Rational Multi-Modal Transformers for TCR-pMHC Prediction

Jiarui Li
jli78@tulane.edu
Department of Computer Science
Tulane University
New Orleans, Louisiana, USA

Zixiang Yin
zyin@tulane.edu
Department of Computer Science
Tulane University
New Orleans, Louisiana, USA

Zhengming Ding
zding1@tulane.edu
Department of Computer Science
Tulane University
New Orleans, Louisiana, USA

Samuel J. Landry
landry@tulane.edu
Department of Biochemistry and Molecular Biology
Tulane University School of Medicine
New Orleans, Louisiana, USA

Ramgopal R. Mettu
rmettu@tulane.edu
Department of Computer Science
Tulane University
New Orleans, Louisiana, USA

Abstract

T cell receptor (TCR) recognition of peptide-MHC (pMHC) complexes is fundamental to adaptive immunity and central to the development of T cell-based immunotherapies. While transformer-based models have shown promise in predicting TCR-pMHC interactions, most lack a systematic and explainable approach to architecture design. We present an approach that uses a new post-hoc explainability method to inform the construction of a novel encoder-decoder transformer model. By identifying the most informative combinations of TCR and epitope sequence inputs, we optimize cross-attention strategies, incorporate auxiliary training objectives, and introduce a novel early-stopping criterion based on explanation quality. Our framework achieves state-of-the-art predictive performance while simultaneously improving explainability, robustness, and generalization. This work establishes a principled, explanation-driven strategy for modeling TCR-pMHC binding and offers mechanistic insights into sequence-level binding behavior through the lens of deep learning.

CCS Concepts

• **Applied computing** → **Molecular structural biology**; **Molecular sequence analysis**; • **Computing methodologies** → *Information extraction*; **Neural networks**.

Keywords

CD4+ T cell response, epitope prediction, explainable AI, multi-modal learning, transformer models, deep learning

ACM Reference Format:
Jiarui Li, Zixiang Yin, Zhengming Ding, Samuel J. Landry, and Ramgopal R. Mettu. 2025. Rational Multi-Modal Transformers for TCR-pMHC Prediction. In *Proceedings of the 16th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '25), October 11–15, 2025, Philadelphia, PA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3765612.3767194

ACM ISBN 979-8-4007-2200-4/2025/10
https://doi.org/10.1145/3765612.3767194

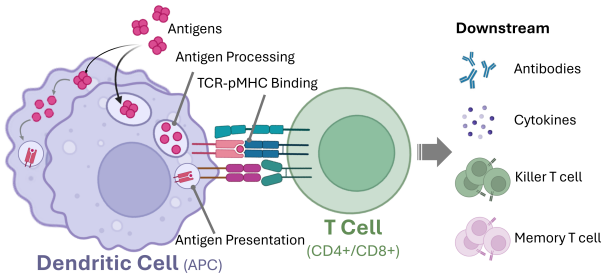

Figure 1: Binding between the peptide-MHC complex and T cell receptors is fundamental to understanding adaptive immune response and especially for developing immunotherapies (figure created in https://BioRender.com).

1 Introduction

T cells are essential components of the adaptive immune system, responsible for recognizing and responding to antigenic proteins from pathogens, such as viruses, bacteria, and cancer cells, as well as self-antigens in autoimmune contexts [12]. A key event in the T cell immune response is the binding between the T cell receptor (TCR) and the peptide-Major Histocompatibility Complex (pMHC), where the MHC molecule presents an antigenic peptide (i.e., epitope) on the surface of antigen presenting cells (APC). This highly specific interaction is foundational to T cell-mediated immunity (see Figure 1) and remains a focal point in both basic immunological research and immunotherapy development. In recent years, understanding and leveraging T cell responses has become a crucial aspect of designing durable vaccines and advancing personalized cancer immunotherapies [27, 28].

Accurate T cell response prediction requires modeling both peptide presentation and TCR recognition [24]. Early computational efforts emphasized peptide-MHCII binding prediction using allele-specific machine learning models [24], such as NetMHCpan [9] and NetMHCcons [13]. More recent approaches incorporate antigen processing through the Antigen Processing Likelihood (APL) algorithm [5, 17, 18], which models the influence of antigen structure and its influences on peptide availability for MHCII binding.

The TCR-pMHC binding prediction problem can be formulated as a binary classification task: given a TCR sequence (in whole or selected components) and an antigenic peptide (with known MHC allele) as input, we must predict whether the TCR will bind to the pMHC complex. Both unsupervised and supervised approaches have been explored [10]to analyze sequencing data from TCR-pMHC assays. Earlier unsupervised methods cluster TCR repertoires via dimensionality reduction and CDR-based similarity metrics (e.g., TCRdist3 [21]), without requiring binding or epitope labels (e.g., GIANA [35]). Clusters obtained from these analyses are then used for downstream analysis [10]. In contrast, more recent supervised methods leverage labeled TCR-pMHC data from resources such as VDJdb [4], McPAS-TCR [30], and IEDB [32] to directly predict binding. Models such as MixTCRpred [7], NetTCR2.2 [11], and TULIP [22] utilize deep learning architectures to achieve robust predictive performance and generalization.

Experimental data from TCR-pMHC binding assays may include multiple input modalities, such as full TCR sequences, complementarity-determining regions (CDRs), and epitope sequences. Prior studies have established CDR3 as the most critical determinant of binding [8], motivating state-of-the-art models (e.g., MixTCRpred [7], BERTrand [23], Cross-TCR-Interpreter [15], and TULIP [22]) to rely solely on CDR3 and epitope inputs. However, non-CDR3 regions have also been shown to contribute to binding prediction [8]. Moreover, existing models either concatenate all sequences into a single input [7, 23] or apply cross-attention exhaustively across all input pairs [15, 22], without an explicit structural organization.

