

Structure-Regularized Interpretable TCR-Epitope Prediction

Jiarui Li¹, Zixiang Yin¹, Yunbei Zhang¹, Janet Wang¹
Samuel J. Landry², Zhengming Ding¹, Ramgopal R. Mettu¹

¹Department of Computer Science, Tulane University

²Department of Biochemistry and Molecular Biology, Tulane University School of Medicine
{jli78, zyin, yzhang111, swang47, landry, zding1, rmettu}@tulane.edu

<https://tcrsrim.jiarui.li/>

Abstract: T cell receptor (TCR)-epitope binding prediction is essential for understanding adaptive immunity and developing immunotherapies. Existing sequence- and structure-based models often generalize poorly to unseen epitopes and provide limited interpretability. Furthermore, the impact of generated structures on model learning remains unclear. We present TCR-SRIM, a structure-regularized interpretable-by-design model that combines protein language model embeddings with interpretable contact prototypes to capture residue-level TCR-epitope interactions. TCR-SRIM achieves state-of-the-art predictive performance and improved interpretation quality on the TCR-XAI benchmark. Using its inherent interpretability, we further evaluate the effect of generated structures on model learning. While structures predicted by AlphaFold3, TCRModel2, and tFold-TCR yield competitive performance, they lead to less accurate interaction patterns and reduced binding-site diversity than experimentally-resolved structures. Our results highlight limitations of current structure prediction models for TCR-epitope learning and demonstrate the value of interpretable-by-design models for studying generated biological structures.

1 Introduction

In the adaptive immune system, T cells play a pivotal role in recognizing and responding to antigens derived from pathogens, such as viruses and bacteria [18], as well as in autoimmune settings. A key step of T cell activation is mediated by the binding between a peptide-Major Histocompatibility Complex (pMHC) and the T cell receptor (TCR). The specificity of this interaction underlies T cell-mediated immunity and is a central focus of therapeutic design and fundamental immunology. A comprehensive understanding of T cell responses is crucial for the development of vaccines and personalized cancer immunotherapies [38, 36].

As illustrated in Figure 1, antigens are first processed by antigen-presenting cells (APCs) and loaded onto MHC I or MHC II molecules, which are then presented on the APC surface as epitopes for recognition by CD8+ or CD4+ T cells, respectively [10, 31]. Subsequently, TCRs bind these pMHC complexes, initiating T cell activation. TCR recognition is mediated by the alpha and beta chains, each comprising variable (V), joining (J), and constant (C) regions, with the beta chain additionally containing a diversity (D) region [4]. Accurate prediction of T cell responses necessitates modeling antigen processing, presentation and TCR-epitope binding [34, 32].

While effective tools for MHC I and MHC II binding/presentation have been developed over the last two decades, the prediction of TCR-pMHC binding remains a central challenge in quantitative immunology and adaptive immune response modeling [15]. Existing approaches are primarily sequence-, structure-based, or a combination of both. Sequence-based methods include both unsupervised and supervised strategies [15, 16]: unsupervised approaches use similarity metrics such as TCRdist3 [28] on CDRs to cluster TCRs without binding or epitope labels (e.g., GIANA [47],

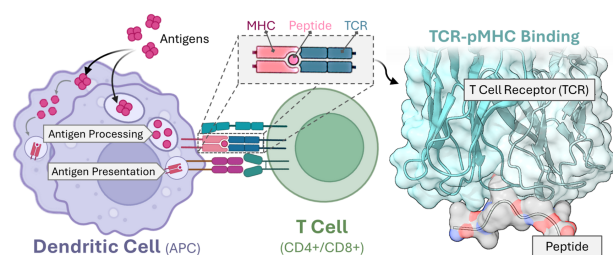


Figure 1: The interaction between peptide-MHC complexes and T cell receptors is central to adaptive immune responses and critical for the design of vaccines and immunotherapies (figure created in <https://www.biorender.com/>).

GLIPH2 [14]), whereas supervised models use curated TCR-epitope datasets [15] such as VDJdb [3], McPAS-TCR [43], and IEDB [44], together with deep learning methods (e.g., MixTCRpred [8], NetTCR2.2 [17], TULIP [29]).

Because experimentally resolved TCR-epitope structures remain scarce in databases such as TCR3D2.0 [23] and STCRDab [20], recent structure-based and sequence-structure hybrid models, including NetTCR-struc [11] and STAG-LLM [41], rely on predicted structures from models (e.g., TCRModel2 [46], tFold-TCR [45], and AlphaFold3 [2]). Despite utilizing structural information, these models still exhibit limited generalization on unseen-epitope benchmarks, while their black-box nature hinders the diagnosis of failure modes and systematic model improvement. To address interpretability, we recently introduced EGM [21], a state-of-the-art TCR-epitope predictor developed with QCAI [22], a post-hoc interpretability framework for multimodal TCR-epitope transformers. While EGM shows that interpretability can guide performance gains, it does not incorporate structural information. An alternative is interpretable-by-design models, which embed interpretability directly into their architectures [39]. Among these, prototype-based networks [6] learn representative patterns that both support prediction and provide faithful interpretations [6, 30]. PISTE [12] adopts this paradigm, but models interactions using predefined biological rules rather than structural information.

To integrate structural information with inherent interpretability, we propose TCR-SRIM, an interpretable-by-design model that bridges TCR-epitope sequence and structural data for strong generalization and built-in interpretability. TCR-SRIM embeds sequences with protein language models including ProteinBERT [5], ESM-1b [37], and ESM-2 [24]. It explicitly models residue-level CDR3-peptide interactions via *contact prototypes*, which are regularized with a small set of real or predicted structures while the model is trained on large-scale sequence data. TCR-SRIM outperforms all baselines: paired with ProteinBERT and regularized with real structures, it improves top-100 ROC-AUC by over 9% relative to TULIP and MixTCRpred, achieves comparable performance on a recent comprehensive TCR-epitope prediction benchmark [26], and surpasses existing interpretable methods on our TCR-XAI benchmark by $\sim 10\%$ in BRHR, a metric rewarding correct attribution to proximal residue contacts [22]. We further find that predicted structures match real ones in benchmark performance, yet their binding sites diverge from the truth and show homogeneous patterns lacking the diversity of real structures, which limits the generalization of models trained on them. Overall, TCR-SRIM achieves state-of-the-art prediction and interpretability while revealing a meaningful gap between predicted and real structures for model learning.

1.1 Problem Definition

This section introduces both sequence- and structure-based formulations for the TCR-epitope binding problem. TCR-epitope binding prediction is a binary classification task: given the CDR3a, CDR3b, and epitope, the model predicts whether they form a binding complex. We omit the MHC molecule, as its identity/sequence is not typically available. When provided sequences as input, the CDR3a, CDR3b, and epitope peptide are denoted $\alpha \in \mathbb{A}^n$, $\beta \in \mathbb{A}^n$, and $e \in \mathbb{A}^n$, respectively, where $\mathbb{A}$ is the amino acid alphabet and n is the maximum sequence length where inputs are padded as needed. The binding probability is defined as: $p_{\text{bind}} = P(\text{bind} \mid \alpha, \beta, e)$, and a pair is classified as positive if $p_{\text{bind}} > \eta$ with $\eta \in [0, 1]$.

When provided input structures, the CDR3a, CDR3b, and epitope peptide are denoted as $s^a, s^b, s^e \in \Theta$, with Θ representing the coordinate space for structures of size n . Each residue i is represented by the Cartesian coordinates of its atoms, $\theta_i = [(x_{i,j}, y_{i,j}, z_{i,j})]_{j=1}^{n_i}$, where n_i is the atom count of residue i . The structural resolution $r \in \mathbb{R}^+$ indicates the quality of the structure, with smaller r corresponding to higher resolution.

2 Our Approach

As shown in Figure 2, our approach first enhances and fuses features through cross-attention between inputs (i.e., TCR and peptide sequences), then models their contacts using contact prototype layers, and finally regularizes these prototypes with experimentally determined binding structures. The model alternates between standard training on sequence-only data and regularization with structures, thereby incorporating structural knowledge into the prediction model without requiring structures at training or inference. These components can be directly attached to protein language model (PLM) backbones, which provide embeddings for CDR3a, CDR3b, and peptide sequences, denoted as $\mathbf{E}_a \in \mathbb{R}^{n \times d}$, $\mathbf{E}_b \in \mathbb{R}^{n \times d}$, and $\mathbf{E}_e \in \mathbb{R}^{n \times d}$, where d is the embedding dimension.

2.1 Protein Language Models

To obtain richer representations of protein amino acid sequences, several pretrained, self-supervised, transformer-based foundation protein language models (PLMs) have been developed. ProteinBERT is trained on protein sequences and functional annotations, capturing both local and global features for downstream prediction tasks [5]. ESM-1b is a large-scale transformer pretrained on UniRef50, providing contextualized protein embeddings widely used for structure and function prediction [37]. ESM-2 improves upon this family with larger architectures and expanded pretraining, yielding stronger representations across diverse biological applications [24].

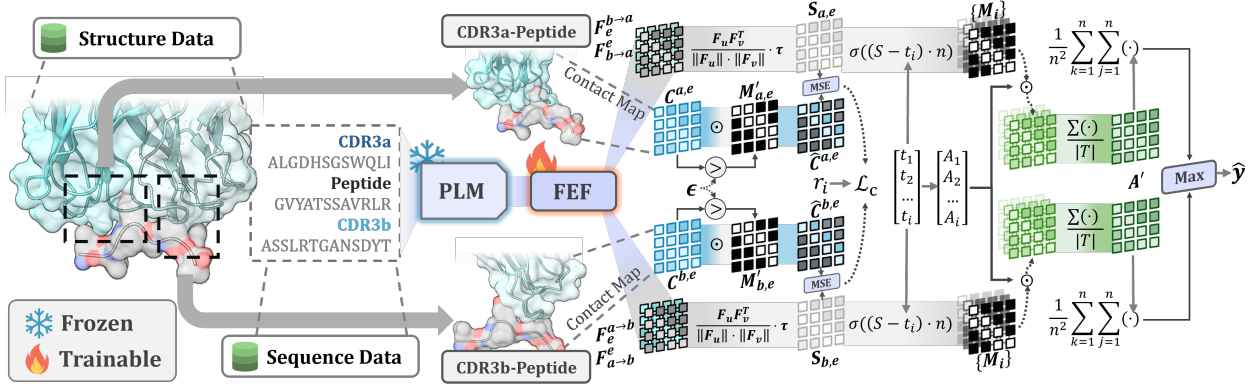


Figure 2: Overview of the proposed TCR-SRIM framework for regularizing contact prototypes using structures.

