An Application of Attention for the Prediction of TCR-Epitope Binding Affinity

by

Michael Cai

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved November 2021 by the
Graduate Supervisory Committee:

Heewook Lee, Chair
Seojin Bang
Chitta Baral

ARIZONA STATE UNIVERSITY

December 2021

ABSTRACT

T-cells are an integral component of the immune system, enabling the body to distinguish between pathogens and the self. The primary mechanism which enables this is their T-cell receptors (TCR) which bind to antigen epitopes foreign to the body. This detection mechanism allows the T-cell to determine when an immune response is necessary. The computational prediction of TCR-epitope binding is important to researchers for both medical applications and for furthering their understanding of the biological mechanisms that impact immunity. Models which have been developed for this purpose fail to account for the interrelationships between amino acids and demonstrate poor out-of-sample performance. Small changes to the amino acids in these protein sequences can drastically change their structure and function. In recent years, attention-based deep learning models have shown success in their ability to learn rich contextual representations of data. To capture the contextual biological relationships between the amino acids, a multi-head self-attention model was created to predict the binding affinity between given TCR and epitope sequences. By learning the structural nuances of the sequences, this model is able to improve upon existing model performance and grant insights into the underlying mechanisms which impact binding.

TABLE OF CONTENTS

Page

Chapter 1

INTRODUCTION

1.1 Biological Background

The adaptive immune system is the body's innate way of recognizing and defending against pathogens (Wilson and Hunt, 2002). One of the defining mechanisms used by the adaptive immune system is the T-cell. A primary function of these cells is distinguishing foreign invaders from the cells which comprise the self. Cells produce antigens as a byproduct of normal cell metabolism, because of this each antigen is unique to the process by which it was created. This means that pathogen-infected cells produce different antigens than those normally produced by healthy cells in the body. Each cell presents its created antigens using a major histocompatibility complex (MHC) protein which binds to these fragments, carries them to the cell surface, and presents them there.

The exposed antigens on the cell surface enable T-cells to investigate if a particular cell has been infected with a virus. As T-cells patrol the body their receptors come into contact with the binding regions on the other cell's antigens, also known as an epitope. If a binding occurs between the TCR and epitope, an immune response is started to kill the target pathogen-infected cell. However, the immune system must monitor for a plethora of diseases and their respective epitopes. To accommodate for this the immune system creates millions of distinct TCRs using the variable (V), diversity (D), and joining (J) genes, creating immune cover through VDJ recombination (Sewell, 2012). These TCRs are also cross-reactive with the ability to recognize multiple epitopes.

Utilizing the adaptive immune system is the primary focus of immunotherapy, a treatment that harnesses the body's innate defenses against novel diseases (Schumacher *et al.*, 2019). One common immunotherapy target is cancer, a disease caused by somatic mutations leading to abnormal cell growth and function. Cancer can evade the body's immunosurveillance and remain undetected through various mechanisms making it difficult to deal with unaided (Seliger, 2005). The immune system, however, is not rendered completely useless. The mutations which create cancer cells cause them to produce neoantigens, antigens that contain peptides that are absent from the human genome (Schumacher *et al.*, 2019).

Immunotherapy has been shown to be an effective treatment due to its long-lasting effects and selectivity (Koury *et al.*, 2018). One avenue of treatment immunotherapy uses is the transfusion of T-cells with receptors that bind to a patient's cancer cell neoantigens. This enables their immune system to identify cancer cells as foreign to the body. Immunology researchers start this process by determining the neoantigen epitopes produced by the patient's cancer cells. Once the epitope sequences have been determined, they must find or engineer T-cells with the cognate TCRs required for treatment.

The TCR screening process however is challenging and costly. There are over 10^{15} possible rearrangements of a T-cell's VDJ genes, the genes which ultimately influence the epitopes a T-cell binds to (Lythe *et al.*, 2016). The scale of TCR candidates to test against the patient's epitopes makes manual laboratory approaches for determining which TCRs bind to a target neoantigen epitope expensive and time-consuming. A computational solution is necessary to reduce the number of candidates for the development of a patient-specific course of treatment.

1.2 Computational Background

In recent years, several antigen-specific databases have been created including VDJDB, McPAS, and IEDB (Shugay *et al.*, 2018; Tickotsky *et al.*, 2017; Vita *et al.*, 2019). The new availability of data has enabled machine learning approaches for predicting the specificity between TCRs and epitopes. A few different computational models have been proposed for this problem. Solutions include netTCR which utilizes convolutional neural networks (CNN) to determine the interactions between TCRs and epitopes in the most common human allele HLA-A*02:01 (Jurtz *et al.*, 2018). ERGO on a similar note experiments with long short-term memory (LSTM) and autoencoder (AE) structures to build a unified prediction model (Springer *et al.*, 2020). TCRex's approach involves creating a random forest model to build a series of decision trees for each epitope (Gielis *et al.*, 2019). Another solution, TCRGP utilizes both the TCRα and TCRβ regions to determine which regions are important for epitope recognition (Jokinen *et al.*, 2019). However, these methodologies suffer from several major problems.

The first problem arises with methods such as TCRGP and TCRex which both propose antigen-specific models for each epitope. This limits their prediction models to only epitopes which have a sufficient number of known cognate TCRs, severely inhibiting their practicality for out-of-sample predictions. Next, all of these models are black-box models which suffer from a lack of interpretability, leaving the biological explanations behind binding still unknown. The final problem shared by these models is the loss of positional and contextual information from the TCR and epitope sequences in the models. Their chosen structures fail to learn how amino acids impact the structure and function of others in the same sequence.

Attention-based models have shown success in their ability to encode contextual information in natural language processing (NLP) (Vaswani *et al.*, 2017). Similar to sentences, protein sequences also share hidden contextual relationships between their amino acids which impact their ability to bind to other sequences (Serçinoğlu and Ozbek, 2020). In this paper, I present a new model which utilizes a multi-head self-attention mechanism. The attention layer helps the model learn the biological contextual representations of epitopes and TCRs and understand how each of the amino acids determines their binding affinity to each other.

