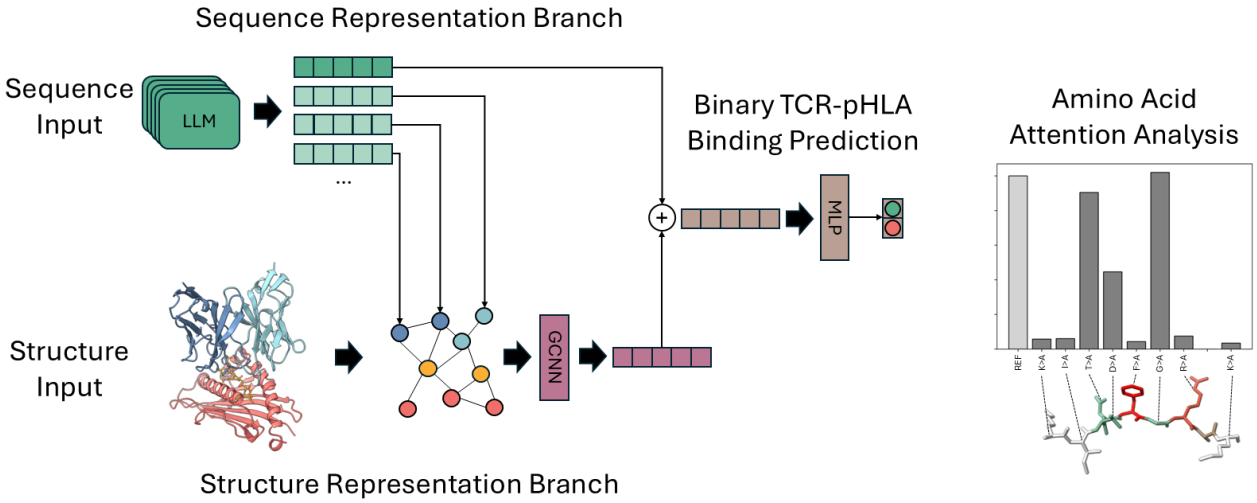
Graphical Abstract

STAG-LLM: Predicting TCR-pHLA Binding with Protein Language Models and Computationally Generated 3D Structures

Jared K. Slone, Minying Zhang, Peixin Jiang, Amanda Montoya, Emily Bontekoe, Barbara Nassif Rausseo, Alexandre Reuben, Lydia E. Kavraki



Highlights

STAG-LLM: Predicting TCR-pHLA Binding with Protein Language Models and Computationally Generated 3D Structures

Jared K. Slone, Minying Zhang, Peixin Jiang, Amanda Montoya, Emily Bontekoe, Barbara Nassif Rausseo, Alexandre Reuben, Lydia E. Kavraki

- Identifying which TCRs will bind strongly to which pHLAs, can serve as a first step in designing personalized immunotherapy treatments.

- STAG-LLM combines sequence data and 3D structural information using a protein language model and geometric deep learning to predict TCR-pHLA binding specificity.

- The incorporation of 3D protein structure data leads to improved TCR-pHLA binding predictions.

- The residue-level attention values produced by our model correlate with *in vitro* experimental results.

STAG-LLM: Predicting TCR-pHLA Binding with Protein Language Models and Computationally Generated 3D Structures

Jared K. Slone[a], Minying Zhang[b], Peixin Jiang[b], Amanda Montoya[b], Emily Bontekoe[b], Barbara Nassif Rausseo[b], Alexandre Reuben[b], Lydia E. Kavraki[a,c,1]

[a]*Computer Science, Rice University, , Houston, 77005, TX, USA*
[b]*Thoracic/Head & Neck Medical Oncology, MD Anderson Cancer Center, , Houston, 77030, TX, USA*
[c]*The Ken Kennedy Institute, Rice University, , Houston, 77005, TX, USA*

Abstract

Background: Strong binding between T cell receptors (TCRs) and peptide–HLA (pHLA) complexes is important for triggering the adaptive immune response. Binding specificity prediction, identifying which TCRs will bind strongly to which pHLAs, can serve as a first step in designing personalized immunotherapy treatments. Existing machine learning (ML) methods to predict binding specificity rely primarily on the amino acid sequences of TCRs and pHLAs to make predictions. However, incorporating the 3D structure and geometry of the TCR-pHLA complex as an additional data modality alongside protein sequence offers a promising approach to improving ML methods for predicting TCR-pHLA binding specificity. Modern computational modeling tools present unprecedented opportunities to incorporate structure data into ML pipelines. We utilize such computational tools to incorporate 3D data into this work.

Results: We present STAG-LLM, a multimodal ML model for predicting TCR-pHLA binding specificity that leverages sequence data and computationally generated 3D protein structures. We show that by combining a protein language model with a geometric deep learning architecture, our method outperforms existing methods even when trained on 3x smaller datasets. To further validate our model, we conduct *in vitro* alanine scanning experiments for four peptides and demonstrate a correlation with the attention weights learned by our model and *in vitro* results. We also seek to address three key challenges that arise from using computationally generated 3D structures in ML pipelines: increased inference costs arising from the need to generate 3D structures, limited training data, and robustness to noise in the generated structures.

Conclusions: STAG-LLM shows tremendous potential for structure-based TCR-pHLA binding prediction methods, offering a foundation for further advancements in using modeled 3D structures to solve problems in immunology and proteomics. We anticipate that the usefulness of STAG-LLM and similar tools will increase in coming years as both protein structure prediction models and large language models continue to advance.

Keywords: structural bioinformatics, proteomics, immunology, protein language model, geometric deep learning, TCR, HLA

1. Background

Accurately predicting TCR-pHLA binding specificity could significantly advance immunotherapy treatments. Cellular immunotherapy, now a core pillar of cancer treatment, relies primarily on the ability of T cells to target and destroy cancer cells as part of the adaptive immune response (1). This process revolves around two key proteins/complexes: the T cell receptor (TCR) and the peptide-human leukocyte antigen (peptide-HLA or pHLA) complex (2). TCRs are surface proteins that enable T cells to distinguish malignant cells from healthy ones. HLAs present a variety of peptides on the surface of cells as pHLA complexes; TCRs recognize and bind to them. Strong binding between the TCR and pHLA is important for triggering the necessary adaptive immune response (2). Binding specificity prediction, identifying which TCRs will bind strongly to which pHLAs, can serve as a first step in designing personalized immunotherapy treatments such as peptide vaccines, adoptive cell therapy, or TCR engineering (3).

Numerous machine learning (ML) methods have been proposed to predict TCR-pHLA binding specificity *in silico*, yet this remains an open problem (3; 4; 5). The immense diversity of the immune system makes this task particularly difficult. It is estimated that there are over 10^{20} possible CD8+ TCRs (6), which can interact with any of 20^9 possible peptides (6) bound to one of the over 28,000 known HLA class I alleles (7). However, at the time of writing, public databases such as McPAS (8) and VDJdb (9) contain fewer than 10^5 unique examples of TCR-pHLA binding that can be used to train ML models. This disparity underscores the complexity of the problem. Adding to the difficulty is the promiscuity of TCR-pHLA interactions; a single TCR can strongly bind to multiple pHLAs, a phenomenon known as T cell cross-reactivity (10; 11). Existing ML methods struggle to accurately predict binding specificity for peptides not represented in their training data, highlighting the limitations of these models given the small scale of available datasets relative to the enormous space of possible TCR-pHLA pairs (4; 5). In addition to further data acquisition, sophisticated

and generalizable methods are needed to produce more accurate TCR-pHLA binding predictions.

Incorporating the 3D structure and geometry of the TCR-pHLA complex as an additional data modality alongside protein sequence offers a promising approach to improving ML methods for predicting TCR-pHLA binding specificity. Most existing computational models focus exclusively on the amino acid sequences of the TCR, peptide, and HLA without utilizing structural features (12; 13; 14; 15; 16). However, considering the geometry and relative orientation of the proteins in TCR-pHLA complexes can be essential, as they have been shown to reveal mechanisms of T cell cross-reactivity that sequence-based analyses fails to capture (17; 18). Historically, the development of structure-based ML methods for TCR-pHLA binding specificity prediction has been hindered by a scarcity of structural data. For example, at the time of writing, the STCRDab database contains only 702 TCR structures complexed with HLA or HLA-like molecules (19). Advances in computational protein modeling now present an opportunity to bridge this gap by enabling accurate *in silico* predictions of TCR-pHLA 3D structures. Tools like TCRmodel2 (20) pave the way for fast structurally informed ML pipelines, which have the potential to significantly improve TCR-pHLA binding specificity predictions.

Most existing approaches that incorporate structural information to predict TCR-pHLA binding focus on analyzing inter-protein contacts between the TCR and the pHLA (21; 22; 23; 24). While statistical methods using contact data have shown encouraging results, the increasing accuracy and reduced noise of predicted 3D structures open the door for leveraging more expressive models, such as deep neural networks, with greater potential for success. In this work we propose a highly expressive architecture, STAG-LLM, that uses both a large language model and a graph convolutional network to predict TCR-pHLA binding specificity. Our method differs from recent works which have sought to incorporate 3D protein structure into their predictions but forwent modeling the entire TCR-pHLA complexes as part of their pipeline, instead choosing to model only the pHLA (25) or the CDR3β loop of the TCR and the peptide (26). These works mention the difficulty of modeling the CDR loops of the TCR (25) and the large amount of public data that only discloses the peptide and CDR3β loop of the TCR (26). Yet, previous works have shown that the CDR3α loop, the HLA, and other CDR loops of the TCR complex are all important when predicting binding specificity so we choose to work with full TCR-pHLA complexes in this work (27; 13; 28; 14).

It has recently been shown that a structure-based deep learning method leveraging wholly modeled TCR-pHLA structures tends to outperform sequence-based ML methods for predicting TCR-pHLA binding affinity when trained and tested on the same datasets (29). Yet, ML approaches in proteomics that rely on modeled 3D structures, such as STAG (Structural TCR And pHLA binding specificity prediction Graph neural network), face three key challenges, which this work addresses. The **first challenge** is the increased inference time due to the computational cost of modeling protein structures. Many state-of-the-art modeling tools often require significant wall time and specialized hardware to generate each structure, making this process resource-intensive and, by extension, expensive (30). The **second challenge** is the inability of structure-based ML methods to learn from partial sequence data effectively. This limitation is potentially significant for TCR-pHLA binding prediction, as the majority of publicly available data include only partial TCR sequences, typically restricted to the CDR3 loops. More information than just the CDR3 amino acid sequences is needed to produce an accurate 3D model of the TCR-pHLA complex, so purely structure-based ML methods cannot make use of this data during training. The **third challenge** is the potential for error propagation. Computational protein modeling of TCR-pHLA complexes is not 100% accurate (30), and inaccuracies in the structural models can lead to inconsistent performance in downstream ML tasks. This work introduces a framework to mitigate these weaknesses of structure-based proteomics ML in the context of TCR-pHLA binding specificity prediction.