We extract features from these pre-trained PLMs to ensure our TCR-SRIM can work with different PLMs: ESM-1b [37], ESM-2 [24], and ProteinBERT [5]. ESM-1b is a 650M parameter model. We used the 8M, 35M, and 650M variants of ESM-2 to investigate how parameter scale affects our TCR-SRIM performance. ProteinBERT provides a single 16M parameter model, a PLM backbone that provides both local and global features.

2.2 Feature Enhancement and Fusion

In [21] we developed an Explanation-Guided Model (EGM), that used post-hoc analyses to inform cross-attention design for TCR-epitope binding prediction. We use the same design for TCR-SRIM and denote it as the feature enhancement and fusion module (FEF). For completeness we outline the design below. Formally, we denote cross-attention from arbitrary modalities q to p as $\mathcal{A}_{q \rightarrow p} : \mathbb{R}^{n \times d} \times \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^{n \times d}$ where q serves as the query and p as the key and value. We first obtain cross-fused representations of CDR3a and CDR3b using: $\mathbf{E}_{a \rightarrow b} = \mathcal{A}(\mathbf{E}_a, \mathbf{E}_b)$ and $\mathbf{E}_{b \rightarrow a} = \mathcal{A}(\mathbf{E}_b, \mathbf{E}_a)$. Subsequently, the peptide embeddings are fused further with $\mathbf{E}_{a \rightarrow b}$ and $\mathbf{E}_{b \rightarrow a}$ to obtain enriched features for TCR-epitope modeling:

$$\mathbf{F}_e^{a \rightarrow b} = \mathcal{A}(\mathbf{E}_e, \mathbf{E}_{a \rightarrow b}), \quad \mathbf{F}_{a \rightarrow b}^e = \mathcal{A}(\mathbf{E}_{a \rightarrow b}, \mathbf{E}_e), \quad \mathbf{F}_e^{b \rightarrow a} = \mathcal{A}(\mathbf{E}_e, \mathbf{E}_{b \rightarrow a}), \quad \mathbf{F}_{b \rightarrow a}^e = \mathcal{A}(\mathbf{E}_{b \rightarrow a}, \mathbf{E}_e). \quad (1)$$

2.3 Contact Prototype Layers

Residue-level contacts between TCR and pMHC are a key determinant of binding specificity. TCRdist, a widely used method for TCR-epitope prediction, defines similarity as a weighted mismatch distance between potential pMHC-contacting loops of two receptors [9]. Similarly, PISTE incorporates TCR-epitope contact rules into the attention mechanism to improve both predictive performance and interpretability [12]. Motivated by these approaches, we design prototype-based layers to explicitly model contacts between TCR and pMHC.

These layers estimate residue contacts between CDR3a and peptide, and between CDR3b and peptide, respectively. For two arbitrary fused embeddings $\mathbf{F}_u \in \mathbb{R}^{n \times d}$ and $\mathbf{F}_v \in \mathbb{R}^{n \times d}$ as defined in Eq. (1), the contact prototype layers take them as inputs and calculate the “contact area” between these chains. Inspired by the cross-attention mechanism, we model contact distance through cosine similarity as contact prototypes:

$$\mathbf{S}_{u,v} = \tau \cdot (\mathbf{F}_u \mathbf{F}_v^T) / (\|\mathbf{F}_u\| \|\mathbf{F}_v\|) \in [0, 1]^{n \times n}, \quad (2)$$

where $\|\cdot\|$ denotes L2 norm and $\tau \in \mathbb{R}^+$ is a learnable temperature parameter controlling the sharpness of the similarity distribution. A user-defined threshold set $T = [t_0, t_1, \dots, t_{|T|}]$ determines contact maps $\mathbf{M}_i := \sigma((\mathbf{S}_{u,v} - t_i) \cdot n) \in [0, 1]^{n \times n}$, with $\sigma(\cdot)$ denoting the sigmoid function. Contacts are aggregated by weighting each threshold with $A = \text{softmax}(T)$ and $A_i \in [0, 1]$ represents its i -th entry:

$$\mathbf{A}' = \left(\sum_{i=1}^{|T|} \sqrt{A_i} \mathbf{M}_i \right) / |T| \in [0, 1]^{n \times n}. \quad (3)$$

The overall contact score between $\mathbf{F}_u$ and $\mathbf{F}_v$ is:

$$w_{u,v} = \left(\sum_{k=1}^n \sum_{j=1}^n A'_{k,j} \right) / n^2 \in [0, 1]. \quad (4)$$

Defining $h : \mathbb{R}^{n \times d} \times \mathbb{R}^{n \times d} \rightarrow [0, 1]$ as the contact prototype function, $h(\mathbf{F}_u, \mathbf{F}_v) = w_{u,v}$, the final TCR-epitope contact score is:

$$\hat{y} = \max [h(\mathbf{F}_e^{b \rightarrow a}, \mathbf{F}_b^{a \rightarrow e}), h(\mathbf{F}_e^{a \rightarrow b}, \mathbf{F}_a^{b \rightarrow e})]. \quad (5)$$

We adopt a max operation to reflect that, because empirically, one chain often dominates peptide recognition and single TCR chain can support effective prediction [40].

2.4 Structure Regularization of Contact Prototype

To regularize the CDR3-peptide contact map, it is necessary to derive the ground-truth contact map from crystallographic structures. For any given pair of structures θ^p, θ^q from the protein structure space Θ , we define the contact between i -th residue θ_i^p in protein θ^p and the k -th residue θ_k^q in protein θ^q as $c_{i,k}^{p,q}$, computed as:

$$c_{i,k}^{p,q} = \min_{j=1}^{n_i} \min_{l=1}^{n_k} \sqrt{[(x_{i,j}^p - x_{k,l}^q)^2 + (y_{i,j}^p - y_{k,l}^q)^2 + (z_{i,j}^p - z_{k,l}^q)^2]} / 3 \in \mathbb{R}. \quad (6)$$

The full contact map between protein modalities p and q is then given by $\mathbf{C}^{p,q} = \{c_{i,k}^{p,q}\}_{i=1,k=1}^n \in \mathbb{R}^{n \times n}$. Using this approach, the ground-truth contact maps for CDR3a-peptide and CDR3b-peptide interactions are obtained as $\mathbf{C}^{a,e}$ and $\mathbf{C}^{b,e}$, respectively. The core region of residue-level contacts between TCRs and pMHCs are critical determinants of binding specificity. Motivated by this observation, we utilize the nearest residues between CDR3 regions and the peptide to regularize the model-derived contact prototype $\mathbf{S}_{u,v}$.

For a structural contact map $\mathbf{C}^{p,q}$, we first normalize its scale to align with that of the model-derived contact prototype and invert its values to transform distance measurements into contact importance. The normalized and inverted contact map $\hat{\mathbf{C}}^{p,q}$ is defined as:

$$\hat{\mathbf{C}}^{p,q} = 1 - [(\mathbf{C}^{p,q} - \min(\mathbf{C}^{p,q})) / (\max(\mathbf{C}^{p,q}) - \min(\mathbf{C}^{p,q}))] \in [0, 1]^{n \times n}. \quad (7)$$

We then identify spatially proximal residues by constructing a binary mask $\mathbf{M}'_{p,q} = \mathbb{1}[\hat{\mathbf{C}}^{p,q} > \epsilon] \in \{0, 1\}^{n \times n}$, where $\epsilon \in [0, 1]$ denotes the regularization threshold, which is a hyperparameter.

Then, we regularize the model-derived contact prototype using the top- ϵ important residues through a masked mean squared error loss:

$$\mathcal{L}^r = \left\{ \sum_{i=1}^n \sum_{j=1}^n [\mathbf{M}'_{p,q} \odot (\mathbf{S}_{u,v} - \hat{\mathbf{C}}^{p,q})^2]_{i,j} \right\} / \left[\sum_{i=1}^n \sum_{j=1}^n (\mathbf{M}'_{p,q})_{i,j} \right] \in \mathbb{R}, \quad (8)$$

where $\odot$ is element-wise product. Since structural resolutions vary across different samples, we incorporate resolution-based weighting to balance the regularization loss within each batch. Consider a batch of samples indexed by $B \subseteq \mathbb{Z}^+$ containing $|B|$ structures. For this batch, the corresponding resolutions are denoted as $R = [r_i]_{i \in B} \in \mathbb{R}^{|B|}$, and the regularization losses for each structure are represented as $\mathcal{L}_B^r = [\mathcal{L}_i^r]_{i \in B} \in \mathbb{R}^{|B|}$.

To incorporate structure quality into training, the resolution-weighted regularization loss ensures that higher-quality structures contribute more to the overall loss, which is defined as:

$$\mathcal{L}^{r'} = \sum_{i \in B} \left[\left(e^{-R} / \sum_{j \in B} e^{-r_j} \right) \odot \mathcal{L}_B^r \right]_i \in \mathbb{R}, \quad (9)$$

where lower-resolution values (i.e., higher-quality structures) correspond to higher importance weights to enable model focus on the high resolution samples. The overall objective function for structure-based regularization is defined as $\mathcal{L} = \mathcal{L}^{r'} + \mathcal{H}_{\text{CE}}(\hat{y}, y)$, where $\mathcal{L}^{r'}$ represents the resolution-weighted structural regularization loss, $\mathcal{H}_{\text{CE}}$ is the cross-entropy classification loss, and $y \in \{0, 1\}$ denotes the ground-truth binding label.

The integrated training procedure alternates between standard supervised training and structure-based regularization. Specifically, prior to each normal training epoch, the contact prototype module is regularized using the structural dataset for a number of epochs to incorporate structural constraints.

3 Experimental Analysis

We evaluate our model from three complementary perspectives: predictive performance, interpretation quality, and the impact of predicted versus experimental structures on accuracy and interpretability. Predictive performance and generalization ability are evaluated using appropriate ROC-AUC metrics on both our compiled datasets and a comprehensive benchmark. Interpretation quality is evaluated quantitatively using the binding region hit rate (BRHR) [22, 21] metric and qualitatively through representative case studies. Since our approach uses 3D structure in the regularization phase, to analyze the contributions of this step we compare the use of experimental structures with those predicted by AlphaFold3, TCRModel2, and tFold-TCR and evaluate performance and interpretation quality of corresponding prototypes to evaluate the impact of predicted structures to prediction models. Finally, to validate the effectiveness of model design choices, we use ablation studies to characterize the impact of structural regularization, mixing strategies, regularization thresholds, and regularization frequency, which are presented in Appendix C.

3.1 Training and Test Datasets with Unseen Epitopes

We constructed a TCR-epitope dataset containing 349,716 paired sequences of TCR alpha and beta chains, covering 2,316 unique peptides, 29,581 CDR3a, and 32,578 CDR3b sequences from Homo sapiens and Mus musculus. Of these, 95.7% are MHC-I and 4.3% MHC-II. Data were collected from VDJdb [3], McPAS-TCR [43], IEDB [44], TBAdB [48], and 10× Genomics [1], retaining only entries with valid CDR3a, CDR3b, and peptide sequences. Following other TCR-epitope training protocols (e.g., MixTCRpred [8], NetTCR-2.2 [17]) we generated negative samples by shuffling TCR-epitope pairs (4:1 ratio); we also directly sampled from 10× Genomics dataset, which contains negative binding data. The dataset was split 95:5 into training and test sets, where the test set (15,503 samples) includes 288 unseen epitopes. To assess generalization to dissimilar epitopes, we sampled evaluation sets with minimal Levenshtein distances between peptides ranging from 1 to 9 following the approach taken by TULIP [29].

