doi.10.1093/bioinformatics/xxxxxx

Advance Access Publication Date: Day Month Year

Manuscript Category

OXFORD

Problem solving protocols

# An attention network for predicting T cell receptor-peptide binding can associate attention with interpretable protein structural properties

**Kyohei KOYAMA [1, 2, 3, 4]*, Kosuke HASHIMOTO [1], Chioko NAGAO [1], and Kenji MIZUGUCHI [1, 3, 4]**

[1] Laboratory for Computational Biology, Institute for Protein Research, Osaka University, Suita, Osaka, Japan

[2] SyntheticGestalt, Shinjuku,Tokyo, Japan.

[3] National Institutes of Biomedical Innovation, Health and Nutrition, Osaka, Japan

[4] Graduate School of Frontier Biosciences, Osaka University, Suita, Osaka, Japan.

* Correspondence: Kyohei KOYAMA

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

## Abstract

**Motivation:** Understanding how a T cell receptor (TCR) recognizes its specific ligand peptide is crucial for gaining insight into biological functions and disease mechanisms. Despite its importance, experimentally determining TCR-peptide interactions can be expensive and time-consuming. To address this challenge, several computational methods have been proposed, but none have incorporated an attention layer from language models in conjunction with TCR-peptide "structural" binding mechanisms.

**Results:** In this study, we developed a machine learning model based on a modified version of the transformer, a cross-attention neural network, to predict TCR-peptide binding solely based on protein sequences. This model achieved state-of-the-art performance as measured by the ROCAUC score on a benchmark dataset of TCR-peptide binding. Furthermore, more importantly, by analyzing the results of binding predictions, we associated the neural network weights with protein structural properties. By classifying the residues into large and small attention groups, we identified statistically significant properties associated with the largely attended residues, such as hydrogen-bond connections within the TCR. This ability to provide an interpretable prediction of TCR-peptide binding should increase our knowledge of molecular recognition and pave the way to designing new therapeutics.

**Contact:** u117209j@ecs.osaka-u.ac.jp

**Supplementary information:** Supplementary data are available at *https://github.com/kyoheikoyama/TCRPrediction* online.

## Introduction

The T cell receptor (TCR) is an antigen receptor composed mostly of alpha (TCR alpha) and beta (TCR beta) chains. The TCR is known for its sequence diversity in the complementarity-determining region (CDR), as is the B cell receptor (antibody). Because T cells are immune cells that are produced by somatic recombination, the sequence patterns of CDRs and TCRs are enormous, at up to $10^{61}$ [1, 2]. The CDRs of TCR bind and react with peptides on the major histocompatibility complex (MHC), the ligands of TCR, to initiate an immune response *in vivo*. These immune responses can lead to various reactions *in vivo*, ultimately leading to disease recovery. Understanding the mechanisms of TCR-peptide recognition is essential for gaining insight into biological functions and disease mechanisms.

In general, experimentally confirming protein function is often difficult, as this process requires substantial resources and a combination of techniques such as biochemistry, genetics, and structural biology. Three-dimensional (3D) structures of protein complexes can often provide a thorough understanding of protein functions [3, 4, 5], and techniques such as X-ray crystallography and cryo-electron microscopy can be used. However, these experiments are costly and time-consuming.

To address this challenge in predictive models of protein functions, deep learning models with different neural networks architectures have been successfully developed for a wide range of biological problems [6, 7, 8, 9, 10, 11]. Specifically, methods have been proposed to predict TCR-peptide binding, MHC-peptide binding, and TCR repertoire prediction [3, 12, 13, 14, 15, 16, 17, 18].

In this study, we aim to contribute to the research field regarding two key points: revealing relationships between the attention weights and protein structural properties, and overcoming the limitations of interpretation within the BERT method. The attention layer proposed in the transformer [19] is a powerful computational method that allows a model to focus on the most relevant parts of a sequence in relation to another sequence. This is achieved through attention weights, which are calculated by learning the conditional relationship between the two sequences, also known as the source and target sequences. Despite the heavy use of attention in prediction tasks, current TCR-peptide studies have not conducted statistical analyses of attention weights to reveal the relationship provided a sequence and to reveal the importance of specific structural properties such as hydrogen bonds (H-bonds). Additionally, previous methods have relied on deep learning models, including BERT-like models with attention networks [20, 21, 22, 23]. However, the BERT model's self-attention mechanism poses several limitations. The attention of BERT covers both two proteins at the same time and at the same layer, which makes it difficult to understand the specific interaction between the two proteins. This means that the BERT model's self-attention mechanism is not specifically limited to the mutual interaction between a TCR and a peptide. Instead, it is dependent on the two entire sequences, which can possibly lead to overgeneralized predictions and a lack of interpretability of the relationship. Therefore, we have restricted the attention mechanism only to the crossing subpart and addressed these limitations through a modified attention layer known as a cross-attention layer or a bimodal layer [24, 25, 26, 27], and an extensive analysis of protein structures. In addition, too many stacked layers of BERT and the complexity of its computation made it difficult to analyze the result.

The aim of this study is to develop a computational method that can predict TCR-peptide binding with high accuracy and to interpret TCR-peptide binding relationships to gain new biological insights into the TCR function. To achieve this purpose, we propose a model, the Cross-TCR-Interpreter, to use a cross-attention mechanism for predicting TCR-peptide binding. Unlike the traditional BERT attention values, our cross-attention mechanism allows for explicit analysis of the mutual relationships of residues. We approached this as a binary classification, and trained our model on a benchmark dataset of TCR-peptide binding. Our model achieved state-of-the-art performance of the ROCAUC score on the benchmark, and by performing statistical tests on the attention values, we successfully identified statistically significant structural properties of largely attended residues, such as hydrogen bonds and residue distance within known protein structures. This approach should allow for a deeper understanding and analysis of the interactions between the proteins and enable an analysis of one protein given the other.

## Materials and Methods

### Model

An overview of the prediction model used in this study is shown in Figure 1. The peptide sequence and the sequence of TCR$\alpha$ and $\beta$ connected with the connection-token (colon ":") were processed separately in the embedding layer and transformer, and then they were input into the cross-attention layer designed for interaction prediction. Cross-attention was used to create a mutual-only layer in the architecture of the neural networks, enabling the model to verify the interactions. Finally, they were concatenated and averaged over the length direction in the output layer. A multi-layer perceptron (MLP) layer had output a single prediction as a real value, known as the confidence value, from 0 to 1, whereas a true binding datum was represented as a binary value of 0 or 1. Cross-entropy was used as the loss function, and the model output was evaluated using the ROCAUC score.