In this paper, we present a principled approach to designing transformers with improved performance and stronger generalization for TCR-pMHC prediction by using a new method for explainability [19] that helps us understand the functional roles of each input and the internal dynamics of transformer-based architectures. Model explanation provides insight into why a deep learning model performs well or poorly, enabling principled analysis of how different architectural choices (i.e., cross-attention) affect model behavior. This forms the basis for a model optimization strategy driven by explainability as shown in Figure 2. We decompose the construction of a transformer-based TCR-pMHC model into four key stages: (1) input modality selection, (2) cross-attention design for multi-modal fusion, (3) loss function strategy design, and (4) training strategy design.

We train and test the models obtained from our approach with several TCR-pMHC datasets [1, 3, 4, 30, 32]. We further evaluate model generalization and explainability on the IMMREP23 [25] sequence benchmark and our TCR-XAI [19] structural benchmark. While CDR regions play the primary role in TCR-pMHC binding prediction [8], our analysis demonstrates that non-CDR regions empower model encoding of the relationships between CDR regions, resulting in enhanced performance. We also explore cross-attention between CDR3b and epitope features and identify patterns of cross-attention that quantitatively improves model understanding of these modalities. Next, we demonstrate the potential that incorporating auxiliary losses and the explanation-based model training strategy can further improve generalization. We use these findings to develop a model ("EGM-2") that achieves state-of-the-art performance and generalization. We achieve approximately 4-6% improvements in AUC over methods such as TULIP, BERTrand and MixTCRPred in 5-fold cross validation and on the IMMREP23 [25] test set. Over our structure-based TCR-XAI [19] benchmark, we use a performance metric called binding region hit rate (BRHR) that relates model explainability with ground truth. EGM-2 achieves about a 10% improvement in BRHR over existing methods on the TCR-XAI benchmark.

2 Background

The TCR-pMHC binding prediction problem can be formulated as a binary classification task: given a TCR composed of alpha (α) and beta (β) chains, an epitope e, and an MHC molecule m, the model predicts whether the pair binds (binder) or does not bind (non-binder). In this section, we introduce transformer-based architectures and highlight their application in TCR-pMHC binding prediction. Finally, we describe the selected post-hoc explainable AI (XAI) techniques used to interpret these models.

2.1 TCR-pMHC Prediction with Transformers

A standard transformer architecture consists of two main components: the encoder and the decoder [31]. The encoder extracts and transforms features from the inputs, while the decoder fuses these features, particularly through cross-attention mechanisms, enabling the modeling of interactions between different modalities.

TCR-pMHC binding prediction inherently involves the interaction between TCR and the pMHC complex. Consequently, encoder-decoder architectures such as TULIP [22] and Cross-TCR-Interpreter [15] have demonstrated strong performance by explicitly modeling such interactions. However, these models typically limit their input to the CDR3 regions and epitope sequence, allowing for straightforward application of cross-attention between each input pair. In contrast, models such as MixTCRpred [7], which incorporate all CDR regions along with the epitope, face increased complexity in applying cross-attention exhaustively between every pair of inputs. While this approach yields good performance, interpretability is difficult to pinpoint the key architectural contributions.

2.2 Post-hoc Explainability for Transformers

As with most deep learning methods, transformers are "black boxes" and pose significant challenges when relating the predictions to input features. Thus post-hoc explainable AI (XAI) is an intense area of study. Initial methods focused on CNNs and other architectures [29]. Recent work has developed methods for transformers [2, 19]. TEPCAM [6], raw attention [33] and a method we have recently developed named QCAI [19], have been used for TCR-pMHC models.

In the context of transformer-based architectures, AttnLRP has demonstrated state-of-the-art performance for encoder-only models [2], while QCAI has proven effective for multi-modal encoder-decoder models [19]. Therefore, in this work, we use these methods to interpret and analyze transformer models applied to TCR-pMHC binding prediction.

3 Our Approach and Results

Since this paper focuses on rational development of models, we interleave method development with experimental results, providing intermediate analyses to justify various aspects of model design. We

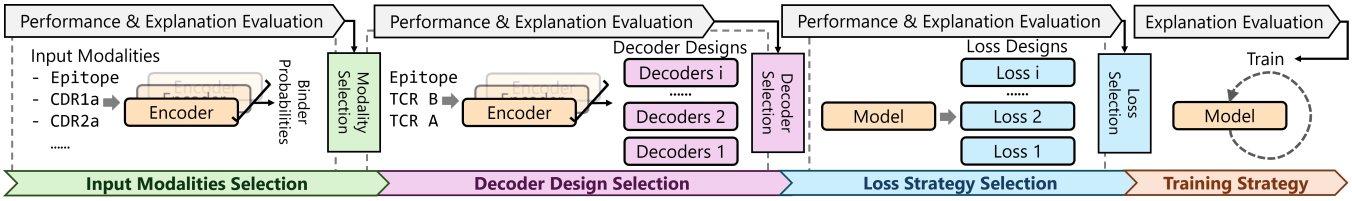

Figure 2: The development of a multi-modal transformer model for TCR-pMHC binding prediction can be systematically decomposed into four key components: input modality selection, decoder architecture design, loss function strategy, and training methodology. We employ post-hoc explainability analysis to evaluate each design choice and formulate an explanation-guided strategy across these components, ultimately resulting in a state-of-the-art predictive model.

decompose the transformer model design for TCR-pMHC binding prediction into four key components: (1) input modality selection, (2) cross-attention design, (3) loss function design, and (4) model training strategy design. For each component, we analyze how various design choices influence model behavior and identify the most effective configurations or improve the model, supported by explainability analyses. Based on our analyses, we obtain pan-allele TCR-pMHC binding models that achieve improved performance, explainability and generalizability over current approaches such as TULIP, BERTrand and MixTCRPred.