3.2 Epitope-wise ROC-AUC Analysis

Following TULIP [29], we report epitope-wise ROC-AUC on our compiled dataset. Unlike the commonly reported overall ROC-AUC (aggregated across all epitopes), epitope-wise ROC-AUC evaluates each epitope and its associated TCRs separately, giving a more informative measure of generalization to unseen epitopes, where the aggregate scores for all models are only about 0.55. We also provide an independent benchmark evaluation in the next section. Following established protocols [33], we constrain the false positive rate to 0.1 per epitope and evaluate on a test set of exclusively unseen peptides; the top- k epitopes by ROC-AUC are reported in Table 1.

We compare against MixTCRpred [8], TULIP [29], and PLMs with linear classifiers trained on our datasets. We exclude PISTE [12], which targets HLA epitope prediction and requires HLA typing unavailable for most of our samples. TCR-SRIM consistently outperforms all baselines, surpassing TULIP and MixTCRpred across every setting and exceeding PLM backbones with linear classifiers by over 10% on average. With ProteinBERT, it attains an ROC-AUC of 0.989 on the top-100 epitopes, and roughly 9% and 17% above MixTCRpred and TULIP, respectively. We also compare against EGM [21], which shares the same design as FEF; even here, TCR-SRIM with ProteinBERT improves by 3-5%. These gains show that structure-based regularization effectively guides the model toward biologically meaningful CDR3-peptide contact patterns, advancing performance in TCR-epitope recognition.

3.3 Evaluation on Independent Comprehensive Benchmark

A large-scale, independent dataset has been proposed to evaluate TCR-epitope binding prediction [26], which we refer to as the “comprehensive” benchmark throughout this paper for clarity. This dataset provides two evaluation settings for seen epitopes: one using only CDR3b, and another incorporating CDR3b along with additional features. Since our method utilizes both CDR3b and CDR3a as inputs, we adopt the latter setting (CDR3b with additional features) for evaluation [26]. We directly use the AUPRC (Area Under the Precision-Recall Curve) and ROC-AUC (Area Under the Receiver Operating Characteristic) reported in Lu et al. [26] for PISTE (PISTE-reftcr) [12], vibtr (vibtr-AB) [13], and MixTCRpred [8].

Table 2 shows that TCR-SRIM with a ProteinBERT backbone and one epoch of structure regularization (TCR-SRIM (ProteinBERT)-1R) achieves an ROC-AUC of 0.571, establishing state-of-the-art performance. Meanwhile, TCR-SRIM with ESM2-150M achieves an AUPRC of 0.587, comparable to vibtr (0.588) and outperforming all other models. We also evaluate TCR-SRIM with other PLM backbones and more models in Appendix A.4.

Table 1: Comparison of ROC-AUC scores with the false positive rate restricted to 0.1 across the top-100, top-150, top-200, and top-250 peptides. TCR-SRIM yields improvements of approximately 9%, and 17% over MixTCRpred and TULIP, and achieves state-of-the-art.

PLM Backbone	Model	Top- k ROC-AUC @ FPR $\leq$ 0.1			
		100	150	200	250
MixTCRpred [8]		0.906	0.773	0.698	0.653
TULIP [29]		0.821	0.706	0.648	0.613
ProteinBERT [5]	Linear	0.772	0.675	0.625	0.595
	TCR-SRIM	0.989	0.871	0.773	0.713
ESM-1b [37]	Linear	0.900	0.795	0.716	0.668
	TCR-SRIM	0.986	0.862	0.766	0.707
ESM2-8M [24]	Linear	0.830	0.719	0.658	0.621
	TCR-SRIM	0.953	0.821	0.734	0.682
ESM2-35M [24]	Linear	0.783	0.685	0.632	0.600
	TCR-SRIM	0.977	0.838	0.747	0.692
ESM2-650M [24]	Linear	0.876	0.762	0.690	0.647
	TCR-SRIM	0.971	0.837	0.746	0.692

Table 2: Performance comparison of TCR-epitope binding prediction models on the comprehensive benchmark.

Models	AUPRC	ROC-AUC
TCR-SRIM (ProteinBERT)-1R	0.575	0.571
TCR-SRIM (ESM2-150M)	0.587	0.564
PISTE	0.567	0.564
vibtr	0.588	0.563
MixTCRpred	0.540	0.546

3.4 Evaluation of Interpretability on TCR-XAI Benchmark

To evaluate interpretability for TCR-epitope binding prediction, we constructed the TCR-XAI benchmark [22] from all currently available experimental TCR-epitope structures in STCRDab [20] and TCR3d 2.0 [23]. Retaining only structures with complete TCR α/β chains, full peptide sequences, intact CDR3 regions, and non-overlapping MHC and peptide chain IDs yields 274 high-resolution TCR-epitope complexes. We randomly split these 1:1 into structure-regularization and evaluation sets. For each evaluation sample, residue-level distances are computed as the minimal atom-wise distance (1) from each CDR residue to the peptide and (2) from each peptide residue to any CDR atom; smaller distances indicate stronger interactions and serve as interpretation ground truth.

Table 3: Comparison of TCR-SRIM on BRHR across different protein language model backbones, where $a \rightarrow b$ denotes the evaluation of residues in chain b with respect to chain a . TCR-SRIM shows improvements comparing to baselines, which demonstrates it effectively transforms structure information to sequence-based interpretable models. The interpretation of baseline models are obtained using post-hoc methods.

Models	Embeddings	Peptide	Peptide	CDR3a	CDR3b
	/ Post-hoc	↓ CDR3a	↓ CDR3b	↓ Peptide	↓ Peptide
MixTCRpred	AttnLRP	0.718	0.723	0.795	0.675
TULIP	QCAI	0.702	0.634	0.798	0.646
EGM	QCAI	0.782	0.734	0.740	0.841
TCR-SRIM	ProteinBERT	0.568	0.996	0.392	0.855
TCR-SRIM	ESM-1b	0.833	0.944	0.818	0.852
TCR-SRIM	ESM2-8M	0.604	0.801	0.804	0.816
TCR-SRIM	ESM2-35M	0.834	0.944	0.782	0.746
TCR-SRIM	ESM2-650M	0.961	0.820	0.788	0.809

We require a metric linking importance scores from an interpretability method to ground-truth binding, and thus construct one correlating the structural proximity of binding residues with importance scores. While proximal residues are not the exclusive drivers of binding, they play a necessary role in mediating TCR recognition; indeed, prior work such as *tcridist3* [28] uses proximity to define binding fitness. We therefore define the *Binding Region Hit Rate* (BRHR) [22] to measure how well an interpretation method identifies true binding residues. For a percentile threshold t , we take the top- t residues ranked by contact scores and the top- t ranked by peptide-TCR distances; a residue is a *hit* if it appears in both. Following prior work, we use $t = 0.25$, a stringent threshold guaranteeing at least one binding residue per instance. BRHR is computed per chain and averaged over all positive samples.

As shown in Table 3, TCR-SRIM consistently improves peptide-CDR3b BRHR across all backbones: for peptide $\rightarrow$ CDR3b, all models exceed 0.8, with ProteinBERT approaching 100%. The results also reveal differing backbone strategies. ProteinBERT is near-perfect on CDR3b-peptide interactions but weak on CDR3a-peptide (below 0.4), whereas ESM2-650M reaches ~ 0.96 on peptide $\rightarrow$ CDR3a despite moderate CDR3b-peptide performance.

3.5 Contact Prototype Visualization

To investigate the interpretations modeled by the contact prototype and their correspondence to true binding across all samples in the TCR-XAI evaluation subset, we extracted and visualized the contact prototypes. For each set of contact prototypes, valid distance values were aligned at the center of the padded contact prototypes and subsequently averaged to reveal the contact patterns captured by the contact prototypes.

As shown in Figure 3, we compared the contact prototypes of different parameter variants of ESM2 and ProteinBERT. All models correctly identified the most critical regions of CDR3b, which are closest to the peptide. For CDR3a, the ESM2 series demonstrated notable accuracy, whereas ProteinBERT failed to capture the correct pattern, consistent with the conclusions from our BRHR analysis. Furthermore, for ESM2 models, increasing the number of parameters led to higher “resolution” in the contact prototypes, indicating that larger models focus on specific regions rather than producing uniform contact prototypes. Notably, our TCR-SRIM with ProteinBERT produced a highly informative and

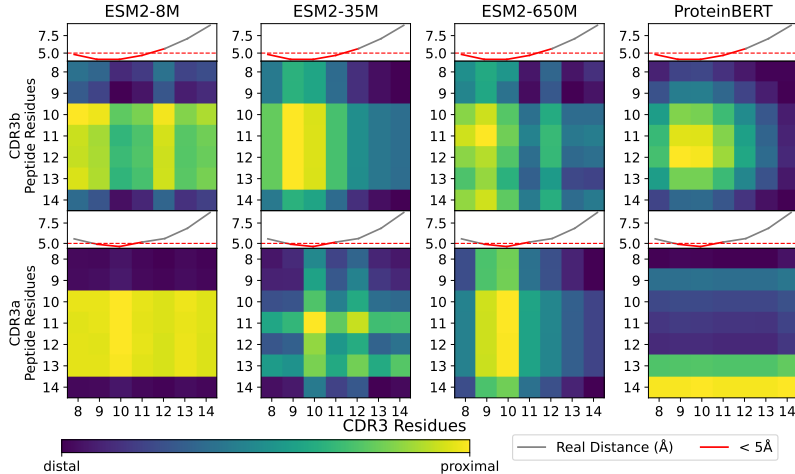


Figure 3: Visualization of contact prototypes from our TCR-SRIM with different PLM backbones. It demonstrates structure regularization effectively transfers contact patterns to sequence-based models.

