Fusion-pMT: Biological Language Modeling for Tri-Molecular Binding in Immunogenicity Prediction

Anonymous Author(s)

Abstract

Recent advancements in multimodal techniques and large language models (LLMs) offer a new perspective on handling biological sequences through biological language modeling. One particularly critical yet underexplored challenge lies in modeling the tripartite interaction among peptide, MHC, and TCR—an essential step in understanding T cell-mediated immunity and improving immunogenicity prediction. In this paper, we propose **Fusion-pMT**, a biological language modeling framework that (1) learns unified representations of the three molecular inputs by leveraging their common structure as amino acid sequences, and (2) fuses the representations of each sequence to enable interaction among heterogeneous molecular inputs, aligning with the stepwise nature of immune recognition. On this foundation, Fusion-pMT effectively supports both pairwise and tripartite interaction modeling among peptide, MHC, and TCR. Moreover, its parameter-sharing design reduces memory usage during inference, making it lightweight and practical for biological applications. To validate its effectiveness, we conduct comprehensive experiments covering both pairwise and tripartite interactions (including out-of-distribution evaluation) and demonstrate that our Fusion-pMT consistently outperforms state-of-the-art baselines across all benchmarks. The model implementation and code, as well as a complete manuscript with technical appendix, are provided on [2] and will be open-sourced upon publication.

CCS Concepts

• **Computing methodologies** → **Machine learning**.

Keywords

Immunogenicity Prediction, Tri-Molecular Binding, Biological Language Modeling, Sequence Fusion

ACM Reference Format:

Anonymous Author(s). 2018. Fusion-pMT: Biological Language Modeling for Tri-Molecular Binding in Immunogenicity Prediction. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 16 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The success of multimodal techniques and large language models (LLMs) has demonstrated the remarkable ability of Transformers to process diverse sequential data, including biological languages [30]. This breakthrough has catalyzed a growing body of research in the data mining and machine learning community to model molecules through bio-sequence representations [49, 50], paving the way for deeper explorations of complex biomolecular interactions in immunology.

Among these problems, one of the critical challenges is understanding how biological sequences dictate immune recognition—particularly, the intricate interactions among the ❶ antigenic **peptide**, ❷ major histocompatibility complex (**MHC**), and ❸ T cell receptor (**TCR**). Although structural data can provide valuable insights into molecular recognition, anonated peptide-MHC–TCR complex structures are extremely scarce, and AlphaFold [34] struggles to accurately model these complexes without docking priors [5, 6, 10]. These limitations motivate us to continue explore the immunogenicity prediction from a language perspective. Particularly, immunogenicity prediction can be viewed as an immunological conversation, where the core principle is self–foreign discrimination [60]. The distinction between non-self and self-peptides parallels the distinguishing of foreign languages from native language [57], as illustrated in Figure 1a.

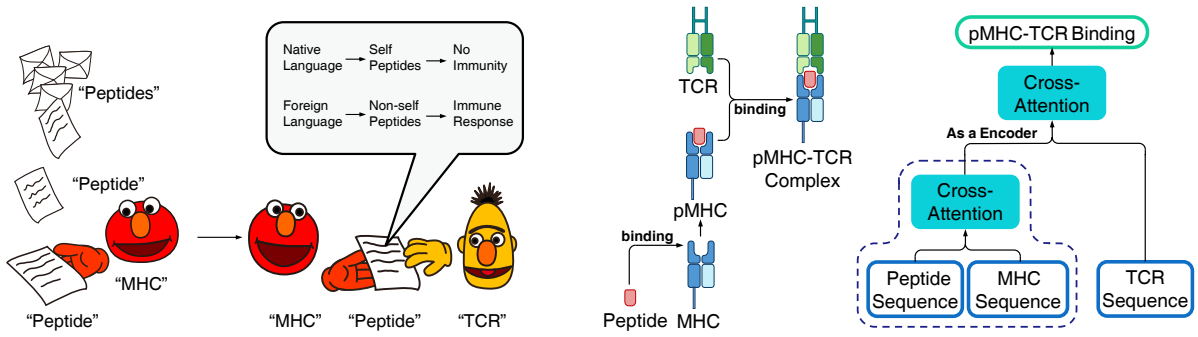
Although a number of prior studies have investigated immunogenicity interaction prediction [15, 46, 62], their discussions were limited to two of the three molecules, i.e., pairwise interactions. This significantly restricts their clinical applicability, as no matter from a biological view, which requires capturing the complete biological process, or from a computational standpoint, where the law of total variance [59] emphasizes the need to account for all sources of variation. This further highlights the urgent need to model the full tripartite interaction, especially given the increasing demand in areas like designing cancer vaccines [53, 63] and guiding personalized immunotherapy [18, 24, 48].

In response to this critical problem, pMTnet [43] is one of the earliest works to explore the interaction between peptides, MHC, and TCRs for immunogenicity prediction. However, due to the significant length difference in TCR sequences, their approach represents TCR as a separate vector, distinct from the other two sequences, leading to suboptimal modeling of the complex tri-molecular binding. Building on this work, PISTE [19] directly encodes the three sequences into a sliding-attention transformer. However, neglecting the real biological process raises concerns about whether it may compromise performance.

Guided by the hope that our model offers better interpretability for tri-molecular binding [45], we aim to provide a model that is capable of ① preserving the unique characteristics of each sequence while ② maintaining the consistency of the model structure with the real biological process. For ①, drawing on the transformative capabilities of LLMs, which handle sequences of varying lengths without losing structural information [17], we propose a unified token embedding technique to simultaneously represent all three

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX



(a) Immunological Conversation. **(b) Biological Process vs Model Architecture.**

Figure 1: The role of pMHC-TCR in adaptive immunity and the correspondence between our model architecture and the biological process. (More details will be introduced in Section 3.1.) (a) The immunological conversation. The self-foreign discrimination relies on the sequences of immunological molecules. (b) The workflow of model training takes peptide-MHC binding as pre-training support for pMHC-TCR binding predictions (right panel) by mimicking the biological process (left panel). The dashed box indicates the pre-training module.

molecules, which can also simplify the model architecture. For ②, we model the tri-molecular binding as shown in Figure 1b. Specifically, immune response activation is driven by two sequential steps: (a) the binding of peptide and MHC, forming the pMHC [38], and (b) the subsequent binding of TCR and pMHC [28]. In addition, we employ a shared embedding layer and encoder for Fusion-pMT to learn unified representations across all three molecular inputs by leveraging their common structure as amino acid sequences. This design enables the model to capture deep underlying relationships among peptides, MHC molecules, and TCRs, thereby improving its capacity for immunogenicity prediction. Moreover, by eliminating the need for separate encoders for each molecular type, this approach reduces the total parameter count by approximately two-thirds, substantially lowering storage requirements and memory footprint. As a result, the compact architecture can run efficiently on CPUs, making it more accessible for practical use in biological research settings without requiring specialized hardware.

Overall, equipped with the language modeling techniques from representation learning, we propose our **Fusion-pMT**, which can efficiently integrates information from all three sequences, offering insights for future models aimed at integrating more complex biological languages. To further demonstrate its efficiency, we conduct extensive experiments and ablation studies on real-world datasets across diverse prediction tasks, including both pairwise and tripartite interactions, as well as out-of-distribution generalization. In summary,

- We develop a new model for peptide-MHC-TCR triad binding, **Fusion-pMT**, which maintains the sequence form during data transform and aligns with the real biological process.
- We introduce representation learning techniques (unified token embedding + multimodal fusion) to model the sequence fusion, improving the performance of immunogenicity prediction.
- We evaluate the effect of the representation learning techniques we adopt, which further validates the effectiveness and versatility of our design from empirical aspects. The analysis improves the practical relevance of our approach and its potential to crack immunogenicity prediction.

2 Related Works

The prediction of interactions among peptides, MHC, and TCR is crucial in immunoinformatics. However, most methods are focusing on TCR-antigen specificity or peptide-MHC Class I binding, with only a few addressing peptide-MHC-TCR triad binding due to its complexity and the scarcity of experimental data. We identify immunological language modeling and molecular binding mechanisms as key factors and review related works accordingly (summarized in Table 1 for the reader's convenience). In addition, we also discuss the opportunities and challenges of incorporating structural information into pMHC–TCR modeling.

Immunological language modeling. Previously, traditional methods represented immunological sequences in non-sequence forms. For instance, Montemurro et al. [46, NetTCR] employed CNN encoders for TCR and antigenic peptides, and models such as Jin et al. [31, DeepAttentionPan] and Kalemati et al. [37, CapsNet-MHC] enhanced traditional CNNs with attention layers to improve feature extraction. With advances in representation learning, sequence modeling techniques have gained popularity. TransPHLA [15, 16] incorporated self-attention modules to capture complex dependencies. Ye et al. [64, STMHCpan] instead modeled peptide-MHC interactions as graphs, introducing graph neural networks to the field.

Molecular binding mechanisms. NetTCR [46] employed direct concatenation of hidden embeddings, while TransPHLA [15] utilized self-attention. UniTCR [20] integrated RNA sequence data with TCR analytics via cross-attention, though its clinical relevance is limited by not accounting for peptide-MHC binding before TCR interaction.

Models for Peptide-MHC-TCR Triad Binding. pMTnet [43] is the first model proposed for directly modeling peptide-MHC-TCR triad binding. This model highly relies on two pre-trained modules

Table 1: Comparison of common models for immunological sequence binding, highlighting differences in model components and concatenation methods.

Model	MHC Modeling	Peptide Modeling	TCR Modeling	Binding Mechanism
STMHCpan (2023)	peptide-MHC Graph	peptide-MHC Graph	N/A	Star-Transformer
TransPHLA (2022, 2021)	Self-Attention	Self-Attention	N/A	Self-Attention
PISTE (2024)	Self-Attention	Self-Attention	Self-Attention	Sliding-attention
netMHCpan (2024)	LSTM	LSTM	N/A	Concat
CcBHLA (2023)	BiLSTM	BiLSTM	N/A	CNN
ESM-2 (2023)	Transformer	Transformer	Transformer	Concat
UniTCR (2024)	N/A	N/A	Self-Attention	Cross-Attention
DeepAIR (2023)	N/A	N/A	Self-Attention	Gate-Based Attention
NetTCR (2021)	N/A	1D CNN	1D CNN	Concat
pMTnet (2021)	LSTM	LSTM	Autoencoder	Concat

netMHCpan [35, 51] and Tessa [67] to encode three sequences, respectively. Specifically, it employs a vector concatenation strategy to model TCR-pMHC interactions. PISTE [19] further introduced directly feeding the three sequences into the sliding-attention transformer for prediction, while leveraging sliding-attention to capture the physics-driven dynamics. Nevertheless, this approach neglects the two sequential steps of the real biological processes involved in immune response, compromising its interpretability.

