

# Machine Learning Prediction of TCR-Epitope Binding

*Julian Faust  
Yun S. Song*



Electrical Engineering and Computer Sciences  
University of California, Berkeley

Technical Report No. UCB/EECS-2022-216

<http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-216.html>

August 17, 2022

Copyright © 2022, by the author(s).  
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

---

# Machine Learning Prediction of TCR-Epitope Binding

by Julian Faust

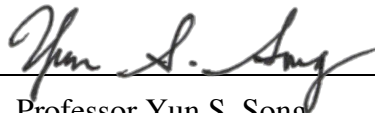
---

## Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences,  
University of California at Berkeley, in partial satisfaction of the requirements for the  
degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

### Committee:



---

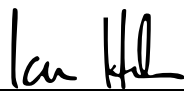
Professor Yun S. Song  
Research Advisor

August 10, 2022

---

(Date)

\* \* \* \* \*



---

Professor Ian Holmes  
Second Reader

August 17, 2022

---

(Date)

Machine Learning Prediction of TCR-Epitope Binding

by

Julian Faust

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Science

in

Electrical Engineering and Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Yun S. Song, Chair

Professor Ian Holmes

Summer 2022

Abstract

Machine Learning Prediction of TCR-Epitope Binding

by

Julian Faust

Master of Science in Electrical Engineering and Computer Science

University of California, Berkeley

Professor Yun S. Song, Chair

Prediction of T-cell receptor (TCR) binding with peptide-MHC complexes remains a difficult problem due to data accuracy, data scarceness, and problem complexity. Here, we compare predictions of TCR-pMHC binding across several approaches of featurizing the TCR, and several different machine learning methods. First, we analyze the available data and discuss the formulation of binder/non-binder designations for our binary classification framework. Next, we compare several featurizations of the TCR across different machine learning methods of varying complexity. We provide an ablation study across different region combinations common in cases with limited data. We show that simpler machine learning methods trained on binders and non-binders of a single epitope can be used to better understand binding factors. Our attention-based neural network directly incorporates peptide and MHC sequence information, and performs similarly on the harder problem of training with binders and non-binders of many epitopes at once. Lastly, we incorporate gene usage data into our prediction framework.

# Contents

<b>Contents</b>	<b>i</b>
<b>List of Figures</b>	<b>ii</b>
<b>List of Tables</b>	<b>iii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Problem Description . . . . .	1
1.3 Related Work . . . . .	2
1.4 Contributions . . . . .	3
<b>2 TCR-Epitope Prediction</b>	<b>4</b>
2.1 Data . . . . .	4
2.2 Methods . . . . .	9
2.3 Results . . . . .	13
<b>3 Conclusion</b>	<b>28</b>
3.1 Summary . . . . .	28
3.2 Future Work . . . . .	28
<b>4 Declarations</b>	<b>30</b>
4.1 Acknowledgements . . . . .	30
4.2 Author Contributions . . . . .	30
<b>Bibliography</b>	<b>31</b>

# List of Figures

2.1	dCODE Dextramer Reagent, figure taken from Immudex [10]	5
2.2	KLGGALQAK UMI Count Distribution Histogram, $\text{Min}(U_1, \dots, U_n) > 10$ and $\text{Min}(U_1, \dots, U_n) > 5 \times \text{Max}(\text{Negative Controls})$	7
2.3	GILGFVFTL UMI Count Distribution Histogram, $\text{Min}(U_1, \dots, U_n) > 10$ and $\text{Min}(U_1, \dots, U_n) > 5 \times \text{Max}(\text{Negative Controls})$	8
2.4	GILGFVFTL Binders Position-Frequency For Different Length CDR3 $\alpha$	9
2.5	GILGFVFTL CDR3 $\beta$ Length Distributions	10
2.6	CDR Attention Model Architecture	13
2.7	GILGFVFTL Random Forests Mean Decrease in Impurity Feature Importance	15
2.8	KLGGALQAK Binder, GILGFVFTL Binder, All TCR AHo Numbered CDR3B Region Amino Acid Distributions	18
2.9	KLGGALQAK Binder, GILGFVFTL Binder AHo Numbered CDR3B Region Amino Acid Log Enrichment	19
2.10	KLGGALQAK Binder, GILGFVFTL Binder AHo Numbered CDR3B Region Amino Acid Log Enrichment, Fisher's Exact Test $< 0.01$	20
2.11	KLGGALQAK Binder, GILGFVFTL Binder AHo Numbered CDR3B Region Amino Acid Log Enrichment, Fisher's Exact Test $< 0.0001$	21
2.12	KLGGALQAK Binders V $\beta$ Gene Distribution	22
2.13	GILGFVFTL Binders V $\beta$ Gene Distribution	23
2.14	General TCR V $\beta$ Gene Distribution	23
2.15	GILGFVFTL Binders Length 14 CDR3 $\alpha$ TRAV27 and TRBV19 Position-Frequency	24
2.16	GILGFVFTL Non-Binders Length 14 CDR3 $\alpha$ TRAV27 and TRBV19 Position-Frequency	25
2.17	GILGFVFTL Binders Length 13 CDR3 $\beta$ TRAV27 and TRBV19 Position-Frequency	25
2.18	GILGFVFTL Binders Length 13 CDR3 $\beta$ TRAV25 and TRBV19 Position-Frequency	27

# List of Tables

2.1	UMI Distribution Statistics . . . . .	6
2.2	UMI Distribution Statistics, $\text{Min}(U_1, \dots, U_n) > 10$ and $\text{Min}(U_1, \dots, U_n) > 5 \times$ Max(Negative Controls) . . . . .	6
2.3	UMI Difference in Identical TCR Sequences Distribution Statistics . . . . .	7
2.4	UMI Difference in Identical TCR Sequences Distribution Statistics, $\text{Min}(U_1, \dots, U_n)$ $> 10$ and $\text{Min}(U_1, \dots, U_n) > 5 \times \text{Max}(\text{Negative Controls})$ . . . . .	8
2.5	UMI Threshold For Binders and Counts . . . . .	10
2.6	Classification Results Regional Features Amino Acid Counts . . . . .	14
2.7	Classification Results Regional Features VHSE Descriptors . . . . .	14
2.8	Classification Results Regional Features VHSE Descriptors Random Forests . .	16
2.9	Classification Results Positional Features VHSE Descriptors . . . . .	16
2.10	Classification Results Attention Model . . . . .	17
2.11	Classification Results Gene Usage Features . . . . .	22
2.12	Average Pairwise Edit Distance for Binders . . . . .	26
2.13	Classification Results Regional Features VHSE Descriptors Random Forest Ac- curacy, TRAV27 and TRBV19 Genes Fixed . . . . .	27



# Chapter 1

## Introduction

### 1.1 Background

T lymphocytes, also known as T cells, are crucial in the cellular immune response [1]. T cell receptors (TCRs) are two-chained ( $\alpha$  and  $\beta$ ) protein complexes on the surface of T cells. They are responsible for recognizing peptide antigens (epitopes) presented on a Major Histocompatibility Complex (MHC). Although the likelihood that a random TCR will bind a random epitope is very low, many different TCRs can recognize the same antigen peptide and many antigen peptides can be recognized by the same TCR. The complementarity-determining region 1 (CDR1) and CDR2 loops of the TCR  $\alpha$  and  $\beta$  chains contact specific regions of the MHC while the hypervariable complementary determining regions (CDR3) interact mainly with the peptide [1]. CDR3 $\alpha$  and CDR3 $\beta$  loops have the highest sequence diversity and are the principal determinants of binding. Accurate prediction of TCR-epitope binding would accelerate the development of numerous therapeutics. and potentially have major implications in cancer and immune research [1]. Many new TCR-based diagnostic and rational immunotherapy design methods would become viable, improving our ability to treat many diseases.

### 1.2 Problem Description

The TCR-pMHC binary classification problem is presented as follows: given the amino acid sequences of a peptide, MHC, and TCR, classify instances as either binders or non-binders. The data was published by 10x Genomics, and is generated from a highly multiplexed experiment with many different pMHC multimers [2]. For each TCR and for each of the 50 pMHC's (6 of which are negative controls), there is an associated integer UMI count value that can be used to distinguish binders from non-binders. UMI is an acronym for Unique Molecular Identifier. These UMI values correspond to some unknown degree with binding affinity. 10x Genomics suggests a threefold decision rule, labeling a binder if a UMI value meets the following criteria: "a UMI count greater than 10 that was also greater than five

times the highest negative control UMI count for that cell. In cases where a cell was assigned more than one specificity, we considered it to be specific only for the pMHC with the highest UMI count.” As there are clones in the data, we have  $n$  TCRs with the same given sequence and let  $U_1, \dots, U_n$  denote their UMI values. Due to variation in  $U_1, \dots, U_n$ , there could be ambiguous binder/non-binder labeling depending on which copy of the TCR sequence is chosen. In order to resolve this ambiguity, we propose a new labeling system based on the following rule: Binders are TCRs with  $\min(U_1, \dots, U_n)$  greater than some pMHC-specific cutoff value and greater than the maximum of all negative control values across the 6 negative controls and  $n$  TCRs. Non-binders are those with  $\max(U_1, \dots, U_n) = 0$ . For binary classification, we ensure that there are no duplicate sequences in the combined training and test sets.

### 1.3 Related Work

There have been many attempts to apply machine learning to predict TCR-pMHC binding, using a variety of approaches. Many papers in this domain have focused on predicting epitope specificity, which is an easier prediction problem since TCRs are filtered so that only TCRs binding to a single epitope across the set (epitope-specific TCRs) are included [3]. There have also been attempts at building a single model which generalizes TCR-epitope predictions across a set of different epitopes. In 2020, Springer et al. [4] attempted this task with autoencoder and LSTM-based approaches to achieve a 0.81 AUROC on a test set of unseen TCRs for a dataset of multiple epitopes presented on the same MHC complex. It is key to note that their network was trained on over 200,000 TCRs, using only CDR3 chains as input. This is because most available TCR data (from McPAS or VDJdb databases) provide only the CDR3 sequences, despite good evidence that using CDR1 and CDR2 sequence information improves the predictive accuracy of various models [5]. Several papers have attempted this task using AUROC as their performance metric despite an uneven positive to negative class ratio (such as 1:5), a choice that compromises the validity of the results since the model could predict predominantly negatives to achieve a high AUROC [4][5][6].

Many papers have used the same 10x Genomics TCR-pMHC single-cell dataset with UMI count values characterizing the binding of TCRs to 44 specific pMHC multimers and 6 negative controls. Fischer et al. [7] opted to use the binder designations provided by 10x Genomics. Sidhom et al. [8] proposed to use their DeepTCR network to regress UMI counts, which they considered a proxy for binding affinity. Zhang et al. [9] proposed to handle multiple UMI count values associated with a single unique TCR sequence by using the median UMI value.

