Doctoral Thesis

Transformer encoders for predicting T cell receptor-peptide binding can associate attention weights with interpretable protein structural properties

by

Kyohei Koyama

supervised by Prof. Kenji Mizuguchi

Graduate School of Frontier Biosciences
Osaka University

February 2024

Abstract

This dissertation presents a novel exploration into T cell receptors (TCRs), crucial components of the immune system that interact with ligand peptides. These peptides, presented by Major Histocompatibility Complex (MHC) molecules, are recognized by TCRs to initiate immune responses. The interaction between a TCR and its corresponding peptide-MHC complex (pMHC) is fundamental to initiating protective immune reactions against pathogens and malfunctioning cells. Understanding TCR-pMHC interactions is, therefore, vital for analyzing immune system mechanics, designing vaccines, and developing targeted immunotherapies. However, current experimental methodologies for studying TCR-pMHC interactions are resource-intensive and time-consuming, with computational models limited to retrospective data analysis and lacking interpretability. Additionally, the prediction of the TCR-pMHC interaction is difficult due to the massive combination patterns of TCRs and peptides. Addressing these challenges, this work introduces a novel approach using a machine learning model with a modified Transformer encoder, employing a source-target-attention neural network, or cross-attention layer. Central to this research is the development of a model that predicts TCR-pMHC interactions from amino acid sequences of the TCR' s complementarity-determining region 3 (CDR3) and peptides. Unique to this study is the utilization of an external prospective dataset and the Transformer encoder layer to examine TCR-pMHC structural properties through attention weights. The model demonstrates superior performance on benchmark test sets and external datasets, surpassing other models in the average precision score, although the score limitation of the model is revealed by visualizing the data distribution difference. A detailed analysis links neural network attention weights to protein structural properties, classifying residues into attended groups to identify statistically significant properties, such as hydrogen bonds within CDR3, not between CDR3s and peptides. Chapters 2 and 3 of the dissertation delve into the cross-attention mechanism's predictive power and its interpretability in TCR-pMHC interactions, with Chapter 3 specifically comparing the efficacy of cross-attention and standard-attention mechanisms. The findings affirm the cross-attention model's superiority in revealing interaction dynamics at the molecular level, thus confirming more interpretability. In summary, this dissertation contributes substantially to bioinformatics and immunological studies using the Transformer-based attention's interpretability, providing a pathway toward more effective and interpretable computational tools. The insights acquired hold significant implications for the prediction of TCR-pMHC interactions, and this research not only enhances our understanding of molecular recognition but also lays the groundwork for developing new therapeutic approaches.

Acknowledgements

I wish to express my deep-seated gratitude to my supervisor, Kenji Mizuguchi, whose insightful guidance, precision, and unerring acumen have been invaluable. His patience, support, and flexibility he demonstrated, particularly in facilitating remote working conditions, have played a crucial role in my academic journey.

Equally, I owe a profound debt of gratitude to the exceptional team members including Kosuke Hashimoto, Chio Nagao, and all those in our lab. Their profound biological knowledge and assistance have been important for this research. Thank you, Kosuke, for helping me a lot on revising the papers submitted and Chio for advising on proteins. Without their collective wisdom and guidance, this achievement would not have been possible.

Furthermore, I extend my earnest appreciation to my colleagues and superiors at SyntheticGestalt, where I have been privileged to pursue my academic degree.

Finally, I extend my deepest gratitude to my wife, Yui. Her support has been the best. I also express my gratitude to my daughter, Nanasa, for reminding me of the importance of curiosity and prioritization. You both have raised me up, and I love both of you.

Part of this dissertation, especially Chapter 2, has been published in the following article [1].

- Kyohei Koyama, Kosuke Hashimoto, Chioko Nagao, and Kenji Mizuguchi. Attention network for predicting T cell receptor-peptide binding can associate attention with interpretable protein structural properties. *Frontiers in Bioinformatics*, Volume 3, 2023. doi: 10.3389/fbinf.2023.1274599.

This work was partly achieved through the use of SQUID at the Cybermedia Center, Osaka University.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

This chapter provides a comprehensive introduction to the interactions between T cell receptors and peptides. It begins with an overview of the fundamental aspects of this interaction. Subsequently, the chapter delves into an examination of previous studies focused on predicting the binding of T cell receptor and peptide. Following this, an overview of self-attention and cross-attention mechanisms is presented, offering a brief look into these concepts. Finally, the chapter ends by outlining the contributions of this dissertation and describing the primary objectives of the study.

1.1 T cell receptor and peptide binding

The T cell receptor (TCR) serves as an antigen receptor, primarily composed of alpha (TCRα) and beta (TCRβ) chains. The TCRs are vital for recognizing antigenic peptides presented by the major histocompatibility complex (MHC) molecule. The molecule recognized by the TCR is called peptide-major histocompatibility complex (pMHC) and the interaction between the two leads to the biological responses.

In a broader context, understanding TCR-pMHC interactions is crucial for effective immune surveillance, enabling the identification and elimination of pathogens and cancer cells. It also provides insights into autoimmune diseases where these interactions malfunction. Crucially, this understanding aids vaccine development, informing the design of antigens that can be effectively recognized by T cells. Additionally, it supports advancements in cancer immunotherapy, such as in designing T cell based therapies, and plays a role in addressing transplant rejection and managing infectious diseases. The study of TCR-pMHC interactions is pivotal in shaping the approach to various medical challenges and enhancing healthcare outcomes.

However, the TCRs have an enormous sequence diversity in their complementarity determining regions 3 (CDR3s), similar to antibodies, B cell receptors. The CDR3s of TCR, found in both α and β chains (CDR3α and CDR3β, respectively), are the most diverse parts that are produced through somatic recombination. The sequence diversity of peptides in nature as a form of pathogens is immense needless to say, therefore, CDR's potential responses with different peptides are almost infinite patterns. Also, determining protein functions experimentally in the lab can be resource-intensive and time-consuming, involving many human resources and lab facilities.

Therefore, predicting and confirming TCR-pMHC interaction, primarily involving CDR3-peptide binding, is difficult, despite its importance. This prediction has the potential to greatly enhance our understanding of biological processes and mechanisms underlying diseases, and it could inform strategies for disease treatment and recovery.

1.2 Related previous work on T cell receptor predictions

Generally, experimentally determining protein functions can be challenging due to the significant resources and the need for various techniques in biochemistry, genetics, and structural biology. Understanding the functions of proteins is often enhanced by examining their three-dimensional (3D) structures [2–4], utilizing methods like X-ray crystallography and cryo-electron microscopy. However, these methods are expensive and require a lot of time.

In response to this challenge, there have been many machine learning (ML) methods created for TCR-pMHC prediction [2, 5–10]. There also exist researches trying to develop the attention models to predict the TCR-pMHC binding [11–14]. Notably, when performing predictions of computational models based on cellular assay data regarding the recognition of pMHC by TCRs, the term, "TCR-pMHC interactions" is appropriate despite the absence of MHC or the non-CDR3 TCR sequence in the computational model inputs.

The Transformer-related models like BERT [15, 16] are famous for their exceptional performance and their interpretability [17–19]. They have shown the effectiveness of the cross-attention mechanism, or source-target-attention, in tasks involving multiple inputs like machine translation or image-text classification [20–22]. There are studies in the bioinformatics field using the source-target-attention [23–26]. Additionally, using cross-attention for two distinct sequences during training is more efficient than applying self-attention to combined sequences, as the computational load of the Transformer's attention escalates quadratically with the input sequence length. However, despite the

Transformer's broad usage, there's a lack of in-depth interpretative analysis for multi-input tasks like the TCR-pMHC protein complex. Few studies have attempted to use the Transformer's source-target-attention model to examine individual residues in CDR3$\alpha\beta$ or peptides, particularly for structural aspects like hydrogen bonds.

For instance, Vig et al.[27] found that attention values are more critical in protein binding sites after training large models and conducting statistical tests. In contrast to their approach, my model leverages pairs of CDR3 and peptide, enhancing the modeling of the relationship. While models like NetTCR-2.0 [5] and ERGO-II [8] show strong predictive capabilities, they rely on convolutional or recurrent neural networks. PanPep[10] uses attention but is limited to CDR3β, ignoring key residues of the α chains and interaction factors related to H-bonds. TCR-BERT[11] considers both α and β chains but lacks peptide training and attended structural analysis with bonds. AttnTAP[13] applies attention but without directly using Transformer attention for both TCR and peptide, and also excludes the α chain. DLpTCR [14] employs ResNet attention but does not use Transformer attention.

To summarize those models in comparison to this study, I provide a rough view of the comparison in Table 1.1. The model in this study presents the pivotal difference of key factors in using Transformer-based attention and its interpretation analysis on structures, with the data both of peptide binding and α chain.

TABLE 1.1: Comparison with other models in terms of key factors.

	Net-TCR2.0	ERGO-II	Pan Pep	TCR-BERT	Attn TAP	DLp TCR	This Study
Uses the Transformer attention	NO	NO	YES	YES	YES	NO	YES
Uses only the cross-attention for interactions	NO	NO	NO	NO	NO	NO	YES
Analyzes structures on each residue	NO	NO	YES	YES	NO	NO	YES
Analyzes structural features statistically	NO	NO	YES*	YES	NO	NO	YES
Predicts TCR-peptide binding	YES	YES	YES	NO	YES	YES	YES
Utilizes α chain's CDR3	YES	YES	NO	YES	NO	YES	YES

*Although Panpep shows statistical test, no research has been made on Hydrogen bonds

1.3 Advancing TCR-Peptide interaction analysis; beyond traditional self-attention to cross-attention

The attention layer proposed in the Transformer [15] represents a great computational strategy that learns the conditional relationship between the two sequences, namely the source and target sequences. Although attention mechanisms are widely used in prediction tasks, studies on TCR-peptide interactions have yet to perform statistical analysis of attention weights. Such analysis could uncover relationships within sequence data provided another sequence and highlight significant structural features like hydrogen bonds (H-bonds).

Current approaches predominantly use deep learning models, such as those similar to BERT, which incorporate attention networks [16–19]. However, I think the self-attention mechanism in the BERT model, when it comes to the residue binding analysis, holds a few limitations. Primarily, the BERT attention covers the two sequences of proteins simultaneously at the same layer, which means the self-attention mechanism of the BERT model does not specifically focus on mutual interactions. The attention of BERT is dependent on the two entire sequences, making it harder to analyze the interaction between the two sequences. Also, it is conceivable that self-attention might predict a positive result with only a specific TCR part without considering its relationship with the peptide. In other words, the prediction from those self-attention values may yield a positive value due to only the self-coincidence relationship inherent to a sequence, using only one side of the information. When one gets a largely attended TCR residue from the self-attention approach of concatenated TCR-peptides analysis, that attention might be due to only TCR side information of TCR-TCR attention weights. Furthermore, the numerous stacked layers in BERT and its computational complexity can complicate the analysis of attention layer results.

Therefore, I have limited the attention mechanism exclusively to the intersecting segment, overcoming these limitations by employing a modified attention layer, referred to as a cross-attention layer or a source-target-attention layer [23–26], along with an extensive analysis of protein structures. Chapter 3 of this dissertation will compare the standard-attention and cross-attention in terms of their predictive performance. The cross-attention framework is designed to address the BERT's limitation, ensuring an attention analysis of TCR-peptide interactions.

1.4 Contribution of the dissertation

In this thesis, by summarizing the published research [1], I aimed to contribute to the TCR research regarding two key points: providing the very competitive prediction model on TCR-pMHC benchmark datasets and revealing relationships between the attention weights and protein structural properties.

The model demonstrated superior performance on benchmark datasets in our previous article [1], particularly when evaluated using the average precision score which is more appropriate for assessing datasets with few positive samples than the ROCAUC score. Additionally, while I acknowledge the limitations in generalizing my model to unseen data, such challenges are not exclusive to my approach but are also present in other models.

A key contribution in the previous work [1] was the application of statistical tests to attention values in protein structures. By performing statistical tests on the attention values over the complex structures, the previous work [1] successfully identified statistically significant structural properties of largely attended residues such as hydrogen bonds and residue distance. This statistical test on attention values was only possible because I used the source-target-attention neural network as the source-target-attention pays attention to a sequence given the other side. The modified cross-attention layer provides clearer insights into the binding relationship between TCR and peptides, avoiding the interpretability issues seen in standard-attention models like BERT.

In this dissertation, I included Chapter 3 in addition to the previous work [1], where I improved the previous model by changing the hyperparameters, compared the cross-attention model with the standard Transformer encoder, and included more PDB structures to analyze the attention weights. In summary, my approach not only advances TCR-pMHC interaction prediction but also enhances the interpretability of these predictions, linking them to structural protein properties in a way that has not been previously explored. This should advance the research of molecule recognition and the design of new therapeutics.

1.5 The main objective of this dissertation

This dissertation aims to develop a new computational methodology for TCR-pMHC interaction analysis. Distinct from existing research, it focuses on developing a method that integrates CDR3α, CDR3β, and peptide sequences. Also, it can provide detailed

residue-wise structural analysis. This approach leverages a Transformer-based attention mechanism applied to sequence data.

I hypothesize that the attention layer can accurately predict TCR-pMHC interaction and provide interpretable biological insights about the CDR3-peptide binding, or TCR function. To achieve this objective, I propose a model, Cross-TCR-Interpreter, which uses the cross-attention layers for predicting TCR-pMHC interaction, the binding between a peptide and CDR3 regions of both the α and β chains.

Following this introduction, Chapter 2 presents our previously published work [1], which lays the groundwork for this research. Chapter 3 delves into a comparative analysis between self-attention and cross-attention mechanisms, exploring the advantages of the cross-attention model in the context of TCR-pMHC interaction prediction. The dissertation ends with the conclusion chapter, which synthesizes the findings and implications of this research.

Chapter 2

Transformer-based model for predicting TCR-peptide binding and interpretability

This chapter is based on our paper "Attention network for predicting T cell receptor-peptide binding can associate attention with interpretable protein structural properties" [1], for which I am the first author. This journal's editorial office has confirmed that, as an author, I retain the copyright for the article including images and I am free to use it in the dissertation. To maintain the academic standards of this work, the contents of the paper are appropriately adapted and integrated within the context of this dissertation. This approach ensures continuity in the narrative and alignment with the broader themes and objectives of my research.

2.1 Abstract

Understanding the way a T cell receptor (TCR) identifies its corresponding ligand peptide is key to gaining insight into biological processes and the mechanisms behind diseases. However, experimentally analyzing interactions between TCR, peptide, and major histocompatibility complex (TCR-pMHC) is both costly and time-consuming. In response, computational techniques have been developed. Yet, these are usually only assessed through internal retrospective validation, and few have integrated and examined an attention layer from language models in the context of structural data.

Therefore, in this research, I developed a machine learning (ML) model using an adapted version of the Transformer, a source-target-attention neural network, to predict the interaction of TCR-pMHC using just the amino acid sequences of TCR's complementarity-determining region (CDR) 3 and the peptide.

The model exhibited competitive performance on both benchmark TCR-pMHC datasets and a new external prospective dataset. Furthermore, I associated the neural network's weights with the structural attributes of proteins. By categorizing residues into groups of high and low attention, I uncovered statistically significant characteristics linked to residues receiving large attention values, like hydrogen bonds within the CDR3. The creation of this dataset and the model's capacity for providing interpretable predictions of TCR-peptide binding are steps forward in enhancing our understanding of molecular recognition and could lead to the development of novel therapeutic solutions.

2.2 Introduction

In this chapter, I introduce our previously published work [1], focusing on the interpretability of the Transformer models of the T cell receptor (TCR). The TCR, serving as a crucial antigen receptor, is primarily composed of α and β chains. It has a remarkable sequence diversity in its complementarity-determining region 3 (CDR3). The CDR3 region of the TCR, located in both the α and β chains (referred to as CDR3α and CDR3β), is remarkably diverse and plays a crucial role in identifying antigenic peptides presented by the major histocompatibility complex (MHC) molecule.

This chapter describes a computational method that can incorporate the protein sequence pairs of CDR3α, CDR3β, and peptide while enabling a residue-wise structural analysis and leveraging a Transformer-based [15, 16] attention mechanism on the protein sequences. I hypothesize that an attention-based neural network can accurately predict TCR-peptide binding and provide interpretable biological insights into the TCR function. I should be able to answer the question of why the interaction happens based on two amino acid sequences. To achieve the interpretation purpose, I created a cross-attention-based model using Transformer encoders. I tested the models on benchmark test sets, a prospective test dataset that is chronologically distinct from the training dataset, and a test dataset derived from a study on Covid-19. The attention weights with crystal structures were analyzed to see any meaningful features on amino acid residues and a paired t-test was conducted to see any property difference between the attended residues and not-attended residues. Also, the input perturbation method was utilized to analyze the effect of the change on the input CR3 sequences of attended residues, where I changed the input CR3 sequences and observed any changes in attention values.

2.3 Method

2.3.1 Model of this study

An overview of the models used in this chapter is shown in Figure 2.1, which is the same as the model Cross-TCR-Interpreter from our previous paper [1]. In the model overview depicted in Figure 2.1, the sequences of the peptide and the connected CDR3α and CDR3β (linked by a colon ":") were processed independently in the embedding layer and Transformer. These layers were designed to extract the sequence information independently in a way not related to each other. They were then fed into a specially designed cross-attention layer for predicting sequence relationships. Two cross-attention layers were utilized to form a layer focused solely on mutual interactions, allowing the

model to assess their relationship. The outputs from the cross-attention layer were merged and then averaged across their length in the output layer. It's worth noting that in Chapter 3, the cross-attention layers will be substituted with a standard-attention layer as an ablation and comparative model study. A multi-layer perceptron (MLP) layer produces a prediction value in the form of a real value, termed the confidence value, ranging from 0 to 1. In contrast, an actual binding data point was stored as a binary value, either 0 or 1. Therefore, the loss function employed was Binary Cross Entropy (BCE), and the output of the model was assessed using both the ROCAUC score and the average precision score.

The model processes only the amino acid sequences of CDR3α, CDR3β, and peptide, focusing on the CDR3s rather than the full TCR sequences. It does not incorporate additional data like gene types. I believe relying purely on sequence data, without the inclusion of domain-specific expertise such as gene or MHC information, is crucial for mimicking interpretability, closely mirroring the natural binding processes of CDR3. The sequences of CDR3 and peptides were encoded using the standard 20 amino acids. Additionally, positional embeddings and padding tokens were integrated into these sequences. Padding at the end of each sequence was performed to ensure the lengths of each CDR3 sequence aligned with the maximum sequence length in the training data; hence, each CDR3α had the same length. This was also performed for CDR3β and peptide. The maximum length and the minimum length for the datasets used in this study are provided in the Results section.

2.3.2 Training and test datasets preparation for sequences

In the realm of TCR-pMHC interaction prediction, particularly concerning CDR3s and peptide binding datasets, I utilized the same datasets as the study [1]. They used The ERGO-II repository [8], which incorporates McPAS [28] and VDJdb [29]. Additionally, I independently sourced and compiled a more recent version of the VDJdb and Covid-19 datasets [30].

In more detail, the sequence datasets used for this dissertation include:

- **McPAS and VDJdb-without10x as benchmark datasets (training set and test set)**: These foundational datasets, McPAS and VDJdb (excluding 10x genomics data, termed VDJdb-without10x), were obtained from the ERGO-II repository. Both datasets comprised training and test sets and encompassed a mix of positive and negative TCR-pMHC interactions.

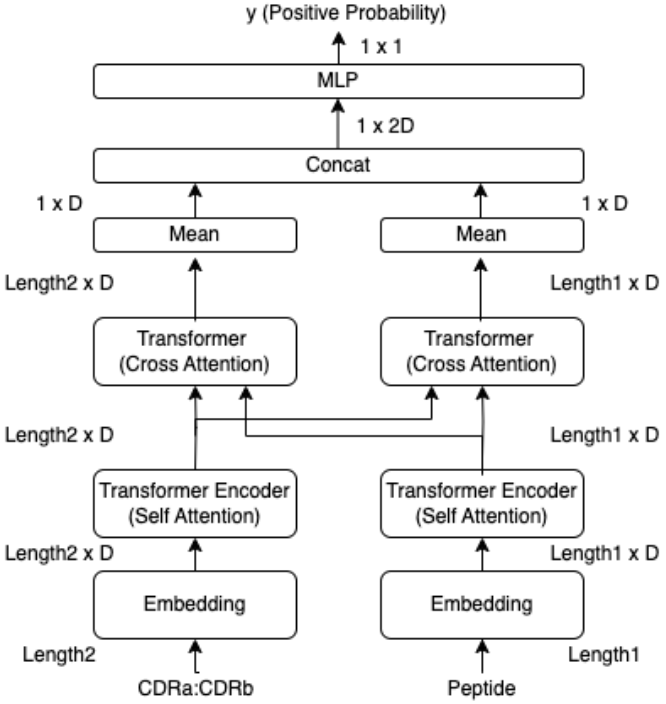

FIGURE 2.1: Overview of the Cross-TCR-Interpreter model, which is reproduced from our published paper [1]. The sizes of data tensors are displayed. The cross-attention layers, located centrally in the diagram, underwent analysis with structural data following training on sequence data. Each embedding layer is designed to receive amino acid sequences as input.

- **Combined data dataset (training set)**: To develop a more extensive dataset for model training, I assembled a merged dataset, referred to as the "combined data". This dataset combined VDJdb-without10x, VDJdb with the 10x genomics data (VDJdb-with10x), and the McPAS dataset. This dataset was instrumental in training the model for subsequent evaluation against the Covid-19 dataset and the recent data test set.

- **Recent data test set (test set)**: Aiming to validate the model's performance on novel and unseen data, I applied the model trained on the combined data to a recent test set from VDJdb downloaded in 2023. To accommodate the dataset's predominantly positive TCR-pMHC interactions, I included randomly selected negative TCR-pMHC pairs. This approach was pivotal in assessing the model's predictive accuracy on the latest TCR-pMHC interactions in the realistic setting.

