino acid composition at specific positions, and a more complex model is needed for better generalization when predicting on new data.

We have created a model which keeps the complete TCR sequence information while still having the input be independent of the CDR3 length. This was obtained by employing in-house scripts in our pipeline, which outputs aligned reconstructed sequences using the CDR3 together with the V and J gene, each with a total length of 138 gapped amino acid sequences. This allows for the inputs to be of constant length while also retaining the original CDR3 sequences and allowing for the discovery of any particular amino acids or positions playing an important role in lineage choice. When using different approaches, e.g., training only on the CDR3 sequence, we would observe a significant drop in performance (data not shown). We show that a more complex model compared to the current models in the field improves the prediction. We chose to implement a convolutional neural network (CNN) model due to CNNs being ideal for detecting local spatial relations.

An Extreme Gradient Boosted decision tree classifier was trained to enable comparison between the CNN and the results in the SCS paper. The encoding of the data and training of the XGBoost model was kept as close as possible to the originally stated setup in (5). However, few changes were made to enable comparison between methods. We expanded the encoding space to enable testing of the trained model on the VDJdb-McPAS combined dataset. It should be mentioned that although we expanded the encoding space for the XGBoost, this did not show an effect on the results. The model obtained the same results when using the original encoding space on the SCS data as when using the expanded encoding scheme (results not included), and we, therefore, considered it acceptable to use the expanded encoding space going forward. The way the data was divided also diverged from the original paper. We chose to cluster the data based on 80% sequence similarity prior to splitting the data, where sequences of 80% or higher similarity are clustered together. Unlike the original paper, we chose to cluster the data to reduce redundancy that may be present in the data since this can have a negative impact on the ability of the method to generalize and thus predict well on new data. The data was split so that sequences from the same 80% or higher similarity thresholds would be present in the same splits, while each split would contain similar CD4/CD8 ratios. The original paper used StratifiedKFold, where the main idea is to generate datasets, where each set contains as close as possible the same distribution of classes. However, this method does not account for sequence similarity. Lastly, we trained the XGBoost model on a bigger dataset compared to the original article; this due to the authors choosing to only train and test on a unique set of TCR sequences and removed sequences that could be found as both CD4+ and CD8+ T cells.

In our results, we observed logo plots that were similar between CD4+ and CD8+ T cells. A potential argument for the seeming absence of a clear distinction between TCRs from CD4+ and CD8+ T cells in the logo plots and why we did not observe a strong signal could be that lineage choice is not mainly driven by TCRs having specific patterns that are capable of only interacting with either a MHC I or MHC II bound peptide, but rather other factors. This idea is further strengthened by experiments performed by Matechak *et al.* (18) and Kirberg *et al.* (19), which both showed that supposedly class II specific TCRs do not only generate CD4 T cells but also CD8 T cells, albeit in lower amounts. Matechak *et al.* (18) furthermore showed that in the absence of CD4, cells with class II specific TCRs would differentiate into the CD8 lineage, with amounts comparable to the amount of mature CD4 T cells in the presence of CD4. Interestingly, in certain patients with human immunodeficiency virus (HIV) infection, known for eradicating CD4+ T cells, MHC class II restricted CD8+ T cells have been observed (20). It is known that

the T cell repertoires consisting of CD4 and CD8 T cells are generated via thymic selection in newborns, and that the thymus ceases to function with age (21). We can speculate that plasticity of TCRs in their ability to bind both class I and class II molecules might be functional to the fitness of the adaptive immune repertoire: any type of distortion of the immune system occurring after the thymus has concluded its function cannot be compensated by the production of new T cells, but instead needs to be dealt with by the T cells already produced. For example, in the case of viral infections depleting portions or subtypes of certain T cells, conceivably as theorized by Gunzman and Chen (22), an intrinsic plasticity would allow for the system to employ different strategies to protect the balance and integrity of the adaptive immune system. The potential requirement for plasticity and flexibility in the adaptive immune system could be a possible explanation for the almost indistinguishable difference between CD4+ T cell and CD8+ T cell receptors observed in our explorations.

As mentioned in the results section, curiously, the same TCR sequences were observed as both a CD4+ T cell and CD8+ T cell. This was not only the case for the SCS, but an instance of the same TCR sequence being detected as both a CD4+ T cell and CD8+ T cell, was also present in the VDJdb dataset. If the 3D structure is the same between the CD4+ and CD8+ T cells, this would strengthen the idea that other factors beyond the T cell receptor are what drives lineage choice. All this viewed together indicates concurrence with the idea that interaction determines lineage choice. However, instead of being determined by a TCR which is only capable of interacting with an MHC class I or class II, these lines may be blurred due to potential plasticity in TCR interaction together with other not yet well understood factors that can influence lineage choice, and also promotes the idea that TCRs may have the potential to exhibit cross-reactivity across MHC classes.

The similar logo plots between CD4+ and CD8+ T cells and the small prediction signal indicate that sequence alone might not be enough to predict whether a cell will differentiate into either a CD4+ or CD8+ T cell. However, since structure plays a big role in how molecules interact with each other, it could be feasible that prediction capabilities may be improved upon including structural information as well in the model. Therefore it could be interesting as a future perspective to test whether introducing structure could improve prediction capabilities of lineage choice. In an article by Yin *et al.* (23) they showed that a CD8+ T cell underwent conformational changes depending on whether it was bound to a MHC class I or class II complex. Furthermore, as mentioned earlier, some of the paired TCR sequences were present in the dataset as both a CD4+ T cell or CD8+ T cell. It could be of potential interest to study whether a clear difference is observed between the 3D structure of the same T cell receptors observed on both CD4+ and CD8+ T cells.

Models have been proposed to explain the lineage choice of the bipotential double positive (DP) thymocytes expressing both the CD4 and CD8 co-receptor into either a CD4+ or CD8+ T cell. Two models have initially been proposed to elucidate lineage selection, namely the "instructive" (24) and the "stochastic" (25) model. The "instructive" model is based on the idea that there is a co-engagement of CD4 and CD8 with the TCR, which via distinct intracellular signals directs the development of an immature DP CD4+CD8+ thymocyte into either the CD4 or CD8 lineage (24). In the data used, we found TCRs that had the same sequence but different co-receptor labels. If the 3D structure of the T cell receptors with the same sequences are the same for both CD4+ and CD8+ T cells, then our results will to a certain extent, go against the instructive model. If the lineage choice is dictated by the TCR, we would not expect to observe a significant amount of sequences both labeled as CD4+ and CD8+, thus this observation goes against the instructional model.

Conversely, the "stochastic" model hypothesizes that the expression of either the CD4 or CD8 co-receptor occurs at random. The stochastic model also postulates that after positive selection a second TCR-dependent

"rescue" occurs. Here single positive (SP) thymocytes expressing only the CD4 or CD8 co-receptor, which have a TCR matching the expressed co-receptor, differentiate into mature T cells (25). In our study, we don't see a strong signal when predicting on the internal SCS dataset; however, a signal is still present, indicating that the selection is not random and thus, to a certain degree, goes against the theory of random selection.

As mentioned previously, SCS was performed to obtain the paired TCR sequences. T cells were singularly encapsulated in droplets, which each had a unique droplet barcode. Typing of the T cells was carried out by reading CD4 and CD8 amounts for a given droplet barcode. If the number was more than 0 for one type only, the type was considered true and kept. For barcodes that had no type assigned to them, due to the number being 0 and barcodes having more than 0 for each co-receptor type were discarded. Nevertheless, although only one co-receptor type could be present per barcode for it to be counted true, it should not be ignored that some of these cells may be cells transitioning from one state to another. In a newer model named the kinetic model (26), they propose that lineage choice occurs in sequential steps dictated by TCR signal duration, where the CD8 co-receptor gets downregulated to "audition" for the CD4 lineage before differentiating into a lineage. On the other hand, it has also been stated that residual amounts of CD8 surface protein can be found expressed on cells in this intermediate state (26, 27), thus if these residual amounts are being detected, then these cells would be discarded since both types of co-receptors would be present on the cell.

Predicting what a T cell recognizes and when and with what the T cell will be cross-reactive with is to this day still a very complicated task. It is imaginable that the adaptability in what determines the lineage choice may extend into T cell activation and what a T cell is capable of recognizing.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

Anna-Lisa Schaap-Johansen and Paolo Marcatili conceived and wrote the paper. Anna-Lisa Schaap-Johansen created the figures with additional help from Kamilla Kjærgaard Munk. The twin dataset was processed by Tina Funck. The manuscript was reviewed and corrected by Kamilla Kjærgaard Munk, Martin Closter Jesnen and Vanessa Isabell Jurtz.

FUNDING

Anna-Lisa Schaap-johansen is founded by the 2018 SDC grants.

REFERENCES

1 .Sewell AK. Why must T cells be cross-reactive? *Nature Reviews Immunology 2012 12:9* **12** (2012) 669–677. doi:10.1038/nri3279.

2 .Wooldridge L, Ekeruche-Makinde J, Van Den Berg HA, Skowera A, Miles JJ, Tan MP, et al. A single autoimmune T cell receptor recognizes more than a million different peptides. *The Journal of biological chemistry* **287** (2012) 1168–1177. doi:10.1074/JBC.M111.289488.

3 .Singer A, Adoro S, Park JH. Lineage fate and intense debate: myths, models and mechanisms of CD4- versus CD8-lineage choice. *Nature reviews. Immunology* **8** (2008) 788–801. doi:10.1038/NRI2416.

4 .Li HM, Hiroi T, Zhang Y, Shi A, Chen G, De S, et al. TCR repertoire of CD4+ and CD8+ T cells is distinct in richness, distribution, and CDR3 amino acid composition. *Journal of Leukocyte Biology* **99** (2016) 505–513. doi:10.1189/JLB.6A0215-071RR/-/DC1.

