Computational and Structural Biotechnology Journal 23 (2024) 165–173

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

TCR-ESM: Employing protein language embeddings to predict TCR-peptide-MHC binding

Shashank Yadav [a,1], Dhvani Sandip Vora [b,c,1], Durai Sundar [b], Jaspreet Kaur Dhanjal [c,*]

[a] Department of Biomedical Engineering, University of Arizona, Tucson 85721, AZ, USA
[b] Department of Biochemical Engineering and Biotechnology, Indian Institute of Technology Delhi, New Delhi 110016, India
[c] Department of Computational Biology, Indraprastha Institute of Information Technology, Delhi, New Delhi 110020, India

ARTICLE INFO

Keywords:
T-cell therapy
TCR-pMHC interactions
Protein language models
TCR specificity
Peptide embeddings

ABSTRACT

Cognate target identification for T-cell receptors (TCRs) is a significant barrier in T-cell therapy development, which may be overcome by accurately predicting TCR interaction with peptide-bound major histocompatibility complex (pMHC). In this study, we have employed peptide embeddings learned from a large protein language model- Evolutionary Scale Modeling (ESM), to predict TCR-pMHC binding. The TCR-ESM model presented outperforms existing predictors. The complementarity-determining region 3 (CDR3) of the hypervariable TCR is located at the center of the paratope and plays a crucial role in peptide recognition. TCR-ESM trained on paired TCR data with both CDR3α and CDR3β chain information performs significantly better than those trained on data with only CDR3β, suggesting that both TCR chains contribute to specificity, the relative importance however depends on the specific peptide-MHC targeted. The study illuminates the importance of MHC information in TCR-peptide binding which remained inconclusive so far and was thought dependent on the dataset characteristics. TCR-ESM outperforms existing approaches on external datasets, suggesting generalizability. Overall, the potential of deep learning for predicting TCR-pMHC interactions and improving the understanding of factors driving TCR specificity are highlighted. The prediction model is available at http://tcresm.dhanjal-lab.iiitd.edu.in/ as an online tool.

1. Introduction

The surveillance against pathogens and pathological cells of the body is carried out by the adaptive immune system. A cornerstone of the adaptive immune response system is the presentation of peptides by major histocompatibility complexes (MHC) class I or class II, expressed on the cell surfaces. The human MHCs are also called Human Leukocyte Antigens (HLAs) and are classified in three gene classes based on structure and function of the gene products. Class I gene products, encoded by three distinct genomic loci, HLA-A, HLA-B and HLA-C present endogenous peptides to $CD8^+$ T-cells. The letters "A", "B" or "C" are assigned based on the antigens defined by serology. The peptide-MHC complex presented to T-cells enables recognition of the antigen via the T-cell receptors (TCR). Upon activation, the T cells undergo clonal expansion [1]. A fraction of this clonally expanded repertoire is retained as long-living memory against the antigen [2]. The affinity of the TCR for any peptide is governed by the heterodimeric TCRs consisting of the α and β subunits. Both chains have been reported to affect the binding of the TCR to the peptide-MHC complex (TCR-pMHC), however, prediction of TCR-pMHC binding has been carried out with high accuracy with only the β chain [3].

Within the β chain of the TCR, the three complementarity determining regions (CDRs) make primary contacts with the MHC (CDR1 and CDR2), and the peptide (CDR3), recent studies suggest contributions of the alpha chain and other CDRs as well [4]. In both the α and β chains, the CD3 loops represent the region of the highest sequence diversity and hence, are the regions that determine receptor binding specificity [5,6]. The CDR3 diversity is defined by multiple genomic recombination events in the 'Variable' (V), 'Diversity' (D) and 'Joining' (J) TCR-related genes. The V- and J- recombinations make up the α-chain while the β-chain is a result of the V-, D- and J genes generating a broader diversity. Hence, most of the previous studies have focused only on the β-chain.

The Immune Epitope database, VDJDB and McPAS-TCR primarily

* Corresponding author.
E-mail address: jaspreet@iiitd.ac.in (J.K. Dhanjal).
[1] Equal contribution

https://doi.org/10.1016/j.csbj.2023.11.037
Received 10 September 2023; Received in revised form 19 November 2023; Accepted 20 November 2023
Available online 22 November 2023

contains information on CDR3β and houses a large fraction of the publicly available TCR-pMHC binding data [7–9] . Recent studies illuminate the importance of CDR3 of both α- and β-chains in driving specificity of the TCR [10,11]. While single-cell based high-throughput techniques for assessing TCR-pMHC binding are available, determining the role of individual components remains a time- and resource-intensive pursuit [12,13].

Various studies have reported TCR-pMHC data modeling and prediction techniques, mostly based on the data from VDJDB, IEDB and McPAS-TCR, using either or both CDR3 α- and β-sequences. These previous studies have employed Gaussian processes, position-specific scoring matrices, deep learning methods [14–17] as well as advanced methods employing Natural Language Processing (NLP) [3]. The performance of current methods is limited by the paucity of available data. The applicability in terms of generalizability of the predictors is also hindered by the redundancy in epitope specific TCR sequences. Extracting relevant information from sequence data, labeled and unlabeled, has been demonstrated by NLP-based self-supervised learning algorithms. One such algorithm, Bidirectional Encoder Representations from Transformers (BERT), has been reported to reliably capture biological properties of proteins [18]. Large protein language models such as ESM (Evolutionary Scale Model) [19] and its variants have been used to predict biological structure and function of proteins. Also, the embeddings learned by the model have been fine-tuned for downstream tasks such as protein-drug interactions prediction [20], protein variant effect prediction [21] and gene ontology annotation [22].

Here, we present TCR-ESM, a deep learning-based model to predict TCR-peptide-MHC binding. TCR-ESM is a feedforward neural network trained on embeddings extracted from ESM, a protein language model [23]. ESM1v - a variant in the ESM family of protein language models has been used to generate sequence length-independent embeddings for the three protein moieties involved in the binding. The encoded protein sequence information is fed into a feedforward neural network to predict possibility of interaction. The positive data class consists of experimentally validated pairs of class 1 and class 2 MHC-peptide which bind to the TCRs. The negative data is generated by mismatching the positive dataset and ensuring the new combinations are absent in the positive class. Ablation analysis has also been carried out to qualitatively determine the importance of the TCRα, TCRβ and MHC components in the classifier's predictions. The study has been benchmarked against current TCR-pMHC prediction models such as pEptide tcR matchinG predictiOn (ERGO II) [4] and netTCR2.0 [24] using the netTCR2.0, McPAS and VDJDB datasets [3,9,24,25]. Our model, TCR-ESM has also been tested on the external dataset MIRA, as reported in the netTCR2.0 study as well as an additional independent external test set pMTnet [24, 26]. We report improved prediction capacity based on the Matthews correlation coefficient (MCC) score, which has been established as a more reliable performance metrics in binary classification on imbalanced data rather than precision, recall and ROC AUC [27,28]. The MCC score ranges from 0, indicating poor model predictive performance, to 1, signifying ideal model performance. However, we additionally report model performance on precision-recall, ROC AUC as well as the F1 score. We also observe that fine tuning the embeddings extracted from large protein language models can preserve both local and global information towards specific objectives.