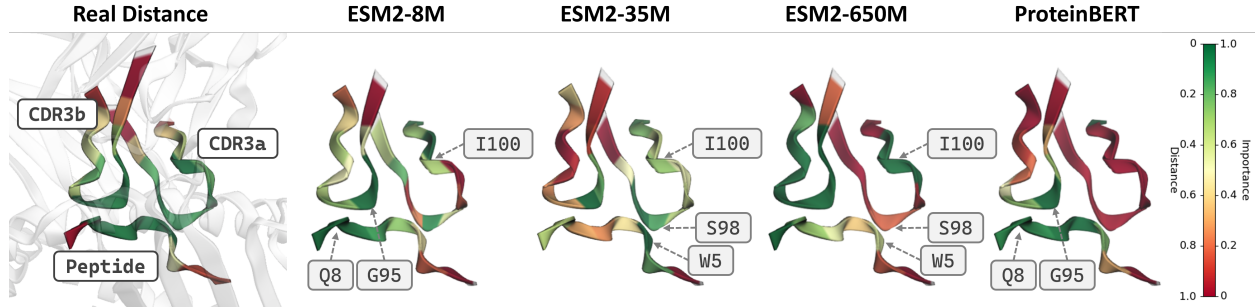


Figure 4: Case study of a tumor-associated peptide analogue with enhanced MHCI binding (2BNQ) [7], illustrating the effects of structure regularization and model-specific contact pattern differences.

precise contact prototype, consistent with the BRHR analysis, where it nearly achieved 100% accuracy. These results show that our method effectively learns the CDR3-peptide contact patterns while providing high-quality interpretability.

3.6 Case Study

To give a concrete example of our approach, we consider a TCR-epitope complex for a tumor-associated peptide analogue with enhanced MHCI binding (PDB: 2BNQ) [7], originally investigated to elucidate mechanisms of TCR-epitope binding strength for epitopes in cancer vaccine design. For this case study, we analyzed the contact prototypes in TCR-SRIM with ESM2 and ProteinBERT backbones. To summarize peptide-TCR interactions for peptide, we aggregate contact scores by averaging those of the CDR3 regions. As shown in Figure 4, all models correctly identified the critical CDR3b residue G95. For the peptide residue Q8, which contacts CDR3b, all models assigned relatively high scores. Except for ESM2-35M, which focused on the CDR3a-peptide contacts at W5 and S98, other models emphasized peptide-CDR3b interactions. TCR-SRIM with ProteinBERT ignored CDR3a, consistent with our BRHR analysis. Additionally, all ESM2 variants highlighted CDR3a residue I100, near the C-terminus of CDR3a. These examples reveal differences in contact pattern modeling across models for a specific, experimentally verified TCR-epitope complex.

3.7 Impact of Generated Structures on Model Learning

Due to the limited availability of experimentally resolved TCR-epitope complex structures, recent structure-aware prediction methods rely on computationally predicted structures to augment structural training data. However, despite incorporating structural information, the performance gains achieved by these approaches remain limited [27]. By integrating an interpretable prototype layer and regularizing prototype learning with predicted structural information, TCR-SRIM enables a systematic investigation of how generated structures influence model learning and predictive performance. We use ESM2-8M as backbone for TCR-SRIM to identify the difference. To this end, we constructed three predicted-structure variants of the TCR-XAI regulation dataset using AlphaFold3 [2], TCRModel2 [46], and tFold-TCR [45]. Structures generated by AlphaFold3 and TCRModel2 were obtained through their publicly available prediction servers. For TCRModel2, structures corresponding to 4Z7W, 4OZH, and 5KSB were excluded after three unsuccessful prediction attempts. For tFold-TCR, we deployed a local version to predict structures.

3.7.1 Impact of Epitope-wise ROC-AUC

We first evaluate the generalization performance of TCR-SRIM when regularized using experimental versus predicted structures. As shown in Table 4, regularization with real structures yields the best performance, achieving a top-100 ROC-AUC of 0.953. Although regularization with structures predicted by AlphaFold3, TCRModel2, and tFold-TCR results in lower performance, all three variants achieve top-100 ROC-AUC scores above 0.910, where they achieve ROC-AUC scores of 0.941, 0.929, and 0.913, respectively ($p < 0.005$ relative to real-structure performance). Notably, this ranking is consistent with the reported accuracy of the structure prediction methods [45], though we note the predictors may have been trained on structures in the TCR-XAI evaluation set. These results suggest that predicted structures can serve as effective substitutes when experimentally resolved structures are unavailable, yielding performance comparable to that obtained with real structures. Nevertheless, a consistent performance gap remains, indicating that inaccuracies in generated structures may limit their ability to fully capture the structural information required for optimal TCR-epitope binding prediction.

Table 4: Comparison of ROC-AUC scores with the false positive rate restricted to 0.1 across the top-100, 150, 200, 250, and 300 peptides among TCR-SRIM regularized with predicted structures.

Structures	Top- k ROC-AUC @ FPR $\leq$ 0.1				
	100	150	200	250	300
Real	0.953	0.821	0.734	0.682	0.655
AlphaFold3	0.941	0.804	0.721	0.672	0.646
TCRModel2	0.929	0.790	0.711	0.663	0.638
tFold-TCR	0.913	0.778	0.702	0.656	0.632

3.7.2 Impact of Interpretability

To assess the impact on the choice of regularization structures on interpretability, we evaluate TCR-SRIM’s BRHR on the TCR-XAI evaluation set when using experimental versus predicted structures. To do this, we substitute the experimental structures used with those predicted by AlphaFold3, TCRModel2, or tFold-TCR for the same TCR-epitope complex. As shown in Table 5, models regularized with predicted structures yield substantially lower peptide-CDR3b BRHR than the corresponding experimental structure, falling below 0.400 for peptide $\rightarrow$ CDR3b across all predicted variants, while still exceeding 0.780 for peptide $\rightarrow$ CDR3a. Thus predicted-structure regularization preferentially captures peptide-CDR3a patterns while inadequately modeling peptide-CDR3b interactions. Given CDR3b’s central role in antigen recognition, this deficiency may explain the performance gap relative to real-structure regularization.

To characterize the effects of structural regularization, we visualize the contact maps learned by prototypes regularized with different structure sources (Figure 5). For peptide-CDR3a prototypes, only models regularized with tFold-TCR and experimentally resolved structures assign high importance to residue 10, matching the ground truth computed from real structures; AlphaFold3 and TCRModel2 prototypes focus elsewhere and miss it. For peptide-CDR3b prototypes, predicted structures produce two notable artifacts: shifted binding-site localization and overly smooth contact distributions. Their highest-intensity regions concentrate around residues 13-14, whereas experimentally resolved structures focus on residues 8, 9, and 12, which are more consistent with the observed contact-core spanning residues 8-12. This shift offers an explanation for the reduced peptide-CDR3b BRHR under predicted-structure regularization. The smoother contact maps further indicate lower spatial variance and less localized signals, suggesting that structural prediction errors introduce an over-smoothing effect that captures generic patterns while losing residue-specific contacts. To probe this, we analyzed the structural variability of CDR loops. For each structure, we computed the CDR-peptide distance matrix, then the pairwise similarity between these matrices as a measure of binding-site diversity, where higher similarity indicates more varied contact patterns. Real structures show significantly similar ($p < 0.005$) than most predicted datasets (1.00 for real CDR3b versus 0.70-0.72 for predicted CDR3b), the sole exception being tFold-TCR on CDR3a (0.99; full results in Appendix B.3). Structure prediction methods tend to generate similar CDR conformations despite these regions’ inherent flexibility, losing the binding-mode diversity of real complexes. This potentially limits the capacity to learn varied interaction patterns, possibly leading to the weaker generalization under predicted-structure regularization.

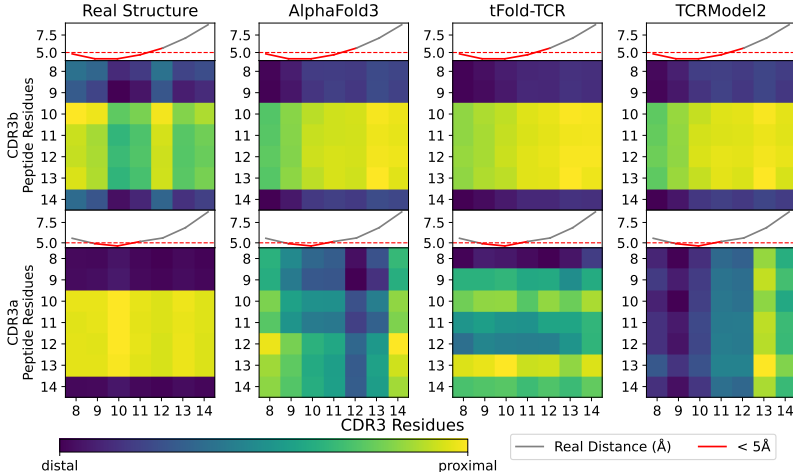


Figure 5: The contact prototypes regularized with predicted structures.

4 Conclusion

In summary, we present TCR-SRIM (structure-regularized interpretable model), an interpretable framework that bridges sequence- and structure-based modeling through contact prototypes while requiring only sequence inputs at inference. Using PLMs and a small set of high-resolution crystal structures for contact-prototype regularization, TCR-SRIM learns CDR3-peptide contact patterns from sequences alone and outperforms state-of-the-art models, showing that structural regularization effectively transfers structural knowledge into sequence models to improve both prediction and, perhaps more importantly, interpretability. Beyond prediction, TCR-SRIM offers a novel framework for analyzing how generated structures shape model learning. Although structures from AlphaFold3, TCRModel2, and tFold-TCR yield competitive performance, they show reduced binding-site diversity and less accurate peptide-CDR3b interaction patterns than experimentally resolved structures, highlighting a key limitation of current structure-prediction models for TCR-epitope learning and the value of interpretable-by-design approaches for evaluating predicted structures.

Acknowledgments: Harold L. and Heather E. Jurist Center of Excellence for Artificial Intelligence at Tulane University.