- **Covid-19 dataset (test set)**: As the last and most stringent dataset to provide the assessment of the combined-data-trained model, I created a dataset derived from the Covid-19 study[30]. This dataset poses a significant challenge to the model that was trained using the combined data dataset.

This research, centered on a binary classification framework, necessitated the inclusion of negative label data for model training. Given that the majority of available TCR and peptide interaction data are positively labeled, the methodology mirrored the approach of the ERGO-II method that generated random CDR3-peptide pairs and assigned negative labels to adjust the positive-negative ratio. Notably, the volume of negative data incorporated was five-fold that of the positive data. Therefore, each data record comprised a tuple of CDR3α, CDR3β, peptide with an assigned binary interaction label. Data records lacking either CDR3α or CDR3β sequences were excluded from the training set or from the test set to maintain data integrity and ensure accuracy in binary interaction labeling.

Furthermore, to avoid data redundancy in the analysis, any CDR3α, CDR3β, peptide combinations found in both the training and test sets were meticulously removed from the test set. Despite these efforts, it's important to note that there may be instances of repeated CDR3α, CDR3β pairs or individual CDR3 sequences or peptides in both sets. This occurrence is due to the representation of identical TCRs in both sets, which may be associated with different peptides. The extent and implications of these potential duplications within the McPAS and VDJdb datasets are further explored in the Results section.

2.3.3 Benchmark dataset and experiment

The validity of the cross-attention model was assessed by benchmarking it against established datasets, specifically McPAS and VDJdb excluding 10x Genomics data (referred to as VDJdb-without10x). The performance of the model was compared to that of notable benchmark models such as ERGO-II [8] and NetTCR2.0 [5], which utilize both CDR3α and CDR3β chains. Additionally, models focusing exclusively on the CDR3β chain, including NetTCR2.0[5], PanPep [10], AttnTAP [13], and DLpTCR [14], were also considered for comparison. In these datasets, binary labels were attributed to the pairs of CDR3β and peptides.

The performance of the proposed model was thoroughly evaluated, not just on the complete test set, but also on a per-peptide basis within the test set. The benchmark datasets utilized in the ERGO-II study were constructed by integrating assumed negative instances and subsequently dividing them into training and test datasets. This method, however, might oversimplify the problem due to the potential overlap of peptides or CDRs in both training and test datasets.

The process for creating the detailed benchmark dataset of McPAS and VDJdb involved:

- Step 1: Download the test and training sets from ERGO-II, removing any data records that lack CDR3α, CDR3β, or peptide.
- Step 2: Exclude from the test set any records with duplicated pairs that appear in the training set.

During the training phase, the focus was on minimizing the binary cross-entropy for the benchmark datasets. The model underwent hyperparameter optimization using a validation dataset. Training ceased if there was no improvement in binary cross entropy after 10 consecutive updates. The optimal model was then determined based on the weights that yielded the minimum binary cross-entropy value.

2.3.4 The combined data dataset and the recent data test set

Following the validation of the model's performance on benchmark datasets, I retrained the model with a whole dataset referred to as the "combined data" dataset. This dataset combined McPAS, VDJdb-without10x, and VDJdb-with10x. The objective behind employing the combined data dataset was to delve deeper into the TCR-pMHC interactions, aiming to clarify the binding dynamics and derive insightful interpretations from the model.

The use of this combined data dataset was intended to facilitate the learning of relationships between sequences within the attention layer of the model. Notably, the inclusion of the 10x dataset [31], which was previously excluded from the benchmark experiments, was a strategic choice. This incorporation aimed to integrate an extensive range of binding-related information into the model, thereby enabling a thorough analysis of the attention weights in the trained model.

For testing the model, I selected the most up-to-date data from VDJdb, constituting the "recent data" test set, which comprised data downloaded between 2022 and 2023. In contrast, the training set consisted of VDJdb data acquired before 2022 and the McPAS data. Subsequent to data acquisition, the recent data test set was augmented with a 5 times greater volume of negative data records, exclusively sampled from the test set itself and not from the combined data dataset.

This approach of using the recent data test set aimed to simulate a real-world scenario where the model undergoes prospective validation, thus evaluating its performance in a forward-looking manner and non-retrospectively. The process of assembling the combined data dataset and the recent data test set involved the following steps:

- Step 1: Download data from McPAS, VDJdb-without10x, and VDJdb-with10x from the ERGO-II repository. Perform concatenation and eliminate records missing any of the components: CDR3α, CDR3β, or peptide.
- Step 2: Remove any duplicated pairs within the dataset (the combined data dataset).
- Step 3: Acquire the latest VDJdb data as of June 2023, and construct tuples of CDR3α, CDR3β, and peptide (forming the recent data test set).
- Step 4: Remove, from the recent data test set, any records with duplicated pairs that overlap with the training set.
- Step 5: Enrich the recent data test set with fivefold more negative data records.

To describe and understand how diverse the recent data and the combined data dataset were, the sequence-sequence pairwise distance matrix was calculated using Clustal Omega software [32] for sequence space analysis. A key distinction between the recent test and benchmark test sets is the timing of the data split, especially before the addition of assumed negative samples in the recent data test set. This was to avoid the oversimplification problem often encountered in such datasets. To illustrate the diversity within the recent data and the combined data dataset, I conducted a sequence-space analysis using a pairwise distance matrix calculated via Clustal Omega software [32].

2.3.5 Covid-19 data and experiment

To further assess the real-world applicability of the model trained on the combined data, I conducted an evaluation using a Covid-19 dataset derived from a recent study [30]. This experiment aimed to test the model's accuracy in scenarios where the peptides involved are previously unknown, akin to the conditions presented by the Covid-19 dataset.

For this purpose, I constructed a simulated dataset using TCR pairs and peptides associated with the S(Spike) protein from the Covid-19 study. In the original research, the interaction between these peptides and TCRs was determined through a reporter cell assay, which involved measuring green fluorescent protein expression indicative of TCR activation. The peptides in that study were synthesized with a 15-residue window, shifting by four residues each time.

In my virtual replication, I adjusted this approach to better align with the data characteristics of the combined data dataset. Specifically, I created peptides using a 9-residue window, which reflects the median peptide length in the combined data dataset and

shifted by one residue at a time. This ensured no overlap between the peptides in the combined data dataset and the 9-residue peptides from the Covid-19 dataset.

To quantify and demonstrate the diversity of the peptides within the Covid-19 dataset, I employed the same sequence-sequence pairwise distance matrix analysis as used for the combined data dataset. This analysis was instrumental in highlighting the distinct nature of the Covid-19 peptides compared to those in the combined data dataset.

2.3.6 Hyperparameters

The model's hyperparameters are fine-tuned using the Optuna package [33] and oprimized by using random 20% of the training records. Apart from hyperparameter tuning, the actual training was carried out on a single A100 GPU node at Osaka University's SQUID cluster, taking about 3 hours. The inference process was completed on a 2.6 GHz 6-Core Intel Core i7 CPU, taking roughly 2 hours. The best parameters are $learning\_rate$ = 9.387e-05, d_ff of the final MLP layer = 84, $dropout\_rate$ = 7.651e-05, and dim in the model = 256. The repeated count of Transformer layers for the self-attention layers in Figure 2.1 is two on both sides of the CDR3s and the peptide. The n_head, number of heads in the Transformer encoders is four. The number of cross-attention layers is one.

For the previously published model [1], the positive weight applied in the Binary Cross-Entropy (BCE) calculation was 15.0. In terms of sequence length, the maximum lengths for TCRα and TCRβ were padded to 62, and for the peptide, to 26. While more extensive and exhaustive searches might yield different outcomes, initial explorations into pre-training and transfer learning did not result in any significant improvements in the final score.

2.3.7 Analysis of residues with high attention values using 3D structural data

Following the model's training on the combined data dataset, I gained the capability to extract attention matrices for any given residue. I argue that it makes sense to analyze the model since I used the data that are correctly predicted. The objective was not to selectively pick data but to examine and interpret the significant features as identified by the model.

By categorizing the residues into two groups, those with high and low attention, it became feasible to examine the attention values in detail. For each head where CDR3s

were given attention provided by a peptide, the definition of the CDR3s residue indices of large attention values is $R_{large,h}$ in Equation 2.1.

$$R_{large,h} = \left\{ t | \max_{p} a_{t,p} > \bar{a} + \gamma \cdot \sigma \right\}$$
$$\text{where } h \text{ denotes head}$$
$$\text{and } a_{t,p} \text{ denotes an attention value} \tag{2.1}$$
$$\text{of CDR3 residue index } t \text{ and peptide residue index } p.$$

$$R_{large,all} = Concat_h(R_{large,h}) \tag{2.2}$$

The TCR side attention is described in Equation 2.1. Given a head h in the cross-attention layer, let A_h be a TCR side attention matrix, with elements $a_{t,p}$. Notably, in the definition of Equation 3.1, $\sum_t a_{t,p}$ can be the one-dimensional all-one vector having the length of P, the peptide's length. The one dimensional all-one vector is $(1_1, 1_2, \cdots, 1_p, \cdots, 1_P)$. This definition of A_h is the TCR-side attention because each p assigns the attention to TCRs as a sum of one. The function $\max_p$ selects the highest value along the peptide axis. The symbol $\bar{a}$ represents the average of the attention values in A_h, and σ denotes the standard deviation (STD) of A_h. The factor γ is used to empirically determine the criteria for classifying values as large or small, as further detailed in the Results section subsection 2.4.6. For calculating the peptide residues that received significant attention, I interchanged the notation of t and p. In determining the residues with less attention, I substituted the in-equation operator with a less-than symbol ("<"). Equation 2.2 illustrates the highly attended TCR side residues when all heads are combined.

2.3.8 Analysis of structural data from the Protein Data Bank (PDB)

In this study, I focused on analyzing attended residues by attention using TCR-pMHC complex structures sourced from the Protein Data Bank (PDB) [34]. To compile relevant TCR structures, I utilized the PDB Search and the SCEptRe server [4], which specializes in collecting TCR complex structures. The SCEptRe dataset referenced here was accessed on June 2, 2021. Through this process, 65 structures featuring both α and β chains were identified, with the Anarci tool [35] being instrumental in extracting the CDR3 segments of these structures.

Subsequently, I refined this selection to 55 from 65 structures by applying the lengths criteria of TCRs and peptide sequences. Within these 55 structures, I identified eight pairs that shared identical sequences in their CDR3s and peptides. Consequently, the final analysis was conducted on the distinct sequences of 47 structures. Please note that the same sequences produce identical attention matrix and y-probability, and the same sequences would bias the statistical test, thus duplicated sequences were eliminated.

To statistically evaluate the differences between the two groups of residues -those receiving substantial attention and those that did not -I employed a paired student's t-test, also known as a dependent t-test. This statistical method is designed to compare the means of two interrelated groups. In the context of this research, the TCR-pMHC complex structures served as the subjects for this t-test. The variables tested in the t-test included properties like the percentage of TCR residues forming hydrogen bonds with the peptide and whether a particular residue was involved in a hydrogen bond, etc. The extraction and analysis of these structural properties were facilitated using BioPython [36] and LIGPLOT [37].

2.3.9 Input perturbation

In an effort to delve deeper into specific instances, I applied the input perturbation technique, mirroring the same approach used in our prior study [1]. This method is designed to test the model's sensitivity to alterations in its inputs, serving as a fine-grained analysis that enhances the broader insights gained from the paired t-test on grouped data.

The process of input perturbation entails the strategic substitution of certain amino acids at key positions with different amino acids and then observing the subsequent shifts in the model's prediction and attention values. This approach would be thought of as a similar approach to the Alanine scan, which examines a specific protein residue functionality by substituting it with Alanine. Through this method, I evaluated how the model reacts to changes in the residues that received significant attention. This analysis provided valuable insights into the dynamics of the model's predictions and attention mechanisms in response to these alterations.

2.4 Result

2.4.1 Overview of the study and types of experiments conducted

Three distinct experiments were conducted to demonstrate the effectiveness of the Cross-TCR-Interpreter model developed in this study.

The initial experiment involved training and validating the model with established benchmark datasets. This phase included a comparative analysis of the model's performance against previously developed models in other studies.

For the second experiment, the focus shifted to external prospective validation of TCR-pMHC interactions. Here, the model was retrained using the "combined data" dataset and subsequently validated against two specific datasets: the Covid-19 dataset and the recent data test set.

The third and final experiment was geared towards elucidating the explainability aspect of the model. In this phase, the model, trained with combined data, was applied to a dataset of TCR-pMHC complexes with known 3D structures. This application facilitated a detailed statistical analysis of the cross-attention values, shedding light on the biochemical binding events between CDR3 and peptides. Additionally, input perturbation analysis was incorporated to monitor the changes in attention and y-probability driven by alterations in the input. This experiment underscored that, although the model's training was solely based on sequence data, its predictive capabilities could be effectively interpreted and augmented using structural data.

2.4.2 The presence of unique element overlap and record-wise overlap in datasets can account for the difficulties of datasets

I have used the same sequence dataset as the previous research [1]. The key statistics summary of the sequence datasets are described in Table 2.2. The training records of benchmark datasets are 19,526 for VDJdb-without10x and 23,363 for McPAS. The records of test sets are 4,010 for VDJdb-without10x and 4,729 for McPAS with no duplicated records between the test set and the training dataset. Additionally, the sequence length for each sequence dataset is described in Table 2.1, and Figure 2.2 shows the distribution of the length for each dataset.

In Table 2.2, the unique counts of CDR3s, peptides, and their combinations are presented. For example, the McPAS training set contains 23,363 records, comprising 3,181 distinct CDR3s sequences and 316 distinct peptides, with a positive rate of 16.67%. Of

TABLE 2.1: Sequence length for each dataset, which is reproduced from our published paper [1]. For the median and mean, the data record was counted on each record (=pair) basis. The distribution of the length is provided in Figure 2.2.

Dataset	Sequence	Max	Min	Median	Mean
McPAS	CDR3α	26	6	13	13.27
	CDR3β	21	7	14	13.79
	Peptide	25	8	9	9.761
VDJdb-without10x	CDR3α	22	5	13	13.37
	CDR3β	21	8	13	13.76
	Peptide	20	8	9	9.462
The combined data dataset	CDR3α	26	5	14	13.61
	CDR3β	26	7	14	14.37
	Peptide	25	7	9	9.520
The Covid-19 dataset	CDR3α	20	6	14	13.69
	CDR3β	21	10	15	14.60
	Peptide	9	9	9	9.00

these sequences, 833 CDR3s and 190 peptides also appear in the test dataset, yet no identical CDR3-peptide pairs are repeated in the test set. Theoretically, under ideal circumstances, if every unique CDR3 sequence were paired with every unique peptide, it would result in 1,005,196 possible combinations ($= 3181 \cdot 316$). However, due to inherent limitations in the available data, this comprehensive pairing is not achieved in reality.

Moreover, there are minimal duplicated CDR3s and peptides shared between the combined data dataset and both the recent data test set and the Covid-19 dataset. This lack of overlap in the sequences contributes to the challenge in accurately predicting TCR-pMHC interactions for the recent data test set and the Covid-19 dataset.

Figure 2.3 and Table A.2 detail the pair-wise duplication in each test dataset. This duplication means the count of records where either the peptide or CDRs overlap with those in the training dataset. The data reveals that the test set records of McPAS and VDJdb-without10x consist primarily of peptides already encountered in the training dataset. Conversely, 14.8% of the records in the recent data test set include known peptides, while none of the Covid-19 dataset peptides appear in the combined data dataset. Both the recent data test set and the Covid-19 dataset predominantly feature records with CDRs or peptides that have not been seen before.

For example, the McPAS test set contains 4,729 records, of which 4,683 records include peptides that are already present in the training set, and only 46 records hold entirely new peptides. In contrast, the recent data test set comprises 33,360 records, but only 4,938 of these include peptides that were seen in the training set.

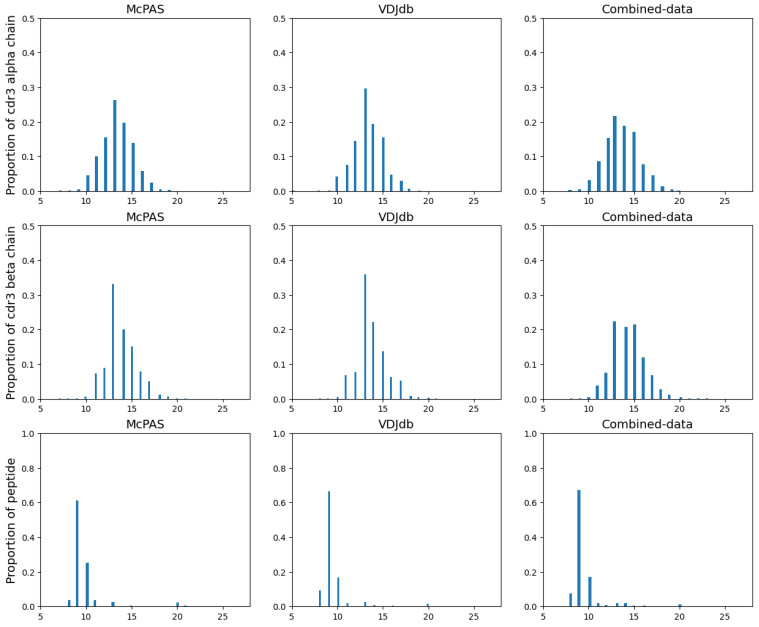

FIGURE 2.2: The distribution of the length for each dataset, which is reproduced from our published paper [1].

TABLE 2.2: Statistics of datasets which are reproduced from our published paper [1]. The "Record" column indicates the unique count of CDR3α, CDR3β, Peptide pairs, while CDR3$\alpha\beta$ refers to the unique count of CDR3α and CDR3β pairs. The "in duplication" row under the "Unique count" column represents the count of unique data shared between training and test sets, i.e. overlapped data. The "Pos. Rate" column specifies the ratio of positive instances in the binary label.

Dataset name	Unique count	CDR3$\alpha\beta$	Peptide	Record	Pos. Rate
McPAS	in training	3181	316	23363	0.1665
McPAS	in test	833	190	4729	0.1512
-	in duplication b/w training and test	132	171	0	N/A
VDJdb-without10x	in training	2902	175	19526	0.1670
VDJdb-without10x	in test	689	120	4010	0.1504
-	in duplication b/w training and test	111	111	0	N/A
The combined data dataset (A)	in training	23299	478	119046	0.1400
The recent data test set (B)	in test	33183	838	33360	0.1667
The Covid-19 dataset (C)	in test	1676	1265	2120140	$1.887 \cdot 10^{-5}$
-	in duplication b/w (A) and (B)	18	44	0	N/A
-	in duplication b/w (A) and (C)	1	0	0	N/A

2.4.3 Superior performance of the model in the benchmark datasets

To evaluate the performance of the model, I used the same training and test datasets as in the previous research [1] that was inspired by those of ERGO-II [8]. Two benchmark datasets, McPAS and VDJdb without 10x Genomics data (VDJdb-without10x), were prepared.

When assessed using both the ROCAUC score and the average precision score, the Cross-TCR-Interpreter model demonstrated robust and competitive performance in comparison to other models across benchmark datasets, specifically within models based only on sequence features (Table 2.3).

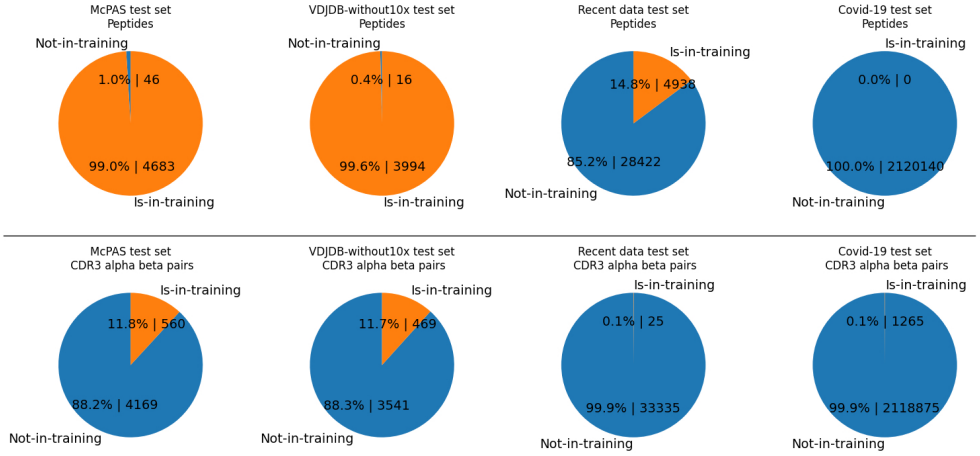

FIGURE 2.3: The pie charts display the proportion and count of test set data records containing elements previously seen in the training data and unseen elements, as detailed in our published paper [1]. Contrasting with the data outlined in Table 2.2, these figures represent the counts of test records that share either peptides or CDRs with the training set. Each pie chart shows the number of test records with duplicated CDR3 pairs or peptides from the training dataset. 'Is-in-training' denotes that the peptides or CDR3s are from the training dataset, while 'Not-in-training' indicates they are unique to the test set, and not found in the training. The total counts of test data records for McPAS, VDJdb-without10x, the recent data test set, and the Covid-19 test set are 4,729, 4,010, 33,360, and 2,120,140 respectively. Displayed in two rows, the upper row represents peptides and the lower row shows CDR3$\alpha\beta$ data. Each column corresponds to a different dataset, starting from the left with the McPAS test set, VDJdb-without10x test set, recent data test set, and finally the Covid-19 test set. For instance, of the 4,729 records in the McPAS test set, 4,683 involve peptides previously seen in training, while 46 feature entirely new peptides. In terms of CDR3s, 560 out of 4,729 records involve unseen CDR3s, while the remaining are composed of CDR3s encountered during training.

