Interpretable Prediction of SARS-CoV-2 Epitope-Specific TCR Recognition Using a Pre-Trained Protein Language Model

Sunyong Yoo, Myeonghyeon Jeong, Subhin Seomun, Kiseong Kim, and Youngmahn Han

***Abstract*—The emergence of the novel coronavirus, designated as severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), has posed a significant threat to public health worldwide. There has been progress in reducing hospitalizations and deaths due to SARS-CoV-2. However, challenges stem from the emergence of SARS-CoV-2 variants, which exhibit high transmission rates, increased disease severity, and the ability to evade humoral immunity. Epitope-specific T-cell receptor (TCR) recognition is key in determining the T-cell immunogenicity for SARS-CoV-2 epitopes. Although several data-driven methods for predicting epitope-specific TCR recognition have been proposed, they remain challenging due to the enormous diversity of TCRs and the lack of available training data. Self-supervised transfer learning has recently been proven useful for extracting information from unlabeled protein sequences, increasing the predictive performance of fine-tuned models, and using a relatively small amount of training data. This study presents a deep-learning model generated by fine-tuning pre-trained protein embeddings from a large corpus of protein sequences. The fine-tuned model showed markedly high predictive performance and outperformed the recent Gaussian process-based prediction model. The output attentions captured by the deep-learning model suggested critical amino acid positions in the SARS-CoV-2 epitope-specific TCRβ sequences that are highly associated with the viral escape of T-cell immune response.**

***Index Terms*—Attention mechanism, deep learning, epitope, SARS-CoV-2, T-cell receptor.**

Manuscript received 13 June 2023; revised 16 December 2023; accepted 18 February 2024. Date of publication 21 February 2024; date of current version 5 June 2024. This work was supported by the Korea Institute of Science and Technology Information (KISTI), National Research Foundation of Korea(NRF), in part by the Korea government (MSIT) under Grant RS-2023-00217317, in part by the Korea Bio Data Station (K-BDS) with computing resources including technical support, and in part by the MSIT (Ministry of Science and ICT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program under Grant IITP-2023-RS-2022-00156287, in part by the IITP (Institute for Information & communications Technology Planning & Evaluation). *(Sunyong Yoo and Myeonghyeon Jeong are co-first authors.) (Corresponding author: Youngmahn Han.)*

Sunyong Yoo, Myeonghyeon Jeong, and Subhin Seomun are with the Department of ICT Convergence System Engineering, Chonnam National University, Gwangju 61186, South Korea (e-mail: syyoo@jnu.ac.kr; dureelee01@gmail.com; 216129@jnu.ac.kr).

Kiseong Kim is with the R&D center, BioBrain Inc., Daejeon 34013, South Korea (e-mail: ks@biobrain.kr).

Youngmahn Han is with the Supercomputing Application Research Center, Korea Institute of Science and Technology Information, Daejeon 02792, South Korea (e-mail: hans@kisti.re.kr).

This article has supplementary downloadable material available at https://doi.org/10.1109/TCBB.2024.3368046, provided by the authors.

Digital Object Identifier 10.1109/TCBB.2024.3368046

I. Introduction

THE emergence of severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) has caused high rates of transmission and many lives to be lost due to coronavirus disease 2019 (COVID-19) worldwide [1]. These outbreaks have threatened public health and socioeconomics worldwide, and efforts, such as lockdowns, quarantines, and social distancing, have been implemented to reduce the impact [2]. Since the World Health Organization (WHO) declared SARS-CoV-2 as a pandemic on March 11, 2020, there have been 676,609,955 confirmed cases and 6,881,955 deaths worldwide [3]. Over 13.3 billion vaccinations have been administered, curbing hospitalizations and deaths in COVID-19-infected populations [4]. However, the emergence of SARS-CoV-2 variants, which are associated with high transmission rates, increased disease severity, and viral escape from humoral immunity, can potentially change the profile of the outbreak and threaten public health [5]. In particular, the emergence of the novel variant designated as SARS-CoV-2 B.1.1.529 (Omicron) [6], which spreads rapidly and has been reported to significantly reduce susceptibility to the neutralizing antibodies induced by vaccination [7]. Many countries and the global scientific community are developing effective vaccines and appropriate therapies in response to these variants.

In addition to the virus-neutralizing antibodies produced by B-cells, cytotoxic $CD8^+$ T-cells and helper $CD4^+$ T-cells are essential for viral clearance. T-cells circulating in the blood are the first in the adaptive immune system to respond to a virus: they detect infected cells and mount an immune response or directly clear the infected cells, often before symptoms appear [8], [9], [10]. Therefore, developing effective SARS-CoV-2 vaccines depends on identifying T-cell epitopes that can induce T-cell immune responses.

Peptide-major histocompatibility complexes (MHCs) on the cell surface are recognized by T-cells by a dimeric surface protein, the T-cell receptor (TCR), consequently leading to T-cell activation and proliferation by clonal expansion [11]. TCR recognition of a T-cell epitope is crucial for determining the immunogenicity of the epitope. TCRs are generated by genomic rearrangement of the germline TCR loci from a large collection of variable (V), diversity (D), and joining (J) gene segments. During T cell development, most TCRs are formed independently by a pair of α- and β-chains (90–95% of T cells) via V(D)J recombination of each locus. This rearrangement is estimated to

generate 10^{18} different TCRs, providing an enormous diversity of epitope-specific T-cell repertoires [12], [13]. Despite this TCR diversity, recent studies have found that TCRs recognizing a specific target epitope often share common sequence features. Glanville et al. [14] and Dash et al. [15] have shown a clear signature of the amino acid motif in the complementarity-determining region 3 (CDR3) of TCRβ and TCRα that interacts with specific peptides presented by specific MHC molecules. Furthermore, concerted data collection efforts [16], [17], [18], [19] and advances in high-throughput TCR sequencing technologies have demonstrated T-cell specificity [20], [21], allowing the development of data-driven models for predicting epitope-specific TCR recognition [22]. Several methods using position-specific scoring matrices [14], Gaussian processes [23], random forests [24], convolutional neural networks [25], deep generative models [26], [27], and natural language process (NLP)-based deep learning models [28] have been proposed. However, increasing the predictive power of a machine-learning (or deep learning) model remains challenging because of the scarcity of training data. As of October 2019, the VDJdb [16] and McPAS-TCR [19] databases contained about 20,000 and 55,000 epitope-specific TCR sequences, respectively.

Recent advances in NLP have demonstrated that self-supervised learning can be a powerful tool for extracting useful information from unlabeled sequence data [29], [30], [31]. One successful approach, Bidirectional Encoder Representations from Transformers (BERT), is a language model pre-trained using a huge amount of unlabeled text data via two self-supervised tasks: masked token prediction and next sentence prediction [29]. BERT models, fine-tuned using a small number of datasets, have shown ground-breaking results in 11 NLP downstream tasks. The self-supervised transfer learning strategy constructs the final model by fine-tuning the self-supervised pre-trained model from a large amount of unlabeled data, using a small amount of labeled data in the downstream task. This strategy help increase the predictive power of a deep learning model when there is scarce training data. Self-supervised transfer learning has been demonstrated to help learn protein sequence patterns [32], [33], [34]. The Tasks Assessing Protein Embeddings (TAPE) [34] model was pre-trained on 31 million unlabeled protein sequences derived from the Pfam database [35] via two protein-specific self-supervised tasks: amino acid contact prediction and remote homology detection. The TAPE pre-trained model helps improve the predictive performance in supervised downstream tasks such as secondary structure prediction, amino acid contact prediction, remote homology detection, fluorescence landscape prediction, and protein stability landscape prediction. BERTMHC, a deep learning model generated by fine-tuning the pre-trained TAPE model, has shown reliable performance in predicting peptide-MHC-II binding and presentation [36].

Many sequence-based methods for modeling epitope-specific TCR recognition have used a multiple sequence alignment (MSA) of the TCR sequences to identify position-specific amino acid motifs. This makes it difficult to find the critical amino acid positions in the epitope and the TCR sequence, which are highly relevant in TCR recognition [14], [15], [26], [27], [28], [37]. A recent study of protein language models has shown that the output attentions of BERT-based protein models can capture biologically relevant protein properties [38]. An attention-based deep learning model for peptide-MHC-I binding predictions has shown that the attentions learned by the predictive model can capture critical amino acid positions of the peptides, which help stabilize the peptide-MHC-I bindings [39].