## 1.4 Contributions

In this work, we compare various approaches for TCR-pMHC binding prediction. We structure this thesis as follows:

- In Section 2.1, we provide a detailed analysis of the data and the potential challenges in designating accurate binary labels.
- In Section 2.2, we present an overview of the different featurization schemes and machine learning methods used.
- In Section 2.3, we demonstrate the performance of our models, comparing results across various methods.

## Chapter 2

# TCR-Epitope Prediction

## 2.1 Data

### UMI Values

The degree to which a TCR-pMHC pair's UMI value is correlated with the binding affinity is unknown. In order to better understand the TCR-pMHC UMI data, we proceed with a brief description of the multiplexed binding and sequencing experiment used to generate our data.

First, a pool of dCODE Dextramer reagents is created. Each reagent, as depicted in Figure 2.1 (created by Immudex), is composed of a flexible dextran backbone with coupled fluorophores and pMHC complexes [10]. The number of pMHC complexes loaded onto the dextran backbone was optimized by Immudex to increase the avidity of the interaction between the reagent and interacting TCRs while minimizing the effect of other dCODE Dextramer reagents. There are multiple fluorophores that boost the brightness of the reagent and improve the signal-to-noise ratio. There is also a barcode for use in bulk or single-cell sequencing. Each reagent has a unique pMHC complex and a unique barcode. Equal amounts of each reagent (160 nM) are combined in the pool so they are equally represented in the solution according to the Immudex cell staining profile. T cells are then stained with this pool of reagents. Fluorescence-activated cell sorting (FACS) antibodies and sequencing antibodies are added at this phase. FACS antibodies apply a gating strategy and effectively work as cell sorters. They are able to separate the cells specific to each dCODE Dextramer reagent with its distinct phycoerythrin (PE) fluorophore. Since sequencing is costly and 10x sequencing instruments can only load around 10,000 single cells at a time, this step is important to ensure that primarily antigen-specific cells are sequenced. Then, the 10x Chromium microfluidic system is used to load single cells into gel beads in emulsion (GEMs), each of which contains a single barcoded 10x bead and a single cell. After the cell is lysed, the gel bead is coated in the barcodes, which can be isolated and sequenced. On each individual barcode on each individual reagent, there is a unique UMI. After sequencing, the number of UMIs associated with a T cell is the number of molecules of the reagent that were bound to that T cell.

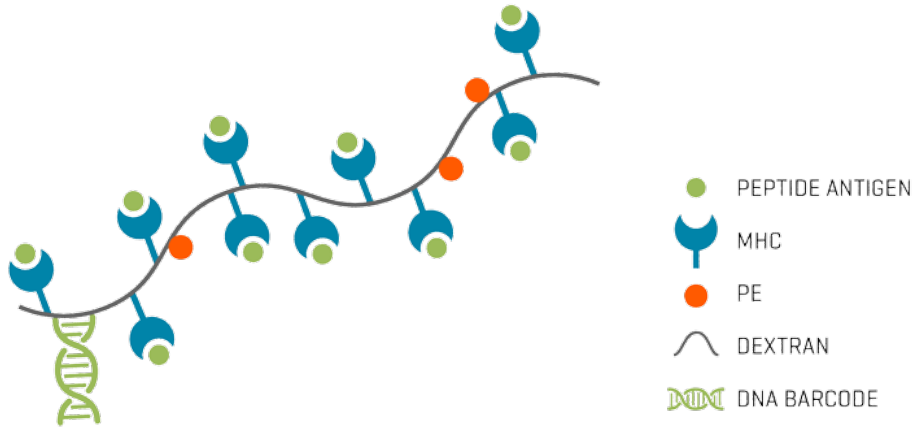


Figure 2.1: dCODE Dextramer Reagent, figure taken from Immudex [10]

The number of dCODE Dextramer reagents that bind to a T cell is dependent on multiple factors, including the avidity (correlated with the affinity of the interaction between the pMHC and TCR) as well as the number of TCRs on the surface (correlated with T cell activation status). For these reasons and any other experimental noise that may be present, the number of UMIs is not directly correlated with the affinity of the TCR-pMHC interaction alone.

In order to differentiate cases of non-specific binders from true binders, it is recommended to make use of negative control dCODE Dextramer reagents. These can be MHC allele-matched (preferably) or allele mismatched to experimental reagents, and carry irrelevant peptides to the samples that should not result in binding. Binding to negative controls can be used to approximate non-specific binding in a cell, so any instance of specific binding should have a significantly higher UMI count than the maximum negative control UMI count.

We filter sequences based on “Productive” and “High Confidence” filters provided by 10x Genomics, including only cells with exactly one  $TCR\alpha$  sequence and one  $TCR\beta$  sequence. After filtering, there are a total of 74,320 paired TCRs with UMI values for 50 pMHCs (including negative controls), and 28,788 unique paired TCR sequences. The UMI distributions vary by pMHC in shape and scale. In Table 2.1, we provide summary statistics for the UMI distributions of unique TCR sequences for the 9 epitopes in the dataset with sufficient binders to be analyzed with our methods. For each unique TCR sequence with  $n$  associated UMI values,  $\min(U_1, \dots, U_n)$  is used as the representative value. As zero counts and lower UMI values dominate the data, we present the same results in Table 2.2 conditioned on  $\min(U_1, \dots, U_n)$  being greater than 10, and greater than the maximum of all negative control values across the 6 negative controls and  $n$  TCRs with the same sequence.

Table 2.1 and Table 2.2 illustrate the vast differences in UMI value distributions across epitopes. In Figure 2.2 and Figure 2.3, we present histograms for the UMI distributions of CMV epitope KLGALQAK (bound to HLA-A\*0301) and Influenza peptide GILGFVFTL

Table 2.1: UMI Distribution Statistics

PEPTIDE SEQ.	COUNT	UMI MEAN	UMI MEDIAN	UMI MODE	UMI STD.
AVFDRKSDAK	28788	3.9	1	0	9.6
ELAGIGILTV	28788	0.2	0	0	2.2
FLYALALLL	28788	0.1	0	0	1.4
GILGFVFTL	28788	1.7	0	0	14.3
GLCTLVAML	28788	0.1	0	0	2.8
IVTDFSVIK	28788	2.9	0	0	17.5
KLGGALQAK	28788	7.1	1	0	14.3
RAKFKQLL	28788	1.1	0	0	9.7
RLRAEAQVK	28788	2.6	0	0	5.4

Table 2.2: UMI Distribution Statistics,  $\text{Min}(U_1, \dots, U_n) > 10$  and  $\text{Min}(U_1, \dots, U_n) > 5 \times \text{Max}(\text{Negative Controls})$ 

PEPTIDE SEQ.	COUNT	UMI MEAN	UMI MEDIAN	UMI MODE	UMI STD.
AVFDRKSDAK	3263	22.7	19	11	18.9
ELAGIGILTV	147	25.8	19	16	16.2
FLYALALLL	25	41.8	37	26	22.9
GILGFVFTL	530	86.0	71	31	62.6
GLCTLVAML	55	46.0	29	43	43.7
IVTDFSVIK	1663	29.0	16	11	67.2
KLGGALQAK	5158	31.5	26	11	19.7
RAKFKQLL	433	63.3	49	13	48.0
RLRAEAQVK	2079	18.4	16	11	7.6

(bound to HLA-A\*0201). For certain epitopes like GILGFVFTL, the majority of TCRs with a UMI greater than 10 have a UMI much greater than 10. For other epitopes like KLGGALQAK, a large proportion of TCRs with a UMI greater than 10 have a UMI only slightly greater than 10. Due to the shape of this distribution, this holds true not just for the threshold value of 10 in particular, but for any reasonable threshold value.

## Analysis of Clones

The majority of the unique sequences in the data have only a single copy. These cases make up 26019 out of the 28788 unique TCR sequences in the dataset. The maximum number of copies of a unique TCR sequence was 5669. In order to understand the degree of variance in UMI values, we sample the distribution of the difference in UMI value between any two clones (TCRs with identical sequences) for each epitope. We present some summary statistics for these distributions in Table 2.3. We also present summary statistics for these distributions in Table 2.4, this time conditioning on  $n$  TCRs having  $\text{min}(U_1, \dots, U_n)$  being greater than 10, and greater than the maximum of all negative control values across the 6 negative controls. It is important to note that TCR sequences with more copies are overrepresented

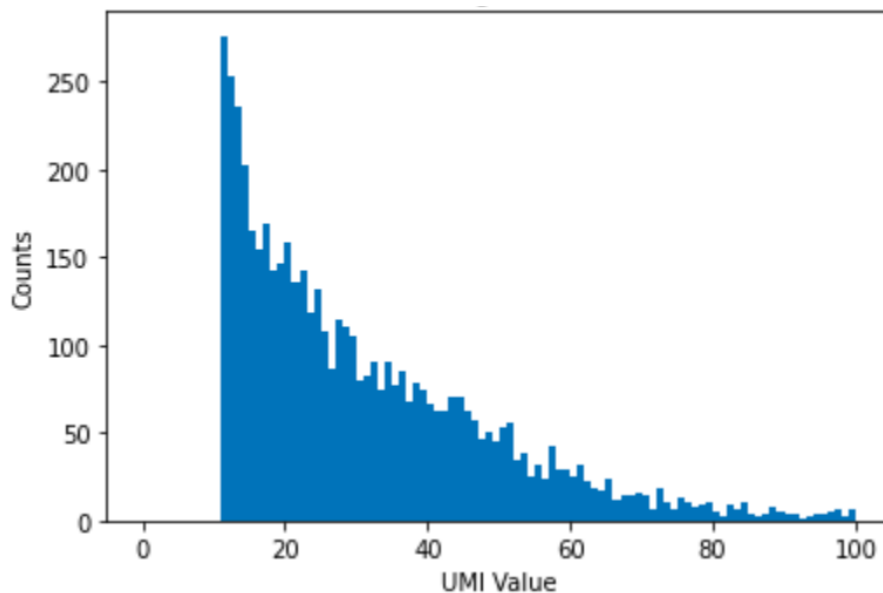


Figure 2.2: KLGALQAK UMI Count Distribution Histogram,  $\text{Min}(U_1, \dots, U_n) > 10$  and  $\text{Min}(U_1, \dots, U_n) > 5 \times \text{Max}(\text{Negative Controls})$

in the distribution of pairwise identical sequence UMI differences, as the number of pairwise differences possible is quadratic in the number of copies of a TCR sequence. Since the degree of TCR expansion may correlate with the UMI values of some epitopes, the UMI values from which the differences are computed may be higher than those from Table 2.1 and Table 2.2.