The CDR and peptide sequences were represented by 20 amino acid residues. Positional embedding and padding tokens were also added to the sequences. Padding was performed so that the lengths of each TCR sequence aligned with the maximum sequence length in the training data. The same was performed for the peptide side.

The cross-attention layer is a modified model of transformer attention. In particular, it takes two amino acid sequences as input values and allows meaningful information to be extracted from the entire information of the other sequence based on the entire information of one sequence, which means it is beneficial for interaction predictions.


The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

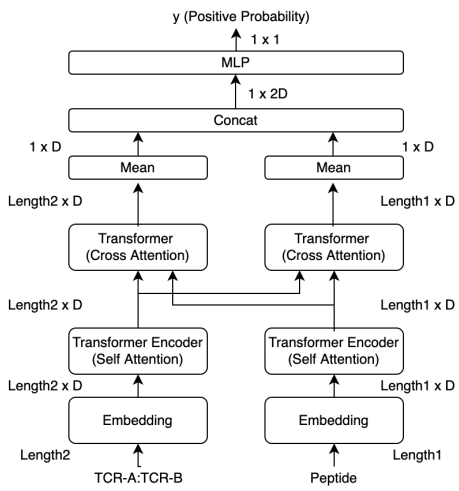


Fig. 1: Overview of our Cross-TCR-Interpreter model. Data tensor sizes are denoted.

The attention layer is specified by the following Equation 1.

$$Attention(Q, K, V) = Softmax(QK^T/d)V \tag{1}$$

In Equation 1 for the attention layer, *Q, K*, and *V* are the data matrices, and *d* is the scaling factor. $K^T$ denotes transposed matrix of $K$, where the sizes of arrays are $Q : L_1 \times D$, $K : L_2 \times D$, and $V : L_2 \times D$. $D$ is the embedding dimension. When $Q = K = V$ and $L_1 = L_2$, this is a self-attention layer.

In the cross-attention layer, $K(= V)$ and $Q$ represent two different inputs; a connected sequence of TCR$\alpha$-$\beta$ and a peptide, respectively. In addition, we defined four heads for each side in the cross-attention layer and those heads were concatenated as in a typical transformer. The softmax function defines the weights to $V$ when matrix $Q$ is 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 attrition and is used for the analysis and visualization, suggesting the residue positions that are important within the length of $V$.

Visualization and analysis of the cross-attention layer for interactions allow interpretation of the interaction across sequences (e.g. Figure 2). The cross-attention layer uses peptides as inputs and assigns specific weights to each residue of TCR to learn the important sites of TCR, and vice versa. This enabled us to analyze each side of the two areas of attentions separately.

All hyperparameters of the model are reported in Supplementary Information. Except for the hyperparameter tuning, the training was completed with one A100 GPU node at the Osaka University SQUID cluster for approximately 3 hours, and the inference was completed with a 2.6 GHz 6-Core Intel Core i7 CPU for approximately 2 hours.

## Benchmark datasets and experiment

We used three sequence-based datasets of TCR-peptide binding for model training (McPAS, VDJdb-without10x, and the entire data). McPAS [28] and VDJdb [29] are large TCR-peptide pair datasets. The validity of the cross-attention model was confirmed by comparing the test scores of the sequence data with the existing benchmark study ERGO-II [18] using McPAS and VDJdb-without10x, where the binary labels were assigned to TCR and peptide pairs. When either the TCR$\alpha$ or TCR$\beta$ sequence was missing, the interaction was removed and not used for training.

Because this study is a binary classification problem, negative label data were needed to train the model. However, as most of the TCR and peptide response data are positive-labeled, this study followed the same configuration and data of the existing study [18] that generated random TCR-peptide pairs and assigned negative labels to adjust the positive-to-negative ratio. The size of the negative data was five times larger than that of the positive data.

## The entire dataset and experiment

After confirmation of the model's performance, we trained the model again with the whole dataset (namely, the "entire data" dataset) that included McPAS, VDJdb-without10x, and VDJdb-with10x. The entire data includes 10x dataset [30] that was omitted in the ERGO-II model. The training for the entire data was performed by optimizing the validation loss and not defining the test datasets. We used 20% of the interactions as the validation dataset, whereas the renaming interactions were used for the training data. By using all the data, we attempted to incorporate as much information related to binding as possible into the model.

## Attended residue analysis with attention values on 3D structures

After applying the trained model to whole the entire dataset, we were able to acquire any attention matrix on arbitrary residues. Dividing the residues into two groups of large and small attention made it possible to analyze the attention values. For each head of TCR attention provided a peptide, we defined the residue indices of large TCR as $R_{large,h}$ in Equation 2.

$$R_{large,h} = \left\{ t | \max_p a_{t,p} > \bar{a} + \gamma \cdot \sigma \right\} \tag{2}$$

where $h$ denotes head and $a_{t,p}$ denotes an attention value of TCR residue index $t$ and peptide residue index $p$.

$$R_{large,all} = Concat_h(R_{large,h}) \tag{3}$$

Equation 2 shows the TCR side attention. Given a head $h$ of the cross-attention layer, let $A_h$ be an attention matrix of the TCR side with elements $a_{t,p}$. Please note that, by the definition in Equation 1, $\sum_t a_{t,p}$ is one dimensional all-one vector, $(1_1, 1_2, , 1_p, , 1_P)$, where $P$ is the peptide's length. We defined this as TCR-side attention because each $p$ assigns the attention to TCRs as a sum of one. $\max_p$ takes the maximum value to the peptide axis. $\bar{a}$ is the mean of the attention values of $A_h$, and $\sigma$ is the standard deviation of $A_h$. $\gamma$ is a factor that defines the large or small definition that is empirically expressed later in Result section . When computing the largely attended peptide residues, we exchanged the notation of $t$ and $p$. When computing the not-largely attended residues, we replaced the in-equation operator with smaller-than. Equation 3 shows the attended residues of the TCR side when all heads are concatenated.