Chapter 2

METHODS

Data was collected and cleaned from three databases for training and testing. After analysis of procured data, the input to the multi-head self-attention model was chosen to be a pair of sequences consisting of a sequence from the TCRβ chain and an epitope sequence. Given the two sequences, the model calculates and outputs the binding affinity between them as a value between 0 and 1. The prediction model was implemented using PyTorch (Paszke *et al.*, 2019). The model's architecture consists of two separate encoder layers for each type of sequence, which then both feed into a unified decoder layer. The encoders both contain multi-head attention layers, while the decoder consists of a dense linear layer.

Prior to being input into the model, the sequences are transformed into embeddings using an embedding layer. The combination of the embedding layer and multi-head attention layer in the encoder enables the model to learn specific amino acid latent features and relationships, as well as their impacts on binding affinity. The model was trained on the collected data which was split using two different methodologies designed to test the model's performance on out-of-sample data. Hyperparameter selection and performance assessment were done using nested-cross validation over five folds. In order to test the multi-head-attention model's performance, netTCR, ERGO-LSTM, and ERGO-AE were chosen to be baseline models (Tickotsky *et al.*, 2017; Jurtz *et al.*, 2018). These three models are trained on the same dataset and splits as the multi-head-attention model to compare their performances.

2.1 Data Procurement and Processing

Data for known TCR-epitope binding pairs was collected from three databases VDJDB, McPAS, and IEDB in December 2020 (Shugay *et al.*, 2018; Tickotsky *et al.*, 2017; Vita *et al.*, 2019). Initially, the combination of the three databases contained 284,859 pairs before pre-processing. Entries in the databases represented the TCRs and epitopes using strings of characters indicating the amino acids in the protein sequences. Furthermore, they included information for the complementarity-determining region 3 (CDR3) sequences from both TCRα and TCRβ regions. The CDR3 is used across all three of the databases to represent an individual TCR due to its importance as a determinant for antigen recognition (Tickotsky *et al.*, 2017). Figure 2.1 shows the distribution of TCR sequence information available in the collected data. Approximately 68.8% of TCR-epitope pairs were missing a reported sequence from the TCRα chain. Inferring these sequences would likely introduce errors due to the large sequence space over which the TCRα chain can exist. Consequently, the TCRα chain sequences were dropped from the data rather than imputed.

The databases also included epitopes from both MHC class I and II proteins, however, the data was mostly skewed towards MHC class I. Figure 2.2 highlights the lack of MHC class II epitopes in the databases. Both classes of MHC proteins are structurally similar but contain some differences that enable them to bind to specific types of T-cells (Wilson and Hunt, 2002). These small structural changes make it difficult to extrapolate information from one class to another. With very few samples, it is unlikely the model will learn significant binding determinants for MHC class II epitopes. Furthermore, most nucleated cells present class I MHC proteins whereas only some specialized cells present class II MHC proteins. This makes predictions on class I MHC proteins more desirable than their counterparts. The aforementioned

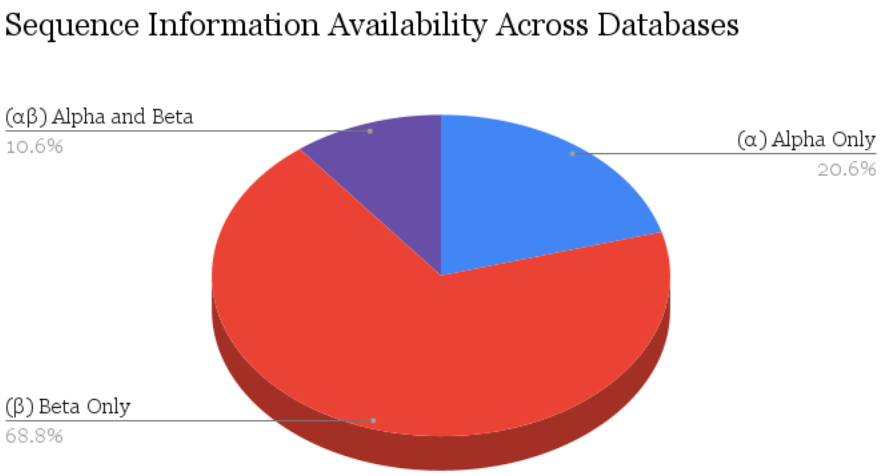

Figure 2.1: The availability of CDR3 protein sequences across VDJDB, McPAS, and IEDB.

combination of factors led to the data being filtered to only contain MHC I epitope sequences.

After the two previously mentioned filters for missing data were applied, the rest of the dataset was pre-processed into a unified format consisting of the CDR3 sequence from the TCRβ chain and the epitope sequence for the target antigen. Several additional quality control filters were also placed on the remaining data to remove pairs that were not described using protein sequences or contained unknown amino acid symbols. Once these filters were applied the remaining data consisted of 6,388 pairs from VDJDB, 11,936 pairs from McPAS, and 169,223 pairs from IEDB. The three databases did contain some repeated samples and overlapping data, requiring the removal of duplicates after they were combined. After all pre-processing steps, the final data count came to 150,008 unique TCR-epitope binding pairs. Inside of which 982 unique epitopes and 140,675 TCRs were observed.