2. Results

The TCR-ESM prediction model was trained on the peptide, TCR and MHC embeddings as detailed in Fig. 1. The peptide, TCR and MHC sequence information was fed into the neural network after embeddings were generated from the penultimate layer of the ESM1v model (Fig. 1**a**). The feedforward neural network was trained to predict TCR-peptide binding pairs as 1, and non-interacting pairs as 0. Performance was evaluated over MCC scores, area under the receiver operating characteristic (AUC-ROC), area under the Precision-Recall curve (AU-PR) and F1 scores. The choice of model architecture was governed by the feature set in consideration. For instance, for predicting CDR3α-CDR3β-peptide binding the model illustrated in Fig. 1**b** was used. Similarly, Fig. 1**c** illustrates the model used for CDR3α-CDR3β-peptide-MHC binding. For CDR3α-peptide, CDR3β-peptide, CDR3α-peptide-MHC, CDR3β-peptide-MHC binding prediction, the model architectures are illustrated in Supplementary Figure 1a, 1b, 1c and 1d, respectively.

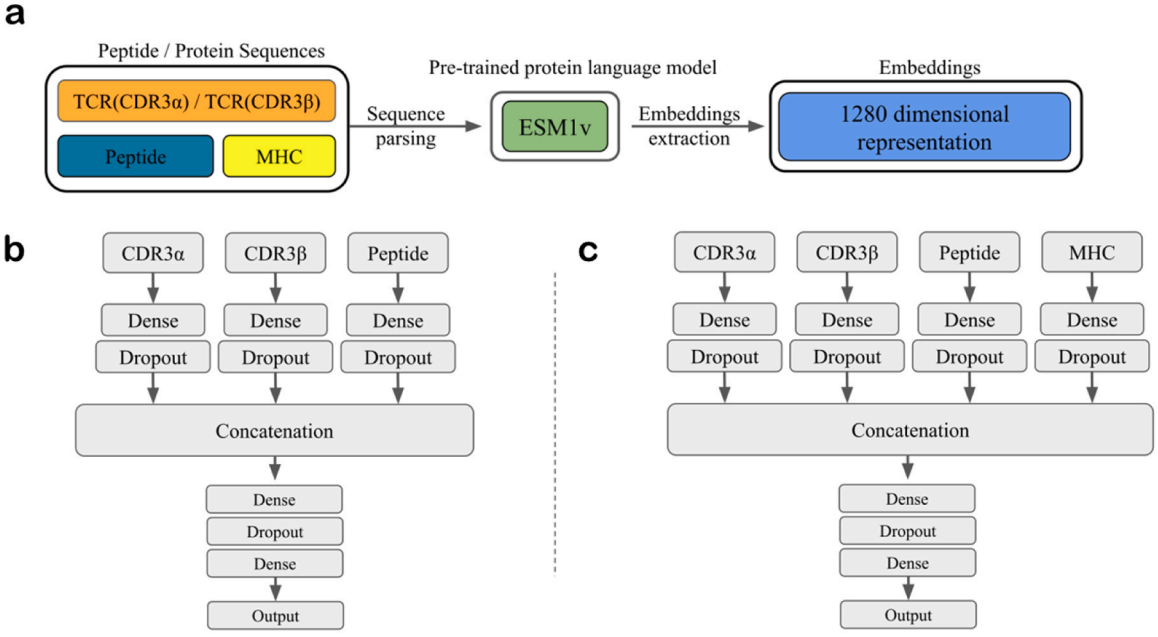

Fig. 1. : Schematic of the method followed. (a) The ESM1v protein language model was employed to extract the sequence information in the form of 1280-dimensional embeddings for each of the three components- TCR, peptide and MHC. (b) Architecture of the neural network employed for the TCR(CDR3α+CDR3β)-peptide binding classification task. (c) The neural network architecture employed for the TCR(CDR3α+CDR3β)-peptide-MHC binding prediction task.

2.1. Model performance evaluation on CDR3β data

The probability of the TCR-peptide binding predicted by TCR-ESM on the netTCR2.0 CDR3β dataset was evaluated. A side-by-side comparison of TCR-ESM with the 1D CNN-based netTCR2.0 was carried out. The data was prepared and partitioned the same as the netTCR2.0 study [24]. The netTCR2.0, ERGO-AE, ERGO-LSTM and the classifier presented in this study- TCR-ESM were trained, and cross-validation was carried out on the CDR3β data. Same as netTCR2.0, five different MIRA datasets were obtained by imposing a separation from the training set of 90 %, 92 %, 94 %, 99 %, and 100 % similarity. That is, MIRA 94 % TCRs do not share more than 94% Levenshtein similarity to any of the TCRs in the training set. These datasets were used as independent test sets. Our model outperformed netTCR2.0 on all similarity-based partition thresholds (Fig. 2**a**, Supplementary Table S1a, Supplementary Figs. S2, S3 and S4.

We then compared the peptide-specific outperformance for the two most abundant peptides, GILGFVFTL (GIL) from Influenza A virus and GLCTLVAML (GLC) from Human herpesvirus 4 (Epstein–Barr virus) for all partitioning thresholds (Figs. 2**b**, 2**c**, Supplementary Table S1a).

2.2. Model performance evaluation and ablation analysis on paired CDR3α and β

In addition to the CDR3β information, to check whether CDR3α information or paired CDR3αβ information is beneficial for the prediction of peptide binding by the TCR, the TCR-ESM model was cross-validated and benchmarked against netTCR2.0. Fig. 2**d** shows the TCR-ESM and netTCR2.0 models evaluated with 5-fold cross-validation run 10 times independently on the sequence similarity-based partitioned datasets as is from the netTCR2.0 study with the CDR3 α-chain, CDR3 β-chain, and both α- and β-chains of CDR3. The TCR-ESM model significantly