Incorporating Structural Information: Opportunities and Challenges. While combining sequence and structure information is generally advantageous in protein modeling, structural annotations for pMHC–TCR complexes are extremely scarce (≈350 known 3D structures [36], covering 1% of our sequence data). Furthermore, to the best of our knowledge, only a single study [5] has attempted to incorporate structural information into tripartite immunogenicity prediction. The authors used AlphaFold [34] to predict pMHC–TCR complex structures, with 12 known docking templates guiding the folding process and thereby introducing strong structural priors. Given that the method was validated on 130 samples, its applicability to broader scenarios remains uncertain. Although structure prediction tools such as AlphaFold [1, 34] have achieved remarkable success in modeling individual protein structures, accurately capturing triad pMHC-TCR complexes remains a challenge. In this study, we therefore focus on sequence-level modeling across peptide, MHC, and TCR, aiming to prioritize biologically meaningful candidates for downstream immunological investigation and targeted experimental validation.

3 Preliminaries

In this section, we first outline the core molecules of the immune system in Section 3.1, then discuss the computational challenges of multi-sequence biological problems in Section 3.2, and finally introduce the cross-attention mechanism in Section 3.3.

3.1 Biological Sequences in Immune Systems

Adaptive immunity hinges on two key processes: **antigen presentation** and **antigen recognition**. Antigen-presenting cells (APCs) first bind antigenic peptides to MHC, forming peptide-MHC (pMHC) complexes, and then T cells **recognize** these complexes via the T-cell receptor (TCR), forming pMHC-TCR complexes [28]. T cells harbor highly diverse TCRs composed of α and β chains, wherein β-chain diversity is central to distinguishing self from non-self antigens [47].

Consequently, successful immune responses require both MHC-mediated presentation and TCR-mediated recognition. A broader overview of immune molecules is provided in [2, Appendix B.1], and the embedding methods of amino acids are deferred to [2, Appendix B.2].

3.2 Molecular Binding Tasks

Research on immunological sequence binding highlights several challenges: ❶ **Peptide-MHC binding** is crucial for antigen presentation, and useful for vaccine development. However, not all peptides binding to MHC can form a pMHC complex that also binds to TCR, limiting the model's reflection of entire cellular immunity. ❷ **Peptide-TCR binding** is critical for T-cell activation and T-cell therapies. However, we notice Peptide-TCR binding requires a suitable MHC, which is missed in this task. Therefore, we specifically dismiss this task. ❸ **Peptide-MHC-TCR binding** is vital for understanding cellular immunity, and useful for vaccine development. It offers a holistic view of immune recognition, facilitating better disease-combating strategies. Our paper is therefore committed to this holistic challenge ❸, and a quick adaptation can be applicable to task ❶ (see the discussion of Fusion-pM / Stage 1 in Section 4.3) due to the alignment of our method with the real biological process.

In general, for the input sequences $\boldsymbol{X}_p, \boldsymbol{X}_M, \boldsymbol{X}_T$ (corresponding to peptide, MHC, and TCR sequences respectively), the binding task ❸ can be formulated as modeling the following probability:

$$\mathbb{P}\left\{\boldsymbol{X}_p, \boldsymbol{X}_M, \boldsymbol{X}_T \text{ altogether trigger an immune response}\right\},$$

which matches the modeling in language tasks and justifies the usage of representation learning techniques within. Similarly, task ❶ induces the modeling of the probability

$$\mathbb{P}\left\{\boldsymbol{X}_p, \boldsymbol{X}_M \text{ form a pMHC complex}\right\},$$

which corresponds to "antigen presentation" and is the necessary step for the tri-molecular binding process. Moreover, real-world immunogenicity experiments provide non-binary and probabilistic assessments as supervised signals, different from regular deep learning tasks; nonetheless, common cross-entropy loss can already cater such irregular labels, considering that cross-entropy loss refers to the KL divergence between the label distribution (i.e., non-binary and probabilistic assessments) and the model estimation distribution.

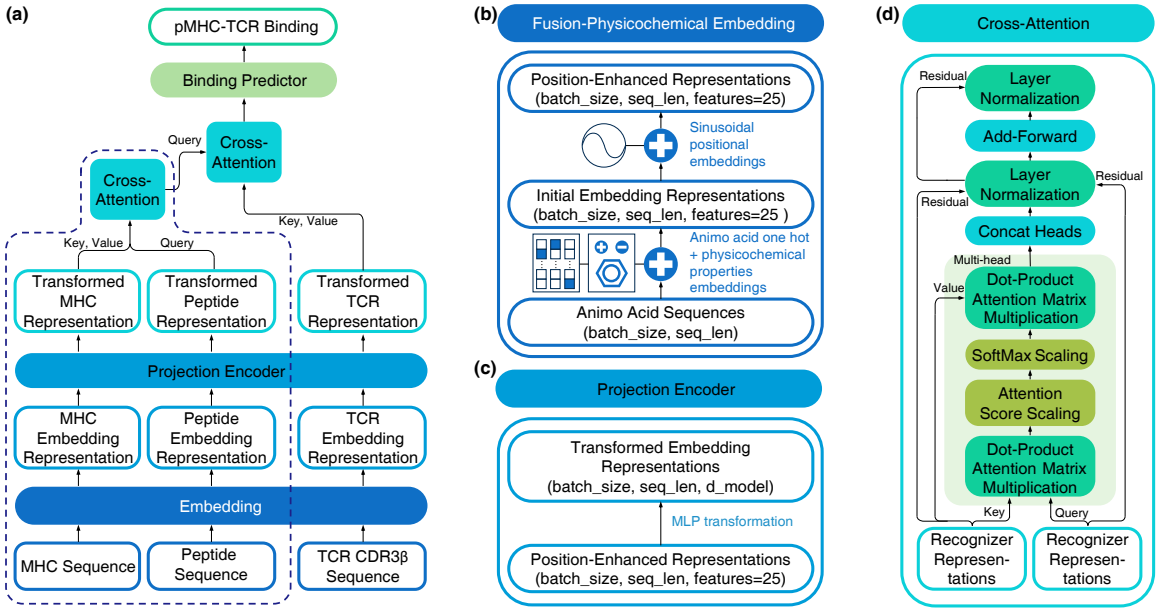


Figure 2: An overview of our model structure. The figure illustrates our complete model architecture for remodeling peptide-MHC-TCR triad binding as a representation learning and sequence fusion task. Notably in (a), the modules in the dashed box that involve peptide and MHC are first pre-trained on a peptide–MHC binding task (c.f. the details in Section 4.3), which provides a robust initialization prior to the core fine-tuning (which concerns all the modules in (a)) for pMHC–TCR binding predictions. Additionally, (b) illustrates the "Fusion-Physicochemical Embedding" module, (c) demonstrates the "Projection Encoder" module, and (d) depicts the "Cross-Attention" module. These components collaboratively enhance feature extraction and cross-sequence interactions.

For the reader's convenience, Table 4 in [2, Appendix B.1] provides a comprehensive overview of the key molecules involved in antigen presentation, detailing their cellular locations, structural properties, primary functions, theoretical diversity, and sequence homology.

3.3 Cross-Attention

Cross-attention [9, 27] gains prominence in various sequence interaction tasks, including text translation [21], image captioning [66], and voice recognition [54]. Its core advantage lies in enabling one sequence to selectively attend to relevant parts of another, thus enhancing interaction modeling [32, 33]. This characteristic happens to parallel the biological selectivity and specificity of immune responses, aiding in accurate prediction of binding affinities and antigen presentation [39].

Formally, a cross-attention module is given as:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\mathbf{Q}\mathbf{K}^{\mathsf{T}}/\sqrt{d}\right)\mathbf{V}, \tag{1}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the query, key, and value matrices derived from the input sequences with dimension d. Following Chen et al. [11], $\mathbf{K}$ and $\mathbf{V}$ typically **originate from the same sequence**, enabling the model to align context across biological sequences such as TCRs, MHCs, and antigens.

4 Remodeling Peptide-MHC-TCR Triad Binding as Sequence Fusion

To comprehensively understand and predict peptide-MHC-TCR interactions, accurate representation of protein sequences is indispensable. This section delineates our approach to capturing both the spatial relations (Section 4.1) and inherent characteristics of amino acids (Section 4.2) among distinct sequences. Through the new model proposed in Section 4.3, we aim to preserve the innate sequential characteristics of proteins, which are crucial for understanding their biological functions and interactions. Important implementation details are discussed in Section 4.4.

4.1 Representing Biological Sequences

In this subsection, we outline protein sequence representation methods used within our model, emphasizing the critical need to accurately capture both the amino acids and their positional information. In general, our approach preserves the intrinsic sequential nature of biological sequences throughout the encoding and transformation processes; the comparison with the traditional vector representation will be dissected below, and more details on position encoding and context-aware embeddings we use are provided in Appendix A.

Issues for a vector representation. The transformation of protein sequences into vector representations, adopted by [43] and other pioneering works, poses several challenges. One major issue is the **potential loss of sequential context and structural information**, which are critical for understanding protein functionality.

Traditional vectorization methods often flatten the sequence, treating it as a mere collection of features without considering the natural order and interactions between amino acids. This results in significant information loss, particularly in cases where the spatial arrangement and chemical properties of residues dictate their interactions and functions. A notable limitation of pMTnet [43] is its reliance on LSTM-based encoding, which, while effective for capturing local sequential dependencies, struggles to preserve long-range dependencies and structural interactions crucial for peptide-MHC binding prediction.

Importance of Sequence Form. The structural form of a protein sequence–its sequence of amino acids and their respective positions–plays a pivotal role in determining its biological function. Proper representation of these sequences is crucial for computational models to predict protein interactions.

Our method emphasizes the maintenance of the sequential integrity of protein sequences to ensure that both local and global structural characteristics are accurately represented, which is essential for predicting interactions. As **an empirical justification**, we verify the proposal of maintaining the sequence form for biological sequences, through ablation studies on TCR-Peptide binding in Section 5.3.

4.2 Unified Encoders for Heterogeneous Sequences

In a representation learning study, Chen et al. [12] proposed that similar representations can make effective use of the attention mechanism. Following this observation, we accordingly suggest each MHC and antigenic peptide sequence share the same encoder, so that the cross-attention mechanism we propose can be more effective in modeling sequence fusion.

In more details, for a sequence one-hot embedding matrix $\boldsymbol{X}$, no matter whether it indicates peptide, MHC, or TCR, **the linear transform matrix $\boldsymbol{W}$ will be identical** (as shown in Figure 2(a)). We note that this technique enforces the same encoding for different biological sequences, which aligns with the real binding process, considering that the low-level amino acids are identical in arbitrary biological sequences. In Section 5.6, we ablate the usage of the unified encoder against distinct encoders to empirically verify the design.

