universität innsbruck
Department of Computer Science

Neo Epitope Prediction for Personalized Cancer Immunotherapy

Matthias Leitner

M.Sc. Thesis

Supervisor: Justus Piater
Hubert Hackl
24th February 2020

Abstract

With advances in genomic sequencing technologies, personalized immunotherapy in cancer has gained significant traction and various computational approaches based on different analytical approaches have emerged to confidently identify surface antigens of individual tumors to use as therapeutic targets. In order to provide a high fidelity and easy to apply solution for neo epitope identification for cancer treatment, this thesis provides a new analytical workflow designed to run directly on raw sequence data without requiring additional preprocessing steps or other forms of user interaction during the analysis process. Overcoming the shortcomings of various other workflows, mutational information is derived from single nucleotide variations, indels and frameshift mutations and phase information, with respect to both somatic and germline mutations, is considered in the workflow provided in this thesis. With the goal of identifying features relating to the identified antigen's ability to lead to tumor rejection, data on 21,148 in vitro tested antigenic sequences has been collected and analysed. Out of the various models tested, a random forest approach based on BLOMAP encoding of the individual amino acids was found to provide the best predictive abilities and as such was included to provide an additional parameter for peptide evaluation for therapeutic application.

Contents

Abstract i

Contents iii

List of Figures v

List of Tables vii

Acronyms ix

Declaration xi

1 Introduction 1

2 Immunotherapy 3
2.1 Immune Checkpoint Therapy . . . 3
2.2 Cancer Vaccines . . . 4

3 Biological Background 7
3.1 Central Dogma of Molecular Biology . . . 7
3.2 Mutations . . . 8
3.3 Antigen Presentation . . . 10
3.4 Cytotoxic T-Cells . . . 11

4 Analytical Workflow 13
4.1 Motivation . . . 13
4.2 State of the Art . . . 15
4.3 Neo Epitope Prediction . . . 16
4.3.1 Data Preprocessing and Variant Calling . . . 18
4.3.2 Neo Epitope Identification . . . 21
4.4 Pipeline Execution . . . 30

5 Immunogenicity Prediction 33
5.1 Motivation . . . 33
5.2 State of the Art . . . 33
5.3 Experimental Setup . . . 38
5.3.1 Dataset . . . 38
5.3.2 Random Forest . . . 42
5.3.3 Kernel Support Vector Machines . . . 45
5.3.4 Model Evaluation . . . 48

6 Future Research **53**
6.1 Workflow Extensions 53
6.2 Predictive Modelling Advancements 54

7 Conclusion **55**

Bibliography **57**

A Software Availability **67**

B Materials and Methods **69**

List of Figures

2.1 Cancer Vaccine Development Process 5

3.1 Central Dogma of Molecular Biology 7
3.2 Mutation Types 9
3.3 MHC-I Antigen Presentation Pathway 10
3.4 T-cell Negative Selection 12

4.1 Workflow 17
4.2 Neo Epitope Identification 22
4.3 Read Backed Phasing 25
4.4 Peptide Generation Process 26

5.1 Amino Acid Structure 36
5.2 Data Distribution 40
5.3 Distribution of Binding Affinity 41
5.4 Random Forest Model Estimation 42
5.5 Random Forest Prediction 43
5.6 Kernel Support Vector Machine 46
5.7 Bootstrap AUC results 49
5.8 Random Forest Variable Importance 50
5.9 Repeated Measures Anova 52

List of Tables

4.1 Result values 29
4.2 Pipeline Input Parameters 30
4.3 Epitope Prediction Program Parameters 31

5.1 Data Sources 39
5.2 Data Set Distribution 40
5.3 Peptide Encodings 45
5.4 Five Point Summary of Bootstrap Results 49
5.5 Shapiro-Wilk Normality Test results 51
5.6 Mauchly's Test for Sphericity 51
5.7 Bonferroni adjusted post-hoc two sample T-test 52

A.1 Pipeline dependencies 67

B.1 5 dimensional BLOMAP encoding 70
B.2 Blosum 35 71
B.3 Blosum 62 72
B.4 BLOSUM 62-2 73

Acronyms

A Adenine.

AAC Amino Acid Composition.

ANOVA Analysis of Variance.

APseAAC Amphiphilic Pseudo-Amino Acid Composition.

AUC Area under the Curve.

BLOSUM Blocks Substitution Matrix.

BQSR Base Quality Score Recalibration.

BWA Burrows Wheeler Aligner.

C Cytosine.

CDF Cumulative Distribution Function.

CTL Cytotoxic T Lymphocyte.

CTLA-4 Cytotoxic T Lymphocyte-Associated Protein 4.

DAI Differential Agretopicity Index.

DNA Desoxyribonucleic Acid.

ELISpot Enzyme Linked Immuno Spot.

ER Endoplasmic Reticulum.

FDA Food and Drug Administration.

G Guanine.

GATK Genome Analysis Toolkit.

HLA Human Leukocyte Antigen.

HPC High Performance Computing.

KSVM Kernel Support Vector Machine.

MHC Major Histocompatibility Complex.

MHC-I Major Histocompatibility Complex class 1.

mRNA Messenger RNA.

NGS Next Generation Sequencing.

NSCLC Non-Small-Cell Lung Carcinoma.

PAM Point Accepted Mutation.

PD-1 Programmed Cell Death 1.

PseAAC Pseudo Amino Acid Compositon.

PTM Post-Translational Modification.

QSO Quasi Sequence Ordering.

RBF Radial Basis Function.

RF Random Forest.

RNA Ribonucleic Acid.

SGE Son of Grid Engine.

SNV Single Nucleotide Variation.

STAR Spliced Transcripts Alignment to a Reference.

SVM Support Vector Machine.

T Thymine.

TAP Transporter Associated with Antigen Processing.

TCR T-cell Receptor.

TPM Transcripts per Million.

TSA Tumor Specific Antigen.

VAF Variant Allele Frequency.

VEP Variant Effect Predictor.

WES Whole Exome Sequencing.

WGS Whole Genome Sequencing.

Declaration

By my own signature I declare that I produced this work as the sole author, working independently, and that I did not use any sources and aids other than those referenced in the text. All passages borrowed from external sources, verbatim or by content, are explicitly identified as such.

Signed: ..... Date: .....

Chapter 1

Introduction

In the search for new and alternative treatments for cancer, the area of immunotherapy has become a promising field of research. Using a patient's own immune system to systematically search for and destroy cancerous cells, immunotherapeutic approaches provide an alternative to conventional cancer treatments such as chemo therapy or surgery and harbour the potential of overall causing fewer side effects (Rosenblum et al., 2015).

Many of such approaches in immunotherapy first require the identification of the set of all mutations which are specific to the tumor cell. Due to cell's internal machinery, these tumor specific mutations contain the potential to give rise to characteristic peptide sequences which will only be displayed on the surface of cancerous cells as antigens and, therefore, be completely absent from healthy cells. As such, these antigens - which are due to their tumor restrictedness also often referred to as neoantigens - provide the central target for many immunotherapeutic approaches. As reported by Alcazer et al. (2019) the specific antigenic sequences may vary strongly between individual tumors and as such require a personalized approach.

It should be noted at this stage that throughout scientific literature the terms antigen and epitope are often used interchangeably. As such, both terms will also be used synonymously in this thesis.

As identification and evaluation of such cancer antigens in a laboratory presents a very costly and time consuming approach, as stated by Karasaki et al. (2017), computational methods provide a more cost effective and less time intensive alternative for identification and preselection of such neoantigens.

Due to technological advances in the area of genome sequencing, allowing for cheaper execution and resulting in overall more accurate results, various neoantigen prediction pipelines have emerged. Relying on different analytical approaches, certain limitations and shortcomings of the individual solutions can be identified to varying degrees.

One commonly shared limitation found across a large range of pipelines is that they only consider Single Nucleotide Variations (SNVs) as a source for neoantigens, disregarding more complex mutational forms. As argued by Smith et al. (2019b), more complex mutation types should be considered as they bear the potential to result in antigens which are more different from healthy sequences and as such with greater ability to aid in tumor rejection as opposed to SNV derived antigens.

Furthermore, many of the existing approaches analyze identified mutations in isolation, hence as implicated by Hundal et al. (2019), they implicitly assume that the positions surrounding a mutation are equal to the reference genome. Proximal mutations, i.e. mutations which lie on the same copy of a gene and in close proximity to one another, may, when analyzed in combination, result in a different antigen than when investigated alone. Such phase information, either from other cancer mutations, or mutations specific to the given patient, is hence often disregarded, resulting in potentially different predicted antigens.

Another commonly encountered trait in existing approaches is that varying amounts of preprocessing are required before the analysis can be performed. Whether operating on identified mutations or requiring them to further be annotated, by additional programs, only few workflows have internalized preprocessing steps and are able to operate on raw sequence data directly.

In order to overcome these limitations, this thesis presents a new analytical approach for computational neoantigen identification and evaluation for which the inclusion of following features has been of central importance:

- Consideration of Single Nucleotide Variations (SNVs), indels and frameshift mutations as potential sources for cancer neoantigens.
- Inclusion of patient specific mutations which are also present in healthy cells
- Incorporation of phase information for evaluation of proximal mutations
- Machine learning driven prediction of antigen's immunogenicity
- Ease of application in a clinical setting by operating directly on raw sequence data without further user interaction
- Enabling of customization towards varying qualities of used data by exposing filtering parameters to the user.

The subsequent parts of this thesis is organized as follows:
In the first part a general overview of the field of immunotherapy will be given in Chapter 2. Next, in Chapter 3 the underlying biological processes governing how tumor specific mutations give rise to cancer specific neoantigens and neo epitopes are reviewed. Chapter 4 will present in detail how the here proposed workflow is structured and how a custom software solution was designed to take into account all the above mentioned information to derive antigenic sequences from identified mutations. In the next part, Chapter 5, various approaches to modelling an antigen's likelihood of invoking an immune response are explored, and statistical evaluation of the generated models is provided. Chapter 6 will then outline potential areas which can be improved upon in the future, and Chapter 7 will conclude this thesis.

Chapter 2

Immunotherapy

The most common methods to treat cancer involve surgery, radiation and chemotherapy. Despite these methods having been widely established over the years, Boopathi et al. (2019) state that such therapeutic methods are often very expensive and prone to heavy side effects as they fail to effectively discriminate cancerous cells from healthy host cells. Further concerns with classical approaches in cancer treatment are indicated by Holohan et al. (2013) noting that chemotherapy as well as other molecularly targeted therapies harbour the risk of creating resistances in cancer cells to these types of approaches. Therefore, in search of less side effect prone treatment options, immunotherapy has become a promising field of research. Based on the idea of activating the patient's on immune system, immunotherapeutic approaches hinge on the idea of provoking a strong and targeted attack on only cancerous cells while ignoring healthy host cells.

Cells of the human immune system already possess the ability to identify and eliminate cancerous cells by means of the antigen presentation mechanism (see Chapter 3). However, abusing various strategies, cancerous cells can stay hidden from the immune system or even deactivate cells about to launch an attack. As such, immunotherapeutic approaches can be used to aid immune cells in overcoming such strategies and help them to successfully identify and kill cancerous cells.

While built on the same underlying idea, various immunotherapeutic approaches have emerged over the years. This thesis will provide a brief overview of immune checkpoint therapy - as this form of therapy also has the potential to provide synergetic effects when used in conjunction with other approaches - before moving on to the development of cancer vaccines - the development of which the workflow presented in this thesis is designed for.

2.1 Immune Checkpoint Therapy

In order to be able to control an ongoing immune response and prevent the immune system from attacking the host's own cells, specialized immune cells - so called T-cells (see Chapter 3 - are equipped with a specific set of receptor molecules on their surface which, when interacted with, deactivate the respective immune cell, and hence prevent the destruction of the targeted cell. These surface molecules are referred to as immune checkpoints and the exploitation of these molecules presents a possible way for cancerous cells to avoid destruction by the host's immune system. (Marin-Acevedo et al., 2018).

Immune checkpoint inhibitors are specialized antibodies that when administered to a patient

bind to these surface molecules and hence prevent their activation by cancerous cells. Popular targets for this inhibitory approach include surface molecules such as Programmed Cell Death 1 (PD-1) or Cytotoxic T Lymphocyte-Associated Protein 4 (CTLA-4). The inhibition of these checkpoint molecules has already been shown promising results in overcoming a tumor's immunosuppressive abilities (Pardoll, 2012), (Hassel, 2016).

Every since its first application, immune checkpoint therapy has shown very promising results in cancer treatment. A case study following multiple melanoma patients, presented by Ott et al. (2019), shows that in two out of the five examined patients anti-PD-1 therapy was able to lead to the complete and durable regression of the tumor. Due to such encouraging results immune checkpoint therapy has already become a vital part in cancer treatment and as indicated by Narang et al. (2019) has even been approved by the Food and Drug Administration (FDA) as a treatment for various forms of cancer, including melanoma or Non-Small-Cell Lung Carcinoma (NSCLC).

2.2 Cancer Vaccines

An alternative and more personalized immunotherapeutic approach - which will also be the focus of this thesis - is the generation of cancer vaccines.

Mutations that lead to the emergence of cancer, can allow those cells to be distinguished from healthy cells by means of mutated antigens presented on their surface. In targeting those particular antigens, a highly specific immune response can be launched against cells displaying such peptides while healthy cells which don't present those particular sequences are left alone.

The underlying idea of cancer vaccination relies on exploiting knowledge of which antigens will only be present on the surface of tumor cells and as denoted by Hundal et al. (2016) to selectively boost the immune cells reactive to those particular tumor antigens. In doing so, only a selected part of the immune system will be boosted increasing the anti cancer response, specifically.

Considering that only part of the immune system will be affected by administration of such vaccines, Rosenblum et al. (2015) note that this form of interventional therapy is theoretically more safe than various alternatives as healthy cells which are not capable of producing the respective antigens will not be targeted by the boosted part of the immune system.

Application of cancer vaccines does, however, not have to happen in isolation. As Alcazer et al. (2019) note, synergetic effects between different immunotherapeutic approaches exist and additional application of checkpoint inhibitors alongside vaccination harbours the potential of further increasing tumor rejective effects.

In terms of clinical application, studies implicate that overall accomplishments have been somewhat limited. Alcazer et al. (2019) argue that the percentage of observed objective clinical responses lie below than 7% and the overall clinical benefit is estimated around 20%. In spite of this low rate of success, many case studies have been presented outlining the potential achievements that could be realized through cancer vaccination.

With melanoma being the most prominent cancer form to which vaccines are applied, Ott et al. (2019) present a case study on five patients with advanced state melanoma. Two of which

responded immediately after being treated with vaccines alone. A third patient was found to experience complete regression after having additionally received anti-PD-1 treatment following the vaccination. Another study on melanoma patients presented by Ott et al. (2017) found that out of six vaccinated patients, four stayed free of any recurrence even 25 months after the initial administration, whereas the other two patients would after an additional dose of anti-PD-1 experience complete tumor regression. This case study gives also an indication of potential side effects of vaccination therapy, as the authors report treatment induced side effects to only consist of mild flu-like symptoms or reactions at the injection site such as rashes, supporting the notion that cancer vaccines are less side effect prone than conventional therapies.

With its comparatively high amount of somatic mutations, melanoma has become a popular target for cancer vaccination therapy. Such approaches, however, are not limited to melanoma alone. Johanns et al. (2019), for example, demonstrate that even in tumors with a lower number of mutations such as glioblastoma personalized neoantigen vaccines were successfully able to infiltrate tumors of the central nervous system and therefore enhance the immune response against the cancerous cells.

The central target of such cancer vaccines are as indicated antigens specific to the respective tumor cells. As such Tumor Specific Antigens would not be able to be presented by any healthy cell and are from the point of view of the immune system completely new, they are often also referred to as neoantigens. Such neoantigens are generated as a result of the mutations that lead to the emergence of cancer in the first place, and as studies like Zhou et al. (2019a) show, the specific mutations may diverge largely across individuals as well as differ substantially depending on the specific types of tumor. As a result of this, the authors also argue that cancer vaccination requires assessment of each patient's tumor individually - requiring a very personalized approach.

A general overview of the steps involved in the creation of a personalized cancer vaccine is depicted in Figure 2.1 and will be briefly outlined below.

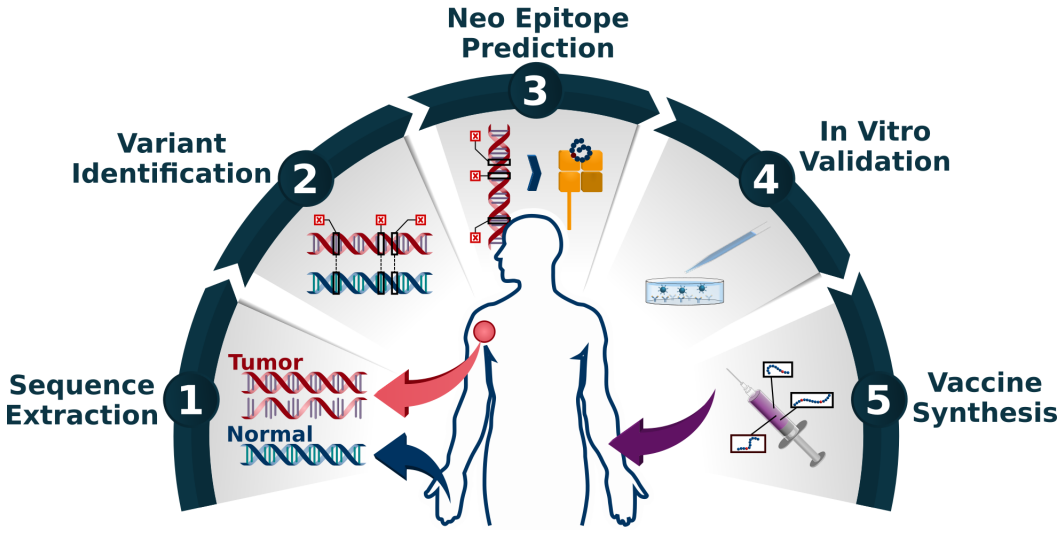

Figure 2.1: Cancer Vaccine Development Process

1. As neoantigens are derived from mutations in the genomic code, the starting point for all tumor specific analyses is given by the genomic sequences of both tumor cells and normal cells. So, in a first step, corresponding tissue needs to be extracted from the patient

and it's genomic code needs to be sequenced using techniques from the area of NGS. As indicated by Schmidt and Lill (2019), an additional step often conducted is to also generate sequences from tumor RNA which allows for an overview of the tumor's transcriptomic profile, alongside its genomic one.

2. After extraction and preprocessing of the raw sequence information a vital part of the analysis lies in the identification of mutations that are specific to the given tumor by comparing the respective genomic sequences to those of healthy unmutated cells.

3. Having identified the tumor's mutational landscape, the respective mutations can be translated into the mutated proteins arising from their inclusion in a cell's genetic code and potential neoantigens originating from these mutated sequences can be inferred. Also, through the usage of machine learning methods, the likelihood of the generated peptides to be presented on an MHC-I complex as well as subsequently inducing an immune response can be evaluated.

4. After identification and potential prioritization by computational methods, before their therapeutic application, the predicted sequences need to be evaluated experimentally in a laboratory through immunological assays such as the Enzyme Linked Immuno Spot (ELISpot) assay (Czerkinsky et al., 1983).

5. Lastly, the in vitro confirmed epitopes can be used for the synthesis of a vaccine specifically tailored to the analyzed tumor which can then be administered to the patient.

As indicated by Hundal et al. (2016) the most expensive and time consuming part of this overall process consists of the manufacturing and testing of antigenic peptides. Being comparatively cheap to execute, computational methods have a large incentive to provide an accurate selection of potential cancer rejective candidates such as to limit the amount of possible peptides to check in vitro. Therefore, a crucial part of in silico analysis and prediction of neo eptiope candidates is to employ various filters such as to limit the number of false positives and negatives to substantially alleviate the burden on in vitro evaluation.

It should also be noted at this point that alternatives to computational identification of neoantigens such as tandem mass spectrometry exist. However, as stated by Smith et al. (2019b) such methods are more error prone and less sensitive making computational methods still the most widely used approach for identifying tumor specific neoantigens.

Chapter 3

Biological Background

As cancer immunotherapy is based on the idea of harnessing the patient's own immune system to elicit cancer rejection, this aim of this section is to provide a general overview of the fundamental mechanisms by which mutations in a cell's DNA give rise to Tumor Specific Antigens (TSAs) which can aid in the identification and subsequent elimination of cancerous cells.

3.1 Central Dogma of Molecular Biology

As a cell's most central component, the double stranded DNA molecule contains information that allow for the synthesis of various proteins a cell requires in order to fulfill its function. This genetic code is provided in form of a sequence of four possible complementary bases: Adenine (A), Thymine (T), Cytosine (C) and Guanine (G). The process by which cells convert this genetic code into proteins is governed by the central dogma of molecular biology and can be split into two distinct steps (see Figure 3.1).

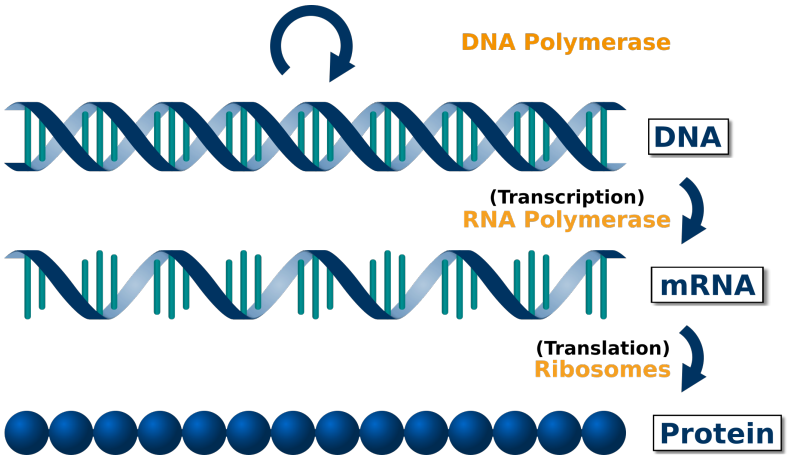

Figure 3.1: Central Dogma of Molecular Biology

The first step - transcription - occurs in a cell's nucleus and refers to the process of using the the double stranded DNA as a template and synthesizing a specialized single stranded molecule called Messenger RNA (mRNA).

After processing, the mRNA molecule leaves the nucleus and travels into the cytoplasm where it binds to ribosomes in order for the second stage to take place. During the translation phase, the sequence of the mRNA is read three bases at a time and for each of these non-overlapping

consecutive three-base sequences - referred to as codons - one of a possible set of 20 amino acid is integrated into a directed chain, thus forming a protein.

As each position in a codon can contain one of four possible bases, overall $4^3 = 64$ distinct codons are possible each corresponding to one of 20 possible amino acids, indicating that the same amino acid can be encoded by multiple different three base sequences. Two special types of codons that should be noted are **ATG** which encodes for the amino acid **Methionine** and also doubles as a start codon highlighting the position in the mRNA at which translation should begin. Correspondingly, the sequences **TAA**, **TAG** and **TGA** serve as stop codons, not encoding any amino acid themselves but instead marking the termination of the translation process.

3.2 Mutations

Since DNA contains the coding sequences for the proteins which are to be synthesized, modifications to this genetic code may therefore result in the creation of different proteins with potentially altered functionality. Accumulation of such mutations in a cell's DNA is, as denoted by Karasaki et al. (2017), a characteristic of the development of cancerous cells.

Mutations that change the genomic sequence and therefore result in modified proteins may come in various forms. The following types of mutations will be the subject of this thesis:

- **Single Nucleotide Variations (SNVs)**
  Probably the most simplistic and as indicated by Smith et al. (2019b) most researched form of mutations are Single Nucleotide Variations. These types are characterized by a single substituted base in the coding sequence and as such do not influence the overall length or reading frame. Depending on the particular replacement, SNVs can be further divided into synonymous (or silent) mutations and non-synonymous (or missense) mutations. As many amino acids are encoded by multiple three base sequences, a codon affected by a SNV mutation may still result in the same amino acid being introduced in the generated protein and is hence considered synonymous. In contrast, if the mutation leads to a different amino acid being encoded, the mutation is considered to be non-synonymous.