Table 2.3: UMI Difference in Identical TCR Sequences Distribution Statistics

PEPTIDE SEQ.	COUNT	UMI MEAN	UMI MEDIAN	UMI MODE	UMI STD.
AVFDRKSDAK	20023912	2.8	1	0	9.0
ELAGIGILTV	20023912	0.1	0	0	0.3
FLYALALLL	20023912	0.1	0	0	0.4
GILGFVFTL	20023912	1.1	0	0	9.5
GLCTLVAML	20023912	0.1	0	0	1.1
IVTDFSVIK	20023912	14.9	1	0	65.1
KLGGALQAK	20023912	4.2	2	1	5.6
RAKFKQLL	20023912	31.7	23	0	33.0
RLRAEAQVK	20023912	2.0	1	0	2.9

There are many epitopes for which the actual UMI values and the difference in UMI values among identical TCRs are of similar magnitude. This highlights the high variance of UMI values and suggests that the correlation with binding affinity is quite weak in the

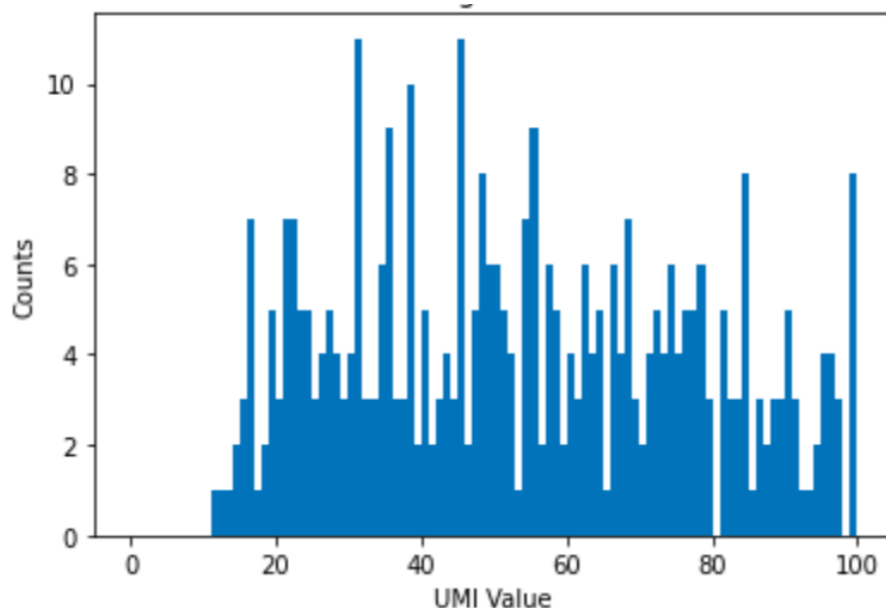


Figure 2.3: GILGFVFTL UMI Count Distribution Histogram,  $\text{Min}(U_1, \dots, U_n) > 10$  and  $\text{Min}(U_1, \dots, U_n) > 5 \times \text{Max}(\text{Negative Controls})$

Table 2.4: UMI Difference in Identical TCR Sequences Distribution Statistics,  $\text{Min}(U_1, \dots, U_n) > 10$  and  $\text{Min}(U_1, \dots, U_n) > 5 \times \text{Max}(\text{Negative Controls})$

PEPTIDE SEQ.	COUNT	UMI MEAN	UMI MEDIAN	UMI MODE	UMI STD.
AVFDRKSDAK	3269	231.9	180	6	197.8
ELAGIGILTV	0	0	0	0	0
FLYALALLL	953	33.4	28	7	25.1
GILGFVFTL	99358	87.6	51	1	528.3
GLCTLVAML	2484	33.3	24	14	32.1
IVTDFSVIK	187	167.0	107	8	174.4
KLGGALQAK	884	28.1	18	2	31.0
RAKFKQLL	19002	53.5	41	11	45.9
RLRAEAQVK	79	10.7	7	2	11.3

case of certain epitopes. This variance is problematic when determining a binding threshold value for epitopes with UMI distributions similar to KLGGALQAK. Due to the distribution shape illustrated in Figure 2.2, a large proportion of UMI values greater than any reasonable threshold value  $x$  will be very close to  $x$ . For this epitope, the median UMI value with conditioning from Table 2.2 is 26, while the median UMI difference of identical TCR sequences with conditioning from Table 2.4 is 18. From our analyses, including the suggested criteria for negative controls does not significantly alter the UMI distributions or reduce the UMI



variance of TCRs with the same sequence.

## Varying Region Lengths and the Alignment Problem

We describe a positional alignment issue, imposed by the length variations of different regions of the TCR. We found some degree of length variation in every CDR and Framework region. The greatest diversity in lengths occurs in the CDR3 regions, due to non-templated insertions and deletions during V(D)J recombination. Furthermore, the position-amino acid frequency distributions of CDRs of different lengths are similar at or very close to the ends of the sequence. The position-frequency maps in Figure 2.4 suggest gap character insertion(s) somewhere in the middle of the CDR3 $\alpha$  might better align the positions with similar amino acid distributions. The AHo alignment scheme tends to align CDR regions in this way, with the ends aligned and gap characters from the middle out [11]. However, many solutions to featurizing the TCR have simply left, middle or right padded to deal with the varying lengths, which seems to improperly align the positions [8][12]. While aligning CDRs of different lengths seems natural, the amino acid distributions still do not align perfectly with the insertion of a gap character in the optimal location(s). The length of the CDR sequence is also closely related to the loop function, and certain lengths are “preferred” by positive binders of certain epitopes. An example of this can be seen in the histogram in Figure 2.5. Ideally, there would be enough data that a different distribution could be learned for each different set of lengths.

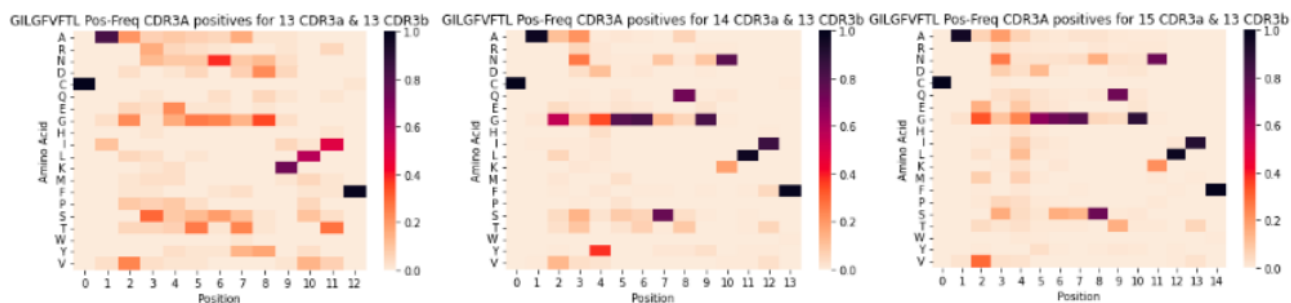


Figure 2.4: GILGFVFTL Binders Position-Frequency For Different Length CDR3 $\alpha$

## 2.2 Methods

For each epitope, equal numbers of positive and negative binders are used for training and testing. There are no duplicate sequences in the combined training and test sets. The train/test ratio is 80/20, and 5-fold cross-validation was used to select the hyperparameters for the Random Forest Classifier. Logistic Regression with L2 Regularization was chosen. In

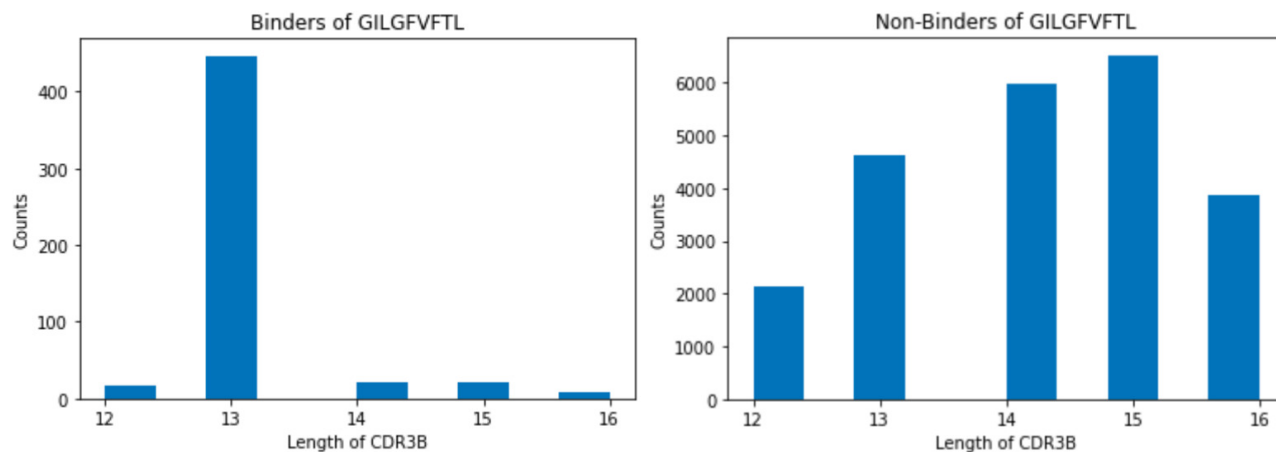
Figure 2.5: GILGFVFTL CDR3 $\beta$  Length Distributions

Table 2.5, we present the pMHC threshold values used to determine positive binders along with the maximum negative control condition described earlier.

Table 2.5: UMI Threshold For Binders and Counts

PEPTIDE SEQ.	UMI THRESHOLD	POSITIVES
AVFDRKSDAK	30	139
ELAGIGILTV	10	147
FLYALALLL	10	25
GILGFVFTL	20	504
GLCTLVAML	15	47
IVTDFSVIK	30	164
KLGGALQAK	30	2123
RAKFKQLL	30	312
RLRAEAQVK	30	147

## Regional Methods

With limited binders, we first turn towards two simple featurizations that aggregate information across each of the 6 CDR and 8 framework regions. While the positional information within a region is lost with these featurizations, they contain important regional information and set a prediction baseline for position-based methods. With the following featurization methods, we train a separate Random Forest and Logistic Regression classifier for the positive/negative binders of each epitope.

### Amino Acid Counts

The first featurization method we used is the amino acid counts for each of the 14 regions, along with the region length. The final feature vector has 294 features, the counts for 20 amino acids in each of the 14 regions, plus 14 features corresponding to the lengths of those regions. We also experimented with a closely related featurization method, which divides these count values by the region length (each position now represents the frequency of the amino acid within the region). An advantage of using amino acid count features is that the feature importances are more easily interpretable than with amino acid descriptors. This method yielded nearly identical results to the amino acid counts method.

### Amino Acid Descriptors

The second featurization method used the set of 8 amino acid descriptors called VHSE [13]. We computed the average VHSE vector for each region, along with the lengths of the 14 regions as before. The final feature vector has 126 features. Amino acid descriptors require fewer features than amino acid counts and encode similarities between amino acids. This is helpful for generalization to unseen TCRs.