This study presents a BERT-based model employing self-supervised transfer learning for predicting SARS-CoV-2 T-cell epitope-specific TCR recognition. The predictive model was generated by fine-tuning the pre-trained TAPE model using epitope-specific TCR CDR3β sequence datasets. The fine-tuned model showed markedly high predictive performance for two independent evaluation datasets containing SARS-CoV-2 epitope-specific TCRβ sequences and outperformed the recent Gaussian process-based prediction model. In particular, the model found critical amino acid positions in the epitope and CDR3β sequences, which potentially contribute to the TCR recognition of an epitope, and these amino acid positions can be captured using the output attention weights of the model. The findings of this study will provide new frameworks for constructing a reliable model for predicting the immunogenic T-cell epitopes using limited training data and help accelerate the development of an effective vaccine in response to SARS-CoV-2 variants by identifying potential amino acid motifs that are highly relevant to the epitope-specific TCR recognition.

II. Materials and Methods

A. Training Process and Model Architecture

Fig. 1 is a schematic representation of the training process of the proposed model. The initial model was cloned from the pre-trained BERT-based TAPE model, with an added classification layer at the end. Recent NLP research has been concerned about biases of pre-trained models and whether it actually affects the downstream task [40], [41]. To minimize the pre-trained bias and increase the performance of SARS-CoV-2 epitope specific CDR3β prediction, we performed fine tuning on the model. First, the initial TAPE model was fine-tuned using general epitope-specific CDR3β sequence data while freezing the embedding layer and the top two encoding layers. Next, the final model was fine-tuned using SARS-CoV-2 epitope-specific CDR3β sequence data derived from the Immune Epitope Database (IEDB) while freezing the embedding layer and the top six encoding layers. For each round of fine-tuning, the decision to freeze embedding layers and specific encoding layers referred to previous studies of the transformer model [42], [43], [44], [45]. This strategy is crucial for retaining essential pre-trained knowledge while allowing the model to adapt to the downstream task of our dataset. Furthermore, it prevents overfitting and maintain a stable base for the model. Fig. 2 shows the proposed model architecture. Input amino acid sequences concatenated by epitope and CDR3β sequences were first encoded into tokens using a tokenizer, where each token was an integer code for a single amino acid. Each token was then embedded into a 768-dimensional vector in the pre-trained TAPE model based on the BERT model, which has 12 encoding layers with 12 self-attention heads in each layer. The TAPE model was pre-trained using 31 million unlabeled protein sequences via next-token prediction and bidirectional masked-token prediction tasks and

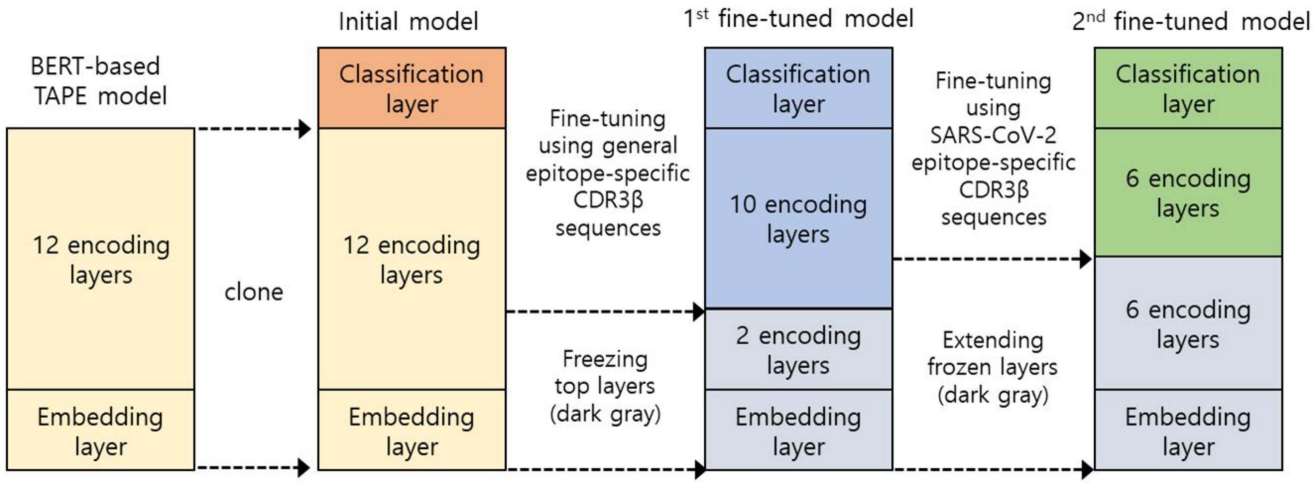

Fig. 1. Training process for the proposed model. The initial model was cloned from the pre-trained TAPE model, with an added classification layer at the end. The pre-trained model was fine-tuned in two rounds in a progressively specialized manner while extending the frozen layers between the rounds.

then underwent further supervised training via protein-specific tasks, contact prediction, and remote homology detection. The output of the pre-trained TAPE model was the hidden states of the [CLS] token. This [CLS] token represents the entire input sequence and can be TCR-epitope pairs bind or not. This classification head is a used in classification tasks. This is used as input to the classification head, which is a multi-layer perceptron consisting of a single dense layer and an output layer. The number of nodes in a dense layer is 512, and the output layer predicts whether crucial element as it transforms the rich and contextualized embeddings generated by the pre-trained TAPE model into specific classification task. The parameters are updated along with the pre-trained parameters during fine-tuning.

B. Datasets

1) Fine-Tuning Datasets: For the first fine-tuning round, the positive dataset containing epitope-specific TCR CDR3β sequences was compiled in May 2021 from three data sources: Dash et al. [15], which provided the epitope-specific paired TCRα and TCRβ chains for three human epitopes and seven mouse epitopes, and two manually curated databases, which provided the pathology-associated TCR sequences: VDJdb [16] and McPAS-TCR [19]. All the VDJdb entries had confidence scores: 0, critical information missing; 1, medium confidence; 2, high confidence; 3, very high confidence. The VDJdb entries with a confidence score of at least 1 were selected. For the second fine-tuning round, SARS-CoV-2 T-cell epitope-specific CDR3β sequence data were obtained from the Immune Epitope Database in June 2021 [46]. After selecting the epitopes with at least 20 CDR3β sequences and removing the duplicates with the same combination of epitope and CDR3β sequences from each of the fine-tuning datasets, the datasets for the first and second fine-tuning rounds contained 12,569 positive data points covering 78 epitopes and 49,282 positive data points covering 145 epitopes, respectively. The integration of epitope or amino acid sequences from multiple sources that are not completely redundant but have high similarity could lead to bias in the dataset. To confirm this potential problem, we assessed the consistency of the datasets by calculating average similarity between samples. For this, we used the Needleman-Wunsch algorithm, a common method that can calculate the similarity score (S_{nw}) between the amino acid sequences of two proteins [47]. The result indicated that the average similarity between samples with concatenated epitope and CDR3β sequences did not increase before ($S_{nw} = 0.4258$-0.4931) and after integrating multiple sources ($S_{nw} = 0.4286$). Also, the average similarity of samples with only epitope sequences did not increase before ($S_{nw} = 0.2396$-0.2701) and after integrating multiple sources ($S_{nw} = 0.2411$). Therefore, we confirmed that there are very few potential sequence redundancies that are not filtered during the data deduplication process.

In our dataset, there were only positive samples of epitope-specific TCR CDR3β sequences, and no corresponding negative samples. To increase the specificity of the model, it was necessary to add negative samples. While it is ideal to use data with experimentally validated non-binding TCR-epitope pairs as negatives, most experimentally validated information is for binding TCR-epitope pairs. To complement this, previous studies have created a silver-standard for the negative set in two ways; 1) TCRs and epitopes included in the positive dataset are randomly paired; and 2) randomly pairing TCRs from an ambient set of TCRs obtained through high-throughput sequencing of the human immune repertoire [23], [25], [48], [49], [50]. In this study, we used TCR sequences obtained by high-throughput sequencing from the blood of healthy donors [51]. First, we constructed a background dataset consisting of CDR3β sequences obtained by Howie et al., who collected blood from two healthy donors. Then, we randomly sampled CDR3β sequences from the background dataset that were not positive samples of epitope-specific TCR CDR3β sequences in the training dataset. Negative samples were generated in the same proportion as positive samples for each epitope-specific TCR CDR3β sequence. Summarizes of the final epitope-specific