After defining the largely attended residues, the results were examined using a dataset of TCR-peptide complex structures taken from the Protein Data Bank (PDB) [31]. We collected TCR-related structures from the PDB Search and the SCEptRe server [5], which gathers the complex structures of TCRs. The SCEptRe data used here in this study were downloaded on June 2, 2021. Then, the alpha and beta chains were identified with PDB headers, and Anarci [32] was used to extract the CDR portion of these 80 structures. In addition, we limited our peptide length to 25 residues, which was the maximum length of training data. When the methods were not applicable, the structure was skipped. Then, a number of structural properties were calculated, such as the closest distance to the partner molecule (e.g., the peptide when examining a TCR residue), whether the residue was engaged in an H-bond, and how many H-bonds the residue had, using BioPython [33] and LIGPLOT [34].

### Paired t-test on each PDB structure

To assess the differences between the two groups, we performed a paired student's t-test (also called the dependent t-test). The paired t-test is a statistical method used to compare the means of the two groups that are dependent on each other. In this study, 39 TCR-peptide complex structures were used as subjects and the test values were properties such as the proportion of TCR residues that were hydrogen-bonded to the peptide in the large and small attention groups, respectively.

### Covid-19 datasets and experiment

To evaluate how accurately our model would perform in a realistic situation, we applied it to prediction tasks of a real-world Covid-19 dataset generated from the Covid-19 study [35]. A virtual dataset was created using the TCRs pairs and peptides taken from the S(Spike) protein. In the original study, the peptides of the S protein were created with a 15-length residue window of amino acid residues by moving four strides of residues. We adopted the same procedure virtually to create the peptides of the S protein, by creating a 10-length residue window and moving one residue stride, as the median length of peptides in the training dataset was 10. For sequence space analysis, the sequence-sequence pairwise distance matrix was calculated by Clustal Omega software [36].

## Results

### Study overview and experiment types

We built an attention network model and named it "Cross-TCR-Interpreter", as shown in Figure 1, for predicting whether a given pair of TCR and peptide would bind. The model uses only a pair of sequences of TCRs and a peptide as inputs. Whereas previous studies have used additional features including gene, MHC, and T cell types, the available data records for those are still limited, and using these features could make the model overfitting to the existing data. Additionally, the use of only sequences without human knowledge, such as gene information, is important to provide interpretability close to nature.

We performed three experiments to validate the performance and usefulness of our proposed cross-attention model. In the first experiment, we trained and validated the model using an existing benchmark set, comparing its performance with that of a previously proposed model. In the second experiment, we retrained the model using all available data and applied it to a dataset of TCR-peptide pairs of known 3D structures, performing statistical analyses of cross-attention values to detail the biochemical binding event. Finally, in the third experiment, the model was applied to a recently published Covid-19 dataset [35] to assess its generalizability in a real-world setting.

### Benchmark experiment results

To evaluate the performance of our model, we used training and test datasets identical to those in the previous study [18], in which the authors developed the model ERGO-II. Two benchmark datasets, McPAS and VDJdb without 10x Genomics data (VDJdb-without10x), were prepared for this experiment, and the dataset statistics are listed in Table 1. In this task, the test set was a randomly chosen subset of the available TCR-peptide pairs.

Using the ROCAUC score as in the previous study, our model outperformed ERGO-II for both benchmark datasets in the sequence-feature-only setting and achieved a score comparable to that of the best model (Table 2). The features with the best scores in the Table incorporated sequences, VJ genes, and MHC-type as inputs, and the model architecture was relatively complicated in terms of the loss function and pre-training. Although there might be a possibility that other features can explain more of the TCR-peptide binding, we thought it was difficult to address those features inside the cross-attention model architecture. It was also difficult to integrate them into the interpretability. Thus, we used only the simplest feature, the sequence.

### The entire data dataset statistics and validation scores

After confirmation of the model performance, we retrained the model with a larger dataset (herein referred to as "the entire data") consisting of McPAS and the whole VDJdb including the 10x Genomics dataset [30], which was not used in the ERGO-II model. In the entire data, there were

Table 1. Dataset statistics.

| Dataset name | TCR$\alpha$ | TCR$\beta$ | Peptide | Interaction |
|---|---|---|---|---|
| The entire dataset | 17954 | 19162 | 478 | 175211 |
| McPAS, training | 2423 | 2560 | 316 | 23363 |
| McPAS, test | 721 | 716 | 190 | 5611 |
| VDJdb-without10x, training | 2151 | 2171 | 175 | 19526 |
| VDJdb-without10x, test | 573 | 576 | 120 | 4485 |

Table 2. Result of benchmark dataset.

| Model | Dataset | Features | ROCAUC Score |
|---|---|---|---|
| Our model | McPAS | Only Sequences | **0.933** |
| Our model | VDJdb-without10x | Only Sequences | **0.925** |
| ERGO-II, LSTM | McPAS | Only Sequences | 0.855 |
| ERGO-II, LSTM | VDJdb-without10x | Only Sequences | 0.800 |
| ERGO-II, LSTM | McPAS | Sequences, VJ genes and MHC Type | 0.939 |
| ERGO-II, LSTM | VDJdb-without10x | Sequences, VJ genes and MHC Type | 0.849 |

175,211 interactions. Of those, there were 17,954, 19,162, and 478 unique TCR$\alpha$, TCR$\beta$ data, and peptides, respectively (Table 1).

The 20% interactions were used for validation, resulting in 140,082 interactions for training and 35,129 for validation. The training for the entire data was conducted without defining the test datasets. This means the training parameter updates were stopped by measuring validation set scores. The model with the best validation performance was selected as the final model. After the training, the ROCAUC of the training dataset was 0.952 and that of the validation dataset was 0.825, which indicated that the model was well adapted to the entire data.

## Prediction on 3D data by the entire-data-trained model

Using the procedure described in the Methods section, we started with 80 TCR-related structures from the PDB Search and the SCEptRe server and obtained 55 structures. The 55 structures contained eight pairs with identical sequences for the CDR and peptide and therefore, the prediction was made based on the sequences of 47 structures. 39 structures out of 47 were predicted as positive interactions, and the attention layers were analyzed further in these 39 cases (See Supplementary Information); the precision score was 83% (=39/47). Attention values in the cross-attention layer were created with respect to each protein structure.

## Attention analysis and grouping

The attention values were considered "large" when they exceeded the threshold of MEAN + 5.5 STD on the peptide side and MEAN + 4.5 STD on the TCR side (4.5 and 5.5 are $\gamma$ in Equation 2). Approximately 20% of the residues were identified as large on each side, using $\gamma$ as a result of the total sum of the four heads. The thresholds were determined through empirical evaluation, and the residue count generated by changing $\gamma$ is provided in the Supplementary Information. The chosen thresholds were found to be effective in differentiating between large and small attention values. Note that the threshold for large attention values varies for each PDB entry or head due to differences in the distribution of attention values.