References

- [1] 10x Genomics. A new way of exploring immunity: Linking highly multiplexed antigen recognition to immune repertoire and phenotype. Technical report, 10x Genomics, 2022.
- [2] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.
- [3] 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. Vdjdb in 2019: database extension, new analysis infrastructure and a t-cell receptor motif compendium. *Nucleic acids research*, 48(D1):D1057–D1062, 2020.
- [4] Rémy Bosselut. T cell antigen recognition: Evolution-driven affinities. *Proceedings of the National Academy of Sciences*, 116(44):21969–21971, 2019.
- [5] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.
- [6] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.
- [7] Ji-Li Chen, Guillaume Stewart-Jones, Giovanna Bossi, Nikolai M Lissin, Linda Wooldridge, Ed Man Lik Choi, Gerhard Held, P Rod Dunbar, Robert M Esnouf, Malkit Sami, et al. Structural and kinetic basis for heightened immunogenicity of t cell vaccines. *The Journal of experimental medicine*, 201(8):1243–1255, 2005.
- [8] Giancarlo Croce, Sara Bobisse, Dana Léa Moreno, Julien Schmidt, Philippe Guillame, Alexandre Harari, and David Gfeller. Deep learning predictions of tcr-epitope interactions reveal epitope-specific chains in dual alpha t cells. *Nature Communications*, 15(1):3211, 2024.
- [9] Pradyot Dash, Andrew J Fiore-Gartland, Tomer Hertz, George C Wang, Shalini Sharma, Aisha Souquette, Jeremy Chase Crawford, E Bridie Clemens, Thi HO Nguyen, Katherine Kedzierska, et al. Quantifiable predictive features define epitope-specific t cell receptor repertoires. *Nature*, 547(7661):89–93, 2017.
- [10] Mark M Davis and Pamela J Bjorkman. T-cell antigen receptor genes and t-cell recognition. *Nature*, 334(6181):395–402, 1988.
- [11] Sebastian Nymann Deleuran and Morten Nielsen. Netcr-struct, a structure driven approach for prediction of tcr-pmhc interactions. *Frontiers in Immunology*, 16:1616328, 2025.
- [12] Ziyang Feng, Jingyang Chen, Youlong Hai, Xuelian Pang, Kun Zheng, Chenglong Xie, Xiujuan Zhang, Shengqing Li, Chengjuan Zhang, Kangdong Liu, et al. Sliding-attention transformer neural architecture for predicting t cell receptor–antigen–human leucocyte antigen binding. *Nature Machine Intelligence*, 6(10):1216–1230, 2024.
- [13] Filippo Grazioli, Pierre Machart, Anja Mösch, Kai Li, Leonardo V Castorina, Nico Pfeifer, and Martin Renqiang Min. Attentive variational information bottleneck for tcr–peptide interaction prediction. *Bioinformatics*, 39(1):btac820, 2023.
- [14] Huang Huang, Chunlin Wang, Florian Rubelt, Thomas J Scriba, and Mark M Davis. Analyzing the mycobacterium tuberculosis immune response by t-cell receptor clustering with glyph2 and genome-wide antigen screening. *Nature biotechnology*, 38(10):1194–1202, 2020.
- [15] Dan Hudson, Ricardo A Fernandes, Mark Basham, Graham Ogg, and Hashem Koohy. Can we predict t cell specificity with digital biology and machine learning? *Nature Reviews Immunology*, 23(8):511–521, 2023.
- [16] Dan Hudson, Alex Lubbock, Mark Basham, and Hashem Koohy. A comparison of clustering models for inference of t cell receptor antigen specificity. *ImmunoInformatics*, 13:100033, 2024.
- [17] Mathias Fynbo Jensen and Morten Nielsen. Netcr 2.2-improved tcr specificity predictions by combining pan-and peptide-specific training strategies, loss-scaling and integration of sequence similarity. *bioRxiv*, pages 2023–10, 2023.

- [18] Alok V Joglekar and Guideng Li. T cell antigen discovery. *Nature methods*, 18(8):873–880, 2021.
- [19] Emmi Jokinen, Alexandru Dumitrescu, Jani Huuhtanen, Vladimir Gligorijević, Satu Mustjoki, Richard Bonneau, Markus Heinonen, and Harri Lähdesmäki. Tcrconv: predicting recognition between t cell receptors and epitopes using contextualized motifs. *Bioinformatics*, 39(1):btac788, 2023.
- [20] Jinwoo Leem, Saulo H P de Oliveira, Konrad Krawczyk, and Charlotte M Deane. Stcrdab: the structural t-cell receptor database. *Nucleic acids research*, 46(D1):D406–D412, 2018.
- [21] Jiarui Li, Zixiang Yin, Zhengming Ding, Samuel J Landry, and Ramgopal R Mettu. Rational multi-modal transformers for tcr-pmhc prediction. In *Proceedings of the 16th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10, 2025.
- [22] Jiarui Li, Zixiang Yin, Haley Smith, Zhengming Ding, Samuel J Landry, and Ramgopal R Mettu. Quantifying cross-attention interaction in transformers for interpreting tcr-pmhc binding. *arXiv preprint arXiv:2507.03197*, 2025.
- [23] Valerie Lin, Melyssa Cheung, Ragul Gowthaman, Maya Eisenberg, Brian M Baker, and Brian G Pierce. Tcr3d 2.0: expanding the t cell receptor structure database with new structures, tools and interactions. *Nucleic Acids Research*, 53(D1):D604–D608, 2025.
- [24] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.
- [25] Tianshi Lu, Ze Zhang, James Zhu, Yunguan Wang, Peixin Jiang, Xue Xiao, Chantale Bernatchez, John V Heymach, Don L Gibbons, Jun Wang, et al. Deep learning-based prediction of the t cell receptor–antigen binding specificity. *Nature machine intelligence*, 3(10):864–875, 2021.
- [26] Yanping Lu, Yuyan Wang, Meng Xu, Bingbing Xie, Yumeng Yang, Haodong Xu, and Shengbao Suo. Assessment of computational methods in predicting tcr–epitope binding recognition. *Nature Methods*, 23(1):248–259, 2026.
- [27] Claus Lundegaard, Kasper Lamberth, Mikkel Harndahl, Søren Buus, Ole Lund, and Morten Nielsen. Netmhc-3.0: accurate web accessible predictions of human, mouse and monkey mhc class i affinities for peptides of length 8–11. *Nucleic acids research*, 36(suppl_2):W509–W512, 2008.
- [28] 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. Tcr meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, hla-restricted clusters of sars-cov-2 tcrcs. *Elife*, 10:e68605, 2021.
- [29] Barthelemy Meynard-Piganeau, Christoph Feinauer, Martin Weigt, Aleksandra M Walczak, and Thierry Mora. Tulip: A transformer-based unsupervised language model for interacting peptides and t cell receptors that generalizes to unseen epitopes. *Proceedings of the National Academy of Sciences*, 121(24):e2316401121, 2024.
- [30] Meike Nauta, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. Pip-net: Patch-based intuitive prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2744–2753, 2023.
- [31] Jacques Neefjes, Marlieke LM Jongsma, Petra Paul, and Oddmund Bakke. Towards a systems understanding of mhc class i and mhc class ii antigen presentation. *Nature reviews immunology*, 11(12):823–836, 2011.
- [32] Morten Nielsen, Massimo Andreatta, Bjoern Peters, and Søren Buus. Immunoinformatics: predicting peptide–mhc binding. *Annual review of biomedical data science*, 3(1):191–215, 2020.
- [33] Morten Nielsen, Anne Eugster, Mathias Fynbo Jensen, Manisha Goel, Andreas Tiffeau-Mayer, Aurelien Pelissier, Sebastiaan Valkiers, María Rodríguez Martínez, Barthélémy Meynard-Piganeau, Victor Greiff, et al. Lessons learned from the immrep23 tcr-epitope prediction challenge. *Immunoinformatics*, 16:100045, 2024.
- [34] Bjoern Peters, Morten Nielsen, and Alessandro Sette. T cell epitope predictions. *Annual Review of Immunology*, 38(1):123–145, 2020.

- [35] 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. epitcr: a highly sensitive predictor for tcr-peptide binding. *Bioinformatics*, 39(5):btad284, 2023.
- [36] Mansour Poorebrahim, Niloufar Mohammadkhani, Reza Mahmoudi, Monireh Gholizadeh, Elham Fakhr, and Angel Cid-Arregui. Tcr-like cars and tcr-cars targeting neoepitopes: an emerging potential. *Cancer gene therapy*, 28(6):581–589, 2021.
- [37] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.
- [38] Luis A Rojas, Zachary Sethna, Kevin C Soares, Cristina Olcese, Nan Pang, Erin Patterson, Jayon Lihm, Nicholas Ceglia, Pablo Guasp, Alexander Chu, et al. Personalized rna neoantigen vaccines stimulate t cells in pancreatic cancer. *Nature*, 618(7963):144–150, 2023.
- [39] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.
- [40] Aisha Shah, Raphael Genolet, Aymeric Auger, Dana Léa Moreno, Yan Liu, Giancarlo Croce, Julien Racle, Alexandre Harari, and David Gfeller. Unpaired tcr α + tcr β sequencing is sufficient for training machine learning tcr-epitope recognition predictors. *bioRxiv*, pages 2026–03, 2026.
- [41] Jared K Slone, Mingyong Zhang, Peixin Jiang, Amanda Montoya, Emily Bontekoe, Barbara Nassif Rausseo, Alexandre Reuben, and Lydia E Kavraki. Stag-llm: Predicting tcr-phla binding with protein language models and computationally generated 3d structures. *Computational and Structural Biotechnology Journal*, 2025.
- [42] Ido Springer, Nili Tickotsky, and Yoram Louzoun. Contribution of t cell receptor alpha and beta cdr3, mhc typing, v and j genes to peptide binding prediction. *Frontiers in immunology*, 12:664514, 2021.
- [43] Nili Tickotsky, Tal Sagiv, Jaime Prilusky, Eric Shifrut, and Nir Friedman. Mcpas-tcr: a manually curated catalogue of pathology-associated t cell receptor sequences. *Bioinformatics*, 33(18):2924–2929, 2017.
- [44] Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters. The immune epitope database (iedb): 2018 update. *Nucleic acids research*, 47(D1):D339–D343, 2019.
- [45] Fandi Wu, Yu Zhao, Yang Xiao, Chenchen Qin, Fang Wang, Zihan Wu, Long-Kai Huang, Xiao Liu, Jiangning Song, Bing He, et al. Fast and accurate modeling of tcr-peptide-mhc complexes using tfold-tcr. *bioRxiv*, pages 2025–01, 2025.
- [46] Rui Yin, Helder V Ribeiro-Filho, Valerie Lin, Ragul Gowthaman, Melyssa Cheung, and Brian G Pierce. Tcrmodel2: high-resolution modeling of t cell receptor recognition using deep learning. *Nucleic Acids Research*, 51(W1):W569–W576, 2023.
- [47] Hongyi Zhang, Xiaowei Zhan, and Bo Li. Giana allows computationally-efficient tcr clustering and multi-disease repertoire classification by isometric transformation. *Nature communications*, 12(1):4699, 2021.
- [48] Wei Zhang, Longlong Wang, Ke Liu, Xiaofeng Wei, Kai Yang, Wensi Du, Shiyu Wang, Nannan Guo, Chuanchuan Ma, Lihua Luo, et al. Pird: pan immune repertoire database. *Bioinformatics*, 36(3):897–903, 2020.

Structure-Regularized Interpretable TCR-Epitope Prediction

(Supplementary)

Jiarui Li¹, Zixiang Yin¹, Yunbei Zhang¹, Janet Wang¹
Samuel J. Landry², Zhengming Ding¹, Ramgopal R. Mettu¹

¹Department of Computer Science, Tulane University

²Department of Biochemistry and Molecular Biology, Tulane University School of Medicine

{jli78, zyin, yzhang111, swang47, landry, zding1, rmettu}@tulane.edu

<https://tcrsrim.jiarui.li/>

A Performance Evaluation

In this section, we introduce the baseline models compared to our model. Then, we extend the epitope-wise ROC-AUC and other metrics analysis and comprehensive benchmark to compare additional TCR-SRIM variants against other models, and we evaluate TCR-SRIM on the IMMREP23 benchmark.