3.1 Model and Training Configuration

To control for confounding variables, we constructed all models using standard, non-pretrained BERT modules from the Huggingface transformers library. For the input modality selection, we used encoder-only transformers. For each input modality combination, an independent and identical encoder with masked language modeling loss (MLM) is applied to each input modality, and the output features are concatenated together and transformed by linear layers to predict binder or non-binder. For cross-attention design, the input features are processed by encoders following the same configuration as in input modality selection. For loss strategy design, each auxiliary loss is linked to the independent linear layers transforming the transformers' output features.

Each encoder and decoder module consists of two hidden layers with 128-dimensional hidden states and a single attention head to minimize computational overhead. All models were trained for 500 epochs using the AdamW optimizer with a learning rate of $1E^{-4}$. When considering the IMMREP and TCR-XAI test sets, we examine an explanation-based training strategy (Section 3.6). All training was performed on a machine equipped with two NVIDIA A2000 GPUs and two Intel E5 CPUs.

3.1.1 Datasets. To train and evaluate model performance, we collected a positive dataset following the procedure described in MixTCRpred [7], aggregating TCR-pMHC binding data from both *Homo sapiens* and *Mus musculus* across multiple sources: VDJdb [4], McPAS-TCR [30], IEDB [32], 10X Genomics [1], Andreatta et al. [3], and Zander et al. [34]. Negative samples were generated by pairing TCRs with non-binding pMHCs, maintaining a 1:1 ratio of negative to positive examples. For each epitope, we sampled an equal number of negative pairs to ensure class balance. Model performance was evaluated using 5-fold cross-validation.

3.2 Evaluation Metrics

For each model, we first evaluate its performance using 5-fold cross validation on the compiled training dataset. Then, to assess generalization, we train models on the full training set and evaluate them on IMMREP23 [25], a public benchmark for TCR-epitope specificity prediction that includes peptides unseen during training. To probe the internal mechanisms of the models and understand how they interpret input features, we applied post-hoc explainability methods: AttnLRP [2] for encoder-only models and QCAI [19] for encoder-decoder models. We only use binder classification loss to generate explanations for encoder-only models and use the training loss for encoder-decoder models. To generate attention weights, we consider all "not available" values (NA) as 0, and apply a smoothing operation using a convolution operation with core $[1/3, 1/3, 1/3]$ to tolerate one residue offset.

3.2.1 TCR-XAI Benchmark. Explanation quality was assessed using our recently developed TCR-XAI benchmark [19], which quantifies how well model-generated importance scores align with structural ground truth. It consists of 274 high-resolution crystal structures of TCR-pMHC complexes from the STCRDab [16] and TCR3d 2.0 [20] datasets. The availability of these structures gives us a means to objectively evaluate both accuracy and explainability. We use the Binding Region Hit Rate (BRHR) [19] to assess explanation quality. This score reflects how effectively the explanation method identifies actual binding residues based on structural proximity. Intuitively, BRHR compares top-ranked residues by explanation score against top interacting residues by distance. To calculate BRHR, we choose a percentile threshold $t \in (0, 1]$ and select the top t fraction of residues with the highest importance scores $\mathbf{S}$. A residue is counted as a *hit* if its structural interaction distance also falls within the top t fraction. For each sequence type of each sample that is predicted as a binder by a given model, we compute the individual hit rate, then average these values across the dataset (TCR-XAI) to produce the final BRHR for that model. In this paper, we use $t = 0.25$ to decide whether a residue is correctly identified as involved in binding; this is the most strict threshold that produces at least one binding region in every sample in the TCR-XAI set. We have tested other thresholds exhaustively and find similar results for all experiments in this paper.

3.3 Input Modality Selection

Most existing models use only the CDR3 regions and epitope sequences as input [15, 22]. Although the CDR3b region is widely acknowledged as a key determinant of TCR-pMHC interaction, recent studies suggest that non-CDR3 and even non-CDR regions may also contribute meaningfully to TCR-pMHC binding [8]. Therefore, the contributions of other TCR components have not been thoroughly investigated.

To address this gap, we conduct two sets of experiments: one to assess how different CDR regions affect model behavior, and another to examine the impact of both CDR and non-CDR regions on TCR-pMHC binding prediction performance.

Table 1: The ROC-AUCs of transformer models evaluated across different combinations of input modalities. Boldfaced values (0.902 and 0.755) denote the best 5-fold and test performance; the need to include both CDR3b and epitope sequences is evident.

Input Modalities	5-Fold	Test
CDR3b only	0.488±0.007	0.505
Epitope only	0.507±0.005	0.500
CDR3b + Epitope	**0.704±0.008**	0.517
All CDR3s + Epitope	0.756±0.010	0.607
All CDRs + Epitope	0.847±0.004	0.694
TCR A + All CDRbs + Epitope	0.899±0.005	0.751
TCR B + All CDRas + Epitope	0.893±0.004	**0.755**
TCRs + Epitope	0.867±0.003	0.753
TCRs + All CDRs + Epitope	**0.902±0.003**	0.738

3.3.1 CDR Regions. To evaluate how various CDR regions influence model behavior, we trained five encoder-only models using different combinations of CDRs and epitope. This design mirrors input modality combinations commonly adopted in prior work. The input modalities and the related representative TCR-pMHC prediction models are as follows: CDR3b only (e.g., TCRdist3 [21], GIANA [35]), Epitope only, CDR3b + Epitope (e.g., BERTrand [23], epiTCR [26]), CDR3s + Epitope (e.g., TULIP [22], TSpred-Attention [14]), and All CDRs + Epitope (e.g., MixTCRPred [7], NetTCR-2.2 [11], TSpred-CNN [14]). For each configuration, we used separate encoder modules to extract features from each modality. The resulting embeddings were then concatenated and passed through a linear classification layer to predict binding outcomes.

As shown in Table 1, the ROC-AUC results from both 5-fold cross-validation and independent test set evaluation indicate that incorporating additional input modalities than only using CDR3b and epitope improves model performance from 0.704 to 0.847 and generalization ability from 0.517 to 0.694. Notably, the model requires at least both the epitope and CDR3b as inputs to develop a valid understanding of TCR-pMHC binding.