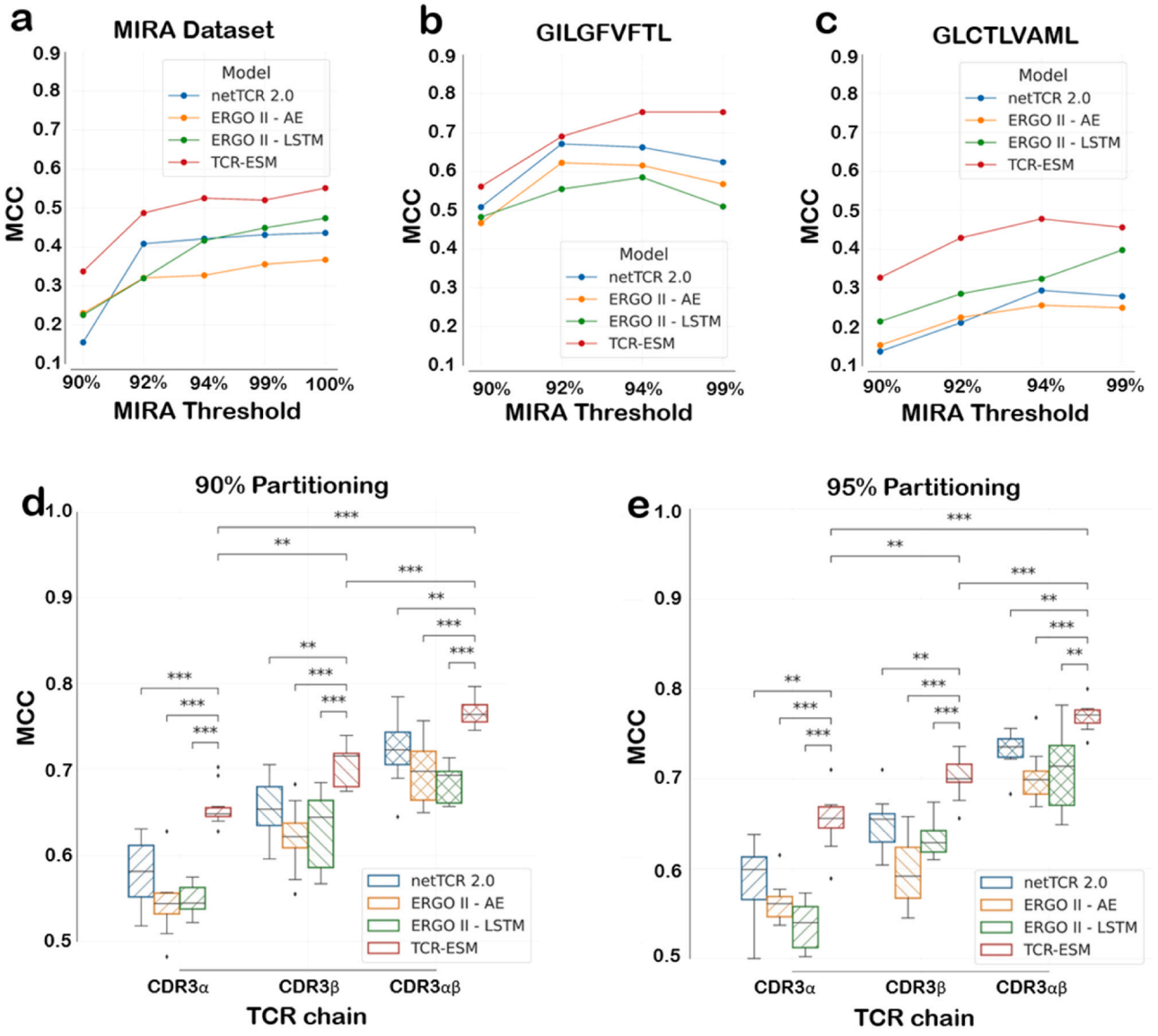

Fig. 2. : Model performance benchmarking. (a) Comparison of TCR-ESM with netTCR2.0 on the external MIRA dataset as test set for different partitioning thresholds. (b) Comparison of TCR-ESM with netTCR2.0 on the MIRA dataset as test set specific to the most common peptide in the dataset, 'GILGFVFTL'. (c) Comparison of TCR-ESM with netTCR2.0 on the MIRA dataset as test set specific to the second most common peptide 'GLCTLVAML'. 5-fold cross validation is run independently 10 times to compare the model performance and the distribution is represented by the box-whisker plots to compare the performance of TCR-ESM with netTCR2.0 at (d) 90% partitioning threshold and (e) 95% partitioning threshold for different prediction tasks such as TCR(CDR3α)+peptide, TCR(CDR3β)+peptide, and TCR(CDR3α+CDR3β)+peptide. All statistical comparisons were done using Mann-Whitney U test with Benjamini-Hochberg correction (ns: $p > 5.00e\text{-}02$, *: $1.00e\text{-}02 < p <= 5.00e\text{-}02$, **: $1.00e\text{-}03 < p <= 1.00e\text{-}02$, ***: $1.00e\text{-}04 < p <= 1.00e\text{-}03$, ****: $p <= 1.00e\text{-}04$).

outperforms (p-value <= 0.05) the netTCR2.0 model on all three datasets (CDR3α only, CDR3β only, and paired). We also obtained similar outperformance for the 95% partitioned dataset (Fig. 2**e,** Supplementary Table S1b). We validated this observation by comparing netTCR2.0 model performance on the three datasets (CDR3α only, CDR3β only, and paired) and found that models trained on paired chain information outperformed models trained only on single CDR3 chains. (Supplementary Fig. S5).

To further validate the performance of the TCR-ESM model, we compared our model with the ERGO II model. The ERGO II model is available in two types, namely ERGO II-Autoencoder (ERGO II-AE) and ERGO II-Long Short-Term Memory (ERGO II-LSTM). Since ERGO II-AE and ERGO II-LSTM were originally benchmarked on the McPAS and VDJDB data, we utilized these datasets with exactly the same train-test splits as provided. We observed that our model outperforms netTCR2.0, ERGO II-AE and ERGO II-LSTM model for both McPAS and VDJDB paired chain datasets (10-fold cross validation, p-value<=0.05, p-value correction with Benjamini-Hochberg method) (Fig. 3**a,** Supplementary Tables S2**,** Supplementary Figs. S6, S7, S8. Both McPAS and VDJDB subsets 1 lack MHC information while subsets 2 include MHC information. We cross-validated these models on the McPAS data subset-1. We also observed a similar trend on test-set MCC where our model performed better when compared to the other three models (Fig. 3**b**). We observe a similar trend of TCR-ESM outperforming in the VDJDB data subset-1 on both cross validation and test setting as illustrated in Fig. 3**c** and Fig. 3**d**.

Further supporting analysis on whether CDR3 paired chain information can predict the TCR-peptide binding better than single chains was carried out. We performed 10-fold cross-validation for all the four models (netTCR2.0, ERGO II-AE, ERGO II-LSTM and our TCR-ESM) on paired CDR3 in the McPAS data subset-1 as well single chains, and observed that paired CDR3 information improves the MCC significantly (t-test p-value<=0.05, p-value correction with Benjamini-Hochberg method) when compared to the model trained only on CDR3α and CDR3β **(**Supplementary Figs. S9, S10**)**. We observe a similar trend of model outperformance with paired CDR3 information on the VDJDB data subset-1 (Supplementary Figs. S6, S7, S8).

2.3. Model performance evaluation on paired CDR3 with MHC information and ablation analysis

The contribution of different features was further illustrated by performing ablation analysis on McPAS data subset-2 and VDJDB data subset-2, which contains MHC information. Since there are three components, CDR3α, CDR3β, and MHC, six feature sets were constructed-CDR3α only, CDR3β only, paired CDR3α-CDR3β, CDR3α-MHC, CDR3β-MHC and combined CDR3α-CDR3β-MHC. The TCR-ESM-MHC model was compared with ERGO II-AE and ERGO II-LSTM models on McPAS data subset-2 (containing MHC information) for each of the six feature sets. TCR-ESM-MHC is shown to outperform the ERGO-AE model on all