While I was able to access the performance metrics for ERGO-II's top model from their published research repository, their cessation of weight updates using the test set, presumably to enhance the utility of their code repository, prevented me from precisely replicating the predictions of their highest-performing model. This limitation posed a challenge in conducting an equitable comparison based on average precision.

To further elaborate on the model's performance, particularly on a per-peptide basis, I computed and detailed the scores for the eight most prevalent peptides in the test sets (Figure 2.4). This analysis using the Cross-TCR-Interpreter model revealed that it holds up well in comparison to the NetTCR2.0[5] model, especially in the context of per-peptide performance metrics.

Given the significant impact that sequence similarity can have on protein prediction accuracy, it is crucial to eliminate sequences in the test set that closely resemble those in the training set. Addressing this, I have included an analysis in the Discussion section that explores the impact of TCR sequence distance on performance variation.

TABLE 2.3: Result of benchmark dataset of McPAS and VDJdb, which is reproduced from our published paper [1]. APS stands for the average precision score.

Dataset	Model	Features in addition to peptides	ROCAUC	APS
McPAS	Cross-TCR-Interpreter [1]	CDR3s of α and β chains	0.9154	0.6211
	NetTCR2.0	CDR3s of α and β chains	0.9204	0.5808
	PanPep	CDR3 Sequence of β chain with biochemical features	0.8374	0.4519
	AttnTAP [1]	CDR3 Sequence of β chain	0.840	-
	DLpTCR [1]	CDR3 Sequence of β chain	0.633	-
	ERGO-II, LSTM [2]	CDR3s of α and β chains	0.855	-
	ERGO-II, LSTM [2]	CDR3s of α and β chains, VJ genes and MHC Type	0.939	-
VDJdb	Cross-TCR-Interpreter [1]	CDR3s of α and β chains	0.9445	0.7600
	NetTCR2.0	CDR3s of α and β chains	0.9492	0.7262
	PanPep	CDR3 Sequence of β chain with biochemical features	0.9009	0.6435
	AttnTAP [1]	CDR3 Sequence of β chain	0.894	-
	DLpTCR [1]	CDR3 Sequence of β chain	0.622	-
	ERGO-II, LSTM [2]	CDR3s of α and β	0.800	-
	ERGO-II, LSTM [2]	CDR3s of α and β chains, VJ genes and MHC Type	0.866	-

1. The numbers were derived from the AttnTAP paper because my implementation only achieved a maximum AUC value of 0.6. Hence, to avoid potential misinterpretation due to poor scores, I have opted not to display the average precision score in this context.
2. The scores for ERGO-II's best model were obtained directly from their research paper. Their ROCAUC for McPAS and VDJdb were 0.939 and 0.866, respectively. However, their ceased weight updates prevented me from replicating their top-performing model's predictions accurately, thus hindering a comparison on the average precision.

2.4.4 The models exhibit limited performance in the recent data test set

Subsequent to establishing the model's performance, I undertook the retraining of the model using a combined dataset, hereafter referred to as the "combined data dataset," which included both McPAS and the complete VDJdb along with the 10x Genomics dataset [31]. The model, having been trained with the combined data, was then applied to the most recent dataset, termed the "recent data test set," to evaluate their performance in real-world TCR-pMHC interaction prediction scenarios.

The aim of using the combined data dataset was primarily to discover pertinent relationships and closely emulate the binding dynamics of TCR-pMHC interactions or CDR3-peptide binding. However, as indicated in Table 2.4, the majority of the models, including mine, struggled to surpass a 0.55 ROCAUC score with the purely recent data dataset, unlike in the benchmark tests where the models achieved 0.9 ROCAUC. This

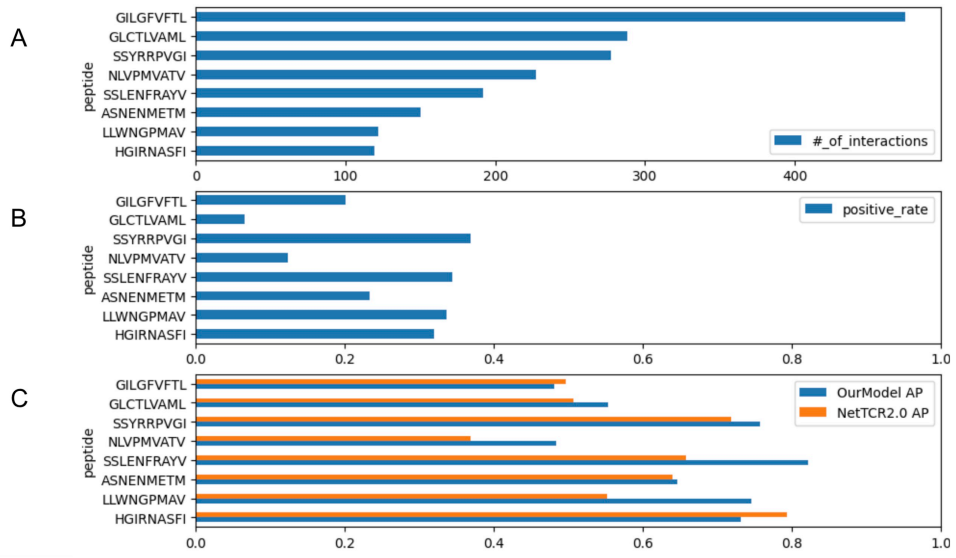

FIGURE 2.4: Performance comparison of two models on the frequent peptide subsets (Reproduced from our published paper [1]). (A) The number of records for each peptide, (B) the positive rate inside the records (the ratio of positively recorded CDR3$\alpha\beta$), and (C) the average precision scores (AP) of Cross-TCR-Interpreter [1] and the NetTCR2.0 model on the benchmark test set of McPAS.

limitation can be attributed to the challenges posed by the inclusion of records with previously unseen CDRs or peptides.

After the training, the Cross-TCR-Interpreter model achieved an ROCAUC score of 0.952 and an average precision of 0.7952 on the training dataset. Yet, when applied to the recent data test set (Table 2.4), there was a noticeable drop in performance, with the ROCAUC and average precision decreasing to 0.5362 and 0.1855, respectively. However, as also discussed in [1], when the test set was limited to known peptides, there was a relative improvement in average precision, rising to 0.3318 (Table 3.2). Conversely, focusing solely on new peptides unseen in the training dataset led to a further decline in average precision to 0.1707.

This trend of underperformance was not exclusive to the Cross-TCR-Interpreter model; other models like NetTCR2.0 and PanPep exhibited similar challenges in the recent data test set and its subcategories. In the case of the PanPep model, despite its zero-shot setting for unseen peptides and a majority setting for known peptides, the average precision score only slightly improved. The performance for the test data subset of known peptides was even lower than for the new peptide subset.

These findings highlight that, despite advances, the current computational models, including this research, are still not sufficient to replace wet lab experiments fully. It

underscores the significant hurdle in accurately predicting CDR3-peptide interactions involving peptides that were not present in the training dataset.

TABLE 2.4: Result of the recent data test dataset (Reproduced from our published paper [1]). APS stands for the average precision score.

Model	Dataset	ROCAUC	APS	Number of data records	Pos. Rate
Cross-TCR-Interpreter	Recent data test set	0.5362	0.1855	33360	0.1667
	- Recent data test set of new peptide subset	0.5085	0.1707	28422	0.1662
	- Recent data test set of known peptide subset	0.6598	0.3318	4938	0.1692
	- Recent data test set of new CDR3s subset	0.5355	0.1844	33335	0.1660
NetTCR2.0	Recent data test set	0.5274	0.1808	33360	0.1667
	- Recent data test set of new peptide subset	0.5113	0.1705	28422	0.1662
	- Recent data test set of known peptide subset	0.6327	0.3008	4938	0.1692
	- Recent data test set of new CDR3s subset	0.5267	0.1798	33335	0.1660
PanPep *	Recent data test set	0.5337	0.1897	30221	0.1745
	- Recent data test set of new peptide subset	0.5359	0.1908	25661	0.1739
	- Recent data test set of known peptide subset	0.5199	0.1852	4560	0.1779
	- Recent data test set of new CDR3s subset	0.5374	0.1923	29145	0.1752

The scores for the test set comprising only known CDR3s could not be computed as all the data records are positive.
However, when setting a threshold at 0.5, the model achieves a recall score of 0.56, in comparison to NetTCR2.0's score of 0.44 and PanPep's 0.59.
* The datasets in the Cross-TCR-Interpreter and NetTCR2.0 were identical, but the dataset utilized in PanPep was different due to its exclusive use of CDR3β. Consequently, by eliminating duplicates of the β chain CDR3 from the test set, the number of data records was reduced.

2.4.5 The Cross-TCR-Interpreter model does not exhibit satisfactory performance for the Covid-19 dataset

This subsection assesses the Cross-TCR-Interpreter model's performance in dealing with novel peptide scenarios, using the recently released Covid-19 dataset [30] as a test case. The model, trained with the combined data, was applied to gauge its capability in accurately predicting TCR-pMHC interactions in a practical scenario. As outlined in the Methods section, this dataset involved peptides from SARS-CoV-2 proteins, each featuring a 9-residue length, and generated by shifting one residue at a time. It is important to note that none of the peptides in the Covid-19 dataset were included in the combined data dataset.

The total number of data records was 2,120,140, of which 2,120,100 were negative data records and only 40 data records were positive. In the 2,120,140 data records, there were 1,676 unique CDR3 alpha-beta pairs and 1,265 unique peptides ($1,676 \cdot 1,265 = 2,120,140$, shown in Table 2.2). Out of the 40 positive records, I found 10 unique CDR3$\alpha\beta$ pairs and 24 unique peptides. Consequently, this means that the remaining 200 records, composed of these specific CDR3s and peptides, are classified as negative records ($10 \cdot 24 - 40 = 200$).

By maximizing the F1-score of this prediction task, the model achieved a precision of $2.501 \cdot 10^{-5}$, and a recall of 0.600. The confusion matrix revealed 24 True Positives, 16 False Negatives, 959,512 False Positives, and 1,160,588 True Negatives. The calculated ROCAUC score stood at 0.5461, and the average precision score was a mere 2.032×10^{-5}. Considering the exceedingly low occurrence rate of positive records at 1.887×10^{-5} (= 40/2120140), the model demonstrated an ability to identify positive records only 1.326 times (= 2.501/1.887) more effectively than random chance. However, the specificity of the model was insufficient to consider it a viable alternative to traditional wet lab experiments.

Table 2.5: PDB identifiers of structural data. The third row shows the 39 PDB IDs used in the attention analysis.

55 PDBIDs prior to processing	1D9K, 1FYT, 1J8H, 1LP9, 1U3H, 2BNQ, 2BNR, 2ICW, 2J8U, 2NX5, 2UWE, 2VLJ, 2VLK, 2VLR, 2YPL, 2Z31, 3MBE, 3MV7, 3MV8, 3MV9, 3PQY, 3QIU, 3VXR, 3VXS, 3VXU, 3W0W, 4JFD, 4JFE, 4JRX, 4JRY, 4MJI, 4OZF, 4OZG, 4OZH, 4P2O, 4P2Q, 4P2R, 4QOK, 4Z7V, 5BRZ, 5D2L, 5ISZ, 5KS9, 5MEN, 5NHT, 5TEZ, 5WKF, 5WKH, 5WLG, 6AVF, 6AVG, 6EQA, 6EQB, 6Q3S, 6RPB
47 PDBIDs having no duplicated sequences	1D9K, 1J8H, 1U3H, 2BNR, 2ICW, 2J8U, 2NX5, 2VLK, 2VLR, 2YPL, 2Z31, 3MBE, 3MV8, 3PQY, 3QIU, 3VXR, 3VXS, 3VXU, 4JFD, 4JFE, 4JRX, 4JRY, 4MJI, 4OZF, 4OZG, 4OZH, 4P2O, 4P2Q, 4P2R, 4QOK, 4Z7V, 5BRZ, 5D2L, 5ISZ, 5KS9, 5MEN, 5NHT, 5TEZ, 5WKF, 5WKH, 5WLG, 6AVF, 6AVG, 6EQA, 6EQB, 6Q3S, 6RPB
39 PDB IDs having positive predictions	1D9K, 1J8H, 1U3H, 2BNR, 2J8U, 2NX5, 2VLK, 2VLR, 2YPL, 2Z31, 3MBE, 3MV8, 3PQY, 3QIU, 3VXR, 3VXS, 3VXU, 4JFD, 4JFE, 4JRX, 4JRY, 4MJI, 4OZF, 4OZG, 4OZH, 4P2O, 4P2Q, 4P2R, 4QOK, 4Z7V, 5D2L, 5MEN, 5NHT, 5TEZ, 5WKF, 5WKH, 6EQA, 6EQB, 6Q3S
9 PDB IDs that are not in the combined data dataset	2J8U, 3VXU, 4JFD, 4QOK, 4JFE, 5NHT, 5MEN, 6EQA, 6EQB

2.4.6 Structural data categorized residues based on their level of attention, into groups of highly attended and less attended ones

While I didn't achieve perfect generalizability, I focused on interpreting the model using 39 complex structures where the model showed adequate performance for analysis. As outlined in the Methods section, I began by selecting 47 TCR-related structures from the SCEptRe server. Out of these, the model identified 39 as showing positive TCR-pMHC interactions, based on a cutoff point of 0.5.

Interestingly, 30 of these 39 structures had sequences that were also present in the combined data dataset. My analysis particularly concentrated on these 39 cases, as I believed the model's predictions and interpretations to be reliable for them. This approach is somewhat akin to performing a regression analysis where I explore the influence of certain variables on the outcome, with the aim here being to pinpoint key features that the model recognizes, specifically the characteristics of the most focused-on amino acid residues.

Detailed information about these 39 structures is provided in Table 2.5. For this analysis, there are 39 PDB IDs involved. Of these, 9 were not part of the combined data dataset, while the remaining 30 were included.

I defined the attention values as "large" if they were higher than the average (MEAN) plus 5.5 times the standard deviation (STD) for the peptide side and MEAN plus 4.5 STD for the TCR side. These numbers, 5.5 and 4.5, are referred to as γs in Equation 2.1. Using this method, I identified about 20% of the residues as having large attention values on each side. The thresholds were determined through empirical evaluation, and the residue count generated by changing γ is provided in the Appendix. For each PDB entry or head, the threshold for what counts as large attention can be different. You can find more about how I picked these thresholds and the number of residues they identified in the Appendix.

I carried out this analysis for each of the four heads in the cross-attention layer (heads 0 to 3) of both the TCR and the peptide, looking at each head on its own. The cross-attention layer works between a sequence of CDR3$\alpha\beta$ and a peptide. This creates an attention matrix, the size of the lengths of the peptide and the CDR3$\alpha\beta$ residues. It' s possible for a specific residue to get attention in one head but not in others, as shown in Equation 2.1.

For example, in our previous work [1], I looked at the attention values for the TCR-pMHC complex with the PDB entry 5TEZ. You can see this as eight heatmaps in Figure 2.5. The 5TEZ structure includes MHC class I HLA-A2, influenza A virus, and

TCRs [38]. I highlighted the residues getting a lot of attention in different colors. In Figure 2.6, you can see the 3D structure of this TCR-pMHC complex. The peptide, CDR3α, and CDR3β sequences for this complex are `GILGFVFTL`, `CAASFIIQGAGKLVF`, and `CASSLLGGWSEAFF`, respectively.

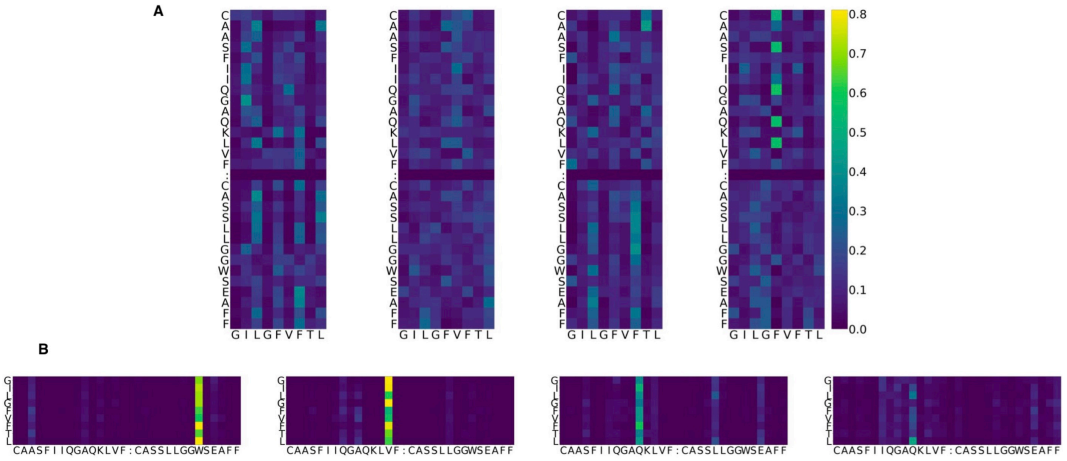

FIGURE 2.5: Example visualization of attention values for PDB ID 5TEZ [38] as reproduced from our published paper [1]. (A) The upper figures display attention values of a peptide provided a CDR3$\alpha\beta$ pair, normalizing the sum to 1 across the peptide. The X-axis indicates the peptide residues, and the Y-axis shows CDR3 pair residues. (B) The lower portion reflects the attention of a CDR3 pair provided a peptide, also normalized to sum to 1 across the CDR3s. Here, the X-axis denotes CDR3 residues, and the Y-axis corresponds to peptide residues. Both visualizations ensure the attention sum over the x-direction equals 1. Each column in the four columns corresponds to one head in the multi-head attention layer. The color scale ranges from dark blue for lower attention values to bright yellow for higher values, with green indicating intermediate values. The lower image shows a specific cell of the first head (the leftmost figure) at the intersection of peptide position L_8 (the last row) and CDR3β position W_{24} (sixth column from the right), indicating the significance of CDR3 W_{24} when associated with peptide L_8. Its bright yellow color suggests a high attention value, indicating these residues might have an important biological role during prediction. This value exceeds the set threshold of MEAN + γ STD, which is individually determined for each PDB ID and each head.

2.4.7 Statistical analysis shows largely attended residues form H-Bonds with CDR3

I classified TCR residues into two categories, "large" and "small" attention groups, based on their attention values. This was done using the γ factor of 4.5 as mentioned in Equation 2.1, particularly for the TCR residues' cross-attention when given a peptide. To understand the characteristics of these groups, I looked at their structural properties.

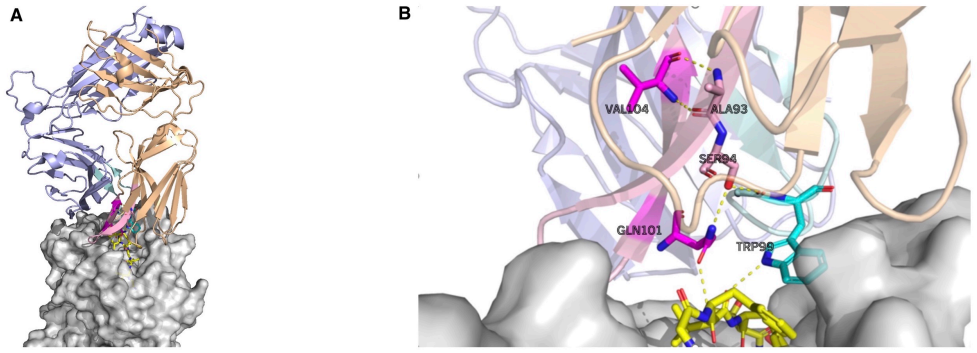

FIGURE 2.6: The visualization highlights attended residues in the TCR interacting with an influenza virus epitope and HLA complex (PDB ID: 5TEZ), with the CDR3 sequences of TCRα and β being `CAASFIIQGAQKLVF` and `CASSLLGGWSEAFF` respectively, as reproduced from our published paper [1]. The left image (A) depicts the overall structure of the complex, while the right image (B) zooms in on the interactions involving largely attended residues: VAL104 (14th Val(V) of TCRα), GLN101 (11th Gln(Q) of TCRα), and TRP99 (9th Trp(W) of TCRβ chain). The TCRα chain is colored wheat, the TCRβ chain in light-blue, the CDR3 of TCRα in light-pink, and the CDR3 of TCRβ in pale-cyan. The largely attended residues in CDR3α and CDR3β are highlighted in magenta and cyan respectively, with MHC in grey. These residues and their interacting partners are shown as sticks and hydrogen bonds are illustrated with yellow dotted lines. VAL104 forms two hydrogen bonds with TCRα's ALA93 (3rd Alanine Ala(A) of TCRα), contributing to the stability of the CDR3 loop conformation. GLN101 forms a hydrogen bond with TCRα SER94, which in turn is bonded to TCRβ, helping to maintain the structure between the α and β chains. Both GLN101 of TCRα and TRP99 of TCRβ form hydrogen bonds with the epitope. The figure was generated using PyMol [39].

To check for differences between these groups, I used a paired t-test. This test helped remove variations that might come from each structure's unique features. For this, 39 TCR-pMHC complex structures were studied, comparing the structural properties linked to both large and small attention groups. The paired t-test aimed to test whether the average difference between these pairs was zero. The test value, like the likelihood of being H-bonded to any peptide residue, was calculated using a formula where $P = A_h/B_h$, with A_h being the number of residues with at least one H-bond of the specified type within the large attention group, and B_h being the number of residues of large attention values, where h denotes the head.

Each of the four heads was analyzed separately, and they showed varying results. The combined results for all heads are in Table 2.6. The individual results for each head are also reported in the Table A.3 and Table A.4 of the Appendix. Since a TCR sequence includes both CDR3 and non-CDR3 parts, I divided the residues into these two categories to measure H-bond properties.