4.3 Complete Learning Mechanisms

We incorporate the representation learning techniques above and introduce the entire process of our proposed model Fusion-pMT, which further elaborates Figure 2. On the learning side, we follow a common "pre-training + fine-tuning" paradigm to handle the three input sequences; on the architecture side, we intentionally first fuse peptide and MHC and then turn to the fusion with TCR sequences, which not only resembles the biological process but also effectively utilize the abundant data for peptide-MHC interactions. Notably, the pre-trained model in Stage 1 below can be directly applied to the peptide–MHC binding task; we denote it as Fusion-pM and will evaluate its performance in Section 5.3, to further justify the design in Stage 1.

Stage 1: Pre-training via peptide–MHC binding. As depicted in Figure 2(a), we first solely train partial modules in the dashed box that involve the peptide and MHC sequences with a peptide–MHC binding prediction task. Specifically, we take the pre-training data (the peptide and the MHC sequences along with their binding labels) from Chu et al. [15] and feed the input sequences into the (partial) model. The model transforms peptide and MHC sequences (in one-hot encoding) into high-dimensional embeddings using linear transformations $\boldsymbol{W}$, and further incorporate the structural information via sinusoidal positional encodings (c.f. Appendix A.2), which help preserve sequence integrity and temporal dynamics. The projection encoder (illustrated in Figure 2(c)) further lifts the sequence dimension to d (so that align the dimensionality of different sequences) through an MLP module and provides extra flexibility. Ultimately, we obtain the query matrix $\boldsymbol{Q}_p$ (from the peptide sequence $\boldsymbol{X}_p$), and the key/value matrix $\boldsymbol{K}_M, \boldsymbol{V}_M$ (from the MHC sequence $\boldsymbol{X}_M$). A **cross-attention** module then takes those three matrices as inputs and dynamically integrates both peptide and MHC sequences through

$$\mathrm{Attn}(\boldsymbol{Q}_p, \boldsymbol{K}_M, \boldsymbol{V}_M) = \mathrm{softmax}\Big(\boldsymbol{Q}_p \boldsymbol{K}_M^\top / \sqrt{d}\Big)\boldsymbol{V}_M,$$

The sequence matrix $\mathrm{Attn}(\boldsymbol{Q}_p, \boldsymbol{K}_M, \boldsymbol{V}_M)$ (which shares the same shape as the peptide query matrix $\boldsymbol{Q}_p$) output by the last cross-attention layer is then refined through normalization and feed-forward layers and undergoes a mean pooling; ultimately, the resulting vector is fed into a classifier along with the cross-entropy loss for training.

Remark. This design effectively utilizes the objectively available peptide–MHC binding data, which is more abundant than the peptide-MHC-TCR triad binding data; this practice is adopted by Lu et al. [43] as well. In Section 5.3, we observe our sequence modeling effectively mines the information behind the peptide–MHC binding data, and Fusion-pM, the side model produced in Stage 1, attains high prediction performance (see Figure 3).

Stage 2: full parameter fine-tuning for peptide-MHC-TCR binding. We then start to train the whole model. Here is how we handle the three sequences: As shown in Figure 2(a), we pass peptide and MHC sequences to the pre-trained model above in Stage 1, wherein the aforementioned mean-pooling module and the classifier are removed; the sequence matrix output by the last cross-attention layer in the pre-trained peptide–MHC part, this time will be transformed into a query matrix $\boldsymbol{Q}_{\mathrm{pM}}$ and interact with the TCR key/value matrix $\boldsymbol{K}_T, \boldsymbol{V}_T$ in a cross-attention module, which returns

$$\mathrm{Attn}(\boldsymbol{Q}_{\mathrm{pM}}, \boldsymbol{K}_T, \boldsymbol{V}_T) = \mathrm{softmax}\Big(\boldsymbol{Q}_{\mathrm{pM}} \boldsymbol{K}_T^\top / \sqrt{d}\Big)\boldsymbol{V}_T,$$

Remark. This model design notably reflects the representation learning techniques we mentioned before. In particular, we maintain the sequence form for both TCR and the product of the peptide–MHC interaction, as discussed in Section 4.1; moreover, we apply the unified encoder to all of the three sequences, as per Section 4.2.

The two-stage paradigm allows for a realistic interaction modeling between the peptide, MHC, and TCR sequences. Initially, the peptide and MHC representations interact to produce an intermediate sequence matrix, which is then used to interact with the TCR representation, capturing the complex dependencies between these biological sequences.

4.4 Implementation details

For the architecture of Fusion-pMT, we set the embedding dimension to 64, use 4 attention heads. Here, we would like to further discuss

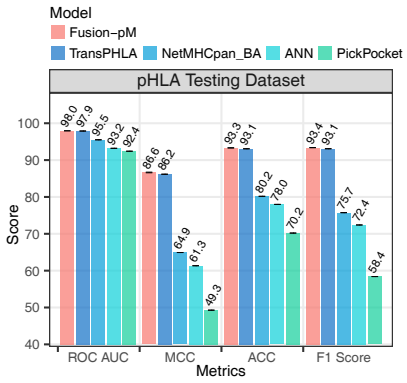


Figure 3: Comparison of peptide–MHC binding models based on multiple evaluation metrics (%), including ROC AUC, MCC, ACC, and F1 Score. The models compared are Fusion-pM, TransPHLA, NetMHCpan_BA, ANN, and PickPocket, evaluated on "pHLA Testing Dataset", the testing dataset of TransPHLA [15].

the practical issues and the related implementation details crucial to the model performance, as a closing remark to this methodology section.

Gradient Vanishing. To mitigate the issue of gradient vanishing, we employ the LeakyReLU activation function [29] in both the intermediate layers and the feedforward layers. Additionally, we implemented residual connections that bypass the attention mechanism by directly connecting the encoded sequence information to the fully connected layers, which reduces the risk of gradient vanishing.

5 Experiment Results

This section includes three main parts. We first describe the experimental setup in Section 5.1, followed by the baseline models and evaluation benchmarks in Section 5.2. Then, we present the results on two core tasks: Peptide–MHC Binding and Peptide–MHC–TCR Binding (respectively in Sections 5.3 and 5.4), which involve out-of-distribution (OOD) evaluation (detailed in Section 5.1) to assess generalization. Finally, we conduct ablation studies in Sections 5.5 and 5.6 to examine the contributions of key model components, including sequence representation design and unified encoder design.

5.1 Experiment Setups

The experiments along this section are mainly conducted to examine the performance of two important variants featured with our proposed techniques: **Fusion-pM** (the model for predicting **peptide–MHC binding**) and **Fusion-pMT** (the model for predicting tripartite **peptide-MHC-TCR binding**). Overall, both of the models were implemented in PyTorch and conducted on an NVIDIA 3090 GPU cluster with 8 GPUs. In training, we used a batch size of 64 and a learning rate of 0.01, an Adam optimizer, and trained the models for 300 epochs.

Peptide–MHC binding. For the model training and testing, we align our testing protocols and datasets with the ones from the previous work TransPHLA [15]. Specifically, TransPHLA organizes the data into multiple partitions: Training, Validation, Testing (called `Independent` in [15]), where the testing set is used to test the model's generalization to unseen alleles, and designated as **pHLA Testing Dataset**.

Peptide–MHC–TCR binding. For model training, validation and testing, we used a dataset derived from Lu et al. [43], referred to as the **pMT training**, **pMT validation** and **pMT testing**, with a positive-to-negative sample ratio of 1:10. For details, we closely follow the experimental setup used in pMTnet [43] and pMTnet-omni [23]. We adhere to the original preprocessing protocols, including sequence amendments and negative pair generation, to ensure consistency and comparability with prior work, i.e., the positive samples were further augmented tenfold, resulting in an effective 1:1 ratio between positive and negative samples during training. Specifically, this dataset includes 28,604 unique TCR CDR3 sequences, 426 peptides, and 63 HLA types (MHC alleles). Aligning with pMTnet [43], we utilized the same **pMT testing dataset** from the GitHub of Lu et al. [43], which contains 272 TCR sequences, 224 peptide sequences, and 24 MHC sequences, selected to ensure that the peptides are entirely unseen in the training or validation data. Overall, the dataset contains 32,607 pMHC–TCR positive pairs and a significantly larger set of generated negative pairs. Similarly to Lu et al. [43], we then partitioned the training data into a training set and a validation set using an 80:20 split to support model selection and hyperparameter tuning.

To further evaluate the model's out-of-distribution (OOD) generalization, we curated an additional **OOD testing dataset** using samples collected from VDJdb [22], which can be regarded as arising from a distinct sampling process compared to the original datasets [43]. This OOD set consists of 1,346 TCR sequences, 239 peptide sequences, and 53 MHC sequences, and was designed to challenge the model's robustness when applied to data distributions not seen during training.

In both test datasets, negative samples were randomly generated (the negative samples in **pMT testing dataset** were copied down from the official GitHub of Lu et al. [43]), and all positive samples previously included in the training or validation sets were explicitly excluded. This results in an imbalanced testing setup, maintaining a roughly 1:10 positive-to-negative ratio, which reflects real-world biological conditions [43].

5.2 Baselines and Benchmarks

Despite the growing interest in computational immunology, the field still lacks standardized datasets and benchmark protocols for evaluating interactions among T cell receptors (TCRs), major histocompatibility complex (MHC) molecules, and antigenic peptides. This absence of widely accepted benchmarks hinders robust comparison, reproducibility, and scientific progress. Establishing reliable baselines and benchmarks is thus crucial for advancing the predictive capabilities and practical impact of computational models in immunology.

For the peptide–MHC (pMHC) binding task, we compare against several strong baselines, including TransPHLA [15], netMHCpan [4], ANN [14], PickPocket [65], CcBHLA [61], and STMHCpan [64]. We evaluate model performance using multiple standard metrics,

Table 2: Performance (%) Metrics of Fusion-pMT Model: Mean and Standard Deviation of ROC AUC, ACC, MCC, F1 Score, Sensitivity, Specificity, Precision, Recall, and PR AUC values on pMT validation, training, Testing and OOD Datasets.