## Positional Methods

In this section, we used ANARCI to align TCRs according to the AHO numbering scheme [11][14]. The AHO numbering system is based on the spatial alignment of known three-dimensional structures of immunoglobulin domains and places alignment gaps in a way that minimizes the average deviation from the averaged structure of the aligned domains. AHO numbering aligns the start positions and end positions of different length CDR3 regions, adding gaps from the middle. Although different length CDRs can have different amino acid distributions that may not be perfectly alignable with a properly placed gap character, AHO numbering seems to be an appropriate way of aligning CDR3 regions based on the visual presented in Figure 2.4. The AHO aligned sequence length (with gaps) is 150 for each complete TCR chain, for a total of 300 positions. Of these 300 positions, 140 are in CDR regions while 160 are in framework regions.

### Amino Acid Descriptors

For each position, we represent each amino acid by a vector of its corresponding 8 VHSE descriptors. For gap characters, we use a vector of all zeros. As there are significantly more features without any aggregation over the region, we use only the 140 CDR positions. The final feature vector has a total of 1120 features.

## Attention Models

In addition to applying these simpler machine learning methods with separate models for each epitope, we wanted to train a neural network that works across all the epitopes. Our network trains on triplets of TCR, epitope, and MHC sequences, along with a label 1 for positive binders and a label 0 for negative binders. Attention-based architectures are chosen based on the high-level principle of the function of attention layers. The UMI thresholds, train/test, and positive/negative ratios remain the same from the earlier experiments.

Attention layers generate attention maps of scores between 0 and 1, representing the interaction of positions of the first input and the second. In our case, these are maps representing the interaction between CDR sequence and epitope sequence or maps representing the interaction between CDR sequence and MHC sequence. We use “multi-headed” attention layers rather than simply scaled dot product attention. Consider attention in the context where the first input of the attention layer is both the key and the value vector, and the second input is the query vector. We have query, key, and value weight matrices  $W^Q$ ,  $W^K$ ,  $W^V$ , respectively. These are multiplied by the input key, query, and value vectors. Scaled dot product attention is described by the equation  $Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$ , where  $\sqrt{d_k}$  is the dimension of the key vector  $k$  and query vector  $q$ . Multi-head attention modifies this in the following way:  $MultiHead(Q, K, V) = Concat(head_1, \dots, head_n)W^O$ , where  $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$  [15]. As these operations are asymmetric, we implement all attention layers a second time with the first and second input order reversed. Additionally, we upsample positive examples in the training set to train on more diverse negatives, which we found to increase performance slightly. While there is a 1:1 ratio of positive to negative labels in the training data, the ratio of unique positive samples to unique negative samples is 1:3. There is no upsampling for the test set, where positives and negatives are represented at a 1:1 ratio as before. We tested several neural network architectures that use attention between features from the TCR and features from the epitope or MHC. The sequences for the MHC alleles were found using the IPD-IMGT/HLA database [16]. We present only our best-performing network.

Our network takes in the TCR as 6 separate inputs, each of which corresponds to a CDR sequence. Each CDR is aligned with AHO numbering, and each amino acid position is encoded by its VHSE descriptors, where gap characters are represented by all zeros. This injects some domain knowledge into our design since CDR1 and CDR2 loops are known to contact the MHC, while the CDR3 loops are in contact with the peptide. The output of all 12 attention layers is flattened before being concatenated and passed through a fully connected network. The architecture is shown in Figure 2.6. Since there are 6 CDR loops, there are 6 attention layers for which the CDR is the first input and 6 attention layers for which the CDR is the second input for a total of 12 attention layers (rather than the 6 shown in the simplified Figure 2.6). The network has a total of 136,745 parameters.

The attention layers were implemented with the Keras MultiHeadAttention layers, with 3 attention heads and  $key\_dim = 3$ . The output of all attention layers is then flattened, concatenated, and then fed into a fully connected network. The network is trained with the

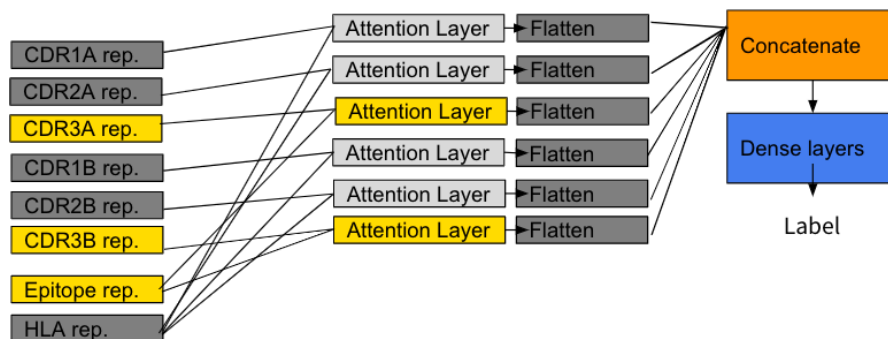


Figure 2.6: CDR Attention Model Architecture

Adam optimizer and standard binary cross-entropy loss for 3 epochs, with the following sizes for fully connected layers: [100, 40, 1].

The CDR, peptide, and MHC sequences were featurized by converting each amino acid position with a vector of its corresponding amino acid descriptors (VHSE). While the total length of the main MHC chain is hundreds of amino acids long, we only consider a discontinuous slice of 60 positions that previous studies have shown to interact with the TCR and peptide [17]. We experimented with featurizing each amino acid in the TCR, peptide, and MHC with one-hot encoding and ProtBERT-BFD embeddings, and found that the VHSE descriptors performed better. Sin-cos positional encoding is added to all features before passing into attention layers.

## 2.3 Results

We use accuracy and area under the precision-recall curve (AUC-PR) on the test set as measures of the predictive ability of our models. Since there is a 1:1 ratio of binders to non-binders in the training and test sets for every epitope, accuracy is a valid metric to consider. We provide accuracy and AUC-PR across all epitopes for different featurizations and methods. As we have ample negatives, we average accuracy and AUC-PR values across results obtained with 3 different non-overlapping sets of non-binders.

### Regional Methods

#### Amino Acid Counts

The overall Random Forests accuracy with amino acid counts was 0.761, and the overall Logistic Regression accuracy with amino acid counts was 0.755.

Table 2.6: Classification Results Regional Features Amino Acid Counts

PEPTIDE SEQ.	RF ACC.	LR ACC.	RF AUC-PR	LR AUC-PR
AVFDRKSDAK	0.526	0.561	0.621	0.533
ELAGIGILTV	0.949	0.887	0.948	0.945
FLYALALLL	0.833	0.933	0.883	0.988
GILGFVFTL	0.954	0.906	0.964	0.964
GLCTLVAML	0.929	0.895	0.942	0.961
IVTDFSVIK	0.666	0.621	0.796	0.766
KLGGALQAK	0.609	0.606	0.728	0.643
RAKFKQLL	0.853	0.829	0.893	0.921
RLRAEAQVK	0.531	0.559	0.618	0.533
OVERALL	0.761	0.755	0.822	0.807

### Amino Acid Descriptors

The overall Random Forests accuracy with amino acid descriptors was 0.773, and the overall Logistic Regression accuracy with amino acid descriptors was 0.761.

Table 2.7: Classification Results Regional Features VHSE Descriptors

PEPTIDE SEQ.	RF ACC.	LR ACC.	RF AUC-PR	LR AUC-PR
AVFDRKSDAK	0.515	0.570	0.606	0.520
ELAGIGILTV	0.949	0.881	0.948	0.939
FLYALALLL	0.900	0.900	0.928	0.973
GILGFVFTL	0.950	0.909	0.963	0.966
GLCTLVAML	0.930	0.877	0.939	0.975
IVTDFSVIK	0.657	0.687	0.785	0.777
KLGGALQAK	0.589	0.588	0.715	0.602
RAKFKQLL	0.853	0.848	0.894	0.930
RLRAEAQVK	0.616	0.576	0.686	0.521
OVERALL	0.773	0.761	0.829	0.800

While Random Forests and Logistic Regression generally seem to predict the same epitopes with similar accuracy, Random Forests has a slightly higher overall accuracy and AUC-PR. Average region VHSE descriptors also slightly outperformed region amino acid counts. Both Random Forests and Logistic Regression models can be interpreted to look for significant features in the data, through features importances and coefficients respectively. We provide an example of how the Random Forests importance weights of the amino acid counts model can inform a better understanding of the significant region-amino acid combinations differentiating the positive binders of GILGFVFTL in Figure 2.7.

The feature importance diagram suggests that many significant features for prediction reside in the framework regions for this epitope. Interestingly, using either CDR or frame-

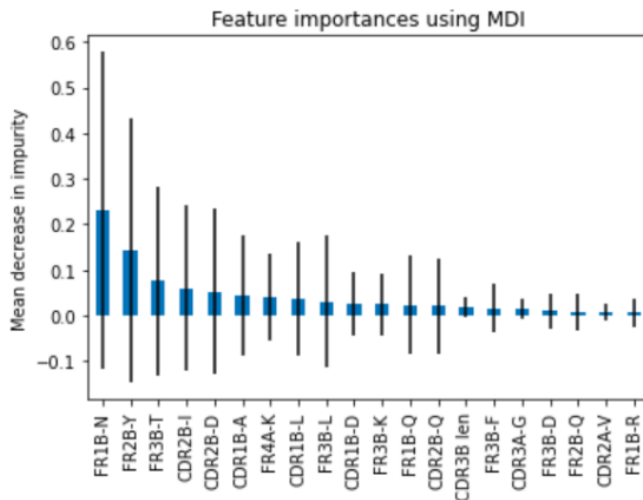


Figure 2.7: GILGFVFTL Random Forests Mean Decrease in Impurity Feature Importance

work regions made predictions nearly as good as using both, a result that will be explained further in our ablation section. CDR3 $\beta$  length also seems to be an important distinguishing factor for binding this epitope based on the feature importances. Looking at the CDR3 $\beta$  length distributions for GILGFVFTL binders and non-binders in Figure 2.5, we see that the majority of binders have a CDR3 $\beta$  length of 13. In fact, we were able to make an 81.3 percent accurate Random Forests classifier for GILGFVFTL on the test set with the 14 region lengths as the sole features.

## Ablation

We provide an ablation study in Table 2.8, using average region VHSE descriptors across the following combinations of regions: CDR and framework, CDR only, framework only, CDR3 only, CDR3 $\beta$  only,  $\alpha$  chain only,  $\beta$  chain only. A TCR can be divided into its CDR and framework regions, or into its  $\alpha$  and  $\beta$  chains.