An interesting observation is that while combining CDR3b and epitope inputs leads to a substantial performance improvement over using either modality alone in 5-fold cross-validation, the generalization ability improves by only 1.7%. However, when using both CDR3a and CDR3b as input, generalization ability improves more than 10%. To understand this discrepancy, we analyzed attention intensity and model explanation results (Figure 3). The analysis reveals that the inclusion of epitope and CDR3b as inputs lead to model bias attention on CDR3b and limited model interpretability. However, using both CDR3a and CDR3b as input significantly enhances the model's ability to interpret the CDR3b region. This suggests that while epitope and CDR3s are essential for the model to learn meaningful binding representations, the inclusion of additional CDRs (particularly CDR3a) enables the model to better learn the binding pattern of the epitope and improves mutual understanding between CDR3a and CDR3b under binding scenarios.

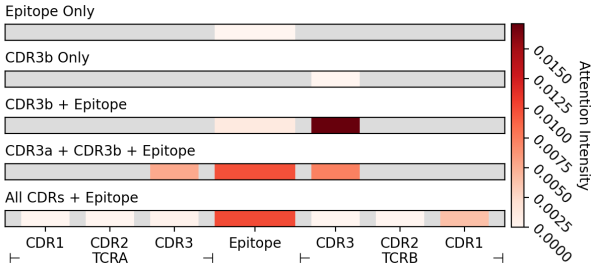

Figure 3: The sample-wise average attention intensities across different TCR and epitope regions from transformers with various CDR input modalities.

However, when all CDR regions are used as inputs, the average attention intensity on CDR regions decreased and explanation average BRHR slightly decline about 0.01. The model appears to struggle with encoding the relationship between CDR regions effectively. Although this configuration achieves improved performance and generalization, the explainability, particularly in terms of epitope attention and explanation BRHR, declined by 0.08. These findings suggest that if we can better structure the input from all CDR regions, it may be possible to further improve both model performance and generalization.

3.3.2 Full TCR Sequences. To improve the model's ability to organize input from the CDR regions and to investigate the contribution of non-CDR regions to TCR-pMHC binding prediction, we extend the input modalities to include full TCR sequences. These models follow the same configuration as those used in the CDR region experiments. The input modalities for each configuration are as follows: (1) All CDRs + Epitope, (2) TCR A sequence + All CDRb regions + Epitope, (3) TCR B sequence + All CDRa regions + Epitope, (4) Full TCR sequences + Epitope, and (5) Full TCR sequences + All CDR regions + Epitope.

As shown in Table 1, while the 5-fold performance shows only moderate improvement to 0.867 from 0.847, the generalization ability increases substantially to 0.753 from 0.694. Notably, using only the full TCR A or TCR B chain still achieves ROC-AUCs of 0.751 and 0.755 respectively on the independent test set. However, when both full TCR sequences and all CDR regions are included as inputs, the model achieves an ROC-AUC of 0.902 on the 5-fold validation but a lower ROC-AUC of 0.738 on the independent dataset, indicating overfitting. These findings suggest that incorporating either TCR A or TCR B full sequence is sufficient to enhance both performance

and generalization. To better understand the underlying mechanism, we further analyzed the explanation quality and attention intensity of these models.

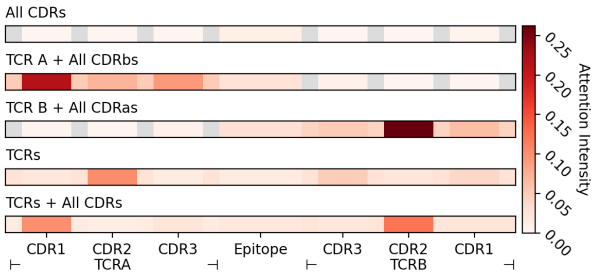

Figure 4: The sample-wise average attention intensities across different regions of TCRs and the epitope from transformers with different TCR input modalities.

As shown in Figure 4 and Table 1, incorporating full TCR chains allows the model to assign it higher attention intensity and better understand the epitope-TCR interaction with 0.04 BRHR improvement. In particular, the full TCR B chain enables the model to capture TCR B-epitope interaction more effectively with an 0.06 BRHR increase. However, the model gains a worse understanding of interactions between TCR A and TCR B with decreases of 0.02 and 0.05 in BRHR, respectively.

Consistent with previous findings, these results indicate that although full TCR sequences help the model learn more about the epitope and the model primarily relies on CDR regions for accurate TCR-pMHC binding prediction. However, the increased sequence length and complexity make it difficult for the model to process and organize all the information effectively. Therefore, it is crucial to develop strategies that structure and prioritize input information to improve model explainability and generalization.

3.4 Cross-Attention for Multi-Modal Fusion

To address complexity of solving feature relationships and dependencies, explicitly modeling the interactions between input features using a structured design can lead to improved performance and generalization. One effective approach to achieve this is by employing a decoder with cross-attention [31], which enables controlled and explainable information flow between different input components. Before constructing such a model, it is essential to understand the cross-attention mechanism within the decoder architecture.

3.4.1 Analysis of Decoder Cross-Attention. The decoder is composed of multiple layers, each containing a self-attention (encoder) layer and a cross-attention layer. The cross-attention mechanism allows one input embedding (the query) to attend to and extract information from another input or a concatenated set of inputs (the keys and values). This enables explicit modeling of interactions between distinct input modalities. However, it remains unclear whether cross-attention truly enhances the model's understanding of interactions between the query and the attended inputs, and what specific information is ultimately propagated through this mechanism. To investigate this, we design an experiment using CDR3b and epitope sequences as inputs, aiming to elucidate how cross-attention captures and represents their interaction.