To address the three challenges listed and improve prediction accuracy, this work combines geometric deep learning with large language models to create a structurally aware TCR-pHLA binding specificity prediction model. Our method outperforms existing sequence and structure-based methods in terms of accuracy (ROC-AUC), even when existing sequence-based methods are trained on over 3x as much data. Additionally, we show that our model can, for specific cases, replicate experimental alanine scans *in silico*. These advancements mark a significant step forward in the development of robust and effective TCR-pHLA binding specificity prediction tools.

2. Methods

2.1. Model Architecture

STAG-LLM integrates learned sequence embeddings with 3D protein structures by combining a graph convolutional neural network (GCNN) and a transformer-based large language model (LLM). GCNNs can capture critical structural and physicochemical relationships within proteins, translating to outstanding performance across diverse proteomics problems, including protein-protein docking (31), protein-ligand docking (32), protein-ligand binding affinity prediction (33), and protein function prediction (34; 35). GCNNs encode a protein's 3D structure as a graph, where nodes represent atoms or amino acids, and edges are often defined according to spatial distances (36). GCNNs aggregate information across connected nodes via message-passing convolutions, allowing them to derive local and global information about the graph when making predictions. LLMs trained on protein sequences, known as Protein Language Models (PLMs), can capture patterns and key relationships among amino acids. PLMs are typically foundation models trained on vast amounts of unlabeled data, allowing them to implicitly learn relationships governed by physical and evolutionary principles (37; 38). Such PLMs can then be fine-tuned to specific tasks, where they have demonstrated success in proteomics applications such as protein folding (38), protein-ligand binding affinity prediction (39), and protein function prediction (34; 35). STAG-LLM builds on previous efforts that

combined PLMs with GCNNs to achieve superior performance compared to either PLMs or GCNNs alone (34; 35; 40; 41; 42).

STAG-LLM is a multimodal deep learning model that combines a PLM with a GCNN to predict TCR-pHLA binding specificity. The model is comprised of three interconnected components: a PLM, a structure representation branch, and a sequence representation branch. The model's architecture is illustrated in Figure 1.

2.1.1. Sequence Representation Branch

The sequence representation branch is depicted in the top half of Figure 1. A classification token ($[CLS]$) is prepended to the concatenated TCR-pHLA sequence, which is passed through the protein language model, ESM-2. The embedding vector corresponding to the ($[CLS]$) token is treated as an encoding of the entire sequence, as has been done in previous work (43). This vector is then passed through the MLP to produce a binary prediction. This branch complements the structural insights provided by the GCNN module as it considers only the amino acid sequences of the TCR and peptide-HLA.

Protein Language Model. The PLM utilized in STAG-LLM is the 8M-parameter edition of ESM-2 (`esm2_t6_8M_UR50D`) (38). While larger PLMs, such as the 15B-parameter edition of ESM-2, may outperform this smaller model, such exploration is beyond the scope of this work. The ESM-2 model is loaded with its pre-trained weights using the Hugging Face API (44) and is first fine-tuned on a masked token prediction task using concatenated TCR-pHLA sequences separated by a designated token ($[SEP]$). Then, this fine-tuned ESM-2 model is integrated with the STAG-LLM architecture for binding specificity prediction. The ESM-2 model is further fine-tuned for the task of TCR-pHLA binding specificity prediction using a protocol similar to that of LM-GVP (35).

2.1.2. Structure Representation Branch

The structure representation branch is depicted in the bottom half of Figure 1. The structure representation branch encodes the 3D structure of a modeled TCR-pHLA complex as a graph. It uses a GCNN to derive a vector representation of the 3D structure than can be used to predict TCR-pHLA binding specificity.

Graph Construction. The graph construction encodes the 3D structure of the complex in a manner similar to previous works (29; 31), with nodes centered on the carbon-α atoms of each amino acid and connectivity determined by a radius graph of 10.0Å. Edge features are encoded using a radial basis function that captures pairwise distances between connected nodes. The node features are derived from the PLM, as has been done in prior work (35; 40). ESM-2's encoding layer generates embeddings for each amino acid in the concatenated TCR-pHLA sequence. These embeddings are used as node features. The resulting graph is a computational representation of the 3D structure of the TCR-pHLA complex. By nature of its construction, this representation is invariant to rotation and translation of the input complex (29).

GCNN Module. Once the input structure has been encoded as a graph, a GCNN is applied to that graph to predict binding specificity between the TCR and pHLA. This GCNN consists of three heterogeneous transformer convolutions with separate weights learned for different types of edges (e.g. TCR-TCR, TCR-peptide, peptide-HLA). Residual connections follow each convolutional layer. A global max pooling operation aggregates the graph into a single vector, representing the complex's 3D structure. We finalized this architecture after taking inspiration from previous works (29; 35; 42). We also experimented with different hyperparameters in earlier versions of our model. Utilizing a 25% fraction of the total dataset, we performed 5-fold cross validation for versions of STAG-LLM model with different hyperparameters and selected the best performing model configuration according to average AUPRC. The results of these experiments influenced design choices such as our choice in pooling operator (max pooling), our choice in graph convolutional operator (transformer convolution), and the number of convolutional layers in the GCNN (three).

2.1.3. Classification Branch

Once we have obtained a vector from our sequence representation branch and a vector from our structure representation branch, these vectors are averaged to create a single vector representation of the TCR-pHLA pair. This vector is then passed through the multi-layer perceptron (MLP) to obtain a binary prediction. Using average pooling to combine the sequence and structure representations of the TCR-pHLA complex, instead of other possible methods like concatenation, was inspired by previous work (42) and empirical results.

2.2. Data Curation

The dataset used in this work consist of TCR-pHLA pairings and binary labels indicating whether the TCR and the pHLA are a strong or a weak binding pair. Positive samples, or strong binding TCR-pHLA pairs, were curated from the McPAS (8), VDJ (9), and IEDB (45) databases as well as from the 10x genomics public datasets (46). Negative data points were both sampled from the 10x genomics datasets and generated through randomly swapping TCRs, as has been done in previous work (13; 12; 47; 29).

The dataset includes the full amino acid sequences of the TCR, peptide, and HLA, as well as a computationally modeled 3D structural complex for each TCR-pHLA pair. All TCR-pHLA pairings in the dataset were derived from single-cell sequenceing. TCRs are often described in public databases using their V, D, J, and CDR3 regions as opposed to their full amino acid sequences. For such cases, STITCHR (48) was used to translate V, D, J, and CDR3 information into full amino acid sequences. TCRmodel2 (20) was then used to model 3D protein structures from the full sequences. TCRmodel2 produced five 3D conformations for each input complex. Conformations with incident angles outside the range $[0°, 45°]$ or crossing angles outside the range $[5°, 95°]$ were excluded based on established analyses of solved TCR-pHLA structures (49). The remaining 3D conformation with the highest PLDDT score was

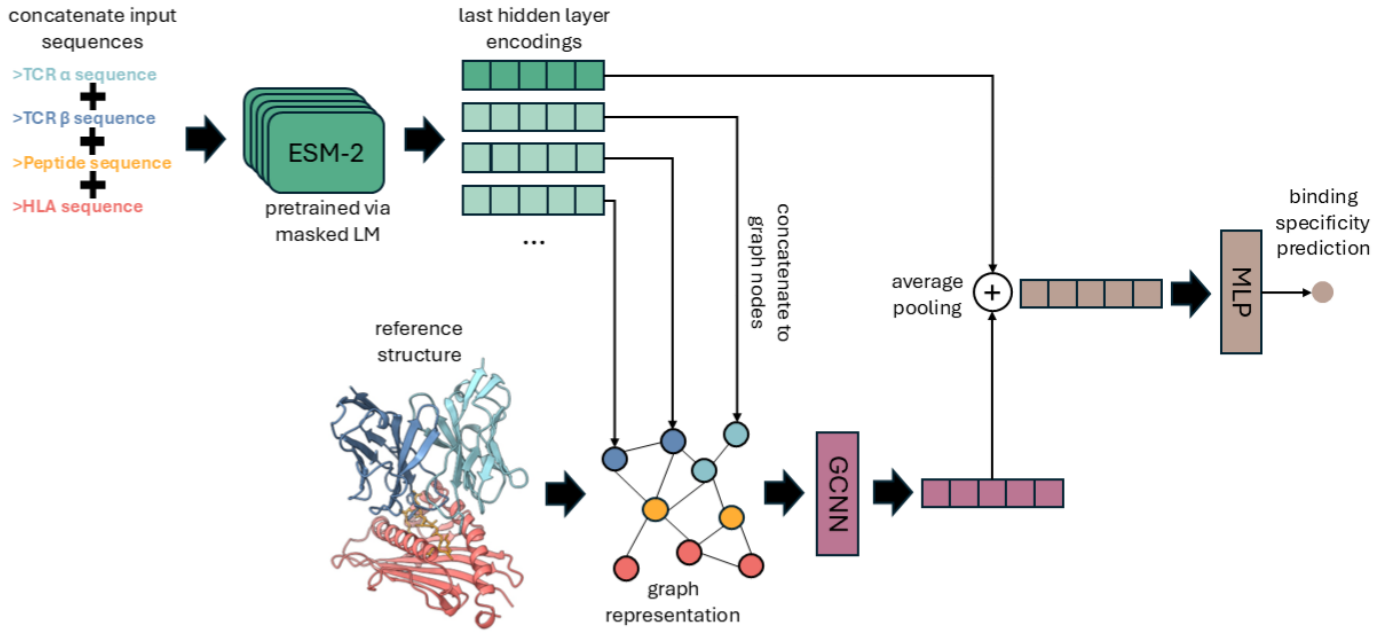


Figure 1: A visualization of STAG-LLM, a deep learning framework for predicting TCR-pHLA binding specificity. First, the ESM-2 model, our protein language model, is trained to predict masked tokens on a large unlabeled set of concatenated TCR, peptide, and HLA sequences. Then this pre-trained model is utilized in STAG-LLM to obtain token-wise embeddings of input sequences. The *sequence representation branch* (top) utilizes the embedding corresponding to the ($[CLS]$) token as the vector representation of the TCR-pHLA sequence. The *structure representation branch* (bottom) utilizes a reference structure generated by TCRmodel2 to construct a graph representation of the geometry of the TCR-pHLA complex. The node features of this graph are derived from the vector encodings produced by the ESM-2 model. A GCNN is used to distill the graph representation of the reference structure into a vector representation. The sequence and structure representations are then fused via average pooling and the TCR-pHLA complex is assigned a binary classification using an MLP.

used in our dataset (50). TCR-pHLA pairs for which no conformation produced acceptable incident and crossing angels were excluded from the final dataset. In total, the primary dataset contains 7,412 positive examples and 38,797 negative examples. A positive example is a TCR-pHLA pair that have been experimentally shown to bind to each other. A negative example is a TCR-pHLA pair that have been experimentally shown not to bind, or a TCR randomly paired with a pHLA for which binding is unlikely (47).