5 .Carter JA, Preall JB, Grigaityte K, Goldfless SJ, Jeffery E, Briggs AW, et al. Single T Cell Sequencing Demonstrates the Functional Role of $\alpha\beta$ TCR Pairing in Cell Lineage and Antigen Specificity. *Frontiers in Immunology* **10** (2019). doi:10.3389/FIMMU.2019.01516/FULL.

6 .Atchley WR, Zhao J, Fernandes AD, Drüke T. Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences of the United States of America* **102** (2005) 6395. doi:10.1073/PNAS.0408677102.

7 .Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, Crawford JC, Dolton G, et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Research* **46** (2018) D419–D427. doi:10.1093/NAR/GKX760.

8 .Bagaev DV, Vroomans RM, Samir J, Stervbo U, Rius C, Dolton G, et al. VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Research* **48** (2020) D1057–D1062. doi:10.1093/NAR/GKZ874.

9 .Tickotsky N, Sagiv T, Prilusky J, Shifrut E, Friedman N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics (Oxford, England)* **33** (2017) 2924–2929. doi:10.1093/BIOINFORMATICS/BTX286.

10 .Rubelt F, Bolen CR, McGuire HM, Heiden JA, Gadala-Maria D, Levin M, et al. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nature communications* **7** (2016). doi:10.1038/NCOMMS11112.

11 .Klausen MS, Anderson MV, Jespersen MC, Nielsen M, Marcatili P. LYRA, a webserver for lymphocyte receptor structural modeling. *Nucleic acids research* **43** (2015) W349–W355. doi:10.1093/NAR/GKV535.

12 .Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, et al. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic acids research* **34** (2006). doi:10.1093/NAR/GKJ088.

13 .Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics* **16** (2000) 412–424. doi:10.1093/BIOINFORMATICS/16.5.412.

14 .Hauser M, Steinegger M, Söding J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics* **32** (2016) 1323–1330. doi:10.1093/BIOINFORMATICS/BTW006.

15 .Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* **89** (1992) 10915. doi:10.1073/PNAS.89.22.10915.

16 .Tareen A, Kinney JB. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36** (2020) 2272–2274. doi:10.1093/BIOINFORMATICS/BTZ921.

17 .Vacic V, Iakoucheva LM, Radivojac P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics (Oxford, England)* **22** (2006) 1536–1537. doi:10.1093/BIOINFORMATICS/BTL151.

18 .Matechak EO, Killeen N, Hedrick SM, Fowlkes BJ. MHC class II-specific T cells can develop in the CD8 lineage when CD4 is absent. *Immunity* **4** (1996) 337–347. doi:10.1016/S1074-7613(00)80247-2.

19 .Kirberg J, Baron A, Jakob S, Rolink A, Karjalainen K, Von Boehmer H. Thymic selection of CD8+ single positive cells with a class II major histocompatibility complex-restricted receptor. *The Journal*

of experimental medicine **180** (1994) 25–34. doi:10.1084/JEM.180.1.25.

20 .Ranasinghe S, Lamothe PA, Soghoian DZ, Kazer SW, Cole MB, Shalek AK, et al. Antiviral CD8 + T Cells Restricted by Human Leukocyte Antigen Class II Exist during Natural HIV Infection and Exhibit Clonal Expansion. *Immunity* **45** (2016) 917–930. doi:10.1016/J.IMMUNI.2016.09.015.

21 .Goronzy JJ, Lee WW, Weyand CM. Aging and T-cell diversity. *Experimental gerontology* **42** (2007) 400–406. doi:10.1016/J.EXGER.2006.11.016.

22 .Guzman MP, Chen Z. Conversion of the CD8 lineage to CD4 T cells. *Oncotarget* **6** (2015) 20748–20749. doi:10.18632/ONCOTARGET.5235.

23 .Yin L, Huseby E, Scott-Browne J, Rubtsova K, Pinilla C, Crawford F, et al. A single T cell receptor bound to major histocompatibility complex class I and class II glycoproteins reveals switchable TCR conformers. *Immunity* **35** (2011) 23–33. doi:10.1016/J.IMMUNI.2011.04.017.

24 .Robey EA, Fowlkes BJ, Gordon JW, Kioussis D, von Boehmer H, Ramsdell F, et al. Thymic selection in CD8 transgenic mice supports an instructive model for commitment to a CD4 or CD8 lineage. *Cell* **64** (1991) 99–107. doi:10.1016/0092-8674(91)90212-H.

25 .Davis CB, Killeen N, Crooks ME, Raulet D, Littman DR. Evidence for a stochastic mechanism in the differentiation of mature subsets of T lymphocytes. *Cell* **73** (1993) 237–247. doi:10.1016/0092-8674(93)90226-G.

26 .Brugnera E, Bhandoola A, Cibotti R, Yu Q, Guinter TI, Yamashita Y, et al. Coreceptor reversal in the thymus: signaled CD4+8+ thymocytes initially terminate CD8 transcription even when differentiating into CD8+ T cells. *Immunity* **13** (2000) 59–71. doi:10.1016/S1074-7613(00)00008-X.

27 .Singer A. New perspectives on a developmental dilemma: the kinetic signaling model and the importance of signal duration for the CD4/CD8 lineage decision. *Current Opinion in Immunology* **14** (2002) 207–215. doi:10.1016/S0952-7915(02)00323-0.

8 Global energy terms for improved TCR-pMHC binding prediction

In this chapter we present a work in progress. Many T cell based immunotherapies focus on CD8+ T cell interaction to elicit an immune response to help fight diseases in patients. However, we still do not fully understand what will provoke a CD8+ T cell response.

The main goal of this project was to investigate whether global energy terms calculated on modeled TCR-pMHC complexes have any prediction power and if they can be used to add additional information to a model predicting T cell recognition of MHC presented peptides.

We show that global energy terms by themselves do carry some predictive power when used in isolation - however, they have limited impact when used in conjunction with other sequence-derived features. We also show that it is a difficult task to predict T cell recognition of peptide-MHC complexes across peptides, since there is no clear sequence pattern differentiating binders from non-binders.

Global energy terms for improved TCRpMHC binding prediction

Anna-Lisa Schaap-Johansen[1] **and Paolo Marcatili**[1]

[1]**Department of Health Technology, Technical University of Denmark, Lyngby, Denmark**

ABSTRACT

Not all peptides presented by the major histocompatibility complex (MHC) will elicit an immune response from T cells. However, predicting which MHC presented peptides a T cell will recognize remains challenging. The majority of methods utilize the T cell receptor sequence to predict their interaction, but this may not contain substantial information to make a clear distinction between complexes. From the limited data available, it is evident that when looking at the sequences alone, it is difficult to find overall generalizing patterns that can be used across different peptides, as will also be shown in this study. Therefore, in this study, we investigate whether structural energy terms calculated for the overall interacting complex carry information that potentially can add additional predicting power.

Keywords: TCR, MHC, peptide, epitope, neoepitope, TCRpMC, TCR-pMC

INTRODUCTION

T-cells are a part of the adaptive immune response and play a vital role in recognizing infected cells or abnormal cells arising such as can occur with cancer [1]. It is known that T-cells utilize their T-cell receptors (TCRs) to survey whether non-self peptides, such as epitopes or in the case of cancer-specific peptides called neoepitopes, are presented by the major histocompatibility complex (MHC) on the surface of a cell. If T-cell recognition of a peptide-MHC (pMHC) complex occurs, it can induce an immune response, thereby helping the body defend against potentially foreign invaders and unhealthy cells. A deeper understanding of what promotes an immune response from a T-cell could help further the development of different immunotherapies such as T-cell therapy and T-cell vaccines [2].

TCRs are hetro-dimeric proteins, consisting of two-membrane bound chains, where these can either be α and β chains or γ and δ chains. The majority of T-cells express TCRs comprising α and β chains [3], and this group of TCRs can further be divided into whether they recognize peptides bound to MHC class I (MHC I) or MHC class II (MHC II). T-cells interacting with peptides bound to MHC I are called cytotoxic T-cells, and are known to directly kill infected cells. On the other hand, T-cells interacting with peptides bound to MHC II are known as T-helper cells, which activates other immune cells to act against the compromised cells, which is done either directly or indirectly. TCRs mainly interact with pMHCs through six loops situated in the TCR α and β chain. These loops are generally known as complementarity determining regions (CDRs) and individually denoted as CDR1, CDR2 and CDR3.

The majority of studies published that predict the interaction between TCRs and pMHCs (TCR-pMHC) have mainly focused on utilizing the amino acid sequences of the complexes. There are currently only a scarce number of models which introduce structure into their models when predicting TCR peptide interaction [4]. Complexes, where T cells have been measured to bind to the MHC, presented peptides may be more stable than non-binders. Therefore, it is possible that calculating structural energy may provide additional information, which can help differentiate binders from non-binders.

In this paper, we, therefore, introduce structural information by calculating the overall energy from modeled TCR-pMHC structures to investigate their predictive power in a machine learning setup. We also study whether the addition of global energy terms to an already existing model available for predicting T cell peptide recognition, namely NetTCR2, can improve prediction performance, as well as what can make this a challenging task to predict.

MATERIALS AND METHODS

The dataset

Paired CDR3 and sequences are obtained from the paper by Montemurro et al. [5]. The dataset consists of positive and negative data in the sense that positive binders entail TCR binding to the pMHC complex and vice versa. The TCRs in the dataset are restricted to TCRs that bind to HLA-A*02:01-specific peptides of length 9, where both CDR3 α and CDR3 β are available. Details surrounding the original setup of this dataset can be found in the paper from where the dataset was obtained. The dataset consists of a total of 1783 paired sequences.