Non-binding pairs are not readily available as the consulted databases only record binding pairs. A method for negative data generation was developed using random

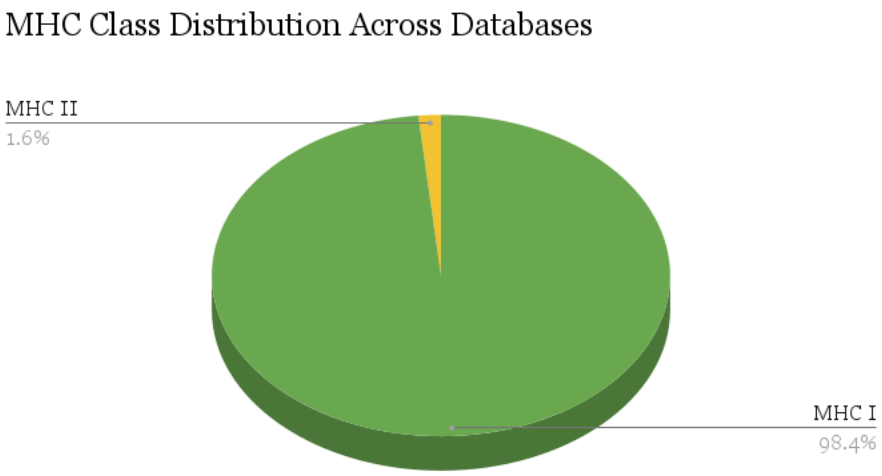

Figure 2.2: The distribution of MHC classes across VDJDB, McPAS, and IEDB.

recombination of the positive data. One consideration for random recombination is the possibility for a single TCR to bind to multiple epitopes and multiple TCRs to a single epitope (Sewell, 2012). However, the space of possible TCR and epitope sequences makes it unlikely that two randomly selected sequences will have an affinity for binding with each other. With this consideration, the random recombination method was used to supplement the positive data. The method works by duplicating an existing positive pair and replacing the TCR in the duplicate with another randomly selected TCR from the positive data. If this generated a pair already present in the dataset another TCR was chosen until a unique pair was created. This process generated a 1:1 ratio of negative data to be utilized for training, validation, and testing. Additionally, since each original pair's epitope was used during the random recombination, the distribution of epitopes in the data remained the same as before.

The recent severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic has highlighted the need for antigen-specific treatments such as vaccines. An additional SARS-CoV-2 dataset was pulled from IEDB in July 2021 (Vita *et al.*,

2019). This data was kept separate from the rest of the training data for use as an independent out-of-sample testing dataset. The pulled data featured 332 newly added TCR-epitope pairs for SARS-CoV-2. The epitopes and TCRs in the SARS-CoV-2 dataset were not present in the training data. The same quality control filters were applied to ensure all data was described using protein sequences. After filtering, two epitopes were present in the SARS-CoV-2 dataset. The first, YLQPRTFLL, had 304 recorded cognate TCRs while the other epitope, RLQSLQTYV, had 28 recorded cognate TCRs.

2.2 Training and Testing Set Split

The possible sequence space for both epitopes and TCRs is extremely large. It is estimated that the body harbors 10^{10} distinct TCR clonotypes at any time to screen for the various diseases humans encounter regularly (Lythe *et al.*, 2016). It is also possible for a pathogen to escape recognition by a particular TCR through a series of mutations that cause sufficient antigen changes to render it non-binding (Sewell, 2012). Therefore it is of interest to predict binding affinities for sequences not seen by the model before. To simulate predictions on out-of-sample sequences, two strategies were devised to split the training and testing data. These strategies were used to evaluate the model's ability to generalize on unseen TCRs and epitopes.

- TCR Split: The data was split such that any TCRs that are in the testing set are not found in the training set. This split of the data models a situation where additional TCRs are being screened for epitopes with known binding TCRs. The goal of training and testing on this split is to evaluate the model's prediction performance on out-of-sample TCRs.

- Epitope Split: The data was split such that any epitopes that are in the testing

set are not found in the training set. This split of the data models a situation where the antigens and therefore epitopes for a specific disease have not been encountered before. This is possibly due to the appearance of a novel disease or a series of mutations to a known disease. The goal of training and testing on this split is to evaluate the model's prediction performance on out-of-sample epitopes.

A random sampling method was originally considered as a part of the experiment on model performance. However, upon further examination of the collected data, it was hypothesized that a random split would have a similar performance to a TCR split. The vast majority of TCR sequences in the dataset only appear once, meaning a random sampling would already have minimal TCR sequence overlap between the training and testing set. To test this hypothesis and compare the two methods, 10 random splits of the data were created. Each of the splits was then converted into a TCR split by correcting the folds to not overlap on any TCR sequences.

An overlapping fold was corrected by exchanging offending TCR sequences with other folds. A fold removing an offending TCR would, when possible, exchange for a sequence that was already present in its fold. Otherwise, it would select a random TCR which was unique with no duplicates in the other fold. After all corrections were made, each split of the random split was compared to its respective corrected TCR split. On average there was a 91.22% similarity in TCR sequence membership between the two kinds of splits. Due to the relative similarity between the splits generated by both methods their performance would likely also be similar. The random sampling method was not utilized in the rest of the experiments to reduce result redundancy and the overall amount of training required.

2.3 Multi-head Attention Model

The created binding affinity prediction model consists of two encoders that encode the TCR and epitope sequences separately and a linear decoder that determines the affinity between the two sequences. The defining characteristic of this model is the multi-head self-attention mechanism located inside the encoders (Vaswani *et al.*, 2017). The multi-head self-attention mechanism selectively focuses the attention of the model based on the strength of relationships between amino acids in the sequences. The arrangement of the amino acids in TCR and epitope sequences plays a vital role in determining their structure and function (Serçinoğlu and Ozbek, 2020). The attention layer helps the model learn the biological context behind these arrangements that ultimately impact their ability to bind to other sequences.

Prior to being input into the model, each input sequence is padded to a constant length. These two lengths p_T and p_E for TCR and epitope sequences respectively are a part of the hyperparameter set. After this pre-processing step, each input sequence to the model is fed into an initial embedding layer to obtain embeddings of size p. Figure 2.3 shows an example of a padded TCR sequence being transformed into embeddings of size $p_T \times p$. The individual amino acid symbols were treated as words in this case, with each symbol being one of p entries in the embedding dictionary. The intuition behind this layer is to learn the latent features for each amino acid and determine which have similar properties for binding.