The analysis was performed separately for each of the four heads in the cross-attention layer (heads 0 to 3) on both the CDR and the peptide sides, with each head being analyzed separately. The cross-attention layer was defined on a CDR sequence and a peptide sequence, resulting in an attention matrix with a shape determined by the length of both peptide and CDR residues in each head. It was possible for a particular residue to have a large attention value in head 0 but not in the other heads (as seen in Equation 2).

As an example, the attention values for the TCR-peptide complex of PDB entry 5TEZ [37] are shown as heatmaps in Figure 2. Eight heatmaps are shown. The corresponding 3D structure of the TCR-peptide complex is shown in Figure 3, in which the amino acid sequences of the peptide, TCR$\alpha$, and TCR$\beta$ are `GILGFVFTL`, `CAASFIIQGAGKLVF`, and `CASSLLGGWSEAFF`, respectively.

## Statistical analysis on structural data with attention groups

Using the $\gamma$ factor of 4.5 defined in Equation 2, we classified the TCR residues into two groups based on their attention values, "large" and "small," for the cross-attention of the TCR side given a peptide. To gain insight into the characteristics of each group of residues, we analyzed their structural properties.

To assess differences between the two groups, we performed a paired t-test to remove variations arising from individual structural factors. In this study, 39 TCR-peptide complex structures were used as subjects and structural properties associated with large or small attention groups were the two measurements. The purpose of the test was to examine the null hypothesis that the mean difference between the pairs of measurements is zero. The proportion of a property (e.g. H-bonded to any peptide residue), is calculated by $P = A_h/B_h$, where $A_h$ is the number of residues with one or more H-bonds of the specified type within the residues of large attention values, and $B_h$ is the number of residues of large attention values, where $h$ denotes the head (See Figure 4).

The results of the statistical tests are shown in Table 3. Although each head was analyzed equally and separately, they showed different results

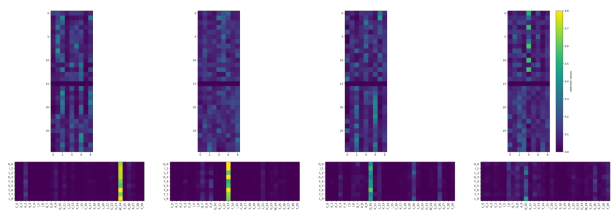

Fig. 2: Example of attention value visualization of PDB ID 5TEZ [37]. Upper half: attention values of a peptide so that the sum is 1 over the peptide given a TCR pair. The X-axis represents the residue of the peptide, while Y-axis represents the residue of TCRs. Lower half: attention values of a TCR pair so that the sum is 1 over the TCR given a peptide. The X-axis is the residue of TCRs while Y-axis is the residue of peptides. The sum over the x-axis direction is 1 for both images. Four columns denote the heads of the multi-head attention layer. Colors denote the magnitude of attention value: dark blue represents smaller attention, yellow represents larger, and green is in the middle. Basically, residues with large attention are highlighted in bright colors. In the lower figure, the cell corresponding to peptide position $L_8$ (the last row) and TCR$\beta$ position $W_{24}$ (sixth column from the right) represents the weight of how important the TCR $W_{24}$ is given the peptide $L_8$. Its color is bright yellow, which means that the attention value is large (when the values are larger than the MEAN + $\gamma$ STD threshold for each PDB ID and each head) and the two residues might play a potentially biologically important role during predictions.

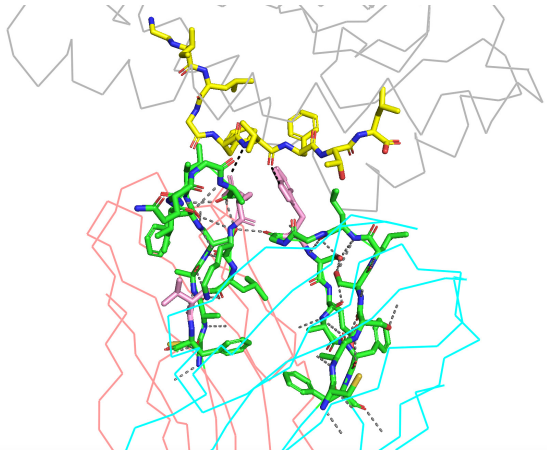

Fig. 3: PyMol visualization of large attention residues on PDB 5TEZ. Residues are displayed with solid lines and bonds are displayed by dotted lines. Green thick line: CDR residues of TCR$\alpha$ or TCR$\beta$. Pink thick line: Large attention CDR residues. Three amino acid residues, $W_{24}$ of TCR$\beta$ and $Q_{10}$ and $V_{13}$ of TCR$\alpha$, were identified as attentions with large values. Yellow thick line: Peptide residues derived from influenza A viruses. Orange line: TCR$\alpha$. Sky Blue line: TCR$\beta$. Grey line: MHC residues. Grey dotted line: polar contact between CDR and TCR. Black dotted line: polar contact between CDR and peptide.

(see Supplementary Information). Head 0 happened to show the most significant differences, thus the results of head 0 and all concatenated heads are shown in Table 3. The results of the other heads are reported in the Supplementary Information.

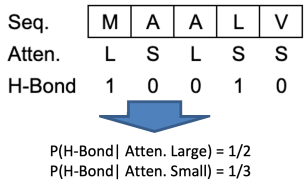


Fig. 4: Example of the calculation of proportion for a residue to have an H-bond given the large attention value in a single protein structure. Given a single PDB structure, the number of H-bonds per residue, proportion, is 1/2 for the large attention group and 1/3 for the small attention group.

Because a TCR sequence includes both CDR and non-CDR portions, the H-bond properties were measured by dividing the residues into CDR and non-CDR portions. Whereas the proportion of residues that are H-bonded to any TCR residue (i.e., H-bonds within the TCR chains) showed no difference between the large and small attention groups, the residues with large attention values had a more significant proportion of having an H-bond with the CDR portions. This significant difference was observed both in head 0 and in all heads concatenated. A natural consequence of those observations is that the largely attended residues are less likely to be H-bonded to the non-CDR portions, compared to the residues with small attention values. The most significant property difference in all concatenated heads occurred in the proportion of H-bonded to any non-CDR TCR residue. This means that the largely attended residues are highly likely to avoid the H-bonded to the non-CDR TCR part.