Table 2: The ROC-AUCs of transformer models with various cross-attention designs for epitope and CDR3b. The cross-attention $a \rightarrow b$ only preserves information from b, because ROC-AUC of $a \rightarrow b + a$ is about 0.72 while $a \rightarrow b(+b)$ is near random.

Cross-Attentions	5-Fold
Epitope→CDR3b	0.520±0.008
Epitope→CDR3b + CDR3b	0.522±0.006
Epitope→CDR3b + Epitope	**0.732±0.006**
CDR3b→Epitope	0.484±0.004
CDR3b→Epitope + CDR3b	**0.718±0.007**
CDR3b→Epitope + Epitope	0.478±0.004
CDR3b↔Epitope	**0.718±0.007**

Table 3: The Binding Region Hit Rate (BRHR) for transformers with different cross-attention designs between the epitope and CDR3b. These results highlight the ability of cross-attention to directionally enhance the model's understanding of inter-modality interactions.

Modalities	Interact with	Epitope ↓ CDR3b	CDR3b ↓ Epitope	CDR3b + Epitope
Epitope	TCR B	0.7080	**0.7638**	0.6636
	CDR1b	0.6722	**0.7248**	0.6441
	CDR2b	0.6657	**0.7610**	0.6924
	CDR3b	0.7211	**0.7886**	0.6805
CDR3b	Epitope	**0.7369**	0.6122	0.7286

We investigate the directional behavior of the decoder's cross-attention by applying it between two input modalities: CDR3b and epitope. Specifically, we construct models in which one input modality serves as the query to attend to the other (e.g., CDR3b → epitope), and analyze performance both when using only the decoder output, where we denote $a \rightarrow b$ as using a as the query to attend to b. As shown in Table 2, using cross-attention alone between CDR3b and epitope yields similar performance to using either input independently around 0.5, essentially random guessing, suggesting that cross-attention tends to preserve information primarily from only one input.

To determine which input's information is retained, we include the original features of either the query or the attended input in the final prediction layer. We observe that combinations like epitope → CDR3b (with CDR3b features) and CDR3b → epitope (with epitope features) significantly improve performance to 0.732 and 0.718 respectively, whereas using the query features yields lower performance. This indicates that, in the cross-attention $a \rightarrow b$, the decoder primarily retains information from b, the attended modality.

To further investigate the explainability of cross-attention, we analyze explanation quality using binding region hit rate (BRHR) as

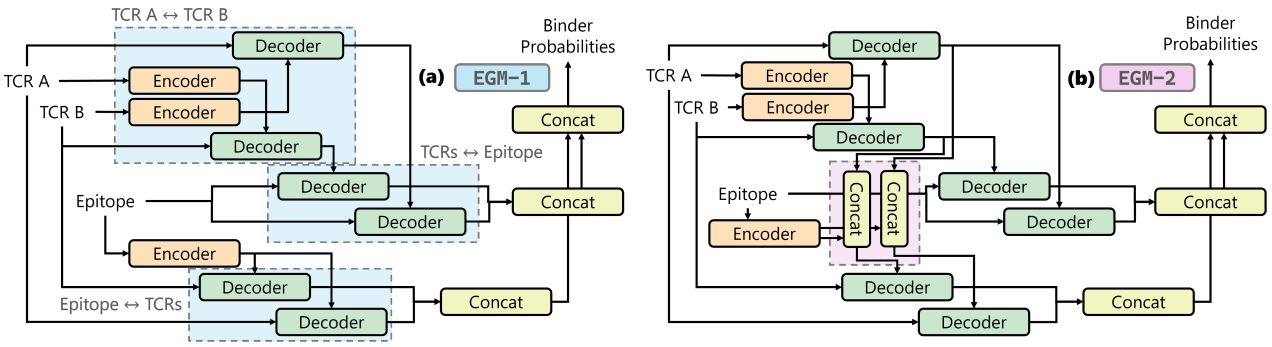

Figure 5: The architectures of the explanation-guided models (EGM). EGM-1 (a) includes additional decoders to capture internal TCR interactions first and independent decoders to capture epitope-TCR interactions. EGM-2 (b) is an improved version of EGM-1 with extra information from the other TCR chain and was developed by analyzing explanability.

shown in Table 3. For epitope → CDR3b, the model shows improved understanding from CDR3b to epitope with BRHR achieving 0.7369, and conversely, for CDR3b → epitope, the model better understands epitope to CDR3b interactions with BRHR reaching 0.7638. This directional explainability suggests that cross-attention $a \rightarrow b$ enhances the model's ability to capture interactions from b to a. Based on these observations, we propose to leverage directional cross-attention in encoder-decoder architectures to explicitly guide and enhance interaction modeling between TCR and pMHC, thereby improving both performance and generalization capability.

3.4.2 Explanation-Guided Cross-Attention Design. As demonstrated in our exploration of input modalities, incorporating full TCR chains enhances the model's understanding of both the epitope and the TCR itself. Therefore, we adopt TCR A, TCR B, and the epitope as input modalities for the design of our encoder-decoder architecture. The simplest approach to constructing such a model (EGM-0) is to apply direct cross-attention from one modality to the other two, following the design principle of TULIP [22]. This design enables the model to enhance its representation of a given modality by attending to complementary contextual information from the others.

Table 4: ROC-AUCs of explanation-guided models (EGM-1, EGM-2) versus baselines [7, 22, 23]. Our models consistently outperform baselines in 5-fold cross-validation and test set evaluation, demonstrating enhanced predictive performance and generalization.

Models	5-Fold	Test
CDR3s + Epitope (BERTrand [23])	0.704±0.008	0.517
CDR3s↔Epitope (TULIP [22])	0.803±0.002	0.566
All CDRs + Epitope (MixTCRPred [7])	0.847±0.004	0.694
EGM-0 (TCRs↔Epitope)	0.879±0.006	0.750
EGM-1	0.885±0.003	0.760
EGM-2	**0.888**±0.002	**0.765**

Table 5: The Binding Region Hit Rate (BRHR) for explanation-guided models. EGM-1 demonstrates improvement on inter-TCR interaction understanding and EGM-2 increases understanding among all interactions.