During each cross-validation procedure, the dataset was randomly partitioned into 3 mutually exclusive and collectively exhaustive groups: training (60%), validation (20%), and testing (20%). To prevent data leakage, it was ensured that the Levenstein similarity ratio between the CDR3β sequences of two TCRs paired with the same peptide in different partitions did not exceed 0.9 (Figure 2D). Like with other TCR-pHLA datasets, the dataset we use in this paper is extremely long tailed (see Figure 2B) (51; 52). Although there are over 370 unique peptides in our dataset, the 10 most common peptides account for 82% of the entries. Likewise, some HLA alleles are more common in our dataset than others. HLA-A makes up 80.8% of the dataset, HLA-B makes up 18.9% of the dataset, HLA-C makes just up 0.2% of the dataset and HLA-E makes up the final 0.1% of the dataset. While it is infeasible to eliminate all sources of bias, taking repeated measurements on different partitions of the dataset and using the same partitions to benchmark each model helps to mitigate biases.

2.3. Alanine Scan Protocols

Alanine scanning identified peptide residues that are critical in for T cell activation. Alanine-substituted immunogenic peptides (purity > 85%) were obtained from GeneScript USA, Inc. Target tumor cells were peptide-pulsed for 4 hours, co-cultured with TCR-T cells at a 5:1 ratio, and analyzed by either ELISPOT methods or MIP-1β ELISA methods.

ELISPOT protocol. TCR-T cell reactivity was assessed using IFN-γ Enzyme-link Immunospot (ELISPOT) (53). Plates were coated with anti-human IFN-γ capture antibody, washed, blocked and washed. TCR-T cells were co-cultured at a 5:1 ratio with mapped alanine peptide pulsed target tumor cells for 15-18 hours. Then, IFN-γ monoclonal antibody (Mabtech, 3420-6-1000) was added to the plate and incubated for 1 hr. ExtrAvidine-Alkaline Phosphatase solution (Sigma-Aldrich, E2636) was added and incubated for 1.5 hrs. Spots were observed by adding BCIP/NBT Membrane Alkaline Phosphatase Substrate (Sigma, 11697471001) solution. ELISPOT plates were scanned and counted using ImmunoSpot ELISPOT analyzer (Cellular Technology Ltd).

MIP-1β ELISA protocol. MIP-1β secretion was measured using the Human MIP-1β ELISA Kit (Invitrogen , BMS2030INST). TCR-T cells and peptide-pulsed tumor cells were co-cultured at a 2:1 ratio for 16-18 hrs. Supernatant was collect and added to provided ELISA plates, according to manufacturer instructions, and read on a Synergy microplate

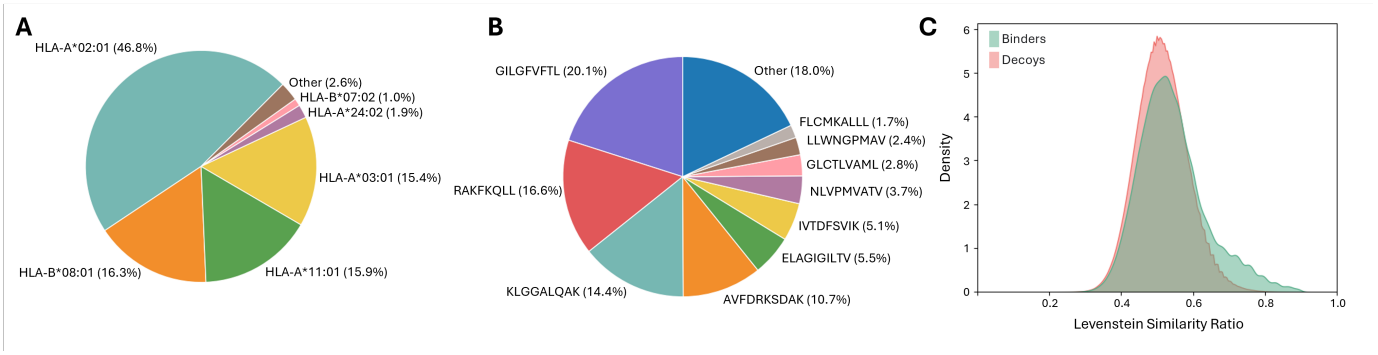


Figure 2: **A: HLA Composition of Our Dataset** - The pie chart shows the percentage makeup of the HLAs in our dataset. HLAs that make up at least 1% of the dataset are listed by name. **B: Peptide Composition of Our Dataset** -The pie chart shows the percentage makeup of the peptides in our dataset. The 10 most common peptides are listed by name. **C: Levenshtein Similarity Between Train/Test Sequences** – For each data partition, we measured how similar TCRs in our test sets were to TCRs in our training/validation datasets. KDE plots show the Levenstein Similarity Ratio for the CDR3β regions of TCRs matched to the same peptide (such that one of these TCRs appears in our training/validation data and one appears in our testing data). Binders are plotted in green while decoys are plotted in red. Note that a large portion of TCRs in the testing sets were paired with peptides that did not occur in the training dataset. These TCRs are excluded from this visualization. More information on the data curation process can be found in *Section 2.2*

reader (BioTek). MIP-1β concentrations were calculated according to the standard curve.

3. Results

3.1. STAG-LLM Outperforms Existing TCR-pHLA Binding Specificity Prediction Models

The six models considered in this work are as follows:

- **STAG-LLM**: Trained on 3D structure plus full TCR, peptide, and HLA sequences
- **STAG**: Trained exclusively on 3D structures (29)
- **NetTCR 2.2**: Trained on amino acid sequences from all six CDR loops (14)
- **TCR-ESM**: Trained on CDR3α and CDR3β sequences (12)
- **ERGO II AE**: Trained on CDR3α and CDR3β sequences (13)
- **ERGO II LSTM**: Trained on CDR3α and CDR3β sequences (13)

All methods evaluated in this section were benchmarked on the same dataset using the same repeated cross-validation framework. The dataset used for the analysis in Figure 3A. was the dataset described in Section 2.2 consisting of 46,209 TCR-pHLA pairs, each with full amino acid sequences and modeled 3D structures. Benchmarking all methods on the same dataset is critical, as the performance of TCR-pHLA binding specificity predictors has been shown to vary widely between datasets (4). Each model was trained, validated, and tested using 15 different Train-Validation-Test partitions of the primary dataset, resulting in 15 distinct ROC-AUC measurements. The training and validation procedures for each model were as suggested in their initial publications.

The results of the repeated cross-validation on the primary dataset are shown in Figure 3A. The colored bars correspond to the median ROC-AUC score recorded for each classifier during cross-validation. Swarm plots show the 15 individual ROC-AUC scores measured for each classifier. The median ROC-AUC scores for each of the six classifiers were: STAG-LLM (0.815), STAG (0.800), NetTCR 2.2 (0.790), TCR-ESM (0.754), ERGO II AE (0.711), ERGO II LSTM (0.691). These results indicate that the methods STAG and STAG-LLM, which leverage the predicted 3D structure of the TCR-pHLA complexes, significantly outperformed the sequence-based methods: NetTCR 2.2 (14), TCR-ESM (12), and ERGO II (13). Additionally, the combination of the LLM and the structure-based graph neural network, STAG-LLM, outperformed the graph neural network model, STAG. We used the Welch's T-test to estimate the statistical significance of the difference in performance for all methods (54). All p-values were below 0.05 and are shown in Table A.4. These results suggest that incorporating structural information enhances the accuracy of TCR-pHLA binding specificity predictions.

3.2. Evaluating the Impact of Training Data Quantity and Composition on Classifier Performance

A limitation of structure-based methods like STAG, is that they cannot be trained on partial protein sequences because the full protein sequence is needed to produce a 3D model. Since most publicly available data on TCR-pHLA binding provides only partial TCR sequences, often just the CDR3 loops, assessing how the inclusion or exclusion of additional training data impacts classifier performance is important. To address this, we conducted an experiment comparing the performance of structure-based and sequence-based models as the size and composition of the training dataset is varied. The results are shown in Figure 3B.

For the experiments shown in Figure 3B, the testing and validation datasets were the same ones sampled from the dataset corresponding to the pan-peptide prediction results shown in