Datasets	Metrics Score (mean ± std)								
	ROC AUC	ACC	MCC	F1 Score	Sensitivity	Specificity	Precision	Recall	PR AUC
pMT Validation	88.36 ± 0.25	79.04 ± 0.35	59.48 ± 0.84	80.96 ± 0.43	89.39 ± 2.43	68.75 ± 2.68	74.04 ± 1.16	89.39 ± 2.43	87.25 ± 0.19
pMT Training	92.70 ± 0.23	83.53 ± 0.41	68.01 ± 0.61	84.73 ± 0.27	91.43 ± 2.25	75.64 ± 2.86	79.02 ± 1.56	91.43 ± 2.25	92.40 ± 0.20
pMT Testing	88.63 ± 0.27	79.38 ± 0.34	60.01 ± 0.82	81.35 ± 0.43	89.45 ± 2.51	69.18 ± 2.79	74.66 ± 1.21	89.45 ± 2.51	87.85 ± 0.21
OOD Testing	88.35 ± 0.21	78.93 ± 0.18	59.19 ± 0.55	80.86 ± 0.38	89.05 ± 2.75	68.81 ± 3.04	74.13 ± 1.28	89.05 ± 2.75	87.40 ± 0.15

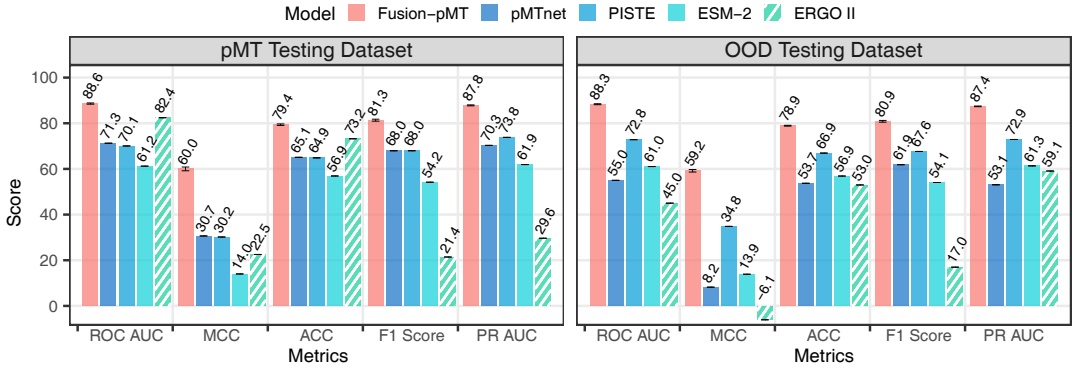


Figure 4: Performance (%) of the five tri-molecular binding prediction models on pMT and OOD testing datasets in terms of ROC AUC, ACC and MCC. Here, ERGO II was developed and as well evaluated solely for peptide-TCR binding prediction, serving as a sanity check that incorporating more molecules is beneficial; other methods were tested on peptide-MHC-TCR triad binding prediction. "pMT Testing Dataset" is the original testing dataset in Lu et al. [43]. "OOD Testing Dataset" is a new dataset collected from Goncharov et al. [22]. Notably, the calculation of the metric "MCC" hinges on correlation, which thus can be negative.

including Accuracy (ACC), F1 Score, Matthews Correlation Coefficient (MCC), and Receiver Operating Characteristic Area Under the Curve (ROC AUC).

For the peptide–MHC–TCR binding task, we closely follow the experimental setups defined in pMTnet [43] and pMTnet-omni [23], using the same datasets, preprocessing protocols, and data splits as described in Section 5.1. In this context, we benchmark our models against four state-of-the-art baselines specifically designed for peptide–MHC–TCR binding prediction or general protein interaction modeling: pMTnet [43], PISTE [19], ESM-2 [41], and ERGO II [46]. These baselines reflect diverse modeling approaches, including LSTM-based architectures, sliding and cross attention mechanisms, and large-scale pretrained protein language models. We evaluate model performance using multiple standard metrics, including Accuracy (ACC), F1 Score, Matthews Correlation Coefficient (MCC), Receiver Operating Characteristic Area Under the Curve (ROC AUC), and Precision–Recall Area Under the Curve (PR AUC).

In addition, we provide the complete evaluation results of our Fusion-pMT on the peptide–MHC–TCR binding task across four distinct sets—pMT training, validation, testing, and OOD testing sets—not only covering the five core metrics mentioned above but also including sensitivity, specificity, precision, and recall to offer a more comprehensive assessment of our model performance.

5.3 Immune Presentation Prediction (Peptide–MHC Binding)

Our proposed peptide–MHC binding model, **Fusion-pM**, demonstrates superior performance compared to existed models, as depicted in Figure 3. Our model achieves a marginally higher accuracy (ACC) of 93.30% compared to TransPHLA's 93.08% and an improved F1 score of 93.40% over TransPHLA's 91.40%. More notably, our model outperforms TransPHLA in Matthews Correlation Coefficient (MCC), achieving 86.60 versus 86.20, and in ROC AUC, scoring 98.00 compared to 97.90. These results highlight the robustness and generalization capability of our approach.

Fusion-pM integrates two key architectural innovations to enhance both performance and efficiency: a *Unified Encoder* for peptides and MHCs, and a *Cross-Attention Sequence Fusion Block*. The unified encoder enables consistent and structured feature extraction across heterogeneous sequence inputs, while the cross-attention mechanism explicitly models dynamic interactions between peptide and MHC sequences, enriching the contextual representation of binding interactions. Together, these components enable a parameter-sharing design that results in a remarkably compact model with fewer than 700k total parameters, significantly reducing computational overhead. Unlike prior models such as TransPHLA [15], which rely on multi-layer self-attention, Fusion-pM achieves comparable or superior predictive accuracy with a much smaller parameter footprint. Importantly, this lightweight design enables rapid

Table 3: Statistical Comparison of Fusion-pMT Models: T-value and P-value for ACC, ROC AUC, and PR AUC, on Peptide-MHC-TCR binding task. ^{N} indicates the model is equipped with the fusion technique from netMHCpan [35].

Metrics	Fusion-pMT^{N} vs. pMTnet		Fusion-pMT vs Fusion-pMT^{N}	
	t-value	p-value	t-value	p-value
PR AUC	55.00	0.000330	57.74	0.000001
ROC AUC	25.69	0.001512	10.32	0.008406
ACC	42.21	0.000561	85.98	0.000001

inference—achieving predictions on CPU within seconds—making Fusion-pM not only accurate but also highly practical for integration into biological research pipelines and real-world applications.

5.4 Immunogenicity Prediction (Peptide-MHC-TCR Binding)

pMT testing dataset. Figure 4 presents the performance of the five Peptide-MHC-TCR triad binding prediction models on pMT testing datasets. Specifically, Fusion-pMT attains a ROC AUC of 88.63%, representing a notable improvement over pMTnet (71.28%), and achieves a PR AUC of 87.85%, significantly higher than pMTnet's 70.30%. These improvements reflect the model's enhanced ability to capture the complex binding relationships among peptides, MHC molecules, and TCR sequences. Importantly, Fusion-pMT achieves a Matthews Correlation Coefficient (MCC) of 60.0%, doubling the MCC achieved by pMTnet (30.7%). Since MCC accounts for true positives, true negatives, false positives, and false negatives in a balanced manner, this result indicates that our model not only performs well in terms of accuracy but also truly learns meaningful immunogenicity patterns, avoiding overfitting or biased predictions. A high MCC suggests that the model generalizes better across both positive and negative classes, reflecting a deeper understanding of the underlying biological interactions.

Although ERGO II achieves relatively high ROC AUC (82.4%) and ACC (73.2%) on the pMT testing dataset, its performance on MCC (14.2%), F1 Score (24.1%), and PR AUC (26.6%) is substantially worse. This discrepancy suggests that ERGO II's predictions are heavily biased toward the negative class, exploiting the dataset's natural 90% negative-to-positive imbalance. As a result, the model achieves deceptively high ACC and ROC AUC by predominantly ranking and predicting negative samples, while failing to meaningfully capture positive binding interactions. In contrast, Fusion-pMT achieves balanced performance across all metrics, demonstrating its superior capacity to learn biologically relevant binding patterns.

To further guard against such misleading effects and provide a comprehensive assessment, we additionally report the sensitivity, specificity, precision, and recall of Fusion-pMT in Table 2. Fusion-pMT achieves a sensitivity of 89.45%, indicating its strong ability to correctly identify true binding cases (true positive rate). Its specificity reaches 69.18%, reflecting its capability to correctly reject non-binding cases (true negative rate). Together, these metrics offer a nuanced picture of the model's performance across different aspects of immunological prediction, confirming that Fusion-pMT's robust and consistent performance.

Out-of-distribution testing dataset. Figure 4 presents the performance of the five Peptide-MHC-TCR triad binding prediction models on OOD testing datasets. As shown, the proposed Fusion-pMT consistently outperforms pMTnet across all metrics. For example, on the pMT testing dataset, Fusion-pMT achieves a mean ROC AUC of 88.36%, surpassing pMTnet's 78.91%. In terms of accuracy (ACC), Fusion-pMT achieves an accuracy of 78.9%, while pMTnet had 53.7, a value close to random guessing in a binary classification setting. This observation is further supported by pMTnet's Matthews Correlation Coefficient (MCC) of just 8.2, indicating that the model fails to capture meaningful predictive patterns and largely operates at a chance level. In contrast, Fusion-pMT achieves similar results on both the out-of-distribution (OOD) testing set (i.e., from Goncharov et al. [22]) and the in-distribution pMT testing set (i.e., from Lu et al. [43]), demonstrating that our model has truly learned biologically meaningful features essential for immunogenicity prediction, rather than merely memorizing training data. Remarkably, these strong generalization capabilities are achieved with a highly compact architecture containing fewer than 700k parameters, underscoring the efficiency and practicality of our design compared to larger, less efficient baseline models [43]. This performance margin is further supported by the statistical comparison shown in Table 3, where Fusion-pMT significantly outperforms pMTnet across all key metrics (ACC, ROC AUC, and PR AUC), with t-values exceeding 25 and p-values consistently below 0.002. These results indicate that the observed improvements are not merely numerical fluctuations but reflect statistically significant gains.

5.5 Ablation studies: sequence representation

Compared with pMTnet using a bottleneck autoencoder model to encode the TCR sequence, we propose a cross-attention-based transformer, Fusion-pMT (netMHCpan), to represent the TCR sequence and preserve the sequence form until the binding prediction block (thus the pre-trained module netMHCpan is kept, as in pMTnet, for fair comparison). The model architecture is in Section 4. Fusion-pMT (netMHCpan) shows significant improvements in performance metrics over pMTnet, as evidenced from both numerical comparisons (e.g., ACC, PR AUC, and ROC AUC in Figure 5) and statistical significance tests (e.g., t-tests in Table 3). These considerable improvements demonstrate the critical importance of maintaining sequence integrity in our model, which enables more effective capturing of complex, sequence-dependent interactions crucial for accurate binding predictions.

5.6 Ablation studies: unified encoders

To ablate the usage of unified encoders, we introduce a model variant, Fusion-pM (w/o unified), which employs *distinct* encoders for peptides and MHC sequences and similarly incorporates a cross-attention sequence fusion block. This specification allows a fair comparison with the base Fusion-pM (w/ unified) variant, which we recall employs a unified encoder for both peptides and MHCs.