Modalities	Interact with	EGM-0 (TCR↔Epitope)	EGM-1	EGM-2
Epitope	TCR A	0.7019	0.7456	**0.7821**
	TCR B	0.6394	0.7207	**0.7341**
TCR A	Epitope	0.7981	0.7320	0.7404
	TCR B	0.7309	**0.7750**	**0.8024**
TCR B	Epitope	0.6457	0.6798	**0.8413**
	TCR A	0.6459	**0.6809**	**0.7742**

As shown in Table 4, applying direct cross-attention between TCR sequences and epitopes enhances 5-fold cross-validation performance to 0.879 but does not improve generalization ability, which keep 0.75. To investigate this, we leverage model explanations. According to Table 5, the model exhibits limited improved understanding of the interaction from TCR B to TCR A, which is smaller than 0.05. In addition, its understanding of interactions from TCR A and TCR B to epitope declines from 0.9109 to 0.7981 and from 0.9476 to 0.6457 respectively. These findings suggest two potential directions for further model improvement: (1) enhancing the model's ability to capture interactions from epitope to TCRs, and (2) improving its understanding of the interactions from TCRs to the epitope.

We find that EGM-0 exhibits insufficient understanding of the interaction between TCR A and TCR B, as indicated by BRHR scores of 0.65 (TCR B → TCR A) and 0.73 (TCR A → TCR B). Following the first strategy, we designed the initial version of our explanation-guided model (EGM-1) architecture, as illustrated in Figure 5a. To enhance the model's understanding of epitope-TCR interactions with extended representational capacity, we first apply cross-attention between TCR A and TCR B chains. Subsequently, the epitope sequence is used to query each TCR chain independently. This enables epitope-TCR interactions to be processed with independent decoders. Also, to further expand the representational capacity and improve predictive performance, we employ separate

decoders to perform cross-attention from both TCR A and TCR B to the epitope.

As shown in Table 4, this architecture effectively improves both cross-validation and generalization performance to 0.885 and 0.76 respectively. Explanation based analysis further reveals that the model develops a stronger understanding of TCR A-TCR B, epitope-TCR A, and epitope-TCR B with 4%, 3.5%, and 8% BRHR improvement. However, due to the use of independent decoders for TCR-to-epitope attention, the model's ability to capture joint TCR-epitope interactions is reduced.

Although EGM-1 improved modeling of TCR inter-chain interactions, its understanding of epitope-TCR interactions remains limited: the BRHR from TCR B to the epitope is only 0.68, significantly lower than other interaction BRHR scores. Building upon EGM-1 and incorporating the second improvement strategy, we designed a second version of the model, EGM-2, illustrated in Figure 5b. In EGM-2, to enhance the model's understanding of epitope interactions, we modify the decoder responsible for cross-attention from the TCR chains to the epitope. Specifically, we integrate additional features from the complementary TCR chain during cross-attention. This design allows the model to contextualize each TCR chain's interaction with the epitope in the presence of the other chain's information, thereby fostering a more comprehensive understanding of TCR-epitope binding patterns. According to the Table 4, it achieves 0.765 ROC-AUC on test dataset. With respect to explainability, as shown in Table 5, it demonstrates 10% BRHR improvement for all interactions in average. In particular, the BRHR of interaction between TCR B and epitope increases 20% to 0.8413.

3.5 Loss Strategies

Loss strategies are a significant component of transformer model design, guiding model optimization during training and influencing the representational capacity of the model. For all previously discussed models, we employed two types of loss functions: masked language modeling (MLM) loss and binder classification loss. To identify the most effective loss strategy for TCR-pMHC prediction, we conducted a two-step investigation: (1) evaluating the role of MLM loss, and (2) exploring potential auxiliary losses to enhance model performance. We used EGM-2 to explore loss strategies.

Table 6: Binding Region Hit Rate (BRHR) of EGM-2 under different loss strategies. MLM loss improves modality-level understanding, while MHC and TRVJ allele classification losses enhance interpretability for epitope recognition and inter-TCR interactions respectively.

Loss	MLM	-	✓	✓	✓
	Auxiliary	-	-	MHC	V/J
Modalities	Interact with				
Epitope	TCR A	0.7533	0.7821	**0.8166**	0.7581
	TCR B	**0.8075**	0.7341	0.5817	0.7018
TCR A	Epitope	**0.7498**	**0.7404**	0.6307	0.5508
	TCR B	0.6841	0.8024	0.8086	**0.8321**
TCR B	Epitope	0.7606	**0.8413**	**0.7910**	0.6869
	TCR A	0.6627	**0.7742**	0.7140	**0.7515**

3.5.1 Masked Language Modeling Loss. The masked language modeling (MLM) loss masks parts of the input and uses cross-entropy to evaluate how well the model can recover the masked tokens. In our previous models, we applied MLM loss to both encoders and decoders. To assess how MLM loss affects the binder classification task, we trained the designed model with encoders regularized by MLM loss, but decoders optimized solely with the binder classification loss. Although removing MLM loss from the decoders little decrease 5-fold validation and generalization ROC-AUC within 0.001 and 0.05 respectively. Explanation analysis in the Table 6 indicates that removing decoder MLM loss substantially impaired the model's understanding of the interaction from epitope to TCR A and from TCR B to epitope and TCR A. These results demonstrate that MLM loss enhances the decoder's ability to understand the input data, ultimately improving model robustness.