- **Insertions**
  Changes in the coding sequence that affect its length and subsequently the length of the protein generated are referred to as Indels - as a portmanteau derived from insertions and deletions. Insertion mutations, therefore, specify changes in which additional bases are included into the coding sequence. For easier distinguishability, throughout this thesis, only mutations through which a multiple of three bases is added into the DNA sequence and hence at most two contiguous codons are affected while leaving all subsequent ones intact are referred to as insertions.

- **Deletions**
  The second type of indel mutation consists of deletions. As opposed to their insertion counterparts, deletions refer to the loss of coding bases from the original sequence, hence resulting in a shorter protein. Analogously to insertions, the term deletion will only be used to refer to removal of a multiple of three bases from the coding sequence.

- **Frameshift Mutations**
  The last type of mutation included in this analysis are so called frameshift mutations which can be considered special cases of insertions or deletions and are as such also able

to affect the length of the synthesized protein. Frameshift mutations denote insertions or deletions in which bases are added or removed in non-multiples of three. As a protein's coding sequence is translated in groups of three bases at a time into their corresponding amino acids, these types of mutations, are not only able to influence the codon in which they occur but affect the grouping of all subsequent bases into codons as well. As such they change the whole reading frame of the coding sequence leading to potentially vastly different proteins being synthesized.

A graphical representation of these types of mutations is presented in Figure 3.2.

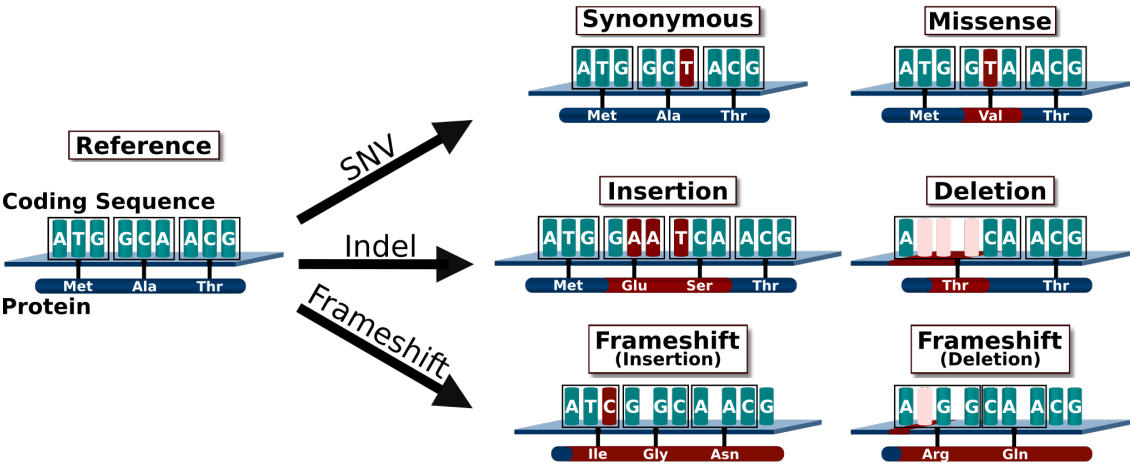

Figure 3.2: Mutation Types

It should be noted at this point that the above list of mutations is not exhaustive and other types such as the fusion of genes can result in the generation of neoantigens. Additionally, altered proteins may not exclusively emerge due to changes in the underlying coding sequence. As indicated by Zhou et al. (2019b), additional sources can be identified such as translation of non-coding regions within the DNA sequence, edits to the mRNA like alternative splicing, or even changes to the protein which occur after the coding sequence has been translated - so called Post-Translational Modifications (PTMs) - can give rise to different proteins.

In the context of this thesis, however, in terms of the generation of mutated proteins, only the above mentioned four types of mutations are considered.

An important distinction that has to be made is that based on their occurrence mutations can be grouped into either germline or somatic mutations. Germline mutations refer to the DNA changes that occur in reproductive cells and are therefore passed down from parents to their children. As such they can be found in all of an individual's healthy cells and are in their particular configuration specific to an individual.

Somatic mutations, on the other hand, name the types of mutations which occur in singular cells due to internal factors such as errors in the DNA duplication process or external influences including smoking or UV radiation. As opposed to their germline counterparts which are usually benign, somatic mutations are initially only found in individual cells and may result in these cells eventually turning cancerous. As such, the main focus of therapeutic cancer intervention is the identification and evaluation of such somatic mutations. However, in order to provide a more accurate representation of a patient's set of proteins, both healthy and mutated, the

analytical pipeline outlined in this thesis also puts a high emphasis on the inclusion of germline mutations alongside somatic mutations.

3.3 Antigen Presentation

Cells that have acquired somatic mutations and thus synthesize altered version of proteins can notify the immune system by displaying small parts of these mutated proteins as antigens on their surface using class 1 molecules of the Major Histocompatibility Complex (MHC). These MHC-I molecules are found on the surface of all nucleated cells of most vertebrates (Starr et al., 2003), including humans, where they are also often referred to as Human Leukocyte Antigen (HLA) which is why both abbreviations will be used interchangeably in this thesis.

The general process by which antigens are generated from synthesized proteins and presented on the cell's surface will be briefly outlined below and is depicted in Figure 3.3.
A more in-depth explanation of the MHC-I antigen presentation process can be found in Yewdell et al. (2003).

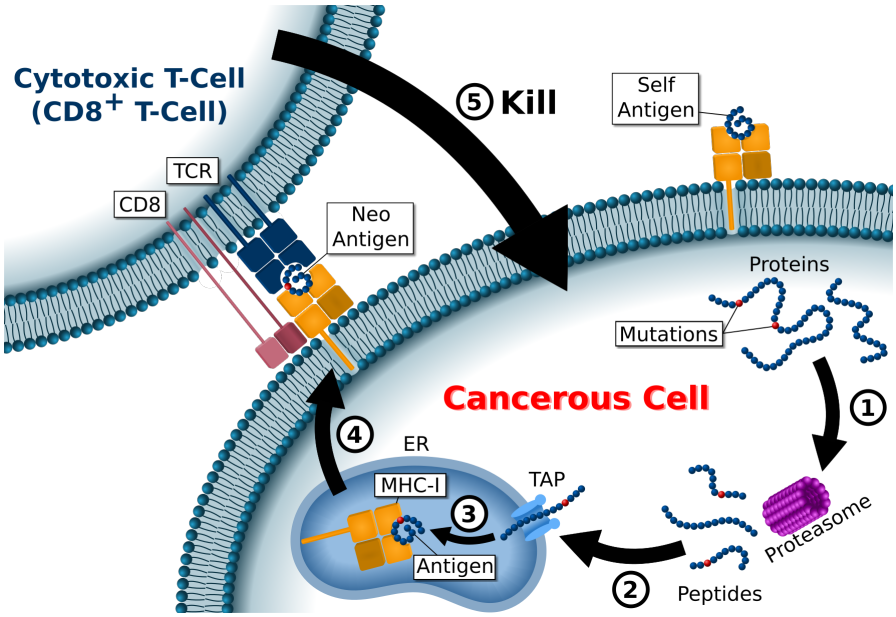

Figure 3.3: MHC-I Antigen Presentation Pathway

In the cell's cytosol, enzymes such as the proteasome are responsible for degrading both normal as well as mutated proteins and cutting them into shorter fragments. The so generated short protein fragments, also referred to as peptides, dissolve towards the Endoplasmic Reticulum (ER) on the surface of which the Transporter Associated with Antigen Processing (TAP) molecule is located. Upon interaction between these small peptides and the TAP, peptide fragments are transported to the inside of the ER - the ER lumen.

It is here that those peptides are then further processed and cut into even smaller fragments such that they are between eight and eleven amino acids long. It also occurs within the ER lumen that the MHC-I molecule is assembled and ready to interact with a peptide for subsequent presentation. Peptides that display a high binding affinity towards the MHC-I complex, i.e. closely fit the binding grove, are loaded onto the MHC-I molecule and are then transported

to the cell's surface where they can be examined by cells of the host's immune system.

As indicated by Buhler and Sanchez-Mazas (2011), genes encoding the HLA molecules are among the most variable positions in the human genome, implying that the specific types of HLA molecules may differ substantially between individuals. As such, a major contributing factor to a peptide's binding affinity is the individual's combination of HLA types, indicating that the specific set of presentable peptides may also vary largely from person to person.

This overall process of protein degradation and possible loading onto an MHC-I complex is applied to both normal and mutated proteins alike. As proteins containing somatic mutations have the potential of being cut into peptide fragments which would be completely absent from healthy cells, antigens arising from such mutated proteins are usually referred to as neoantigens or Tumor Specific Antigens (TSAs).

3.4 Cytotoxic T-Cells

Cytotoxic T Lymphocites (CTLs), also often referred to as T-Killer cells or CD8 positive (CD8+) T-cells - due to the presence of the CD8 receptor on their surface - are able to interact with MHC-I complexes and inspect the presented antigen using a specialized receptor on their surface - the T-cell Receptor (TCR).

In order for a presented antigen to be recognized by the immune system and therefore result in subsequent destruction of the distressed cell, a CD8+ T-cell with a T-cell Receptor matching the presented antigen is required. If a CTL with a matching TCR binds to the presented antigen, specialized molecules are released, resulting in the death of the targeted cell.

The process by which individual TCRs are generated is highly stochastic as to allow the immune system to identify a large variety of foreign peptide sequences (Livak et al., 2000). Due to their random generation, TCRs reacting to unmutated antigens presented by healthy cells will inevitably be generated which would result in the immune system attacking healthy host cells. Therefore, a crucial step during their development in the thymus, is negative selection - also referred to as central tolerance - in which T-cells falsely identifying normal self peptides as foreign are eliminated and only cells possessing the ability to correctly distinguish foreign peptides form self survive (Hogquist et al., 1994) (see Figure 3.4).

Although a T-cell possesses multiple copies of only a single specific TCR on its surface, the receptor is able to recognize multiple even seemingly different peptide fragments - an ability which was termed polyspecificity (Wucherpfennig et al., 2007). It is due to this polyspecificity that certain mutations, albeit resulting in antigens completely absent from healthy cells yet similar enough to normal peptides, will not be recognizable by the host's immune system, as any TCR capable of its recognition would have been deleted during thymic development (Duan et al., 2014). Calis et al. (2012) indicate that about one third of peptides generated from sources completely absent from healthy tissue are too similar to self peptides and would not be capable of eliciting an immune response leading to the destruction of cancerous cells.

As noted by Frankild et al. (2008), due to this polyspecificity, certain holes are generated in the set of recognizable antigens which a tumor can use to remain undetected from the immune system. Conversely, any mutation producing a drastically different neoantigen would likely be recognized by the host's immune system leading to the cell's death. As such, Dunn et al. (2002)

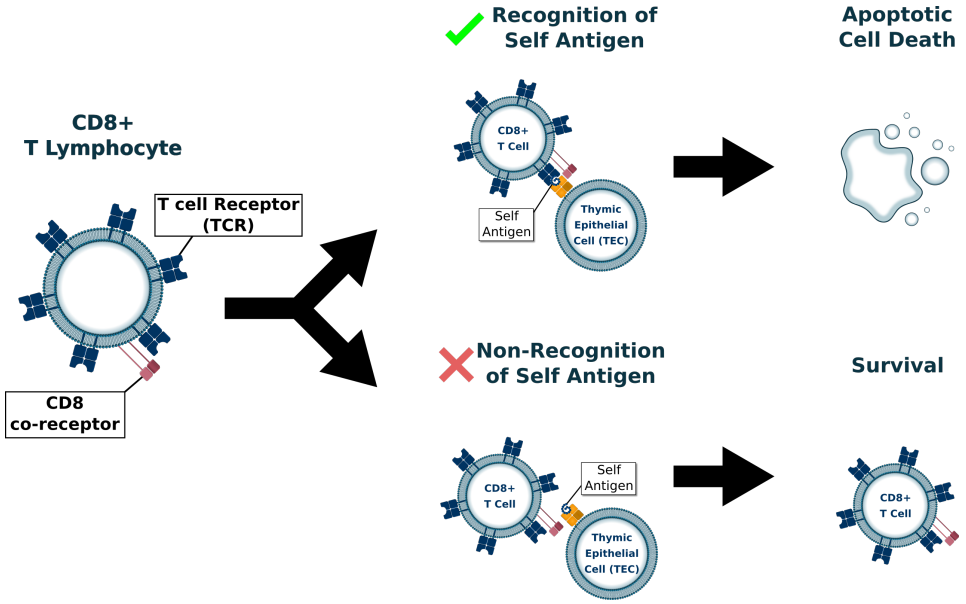

Figure 3.4: T-cell Negative Selection

note that the mutational landscape of tumors having survived the immune system have likely been influenced as to exploit these holes in the repertoire of TCRs - a process which is called immunoediting.

Due to their innate ability of recognizing and eliminating rogue cells, CTLs play a major role in targeted cancer immunotherapy. As only cancerous cells would possess the ability to display specific neoantigens, therapies targeted towards those characteristic sequences would in theory allow for confident eradication of tumor cells while at the same time ignoring healthy cells.

Chapter 4

Analytical Workflow

As outlined above, with laboratory based evaluation methods and chemical peptide synthesis still presenting a major cost intensive and time consuming aspect of the overall process, computational identification of tumor restricted neoantigens has become a vital factor in the development of therapeutic cancer vaccines.

With various computational workflows already in existence, this section will first provide a high level overview of the general analysis process involved in epitope prediction and restate the goals for an easily applicable high confidence analysis. Furthermore, current analytical approaches are reviewed and shortcomings with respect to the stated goals are identified before the custom workflow developed in this thesis is presented in detail.

4.1 Motivation

Being based on information contained in genomic sequences, neo epitope identification workflows have largely benefited from recent advances in sequencing as well as the creation of various Next Generation Sequencing (NGS) technologies. As such, identification of DNA and RNA sequences has become increasingly cheaper to conduct and has become more and more able to provide much higher confidence results.

Due to these advancements in sequencing technology and the gaining influence of immunotherapy in cancer treatment, various analytical workflows for the identification of cancer mutations have been proposed. While building on similar underlying principles, Karasaki et al. (2017) argue that the amount and the particular sequences of the identified neoantigens, may be greatly dependent on the specific strategies involved in their identification.

The overall process of deriving antigenic peptides from raw sequence data can be described as follows:

As a basis for the analysis, data derived from Whole Genome Sequencing (WGS) or Whole Exome Sequencing (WES), containing the whole genome or just the protein coding sequences, respectively, from both healthy patient tissue and cells taken from a tumor sample, is used. Often, in addition to the cancer's genomic profile, tumor RNA - specifically mRNA - is included as well giving insight into the tumor's transcriptomic profile - i.e. which genes of the DNA sequence will actually be transcribed and eventually used to synthesize proteins. An example for the importance of such transcriptomic information is given by Narang et al. (2019) finding that only 51% of their identified neo epitopes were actually expressed at RNA level, meaning

that roughly half of their found neo epitopes could be disregarded in further analysis.

The two DNA samples from tumor and healthy tissue, respectively, are in a next step compared against one another in a process called variant calling and mismatches between the two can be identified as somatic mutations, i.e. mutations that are only contained within tumor cells and hence bear the potential for tumor specific mutated proteins being created.

Subsequently, the identified mutations are incorporated into the respective unmutated reference coding sequence and mutated proteins can be generated computationally by mimicking the underlying biological processes. The so obtained mutated protein sequences then build the basis for the identification of neoantigens, and neo epitopes, in that they can be broken down into shorter fragments each of which can further be analyzed with respect to binding affinity towards molecules of the MHC-I complex or their potential of eliciting tumor rejective capabilities, before being validated in vitro by means of biological assays.

While the overall process shares large amounts of similarities throughout many proposed analytical pipelines, depending which particular parts of the information contained in the biological sequences are used for further downstream analysis and which filtering strategies are used, results generated by different workflows may lead to drastically divergent results.

Due to this lack of a standardized approach, a workflow is presented in this thesis based on the following principles to provide results as close to reality as possible while at the same time allowing for easy applicability in a therapeutic setting:

- **Usage of Multiple Types of Mutations**
  A common limitation found in many analytical workflows is that only effects of Single Nucleotide Variations are considered for the analysis. This shortcoming is also implicated by Koşaloğlu-Yalçın et al. (2018) indicating that more complex forms of mutations are expected to produce more immunogenic results, due to their capability of introducing more excessive alterations into the coding sequence resulting in antigens differing more substantially from their healthy counterparts. As such, the types of mutations included in this workflow are as follows: **SNVs**, **Insertions**, **Deletions** and **Frameshift** mutations.

- **Inclusion of Germline context**
  Hundal et al. (2019) indicate that apart from sites of identified somatic mutations, surrounding sites are often implicitly expected to be equal to a reference genome. As every individual's genome slightly differs from one another by means of germline mutations, an additional step is introduced into the overall process and an individual's germline mutations are called alongside tumor specific ones and considered in downstream analysis. In doing so, a more accurate representation of both mutated and non-mutated peptide sequences can be provided.

- **Incorporation of Phase Information**
  Commonly, identified mutations are solely evaluated in isolation. As mutations occurring in close proximity to one another may affect the same protein, prior to protein generation, identified mutations - both germline and somatic - will be phased to evaluate whether they should be considered in combination for downstream analysis.

- **Immunogenicity Prediction**
  Many analytical workflows indicate a peptide's fitness for further therapeutic usage by performing MHC-I binding affinity predictions. As indicated above, while being necessary

in terms of application, strong binding affinity is not sufficient for a neoantigen to elicit tumor rejective capabilities. As such in addition to evaluating the peptides' binding affinities, a custom machine learning approach is used for evaluating each antigen's potential to invoke an immune response (a detailed explanation is given in Chapter 5), allowing for further reduction of the set of potential vaccination candidates.

- **End to end**
  A major goal in terms of usability is to provide an analytical workflow which can be run directly from raw sequence data without the need for explicit data preprocessing or requiring further user interaction. Using a specialized workflow engine in conjunction with container technologies, all preprocessing steps are integrated in the overall process such that the proposed workflow can be run on any High Performance Computing (HPC) system directly on raw sequence data.

- **Adjustability of Filtering Parameters**
  In order to be able to accommodate data sources of different levels of quality, parameters used to filter the number of false positives and false negatives are exposed to the user to allow for tailoring the filtering process to a specific application.

4.2 State of the Art

Ever since the idea of cancer immunotherapy has been presented and computational analysis has gained momentum due to advances in sequencing technologies, various analytical pipelines for the computational identification of neo epitopes have been generated.

As direct evaluation of the accuracy of such analytical processes is notoriously hard given that data from conducted vaccination studies are few and far between, the aim of this section is to give a brief overview over existing workflows and indicate differences with respect to the goals for this thesis as outlined above.

A major shortcoming shared by many approaches that aim at deriving neo epitopes from Tumor Specific Antigens (TSAs) is as indicated by Smith et al. (2019b) their restrictiveness in terms of mutation types considered. Many pipelines solely rely on SNVs. However, as also described above, different types of mutations harbour larger potential in terms of leading to immunogenic neo epitopes.

An approach considering such more complex mutation types is presented by Zhang et al. (2017). The program presented by the authors - INTEGRATE-Neo - is specifically designed to identify neo epitopes based on gene fusions. However, as the authors do not provide a standalone pipeline, additional steps such as sequence alignment and variant calling are required to be handled by the user beforehand.

This lack of a standardized pre processing protocol is also shared by several other approaches. PVAC-Seq (Hundal et al., 2016), for example, also requires an already preprocessed list of mutations as well as a list of the amino acid changes resulting from these mutations and transcript sequences for their analysis.

To the best of my knowledge, at the current stage, the only pipeline designed to run end to end without requiring user interaction or heavy preprocessing is presented by ProTECT (Rao et al., 2019). However, this pipeline exhibits another shortcoming which is commonly observed,

in that patient specific germline context is not taken into account. Other popular pipelines such as MuPeXI Bjerregaard et al. (2017a) or the above mentioned pVAC-Seq share the same problematic in that either germline context is not considered or in-phase variants are not considered, and the sites around mutated positions are implicitly assumed to be equal to the reference genome.

While most analytical workflows are built around the same idea of using tumor and normal sample DNA to identify somatic mutations, a different approach was presented by Duan et al. (2014). In their pipeline called Epi-Seq the authors derive mutations from RNA sequences from cancerous cells. A major advantage of an RNA based approach over a DNA based analysis is that a mutated protein's coding sequence can directly be inferred from the RNA as it already incorporates any type of alteration occurring prior to protein synthesis. Therefore, such an RNA based approach would eliminate the need for additional germline variant calling or phasing steps.

While such an approach has the potential of incorporating a wide range of mutations without requiring additional analysis steps, it should be noted that, as indicated by Xu (2018), RNA sequencing is a different process than DNA sequencing. As the authors state, due to this difference in sequencing protocols, additional sources for errors exist in RNA sequencing such that the obtained data is usually less reliable and therefore bears a higher error rate.

As a result, throughout many analytical workflows, mutations are typically called from DNA samples and RNA sequences' predominant use - if at all included - is to provide an overview of which genes are actually transcribed and used for protein creation.

A slightly different usage for RNA-seq data is found in the analytical pipeline Vaxrank as presented by Rubinsteyn et al. (2017). Starting from already called mutations, the authors use RNA data to confirm the presence of the presented mutations in the tumor's transcriptomic profile and as such use it as an additional filtering process where any mutation not present in a required amount of transcripts is disregarded.

While the above mentioned pipelines do not present an exhaustive list of all existing analytical workflows, it provides an indication that existing approaches differ largely among one another and, to the best of my knowledge, no singular pipeline comprising all the above laid out requirements exists at this point.

Therefore, this thesis presents a novel analytical pipeline which runs end to end on raw sequencing data, provides fully customizable filter parameters, includes Single Nucleotide Variations, insertions, deletions and frameshift mutations and takes proximal variants, both germline and somatic into account. The full analysis process is outlined in the section below.

4.3 Neo Epitope Prediction

The process of neo epitope prediction and subsequent assessment can be split into two parts:

1. Preprocessing of raw sequence data and variant calling
2. Neo epitope generation and evaluation with respect to binding affinity and immunogenicity

The first part will be explained in further detail in Section 4.3.1 while the second part will be explored in Section 4.3.2. A graphical overview of the whole process is given in Figure 4.1.