A.1 Baseline Models

We consider two categories of comparable models in our evaluations. First, to evaluate PLM-based models, we constructed a standard linear classifier for PLM features. Specifically, we added two fully connected layers with hidden dimension three times the feature dimension and ReLU activation. Each classifier takes concatenated global representations of CDR3a, CDR3b, and the peptide as input and outputs a prediction score. For ProteinBERT, we used the provided global features, and for the ESM models, we averaged local residue-level features to obtain global representations. Second, we compared our models with two recent transformer-based TCR-epitope prediction methods. MixTCRpred [8], one of the most widely used TCR-epitope models, utilizes all CDR regions as input. TULIP [29] is another recent model that outperforms the widely used NetTCR-2.2 [17] baseline in terms of accuracy and generalization ability. EGM [21] is a TCR-epitope prediction model designed with post-hoc analysis guidance. We utilized EGM2 architecture, but change the inputs from full TCR to only CDR3 regions.

A.2 Extended Epitope-wise ROC-AUC Analysis

As shown in the Figure 6, we demonstrate the epitope-wise ROC-AUC ($\text{FPR} < 0.1$) for all epitopes. Our TCR-SRIM achieves strong performance across a wide range of epitopes, with approximately 20% reaching ROC-AUC values above 0.8, which shows superior generalization ability and predictive performance.

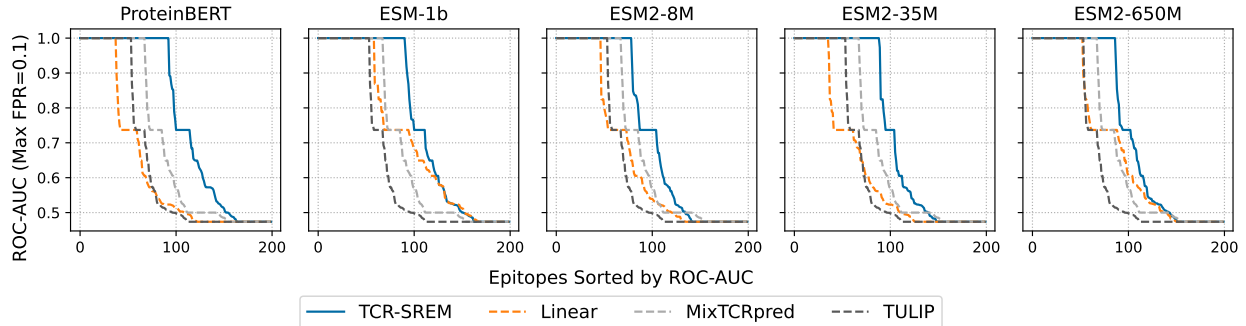


Figure 6: The epitope-wise ROC-AUC ($\text{FPR} < 0.1$) evaluated on TCR-SRIM and baseline models. TCR-SRIM consistently achieves higher predictive ROC-AUC for more epitopes comparing against MixTCRpred and TULIP, where $\sim 20\%$ epitopes can achieve above 0.8.

A.3 Epitope-wise Precision, Recall, and F1 Analysis

Considering that the test set is imbalanced, with substantially more negative than positive samples, we additionally report precision, recall, and F1 scores. Because these metrics require deterministic prediction labels, we apply a threshold of 0.5 for positive classification. As shown in Table 6, models with structure regularization consistently achieve higher recall, reaching values above 0.84. In contrast, models without structure regularization obtain higher precision but suffer from significantly reduced recall, with decreases of approximately 10-20%.

PLM Backbone	Metric	Top- k @ Threshold = 0.5			
		Top-100	Top-150	Top-200	Top-250
ESM-1b	Precision	0.634	0.382	0.349	0.342
	Recall	0.891	0.544	0.482	0.472
	F1	0.741	0.449	0.405	0.397
ESM2-8M	Precision	0.522	0.306	0.270	0.266
	Recall	0.877	0.588	0.512	0.506
	F1	0.655	0.403	0.354	0.349
ESM2-35M	Precision	0.589	0.307	0.305	0.300
	Recall	0.844	0.410	0.406	0.396
	F1	0.694	0.351	0.348	0.342
ESM2-650M	Precision	0.594	0.324	0.316	0.310
	Recall	0.887	0.441	0.426	0.415
	F1	0.712	0.373	0.363	0.355
ProteinBERT	Precision	0.705	0.566	0.382	0.371
	Recall	0.925	0.536	0.380	0.369
	F1	0.800	0.550	0.381	0.370

Table 6: Precision, recall, and F1 of structure-regularized TCR-SRIM (threshold 0.5) across the Top-100, 150, 200, and 250 peptides, for each PLM backbone.

A.4 Extended Evaluation on Independent Comprehensive Benchmark

Table 7: AUPRC and ROC-AUC comparison of TCR-epitope binding prediction models on the Comprehensive benchmark. Our models based on different PLM backbones achieve competitive performance, with TCR-SRIM (ESM2-150M) and vibtr obtaining the highest AUPRC, and TCR-SRIM (ProteinBERT)-1R (TCR-SRIM (ProteinBERT) with one epoch regularization) achieving the best ROC-AUC.

Models	AUPRC	ROC-AUC	Models	AUPRC	ROC-AUC
TCR-SRIM (ProteinBERT)-1R	0.575	0.571	TCR-SRIM (ESM2-35M)	0.571	0.538
TCR-SRIM (ESM2-150M)	0.587	0.564	TCR-SRIM (ProteinBERT)-0R	0.509	0.532
PISTE	0.567	0.564	epiTCR	0.499	0.527
vibtr	0.588	0.563	pMTnet	0.513	0.524
TCR-SRIM (ProteinBERT)	0.550	0.541	TCRconv	0.548	0.519
MixTCRpred	0.540	0.546	TCR-SRIM (ESM2-650M)	0.544	0.505
ERGOII	0.540	0.546	TCR-SRIM (ESM2-8M)	0.502	0.475
TCR-SRIM (ESM1b)	0.557	0.540			

We compare TCR-SRIM with different PLM backbones and more models including ERGOII [42], epiTCR [35], TCRconv [19], and pMTnet [25] in Table 7. Notably, using the same backbone with different numbers of structure regularization epochs leads to distinct performance. Without structure regularization, the AUPRC drops to 0.509; with one epoch, it improves to 0.575; however, with three epochs, it decreases to 0.550. This indicates that generalization is sensitive to the extent of structure regularization. A limited number of regularization epochs helps guide the model to learn meaningful contact patterns in TCR-peptide binding, whereas excessive regularization degrades generalization performance.

A.5 IMMREP23 ROC-AUC Analysis

To further validate our findings, we evaluated all models on the IMMREP23 benchmark [33] and compared their performance against MixTCRpred. All models were trained using our compiled training dataset. The ESM-1b backbone achieves the best overall performance (ROC-AUC @ FPR<0.1 = 0.602), whereas ProteinBERT performs the worst (ROC-AUC @ FPR<0.1 = 0.546). Notably, all PLM backbones integrated with TCR-SRIM consistently outperform MixTCRpred (ROC-AUC @ FPR<0.1 = 0.482). This performance difference can be partially attributed to the modeling focus of the respective backbones: ESM-1b places greater emphasis on CDR3a-peptide interactions, while ProteinBERT primarily captures CDR3b-peptide interactions. In contrast, the relatively weaker performance of all these models may stem from several factors, including the presence of non-human and non-murine samples in our training dataset, the absence of epitope-specific optimization for the IMMREP23 task. For MixTCRpred, the exclusion of samples lacking complete CDR regions is the other important reason.

Table 8: Comparison of ROC-AUC scores with the false positive rate restricted to 0.1 on IMMREP23 benchmark for TCR-SRIM with various PLM backbones and MixTCRpred trained on our training dataset.

Backbones	ProteinBERT	ESM-1b	ESM2-8M	ESM2-35M	ESM2-650M	MixTCRpred
ROC-AUC	0.546	0.602	0.575	0.572	0.577	0.482

B Impact of Generated Structures

In this section, we first describe the preparation of predicted structures. We then provide extended analysis showing that real-structure regularization significantly improves performance over predicted structures, and report detailed per-sample pairwise normalized RMSD to illustrate the binding-site diversity difference between real and predicted structures.

B.1 Preparation of Predicted Structure

To prepare the predicted-structure dataset, we first extracted sequences from the TCR-XAI regularization dataset. Each entry includes TCR α , TCR β , peptide, and MHC sequences (a two-chain MHC is recorded as two separate chains). For each method, we define a proxy resolution that rescales the model’s per-residue confidence to a range comparable to experimental resolution:

$$r = \frac{100 - \bar{\phi}_{\min}}{10}, \quad \bar{\phi}_{\min} = \min\left(\frac{\sum \phi_a}{|\phi_a|}, \frac{\sum \phi_b}{|\phi_b|}, \frac{\sum \phi_e}{|\phi_e|}\right),$$

where ϕ_a , ϕ_b , and ϕ_e are the per-residue confidence scores (pLDDT or equivalent) of the CDR3a, CDR3b, and peptide chains, and $\bar{\phi}_{\min}$ is the lowest of the three chain-averaged confidences.

AlphaFold3. We used the AlphaFold3 server (<https://alphafoldserver.com/>), treating each component of the TCR-epitope complex as an independent chain, and took the rank-1 model as the predicted structure. Proxy resolution follows the equation above with ϕ taken as pLDDT.

tFold-TCR. We deployed a local instance following <https://github.com/TencentAI4S/tfold> and ran inference on CPU. Here confidence is the B-factor (or equivalent) reported in the output PDB, normalized to $[0, 1]$, so proxy resolution is computed as

$$r = 10(1 - \bar{\phi}_{\min}).$$

TCRModel2. We used the web service (<https://tcmodel.ibbr.umd.edu/>) with ϕ taken as pLDDT. For MHC-I, the server accepts only a single chain, so we omitted the β_2 -microglobulin sequence when present; for MHC-II, the server truncates sequences longer than 11 residues, and we padded the truncated positions with the nearest values when computing contact maps. Three structures were omitted due to server errors: 4Z7W and 4OZH returned an unknown error, and 5KSB returned “cannot find core.”

B.2 Test of Significant for Impact of Epitope-wise ROC-AUC

Because the improvement between regularizing with real versus AlphaFold3-predicted structures is only 1.2%, we report a significance test in Table 9. All p -values are below 0.005, confirming that the improvements are statistically significant.

Table 9: ROC-AUC scores (FPR restricted to 0.1) for TCR-SRIM regularized with predicted structures, across the Top-100, 150, 200, 250, and 300 peptides. Significance versus real-structure regularization shown in parentheses (p -value).