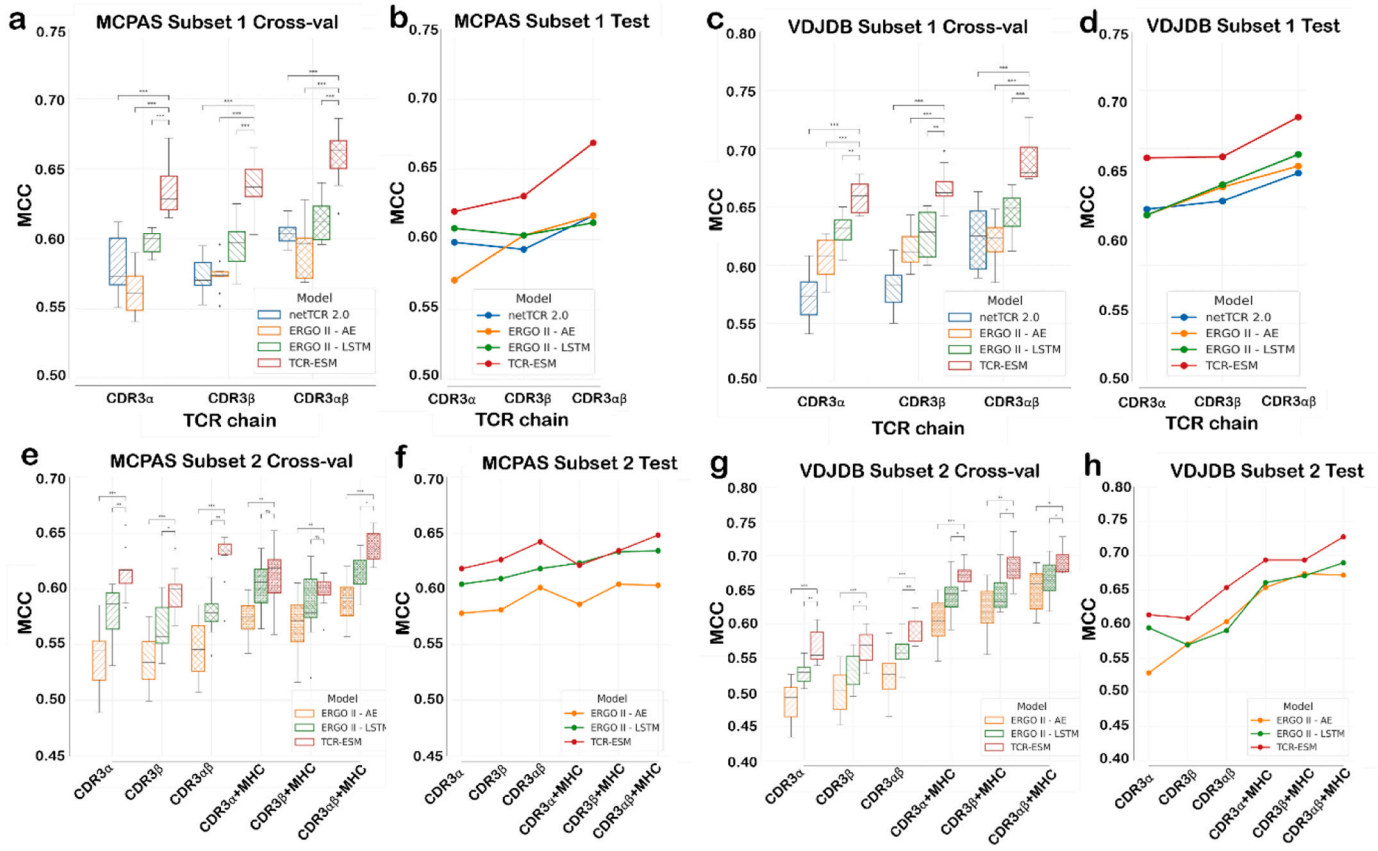

Fig. 3. : Feature importance as determined by ablation. The box-whisker plots show model performance distribution on 10-fold cross-validation using the TCR-ESM predictor, netTCR2.0, ERGO II-AE and ERGO II-LSTM for different prediction tasks- TCR(CDR3α)+peptide, TCR(CDR3β)+peptide, and TCR(CDR3α+CDR3β)+peptide on (a) McPAS data subset-1, (c) VDJDB data subset-1. Similarly, hold-out testing set comparison of TCR-ESM model with netTCR2.0, ERGO II-AE and ERGO II-LSTM model for the three different prediction tasks on (b) McPAS data subset-1, and (d) VDJDB data subset-1. 10-fold cross validated comparison of TCR-ESM model with netTCR2.0, ERGO II-AE and ERGO II-LSTM model for different prediction tasks such as TCR(CDR3α)+peptide, TCR(CDR3β)+peptide, TCR(CDR3α + CDR3β) + peptide, TCR(CDR3α)+peptide+MHC, TCR(CDR3β)+peptide+MHC and TCR(CDR3α + CDR3β) + peptide+MHC on (e) McPAS data subset-2 and (g) VDJDB data subset-2. Similarly, hold-out testing set comparison of TCR-ESM model with netTCR2.0, ERGO II-AE and ERGO II-LSTM model was performed for the different prediction tasks on (f) McPAS data subset-2 and (h) VDJDB data subset-2. All statistical comparisons were done using Mann-Whitney U test with Bejamini-Hochberg correction (ns: p > 5.00e-02, *: 1.00e-02 < p <= 5.00e-02, **: 1.00e-03 < p <= 1.00e-02, ***: 1.00e-04 < p <= 1.00e-03, ****: p <= 1.00e-04).

six feature sets during cross-validation and on the test set, while also outperforming ERGO-LSTM on four of the six sets during cross-validation and on the test set (Figs. 3**e,** 3**f**). TCR-ESM-MHC also outperformed on the VDJDB data subset-2 (VDJDB with MHC information) as compared to both ERGO II-AE and ERGO II-LSTM model on 10-fold cross validation and test set (Fig. 3**g,** Fig. 3**h**). For the ablation experiment, the ERGO II-AE, ERGO II-LSTM and TCR-ESM-MHC models were trained individually on the six feature sets. The feature importance of the MHC based on increase in MCC scores is dataset- and method-dependant. For the McPAS data subset- 2, when analysed with ERGO II-AE, addition of the MHC features significantly improved model performance indicating high feature importance. However, when the McPAS data subset-2 was analysed with ERGO II-LSTM and the TCR-ESM-MHC models, inclusion of the MHC information did not significantly improve performance reflecting low feature importance (Supplementary Figs. S8, S9, S10). Contrarily, for the VDJDB data subset-2, there was a statistically significant improvement in performance upon inclusion of the MHC features for all the models trained- ERGO II-AE, ERGO II-LSTM and TCR-ESM-MHC. The observed increase in MCC scores indicated that the MHC sequence information plays an important role in driving model output and therefore is an important feature (Supplementary Figs. S8, S9, S10).

2.4. Analysis of embeddings learned by the TCR-ESM model

The outputs from different layers of the TCR-ESM-MHC model were extracted to understand how the output of the model is driven, t-SNE was used to reduce the layer output to two dimensions. t-SNE enables capturing local relationships while also capturing non-linear relationships in the data. This is significant to determine if the embeddings can be repurposed for related tasks. We compared the embeddings of binding TCR to non-binding TCR in the input layer, concatenation layer and penultimate dense layer. For the McPAS dataset, the most common peptide 'GILGFVFTL' was selected from the test set. As one progresses through the layers of our model, the outputs generated become more abstract and less directly tied to specific features in the input data. Later layers then use these abstractions to construct more sophisticated features that are better suited for specific tasks. This can lead to the outputs of the later layers being more separable, or easier to distinguish from one another, compared to the outputs of the earlier layers. The results showed that the learned embeddings became more distinct as the model was trained, with positive TCR interactions and negative TCR interactions in the independent dataset for specific peptides being mixed at the input layer (Figs. 4**a,** 4**b**). The subsequent neural network layers learn the embeddings of the multiple inputs jointly, following the input layer where the information is supplied separately. Gradually the learned embeddings can be distinguished at the concatenation and penultimate dense layers which also capture the information of binding and non-binding CDR3α and CDR3β sequences (Figs. 4**c,** 4**d**). Similarly, for the VDJDB dataset, we picked the most common 'NLVPMVATV' peptide and performed a similar analysis of the input, concatenation and penultimate dense layer embeddings of binding and non-binding CDRα and CDR3β sequences, as illustrated in Supplementary Fig. S11**.** A similar analysis was performed for two other randomly selected peptides, also demonstrating separability as shown in Supplementary Fig. S12**.** A layer-wise analysis of the separation obtained is studied by employing a random forest classifier at each layer and evaluating the MCC scores (Supplementary Fig. S13). We observed that the model learns a joint embedding for CDR3α and CDR3β and shows separation between positive and negative TCR samples for specific peptides.