The residues with large attention values often form hydrogen bonds with CDR3 sections of their chain. However, their likelihood of hydrogen bonding with other TCR residues isn't significantly different from less attended residues.

Consequently, those observations indicate that the residues with large attention values tend to avoid hydrogen bonding with the non-CDR3 portions, compared to the residues with small attention values. The most notable difference across all attention heads was observed in the proportion of hydrogen bonds formed with non-CDR3 TCR residues. Residues with high attention values showed a tendency to avoid hydrogen bonding with non-CDR3 regions of the TCR. This indicates that the highly attended residues are more likely to avoid the H-bonded to the non-CDR3 part of TCR, whereas they tend to form H-bonds with the CDR3 portions.

To avoid pitfalls associated with multiple p-values in the statistical analysis, I executed the Benjamini Hochberg (BH) procedure and adjusted the p-values. Here, the *False Discovery Rate (FDR) for BH* represents the likelihood of incurring a Type I error among all rejected null hypotheses. At an FDR threshold of 0.05, only the "H-bonded to any non-CDR3 TCR residue" hypothesis was rejected, demonstrating the rigor of this threshold. While many assertions in my study might be substantiated when considering average metrics, they may not attain statistical significance at this level. Meanwhile, modifying FDR to 0.1 led to the rejection of two hypotheses: "H-bonded to any non-CDR3 TCR residue" and "H-bonded to any CDR3 residue of own chain", which was additionally highlighted by the symbol "**" in Table 2.6. Further increasing FDR to 0.15 expanded the rejections to four hypotheses, adding "H-bonded to any CDR3 residue of opposite chain" and "H-bonded to any CDR3 residue", which are designated by "*" in Table 2.6. Collectively, these statistical evaluations lend support to the hypothesis that attended residues significantly avoid H-bonds with non-CDR3 TCR regions, favoring H-bonds within the CDR3 regions.

Contrary to expectations, there was no significant difference in the proportion of highly attended TCR residues forming hydrogen bonds with any peptide residues across all attention heads. This unexpected finding was surprising, adding value to the analysis. Additionally, when examining the closest distances between TCR residues and any peptide residues, no notable differences were observed.

On the peptide side, as shown in Table 2.7, residues with large attention values demonstrated shorter distances to the nearest TCR residues. This pattern, intriguingly, did not appear on the TCR side, though the findings were not statistically significant. This observation introduces an interesting structural dimension to the study.

TABLE 2.6: TCR-side attention analysis (Reproduced from our published paper [1]). Comparative analysis of structural properties between the residue groups receiving large and small attention is presented. The p-adjusted column shows the adjusted p-value by the Benjamini Hochberg (BH) procedure. The symbol "***" denotes the significant difference based on a False Discovery Rate (FDR) of 0.05 in the BH procedure; the symbol "**" indicates significance with the FDR of 0.10; the symbol "*" indicates the FDR of 0.15. The numbers in the Large Attention or Small Attention columns are the average and standard deviation.

Property	**Large Attention** [1]	**Small Attention** [1]	**p-value**	**p-adjusted**	
H-bonded to any peptide residue	0.0862 ±0.1368	0.0805 ±0.0675	0.828	0.9108	
H-bonded to any CDR3 residue	0.4846 ±0.2216	0.4103 ±0.1040	0.0478	0.1315	*
H-bonded to any non-CDR3 TCR residue	0.2940 ±0.1923	0.4672 ±0.0846	3.880e-05	4.268e-04	***
H-bonded to any TCR residue	0.6845 ±0.1650	0.7294 ±0.0880	0.0987	0.2145	
H-bonded to any CDR3 residue of own chain	0.4643 ±0.2180	0.3752 ±0.0922	0.0107	0.05885	**
H-bonded to any TCR residue of own chain	0.6013 ±0.1999	0.6561 ±0.0880	0.117	0.2145	
H-bonded to any TCR residue of opposite chain	0.1679 ±0.1714	0.1497 ±0.0793	0.562	0.7199	
H-bonded to any CDR3 residue of opposite chain	0.0306 ±0.0857	0.0672 ±0.0743	0.0369	0.1315	*
In the edge [2]	0.6434 ±0.2064	0.5928 ±0.0570	0.218	0.3426	
Closest distance to peptide (Å) [3]	8.4072 ±2.2892	8.4122 ±0.9592	0.988	0.988	
Number of H-bonds formed [3]	2.0234 ±0.9370	2.0875 ±0.6685	0.589	0.7199	

1. Mean and standard deviation (for the 39 structures) of the proportion of residues that satisfy the property shown in the first column.
2. Four residues from the beginning and four from the end of the CDR.
3. In the last two properties, per-residue averages were used instead.

2.4.8 Influence of attended residues on model behaviors: analyzing through input perturbation method

In this subsection, the focus is on understanding how changes in input sequences influence prediction outcomes and attention metrics, a method detailed in our earlier work [1]. This technique was utilized on the training data, PDBID 5TEZ. Additionally, this method was adapted for a mutation study not included in the training data, as described

TABLE 2.7: Peptide side attention analysis (Reproduced from our published paper [1]). Comparative analysis of structural properties between the residue groups receiving large and small attention is presented. The p-adjusted column shows the adjusted p-value by the Benjamini Hochberg procedure.

Property	**Large Attention** [1]	**Small Attention** [1]	**p-value**	**p-adjusted**
H-bonded to any peptide residue	0.0495 ±0.1443	0.0659 ±0.1206	0.458	0.56
H-bonded to any CDR3 residue	0.2050 ±0.3024	0.1682 ±0.0982	0.48	0.56
H-bonded to any TCR residue	0.3401 ±0.3714	0.2372 ±0.1184	0.151	0.3523
H-bonded to any non-CDR3 TCR residue	0.1712 ±0.3112	0.1118 ±0.1283	0.355	0.56
In the edge [2]	0.4459 ±0.4097	0.5874 ±0.1232	0.0795	0.3523
Closest distance to peptide (Å) [3]	4.6398 ±1.7149	5.1926 ±1.2647	0.141	0.3523
Number of H-bonds formed [3]	2.1126 ±1.4959	2.0031 ±0.9051	0.668	0.668

1. Mean and standard deviation (for the PDB structures) of the proportion of residues that satisfy the property shown in the first column.
2. Three residues from the beginning and three from the end of the peptide.
3. In the last two properties, per-residue averages were used instead.

in [40]. This research involved altering the CDR3 β loop sequence in A6-TCR and evaluating its binding efficacy with the TAX peptide, derived from Human T cell leukemia virus type I associated with MHC class I HLA-A2. The sequences and their resulting structures, pre and post-mutation, were documented in the PDB with IDs PDBID 1AO7 (pre-mutation) and PDBID 4FTV (post-mutation).

For the PDB structure 5TEZ, three residues were identified with large attention values: the 11th Gln(Q) and 14th Val(V) in CDR3α, and the 9th Trp(W) in CDR3β (as shown in Table 2.8). I evaluated the impact on predictions and attention values when these residues were replaced with different amino acids. The CDR3α sequence in 5TEZ is *CAASFIIQGAQKLVF*, and the CDR3β sequence is *CASSLLGGWSEAFF*. Notably, while the 11th Gln(Q), 14th Val(V), and 9th Trp(W) all participate in H-bond formation, only the 14th Val(V) of the α chain creates two H-bonds within the TCR' s own CDR3 chain.

Substituting the 14th Val(V) in CDR3α significantly influenced predictions, usually dropping the "unbound" prediction below 0.9 (Figure 2.7), likely due to this residue's dual H-bonding within the CDR. Changes to the 11th Gln(Q) in α had less impact, while modifications to the 9th Trp(W) in β altered predictions but still maintained positive

outcomes with various amino acid substitutions. These findings, further corroborated by Figure 2.6, suggest the importance of internal H-bonding in CDR3 for peptide binding.

In the structures of 1AO7 (pre-mutation) and 4FTV (post-mutation), a previous study [40] observed that altering four residues in the CDR3β chain of the TCR improved peptide binding by nearly a thousand times. My analysis focused on how these mutations affected predicted y-values, particularly in these two structurally modified examples. As detailed in Table 2.8, 1AO7 and 4FTV share the CDR3α sequence CAVTTDSWG. However, CDR3β differs between them: in 1AO7, it is CASRPGLAGGRP, and in 4FTV, it is CASRPGLMSAQP. The mutation in 4FTV changed residues 8-11 from AGGR to MSAQ, notably enhancing binding affinity. Interestingly, the Cross-TCR-Interpreter model focused on the mutated 10th Ala(A) and 11th Gln(Q) in the β chain, predicting both as positive. The study [40] also noted that the mutation resulted in one fewer hydrogen bond with the peptide, yet overall affinity increased, hinting at indirect factors contributing to binding beyond TCR-peptide H-bonds. This observation supports my assertion that CDR3's internal H-bonded structure is crucial for peptide binding and underscores the biological relevance of attention values in the model.

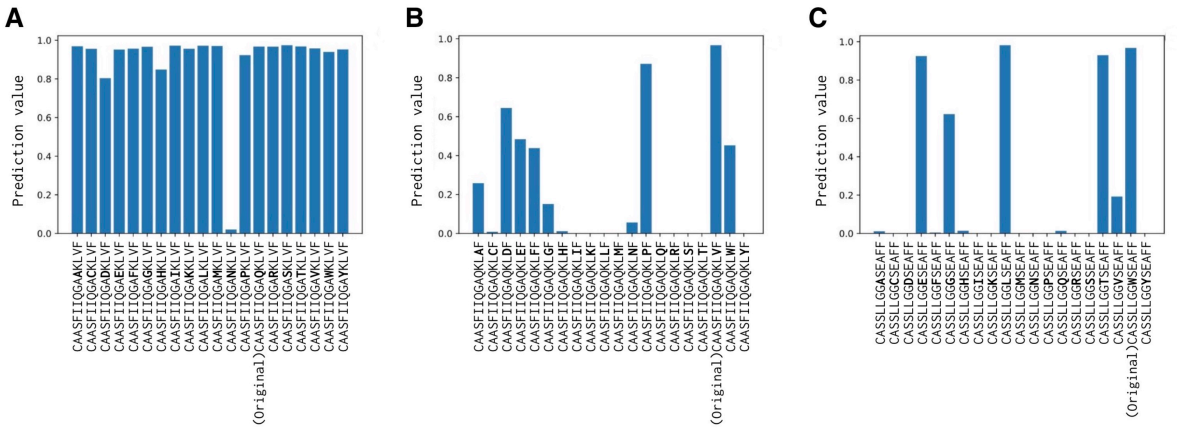

FIGURE 2.7: Input Perturbation Analysis (Reproduced from our published paper [1]). (A) The initial diagram illustrates the prediction shift when substituting the Glutamine(Q) in the CDR3α sequence from CAASFIIQGAQKLVF to CAASFIIQGA(*)KLVF, with (*) indicating the substituted amino acid. (B) The middle diagram illustrates the prediction shift for Valine(V) in CDR3α, changing from CAASFIIQGAQKLVF to CAASFIIQGAQKL(*)F. (C) The last diagram shows the change of Tryptophan(W) of CDR3β, changing from CASSLLGGWSEAFF to CASSLLGG(*)SEAFF. Remarkably, it is the Valine(V) of CDR3α, forming two H-bonds within the same α chain, that exhibits the most substantial impact upon substitution, suggesting the importance of H-bonds within the CDRs.

TABLE 2.8: CDR3 chain analysis for 5TEZ, 1AO7 (Before mutation) and 4FTV (After mutation). The tables are reproduced from our published paper [1]. The mutation in 4FTV changed residues 8-11 from AGGR to MSAQ, notably enhancing binding affinity. Interestingly, the Cross-TCR-Interpreter model focused on the mutated 10th Ala(A) and 11th Gln(Q) in the β chain.

5TEZ	α chain	β chain
AA types	CAASFIIQGAQKLVF	CASSLLGGWSEAFF
Attention Large or Small	SSSSSSSSSS***L***SS***L***S	SSSSSSSS***L***SSSSS
# of Hbonds	212421222253322	22453423486223
# of Hbonds with self CDR3	102120200213120	10220300140020
# of Hbonds with peptide	000000000010000	00000000100000

1AO7	α chain	β chain
AA types	CAVTTDSWG	CASRPGL***AGGR***P
Attention Large or Small	SS***L***SSSSS***L***	SSSSSSSSSSSS
# of Hbonds	122325223	124610210212
# of Hbonds with self CDR3	102013011	002200010012
# of Hbonds with peptide	000000200	000100100000

4FTV	α chain	β chain
AA types	CAVTTDSWG	CASRPGL***MSAQ***P
Attention Large or Small	SS***L***SSSSS***L***	SSSSSSSSS***LL***S
# of Hbonds	222428312	224711101112
# of Hbonds with self CDR3	102114011	002200001012
# of Hbonds with peptide	000000200	000000000000

2.5 Discussion

2.5.1 Prediction interpretability with proteins

Interpreting the outcomes of machine learning models, especially in biological sequence binding predictions, can be challenging. However, for neural network models used in this context, it's vital to understand how the model arrives at its predictions. The ability to pinpoint specific residue positions is crucial for comprehending the model's decisions and is essential for their application in advanced stages. My project, the Cross-TCR-Interpreter, aims to facilitate this interpretative process through the utilization of an attention layer.

Contrary to initial expectations, my analysis revealed that in the machine learning model, CDR residues drawing high attention values did not consistently interact directly with peptide residues. This finding, substantiated statistically in Table 3.3, implies that even with a relatively low frequency of hydrogen bond formation between CDR3s and a peptide, the model can still yield positive predictions.

These findings echo previous research [41] that examined the role of TCR's affinity enhancement towards a peptide. This study posited that an increase in TCR's affinity could potentially induce reactivity and damage specificity, leading to a wider and potentially harmful immune response. The cross-attention model might have more advantages in interaction prediction. Another former study [40] stated the loss of one hydrogen bond with the peptide made the overall affinity stronger.

These results resonate with prior study [41], which explored the impact of heightened TCR affinity for a peptide. This research suggested that enhanced TCR affinity might trigger reactivity, potentially compromising specificity and broadening immune responses, possibly to a detrimental extent. The cross-attention model could offer more effective interaction prediction in this context because the prediction is made by the entangled relationship of TCR-peptide. Additionally, another study [40] observed that a reduction in hydrogen bonding with the peptide actually led to an overall increase in affinity.

Instead of directly interacting with peptides, TCR residues that gained significant attention in the model seemed to reinforce a specific loop structure crucial for peptide binding, forming H-bonds within the CDR3s. Research indicates that residues forming this H-bond network within the TCR are often evolutionarily conserved [42, 43], and the internal organization of the interface plays an important role in protein-protein

interactions [44, 45]. Our findings imply that certain residues may be oriented in a specific direction strategically aligned through internal H-bonds, with the attention layer pinpointing their relevance in TCRs for maintaining binding stability.

This result is intriguing and implies that the extent of connectivity, rather than the surface area of interaction, primarily determines the strength of binding. Protein-protein binding sites are not just clusters of nearby interacting residues; rather, they have a high degree of internal organization. This structure may be essential if the two proteins are to bind to each other in an aqueous environment. Therefore, it appears that the binding results from a sophisticated organization within the binding sites, rather than merely from the matching of surface shapes.

This research also revealed that, on average, the distance between TCR and peptide in their 3D structures decreases when the peptide side receives high attention values. This might be due to the peptide's limited length, making its influence on TCR binding predominantly distance-based. Conversely, higher attention values on the TCR side did not consistently result in reduced distances in 3D, likely owing to the TCR's longer and more complex role in binding.

Although the availability of 3D structural data for all sequence pairs was limited, I conducted experiments with as many existing structures as feasible. These 3D structures were crucial in validating the interpretations derived from the model's attention layer. Looking ahead, future research could involve deploying alternative machine learning models, such as advanced perturbation methods, on a more comprehensively gathered set of sequences.

2.5.2 Performance limitations coming from the dataset

In my investigation of the TCR-pMHC interaction, I simplified my focus to the CDR3 region of the TCR and its peptides. This approach offers computational efficiency and enhanced interpretability by emphasizing the most variable and antigen-specific regions. Also, given the data limitation of the experimental data of the whole sequences, it offers some advantages over the methods required to have the whole sequences. However, this narrowed scope might miss out on integral information from the complete TCR and MHC, potentially leading to overlooked critical interactions vital for binding. For example, my result about no significant difference in H-bonds with peptide on TCR-side attention can possibly mean that there might exist an important bond between the MHC sequence or other TCR residues such as other CDRs, CDR1 or CDR2.

For a more comprehensive view of the entire binding mechanism, methods such as molecular simulations might be more suitable, though they are computationally demanding. The Cross-TCR-Interpreter model, similar to other existing models discussed in the general introduction section, captures a specific aspect of a complex biological process, and its utility must be contextualized based on research objectives and available resources.

Admittedly, my analysis and findings may prompt questions about how one interprets the attention values seen between TCR and peptides. The difference in variety between TCR samples and peptide samples in the dataset might have influenced the notable correlation between high attention and structural features like hydrogen bonds, mainly within the TCRs, not within peptides. This disparity could clarify why my model tends to link certain TCR residues with reactivity to a specific peptide, rather than focusing on attention from the peptide's side.

2.5.3 Prediction difficulties on unseen data of the Covid-19 dataset and the recent dataset

The method used to create negative data for benchmark datasets might result in the test dataset being more similar to the training dataset than it would be in typical real-world, forward-looking evaluation scenarios. This is evidenced by the lower scores in the recent data test set and the results from the Covid-19 dataset. The challenge in accurately predicting in both the recent data test set and the Covid-19 dataset stemmed not from variances in TCR, but from differences in peptides between the Covid-19 data and the combined data dataset. In Figure 2.8, I have illustrated the sequence-sequence pairwise distance matrix using UMAP dimension reduction. The analysis showed no notable variance in the TCR distribution between the combined data dataset and the Covid-19 data. However, a significant difference was observed in the peptide distribution. This observation aligns with findings from prior studies on TCR predictions ([24, 46, 47]), which have highlighted the complexities and challenges associated with generalizing and extrapolating to unseen epitopes.

Furthermore, the positive ratio of the test dataset also influences the metrics. Adjusting the Covid-19 dataset to have a 20% positive ratio resulted in an ROCAUC value of 0.5881 and an average precision score of 0.2305 for the model. Setting the threshold to achieve the maximum F1 score yielded a precision of 0.2892 and a recall of 0.60. This adjustment highlights how the positive ratio can influence the evaluation performance of the model.

Moreover, I aimed to assess the Cross-TCR-Interpreter's performance on the records involving different TCRs or unknown peptides within the test sets of VDJdb and McPAS

data. This involved either eliminating test records that contained peptides of the training dataset or excluding test records with TCRs that were similar to those in the training set.

Despite the training data containing most of the peptides, I found a smaller group within the datasets consisting of 46 McPAS records (14 positives) and 16 VDJdb records (8 positives) featuring new, unseen peptides. The numbers, 46 and 16, are also detailed in Figure 2.3. For these subsets of unseen peptides, the ROCAUC scores were notably lower at 0.721 for McPAS and 0.719 for VDJdb, compared to the higher scores reported in Table 2.3, which have records with familiar peptides. This trend of improved performance with known peptides was similarly noted in the experiments involving the recent data test set as shown in Table 2.4. Figure 2.9 shows a decrease in performance metrics when the test dataset was adjusted to exclude records with TCRs exceeding a certain distance threshold from those in the training set. This pattern highlights the model's sensitivity to variations in TCR diversity and distribution within the test data.

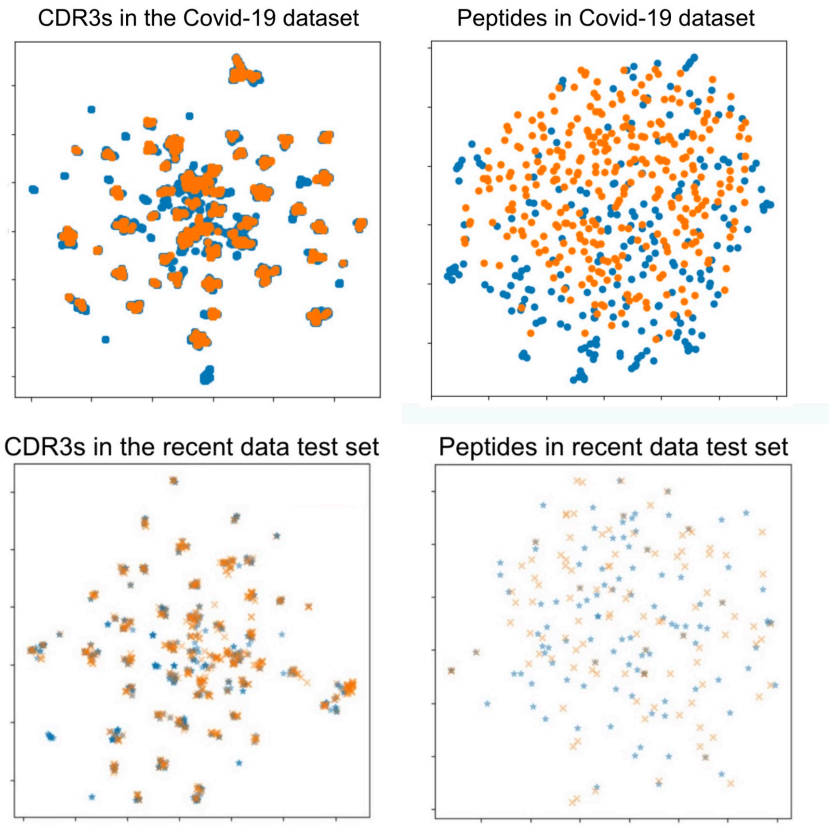

FIGURE 2.8: The upper pair of diagrams depict UMAP visualizations of sequence distance for TCRs and peptides in the Covid-19 dataset, while the lower pair illustrates the same for the recent data test set, as derived from our previously published work [1]. The left diagrams demonstrate TCR sequences (CDR3$\alpha\beta$), and the right diagrams show peptide sequences, with orange indicating either the Covid-19 or recent data test sets, and blue representing the combined data dataset. Left: TCR sequences (CDR3$\alpha\beta$) visualization. Right: Peptide sequences visualization. Orange points: The Covid-19 dataset or the recent data test set. Blue points: The combined data dataset. Each dot represents a sequence, with the two colors in the left side pictures are overlapping, and with the color differentiation in the right diagrams (peptide side) highlighting *minimal* overlap, reflecting the distinct nature of peptides between datasets.