Structures	Top- k ROC-AUC @ FPR $\leq$ 0.1				
	100	150	200	250	300
Real	0.953	0.821	0.734	0.682	0.655
AlphaFold3	0.941 ($2.7e-4$)	0.804 ($1.2e-10$)	0.721 ($2.4e-10$)	0.672 ($3.6e-10$)	0.646 ($4.4e-10$)
TCRModel2	0.929 ($1.7e-7$)	0.790 ($5.5e-14$)	0.711 ($2.2e-13$)	0.663 ($4.6e-13$)	0.638 ($6.7e-13$)
tFold-TCR	0.913 ($2.4e-6$)	0.778 ($1.6e-13$)	0.702 ($5.7e-13$)	0.656 ($1.1e-12$)	0.632 ($1.6e-12$)

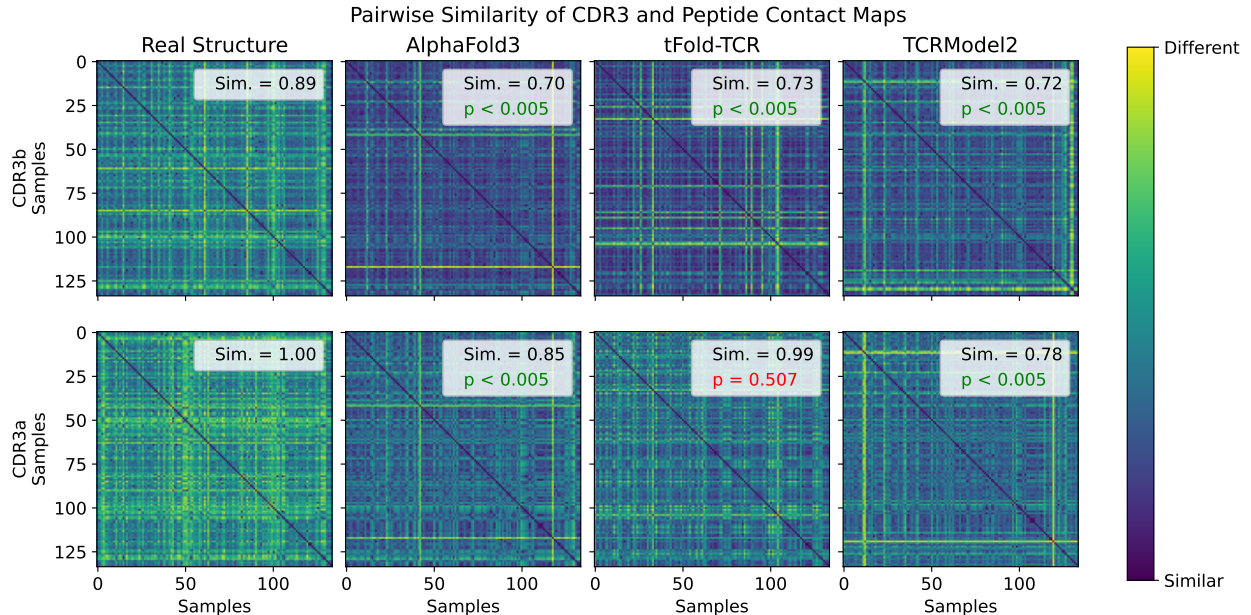


Figure 7: Pairwise RMSD of CDR3 and Peptide Contact Maps

B.3 Binding-Site Diversity Differs between Predicted and Real Structures

As shown in Figure 7, we compute pairwise similarity of the CDR-peptide distance matrices and use averaged similarity to indicate binding site diversity, where the pairwise similarity is RMSD between any pair of z-scored contact distance matrices. Except CDR3a-peptide predicted by tFold-TCR, for all other predicted structures, experimentally resolved structures showing significant higher contact site diversity with $p < 0.005$.

C Ablation Studies

To investigate the effectiveness of structural regularization and the sensitivity of TCR-SRIM to key hyperparameters, we conducted ablation studies using the ESM2-8M backbone, which outperforms baseline models while remaining efficient enough to allow testing of many hyperparameter combinations. Specifically, we examined four factors: (1) *structural regularization*, evaluating its impact on predictive performance and interpretation quality, as well as the relationship between the number of available structures and model performance; (2) *mixing strategy*, comparing maximum versus average pooling for aggregating CDR3a-peptide and CDR3b-peptide contact prototype scores; (3) *regularization threshold*, assessing the effect of varying the threshold ϵ from 0.6 to 0.9 on structural regularization strength; and (4) *regularization frequency*, varying the number of structural regularization epochs applied after each standard training epoch.

C.1 Structure Regularization

To evaluate the impact of structure regularization, we conduct ablation studies to assess its effect on both performance and interpretability. In the absence of structure regularization, the model learns the contact prototypes in an unsupervised

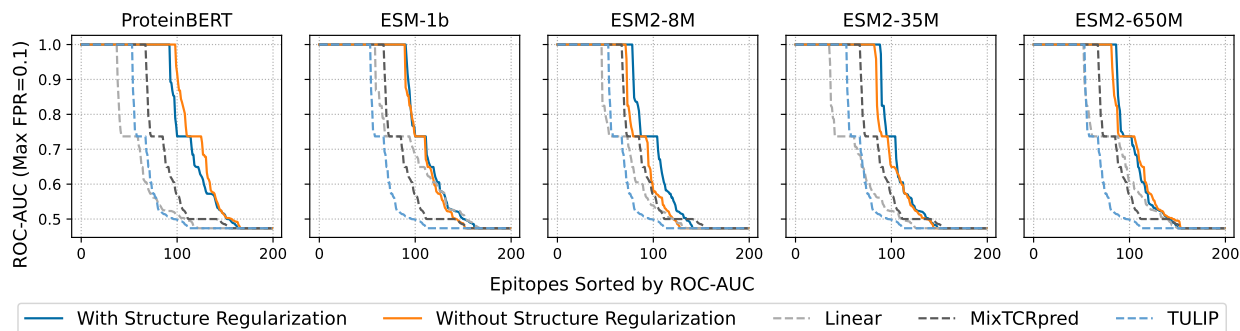


Figure 8: The epitope-wise ROC-AUC (Max FPR = 0.1) evaluated on TCR-SRIM with or without structure regularization and baseline models.

manner and infers contact prototypes of the CDR3-peptide based on only sequence inputs. In practice, this was implemented by omitting the structure regularization epochs following the standard training epochs. Overall, we find that structure regularization yields minor improvements, while interpretability has major, statistically significant, improvements.

Table 10: Comparison of ROC-AUC scores with the false positive rate restricted to 0.1 across the Top-100, Top-150, Top-200, Top-250, and Top-300 peptides among TCR-SRIM with or without structure regularization.

Model	Embedding	Top- k ROC-AUC @ FPR $\leq$ 0.1				
		100	150	200	250	300
Omitting Structure Regularization	ESM-1b	0.982	0.854	0.759	0.702	0.672
	ProteinBERT	0.999	0.895	0.792	0.728	0.695
	ESM2-8M	0.926	0.786	0.708	0.661	0.636
	ESM2-35M	0.960	0.823	0.735	0.683	0.655
	ESM2-650M	0.960	0.836	0.746	0.691	0.663
Structure Regularization	ESM-1b	0.986	0.862	0.766	0.707	0.677
	ProteinBERT	0.989	0.871	0.773	0.713	0.681
	ESM2-8M	0.953	0.821	0.734	0.682	0.655
	ESM2-35M	0.977	0.838	0.747	0.692	0.663
	ESM2-650M	0.971	0.837	0.746	0.692	0.663

Table 11: Precision, recall, and F1 (threshold = 0.5) across Top-100/150/200/250 peptides. For each PLM backbone, rows compare TCR-SRIM without (−) and with (+) structure regularization.

PLM Backbone	Structure	Precision @ 0.5				Recall @ 0.5				F1 @ 0.5			
		T100	T150	T200	T250	T100	T150	T200	T250	T100	T150	T200	T250
ESM-1b	−	0.836	0.498	0.481	0.467	0.631	0.184	0.174	0.167	0.719	0.269	0.255	0.246
	+	0.634	0.382	0.349	0.342	0.891	0.544	0.482	0.472	0.741	0.449	0.405	0.397
ESM2-8M	−	0.729	0.477	0.464	0.373	0.656	0.269	0.256	0.219	0.691	0.344	0.330	0.276
	+	0.522	0.306	0.270	0.266	0.877	0.588	0.512	0.506	0.655	0.403	0.354	0.349
ESM2-35M	−	0.764	0.468	0.449	0.434	0.723	0.269	0.256	0.247	0.743	0.342	0.326	0.315
	+	0.589	0.307	0.305	0.300	0.844	0.410	0.406	0.396	0.694	0.351	0.348	0.342
ESM2-650M	−	0.827	0.513	0.486	0.480	0.613	0.219	0.195	0.187	0.705	0.307	0.278	0.269
	+	0.594	0.324	0.316	0.310	0.887	0.441	0.426	0.415	0.712	0.373	0.363	0.355
ProteinBERT	−	0.931	0.730	0.537	0.524	0.736	0.382	0.199	0.190	0.822	0.502	0.290	0.279
	+	0.705	0.566	0.382	0.371	0.925	0.536	0.380	0.369	0.800	0.550	0.381	0.370

Impact of Regularization on Predictive Performance: Table 10 and Figure 8 shows results of our TCR-SRIM with different backbones with and without structure regularization. We see that regularization generally leads to a performance increase across most PLM backbones. Notably, ESM2-8M, a smaller PLM, experiences an increase of approximately 2.5% in ROC-AUC, while larger models experience smaller improvements. ProteinBERT was an exception, exhibiting a modest improvement without structure regularization (of about 1%), which we attribute to its strong pretraining on UniProtKB/UniRef90 (~ 106 million protein sequences) and with gene ontology (GO) annotation prediction task that empowers it to produce high-quality protein features. We also provide epitope-wise precision, recall, and F1 for reference in Table 11.

Table 12: Comparison of structure regularization and non-regularization on BRHR across different protein language model backbones, where $a \rightarrow b$ denotes the evaluation of residues in chain b with respect to chain a .

Embeddings	Structure	Peptide	Peptide	CDR3a	CDR3b
	(-) without (+) with	↓ CDR3a	↓ CDR3b	↓ Peptide	↓ Peptide
ProteinBERT	(+)	0.568	0.996	0.392	0.855
	(-)	0.842	0.848	0.786	0.743
ESM-1b	(+)	0.833	0.944	0.818	0.852
	(-)	0.876	0.854	0.817	0.788
ESM2-8M	(+)	0.604	0.801	0.804	0.816
	(-)	0.842	0.840	0.784	0.779
ESM2-35M	(+)	0.834	0.944	0.782	0.746
	(-)	0.847	0.857	0.687	0.773
ESM2-650M	(+)	0.961	0.820	0.788	0.809
	(-)	0.861	0.823	0.776	0.789

Impact of Regularization on Interpretation Quality: To evaluate the impact of structure regularization on the quality of contact prototype interpretations, we evaluated the BRHR of TCR-SRIM with and without structure regularization. As shown in Table 12, structure regularization leads to a substantial increase in BRHR for CDR3b-peptide interactions, with ProteinBERT showing an increase of approximately 10% (p-value < 0.001 compared to all baselines). Interestingly, in the absence of structure regularization, the model assigns more balanced weights to CDR3a-peptide and CDR3b-peptide interactions, rather than focusing predominantly on a single interaction as observed in TCR-SRIM with ProteinBERT.