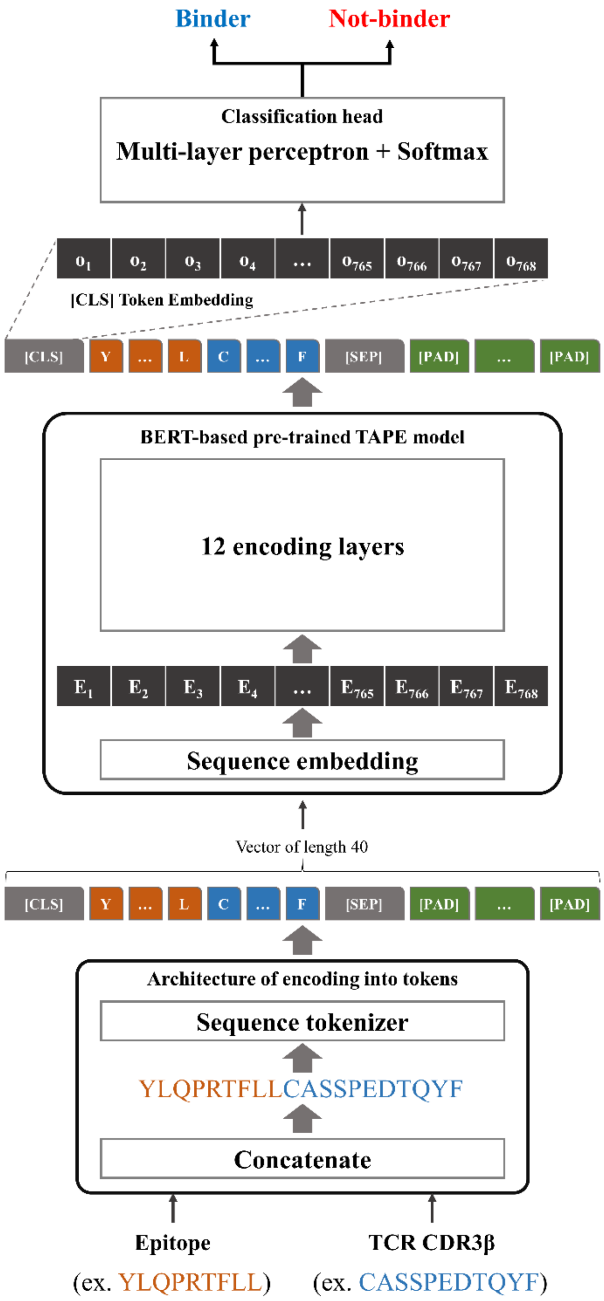

Fig. 2. Proposed model architecture. Input amino acid sequences concatenated by epitope and CDR3β sequences were first encoded into tokens using a tokenizer. Each token was then embedded into a 768-dimensional vector in the pre-trained TAPE model, which has 12 encoding layers with 12 self-attention heads in each layer. The classification head, a 2-layer feed-forward network, was then used to predict either binder or not from the output of the TAPE model.

CDR3β sequence data for each fine-tuning dataset are provided in the Appendix-I, available in the online supplemental material, of this paper.

2) Evaluation Datasets: The final model was evaluated using two independent datasets. The first dataset contained 305 SARS-CoV-2 **S**-protein$_{269\text{-}277}$ T-cell epitope (**YLQPRTFLL**)-specific TCRβs from a recent study by Shomuradova et al. (hereafter referred to as the Shomuradova dataset) with the same number of negative data points [37]. The second dataset (hereafter referred to as the ImmuneCODE dataset) contained 390 **YLQPRTFLL**-specific TCRβs from the ImmuneRACE study launched on June 10, 2020, by Adaptive Biotechnologies and Microsoft (https://immunerace.adaptivebiotech.com) with 328 negative data points are provided in the Appendix-II, available in the online supplemental material, of this paper. Furthermore, we examined whether the training dataset would have potential sequence redundancies with the evaluation dataset. The average similarity between the epitopes utilized for the first round of fine-tuning and those for the second round, both compared to the **YLQPRTFLL** epitopes, was 0.2730 and 0.2736, respectively. These results show that the epitopes in the training dataset have a low similarity to the **YLQPRTFLL** epitope used in the evaluation.

C. Fine-Tuning and Evaluating the Model

The pre-trained model was fine-tuned in two rounds, changing the frozen layers between the rounds in a progressively specialized manner. In the first fine-tuning round the model was trained while freezing the embedding layer and the top two encoding layers, so that the weights of the layers were not updated during the training process. In the second fine-tuning round, the freezing was extended to the top six encoding layers. In each fine-tuning round, the training dataset was split into 80% training and 20% validation subsets. The training and validation was repeated up to 150 and 100 epochs for the first and second fine-tuning rounds, respectively [52]. The training and validation accuracies were measured for each epoch, and the training process was stopped early at epochs where the validation accuracy did not increase for 15 and 10 consecutive epochs in the first and second fine-tuning, respectively. Accuracy is defined as the ratio of correct results to the total number of cases and can represented by (1).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{1}$$

where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives.

For all epochs in both fine-tuning rounds, the batch size was 512, and an Adam optimizer with a learning rate of 0.0001 was used [53]. The PyTorch deep learning library (https://pytorch.org) was used to implement the model. The Python source codes and datasets are available at https://github.com/aidanbio/TCRBert. The final fine-tuned model was evaluated using the Shomuradova and ImmuneCODE datasets. First, a kernel density estimate (KDE) distribution, a non-parametric method used to predict the probability density function of a random variable based on kernels as weights, was checked [54]. The KDE distribution of the prediction scores of the model for the samples with actual labels was examined as positives or negatives. The more separation in the KDE distribution of the prediction scores of the model by the labels in the sample, the higher the classification power of the model was regarded. In addition, a fold enrichment (FE) test was performed to confirm the correlation between the predictions of the model and the actual answers. All samples in the test dataset were sorted in descending order according to the prediction scores of the model and then into bins in order. The FE score was generated, which is the ratio of positive samples in each bin. The FE score was

calculated by the formula:

$$\text{FE score} = \frac{m/n}{M/N}, \tag{2}$$

Where m is the number of positive samples in the bin, n is the number of samples in the bin, M is the total number of positive samples, and N is the total number of samples. In the Shomuradova dataset, N, indicating the total number of samples, was 610, and M, representing the total number of positive samples, amounted to 305. Similarly, in the ImmuneCODE dataset, the total sample count N was 718, with M, the count of positive samples, being 328. Then, a linear regression model was fit to the FE score, and the prediction scores were averaged for each of the generated bins to visualize and validate the correlation. In this study, the number of samples in a bin (n) was set to 10. Lastly, the predictive performance was quantified using the area under a receiver operating characteristic (AUROC) score and the area under the precision-recall curve (AUPR).

D. Interpreting Position-Specific Attention Weights

To identify the critical amino acid positions in the SARS-CoV-2 epitope (**YLQPRTFLL**) and CDR3β sequences, which potentially contribute to TCR recognition of the epitope, the output attention weights of our model were investigated for the **YLQPRTFLL**-CDR3β sequence pairs predicted as a binder in the Shomuradova and ImmuneCODE datasets. CDR3β sequences were selected with the most common lengths of 13 ($n = 159$), 16 ($n = 62$), and 11 ($n = 35$) from the Shomuradova dataset, and 13 ($n = 162$), 14 ($n = 60$), and 16 ($n = 58$) from the ImmuneCODE dataset. The output attention weights had the following dimensions: $\boldsymbol{L, N, H,}$ and $\boldsymbol{S}$. Where $\boldsymbol{L}$ was the number of encoding layers, $\boldsymbol{N}$ was the number of **YLQPRTFLL**-CDR3β sequence pairs, $\boldsymbol{H}$ was the number of attention heads, and $\boldsymbol{S}$ was the fixed length of the sequences. The attention weights were marginalized into a two-dimensional vector of CDR3β sequences and epitope pairs in each layer. The attention score $\mathbf{A}_{\boldsymbol{ij}}^{(\boldsymbol{l})}$ for the CDR3β sequences to the epitope in a given layer was calculated by the formula:

$$A_{ij}^{(l)} = \frac{\sum_k^N \sum_h^H a\,(l, n, h, i, j)}{N \times H} \tag{3}$$