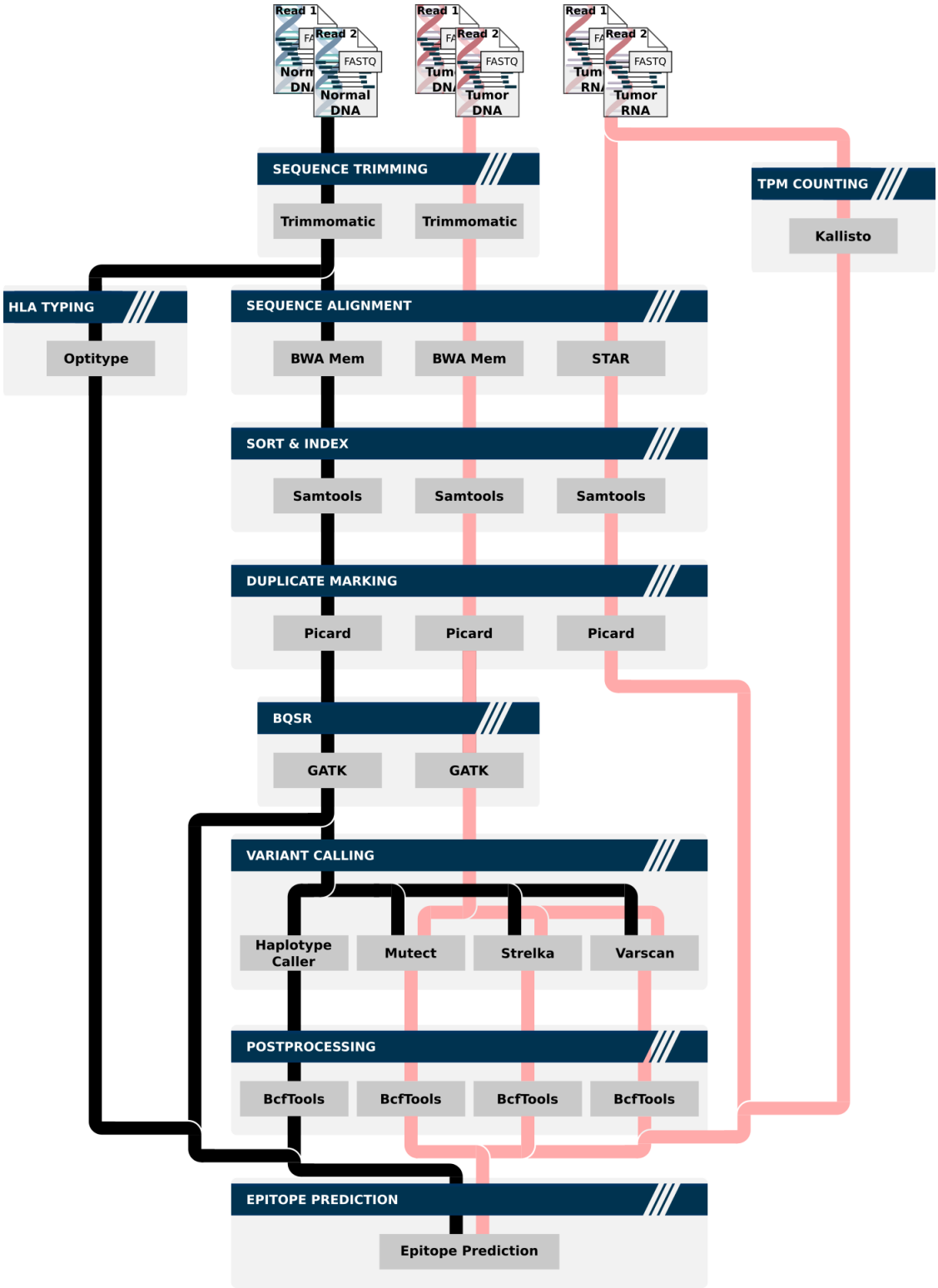

Figure 4.1: Workflow

4.3.1 Data Preprocessing and Variant Calling

As a starting point for the analytical process, input data has to be presented consisting of DNA from healthy cells, Tumor DNA as well as RNA sequences from the cancerous cells - the combination of which is often times referred to as the genomic trio (Rao et al., 2019).

As outlined above, a major consideration in the creation of this pipeline was to be able to directly operate on raw sequence data. Therefore, in a first step, the raw data is preprocessed and brought into a form on which variant callers can operate. This necessary preprocessing step has been well established and a best practices workflow is presented by Van der Auwera et al. (2013). In closely following the best practices steps as indicated, frequently used and well established tools in the scientific community have been combined into a singular pipeline. The steps of this process are outlined below.

Sequence Trimming

As a first step of quality control, raw DNA sequences from both the normal as well as the tumor sample are trimmed using the sequence trimmer **Trimmomatic** (Bolger et al., 2014). The aim of this process is twofold:

First, any residual adapter sequences that were added to the DNA strand prior to sequencing and have been left over are removed. Second, in order to assure the quality of the reported sequences, reads with low confidence in that they have been identified correctly are removed. For the purpose of quality inspection, alongside the raw sequence identified by the sequencer, quality scores are given to every position of the sequence indicative of the sequencer's confidence of having called the position correctly. These quality measures - reported as Phred scores (Ewing et al., 1998) - can be used to remove parts of the identified sequences, if their corresponding score falls below a certain threshold.

It should be noted at this point, that this step of sequence trimming is only performed for DNA reads in this workflow. While DNA sequences will be primarily used for variant calling, the main idea behind the inclusion of RNA sequences is to give an overview of the tumor's transcriptome. As such, Williams et al. (2016) argue that trimming of RNA data usually is applied too eagerly and may negatively impact downstream gene expression analyses. Therefore, RNA trimming is not included in this workflow.

Sequence Alignment

During the sequencing process both DNA and RNA need to be cut into shorter chunks, as sequencing of both types of molecules as a whole is not possible. Therefore, a vital step before the sequences can be further analyzed is to align the short sequence reads to a reference genome.

In this step, each of the obtained reads is compared to a standardized reference sequence and the position it best matches to is evaluated. This step is performed for both DNA and RNA alike, however, due to their different natures, different tools as well as different references have to be used.

For DNA alignment, **Burrows Wheeler Aligner (BWA)** (Li and Durbin, 2009) has been chosen. As RNA is derived from non-contiguous sequences of the human genome, which are spliced together prior to being translated into proteins (Dobin et al., 2013), and still may con-

tain mismatches with respect to a reference genome, the **Spliced Transcripts Alignment to a Reference (STAR)** program (Dobin et al., 2013) is used for alignment.

At the end of this step for both DNA samples as well as RNA reads a file containing the original reads annotated with the position in the genome they have been matched with as well as a score indicating the quality of this match is returned.

Sort and Index

In order to facilitate further processing, the previously position annotated sequences are ordered according to their position in the genome and an index file is created for easier and faster analysis.

For this step the widely employed **Samtools** (Li et al., 2009) has been used.

Duplicate Marking

Having been properly sorted and indexed, the next step for all samples consist of the identification and subsequent marking of duplicate sequences. As indicated by Van der Auwera et al. (2013), a source of error that can occur is given by the fact that during the sequencing process the same molecule is sequenced multiple times from which identical reads are generated which bear no further informational content.

As the presence of such duplicate reads can skew further downstream analyses, the **Picard** tool (Institute, 2019) is used to scan the sorted reads for such identical sequences, and mark them, such that they can be disregarded from downstream analyses.

Base Quality Score Recalibration

Before the deduped DNA reads can be used for variant calling, an additional step which is recommended to be included by the best practices pipeline is to perform Base Quality Score Recalibration (BQSR) (Van der Auwera et al., 2013).

As the authors indicate, sequencers themselves may be subject to systematic errors when computing the certainty with each individual base in a genomic sequence is called. As these quality scores are used in further analysis and also variant callers rely heavily on the accuracy of the reported scores, it is important to limit the possibility of false positives through controlling for any sort of systematic error introduced by the sequencers in their assessment.

For this purpose, the Base Quality Score Recalibration tool from the **Genome Analysis Toolkit (GATK)** (Van der Auwera et al., 2013) is used which leverages machine learning methods to identify and subsequently correct such systematic errors in quality score assessments.

Variant Calling

One of the most vital steps in neoantigen prediction is the identification of the underlying tumor specific mutations that may lead to the emergence of mutated proteins. This process is commonly referred to as (somatic) variant calling and relies on the comparison of both tumor derived and healthy normal cell derived DNA sequences.

While the overall process of comparing already aligned sequences to one another to identify

non matching positions appears to be straight forward, Van der Auwera et al. (2013) note that multiple sources of errors exist which make this process overall more difficult. Among others, as the authors indicate, such errors may stem from biases during the sequencing process as well as errors induced when aligning the reads to a reference genome.

An additional problem which should be noted at this stage is, as indicated by Alcazer et al. (2019), that tumors are, especially once they have reached more advanced stages, often heterogeneous, indicating that only part of the tumor will exhibit certain mutations while they are completely absent from the rest of the tumor, hence making the identification from common mutations more difficult. Another major difficulty arises from the fact that tumor samples are rarely pure. As pointed out by Karasaki et al. (2017), such samples usually contain both cells from healthy tissue as well as mutated cells which poses further risks of errors in identifying tumor specific mutations especially when considering that the sequencing process itself may wrongly identify certain bases which may also pose the risk of being falsely identified as mutations.

Therefore, multiple strategies in variant calling exist to limit the number of falsely presented mutations, and results may vary depending on which variant caller was used. As reported by Ewing et al. (2015), in order to mitigate this potential for errors and improve the overall accuracy, an ensemble of variant callers should be preferred over a single one.

Therefore, in this pipeline the combination of three variant callers is used, namely GATK's **Mutect2** (Cibulskis et al., 2013), **Strelka2** (Saunders et al., 2012), and **Varscan2** (Koboldt et al., 2012), all of which were chosen due to their abundant usage throughout scientific literature. To mitigate false positives reported by a singular variant caller, a somatic mutation is only considered for further analysis if it is confirmed by at least two out of the three programs used.

As a core objective of this workflow is to incorporate germline information into the prediction process, at this point not only somatic mutations are called but also germline mutations are identified using GATK's **Haplotype Caller** (McKenna et al., 2010).

Postprocessing

In order to make comparison between the results obtained from the individual variant callers easier, an additional post processing step is introduced after germline and somatic variants have been identified. For this purpose the **normalize** function provided by the **BcfTools** program (Li et al., 2009) is used.

As both **Strelka** and **Varscan** output two files - one containing SNVs and the other for indels - the **concat** function is used prior to normalization to make subsequent analyses easier.

HLA Typing

As indicated above, a major factor when selecting predicted neo peptides for therapeutic application, is whether they will actually be presented on top of the cancerous cell's surface. As such, it is necessary to assess each peptide's binding affinity towards the respective combinations of MHC-I complexes.

As the particular types of the MHC-I molecules differ largely between individuals, an important

step is to discern the particular combination of HLA types particular to the patient that is being analyzed.

For this purpose, **OptiType** (Szolek et al., 2014) is used to infer the specific HLA types based on the DNA sequences obtained from healthy donor cells, after potential residual adapter sequences as well as low quality base calls have been removed, i.e. after the sequence trimming phase.

TPM Counting

Additionally, as the main purpose of the inclusion of tumor RNA is to get an overview of the tumor's transcriptome, raw RNA sequences are assigned to the respective genes they were generated from in a process called pseudo-alignment.

The pseudo-aligner used in this workflow - **Kallisto** (Bray et al., 2016) - reports counts on how many transcripts are generated from the respective genes in the form of Transcripts per Million (TPM), a metric which can at a later step be used to discern which peptides actually stem from transcribed genes and should be further be considered.

4.3.2 Neo Epitope Identification

Having identified both germline and somatic mutations that are restricted to the tumor in question, as well as the patient's distinct combination of HLA types and the tumor's transcriptional profile, the next step consists in taking this information and inferring the resulting mutated peptide sequences.

While the workflow of preprocessing raw sequence data and subsequent identification of mutations has been quite standardized, no such definitive standard is provided for deriving mutated antigens from mutational data. As such, depending on how germline and phase information is handled, as well as which types of mutations are considered in the analysis, different resulting epitope sequences may be identified.

Due to this lack of standardization, in order to derive neo epitopes from mutational data, a custom approach has been developed and implemented which is integrated into the overall pipeline. A graphical overview on how the mutational data is incorporated to derive neo epitope sequences is laid out in Figure 4.2 and the individual steps will be explored in further detail below.

Variant Preprocessing and Filtering

In a first preprocessing step, results obtained from the individual variant callers are filtered and subsequently both germline and somatic variants are combined into a singular file to alleviate further processing.

A filter applied to somatic and germline mutations alike is to first disregard all mutations occurring in the genomic sequence which could not be assigned to a valid position in the sequence by the variant callers. Therefore, all mutations occurring on unknown chromosomes are dropped from further analysis.

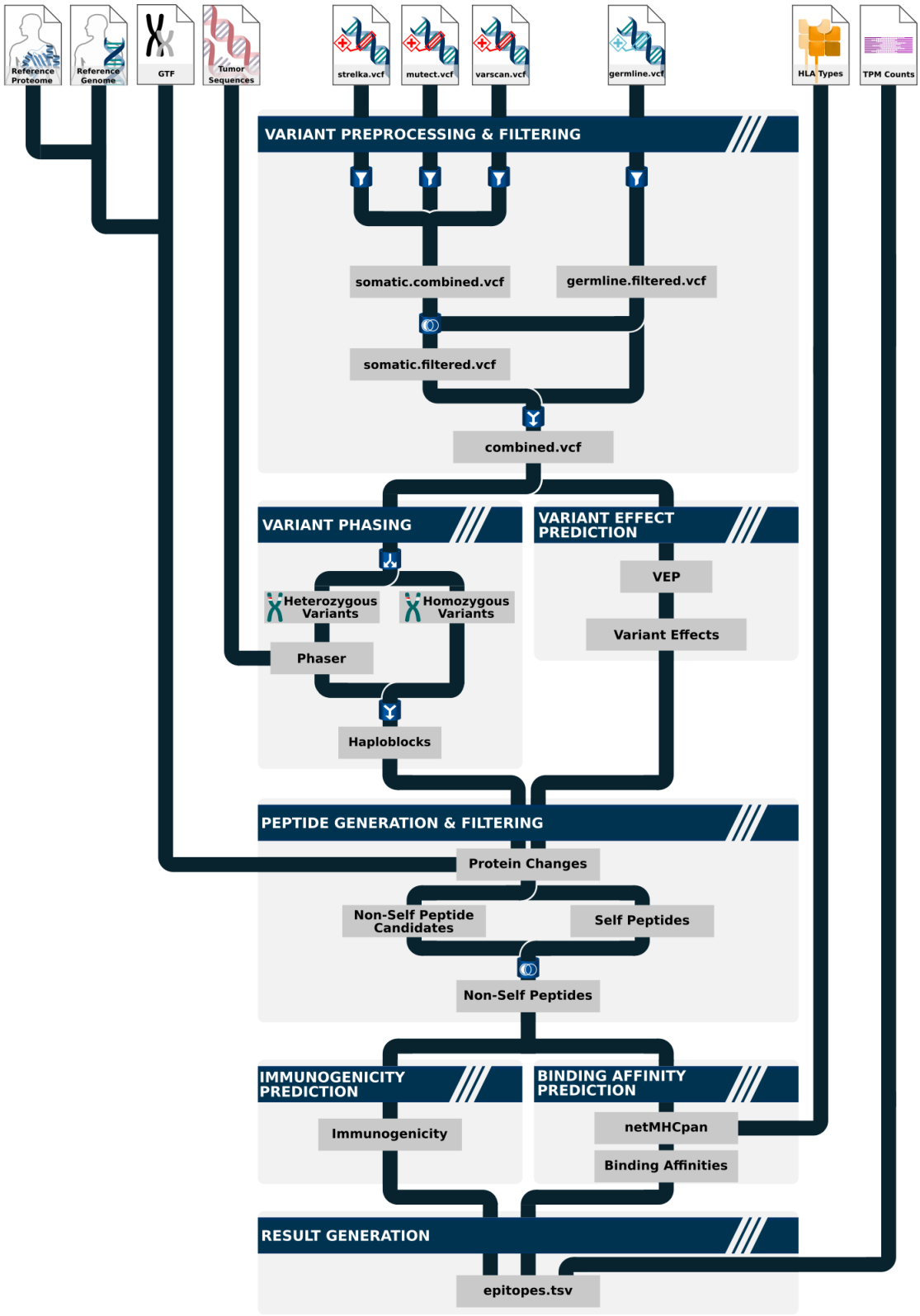

Figure 4.2: Neo Epitope Identification

A commonly applied quality control step is to filter variants based on their Variant Allele Frequency (VAF) (Hundal et al., 2016).

$$\text{VAF} = \frac{\text{Number of reads containing the allele}}{\text{Overall number of reads spanning the given site}} \tag{4.1}$$

Being computed as the number of reads containing a particular mutation divided by the total number of reads spanning the given position, variant caller software typically includes estimates of VAF for both investigated tumor and normal samples. As such, two requirements are posed in this step in that first Variant Allele Frequency for any mutation may not exceed a given threshold in the normal sample while at the same time a minimum value specified for the tumor sample has to be met.

This marks a crucial step for further investigation as this way falsely identified mutations originating from errors in the sequencing process as well as mutations only occurring in a small fraction of the tumor can be eliminated. As such, tuning of the filtering parameters marks, as indicated by Karasaki et al. (2017), an important trade off as thresholds set too lenient will result in many false positives while too strict cutoff values will possibly eliminate potentially viable mutations.

In addition to VAF filtering, for results obtained from both **Varscan** and **Strelka**, an additional check is performed. As both variant callers give an assessment of whether given mutations are considered to be somatic or are identified as germline, potential germline mutation candidates are eliminated from both results.

To further improve the accuracy of the used mutations, as outlined above, a majority voting system is used requiring a variant to be identified by at least two out of the three used variant callers to be considered for further processing.

As a final filtering step, somatic mutations are checked against the list of germline mutations identified by **Haplotype Caller**. As the process of identifying germline variants is only executed on the sample containing unmutated cells, any mutation identified by the somatic variant callers which is also identified by **Haplotype Caller** is likely germline and needs to be treated as such.

Finally, after filtering, both germline and somatic variants are marked accordingly and merged into a single variant file.

Variant Effect Prediction

After both germline and somatic variants have been filtered and marked according to their type, the combined variants are next analyzed with regards to how they affect the proteins that are being synthesized.

Using the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016), the changes introduced into the underlying coding sequence by each mutation is considered in isolation and the alterations introduced into subsequent proteins are identified.

It should be noted at this stage, that a common filter applied when evaluating changes in-

troduced by SNVs is to only consider mutations that introduce non-synonymous changes (see Chapter 3). As even synonymous variants may introduce changes when occurring downstream of other in-phase frameshift mutations, for example, all variants affecting a protein's coding sequence are considered for further analysis. If mutations do not result in the generation of neo peptides, even after being considered in combination with other in-phase mutations, those variants are implicitly disregarded at a later stage in the workflow.

Variant Phasing

A major shortcoming of various neo epitope prediction workflows is as indicated by Hundal et al. (2019) the implicit assumption that sites surrounding mutated positions are equal to the reference genome. As such, a common limitation is that for protein generation, effects of individual mutations are only considered in isolation and other somatic or patient specific germline variants in close proximity are ignored. As the authors indicate, not considering mutations in close proximity together when evaluating their overall effect on protein generation may give an inadequate image of the neoantigen landscape in that it can lead to both the occurrence of false positives and false negatives.

The importance of such information is also highlighted by Wood et al. (2019). In developing their own predictive pipeline - **neoepiscope** - the authors estimate that about 5% of neo epitopes arise from more than one variant in their coding sequence and hence require accurate processing of the proximal variants to not lead to wrong results.

However, incorporating this information of closely occurring variants is not quite straight forward. As specified by Buckley et al. (2019), every individual inherits one chromosome maternally and the other one paternally, therefore, every gene is present in every individual in two forms and any mutation - somatic or germline - can be present on either both copies of the gene - therefore be **homozygous** - or on just one copy - and be called **heterozygous**. Information on whether a given mutation is considered to be homozygous or heterozygous is typically generated by the variant caller and included alongside the called variants.

While **homozygous** variants are straight forward to incorporate as they occur on both copies, two **heterozygous** variants can either lie on the same chromosome - i.e. be proximal - or on the respective opposing chromosomes - i.e. be distal.

As pointed out by Buckley et al. (2019), standard sequencing methods do not assign reads to individual chromosomes, as such it is not directly possible to discern for two heterozygous variants that affect the same gene whether they should be considered in conjunction or in isolation in downstream analyses.

The identification process of which mutations occur together on the same gene is called variant phasing and its importance in the process of identifying neo epitopes has been indicated by several studies. Hundal et al. (2019) find that in their data as much as 5% of variants were in-phase with another missense variant. Buckley et al. (2019) denote that of all the samples considered in their analysis, in 88.3%, at least two missense mutations were found to lie in-phase with one another. And both Castel et al. (2016) and Buckley et al. (2019) highlight the importance of including such phase information into the analysis, as otherwise potentially different peptide sequences will be identified.

While various approaches for variant phasing exist, Castel et al. (2016) indicate that they often

times come with serious restrictions in terms of their applicability. The authors indicate that some of these techniques rely on specialized sequencing techniques, while others are only applicable if not only the patient's but also their parents' sequences are available, hence making them overall unfit for general application.

A different approach introduced by Buckley et al. (2019), termed VAF phasing, while presenting an easy applicable approach is, however, as indicated by the authors, only applicable to regions of somatic copy number alterations.

For those reasons, in this workflow, **phaser** (Castel et al., 2016) is used to discriminate between proximal and distal variants. The authors build upon the idea of read backed phasing (Yang et al., 2013) - i.e. using reads which span more than one site which has been indicated by the variant caller to be mutated and evaluating which of the respective mutations occur on individual reads together. (see Figure 4.3)

Figure 4.3: Read Backed Phasing

As the authors indicate, a common limitation in read backed phasing is that longer sequences spanning greater genomic distances are required in order to provide accurate results. As further noted by the authors, creation of such longer sequence reads is usually more costly. As an alternative, in their presented program - **phaser** - the authors note that read backed phasing is conducted on the basis of RNA sequences, as they are spliced together from positions in the genome which potentially lie far apart, hence spanning longer distances. A further reason for why **phaser** has been chosen is given by the fact that also DNA data can be included to improve overall accuracy, as well as that it allows for phasing of indels as well as SNVs which is not commonly supported by other read backed phasing tools such as the one provided in the Genome Analysis Toolkit (McKenna et al., 2010).

Therefore, all germline and somatic variants are split according to whether they are found on one copy of the gene or both, and the heterozygous mutations are then phased using the **phaser** tool generating blocks of mutations - so called Haploblocks - which occur on the same copy of the respective gene they affect.

Combining the output of the variant phasing process with information obtained from VEP which indicates which protein is affected by any given variant by means of a unique protein id, a look up table is generated, assigning each protein multiple blocks of in-phase mutations which jointly affect its coding sequence.

Peptide Generation and Filtering

With the information obtained from phasing and identifying in-phase changes on a protein basis, the next step consists of integrating these changes and generating the resulting mutated proteins.

A graphical representation of the individual steps of deriving a set of neo peptides from mutational information is provided in Figure 4.4

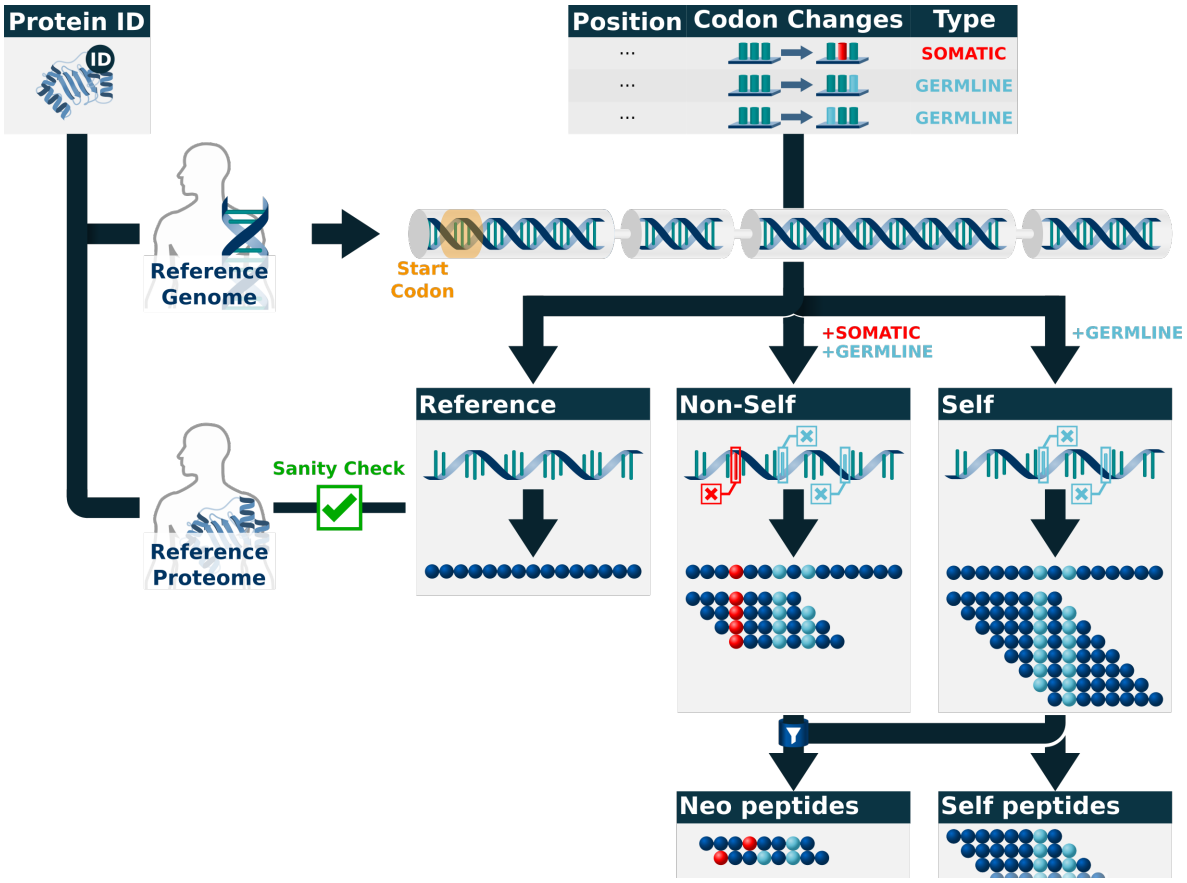

Figure 4.4: Peptide Generation Process

In a first step, the output generated by the Variant Effect Predictor is used and IDs of all transcripts which are affected by mutations are collected. Using an additional file containing positional information on the human genome, the sequence of all exons - i.e. the parts of the genomic sequence coding for the respective protein - is extracted from the reference genome. An important fact to note at this stage is that instead of directly using the reference coding sequence for a protein, which has already been trimmed to the start and stop codon, the underlying exonic sequences are used and manually spliced together. The reason for this approach is that mutations may also occur in start or stop codons resulting in the loss of the same. In terms of a mutated stop codon, this would mean that the translation process would not terminate but rather continue until the transcript sequence is exhausted. Additionally, mutations may result in the additional introduction of a stop codon causing the translation process to terminate prematurely and result in a shorter protein.