2.5. External testing analysis

The level of dissimilarity between CDR3β and peptide sequences in

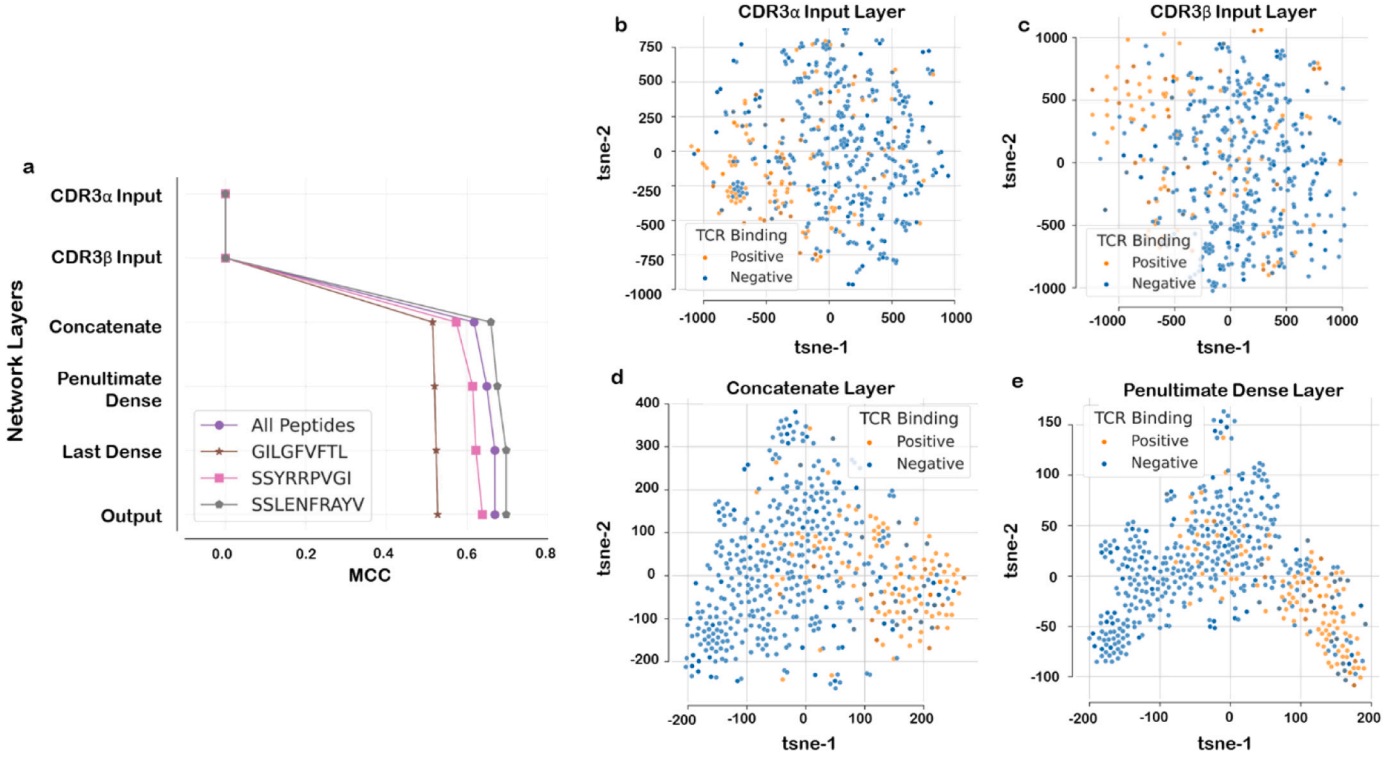

Fig. 4. : (a) Predictive capacity of the different layers of the TCR-ESM classifier for randomly selected peptides 'GILGFVFTL', 'SSYRRPVGI' and 'SSLENFRAYV'. Comparison of two-dimensional t-SNE embeddings for positive and negative (b) TCR(CDR3α) and (c) TCR(CDR3β) for the most common peptide 'GILGFVFTL' from the Input Layer of model (TCR(CDR3α + CDR3β) + peptide) trained on McPAS dataset. Comparison of jointly learned two-dimensional T-SNE embeddings for positive and negative TCR(CDR3α + CDR3β) for the most common peptide 'GILGFVFTL' from the (d) Concatenation Layer and (e) Penultimate Dense Layer of model (TCR(CDR3α + CDR3β) + peptide) trained on McPAS dataset.

the pMTnet dataset with the McPAS and VDJDB dataset was determined. For CDR3β sequences, pMTnet had only one (1 in 272; 0.4% identical) CDR3β sequence which was common with CDR3β in McPAS dataset. Similarly, for the VDJDB dataset, we found that the pMTnet dataset has 39 (39 in 272; 14.7% identical) common CDR3β sequences. Comparing peptides, pMTnet data has 21 (21 in 224; 9.38% identical) common peptides with McPAS data and 42 (42 in 224; 19.6% identical) common peptides with VDJDB data.

Separate models were trained on the McPAS and VDJDB dataset and tested on the pMTnet dataset to evaluate how the models perform in an external setting (Fig. 5**a**). Since there were some common peptides between models trained on McPAS and VDJDB with the external pMTNet test set, the models were evaluated after removing the common peptides as well (Fig. 5**b**). Further stringent testing was carried out by removing peptides with 90% and 80% sequence similarity to any peptide in train sets (Fig. 5**c, d**). The TCR-ESM model was shown to perform significantly better than the netTCR2.0, ERGO II-AE and ERGO II-LSTM models on both McPAS and VDJDB datasets. TCR-ESM also showed a higher performance on the test sets when trained on VDJDB as compared to McPAS, even after filtering similar sequences (Fig. 5**,** Supplementary