The sequence embeddings $\boldsymbol{T} \in R^{p_T \times p}$ and $\boldsymbol{E} \in R^{p_E \times p}$ are then fed into their corresponding encoders. The attention mechanisms in each encoder f_T and f_E process the sequences separately to learn their contextual representations. The encoder then quantifies and learns the strength of linear relationships between amino acid and their positions in the sequence. These learned relationships then become the basis

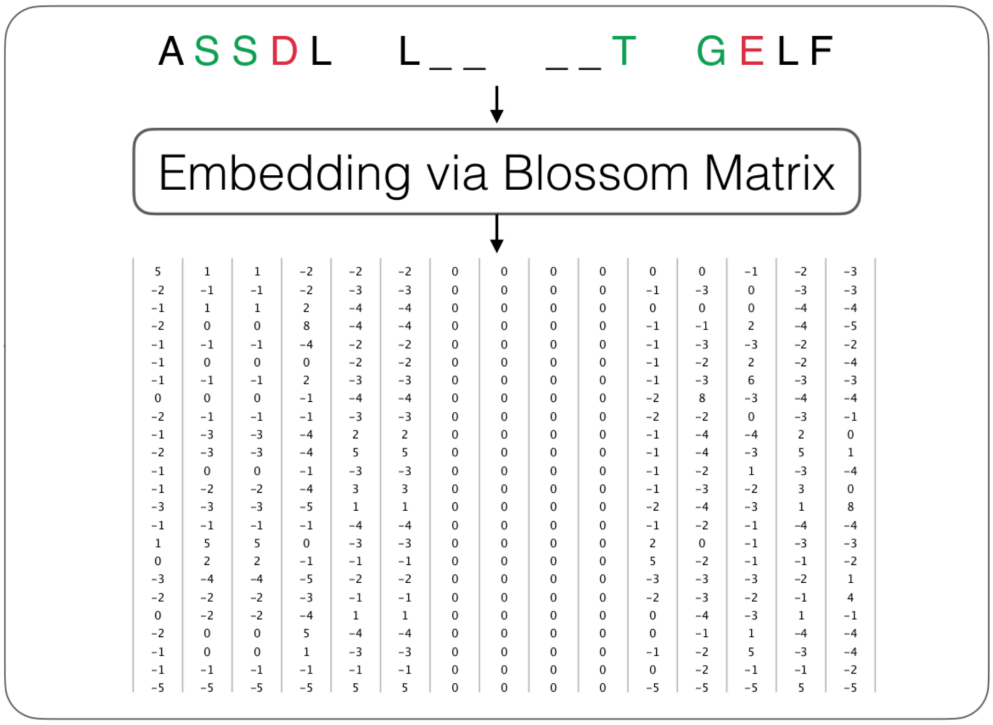

Figure 2.3: An example of a padded sequence being transformed into embeddings

for the amount of attention the model places on other positions when transforming the input. The new representation of each sequence becomes a linear average of the input representation weighted by the learned attention values from the strength of the input's positional relationships. More precisely, a TCR sequence $\boldsymbol{T}$ is fed into three linear layers returning (linear-transformed) *key* ($\boldsymbol{K}$), *query* ($\boldsymbol{Q}$), and *value* ($\boldsymbol{V}$) matrices as follows:

$$\boldsymbol{K} = \boldsymbol{T}\boldsymbol{W}^K,\ \boldsymbol{Q} = \boldsymbol{T}\boldsymbol{W}^Q,\ \boldsymbol{V} = \boldsymbol{T}\boldsymbol{W}^V \tag{2.1}$$

The relationship strength between i-th amino acid and the others, denoted as $\boldsymbol{w}_i$, is determined by the scaled dot-product of the i-th row of $\boldsymbol{Q}$ with all rows of $\boldsymbol{K}$ as follows:

$$\boldsymbol{w}_i = Softmax\left(\frac{\boldsymbol{q}_i\boldsymbol{K}^T}{\sqrt{p_T}}\right) \tag{2.2}$$

Where $\boldsymbol{q}_i$ is the i-th row of $\boldsymbol{Q}$. The contextual representation of i-th amino acid

is then defined as a linear sum of all amino acid vectors weighted by $\boldsymbol{v}_i$.

$$\begin{aligned} \boldsymbol{t}_i^* &= \boldsymbol{w}_i \boldsymbol{V} = \boldsymbol{w}_{i1}\boldsymbol{v}_1 + \cdots + \boldsymbol{w}_{ip_T}\boldsymbol{v}_{p_T}, \\ \boldsymbol{T}^* &= \boldsymbol{W}\boldsymbol{V} \end{aligned} \tag{2.3}$$

Where $\boldsymbol{w}_i$ is the i-th row of $\boldsymbol{W}$. Each element of $\boldsymbol{w}_i$ can be interpreted as an importance score of each amino acid in the sequence for determining the new representation of the i-th amino acid. This attention mechanism is called self-attention. The cognitive load of the attention mechanism is spread out by concatenating and passing multiple self-attention outputs through a single dense layer. The epitope sequence is similarly processed through a separate multi-head attention layer to obtain a new representation matrix. These two encoders are distinct to separately learn the structures of both sequences. Finally, the output of both multi-head attention layers forms the expected dimensions of TCR and epitope representations ($\boldsymbol{T}^*$ and $\boldsymbol{E}^*$).

The encoded sequence representations are concatenated and fed into the decoder f_d. The concatenated representation is then passed through several linear transformations decreasing in size using the Sigmoid Linear Unit activation function, also known as the Swish function (Ramachandran *et al.*, 2017). The output of the decoder is then fed into a Sigmoid activation function to receive the binding affinity score between the two sequences as follows:

$$Score(\boldsymbol{T}, \boldsymbol{E}) = \frac{1}{1 + e^{(-f_d(\boldsymbol{T}^*, \boldsymbol{E}^*))}} \tag{2.4}$$

The outputted binding affinity is a real number between 0 and 1 representing the likelihood of a binding between the two sequences. This value is rounded to the nearest integer to receive either a 0 or a 1 indicating a negatively or positively predicted binding respectively. The model is trained using the Adam algorithm to minimize the binary cross-entropy loss (Kingma and Ba, 2014). During the creation of the model, five hyperparameters were identified for tuning.