In contrast, contrary to expectations, the proportion of largely attended TCR residues to form an H-bond with any peptide residues was not significant in all heads. Also, it is statistically significantly smaller than that for small-attended residues in head 0 with p-value $0.043$. This highlighted a surprising and counter-intuitive finding in our analysis. We also examined the closest distance from a given TCR residue to any peptide residue, however, no significant difference was observed.

We also performed a similar analysis on the peptide side (Table 4), and observed that amino acid residues with large attention values had significantly smaller distances to the closest TCR residues. This property was not observed in the TCR side attention and it poses an interesting structural aspect.

## Covid-19 prediction experiment result

We applied our predictive model to a recently published Covid-19 dataset [35], evaluating its efficacy in predicting TCR-peptide interactions in a real-life setting. As described in the Methods section, peptides from each SARS-CoV-2 protein were created with a 10-length residue window by moving one stride. Ten unique TCR alpha-beta pairs were confirmed to be interacting with the S protein by wet experiments. The total number of interactions was 2,118,464, of which 2,118,434 were negative interactions and only 30 interactions were positive. In the 2,118,464 interactions, there were 1676 unique TCR alpha-beta pairs and 1264 unique peptides. In the

Table 3. TCR-side attention analysis. Structural property comparisons between the large and small attention residue groups are shown.

| 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 CDR residue | 0.6068 ± 0.4262 | 0.4135 ± 0.1061 | 0.0111 | 0 | *** |
| H-bonded to any non-CDR TCR residue | 0.1966 ± 0.3734 | 0.4479 ± 0.0738 | 0.000482 | 0 | *** |
| H-bonded to any TCR residue | 0.7308 ± 0.3897 | 0.7193 ± 0.0903 | 0.864 | 0 | |
| H-bonded to any CDR residue of own chain | 0.6068 ± 0.4262 | 0.3792 ± 0.0991 | 0.0027 | 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 CDR residue of opposite chain | 0.0000 ± 0.0000 | 0.0642 ± 0.0673 | 8.37e-07 | 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 | 0.0862±0.1368 | 0.0805±0.0675 | 0.828 | all | |
| H-bonded to any CDR residue | 0.4846±0.2216 | 0.4103±0.1040 | 0.0478 | all | *** |
| H-bonded to any non-CDR TCR residue | 0.2940±0.1923 | 0.4672±0.0846 | 3.88e-05 | all | *** |
| H-bonded to any TCR residue | 0.6845±0.1650 | 0.7294±0.0880 | 0.0987 | all | |
| H-bonded to any CDR residue of own chain | 0.4643±0.2180 | 0.3752±0.0922 | 0.0107 | 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 CDR residue of opposite chain | 0.0306±0.0857 | 0.0672±0.0743 | 0.0369 | all | *** |
| In the edge [2] | 0.6434±0.2064 | 0.5928±0.0570 | 0.218 | all | |
| Closest distance to peptide (Å) | 8.4072±2.2892 | 8.4122±0.9592 | 0.988 | all | |
| Number of H-bonds formed | 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. In the last two properties, per-residue averages were used instead.

2. Four residues from the beginning and four from the end of the CDR.

Table 4. Peptide side attention analysis. Structural property comparisons between the large and small attention residue groups are shown. Only head 3 results and all head results are shown since head 3 happened to be the most significant result.

| Property | Large Attention [1] | Small Attention [1] | p-value | Head | |
|---|---|---|---|---|---|
| H-bonded to any peptide residue | 0.1200±0.3250 | 0.0620±0.1108 | 0.31 | 3 | |
| H-bonded to any CDR residue | 0.2800±0.4490 | 0.1714±0.0963 | 0.196 | 3 | |
| H-bonded to any TCR residue | 0.4000±0.4899 | 0.2491±0.1120 | 0.0866 | 3 | |
| In the edge [2] | 0.2600±0.4271 | 0.5771±0.1191 | 0.00591 | 3 | *** |
| 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 peptide residue | 0.0495±0.1443 | 0.0659±0.1206 | 0.458 | all | |
| H-bonded to any CDR 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 | |
| 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. In the last two properties, per-residue averages were used instead.

2. Three residues from the beginning and four from the end of the peptide.

positive 30 interactions, there were 10 unique TCR alpha-beta pairs and 18 unique peptides.

By maximizing the precision of this prediction task, we adjusted a threshold to 0.9569, achieving a precision of $3.303\dot{1}0^{-5}$, and a recall of 0.2333. Regarding a confusion matrix, the True Positive count was 7, the False Negative count was 23, the False Positive count was 213,151, and the True Negative count was 1,905,283. The ROCAUC score was 0.496. Given the fact that positive interactions exist at a rate of $1.416\dot{1}0^{-5}$ ($= 30/2118464$), we can claim that the model can detect positive interactions 2.33 times ($= 3.303/1.416$) better than the random selection, but its specificity was not adequate enough to replace wet experiments.

## Discussion

### Interpretation of prediction in protein sequences

Prediction results in machine learning are generally difficult to interpret, but when a model is used for binding predictions of biological sequences, the interpretability of the neural network model is essential. Highlighting residue positions is useful for understanding model predictions and is critical for utilizing later-stage applications. The cross-attention layers can enable this interpretation.

Unexpectedly, the TCR residues with large attention values were not found to interact directly with peptide residues (Table 3). Instead, these

residues appeared to stabilize a specific loop conformation required for the peptide binding, by forming H-bonds within the CDRs. Accordingly, researchers have observed that residues that comprise the H-bond network within the TCR may be evolutionarily conserved [38, 39] and that the internal organization of the interface plays an important role in protein-protein interactions [40, 41]. Our findings suggest that certain residues may be oriented in a specific direction with internal H-bonds and that the attention layer may emphasize their importance in TCRs in terms of binding stability.

Our findings also showed that the average distance between the TCR and peptide in 3D decreased when the attention value on the peptide side of cross-attention was large. This may be because the peptide, being a short sequence, has a limited contribution to TCR binding that is related solely to distance. However, the larger attention values on the TCR side did not necessarily correspond to smaller distances in the 3D structure, potentially because TCRs are longer and more complicated in their binding role.