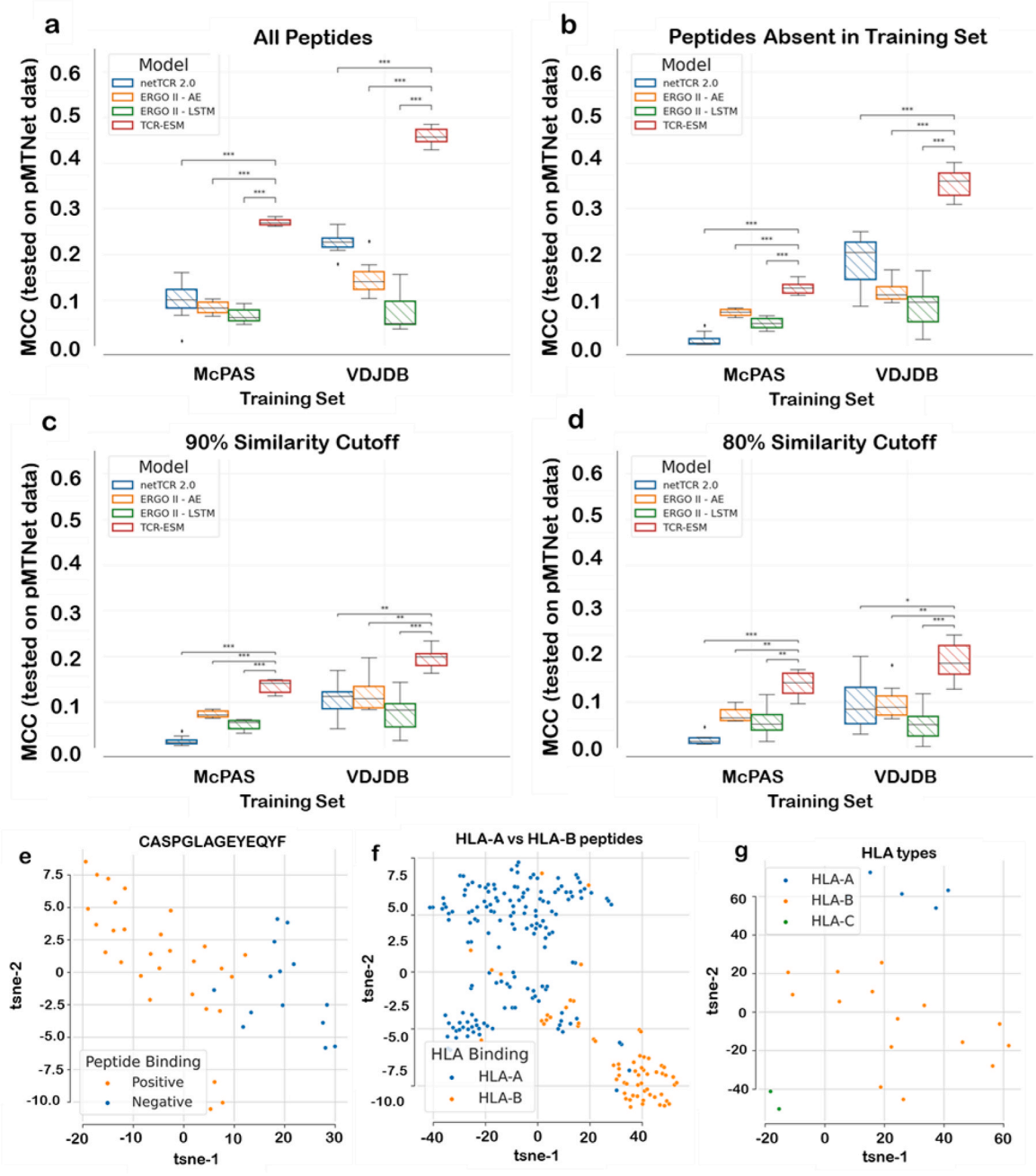

Fig. 5. : Comparison on external test set performance for netTCR2.0, ERGO II-AE, ERGO II-LSTM and TCR-ESM models. Models were trained on McPAS and VDJDB datasets and tested on PMTNet dataset (a), after removing peptides from the test set that are common to the training sets (b), after removing peptides that share 90% similarity with any peptide in the train sets (c) and after filtering 80% similar peptides (d). (e) Two-dimensional t-SNE embeddings for positive and negative peptides for an external dataset TCR(CDR3β) 'CASPGLAGEYEQYF' from the Penultimate Dense Layer of model (TCR(CDR3β) + peptide). Two-dimensional t-SNE of ESM-generated embeddings of (f) test set peptides which bind to different MHC-I types (HLA-A vs HLA-B), and (g) different MHC-I subtypes (HLA-A, HLA-B, and HLA-C). All statistical comparisons were done using Mann-Whitney U test using Benjamini-Hochberg correction (ns: $p <= 1.00e+00$, *: $1.00e-02 < p <= 5.00e-02$, **: $1.00e-03 < p <= 1.00e-02$, ***: $1.00e-04 < p <= 1.00e-03$, ****: $p <= 1.00e-04$).

Fig. S14). Overall, the TCR-ESM-MHC model showed improved performance on an external dataset, indicating that the approach has promising generalization capabilities. The results are encouraging because it suggests that the model is not simply memorizing the training data, but rather learning more generalizable features that can be applied to different types of data. Further testing on a variety of external datasets will be necessary to confirm the robustness of this approach, nevertheless, the initial results seem promising.

Next, the embeddings learned by the model in this external set case were evaluated. One particular case was identified where the model generalizes for the CDR3β sequence 'CASPGLAGEYEQYF' which is not present in the training set. The embeddings learned by the penultimate layer of the TCR-ESM-MHC model were able to reliably differentiate between positive and negative peptides which can bind to 'CASPGLAGEYEQYF' (Fig. 5**e**).

The embeddings learned by the model were observed to help differentiate between peptides which bind to HLA-A versus peptides which bind to HLA-B (Fig. 5**f**). The embeddings were able to capture this information and encode it in a way that allows the model to accurately predict the binding status of a given peptide. It suggests that models trained on embeddings learned by large protein language models can also be used to predict the immunogenicity of different peptides and other related tasks. The nature of the HLA subtype with which a peptide is presented to T cells determines the immunogenicity, the Fig. 5 shows that the embeddings can learn the differences in these HLA subtypes. While it is true that the immunogenicity of a peptide cannot be predicted directly by the embeddings of the HLA, extracting meaningful features is a key step in building powerful prediction models. Also, the embeddings could distinguish between the HLA types for the positive peptides present in the test data (Fig. 5**g**).

3. Discussion

One major obstacle in the creation of T-cell therapies is the difficulty of identifying the specific targets, known as cognate targets, that are recognized by T-cell receptors (TCRs). This is a crucial step in the development of these therapies because the TCRs are responsible for recognizing and binding to these targets in order to initiate an immune response. Without the ability to identify and target these cognate targets, it is difficult to effectively design and implement T-cell therapies. By creating models that can anticipate TCR-pMHC interactions based on the amino acid sequences of the peptide and CDR3 region of the TCR chains, we present a study that aims to address this bottleneck using learned representations of peptides extracted using large protein language models. There were several model designs examined, ranging from single chain CDR3α-peptide binding prediction to paired CDR3α-CDR3β-peptide-MHC binding. The models were built utilizing rigorous data-redundancy reduction guidelines, trained using cross-validation, and verified using independent assessment data.

Models that used data from paired TCR category, included both CDR3α and CDR3β information, performed significantly better when compared to models trained on data with only CDR3β information. The results from the proposed study support the idea that both TCR chains contribute to TCR specificity, and that their relative importance varies depending on the specific pMHC being targeted. Considering that the datasets of McPAS and VDJDB are vastly different, the results comparing model performance on the two sets should be taken with a pinch of salt. However, on the same dataset, for example, McPAS, the TCR-ESM model does not show significant improvement in performance as opposed to ERGO II, which shows significant improvement upon inclusion of MHC data. This could be attributed to the nature of the prediction model itself. TCR-ESM is able to learn the distinguishing properties between the two classes completely based on the CDR3 sequence information. The McPAS dataset has more variable MHC chain information and the MCC score of TCR-ESM is lesser than that for the MCC on the VDJDB set, with less variable MHC information. This could be a case of the model being overfit on the limited data.

Furthermore, the inclusion of MHC information in the prediction task of TCR-peptide binding may improve the performance of the model. The impact of MHC on TCR-peptide binding is highly dataset specific and can vary depending on the characteristics of the dataset being used. The role of MHC in TCR recognition is highly complex and context-dependent, influenced by factors such as MHC polymorphism, peptide binding motifs, and peptide-MHC interactions. These factors can introduce variability in TCR-peptide binding across different datasets. Although the dataset specificity of MHC influence presents a challenge, our proposed method aims to capture and model the general principles underlying TCR-peptide binding, while acknowledging the dataset-specific nuances. By incorporating a diverse range of training data that covers various MHC alleles and peptide sequences, the model developed captures the common features and patterns of TCR-peptide binding. Furthermore, during the development and evaluation of the method, steps have been taken to address dataset specificity. It is ensured that the training dataset comprises a broad representation of MHC alleles to account for the variability in MHC-specific effects. This allows the model to learn generalizable features that are not overly biased towards any specific MHC allele.