As in comparison with the reference peptidome some discrepancies could be found which may be due to exceptions in translation specific to certain proteins or due to some protein sequences not having been analyzed in full detail, a sanity check is introduced at this stage. Therefore, the reference raw exonic sequence is translated from the start codon onwards until the first stop codon is encountered. The so obtained protein is then compared with the reference proteome and if both sequences do not match, the protein is discarded from further analysis to ensure high confidence in the reported neoantigens.

After the validity of the unmutated sequence has been established in a next step mutations are introduced into the coding sequence and for each protein affected by at least one somatic mutation two versions of the resulting coding sequence are generated.

- The first version is meant to represent the protein as it would occur in a healthy cell and is as such generated from the coding sequence into which only germline mutations are introduced - hence forth referred to as the self protein.
- The second version, is indicative of what the resulting mutated protein within cancer cells would look like and as such is generated by introducing both somatic as well as germline mutations into the coding sequence - further referred to as non-self protein.

As each mutation has so far been analyzed in isolation and therefore affected positions both in the coding sequence as well as in the resulting proteins are denoted with respect to the reference coding sequence, mutations altering the length of the resulting protein such as indels or frameshifts, may invalidate this positional information. Therefore, for both coding sequences, mutations are introduced from back to front and in case of length altering changes, positional information of already included somatic mutations is adjusted accordingly as this information will later be used to link the mutated peptide sequence back to the mutations that it originated from.

The so obtained mutated coding sequences are then translated beginning at the start codon until the first stop codon is found. If no such stop codon is encountered, due to mutations affecting the same, the translation process is continued until the coding sequence is exhausted.

As the main goal is to identify peptides which are unique to the cancerous cells, in a first step peptides of lengths eight, nine, ten, and eleven which contain at least one position altered by a somatic mutation are generated from the non-self protein, and are compared against the set of all possible substrings of the respective lengths from the protein only containing germline mutations. Only peptide sequences absent from the healthy self protein are considered in further analysis as possible neo epitope candidates.

As the overall number of possible peptide sequences of the given lengths is finite, it is still possible that an amino acid originating from a mutated protein may be part of another non mutated healthy protein. Therefore, from all germline proteins, all possible amino acid sequences of lengths eight, nine, ten and eleven are generated and collected such that in the next stage potential neo peptide candidates can be checked against the full set of all self peptides and eliminated from further analysis if a match is found.

In case of the absence of somatic mutations for a given protein but the presence of germline mutations, only one version of the respective protein is generated and corresponding self peptide sequences are generated. In the absence of mutations altogether the reference protein is

used directly without requiring deliberate translation of the coding sequence and substrings are extracted directly.

Binding Affinity and Immunogenicity Evaluation

Having identified a list of peptide sequences that are created due to somatic mutations and will only be present in tumor cells, the next step lies in the evaluation of their potential to be used in a therapeutic setting.

A first step in this analysis is given by evaluating whether those peptide sequences will actually be presented as antigens to the immune system on top of MHC-I complexes - i.e. their binding affinity to these complexes needs to be evaluated.

For this purpose, throughout scientific literature **NetMHCpan** (Jurtz et al., 2017) has become the most widely applied program in terms of MHC-I binding affinity prediction and is therefore also included in this workflow to evaluate the binding affinity of each identified peptide with respect to the above identified particular HLA types.

While high binding affinity is a necessary factor for therapeutic application, it does not guarantee a neoantigen's ability to elicit tumor rejection. As no definitive process for immunogenicity evaluation exists at this point and feature choices and predictive accuracy varies widely across literature in this workflow a custom predictive model has been included which is described in more detail in Chapter 5.

Result Generation

In addition, another important factor when selecting peptides for vaccination therapy is whether the gene from which the mutated sequences originate is actually expressed in the cancerous cells, i.e. proteins are actually generated from the coding sequence. For this purpose the transcriptional information as generated by **Kallisto** - as mentioned in the previous section - is used. While some workflows introduce hard filtering at this stage and only further consider peptides if their underlying genes meet a specified TPM cutoff, no such hard cutoff is used in this approach and the information is merely reported in the overall output file, and for the user to discern whether their requirements are met.

The rationale for this approach is as denoted by Karasaki et al. (2017) the fact that data obtained from RNA sequencing does not necessarily indicate that the mutated mRNA is actually present in the cancer cell. As tumors are rarely pure, and hence contain genomic information from both cancerous and healthy cells, part of the TPM counts may be derived from healthy cells contained in the sample hence not giving any information on the tumor's transcriptome. The authors further indicate that TPM counting is not subject to variant phasing. As such, counts are generated from both chromosomes and the exact number of transcripts generated from the mutated position cannot be easily discerned.

At the end of this overall process, a file containing the mutated sequences alongside information on the mutations they originated from, as well as an evaluation of their binding affinities and potential to cause an immune response is returned.

A summary overview of all the features reported for each identified neoantigen is given in Table 4.1.

Parameter	Type	Description
Peptide	String	Amino acid sequence
		Variant
Variants	String	Comma separated list of mutations that are contained in the sequence. Each mutation is given in the form: **chr<Id>:<Position>:<Reference>:<Mutated>**
VariantTypes	String	Comma separated list variant types. Each can be one of: **SNV**, **Insertion**, **Deletion** or **Frameshift**
VariantCallers	String	Comma separated list of number of variant callers that confirmed the respective mutations.
DNA VAF Normal	String	Comma separated list of variant VAFs in normal DNA sample (Each averaged over reporting variant callers)
DNA VAF Tumor	String	Comma separated list of variant VAFs in tumor DNA sample (Each averaged over reporting variant callers)
RNA VAF	String	Comma separated list of variant VAFs in tumor RNA
Read Count DNA Normal	Float	Comma separated list of variant read counts in normal DNA sample (Each averaged over reporting variant callers)
Read Count DNA Tumor	Float	Comma separated list of variant read counts in tumor DNA sample (Each averaged over reporting variant callers)
Read Count RNA Tumor	Integer	Comma separated list of variant read counts in tumor RNA sample
		Protein
gene	String	Name of the gene the variants occurred in
Protein	String	Ensembl-ID of the mutated protein
		Binding Affinity
HLA	String	HLA type for which binding affinity is reported
BindingCore	String	9-mer directly bound to MHC complex
ICore	String	Interaction core
RawPredictionScore	Float	Raw binding affinity prediction score
Affinity (nM)	Float	Binding affinity reported in nano molar
%Rank	Float	Rank of binding affinity
Exp	Integer	NetMHCpan associated value
Strength	String	**WB** if %Rank < 2, **SB** if %Rank < 0.5, empty otherwise
		Transcription
TPM	Float	Transcripts per Million of the respective gene
expressed	String	**yes** if TPM $>$ minTpmCount - **no** otherwise
		Immunogenicity
ImmunogenicityScore	Float	Raw predicted immunogenicity score ranging from 0 to 1
Immunogenic	String	**yes** if ImmunogenicityScore > 0.5 - **no** otherwise

Table 4.1: Result values

4.4 Pipeline Execution

In order to allow for easy applicability and portability of the presented workflow, the **Cromwell** workflow engine (Voss et al., 2017) which has found a wide range of applications in the field of bioinformatics has been used. Using its workflow description language, each step of the process has been modelled as an individual task and a respective container image has been specified such that the pipeline is portable between systems and no additional software apart from the **Cromwell** execution engine as well as a respective container technology is required.

For its execution, apart from the workflow description file, a configuration file needs to be specified containing information on how the individual tasks should be executed, i.e. specifying, for example, commands for submitting the tasks to a HPC job scheduler. In addition, the respective input parameters are passed into the process in form of a json file. A full list of the required parameters is presented in Table 4.2. Note, that in the input file all parameters need to be specified as "**EpitopePrediction.**<Parameter>".

Parameter	Description
	Raw Sequences
seqNormal1	Forward DNA reads from normal sample (**.fastq.gz**)
seqNormal2	Reverse DNA reads from normal sample (**.fastq.gz**)
seqTumor1	Forward DNA reads from tumors sample (**.fastq.gz**)
seqTumor2	Reverse DNA reads from tumor sample (**.fastq.gz**)
seqRNA1	Forward RNA reads from tumor sample (**.fastq.gz**)
seqRNA2	Reverse RNA reads from tumor sample (**.fastq.gz**)
	References
seqAdapters	File containing adapter sequences introduced in the sequencing process (**.fa**)
cropLength	Maximum length to trim all raw sequences to.
refGenomeFolder	Path to folder containing the reference genome
refGenomeName	Filename of the reference genome used
refGenomeBWAFolder	Path to folder containing the reference genome files to be used by BWA
refGenomeBWAPrefix	Common name prefix for all files used by BWA
knownSitesFolder	Path to folder containing the three subsequent files
refSNP	Reference SNV file used as **known-sites** for BQSR
refGoldIndels	Reference indel file used as **known-sites** for BQSR
refKnownIndels	Reference indel file used as **known-sites** for BQSR
cdnaIndex	Index used by Kallisto for pseudo-alignment
gtf	File containing genomic positions (**.gtf**) (See http://www.ensembl.org/info/data/ftp/index.html)
proteins	File containing reference proteins (**.fa**) (See http://www.ensembl.org/info/data/ftp/index.html)
STARGenomeDir	Path to reference genome for usage by STAR
vepCacheDir	Path to cache used by VEP

Table 4.2: Pipeline Input Parameters

As is indicated by the above table, the analytical workflow was tailored to be executed on

sequencing data stemming from a paired end sequencing process. However, if usage for single end sequencing data is required, the commands for the programs specified to specifically run on paired end sequencing data can easily be adjusted in the workflow description file accordingly.

With respect to the generation of eptiope sequences from mutational information, a custom software solution has been generated and incorporated into the workflow. With the goal of providing easily adjustable filtering depending on the input data and the underlying application, several cutoff parameters have been exposed as program arguments. A full list of all input parameters also containing the respective cutoff parameters, which are used for the neo eptiope generation step are presented in Table 4.3.

Argument	Type	Default	Description
		Required Input Parameters	
--dna	String	–	Tumor DNA - aligned, sorted and indexed (**.bam**)
--rna	String	–	Tumor RNA - aligned, sorted and indexed (**.bam**)
--somatic	List	–	Space separated list of vcf files containing somatic variants (**.vcf** | **.vcf.gz**)
--vc	List	–	Space separated list of name of variant caller for each vcf file given with **--somatic**. Each can be any of: **Mutect2**, **Strelka**, **Varscan**
--germline	String	–	File containing variants called by GATK Haplotypecaller (**.vcf** | **.vcf.gz**)
--proteins	String	–	Ensembl reference protein file (**.fa**)
--gtf	String	–	File containing genomic positions (**.gtf**)
--ref	String	–	Reference Genome (**.fa**)
--hla	String	–	Output file derived from Optitype
--tpm	String	–	Output file derived from Kallisto (**.tsv**)
--phaser	String	–	Path to **phaser.py**
--vepCacheDir	String	–	Path to cache for Variant Effect Predictor
		Optional Filter Parameters	
--minVcCount	Integer	2	Minimum number of variant callers required to confirm a mutation
--maxNormalVaf	Float	0.05	Maximum VAF for variant in normal sample
--minTumorVaf	Float	0.05	Minimum VAF for variant in tumor sample
--minRcAbs	Integer	1	Minimum absolute number of reads required to confirm a variant
--minRcRel	Integer	0.2	Minimum fraction of reads required to confirm a variant
--minTpmCount	Float	1.0	Minimum TPM for a gene to be considered expressed
--dnaMapQ	Integer	1	Minimum DNA mapping quality for phaser
--rnaMapQ	Integer	255	Minimum RNA mapping quality for phaser
		Flags	
--help, -h	–	–	Produce help message
--dirty	–	–	Keep temporary files

Table 4.3: Epitope Prediction Program Parameters

While not being directly specified as input parameters, the phasing process requires a corresponding index file for both tumor DNA and tumor RNA sequences to be present in the same folder. Both index files need to be named after the corresponding sequence files, followed by the extension **.bai**.

It should be noted that during the process of filtering identified neoantigen sequences with respect to the generated self peptides, a full set of self peptides stemming from germline mutations is generated as a temporary file. While those temporary files will be deleted at the end of the analysis, through setting of the **--dirty** flag, those files can be retained and potentially used for further analysis such as the evaluation of similarity between predicted neoantigens and the whole set of self peptides.

Chapter 5

Immunogenicity Prediction

5.1 Motivation

As mentioned in previous sections, a peptide's ability to be presented by molecules of the MHC-I is a necessary yet insufficient requirement to induce tumor rejection. In fact, experimental evidence presented by Kristensen (2017) show that only a very small percentage of about 1% of in silico predicted peptide-MHC complexes is able to elicit an immune response, i.e. be immunogenic and can hence be selected for therapeutic intervention. As a reasoning for this limited amount of immunogenic neo peptides Koşaloğlu-Yalçın et al. (2018) argue that mutation-derived neoantigens are in contrast to entirely foreign peptide sequences, such as the ones derived from viruses, in general highly similar to their non mutated versions which makes any potential T-cell Receptor reacting to these neoantigens subject to central tolerance (see Chapter 3)

In contrast to MHC binding, factors that allow peptides to be successfully recognized by the host's immune system still remain vastly unclear and throughout scientific literature many competing views and models are found resulting in a lack of standardization in terms of immunogenicity predictions. Due to the challenging nature of in silico immunogenicity prediction, the main focus of most neoantigen prediction approaches lies with identifying antigens that will likely be presented on top a cell's surface rather than whether they are actually able to cause an immune response (Smith et al., 2019a).

A major limiting factor in terms of immunogenicity modelling is the lack of availability of experimentally validated data. As such, Koşaloğlu-Yalçın et al. (2018) find that many immunogenicity studies only provide a very limited view on the overall contributing factors, as their analyses are often constrained to peptides of specific length - such as the abundant 9-mer peptides - or only focus their analyses on the most frequent types of MHC-I complexes.

Due to this lack of uniformity in the area of antigen immunogenicity prediction, this chapter aims at presenting an overview of the most widely used factors and also, based on data collected from various scientific studies, provide an empirical evaluation of different approaches using two commonly employed machine learning techniques - Random Forests (RFs) and Kernel Support Vector Machines (KSVMs).

5.2 State of the Art

While the overall landscape of immunogenicity predictors differs vastly among publications, a brief overview of the most commonly employed measures will be given in this chapter.

MHC-I Binding Affinity

As the evaluation of the binding strength of potential neo epitope candidates to the MHC-I molecules is an integral part of most neo epitope prediction pipelines, this measure is often a common starting ground for assessment of an antigen's ability to provoke an immune response.

Studies like Koşaloğlu-Yalçın et al. (2018) argue that the predictability of an antigen's immunogenicity is strongly determined by the strength with which it binds to the MHC-I complex, finding that strong binders also are the ones most likely to be recognized by the immune system. Such results, however, are not unquestioned, and the authors themselves indicate that other studies exist claiming that the overall correlation should be negative, meaning that strongly binding peptides would have only a very low chance of being immunogenic. The argumentation lies with the fact that strongly binding antigens would be presented during thymic development and any T-cell that would recognize this strongly binding peptide would be destroyed due to the mechanisms of central tolerance.

Despite the popularity of the binding affinity as a measure of immunogenicity, the predictive use of binding affinity does not go unquestioned regardless of the direction of the correlation. Duan et al. (2014) find in their evaluation that the predictive ability of binding affinity scores is overall very poor.

It should be noted at this point, however, that, as outlined above, a major limiting factor in immunogenicity in general - and assessment of the predictive capability of binding strength scores, in particular - is the availability of experimentally evaluated data. Neo epitope data, in particular sequences that have been shown to provoke an immune response, are typically derived from vaccination studies. As MHC-I binding is a vital factor when choosing potential vaccination candidates, positively tested sequences from vaccination studies present a possible source for biasedness towards high immunogenicity of strongly binding peptides.

Differential Agretopicity Index

An extension to using the binding strength of a mutated peptide sequence is provided by Duan et al. (2014) introducing the Differential Agretopicity Index (DAI) - agretopicity referring to a molecules MHC binding affinity. As its name implies the DAI can be computed by taking the difference in binding affinity scores between a given peptide and its unmutated counterpart also referred to as a peptide's wild type.

The rationale behind this measure is given by the assumption of high sequence similarity between a mutated peptide and its unmutated wild type as well as T-cell Receptor's polyspecificity. As argued by Duan et al. (2014), during thymic development of T-cells, the wild type peptide will be presented during negative selection if it shows a strong binding affinity. Therefore, due to the mechanisms of central tolerance, any Cytotoxic T Lymphocyte (CTL) reactive to this peptide, and by means of polyspecificity also to the corresponding mutated sequence, would consequently be destroyed.

As such, the authors propose that if mutations cause low affinity wild type peptides to bind more strongly to the MHC complex and would, therefore, be presented by cancerous cells, CTLs reactive to this antigen would likely not have been subject to central tolerance, thus, providing a potential target for immunotherapy.

Even though the proposing authors of this measure find statistical significance of the DAI and argue that it provides better results than using MHC-I binding affinity alone, they indicate that most sequences identified as immunogenic would upon clinical evaluation still fail to elicit immune responses. Other studies such as Koşaloğlu-Yalçın et al. (2018), however, dispute the usefulness of the Differential Agretopicity Index as a predictive measure. The authors find that in their dataset binding strength alone was able to provide better results than the proposed DAI. Similar results were also reported by Bjerregaard et al. (2017b) stating that MHC-I binding affinity alone was able to produce more accurate predictions than DAI, when both were considered independently.

Another major limiting factor in the application of the Differential Agretopicity Index is that evaluation of any peptide requires also the identification of its unmutated counterpart. And while many studies focus on the evaluation of SNV type mutations for which wild type peptides are straight forward to compute, application to more complex mutations such as insertions, deletions and frameshift mutations becomes more challenging as for these types the wild type is not clearly defined.

Antigen Similarity

A different approach based on the same rationale is to consider the similarity of the respective neoantigen not only to its wild type sequence but the overall set of self peptides that can be generated by healthy cells. As such, any neo peptide too similar to the set of self peptides would not be recognizable by the immune system due to self tolerance.

Even when restricting the computation of the similarity of a mutated antigen to only its wild type, Bjerregaard et al. (2017b) report better prediction results especially in cases where both sequences, normal and mutated, displayed similar levels of MHC-I binding affinity.

A slight variation of this approach was taken by Kim et al. (2018), computing not the similarity to existent self peptides but the relatedness of the mutated sequence to known pathogenic epitopes which were already known to induce an immune response.

Physiochemical Properties

While the specific features related to the immunogenicity of a particular peptide are still not fully understood and many conflicting views exist, Wang et al. (2019) note that throughout scientific publications a general consensus exists which recognizes the importance of an antigen's foreignness, its accessibility to potential investigation by CTLs and additional chemical properties of the peptide such as its weight and structure.

The 20 different amino acids which represent the core building blocks of proteins, albeit being similar in their molecular constitution, display distinguishing chemical properties depending on their side chain residues. These differences result in the peptide formed by a sequence of amino acids displaying distinguishable chemical properties depending on the set and sequence of amino acids they are made of. These chemical properties can for example be assessed by taking the sequence as a whole into account using subsequences of the investigated peptide or even by assessing the amino acid at each position individually.

The common chemical structure shared by all 20 amino acids which are used for building proteins is depicted in Figure 5.1 and consists of a central carbon atom to which a basic amino

group (–NH_2), an acidic carboxyl group (–COOH) and a hydrogen atom (H) are attached. The fourth group attached to this central carbon atom consists of an organic residue (R) which is specific to each of the 20 amino acids (Berg et al., 2012).

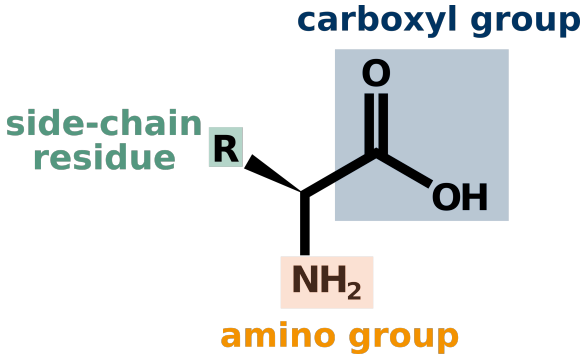

Figure 5.1: Amino Acid Structure

Due to these variations in side chain residues, different amino acids can be attributed chemical characteristics to variant degrees.

One such chemical feature frequently identified as playing a major role in discerning an antigen's immunogenicity is a peptide's hydrophobicity. In investigating peptide sequences as a whole, Teku and Vihinen (2018) observe that immunogenic neo epitopes tend to contain more hydrophobic amino acids than unmutated wild type peptides. Similar findings are also reported by Chowell et al. (2015), from which a neural network - ANN-Hydro - trained on a dataset of 9-mer peptides restricted to HLA-A2, was constructed, discriminating between immunogenic and non-immunogenic peptides only taking a peptide's hydrophobicity value into account.

Another study outlining the importance of antigen hydrophobicity is presented by Chowell et al. (2015). Analyzing intracellular pathogens like viruses, the authors report that the resulting coding sequence for viral antigens were found to be much lower in G+C content than self peptides. As argued by Khrustalev and Barkovsky (2011), this less frequent occurrence of G+C codons results in the incorporation of more hydrophobic amino acids into the protein sequence, making the overall antigen more hydrophobic in contrast to self peptides.

Additional observations on chemical properties of immunogenic antigens conducted by Chowell et al. (2015) reveal that highly polar amino acids were encountered less frequently in immunogenic peptides. While the authors themselves do not report any significant correlation with the bulkiness of the side chain residue, Koşaloğlu-Yalçın et al. (2018) argue that an immune response is more likely caused when the presented antigen contains large and aromatic amino acids in particular at positions which face the TCR.

Positional Information

Chemical properties of antigenic peptides are often times measured as an aggregated score across the whole of the peptide. While finding frequent use, this approach fails to include how individual chemical properties are distributed along the peptide's sequence or how important the positioning of a particular amino acid in the antigen's sequence is for it to be able to trigger an immune response. In fact, analyses conducted by Smith et al. (2019a) find the positioning of individual amino acids as well as mutation induced changes in specific positions of a peptide's

sequence to be among the most predictive features in terms of immunogenicity.

Analogous reasoning in the field of predicting MHC-I binding strength is pointed out by Frankild et al. (2008). Investigating the HLA-A0201 molecule, the authors find that for binding affinity prediction certain positions in the sequence harboured more importance than others. In particular, the authors state that characteristics of amino acids contained at positions two and nine were the most important in terms of MHC binding affinity, while at the same time carrying little to no predictive information on whether an immune response will be elicited. In accordance with their findings, the authors also state that throughout the scientific literature a consensus is apparent with respect to the influence of certain positions in a peptide's sequence. While amino acids at the edge are seen as more important in terms of MHC binding, amino acids towards the center of an antigen appear to bear the most importance with respect to TCR interaction.