- Maximum TCR Size - The size to which TCR sequences are uniformly padded.
- Maximum Epitope Size - The size to which epitope sequences are uniformly padded.
- Initial Embeddings - A blocks substitution matrix (BLOSUM) matrix that describes the initial weights in the embedding layer.
- Hidden Layer Dimensions - The size of the hidden layers present in the linear transformations after concatenation of the two post-attention representations of the sequences.
- Drop Rate - The probability of dropout during the linear transformations in the dense layer. Dropout has been shown to reduce overfitting on datasets of limited sizes (Hinton *et al.*, 2012).

The TCR and epitope size hyperparameters were chosen to identify if particular segments of the sequences were more important than others for binding. During preprocessing both sequence types utilize mid-padding which inserts or deletes symbols in the middle of the sequence until the sequence is a standard length. This means sequences longer than the maximum length will only have symbols from the beginning or end of their sequences preserved. A larger maximum size means more information from longer sequences is preserved, however shorter sequences become more sparse with the insertion of several padding symbols. Several sizes were tested based on the distribution of sequence lengths in the data. Additionally, a size termed ∞ was tested where all of the sequences were padded to match the length of the largest sequence in the dataset.

The initial embeddings hyperparameter tests if integrating BLOSUM matrices can improve model performance. BLOSUM matrices are used during protein sequence

alignment to compare the similarity of two amino acids (Henikoff and Henikoff, 1992). The embedding layer weights are still updated during training, however, the intent of this experiment is to determine if current biological models contain important information for binding which cannot be intuitively learned by the model. The last two hyperparameters were identified in the dense layer of the decoder. Various sizes of hidden layers and dropout rates were tested to maximize model performance and minimize overfitting. Table 2.1 shows the chosen hyperparameter search space for the model. The hyperparameters were tuned via grid search and each set's performance was tested using 5-fold nested cross-validation.

Table 2.1: The hyperparameter search space for the model tuned via grid search.

Hyperparameter	Values Tested	Final Model Value
Maximum TCR Size	$\{10, 15, 18, 20, 22, \infty\}$	20
Maximum Epitope Size	$\{10, 15, 18, 20, 22, \infty\}$	22
Initial Embeddings*	{45, 50, 62, None}	None
Hidden Layer Dimensions	$\{512, 1024, 2048\}$	1024
Drop Rate	$\{25\%, 33\%, 50\%\}$	25%

**The values for the initial embeddings refer to a BLOSUM matrix number.*

2.4 Additional Structures Tested

The described structure was not the only model tested for this task but it was found to be the best performing of all the created architectures. The aforementioned multi-head attention model utilizes just a single attention layer to receive the encoded representation of each protein sequence. However, structures for natural language processing such as ELMO and BERT utilize multiple attention layers to obtain their

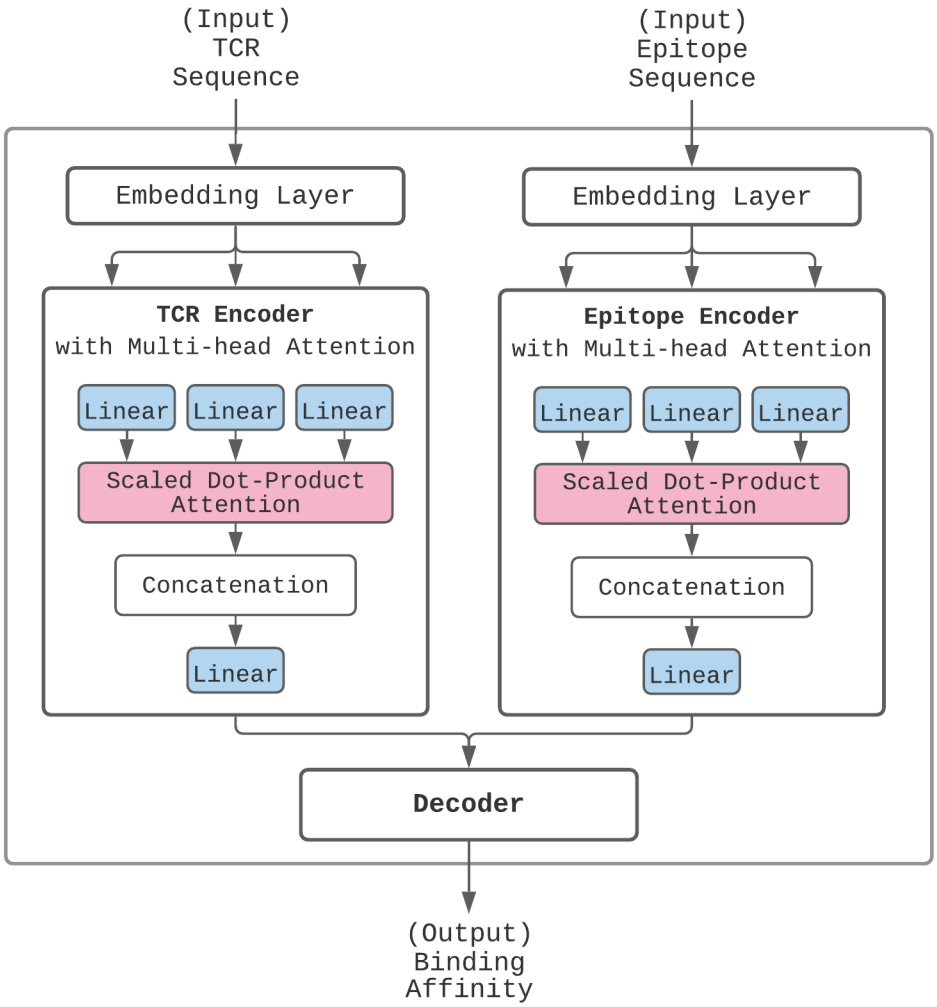

Figure 2.4: A depiction of the multi-head attention model

rich encodings of sentences (Peters *et al.*, 2018; Devlin *et al.*, 2019). An additional multi-head attention model using two attention layers was created and trained on the same hyperparameter sets. This model was found to have no significant differences in performance compared to the single-layer attention model. The single-layer attention model was kept instead to minimize training overhead. Additionally, multiple activation functions were tested for the dense layer in the decoder. The model was trained multiple times using each of the PyTorch implementations of Swish, ReLU, LeakyReLU, GELU, and Tanh across several hidden layer sizes (Paszke *et al.*, 2019).