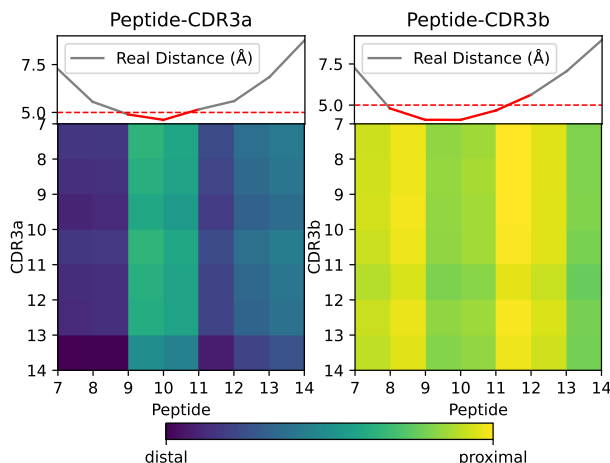


Figure 9: Average centralized contact prototypes on the TCR-XAI benchmark in the absence of structure regularization.

Impact of Regularization on Contact Prototype: We also visualized the contact prototypes of TCR-SRIM with ProteinBERT, omitting structure regularization, on the TCR-XAI benchmark. As shown in Figure 9, the model exhibits reasonably balanced accuracy for both CDR3a-peptide and CDR3b-peptide interactions, rather than focusing predominantly on CDR3b-peptide. However, the resolution and quality of the CDR3b-peptide interaction remain

substantially lower than in the structure-regularized models. This proves our observation from quantitative BRHR comparison between structure regularization and omitting structure regularization.

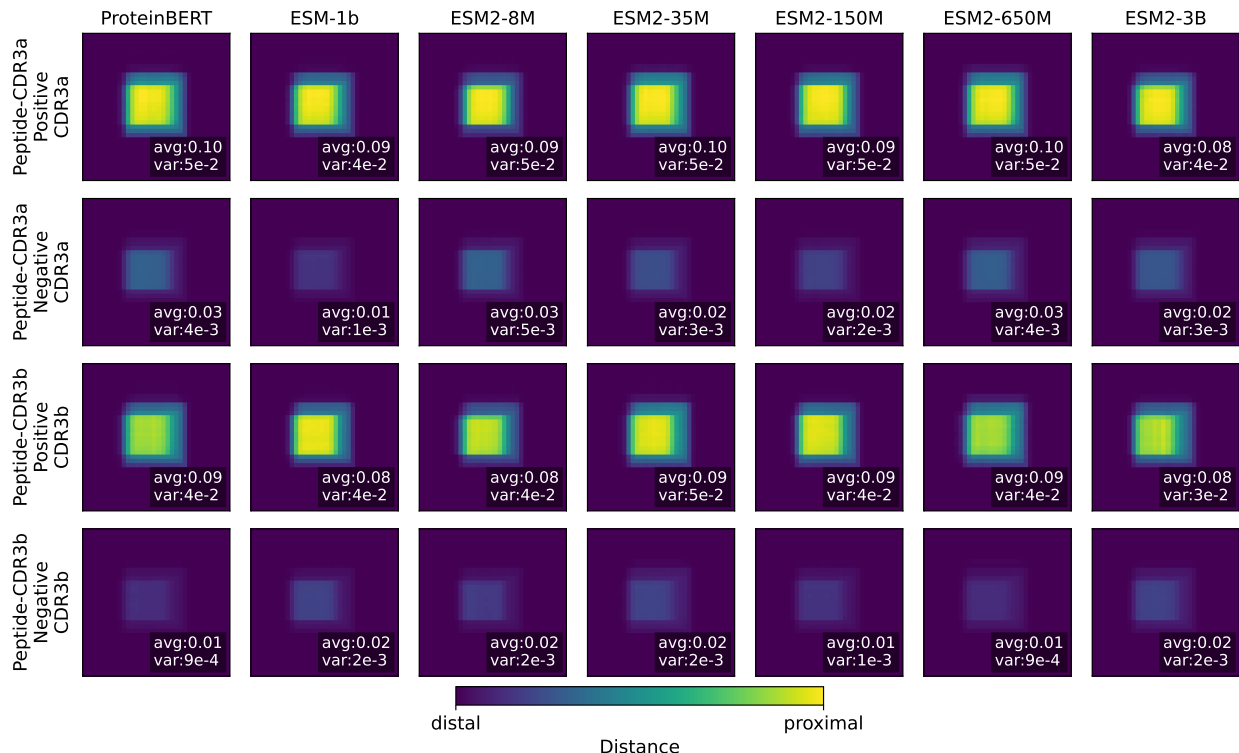


Figure 10: Average centralized contact prototypes on the training dataset in the absence of structure regularization from TCR-SRIM with various PLM backbones.

Finally, we visualized the contact prototypes across all backbones on the entire training dataset. As shown in Figure 10, for TCR-SRIM without structure regularization, the variance of contact prototypes from positive samples is substantially higher than that of negative samples, indicating that the model can still learn meaningful contact patterns from sequence data alone. Nonetheless, structure regularization ensures the learning of higher-quality prototypes.

C.2 Regularization Set Size

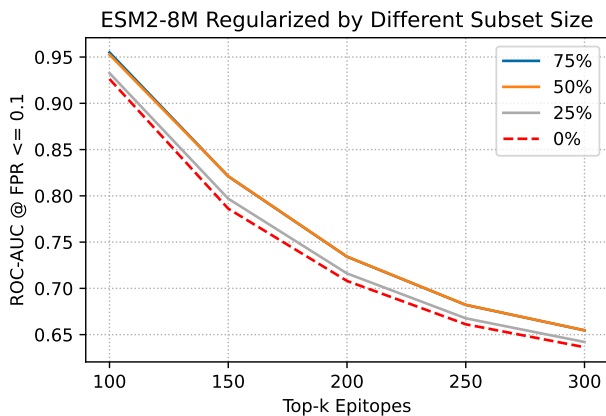


Figure 11: Comparison of ROC-AUC scores with the false positive rate restricted to 0.1 across TCR-SRIM with ESM2-8M regularized by different TCR-XAI subset size including 75%, 50%, and 25%. Structural regularization consistently improves generalization performance across all subset sizes monotonically.

We also examined the effect of the number of structures used for regularization on model performance. We regularized TCR-SRIM with the ESM2-8M backbone using randomly selected subsets of TCR-XAI benchmark structures, comprising 25%, 50%, and 75% of the full set. As shown in Figure 11, structural regularization consistently improves generalization performance across all subset sizes monotonically (0.5%, 2.5%, and 2.5% improvement, respectively). We note that, interestingly, even using 25% of the structures used (68 complexes) yields a minor improvement.

C.3 Mixing Strategy and Regularization Thresholds

To examine how different mixing strategies of CDR3a-peptide and CDR3b-peptide contact scores, as well as regularization thresholds, influence model performance, we conduct ablation studies varying both factors. The mixing strategies include maximum (Max.) and average pooling (Avg.), and the thresholds (ϵ) range from 0.6 to 0.9. As shown in Table 13, the choice of mixing method has a minor impact, with performance differences within 5%. In contrast, the regularization threshold has a pronounced effect: the best results are achieved at $\epsilon = 0.75 - 0.8$, while larger or smaller values reduce performance by approximately 10%.

Table 13: Ablation study on regularization thresholds (ϵ) for max and average pooling, evaluated by ROC-AUC (Max FPR = 0.1) across different top- k peptide subsets.

Regularize Threshold	Mix Method	Top- k ROC-AUC @ FPR ≤ 0.1				
		100	150	200	250	300
$\epsilon = 0.6$	Max.	0.916	0.785	0.707	0.661	0.636
	Avg.	0.896	0.763	0.691	0.647	0.624
$\epsilon = 0.7$	Max.	0.844	0.722	0.660	0.623	0.603
	Avg.	0.893	0.763	0.690	0.647	0.624
$\epsilon = 0.75$	Max.	0.932	0.794	0.714	0.666	0.641
	Avg.	0.961	0.825	0.737	0.684	0.657
$\epsilon = 0.8$	Max.	0.953	0.821	0.734	0.682	0.655
	Avg.	0.948	0.811	0.726	0.676	0.649
$\epsilon = 0.85$	Max.	0.881	0.749	0.680	0.639	0.617
	Avg.	0.897	0.766	0.693	0.649	0.626
$\epsilon = 0.9$	Max.	0.878	0.749	0.681	0.639	0.617
	Avg.	0.922	0.786	0.708	0.661	0.636

C.4 Regularization Frequency

Table 14: Ablation study on regularization frequency evaluated by ROC-AUC (Max FPR = 0.1) across different top- k peptide subsets.

Regularize Epoch	Top- k ROC-AUC @ FPR ≤ 0.1				
	100	150	200	250	300
1	0.899	0.762	0.690	0.647	0.624
3	0.973	0.839	0.748	0.693	0.664
5	0.953	0.821	0.734	0.682	0.655
7	0.908	0.774	0.699	0.654	0.630
9	0.850	0.728	0.664	0.626	0.606

Another key factor influencing structure regularization is the regularization frequency, defined as the number of structure regularization epochs applied after each standard training epoch. We evaluated the impact of varying this frequency from 1 to 9 epochs. As reported in Table 14, the model achieves optimal performance, ROC-AUC exceeding 0.95, when the number of regularization epochs is 3 or 5. Both higher and lower frequencies lead to a substantial performance drop of over 5%.

D Availability and Implementation

All experiments were performed on three Ubuntu servers equipped with two NVIDIA A2000, one NVIDIA RTX3090, and two NVIDIA 6000 Ada GPUs respectively. To facilitate efficient training and evaluation with large protein language models (PLMs), we first extracted and de-duplicated amino acid sequences from CDR3a, CDR3b, and peptides. The

corresponding embeddings were pre-computed using PLMs and stored for downstream use. During model training and evaluation, sequence indices were used to retrieve and assemble these pre-computed representations. Each model was trained for 150 epochs with a batch size of 512 and a learning rate of 1×10^{-3} using the AdamW optimizer. A dropout rate of 0.2 was applied to enhance generalization performance.

E Data & Code Availability:

The code and data used in this paper can be found at GitHub repository: <https://github.com/Tulane-Mettu-Landry-Lab/tcr-sr>.