A further study investigating the importance of specific positions in peptides is presented by Koşaloğlu-Yalçın et al. (2018) who were able to report correlations between the types of amino acids that were included at positions that were presumed to be most likely in contact with binding TCRs.

An examination of the importance of different chemical properties at specific positions is provided by Calis et al. (2013) who in concordance with Frankild et al. (2008) identify the most information on immunogenicity to be contained in the central amino acids. In particular, the authors identify that the presence of amino acids with large and aromatic side chains at positions four,five and six is a reasonable indicator of a peptide's immunogenicity. A case for hydrophobicity in certain positions is made by Riley et al. (2019), reasoning that hydrophobic amino acids in positions that are exposed to a TCR facilitate binding through the hydrophobic effect. As stated by the authors, binding of hydrophobic amino acids only requires another hydrophobic environment on the binding TCR and is therefore more favorable than the binding between charged amino acids.

Similar findings are also reported by Brown and Holt (2019) also reporting that positions referred to as anchor adjacent, such as position eight, are of little importance to a peptide's immunogenicity and respective changes would do little to influence a peptide's immunogenic capabilities.

It should be noted, however, that not all studies find that the preference for specific amino acids or amino acids with particular chemical properties is position dependent. Teku and Vihinen (2018) report that irrespective of their position in the peptide, certain amino acids appear to be overall preferred and more likely to invoke an immune response. Accordingly, Wang et al. (2019) find that immunogenic epitopes display an overall preference for the amino acid **Leucine**, while the occurrence of **Tryptophan**, **Histidine** and **Cysteine** in a peptide's sequence is more common in non-immunogenic antigens.

It therefore appears that at least some of the information on whether a peptide will be immunogenic is contained in specific positions of the peptide's sequence, whether the information can be captured by physiochemical factors such as hydrophobicity or the preference lies in the inclusion of specific amino acids themselves.

TCR-Peptide Contact Potential

A drastically different approach to the above mentioned properties was taken by Ogishi and Yotsuyanagi (2019) investigating the concept that immunogenicity is determined by thermodynamic principles. Instead of using a peptide's physiochemical properties in isolation, interactions between antigens and corresponding TCRs have been investigated.

Computationally, this was achieved by extending the pairwise sequence alignment approach by means of a custom substitution matrix specifically tailored to reflect amino acid interactions. The authors argue that in using this approach, the final alignment scores would be indicative of the intramolecular energy potentials between an antigen and the TCR.

Using information from a publicly available TCR database, tested peptides are aligned to all TCRs and the receptor representing the best match would be indicative of the antigen's ability to trigger an immune response.

While as mentioned in Chapter 3, creation of the T-cell Receptors is governed by random gene arrangements, the authors argue that due to the selective nature during thymic development of Cytotoxic T Lymphocites, the total set of recognizable antigens follows similar patterns for every person and therefore displays large similarities among individuals. Following this argumentation, the authors employed the use of a public non-patient-specific TCR database as a reference.

5.3 Experimental Setup

5.3.1 Dataset

As mentioned above, most analytical frameworks stop after evaluating a peptide's MHC-I binding affinity and implicitly assume that higher binding affinity indicates better suitable candidates and no additional evaluation of a peptide's ability to elicit tumor rejection is performed. Furthermore, throughout scientific literature, the decisive factors that indicate whether an antigen is able to provoke the host's immune system are still heavily debated and many conflicting results are found.

Another potential hindrance when aiming to include immunogenicity prediction is encountered by the availability of the training data or the distribution of the same. While many different papers report their model to be able to achieve an Area under the Curve (AUC) of close to 1, upon closer inspection, those models are often times built on skewed datasets. In various of the examined datasets immunogenic peptides are heavily outnumbered by non-immunogenic ones and as a result only make up a very small fraction of the overall dataset. Secondly, datasets from which predictive models are generated were found to be often times very limited in their size making the generalizability of such models highly doubtable. Additionally, often times models are trained on a very limited subset of potential neo epitope candidates such as when restricting the analysis to 9-mer peptides or only focusing on specific HLA types.

Considering these shortcomings, many of the examined models were, due to their limitations, deemed unfit for inclusion in this analytical workflow. Another hindrance encountered was the usage of a generated antigen's unmutated wild type sequence in some models as a predictor for its immunogenicity which is not unambiguously defined for antigenic sequences arising from

mutations that alter the length of the generated protein. As a central goal in the development of the here presented approach was to allow for the inclusion of a variety of mutations, models relying on such unmutated peptide sequences could not be confidently included.

As not to impose any restrictions in terms of a peptide's length or the type of mutation it arises from, a custom model to predict an identified antigen's immunogenicty is generated and included in to the presented workflow. For this purpose data on empirically tested antigens was accumulated from overall 19 sources, containing information on a peptide's sequence as well as a binary indicator whether the peptide was proven to be immunogenic or not. This collected dataset will further present the foundation for machine learning approaches to model a peptide's ability to trigger an immune response.

A comprehensive list of the distribution of immunogenic and non-immunogenic peptides in the raw dataset - irrespective of the HLA types they have been tested in conjunction with - as well as the sources they were collected from - is presented in Table 5.1. The respective counts indicate the number of unique peptide sequences with corresponding immunogenicity indication per source.

Source	Immunogenicity			Reference
	Positive	Negative	Total	
Bassani-Sternberg	2	5	7	Bassani-Sternberg et al. (2016)
Bentzen	9	627	636	Bentzen et al. (2016)
Calis	1017	0	1017	Calis et al. (2013)
Chowell	2912	3813	6725	Chowell et al. (2015)
Cohen	9	418	427	Cohen et al. (2015)
EPIMHC	761	868	1629	Reche et al. (2005)
HCV	88	0	88	Kuiken et al. (2005)
HIV	206	0	206	Llano et al. (2013)
IEDB	4052	8231	12283	Fleri et al. (2017)
IMMA2	557	405	962	Tung et al. (2011)
Lu	2	8	10	Lu et al. (2014)
MHCBN	1151	354	1505	Lata et al. (2009)
McGrannahan	11	980	991	McGranahan et al. (2016)
Ott	17	144	161	Ott et al. (2017)
Rajasagi	3	39	42	Rajasagi et al. (2014)
Robbins	9	216	225	Robbins et al. (2013)
Stronen	6	1049	1055	Strønen et al. (2016)
TANTIGEN	360	0	360	Olsen et al. (2017)
Wick	1	108	109	Wick et al. (2014)

Table 5.1: Data Sources

Disregarding the various HLA types from the raw data has been a necessary step, as immunogenicity of a given peptide has been found not to vary with respect to the different HLA types it has been tested with. As a result, expansion of the individual rows for every corresponding HLA type would lead to the creation of many duplicate peptide-immunogenicity pairs resulting in biased estimates for the overall goodness of a model's performance.

Data extracted from the above sources has further been combined to eliminate duplicate entries

resulting in a dataset of overall **21,148** unique peptide- immunogenicity pairs, of which **5,288** ($\approx \frac{1}{4}$) were tested positively in terms of immunogenicity and **15,860** ($\approx \frac{3}{4}$) were not found to be immunogenic, respectively.

Overall, it can be seen that not all lengths are distributed evenly across the dataset. As commonly observed in immunogenicity studies, peptides of length nine are encountered more frequently than peptides of any other length. This is to be expected as 9-mer peptides are throughout various studies among the most studied length of antigenic peptides due to their tendency of being the most strong binding peptides as indicated by Teku and Vihinen (2018), and as stated by Bjerregaard et al. (2017b) show the highest promise in terms of invoking an immune response. Conversely, amounts of peptides of length eight and eleven are the lowest in the whole dataset.

A graphical representation of the distribution of peptide lengths and their immunogenicity is given in Figure 5.2 and the corresponding absolute counts are presented in Table 5.2.

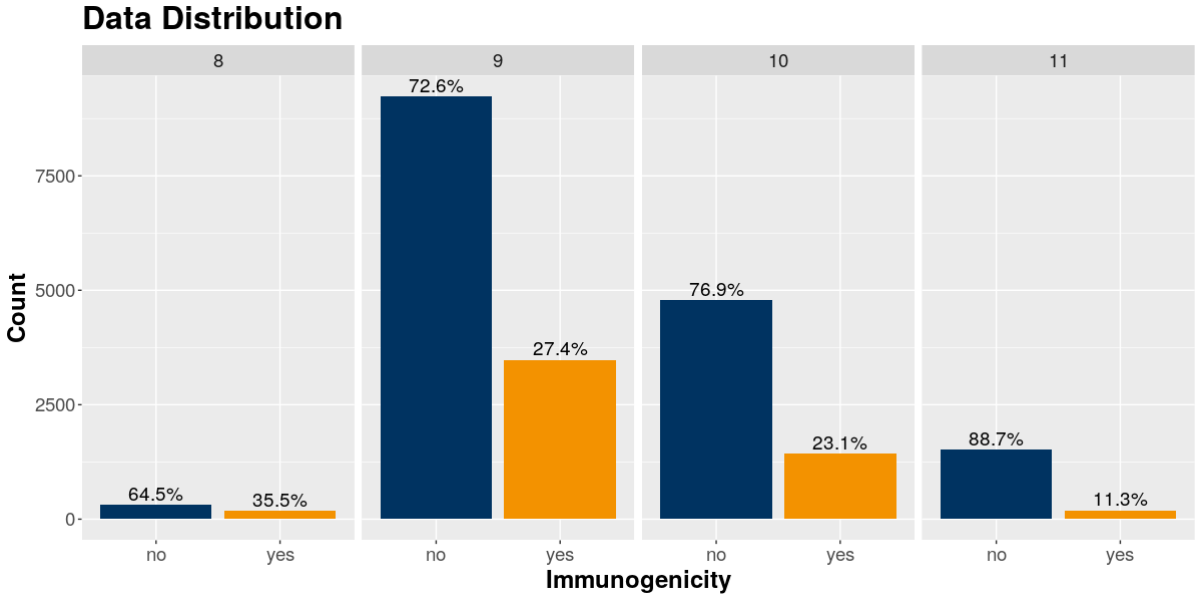

Figure 5.2: Data Distribution

Immunogenicity \ Length	8	9	10	11	Sum
no	321	9229	4785	1525	15860
yes	177	3481	1435	195	5288
Sum	498	12710	6220	1720	21148

Table 5.2: Data Set Distribution

It should be noted at this point that as is the case in many immunogenicity studies, the underlying data exhibits a certain bias towards strong and weakly binding peptides as opposed to sequences presenting no reasonable binding strength. When evaluating the binding affinities of the sequences in the dataset with respect to all the HLA types they have been tested with this biasedness towards strong and weak binders becomes easily apparent.

With binding affinity reported as percentage rank by **NetMHCpan** (Jurtz et al., 2017), where lower values indicate higher binding affinity, the authors classify peptides with a rank value of less than **2** as a weak binder (WB), values even below **0.5** to be indicative of strong binders (SB). Any value above **2** is considered to exhibit insignificant binding affinity, i.e. be a non-binder (NB).

A graphical overview of the distribution of the peptides' binding affinities is given in Figure 5.3. The left plot in this graphic gives an indication of the empirically derived Cumulative Distribution Function (CDF) as estimated from the provided data. Each value on the y-axis in this plot is indicative of the fraction of data points which present a %Rank value less or equal to the corresponding value on the x-axis. As the line presented in the left graphic is highly skewed to the left, it becomes easily apparent that a large fraction of the contained peptides exhibits a very low %Rank value and as such the dataset is skewed towards higher binding affinity.

The second plot provides an overview of the distribution of the above mentioned binding affinity classes. As can be seen from this barplot, roughly 70% of the peptide sequences are considered to exhibit at least weak binding affinity.

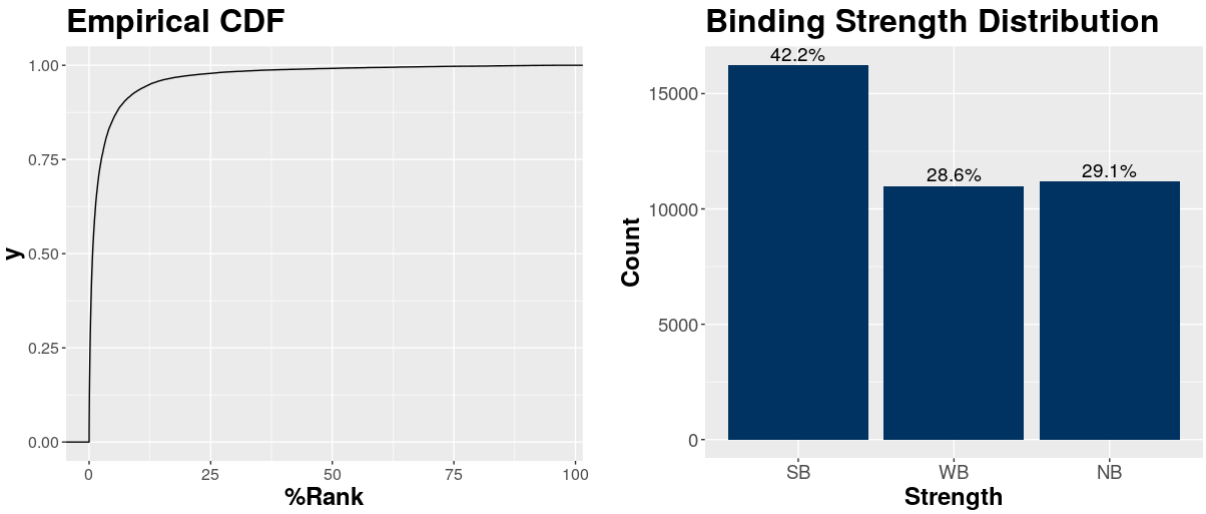

Figure 5.3: Distribution of Binding Affinity

While this biasedness in the dataset should be taken into consideration when trying to derive universal statements with respect to immunogenicity, it can be argued that due to the nature of the underlying application of this analytical pipeline, immunogenicity predictions are only ever taken into consideration in conjunction with a peptide's ability to bind to an MHC-I complex as it presents a necessary requirement and antigens which immunogenic but not presented would serve little purpose in terms of therapeutic application.

As a result, the sole source of information used to discern a peptide's potential to invoke tumor rejection is assumed to be a peptide's sequence of amino acids and in the subsequent section two types of models will be explored using various encoding and measuring approaches - Random Forest (RF) and Kernel Support Vector Machine (KSVM).

5.3.2 Random Forest

The first type of model that will be used in this thesis is Random Forests (Breiman, 2001).

As their core building block random forests are made up of an ensemble of decision trees. Each of these decision trees is able to assign a label to a data point by starting at the root node and depending on the predictor values for a given data point following a specific path down the tree until a leaf node, containing a label, is reached. Each internal node along this way can be thought of as asking a yes or no question concerning the data point, and depending on the answer the path down the tree is continued with the left or right child of the internal node, respectively.

While such decision trees are easy to construct, interpret and to apply to predict new data, a commonly encountered problem with their usage is that they fall short in terms of accuracy. Despite being able to label training data that is used to infer the model parameters with high confidence, labeling of new data often incur heavy predictive errors (Friedman et al., 2001).

In order to mitigate these shortcomings a popular extension for a single decision tree is given by Random Forests (Friedman et al., 2001). When creating a random forest from a dataset, each tree is created from a bootstrapped subset of the original training data. To create this subset, the total dataset is sampled with replacement, meaning that observations in the original dataset will be included into the bootstrapped sample, once, multiple times, or not at all. When creating a decision tree for the given dataset, at each split point a random subset of predictors is examined and the best split point is tested for (see Figure 5.4).

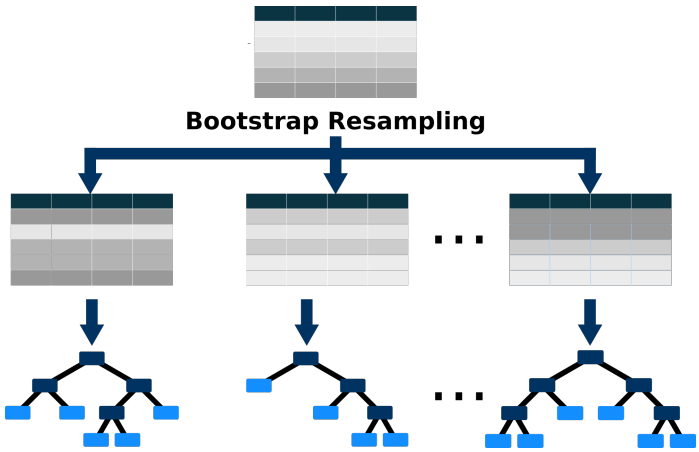

Figure 5.4: Random Forest Model Estimation

When using a random forest for prediction, every decision tree is evaluated independently of one another for the given data point based on its model parameters inferred from the respective bootstrapped dataset, and overall results obtained from all trees are aggregated to form a single prediction. Depending on the application, either the class making up the majority of the predicted labels is used, or in order to allow for the prediction of probabilities the fraction of the number of trees predicting a specific class over the total number of trees tested can be used (see Figure 5.5).

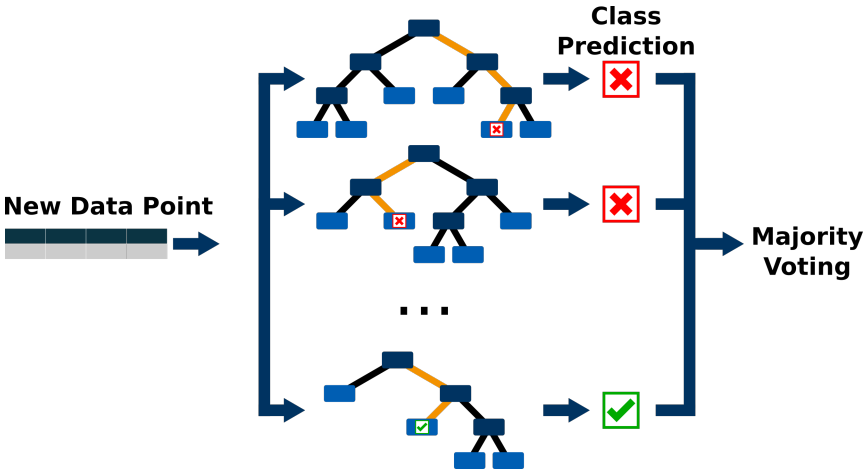

Figure 5.5: Random Forest Prediction

Encodings

In order to be able to estimate models such as Random Forests from epitope data, a crucial part is choosing an appropriate encoding of individual peptides. Ideally, as identified by Zamani and Kremer (2011), an encoding of the data should aim at preserving information about the underlying data which is important for solving the problem at hand while at the same time removing unnecessary background noise in the data.

As potential MHC-I binding antigens are made up of sequences that may be between 8 and 11 characters long, Chou (2009) point out two overall possible encoding categories which can be referred to as sequential representations or discrete ones. As the here examined Random Forest type models are not equipped to handle variable length input data out of the box, only discrete encodings are considered for further analysis.

When trying to encode differing length amino acid sequences into fixed length discrete representations, depending on the granularity of the encoding, two possible strategies can be used.

1. Extract a fixed set of factors from the peptide as a whole
2. Pad or trim all peptides to a common length and encode each amino acid individually

Both of which will be examined in this thesis.

In terms of extracting a fixed length set of features, one of the simplest encoding schemes is provided by Amino Acid Composition (AAC) (Xiao et al., 2015). Using this encoding, any arbitrary length amino acid sequence can be represented as a 20 dimensional feature vector. Every entry of the so constructed vector corresponds the amount the respective amino acid is contained within the overall sequence. In this approach it is possible to use absolute numbers or relative frequencies.

Expansions of this basic approach can be created by instead of counting occurrence of particular amino acids, grouping amino acids together by means of their chemical properties and counting occurrences of each of the so constructed groups. As in this thesis physiochemical properties are modelled using different approaches, the vanilla AAC encoding scheme is employed using relative frequencies.

While presenting a very simple and straight forward encoding, a major limitation of this approach is, as implicated by Chou (2009), that any information relating to a peptide's overall sequence such as the positions of the respective amino acid is completely lost. Therefore, an extension to this basic approach was provided in Chou and Elrod (1999), introducing Pseudo Amino Acid Compositon (PseAAC). PseAAC allows for the encoding of variable length sequences as fixed length vectors, while still retaining some of the information contained in the sequence of amino acids.

As to not completely use sequence information, PseAAC presents a discrete encoding for any length protein sequences which results in a feature vector of more than 20 dimensions. While the first 20 features of this encoding are representative of a peptides amino acids, much like in vanilla AAC encoding, the additional factors are reflective of correlations between contiguous amino acids. In specifying an additional parameter λ which has to be strictly smaller than the length of the protein's sequence, correlation information of all amino acids being up to λ positions apart in the sequence is incorporated, where correlation is measured in terms of their chemical properties, resulting in an overall dimensionality of the encoding schemes of $20 + \lambda$ dimensions. A detailed explanation of this encoding scheme can be found in Chou (2009).

In addition to PseAAC, in this analysis also Amphiphilic Pseudo-Amino Acid Composition (APseAAC) (Chou, 2005) and Quasi Sequence Ordering (QSO) (Chou, 2000) will be evaluated. While all three approaches are based on the idea of capturing sequence information by comparing amino acids which are up to maximum distance of λ positions apart, they differ in terms of how correlation among individual amino acids is assessed as well as the way in which these correlation scores are aggregated over the peptide's sequence and eventually reported. An in depth explanation of APseAAC and QSO can be found in Chou (2005) and Chou (2000), respectively.

As all three of these encoding approaches are similar in that they compute distance measures between amino acids a certain amount of positions apart, all three approaches require a λ factor to be set. As the examined dataset contains peptide sequences of lengths eight, nine, ten and eleven, respectively, λ has been set to 7, for all these encoding schemes in the subsequent analysis.

The second type of encodings used in this thesis is directly conducted for each amino acid individually. A big advantage of this granular approach is that for every amino acid positional information can be retained as individual positions may be of varying importance as outlined above.

The easiest form of encoding individual amino acids which is also frequently used in machine learning approaches is to use an orthogonal or one-hot encoding Baldi et al. (2001). In this encoding each amino acid is represented as a 20 dimensional vector in which all positions are set to 0 apart from a singular 1, the position of which depends on the particular amino acid being encoded.

While being straight forward, Maetschke et al. (2005) point out that one hot encoding implicitly assumes that all amino acids are equally similar or dissimilar to one another, as euclidian distance between two one hot encoded vectors always equates to 2, regardless of the particular amino acids under comparison. Therefore the authors proposed another form of encoding called **BLOMAP**. Based on the BLOSUM62 matrix, the authors provide a 5 dimensional feature vector for each amino acid, which as argued by the authors, is capable of capturing fundamental

similarities in terms of chemical properties between individual amino acids.

While providing a very dense representation of physiochemical similarities between amino acids, the last type of encoding explored for Random Forest models in this thesis is to encode each amino acid in directly terms of their numerical values with respect to various physiochemical properties. To alleviate this process, the amino acid index (Kawashima et al., 1999) presents a central database in which over 500 different chemical properties are recorded for each of the 20 amino acids.

As the amino acid index database provides a large number of chemical features, a crucial part when encoding amino acid sequences in terms of their physiochemical properties is the selection of which specific features to use. The features used in this thesis were chosen with respect to the approach taken by Chowell et al. (2015). The authors used the amino acid's hydrophobicity measured on the Kyte-Doolittle scale (Kyte and Doolittle, 1982), the polarity as introduced by Grantham (1974) and the bulkiness of an amino acid's side chain as measured by Zimmerman et al. (1968). While it is also possible to specify such chemical properties for both PseAAC and APseAAC encodings, in this thesis their respective default values have been used.

A summary overview of all the encodings used to create random forest models is provided in Table 5.3 where the column "Function" is indicative of which package and which function respectively has been used to extract the encoding. In case of **BLOMAP** encoding, the features have been taken from the original paper and a custom mapping function was generated.