As shown in Figure 5b, the results demonstrate that Fusion-pM (w/ unified) statistically significant improvements across all metrics. The ACC scores exhibit minimal (though significant) variation between the two model variants; however, both F1 and MCC metrics indicate more substantial gains with the Fusion-pM (w/ unified)

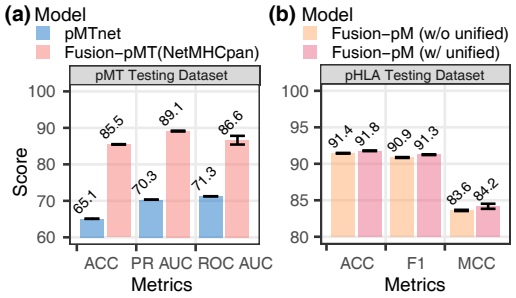


Figure 5: Ablation studies for (a) Peptide-MHC-TCR Traid binding Prediction with Sequence Representaion and (b) peptide–MHC Models with uified encoders. Notably, "pMT / pHLA Testing Dataset" is the original testing dataset in Lu et al. [43] / Chu et al. [15].

configuration, suggesting that employing the unified encoder for sequence fusion not only simplifies the model architecture but may also enhance performance in terms of both prediction precision and class balance handling [8]. These results prompt further investigation into the benefits of encoder uniformity in complex sequence fusion tasks in immunological prediction.

6 Conclusions

In this paper, we have proposed a new model Fusion-pMT for developing the biological language modeling specifically in peptide-MHC-TCR triad binding, through a revisit of sequence fusion mechanisms in representation learning research. In particular, we notice that maintaining the sequence form for each input along the transform aligns with the real biological processes and can significantly improve the performance in immunogenicity prediction. With this insight, we propose to characterize how different sequences interact with each other through both the cross-attention mechanism and the unified embedding vocabulary of amino acids. We evaluate our Fusion-pMT on various real-world datasets and show that it consistently exhibits higher performance than baselines. Furthermore, with fewer than 700k parameters, Fusion-pMT offers a lightweight architecture that can efficiently run on CPUs, making it accessible for practical deployment in biological research without the need for specialized hardware. Overall, we present a new way for understanding peptide-MHC-TCR triad binding with insights from the language modeling techniques, providing beneficial reference for future models aiming to integrate more complex and heterogeneous biological languages.

A Appendix for Protein Sequence Embedding

In this appendix, we introduce the tedious notations for the concrete protein sequence embedding in our models. Our approach is designed to preserve the intrinsic sequential integrity of these sequences throughout the encoding and transformation processes; these details are vital for understanding the complex interactions within biological sequences.

A.1 Protein Sequences

We start with the embedding of protein sequences in this subsection, and then introduce the positional encoding method we adopt in Appendix A.2.

A protein sequence S is composed of L amino acids, which can be mathematically represented as

$$S = [a_1, a_2, a_3, \ldots, a_L] .$$

Here, each a_i indicates a certain amino acid, and the collection of the 21 standard amino acids is denoted as

$$\mathcal{A} := \{\mathsf{A}, \mathsf{C}, \mathsf{D}, \mathsf{E}, \mathsf{F}, \mathsf{G}, \mathsf{H}, \mathsf{I}, \mathsf{K}, \mathsf{L}, \mathsf{M}, \mathsf{N}, \mathsf{P}, \mathsf{Q}, \mathsf{R}, \mathsf{S}, \mathsf{T}, \mathsf{V}, \mathsf{W}, \mathsf{X}, \mathsf{Y}\}$$

in terms of letter abbreviations of amino acids.

One-hot encoding. Following Devlin et al. [17], we first transform a protein sequence into a binary vector representation, which is a common practice in the representation of textual sequences. Here, each amino acid a_i corresponds to a one-hot encoded vector $\boldsymbol{h}(a_i)$ of length $|\mathcal{A}|$:

$$\boldsymbol{h}(a_i) = [h_1, h_2, \ldots, h_{|\mathcal{A}|}]^T,$$

where $h_j = 1$ if a_i is the j-th amino acid in $\mathcal{A}$, and $h_j = 0$ otherwise. The entire protein sequence S is thus

$$\mathbf{H}^T(S) = [\mathbf{h}(a_1), \mathbf{h}(a_2), \ldots, \mathbf{h}(a_L)],$$

where the sequence matrix $\mathbf{H}(S) \in \{0, 1\}^{L \times |\mathcal{A}|}$.

Encoding. Notably, each amino acid a_i can be encoded using another specific numerical representation that captures its chemical properties and contributes to its role within the protein structure. This encoding might utilize techniques ranging from simple categorical encoding schemes to more complex embeddings derived from machine learning models.

Transformation processes. The encoded representations are processed through computational models (e.g., convolutional neural networks or recurrent neural networks) designed to capture the interactions between amino acids and to preserve their positional information.

Aggregation. The transformed representations are ultimately aggregated to form a comprehensive vector representation of the entire sequence. This step may involve methods like pooling.

Overall, these steps, which are well-studied in representation learning literature, ensure that our model not only captures the individual characteristics of each amino acid but also their contextual relationships within the entire sequence.

A.2 Positional Encoding for Protein Sequences

The positional encoding provides the model with information about the relative or absolute position of the tokens in the sequence. Let s denote the position within the biological sequence, i be the dimension within the embedding spaces, and d indicate the dimensionality of the model embeddings. One can then apply the sine and cosine functions for positional encoding as follows Vaswani et al. [56]:

$$p(s, 2i) = \sin\left(\frac{s}{10000^{2i/d}}\right),$$

$$p(s, 2i+1) = \cos\left(\frac{s}{10000^{2i/d}}\right),$$

The encoding mechanism ensures that the model can effectively interpret the sequential order of the sequences for biological interactions, and we apply this positional encoding in computing the cross-attention modules.

GenAI Statement

Statement on Generative Artificial Intelligence (GenAI) Use In this research, the authors utilized ChatGPT-4o (OpenAI, 2024) for language polishing and stylistic refinement to enhance the clarity and professionalism of the manuscript. The tool was employed exclusively for auxiliary tasks at the linguistic level.

References

[1] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charle Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630, 8016 (2024), 493–500. doi:10.1038/s41586-024-07487-w

[2] Anonymous. 2025. Fusion-pMT: Code Repository (Anonymous). *Anonymous 4open* (2025). Available at https://anonymous.4open.science/r/Fusion-pMT-CIKM, expires on 2025-11-20.

[3] Christina M. Arieta, Yushu Joy Xie, Daniel A. Rothenberg, Huitian Diao, Dewi Harjanto, Shirisha Meda, Krisann Marquart, Byron Koenitzer, Tracey E. Sciuto, Alexander Lobo, Adam Zuiani, Stefanie A. Krumm, Carla Iris Cadima Couto, Stephanie Hein, André P. Heinen, Thomas Ziegenhals, Yunpeng Liu-Lupo, Annette B. Vogel, John R. Srouji, Stephanie Fesser, Kaushik Thanki, Kerstin Walzer, Theresa A. Addona, Özlem Türeci, Uğur Şahin, Richard B. Gaynor, and Asaf Poran. 2023. The T-cell-directed vaccine BNT162b4 encoding conserved non-spike antigens protects animals from severe SARS-CoV-2 infection. *Cell* 186, 11 (2023), 2392–2409.e21. doi:10.1016/j.cell.2023.04.007

[4] Piyush Borole and Ajitha Rajan. 2024. Building trust in deep learning-based immune response predictors with interpretable explanations. *Communications Biology* 7, 1 (2024), 279. doi:10.1038/s42003-024-05968-2

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

[6] Patrick Bryant, Gabriele Pozzati, Wensi Zhu, Aditi Shenoy, Petras Kundrotas, and Arne Elofsson. 2022. Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. *Nature communications* 13, 1 (2022), 6028.

[7] Yue Cao, Payel Das, Vijil Chenthamarakshan, Pin-Yu Chen, Igor Melnyk, and Yang Shen. 2021. Fold2Seq: A Joint Sequence(1D)-Fold(3D) Embedding-based Generative Model for Protein Design. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 1261–1271. https://proceedings.mlr.press/v139/cao21a.html

[8] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*. 169–174.

[9] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 357–366.

[10] Hedi Chen, Xiaoyu Fan, Shuqian Zhu, Yuchan Pei, Xiaochun Zhang, Xiaonan Zhang, Lihang Liu, Feng Qian, and Boxue Tian. 2024. Accurate prediction of CDR-H3 loop structures of antibodies with deep learning. *Elife* 12 (2024), RP91512.

[11] Yifan Chen, Devamanyu Hazarika, Mahdi Namazifar, Yang Liu, Di Jin, and Dilek Hakkani-Tur. 2022. Empowering parameter-efficient transfer learning by recognizing the kernel structure in self-attention. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 1375–1388. doi:10.18653/v1/2022.findings-naacl.102

[12] Yifan Chen, Devamanyu Hazarika, Mahdi Namazifar, Yang Liu, Di Jin, and Dilek Hakkani-Tur. 2022. Inducer-tuning: Connecting Prefix-tuning and Adapter-tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 793–808. doi:10.18653/v1/2022.emnlp-main.50

[13] Hongbo Chi, Marion Pepper, and Paul G. Thomas. 2024. Principles and therapeutic applications of adaptive immunity. *Cell* 187, 9 (2024), 2052–2078. doi:10.1016/j.cell.2024.03.037

[14] S-E Choi, C-W Park, Y-H Sohn, S-Y Ko, H-B Oh, G-H Kim, and H Kim. 2011. Artificial neural network weights of residues for the serological specificities of HLA. *International Journal of Immunogenetics* 38, 3 (2011), 269–275.