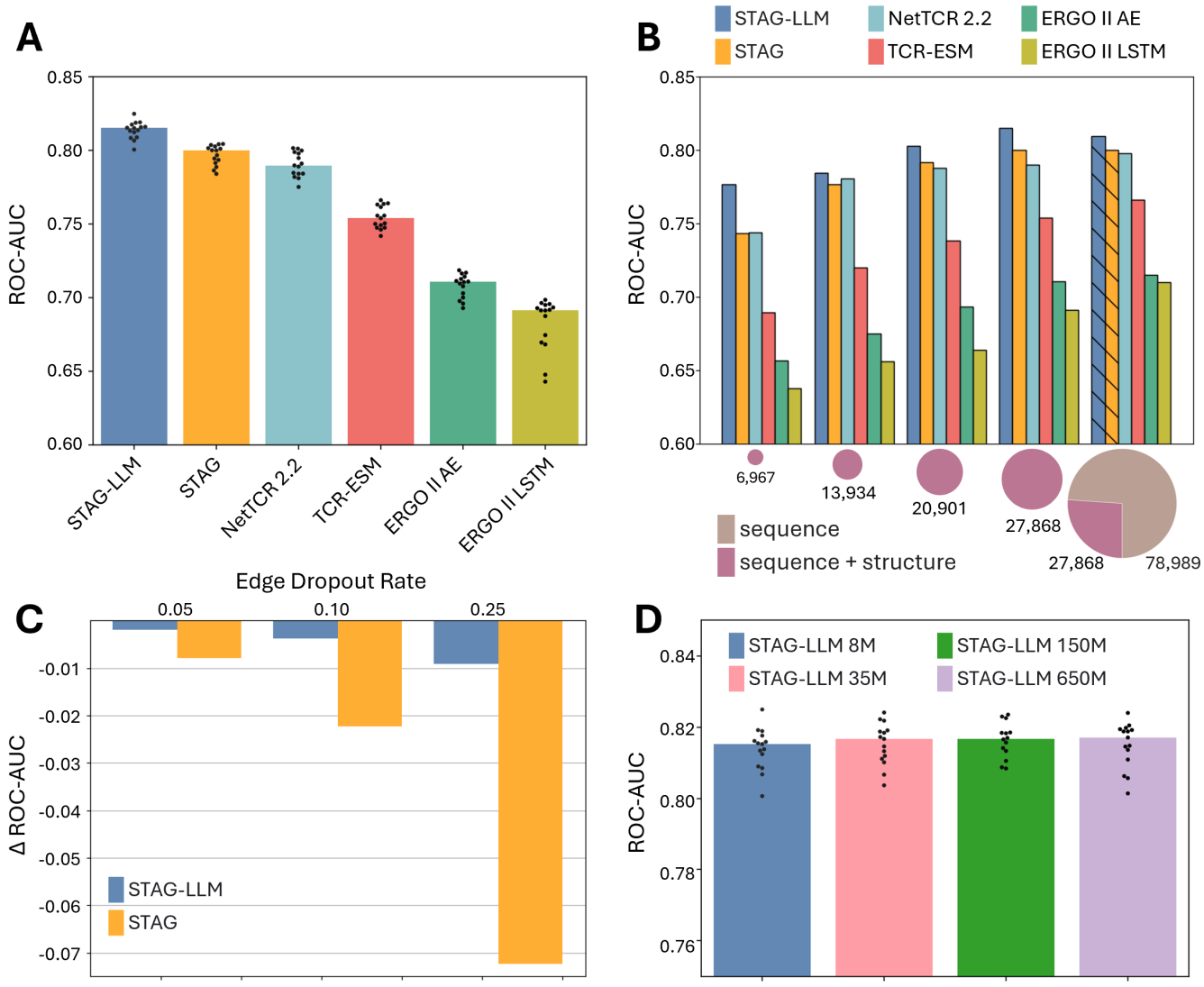


Figure 3: **A: ROC-AUC Scores on Primary Dataset** – Six classifiers (x-axis) were trained and tested on identical partitions of the primary dataset using repeated cross-fold validation. Colored bars show the median ROC-AUC scores (y-axis); swarm plots show the 15 individual scores measured for each classifier. The median ROC-AUC scores were: STAG-LLM (0.815), STAG (0.800), NetTCR 2.2 (0.790), TCR-ESM (0.754), ERGO II AE (0.711), ERGO II LSTM (0.691). Dataset construction is described in *Section 2.2*. The results in Figure 3A are discussed in *Section 3.1*. **B: Classifier Performance vs. Training Data Size** – Five bar clusters (x-axis) represent different training dataset sizes; each cluster includes six bars (one per classifier) showing median ROC-AUC from repeated cross validation. Testing/validation sets were fixed while training data sizes and composition were varied as noted below each cluster. Hatched bars (STAG-LLM and STAG in cluster 5) indicate that the models could not properly train on the extra 78,989 sequence-only samples. The results in Figure 3B are discussed in *Section 3.2*. **C: Classifier Robustness to Noise** – Δ ROC-AUC (y-axis) shown for three dropout rates (x-axis), at which edges were randomly dropped from the graph neural network during inference. A less negative Δ (less negative change in performance) is better. The results in Figure 3C are discussed in *Section 3.3*. **D: LLM Size vs Model Performance** - We experimented with 4 versions of STAG-LLM, each using a different ESM-2 model (eg. with 8M, 35M, 150M, or 650M parameter models). The median ROC-AUC scores were: STAG-LLM 8M (0.815), STAG-LLM 35M (0.817), STAG-LLM 150M (0.817), STAG-LLM 650M (0.817). While the larger LLMs performed slightly better, we did not observe any statistically significant differences between the models. The results in Figure 3D are discussed in *Section 3.4*.

Figure 3A. The training dataset was varied according to Table 1. We conducted experiments down-sampling the training dataset to 25% (resulting in just 6,967 training data points), down-sampling the training dataset to 50% (resulting in 13,934 training data points), down-sampling the training dataset to 75% (resulting in just 20,868 training data points). We also conducted an experiment where we augmented the training dataset with an additional 78,989 sequence-only data points, resulting in over 100,000 training data points for the sequence-only models to learn from.

Train Dataset	Total Training Size	Sequences with Structures	Sequences without Structures
Subsampled 25%	6,967	6,967	0
Subsampled 50%	13,934	13,934	0
Subsampled 75%	20,901	20,901	0
Whole dataset	27,868	27,868	0
Expanded dataset	106,857	27,868	78,989

Table 1: Summary of dataset sizes and the availability of modeled 3D structures.

The results in Figure 3B show that even when the sequence-only methods were trained on datasets more than double or even triple the size of those used by structure-based methods, their performance still lagged behind that of STAG and STAG-LLM. For all methods studied here, increasing the size of the training dataset while keeping the data modalities consistent improved performance. However, incorporating many data points with sequence-only data into the training of STAG-LLM actually diminished its performance, even though reference structures were present in the testing and validation datasets. We attribute this behavior to the network ignoring the modality that is severely missing during training, diminishing the contribution of the structure-based branch. This phenomenon has been seen in other multimodal learning problems (55; 56). The performance of STAG, which cannot utilize sequence-only data, remained unchanged with the inclusion of additional sequence data.

Variable	**Training & Evaluation Configuration**				
	config 1	*config 2*	*config 3*	*config 4*	*config 5*
Fine-tuned LLM	✓	X	✓	✓	X
Ref. struct. in train.	✓	✓	✓	X	X
Ref. struct. in eval.	✓	✓	X	X	X
Median ROC-AUC	**0.815**	0.802	0.797	0.803	0.737
variance	3.23E-5	1.09E-5	8.36E-3	1.35E-4	9.66E-5
p-value with *config 1*	N.A	1.55E-3	7.40E-3	1.63E-3	5.39E-14

Table 2: **Performance Of The STAG-LLM Model With Different Training And Evaluation Settings (Ablation Study)** "Fine-tuned LLM" specifies whether the protein language model (PLM) was fine-tuned for the binding specificity task or used without additional tuning. "Ref. struct. in train." indicates whether reference structures were included during training, enabling the graph-based branch of the model to be trained. "Ref. struct. in eval." indicates whether reference structures were included during evaluation, allowing the graph-based branch of the model to contribute to predictions. Five distinct combinations of these three variables were evaluated, giving five different training/testing configurations. The five configurations represent all meaningful configurations; for example, testing using structure without first training the GCNN portion of the model would not make sense.

3.3. STAG-LLM Shows Greater Resilience to Errors in the Reference Structure Compared to STAG

A major challenge of using modeled 3D structures as inputs to a machine learning pipeline for predicting TCR-pHLA binding specificity that STAG-LLM addresses is the potential for inaccuracies in the 3D models to propagate errors into the final predictions (29; 30). Here, we demonstrate that the STAG-LLM architecture mitigates this error propagation, showing greater resilience to inaccuracies in input structures compared to STAG. Due to the graph construction process used in both STAG and STAG-LLM, inaccuracies in the input structures affect only the edges of the graph, not the composition of the nodes (29). Therefore, to assess resilience, we conducted experiments where edges in the input graph were randomly dropped, simulating small errors in the reference structure. The results, shown in Figure 3C, indicate that STAG-LLM maintains a much higher performance under these conditions than STAG, highlighting its robustness to errors in the initial reference 3D model of the TCR-pHLA complex.

3.4. Model Performance with Different Training and Evaluation Configurations

To evaluate the contribution of each component of the STAG-LLM architecture to its overall performance, we conducted an ablation study. This study investigated the impact of fine-tuning the PLM and its embeddings, as well as the independent contributions of the sequence and structure representation branches. The results are summarized in Table 2. These results indicate that the full STAG-LLM model, incorporating all three components, achieves optimal performance.

Utilizing the ESM-2 model "out-of-the-box" produced good results for the structure-only portion of our architecture, as shown under *config 2* of Table 2. The ROC-AUC values produced when fixing the PLM embeddings and training only the GCNN were roughly equivalent to the ROC-AUC values produced by the original STAG model, which uses a GCNN with physiochemical properties as node features. This suggests that amino acid-wise embeddings produced by ESM-2 are as roughly effective as physiochemical properties for node embeddings in a GCNN trained to predict TCR-pHLA binding specificity.

Fine-tuning the PLM is essential to achieving adequate performance from the sequence representation branch of the architecture. Without this fine-tuning, the sequence-only branch performed poorly. One explanation is that the ([CLS]) token was not trained to represent anything in the original ESM-2. However, other works have gotten around this by averaging the embeddings from the last layer before performing sequence classification, (12). This method provided no statistically significant improvement here. Notably, using just the *Sequence Representation Branch* without first finetuning the PLM underperformed compared to TCR-ESM (12), a model that uses ESM-1v to generate individual embeddings for the CDR3a, CDR3b, and peptide sequences, which are concatenated and classified with an MLP. This result suggests that without fine-tuning, concatenating sequence-specific embeddings from the PLM is more effective than producing a single representation for the entire concatenated sequence. This observation could be due to the fact that ESM-2 was originally trained on homodimers rather than protein complexes. Finally, while fine-tuning the PLM improves performance, increasing the size of the PLM does not have a significant impact on the performance of the overall model (see Figure 3 D.). This could be due to dataset limitations.

The sequence representation branch exhibited significant variability when trained exclusively on data with reference structures. This indicates that our architecture does not ensure proper training of the sequence-only portion of the model if all examples seen during training have reference structures. While the median ROC-AUC for the fully trained model without reference structures during evaluation was a respectable 0.797, results ranged widely, with a minimum ROC-AUC of 0.729. This variability was addressed in the final version of our model by periodically incorporating some training examples without reference structures. For such examples only the *Sequence Representation Branch* was updated during backpropagation. This resulted in more consistent performance from the *Sequence Rep-*

resentation Branch.

Enabling accurate predictions from the sequence-only branch is a major advantage of STAG-LLM over purely structural ML methods like the original STAG. One major limitation of structure-only methods is the computational cost of generating 3D structures for each TCR-pHLA pair. STAG-LLM mitigates this limitation by allowing for initial predictions using only the PLM and sequence branch. TCR-pHLA pairs of interest can be triaged according to the sequence-only predictions before 3D models are selectively generated. Then, final predictions can be made on the 3D models using the complete STAG-LLM architecture.

3.5. Correlation Between Attention Values from STAG-LLM's Structure Representation Branch and in Vitro Alanine Scans

Alanine scanning of peptides in the TCR-peptide-HLA complex provides valuable insights into which residues are most critical for eliciting immune responses and helps assess the risk of T cell cross-reactivity (57). It has been shown in other domains that attention coefficients extracted from trained GCNNs "can reveal hidden node relations and quantify the importance of nodes" (58). Upon investigation, we discovered that in several cases, attention values generated by the GCNN component of the STAG-LLM architecture can be used to identify residues that contribute significantly to TCR-pHLA binding. To further benchmark STAG-LLM and assess the utility of these attention values, we conducted four *in vitro* alanine scanning experiments and compared the results to our model's predictions. We observed favorable results for three out of the four alanine scans, shown in Figure 4. Importantly, none of the native peptides, mutated peptides, or the corresponding TCRs appeared in the model's training dataset.

3.5.1. Spearman Correlation with Attention Values

For the three alanine scans shown in Figure 4, an inverse correlation was observed between the importance assigned to peptide residues by STAG-LLM's attention mechanism and the change in T cell activation, measured as cytokine production, when those residues were substituted with alanine.