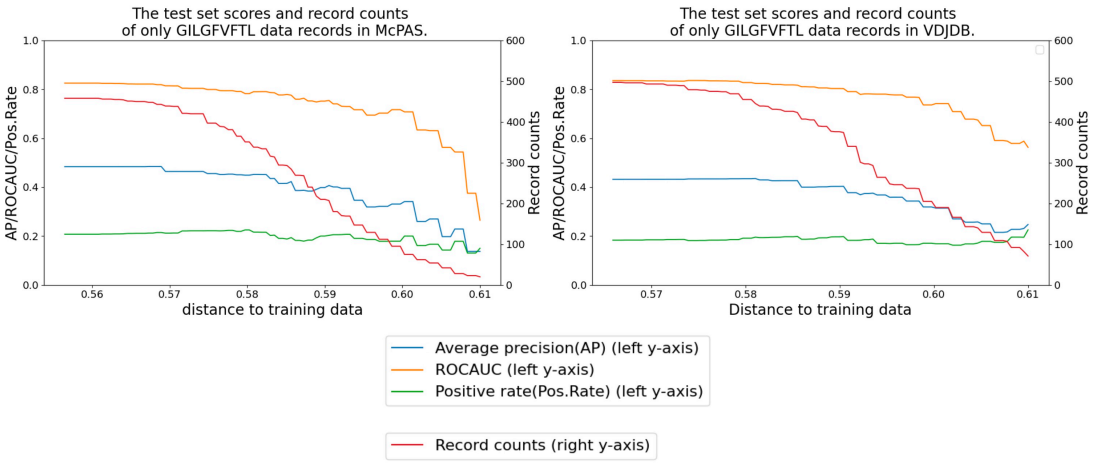

FIGURE 2.9: Performance delegation of the peptide `GILGFVFTL` by removing similar CDR records from the test set (Reproduced from our published paper [1]). The x-axis shows the distance threshold. For instance, if the x-value is 0.58, no test data less than that distance is used for evaluation.

2.6 Conclusion

In this chapter, I have presented the published work [1], illustrating a computational method for predicting the binding of T cell receptors (TCR) and ligand peptides. The study employed a cross-attention mechanism to predict TCR-pMHC interactions and conducted an extensive analysis of available protein structures, thereby providing novel insights into the functional relationships between TCR and peptides.

The machine learning model, integrating an attention layer inspired by language models, demonstrated strong performance on a benchmark dataset for TCR-pMHC interaction, despite facing persistent challenges with the Covid-19 dataset and a recent data test set.

The model's analysis enabled me to link neural network weights to protein 3D structure datasets, revealing significant characteristics of residues that received large attention values, and enabled me to elucidate the binding principles. This was achieved by visualizing and analyzing the cross-attention, source-target-attention layers.

Statistical tests of the attention layer in relation to structural data indicated that residues receiving high attention tended to interact more with their own CDR3 compared to other residues. This sheds light on the mechanisms of CDR3-peptide binding. Proteins form hydrogen bonds, leading to unique structural configurations, which are crucial, particularly in interactions with peptides conditioned to respond to them.

Chapter 3

Comparison between standard-attention model and cross-attention model

This chapter advances our paper "Attention network for predicting T cell receptor-peptide binding can associate attention with interpretable protein structural properties" [1], for which I am the first author. I advance it in terms of model performance and amount of structural data. In addition, this chapter strongly supports the usefulness of the cross-attention model by comparing it with a model with a standard self-attention.

3.1 Abstract

This chapter presents a comprehensive comparison between cross-attention and standard-attention mechanisms in the context of T cell receptor (TCR) and peptide-major histocompatibility complex (pMHC) interaction prediction. Building on the foundational work introduced in Chapter 2, this analysis delves deeper into the computational models' capabilities, focusing on their performance and interpretability in predicting TCR-pMHC interactions.

The core objective of this chapter is to experimentally validate the hypothesis that a cross-attention model, as opposed to a traditional standard-attention model, offers enhanced accuracy and interpretability in modeling the complex interactions between TCRs and peptides. To this end, I have meticulously developed and compared two distinct models: one employing the standard self-attention layers, and the other utilizing the cross-attention framework identical to Chapter 2 and our study[1].

Further advancements from our previous work [1] are incorporated, including adjustments to the models' hyperparameters and the inclusion of additional Protein Data Bank (PDB) structures for a more inclusive analysis of attention weights. The comparative study aims to demonstrate the superiority of the cross-attention model in capturing the nuanced dynamics of TCR-peptide interactions, particularly in terms of predictive performance and the ability to yield biologically relevant insights.

This chapter not only highlights the cross-attention model's effectiveness in dealing with TCR-pMHC interactions but also underscores its potential in advancing computational approaches within the field of bioinformatics, setting a precedent for future research in TCR-pMHC interaction prediction and analysis.

3.2 Introduction

Chapter 3 of this dissertation builds upon the foundational research presented in Chapter 2, where I introduced a novel computational model for predicting T cell receptor (TCR) and peptide-major histocompatibility complex (pMHC) interactions. The focus of this chapter is to conduct a detailed and rigorous comparison between two pivotal mechanisms within the realm of ML models: cross-attention and self-attention. This comparison is crucial for advancing our understanding of the most effective computational approaches in modeling TCR-pMHC interactions.

In the previous chapter, this dissertation showed that a model based on the cross-attention mechanism exhibited superior performance in both accuracy and interpretability compared to existing computational models. This chapter seeks to experimentally validate this hypothesis compared to the standard-attention model. I delve into both approaches, exploring how each model performs on benchmark datasets and interprets complex protein sequence data, and how this impacts their overall effectiveness in predicting TCR-pMHC interactions.

To achieve a comprehensive comparison, this chapter introduces an improved version of the cross-attention model, namely "the improved model", incorporating refined hyperparameters for enhanced performance. Additionally, a new model employing the standard Transformer encoder with self-attention layers has been developed, namely the "standard-attention model". This model should serve as a benchmark to underscore the distinctions and advantages of the cross-attention approach. Further, I have expanded the dataset used for analysis by including a broader range of PDB structures, enabling a more thorough examination of the attention weights and their correlation with protein structural properties.

The primary objective of this chapter is to provide empirical evidence supporting the superiority of the cross-attention mechanism over the standard-attention approach in the context of TCR-pMHC interaction prediction. Through this comparative analysis, I aim to highlight the strengths and limitations of each model and to showcase the cross-attention model's unique ability to yield deeper biological insights into TCR function and peptide binding mechanisms.

This chapter's findings are expected to contribute significantly to the field of bioinformatics, particularly in the study of TCR-pMHC interactions, and to pave the way for the development of more accurate and interpretable computational tools in biomedical research.

3.3 Method

3.3.1 The attention models

The Transformer model introduced in the paper "Attention Is All You Need" [15] represents a significant shift in neural network architecture, particularly for sequence-to-sequence tasks. Central to its architecture is the attention mechanism, specifically the self-attention layer, which allows the model to weigh the importance of different parts of the input sequence when processing each element.

The attention layer works by utilizing three matrices for each amino acid in the sequence: a query matrix (Q), a key matrix (K), and a value matrix (V), as in Equation 3.1. These matrices are derived from the amino acid embeddings, which encapsulate the properties of each amino acid in the sequence. In Equation 3.1 for the attention layer, Q, K, and V are the data matrices of sequences, and d is the scaling factor. When $Q = K = V$, this is a self-attention layer, whereas it is a cross-attention layer when $K = V$ and $Q \neq K$. K^T denotes transposed matrix of K, where the sizes of arrays are $Q : L_1 \times D$, $K : L_2 \times D$, and $V : L_2 \times D$, and D is the embedding dimension. The self-attention mechanism computes attention scores by taking the dot product of the query vector with all the key vectors, which are then scaled, normalized, and passed through a $Softmax$ function to yield weights. The output of the attention layer is a weighted sum of the value vectors, which means the value vectors are weighted on each position as an importance factor.

$$AttentionLayer(Q, K, V) = Softmax(QK^T/d)V \tag{3.1}$$

In the context of my study on TCR-pMHC interactions, the attention mechanism in the Transformer model is utilized to analyze the amino acid sequences of the TCR's complementarity-determining region (CDR) 3 and the peptide. In the cross-attention layer, $K(= V)$ and Q represent two different inputs, i.e., a connected sequence of CDR3α:CDR3β and a peptide, respectively as in Equation 3.2 and Equation 3.3. The $Softmax$ function defines the weights to V when matrix Q is query input, and the weights are allocated so that the sum is 1 over the length direction of V.

This $Softmax(QK^T/d)$ is the attention weights ($=Attention_{TCR-cross}$) and is used for the analysis and visualization in this study, suggesting the residue positions that are important within the position of the sequence V. These attention weights are the TCR-side attention given the query input of peptide. The cross-attention layer uses peptides as inputs and assigns specific weights to each residue of CDR3s to learn the important sites of CDR3s. This enabled me to analyze each side of the two areas of attention

separately. The same operation was done for the peptide side and $H_{Peptide}$ can be obtained as in Equation 3.4 and Equation 3.5. In addition, through the hyperparameter tuning, I defined four heads for each side in the cross-attention layer, and those heads were concatenated similarly in a typical Transformer. Visualization and analysis of the attention layer allow interpretation of the residue interaction across sequences using the output of Equation 3.3, Equation 3.5.

$$H_{TCR} = Attention_{TCR-cross} V_{TCR} \tag{3.2}$$

$$Attention_{TCR-cross} = Softmax(Q_{Peptide} K_{TCR}^{T} / d) \tag{3.3}$$

$$H_{Peptide} = Attention_{Peptide-cross} V_{Peptide} \tag{3.4}$$

$$Attention_{Peptide-cross} = Softmax(Q_{TCR} K_{Peptide}^{T} / d) \tag{3.5}$$

$$y = MLP(Concat(Mean(H_{Peptide}), Mean(H_{TCR})))$$

where the "Mean" layer takes the mean over the length dimension.

On the other hand, my ablation analysis of the standard self-attention model can be expressed by the following equation with the "TCR&Peptide" output from concatenation layer in Figure 3.1, Equation 3.6 and Equation 3.7, where the shape of $Attention_{TCR\&Peptide}$ is the square, the sum of lengths of TCR and Peptide times the same length. Each of Q, K, V having "TCR&Peptide" has the shape of $(L1 + L2) \times D$, where $L1$ is the length of CDRs and $L2$ is that of peptide.

$$H_{TCR\&Peptide} = Attention_{TCR\&Peptide} V_{TCR\&Peptide} \tag{3.6}$$

$$Attention_{TCR\&Peptide} = Softmax(Q_{TCR\&Peptide} K_{TCR\&Peptide}^{T} / d) \tag{3.7}$$

$$y = MLP(Mean(H_{TCR\&Peptide}))$$

where the "Mean" layer takes the mean over the length dimension.

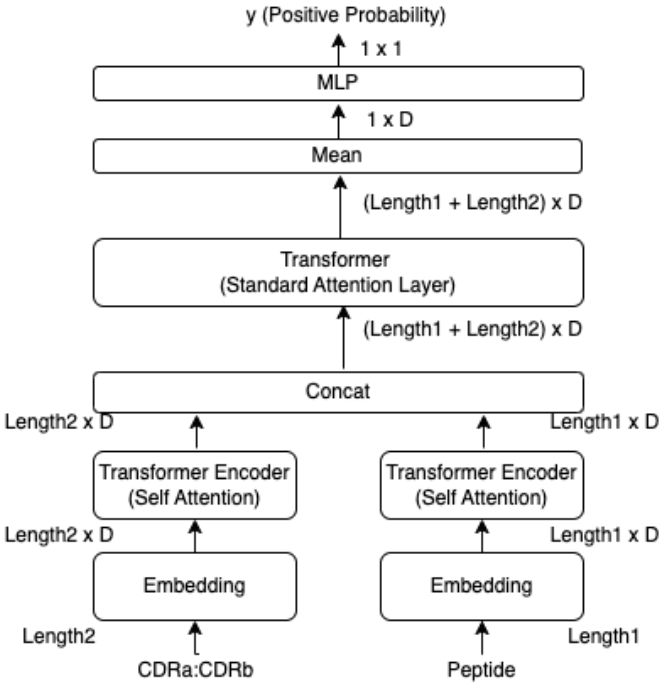

FIGURE 3.1: Overview of the standard-attention model. Data tensor sizes are denoted. The standard-attention layers in the middle of the figure were analyzed using structural data after the training. The difference between Figure 2.1 and Figure 3.1 lies in the standard-attention layer and two cross-attention layers in the middle of the figures.

Obviously, this square-shaped attention layer is dependent on both sequences, meaning the attended TCR residue might be the result of TCR itself. Therefore, using the cross-attention model is helpful in identifying the relationship. One should be able to weight the target sequence given one side of amino acid sequences, query sequence. This cross-attention effect is not possible by the standard-attention because Equation 3.6 and Equation 3.7 did not explicitly limit the attention on the one side of the sequences.

3.3.2 Sequence data to train the model and structural data to analyze the attention weights

I have used the identical sequence dataset to train and evaluate the models, same as the previous research [1] and as Chapter 2. Hence the combined dataset was used as the training to evaluate the recent dataset, whereas the benchmark datasets of McPAS and VDJdb have respective training and test sets. The key statistics of the sequence datasets are the same as in Table 2.2.

Just like in Chapter 2, the TCR-peptide complex structures were used as subjects of the t-test. But this time, I manually searched TCR-peptide complex structures with CDR3s from the Protein Data Bank (PDB, [34]). The SCEptRe server [4] was also used, yielding 82 unique CDR3α-CDR3β-peptide sequence pairs as a total. Please note that identical sequence pairs produce the same outputs and attention values, therefore I used only one structure for identical sequences.

The target values of the t-test were properties such as the proportion of TCR residues forming H-bonds with the peptide were the same as in Chapter 2. BioPython [36], LIGPLOT [37], and Anarci [35] were used again.

Eventually, I performed a paired student's t-test, in more amount of complex structures than Chapter 2, to assess the differences between the two residue groups of attended and not-attended.

3.3.3 Hyperparameters

By following Chapter 2, for both the standard-attention and the improved model, I used the same hyperparameters; d_ff of the final MLP layer = 84, dim in the model = 256, $dropout\_rate$ = 7.651e-05, and $learning\_rate$ = 9.387e-05.

There is one difference in the positive weight on the binary cross entropy; the positive weight this time was 6.0 for both the standard-attention model and the improved model.

3.4 Result

3.4.1 Scores on benchmark datasets and recent data dataset

In addition to the previous paper, two new models, the improved model and the standard-attention model were introduced. I compared the improved model with a normal Transformer-encoder, namely the "standard-attention" model. It used the standard self-attention layer, instead of the cross-attention layers, on concatenated hidden representations after the Transformer encoders of CDR3s and peptides.

As for the benchmark datasets of McPAS and VDJdb as shown in Table 3.1, the standard-attention model shows the highest ROCAUC values (0.9206 for McPAS, and 0.9513 for VDJdb-without-10x). However, when it comes to the average precision score (APS), the improved model outperforms the others in both datasets (0.6349 for McPAS and 0.7760 for VDJdb-without-10x), suggesting a better precision-recall balance. The Cross-TCR-Interpreter model shows competitive, but slightly lower performance compared to the improved model and the standard-attention model in both metrics across the datasets.

As for the recent data test set, the improved model has the highest ROCAUC score (0.5519), whereas the other models have similar but slightly lower ROCAUC scores, with the Cross-TCR-Interpreter and the standard-attention model having almost identical scores. Similarly, the improved model leads in the APS (0.2052), suggesting it has a better precision-recall balance than the others. The APS scores of the other models are closely matched, with PanPep showing a slightly better score than the standard-attention model and Cross-TCR-Interpreter.

Provided that the positive ratio is 16% across both datasets, the APS is better to use to compare and hence the improved model could be the best model among the models. Overall, the improved model seems to outperform the others in both metrics on the recent data test set. However, the overall scores for all models are relatively low, which indicates the challenging nature of the dataset or the task.

3.4.2 Attention analysis

I sought to interpret the model within the 82 complex structures where the models surely perform well enough to analyze. The PDB identifiers used are provided in Table 3.4, which is more than Chapter 2.

TABLE 3.1: Result of benchmark dataset. APS stands for the average precision score.

Dataset	Model	CDR's Features	ROCAUC	APS
McPAS	The improved model	CDR3s of α and β chains	0.9187	0.6349
	Cross-TCR-Interpreter [1]	CDR3s of α and β chains	0.9154	0.6211
	Standard-attention model	CDR3s of α and β chains	0.9206	0.6123
	NetTCR2.0	CDR3s of α and β chains	0.9204	0.5808
	PanPep	CDR3 Sequence of β chain with biochemical features	0.8374	0.4519
VDJdb-without-10x	The improved model	CDR3s of α and β chains	0.9459	0.7760
	Cross-TCR-Interpreter [1]	CDR3s of α and β chains	0.9445	0.7600
	Standard-attention model	CDR3s of α and β chains	0.9513	0.769
	NetTCR2.0	CDR3s of α and β chains	0.9492	0.7262
	PanPep	CDR3 Sequence of β chain with biochemical features	0.9009	0.6435

TABLE 3.2: Result of the recent data test dataset to compare the cross-attention and the standard-attention. APS stands for the average precision score.

Model	Dataset	ROCAUC	APS
The improved model	Recent data test set	0.5519	0.2052
Cross-TCR-Interpreter [1]	Recent data test set	0.5362	0.1855
Standard-attention model	Recent data test set	0.5357	0.1889
PanPep	Recent data test set	0.5337	0.1897

Whereas the models did not fully generalize on the recent data test, I interpreted the attention layers by using reliable predictions as defined in the following manner. Of the 82 complex structures, the cross-attention model identified 52 with positive interactions, while the standard-attention model did 54, using a 0.5 threshold. I paid special attention to these cases in the analysis of attention layers, on the premise that the model's accurate interpretability could be safely assumed for these instances. This is similar to a regression analysis examining the effect of some explanatory variables on target variables, and the goal was to identify the important features that the model learns, i.e., the features of the attended amino acid residues.

Attention values exceeding MEAN + 5.5 STD for peptides and MEAN + 4.5 STD for TCRs were deemed "high attention" in the improved model, while the standard-attention model used the threshold of 4.5 for both (γs in Equation 2.1). Just like in Chapter 2, these thresholds with the cumulative outputs of residue index from the four heads classified approximately 20% of residues as attended on each side of TCR and peptide, being able to effectively distinguish between large and small attention values.

Note that the threshold for large attention values is different for each PDB entry or for head due to differences in the distribution of attention values, just like in the same way in Chapter 2.

The analysis was performed for the concatenated attended residues from each of the four heads in the attention layer (heads 0 to 3) on both the TCR and the peptide sides. The standard-attention layer was defined on an embedded concatenated CDR3$\alpha\beta$ sequence and a peptide sequence, resulting in an attention matrix with a square shape determined by the length of the peptide plus CDR3$\alpha\beta$ residues. When analyzing the standard-attention model, I limit the attention matrix only to the TCR part or the peptide part. On the other hand, cross-attention has an attention matrix with a rectangular shape determined by the length of the peptide and CDR3$\alpha\beta$ residues. It was possible for a particular residue to have a large attention value in a specific head but not in the other heads (as seen in Equation 2.1).

3.4.3 Statistical analysis shows the same conclusion as Chapter 2

Using the γ factor, I classified the TCR residues into two groups and analyzed their structural properties. To assess the differences between the two groups, I performed a paired t-test again just in the same way as Chapter 2.

The results of the statistical tests of all concatenated heads are shown in Table 3.3 for both models of cross-attention model and standard-attention model, and for TCR-side attention and peptide-side attention. The individual results for each head are also reported in the Table A.5 and Table A.6 of the Appendix. As a TCR sequence for the structural analysis includes both CDR3 and non-CDR3 portions, the H-bond properties were measured by dividing the residues into CDR3 and non-CDR3 portions.

As the TCR side analysis result, Table 3.3 shows a statistically significant observation; for both models, the attended CDR3 residues often form H-bonds with general CDR3 or their own chain's CDR3. Unexpectedly, the CDR3 residues forming H-bonds with peptides were not highlighted by attention values in both models (p-values with 0.6278 and 0.1161), though both models generally indicated a higher rate of binding to peptides on average. Only in the cross-attention model, the attended CDR3 residues are significantly closer to the peptides (p=0.0155 and p-adj.=0.0850), aligning with the idea that proximity to peptides and H-bond formation are correlated. Just like in Chapter 2, the adjusted p-values are also provided to avoid the misunderstanding associated with multiple p-values. An FDR of 0.1 led to the rejection of two hypotheses in the cross-attention model TCR-side: "Proportion of residue in edge" and "Distance to the nearest opposite peptide". In the standard-attention model TCR-side, an FDR of 0.1 led to the

TABLE 3.3: Attention analysis to compare the standard-attention with this study. Comparative analysis of structural properties between the residue groups receiving large and small attention is presented. The numbers in the Large Attention or Small Attention columns are the average and standard deviation. The p-adjusted column shows the adjusted p-value by the Benjamini Hochberg (BH) procedure. The symbol "***" denotes the significant difference based on a False Discovery Rate (FDR) of 0.05 in the BH procedure; the symbol "**" indicates significance with the FDR of 0.10; the symbol "*" indicates the FDR of 0.15.