Another limitation of the study being that it is important to carefully evaluate the impact of MHC on model performance for each specific dataset in order to determine the most effective approach for modeling TCR-peptide binding. This may be attributed to multiple peptides being generally presented by the same MHC allotypes and also multiple MHC allotypes presenting similar peptides, depending on the peptide processing and presentation pathway of the host organism.

Both negative and positive samples are essential to train a binary classifier. In the absence of negative instances derived experimentally, data points were generated by shuffling the positive set, as was derived from previous studies as described in the Methods section. However, the approaches assume that TCRs show no cross-reactivity. However, this assumption, due to challenges in obtaining negative training data, may limit the model's utility, especially in predicting the behavior of promiscuous TCRs, with the capacity to bind diverse peptide-MHC complexes. TCRs exhibit varying degrees of cross-reactivity, recognizing structurally similar peptides presented by different MHC molecules. Without accounting for cross-reactivity, predictions for promiscuous TCRs may be inaccurate, struggling to capture nuanced interactions with diverse peptide-MHC ligands. To improve the method's utility, researchers could explore ways to incorporate cross-reactivity, possibly through additional features capturing structural and biochemical properties influencing cross-reactivity. Strategies like transfer learning or advanced architectures capable of learning hierarchical representations could be considered. Including limited experimental data on cross-reactive TCRs may further improve the model's generalization beyond training data.

The power of utilizing embeddings learned by large protein language models was determined. When these embeddings are learned by large language models, they can capture the complex relationships and patterns present in the data. By using these learned embeddings as inputs to simpler machine learning models, such as multi-layer perceptrons (MLPs), we can train these models more efficiently compared to using more complex models like convolutional neural networks (CNNs), autoencoders (AEs), or long short-term memory (LSTM) networks. In the context of peptides, utilizing learned embeddings from large protein language models and then fine-tuning the embeddings can be particularly useful for tasks such as classification. The fine-tuned embeddings can be specific to the problem being addressed, in this case, they capture the interactions between proteins, such as TCRs, antigens, and MHCs. By training a classifier in the feature space of the fine-tuned protein language model, we can learn better representations of these proteins, which may lead to improved classification performance compared to traditional approaches.

Once the embeddings have been learned, they can be used as input to

other machine learning models, such as classifiers or clustering algorithms. These models can then be trained on the compressed, lower-dimensional representation of the data, rather than the original high-dimensional representation. This can lead to more efficient training and faster model convergence, as well as improved performance on downstream tasks. One potential benefit of using embeddings learned by a language model instead of an autoencoder is that the embeddings capture the relationships between elements in the data, such as amino acids in a protein sequence. This can be particularly useful for biological data, where the relationships between elements can be important for understanding the structure and function of the data. In contrast, autoencoders are generally agnostic to the relationships between elements in the data, and simply learn a compressed representation based on the patterns present in the data.

In conclusion, we have developed a model to predict the interactions between TCRs and their cognate peptides and MHC molecules. Our results indicate that accurate predictions can only be achieved through the use of data from paired TCR α and β chains. While the model's current capabilities are limited to a specific set of peptides due to a lack of training data, we expect that its predictive ability will improve as more data becomes available, enabling it to accurately predict interactions with novel peptides. Additionally, the model framework is adaptable and can easily incorporate MHC molecules or TCRα chains when data becomes available, providing a comprehensive approach for predicting TCR-pMHC interactions.

4. Methods

4.1. Data collection

We download four TCR-peptide binding datasets from the GitHub repositories of netTCR2.0 (https://GitHub.com/mnielLab/NetTCR-2.0), ERGO II (https://GitHub.com/IdoSpringer/ERGO II-II) and pMTnet (https://GitHub.com/tianshilu/pMTnet). The ERGO II repository contains McPAS and VDJDB datasets. Since multiple peptides may be presented by the same MHC alleles, the MHC-peptide information may be repeated in the positive and negative classes. However, the peptide-TCR pairs experimentally validated were labeled as the positive class, the negative set was generated by generating random TCR-peptide combinations and ensuring these pairs are absent from the positive dataset. The training and test datasets were obtained from the ERGO II and netTCR2.0 repositories. The external test dataset, MIRA, was also obtained from the netTCR2.0 GitHub repository. The MIRA dataset contained 376 CDR3β-peptide pairs associated with HLA-A* 02:01. We used partitioned data as detailed in netTCR2.0 and ERGO II. The pMTnet dataset was used as an external test dataset. A detailed summary of all the datasets used in this work is provided in Supplementary Table 3 and Supplementary Table 4.

The McPAS and VDJDB datasets were processed to give two working datasets for model training and testing. The peptide and HLA counts of the two datasets are indicated in Supplementary Figure 15. Subset-1 contains paired CDR3α, CDR3β and peptide information only. Subset-2 contains paired CDR3α, CDR3β, peptide and MHC information. We utilized the fair-esm python library provided by the ESM project to extract the embeddings for the TCR, peptide and MHC sequences [19,21]. First, a FASTA file was created for each dataset. The FASTA files were parsed to the ESM1v model to extract the pre-final layer embeddings of size 1280 for each sequence. The embeddings would have the size of 1280 * peptide length (L). Global average pooling was then employed to convert it to a 1 * 1280-dimensional vector.

4.2. Model training and performance evaluation metrics

The ESM1v protein model was employed as mentioned in the original paper without finetuning [19]. ESM1v-extracted embeddings encode the TCR, peptide and MHC information. The embeddings were fed into a feedforward neural network to predict binding. Grid search was used to optimize model hyperparameters such as learning rate (ranging from 10^{-1} to 10^{-4}), dropout rate (ranging from 0.2 to 0.5), number of hidden layers (ranging from 1 to 2) and number of nodes (ranging from 2^2 to 2^7) in each layer. netTCR2.0 and ERGO II models were run on the datasets as baselines to independently calculate the MCC value for the benchmarking task. The performance was compared with the reported performance values for netTCR2.0 and ERGO II, to verify accurate reproduction of results.

The model uses the GELU activation function [29] in all of its hidden layers and the binary cross-entropy loss function for optimization. During training, the learning rate was set to 0.08 and was reduced by 5% if the validation MCC did not improve after 50 epochs. This was done using a learning rate scheduler to adapt to the changing dynamics of the training data and potentially improve model performance.

The model prediction performance was tested mainly based on MCC, which is reported to be a reliable statistical measure over other metrics such as AUC-ROC and AUC-PR since we performed classification on imbalance datasets [27,30]. The MCC is defined as the correlation between the observed and predicted binary classifications. It ranges from − 1–1, with values closer to 1 or − 1 indicating a stronger correlation and a better model performance. A value of 0 indicates no correlation, while negative values indicate an inverse correlation. The MCC results in a high score only when the predictions are reliable on all of the four categories- true and false positives as well as true and false negatives, while being proportional to the size of positive and negative samples in the dataset [31]. The model performance is also tested and reported on the area under the receiver operating characteristic (AUC-ROC), area under the precision-recall curve (AU-PR) as well as F1 scores. The metrics are calculated as:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1score = \frac{Precision \times Recall}{Precision + Recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

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

4.3. Feature importance determination (ablation experiment)

Feature importance was determined through ablation studies, which involve systematically removing or "ablating" each feature from the model one at a time and measuring the impact on model performance. By comparing the model's performance with and without a given feature, one can determine the importance of that feature in driving the model output. In the context of TCR-peptide binding, feature importance determination and ablation studies can be used to identify which CDR3 chains in the TCR has significant contribution towards binding the peptides. In cases when MHC information is added to the model, ablation studies help determine whether the added information is useful or not. For TCR-peptide binding prediction, we created three different feature subsets: CDR3α-peptide, CDR3β-peptide and CDR3α-CDR3β-peptide. Similarly, for TCR-pMHC binding, we created six different feature subsets such as CDR3α-peptide, CDR3β-peptide, CDR3α-CDR3β-peptide, CDR3α-peptide-MHC, CDR3β-peptide-MHC and CDR3α-CDR3β-peptide-MHC to calculate if adding more features improves the performance of the TCR-pMHC binding prediction. Individual models such as ERGO II-Autoencoder (ERGO II-AE), ERGO II-LSTM and TCR-ESM (developed in this study), were trained for these subsets and the

MCC metric was calculated to test the performance. We checked for statistical significance of performance using T-test and used Benjamini-Hochberg (BH) p-value correction to account for multiple testing [32].