- For the peptide *KITDFGRAK*, the Spearman correlation between the attention values and interferon-gamma ($IFN-\gamma$) production measured via ELISpot assays was -0.999 (p-value: 1.4×10^{-24}).
- For the peptide *YLVPIQFPV*, the Spearman correlation with macrophage inflammatory protein-1 beta ($MIP-1\beta$) production measured via ELISA was -0.427 (p-value: 0.251).
- For the peptide *IYTWIEDHF*, the correlation with $MIP-1\beta$ production was -0.783 (p-value: 0.066).

Aggregating the results across all three peptides that STAG-LLM correctly classified as binding to their TCRs, the overall Spearman correlation between the attention values given by STAG and our alanine scan data was -0.493 (p-value: 0.0169), demonstrating that STAG-LLM correctly focuses on residues within the peptide that are most influential in TCR binding.

3.5.2. Insights from Model Attention Weights and Structural Analysis

STAG-LLM consistently assigned high attention values to residues in the middle of peptides that exhibit prominent structural features, such as aromatic rings. 3D models generated by TCRmodel2 suggest that these prominent features may extend toward the TCR when the peptides are in complex with HLAs. It has been shown that the loss of structurally prominent features in the peptide can cause changes in TCR specificity (59; 57). This appears to be the case for the three peptides shown in Figure 4.

- *KITDFGRAK* peptide: STAG-LLM assigns high attention to the phenylalanine at position 5 and the arginine at position 7, which feature prominent side chains that may extend toward the TCR. The alanine scan, Figure 4 (left panel), indicates that the loss of either of these side chains will cause the TCR to no longer react against the peptide, highlighting their importance. Irrespective of the model's attention values, the amino acids at positions 1, 2, and 9 were revealed by the alanine scan to be very important to T cell activation. This could be because these residues occur at or close to the canonical anchor positions for HLA-A*03:01 and affect peptide binding to the HLA, indirectly reducing the potential for T cell activation.

- *YLVPIQFPV* peptide: For this peptide, Figure 4 (center panel), the model assigned high attention to the phenylalanine at position 7, which our alanine scan results show plays a critical role in eliciting an immune response. Our alanine scan data also indicates that the glutamine at position 6 is important for TCR recognition. However, despite its prominent side chain, STAG-LLM assigned low attention to the glutamine at position 6. This may be because the side chain is buried in the HLA cleft in the 3D structure, limiting its relevance to TCR interaction. Contrary to the attention value assigned by the model, the amino acid at position 8 was shown by the alanine scan to be important to T cell activation. This could be because this residue is close to the canonical anchor position for HLA-A*02:01 and affects peptide binding to the HLA, indirectly reducing the potential for T cell activation.

- *IYTWIEDHF* peptide: As shown in Figure 4 (right panel), high attention was given to the tryptophan at position 4, with its aromatic rings, and to the glutamic acid at position 6, a negatively charged residue. Both residues were confirmed by the alanine scan to play key roles in TCR recognition. The histidine at position 8, however, received low attention from STAG-LLM, despite its prominent side chain. This result agrees with the alanine scan, which showed that a loss of this side chain does not result in a significant change in TCR specificity.

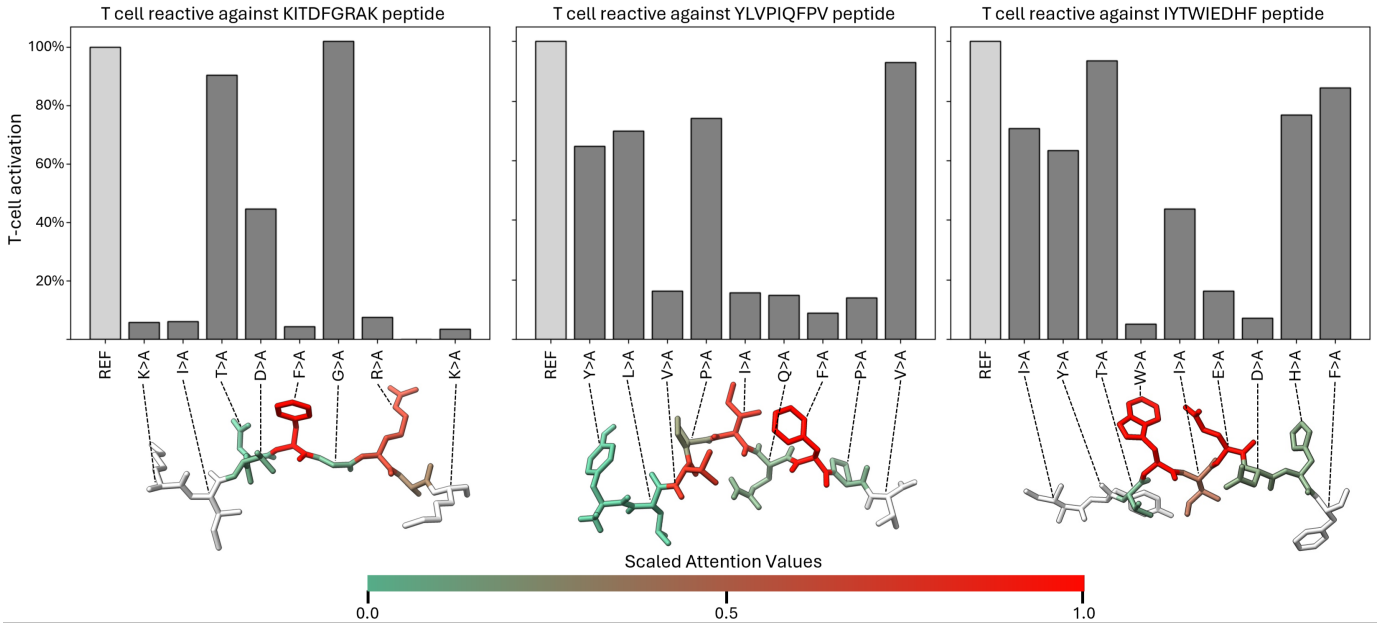


Figure 4: **Comparison Of Model Attention Weights And Alanine Scans** Each panel depicts an alanine scan for a distinct peptide-TCR pair. The peptides are, from left to right: *KITDFGRAK*, *YLVPIQFPV*, *IYTWIEDHF*. The bar charts in each panel show the average measurement of cytokines for the reference peptide (REF) and the mutated peptides across wet lab experiments. All measurements are normalized to the REF peptide for each scan. For the *KITDFGRAK* the panel on the left, $IFN-\gamma$ production was measured via ELISpot assays. In the center and right panels, $MIP-1\beta$ production is measured via ELISA. Underneath each of the bar plots is a 3D model of the peptide colored according to the attention weights from TCR nodes to that peptide node in STAG-LLM's GCNN module. Residues in the peptide whose node had no edges connecting it to the TCR are colored in white and attention values for those nodes are treated as undefined in this analysis. These colors are determined by our computational analysis of the peptides using STAG-LLM and are presented in this figure to be juxtaposed with the experimental results shown in the bar charts.

3.5.3. Misclassification and Observed Limitations

While STAG-LLM correctly predicted binding for the native peptides *KITDFGRAK*, *YLVPIQFPV*, and *IYTWIEDHF* and their corresponding TCRs, even though these peptides were not featured in our training dataset, it failed to classify the *VVGACGVGK* peptide and its corresponding TCR as a binding pair. Additionally, the attention values for the residues in the *VVGACGVGK* peptide showed no correlation with the *in vitro* alanine scan that was conducted for this peptide. Unlike the correctly classified peptides, *VVGACGVGK* lacks residues with prominent structural features, which may have contributed to the misclassification.

Despite the strong correlation between STAG-LLM's attention values and alanine scan data, the model's binding specificity predictions (binary outputs) varied considerably for the mutated peptides. While STAG-LLM correctly predicted strong binding for the native peptides shown in Figure 4, it made incorrect predictions for several mutated peptides, resulting in a lackluster Spearman correlation of 0.227 (p-value: 0.265) between changes in model predictions and *in vitro* alanine scan results. These discrepancies highlight the current limitations of datasets and the challenges of making predictions for unseen peptides and TCRs. While the attention mechanism in STAG-LLM can successfully identify the key residues in TCR-pHLA binding, further research is needed to improve the accuracy of binding specificity prediction methods for novel or mutated peptide-TCR pairs.

3.6. Model Performance on Unseen Peptides Varies

Model \ Peptide	GILGFVFTL	ELAGIGILTV	FLCMKALLL	LLWNGPMAV	AVFDRKSDAK	KLGGALQAK	IVTDFSVIK	NLVPMVATV	RAKFKQLL	GLCTLVAML
STAG-LLM	0.799	0.764	0.628	**0.617**	0.584	0.538	**0.557**	0.507	*0.412*	*0.355*
STAG	**0.845**	**0.815**	0.644	0.585	**0.600**	**0.567**	0.523	**0.532**	**0.444**	*0.396*
NetTCR 2.2	0.534	0.808	**0.662**	*0.422*	0.559	0.554	*0.460*	*0.440*	*0.371*	*0.338*
TCR-ESM	0.587	0.605	*0.452*	*0.398*	0.522	0.506	*0.390*	*0.450*	*0.319*	*0.397*
ERGO II AE	0.620	0.553	0.515	0.525	0.517	0.515	*0.422*	*0.495*	*0.385*	*0.427*
ERGO II LSTM	0.603	0.562	0.522	0.511	0.503	*0.488*	*0.438*	0.510	*0.406*	**0.463**

Table 3: Performance of each ML model on the unseen peptide prediction task for each of the 10 peptides tested. The top performing score is boded for each peptide. ROC-AUC scores that were worse than random guessing (0.5) are highlighted in red.

To evaluate how the models perform on unseen peptides, we conducted an experiment for each of the 10 most common peptides in our dataset. For each of the 10 peptides, we first constructed a dedicated test set containing all instances of that peptide. The remaining data, which did not include any instances of the held-out peptide, was then randomly partitioned into five distinct 80% training and 20% validation pairs. Using these five partitions, we retrained each model five times and measured its performance on the corresponding held-out test set. The median ROC-AUC value for each model on each unseen peptide test set is reported in Table 3.

The models tested demonstrated significant variability in their performance when predicting TCR binding for the unseen peptides. Notably, the structure-based methods, STAG and STAG-LLM, displayed slightly better performance than the sequence-based methods for most unseen peptides. Accuracy on some peptides was high, with models achieving ROC-AUCs greater than 0.8 despite those peptides not being included in the training or validation data. The performance on the majority of

unseen peptides, however, was poor across all models, sometimes even worse than random guessing. These findings are consistent with previously reported challenges in generalizing to unseen peptide-MHC-TCR data (4; 5; 60). While 5 of the 10 peptides considered here have TCR-pHLA crystal structures associated with them in the PDB (*GILGFVFTL ELAGIGILTV NLVPMVATV RAKFKQLL GLCTLVAML*) and the *LLWNGPMAV* peptide has a pHLA crystal structure in the PDB, there is no apparant advantage for the structure based methods on predicting binding for these peptides when they are held out of the training and validation sets, as both structure models performed worse than random on 2 of the 5.