Swish was able to consistently outperform the other activation functions for all layer sizes. This resulted in it being chosen as the primary activation function for the decoder.

2.5 Baseline Comparison Models

Three additional models were chosen as baselines to measure the multi-head attention model's effectiveness. The chosen models; ERGO's LSTM model, ERGO's AE model, and netTCR's CNN model; were trained on the TCR split and epitope split using their best-performing hyperparameters as reported by their corresponding literature (Tickotsky *et al.*, 2017; Jurtz *et al.*, 2018). Training for all three models used the same dataset as the multi-head attention model. The models would all be trained on equivalent ratios of positive and negative samples. This enabled fair comparisons without bias from dataset differences. For measuring performance using cross-validation, the indices for the 5-folds were recorded and utilized across all models to ensure training and testing data remained consistent.

Chapter 3

RESULTS

The multi-head attention model performed better than the other baseline models for the binding affinity prediction task on the TCR split. However, all models performed poorly on the epitope split and were unreliably able to distinguish between binding and non-binding pairs. In further analysis of the multi-head attention model hyperparameters, two were determined to have the most impact on performance. The hidden layer size and the maximum length of the input sequences. An examination of the sequence lengths in the dataset shows that for a majority of sequences, the entire sequence is necessary for determining binding affinity and not just specific regions. However, sequences with larger lengths are not correlated to more accurate predictions.

In further experimentation to improve the performance of the multi-head-attention model on out-of-sample data, the learned attention matrix was analyzed as a confidence measure for predictions. Similar epitopes to the SARS-CoV-2 epitopes were found in the training data and used as baselines. The attention matrices were calculated for the baseline epitope's cognate TCRs. These matrices were then averaged into different groups based on the model's original predictions. The Euclidean distances between the attention matrices for the SARS-CoV-2 TCRs and the baseline attention matrices were compared. It was found that the distances between the attention matrices were significantly closer when a SARS-CoV-2 TCR belonged to a similar prediction group as the baseline TCRs. This led to the conclusion that the attention matrix can be used to establish confidence in out-of-sample predictions.

3.1 Prediction Performance

The model's performance for the binding affinity prediction tasks was evaluated on data splits created using the previously described TCR split and epitope split. Table 3.1 shows that the multi-head attention model outperforms other models in the TCR split for the AUC and recall performance metrics. Only one other model, ERGO's AE model, is able to outperform the multi-head attention model in terms of precision. The multi-head attention model's recall score is an especially distinguishing feature. These prediction models are intended to reduce the pool of TCR candidates which can be considered for treatment development. Recall is an important metric for this reason, as models should still be sensitive to potential TCR candidates even if it comes at some cost of selectivity.

Notably, the results in table 3.2 highlight a massive performance difference between the TCR split and the epitope split for all tested models. Despite being trained on the same pool of data, the exclusion of specific testing epitopes from the training set dramatically decreases performance across the board. The middling performance metrics for each model as well as the high variances indicate that all models essentially became random classifiers. The multi-head-attention model performed similarly on out-of-sample SARS-CoV-2 data as it did on the epitope split with an overall recall of 64.26%. The individual recall for each epitope was 65.13% for YLQPRTFLL and 57.14% for RLQSLQTYV.

Part of this disparity can be explained by the differences in data distribution between the two types of sequences. A large proportion of the TCRs in the database are unique and many share common motifs in their sequences. There are several groups of TCR sequences that only differ by a few amino acid substitutions. These groups often do not bind to the same few epitopes. The addition of the negative data

generated through random recombination allows the model to see several examples of similar TCRs with both binding and non-binding affinities to a particular epitope. This enables the multi-head-attention model and other baseline models to learn the changes to TCRs that impact their epitope-specific binding affinities.

On the other hand, the epitope sequences do not have this same overlap of features and motifs. Most epitopes in the dataset are distinct from others with multiple amino acid changes. This makes it difficult to extrapolate their learned structures to another epitope in the dataset. The epitope split ensures that an epitope in the testing set has never been seen by the model before. Without a diverse set of examples or similar sequences, the learned models struggle to apply their limited information to make an accurate prediction. The near 50% AUC and high variance for all models on the epitope split indicates that key features for determining binding are still missing.

During the hyperparameter tuning stage, two hyperparameters were identified as having the greatest effect on model performance. The first of which was the size of hidden layers within the dense layer. The second was the lengths the sequences were uniformly padded to. The multi-head attention model benefited from having high length bounds on the sequences. Reducing these bounds below 15 amino acids was very detrimental to the performance of the model. Similarly, having unbounded lengths on the sequences and simply padding each sequence to the longest length in the dataset also resulted in a decrease in performance.

On further analysis of the data, the average epitope length was 12.36 and the average TCR length was 14.38. This seems to indicate the entire epitope and TCR sequence may be crucial for binding and not just specific positions. However, in the final model, the same distribution of lengths for TCRs occurred in both the correctly and incorrectly predicted groups. This indicates that TCRs with longer protein sequences do not necessarily encode more information for this task than their

shorter counterparts. Subsequently, initializing the embedding layer using BLOSUM matrices was found to have little to no effect on the performance of the model. On average, a random initialization of the embedding matrix performed better than a BLOSUM initialized embedding matrix.

Table 3.1: Performance comparison of the multi-head attention model to other baseline models on the TCR split.

Model	TCR Split		
	AUC	Recall	Precision
Multi-head Attention	75.76% (±0.26%)	82.52% (±2.43%)	62.58% (±0.91%)
ERGO - AE	75.66% (±0.19%)	74.16% (±1.28%)	65.44% (±0.57%)
ERGO - LSTM	72.40% (±0.27%)	76.24% (±3.35%)	62.12% (±0.67%)
netTCR - CNN	72.94% (±0.84%)	81.35% (±2.74%)	61.43% (±0.67%)

Table 3.2: Performance comparison of the multi-head attention model to other baseline models on the epitope split.