4.4. Analysis of fine-tuned embeddings

The output of the intermediate and penultimate layers was extracted from the TCR-ESM feedforward neural network for the most frequent peptides in both McPAS and VDJDB dataset. t-distributed Stochastic Neighbor Embedding (t-SNE) was then performed on the output embeddings to reduce them to 2D for visualizing if the positive and negative TCRs cluster separately.

4.5. External testing analysis

A crucial component of assessing deep learning models is external testing. The model performance is tested on data samples taken from an external, other than the training data. Model performance on external samples can reveal information about how well it generalizes, or how well it can make predictions about unknown data. The model, when used in settings where it is likely to encounter data that differs from the training data, such as in the real world, external testing is especially important. We checked the generalizability by testing the TCR-ESM model on an independent dataset obtained from pMTnet [26] (Supplementary Table S5), which is a recently reported experimental dataset of TCR-peptide-MHC binding. pMTnet contains CDR3-peptide sequences which are different from CDR3-peptides sequences present in the McPAS and VDJDB datasets, however, there are some common peptides. The TCR-ESM models trained on McPAS and VDJDB were tested on the pMTNet set as is, after removing the common peptides and also after filtering based on 90 % and 80 % sequence similarity (sim) score measured by aligning the peptides pairwise and normalizing the alignment scores by length of the peptide.

$$sim(S_1, S_2) = \frac{\max_{a \in A} score(a)}{\max(len(S_1), len(S_2))}$$

Where A is the set of all local alignments between sequences S_1 and S_2, and a is the alignment with the highest score score(a).

Author Statement

We confirm that the manuscript has been read and approved by all authors.

Declaration of Competing Interest

The authors declare no conflict of interest.

Data availability

Raw data can be downloaded from the GitHub links as mentioned in the methods section. Processed data and the working code are available at https://GitHub.com/dhanjal-lab/tcr-esm.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.11.037.

References

[1] Zhang S-Q, Parker P, Ma K-Y, He C, Shi Q, Cui Z, et al. Direct measurement of T cell receptor affinity and sequence from naïve antiviral T cells. Sci Transl Med 2016;8 (341). 341ra377-341ra377.

[2] Sprent J, Surh CD. T cell memory. Annu Rev Immunol 2002;20(1):551–79.

[3] Springer I, Besser H, Tickotsky-Moskovitz N, Dvorkin S, Louzoun Y. Prediction of specific TCR-peptide binding from large dictionaries of TCR-peptide pairs. Front Immunol 2020;11.

[4] Springer I, Tickotsky N, Louzoun Y. Contribution of T cell receptor alpha and beta CDR3, MHC typing, V and J genes to peptide binding prediction. Front Immunol 2021;12:664514.

[5] Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. Nature 1988;334(6181):395–402.

[6] Krogsgaard M, Davis MM. How T cells' see'antigen. Nat Immunol 2005;6(3): 239–45.

[7] La Gruta NL, Gras S, Daley SR, Thomas PG, Rossjohn J. Understanding the drivers of MHC restriction of T cell receptors. Nat Rev Immunol 2018;18(7):467–78.

[8] Bagaev DV, Vroomans RMA, Samir J, Stervbo U, Rius C, Dolton G, et al. VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. Nucleic Acids Res 2020;48(D1):D1057–62.

[9] Tickotsky N, Sagiv T, Prilusky J, Shifrut E, Friedman N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. Bioinformatics 2017;33(18):2924–9.

[10] Lanzarotti E, Marcatili P, Nielsen M. T-cell receptor cognate target prediction based on paired α and β chain sequence and structural CDR loop similarities. Front Immunol 2019;10:2080.

[11] Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature 2017;547(7661):89–93.

[12] Lu YC, Zheng Z, Lowery FJ, Gartner JJ, Prickett TD, Robbins PF, et al. Direct identification of neoantigen-specific TCRs from tumor specimens by high-throughput single-cell sequencing. J Immunother Cancer 2021;9(7).

[13] Lundegaard C, Lund O, Nielsen M. Predictions versus high-throughput experiments in T-cell epitope discovery: competition or synergy? Expert Rev Vaccin 2012;11(1): 43–54.

[14] Jokinen E, Huuhtanen J, Mustjoki S, Heinonen M, Lähdesmäki H. Predicting recognition between T cell receptors and epitopes with TCRGP. PLOS Comp Biol 2021;17(3):e1008814.

[15] Gielis S, Moris P, Bittremieux W, De Neuter N, Ogunjimi B, Laukens K, et al. Detection of enriched T cell epitope specificity in full T cell receptor sequence repertoires. Front Immunol 2019;10:2820.

[16] Jurtz VI, Jessen LE, Bentzen AK, Jespersen MC, Mahajan S, Vita R, et al. NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. BioRxiv 2018:433706.

[17] Isacchini G, Walczak AM, Mora T, Nourmohammad A. Deep generative selection models of T and B cell receptor repertoires with soNNia. Proc Natl Acad Sci USA 2021;118(14):e2023141118.

[18] Vig J, Madani A, Varshney LR, Xiong C, Socher R, Rajani NF. BERTology meets biology: interpreting attention in protein language models. arXiv Prepr 2020.

[19] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA 2021;118(15):e2016239118.

[20] Kalakoti Y, Yadav S, Sundar D. TransDTI: transformer-based language models for estimating DTIs and building a drug recommendation workflow. ACS Omega 2022; 7(3):2706–17.

[21] Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. Language models enable zero-shot prediction of the effects of mutations on protein function. Adv Neural Inf Process Syst 2021;34:29287–303.

[22] Littmann M, Heinzinger M, Dallago C, Olenyi T, Rost B. Embeddings from deep learning transfer GO annotations beyond homology. Sci Rep 2021;11(1):1–14.

[23] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379(6637): 1123–30.

[24] Montemurro A, Schuster V, Povlsen HR, Bentzen AK, Jurtz V, Chronister WD, et al. NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. Commun Biol 2021;4(1):1060.

[25] Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, Crawford JC, Dolton G, et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. Nucleic Acids Res 2018;46(D1):D419–27.

[26] Lu T, Zhang Z, Zhu J, Wang Y, Jiang P, Xiao X, et al. Deep learning-based prediction of the T cell receptor–antigen binding specificity. Nat Mach Intell 2021; 3(10):864–75.

[27] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom 2020; 21(1):6.

[28] Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Min 2023;16(1):4.

[29] Hendrycks D, Gimpel K. Gaussian error linear units (gelus). arXiv Prepr 2016.

[30] Chicco D, Tötsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min 2021;14(1):1–22.

[31] Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. PLOS One 2017;12(6):e0177678.

[32] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc: Ser B (Methodol) 1995;57(1): 289–300.