Generation of swapped negatives

To avoid having some TCRs present only as positives (meaning binders), thus creating possible biases in the training of our network, additional "swapped" negatives (meaning non-binders) are generated. This is done by adding one swapped negative for each positive by matching the given TCR with a random peptide extracted from the same partition. The swapped combined with the original dataset results in 12,975 entries in total.

TCR sequence reconstruction

The original dataset contains the CDR3 region of both the TCR α and TCR β chain. The dataset only contains the V and J genes for non-binding TCR sequences. We retrieve the V and J genes for the binding TCRs by mapping the sequences to the VDJdb database. TCR sequences are constructed using in-house scripts [6], which takes a CDR3 sequence and its belonging V and J gene as input. After processing the dataset, a total of 11,708 entries remain. The sequence for the MHC molecule HLA-A*02:01:01 was retrieved from the IPD-IMGT/HLA database [7].

Molecular modeling

A Fasta file was created for each of the entries. The fasta file contained the TCRα, TCRβ, peptide, and HLA-A*02:01:01 sequences. For each fasta file, the in-house pipeline TCRpMHCmodels was used to model the complexes. After modeling the fasta files a total of 10,341 complexes are constructed.

Energy calculations

FoldX [8, 9] and Rosetta [10, 11] are used to calculate global energy terms. After calculating energy terms, a total of 9,991 complexes remained.

FoldX Each modeled complex is relaxed in FoldX5.0 using the `RepairPDB` command with the following flags; `ionStrength=0.05`, `pH=7`, `water=CRYSTAL`,`vdwDesign=2`, `out-pdb=1`, `pdbHydrogen=false`. Energy terms are calculated using the AnalyseComplex command. For each of the complexes, we compute the following six interaction energy terms: MHC-peptide, MHC-TCRα, MHC-TCRβ, peptide-TCRα, peptide-TCRβ, TCRα-TCRβ.

Rosetta Models were relaxed in the Rosetta force field energy function 2015, with the following command, `relax.default.linuxgccrelease` with default options. The global energy terms were calculated using the `score_jd2.linuxgccrelease` command.

Logo plots:

Logo plots were created using the tool Seq2Logo [12]. The tool was used with default options. Seq2Logo requires the sequences to be of the same length; therefore, the reconstructed sequences were used, which introduces gaps into the sequences, so they are all of the same lengths. Logo plots were created for binders, non-binders, and non-binders with swapped added to them.

Two sample logo plots:

Two sample logo plots were produced for the CDR3 section of the α and β sequences after being generated using in-house scripts. The plots were constructed using the software from [13], which performs a t-test. The two sample logo plot was created with default options, except for correcting the p-value with the Bonferroni correction. The software requires a "positive" and "negative" input. The binders were set as the "positive" and the non-binders as the "negative" input.

Random forest - Energy impact

A random forest using the scikit-learn library (ver. 0.23.2) was implemented to investigate whether energy has any prediction power. The global energy terms, an array of 75 in length, were used as input.

Baseline model

The original Nettcr2 model was used as a baseline. The data was encoded using the BLOSUM50 matrix [14], meaning that each amino acid residue is presented as a vector of length 20 corresponding to the amino acid row of the BLOSUM50 matrix. All peptides were of length 9, CDR3 sequences were of different length and were therefore zero-padded to a maximum length of 30. This model consists of multiple 1-dimensional CNNs created using pytorch (ver. Anaconda 4.4.0). The peptide and CDR3 were processed separately, each with five differently sized kernels (1,3,5,7,9), and a filter sze of 16, outputting 80 filters in total per input sequence. Kernel weights were initialized with the Glorot normal initializer. The convolutional outputs for the peptide, CDR3a and CDR3b, were max-pooled and concatenated, resulting in a single vector of length 240, which was then fed into a dense layer with 32 hidden neurons. Finally a second dense layer transforms the output from the previous layer to an output of one with a sigmoid activation, to give the probability of peptide-CDR3 pair binding.

Nettcr2 - Energy

The baseline model was used with the addition of global energy terms as an extra variable to the dense layer. The idea is to add extra information to the network at a later point to help guide the network. The global energy terms were run through a batchnorm to normalize the inputs and thereafter concatenated with the convoluted peptide, CDR3a, and CDR3b outputs. The concatenated values were then put through two dense layers as in the original setup.

Nettcr2 - LSTM

The baseline model was used with the addition of an LSTM followed by a dense layer with global energy terms as an extra variable. The convoluted outputs were transposed and then concatenated to be used as input for the LSTM. Before inputting to the LSTM a dropout is used (p=0.1). The LSTM consisted of one layer with 26 hidden units. Outputs from the LSTM were flattened and concatenated together with the global energy terms. The concatenated values were as in the original setup put through two dense layers.

Model training

Models were trained for 300 epochs with early stopping, implementing a patience of 50 epochs using a nested 5-fold cross-validation scheme. The Adam optimizer was used to update the weights, and a learning rate of 0.001 and batch size of 128 was used. Finally binary cross-entropy was used as the loss function.

Performance evaluation

Models were trained for 300 epochs with early stopping, implementing a patience of 50 epochs using a nested 5-fold cross-validation scheme. The Adam optimizer was used to update the weights, and a learning rate of 0.001 and batch size of 128 was used. Finally, binary cross-entropy was used as the loss function.

RESULTS

The main motivation behind this study and the development of our tool was to test whether sequence-only based methods capture enough information in regards to distinguishing which TCR-pMHC complexes will induce an immune response. The global energy terms can provide information regarding how stable a complex is. It is believed that the more stable a complex is, the more likely it is that there is a binding interaction occurring.

Dataset analysis

The dataset used in this study consists of 9991 TCR-pMHC complexes, of which 8265 are non-binders, and 1726 are binders, meaning epitopes inducing an immune response.

An analysis of the dataset shows that it consists of 18 different peptide antigens, which are all 9-mers. There is a bias in regards to the distribution of the different antigens, as can be seen in figure 1. The most

frequent antigen is the antigen from the influenza virus “GILGFVFTL”, which constitutes around 60.3% of the samples. The second and third most frequent antigens are ”GLCTLVAML” with 16.1% of the entries and ”NLVPMVATV”, constituting 11.8% of the antigens, both from the Herpes virus. Furthermore, it can also be gathered from figure 1, that two of the peptides are only present in a non-binding format, and thus these two peptides will only be used for training and not further downstream analysis.

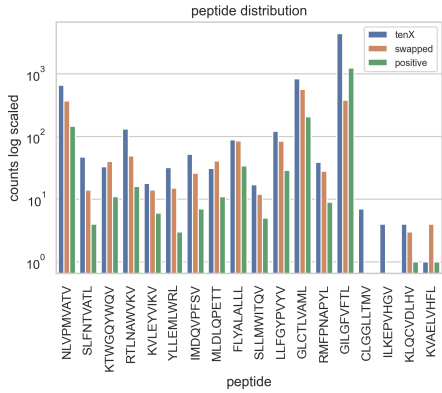


Figure 1. Distribution of the different peptides present in the dataset. The counts are log scaled for easier visualization. The plots show the number of binders, denoted as positive, non-binders, denoted as tenX, and swapped denoted as swapped.

To study whether there is any observable difference between TCRs that have been measured to bind and TCRs that have not been measured to bind, we create a logo plot. As shown in figure 2a and b, there are minor observable distinctions between CDR3β sequences from binders and CDR3β sequences from non-binders. As mentioned previously, in order to avoid having TCRs that are only in the datasets as binders, we created “swapped” non-binders. Figure 2c, shows what the non-binder CDR3 logo plot looks like after adding the “swapped” to the non-binders.

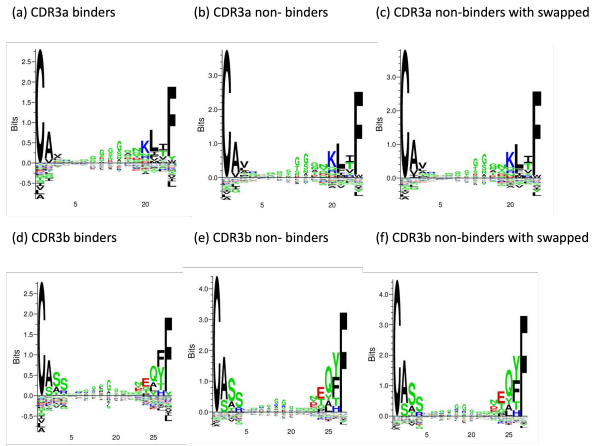


Figure 2. Logo plots created for the CDR3 sequences for binders, non-binders and swapped. Figures a,b and c depict logo plots for CDR3 α sequences. Figures d,e and f show logo plots for CDR3 β sequences.

There may be some of the distinctions which are statistically observable, therefore we create two sample logo plots to study this. As shown in figure 3, there are statistical observable differences between binders and non-binders. Glycine (G) can be found enriched in binders compared to non-binders at multiple positions, especially at position 19, in the CDR3α chain. Glutamine (Q) and asparagine (N) can be seen enriched at positions 18 and 20 in binders. Valine (V) and lysine (K) at position 20 can be seen more commonly in non-binders in the α chain, as can be seen in figure 3a. In the CDR3 β chain, figure 3b, serine (S) is seen enriched at multiple positions, particularly at position 21, in binders. Arginine (R) at position 6 are seen as enriched in binders compared to non-binders. However, as can also be gathered from figure 3, this statistical difference is only present in a subset of the data. The maximum enrichment is in less than 40% of the sequences. Thereby no particular position can discriminate between binders and non-binders.