Where $\boldsymbol{a(l,\ n,\ h,\ i,\ j)}$ is the attention weight, $\boldsymbol{l}$ is the given encoding layer, $\boldsymbol{i}$ is the position of the CDR3β sequences, and $\boldsymbol{j}$ is the position of the epitope. In this study, the attention weights were used to interpret the relationship between amino acids in the CDR3β sequence and those in the epitope sequence. One observation in the BERT model for NLP was that the lower layers (i.e., the encoding layers close to the input) performed the syntactic interpretation of the sentence and the correct classification for most samples [55]. In our model, the lower layers capture the surface level (i.e., dependencies between each amino acid in the sequence), while the higher layers capture the complex and abstract level (i.e., semantic similarities and abstract information between each amino acid in the sequence). The attention weights have different meanings depending on the characteristics of each layer, requiring careful interpretation for the intended purpose. Our interest is to identify which amino acids in the epitope sequence is strongly interact with amino acids in the CDR3β sequence. Therefore, we focused on the interpretation of attention weights at a lower encoding layer. However, the first and second encoding layers were frozen while the pre-trained TAPE model was fine-tuned. These layers would be more of an interpretation related to the protein embedding task rather than a syntactic interpretation of the amino acids associated with the epitope and CDR3β sequence pair. After the seventh layer are the layers that were not frozen in the second fine-tuning round. However, previous research has observed that when fine-tuning a pre-trained BERT model, the lower layers are more invariant and transferable, while the higher layers are optimized for the downstream task [55], [56]. Therefore, the encoding layer after the seventh is focused on SARS-CoV-2 specific classification performance rather than the association between epitope and CDR3β sequence, which is what we are interested in. Despite this layer not specifically learning information related to SARS-CoV-2 epitopes, it can highlight the important relationship between SARS-CoV-2 epitopes and CDR3β in the test set because it learned the broader relationship between epitope and CDR3β. Consequently, we analyzed the third layer, a relatively lower layer that we expected to capture enough relevant information about the amino acid associations between epitope and CDR3β sequence.

III. Result and Discussion

A. Fine-Tuning Results

For each fine-tuning round, we determined the best number of epochs with the highest validation accuracy and examined the model's learning progress. We found that the best accuracy was 0.783 (Fig. 3(a)) at 61 epochs and 0.930 (Fig. 3(b)) at 18 epochs for the first and second fine-tuning rounds, respectively. The first fine-tuning round used a more general training dataset and more trainable encoding layers. In the first fine-tuning round, the validation accuracy was lower, and the difference between the training and validation accuracies was higher. In contrast, the second fine-tuning round used a more specific training dataset and fewer trainable encoding layers. In the second fine-tuning round, the validation accuracy was markedly high, and the difference between the training and validation accuracies was smaller. Fine-tuning the pre-trained model in this progressively specialized manner generated a final model with high predictive performance while avoiding model overfitting.

B. Evaluation Results

The final fine-tuned model was evaluated using two external test datasets containing SARS-CoV-2 epitope (**YLQPRTFLL**)-specific CDR3β sequences. The results of the KDE and the FE test performed on the Shomuradova and ImmuneCODE datasets to confirm the classification performance of the model are shown in Fig. 4. The distributions of the samples with positive and negative labels are separated in the KDE performed on the Shomuradova Fig. 4(a) and ImmuneCODE Fig. 4(b) datasets. In particular, for the positive samples, the prediction scores had

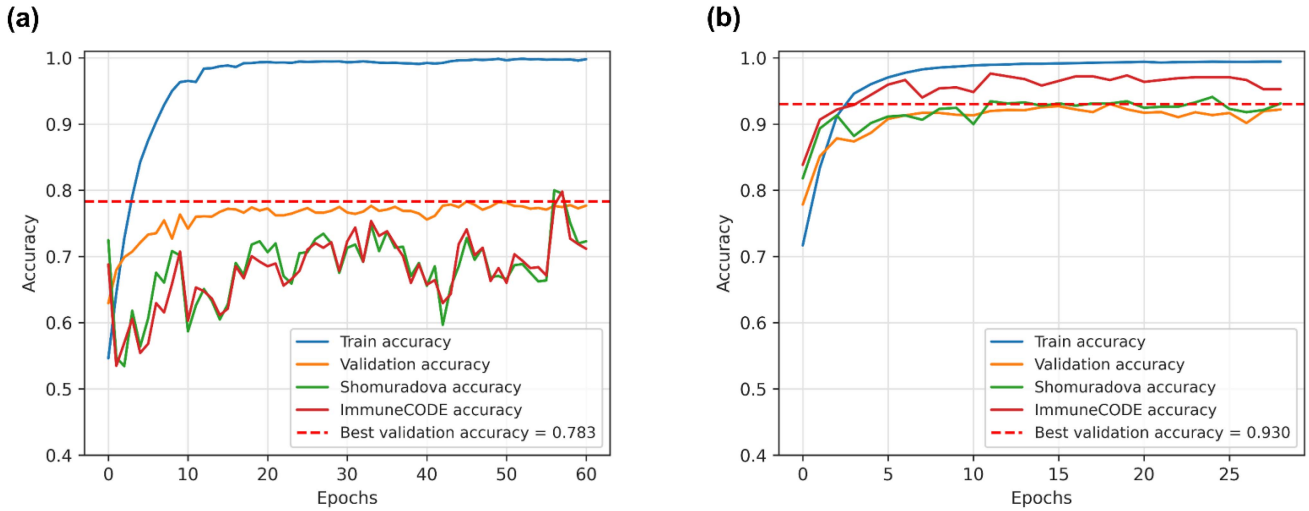

Fig. 3. Fine-tuning of the pre-trained model in two rounds. (a) Results of the training and validation accuracies by the epochs for the first round of fine-tuning the initial TAPE model using the general epitope-specific CDR3β sequence data. (b) Results of the training and validation accuracies by the epochs for the second round of fine-tuning using the SARS-CoV-2 epitope-specific CDR3β sequence data. The final validation accuracies were 0.783 at 45 epochs and 0.930 at 18 epochs for round one and round two of the fine-tuning rounds, respectively. The validation accuracy was increased, and the difference between the training and validation accuracies was reduced in the fine-tuning rounds.

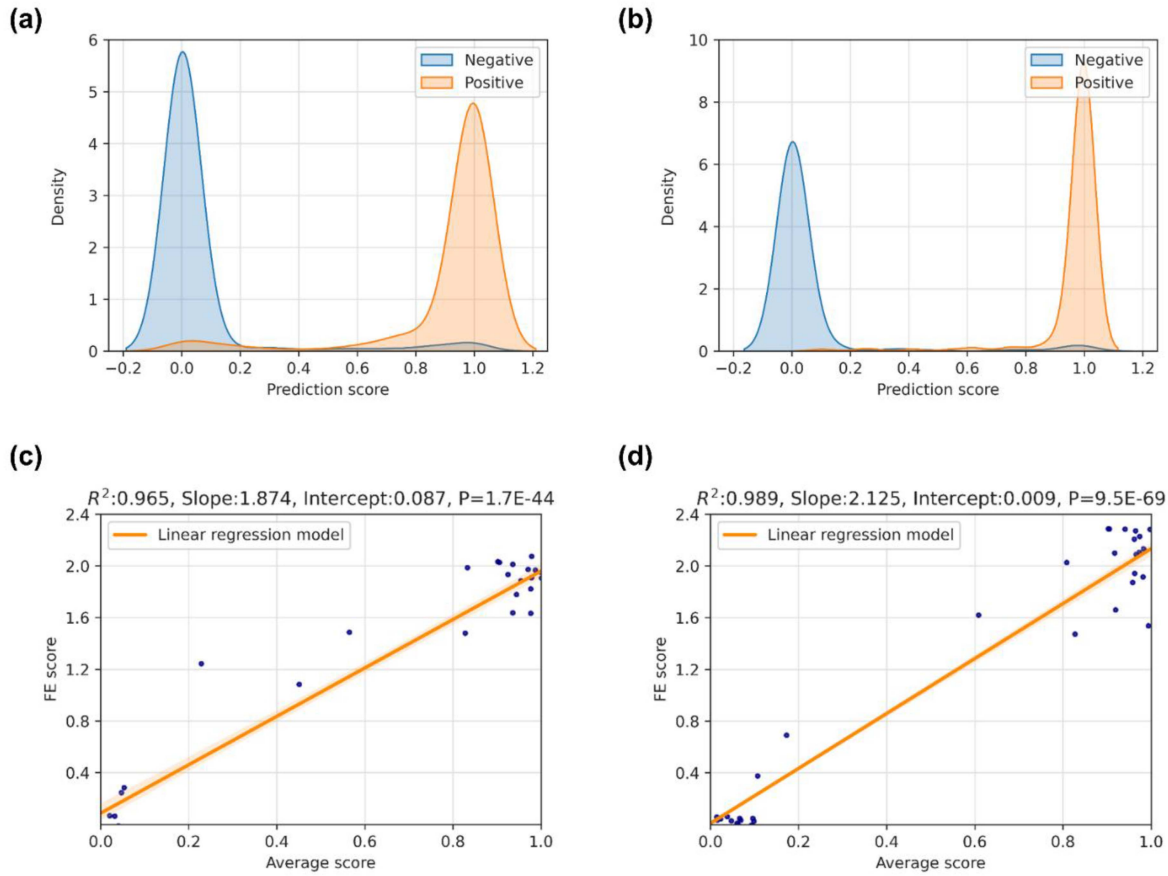

Fig. 4. Performance evaluation of the classifications for SARS-CoV-2 epitope (YLQPRTFLL)-specific CDR3β sequences. (a) The KDE distribution of the prediction scores for the positive and negative samples of the Shomuradova dataset. (b) The KDE distribution of the prediction scores for the positive and negative samples of the ImmuneCODE dataset. (c) The results of the FE test performed on the Shomuradova dataset. (d) The results of the FE test performed on the ImmuneCODE dataset. The linear regression model (orange) was used to fit the distribution of the plots (the orange outline represents the 95% confidence interval).