Not all the sequence paired data were available with 3D structure and we knew that the number was small, but we experimented with as many available structures as possible. We used the 3D structure as the confirmation of the prediction and attention layer interpretation. In future investigations, it may be possible to use a different machine learning model such as a meaningful perturbation method on exhaustively collected sequences.

## Architectural design for interactions and interpretations

The cross-attention layer should be used to understand the interaction of two sequences especially to understand what occurs on one data when interacting with the other, focusing solely on the mutual interaction. The prediction of protein-protein interactions from sequence information is non-trivial. Even if there are two inputs to the model, the model cannot address the binding as a relation between the two datasets: the concatenation of the two datasets is not enough to explain the binding. Even though the BERT model uses two sequences, the interaction is not specifically limited to mutual interaction. Figure 5 shows the attention layer visualization of one head of one layer in a scenario where the BERT would be trained on TCRs and peptides. From the definition of Equation 1, the sum along the x-axis is one. This means that the BERT attention shows both self-attention and cross-attention and that it is difficult to relate this matrix to the binding mechanism. Therefore, the attention interaction between TCR and peptide should be intentionally limited and interpreted by the cross-attention layer.

## Difficulties associated with Covid-19 experiment

The score of the Covid-19 experiment was not prominent; this likely occurred because there were too many negative interactions compared to the baseline test set, and because the distribution of peptides was different from that of the baseline test set.

When the Covid-19 dataset was modified to a positive ratio of 20%, which is the same ratio as the baseline dataset, the model output the

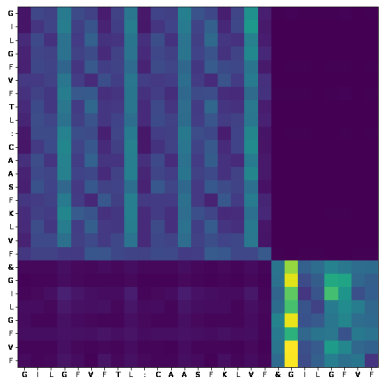

Fig. 5: The attention matrix of BERT as trained on TCR-peptides. The colon is the concatenation token for alpha and beta chains, and the "&" is the separation token for peptide. The attention values are not limited to the mutual mechanism, which makes the interpretation more difficult. Additionally, in a normal powerful BERT architecture, excessive attention heads exist and excessive transformer layers are stacked, which made it almost unattainable to relate attention values with the binding mechanism.

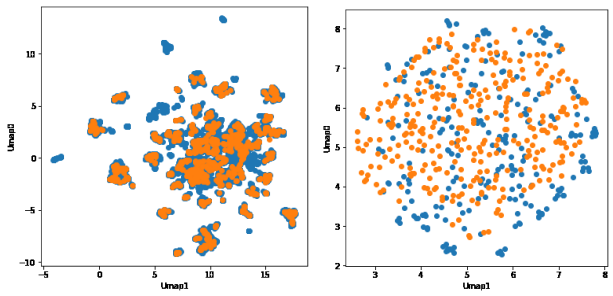


Fig. 6: UMAP visualizations of sequence distance maps of TCRs and peptides. Left: TCR sequences visualization. Right: Peptide sequences visualization. Orange points: Covid19 data. Blue points: Baseline data.

ROCAUC value of 0.5692. When we set the threshold to 0.9977, the precision was 0.6667 and the recall was 0.06667. Clearly, the positive ratio affected the performance of the model evaluation.

In addition, the difficulty of prediction in real Covid-19 data was not caused by differences in TCR, but instead by differences in peptides between the Covid-19 data and baseline data. We plotted the sequence-sequence pairwise distance matrix with UMAP dimension reduction, as shown in Figure 6. There was no difference in the distribution of TCR between the test set data and the Covid-19 data, whereas there was a substantial difference in the distribution of peptides. This discussion is also supported by previous studies on TCR predictions [25, 42, 43], in which the authors stated that generalization and extrapolation to unseen epitopes remain challenging.

Additionally, we sought to evaluate the performance of our model specifically on interactions involving unseen peptides within the benchmark test set of McPAS and VDJDB data, by removing the interactions of peptides in the training dataset. Although the majority of the

**Key Points**

- Our study presents a groundbreaking computational approach for predicting T cell receptor (TCR) binding to specific ligand peptides. Our study predicted TCR-peptide binding with the cross-attention mechanism as well as analyzed the available protein structures comprehensively to gain new insights into TCR-peptide functional relationships.
- By incorporating an attention layer based on language models and structural binding mechanisms, our machine learning model achieved state-of-the-art performance on a benchmark dataset of TCR-peptide binding.
- Our analysis of the model's predictions allowed us to associate neural network weights with protein 3D structure datasets, identify statistically significant properties of largely attended residues, and detail the binding principle through the visualization and analysis of the cross-attention layer. This was made possible because we intentionally limited the attention values specifically on the interaction subpart.
- The statistical analysis revealed that the largely attended residues in the attention layer were more likely to contact their own TCR than normal residues, thereby providing new insights into the TCR-peptide binding mechanisms. It was suggested that proteins create hydrogen bonds in order to form special structures and play special roles when a peptide is conditioned to react with a protein.

peptides were already present in the training data, we identified a subset of 46 interactions (14 positives) for McPAS and 16 interactions (8 positives) for VDJDB that involved unseen peptides. The ROCAUC scores for these unseen peptide interactions were 0.721 for McPAS and 0.719 for VDJDB, which were lower than the scores of 0.933 and 0.925 reported in Table 2 for interactions involving seen peptides.

## Funding

The computational resources of this work were supported by a research proposal based Use of Cybermedia Center, Osaka University.

## Author Contributions

Kyohei Koyama developed the model, analyzed the results and wrote the manuscript as part of his doctoral work. Kosuke Hashimoto, Chioko Nagao, and Kenji Mizuguchi provided advice and Kenji Mizuguchi managed the entire project. All authors have contributed to the manuscript and approved the submitted version.

## Conflict of Interest Statement

The authors declare no conflicts of interest.