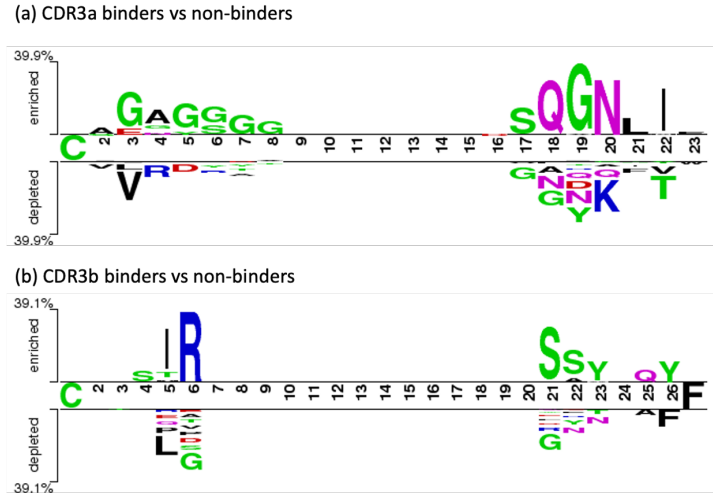


Figure 3. Two sample logo plots created for the CDR3 sequences for binders and non-binders. Enriched here denotes which amino acids are more prevalent in binders compared to non-binders and vice versa. a) shows the two sample logo plot created for the CDR3 α sequences between binders and non binders. b) shows the two sample logo plot created for the CDR3 β sequences between binders and non binders.

We create a two-sample logo plot for binders and non-binders for the three most common peptides to investigate whether there are any statistical similarities between them. Upon observing the two sample logo plots created from CDR3β sequences for the three most common peptides, shown in figure 4, position 21 for peptides GILGFVFTL and GLCTLVAML both have a depletion of leucine (L) to a certain extent in the CDR3β sequences binding these peptides, this is not observed for the NLVPMVATV peptide. For both the NLVPMVATV and GLCTLVAML, there is a depletion of glutamine (Q) in the CDR3β sequences binding these peptides. However, the opposite holds true for the GILGFVFTL peptide, where glutamine is enriched in the CDR3β sequences binding this peptide. Furthermore, for position 2, alanine (A) and serine (S) at position 3 and 4 are enriched for GILGFVFTL but depleted for GLCTLVAM in the CDR3β sequences binding to these peptides. We also see that Asparagine (N) at position 23 is depleted in sequences that bind to GILGFVFTL but enriched in sequences binding to GLCTLVAML. We do not observe one or more amino acids that are being enriched or depleted consistently across all three peptides. This indicates that there is not a clear distinction between what binds a peptide and what does not, across peptides when doing a simple two sample logo plot. As can be gathered from the figure, the enrichments and depletions are only statistically different in a subset of the data, meaning that amino acids showcased as being depleted in CDR3β sequences that bind are also present in some of the CDR3β sequences registered to bind to the specific peptide.

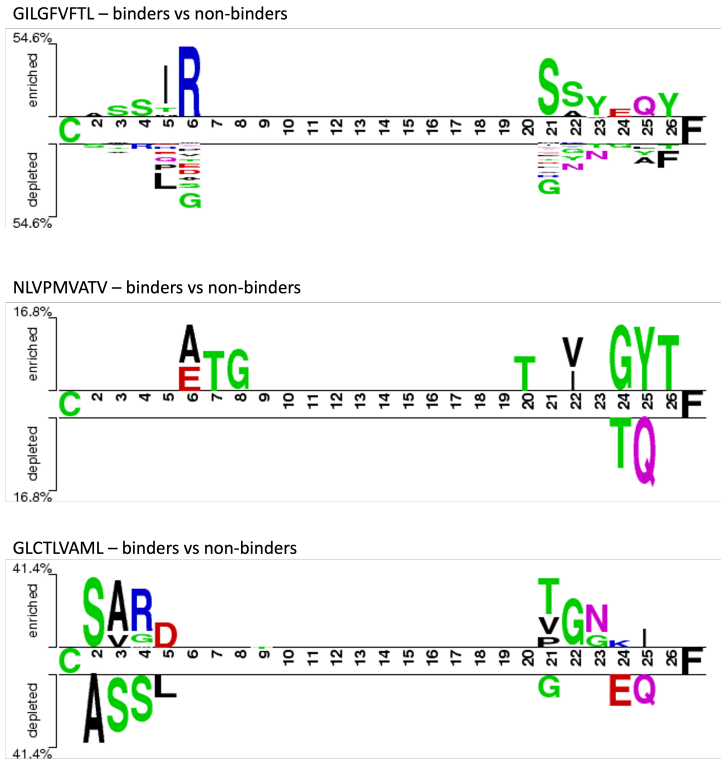


Figure 4. Two sample logo plots created from the CDR3β sequences for binders and non-binders for the three most prevalent peptides in the dataset. Enriched here denotes which amino acids are more prevalent in binders compared to non-binders and vice versa. The peptides are shown in the following order from top to bottom; GILGFVFTL, NLVPMVATV, GLCTLVAML. The plots are made with only the sequences measured to bind or not bind a given peptide.

Machine Learning analysis of global energy term impact

The representation of the data can have an impact on prediction power; therefore, to study whether the global energy terms have any prediction power, we perform a random forest. From the random forest, it was evident that global energy terms have predictive power since an MCC of 0.384 and AUC of 0.791 were obtained with this model.

To investigate whether global energy terms would improve the prediction power for a tool available for T cell peptide recognition prediction, a recently published new rendition of NetTCR from which the data was gathered was utilized. A more complex model may further improve the prediction ability; therefore, we test this by introducing a LSTM layer in the architecture. A benchmark is carried out consisting of the original setup of NetTCR, NetTCR with global energies added to the dense layer, and lastly, NetTCR with an LSTM added and with global energies added to the dense layer. The model was trained with a nested 5-fold cross-validation scheme due to the small dataset and to test the robustness of the models. As illustrated in figure 5, the global energy terms slightly improved the model's predictive power. The original NetTCR2 obtains an AUC of 0.872 and MCC of 0.668, where when energy is added to the dense layer AUC becomes 0.882, and MCC rises to 0.695. The more complex model where LSTM is added to the architecture improves slightly better than the original model, with an AUC of 0.878 and MCC of 0.705. The slightly more complicated model performs slightly worse than the NetTCR2 with global

energy terms added to the dense layer when looking at the AUC and slightly better when looking at the MCC. However, overall no significant difference is observed.

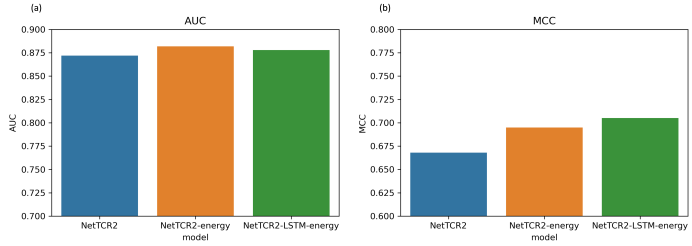


Figure 5. Model performance was calculated using AUC and MCC. The figure shows the AUC and MCC for the three models trained; NetTCR2, NetTCR2 with global eneergy terms added to the dense layer and NetTCR2 with LSTM and global energy terms added to the dense layer. a) shows the AUC values obtained and b) shows the MCC values.

DISCUSSION

In this study we show that global energy terms calculated on three-dimensional models of TCR-pMHC complexes have a predictive power in indicating the ability of the complex to interact and, in principle, to start an immune response. We also show that, to a very minor extent, the energy can improve existing models that are only based on the molecules' sequences. This indicates that including the global energy terms in a model could be of potential interest for future predictive models. In this setup we used both Rosetta and FoldX since they provide different information due to their individual ways of being calculated. Rosetta is based on mathematical and physical assumptions to calculate the energy whereas FoldX is an empirically derived model based on observed energy changes from mutations. The Rosetta energy function is a model which utilizes physical and mathematical assumptions parameterized from small molecule and X-ray crystal structure data. The Rosetta energy function calculates the potential energy by approximating the energy of a biomolecule conformation. This is done by scaling different summed individual energy terms with a weight.

One of the issues with the existing models is connected to the relatively low variability of antigens and MHCs. In this study, all the antigens were presented by HLA*02:01, which is due to the limited availability of data for paired sequences where the HLA is known. This of course limits the general use of these models, but can still provide an indication of their potential use. Because of this, it is worth investigating if structure-based predictions are able to generalize more easily to new MHC molecules, thus increasing the applicability of such tools to Furthermore, not only is there a limited availability of data, but the setup in this study also resulted in a decrease of the data possible to train on, due to the generation of the reconstructed sequences, molecular modeling as well as energy calculations.

A major disadvantage regarding energy calculations is that they are very dependent on the modeling. This means that the same sequence can obtain different energies depending on the modeling of the structure, which is very dependent on the tool used, making the energies less trustworthy. Furthermore, the structures were relaxed before calculating energy, this may potentially make structures from TCR sequences that bind and do not bind more similar, thereby decreasing the information the energy can provide in regards to prediction power.

The addition of the LSTM to the architecture was done to capture short range dependencies from the CNN and long range dependencies from the LSTM and combine this to strengthen the model. A sequence can have contextual information situated close to the position of interest, this should be captured by the CNN, such as the combination of certain amino acids near each other may be more typical for sequences that bind compared to sequences that do not bind. However, long range information may also be present since amino acids not in the nearby vicinity may provide information as well. It was also

possible that the concatenation of the multiple convoluted outputs for each sequence is more complex that what a dense layer is able to interpret well, which drove our decision to test out the addition of LSTMs in the architecture as well. Although the addition did not improve the AUC compared to only using global energies added to the dense layer with no LSTM, the MCC still increased slightly. The dataset is highly imbalanced with many more non-binders than binders, therefore this minor increase in MCC could still indicate that the LSTMs may provide some additional benefits to the model.