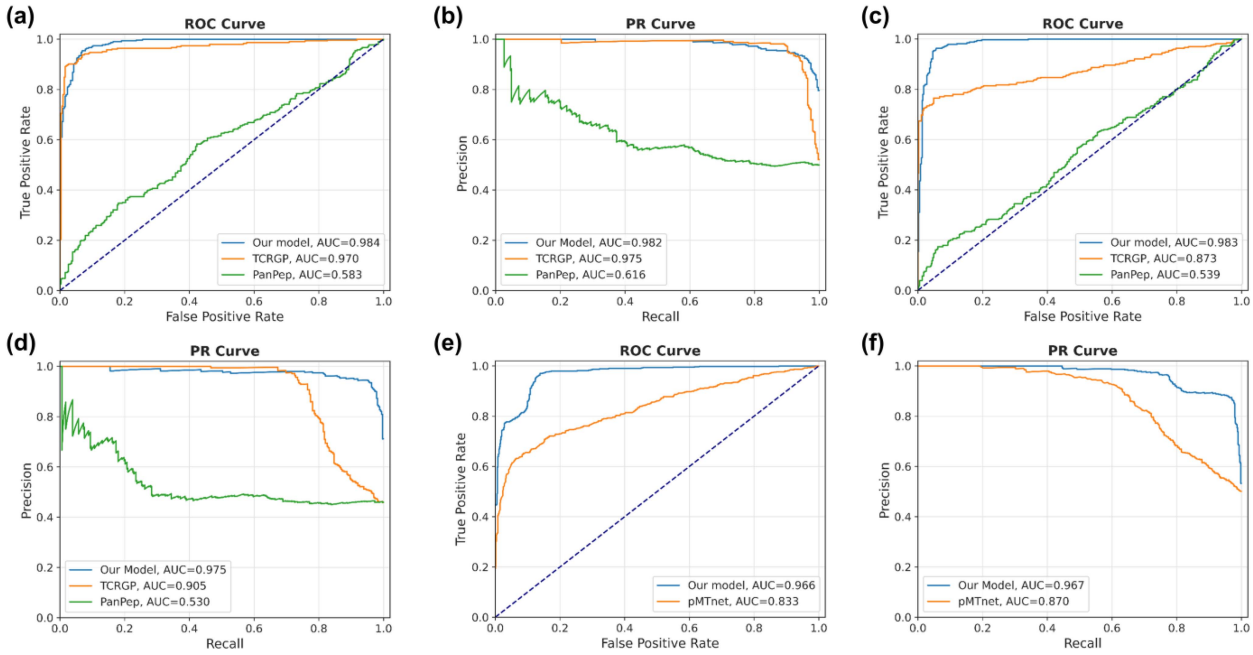

Fig. 5. ROC and PR curves for evaluating each model using each external datasets containing CDR3β sequences specific to the SARS-CoV-2 epitope. (a) ROC and (b) PR curves for each model on the Shomuradova dataset. (c) ROC and (d) PR curves for each model on the ImmuneCODE dataset. (e) ROC and (f) PR curves of our model and pMTnet for the Lu et al.'s dataset.

a high density at 1. In contrast, the negative samples had a high density at 0. Also, the higher the average of prediction scores in each bin, the higher the FE score of that bin, indicating a positive correlation between the predicted scores and correct answers. This result suggests that as the prediction score increases, the accuracy of getting the correct answer increases proportionally. In particular, the R^2 values of the linear regressions fitted to the plots were 0.965 and 0.989 for the Shomuradova Fig. 4(c) and ImmuneCODE Fig. 4(d) datasets, respectively. This result indicates that the positive correlation is highly valid and suggests that the model has significantly high classification performance.

We compared our model with TCRGP [23], a Gaussian process-based model for predicting epitope-TCR binding, and PanPep [48], which combines the concepts of meta-learning and the neural Turing machine Fig. 5(a)–(d) . Both models are known to be able to accurately predict the binding of SARS-CoV-2 epitopes to the TCR. We evaluated their prediction performance on our test sets, the Shomuradova and ImmuneCODE datasets. The TCRGP model must learn about TCR pairs that bind to the **YLQPRTFLL** epitope during the training phase in order to make predictions about the **YLQPRTFLL** epitope. Therefore, to train on the **YLQPRTFLL** epitope, we built two separate models. The model trained on Shomuradova was evaluated for prediction performance on ImmuneCODE, and the model trained on ImmuneCODE was evaluated for prediction performance on Shomuradova. In contrast, our model operates in a zero-shotsetting, having not been specifically trained on the YLQPRTFLL epitope. We further compared it with a previous study, PanPep, which is known to outperform in SARS-CoV-2 peptide-specific TCR recognition in a zero-shot setting.

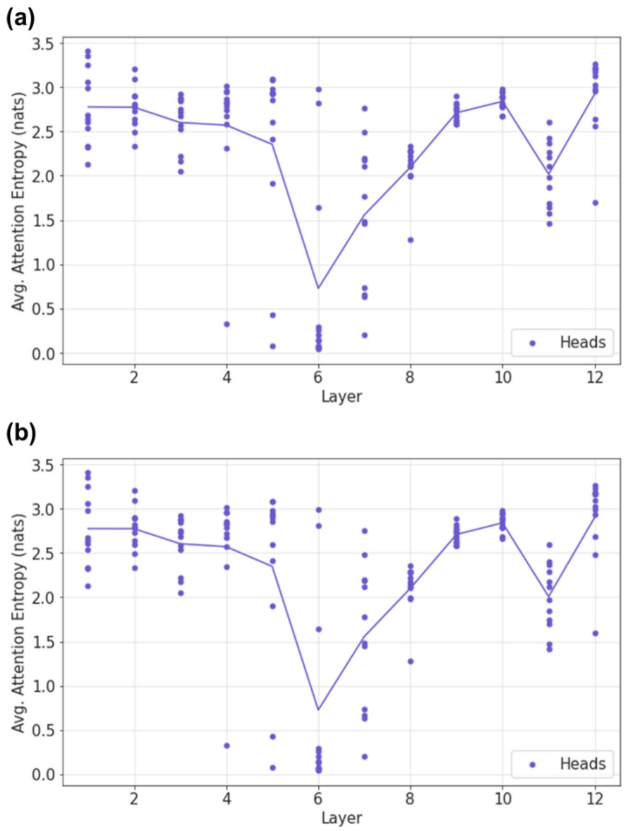

Fig. 6. (a) Average entropies of the attention distribution of the model for the Shomuradova dataset. (b) Average entropies of the attention distribution of the model for the ImmuneCODE dataset. Both datasets had relatively higher entropy of attention distributions in the low and high layers than in the mid layers.