We find that CDR or framework regions alone can be used to make models that predict about as well as models with the combined CDR and framework regions. Using the framework region average descriptors actually delivered a slightly higher Random Forests accuracy on the test set than using the combined CDR and framework regions. On the other hand, there is a larger drop associated with using either the  $\alpha$  or  $\beta$  chains. This suggests that the sequence information contained in the CDR and framework regions could be somewhat redundant from a predictive standpoint, while  $\alpha$  and  $\beta$  chain information are more complementary. While the framework regions have not been characterized as important factors for binding in a biological sense, they are differently expressed in and serve as markers for different genes. For example, framework variants tag the V gene that determines CDR1 and

Table 2.8: Classification Results Regional Features VHSE Descriptors Random Forests

PEPTIDE SEQ.	CDR + FR ACC.	CDR ACC.	FR ACC.	CDR3 ACC.	CDR3 $\beta$ ACC.	$\alpha$ ACC.	$\beta$ ACC.
AVFDRKSDAK	0.515	0.474	0.513	0.482	0.487	0.537	0.516
ELAGIGILTV	0.949	0.949	0.949	0.655	0.633	0.954	0.706
FLYALALLL	0.900	0.833	0.933	0.966	0.833	0.967	0.800
GILGFVFTL	0.950	0.949	0.947	0.894	0.833	0.820	0.950
GLCTLVAML	0.930	0.930	0.912	0.737	0.754	0.842	0.842
IVTDFSVIK	0.657	0.662	0.717	0.606	0.606	0.646	0.652
KLGGALQAK	0.589	0.566	0.603	0.511	0.512	0.575	0.577
RAKFKQLL	0.853	0.861	0.853	0.821	0.749	0.840	0.824
RLRAEAQVK	0.616	0.537	0.582	0.467	0.435	0.559	0.537
OVERALL	0.773	0.751	0.778	0.682	0.649	0.749	0.712

CDR2 regions. As binders preferentially select for certain genes or combinations of genes, framework regions are still useful for prediction of TCR-pMHC binding. We expand on this idea further in our gene usage section.

## Positional Methods

### Amino Acid Descriptors

The overall Random Forests accuracy with amino acid descriptors was 0.766, and the overall Logistic Regression accuracy with amino acid descriptors was 0.771.

Table 2.9: Classification Results Positional Features VHSE Descriptors

PEPTIDE SEQ.	RF ACC.	LR ACC.	RF AUC-PR	LR AUC-PR
AVFDRKSDAK	0.539	0.518	0.680	0.526
ELAGIGILTV	0.921	0.915	0.937	0.948
FLYALALLL	1.00	1.00	1.00	1.00
GILGFVFTL	0.926	0.881	0.947	0.960
GLCTLVAML	0.860	0.860	0.872	0.927
IVTDFSVIK	0.641	0.692	0.766	0.797
KLGGALQAK	0.575	0.579	0.696	0.628
RAKFKQLL	0.885	0.856	0.915	0.891
RLRAEAQVK	0.544	0.556	0.641	0.503
OVERALL	0.766	0.771	0.828	0.798

### Attention Models

The overall Attention Model accuracy was 0.708.



Table 2.10: Classification Results Attention Model

PEPTIDE SEQ.	ACCURACY	AUC-PR
AVFDRKSDAK	0.602	0.719
ELAGIGILTV	0.846	0.881
FLYALALLL	0.777	0.856
GILGFVFTL	0.927	0.955
GLCTLVAML	0.611	0.701
IVTDFSVIK	0.671	0.755
KLGGALQAK	0.574	0.691
RAKFKQLL	0.786	0.837
RLRAEAQVK	0.574	0.707
OVERALL	0.708	0.789

### Comparative Logo Analysis

In Figure 2.8, we provide logo plots for the amino acid distributions of the AHo numbered CDR3 Region for binders of KLGGALQAK and GILGFVFTL, and for the general TCR pool. In Figure 2.9, we plot the log enrichment of each amino acid at every position for the aforementioned epitopes. Log enrichment uses the ratio of the frequency in binders to the frequency of non-binders, with no regards for statistical significance. By plotting only amino acids which have a Fisher’s Exact Test p-value under different thresholds in Figure 2.10 and Figure 2.11, we see that the number of statistically significant log enriched amino acids in different positions in KLGGALQAK binders is lower compared to in GILGFVFTL binders. Across all 9 epitopes tested, we find that epitopes that can be predicted with high accuracy have many statistically significant log enriched amino acids in various positions.

### Gene Usage

Gene usage refers to the categorical V/D/J genes of the TCR, which are included for each sequence in the dataset annotations. Each TCR  $\alpha$  chain has 2 genes (V and J), while each TCR  $\beta$  chain has 3 genes (V, D, and J). These genes are closely related to the TCR sequence, although the sequences of two TCRs with identical genes could vary due to allelic differences. The annotations also specify allele information for each gene. For the purpose of these analyses, we strip away the allele information and use only the broader gene to categorize each TCR. Furthermore, we use only the V and J genes for each chain, ignoring the D gene on the  $\beta$  chain. There are 39  $V\alpha$ , 47  $V\beta$ , 53  $J\alpha$ , and 13  $J\beta$  genes across the dataset.

### Prediction with Gene Usage

First, we assess how well we can predict using gene information alone. We choose to one-hot encode genes for simplicity, concatenating the one-hot vectors of each gene for a total of

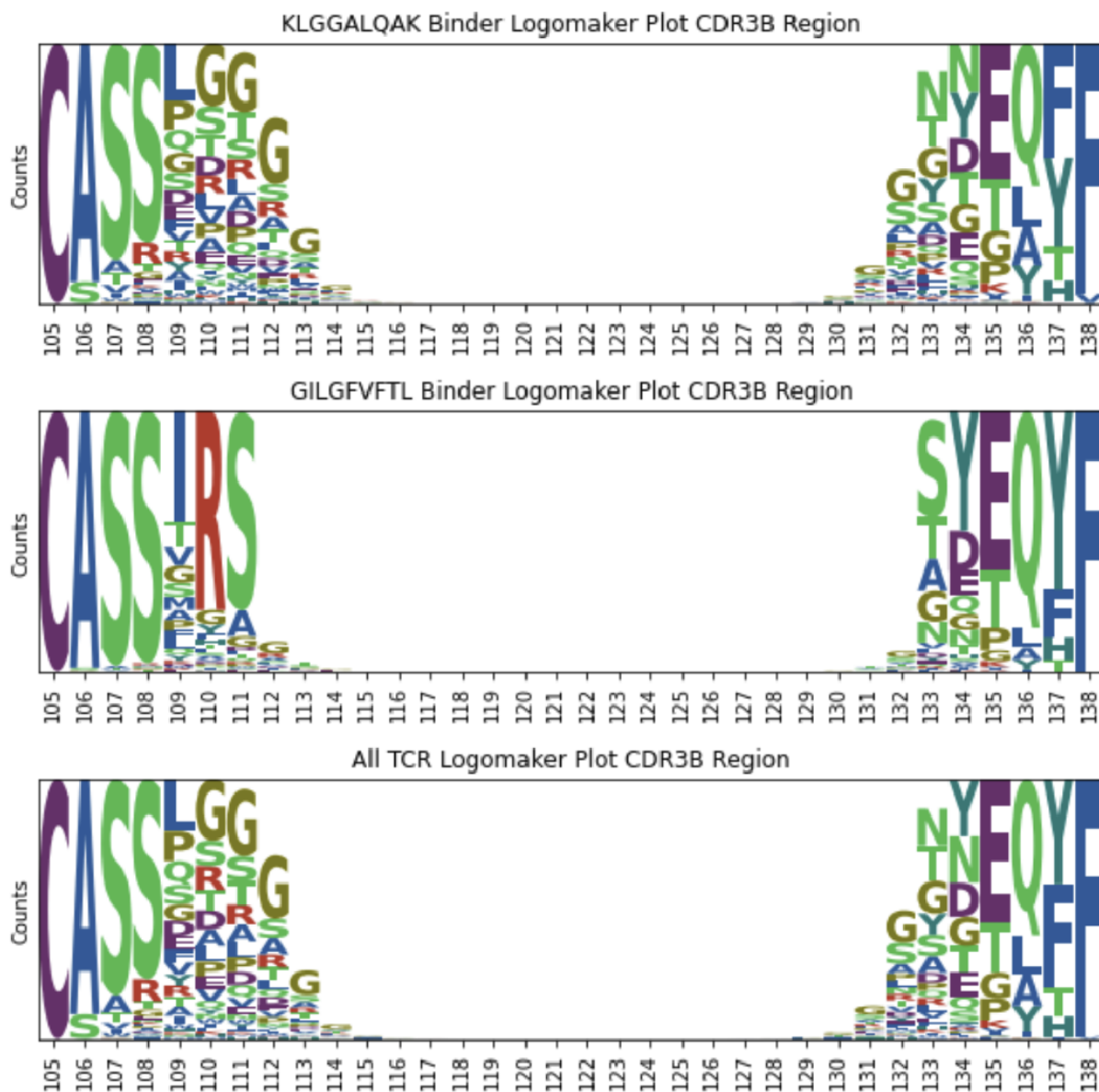


Figure 2.8: KLGALQAK Binder, GILGFVFTL Binder, All TCR AHo Numbered CDR3B Region Amino Acid Distributions

152 features for each TCR, exactly four of which are nonzero. We use the same train/test data as in earlier segments, classifying with Random Forests and Logistic Regression. The results are displayed in Table 2.11. Overall, the results are similar but slightly worse than those achieved by sequence prediction methods. While amino acid methods effectively allow models to capture similarities between genes, this is not possible with the one-hot featuriza-

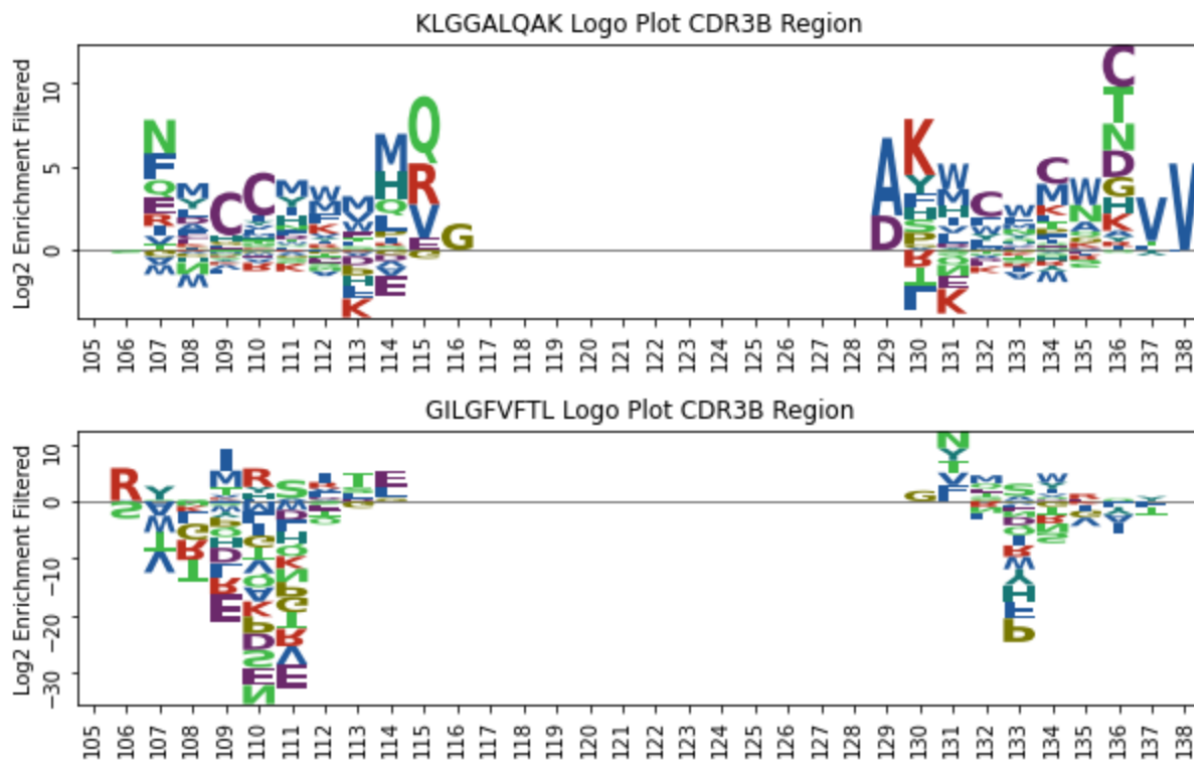


Figure 2.9: KLGGALQAK Binder, GILGFVFTL Binder AHo Numbered CDR3B Region Amino Acid Log Enrichment