As can be gathered from both the logo plots and the two sample logo plots, the sequences are not distinctive from each other in a clear cut manner. When looking at the two sample logo plot for the three most common peptides, as mentioned in the results section, we observe that some amino acids are enriched in sequences binding to one peptide, but the same amino acid at the same position is depleted in amino acids binding to a different peptide. In the case of the most prevalent peptide in the dataset, namely GILGFVTL, there is an enrichment of glutamine at position 25 in CDR3β sequences binding to this peptide. However, the opposite is seen for CDR3β sequences binding to the NLVPMVATV and GLCTLVAML, where glutamine is depleted at this position. All this taken together shows that this is a very difficult task and that finding generalizable features across peptides is not straightforward to do. It should also be mentioned that since sequences binding and not binding to GILGFVTL are present to a higher degree in the dataset; the model will be biased towards this peptide, making it more difficult to discover generalizable features across peptides.

The observations in the logo and two sample logo plots could potentially be due to the limited data availability or how our experiments are conducted. However, it may also be that the inconsistencies and unclear distinctive patterns may just paint the true picture of how TCRs truly behave. TCR-pMHC complexes are of great interest in general and are being sequenced and added to databases. Nevertheless, if the logo plots and two sample logo plots depict the behavior of TCRs in a true manner, then this would indicate that we need more informative inputs than just the sequence by itself. Interestingly, the global energy terms did show some predictive power when used by themselves; therefore, it could be interesting to investigate whether the addition of the global energy terms can help the model generalize on new peptides. This would require a different training setup, such as training a model on the complexes that do not contain a specific peptide and using the excluded complexes to test the model's ability to generalize to new peptides. However, the limited availability and biased data would still cause complications.

In this study we only tested global energy terms, to investigate whether energy terms can have an impact, but also to to avoid having too many parameters the model can overfit on. It is possible to calculate not only global energy terms but also per residue terms when using Rosetta. This can provide more detailed information regarding individual amino acid interactions, instead of just an overall average of how the complex is interacting. It could therefore be interesting to create a model implementing per residue terms as well and study whether this will improve the performance.

Recently the new edition of AlphaFold [15] has shown great improvement in molecular modeling. It could be interesting to model the sequences with AlphaFold and calculate the energy terms on the AlphaFold generated structures instead since these structures may model the complex structure better. Nonetheless, it should be kept in mind that AlphaFold is still not optimal at predicting the structure of complexes. AlphaFold has also shown limitations regarding predicting very variable loops. However, TCR-pMHC is a complex, and TCRs contain very variable CDR loops, which have a big impact on the interaction between TCRs and pMHCs. It would, therefore, still be a difficult task for AlphaFold to model. Nevertheless, it would still be of interest to test whether predictions can be improved by using a different way of modeling the complexes.

As previously mentioned, this is not an easy task to predict. However, we still show in this study that energy calculations may be able to provide additional information that can improve the predictive capabilities of a model in distinguishing binders from non-binders.

REFERENCES

[1] Alex D. Waldman, Jill M. Fritz, and Michael J. Lenardo. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nature reviews. Immunology*, 20(11):651–668, 11 2020.

[2] Harpreet Singh-Jasuja, Niels P.N. Emmerich, and Hans Georg Rammensee. The Tübingen approach: identification, selection, and validation of tumor-associated HLA peptides for cancer therapy. *Cancer immunology, immunotherapy : CII*, 53(3):187–195, 3 2004.

[3] Simon R. Carding and Paul J. Egan. $\gamma\delta$ T cells: functional plasticity and heterogeneity. *Nature Reviews Immunology 2002 2:5*, 2(5):336–345, 2002.

[4] Anna-Lisa Schaap-Johansen, Milena Vujović, Annie Borch, Sine Reker Hadrup, and Paolo Marcatili. T Cell Epitope Prediction and Its Application to Immunotherapy. *Frontiers in Immunology*, 12:2994, 9 2021.

[5] Alessandro Montemurro, Viktoria Schuster, Helle Rus Povlsen, Amalie Kai Bentzen, Vanessa Jurtz, William D. Chronister, Austin Crinklaw, Sine R. Hadrup, Ole Winther, Bjoern Peters, Leon Eyrich Jessen, and Morten Nielsen. NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. *Communications Biology 2021 4:1*, 4(1):1–13, 9 2021.

[6] Michael Schantz Klausen, Mads Valdemar Anderson, Martin Closter Jespersen, Morten Nielsen, and Paolo Marcatili. LYRA, a webserver for lymphocyte receptor structural modeling. *Nucleic acids research*, 43(W1):W349–W355, 2015.

[7] Véronique Giudicelli, Patrice Duroux, Chantal Ginestoux, Géraldine Folch, Joumana Jabado-Michaloud, Denys Chaume, and Marie Paule Lefranc. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic acids research*, 34(Database issue), 2006.

[8] Raphael Guerois, Jens Erik Nielsen, and Luis Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology*, 320(2):369–387, 2002.

[9] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The FoldX web server: an online force field. *Nucleic acids research*, 33(Web Server issue), 7 2005.

[10] Sitao Wu and Yang Zhang. Protein structure prediction. *Bioinformatics: Tools and Applications*, pages 225–242, 2007.

[11] Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of chemical theory and computation*, 13(6):3031–3048, 6 2017.

[12] Martin Christen Frolund Thomsen and Morten Nielsen. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic acids research*, 40(Web Server issue), 7 2012.

[13] Vladimir Vacic, Lilia M. Iakoucheva, and Predrag Radivojac. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics (Oxford, England)*, 22(12):1536–1537, 6 2006.

[14] Steven Henikoff and Jorja G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915, 1992.

[15] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A.A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873*, 596(7873):583–589, 7 2021.

9 Epilogue

This thesis presents projects exploring T cells, their receptors, and how they can be utilized in T cell based immunotherapy. The main focus of this thesis was to study T cell receptors to further our understanding of what may define a given T cell lineage, how T cell receptors interact, and what they recognize, all of which are essential parameters for their potential use in immunotherapy.

The first project included in this thesis is a review that introduces the reader to the adaptive immune system, T cell based immunotherapy, and presents the current tools available for predicting epitopes and neoepitopes for T cell based immunotherapy. This review shows that T cell epitope and neoepitope prediction have received a lot of interest over the years due to their potential, and as a result, a lot of computational tools have been created for that purpose. The majority of the current tools focus on predicting whether a peptide will be presented by an MHC. Although this is an important part of MHC presented peptide recognition, it is only half of the equation. Some tools have been constructed to predict TCRpMHC interaction based on the CDR3 beta sequence, and some newer tools use the paired CDR3 alpha and CDR3 beta sequences. However, only a handful of tools use structural information in their models. This review aims to guide the reader on the tools currently available, what type of input data the individual tools require and what kind of output information the tools are able to produce. In this review, we also present and discuss the strengths and weaknesses of the individual tools as well as potential future perspectives that can be of interest within the field.

The second project presented in this thesis investigates if T cell lineage can be determined based on the T cell receptor sequence alone. It has generally been believed that the role of CD4+ T cells in anti-tumor response is somewhat limited, but newer studies indicate that CD4+ T cells have a more significant impact on anti-tumor immune response than previously thought. This, combined with the fact that it is still unknown if there are any clear differences in their TCRs and potentially based on what the different cells will interact with even though both play a vital role in the adaptive immune system, is the underlying motivation for the second project. In this project, we question how

static a T cell is in its lineage choice and pose the question of whether T cells may have the potential to be cross-reactive across MHC classes. The recent increase of data from single cell sequencing made it possible to study a large number of paired TCRs from both the CD8+ and CD4+ T cell lineage. We discovered with this data that, although there was a signal in the data that could to a certain extent distinguish CD8+ TCR sequences from CD4+ TCR sequences, it was not a clear and strong signal. We also observed that some paired TCR sequences with the exact same V and J genes existed in the data with either a CD8 or CD4 as their label, which led to the idea that T cells may have the potential to be cross-reactive across MHC classes.

The review showed that only a very sparse number of tools use structural information to predict TCRpMHC interaction, which inspired the third project included in this thesis. The third project in this thesis addresses the potential of using structural energies to improve the prediction of TCR recognition of MHC presented peptides. The previous publications do not employ deep learning in their prediction setup. We, therefore, wanted to investigate how much information could be extracted from energy terms using deep learning and to which extent energy terms can improve TCRpMHC interaction prediction. In this study, we observed that global energy terms have prediction power when used by themselves, indicating that energy terms could be of interest in predicting TCR recognition of MHC presented peptides. We also saw that although there is some signal that can help distinguish binders from non-binders, it is difficult to find patterns that are generalizable across different peptides, making it a very difficult task to predict.

9.1 Limitations

One major limitation for both the CD8+ CD4+ lineage prediction project and the CD8+ T cell epitope prediction project is the limited availability of experimental data. The data used in these projects are all gathered from publicly available databases, where the data is collected from multiple different projects and datasets. The use of datasets collected from different experiments can lead to the type of experiment conducted and the quality of data differing between the datasets, which means that not all data may be equally reliable, which can impact the models trained on the data.

Another limitation is that modeling of TCRs and TCRpMHC have proven challenging to perform, especially due to the hypervariability of the CDR loops;

this poses a limitation on how precise the structural energies calculated can be and thereby how well it can represent the data and its potential differences.