Model (Side)	Property	Large Atten.[1]	Small Atten.[1]	p-value	p-adj.	
Improved Model Cross-attn. (TCR Side)	Proportion connecting to peptide	0.0822+-0.1359	0.0732+-0.0478	0.6278	0.6905	
	Proportion connecting to CDR	0.5221+-0.2830	0.4271+-0.1030	0.0310	0.1123	*
	Proportion connecting to their own chain TCR	0.7045+-0.2214	0.6540+-0.0981	0.1634	0.2567	
	Proportion connecting to their own chain CDR	0.4926+-0.2884	0.4016+-0.0988	0.0408	0.1123	*
	Proportion connecting to opposite chain TCR	0.1303+-0.1615	0.1604+-0.0794	0.2707	0.3722	
	Proportion connecting to opposite chain CDR	0.0564+-0.1111	0.0550+-0.0557	0.9225	0.9225	
	Proportion connecting to TCR	0.7716+-0.2323	0.7354+-0.0988	0.3175	0.3881	
	Proportion connecting to not-CDR TCR	0.3771+-0.2599	0.4466+-0.0908	0.1041	0.1908	
	Proportion of residue in edge [2]	0.5256+-0.2145	0.6226+-0.0523	0.0077	0.0850	**
	Distance to the nearest opposite (Peptide) [3]	7.5127+-2.7013	8.4488+-0.9691	0.0155	0.0850	**
	Number of bonds [3]	2.3375+-1.1661	2.0965+-0.7033	0.1026	0.1908	
Improved Model Cross-attn. (Peptide Side)	Proportion connecting to peptide	0.0952+-0.2524	0.0707+-0.1311	0.5394	0.6293	
	Proportion connecting to CDR	0.3075+-0.4044	0.1593+-0.1157	0.0321	0.0562	**
	Proportion connecting to TCR	0.4325+-0.4270	0.2361+-0.1426	0.0113	0.0272	***
	Proportion connecting to not-CDR TCR	0.2857+-0.4051	0.1111+-0.1243	0.0116	0.0272	***
	Proportion of residue in edge[2]	0.5417+-0.4452	0.5611+-0.1171	0.7986	0.7986	
	Distance to the nearest opposite (TCR) [3]	4.2876+-1.8053	5.0500+-1.0295	0.0064	0.0272	***
	Number of bonds [3]	2.5873+-2.2743	2.0902+-0.7780	0.1104	0.1545	
Standard-attn. (TCR Side)	Proportion connecting to peptide	0.0820+-0.1129	0.0534+-0.0771	0.1161	0.2155	
	Proportion connecting to CDR	0.5361+-0.1713	0.4389+-0.1865	0.0046	0.0255	***
	Proportion connecting to their own chain TCR	0.6881+-0.1537	0.6860+-0.1807	0.9498	0.9498	
	Proportion connecting to their own chain CDR	0.5046+-0.1750	0.4239+-0.1838	0.0164	0.0601	**
	Proportion connecting to opposite chain TCR	0.1553+-0.1191	0.1083+-0.1257	0.0565	0.1554	
	Proportion connecting to opposite chain CDR	0.0602+-0.0694	0.0262+-0.0594	0.0046	0.0255	***
	Proportion connecting to TCR	0.7706+-0.1526	0.7472+-0.1914	0.4781	0.5843	
	Proportion connecting to not-CDR TCR	0.4023+-0.1514	0.4392+-0.2231	0.3784	0.5202	
	Proportion of residue in edge [2]	0.6750+-0.1624	0.6085+-0.1854	0.1176	0.2155	
	Distance to the nearest opposite (Peptide) [3]	8.2976+-1.7990	8.3673+-1.8767	0.8350	0.9185	
	Number of bonds [3]	2.1823+-0.8115	2.0340+-0.8724	0.2039	0.3205	
Standard-attn. (Peptide Side)	Proportion connecting to peptide	0.0796+-0.1484	0.0824+-0.1718	0.9130	0.9130	
	Proportion connecting to CDR	0.1679+-0.1386	0.2923+-0.2964	0.0170	0.0397	***
	Proportion connecting to TCR	0.2372+-0.1370	0.3626+-0.2829	0.0105	0.0367	***
	Proportion connecting to not-CDR TCR	0.1075+-0.0979	0.1477+-0.2369	0.2837	0.3971	
	Proportion of residue in edge[2]	0.5899+-0.1469	0.4818+-0.2767	0.0464	0.0813	**
	Distance to the nearest opposite (TCR) [3]	5.1301+-1.1127	4.4726+-1.3541	0.0009	0.0062	***
	Number of bonds [3]	2.1798+-0.9237	2.1143+-1.2355	0.7275	0.8487	

1. Mean and standard deviation (for the PDB structures) of the proportion of residues that satisfy the property shown in the first column.
2. The edge means four residues from the beginning or end of the CDR, whereas three residues for the peptide.
3. In the last two properties, per-residue averages were used instead.

rejection of three hypotheses: "Proportion connecting to CDR", "Proportion connecting to opposite chain CDR", and "Proportion connecting to their own chain CDR".

In more detail, for the cross-attention model of TCR-side analysis, the most significant property difference in all concatenated heads occurred in the "Proportion of residue in edge" of TCR-side, meaning there exist significantly *smaller* amount of residues located in the edge positions when the neural nets paid attention in the CDR sequences (p=0.0077 and p-adj.=0.0850). The edges of CDRs are often far away from the peptide because the CDRs form a bow-like shape, so the attended CDRs are not at the edges and are located in the center on the contrary. This aligns with the significantly smaller distance of attended residues toward the peptides (p=0.0155 and p-adj.=0.0850) because the center of CDRs should be located toward the peptide residues. Interestingly, while the number of H-bonds with peptides isn't significant, it's noteworthy that the attended

Table 3.4: PDB identifiers of structural data used in the attention analysis for the standard-attention model and the cross-attention model

Collected 82 IDs with unique sequences pair	1D9K, 1FYT, 1G6R, 1ZGL, 2AK4, 2BNQ, 2CKB, 2E7L, 2F54 2J8U, 2NX5, 2OI9, 2P5E, 2PXY, 2VLJ, 2VLR, 3DXA, 3E2H 3E3Q, 3GSN, 3HG1, 3MBE, 3MV8, 3PQY, 3QDG, 3QDJ, 3QDM 3QEQ, 3QIB, 3QIU, 3QIW, 3TF7, 3TFK, 3TJH, 3TPU, 3VXR 3VXS, 3W0W, 4JFD, 4JFE, 4JFF, 4JRX, 4JRY, 4L3E, 4MJI 4OZG, 4OZH, 4P2O, 4P2Q, 4P2R, 4PRH, 4PRP, 4QOK, 4Z7W 5BRZ, 5BS0, 5D2L, 5E6I, 5EU6, 5EUO, 5HHO, 5ISZ, 5IVX, 5WKF 5KS9, 5KSA, 5MEN, 5NME, 5NMF, 5NMG, 5SWS, 5SWZ, 5TEZ 5WKH, 5WLG, 6AVF, 6AVG, 6EQA, 6EQB, 6PX6, 6V13, 6V15
52 IDs used in the analysis of the cross-attention model	1D9K, 1G6R, 1ZGL, 2AK4, 2BNQ, 2CKB, 2F54, 2J8U, 2NX5 2P5E, 2PXY, 2VLJ, 2VLR, 3DXA, 3GSN, 3HG1, 3MBE, 3MV8 3PQY, 3QDJ, 3QEQ, 3QIU, 3QIW, 3VXR, 3VXS, 3W0W, 4JFD 4JFE, 4JFF, 4JRX, 4JRY, 4MJI, 4OZG, 4P2O, 4P2Q, 4P2R 4PRH, 4PRP, 4QOK, 4Z7W, 5E6I, 5EUO, 5HHO, 5ISZ, 5MEN 5NMG, 5TEZ, 5WKF, 5WKH, 6EQA, 6EQB, 6V13
54 IDs used in the analysis of the standard-attention model	1D9K, 1FYT, 1G6R, 1ZGL, 2AK4, 2BNQ, 2CKB, 2F54, 2NX5 2P5E, 2PXY, 2VLJ, 2VLR, 3DXA, 3HG1, 3MBE, 3MV8, 3PQY 3QDG, 3QDJ, 3QEQ, 3QIU, 3QIW, 3VXR, 3VXS, 3W0W, 4JFD 4JFE, 4JFF, 4JRX, 4JRY, 4MJI, 4OZG, 4OZH, 4P2O, 4P2Q 4P2R, 4PRH, 4PRP, 4QOK, 4Z7W, 5D2L, 5E6I, 5EU6, 5EUO 5HHO, 5IVX, 5MEN, 5NMG, 5TEZ, 5WKF, 5WKH, 6EQA, 6EQB

CDRs are positioned closer to potential H-bond sites but do not necessarily form these bonds. The residues with large attention values had a more significant proportion of having an H-bond with the CDR3 portions (p=0.0310) or with their own chain's CDR3 (p=0.0408). Nonetheless, the proportion of residues that are H-bonded to any TCR residue (i.e., H-bonds within the TCR chains) showed larger values but no difference between the large and small attention groups (p=0.3175).

For the standard-attention model of TCR-side analysis, the most significant property difference in all concatenated heads occurred in the "Proportion connecting to CDR"(p=0.0046) and "Proportion connecting to opposite chain CDR"(p=0.0046), meaning the neural nets paid attention to the residues connecting to the CDR sequences. Furthermore, highly attended CDR3s are not significantly proximate to the peptide (with a p-value of 0.8350), which is counter-intuitive as one would naturally expect attended CDR3s to be close to the peptide.

Collectively, these statistical evaluations lend support to the hypothesis that attended residues significantly favor H-bonds within the CDR3 regions in the same way as Chapter 2.

For the peptide side analysis, peptide residues with high attention by the cross-attention model were significantly closer to the TCRs and more likely to form H-bonds, offering a clear insight into the model's behavior, and this observation matches the previous CDR observation that attended CDR residues were significantly closer to the peptide.

Also, the peptide residues forming H-bonds with TCR are significantly attended by the cross-attention (with a p-value of 0.0321). Contrarily, the standard-attention model's peptide side attention was less informative. The peptide residues with high attention showed a significantly *smaller* binding proportion to the CDR3 and TCR. Furthermore, their distance to the TCR was significantly greater (p=0.0009).

This result suggests the intuitively accurate consideration of how attention works in the cross-attention model, despite highlighting the limitation of the standard-attention model in which attention values do not showcase the intuitive result. The distinction in attention between the two models might stem from the conditional relationships in their attention layers. The distance to the nearest peptide from the attended CDR was naturally small in the cross-attention, where that behavior was *not* mirrored in the standard-attention model. The distance to the nearest CDRs from the attended peptide is also smaller only in the cross-attention, suggesting that the cross-attention model might better capture the natural interactions. Additionally, if the result implies that peptides need to maintain a distance from CDR3s like in the standard-attention model for effective TCR recognition, it would contradict the known function of MHCs in presenting peptides. Therefore, such a scenario seems unlikely, reinforcing the validity of the observed cross-attention behavior. However, regarding the attended CDRs and the residue property 'proportion connecting to CDR,' the standard-attention model demonstrates a smaller p-value, indicating a more interpretable difference compared to the cross-attention model, and this could still potentially imply that the standard-attention model effectively elucidates the interaction.

Therefore, the observations warrant further investigation in future studies. At this stage, given the overall performance and interpretability, I would like to state that cross-attention provides a more comprehensive understanding of the interactions.

3.5 Discussion

3.5.1 Design framework for interactions and interpretations

The cross-attention layer is pivotal for analyzing the interactions between two sequences, particularly for elucidating changes in one sequence due to its interaction with another, while focusing exclusively on their mutual interaction. Predicting protein-protein interactions based on sequence data alone is a complex task, and often the case models akin to BERT process only a single sequence [27]. Even when such models are fed two separate input sequences, they fall short of capturing the essence of binding as a "relational" phenomenon between the two data sets, and merely concatenating these sequences does not adequately convey the nature of their binding, as illustrated in Equation 3.7.

Although the BERT model processes two sequences, its focus is not exclusively on their mutual interaction. The visualization of the attention layer for one head in a single layer, as shown in Figure 3.2, demonstrates how BERT's attention processes TCRs and peptides when trained on these sequences. Based on Equation 3.1, the sum along the x-axis equals one, indicating that BERT's attention encompasses both self-attention and cross-attention. This complexity makes it challenging to link the attention matrix directly to the binding mechanism. Essentially, the model's self-attention part could produce a positive prediction based solely on the self-coincidence within a sequence, neglecting one side of the information. In other words, it is possible to obtain a positive prediction focused merely on a specific part of the TCR, ignoring its crucial relationship with the peptide. Consequently, to accurately capture the interaction between TCR and peptide, the focus should be selectively placed on the cross-attention aspect for more meaningful interpretation.

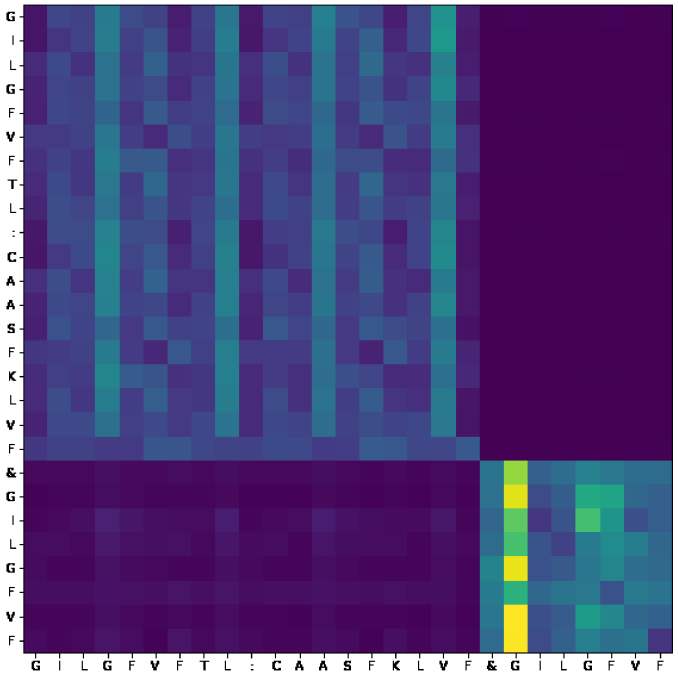

FIGURE 3.2: The attention matrix visualization of standard-attention. The colon symbol is used as a token to concatenate α and β chains, while the "&" symbol is another special separation token for the peptide. This setup doesn't restrict attention values strictly to mutual interactions, complicating their interpretation. Moreover, in a typical robust BERT architecture, the presence of numerous attention heads and multiple stacked Transformer layers complicates the task of correlating attention values directly with the binding mechanism.

3.6 Conclusion

In Chapter 3, I studied an in-depth comparative analysis between cross-attention and standard-attention mechanisms within the framework of predicting TCR-peptide-major histocompatibility complex (TCR-pMHC) interactions. This comparative study was meticulously designed to test the hypothesis that the cross-attention model would not only enhance predictive performance but also offer greater interpretability in understanding TCR-pMHC interactions.

The findings from this chapter have demonstrated the advantages of the cross-attention mechanism in the improved model over traditional standard-attention models. The TCR-side attention statistical trend did not change from the Chapter 2 study, showing H-Bonds within CDRs are attended by the attention layers in both cross-attention and standard-attention models. However, in the peptide-side attention statistical analysis, the advantage of cross-attention over the standard-attention was shown. The experiments revealed that the cross-attention model, with its nuanced approach to analyzing sequences and conditional relationships of sequences, provides a more accurate representation of the intricate dynamics of TCR-pMHC interactions. Like in Chapter 2, the cross-attention model showed a remarkable ability to identify and elucidate key interactions at the residue level, offering insights that were previously unattainable with the standard-attention approaches.

In summary, this chapter has successfully validated the hypothesis that the cross-attention model is superior for predicting TCR-pMHC interactions, both in terms of performance and interpretability. These results underscore the potential of advanced computational methods in enhancing our understanding of biological systems and pave the way for more effective and interpretable models in the study of immunological interactions.

Chapter 4

General Conclusion

This dissertation embarks on a detailed exploration into the complexities of T cell receptor (TCR) interactions with ligand peptides, pivotal for our immune system. It traverses from presenting our published work to executing a comparative analysis between other computational models. The approach leads to profound insights into the interactions between TCRs and peptide-major histocompatibility complex (pMHC), which are essential in immune responses and designing therapeutic methods.

In Chapter 2, I delved into the TCR-pMHC interactions, employing a cross-attention mechanism to predict these interactions accurately, with the best average precision score achieved among other models on the benchmark datasets. Despite facing challenges with specific datasets, like those from the recent dataset and the Covid-19 data, I demonstrated the data distribution and the reason behind the prediction difficulties. Moreover, the proposed model demonstrated a remarkable ability to associate neural network weights with protein 3D structures, uncovering highly attended residues' significant properties such as hydrogen bonds within CDR3. These findings are not just mere data points but are pivotal because I have shown this over the gathered PDB structures. The analysis of available protein structures offered new perspectives on TCR-peptide functional relationships between neural net weights and understanding of molecular immunology.

Chapter 3 furthered this exploration by comparing the efficacy of cross-attention and standard-attention mechanisms in the realm of TCR-pMHC prediction. This rigorous comparative study validated our hypothesis regarding the superiority of the cross-attention model in the benchmark performance. The distinction of interpretability was particularly evident in the peptide-side attention analysis and the distance analysis, which demonstrated the cross-attention model's unique ability to show key interactions at the molecular level.

Collectively, this dissertation has not only demonstrated the predictive power of Transformer-based models in bioinformatics but has also illuminated the pathway toward more interpretable and effective computational tools in immunological studies. The cross-attention mechanism, a central theme in both Chapter 2 and Chapter 3, has emerged as a valid approach to uncovering the mechanism of TCR-pMHC interactions. By unraveling the intricacies of CDR3-peptide binding mechanisms and highlighting the role of specific structural formations like hydrogen bonds, this work paves the way for vaccine development, cancer immunotherapy, and autoimmune disorder understanding.

In conclusion, this dissertation stands as proof of the power of interpretable machine learning methods in understanding complex biological interactions. This work not only significantly enhances the field of bioinformatics but also has the potential to further various areas where complex mechanisms are yet to be fully understood and can be explored through machine learning techniques. My aspiration is that this dissertation helps medicine through deeper understanding and innovative approaches in the TCR-pMHC interaction prediction.

Appendix A

An Appendix

Occurrence comparison of types of amino acid residues in the large-valued and small-valued attention groups

Typically, CDR3 sequences predominantly consist of serine (S), alanine (A), and phenylalanine (F). However, those with high attention values frequently exhibit valine (V), glycine (G), and glutamic acid (E). As indicated in the right panel of Figure A.1, methionine (M), valine (V), and lysine (K) are often associated with larger attention values. Notably, the side chains of M and V are nonpolar and are categorized under the aliphatic group.

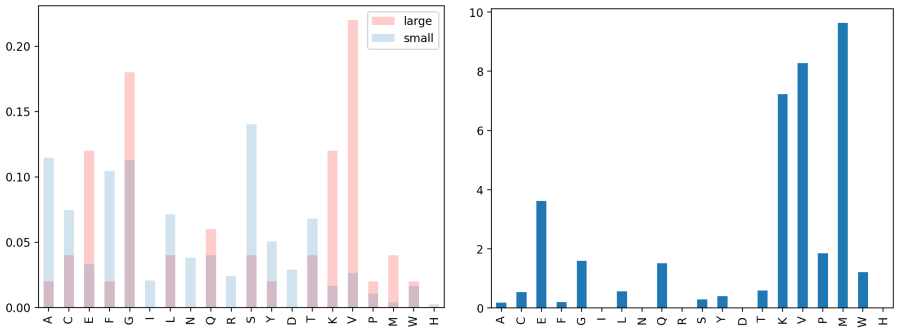

FIGURE A.1: Comparison of occurrence of amino acid types. Left: Distribution of residue types in CDR groups with large and small attention values. Right: The ratio of occurrence in the large attention group to the small attention group (large divided by small), highlighting the tendency of residues to have large attention values.

The influence of the factor γ

In the original research [1], the attention values were deemed "large" if they surpassed MEAN + 5.5 STD for peptides and MEAN + 4.5 STD for TCRs (with 5.5 and 4.5 being γ factors). Applying these thresholds, around 20% of residues on each side were classified as having large attention following the concatenation of the four head classifications. The counts of each residue by changing γ are shown in Table A.1.

TABLE A.1: When changing the factor γ, count of large-valued attention or small-valued attention can vary. Four heads are merged.

	TCRs Large Attention	**TCRs Small Attention**	**Peptides Large Attention**	**Peptides Small Attention**
count (γ=2.0)	569	468	372	65
count (γ=2.5)	441	596	325	112
count (γ=3.0)	364	673	264	173
count (γ=3.5	299	738	219	218
count (γ=4.0)	244	793	172	265
count (γ=4.5	196	841	137	300
count (γ=5.0)	153	884	111	326
count (γ=5.5	118	919	81	356
count (γ=6.0)	97	940	48	389

Statistics of additional dataset

Table A.2 displays unique count statistics. The record column represents the unique count of CDR3α, CDR3β, Peptide pairs. The duplication count in rows refers to the quantity of unique data overlapping between the training and test sets.

TABLE A.2: Unique count statistics.