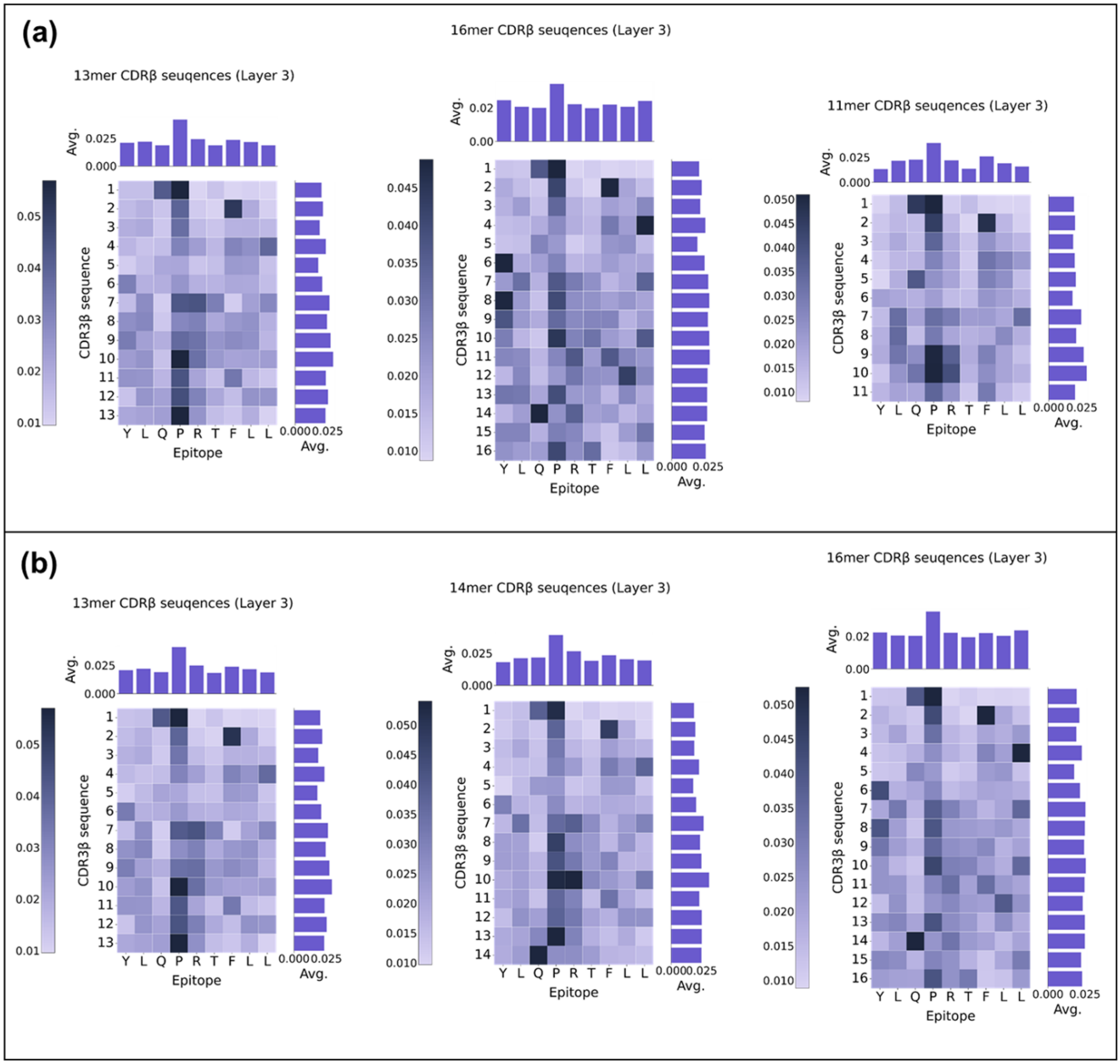

Fig. 7. Heatmap of attention weights by the position of the CDR3β sequences for YLQRPTFLL in layer three. (a) The CDR3β sequences with lengths of 13, 16, and 11 from the Shomuradova dataset. (b) The CDR3β with lengths of 13, 14, and 16 from the ImmuneCODE dataset.

The AUROC and AUPR of our model are 0.984 and 0.982 for Shomuradova, respectively, and 0.983 and 0.975 for ImmuneCODE, respectively. Additionally, our model outperformed the TCRGP and PanPep models in Shomuradova (AUROC = 0.970, 0.583, AUPR = 0.975, 0.616) and ImmuneCODE (AUROC = 0.873, 0.539, AUPR = 0.905, 0.530). We further compared the performance of our model with pMTnet, which has a high generalization ability in predicting pMHC-TCR binding based on transfer learning Fig. 5(e) and (f) [57]. pMTnet requires information about human leukocyte antigen (HLA) alleles in addition to CDR3β and peptide sequence as input. However, ImmuneCODE of our test sets does not include HLA allele information. Therefore, we compared the performance of pMTnet and our model on the independent experimental data used by Lu, et al. [57]. The AUROC of our model and pMTnet were 0.967 and 0.870, respectively, and the AUPR were 0.966 and 0.833, respectively. These results confirmed that our model outperforms pMTnet. Furthermore, although pMTnet requires HLA-allele information to predict the binding of pMHC-TCR, it is limited in its downstream applications due to limited information. In contrast, our model can predict binding with only peptide sequence and CDR3β sequence information. Consequently, we confirmed that our model had a high level of generalization ability in predicting the binding of SARS-CoV-2-specific epitopes to the TCR using a pre-trained protein language model.

C. Position-Wise Attention Weight Analysis

To identify the critical amino acid positions in the YLQPRTFLL and CDR3β sequences, the output attention weights of the model were investigated for the YLQPRTFLL-CDR3β sequence pairs that were predicted to be a binder from the Shomuradova and ImmuneCODE datasets. First, the layers containing attention weights that allowed for identifying critical amino acid positions in the YLQPRTFLL-CDR3β sequence pairs were investigated. Previous linguistic BERT models have characterized layers by identifying the attended words of each layer head. Compared to language, which is comprehended syntactically or semantically, the relationships among attended amino acids in a protein sequence have biological meanings that are not fully understood, leading to difficulties in characterizing each layer. To overcome this, methods for interpretability in NLP should be applied to protein sequence modeling [38]. In the case of linguistic BERT models, it has been observed that different layers have different characteristics of capturing the input sentences [55], [56], [58]. To apply these findings to this model, the characteristics of the layers in linguistic BERT models were compared with this model. In a previous linguistic BERT model, the interpretation for determining the tendency of an attention head was to measure the average entropy of each head's attention distribution [58]. In order to confirm the tendency in this model, the average entropy was measured for each head's attention distribution in the Shomuradova and ImmuneCODE datasets. The results are shown in Fig. 6.

The tendency of the average entropy of each head's attention in each layer was similar to the previous linguistic BERT model [58]. This result suggests that the implications of the attention weights in each model layer will be very similar to the behaviors found in linguistic NLP models. Next, the attention weights for the third layer were analyzed, which allowed for the interpretation of critical amino acid positions in the YLQPRTFLL-CDR3β sequence pair Fig. 7. For the Shomuradova dataset, the proline at position 4 (**P4**) in the epitope had a relatively high attention weight in layer three, indicating that **P4** may have a critical contribution to the TCR recognition of the epitope Fig. 7(a). A recent experimental study of SARS-CoV-2 variants found that CD8+ T-cells from a cohort of convalescent patients, comprising more than 120 different TCRs, failed to respond to the **P272L** variant corresponding to **P4** [59]. In addition, sizable populations of CD8+ T cells from individuals immunized with the currently approved COVID-19 vaccines failed to bind to the **P272L** reagent. The attention weight of the CDR3β sequences for P4 was higher at the ends than those at the central positions, indicating that the TCR amino acids at these positions may interact strongly with the proline at **P4** of the epitope, thereby substantially contributing to the TCR recognition of the epitope.

Similar attention weight patterns were observed for the ImmuneCODE dataset Fig. 7(b) for the epitope and CDR3β sequences. There were relatively high attention weights at **P4** in the epitope and the end positions in the CDR3β sequences for **P4**. These attention-based results are similar to those of MSA-based approaches, which suggests that the conserved positions are highly relevant for epitope- specific TCR recognition (The sequence logos of the MSAs for the CDR3β sequences are provided in the Supplementary Fig. 1, available in the online supplemental material, of this paper.) [60].

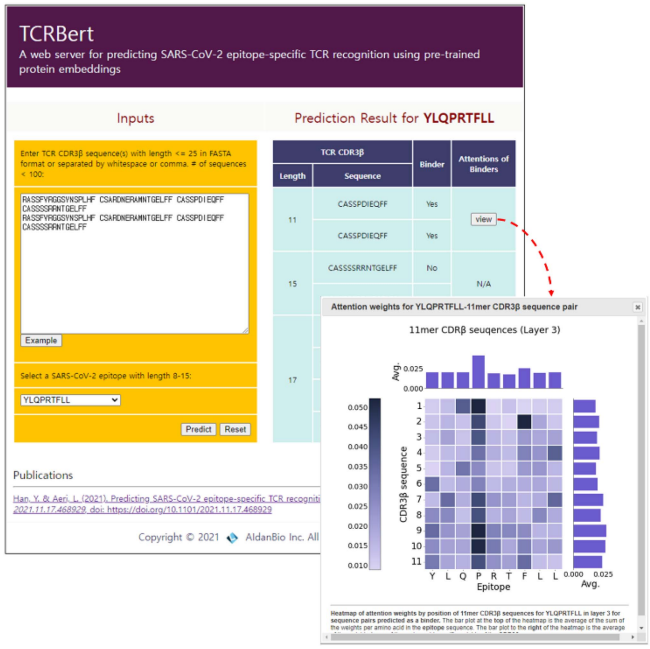

Fig. 8. TCRBert web server. The main web interface consists of the input form panel (left) and the result list panel (right). Users can submit multiple TCR CDR3β sequences and a specific epitope in the input form panel. Once the prediction process is completed, the user can see a list of the prediction results for the input CDR3β sequences grouped by sequence lengths in the result list panel. For each prediction results by CD3β sequence length, the user can also see the marginalized position-wise attention weights captured by our model for the epitope-specific CDR3β sequence pairs predicted as a binder via a pop-up panel.