4. Discussion

In this study, we approach TCR-pHLA binding specificity prediction as a supervised machine learning task as opposed to an unsupervised task. Unsupervised methods in this domain range from clustering techniques (61) to the use of uncertainty quantification metrics derived from general protein structure prediction tools such as AlphaFold (62; 63). Regarding binding prediction for novel or unseen peptides, supervised ML models typically exhibit lackluster performance (4; 29), likely due to the limited amount of available training data relative to the space of all possible TCR-pHLA pairings (6). Unsupervised methods tend to show more consistent performance on unseen peptides (64), yet they still fall short in terms of overall predictive accuracy (65). Additionally, the performance of unsupervised predictors still varies across peptides (64) and typically requires some amount of labeled data for calibration. Given the limitations inherent in both supervised and unsupervised approaches, we advocate for the continued development of supervised models, such as STAG-LLM. These models can be used both as standalone tools and in combination with unsupervised techniques to improve prediction accuracy in clinically relevant scenarios.

With the distinction between supervised and unsupervised classifiers in mind, we only compared STAG-LLM to other supervised classifiers in this work. To ensure fairness in these comparisons, all models were retrained, validated, and tested using the same datasets. When constructing the train, validation, test splits for the repeated cross fold validation, we did not explicitly enforce peptide-strict partitioning. Nonetheless, up to 41.5% of the TCR-pHLA pairings in the test sets contained peptides that did not appear at all in the training or validation sets. As a result, each model's performance on unseen peptides is partially reflected in the reported ROC-AUC scores from our repeated cross-validation experiments (see Figure 3A). However, performance on these unseen peptides varied wildly, both per-peptide and across training splits, making it difficult to draw firm conclusions regarding generalizability to novel peptides. For this reason, we focus our empirical analysis on the 15x repeated cross-validation results on our large dataset of over 46,200 TCR-pHLA pairs. All classifier comparisons in this setting yielded statistically significant differences ($p - values < 0.05$) (see Table A.4), providing strong support for the effectiveness of STAG-LLM in the supervised setting. Finally, we posit that as larger datasets become available, supervised methods, such as deep learning, have the potential to outperform unsupervised approaches in TCR-pHLA binding prediction, mirroring the trajectory observed in protein folding and contact prediction, where deep learning surpassed statistical techniques as data became more ubiquitous and methodologies more sophisticated (66; 67; 68; 50).

5. Conclusion

Predicting TCR-pHLA binding specificity remains a significant challenge. In this work we introduced STAG-LLM, a multimodal deep learning model that leverages both sequence and 3D structure data to outperform existing methods. We addressed three critical limitations of structure-based models: high inference costs, limited training data, and sensitivity to errors in reference 3D structures. STAG-LLM allows users to partially circumvent the cost of generating 3D models, as TCR-pHLA pairs of interest can first be triaged using just the sequence representation branch. We demonstrated that STAG-LLM outperforms existing sequence-based methods, even when those methods are trained on over three times more data. We also showed that STAG-LLM is more resilient to errors in the initial 3D model than previous structure-based deep learning methods for TCR-pHLA binding specificity prediction. As part of this work, we conducted alanine scans and compared the results to predictions made by STAG-LLM and the attention weights assigned by the model to individual amino acids during inference. We observed a correlation between attention weights observed in the STAG-LLM model and the outcome of *in vitro* alanine scans.

The results of our experiments reinforce the importance of incorporating 3D structure data in predicting TCR-pHLA binding specificity. Structural information improves the accuracy of binding predictions by capturing the spatial and physicochemical features of TCR-pHLA interactions that sequence alone cannot fully convey. As more wet lab data becomes available for training, and computational tools for modeling TCR-pHLA structures continue to become more precise, we expect structure-based ML models to further improve in performance and generalizability. The developments presented in this work pave the way for further advancements in using modeled 3D structures to solve problems in immunology and proteomics.

As future work, we anticipate that advancements in in modeling protein flexibility, especially within the highly dynamic CDR loops of the TCR, will be essential for refining predictions. Capturing conformationally diverse binding modes between the TCR and pHLA through molecular dynamics simulations or other computational methods could even shed light on the underlying physiochemical mechanisms for T cell activation in addition to improving the accuracy of *in-silico* binding predictions.

6. List of abbreviations

- TCR: T-cell receptor

- MHC: Major Histocompatability Complex
- pMHC: peptide-MHC
- HLA: Human Lucocyte Antigen
- pHLA: peptide-HLA
- ELISA: ELISA Enzyme-Linked Immunosorbent Assay
- ELISpot: Enzyme-Linked ImmunoSpot
- ML: Machine Learning
- LLM: Large Language Model
- PLM: Protein Language Model
- ESM: Evolutionary Scale Modeling
- GNN: Graph Neural Network
- CNN: Convolutional Neural Network
- GCNN: Graph Convolutional Neural Network
- MLP: Multi-Layer Perceptron
- ROC: Receiver Operating Characteristic
- ROC-AUC: Area Under the ROC Curve
- AUPRC: Area Under Precision-Recall Curve
- KDE: kernel density estimate
- 3D: three dimensional
- STAG: Structural TCR And pMHC binding specificity prediction Graph neural network

7. Availability of Data and Materials

All datasets used in this article, the code for the STAG-LLM architecture, and the trained model weights are available on GitHub at `https://github.com/KavrakiLab/STAG-LLM` . A google colab notebook that can be used to easily run the STAG-LLM model given PDB inputs is also linked to the github page.

8. Competing Interests

No competing interest is declared.

9. Funding

J.K.S. is supported by a training fellowship from the Gulf Coast Consortia on the Training Program in Biomedical Informatics, National Library of Medicine 5T15LM007093. L.E.K. is supported in part by National Institutes of Health NIH U01CA258512. A.R. is supported by the Exon 20 Group, Rexanna's Foundation for Fighting Lung Cancer, the Waun Ki Hong Lung Cancer Research Fund, MD Anderson's Lung Cancer Moon Shot, the Petrin Fund, the Salgado Family Charitable Fund, the Troper Wojcicki Foundation, an NIH/NCI R21 (FP00017376), a US Department of Defense Lung Cancer Research Program Idea Development Award (FP00017091), an AACR Career Development Award (FP00018394), the University Cancer Foundation via the Institutional Research Grant program at the University of Texas MD Anderson Cancer Center, the Happy Lungs Project, an EGFR Resisters/LUNGevity Foundation EGFR-positive research award (FF2024-00063001), a RETpositive/LUNGevity Translational Research Award Program (FF2022-00061026), a Cancer Prevention & Research Institute of Texas High Impact High Reward Award (RP210137) and a Cancer Prevention & Research Institute of Texas Individual Investigator Research Award for Computational Systems Biology of Cancer (RP230363). This work was supported in part by the NOTS cluster operated by Rice University's Center for Research Computing (CRC).

10. CRediT authorship contribution statement

Jared K. Slone: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing, review and editing. **Minying Zhang**: Resources. **Peixin Jiang**: Resources, Investigation. **Amanda Montoya**: Investigation, Writing - review and editing. **Emily Bontekoe**: Investigation, Writing - original draft, Writing review and editing. **Barbara Nassif Rausseo**: Investigation. **Alexandre Reuben**: Conceptualization, Funding Acquisition, Supervision, Writing review and editing. **Lydia E. Kavraki**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing - original draft, Writing review and editing.

11. Acknowledgments

The authors thank Alex Bock and Felix Quintana from the Kavraki Lab, Jefferson Hernandez from Rice University, and Mauricio M. Rigo from the Hackensack Meridian Center for Discovery and Innovation for many helpful discussions. ChimeraX (69) was used to visualize and analyze protein structures.

Appendix A. Statistical Significance of Results

To assess the statistical significance of the results in section 3.1, Welch's t-test for independence was applied to the distributions of ROC-AUC values obtained from repeated cross-validation (15 values per model). This analysis was conducted

for each of the six machine learning models tested on the primary dataset. All p-values were below 0.05.

	STAG-LLM	STAG	NetTCR 2.2	TCR-ESM	ERGO II AE	ERGO II LSTM
STAG-LLM		5.86E-08	1.29E-09	4.54E-20	1.58E-26	1.35E-21
STAG	5.86E-08		2.63E-02	1.20E-15	7.70E-24	8.17E-20
NetTCR 2.2	1.29E-09	2.63E-02		9.48E-13	9.03E-22	9.12E-19
TCR-ESM	4.54E-20	1.20E-15	9.48E-13		1.07E-15	2.07E-14
ERGO II AE	1.58E-26	7.70E-24	9.03E-22	1.07E-15		2.59E-05
ERGO II LSTM	1.35E-21	8.17E-20	9.12E-19	2.07E-14	2.59E-05	

Table A.4: **p-values from Welch's t-test for independence** Welch's t-test for independence was performed on the distributions of ROC-AUC values measured during repeated cross validation for each ML model tested on the primary dataset, 6 models and 15 measurements per model.

To assess the statistical significance of the results in section 3.4, Welch's t-test for independence was applied to the distributions of ROC-AUC values obtained from repeated cross-validation (15 values per model). This analysis was conducted for four different versions of STAG-LLM, each with a different sized version of ESM2. None of the observed differences were statistically significant.

	STAG-LLM 8M	STAG-LLM 35M	STAG-LLM 150M	STAG-LLM 650M
STAG-LLM 8M		0.505	0.234	0.620
STAG-LLM 35M	0.505		0.619	0.888
STAG-LLM 150M	0.234	0.619		0.536
STAG-LLM 650M	0.620	0.888	0.536	

Table A.5: **p-values from Welch's t-test for independence** Welch's t-test for independence was performed on the distributions of ROC-AUC values measured during repeated cross validation for each version of STAG-LLM tested on the primary dataset, 4 model sizes and 15 measurements per model.