3.5.2 Auxiliary Loss. Based on the dataset composition, we identified two potential auxiliary losses: (1) MHC categories and alleles (MHC loss), and (2) V and J region alleles of TCRs (TRVJ loss). The MHC category can be formulated as a binary classification task (MHC-I vs. MHC-II), while both MHC alleles and the V/J region alleles of TCRs can be framed as auxiliary multi-class classification tasks. Incorporating these auxiliary losses slightly improves ROC-AUC on the test dataset about 0.005, which is minor and not significantly different ($p > 0.5$) to the model without auxiliary losses. However, according to the explanation evaluation in the Table 6, the TRVJ loss improves model understanding between TCRs and MHC loss enhances model's understanding from epitope to TCR A and TCR B to epitope. This suggests that these auxiliary objectives mediate models' behavior and affect the way models capture the interaction between TCRs and epitope. These experiments demonstrate that the MLM loss improves model understanding among all input modalities, auxiliary classification loss for MHC enhances the model's understanding of TCR A and epitope interaction, and V/J alleles auxiliary classification loss boosts the model's explanation between TCRs.

3.6 Training Strategy

We compared two model selection strategies: loss-based, choosing the model with the lowest training loss, and explanation-based, selecting the model with the highest explanation quality measured by bidirectional TCR-epitope interactions. While loss-based selection risks overfitting and validation-based stopping may fail under distributional shift, explanation-based metrics offer a promising alternative. From epoch 300 onward, both strategies performed similarly, but after convergence (epoch 350), explanation-based selection consistently yielded models with superior generalization and robustness. These results suggest that explanation-based metrics provide an effective stopping criterion for training, leading to models with improved generalization.

4 Conclusion

In this paper, through comprehensive experimentation guided by analysis of model explainability, we have identified and validated novel model designs that achieve state-of-the-art performance for TCR-pMHC prediction. We believe that the understanding gained in analyzing directional attention mechanisms will enable us to

build models based on *concepts* (i.e., an explainable sub-structure in the model) that effectively capture the interaction between input modalities.

Code and Data Availability: The code, models, and data introduced in this paper can be found at https://github.com/Tulane-Mettu-Landry-Lab/tcr-rational.

Acknowledgments

We thank the anonymous reviewers, area chairs, and program chairs for their valuable feedback. This work was supported by National Institutes of Health (U54-CA260581) "Tulane University COVID Antibody and Immunity Network (TUCAIN)", AWS Cloud Research Credits, and the Harold L. and Heather E. Jurist Center of Excellence for Artificial Intelligence at Tulane University.

References

[1] 10x Genomics. 2019. A new way of exploring immunity–linking highly multiplexed antigen recognition to immune repertoire and phenotype. *Tech. rep* 2019 (2019).

[2] Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2024. AttnLRP: attention-aware layer-wise relevance propagation for transformers. In *Int. Conf. Mach. Learn. (ICML'24)*. JMLR.org, Vienna, Austria, 34.

[3] Massimo Andreatta, Ariel Tjitropranoto, Zachary Sherman, Michael C Kelly, Thomas Ciucci, and Santiago J Carmona. 2022. A CD4+ T cell reference map delineates subtype-specific adaptation during acute and chronic viral infections. *Elife* 11 (2022), e76339.

[4] Dmitry V Bagaev, Renske MA Vroomans, Jerome Samir, Ulrik Stervbo, Cristina Rius, Garry Dolton, Alexander Greenshields-Watson, Meriem Attaf, Evgeny S Egorov, Ivan V Zvyagin, et al. 2020. VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *NAR* 48, D1 (2020), D1057–D1062.

[5] Tysheena Charles, Daniel L Moss, Pawan Bhat, Peyton W Moore, Nicholas A Kummer, Avik Bhattacharya, Samuel J Landry, and Ramgopal R Mettu. 2022. CD4+ T-Cell Epitope Prediction by Combined Analysis of Antigen Conformational Flexibility and Peptide-MHCII Binding Affinity. *Biochem.* 61, 15 (2022), 1585–1599.

[6] Junwei Chen, Bowen Zhao, Shenggeng Lin, Heqi Sun, Xueying Mao, Meng Wang, Yanyi Chu, Liang Hong, Dong-Qing Wei, Min Li, et al. 2024. TEPCAM: Prediction of T-cell receptor–epitope binding specificity via interpretable deep learning. *Protein Sci.* 33, 1 (2024), e4841.

[7] Giancarlo Croce, Sara Bobisse, Dana Léa Moreno, Julien Schmidt, Philippe Guillame, Alexandre Harari, and David Gfeller. 2024. Deep learning predictions of TCR-epitope interactions reveal epitope-specific chains in dual alpha T cells. *Nat. Commun.* 15, 1 (2024), 3211.

[8] James Henderson, Yuta Nagano, Martina Milighetti, and Andreas Tiffeau-Mayer. 2024. Limits on inferring T cell specificity from partial information. *Proc. Natl. Acad. Sci.* 121, 42 (2024), e2408696121.

[9] Ilka Hoof, Bjoern Peters, John Sidney, Lasse Eggers Pedersen, Alessandro Sette, Ole Lund, Søren Buus, and Morten Nielsen. 2009. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* 61 (2009), 1–13.

[10] Dan Hudson, Alex Lubbock, Mark Basham, and Hashem Koohy. 2024. A comparison of clustering models for inference of t cell receptor antigen specificity. *ImmunoInformatics* 13 (2024), 100033.

[11] Mathias Fynbo Jensen and Morten Nielsen. 2023. NetTCR 2.2-Improved TCR specificity predictions by combining pan-and peptide-specific training strategies, loss-scaling and integration of sequence similarity. *eLife* 12 (2023).

[12] Alok V Joglekar and Guideng Li. 2021. T cell antigen discovery. *Nat. Methods* 18, 8 (2021), 873–880.

[13] Edita Karosiene, Claus Lundegaard, Ole Lund, and Morten Nielsen. 2012. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* 64 (2012), 177–186.

[14] Ha Young Kim, Sungsik Kim, Woong-Yang Park, and Dongsup Kim. 2024. TSpred: a robust prediction framework for TCR–epitope interactions using paired chain TCR sequence data. *Bioinformatics* 40, 8 (2024), btae472.