Encoding	Granularity	Dimensions	Function	Parameters
AAC	Peptide	20	protr:::extractAAC	–
PseAAC	Peptide	27	protr:::extractPAAC	$\lambda = 7$
APseAAC	Peptide	34	protr:::extractAPAAC	$\lambda = 7$
QSO	Peptide	54	protr:::extractQSO	$\lambda = 7$
Factor (one-hot)	Amino Acid	11	–	–
Hydro	Amino Acid	33	bio3d:::aa2index	KYTJ820101 ZIMJ680102 GRAR740102
BLOMAP	Amino Acid	55	–	–

Table 5.3: Peptide Encodings

Note that as the statistical software R has been used for all subsequent analyses factor encoding is represented as 11 dimensional. However, as mentioned above, this encoded in the background by means of a one hot vector. Additionally, as peptides shorter than 11 amino acids long have been padded to length 11 with dummy amino acids, while positions 1-8 are represented in R as factor variables with 20 possible levels, the residual positions are encoded as factors with 21 levels - 20 standard amino acids plus a dummy amino acid **X**. For both other amino acid based encodings the corresponding values for this dummy amino acid have been set to 0.

5.3.3 Kernel Support Vector Machines

While the rationale for using a random forest model lies with identifying underlying sequence characteristics that would be indicative of whether an antigen was able to trigger the host's immune system, the second type of model builds on the notion of exploiting the concept of

polyspecificity of T-cell Receptors.

As mentioned in Chapter 3, any CTL reactive to a peptide sequence which is considered too similar to self peptides would during thymic development be deleted. As such a straight forward approach is to consider sequence similarity with peptides which have provably failed to elicit an immune response as a negative predictor, while close relation to already known immunogenic peptides would be indicative of a high likelihood of immunogenicity.

A common machine learning model used for such classification tasks are Support Vector Machines (SVMs) (Vapnik, 1995). As noted by Hofmann (2006), SVMs are constructed simply enough to be analyzable by mathematical means while at the same time containing the potential to fit very complex models as are often required in real world applications.

The objective of such a Support Vector Machine is to find a hyperplane which separates data into the given classes with as much confidence as possible. As such, one major limitation by such a classical SVM approach is that only data which is linearly separable can be modelled accurately.

In light of this limitation, an extension of vanilla SVMs is given by Kernel Support Vector Machines (KSVMs). By defining a mapping function, data which is originally not linearly separable in its input space, can be mapped into a higher dimensional representation, in the space of which construction of such a separating plane may be possible (see Figure 5.6).

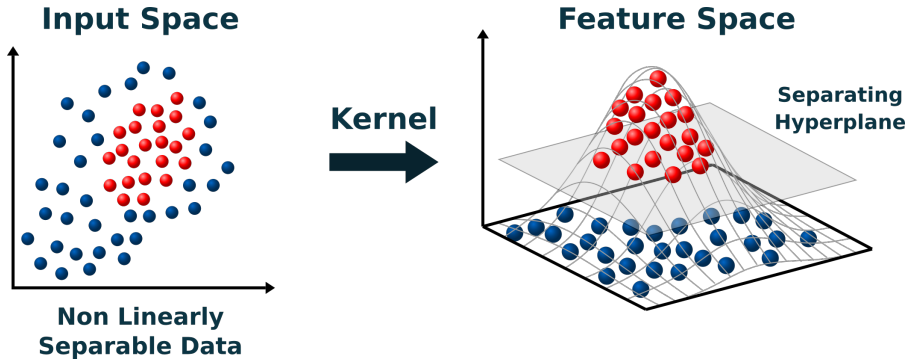

Figure 5.6: Kernel Support Vector Machine

However, in particular when discriminating classes of amino acid sequences which may be of different lengths, already mapping these peptides to individual data points is not straight forward, let alone finding an even higher dimensional feature space in which those sequences can be easily discriminated.

In using the **Kernel Trick** as outlined by Hofmann (2006), however, explicit computation of such higher dimensional representations can be avoided and a predictive model can be fit to the data points regardless. As the author states, the optimization function used to infer the hyperplane's parameters, does not rely on the coordinates - be it input space or feature space - of the data points directly. It only does so by means of the dot product which needs to be defined for the data the model is to be fit on. Therefore, when such a dot product function, i.e. kernel function, is defined, and can therefore when applied to a pair of data points return a numerical value indicative of their overall similarity to one another, explicit computing a vector representation of the given data points can be avoided.

A crucial step when creating a KSVM based model, therefore, lies with the identification of such a similarity measure capable of extracting similarities with respect to the classification task. While Kernel Support Vector Machines are frequently used in the field of bioinformatics and are even frequently encountered in immunogenicity prediction studies, most notably, vanilla SVMs or RBF kernels are used. However, due to the kernel trick mentioned above, KSVM models can be created using custom similarity measures tailored specifically to the classification problem at hand.

A common starting point for similarity measures between protein sequences is, as indicated by Zamani and Kremer (2011) to employ the usage of substitution matrices such as Point Accepted Mutation (PAM) (Dayhoff, 1972) or Blocks Substitution Matrix (BLOSUM) (Henikoff and Henikoff, 1992). As indicated by the authors, such matrices are contain scores of substitutability of amino acids and are derived from biological measurements hence finding frequent use for protein alignment. Similarly, Zamani and Kremer (2011) indicate that substitution matrices provide differences between individual amino acids in a natural and meaningful way. As such, to provide a meaningful similarity measure between individual peptide sequences Blosum matrices will be used in this model.

It should be noted at this point, however, that not just a singular BLOSUM matrix exists, but, rather a set of individual BLOSUM matrices exist which are each indicated by a particular number. As presented by Henikoff and Henikoff (1992), during construction of a BLOSUM matrix sequences which are too similar to one another are clustered together such as to not skew the substitution frequencies by too many highly similar sequences. As such, a threshold needs to be set which, if two sequences are equal to each other by at least that amount of amino acids, they are clustered together. This percentage threshold is then reflected in the name of the BLOSUM matrix.

The means by which these substitution matrices can be incorporated in a similarity measure is by applying the Smith-Waterman algorithm as presented in Smith et al. (1981). In this thesis the local alignment method is used as a similarity measure using BLOSUM35 and BLOSUM62, respectively. While the BLOSUM62 matrix has been included as it is found to be widely applied throughout scientific as the standard for protein alignment, including, for example, its usage by Koşaloğlu-Yalçın et al. (2018), the additional usage of the BLOSUM35 substitution matrix is based on the results obtained by Frankild et al. (2008). Specifically investigating features that allow for the predictability of T-cells' polyspecificity, the authors report predictability to some degree using the BLOSUM35 matrix.

Another frequently used similarity measure for peptide sequences is the kernel similarity measure presented in Shen et al. (2012). Based on the BLOSUM62 matrix, the authors identify three levels of similarity kernels:

$$K^1(x,y) = ([\text{BLOSUM62-2}](x,y))^\beta \tag{5.1}$$

$$K_k^2(u,v) = \prod_{i=1}^{k} K^1(u_i, v_i) \tag{5.2}$$

$$K^3(f,g) = \sum_{\substack{u \subset f, v \subset g \\ |u|=|v|=k \\ all k=1,2,\ldots}} K_k^2(u,v) \tag{5.3}$$

The similarity (**K3**) for two potentially in length differing peptides **f** and **g** is defined as the sum of all pairwise similarities **K2** over all pairs of equal length substrings **u** and **v**, respectively computed for all possible substring lengths. These equal length substring pairs are then compared to one another based on the BLOSUM62-2 matrix - which the authors derived from the BLOSUM62 matrix - followed by a Hadamard power.

In this thesis, three similarity kernel methods were used: Two local alignment scores based on BLOSUM35 and BLOSUM62, using the Smith-Waterman local alignment algorithm, respectively, and the kernel similarity measure as proposed by Shen et al. (2012), which itself is also based on the BLOSUM62 matrix. Local alignments with both substitution matrices were conducted using a penalty of gap opening and extension of 10 and 0.5, respectively, as these penalties are used as default parameters by EMBL-EBI.

All three similarity measures have further been normalized to be within the range of 0 and 1 - where 1 indicates a perfect match - by dividing the similarity measure by the square root of the product of the similarity values of computed for each sequence with itself:

$$\hat{K}(x, y) = \frac{K(x, y)}{\sqrt{K(x, x)K(y, y)}} \tag{5.4}$$

5.3.4 Model Evaluation

In order to evaluate which model from a set of evaluated classifiers provides the consistently best performance a commonly used method is to apply cross validation. However, as in the process of cross validation the original dataset is split into several distinct subsamples, Hothorn et al. (2005) point out that in the process of cross validation, dependencies among the individual generated subsamples is introduces. Therefore, as sample independence is a fundamental assumption of many statistical tests, easy statistical evaluation from a cross validated model is not possible.

Therefore, for the evaluation of the above mentioned 7 Random Forest models, and the 3 KSVM models, a non-parametric bootstrap (Tibshirani and Efron, 1993) will be used. The process of bootstrap sample generation for classifier evaluation is essentially the same as is applied for random forest model estimation. From the original dataset, observations are randomly selected with replacement to form the in-bag dataset which will be used to learn the model. All observations that have not been sampled into the in-bag dataset are considered out-of-bag and the learned classifier will be evaluated on out-of-bag data alone as during the learning process the model had no access to these observations.

In order to produce statistically evaluatable results, the process has been repeated 100 times. During each iteration a random in-bag and out of bag dataset was generated and all models were learned and evaluated on the same respective datasets. Due to the randomness of this evaluation procedure, for easy reproducibility the seed for the random number generator was set to **01015742**. As measure of the overall goodness of a classifier, the heavily used Area under the Curve (AUC) measure (Baesens et al., 2003) has been chosen.

A graphical representation of the resulting AUC values is presented in Figure 5.7 and a corresponding numerical summary of the bootstrap evaluation in terms of minimum, maximum, median, first and third quartile on a per model basis is reported in Table 5.4.

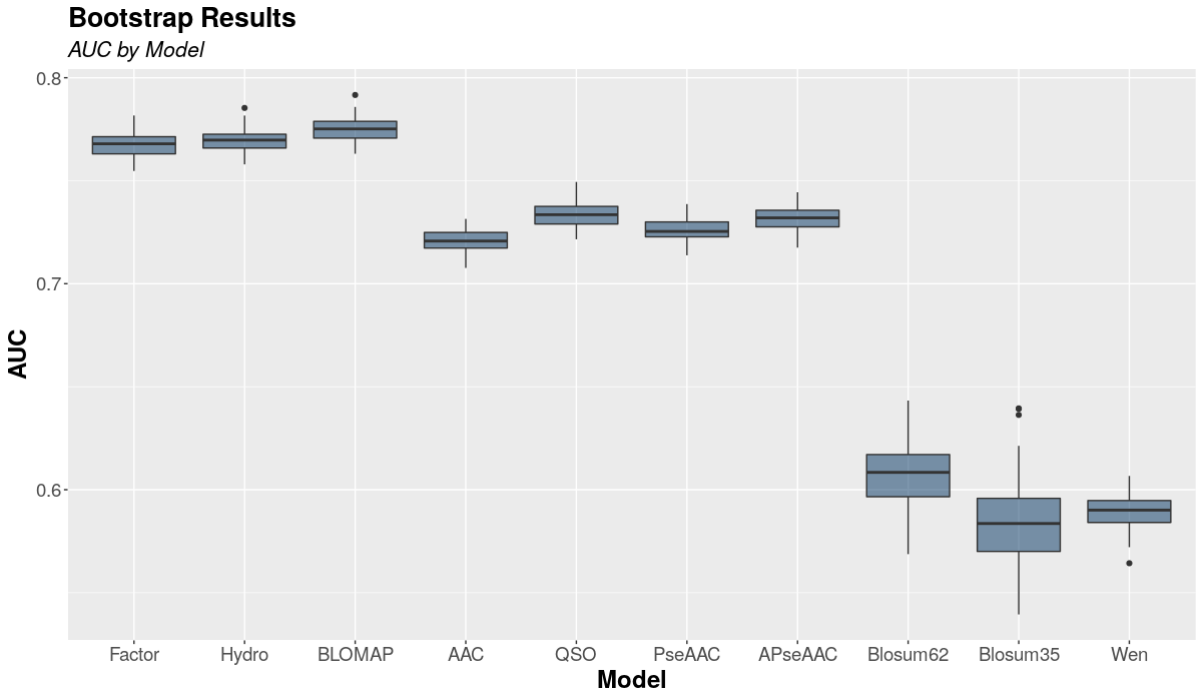

Figure 5.7: Bootstrap AUC results

Model	Minimum	Q1	Median	Q3	Maximum
Factor	0.755	0.763	0.768	0.771	0.782
Hydro	0.758	0.766	0.770	0.773	0.785
BLOMAP	0.763	0.771	0.775	0.779	0.792
AAC	0.708	0.717	0.721	0.725	0.731
QSO	0.722	0.729	0.734	0.738	0.749
PseAAC	0.714	0.723	0.725	0.730	0.739
APseAAC	0.718	0.727	0.732	0.736	0.744
Blosum62	0.569	0.596	0.608	0.617	0.643
Blosum35	0.539	0.570	0.584	0.596	0.639
Wen	0.564	0.584	0.590	0.595	0.607

Table 5.4: Five Point Summary of Bootstrap Results

Investigating the distributions of the resulting AUC values, an overall segmentation into three groups of classifiers, the members of which each appear to exhibit comparable performance, can be observed. The three amino acid based encoding schemes, **Factor**, **BLOMAP** and **Hydro** - indicating the encoding using physiochemical properties such as hydrophobicity - appear to return AUC values ranging on average from 0.75 to 0.80. The other random forest models using encoding based on **AAC** or the three other sequence based encodings **QSO**, **PseAAC** and **APseAAC**, respectively, can be found to exhibit the second highest values ranging between 0.70 and 0.75. By far the worst results in terms of AUC values is returned by the three KSVM models ranging around 0.6, hence being hardly better than random.

From these results it can be seen that models based on Random Forests using encoding schemes directly operating on each position in the peptide's sequence individually, provide the best predictive performance, indicating that information on a peptides immunogenicity may indeed be contained in specific positions along its sequence.

Estimating those three best performing models on the full dataset, importance measures of individual variables were extracted. The resulting importance measures were grouped by the position in the peptide sequence they relate to to make comparison across the individual models easier. A graphical overview of this distribution is provided in Figure 5.8.

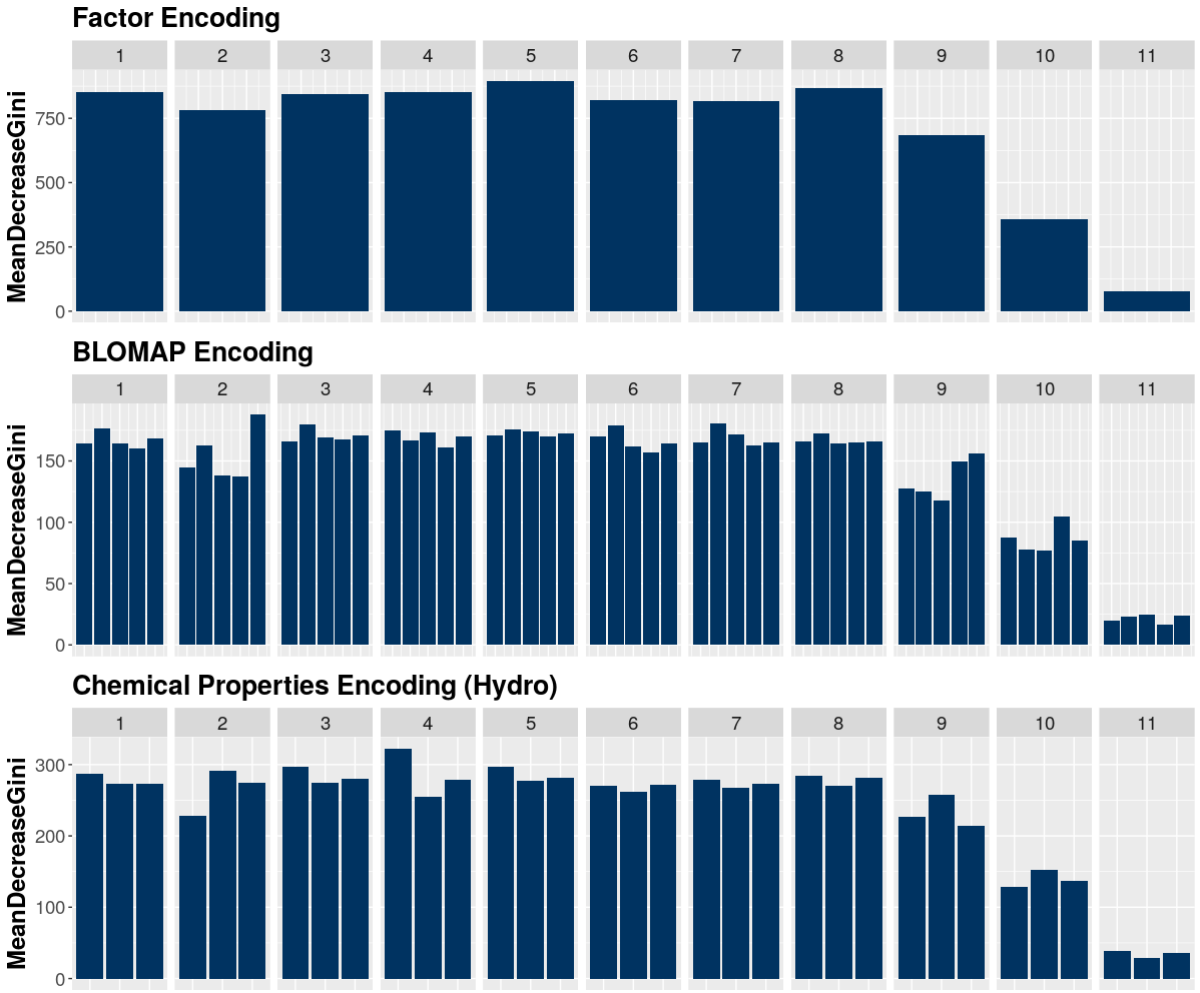

Figure 5.8: Random Forest Variable Importance

As can be seen by examining the variable importance across the peptide sequence, no clear one position can be found to systematically be indicative of a peptide's immunogenicity.

A common notion portrayed by all three models, however, is that from position nine onward importance tends to drop drastically and the amino acid at the last position of a peptide sequence appears to have very little overall importance. This may in part be influenced by the skewed distribution of peptide lengths in the whole training data set, leading to many of the investigated sequences being artificially padded to length eleven. However, as the most abundant

peptide length in the dataset are 9-mer peptides, it can be argued in correspondence with above mentioned research that position 9 marks an anchor position and as such influences the stability of a peptide MHC complex but bears little importance in terms of invoking an immune response.

Additionally, in particular when examining the **BLOMAP** encoding, a strong decrease in importance can be seen at position 2 which is consistent along all the features at this position except for the last one which appears to bear the highest overall importance in this particular encoding. In the other two models, a slight dip with respect to their neighboring positions can be observed at position 2 as well although not as prominent. Such a decrease in importance at position 2 would again be in accordance with previous research noting that like position 9, position 2 bears little to no predictive information for a peptide's immunogenicity.

While the three amino acid encoding models clearly outperform other approaches, it is from the above plot not clearly evident whether all three models would perform comparably well or whether significant differences in predictive performance can be observed.

As during the non parametric bootstrap each classifier was evaluated on the same out of bag sample, and therefore, the samples of classifier evaluations can be considered as paired, a repeated measures ANOVA (Japkowicz and Shah, 2011) will be applied to test the null hypothesis that all three classifiers perform equally well versus the alternative that at least one performs significantly better or worse. In order for a repeated measures ANOVA being applicable, two underlying assumptions need to be met: Normality of the individual samples, and Sphericity.

For all three model samples, normality has been investigated using the test proposed by Shapiro and Wilk (1965). As can be seen from the results of these tests as presented in Table 5.5, the assumption of the data following a normal distribution cannot be rejected at any reasonable significance level.

Model	Test-Statistic	p - value
BLOMAP	0.989	0.570
Factor	0.983	0.237
Hydro	0.989	0.594

Table 5.5: Shapiro-Wilk Normality Test results

The assumption of sphericity has been tested using the commonly employed Mauchly test for sphericity (Mauchly, 1940). As presented in Table 5.6, reporting a p-value of $1.77e-07$ indicates that the sphericity assumption is clearly violated. In order to compensate for this violation the Greenhouse - Geisser correction (Greenhouse and Geisser, 1959) will be applied and the respective degrees of freedom are adjusted.

W	p - value	$p < 0.05$
0.728	$1.77e-07$	*

Table 5.6: Mauchly's Test for Sphericity

Having accounted for this violation by adapting the degrees of freedom, a Fisher F-Test

(Fisher et al., 1934) is carried out returning a value of the test statistic of 308.875 on both 1.57 and 155.67 degrees of freedom resulting in a p-value of $4.87e-49$, hence the null hypothesis of the three classifiers performing equally well can be rejected with an error probability close to 0.

Having rejected the hypothesis of equal predictive performance, in a next step to all pairs of models a paired two sample T-test (Xu et al., 2017) is applied to to ascertain which models differ significantly from one another. In order to adjust for multiple testing, p-values are adjusted using the Bonferroni correction (Abdi, 2007). The result of this post-hoc test is graphically presented in Figure 5.9 and the respective values of the test statistic are presented in Table 5.7.

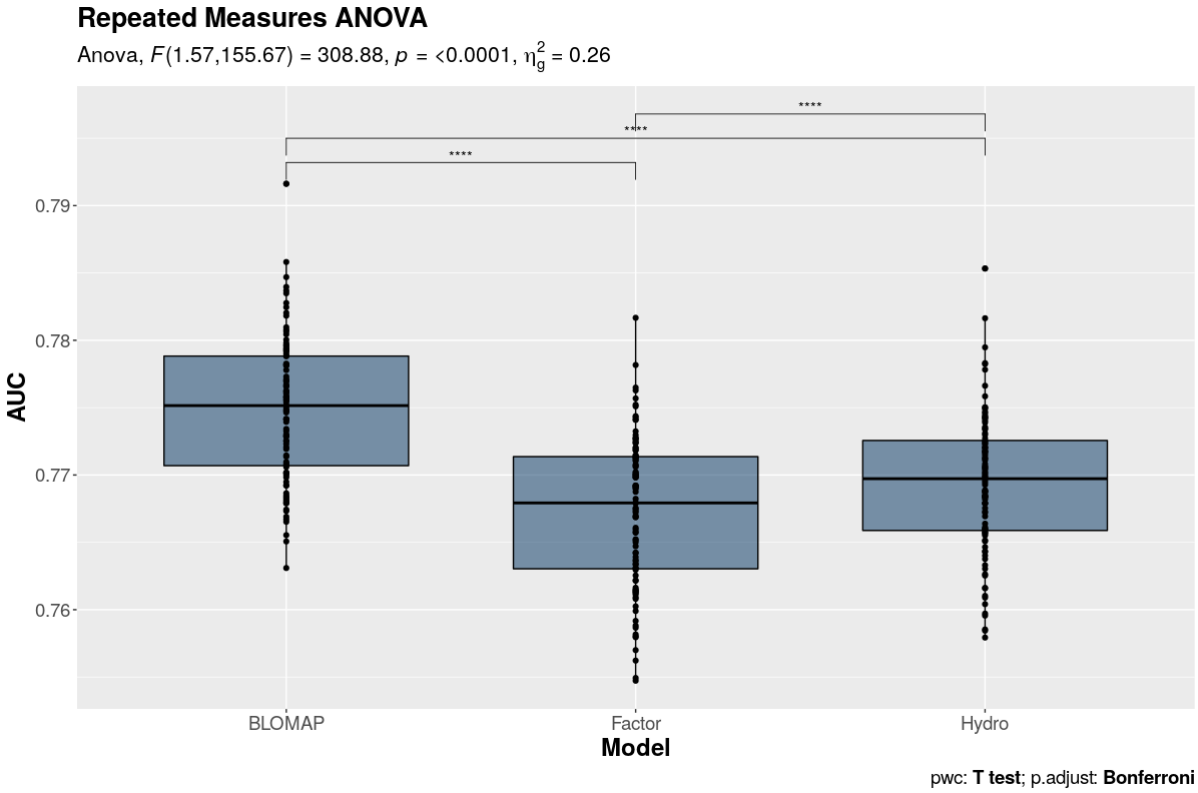

Figure 5.9: Repeated Measures Anova

group1	group2	statistic	df	p	p.adj	p.adj.signif
BLOMAP	Factor	24.7	99	3.49e-44	1.05e-43	****
BLOMAP	Hydro	22.9	99	2.29e-41	6.87e-41	****
Factor	Hydro	-5.47	99	3.45e-7	1.03e-6	****

Table 5.7: Bonferroni adjusted post-hoc two sample T-test

As can be seen from both sources, all two sample tests reject the null hypothesis with p-values far smaller than 0.01, indicating that all three classifiers significantly differ in their performance form each other. As can be easily seen from the respective graphics a Random Forest model in conjunction with **BLOMAP** encoding consistently outperforms all other evaluated models in terms of measured AUC. As such in the analytical workflow as presented in this paper, a Random Forest model with **BLOMAP** encoding is used to discern an antigen's probability of invoking an immune response.