tion described. The gene distributions for binders and non-binders of certain epitopes that can be predicted well with one-hot encoded gene features (ELAGIGILTV, FLYALALLL, GILGFVFTL, GLCTLVAML, RAKFKQLL) tend to have a few predominant genes by which binders can be distinguished with high likelihood. On the other hand, while not identical, the gene distributions for binders and non-binders of the poorly predicted epitopes (AVFDRKSDAK, IVTDFSVIK, KLGGALQAK, RLRAEAQVK) have more overlap and the differences are less statistically significant. For illustration, see the comparison of  $V\beta$  distributions between KLGGALQAK binders, GILGFVFTL binders, and the general pool of TCRs in the data in Figure 2.12, Figure 2.13, and Figure 2.14.

Interestingly, 10x Genomics reported observing “clonotypes with apparently cross-reactive binding” for the aforementioned set of 4 poorly predicted epitopes [2]. Makowski et al. [18] describe strong tradeoffs between the properties of affinity and specificity in antibodies, stating that “increases in affinity along the co-optimal Pareto frontier require progressive reductions in specificity.” Our models are able to differentiate binders of the highly specific epitopes, while they struggle to predict the binding of the cross-reactive epitopes. While UMI values only correlate weakly with binding affinity, the differences in affinity between

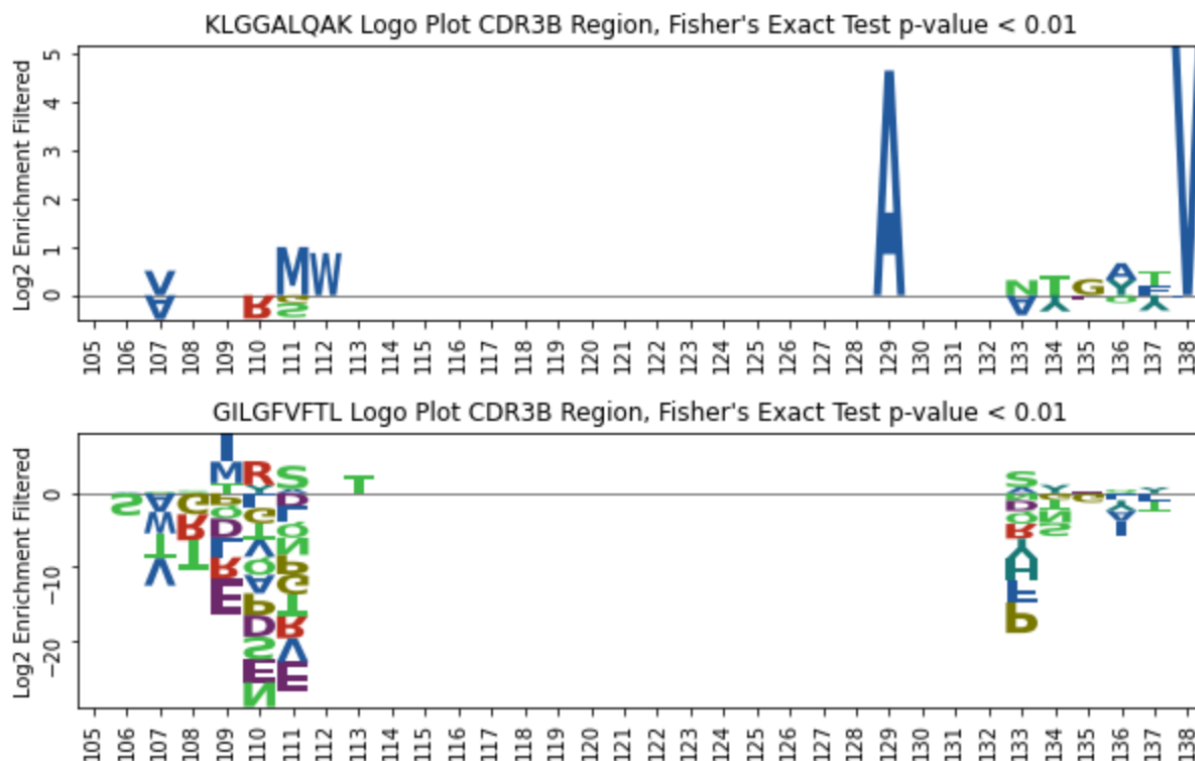


Figure 2.10: KLGGALQAK Binder, GILGFVFTL Binder AHO Numbered CDR3B Region Amino Acid Log Enrichment, Fisher’s Exact Test < 0.01

specific and cross-reactive epitopes may connect to the distributional differences exemplified in Figure 2.2 and Figure 2.3. More analysis is required to understand the link between cross-reactivity and the variance in UMI count values of the apparent cross-reactive epitopes.

Another way of conceptualizing this is through the sequence diversity of the binding TCRs. In Table 2.12, we compute the average pairwise Levenshtein edit distance between different regions of positive binders of each epitope. Positive binders were defined by the same UMI thresholds set in Table 2.5, as well as the negative control condition. “General” in this context refers to the average pairwise edit distance between different regions for the full dataset of TCRs. In general, the edit distances for each epitope correlate inversely with the accuracy achieved by our machine learning models. The epitopes with edit distances close to that of the general pool of TCRs in the dataset were predicted nearly at random by our models. TCRs sharing any gene information will have a much lower edit distance in the associated regions since allelic differences are minor. This explains why ELAGIGILTV and GILGFVFTL binders, with a few predominant  $V\alpha$  and  $V\beta$  genes respectively, have lower edit distances in the regions in their respective  $V\alpha$  and  $V\beta$  genes. This also matches with



Figure 2.11: KLGGALQAK Binder, GILGFVFTL Binder AHo Numbered CDR3B Region Amino Acid Log Enrichment, Fisher's Exact Test < 0.0001

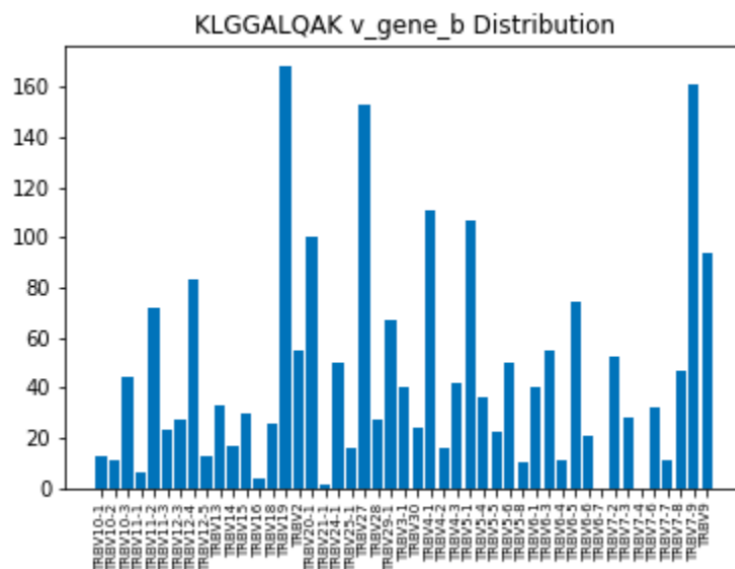
results from the ablation study in Table 2.8, where we see that the  $\alpha$  chain accuracy for ELAGIGILTV is much higher than the  $\beta$  chain accuracy. For GILGFVFTL, although there was a much lower pairwise edit distance in the  $\beta$  chain overall, the edit distance is also quite low in the CDR3 $\alpha$  and framework 4 regions (J gene) of the  $\alpha$  chain. The associated J $\alpha$  gene bias helps explain why the accuracy is high for either chain in the ablation. For the Epstein-Barr epitope FLYALALLL (bound to HLA-A\*0201) with only 25 positive binders, the edit distances show the very high pairwise similarity in both  $\alpha$  and  $\beta$  chains. Unsurprisingly, binders have a significant gene bias in both V and J genes.

### Prediction in Fixed Gene Contexts

Next, we take on the challenge of prediction in a fixed V gene context. This means that an equal number of binders and non-binders are drawn from the exact same V $\alpha$  and V $\beta$  genes, although no restriction is imposed on the J genes. We were only able to find one V gene combination with at least 20 binders and 20 non-binders for any epitope in the dataset. There are 153 GILGFVFTL binders and 47 non-binders with both TRAV27 and TRBV19

Table 2.11: Classification Results Gene Usage Features

PEPTIDE SEQ.	RF ACC.	LR ACC.	RF AUC-PR	LR AUC-PR
AVFDRKSDAK	0.531	0.579	0.687	0.464
ELAGIGILTV	0.932	0.848	0.933	0.928
FLYALALLL	1.00	1.00	1.00	1.00
GILGFVFTL	0.965	0.946	0.979	0.967
GLCTLVAML	0.789	0.842	0.808	0.934
IVTDFSVIK	0.621	0.621	0.773	0.785
KLGGALQAK	0.587	0.565	0.735	0.583
RAKFKQLL	0.872	0.840	0.933	0.884
RLRAEAQVK	0.474	0.593	0.570	0.603
OVERALL	0.753	0.758	0.824	0.794

Figure 2.12: KLGGALQAK Binders  $V\beta$  Gene Distribution

genes. Our combined training and test sets contain only 47 binders and 47 non-binders. As the number of samples is reduced due to conditioning on one specific gene combination, we use our smallest regional featurization with VHSE descriptors and Random Forest. This task is made harder due to the elimination of V gene bias and the decrease in the number of samples. In the absence of V gene bias, the CDR1 and CDR2 regions (in the V region) add little predictive information differentiating the binders from non-binders. The same is true for the Framework 1, 2, and 3 regions. However, the prediction accuracy with just the CDR3 or the Framework 4 regions is high, as shown in Table 2.13. This is attributable to the significant differences in  $J\alpha$  and  $J\beta$  distributions for binders and non-binders.