Appendix B. Attention Values and Alanine Scan Data

The table shows the results for the 4 alanine scans presented in this paper. "amino acid" refers to the amino acids in each peptide. "INFγ" refers to the ELISPOT counts as described in section 2.3. "MIP-1β" refers to the measured MIP-1β secretion as described in section 2.3. "INFγ ratio" and "MIP-1β ratio" refer to the measured ratio of the cytokines measured for the alanine scan peptides vs the reference peptide. "untrained LLM attention" refers to the average attention values returned by the ESM-2 model out of the box. "fine tuned LLM attention" refers to the average attention values returned by the ESM-2 model after it was fine tuned as part of the STAG-LLM architecture. "GCNN TCR attention" refers to the attention weights from TCR nodes to that peptide-amino-acid node in STAG-LLM's GCNN module. The TCRs that were paired with the pHLAs in our experiments are identified by their CDR3β regions and are given in the following table.

amino acid	K	I	T	D	F	G	R	A	K
INFγ	129	132	1.98e+03	979	99	2.23e+03	166	N.A.	76
MIP-1β	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
INFγ ratio	0.0590	0.0603	0.905	0.447	0.0452	1+	0.0759	N.A.	0.0347
MIP-1β ratio	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
Δ STAG-LLM pred	-0.592	-0.796	0.452	0.417	-0.495	-1.70	-0.479	0	2.47
untrained LLM attention	2.33e-03	2.42e-03	2.21e-03	2.27e-03	2.01e-03	2.18e-03	2.02e-03	1.90e-03	2.22e-03
fine tuned LLM attention	2.39e-03	2.22e-03	2.38e-03	2.22e-03	2.12e-03	2.05e-03	2.44e-03	2.22e-03	2.41e-03
GCNN TCR attention	N.A.	N.A.	0.0331	0.101	0.675	0.031	0.216	N.A.	N.A.
amino acid	Y	L	V	P	I	Q	F	P	V
INFγ	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
MIP-1β	298	321	75	341	72	68	40	65	428
INFγ ratio	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
MIP-1β ratio	0.649	0.699	0.163	0.742	0.157	0.148	0.0871	0.142	0.932
Δ STAG-LLM pred	0.452	-0.825	0.0715	-0.883	-0.582	-0.202	-0.872	-0.596	-0.0567
untrained LLM attention	2.37e-03	2.23e-03	2.26e-03	2.19e-03	2.22e-03	2.03e-03	2.24e-03	2.10e-03	2.00e-03
fine tuned LLM attention	2.17e-03	2.22e-03	2.12e-03	2.45e-03	2.36e-03	2.17e-03	2.18e-03	2.31e-03	1.92e-03
GCNN TCR attention	0.015	0.01	0.33	0.116	0.34	0.074	0.461	0.06	N.A.
amino acid	I	Y	T	W	I	E	D	H	F
INFγ	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
MIP-1β	513	488	713	40	334	124	55	574	697
INFγ ratio	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
MIP-1β ratio	0.670	0.638	0.944	0.0575	0.467	0.166	0.0745	0.733	0.911
Δ STAG-LLM pred	0.487	0.777	0.327	0.664	0.987	0.04	0.85	-0.636	0.278
untrained LLM attention	2.85e-03	2.36e-03	2.38e-03	2.77e-03	2.94e-03	2.15e-03	2.16e-03	1.95e-03	2.11e-03
fine tuned LLM attention	2.69e-03	1.74e-03	2.72e-03	2.71e-03	2.75e-03	2.16e-03	2.31e-03	2.08e-03	2.17e-03
GCNN TCR attention	N.A.	N.A.	0.028	0.25	0.111	0.25	0.048	0.046	N.A.
amino acid	V	V	G	A	C	G	V	G	K
INFγ	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
MIP-1β	335	0	0	N.A.	552	41	23	473	0
INFγ ratio	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
MIP-1β ratio	0.704	0	0	N.A.	1+	0.0861	0.0483	0.994	0
Δ STAG-LLM pred	-1.19e-07	-5.9e-08	-1.19e-07	0	5.96e-08	-1.79e-07	-2.38e-07	-1.79e-07	4.59e-06
untrained LLM attention	2.45e-03	2.07e-03	2.33e-03	1.96e-03	2.34e-03	1.89e-03	1.91e-03	2.07e-03	2.29e-03
fine tuned LLM attention	2.53e-03	3.26e-03	2.17e-03	2.29e-03	2.52e-03	1.83e-03	2.57e-03	1.97e-03	2.47e-03
GCNN TCR attention	N.A.	N.A.	0.014	0.014	0.023	0.076	0.297	0.019	0.24

Table C.6: CR CDR3B sequences, and corresponding patent numbers for the TCRs paired with each of the four peptides referenced in table section 3.5.

Peptide	**TCR's CDR3β**	**Patent Number**
KITDFGRAK	CASSYSRDSIREQYF	PCT/US2024/013763
YLVPIQFPV	CASSYAGPGELFF	PCT/US2023/082847
IYTWIEDHF	CASSLGGRSQETQYF	PCT/US2024/036413
VVGACGVGK	CASSFSRGPETQYF	PCT/US2024/010342

Appendix C. TCR Identifiers

Here we provide identifiers for the four TCRs used in the alanine scan experiments.

References

[1] M. Y. Want, Z. Bashir, R. A. Najar, T Cell Based Immunotherapy for Cancer: Approaches and Strategies, Vaccines 11 (4) (2023) 835.

[2] S. A. Rosenberg, N. P. Restifo, J. C. Yang, R. A. Morgan, M. E. Dudley, Adoptive cell transfer: a clinical path to effective cancer immunotherapy, Nature Reviews Cancer 8 (4) (2008) 299–308.

[3] D. Hudson, R. A. Fernandes, M. Basham, G. Ogg, H. Koohy, Can we predict T cell specificity with digital biology and machine learning?, Nature Reviews Immunology 23 (8) (2023) 511–521. doi:10.1038/s41577-023-00835-3.

[4] F. Grazioli, A. Mösch, P. Machart, K. Li, I. Alqassem, T. J. O'Donnell, M. R. Min, On TCR binding predictors failing to generalize to unseen peptides, Frontiers in Immunology 13 (2022). doi:10.3389/fimmu.2022.1014256.

[5] M. Nielsen, A. Eugster, M. F. Jensen, M. Goel, A. Tiffeau-Mayer, A. Pelissier, S. Valkiers, M. R. Martínez, B. Meynard-Piganeeau, V. Greiff, T. Mora, A. M. Walczak, G. Croce, D. L. Moreno, D. Gfeller, P. Meysman, J. Barton, Lessons learned from the IMMREP23 TCR-epitope prediction challenge, ImmunoInformatics 16 (2024) 100045. doi:https://doi.org/10.1016/j.immuno.2024.100045.

[6] C. Soto, R. G. Bombardi, M. Kozhevnikov, R. S. Sinkovits, E. C. Chen, A. Branchizio, N. Kose, S. B. Day, M. Pilkinton, M. Gujral, et al., High frequency of shared clonotypes in human T cell receptor repertoires, Cell reports 32 (2) (2020).

[7] J. Robinson, D. J. Barker, S. G. Marsh, 25 years of the IPD-IMGT/HLA database, HLA 103 (6) (Jun 2024). doi:10.1111/tan.15549.

[8] N. Tickotsky, T. Sagiv, J. Prilusky, E. Shifrut, N. Friedman, McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences, Bioinformatics 33 (18) (2017) 2924–2929.

[9] M. Shugay, D. V. Bagaev, I. V. Zvyagin, R. M. Vroomans, J. C. Crawford, G. Dolton, E. A. Komech, A. L. Sycheva, A. E. Koneva, E. S. Egorov, et al., VDJdb: a curated database of T-cell receptor sequences with known antigen specificity, Nucleic acids research 46 (D1) (2018) D419–D427.

[10] G. Petrova, A. Ferrante, J. Gorski, Cross-reactivity of T cells and its role in the immune system, Critical Reviews™ in Immunology 32 (4) (2012).

[11] S.-H. Chiou, D. Tseng, A. Reuben, V. Mallajosyula, I. S. Molina, S. Conley, J. Wilhelmy, A. M. McSween, X. Yang, D. Nishimiya, et al., Global analysis of shared t cell specificities in human non-small cell lung cancer enables hla inference and antigen discovery, Immunity 54 (3) (2021) 586–602.