D. Web Server

A web server was developed (TCRBert; http://tcrbert.aidanbio.com:5000) to provide user-friendly web interfaces for predicting SARS-CoV-2 epitope-specific TCR recognition using the predictive model. The main web interface consists of the input form panel (left) and the result list panel (right), as shown in Fig. 8. Users can submit multiple CDR3β sequences and specific epitopes in the input form panel. Once the prediction process is completed, the user will see a list of predictions for the inputted CDR3β sequences grouped by sequence lengths in the result list panel. For each prediction results by CD3β sequence length, the user can also see the marginalized position-wise attention weights captured by the model for the epitope-specific CDR3β sequence pairs predicted as a binder via a pop-up panel.

IV. Conclusion

This study developed a BERT-based model employing self-supervised transfer learning to predict SARS-CoV-2 epitope-specific TCR recognition. The predictive model was generated by fine-tuning the pre-trained TAPE model using epitope-specific TCR CDR3β sequence datasets in a progressively specialized manner. The fine-tuned model demonstrated markedly high predictive performance for two evaluation datasets containing SARS-CoV-2 **S**-protein$_{269\text{-}277}$ epitope

(**YLQPRTFLL**)-specific CDR3β sequences and outperformed the recent Gaussian process-based model, TCRGP, for the ImmuneCODE dataset. In particular, the output attention weights of this model suggest that the proline at **P4** in the epitope may contribute to TCR recognition. A recent experimental study of SARS-CoV-2 variants has demonstrated that $CD8^+$ T-cells failed to respond to the **P272L** variant corresponding to **P4**. Further, CDR3β-sequence amino acids at the end positions may contribute to the TCR recognition of the epitope at **P4**. This attention-based approach, which can capture all the motifs in the epitope and CDR3β sequences in epitope-specific TCR recognition, maybe more helpful in predicting immunogenic changes in T-cell epitopes derived from SARS-CoV-2 mutations than MSA-based approaches, which depend entirely on TCR sequences.

Our model has demonstrated high performance with interpretability in predicting SARS-CoV-2 epitope-specific TCR3β sequences. However, it has several potential limitations. First, while we fine-tuned the model using data collected from multiple sources to minimize pre-trained bias, there is still uncertainty about the persistence of pre-trained biases [61]. Therefore, more empirical research is needed to understand the potential biases introduced by pre-trained protein language models. Second, the attention weights in higher layer provided by our model also contain information, but we are uncertain what information it contains. Therefore, the attention weights of the higher layer, as provided by our model, will require additional investigation in the future to yield useful insights.

In future studies, sequence data related to the interactions between TCRα chains and MHC molecules will be integrated into the framework to predict global interaction patterns in TCR recognition of peptide-MHC complexes. These findings will provide new frameworks for constructing a reliable data-driven model for predicting the immunogenic T-cell epitopes using limited training data and help accelerate the development of an effective vaccine for the response to SARS-CoV-2 variants by identifying critical amino acid positions that are important in epitope-specific TCR recognition.

Competing interests: The authors declare that they have no competing interests.

ACKNOWLEDGMENT

The authors would like to thank W. Jeon for the helpful discussions and comments.

REFERENCES

[1] V. Chidambaram et al., "Factors associated with disease severity and mortality among patients with COVID-19: A systematic review and meta-analysis," *PLoS One*, vol. 15, no. 11, 2020, Art. no. e0241541.
[2] L. Zhou, S. K. Ayeh, V. Chidambaram, and P. C. Karakousis, "Modes of transmission of SARS-CoV-2 and evidence for preventive behavioral interventions," *BMC Infect. Dis.*, vol. 21, no. 1, pp. 1–9, 2021.
[3] S. Elbe and G. Buckland-Merrett, "Data, disease and diplomacy: GISAID's innovative contribution to global health," *Glob. Challenges*, vol. 1, no. 1, pp. 33–46, 2017.
[4] M. Biancolella et al., "COVID-19 2022 update: Transition of the pandemic to the endemic phase," *Hum. Genomic.*, vol. 16, no. 1, 2022, Art. no. 19.
[5] K. Tao et al., "The biological and clinical significance of emerging SARS-CoV-2 variants," *Nature Rev. Genet.*, vol. 22, no. 12, pp. 757–773, 2021.
[6] W. H. Organization, "Classification of Omicron (B. 1.1. 529): SARS-CoV-2 variant of concern. 2021," 2022. [Online]. Available: https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern
[7] C. Jung et al., "Omicron: What makes the latest SARS-CoV-2 variant of concern so concerning?," *J. Virol.*, vol. 96, no. 6, 2022, Art. no. e02077-21.
[8] R. Channappanavar, J. Zhao, and S. Perlman, "T cell-mediated immune response to respiratory coronaviruses," *Immunol. Res.*, vol. 59, pp. 118–128, 2014.
[9] H.-L. J. Oh, S. Ken-En Gan, A. Bertoletti, and Y.-J. Tan, "Understanding the T cell immune response in SARS coronavirus infection," *Emerg. Microbes Infections*, vol. 1, no. 1, pp. 1–6, 2012.
[10] L. Yang et al., "Persistent memory CD4+ and CD8+ T-cell responses in recovered severe acute respiratory syndrome (SARS) patients to SARS coronavirus M antigen," *J. Gen. Virol.*, vol. 88, no. Pt 10, 2007, Art. no. 2740.
[11] M. G. Rudolph, R. L. Stanfield, and I. A. Wilson, "How TCRs bind MHCs, peptides, and coreceptors," *Annu. Rev. Immunol.*, vol. 24, pp. 419–466, 2006.
[12] C. H. Bassing, W. Swat, and F. W. Alt, "The mechanism and regulation of chromosomal V (D) J recombination," *Cell*, vol. 109, no. 2, pp. S45–S55, 2002.
[13] H. S. Robins et al., "Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells," *Blood J. Amer. Soc. Hematol.*, vol. 114, no. 19, pp. 4099–4107, 2009.
[14] J. Glanville et al., "Identifying specificity groups in the T cell receptor repertoire," *Nature*, vol. 547, no. 7661, pp. 94–98, 2017.
[15] P. Dash et al., "Quantifiable predictive features define epitope-specific T cell receptor repertoires," *Nature*, vol. 547, no. 7661, pp. 89–93, 2017.
[16] D. V. Bagaev et al., "VDJdb in 2019: Database extension, new analysis infrastructure and a T-cell receptor motif compendium," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D1057–D1062, 2020.
[17] T. Borrman et al., "ATLAS: A database linking binding affinities with structures for wild-type and mutant TCR-pMHC complexes," *Proteins: Struct. Function Bioinf.*, vol. 85, no. 5, pp. 908–916, 2017.
[18] S. Mahajan et al., "Epitope specific antibodies and T cell receptors in the immune epitope database," *Front. Immunol.*, vol. 9, 2018, Art. no. 2688.
[19] N. Tickotsky, T. Sagiv, J. Prilusky, E. Shifrut, and N. Friedman, "McPAS-TCR: A manually curated catalogue of pathology-associated T cell receptor sequences," *Bioinformatics*, vol. 33, no. 18, pp. 2924–2929, 2017.
[20] A. K. Bentzen et al., "Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes," *Nature Biotechnol.*, vol. 34, no. 10, pp. 1037–1045, 2016.
[21] M. Klinger et al., "Multiplex identification of antigen-specific T cell receptors using a combination of immune assays and immune receptor sequencing," *PLoS One*, vol. 10, no. 10, 2015, Art. no. e0141561.
[22] I. V. Zvyagin, V. O. Tsvetkov, D. M. Chudakov, and M. Shugay, "An overview of immunoinformatics approaches and databases linking T cell receptor repertoires to their antigen specificity," *Immunogenetics*, vol. 72, pp. 77–84, 2020.
[23] E. Jokinen, J. Huuhtanen, S. Mustjoki, M. Heinonen, and H. Lähdesmäki, "Predicting recognition between T cell receptors and epitopes with TCRGP," *PLoS Comput. Biol.*, vol. 17, no. 3, 2021, Art. no. e1008814.
[24] S. Gielis et al., "Detection of enriched T cell epitope specificity in full T cell receptor sequence repertoires," *Front. Immunol.*, vol. 10, 2019, Art. no. 2820.
[25] V. I. Jurtz et al., "NetTCR: Sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks," *BioRxiv*, pp. 433706, 2018, doi: 10.1101/433706.
[26] G. Isacchini, A. M. Walczak, T. Mora, and A. Nourmohammad, "Deep generative selection models of T and B cell receptor repertoires with soNNia," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 14, 2021, Art. no. e2023141118.
[27] J.-W. Sidhom, H. B. Larman, D. M. Pardoll, and A. S. Baras, "DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires," *Nature Commun.*, vol. 12, no. 1, 2021, Art. no. 1605.
[28] I. Springer, H. Besser, N. Tickotsky-Moskovitz, S. Dvorkin, and Y. Louzoun, "Prediction of specific TCR-peptide binding from large dictionaries of TCR-peptide pairs," *Front. Immunol.*, vol. 11, 2020, Art. no. 1803.
[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[30] M. E. Peters et al., "Knowledge enhanced contextual word representations," 2019, *arXiv:1909.04164*.
[31] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, 2019, Art. no. 9.