[15] Yanyi Chu, Yan Zhang, Qiankun Wang, Lingfeng Zhang, Xuhong Wang, Yanjing Wang, Dennis Russell Salahub, Qin Xu, Jianmin Wang, Xue Jiang, Yi Xiong, and Dong-qing Wei. 2022. A transformer-based model to predict peptide–HLA class I binding and optimize mutated peptides for vaccine design. *Nature Machine Intelligence* 4, 3 (2022), 300–311. doi:10.1038/s42256-022-00459-7

[16] Yanyi Chu, Yan Zhang, Qiankun Wang, Lingfeng Zhang, Xuhong Wang, Yanjing Wang, Jianmin Wang, Xue Jiang, Dennis Salahub, Yi Xiong, and Dong-qing Wei. 2021. TransMut: a program to predict HLA-I peptide binding and optimize mutated peptides for vaccine design by the Transformer-derived self-attention model. *Research Square* (2021). doi:10.21203/rs.3.rs-785618/v1

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423

[18] Sandra P D'Angelo, Dejka M Araujo, Albiruni R Abdul Razak, Mark Agulnik, Steven Attia, Jean-Yves Blay, Irene Carrasco Garcia, John A Charlson, Edwin Choy, George D Demetri, Mihaela Druta, Edouard Forcade, Kristen N Ganjoo, John Glod, Vicki L Keedy, Axel Le Cesne, David A Liebner, Victor Moreno, Seth M Pollack, Scott M Schuetze, Gary K Schwartz, Sandra J Strauss, William D Tap, Fiona Thistlethwaite, Claudia Maria Valverde Morales, Michael J Wagner, Breelyn A Wilky, Cheryl McAlpine, Laura Hudson, Jean-Marc Navenot, Tianjiao Wang, Jane Bai, Stavros Rafail, Ruoxi Wang, Amy Sun, Lilliam Fernandes, Erin Van Winkle, Erica Elefant, Colin Lunt, Elliot Norry, Dennis Williams, Swethajit Biswas, and Brian A Van Tine. 2024. Afamitresgene autoleucel for advanced synovial sarcoma and myxoid round cell liposarcoma (SPEARHEAD-1): an international, open-label, phase 2 trial. *The Lancet* 403, 10435 (2024), 1460–1471. doi:10.1016/S0140-6736(24)00319-2

[19] Ziyan Feng, Jingyang Chen, Youlong Hai, Xuelian Pang, Kun Zheng, Chenglong Xie, Xiujuan Zhang, Shengqing Li, Chengjuan Zhang, Kangdong Liu, et al. 2024. Sliding-attention transformer neural architecture for predicting T cell receptor–antigen–human leucocyte antigen binding. *Nature Machine Intelligence* 6, 10 (2024), 1216–1230.

[20] Yicheng Gao, Kejing Dong, Yuli Gao, Xuan Jin, Jingya Yang, Gang Yan, and Qi Liu. 2024. Unified cross-modality integration and analysis of T cell receptors and T cell transcriptomes by low-resource-aware representation learning. *Cell Genomics* (2024). doi:10.1016/j.xgen.2024.100553

[21] Mozhdeh Gheini, Xiang Ren, and Jonathan May. 2021. Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1754–1765. doi:10.18653/v1/2021.emnlp-main.132

[22] Mikhail Goncharov, Dmitry Bagaev, Dmitrii Shcherbinin, Ivan Zvyagin, Dmitry Bolotin, Paul G. Thomas, Anastasia A. Minervina, Mikhail V. Pogorelyy, Kristin Ladell, James E. McLaren, David A. Price, Thi H. O. Nguyen, Louise C. Rowntree, E. Bridie Clemens, Katherine Kedzierska, Garry Dolton, Cristina Rafael Rius, Andrew Sewell, Jerome Samir, Fabio Luciani, Ksenia V. Zornikova, Alexandra A. Khmelevskaya, Saveliy A. Sheetikov, Grigory A. Efimov, Dmitry Chudakov, and Mikhail Shugay. 2022. VDJdb in the pandemic era: a compendium of T cell receptors specific for SARS-CoV-2. *Nature Methods* 19, 9 (2022), 1017–1019. doi:10.1038/s41592-022-01578-0

[23] Yi Han, Yuqiu Yang, Yanhua Tian, Farjana J Fattah, Mitchell S von Itzstein, Yifei Hu, Minying Zhang, Xiongbin Kang, Donghan M Yang, Jialiang Liu, Yaming Xue, Chaoying Liang, Indu Raman, Chengsong Zhu, Olivia Xiao, Jonathan E Dowell, Jade Homsi, Sawsan Rashdan, Shengjie Yang, Mary E Gwin, David Hsiehchen, Yvonne Gloria-McCutchen, Ke Pan, Fangjiang Wu, Don Gibbons, Xinlei Wang, Cassian Yee, Junzhou Huang, Alexandre Reuben, Chao Cheng, Jianjun Zhang, David E Gerber, and Tao Wang. 2023. pan-MHC and cross-Species Prediction of T Cell Receptor-Antigen Binding. *bioRxiv* (2023). doi:10.1101/2023.12.01.569599

[24] Jessica C Hassel, Sophie Piperno-Neumann, Piotr Rutkowski, Jean-Francois Baurain, Max Schlaak, Marcus O Butler, Ryan J Sullivan, Reinhard Dummer, John M Kirkwood, Marlana Orloff, Joseph J Sacco, Sebastian Ochsenreither, Anthony M Joshua, Lauris Gastaud, Brendan Curti, Josep M Piulats, April K S Salama, Alexander N Shoushtari, Lev Demidov, Mohammed Milhem, Bartosz Chmielowski, Kevin B Kim, Richard D Carvajal, Omid Hamid, Laura Collins, Koustubh Ranade, Chris Holland, Constance Pfeiffer, and Paul Nathan. 2023. Three-Year overall survival with Tebentafusp in metastatic uveal melanoma. *New England Journal of Medicine* 389, 24 (2023), 2256–2266. doi:10.1056/NEJMoa2304753

[25] Jonas S. Heitmann, Tatjana Bilich, Claudia Tandler, Annika Nelde, Yacine Maringer, Maddalena Marconato, Julia Reusch, Simon Jäger, Monika Denk, Marion Richter, Leonard Anton, Lisa Marie Weber, Malte Roerden, Jens Bauer,

Jonas Rieth, Marcel Wacker, Sebastian Hörber, Andreas Peter, Christoph Meisner, Imma Fischer, Markus W. Löffler, Julia Karbach, Elke Jäger, Reinhild Klein, Hans-Georg Rammensee, Helmut R. Salih, and Juliane S. Walz. 2022. A COVID-19 peptide vaccine for the induction of SARS-CoV-2 T cell immunity. *Nature* 601, 7894 (2022), 617–622. doi:10.1038/s41586-021-04232-5

[26] Jonas S. Heitmann, Claudia Tandler, Maddalena Marconato, Annika Nelde, Timorshah Habibzada, Susanne M. Rittig, Christian M. Tegeler, Yacine Maringer, Simon U. Jaeger, Monika Denk, Marion Richter, Melek T. Oezbek, Karl-Heinz Wiesmüller, Jens Bauer, Jonas Rieth, Marcel Wacker, Sarah M. Schroeder, Naomi Hoenisch Gravel, Jonas Scheid, Melanie Märklin, Annika Henrich, Boris Klimovich, Kim L. Clar, Martina Lutz, Samuel Holzmayer, Sebastian Hörber, Andreas Peter, Christoph Meisner, Imma Fischer, Markus W. Löffler, Caroline Anna Peuker, Stefan Habringer, Thorsten O. Goetze, Elke Jäger, Hans-Georg Rammensee, Helmut R. Salih, and Juliane S. Walz. 2023. Phase I/II trial of a peptide-based COVID-19 T-cell activator in patients with B-cell deficiency. *Nature Communications* 14, 1 (2023), 5032. doi:10.1038/s41467-023-40758-0

[27] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. 2019. Cross attention network for few-shot classification. In *Advances in Neural Information Processing Systems*, Vol. 32.

[28] Johannes B Huppa, Markus Axmann, Manuel A Mörtelmaier, Björn F Lillemeier, Evan W Newell, Mario Brameshuber, Lawrence O Klein, Gerhard J Schütz, and Mark M Davis. 2010. TCR–peptide–MHC interactions in situ show accelerated kinetics and increased affinity. *Nature* 463, 7283 (2010), 963–967. doi:10.1038/nature08746

[29] Kanchan Jha, Sriparna Saha, and Hiteshi Singh. 2022. Prediction of protein–protein interaction using graph neural networks. *Scientific Reports* 12, 1 (2022), 8360.

[30] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. 2021. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* 37, 15 (2021), 2112–2120. doi:10.1093/bioinformatics/btab083

[31] Jing Jin, Zhonghao Liu, Alireza Nasiri, Yuxin Cui, Stephen-Yves Louis, Ansi Zhang, Yong Zhao, and Jianjun Hu. 2021. Deep learning pan-specific model for interpretable MHC-I peptide binding prediction with improved attention mechanism. *Proteins: Structure, Function, and Bioinformatics* 89, 7 (2021), 866–883.

[32] Zhi Jin, Tingfang Wu, Taoning Chen, Deng Pan, Xuejiao Wang, Jingxin Xie, Lijun Quan, and Qiang Lyu. 2023. CAPLA: improved prediction of protein–ligand binding affinity by a deep learning approach based on a cross-attention mechanism. *Bioinformatics* 39, 2 (2023), btad049. doi:10.1093/bioinformatics/btad049

[33] Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou. 2021. Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 4395–4405. doi:10.18653/v1/2021.emnlp-main.360

[34] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 7873 (2021), 583–589. doi:10.1038/s41586-021-03819-2

[35] Vanessa Jurtz, Sinu Paul, Massimo Andreatta, Paolo Marcatili, Bjoern Peters, and Morten Nielsen. 2017. NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *The Journal of Immunology* 199, 9 (2017), 3360–3368. doi:10.4049/jimmunol.1700893

[36] Quentin Kaas, Manuel Ruiz, and Marie-Paule Lefranc. 2004. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic acids research* 32, suppl_1 (2004), D208–D210.

[37] Mahmood Kalemati, Saeid Darvishi, and Somayyeh Koohi. 2023. CapsNet-MHC predicts peptide-MHC class I binding based on capsule neural networks. *Communications Biology* 6, 1 (2023), 492.

[38] Thomas Kammertoens and Thomas Blankenstein. 2013. It's the Peptide-MHC Affinity, Stupid. *Cancer Cell* 23, 4 (2013), 429–431. doi:10.1016/j.ccr.2013.04.004

[39] Hiroyuki Kurata and Sho Tsukiyama. 2022. ICAN: interpretable cross-attention network for identifying drug and target protein interactions. *PLoS ONE* 17, 10 (2022), e0276609. doi:10.1371/journal.pone.0276609

[40] Taeheon Lee, Sangseon Lee, Minji Kang, and Sun Kim. 2021. Deep hierarchical embedding for simultaneous modeling of gpcr proteins in a unified metric space. *Scientific Reports* 11, 1 (2021), 9543.

[41] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 6637 (2023), 1123–1130.