Chapter 6

Future Research

Building on the above laid out structure of the overall workflow and the predictive performance obtained, two main areas in which future research is warranted in particular can be identified.

6.1 Workflow Extensions

As indicated above, a major design factor was to include various mutational sources that may potentially give rise to TSAs. While already accounting for SNVs, indels and frameshift mutations, these do, however, not provide an exhaustive list of sources for tumor antigens and should hence be expanded on further in the future.

One such limitation was incurred by restricting the analysis to mutations lying on coding segments of the DNA, i.e. regions which would eventually be translated into a protein sequence. The rationale for also including mutations occurring in intronic regions, i.e. segments in the DNA sequence which do not code for amino acids, is provided by Apcher et al. (2013). The authors find that part of the peptides displayed by cells on top of MHC-I complexes were derived from intronic sequences hence stating that antigen sequences are generated at a stage before such non-coding regions are removed from the transcript. As the approach presented in this thesis specifically operates on exonic sequences, inclusion of intron mutations should in the future be considered.

Another potential source for neo epitopes is presented by Kanaseki and Torigoe (2019). In this thesis proteins derived from the translation of mutated sequences were split into peptide fragments and directly considered as potential candidates. But as the authors indicate, proteasomes also have the capability of joining two peptide fragments together creating potentially new sequences which were not directly coded for in the cell's genome. As such further investigation of such proteasomal splicing derived peptides would provide an additional way to expand the list of identified potential neo epitopes.

Furthermore, antigens generated by more complex forms of mutations such as genomic fusions have not been included at this stage. As applications specifically developed for neo epitope prediction from such types of mutations - like INTEGRATE-Neo (Zhang et al., 2017) - are already existent, including such a program alongside the custom developed neo eptiope generation approach in the overall workflow or alternatively supporting these types of mutations natively, is desired in a future iteration.

In terms of providing a more succinct view on the epitope landscape and hence further limit

the number of false positives reported by analytical workflows, further filtering based on the work of Pearson et al. (2016) may be considered. In the current state, proteins originating from all genes have been considered equally, and marked according to their evaluated transcriptional profile based on mRNA data. However, as argued in their paper, not all synthesized proteins have the potential of eventually being degraded into presentable antigens. As the authors argue, not all expressed genes which will be translated into proteins will also result in the generation of peptides being presented on top of MHC-I molecules. Therefore, such a gene based preselection would provide a potential additional filtering approach to better identify potential neoantigen targets for immunotherapeutic approaches.

6.2 Predictive Modelling Advancements

The second main area which is to be expanded on in future is to improve overall predictiveness in terms of immunogenicity evaluation.

A major limiting factor for the modelling approaches explored in this thesis is the availability of epitope data and the features reported on these data points. Due to the lack of additional features, predictive models were designed to be solely based on information contained in a peptide's sequence, which is to be extended on in the future.

Expanding on the idea of the mechanism of central tolerance (see Chapter 3), a major part to be included in future is to include sequence comparison of the derived neo peptides with the whole set of the sequences of a patient's self peptides. While at the current stage, a hard filter was applied and all sequences matching any of the patient's self peptides were disregarded, future improvements would in absence of a perfect match also consider the most closely related self peptide sequence and report some form of similarity measure.

Additionally, as more vaccination studies are conducted and full information on the patient's genomic landscape as well as experimentally tested antigen sequences become available, a central aim is to explore the predictive ability of such self similarity measures based on the whole of a patient's self antigens as derived from their individual germline mutations. This approach can further be extended by taking into account a patient's specific combination of HLA types. As indicated by Migalska et al. (2019), depending on the particular combination of HLA types encoded in a patient's genome, the set of available TCRs may be shaped substantially.

Another open problem relates to the computation of sequence similarity. Despite the overall lack in predictive ability of the models based on sequence similarity, as explored here, only three proposed Kernel methods were explored and in the future more similarity measures are to be evaluated.

Chapter 7

Conclusion

A critical part of many cancer immunotherapy approaches relies first on confidently identifying epitopes which are specific to cancer cells and hence absent from healthy host cells.

While due to advances in sequencing technologies throughout recent years a multitude of analytical approaches for the identification of cancer specific neoantigens have emerged, many of these workflows show serious limitations. Many of these analytical approaches, for example, only consider a subset of possible mutation types as sources for neoantigens. Moreover, information on a patient's germline mutations as well as phasing of both mutation types is rarely included in such analytical approaches, therefore potentially resulting in the identification of wrong peptide sequences.

The analytical pipeline presented in this thesis aims at overcoming such limitations by using Single Nucleotide Variations, indels as well as frameshift mutations as a basis for the derivation of Tumor Specific Antigens. By combining existing and well established tools alongside a custom developed software solution into a comprehensive workflow, an analytical pipeline for neo epitope identification is generated which can directly run on raw unprocessed sequence data without requiring additional user input.

In terms of an antigen's ability to aid in tumor rejection, existing analytical workflows often resort to a peptide's affinity to bind to molecules of the MHC complex and do not carry out explicit immunogenicity predictions.

Aiming to extend upon solely relying on binding affinity prediction, epitope data has been collected from various sources to allow for the creation of a custom predictive model in terms of a peptide's potential to invoke an immune response.

Creating both Random Forest and Kernel Support Vector Machine models from the data using various encoding approaches, benchmark analysis reveals that encoding peptides using the BLOMAP approach and modelling their immunogenicity with a Random Forest, the best predictive performance as measured in terms of their AUC could be achieved. As such, this model is integrated into the overall workflow and in conjunction with binding affinity scores provides an additional estimate of an identified antigen's suitability for subsequent therapeutic application.

Further exploration of the individual feature importances for the so constructed model reveals that, in consensus with prior research, certain positions in an amino acid sequence appear to carry more importance with respect to their ability to trigger an immune response than others.

While, in accordance with previous studies, an overall decrease in importance can be observed for amino acids at position nine of a peptide's sequence onward, no such harsh decline in importance with respect to the second position in the sequence can be observed. Additionally, an increase in variable importance towards the more central positions, as would have been expected, is also not distinctly discernible.

Overall, the notion that information on a peptide's immunogenicity is contained in specific positions is strongly supported in this thesis, as models being built on retaining positional information by encoding each amino acid in the sequence individually consistently provide better results than other models examined in this thesis which rely on different types of encodings.

Bibliography

Hervé Abdi. Bonferroni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3:103–107, 2007.

Vincent Alcazer, Paola Bonaventura, Laurie Tonon, Sandrine Wittmann, Christophe Caux, and Stéphane Depil. Neoepitopes-based vaccines: challenges and perspectives. *European Journal of Cancer*, 108:55–60, 2019.

Sébastien Apcher, Guy Millot, Chrysoula Daskalogianni, Alexander Scherl, Bénédicte Manoury, and Robin Fåhraeus. Translation of pre-spliced rnas in the nuclear compartment generates peptides for the mhc class i pathway. *Proceedings of the National Academy of Sciences*, 110 (44):17951–17956, 2013.

Markus Außerhofer. Development and validation of a pipeline for tumor neoantigen prediction. 2020.

Bart Baesens, Tony Van Gestel, Stijn Viaene, Maria Stepanova, Johan Suykens, and Jan Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, 54(6):627–635, 2003.

Pierre Baldi, Søren Brunak, and Francis Bach. *Bioinformatics: the machine learning approach*. MIT press, 2001.

Michal Bassani-Sternberg, Eva Bräunlein, Richard Klar, Thomas Engleitner, Pavel Sinitcyn, Stefan Audehm, Melanie Straub, Julia Weber, Julia Slotta-Huspenina, Katja Specht, et al. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nature communications*, 7(1):1–16, 2016.

Amalie Kai Bentzen, Andrea Marion Marquard, Rikke Lyngaa, Sunil Kumar Saini, Sofie Ramskov, Marco Donia, Lina Such, Andrew JS Furness, Nicholas McGranahan, Rachel Rosenthal, et al. Large-scale detection of antigen-specific t cells using peptide-mhc-i multimers labeled with dna barcodes. *Nature biotechnology*, 34(10):1037, 2016.

Jeremy Mark Berg, John L Tymoczko, Lubert Stryer, et al. Biochemistry/jeremy m. berg, john l. tymoczko, lubert stryer; with gregory j. gatto, jr., 2012.

Grant B.J., Rodrigues A.P.C., ElSawy K.M., McCammon J.A., and Caves L.S.D. Bio3d: An r package for the comparative analysis of protein structures. *Bioinformatics*, 22:2695–2696, Nov 2006.

Anne-Mette Bjerregaard, Morten Nielsen, Sine Reker Hadrup, Zoltan Szallasi, and Aron Charles Eklund. Mupexi: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunology, Immunotherapy*, 66(9):1123–1130, 2017a.

Anne-Mette Bjerregaard, Morten Nielsen, Vanessa Jurtz, Carolina M Barra, Sine Reker Hadrup, Zoltan Szallasi, and Aron Charles Eklund. An analysis of natural t cell responses to predicted tumor neoepitopes. *Frontiers in immunology*, 8:1566, 2017b.

Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

Vinothini Boopathi, Sathiyamoorthy Subramaniyam, Adeel Malik, Gwang Lee, Balachandran Manavalan, and Deok-Chun Yang. macppred: A support vector machine-based meta-predictor for identification of anticancer peptides. *International journal of molecular sciences*, 20(8): 1964, 2019.

Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5):525, 2016.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Scott D Brown and Robert A Holt. Neoantigen characteristics in the context of the complete predicted mhc class i self-immunopeptidome. *Oncoimmunology*, 8(3):1556080, 2019.

Alexandra R Buckley, Trey Ideker, Hannah Carter, and Nicholas J Schork. Rare variant phasing using paired tumor: normal sequence data. *BMC bioinformatics*, 20(1):265, 2019.

Stephane Buhler and Alicia Sanchez-Mazas. Hla dna sequence variation among human populations: molecular signatures of demographic and selective events. *PloS one*, 6(2), 2011.

Jorg JA Calis, Rob J De Boer, and Can Keşmir. Degenerate t-cell recognition of peptides on mhc molecules creates large holes in the t-cell repertoire. *PLoS computational biology*, 8(3): e1002412, 2012.

Jorg JA Calis, Matt Maybeno, Jason A Greenbaum, Daniela Weiskopf, Aruna D De Silva, Alessandro Sette, Can Keşmir, and Bjoern Peters. Properties of mhc class i presented peptides that enhance immunogenicity. *PLoS computational biology*, 9(10):e1003266, 2013.

Stephane E Castel, Pejman Mohammadi, Wendy K Chung, Yufeng Shen, and Tuuli Lappalainen. phaser: Long range phasing and haplotypic expression from rna sequencing. *bioRxiv*, page 039529, 2016.

Kuo-Chen Chou. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochemical and biophysical research communications*, 278(2):477–483, 2000.

Kuo-Chen Chou. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, 21(1):10–19, 2005.

Kuo-Chen Chou. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Current Proteomics*, 6(4):262–274, 2009.

Kuo-Chen Chou and David W Elrod. Prediction of membrane protein types and subcellular locations. *Proteins: Structure, Function, and Bioinformatics*, 34(1):137–153, 1999.

Diego Chowell, Sri Krishna, Pablo D Becker, Clément Cocita, Jack Shu, Xuefang Tan, Philip D Greenberg, Linda S Klavinskis, Joseph N Blattman, and Karen S Anderson. Tcr contact residue hydrophobicity is a hallmark of immunogenic cd8+ t cell epitopes. *Proceedings of the National Academy of Sciences*, 112(14):E1754–E1762, 2015.

Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3):213, 2013.

Cyrille J Cohen, Jared J Gartner, Miryam Horovitz-Fried, Katerina Shamalov, Kasia Trebska-McGowan, Valery V Bliskovsky, Maria R Parkhurst, Chen Ankri, Todd D Prickett, Jessica S Crystal, et al. Isolation of neoantigen-specific t cells from tumor and peripheral lymphocytes. *The Journal of clinical investigation*, 125(10):3981–3991, 2015.

Cecil C Czerkinsky, Lars-Åke Nilsson, Håkan Nygren, Örjan Ouchterlony, and Andrej Tarkowski. A solid-phase enzyme-linked immunospot (elispot) assay for enumeration of specific antibody-secreting cells. *Journal of immunological methods*, 65(1-2):109–121, 1983.

Margaret O Dayhoff. A model of evolutionary change in proteins. *Atlas of protein sequence and structure*, 5:89–99, 1972.

Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

Fei Duan, Jorge Duitama, Sahar Al Seesi, Cory M Ayres, Steven A Corcelli, Arpita P Pawashe, Tatiana Blanchard, David McMahon, John Sidney, Alessandro Sette, et al. Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. *Journal of Experimental Medicine*, 211(11):2231–2248, 2014.

GP Dunn, AT Bruce, H Ikeda, and L Old. J.; schreiber, rd cancer immunoediting: From immunosurveillance to tumor escape. *Nat. Immunol*, 3(11):991–998, 2002.

EMBL-EBI. Emboss water. `https://www.ebi.ac.uk/Tools/psa/emboss_water/`. Accessed: 2020-01-20.

Adam D Ewing, Kathleen E Houlahan, Yin Hu, Kyle Ellrott, Cristian Caloian, Takafumi N Yamaguchi, J Christopher Bare, Christine P'ng, Daryl Waggott, Veronica Y Sabelnykova, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nature methods*, 12(7):623, 2015.

Brent Ewing, LaDeana Hillier, Michael C Wendl, and Phil Green. Base-calling of automated sequencer traces usingphred. i. accuracy assessment. *Genome research*, 8(3):175–185, 1998.

Ronald Aylmer Fisher et al. Statistical methods for research workers. *Statistical methods for research workers.*, (5th Ed), 1934.

Ward Fleri, Sinu Paul, Sandeep Kumar Dhanda, Swapnil Mahajan, Xiaojun Xu, Bjoern Peters, and Alessandro Sette. The immune epitope database and analysis resource in epitope discovery and synthetic vaccine design. *Frontiers in immunology*, 8:278, 2017.

Sune Frankild, Rob J De Boer, Ole Lund, Morten Nielsen, and Can Kesmir. Amino acid similarity accounts for t cell cross-reactivity and for "holes" in the t cell repertoire. *PloS one*, 3(3):e1831, 2008.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Richard Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185(4154):862–864, 1974.

Samuel W Greenhouse and Seymour Geisser. On methods in the analysis of profile data. *Psychometrika*, 24(2):95–112, 1959.

Jessica C Hassel. Ipilimumab plus nivolumab for advanced melanoma. *The Lancet Oncology*, 17 (11):1471–1472, 2016.

Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

Martin Hofmann. Support vector machines-kernels and the kernel trick. *Notes*, 26(3), 2006.

Kristin A Hogquist, Stephen C Jameson, William R Heath, Jane L Howard, Michael J Bevan, and Francis R Carbone. T cell receptor antagonist peptides induce positive selection. *Cell*, 76(1):17–27, 1994.

Caitriona Holohan, Sandra Van Schaeybroeck, Daniel B Longley, and Patrick G Johnston. Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer*, 13(10):714–726, 2013.

Torsten Hothorn, Friedrich Leisch, Achim Zeileis, and Kurt Hornik. The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, 14(3):675–699, 2005.

Jasreet Hundal, Beatriz M Carreno, Allegra A Petti, Gerald P Linette, Obi L Griffith, Elaine R Mardis, and Malachi Griffith. pvac-seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome medicine*, 8(1):11, 2016.

Jasreet Hundal, Susanna Kiwala, Yang-Yang Feng, Connor J Liu, Ramaswamy Govindan, William C Chapman, Ravindra Uppaluri, S Joshua Swamidass, Obi L Griffith, Elaine R Mardis, et al. Accounting for proximal variants improves neoantigen prediction. *Nature genetics*, 51(1):175, 2019.

Broad Institute. Picard toolkit. `http://broadinstitute.github.io/picard/`, 2019.

Nathalie Japkowicz and Mohak Shah. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.

Tanner M Johanns, Christopher A Miller, Connor J Liu, Richard J Perrin, Diane Bender, Dale K Kobayashi, Jian L Campian, Michael R Chicoine, Ralph G Dacey, Jiayi Huang, et al. Detection of neoantigen-specific t cells following a personalized vaccine in a patient with glioblastoma. *OncoImmunology*, 8(4):e1561106, 2019.

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. *The Journal of Immunology*, 199(9):3360–3368, 2017.

Takayuki Kanaseki and Toshihiko Torigoe. Proteogenomics: advances in cancer antigen research. *Immunological medicine*, 42(2):65–70, 2019.

Takahiro Karasaki, Kazuhiro Nagayama, Hideki Kuwano, Jun-ichi Nitadori, Masaaki Sato, Masaki Anraku, Akihiro Hosoi, Hirokazu Matsushita, Masaki Takazawa, Osamu Ohara, et al. Prediction and prioritization of neoantigens: integration of rna sequencing data with whole-exome sequencing. *Cancer science*, 108(2):170–177, 2017.

Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004. URL `http://www.jstatsoft.org/v11/i09/`.

Alboukadel Kassambara. *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*, 2019. URL `https://CRAN.R-project.org/package=rstatix`. R package version 0.3.1.

Shuichi Kawashima, Hiroyuki Ogata, and Minoru Kanehisa. Aaindex: amino acid index database. *Nucleic acids research*, 27(1):368–369, 1999.

Vladislav V Khrustalev and Eugene V Barkovsky. Percent of highly immunogenic amino acid residues forming b-cell epitopes is higher in homologous proteins encoded by gc-rich genes. *Journal of theoretical biology*, 282(1):71–79, 2011.

Sora Kim, Han Sang Kim, E Kim, MG Lee, E-C Shin, S Paik, and S Kim. Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Annals of Oncology*, 29(4):1030–1036, 2018.

Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3):568–576, 2012.

Zeynep Koşaloğlu-Yalçın, Manasa Lanka, Angela Frentzen, Ashmitaa Logandha Ramamoorthy Premlal, John Sidney, Kerrie Vaughan, Jason Greenbaum, Paul Robbins, Jared Gartner, Alessandro Sette, et al. Predicting t cell recognition of mhc class i restricted neoepitopes. *Oncoimmunology*, 7(11):e1492508, 2018.

Vessela N Kristensen. The antigenicity of the tumor cell-context matters. *New England Journal of Medicine*, 376(5):491–493, 2017.

Carla Kuiken, Karina Yusim, Laura Boykin, and Russell Richardson. The los alamos hepatitis c sequence database. *Bioinformatics*, 21(3):379–384, 2005.

Gregory M Kurtzer, Vanessa Sochat, and Michael W Bauer. Singularity: Scientific containers for mobility of compute. *PloS one*, 12(5), 2017.

Jack Kyte and Russell F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132, 1982.

Sneh Lata, Manoj Bhasin, and Gajendra PS Raghava. Mhcbn 4.0: A database of mhc/tap binding peptides and t-cell epitopes. *BMC research notes*, 2(1):61, 2009.

Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3): 18–22, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

Ferenc Livak, Douglas B Burtrum, Lee Rowen, David G Schatz, and Howard T Petrie. Genetic modulation of t cell receptor gene segment usage during somatic recombination. *The Journal of experimental medicine*, 192(8):1191–1196, 2000.

A Llano, A Williams, A Olvera, S Silva-Arrieta, and C Brander. Best-characterized hiv-1 ctl epitopes: the 2013 update. *HIV molecular immunology*, 2013:3–25, 2013.

Yong-Chen Lu, Xin Yao, Jessica S Crystal, Yong F Li, Mona El-Gamil, Colin Gross, Lindy Davis, Mark E Dudley, James C Yang, Yardena Samuels, et al. Efficient identification of mutated cancer antigens recognized by t cells associated with durable tumor regressions. *Clinical Cancer Research*, 20(13):3401–3410, 2014.

Stefan Maetschke, Michael Towsey, and Mikael Boden. Blomap: an encoding of amino acids which improves signal peptide cleavage site prediction. In *Proceedings of the 3rd Asia-Pacific bioinformatics conference*, pages 141–150. World Scientific, 2005.

Julian A Marin-Acevedo, Bhagirathbhai Dholaria, Aixa E Soyano, Keith L Knutson, Saranya Chumsri, and Yanyan Lou. Next generation of immune checkpoint therapy in cancer: new developments and challenges. *Journal of hematology & oncology*, 11(1):39, 2018.

John W Mauchly. Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, 11(2):204–209, 1940.

Nicholas McGranahan, Andrew JS Furness, Rachel Rosenthal, Sofie Ramskov, Rikke Lyngaa, Sunil Kumar Saini, Mariam Jamal-Hanjani, Gareth A Wilson, Nicolai J Birkbak, Crispin T Hiley, et al. Clonal neoantigens elicit t cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*, 351(6280):1463–1469, 2016.

Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome biology*, 17(1):122, 2016.

Magdalena Migalska, Alvaro Sebastian, and Jacek Radwan. Major histocompatibility complex class i diversity limits the repertoire of t cell receptors. *Proceedings of the National Academy of Sciences*, 116(11):5021–5026, 2019.

Pooja Narang, Meixuan Chen, Amit A Sharma, Karen S Anderson, and Melissa A Wilson. The neoepitope landscape of breast cancer: implications for immunotherapy. *BMC cancer*, 19(1): 200, 2019.

Masato Ogishi and Hiroshi Yotsuyanagi. Quantitative prediction of the landscape of t cell epitope immunogenicity in sequence space. *Frontiers in immunology*, 10:827, 2019.

Lars Rønn Olsen, Songsak Tongchusak, Honghuang Lin, Ellis L Reinherz, Vladimir Brusic, and Guang Lan Zhang. Tantigen: a comprehensive database of tumor t cell antigens. *Cancer Immunology, Immunotherapy*, 66(6):731–735, 2017.

Patrick A Ott, Zhuting Hu, Derin B Keskin, Sachet A Shukla, Jing Sun, David J Bozym, Wandi Zhang, Adrienne Luoma, Anita Giobbie-Hurder, Lauren Peter, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*, 547(7662):217, 2017.

Patrick A Ott, Gianpietro Dotti, Cassian Yee, and Stephanie L Goff. An update on adoptive t-cell therapy and neoantigen vaccines. *American Society of Clinical Oncology Educational Book*, 39:e70–e78, 2019.

Drew M Pardoll. The blockade of immune checkpoints in cancer immunotherapy. *Nature Reviews Cancer*, 12(4):252, 2012.

Hillary Pearson, Tariq Daouda, Diana Paola Granados, Chantal Durette, Eric Bonneil, Mathieu Courcelles, Anja Rodenbrock, Jean-Philippe Laverdure, Caroline Côté, Sylvie Mader, et al. Mhc class i–associated peptides derive from selective regions of the human genome. *The Journal of clinical investigation*, 126(12):4690–4701, 2016.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL `https://www.R-project.org/`.

Mohini Rajasagi, Sachet A Shukla, Edward F Fritsch, Derin B Keskin, David DeLuca, Ellese Carmona, Wandi Zhang, Carrie Sougnez, Kristian Cibulskis, John Sidney, et al. Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. *Blood, The Journal of the American Society of Hematology*, 124(3):453–462, 2014.