[15] Kyohei Koyama, Kosuke Hashimoto, Chioko Nagao, and Kenji Mizuguchi. 2023. Attention network for predicting T-cell receptor–peptide binding can associate attention with interpretable protein structural properties. *Front. bioinform.* 3 (2023), 1274599.

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

[17] Jiarui Li, Samuel J Landry, and Ramgopal R Mettu. 2024. GPU Acceleration for Markov Chain Monte Carlo Sampling. In *Proc. 4th Int. Conf. AIML Syst.* ACM, New York, NY, USA, Article 14, 8 pages.

[18] Jiarui Li, Samuel J Landry, and Ramgopal R Mettu. 2024. GPU Acceleration of Conformational Stability Computation for CD4+ T-cell Epitope Prediction. In *IEEE Int Conf Bioinformatics Biomed.* IEEE, Lisbon, Portugal, 191–196.

[19] Jiarui Li, Zixiang Yin, Haley Smith, Zhengming Ding, Samuel J Landry, and Ramgopal R Mettu. 2025. Quantifying Cross-Attention Interaction in Transformers for Interpreting TCR–pMHC Binding. *arXiv preprint arXiv:2507.03197* 2025, 2507.03197 (2025).

[20] Valerie Lin, Melyssa Cheung, Ragul Gowthaman, Maya Eisenberg, Brian M Baker, and Brian G Pierce. 2025. TCR3d 2.0: expanding the T cell receptor structure database with new structures, tools and interactions. *NAR* 53, D1 (2025), D604–D608.

[21] Koshlan Mayer-Blackwell, Stefan Schattgen, Liel Cohen-Lavi, Jeremy C Crawford, Aisha Souquette, Jessica A Gaevert, Tomer Hertz, Paul G Thomas, Philip Bradley, and Andrew Fiore-Gartland. 2021. TCR meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs. *Elife* 10 (2021), e68605.

[22] Barthelemy Meynard-Piganeau, Christoph Feinauer, Martin Weigt, Aleksandra M Walczak, and Thierry Mora. 2024. TULIP: A transformer-based unsupervised language model for interacting peptides and T cell receptors that generalizes to unseen epitopes. *Proc. Natl. Acad. Sci.* 121, 24 (2024), e2316401121.

[23] Alexander Myronov, Giovanni Mazzocco, Paulina Król, and Dariusz Plewczynski. 2023. BERTrand—peptide: TCR binding prediction using Bidirectional Encoder Representations from Transformers augmented with random TCR pairing. *Bioinformatics* 39, 8 (2023), btad468.

[24] Morten Nielsen, Massimo Andreatta, Bjoern Peters, and Søren Buus. 2020. Immunoinformatics: predicting peptide–MHC binding. *Annu. Rev. Biomed* 3, 1 (2020), 191–215.

[25] Morten Nielsen, Anne Eugster, Mathias Jensen Fynbo, Manisha Goel, Andreas Tiffeau-Mayer, Aurelien Pelissier, Sebastiaan Valkiers, Rodríguez María Martínez, Barthélémy Meynard-Piganeeau, Victor Greiff, Thierry Mora, M. Aleksandra Walczak, Giancarlo Croce, L Dana Moreno, David Gfeller, Pieter Meysman, and Justin Barton. 2023. IMMREP23: TCR Specificity Prediction Challenge. https://kaggle.com/competitions/tcr-specificity-prediction-challenge. Kaggle.

[26] My-Diem Nguyen Pham, Thanh-Nhan Nguyen, Le Son Tran, Que-Tran Bui Nguyen, Thien-Phuc Hoang Nguyen, Thi Mong Quynh Pham, Hoai-Nghia Nguyen, Hoa Giang, Minh-Duy Phan, and Vy Nguyen. 2023. epiTCR: a highly sensitive predictor for TCR–peptide binding. *Bioinformatics* 39, 5 (2023), btad284.

[27] Mansour Poorebrahim, Niloufar Mohammadkhani, Reza Mahmoudi, Monireh Gholizadeh, Elham Fakhr, and Angel Cid-Arregui. 2021. TCR-like CARs and TCR-CARs targeting neoepitopes: an emerging potential. *Cancer Gene Ther.* 28, 6 (2021), 581–589.

[28] Luis A Rojas, Zachary Sethna, Kevin C Soares, Cristina Olcese, Nan Pang, Erin Patterson, Jayon Lihm, Nicholas Ceglia, Pablo Guasp, Alexander Chu, et al. 2023. Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer. *Nature* 618, 7963 (2023), 144–150.

[29] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE Int. Conf. Comput. Vis.* IEEE, Los Alamitos, CA, USA, 618–626.

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

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Adv Neural Inf Process Syst* 30 (2017).

[32] Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters. 2019. The immune epitope database (IEDB): 2018 update. *NAR* 47, D1 (2019), D339–D343.

[33] Kevin E Wu, Kathryn Yost, Bence Daniel, Julia Belk, Yu Xia, Takeshi Egawa, Ansuman Satpathy, Howard Chang, and James Zou. 2024. TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-binding analyses. In *Mach. Learn. Comput. Biol.*, Vol. 240. PMLR, Seattle, WA, USA, 194–229.

[34] Ryan Zander, Achia Khatun, Moujtaba Y Kasmani, Yao Chen, and Weiguo Cui. 2022. Delineating the transcriptional landscape and clonal diversity of virus-specific CD4+ T cells during chronic viral infection. *Elife* 11 (2022), e80079.

[35] Hongyi Zhang, Xiaowei Zhan, and Bo Li. 2021. GIANA allows computationally-efficient TCR clustering and multi-disease repertoire classification by isometric transformation. *Nat. Commun.* 12, 1 (2021), 4699.

Received 07 July 2025; revised 10 September 2025; accepted 27 August 2025