[32] M. Heinzinger et al., "Modeling aspects of the language of life through transfer-learning protein sequences," *BMC Bioinf.*, vol. 20, no. 1, pp. 1–17, 2019.
[33] A. Nambiar, M. Heflin, S. Liu, S. Maslov, M. Hopkins, and A. Ritz, "Transforming the language of life: Transformer neural networks for protein prediction tasks," in *Proc. 11th ACM Int. Conf. Bioinf. Comput. Biol. Health Inform.*, 2020, pp. 1–8.
[34] R. Rao et al., "Evaluating protein transfer learning with TAPE," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019.
[35] S. El-Gebali et al., "The Pfam protein families database in 2019: Nucleic acids res," *Nucleic Acids Res.*, vol. 47, pp. D427–D432, 2019.
[36] J. Cheng, K. Bendjama, K. Rittner, and B. Malone, "BERTMHC: Improved MHC–peptide class II interaction prediction with transformer and multiple instance learning," *Bioinformatics*, vol. 37, no. 22, pp. 4172–4179, 2021.
[37] A. S. Shomuradova et al., "SARS-CoV-2 epitopes are recognized by a public and diverse repertoire of human T cell receptors," *Immunity*, vol. 53, no. 6, pp. 1245–1257.e5, 2020.
[38] J. Vig, A. Madani, L. R. Varshney, C. Xiong, R. Socher, and N. F. Rajani, "BERTology meets biology: Interpreting attention in protein language models," 2020, *arXiv:2006.15222*.
[39] J. Jin et al., "Deep learning pan-specific model for interpretable MHC-I peptide binding prediction with improved attention mechanism," *Proteins: Struct. Function Bioinf.*, vol. 89, no. 7, pp. 866–883, 2021.
[40] R. Steed, S. Panda, A. Kobren, and M. Wick, "Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models," in *Proc. 60th Annual Meet. Associat. Comput. Linguist. (Volume 1: Long Papers)*, 2022, pp. 3524–3542, doi: 10.18653/v1/2022.acl-long.247.
[41] H. Zhang, A. X. Lu, M. Abdalla, M. McDermott, and M. Ghassemi, "Hurtful words: Quantifying biases in clinical contextual word embeddings," in *Proc. ACM Conf. Health Inference Learn.*, 2020, pp. 110–120.
[42] N. Houlsby et al., "Parameter-efficient transfer learning for NLP," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2790–2799.
[43] J. Lee, R. Tang, and J. Lin, "What would elsa do? freezing layers during transformer fine-tuning," 2019, *arXiv:1911.03090*.
[44] S. Li, G. Yuan, Y. Dai, Y. Zhang, Y. Wang, and X. Tang, "SmartFRZ: An Efficient Training Framework using Attention-Based Layer Freezing,".
[45] R. Tinn et al., "Fine-tuning large neural language models for biomedical natural language processing," *Patterns*, vol. 4, no. 4, 2023, Art. no. 100729.
[46] R. Vita et al., "The immune epitope database (IEDB) 3.0," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D405–D412, 2015.
[47] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, vol. 48, no. 3, pp. 443–453, 1970.
[48] Y. Gao et al., "Pan-peptide meta learning for T-cell receptor–antigen binding recognition," *Nature Mach. Intell.*, vol. 5, no. 3, pp. 236–249, 2023.
[49] A. M. Luu, J. R. Leistico, T. Miller, S. Kim, and J. S. Song, "Predicting TCR-epitope binding specificity using deep metric learning and multimodal learning," *Genes*, vol. 12, no. 4, 2021, Art. no. 572.
[50] I. Springer, H. Besser, N. Tickotsky-Moskovitz, S. Dvorkin, and Y. Louzoun, "Prediction of specific TCR-peptide binding from large dictionaries of TCR-peptide pairs," *Front. Immunol.*, vol. 11, 2020, Art. no. 1803.
[51] B. Howie et al., "High-throughput pairing of T cell receptor α and β sequences," *Sci. Transl. Med.*, vol. 7, no. 301, pp. 301ra131–301ra131, 2015.
[52] L. Prechelt, "Early stopping-but when?," in *Neural Networks: Tricks of the Trade*, Springer, 2002, pp. 55–69.
[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[54] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Boca Raton, FL, USA: CRC Press, 1986.
[55] I. Tenney, D. Das, and E. Pavlick, "BERT rediscovers the classical NLP pipeline," 2019, *arXiv:1905.05950*.
[56] Y. Hao, L. Dong, F. Wei, and K. Xu, "Visualizing and understanding the effectiveness of BERT," 2019, *arXiv:1908.05620*.
[57] T. Lu et al., "Deep learning-based prediction of the T cell receptor–antigen binding specificity," *Nature Mach. Intell.*, vol. 3, no. 10, pp. 864–875, 2021.
[58] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does bert look at? An analysis of bert's attention," 2019, *arXiv:1906.04341*.
[59] G. Dolton et al., "Emergence of immune escape at dominant SARS-CoV-2 killer T cell epitope," *Cell*, vol. 185, no. 16, pp. 2936–2951.e19, 2022.
[60] M. C. F. Thomsen and M. Nielsen, "Seq2Logo: A method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion," *Nucleic Acids Res.*, vol. 40, no. W1, pp. W281–W287, 2012.
[61] A. Wang and O. Russakovsky, "Overwriting pretrained bias with finetuning data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3957–3968.

Sunyong Yoo received the PhD degree in bio and brain engineering from the Korea Advanced Institute of Science and Technology (KAIST), Korea, in 2018. He is currently an associated professor with the Department of ICT Convergence System Engineering, Chonnam National University, Gwangju, Korea. His research interests include bioinformatics, cheminformatics, systems biology, and data mining.

Myeonghyeon Jeong received the MS degree from the Chonnam National University, Gwangju, Korea, in 2023. He is currently working toward the PhD degree with the Department of ICT Convergence System Engineering, Chonnam National University, Gwangju, Korea. His research interests include bioinformatics, machine learning, and deep learning.

Subhin Seomun received the MS degree in biological sciences and biotechnology from Chonnam National University, Gwangju, Korea, in 2022. She is currently working toward the PhD degree with the Department of ICT Convergence Systems Engineering, Chonnam National University, Gwangju. Her research interests include bioinformatics and data mining.

Kiseong Kim received the PhD degree in bio and brain engineering from the Korea Advanced Institute of Science and Technology (KAIST), Korea, in 2018. He is currently CEO in BioBrain Inc., Daejeon, Korea. His research interests include biomedical imaging, bio-signal processing, bioinformatics, neuro design, and artificial intelligence.

Youngmahn Han received the PhD degree in bio and brain engineering from the Korea Advanced Institute of Science and Technology (KAIST), Korea, in 2018. He is currently a principal researcher with the Supercomputing Application Research Center of Korea Institute of Science and Technology Information (KISTI). He is also CEO in AIdanBio Inc., Daejeon, Korea. His research interests include bioinformatics, immunoinformatics, chemoinformatics, genome analysis, and artificial intelligence.