[42] Guangna Liu, Hua Chen, Xingyu Cao, Lemei Jia, Wei Rui, Hongli Zheng, Daosheng Huang, Fang Liu, Yue Liu, Xueqiang Zhao, Peihua Lu, and Xin Lin. 2022. Efficacy of pp65-specific TCR-T cell therapy in treating cytomegalovirus infection after hematopoietic stem cell transplantation. *American Journal of Hematology* 97, 11 (2022), 1453–1463. doi:10.1002/ajh.26708

[43] Tianshi Lu, Ze Zhang, James Zhu, Yunguan Wang, Peixin Jiang, Xue Xiao, Chantale Bernatchez, John V Heymach, Don L Gibbons, Jun Wang, Lin Xu, Alexandre Reuben, and Tao Wang. 2021. Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nature Machine Intelligence* 3, 10 (2021), 864–875. doi:10.1038/s42256-021-00383-2

[44] Ariel Madrigal, Tianyuan Lu, Larisa M Soto, and Hamed S Najafabadi. 2024. A unified model for interpretable latent embedding of multi-sample, multi-condition single-cell data. *Nature Communications* 15, 1 (2024), 6573.

[45] Christoph Molnar. 2020. *Interpretable machine learning*. Lulu. com.

[46] Alessandro Montemurro, Viktoria Schuster, Helle Rus Povlsen, Amalie Kai Bentzen, Vanessa Jurtz, William D Chronister, Austin Crinklaw, Sine R Hadrup, Ole Winther, Bjoern Peters, et al. 2021. NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. *Communications Biology* 4, 1 (2021), 1060. doi:10.1038/s42003-021-02610-3

[47] Thierry Mora and Aleksandra M Walczak. 2019. Quantifying lymphocyte receptor diversity. In *Systems Immunology: An Introduction to Modeling Methods for Scientists*, Jayajit Das and Ciriyam Jayaprakash (Eds.). CRC Press, Taylor & Francis Group, Boca Raton FL, United States, 183–198. doi:10.1201/9781315119847

[48] Asher Mullard. 2022. FDA Approval of immunocore's first-in-class TCR therapeutic broadens depth of the T cell engager platform. *Nature Reviews Drug discovery* 21, 3 (2022), 170. doi:10.1038/d41573-022-00031-3

[49] Jun-Hyung Park, Yeachan Kim, Mingyu Lee, Hyuntae Park, and SangKeun Lee. 2024. MolTRES: Improving Chemical Language Representation Learning for Molecular Property Prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 14241–14254. https://aclanthology.org/2024.emnlp-main.788

[50] Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. 2024. BioT5+: Towards Generalized Biological Understanding with IUPAC Integration and Multi-task Tuning. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 1216–1240. doi:10.18653/V1/2024.FINDINGS-ACL.71

[51] Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen. 2020. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Research* 48, W1 (2020), W449–W454. doi:10.1093/nar/gkaa379

[52] Kenneth L Rock, Eric Reits, and Jacques Neefjes. 2016. Present Yourself! By MHC Class I and MHC Class II Molecules. *Trends in Immunology* 37, 11 (2016), 724–737. doi:10.1016/j.it.2016.08.010

[53] Luis A Rojas, Zachary Sethna, Kevin C Soares, Cristina Olcese, Nan Pang, Erin Patterson, Jayon Lihm, Nicholas Ceglia, Pablo Guasp, Alexander Chu, Rebecca Yu, Adrienne Kaya Chandra, Theresa Waters, Jennifer Ruan, Masataka Amisaki, Abderezak Zebboudj, Zagaa Odgerel, George Payne, Evelyna Derhovanessian, Felicitas Müller, Ina Rhee, Mahesh Yadav, Anton Dobrin, Michel Sadelain, Marta Łuksza, Noah Cohen, Laura Tang, Olca Basturk, Mithat Gönen, Seth Katz, Richard Kinh Do, Andrew S Epstein, Parisa Momtaz, Wungki Park, Ryan Sugarman, Anna M Varghese, Elizabeth Won, Avni Desai, Alice C Wei, Michael I D'Angelica, T Peter Kingham, Ira Mellman, Taha Merghoub, Jedd D Wolchok, Ugur Sahin, Özlem Türeci, Benjamin D Greenbaum, William R Jarnagin, Jeffrey Drebin, Eileen M O'Reilly, and Vinod P Balachandran. 2023. Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer. *Nature* 618, 7963 (2023), 144–150. doi:10.1038/s41586-023-06063-y

[54] Licai Sun, Bin Liu, Jianhua Tao, and Zheng Lian. 2021. Multimodal cross-and self-attention network for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4275–4279. doi:10.1109/ICASSP39728.2021.9414654

[55] Xinming Tu, Zhi-Jie Cao, Sara Mostafavi, Ge Gao, et al. 2022. Cross-linked unified embedding for cross-modality representation learning. *Advances in Neural Information Processing Systems* 35 (2022), 15942–15955.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, Vol. 30.

[57] Mai Ha Vu, Philippe A Robert, Rahmad Akbar, Bartlomiej Swiatczak, Geir Kjetil Sandve, Dag Trygve Truslew Haug, and Victor Greiff. 2024. Linguistics-based

formalization of the antibody language as a basis for antibody language models. *Nature Computational Science* (2024), 1–11.

[58] Jeffrey S. Weber, Matteo S. Carlino, Adnan Khattak, Tarek Meniawy, George Ansstas, Matthew H. Taylor, Kevin B. Kim, Meredith McKean, Georgina V. Long, Ryan J. Sullivan, Mark Faries, Thuy T. Tran, C. Lance Cowey, Andrew Pecora, Montaser Shaheen, Jennifer Segar, Theresa Medina, Victoria Atkinson, Geoffrey T. Gibney, Jason J. Luke, Sajeve Thomas, Elizabeth I. Buchbinder, Jane A. Healy, Mo Huang, Manju Morrissey, Igor Feldman, Vasudha Sehgal, Celine Robert-Tissot, Peijie Hou, Lili Zhu, Michelle Brown, Praveen Aanur, Robert S. Meehan, and Tal Zaks. 2024. Individualised neoantigen therapy mRNA-4157 (V940) plus pembrolizumab versus pembrolizumab monotherapy in resected melanoma (KEYNOTE-942): a randomised, phase 2b study. *The Lancet* 403, 10427 (2024), 632–644. doi:10.1016/S0140-6736(23)02268-7

[59] Neil A Weiss, Paul T Holmes, and Michael Hardy. 2006. *A course in probability*. Pearson Addison Wesley Boston, MA, USA:.

[60] Inge M. N. Wortel, Can Keşmir, Rob J. de Boer, Judith N. Mandl, and Johannes Textor. 2020. Is T Cell Negative Selection a Learning Algorithm? *Cells* 9, 3 (2020). doi:10.3390/cells9030690

[61] Yejian Wu, Lujing Cao, Zhipeng Wu, Xinyi Wu, Xinqiao Wang, and Hongliang Duan. 2023. CcBHLA: pan-specific peptide-HLA class I binding prediction via Convolutional and BiLSTM features. *bioRxiv* (2023). doi:10.1101/2023.04.24.538196

[62] Minghao Yang, Zhi-An Huang, Wei Zhou, Junkai Ji, Jun Zhang, Shan He, and Zexuan Zhu. 2023. MIX-TPI: a flexible prediction framework for TCR-pMHC interactions based on multimodal representations. *Bioinformatics* 39, 8 (2023), btad475. doi:10.1093/bioinformatics/btad475

[63] Mark Yarchoan, Edward J Gane, Thomas U Marron, Renzo Perales-Linares, Jian Yan, Neil Cooch, Daniel H Shu, Elana J Fertig, Luciane T Kagohara, Gabor Bartha, Josette Northcott, John Lyle, Sarah Rochestie, Joann Peters, Jason T Connor, Elizabeth M Jaffee, Ildiko Csiki, David B Weiner, Alfredo Perales-Puchalt, and Niranjan Y Sardesai. 2024. Personalized neoantigen vaccine and pembrolizumab in advanced hepatocellular carcinoma: a phase 1/2 trial. *Nature Medicine* 30, 4 (2024), 1044–1053. doi:10.1038/s41591-024-02894-y

[64] Zheng Ye, Shaohao Li, Xue Mi, Baoyi Shao, Zhu Dai, Bo Ding, Songwei Feng, Bo Sun, Yang Shen, and Zhongdang Xiao. 2023. STMHCpan, an accurate Star-Transformer-based extensible framework for predicting MHC I allele binding peptides. *Briefings in Bioinformatics* 24, 3 (2023), bbad164. doi:10.1093/bib/bbad164

[65] Hao Zhang, Ole Lund, and Morten Nielsen. 2009. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics* 25, 10 (2009), 1293–1299.

[66] Jing Zhang, Yingshuai Xie, Weichao Ding, and Zhe Wang. 2023. Cross on cross attention: Deep fusion transformer for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 8 (2023), 4257–4268. doi:10.1109/TCSVT.2023.3243725

[67] Ze Zhang, Danyi Xiong, Xinlei Wang, Hongyu Liu, and Tao Wang. 2021. Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. *Nature Methods* 18, 1 (2021), 92–99. doi:10.1038/s41592-020-01020-3

[68] Yu Zhao, Bing He, Fan Xu, Chen Li, Zhimeng Xu, Xiaona Su, Haohuai He, Yueshan Huang, Jamie Rossjohn, Jiangning Song, et al. 2023. DeepAIR: A deep learning framework for effective integration of sequence and 3D structure to enable adaptive immune receptor analysis. *Science Advances* 9, 32 (2023), eabo5128. doi:10.1126/sciadv.abo5128

B Useful Facts

B.1 Biological Molecules of Adaptive Immunity

In adaptive immunity, the major players are the highly diverse B and T cells, with unique surface receptors known as B cell receptors (BCRs) and T cell receptors (TCRs), respectively. These cells recognize specific parts of an antigen, referred to as epitopes. However, the mechanisms of antigen recognition differ between B and T cells. B cells target a fragment of the antigen known as a B cell epitope. Recognition by BCRs primarily depends on three-dimensional conformational information from the fragment, which contains mainly non-contiguous amino acid residues. On the other hand, T cell epitopes, recognized by TCRs, depend on their binding to major histocompatibility complex (MHC) molecules. These epitopes are linear, formed by contiguous amino acid residues.

Major Histocompatibility Complex (MHC). The **major histocompatibility complex (MHC)** is a type of cell surface proteins essential for the adaptive immunity. In humans, MHC genes are called human leukocyte antigens (*HLA*s). The MHC class I molecules present endogenous peptides from proteins self-generated intracellularly, while The MHC class II molecules are mainly expressed on antigen presenting cells. The MHC class I molecules contain an α chain from *MHC* class I genes and β_2 microglobulin ($\beta_2 m$), which can present peptides ranging from 8 to 12 amino acids. MHC class II molecules consist of one α and one β chain, allowing the binding of longer peptides ranging from 9 to 25 residues, or even longer. The MHC class I and MHC class II are also highly diverse, with approximately $6 \times 20^{6-7}$ and 12×20^{10} alleles, respectively [52]. The MHC class I molecules present endogenous peptides from proteins self-generated intracellularly, while The MHC class II molecules are mainly expressed on antigen presenting cells.