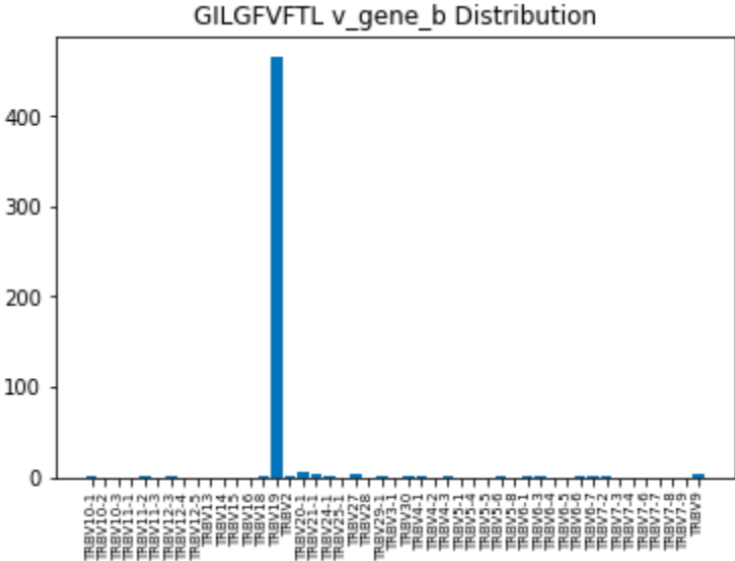


Figure 2.13: GILGFVFTL Binders Vβ Gene Distribution

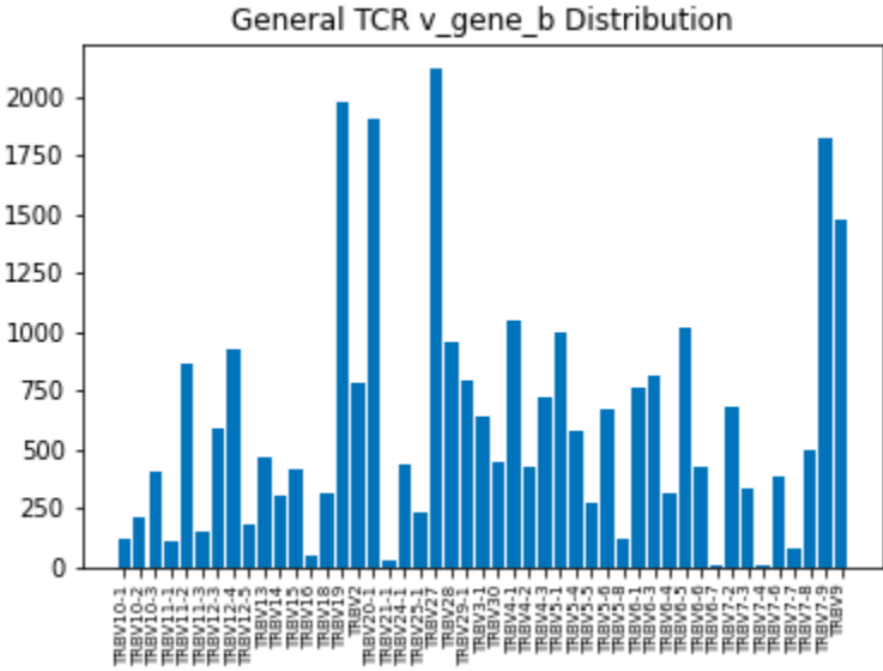


Figure 2.14: General TCR Vβ Gene Distribution

Since the context around the CDR3 is partly defined by the V genes, there are fewer possible conformations the CDR3 can take on. Different V gene contexts also seem to select for different lengths of CDR3. As TCRs with the same V genes have a high degree of sequence similarity, the differences between binders and non-binders are clearer when comparing CDR3 regions, as illustrated by Figure 2.15 and Figure 2.16. Comparing Figure 2.17 and Figure 2.18, we see that a different  $V\alpha$  being fixed can lead to noticeable changes in the amino acid distributions for binders in the CDR3 $\beta$ . This highlights the complexity of the problem at hand, as CDR3 distributions are affected by the context around them. Fixing gene information to isolate a more homogeneous group of binder and non-binders may be a useful technique in the future for TCR design.

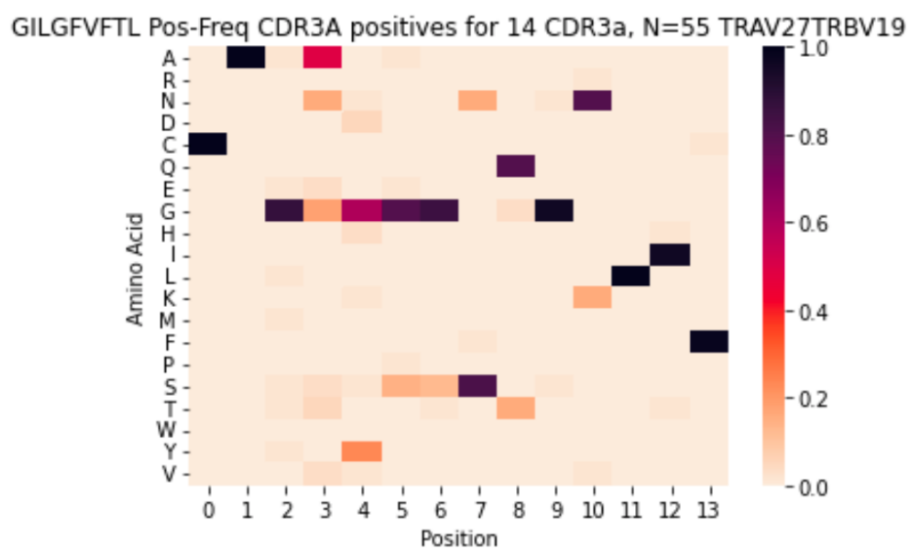


Figure 2.15: GILGFVFTL Binders Length 14 CDR3 $\alpha$  TRAV27 and TRBV19 Position-Frequency



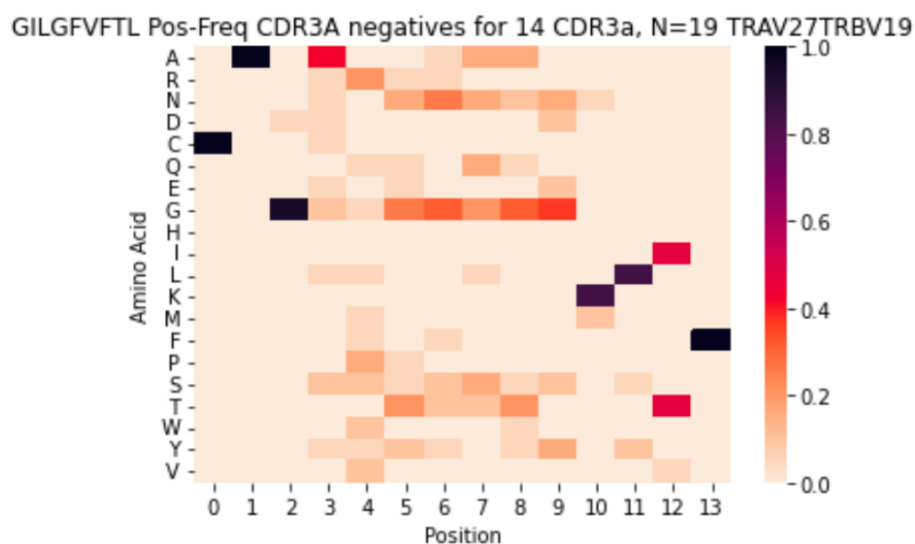


Figure 2.16: GILGFVFTL Non-Binders Length 14 CDR3 $\alpha$  TRAV27 and TRBV19 Position-Frequency

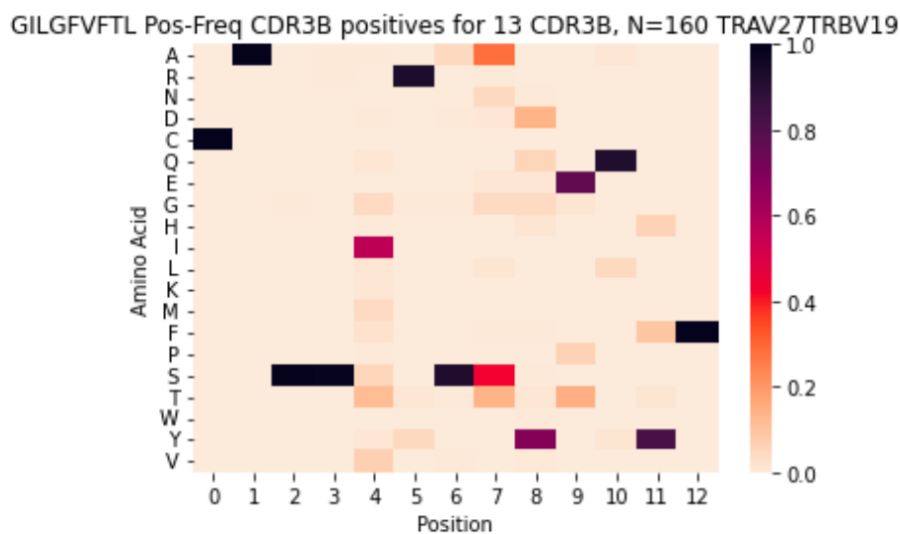


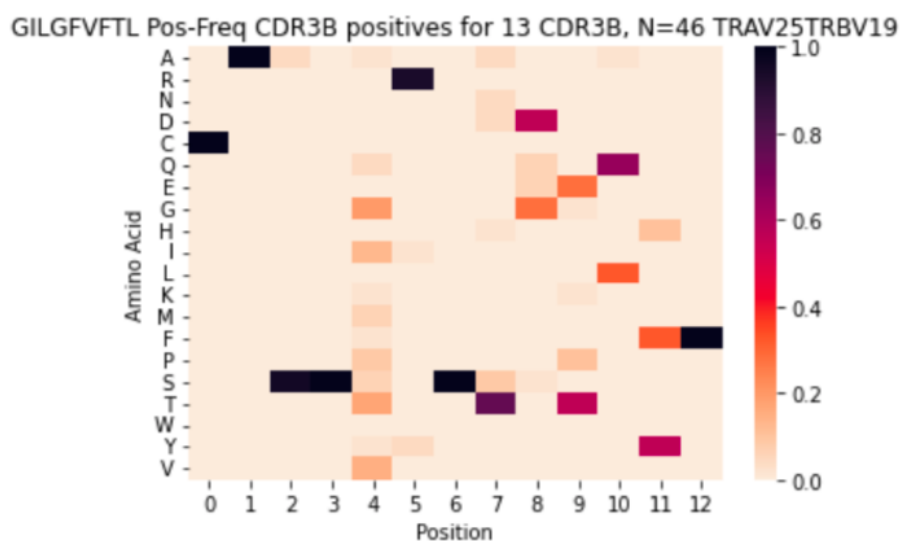
Figure 2.17: GILGFVFTL Binders Length 13 CDR3 $\beta$  TRAV27 and TRBV19 Position-Frequency