Global energy terms were used to add structural information in the CD8+ T cell epitope prediction project. Although the use of global energy terms creates a smaller model with fewer parameters, thereby reducing the extent of possible overfitting, the information it provides may not be detailed enough.

9.2 Future perspectives

Rosetta is not only capable of calculating global energy terms but per residue energy terms as well. The per residue energy terms contain information regarding how individual amino acids behave. Individual amino acids can have signals that can be of interest for the model to pick up on, which may be lost when only using global energy terms since these terms account for the total structure. Therefore, it could be of interest to study whether the per residue terms can improve the predictive capability of a model for predicting TCRpMHC interaction.

The field of structural modeling of protein sequences has seen advancements with the newest edition of the AlphaFold model (54). This model has been shown to model protein structures better than other models. A model which predicts structures that are more precise could improve how accurate the energy calculations will be. Furthermore, any differences in energy between binders and non-binders may become clearer with better modeling. However, one major limitation of this model is that it is still not optimal at predicting protein complexes and proteins with highly variable loops. However, it could still be interesting to test how well this model can represent the structures and whether this can improve the accuracy of the energy calculations.

Bibliography

[1] Ken Murphy, Paul Travers, and Mark Walport. *Janeway's immunobiology.* Garland Science, New York, 7th edition edition, 2008. 1, 5, 6, 9

[2] K. Esfahani, L. Roudaia, N. Buhlaiga, S. V. Del Rincon, N. Papneja, and Wilson H. Miller. A review of cancer immunotherapy: from the past, to the present, to the future. *Current Oncology*, 27(Suppl 2):S87, 2020. 2

[3] Masayuki Hirano, Sabyasachi Das, Peng Guo, and Max D. Cooper. The evolution of adaptive immunity in vertebrates. *Advances in immunology*, 109:125–157, 2011. 6

[4] Ronald N. Germain. T-cell development and the CD4–CD8 lineage decision. *Nature Reviews Immunology 2002 2:5*, 2(5):309–322, 2002. 6

[5] Charles A. Janeway. The T cell receptor as a multicomponent signalling machine: CD4/CD8 coreceptors and CD45 in T cell activation. *Annual review of immunology*, 10:645–674, 1992. 6

[6] Andrea C. Carpenter and Rémy Bosselut. Decision checkpoints in the thymus. *Nature Immunology 2010 11:8*, 11(8):666–673, 7 2010. 7

[7] Simon R. Carding and Paul J. Egan. $\gamma\delta$ T cells: functional plasticity and heterogeneity. *Nature Reviews Immunology 2002 2:5*, 2(5):336–345, 2002. 7

[8] Daniel J. Laydon, Charles R.M. Bangham, and Becca Asquith. Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1675), 8 2015. 9

[9] Vanessa Jurtz, Sinu Paul, Massimo Andreatta, Paolo Marcatili, Bjoern Peters, and Morten Nielsen. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *Journal of immunology (Baltimore, Md. : 1950)*, 199(9):3360–3368, 11 2017. 10

[10] Stephanie Gras, Zhenjun Chen, John J. Miles, Yu Chih Liu, Melissa J. Bell, Lucy C. Sullivan, Lars Kjer-Nielsen, Rebekah M. Brennan, Jacqueline M. Burrows, Michelle A. Neller, Rajiv Khanna, Anthony W. Purcell, Andrew G. Brooks, James McCluskey, Jamie Rossjohn, and Scott R. Burrows. Allelic polymorphism in the T cell receptor and its impact on immune responses. *The Journal of Experimental Medicine*, 207(7):1555, 7 2010. 10

[11] Marek Wieczorek, Esam T. Abualrous, Jana Sticht, Miguel Álvaro-Benito, Sebastian Stolzenberg, Frank Noé, and Christian Freund. Major histocompatibility complex (MHC) class I and MHC class II proteins: Conformational plasticity in antigen presentation. *Frontiers in Immunology*, 8(MAR):292, 3 2017. 10

[12] Wing Ki Wong, Jinwoo Leem, and Jinwoo Leem. Comparative Analysis of the CDR Loops of Antigen Receptors. *Frontiers in Immunology*, 10:2454, 10 2019. 11

[13] Michael Schantz Klausen, Mads Valdemar Anderson, Martin Closter Jespersen, Morten Nielsen, and Paolo Marcatili. LYRA, a webserver for lymphocyte receptor structural modeling. *Nucleic Acids Research*, 43(W1):W349–W355, 7 2015. 11

[14] Kamilla Kjærgaard Jensen, Vasileios Rantos, Emma Christine Jappe, Tobias Hegelund Olsen, Martin Closter Jespersen, Vanessa Jurtz, Leon Eyrich Jessen, Esteban Lanzarotti, Swapnil Mahajan, Bjoern Peters, Morten Nielsen, and Paolo Marcatili. TCRpMHCmodels: Structural modelling of TCR-pMHC class I complexes. *Scientific Reports 2019 9:1*, 9(1):1–12, 10 2019. 11

[15] Andrej Šali and Tom L. Blundell. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*, 234(3):779–815, 12 1993. 11

[16] Lu Deng and Roy A. Mariuzza. Recognition of self-peptide-MHC complexes by autoimmune T-cell receptors. *Trends in biochemical sciences*, 32(11):500–508, 11 2007. 11

[17] Pedro Larrañaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Iñaki Inza, José A. Lozano, Rubén Armañanzas, Guzmán Santafé, Aritz Pérez, and Victor Robles. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112, 3 2006. 13

[18] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 1 2015. 13

[19] J. R. Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27(3):221–234, 9 1987. 14

[20] Bjoern H. Menze, B. Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A. Hamprecht. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 10(1):1–16, 7 2009. 14

[21] Yoshua Bengio, Olivier Delalleau, and Clarence Simard. DECISION TREES DO NOT GENERALIZE TO NEW VARIATIONS. *Computational Intelligence*, 26(4):449–467, 11 2010. 14

[22] Tin Kam Ho. Random decision forests. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 1:278–282, 1995. 14

[23] Leo Breiman. Random Forests. *Machine Learning 2001 45:1*, 45(1):5–32, 10 2001. 14

[24] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`. 15, 18

[25] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature 2015 521:7553*, 521(7553):436–444, 5 2015. 16

[26] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics 1943 5:4*, 5(4):115–133, 12 1943. 16

[27] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 11 1958. 16

[28] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015. 18

[29] Yann Lecun and Yoshua Bengio. Convolutional Networks for Images, Speech, and Time-Series. In M.A. Arbib, editor, *The handbook of brain theory and neural networks*. MIT Press, 1995. 19

[30] Ji Young Lee and Franck Dernoncourt. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, pages 515–520, 3 2016. 19

[31] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 11 1997. 21

[32] Jeffrey L. Elman. Finding Structure in Time. *Cognitive Science*, 14(2):179–211, 3 1990. 22

[33] Xue Ying. An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, 1168(2):022022, 2 2019. 26

[34] Lutz Prechelt. Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, 11(4):761–767, 6 1998. 26

[35] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014. 26

[36] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *32nd International Conference on Machine Learning, ICML 2015*, 1:448–456, 2 2015. 26

[37] Steven Henikoff and Jorja G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915, 1992. 27

[38] Raphael Guerois, Jens Erik Nielsen, and Luis Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology*, 320(2):369–387, 2002. 28

[39] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The FoldX web server: an online force field. *Nucleic acids research*, 33(Web Server issue), 7 2005. 28

[40] Sitao Wu and Yang Zhang. Protein structure prediction. *Bioinformatics: Tools and Applications*, pages 225–242, 2007. 28

[41] Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of chemical theory and computation*, 13(6):3031–3048, 6 2017. 28

[42] Pierre Baldi, Søren Brunak, Yves Chauvin, Claus A.F. Andersen, and Henrik Nielsen. Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*, 16(5):412–424, 2000. 30

[43] Maria Hauser, Martin Steinegger, and Johannes Söding. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics*, 32(9):1323–1330, 5 2016. 30

[44] Rachel A. Woolaver, Xiaoguang Wang, Alexandra L. Krinsky, Brittany C. Waschke, Samantha M.Y. Chen, Vince Popolizio, Andrew G. Nicklawsky, Dexiang Gao, Zhangguo Chen, Antonio Jimeno, Xiao Jing Wang, and Jing Hong Wang. Differences in TCR repertoire and T cell activation underlie the divergent outcomes of antitumor immune responses in tumor-eradicating versus tumor-progressing hosts. *Journal for ImmunoTherapy of Cancer*, 9(1):e001615, 1 2021. 30

[45] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22(13):1658–1659, 7 2006. 31

[46] Alex D. Waldman, Jill M. Fritz, and Michael J. Lenardo. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nature Reviews Immunology 2020 20:11*, 20(11):651–668, 5 2020. 33

[47] Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, Samuel A.J.R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, Niccolò Bolli, Ake Borg, Anne Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P. Butler, Carlos Caldas, Helen R. Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John A. Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinsk, Natalie Jäger, David T.W. Jones, David Jonas, Stian Knappskog, Marcel Koo, Sunil R. Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C. Munshi, Hiromi Nakamura, Paul A. Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V. Pearson, Xose S. Puente, Keiran Raine, Manasa Ramakrishna, Andrea L. Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N. Schumacher, Paul N. Span, Jon W. Teague, Yasushi Totoki, Andrew N.J. Tutt, Rafael Valdés-Mas, Marit M. Van Buuren, Laura Van 'T Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R. Yates, Jessica Zucman-Rossi, P. Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M. Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M. Pfister, Peter J. Campbell, and Michael R. Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013. 33

[48] Harpreet Singh-Jasuja, Niels P.N. Emmerich, and Hans Georg Rammensee. The Tübingen approach: identification, selection, and validation of tumor-associated HLA peptides for cancer therapy. *Cancer immunology, immunotherapy : CII*, 53(3):187–195, 3 2004. 33