## Data Availability

The data and source code used in this study was stored in the GitHub repository: *https://github.com/kyoheikoyama/TCRPrediction*

## References

1. Jayajit Das and Ciriyam Jayaprakash. *Systems immunology: An introduction to modeling methods for scientists*. CRC Press, 2018.
2. Yuval Elhanati, Zachary Sethna, Curtis G Callan Jr, Thierry Mora, and Aleksandra M Walczak. Predicting the spectrum of tcr repertoire sharing with a data-driven model of recombination. *Immunological reviews*, 284(1):167–179, 2018.
3. Ragul Gowthaman and Brian G Pierce. Tcr3d: The t cell receptor structural repertoire database. *Bioinformatics*, 35(24):5323–5325, 2019.
4. 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.
5. 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.
6. John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.
7. Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.
8. Mostafa Karimi, Shaowen Zhu, Yue Cao, and Yang Shen. De novo protein design for novel folds using guided conditional wasserstein generative adversarial networks. *Journal of chemical information and modeling*, 60(12):5667–5681, 2020.
9. Namrata Anand and Possu Huang. Generative modeling for protein structures. *Advances in neural information processing systems*, 31, 2018.
10. Xiaoyong Pan, Jasper Zuallaert, Xi Wang, Hong-Bin Shen, Elda Posada Campos, Denys O Marushchak, and Wesley De Neve. Toxdl: deep learning using primary structure and domain embeddings for assessing protein toxicity. *Bioinformatics*, 36(21):5159–5168, 2021.

11. Paul Morris, Yahchayil DaSilva, Evan Clark, William Edward Hahn, and Elan Barenholtz. Convolutional neural networks for predicting molecular binding affinity to hiv-1 proteins. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 220–225, 2018.
12. Vanessa Isabell Jurtz, Leon Eyrich Jessen, Amalie Kai Bentzen, Martin Closter Jespersen, Swapnil Mahajan, Randi Vita, Kamilla Kjærgaard Jensen, Paolo Marcatili, Sine Reker Hadrup, Bjoern Peters, et al. Nettcr: sequence-based prediction of tcr binding to peptide-mhc complexes using convolutional neural networks. *BioRxiv*, page 433706, 2018.
13. 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 TCR$\alpha$ and $\beta$ sequence data. *Communications Biology*, 4(1):1–13, September 2021.
14. 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.
15. 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.
16. 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.
17. 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.
18. 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, 2021.
19. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
20. 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.
21. 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. *arXiv preprint arXiv:1905.09418*, 2019.
22. 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.
23. 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.
24. 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.
25. 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.
26. Kyohei Koyama, Kotaro Kamiya, and Koki Shimada. Cross a ention dti: Drug-target interaction prediction with cross a ention module in the blind evaluation setup. *BIOKDD2020*, 2020.
27. 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.
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. 10x Genomics. A new way of exploring immunity–linking highly multiplexed antigen recognition to immune repertoire and phenotype. *Tech. rep*, 2019.
31. Helen Berman, Kim Henrick, and Haruki Nakamura. Announcing the worldwide protein data bank. *Nature Structural & Molecular Biology*, 10(12):980–980, 2003.
32. James Dunbar and Charlotte M Deane. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016.
33. Brad Chapman and Jeffrey Chang. Biopython: Python tools for computational biology. *ACM Sigbio Newsletter*, 20(2):15–19, 2000.
34. 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.
35. 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. e20211327.

36. 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.
37. 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.
38. 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.
39. 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.
40. 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.
41. 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.
42. 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.
43. 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.
44. 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.

# 1 Supplementary Information

## 1.1 Hyperparameters of the model

The parameter of the model was optimized by using a hyperparameter optimization package, Optuna [44]. The best parameters are: $batch_size$ = 2949, $d_ff$ of the final MLP layer = 84, $dim$ in the model = 256, $dropout_rate$ = 7.651e-05, $learning_rate$ = 9.387e-05.

The number of transformer layers for self-attention is two on each side of the TCR and the peptide. The $n_head$ of the transformer encoder is four. The number of cross-attention layers is one. The maximum length of TCR$\alpha$ and TCR$\beta$ in the dataset was 62, and that of the peptide was 26. Although an intensive and exhaustive search may have different results, initial searches on pre-training and transfer learning did not contribute to improving the final score.

## 1.2 Comparison of occurrence of amino acid types of residues in the large and small attention valued groups

In general, TCRs have, in order of abundance, serine(S), alanine (A), and phenylalanine (F). On the other hand, TCRs with large attention values have valine (V), glycine (G), and glutamic acid (E). As the right figure of Figure 7 shows, methionine (M), valine (V), and lysine (K) tend to have relatively large attention values. The side chains of M and V are known as polarity-free and belong to the aliphatic group.

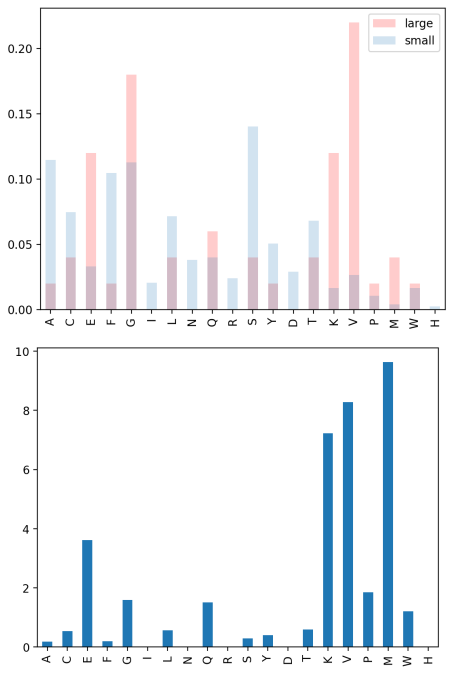


Fig. 7: Upper: Residue type distribution of large attention TCR group and small attention TCR group. Lower: the value of the large group divided by the value of the small group, indicating the occurrence tendency of large attention value.

## 1.3 Example of input perturbation by changing the input sequence

This section is to check how the prediction y-value changes when the input sequence is changed. In the 5TEZ PDB structure, there are three residues with large attention values, (Q11 of TCR$\alpha$, V14 of TCR$\alpha$, and W9 of TCR$\beta$). We checked how the predicted y-value changes when these amino acid residues are substituted with something else. The TCR$\alpha$ sequence of 5TEZ is CAASFIIQGAQKLVF and TCR$\beta$ is CASSLLGGWSEAFF. Note that Q11, V14, and W9 each make H-bonds, but only V13 has an H-bond with the main chain of the internal TCR residue.