Dataset name	CDR3α	CDR3β	Record
McPAS, training	2423	2560	23363
McPAS, test	718	714	4729
McPAS, duplication count	218	198	0
VDJdb-without10x, training	2151	2171	19526
VDJdb-without10x, test	570	572	4010
VDJdb, duplication count	198	196	0
The combined data dataset	17954	19162	119046
The recent data test set	4782	5174	33360
The duplication count of the two above	444	148	0

Attention layer weights can detect indicators of positive binding

In a supplementary experiment, I evaluated the effectiveness of the model's attention mechanism using mock data. For every CDR3α that has a positive interaction, I appended two alanine (A) residues as a distinct marker of positive binding. After training the model with this artificial data from McPAS, it scored a perfect ROCAUC of 1.0. When examining the attention layer of the trained model, it was evident that the markers for positive interaction, the double alanine residues, received high attention values, as illustrated in Figure A.2. Thus, attention values served as a means to interpret the binding mechanism.

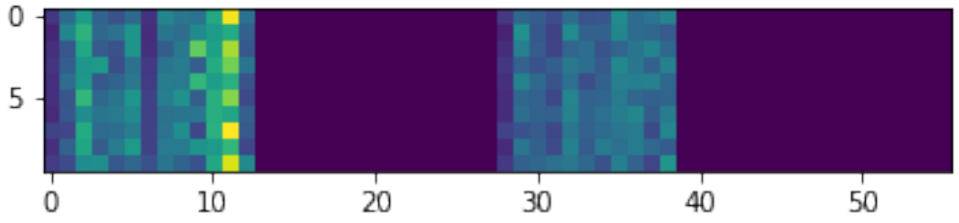

FIGURE A.2: Visualization of attention values for a positive record in the mock dataset. The 11th residue position, the first Alanine added to the TCR alpha chain, received high attention. The X-axis represents residues of CDR3s, and the Y-axis represents residues of peptides. Dark blue marks zero attention values (padding), while yellow shows large attention values.

Comparison of models on the recent test dataset

To address the query about the performance of the model trained on the combined data dataset in comparison to models trained on smaller datasets, I compared three models: The "combined-data-trained model", the "McPAS-model-trained" and the "VDJdb-model". The question was essentially about whether or not training on a larger combined dataset degrades the model's performance.

Each model was evaluated on a recent test dataset. The combined-data-trained model was developed using a larger training dataset that combines data from both the McPAS and VDJdb databases. In contrast, the McPAS-model and VDJdb-model were trained using smaller, individual datasets.

The results are as follows:

- **Combined-data-trained model:** This model achieved the average precision (AP) of 0.1855 and an ROCAUC score of 0.5362 as stated in the table of the main manuscript.
- **McPAS-trained model:** Evaluated scores were 0.1797 for AP and 0.5292 for ROCAUC. These scores were slightly lower than the combined-data-trained model.
- **VDJdb-trained model:** This model returned an ROCAUC of 0.5235, which is below the combined model. However, its AP was 0.1860, comparable to the combined-data-trained model. Given that the recent test dataset originates from the VDJdb database, this model might have had a slight advantage.

The combined-data-trained model performed at least as well as the VDJdb-trained model and outperformed the McPAS-model. It's crucial to note that the "recent data test set" for the combined dataset is temporally distinct from its training dataset, providing a stringent testing condition. In contrast, test sets for VDJdb and McPAS were not temporally distinct. I believe this temporal distinctness explains the lower scores on the recent data test set, rather than the influence of the larger training dataset.

In conclusion, training on a larger combined dataset does not degrade the model's performance. In fact, the combined-data-trained model proves to be competitive, if not superior, when compared to models trained on individual datasets.

Results of statistical tests on structures with different attention heads

The results of the paired t-tests are shown. Although each head was analyzed equally and separately, they showed different results. Table A.3 and Table A.4 show the individual head result for the cross-attention model discussed in Chapter 2, being referenced from the published paper [1].

In Table A.5 and Table A.6, I showcased the structural property comparison between attention-highlighted residues and not-highlighted ones by specifying each head of attention layers obtained from Chapter 3.

TABLE A.3: Peptide side attention analysis on all heads of Cross-TCR-Interpreter. In a cell, the left number is the mean value over the PDBIDs and the right number is the STD value.

Property	Large Attention [1]	Small Attention [1]	p-value	Head
Closest distance to peptide (Å)	4.1233±1.3970	5.0796±1.1194	0.0703	0
Number of H-bonds formed	1.6667±1.5986	2.0125±0.7579	0.84	0
H-bonded to any CDR3 residue	0.0833±0.2764	0.1801±0.1086	0.409	0
H-bonded to any peptide residue	0.0833±0.2764	0.0656±0.1218	1	0
H-bonded to any TCR residue	0.3333±0.4714	0.2559±0.1198	0.552	0
H-bonded to any non-CDR3 TCR residue	0.2500±0.4330	0.1167±0.1116	0.318	0
In the edge [2]	0.2500±0.4330	0.5637±0.0951	0.0504	0
Closest distance to peptide (Å)	5.1794±2.4525	5.0293±1.0674	0.895	1
Number of H-bonds formed	2.4348±1.7087	1.9672±0.7934	0.154	1
H-bonded to any CDR3 residue	0.3261±0.4570	0.1710±0.0996	0.215	1
H-bonded to any peptide residue	0.0000±0.0000	0.0673±0.1226	0.0711	1
H-bonded to any TCR residue	0.4130±0.4812	0.2499±0.1138	0.168	1
H-bonded to any non-CDR3 TCR residue	0.1739±0.3790	0.1162±0.1197	0.457	1
In the edge [2]	0.4783±0.4995	0.5610±0.1018	0.566	1
Closest distance to peptide (Å)	4.5241±1.8359	5.0844±1.0686	0.0771	2
Number of H-bonds formed	1.6304±1.0448	2.0183±0.7681	0.478	2
H-bonded to any CDR3 residue	0.1522±0.3437	0.1791±0.1188	0.653	2
H-bonded to any peptide residue	0.0435±0.2039	0.0659±0.1220	0.527	2
H-bonded to any TCR residue	0.2174±0.4125	0.2574±0.1175	0.827	2
H-bonded to any non-CDR3 TCR residue	0.0652±0.2238	0.1206±0.1151	0.664	2
In the edge [2]	0.4783±0.4773	0.5588±0.0997	0.688	2
Closest distance to peptide (Å)	3.7267±1.0947	5.1352±1.0994	5.15e-05	3
Number of H-bonds formed	1.7200±1.2496	2.0264±0.7809	0.503	3
H-bonded to any CDR3 residue	0.2800±0.4490	0.1714±0.0963	0.196	3
H-bonded to any peptide residue	0.1200±0.3250	0.0620±0.1108	0.31	3
H-bonded to any TCR residue	0.4000±0.4899	0.2491±0.1120	0.0866	3
H-bonded to any non-CDR3 TCR residue	0.1200±0.3250	0.1197±0.1162	0.697	3
In the edge [2]	0.2600±0.4271	0.5771±0.1191	0.00591	3
H-bonded to any peptide residue	0.0495±0.1443	0.0659±0.1206	0.458	all
H-bonded to any CDR3 residue	0.2050±0.3024	0.1682±0.0982	0.48	all
H-bonded to any TCR residue	0.3401±0.3714	0.2372±0.1184	0.151	all
H-bonded to any non-CDR3 TCR residue	0.1712±0.3112	0.1118±0.1283	0.355	all
In the edge [2]	0.4459±0.4097	0.5874±0.1232	0.0795	all
Closest distance to peptide (Å)	4.6398±1.7149	5.1926±1.2647	0.141	all
Number of H-bonds formed	2.1126±1.4959	2.0031±0.9051	0.668	all

1. Mean and standard deviation (for the 39 structures) of the proportion of residues that satisfy the property shown in the first column.
2. Three residues from the beginning and four from the end of the peptide. 3. In the last two properties, per-residue averages were used instead.

TABLE A.4: TCR side attention analysis on all heads of Cross-TCR-Interpreter. In a cell, the left number is the mean and the right number is the STD over the PDBIDs.

Property	Large Attention [1]	Small Attention [1]	p-value	Head
H-bonded to any peptide residue	0.0256 ± 0.1581	0.0853 ± 0.0577	0.043	0
H-bonded to any CDR3 residue	0.6068 ± 0.4262	0.4135 ± 0.1061	0.0111	0
H-bonded to any TCR residue of own chain	0.7051 ± 0.4038	0.6408 ± 0.0888	0.353	0
H-bonded to any CDR3 residue of own chain	0.6068 ± 0.4262	0.3792 ± 0.0991	0.0027	0
H-bonded to any TCR residue of opposite chain	0.1111 ± 0.2833	0.1565 ± 0.0756	0.371	0
H-bonded to any CDR3 residue of opposite chain	0.0000 ± 0.0000	0.0642 ± 0.0673	8.37e-07	0
H-bonded to any TCR residue	0.7308 ± 0.3897	0.7193 ± 0.0903	0.864	0
H-bonded to any non-CDR3 TCR residue	0.1966 ± 0.3734	0.4479 ± 0.0738	0.000482	0
In the edge [2]	0.8376 ± 0.3235	0.5904 ± 0.0387	5.1e-05	0
Closest distance to peptide (Å)	9.1216 ± 3.1274	8.3528 ± 1.0097	0.119	0
Number of H-bonds formed	1.7479 ± 1.1003	2.0919 ± 0.6806	0.0397	0
H-bonded to any peptide residue	0.1453 ± 0.3162	0.0773 ± 0.0564	0.2	1
H-bonded to any CDR3 residue	0.5171 ± 0.4333	0.4147 ± 0.1077	0.183	1
H-bonded to any TCR residue of own chain	0.5427 ± 0.4315	0.6488 ± 0.0892	0.152	1
H-bonded to any CDR3 residue of own chain	0.5085 ± 0.4335	0.3817 ± 0.0999	0.0913	1
H-bonded to any TCR residue of opposite chain	0.1496 ± 0.3198	0.1510 ± 0.0714	0.979	1
H-bonded to any CDR3 residue of opposite chain	0.0171 ± 0.1054	0.0616 ± 0.0678	0.0467	1
H-bonded to any TCR residue	0.6368 ± 0.4480	0.7211 ± 0.0881	0.264	1
H-bonded to any non-CDR3 TCR residue	0.1581 ± 0.2995	0.4511 ± 0.0648	5.28e-07	1
In the edge [2]	0.5513 ± 0.4388	0.6055 ± 0.0443	0.477	1
Closest distance to peptide (Å)	7.7849 ± 3.3952	8.4412 ± 0.9215	0.15	1
Number of H-bonds formed	1.9957 ± 1.5568	2.0725 ± 0.6687	0.742	1
H-bonded to any peptide residue	0.1127 ± 0.2935	0.0805 ± 0.0585	0.654	2
H-bonded to any CDR3 residue	0.3775 ± 0.4470	0.4247 ± 0.0992	0.614	2
H-bonded to any TCR residue of own chain	0.4853 ± 0.4615	0.6499 ± 0.0851	0.067	2
H-bonded to any CDR3 residue of own chain	0.3480 ± 0.4379	0.3919 ± 0.0947	0.633	2
H-bonded to any TCR residue of opposite chain	0.1324 ± 0.3278	0.1557 ± 0.0707	0.759	2
H-bonded to any CDR3 residue of opposite chain	0.0588 ± 0.2353	0.0614 ± 0.0678	0.933	2
H-bonded to any TCR residue	0.5882 ± 0.4451	0.7249 ± 0.0874	0.127	2
H-bonded to any non-CDR3 TCR residue	0.3235 ± 0.4021	0.4399 ± 0.0748	0.144	2
In the edge [2]	0.4706 ± 0.4705	0.6060 ± 0.0390	0.126	2
Closest distance to peptide (Å)	7.9293 ± 4.0178	8.3982 ± 1.0153	0.588	2
Number of H-bonds formed	2.1765 ± 1.8190	2.0746 ± 0.6642	0.607	2
H-bonded to any peptide residue	0.0877 ± 0.2470	0.0818 ± 0.0572	0.909	3
H-bonded to any CDR3 residue	0.4342 ± 0.4318	0.4234 ± 0.1013	0.852	3
H-bonded to any TCR residue of own chain	0.5921 ± 0.4270	0.6499 ± 0.0865	0.466	3
H-bonded to any CDR3 residue of own chain	0.3947 ± 0.4161	0.3908 ± 0.0969	0.927	3
H-bonded to any TCR residue of opposite chain	0.2237 ± 0.2993	0.1476 ± 0.0693	0.116	3
H-bonded to any CDR3 residue of opposite chain	0.0526 ± 0.1916	0.0615 ± 0.0631	0.767	3
H-bonded to any TCR residue	0.6842 ± 0.3723	0.7236 ± 0.0884	0.57	3
H-bonded to any non-CDR3 TCR residue	0.3684 ± 0.3590	0.4391 ± 0.0678	0.263	3
In the edge [2]	0.7281 ± 0.3735	0.5962 ± 0.0430	0.0464	3
Closest distance to peptide (Å) [3]	8.6919 ± 3.5172	8.4100 ± 1.0378	0.597	3
Number of H-bonds formed [3]	2.1535 ± 1.4009	2.0690 ± 0.6701	0.622	3
H-bonded to any peptide residue	0.0862±0.1368	0.0805±0.0675	0.828	all
H-bonded to any CDR3 residue	0.4846±0.2216	0.4103±0.1040	0.0478	all
H-bonded to any TCR residue of own chain	0.6013±0.1999	0.6561±0.0880	0.117	all
H-bonded to any CDR3 residue of own chain	0.4643±0.2180	0.3752±0.0922	0.0107	all
H-bonded to any TCR residue of opposite chain	0.1679±0.1714	0.1497±0.0793	0.562	all
H-bonded to any CDR3 residue of opposite chain	0.0306±0.0857	0.0672±0.0743	0.0369	all
H-bonded to any TCR residue	0.6845±0.1650	0.7294±0.0880	0.0987	all
H-bonded to any non-CDR3 TCR residue	0.2940±0.1923	0.4672±0.0846	3.88e-05	all
In the edge [2]	0.6434±0.2064	0.5928±0.0570	0.218	all
Closest distance to peptide (Å) [3]	8.4072±2.2892	8.4122±0.9592	0.988	all
Number of H-bonds formed [3]	2.0234±0.9370	2.0875±0.6685	0.589	all

1. Mean and standard deviation (for the 39 structures) of the proportion of residues that satisfy the property shown in the first column.
2. Four residues from the beginning and four from the end of the CDR. 3. In the last two properties, per-residue averages were used instead.

TABLE A.5: For the improved model, the structural property comparison results between the high and low attention residue groups by each head.

Model	Side	Property	Large Atten.	Small Atten.	P Value	Head	
Improved Model	TCR	Proportion is_connecting_to_pep	0.0435+-0.2039	0.0779+-0.0536	2.570700e-01	0	
	TCR	Proportion is_connecting_to_cdr	0.3913+-0.4768	0.4446+-0.1034	4.817425e-01	0	
	TCR	Proportion is_connecting_to_ownchain_cdr	0.3913+-0.4768	0.4136+-0.0985	7.693818e-01	0	
	TCR	Proportion is_connecting_to_tcr	0.7065+-0.4493	0.7433+-0.0912	5.806234e-01	0	
	TCR	Proportion digit4_is_in_edge	0.2609+-0.3864	0.6170+-0.0416	5.575990e-07	0	***
	TCR	distance_value	6.3723+-3.5671	8.3285+-0.9984	6.797449e-04	0	***
	TCR	num_bonds	1.6522+-1.1603	2.1445+-0.6957	1.112145e-03	0	***
	TCR	Proportion is_connecting_to_pep	0.0952+-0.2586	0.0750+-0.0479	5.818289e-01	1	
	TCR	Proportion is_connecting_to_cdr	0.5918+-0.4442	0.4304+-0.0973	1.651444e-02	1	***
	TCR	Proportion is_connecting_to_ownchain_cdr	0.5408+-0.4461	0.4067+-0.0931	4.618516e-02	1	***
	TCR	Proportion is_connecting_to_tcr	0.7959+-0.3584	0.7398+-0.0946	2.959348e-01	1	
	TCR	Proportion digit4_is_in_edge	0.6463+-0.4462	0.6017+-0.0383	5.101961e-01	1	
	TCR	distance_value	7.4668+-3.8247	8.3152+-0.9455	1.318983e-01	1	
	TCR	num_bonds	2.5170+-1.4083	2.1075+-0.6953	1.967807e-02	1	***
	TCR	Proportion is_connecting_to_pep	0.1496+-0.3393	0.0720+-0.0451	1.410653e-01	2	
	TCR	Proportion is_connecting_to_cdr	0.4872+-0.4472	0.4268+-0.1006	4.277684e-01	2	
	TCR	Proportion is_connecting_to_ownchain_cdr	0.4744+-0.4538	0.3991+-0.0927	3.262043e-01	2	
	TCR	Proportion is_connecting_to_tcr	0.8846+-0.2646	0.7195+-0.0961	9.987631e-04	2	***
	TCR	Proportion digit4_is_in_edge	0.5598+-0.4505	0.6034+-0.0380	5.661873e-01	2	
	TCR	distance_value	8.2835+-4.4946	8.2558+-0.9384	9.638520e-01	2	
	TCR	num_bonds	2.3291+-1.5019	2.1180+-0.6936	1.248931e-01	2	
	TCR	Proportion is_connecting_to_pep	0.1111+-0.2893	0.0778+-0.0544	4.609290e-01	3	
	TCR	Proportion is_connecting_to_cdr	0.5781+-0.4123	0.4314+-0.0965	1.993077e-02	3	***
	TCR	Proportion is_connecting_to_ownchain_cdr	0.5365+-0.4180	0.4044+-0.0902	3.486602e-02	3	***
	TCR	Proportion is_connecting_to_tcr	0.7014+-0.4124	0.7436+-0.0906	4.879866e-01	3	
	TCR	Proportion digit4_is_in_edge	0.5017+-0.4244	0.6136+-0.0437	9.306497e-02	3	*
	TCR	distance_value	6.8575+-3.4858	8.3529+-0.9824	4.946090e-03	3	***
	TCR	num_bonds	2.4514+-1.6116	2.1082+-0.6999	1.103505e-01	3	
	Peptide	Proportion is_connecting_to_cdr	0.2381+-0.3970	0.1581+-0.0830	3.624122e-01	0	
	Peptide	Proportion is_connecting_to_tcr	0.2857+-0.4246	0.2458+-0.1152	6.831083e-01	0	
	Peptide	Proportion pepres___is_in_edge	0.5000+-0.4629	0.5249+-0.1059	8.148104e-01	0	
	Peptide	pepres___distance_value	4.0704+-2.2471	4.9875+-1.0222	3.627001e-02	0	***
	Peptide	pepres___num_bonds	1.6905+-1.1178	2.1557+-0.7791	7.297248e-01	0	
	Peptide	Proportion is_connecting_to_cdr	0.3125+-0.4635	0.1363+-0.0894	1.306443e-01	1	
	Peptide	Proportion is_connecting_to_tcr	0.4375+-0.4961	0.2424+-0.1005	1.324123e-01	1	
	Peptide	Proportion pepres___is_in_edge	0.6250+-0.4841	0.5331+-0.1071	4.829076e-01	1	
	Peptide	pepres___distance_value	4.6176+-2.1006	4.9592+-0.9888	2.590448e-01	1	
	Peptide	pepres___num_bonds	2.8750+-1.5360	2.1193+-0.7922	4.432649e-02	1	***
	Peptide	Proportion is_connecting_to_cdr	0.3056+-0.4453	0.1926+-0.1361	3.289011e-01	2	
	Peptide	Proportion is_connecting_to_tcr	0.4722+-0.4556	0.2692+-0.1622	1.180892e-01	2	
	Peptide	Proportion pepres___is_in_edge	0.5278+-0.4851	0.5172+-0.0948	9.269207e-01	2	
	Peptide	pepres___distance_value	3.8094+-1.3488	4.9931+-1.0577	4.122880e-03	2	***
	Peptide	pepres___num_bonds	2.7778+-2.0630	2.1155+-0.7743	6.390261e-02	2	*
	Peptide	Proportion is_connecting_to_cdr	0.3478+-0.4763	0.1705+-0.1379	1.323572e-01	3	
	Peptide	Proportion is_connecting_to_tcr	0.3913+-0.4880	0.2532+-0.1588	2.634958e-01	3	
	Peptide	Proportion pepres___is_in_edge	0.4783+-0.4995	0.5407+-0.1233	6.112612e-01	3	
	Peptide	pepres___distance_value	5.0025+-3.7390	4.9582+-0.9705	9.291060e-01	3	
	Peptide	pepres___num_bonds	2.4783+-2.3750	2.1383+-0.7557	5.587625e-01	3	

TABLE A.6: For the standard attention model, the structural property comparison results between the high and low attention residue groups by each head