[49] Ido Springer, Hanan Besser, Nili Tickotsky-Moskovitz, Shirit Dvorkin, and Yoram Louzoun. Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs. *Frontiers in Immunology*, 11:1803, 8 2020. 33

[50] Ido Springer, Nili Tickotsky, and Yoram Louzoun. Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction. *Frontiers in Immunology*, 12:1436, 4 2021. 33

[51] Pieter Moris, Joey De Pauw, Anna Postovskaya, Sofie Gielis, Nicolas De Neuter, Wout Bittremieux, Benson Ogunjimi, Kris Laukens, and Pieter Meysman. Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. *Briefings in Bioinformatics*, 22(4):1–12, 7 2021. 33

[52] Anna Weber, Jannis Born, and Mariá Rodriguez Martínez. TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics*, 37(Supplement_1):i237–i244, 7 2021. 33

[53] Alessandro Montemurro, Viktoria Schuster, Helle Rus Povlsen, Amalie Kai Bentzen, Vanessa Jurtz, William D. Chronister, Austin Crinklaw, Sine R. Hadrup, Ole Winther, Bjoern Peters, Leon Eyrich Jessen, and Morten Nielsen. NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. *Communications Biology 2021 4:1*, 4(1):1–13, 9 2021. 33

[54] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A.A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873*, 596(7873):583–589, 7 2021. 75

Paper II: Appendix

Supplementary Material

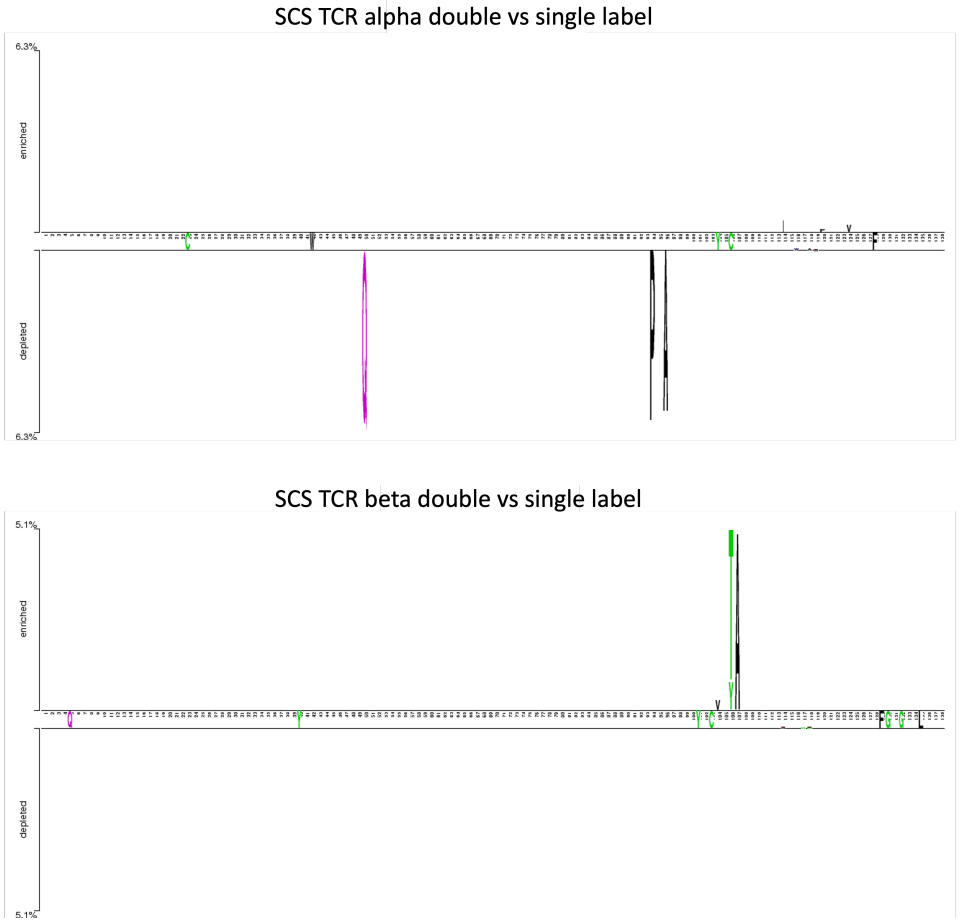


Figure S1. Two sample logo plot created with SCS double labels as ”positive” and single labels as ”negative”. Here double label means that paired TCRs have been found both on a CD4+ and CD8+ T cell. The single label denotes paired TCRs that have only been found on either a CD4+ or CD8+ T cell.

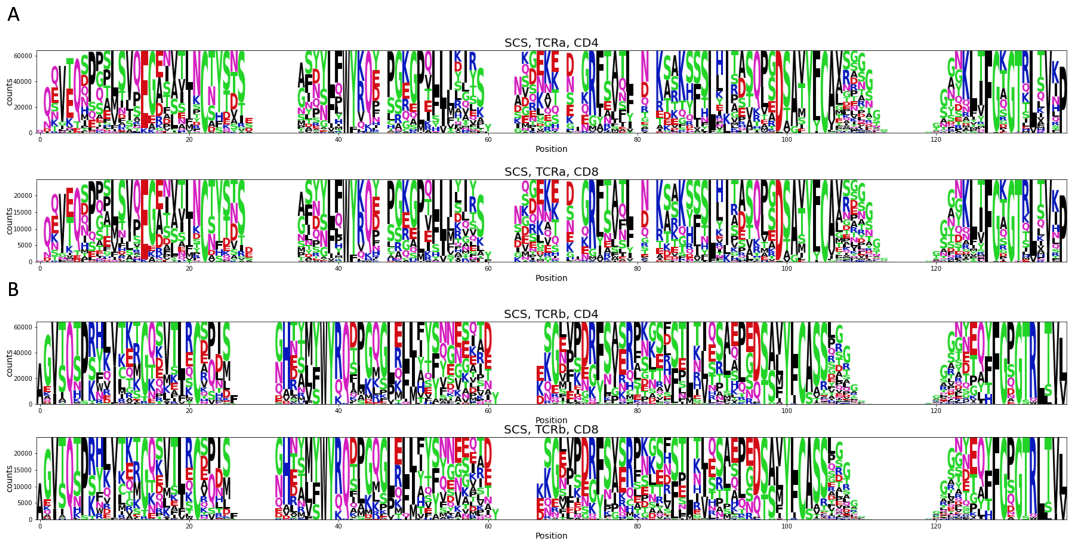


Figure S2. Differences between CD4 and CD8 TCRs in the single cell dataset. Logo plots showing the difference in the TCR of the CD4 and CD8 α chain (A) and CD4 and CD8 β chain (B).

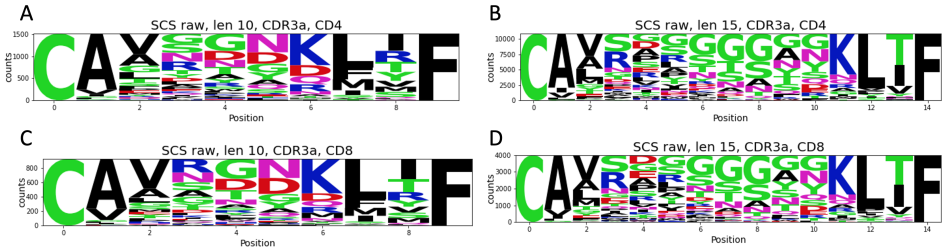


Figure S3. Logo plots showing the similarities and differences in the CDR3 α chain of TCR sequences of length 15 and 10. SCS CDR3 α sequences for CD4 of length 15 (A), SCS CDR3 α sequences for CD4 of length 10 (B), SCS CDR3 α sequences for CD8 of length 15 (C), SCS CDR3 α sequences for CD8 of length 10 (D).

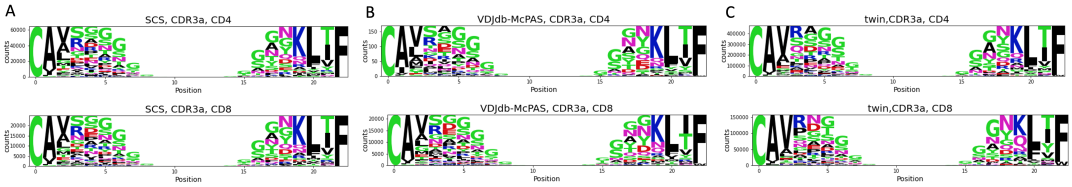


Figure S4. Differences and similarities between the CDR3α chain within different datasets. Comparing logo plots between the single cell dataset (A), the VDJdb-McPAS dataset (B) and the twin dataset (C).