T Cell Receptor (TCR). The **T cell receptor (TCR)** is a type of protein complex on the surface of T cells responsible for recognizing fragments of antigen as peptides bound to MHCs. Classically, the TCR consists of an α chain and a β chain, which are encoded by gene *TRA* and *TRB*, respectively. The high diversity of TCR is generated by rearrangements of the V and J segments of the *TRA* gene and V, D, and J segments of the *TRB* gene in the thymus, with 10^{23} possible rearrangements theoretically [47]. Within the TCRs, the indices for α and β chains have been separately estimated to be 10^9 and 10^{14} [47]. Consequently, β chains garner a greater degree of attention and are the focus of significant experiments in TCR sequencing, making β chains a core component in data-driven modeling. In another estimation, the number of potential rearrangements can be up to 10^{61} [13]. While at one moment, there are around 10^{11} per human with around 10^9 distinct TCRs [13], which requires the highly precise prediction of pMHC-TCR for further drug development based on TCRs.

B Cell Receptor (BCR). The **B cell receptor (BCR)** from B cells contains multiple forms, including the secreted form and the membrane-bound form. Secreted BCRs are usually called **antibody** (**Ab**), while both membrane-bound and secreted BCRs can be called **immunoglobulin** (**Ig**). BCRs are arranged in three globular regions that roughly form a Y shape. In humans, one BCR unit consists of four chains, two heavy chains (H) and two light chains (L). Each heavy chain's variable region is approximately 110 amino acids in length. There are five types of mammalian BCR heavy chains denoted by Greek letters: α, δ, ε, γ and μ. These chains are found in **IgA**, **IgD**, **IgE**, **IgG**, and **IgM** antibodies, respectively. Heavy chains differ in size and composition. α and γ contain approximately 450 amino acids, while ε and μ have about 550 amino acids. In mammals, there are only two types of light chains, λ and κ, which have minor differences in the sequence. A light chain has two successive domains, constant (C_L) and variable (V_L). The approximate length of a light chain is 211–217 amino acids. The diversity of BCR is generated from V(D)J recombination and somatic hypermutation, with 10^{21} possible rearrangements theoretically [47]. Another estimation suggested that the total paired-sequence diversity is 10^{16-18}, while there are 5×10^9 B cells in the peripheral blood of a healthy human.

Table 4: Overview of Molecules Involved in Antigen Presentation.

Molecule	Location	Total Length/aa	Active Length	Main Function	Theoretic Diversity	Homology
MHC Class I	Cell Surface	α: 360 / $\beta_2 m$: 120	Relevant: α chain	Present peptides to $CD8^+$ T cells	$6 \times 20^{6-7}$	Varies
MHC Class II	Cell Surface	α & β: 260-280	Relevant: α_1 and β_1 domains	Present peptides to $CD4^+$ T cells	12×20^{10}	Varies
T Cell Receptor	T Cell Surface	α: 223 / β: 247	Variable regions: 110-120 each chain	Recognize peptide–MHC complexes	10^{23}	Low
B Cell Receptor	B Cell Surface or Secreted Form	Light: 211-217 Heavy: 450/ 550	Variable domain: 110	Recognize antigens	10^{21}	Low

B.2 Amino Acids Embedding

Accurately embedding amino acids is key to modeling protein interactions, as it preserves each residue's biochemical properties and positional context. Approaches include position-specific scoring matrices [44] and deep learning-based embeddings [7, 40, 55], which help maintain the structural and functional integrity of the sequence. Advanced embeddings often integrate attention mechanisms to capture long-range dependencies among residues, thereby improving representation of spatial relationships crucial for protein functionality [51].

B.3 Protein Interactions and Prediction Methods

In our computational study, we developed a specialized neural network model, termed Fusion-pMT, to understand the interactions within the peptide-MHC-TCR complex. The model's architecture leverages a custom-built submodule, which employs an advanced multi-head attention mechanism (with eight attention heads and a dropout rate of 0.1) to process and integrate features from peptide and MHC sequences. The sequences are embedded into a 64-dimensional space, facilitating a detailed representation of their complex biological characteristics.

The model encapsulates the dynamics of peptide–MHC interactions through its cross-attention mechanism, which is crucial for capturing the nuanced dependencies between these biomolecules. Further processing is performed by a fully connected neural network, which integrates the attention outputs with flattened peptide and MHC sequence features. This integration feeds into a deep learning pipeline that includes multiple layers of nonlinear transformations and dropout regularization, aiming to predict interaction outcomes robustly.

Training of the pMHC Model is meticulously orchestrated over 200 epochs, employing a binary cross-entropy loss function optimized via stochastic gradient descent with a learning rate of 0.1. This training regimen includes a patience mechanism set to 10 epochs to prevent overfitting and ensure model generalizability. Model performance is evaluated through both training and validation phases, with checkpoints saved upon achieving new best validation accuracies, underscoring the model's progressive learning capability.

C Miscellanies

C.1 Impacts on immunology and medicine

By using cross attention to address multi-sequence biological problems, the prediction of pMHC-TCR has implications in various fields. In immunology, an AI4Sci model understanding the TCR-pMHC interaction can help in the study of diseases, including autoimmune diseases, infections, and cancers. In medicine, AI-empowered predictions of TCR-pMHC interactions can potentially lead to individualized treatments in precision medicine.

C.2 Impacts in Healthcare

In healthcare, the prediction of pMHC-TCR interactions has significant implications in the development of advanced therapies against cancers or infectious diseases. (Figure 6) .

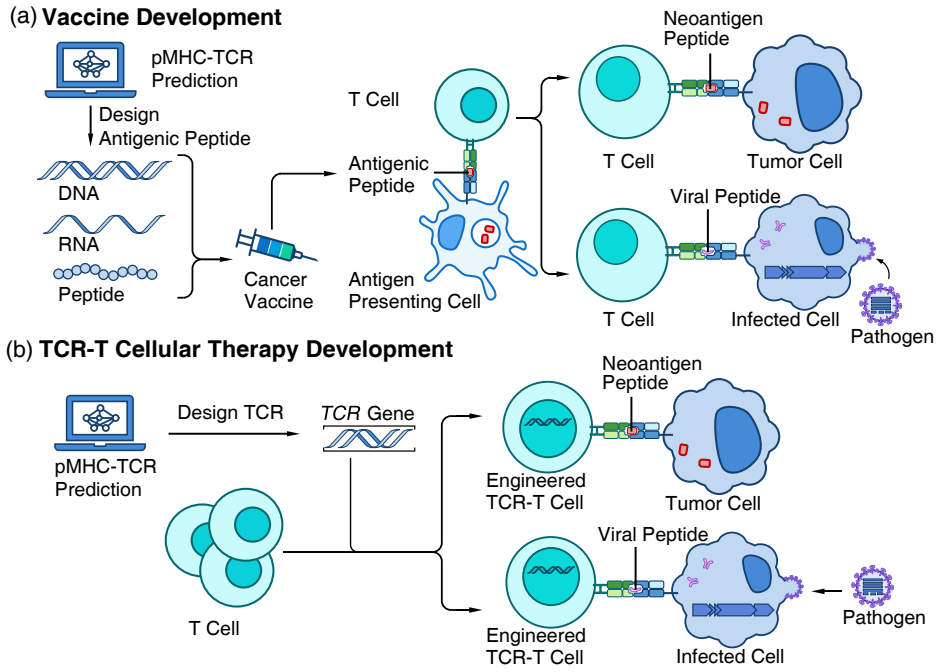


Figure 6: The Application of pMHC-TCR Binding Prediction in Healthcare (a) Schematic Representation of the Therapeutic Cancer Vaccine. (b) Schematic Representation of the Engineered TCR-T Cell Therapy.

Vaccine Development. Understanding which peptides can bind to MHC molecules and be recognized by TCRs can help in the design of more effective vaccines, especially the neoantigen-based cancer vaccine. Neoantigens are newly generated peptides from somatic mutations that can be recognized by TCRs of tumor-specific T cells. Once all mutations are identified, they must be computationally predicted from matched tumor-normal sequencing data, and then ranked according to their predicted capability in stimulating a T cell response. Neoantigen-based cancer has shown promising results in a phase IIb study [58]. This selection of effective neoantigen candidates relies on the precise prediction of pMHC-TCR interactions [53, 63].

In addition to cancers, pMHC-TCR prediction can also accelerate the development of infectious disease vaccines. During the COVID-19 pandemic, T-cell-directed vaccines has been designed in the form of peptides [25] and mRNA [3]. More precise prediction of pMHC-TCR interactions can improve the development of T-cell-directed vaccines in patients with immunodeficiency in phase I/II study [26] (Figure 6a).

TCR-T Cellular Therapy. The prediction of pMHC-TCR interactions can also aid in the development of T cell therapies, where the goal is to enhance the immune system's ability to recognize and destroy abnormal cells. The development of TCR-T therapies against cancer involves identifying a specific TCR that recognizes the tumor antigen by analyzing TCR sequencing data. Subsequently, this *TCR* gene can be manipulated to be expressed in autologous T cells. These engineered tumor-specific T cells can be expanded to induce tumor killing by recognizing pMHC on tumor cells [18, 24]. Two drug based on this therapy has been approve by the Food and Drug Administration of USA on January 25, 2022 [48] and August 2, 2024, respectively. Additionally, the proof-of-concept for using TCR-T therapies against infectious diseases have been validated in treating cytomegalovirus infection after hematopoietic stem cell transplantation [42], which sheds light on the broader application of TCR-T cellular therapies(Figure 6b).

C.3 Asset Licenses

- TransPHLA-AOMP [15]: TransPHLA-AOMP (transformer-based model for pHLA binding prediction and the automatically optimized mutated peptides program) is an algorithm designed to predict peptide and HLA binding. TransPHLA-AOMP is licensed under the GNU GENERAL PUBLIC license 3.0.
- pMTnet [43]: pMTnet (the pMHC–TCR binding prediction network) is an algorithm to predict TCR binding specificities of the neoantigens and T cell antigens in general presented by class I major histocompatibility complexes. pMTnet is licensed under the GNU GENERAL PUBLIC license 2.0.
- NetMHCpan [51]: NetMHCpan (pan-specific binding of peptides to MHC class I proteins of known sequence) is an algorithm to predict the binding of peptides to any MHC molecule of a known sequence using artificial neural networks. NetMHCpan is licensed under the GNU GENERAL PUBLIC license 3.0.
- VDJdb [22]: VDJdb is a curated database of T-cell receptor sequences of known antigen specificity. This database is licensed under the Attribution-NoDerivatives 4.0 International.