Arjun A Rao, Ada A Madejska, Jacob Pfeil, Benedict Paten, Sofie Salama, and David Haussler. Protect: Prediction of t-cell epitopes for cancer therapy. *bioRxiv*, page 696526, 2019.

Pedro A Reche, Hong Zhang, John-Paul Glutting, and Ellis L Reinherz. Epimhc: a curated database of mhc-binding peptides for customized computational vaccinology. *Bioinformatics*, 21(9):2140–2141, 2005.

Timothy P Riley, Grant LJ Keller, Angela Smith, Jason R Devlin, Lauren M Davancaze, Alyssa Arbuiso Arbuiso, and Brian M Baker. Structure based prediction of neoantigen immunogenicity. *Frontiers in immunology*, 10:2047, 2019.

Paul F Robbins, Yong-Chen Lu, Mona El-Gamil, Yong F Li, Colin Gross, Jared Gartner, Jimmy C Lin, Jamie K Teer, Paul Cliften, Eric Tycksen, et al. Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive t cells. *Nature medicine*, 19(6):747, 2013.

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011.

Michael D Rosenblum, Kelly A Remedios, and Abul K Abbas. Mechanisms of human autoimmunity. *The Journal of clinical investigation*, 125(6):2228–2233, 2015.

Alex Rubinsteyn, Isaac Hodes, Julia Kodysh, and Jeffrey Hammerbacher. Vaxrank: a computational tool for designing personalized cancer vaccines. *bioRxiv*, page 142919, 2017.

Christopher T Saunders, Wendy SW Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817, 2012.

Maike Schmidt and Jennie R Lill. Mhc class i presented antigens from malignancies: A perspective on analytical characterization & immunogenicity. *Journal of proteomics*, 191:48–57, 2019.

Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

Wen-Jun Shen, Hau-San Wong, Quan-Wu Xiao, Xin Guo, and Stephen Smale. Towards a mathematical foundation of immunology and amino acid chains. *arXiv preprint arXiv:1205.6031*, 2012.

Christof C Smith, Shengjie Chai, Amber R Washington, Samuel J Lee, Elisa Landoni, Kevin Field, Jason Garness, Lisa M Bixby, Sara R Selitsky, Joel S Parker, et al. Machine-learning prediction of tumor antigen immunogenicity in the selection of therapeutic epitopes. *Cancer immunology research*, 7(10):1591–1604, 2019a.

Christof C Smith, Sara R Selitsky, Shengjie Chai, Paul M Armistead, Benjamin G Vincent, and Jonathan S Serody. Alternative tumour-specific antigens. *Nature Reviews Cancer*, 19(8): 465–478, 2019b.

Temple F Smith, Michael S Waterman, et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

Timothy K Starr, Stephen C Jameson, and Kristin A Hogquist. Positive and negative selection of t cells. *Annual review of immunology*, 21(1):139–176, 2003.

Erlend Strønen, Mireille Toebes, Sander Kelderman, Marit M Van Buuren, Weiwen Yang, Nienke Van Rooij, Marco Donia, Maxi-Lu Böschen, Fridtjof Lund-Johansen, Johanna Olweus, et al. Targeting of cancer neoantigens with donor-derived t cell receptor repertoires. *Science*, 352 (6291):1337–1341, 2016.

András Szolek, Benjamin Schubert, Christopher Mohr, Marc Sturm, Magdalena Feldhahn, and Oliver Kohlbacher. Optitype: precision hla typing from next-generation sequencing data. *Bioinformatics*, 30(23):3310–3316, 2014.

Gabriel N Teku and Mauno Vihinen. Pan-cancer analysis of neoepitopes. *Scientific reports*, 8, 2018.

Robert J Tibshirani and Bradley Efron. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57:1–436, 1993.

Chun-Wei Tung, Matthias Ziehm, Andreas Kämper, Oliver Kohlbacher, and Shinn-Ying Ho. Popisk: T-cell reactivity prediction using support vector machines and string kernels. *BMC bioinformatics*, 12(1):446, 2011.

Geraldine A Van der Auwera, Mauricio O Carneiro, Christopher Hartl, Ryan Poplin, Guillermo Del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, et al. From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, 43(1):11–10, 2013.

Vladimir N Vapnik. The nature of statistical learning. *Theory*, 1995.

Kate Voss, Geraldine Van der Auwera, and Jeff Gentry. Full-stack genomics pipelining with gatk4+ wdl+ cromwell. *F1000Research*, 6, 2017.

Guangzhi Wang, Huihui Wan, Xingxing Jian, Jian Ouyang, Yuyu Li, Xiaoxiu Tan, Yong Xu, Yong Zhao, Yong Lin, and Lu Xie. Ineo-epp: T-cell hla class i immunogenic or neoantigenic epitope prediction via random forest algorithm based on sequence related amino acid features. *BioRxiv*, page 697011, 2019.

Darin A Wick, John R Webb, Julie S Nielsen, Spencer D Martin, David R Kroeger, Katy Milne, Mauro Castellarin, Kwame Twumasi-Boateng, Peter H Watson, Rob A Holt, et al. Surveillance of the tumor mutanome by t cells during progression from primary to recurrent ovarian cancer. *Clinical Cancer Research*, 20(5):1125–1134, 2014.

Claire R Williams, Alyssa Baccarella, Jay Z Parrish, and Charles C Kim. Trimming of sequence reads alters rna-seq gene expression estimates. *BMC bioinformatics*, 17(1):103, 2016.

Mary A Wood, Austin Nguyen, Adam J Struck, Kyle Ellrott, Abhinav Nellore, and Reid F Thompson. neoepiscope improves neoepitope prediction with multi-variant phasing. *BioRxiv*, page 418129, 2019.

Kai W Wucherpfennig, Paul M Allen, Franco Celada, Irun R Cohen, Rob De Boer, K Christopher Garcia, Byron Goldstein, Ralph Greenspan, David Hafler, Philip Hodgkin, et al. Polyspecificity of t cell and b cell receptor recognition. In *Seminars in immunology*, volume 19, pages 216–224. Elsevier, 2007.

Nan Xiao, Dong-Sheng Cao, Min-Feng Zhu, and Qing-Song Xu. protr/protrweb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics*, 31(11):1857–1859, 2015.

Chang Xu. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and structural biotechnology journal*, 16:15–24, 2018.

Manfei Xu, Drew Fralick, Julia Z Zheng, Bokai Wang, Xin M Tu, and Changyong Feng. The differences and similarities between two-sample t-test and paired t-test. *Shanghai archives of psychiatry*, 29(3):184, 2017.

Wen-Yun Yang, Farhad Hormozdiari, Zhanyong Wang, Dan He, Bogdan Pasaniuc, and Eleazar Eskin. Leveraging reads that span multiple single nucleotide polymorphisms for haplotype inference from sequencing data. *Bioinformatics*, 29(18):2245–2252, 2013.

Jonathan W Yewdell, Eric Reits, and Jacques Neefjes. Making sense of mass destruction: quantitating mhc class i antigen presentation. *Nature Reviews Immunology*, 3(12):952–961, 2003.

Masood Zamani and Stefan C Kremer. Amino acid encoding schemes for machine learning methods. In *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, pages 327–333. IEEE, 2011.

Jin Zhang, Elaine R Mardis, and Christopher A Maher. Integrate-neo: a pipeline for personalized gene fusion neoantigen discovery. *Bioinformatics*, 33(4):555, 2017.

Chi Zhou, Zhiting Wei, Zhanbing Zhang, Biyu Zhang, Chenyu Zhu, Ke Chen, Guohui Chuai, Sheng Qu, Lu Xie, Yong Gao, et al. ptuneos: prioritizing tumor neoantigens from next-generation sequencing data. *Genome medicine*, 11(1):1–17, 2019a.

Chi Zhou, Chenyu Zhu, and Qi Liu. Toward in silico identification of tumor neoantigens in immunotherapy. *Trends in molecular medicine*, 2019b.

JM Zimmerman, Naomi Eliezer, and R Simha. The characterization of amino acid sequences in proteins by statistical methods. *Journal of theoretical biology*, 21(2):170–201, 1968.

Appendix A

Software Availability

The analytic workflow presented in this thesis has been made publicly available and is accessible at: https://github.com/csam5596/EpitopePrediction.

The overall pipeline was implemented using the Workflow Description Language provided by Voss et al. (2017) and every task defined has been assigned a container image to allow for portable use. A summary of the used programs alongside their respective versions and docker files is presented in Table A.1.

Program	Version	Docker File
Optitype	1.3	fred2/optitype:release-v1.3.1
Kallisto	latest(0.46.1)	insilicodb/kallisto:1.0.0
Trimmomatic	0.36	comics/trimmomatic:0.36
BWAMem	0.7.13	biocontainers/bwa:v0.7.17-3-deb_cv1
STAR	latest(2.7.3a)	dceoy/star:latest
Samtools	1.9	biocontainers/samtools:v1.9-4-deb_cv1
Picard	2.21.2	broadinstitute/picard:2.21.2
GATK	4.1.4.0	broadinstitute/gatk:4.1.4.0
Strelka	2.9.9	mgibio/strelka:2.9.9
Varscan	2.4.2	mgibio/varscan:v2.4.2
BCFTools	1.9.1	biocontainers/bcftools:v1.9-1-deb_cv1

Table A.1: Pipeline dependencies

Alongside the pipeline's description file, also an input template file - **input.json** - as well as a default configuration file - **configuration.conf** - specifically tailored to be used on an HPC architecture using SGE as a job scheduler and Singularity (Kurtzer et al., 2017) as container technology, is provided.

It should be noted at this stage that the final task defined in the workflow, i.e. the generation and evaluation of epitopes, which has been custom created for this purpose, does at this stage not yet have a container file specified on docker hub and as such requires it to be generated manually as **epi.sif** and be present in the folder from which the pipeline is executed.

The custom created epitope prediction program which has been integrated as part of the pipeline has been created using C++ 17 and is also freely available at

https://github.com/csam5596/Epi alongside which a singularity image file is provided to build the needed container file.

An important part to point out, however, is that, as outlined in the sections above, for binding affinity evaluation **NetMHCpan** has been used. The inclusion of which is intended for academic use and for commercial application requires contacting the authors directly. (See http://www.cbs.dtu.dk/cgi-bin/nph-sw_request?netMHCpan).

Appendix B

Materials and Methods

Data Availability

The raw data used to create the presented statistical analyses was composed from various sources and compiled by Außerhofer (2020) and is made publicly available at https://github.com/csam5596/MasterThesis. Alongside the datasets, all R scripts containing the code used for the creation and evaluation of the individual models and the generation of the presented graphics were made available as well. Additionally, also the results of all bootstrap iterations have been collected and can be accessed at the above link.

Data Analysis

Empirical models in this thesis were built and analyzed using **R version 3.6.2 (2019-12-12) – "Dark and Stormy Night"** (R Core Team, 2019).

Random Forest

All Random Forest models in this thesis were built using the **randomForest** package (Liaw and Wiener, 2002) in version **4.6.14**.

In terms of encoding peptide sequences, to encode individual amino acids with respect to their physiochemical properties, the package **bio3d** (B.J. et al., 2006) in version **2.4.0** has been used. For the particular encoding used in this thesis, the method **aa2Index** was employed, using "KYTJ820101", "ZIMJ680102" and "GRAR740102" as index parameters with a window size of 1, reflective of the hydrophobicity on the Kyte-Doolittle scale (Kyte and Doolittle, 1982), polarity as denoted by Grantham (1974) and the amino acid's side chain mass as specified by Zimmerman et al. (1968), respectively.

Encodings AAC, PseAAC, QSO and APseAAC were constructed using the package **protr** (Xiao et al., 2015) in version **1.6.2**. The vanilla AAC encoding was constructed with the method **extractAAC** with no parameters. The other feature encodings PseAAC, QSO and APseAAC were generated using the methods **extractPAAC**, **extractQSO** and **extractAPAAC**, respectively, each with the λ parameter set to 7.

Factor encoding, as noted in previous sections is constructed by default therefore requiring no additional packages or methods.

The last encoding, the **BLOMAP** encoding, was constructed manually by mapping each amino acid to a feature vector. The encoded values for each amino acid were given by Maetschke et al. (2005) and are presented in Table B.1.

Amino Acid					
A	-0.57	0.39	-0.96	-0.61	-0.69
R	-0.4	-0.83	-0.61	1.26	-0.28
N	-0.7	-0.63	-1.47	1.02	1.06
D	-1.62	-0.52	-0.67	1.02	1.47
C	0.07	2.04	0.65	-1.13	-0.39
Q	-0.05	-1.50	-0.67	0.49	0.21
E	-0.64	-1.59	-0.39	0.69	1.04
G	-0.90	0.87	-0.36	1.08	1.95
H	0.73	-0.67	-0.42	1.13	0.99
I	0.59	0.79	1.44	-1.90	-0.93
L	0.65	0.84	1.25	-0.99	-1.90
K	-0.64	-1.19	-0.65	0.68	-0.13
M	0.76	0.05	0.06	-0.62	-1.59
F	1.87	1.04	1.28	-0.61	-0.16
P	-1.82	-0.63	0.32	0.03	0.68
S	-0.39	-0.27	-1.51	-0.25	0.31
T	-0.04	-0.3	-0.82	-1.02	-0.04
W	1.38	1.69	1.91	1.07	-0.05
Y	1.75	0.11	0.65	0.21	-0.41
V	-0.02	0.30	0.97	-1.55	-1.16

Table B.1: 5 dimensional BLOMAP encoding

Kernel Support Vector Machine

Kernel Support Vector Machine models presented in this thesis were generated using the **kernlab** package (Karatzoglou et al., 2004) in version **0.9.29**.

As for the estimation of the KSVM similarity kernels were generated based on a local alignment approach using the Smith-Waterman algorithm (Smith et al., 1981) and both, BLOSUM substitution matrices as well as a kernel similarity measure, proposed by Shen et al. (2012) were used. The two substitution matrices applied in this thesis, BLOSUM35 and BLOSUM62, were taken from https://ftp.ncbi.nih.gov/blast/matrices/ and are presented in Table B.2 and B.3, respectively. Gap opening and gap extension penalties have been set to 10 and 0.5, respectively the values of which where taken from the default parameters used by EMBL-EBI.

The third similarity measure used was proposed by Shen et al. (2012) and is based upon an extension of the above shown BLOSUM62 matrix. This extended form of the substitution matrix is referred to by the authors as BLOSUM62-2 and is presented in Table B.4.

In order to allow for faster computation of the Kernel Support Vector Machine models, all

pairwise similarities using BLOSUM62, BLOSUM35 and the proposed kernel similarity measure have been precomputed and are made publicly accessible alongside the raw data.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	5	-1	-1	-1	-2	0	-1	0	-2	-1	-2	0	0	-2	-2	1	0	-2	-1	0	-1	-1	0	-5
R	-1	8	-1	-1	-3	2	-1	-2	-1	-3	-2	2	0	-1	-2	-1	-2	0	0	-1	-1	0	-1	-5
N	-1	-1	7	1	-1	1	-1	1	1	-1	-2	0	-1	-1	-2	0	0	-2	-2	-2	4	0	0	-5
D	-1	-1	1	8	-3	-1	2	-2	0	-3	-2	-1	-3	-3	-1	-1	-1	-3	-2	-2	5	1	-1	-5
C	-2	-3	-1	-3	15	-3	-1	-3	-4	-4	-2	-2	-4	-4	-4	-3	-1	-5	-5	-2	-2	-2	-2	-5
Q	0	2	1	-1	-3	7	2	-2	-1	-2	-2	0	-1	-4	0	0	0	-1	0	-3	0	4	-1	-5
E	-1	-1	-1	2	-1	2	6	-2	-1	-3	-1	1	-2	-3	0	0	-1	-1	-1	-2	0	5	-1	-5
G	0	-2	1	-2	-3	-2	-2	7	-2	-3	-3	-1	-1	-3	-2	1	-2	-1	-2	-3	0	-2	-1	-5
H	-2	-1	1	0	-4	-1	-1	-2	12	-3	-2	-2	1	-3	-1	-1	-2	-4	0	-4	0	-1	-1	-5
I	-1	-3	-1	-3	-4	-2	-3	-3	-3	5	2	-2	1	1	-1	-2	-1	-1	0	4	-2	-3	0	-5
L	-2	-2	-2	-2	-2	-2	-1	-3	-2	2	5	-2	3	2	-3	-2	0	0	0	2	-2	-2	0	-5
K	0	2	0	-1	-2	0	1	-1	-2	-2	-2	5	0	-1	0	0	0	0	-1	-2	0	1	0	-5
M	0	0	-1	-3	-4	-1	-2	-1	1	1	3	0	6	0	-3	-1	0	1	0	1	-2	-2	0	-5
F	-2	-1	-1	-3	-4	-4	-3	-3	-3	1	2	-1	0	8	-4	-1	-1	1	3	1	-2	-3	-1	-5
P	-2	-2	-2	-1	-4	0	0	-2	-1	-1	-3	0	-3	-4	10	-2	0	-4	-3	-3	-1	0	-1	-5
S	1	-1	0	-1	-3	0	0	1	-1	-2	-2	0	-1	-1	-2	4	2	-2	-1	-1	0	0	0	-5
T	0	-2	0	-1	-1	0	-1	-2	-2	-1	0	0	0	-1	0	2	5	-2	-2	1	-1	-1	0	-5
W	-2	0	-2	-3	-5	-1	-1	-1	-4	-1	0	0	1	1	-4	-2	-2	16	3	-2	-3	-1	-1	-5
Y	-1	0	-2	-2	-5	0	-1	-2	0	0	0	-1	0	3	-3	-1	-2	3	8	0	-2	-1	-1	-5
V	0	-1	-2	-2	-2	-3	-2	-3	-4	4	2	-2	1	1	-3	-1	1	-2	0	5	-2	-2	0	-5
B	-1	-1	4	5	-2	0	0	0	0	-2	-2	0	-2	-2	-1	0	-1	-3	-2	-2	5	0	-1	-5
Z	-1	0	0	1	-2	4	5	-2	-1	-3	-2	1	-2	-3	0	0	-1	-1	-1	-2	0	4	0	-5
X	0	-1	0	-1	-2	-1	-1	-1	-1	0	0	0	0	-1	-1	0	0	-1	-1	0	-1	0	-1	-5
*	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1

Table B.2: Blosum 35

Model Evaluation

In order to compare the individual models, the package **pROC** (Robin et al., 2011) in version **1.15.3** has been used to compute the Area under the Curve for all generated models in each iteration.

Subsequent statistical evaluation has been conducted using the **rstatix** package (Kassambara, 2019) - version **0.3.1**.

From this package, first the normality test proposed by Shapiro and Wilk (1965) was computed using the **shapiro_test** method. The Repeated Measures ANOVA has been created using the **anova_test** method which also implicitly computed the Mauchly sphericity test (Mauchly, 1940) and adjusted the degrees of freedom using the correction proposed by Greenhouse - Geisser (Greenhouse and Geisser, 1959).

Lastly, the paired two-sample T-test was executed using the method **pairwise_t_test** setting the paired parameter to "True" and the adjustment method to "bonferroni".

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

Table B.3: Blosum 62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	3.9029	0.6127	0.5883	0.5446	0.8680	0.7568	0.7413	1.0569	0.5694	0.6325	0.6019	0.7754	0.7232	0.4649	0.7541	1.4721	0.9844	0.4165	0.5426	0.9365
R	0.6127	6.6656	0.8586	0.5732	0.3089	1.4058	0.9608	0.4500	0.9170	0.3548	0.4739	2.0768	0.6226	0.3807	0.4815	0.7672	0.6778	0.3951	0.5560	0.4201
N	0.5883	0.8586	7.0941	1.5539	0.3978	1.0006	0.9113	0.8637	1.2220	0.3279	0.3100	0.9398	0.4745	0.3543	0.4999	1.2315	0.9842	0.2778	0.4860	0.3690
D	0.5446	0.5732	1.5539	7.3979	0.3015	0.8971	1.6878	0.6343	0.6786	0.3390	0.2866	0.7841	0.3465	0.2990	0.5987	0.9135	0.6948	0.2321	0.3457	0.3365
C	0.8680	0.3089	0.3978	0.3015	19.5766	0.3658	0.2859	0.4204	0.3550	0.6535	0.6423	0.3491	0.6114	0.4390	0.3796	0.7384	0.7406	0.4500	0.4342	0.7558
Q	0.7568	1.4058	1.0006	0.8971	0.3658	6.2444	1.9017	0.5386	1.1680	0.3829	0.4773	1.5543	0.8643	0.3340	0.6413	0.9656	0.7913	0.5094	0.6111	0.4668
E	0.7413	0.9608	0.9113	1.6878	0.2859	1.9017	5.4695	0.4813	0.9600	0.3305	0.3729	1.3083	0.5003	0.3307	0.6792	0.9504	0.7414	0.3743	0.4965	0.4289
G	1.0569	0.4500	0.8637	0.6343	0.4204	0.5386	0.4813	6.8763	0.4930	0.2750	0.2845	0.5889	0.3955	0.3406	0.4774	0.9036	0.5793	0.4217	0.3487	0.3370
H	0.5694	0.9170	1.2220	0.6786	0.3550	1.1680	0.9600	0.4930	13.5060	0.3263	0.3807	0.7789	0.5841	0.6520	0.4729	0.7367	0.5575	0.4441	1.7979	0.3394
I	0.6325	0.3548	0.3279	0.3390	0.6535	0.3829	0.3305	0.2750	0.3263	3.9979	1.6944	0.3964	1.4777	0.9458	0.3847	0.4432	0.7798	0.4089	0.6304	2.4175
L	0.6019	0.4739	0.3100	0.2866	0.6423	0.4773	0.3729	0.2845	0.3807	1.6944	3.7966	0.4283	1.9943	1.1546	0.3711	0.4289	0.6603	0.5680	0.6921	1.3142
K	0.7754	2.0768	0.9398	0.7841	0.3491	1.5543	1.3083	0.5889	0.7789	0.3964	0.4283	4.7643	0.6253	0.3440	0.7038	0.9319	0.7929	0.3589	0.5322	0.4565
M	0.7232	0.6226	0.4745	0.3465	0.6114	0.8643	0.5003	0.3955	0.5841	1.4777	1.9943	0.6253	6.4815	1.0044	0.4239	0.5986	0.7938	0.6103	0.7084	1.2689
F	0.4649	0.3807	0.3543	0.2990	0.4390	0.3340	0.3307	0.3406	0.6520	0.9458	1.1546	0.3440	1.0044	8.1288	0.2874	0.4400	0.4817	1.3744	2.7694	0.7451
P	0.7541	0.4815	0.4999	0.5987	0.3796	0.6413	0.6792	0.4774	0.4729	0.3847	0.3711	0.7038	0.4239	0.2874	12.8375	0.7555	0.6889	0.2818	0.3635	0.4431
S	1.4721	0.7672	1.2315	0.9135	0.7384	0.9656	0.9504	0.9036	0.7367	0.4432	0.4289	0.9319	0.5986	0.4400	0.7555	3.8428	1.6139	0.3853	0.5575	0.5652
T	0.9844	0.6778	0.9842	0.6948	0.7406	0.7913	0.7414	0.5793	0.5575	0.7798	0.6603	0.7929	0.7938	0.4817	0.6889	1.6139	4.8321	0.4309	0.5732	0.9809
W	0.4165	0.3951	0.2778	0.2321	0.4500	0.5094	0.3743	0.4217	0.4441	0.4089	0.5680	0.3589	0.6103	1.3744	0.2818	0.3853	0.4309	38.1078	2.1098	0.3745
Y	0.5426	0.5560	0.4860	0.3457	0.4342	0.6111	0.4965	0.3487	1.7979	0.6304	0.6921	0.5322	0.7084	2.7694	0.3635	0.5575	0.5732	2.1098	9.8322	0.6580
V	0.9365	0.4201	0.3690	0.3365	0.7558	0.4668	0.4289	0.3370	0.3394	2.4175	1.3142	0.4565	1.2689	0.7451	0.4431	0.5652	0.9809	0.3745	0.6580	3.6922

Table B.4: BLOSUM 62-2