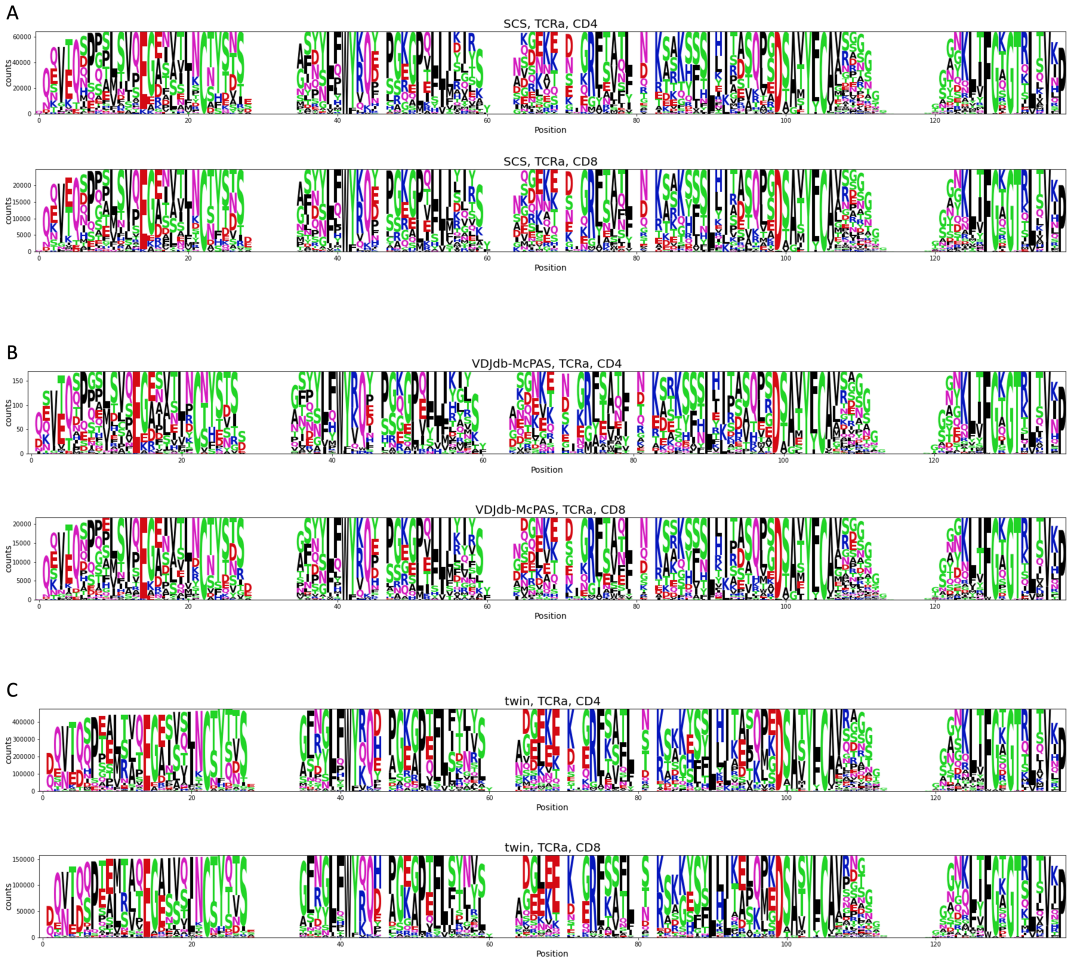


Figure S5. Differences and similarities between the TCRα chain within different datasets. Comparing logo plots between the single cell dataset (A), the VDJdb-McPAS dataset (B) and the twin dataset (C).

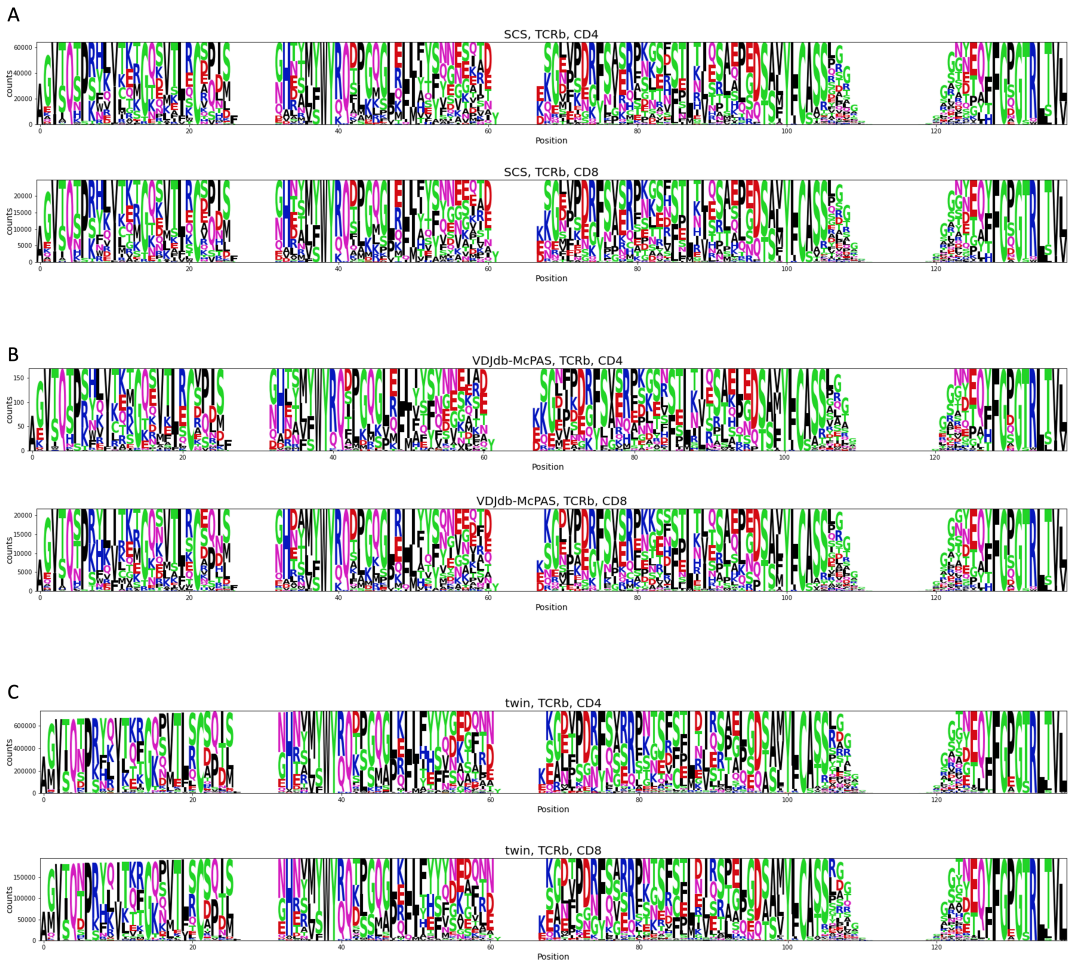


Figure S6. Differences and similarities between the TCRβ chain within different datasets. Comparing logo plots between the single cell dataset (A), the VDJdb-McPAS dataset (B) and the twin dataset (C).

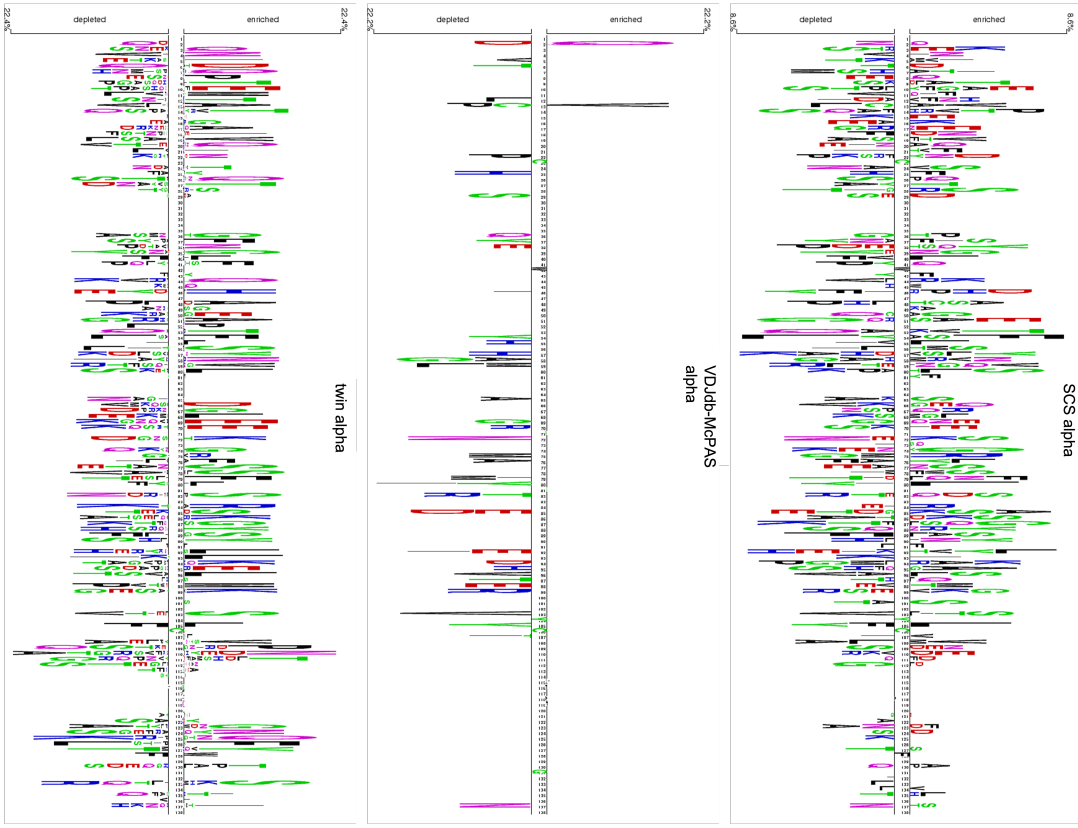


Figure S7. Two-sample logo plot showing the differences in the TCRα sequences from CD4+ and CD8+ T cells within the different datasets. Comparing two-sample logo plots between the single cell dataset (A), the VDJdb-McPAS dataset (B) and the twin dataset (C). Here an enrichment indicates that a given amino acid at a given position is upregulated in CD8+ T cell TCR sequences and vice versa.)

DTU Health Tech
Department of Health Technology

Technical University of Denmark
Kemitorvet, Building 204
2800 Kongens Lyngby

www.healthtech.dtu.dk