Model	Epitope Split		
	AUC	Recall	Precision
Multi-head Attention	51.58% (±5.06%)	59.58% (±18.50%)	49.01% (±3.33%)
ERGO - AE	50.25% (±1.39%)	56.84% (±8.73%)	47.89% (±2.49%)
ERGO - LSTM	49.02% (±3.47%)	57.84% (±8.85%)	49.19% (±1.76%)
netTCR - CNN	54.17% (±3.16%)	62.17% (±15.29%)	52.63% (±1.37%)

3.2 Attention as a Confidence Measure

The multi-head attention model and baseline models all demonstrate limited performance when trained on the epitope split. Furthermore, the multi-head attention

model shows similar results on the out-of-sample SARS-CoV-2 data. A key feature is still missing from each of the models when it comes to predicting binding for epitopes that have never been seen by them. In this section, I show that the multi-head attention model's attention matrix $\boldsymbol{W}$ can be used to further improve prediction performance in such situations.

Intuitively it can be assumed that two TCR sequences that bind to the same epitope share similar positional inter-relationships between their amino acids. It was further assumed that two similar epitopes' cognate TCRs would also share similar amino acid relationships. This would result in the TCRs having similar attention matrix representations in the multi-head attention model. While on the other hand, a TCR which binds to a distant epitope would have distinct amino acid relationships and thus a dissimilar attention matrix. Based on these assumptions, it was hypothesized that previous predictions for epitopes could be used to validate future predictions on similar out-of-sample epitopes by using the attention matrices for their cognate TCRs.

To test this hypothesis, two baseline epitopes were found by comparing the lengths of the longest common subsequences between the training data epitopes and the SARS-CoV-2 epitopes. Pairs that contained these two baseline epitopes were then retrieved from the original training data. For each baseline epitope, its paired TCRs were then sorted into one of four confusion matrix categories based on the multi-head attention model's binding affinity predictions: true-positive (TP), false-positive (FP), true-negative (TN), and false-negative (FN). This resulted in two groups of four categories, one group for each baseline epitope. For each of these categories, a reference attention matrix was calculated by averaging the attention matrices for each of the TCRs in that particular category.

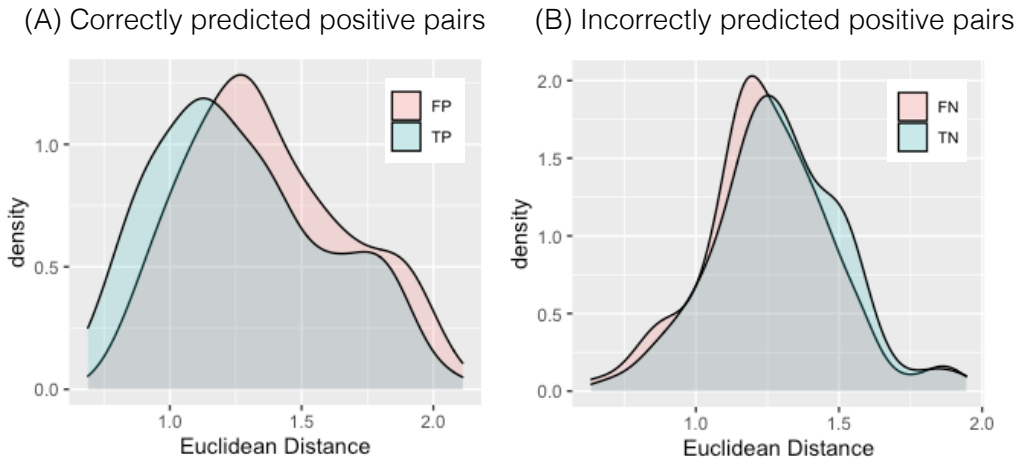

Figure 3.1: Distribution of the distance between (A) the correctly predicted TCRs and the FP/TP reference matrices, and (B) the incorrectly predicted TCRs and the FN/TN reference matrices.

If a TCR was predicted as binding to an epitope in the SARS-CoV-2 dataset, then the attention matrix for that TCR was compared to the true-positive and false-positive reference matrices for the corresponding baseline epitope. If a TCR has a closer attention matrix to that of the true-positive reference matrix, then it can be affirmed that it is a binding TCR. Otherwise, if it shares more similarity with the false-positive reference matrix, it should be reconsidered as a non-binding TCR. In a similar manner, a TCR which was predicted as non-binding to a SARS-CoV-2 epitope should be compared to the true-negative and false-negative reference matrices for the corresponding baseline epitope. It follows such that, similarity with the true-negative matrix will affirm the prediction as non-binding, while similarity with the false-negative matrix indicates the TCR should be reconsidered as binding.

The Euclidean distance was utilized as the distance metric to calculate the similarity of the attention matrices and reference matrices. Figure 3.1 shows the comparison between the Euclidean distances for correctly predicted pairs on the left and incorrectly predicted pairs on the right. In a paired t-test it was observed that TCRs that were predicted to bind to the SARS-CoV-2 epitope had attentions significantly closer

($p < 2.2 \times 10^{-16}$) to that of the true-positive TCRs than those which were false-positives. Similarly, the incorrectly predicted TCRs have attentions significantly closer ($p < 2.2 \times 10^{-16}$) to the false-negative TCRs than the true-negative TCRs. These distances were also found to not be significantly correlated with the model scores (correlation values between $0.05 - 0.23$). This indicates that the distance to each reference matrix is a distinct measure from the model that can be used for evaluating the confidence of predictions on out-of-sample data.

Chapter 4

FUTURE WORK

Although the attention matrices can be used as confidence measures for predictions on out-of-sample data, the epitope split and SARS-CoV-2 dataset performance indicate that additional features might be necessary for binding predictions. The protein sequences in this model are represented as strings of amino acid symbols. But in reality, both TCRs and epitopes are 3-dimensional structures with various grooves and loops which impact their ability to adhere to each other in a binding (Serçinoğlu and Ozbek, 2020). The model is forced to infer the amino acid relationships between the two protein sequences without this knowledge of their structure. Adding these structural features to the data may result in large boosts to model performance.