Model	Side	Property	Large Atten.	Small Atten.	P-Value	Head	
Standard	TCR	Proportion is_connecting_to_pep	0.1152+-0.1708	0.0446+-0.0608	5.128678e-03	0	***
	TCR	Proportion is_connecting_to_cdr	0.5474+-0.2769	0.4469+-0.1700	3.338341e-02	0	***
	TCR	Proportion is_connecting_to_ownchain_cdr	0.5142+-0.2836	0.4228+-0.1708	5.272380e-02	0	*
	TCR	Proportion is_connecting_to_tcr	0.8325+-0.2087	0.7122+-0.1540	2.316844e-04	0	***
	TCR	Proportion digit4_is_in_edge	0.6345+-0.2693	0.6353+-0.0874	9.853132e-01	0	
	TCR	distance_value	8.0513+-2.8446	8.3693+-1.4209	4.264827e-01	0	
	TCR	num_bonds	2.4717+-1.1598	1.9728+-0.7684	3.657217e-04	0	***
	TCR	Proportion is_connecting_to_pep	0.0510+-0.1386	0.0752+-0.0818	3.009511e-01	1	
	TCR	Proportion is_connecting_to_cdr	0.6132+-0.3508	0.4490+-0.1323	2.050789e-03	1	***
	TCR	Proportion is_connecting_to_ownchain_cdr	0.5726+-0.3585	0.4258+-0.1362	5.497273e-03	1	***
	TCR	Proportion is_connecting_to_tcr	0.7639+-0.2941	0.7506+-0.1430	7.696887e-01	1	
	TCR	Proportion digit4_is_in_edge	0.7104+-0.3007	0.6214+-0.0948	8.777165e-02	1	*
	TCR	distance_value	8.3724+-2.2920	8.2474+-1.3273	6.275842e-01	1	
	TCR	num_bonds	2.1236+-1.0459	2.1213+-0.7447	7.830440e-01	1	
	TCR	Proportion is_connecting_to_pep	0.0455+-0.1249	0.0781+-0.0641	1.723829e-01	2	
	TCR	Proportion is_connecting_to_cdr	0.4798+-0.3989	0.5163+-0.1307	5.742949e-01	2	
	TCR	Proportion is_connecting_to_ownchain_cdr	0.4646+-0.3882	0.4946+-0.1266	6.421815e-01	2	
	TCR	Proportion is_connecting_to_tcr	0.7525+-0.3644	0.7543+-0.1283	9.766426e-01	2	
	TCR	Proportion digit4_is_in_edge	0.7020+-0.3500	0.6113+-0.0710	1.991390e-01	2	
	TCR	distance_value	8.0788+-2.8861	8.2790+-1.3446	7.489801e-01	2	
	TCR	num_bonds	1.9141+-0.9463	2.1240+-0.7434	9.369856e-02	2	*
	TCR	Proportion is_connecting_to_pep	0.0311+-0.0987	0.0741+-0.0706	1.045949e-02	3	***
	TCR	Proportion is_connecting_to_cdr	0.4644+-0.2859	0.4990+-0.1200	4.243659e-01	3	
	TCR	Proportion is_connecting_to_ownchain_cdr	0.4414+-0.2936	0.4746+-0.1285	4.632827e-01	3	
	TCR	Proportion is_connecting_to_tcr	0.6701+-0.3401	0.7643+-0.1599	1.503970e-01	3	
	TCR	Proportion digit4_is_in_edge	0.7264+-0.2799	0.6128+-0.1218	3.770662e-02	3	***
	TCR	distance_value	8.9846+-2.9062	8.1733+-1.4200	5.419123e-02	3	*
	TCR	num_bonds	1.9280+-1.0784	2.1002+-0.7389	2.476485e-01	3	
	Peptide	Proportion is_connecting_to_cdr	0.1068+-0.1983	0.2245+-0.1388	5.888566e-04	0	***
	Peptide	Proportion is_connecting_to_tcr	0.2250+-0.2718	0.2978+-0.1430	9.741488e-02	0	*
	Peptide	Proportion pepres_is_in_edge	0.6428+-0.3077	0.5474+-0.1348	5.579167e-02	0	*
	Peptide	pepres_distance_value	5.2700+-1.6018	4.7711+-1.0923	5.640764e-02	0	*
	Peptide	pepres_num_bonds	2.1119+-1.3299	2.1817+-0.9469	7.389934e-01	0	
	Peptide	Proportion is_connecting_to_cdr	0.1628+-0.2243	0.1842+-0.1401	5.633909e-01	1	
	Peptide	Proportion is_connecting_to_tcr	0.3006+-0.3143	0.2788+-0.1527	6.735949e-01	1	
	Peptide	Proportion pepres_is_in_edge	0.5342+-0.3218	0.5577+-0.1529	6.809043e-01	1	
	Peptide	pepres_distance_value	4.9643+-1.8243	4.9214+-0.9025	9.525252e-01	1	
	Peptide	pepres_num_bonds	2.1335+-1.4119	2.1542+-0.9025	8.875797e-01	1	
	Peptide	Proportion is_connecting_to_cdr	0.1945+-0.2071	0.1981+-0.1690	9.234416e-01	2	
	Peptide	Proportion is_connecting_to_tcr	0.2447+-0.2200	0.3075+-0.1709	1.366735e-01	2	
	Peptide	Proportion pepres_is_in_edge	0.5816+-0.2457	0.5401+-0.1902	4.016266e-01	2	
	Peptide	pepres_distance_value	5.0151+-1.3035	4.7941+-1.4010	3.751639e-01	2	
	Peptide	pepres_num_bonds	2.2383+-1.0157	2.1754+-1.0662	7.131025e-01	2	
	Peptide	Proportion is_connecting_to_cdr	0.0884+-0.1988	0.2117+-0.1318	5.864652e-04	3	***
	Peptide	Proportion is_connecting_to_tcr	0.2194+-0.2875	0.2941+-0.1333	1.014840e-01	3	
	Peptide	Proportion pepres_is_in_edge	0.6803+-0.3667	0.5381+-0.0951	1.089272e-02	3	***
	Peptide	pepres_distance_value	5.1556+-1.4941	4.8600+-1.0564	1.806034e-01	3	
	Peptide	pepres_num_bonds	2.1207+-1.7253	2.1408+-0.7772	8.883590e-01	3	

PyMol Command for PDB 5TEZ

```
fetch 5TEZ;
set seq_view, 1;
bg_color white;
hide all;
remove waters;
select beta, chain J and not solvent;
select alpha, chain I and not solvent;
select mhc, (chain A or chain B or chain D or chain E) and not solvent;
show cartoon, alpha;
color wheat, alpha;
show cartoon, beta;
color lightblue, beta;
show cartoon,mhc
color grey90, mhc;
create obj_mhc, mhc
show surface, obj_mhc
set transparency=0.2
sel beta_cdr3, (chain J and resi 91:104);
#set cartoon_side_chain_helper, on
#show sticks, beta_cdr3;
#util.cbag beta_cdr3;
color palecyan, beta_cdr3
sel alpha_cdr3, (chain I and resi 91:105);
#set cartoon_side_chain_helper, on
#show sticks, alpha_cdr3;
#util.cbag alpha_cdr3;
color lightpink, alpha_cdr3
select epitope, chain C and not solvent;
show sticks, epitope;
color yellow, epitope
#util.cbay epitope;
#select cdr3, alpha_cdr3 or beta_cdr3;
#select tcr, alpha or beta;
#dist H_cdr_p, cdr3, epitope, mode=2;
#hide labels, H_cdr_p;
#color black, H_cdr_p;
#dist H_cdr_tcr, cdr3, tcr, mode=2;
#hide labels, H_cdr_tcr;
#color grey, H_cdr_tcr;
sel atten_a_head1, (resi 104 and chain I);
#color pink, atten_a_head1;
sel atten_a_head2, (resi 101 and chain I);
#color pink, atten_a_head2;
sel atten_a_head3, (resi 101 and chain I);
#color pink, atten_a_head3;
sel atten_b_head0, (resi 99 and chain J);
#color pink, atten_b_head0;
show sticks, atten_a_head1
show sticks, atten_a_head2
show sticks, atten_a_head3
show sticks, atten_b_head0
color magenta, atten_a_head1
color magenta, atten_a_head2
```

```
color magenta, atten_a_head3
color cyan, atten_b_head0
sel atten_1_int, (resi 93 and chain I)
sel atten_23_int, (resi 94 and chain I)
sel atten_230_int, (resi 6 and chain C)
show sticks, atten_1_int
show sticks, atten_23_int
show sticks, atten_0_int
sel int_int,(resi 98 and chain J)
show sticks, int_int
color atomic,(not elem C)
color gray90, obj_mhc
dist a1_hb,(resi 104 and chain I),(resi 93 and chain I),mode=2
dist a23_hb,(resi 101 and chain I),(resi 94 and chain I),mode=2
dist a23p_hb,(resi 101 and chain I),(resi 6 and chain C),mode=2
dist a0_hb,(resi 99 and chain J),(resi 6 and chain C),mode=2
dist intint,(resi 94 and chain I),(resi 98 and chain J),mode=2
hide labels, a1_hb
hide labels, a23_hb
hide labels, a23p_hb
hide labels, a0_hb
hide labels, intint
```

S Protein sequence of Covid-19

- mfvflvllpl vssqcvnltt rtqlppaytn sftrgvyypd kvfrssvlhs tqdlflpffs nvtwfhaihv sgtngtkrfd npvlpfndgv yfasteksni irgwifgttl dsktqslliv nnatnvvikv cefqfcndpf lgvyyhknnk swmesefrvy ssannctfey vsqpflmdle gkqgnfknlr efvfknidgy fkiyskhtpi nlvrdlpqgf saleplvdlp iginitrfqt llalhrsylt pgdsssgwta gaaayyvgyl qprtfllkyn engtitdavd caldplsetk ctlksftvek giyqtsnfrv qptesivrfp nitnlcpfge vfnatrfasv yawnrkrisn cvadysvlyn sasfstfkcy gvsptklndl cftnvyadsf virgdevrqi apgqtgkiad ynyklpddft gcviawnsnn ldskvggnyn ylyrlfrksn lkpferdist eiyqagstpc ngvegfncyf plqsygfqpt ngvgyqpyrv vvlsfellha patvcgpkks tnlvknkcvn fnfngltgtg vltesnkkfl pfqqfgrdia dttdavrdpq tleilditpc sfggvsvitp gtntsnqvav lyqdvnctev pvaihadqlt ptwrvystgs nvfqtragcl igaehvnnsy ecdipigagi casyqtqtns prrarsvasq siiaytmslg aensvaysnn siaiptnfti svtteilpvs mtktsvdctm yicgdstecs nlllqygsfc tqlnraltgi aveqdkntqe vfaqvkqiyk tppikdfggf nfsqilpdps kpskrsfied llfnkvtlad agfikqygdc lgdiaardli caqkfngltv lppltdemi aqytsallag titsgwtfga gaalqipfam qmayrfngig vtqnvlyenq klianqfnsa igkiqdslss tasalgklqd vvnqnaqaln tlvkqlssnf gaissvlndi lsrldkveae vqidrlitgr lqslqtyvtq qliraaeira sanlaatkms ecvlgqskrv dfcgkgyhlm sfpqsaphgv vflhvtyvpa qeknfttapa ichdgkahfp regvfvsngt hwfvtqrnfy epqiittdnt fvsgncdvvi givnntvydp lqpeldsfke eldkyfknht spdvdlgdis ginasvvniq keidrlneva knlneslidl qelgkyeqyi kwpwyiwlgf iagliaivmv timlccmtsc csclkgccsc gscckfdedd sepvlkgvkl

Bibliography

[1] Kyohei Koyama, Kosuke Hashimoto, Chioko Nagao, and Kenji Mizuguchi. Attention network for predicting T cell receptor-peptide binding can associate attention with interpretable protein structural properties. *Frontiers in Bioinformatics*, Volume 3, 2023. doi: 10.3389/fbinf.2023.1274599.

[2] Ragul Gowthaman and Brian G Pierce. TCR3d: The T cell receptor structural repertoire database. *Bioinformatics*, 35(24):5323–5325, 2019.

[3] Jinwoo Leem, Saulo H P de Oliveira, Konrad Krawczyk, and Charlotte M Deane. STCRDab: the structural T-cell receptor database. *Nucleic acids research*, 46(D1): D406–D412, 2018.

[4] Swapnil Mahajan, Zhen Yan, Martin Closter Jespersen, Kamilla Kjærgaard Jensen, Paolo Marcatili, Morten Nielsen, Alessandro Sette, and Bjoern Peters. Benchmark datasets of immune receptor-epitope structural complexes. *BMC bioinformatics*, 20 (1):1–7, 2019.

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

[6] Pradyot Dash, Andrew J Fiore-Gartland, Tomer Hertz, George C Wang, Shalini Sharma, Aisha Souquette, Jeremy Chase Crawford, E Bridie Clemens, Thi HO Nguyen, Katherine Kedzierska, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, 547(7661):89–93, 2017.

[7] Ido Springer, Hanan Besser, Nili Tickotsky-Moskovitz, Shirit Dvorkin, and Yoram Louzoun. Prediction of specific TCR-peptide binding from large dictionaries of TCR-peptide pairs. *Frontiers in immunology*, page 1803, 2020.

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

[9] Tianshi Lu, Ze Zhang, James Zhu, Yunguan Wang, Peixin Jiang, Xue Xiao, Chantale Bernatchez, John V Heymach, Don L Gibbons, Jun Wang, et al. Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nature machine intelligence*, 3(10):864–875, 2021.

[10] Yicheng Gao, Yuli Gao, Yuxiao Fan, Chengyu Zhu, Zhiting Wei, Chi Zhou, Guohui Chuai, Qinchang Chen, He Zhang, and Qi Liu. Pan-Peptide Meta Learning for T-cell receptor–antigen binding recognition. *Nature Machine Intelligence*, pages 1–14, 2023.

[11] Kevin Wu, Kathryn E Yost, Bence Daniel, Julia A Belk, Yu Xia, Takeshi Egawa, Ansuman Satpathy, Howard Y Chang, and James Zou. TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-xbinding analyses. *bioRxiv*, page 2021.11.18.469186, November 2021.

[12] John-William Sidhom, H Benjamin Larman, Drew M Pardoll, and Alexander S Baras. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nature communications*, 12(1):1–12, 2021.

[13] Ying Xu, Xinyang Qian, Yao Tong, Fan Li, Ke Wang, Xuanping Zhang, Tao Liu, and Jiayin Wang. AttnTAP: A Dual-input Framework Incorporating the Attention Mechanism for Accurately Predicting TCR-peptide Binding. *Frontiers in Genetics*, page 1871, 2022.

[14] Zhaochun Xu, Meng Luo, Weizhong Lin, Guangfu Xue, Pingping Wang, Xiyun Jin, Chang Xu, Wenyang Zhou, Yideng Cai, Wenyi Yang, et al. DLpTCR: an ensemble deep learning framework for predicting immunogenic peptide recognized by T cell receptor. *Briefings in Bioinformatics*, 22(6):bbab335, 2021.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Łukasz Kaiser, Aidan N Gomez, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[17] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the

Rest Can Be Pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580.

[18] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12963–12971, 2021.

[19] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

[20] Mozhdeh Gheini, Xiang Ren, and Jonathan May. Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1754–1765, November 2021. doi: 10.18653/v1/2021.emnlp-main.132. URL https://aclanthology.org/2021.emnlp-main.132.

[21] Srinivas Parthasarathy and Shiva Sundaram. Detecting expressions with multimodal transformers. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 636–643. IEEE, 2021.

[22] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018.

[23] Jiecao Chen, Liu Yang, Karthik Raman, Michael Bendersky, Jung-Jung Yeh, Yun Zhou, Marc Najork, Danyang Cai, and Ehsan Emadzadeh. DiPair: Fast and accurate distillation for trillion-scale text matching and pair modeling. *arXiv preprint arXiv:2010.03099*, 2020.

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

[25] Kyohei Koyama, Kotaro Kamiya, and Koki Shimada. Cross Attention DTI: Drug-Target Interaction Prediction with Cross Attention module in the Blind Evaluation Setup. *BIOKDD2020*, 2020.

[26] Shion Honda, Kyohei Koyama, and Kamiya Kotaro. Cross Attentive Antibody-Antigen Interaction Prediction with Multi-task Learning. *ICML 2020 Workshop on Computational Biology (WCB)*, 2020.

[27] Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. BERTology Meets Biology: Interpreting Attention in Protein Language Models, 2020. URL https://arxiv.org/abs/2006.15222.

[28] Nili Tickotsky, Tal Sagiv, Jaime Prilusky, Eric Shifrut, and Nir Friedman. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics*, 33(18):2924–2929, 2017.

[29] Mikhail Shugay, Dmitriy V Bagaev, Ivan V Zvyagin, Renske M Vroomans, Jeremy Chase Crawford, Garry Dolton, Ekaterina A Komech, Anastasiya L Sycheva, Anna E Koneva, Evgeniy 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.

[30] Xiuyuan Lu, Yuki Hosono, Masamichi Nagae, Shigenari Ishizuka, Eri Ishikawa, Daisuke Motooka, Yuki Ozaki, Nicolas Sax, Yuichi Maeda, Yasuhiro Kato, Takayoshi Morita, Ryo Shinnakasu, Takeshi Inoue, Taishi Onodera, Takayuki Matsumura, Masaharu Shinkai, Takashi Sato, Shota Nakamura, Shunsuke Mori, Teru Kanda, Emi E. Nakayama, Tatsuo Shioda, Tomohiro Kurosaki, Kiyoshi Takeda, Atsushi Kumanogoh, Hisashi Arase, Hironori Nakagami, Kazuo Yamashita, Yoshimasa Takahashi, and Sho Yamasaki. Identification of conserved SARS-CoV-2 spike epitopes that expand public cTfh clonotypes in mild COVID-19 patientsSARS-CoV-2 spike epitopes for public cTfh cells. *Journal of Experimental Medicine*, 218(12), 10 2021. ISSN 0022-1007. doi: 10.1084/jem.20211327. URL https://doi.org/10.1084/jem.20211327. e20211327.

[31] 10x Genomics. A New Way of Exploring Immunity–Linking Highly Multiplexed Antigen Recognition to Immune Repertoire and Phenotype. *Tech. rep*, 2019.

[32] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7(1):539, 2011.

[33] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[34] Helen Berman, Kim Henrick, and Haruki Nakamura. Announcing the worldwide protein data bank. *Nature Structural & Molecular Biology*, 10(12):980–980, 2003.

[35] James Dunbar and Charlotte M Deane. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016.

[36] Brad Chapman and Jeffrey Chang. Biopython: Python tools for computational biology. *ACM Sigbio Newsletter*, 20(2):15–19, 2000.

[37] Andrew C Wallace, Roman A Laskowski, and Janet M Thornton. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein engineering, design and selection*, 8(2):127–134, 1995.

[38] Xinbo Yang, Guobing Chen, Nan-ping Weng, and Roy A Mariuzza. Structural basis for clonal diversity of the human T-cell response to a dominant influenza virus epitope. *Journal of Biological Chemistry*, 292(45):18618–18627, 2017. doi: 10.2210/pdb5tez/pdb.

[39] Schrödinger and LLC and Warren DeLano. PyMOL. *http://www.pymol.org/pymol*, 2020.

[40] David K Cole, Malkit Sami, Daniel R Scott, Pierre J Rizkallah, Oleg Y Borbulevych, Penio T Todorov, Ruth K Moysey, Bent K Jakobsen, Jonathan M Boulter, Brian M Baker, et al. Increased peptide contacts govern high affinity binding of a modified TCR whilst maintaining a native pMHC docking mode. *Frontiers in immunology*, 4:168, 2013.

[41] Lance M Hellman, Kendra C Foley, Nishant K Singh, Jesus A Alonso, Timothy P Riley, Jason R Devlin, Cory M Ayres, Grant LJ Keller, Yuting Zhang, Craig W Vander Kooi, et al. Improving T cell receptor on-target specificity via structure-guided design. *Molecular Therapy*, 27(2):300–313, 2019.

[42] K Christopher Garcia, Massimo Degano, Robyn L Stanfield, Anders Brunmark, Michael R Jackson, Per A Peterson, Luc Teyton, and Ian A Wilson. An $\alpha\beta$ T cell receptor structure at 2.5 Å and its orientation in the TCR-MHC complex. *Science*, 274(5285):209–219, 1996.

[43] Miguel Andrade, Camila Pontes, and Werner Treptow. Coevolutive, evolutive and stochastic information in protein-protein interactions. *Computational and Structural Biotechnology Journal*, 17:1429–1435, 2019.

[44] Shah Md Abdur Rauf, Mohamed Ismael, Kamlesh Kumar Sahu, Ai Suzuki, Riadh Sahnoun, Michihisa Koyama, Hideyuki Tsuboi, Nozomu Hatakeyama, Akira Endou, Hiromitsu Takaba, et al. A graph theoretical approach to the effect of mutation on the flexibility of the DNA binding domain of p53 protein. *Chemical Papers*, 63: 654–661, 2009.

[45] Dana Reichmann, Ofer Rahat, Shira Albeck, Ran Meged, Orly Dym, and Gideon Schreiber. The modular architecture of protein-protein binding interfaces. *Proceedings of the National Academy of Sciences*, 102(1):57–62, 2005.

[46] Pieter Moris, Joey De Pauw, Anna Postovskaya, Sofie Gielis, Nicolas De Neuter, Wout Bittremieux, Benson Ogunjimi, Kris Laukens, and Pieter Meysman. Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. *Briefings in Bioinformatics*, 22(4):bbaa318, 2021.

[47] Ahmed Essaghir, Nanda Kumar Sathiyamoorthy, Paul Smyth, Adrian Postelnicu, Stefan Ghiviriga, Alexandru Ghita, Anjana Singh, Shruti Kapil, Sanjay Phogat, and Gurpreet Singh. T-cell receptor specific protein language model for prediction and interpretation of epitope binding (ProtLM. TCR). *bioRxiv*, pages 2022–11, 2022.

Publication and Presentation

Publication

- 小山恭平.　アテンションネットワークによる T 細胞受容体とペプチド結合の予測・その機能の解釈性に関する研究.　サイバーメディア HPC ジャーナル, No.12,　2022.　doi:10.18910/89341
- Kyohei Koyama, Kosuke Hashimoto, Chioko Nagao, and Kenji Mizuguchi. Attention network for predicting T cell receptor-peptide binding can associate attention with interpretable protein structural properties. Frontiers in Bioinformatics, Volume 3, 2023. doi: 10.3389/fbinf.2023.1274599.

Presentation

- Kyohei Koyama. Interpretability of Protein Function Prediction by Neural Networks. Presented at the Institute for Protein Research Retreat (IPR Retreat), November 2021. (Awarded the Poster Presentation Award)
- 小山恭平. ニューラルネットワークによる蛋白質機能予測の解釈性に関する研究, 大阪大学サイバーメディアセンター 2021 年度公募型利用制度成果報告会, 2022