Table 2.12: Average Pairwise Edit Distance for Binders

PEPTIDE SEQ.	F1 $\alpha$	C1 $\alpha$	F2 $\alpha$	C2 $\alpha$	F3 $\alpha$	C3 $\alpha$	F4 $\alpha$	F1 $\beta$	C1 $\beta$	F2 $\beta$	C2 $\beta$	F3 $\beta$	C3 $\beta$	F4 $\beta$	VDJ $\alpha$	VDJ $\beta$	TOT
AVFDRKSDAK	16.8	5.2	7.7	9.9	16.3	9.1	4.2	14.3	4.4	6.9	12.3	16.2	8.9	2.3	68.6	64.8	133.4
ELAGIGILTV	3.1	1.1	1.6	2.1	3.0	7.6	4.1	15.3	4.4	6.8	12.3	16.5	8.5	2.5	22.5	65.7	88.2
FLYALALLL	5.3	1.9	2.3	3.0	4.8	2.7	0.9	3.1	1.4	1.5	3.6	4.6	4.5	0.9	13.8	20.7	34.5
GILGFVFTL	14.3	4.5	5.8	9.0	14.2	6.8	2.1	5.9	1.4	1.5	3.1	4.2	5.1	1.7	56.4	22.8	79.2
GLCTLVAML	12.7	3.9	5.0	8.0	13.3	7.4	4.1	12.5	3.8	5.9	8.7	12.4	8.4	2.6	54.2	54.9	109.1
IVTDFSVIK	15.9	4.9	6.5	9.5	15.8	8.6	3.8	13.9	3.8	6.1	11.4	14.3	8.3	2.0	64.6	59.6	124.2
KLGGALQAK	16.7	5.1	7.5	9.9	16.4	9.1	4.2	15.2	4.5	6.8	12.5	16.5	8.8	2.3	68.4	66.2	134.6
RAKFKQLL	13.9	4.1	5.8	8.8	13.4	7.7	4.2	12.3	3.6	6.7	10.5	12.8	7.6	2.3	57.4	55.5	112.9
RLRAEAQVK	16.6	5.2	7.6	9.8	16.4	9.2	4.1	14.7	4.3	6.9	12.4	16.5	9.0	2.1	68.2	65.5	133.7
GENERAL	16.7	5.1	7.5	9.9	16.3	9.0	4.2	14.8	4.4	6.7	12.3	16.2	8.5	2.2	68.3	64.8	133.1

Table 2.13: Classification Results Regional Features VHSE Descriptors Random Forest Accuracy, TRAV27 and TRBV19 Genes Fixed

PEPTIDE SEQ.	CDR + FR	CDR	FR	CDR1 + CDR2	CDR3	FR1 + FR2 + FR3	FR4
GILGFVFTL	1.0	0.895	1.0	0.579	0.947	0.631	1.0

Figure 2.18: GILGFVFTL Binders Length 13 CDR3 $\beta$  TRAV25 and TRBV19 Position-Frequency

# Chapter 3

## Conclusion

### 3.1 Summary

Our efforts focused on improving classification accuracy for TCR-pMHC prediction, as the best-known performing models on this problem are still inadequate. Part of the difficulty in predicting TCR-pMHC binding with the available data comes down to the highly variable UMI count values provided by 10x Genomics. In particular, we find significant variance in the UMI values of identical TCR sequences. This variance increases as we condition with higher UMI values. While TCR-pMHC binding is not usually considered a binary phenomenon but rather one that is characterized by binding affinity, the weak correlation between UMI count values and binding affinity rules out the possibility of training a regression model. To determine binders, we set a pMHC-specific UMI count threshold, and compare it with negative control UMI counts. Using regional and positional featurizations, we find that TCR-pMHC binding can be predicted with high accuracy with separate Random Forest or Logistic Regression models. Comparable classification accuracy is obtained using an attention-based neural network architecture incorporating positional information while being trained across all epitopes at once. As incomplete sequence data remains an issue in working with TCR data, we provide an ablation study showing the relative performance of different region subsets of the TCR. Finally, we present results for prediction with gene usage features and prediction across TCRs with the same V genes.

### 3.2 Future Work

In a future with enough high-quality binding data for thousands of different epitopes, neural networks may be able to generalize to unseen sets of epitopes, as well as unseen sets of TCRs. Advances in multiplexed binding and sequencing technology will be required in order to obtain TCR-pMHC binding data with stronger correlations to binding affinity. Current UMI count values weakly correlate with both the number of TCRs on the cell surface and the binding affinity of each TCR. In theory, the number of TCRs on the cell surface could

be found by using an oligo-tagged antibody against the TCR itself. If collected, it may be possible to use this number to regularize UMI values to approximate binding affinity more closely between samples. In addition to TCR sequences, it is likely that TCR-pMHC solved structures will become an important source of data for this problem. Current state-of-the-art structure prediction methods like AlphaFold2 have the least accurate structure predictions for the TCR's flexible loop regions, which are the most relevant for binding [19].

# Chapter 4

## Declarations

### 4.1 Acknowledgements

I would like to thank my advisor Professor Yun S. Song for providing invaluable guidance and support as I worked my way through this project. I also thank Alexander Whatley, Milind Jagota, Nicholas Bhattacharya, William DeWitt, Melissa Thorne, Professor Jimmy Ye, Professor Jennifer Listgarten, and Professor Ian Holmes for all the interesting scientific discussion and for their support.

### 4.2 Author Contributions

Yun S. Song introduced and formulated the problem. Julian Faust wrote code for the analyses, including building, testing, and evaluating the models. Julian Faust wrote the manuscript.

# Bibliography

- [1] Krogsgaard, M., and Davis, M. "How T-cells see antigen." *Nature immunology* 6.3 (2005): 239-245.
- [2] 10x Genomics. "A New Way of Exploring Immunity: Linking Highly Multiplexed Antigen to Recognition to Immune Repertoire and Phenotype" (2020) [https://pages.10xgenomics.com/rs/446-PB0-704/images/10x\\_AN047\\_IP\\_A\\_New\\_Way\\_of\\_Exploring\\_Immunity\\_Digital.pdf](https://pages.10xgenomics.com/rs/446-PB0-704/images/10x_AN047_IP_A_New_Way_of_Exploring_Immunity_Digital.pdf)
- [3] Jokinen E, Huuhtanen J, Mustjoki S, Heinonen M, Lähdesmäki H. Predicting recognition between T cell receptors and epitopes with TCRGP. *PLOS Computational Biology* 17(3) (2021): e1008814. <https://doi.org/10.1371/journal.pcbi.1008814>
- [4] Springer, I., et al. "Prediction of specific TCR-peptide binding from large dictionaries of TCR-peptide pairs." *Frontiers in immunology* 11 (2020): 1803.
- [5] Dash, P., et al. "Quantifiable predictive features define epitope-specific T cell receptor repertoires." *Nature* vol. 547,7661 (2017): 89-93. <https://doi.org/10.1038/nature22383>
- [6] Montemurro, A., Schuster, V., Povlsen, H.R. et al. NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCR and sequence data. *Commun Biol* 4, 1060 (2021). <https://doi.org/10.1038/s42003-021-02610-3>
- [7] Fischer, D., et al. "Predicting antigen specificity of single T cells based on TCR CDR3 regions." *Molecular systems biology* vol. 16,8 (2020): e9416. <https://doi.org/10.15252/msb.20199416>
- [8] Sidhom, JW., Larman, H.B., Pardoll, D.M. et al. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nat Commun* 12, 1605 (2021). <https://doi.org/10.1038/s41467-021-21879-w>
- [9] Zhang Z, Xiong D, Wang X, Liu H, Wang T. Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. *Nat Methods* (2021): (1) 92-99. <https://doi.org/10.1038/s41592-020-01020-3>

- [10] Immudex. “DCODE Dextramer® Reagents Identify Antigen-Specific Populations and Their TCR Clonotypes at Single-Cell Level.” (2019) <https://www.immudex.com/resources/educational-material/single-cell-level-identification-of-antigen-specific-t-cells-with-dcode-dextramer/>
- [11] Honegger A, Plückthun A. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *J Mol Biol.* (2001) 309:657–70. 10.1006/jmbi.2001.4662
- [12] Davidsen, K., et al. ”Deep generative models for T cell receptor protein sequences.” *Elife* 8 (2019): e46935.
- [13] van Westen, Gerard JP, et al. ”Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets.” *Journal of cheminformatics* 5.1 (2013): 1-11. (= <https://www.jmlr.org/papers/volume21/19-755/19-755.pdf>)
- [14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- [15] Dunbar, J., Deane, C. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics* (2016): Volume 32, Issue 2, Pages 298–300, <https://doi.org/10.1093/bioinformatics/btv552>
- [16] Robinson, J., Halliwell, J. A., Hayhurst, J. D., Flicek, P., Parham, P., Marsh, S. G. The IPD and IMGT/HLA database: allele variant databases. *Nucleic acids research*, 43(Database issue), (2015): D423–D431 . <https://doi.org/10.1093/nar/gku1161>
- [17] Zhang, H., et al. ”The contribution of major histocompatibility complex contacts to the affinity and kinetics of T cell receptor binding.” *Scientific reports* 6.1 (2016): 1-11.
- [18] Makowski, E.K., Kinnunen, P.C., Huang, J. et al. Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nat Commun* 13, 3788 (2022). <https://doi.org/10.1038/s41467-022-31457-3>
- [19] Abanades, B., et al. ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation. *Bioinformatics* (2022): Volume 38, Issue 7, Pages 1877–1880, <https://doi.org/10.1093/bioinformatics/btac016>
- [20] Campillo-Davo, D., et al. ”The quest for the best: how TCR affinity, avidity, and functional avidity affect TCR-engineered T-cell antitumor responses.” *Cells* 9.7 (2020): 1720.



- [21] Lu, T., et al. "Deep learning-based prediction of the T cell receptor-antigen binding specificity." *Nature Machine Intelligence* 3.10 (2021): 864-875.
- [22] Sewell AK. Why must T cells be cross-reactive? *Nat Rev Immunol.* (2012): (9):669-77. <https://doi.org/10.1038/nri3279>
- [23] Moris, Pieter, et al. "Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification." *Briefings in Bioinformatics* 22.4 (2021): bbaa318.
- [24] Yin, Y. and Mariuzza, R. "The multiple mechanisms of T cell receptor cross-reactivity." *Immunity* 31.6 (2009): 849-851.
- [25] Zhang, W., et al. "A framework for highly multiplexed dextramer mapping and prediction of T cell receptor sequences to antigen specificity." *Science Advances* 7.20 (2021): eabf5835.
- [26] Zhang, Z., et al. "Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics." *Nature methods* 18.1 (2021): 92-99.