Here, only V13 has an H-bond with other TCRs, not with the peptide, and only when V13 has changed the value of prediction turned to be a "not bound" prediction. The changes of the other Q10 do not largely deteriorate prediction (Figure 8), while the changes of W24 do change the prediction while keeping its positive (prediction to be bound) prediction with the substitution with Valine. This result can support our statement that the internal H-bonded structure of CDR might be important when binding to the peptide.

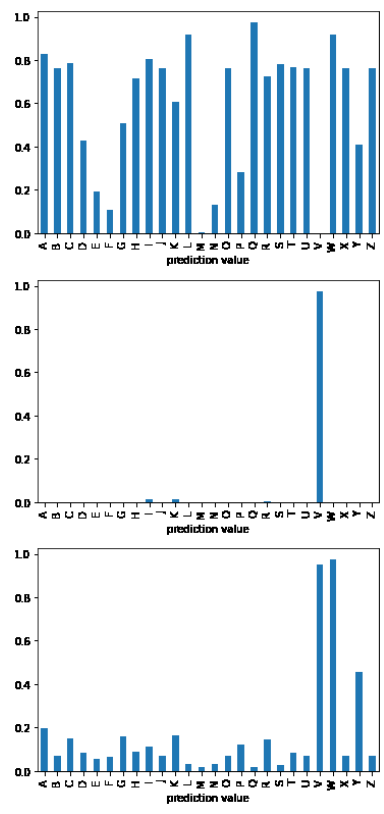


Fig. 8: The first figure shows the prediction change of Glutamine(Q) of TCR$\alpha$ from CAASFIIQGAQKLVF to CAASFIIQGA(-)KLVF. The center shows the prediction change of Valine(V) of TCR$\alpha$ change: from CAASFIIQGAQKLVF to CAASFIIQGAQKL(-)F. The last shows the prediction change of Tryptophan(W) of TCR$\beta$: from CASSLLGGWSEAFF to CASSLLGG(-)SEAFF.

## 1.4 PDB IDs for TCR-peptide crystal

There are 39 PDB ids for the analysis out of the 80 TCR-peptide crystal structures of PDB.

- 39 PDB IDs after post-processing:

- 2VLK, 5WKF, 3PQY, 4MJI, 4P2Q, 2YPL, 1J8H, 4P2R, 5MEN, 3MV8, 4OZF, 3VXR, 3VXS, 4OZG, 5TEZ, 2J8U, 6Q3S, 4JRX, 3VXU, 1U3H, 4JRY, 4Z7V, 4JFE, 4JFD, 3QIU, 2Z31, 2BNR, 3MBE, 4OZH, 2NX5, 5NHT, 4QOK, 5D2L, 1D9K, 4P2O, 5WKH, 6EQB, 2VLR, 6EQA

- 80 PDB IDs before post-processing:

- 2P5E, 2P5W, 2PYE, 5HHO, 5E9D, 4JFD, 4JFF, 4JFE, 4JRX, 4JRY, 3QEQ, 3QDM, 4FTV, 2F54, 2BNQ, 2BNR, 2F53, 1MI5, 3GSN, 5HHM, 2VLJ, 2VLK, 2VLR, 1OGA, 5NMG, 5NME, 5NMF, 4MJI, 2AK4, 3MV8, 4PRI, 4PRH, 3MV9, 3MV7, 4PRP, 5MEN, 2NX5, 5D2N, 4L3E, 2J8U, 2JCC, 2UWE, 1LP9, 5BRZ, 5BS0, 3O4L, 3KPS, 3KPR, 5EUO, 1QRN, 1QSF, 3VXR, 3VXS, 3W0W, 3VXU, 3HG1, 4QOK, 5WKF, 5EU6, 6AVF, 3QDG, 3QDJ, 4QRR, 6AVG, 5ISZ

## 1.5 S Protein of Covid-19

- S Protein: 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 lpplltdemi 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

Table 5. Count of large or small attention residue groups when changing the factor $\gamma$. Four heads are merged.

| | Large Attention of TCRs | Small Attention of TCRs | Large Attention of Peptides | Small Attention of Peptides |
|---|---|---|---|---|
| **count ($\gamma$=2)** | 569 | 468 | 372 | 65 |
| **count ($\gamma$=2.5)** | 441 | 596 | 325 | 112 |
| **count ($\gamma$=3))** | 364 | 673 | 264 | 173 |
| **count ($\gamma$=3.5** | 299 | 738 | 219 | 218 |
| **count ($\gamma$=4))** | 244 | 793 | 172 | 265 |
| **count ($\gamma$=4.5** | 196 | 841 | 137 | 300 |
| **count ($\gamma$=5))** | 153 | 884 | 111 | 326 |
| **count ($\gamma$=5.5** | 118 | 919 | 81 | 356 |
| **count ($\gamma$=6))** | 97 | 940 | 48 | 389 |

Table 6. Peptide side attention analysis (All heads are shown). 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 CDR 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-CDR 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 CDR 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-CDR 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 CDR 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-CDR 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 CDR 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-CDR 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 CDR 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-CDR 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. In the last two properties, per-residue averages were used instead.

2. Three residues from the beginning and four from the end of the peptide.

Table 7. TCR side attention analysis (All heads are shown). 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 | |
|---|---|---|---|---|---|
| H-bonded to any peptide residuetide | 0.0256 ± 0.1581 | 0.0853 ± 0.0577 | 0.043 | 0 | *** |
| H-bonded to any CDR 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 CDR 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 CDR 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-CDR 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 residuetide | 0.1453 ± 0.3162 | 0.0773 ± 0.0564 | 0.2 | 1 | |
| H-bonded to any CDR 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 CDR 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 CDR 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-CDR 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 residuetide | 0.1127 ± 0.2935 | 0.0805 ± 0.0585 | 0.654 | 2 | |
| H-bonded to any CDR 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 CDR 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 CDR 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-CDR 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 residuetide | 0.0877 ± 0.2470 | 0.0818 ± 0.0572 | 0.909 | 3 | |
| H-bonded to any CDR 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 CDR 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 CDR 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-CDR 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) | 8.6919 ± 3.5172 | 8.4100 ± 1.0378 | 0.597 | 3 | |
| Number of H-bonds formed | 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 CDR 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 CDR 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 CDR 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-CDR 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) | 8.4072±2.2892 | 8.4122±0.9592 | 0.988 | all | |
| Number of H-bonds formed | 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. In the last two properties, per-residue averages were used instead.

2. Four residues from the beginning and four from the end of the CDR.