Even without this 3-dimensional structure information, the multi-head attention model's ability to learn sequence interactions can still be further improved. Currently, the model utilizes a naive padding technique for both the TCR and epitope sequences which arbitrarily pads or trims amino acids in the middle of the sequences until they are a standard length. An alignment method that can match the binding regions between TCRs and epitopes would ensure the conservation of amino acids which impact the binding the most.

The success of the multi-head attention model also indicates that other NLP solutions may be applicable to this problem. Although the BLOSUM matrices were unable to enhance model performance, other embedding types can be used for the protein sequences. For example, the embedding layer may be able to be improved using recent NLP embedding models. In lieu of the embedding and attention layers, the embeddings for each sequence can be calculated using ELMO or BERT and di-

rectly fed into the decoder (Peters *et al.*, 2018; Devlin *et al.*, 2019). These models may be able to extract some additional complex characteristics of the sequences and improve model performances.

Chapter 5

CONCLUSION

The multi-head self-attention model shows promising results for the TCR-epitope binding prediction task. The attention mechanism helps the model learn biological contextual representations of both TCR and epitope sequences. By selectively correlating the amino acids in each sequence the model gains insight into the structural changes that occur when a single position's amino acid is changed. It demonstrated its ability to apply this knowledge by outperforming the other baseline models on the TCR split of data. Despite this, both the multi-head attention model and baseline models failed to make accurate predictions on the epitope split of the data. However, the multi-head attention model's learned attention matrix demonstrated its usefulness as an interpretable model. The attention matrix was successfully used as a confidence measure for evaluating out-of-sample predictions. This is evident of the multi-head attention model's ability to learn contextual representations that are translatable across sequences.

REFERENCES

Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", (2019).

Gielis, S., P. Moris, W. Bittremieux, N. De Neuter, B. Ogunjimi, K. Laukens and P. Meysman, "Detection of enriched T cell epitope specificity in full T cell receptor sequence repertoires", Frontiers in Immunology 10, 2820 (2019).

Henikoff, S. and J. G. Henikoff, "Amino acid substitution matrices from protein blocks.", Proceedings of the National Academy of Sciences 89, 22, 10915–10919 (1992).

Hinton, G. E., N. Srivastava, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors", CoRR abs/1207.0580, URL `http://arxiv.org/abs/1207.0580` (2012).

Jokinen, E., J. Huuhtanen, S. Mustjoki, M. Heinonen and H. Lähdesmäki, "Determining epitope specificity of T cell receptors with TCRGP", bioRxiv p. 542332 (2019).

Jurtz, V. I., L. E. Jessen, A. K. Bentzen, M. C. Jespersen, S. Mahajan, R. Vita, K. K. Jensen, P. Marcatili, S. R. Hadrup, B. Peters *et al.*, "NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks", bioRxiv p. 433706 (2018).

Kingma, D. P. and J. Ba, "Adam: A method for stochastic optimization", arXiv preprint arXiv:1412.6980 (2014).

Koury, J., M. Lucero, C. Cato, L. Chang, J. Geiger, D. Henry, J. Hernandez, F. Hung, P. Kaur, G. Teskey and et al., "Immunotherapies: Exploiting the immune system for cancer treatment", Journal of Immunology Research 2018, 1–16 (2018).

Lythe, G., R. E. Callard, R. L. Hoare and C. Molina-París, "How many tcr clonotypes does a body maintain?", Journal of Theoretical Biology 389, 214–224 (2016).

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library", in "Advances in Neural Information Processing Systems 32", pp. 8024–8035 (Curran Associates, Inc., 2019).

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, "Deep contextualized word representations", in "Proc. of NAACL", (2018).

Ramachandran, P., B. Zoph and Q. V. Le, "Searching for activation functions", CoRR abs/1710.05941, URL `http://arxiv.org/abs/1710.05941` (2017).

Schumacher, T. N., W. Scheper and P. Kvistborg, "Cancer neoantigens", Annual Review of Immunology 37, 173–200 (2019).

Seliger, B., "Strategies of tumor immune evasion", BioDrugs 19, 6, 347–354 (2005).

Serçinoğlu, O. and P. Ozbek, "Sequence-structure-function relationships in class i mhc: A local frustration perspective", PLOS ONE 15, 5 (2020).

Sewell, A. K., "Why must t cells be cross-reactive?", Nature Reviews Immunology 12, 9, 669–677 (2012).

Shugay, M., D. V. Bagaev, I. V. Zvyagin, R. M. Vroomans, J. C. Crawford, G. Dolton, E. A. Komech, A. L. Sycheva, A. E. Koneva, E. S. Egorov *et al.*, "VDJdb: a curated database of T-cell receptor sequences with known antigen specificity", Nucleic Acids Research 46, D1, D419–D427 (2018).

Springer, I., H. Besser, N. Tickotsky-Moskovitz, S. Dvorkin and Y. Louzoun, "Prediction of specific tcr-peptide binding from large dictionaries of tcr-peptide pairs", Frontiers in Immunology 11, 1803, URL `https://www.frontiersin.org/article/10.3389/fimmu.2020.01803` (2020).

Tickotsky, N., T. Sagiv, J. Prilusky, E. Shifrut and N. Friedman, "McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences", Bioinformatics 33, 18, 2924–2929, URL `https://doi.org/10.1093/bioinformatics/btx286` (2017).

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need", CoRR abs/1706.03762, URL `http://arxiv.org/abs/1706.03762` (2017).

Vita, R., S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler, A. Sette and B. Peters, "The immune epitope database (IEDB): 2018 update", Nucleic Acids Research 47, D1, D339–D343 (2019).

Wilson, J. H. and T. Hunt, *T Cells and MHC Proteins* (Garland Science, 2002).