[12] S. Yadav, D. S. Vora, D. Sundar, J. K. Dhanjal, TCR-ESM: Employ-

ing protein language embeddings to predict TCR-peptide-MHC binding, Computational and Structural Biotechnology Journal 23 (2024) 165–173.
[13] I. Springer, N. Tickotsky, Y. Louzoun, Contribution of T cell receptor alpha and beta CDR3, MHC typing, V and J genes to peptide binding prediction, Frontiers in Immunology 12 (2021). doi:10.3389/fimmu.2021.664514.
[14] M. F. Jensen, M. Nielsen, Nettcr 2.2 - improved tcr specificity predictions by combining pan- and peptide-specific training strategies, loss-scaling and integration of sequence similarity, eLife (Oct. 2023). doi:10.1101/2023.10.12.562001.
[15] M. Zhang, Q. Cheng, Z. Wei, J. Xu, S. Wu, N. Xu, C. Zhao, L. Yu, W. Feng, Berttcr: a bert-based deep learning framework for predicting cancer-related immune status based on t cell receptor repertoire, Briefings in Bioinformatics 25 (5) (2024) bbae420. doi:10.1093/bib/bbae420.
[16] Z. Yu, M. Jiang, X. Lan, Heterotcr: A heterogeneous graph neural network-based method for predicting peptide-tcr interaction, Communications Biology 7 (1) (2024) 684.
[17] D. K. Sasmal, W. Feng, S. Roy, P. Leung, Y. He, C. Cai, G. Cao, H. Lian, J. Qin, E. Hui, et al., TCR–pMHC bond conformation controls TCR ligand discrimination, Cellular & Molecular Immunology 17 (3) (2020) 203–217.
[18] K. Saotome, D. Dudgeon, K. Colotti, M. J. Moore, J. Jones, Y. Zhou, A. Rafique, G. D. Yancopoulos, A. J. Murphy, J. C. Lin, et al., Structural analysis of cancer-relevant tcr-cd3 and peptide-mhc complexes by cryoem, Nature Communications 14 (1) (2023) 2401.
[19] J. Leem, S. H. P. de Oliveira, K. Krawczyk, C. M. Deane, STCRDab: the structural T-cell receptor database, Nucleic acids research 46 (D1) (2018) D406–D412.
[20] R. Yin, H. V. Ribeiro-Filho, V. Lin, R. Gowthaman, M. Cheung, B.-G. Pierce, TCRMODEL2: High-resolution modeling of T-cell receptor recognition using Deep Learning, Nucleic Acids Research 51 (W1) (2023). doi:10.1093/nar/gkad356.
[21] X. Lin, J. T. George, N. P. Schafer, K. Ng Chau, M. E. Birnbaum, C. Clementi, J. N. Onuchic, H. Levine, Rapid assessment of T-cell receptor specificity of the immune repertoire, Nature Computational Science 1 (5) (2021) 362–373. doi:10.1038/s43588-021-00076-1.
[22] A. Wang, X. Lin, K. N. Chau, J. N. Onuchic, H. Levine, J. T. George, Racer-m leverages structural features for sparse t cell specificity prediction, Science Advances 10 (20) (2024) eadl0161.
[23] L. Gao, Y. Zhang, F. Ge, S. Li, Y. Guo, J. Song, D.-J. Yu, Structure-directed pan-specific t-cell receptor–peptide-major histocompatibility complex interaction prediction, Journal of Chemical Information and Modeling (2025).
[24] V. K. Karnaukhov, D. S. Shcherbinin, A. O. Chugunov, D. M. Chudakov, R. G. Efremov, I. V. Zvyagin, M. Shugay, Structure-based prediction of t cell receptor recognition of unseen epitopes using tcren, Nature Computational Science 4 (7) (2024) 510–521.
[25] J. Ge, J. Wang, Q. Ye, L. Pan, Y. Kang, C. Shen, Y. Deng, C.-Y. Hsieh, T. Hou, Trap: a contrastive learning-enhanced framework for robust tcr–pmhc binding prediction with improved generalizability, Chemical Science 16 (22) (2025) 9881–9894.
[26] H. Ji, X.-X. Wang, Q. Zhang, C. Zhang, H.-M. Zhang, Predicting tcr sequences for unseen antigen epitopes using structural and sequence features, Briefings in Bioinformatics 25 (3) (2024).
[27] A. Montemurro, V. Schuster, H. R. Povlsen, A. K. Bentzen, V. Jurtz, W. D. Chronister, A. Crinklaw, S. R. Hadrup, O. Winther, B. Peters, et al., NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired tcr and sequence data, Communications Biology 4 (1) (2021). doi:10.1038/s42003-021-02610-3.
[28] Y. Han, Y. Yang, Y. Tian, F. J. Fattah, M. S. v. Itzstein, Y. Hu, M. Zhang, X. Kang, D. M. Yang, J. Liu, et al., pan-mhc and cross-species prediction of t cell receptor-antigen binding, bioRxiv (2023) 2023–12.
[29] J. K. Slone, A. Conev, M. M. Rigo, A. Reuben, L. E. Kavraki, TCR-pMHC Binding Specificity Prediction from Structure Using Graph Neural Networks, IEEE Transactions on Computational Biology and Bioinformatics (2024) 1–10doi:10.1109/TCBBIO.2024.3504235.
[30] H. N. Le, M. V. de Freitas, D. A. Antunes, Strengths and limitations of web servers for the modeling of TCRpMHC complexes, Computational and Structural Biotechnology Journal 23 (2024) 2938–2948. doi:https://doi.org/10.1016/j.csbj.2024.06.028.
[31] O.-E. Ganea, X. Huang, C. Bunne, Y. Bian, R. Barzilay, T. S. Jaakkola, A. Krause, Independent SE(3)-equivariant models for end-to-end rigid protein docking, in: International Conference on Learning Representations, 2022.
[32] J. Wang, N. V. Dokholyan, MedusaDock 2.0: Efficient and accurate protein–ligand docking with constraints, Journal of chemical information and modeling 59 (6) (2019) 2509–2515.
[33] S. Li, J. Zhou, T. Xu, L. Huang, F. Wang, H. Xiong, W. Huang, D. Dou, H. Xiong, GIaNt: Protein-Ligand Binding Affinity Prediction via Geometry-Aware Interactive Graph Neural Network, IEEE Transactions on Knowledge and Data Engineering 36 (5) (2024) 1991–2008. doi:10.1109/TKDE.2023.3314502.
[34] V. Gligorijević, P. D. Renfrew, T. Kosciolek, J. K. Leman, D. Berenberg, T. Vatanen, C. Chandler, B. C. Taylor, I. M. Fisk, H. Vlamakis, et al., Structure-based protein function prediction using graph convolutional networks, Nature Communications 12 (1) (May 2021). doi:10.1038/s41467-021-23303-9.
[35] Z. Wang, S. A. Combs, R. Brand, M. R. Calvo, P. Xu, G. Price, N. Golovach, E. O. Salawu, C. J. Wise, S. P. Ponnapalli, et al., LM-GVP: An extensible sequence and structure informed deep learning framework for protein property prediction, Scientific Reports 12 (1) (Apr 2022). doi:10.1038/s41598-022-10775-y.
[36] R. Fasoulis, G. Paliouras, L. E. Kavraki, Graph representation learning for structural proteomics, Emerging Topics in Life Sciences 5 (6) (2021) 789–802.
[37] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, R. Fergus, Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences, PNAS (2019). doi:10.1101/622803.
[38] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, et al., Language models of protein sequences at the scale of evolution enable accurate structure prediction, bioRxiv (2022).
[39] H. Y. I. Lam, J. S. Guan, X. E. Ong, R. Pincket, Y. Mu, Protein language models are performant in structure-free virtual screening, Briefings in Bioinformatics 25 (6) (2024) bbae480. doi:10.1093/bib/bbae480.
[40] F. Wu, L. Wu, D. Radev, J. Xu, S. Z. Li, Integration of pre-trained protein language models into geometric deep learning networks, Communications Biology 6 (1) (2023) 876.
[41] L. M. Blaabjerg, N. Jonsson, W. Boomsma, A. Stein, K. Lindorff-Larsen, Ssemb: A joint embedding of protein sequence and structure enables robust variant effect predictions, Nature Communications 15 (1) (2024) 9646.
[42] M. Li, L. Kang, Y. Xiong, Y. G. Wang, G. Fan, P. Tan, L. Hong, Sesnet: sequence-structure feature-integrated deep learning method for data-efficient protein engineering, Journal of Cheminformatics 15 (1) (2023) 12.
[43] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding (2019).
[44] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Huggingface's transformers: State-of-the-art natural language processing (2020).
[45] R. Vita, J. A. Overton, J. A. Greenbaum, J. Ponomarenko, J. D. Clark, J. R. Cantrell, D. K. Wheeler, J. L. Gabbard, D. Hix, A. Sette, et al., The immune epitope database (IEDB) 3.0, Nucleic acids research 43 (D1) (2015) D405–D412.
[46] 10x Genomics, A new way of exploring immunity–linking highly multiplexed antigen recognition to immune repertoire and phenotype, Tech. rep (2019).
[47] A. Montemurro, L. E. Jessen, M. Nielsen, NetTCR-2.1: Lessons and guidance on how to develop models for TCR specificity predictions, Frontiers in Immunology 13 (2022). doi:10.3389/fimmu.2022.1055151.
[48] J. M. Heather, M. J. Spindler, M. H. Alonso, Y. I. Shui, D. G. Millar, D. S. Johnson, M. Cobbold, A. N. Hata, Stitchr: stitching coding TCR nucleotide sequences from V/J/CDR3 information, Nucleic Acids Research 50 (12) (2022) e68–e68.
[49] B. G. Pierce, Z. Weng, A flexible docking approach for prediction of t cell receptor–peptide–mhc complexes, Protein Science 22 (1) (2012) 35–46. doi:10.1002/pro.2181.
[50] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger,

K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., Highly accurate protein structure prediction with AlphaFold, Nature 596 (7873) (2021) 583–589.

[51] Y. Gao, Y. Gao, Y. Fan, C. Zhu, Z. Wei, C. Zhou, G. Chuai, Q. Chen, H. Zhang, Q. Liu, Pan-Peptide meta learning for T-cell receptor–antigen binding recognition, Nature Machine Intelligence 5 (3) (2023) 236–249. doi:10.1038/s42256-023-00619-3.

[52] M. Cai, S. Bang, P. Zhang, H. Lee, ATM-TCR: TCR-epitope binding affinity prediction using a multi-head self-attention model, Frontiers in Immunology 13 (2022). doi:10.3389/fimmu.2022.893247.

[53] M. Zhang, J. Fritsche, J. Roszik, L. J. Williams, X. Peng, Y. Chiu, C.-C. Tsou, F. Hoffgaard, V. Goldfinger, O. Schoor, et al., Rna editing derived epitopes function as cancer antigens to elicit immune responses, Nature communications 9 (1) (2018) 3919.

[54] B. L. Welch, The significance of the difference between two means when the population variances are unequal, Biometrika 29 (3/4) (1938) 350–362.

[55] W. Yao, K. Yin, W. K. Cheung, J. Liu, J. Qin, Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 16416–16424.

[56] R. Wu, H. Wang, H.-T. Chen, G. Carneiro, Deep multimodal learning with missing modality: A survey, arXiv preprint (2024).

[57] S. J. Turner, K. Kedzierska, H. Komodromou, N. L. La Gruta, M. A. Dunstone, A. I. Webb, R. Webby, H. Walden, W. Xie, J. McCluskey, et al., Lack of prominent peptide–major histocompatibility complex features limits repertoire diversity in virus-specific cd8+ t cell populations, Nature immunology 6 (4) (2005) 382–389.

[58] W. Gu, F. Gao, X. Lou, J. Zhang, Discovering latent node information by graph attention network, Scientific reports 11 (1) (2021) 6967.

[59] D. A. Antunes, M. M. Rigo, M. V. Freitas, M. F. Mendes, M. Sinigaglia, G. Lizée, L. E. Kavraki, L. K. Selin, M. Cornberg, G. F. Vieira, Interpreting T-cell cross-reactivity through structure: implications for TCR-based cancer immunotherapy, Frontiers in immunology 8 (2017) 1210.

[60] Y. Zhang, Z. Wang, Y. Jiang, D. R. Littler, M. Gerstein, A. W. Purcell, J. Rossjohn, H.-Y. Ou, J. Song, Epitope-anchored contrastive transfer learning for paired cd8+ t cell receptor–antigen recognition, Nature Machine Intelligence 6 (11) (2024) 1344–1358.

[61] H. Huang, C. Wang, F. Rubelt, T. J. Scriba, M. M. Davis, Analyzing the mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening, Nature Biotechnology 38 (10) (2020) 1194–1202. doi:10.1038/s41587-020-0505-4.

[62] P. Bradley, Structure-based prediction of T cell receptor: peptide-MHC interactions, Elife 12 (2023) e82813.

[63] S. N. Deleuran, M. Nielsen, Nettcr-struc, a structure driven approach for prediction of tcr-pmhc interactions, bioRxiv (2025) 2025–03.

[64] I. Organizers, Immrep25: Tcr specificity prediction challenge, `https://kaggle.com/competitions/immrep25`, kaggle (2025).

[65] Z. S. Ghoreyshi, J. T. George, Quantitative approaches for decoding the specificity of the human t cell repertoire, Frontiers in Immunology 14 (2023) 1228873.

[66] S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S.-I. Lee, C. J. Langmead, Learning generative models for protein fold families, Proteins: Structure, Function, and Bioinformatics 79 (4) (2011) 1061–1078.

[67] H. Kamisetty, S. Ovchinnikov, D. Baker, Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era, Proceedings of the National Academy of Sciences 110 (39) (2013) 15674–15679.

[68] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, E. Aurell, Improved contact prediction in proteins: using pseudolikelihoods to infer potts models, Physical Review E—Statistical, Nonlinear, and Soft Matter Physics 87 (1) (2013) 012707.

[69] E. C. Meng, T. D. Goddard, E. F. Pettersen, G. S. Couch, Z. J. Pearson, J. H. Morris, T. E. Ferrin, Ucsf chimerax: Tools for structure building and analysis, Protein Science 32 (11) (2023) e4792.