LightMHC: A Light Model for pMHC Structure Prediction with Graph Neural Networks

Antoine P. Delaunay InstaDeep
Yunguan Fu InstaDeep
Nikolai Gorbushin InstaDeep
Robert McHardy InstaDeep
Bachir A. Djermani InstaDeep
Liviu Copoiu InstaDeep
Michael Rooney BioNTech
Maren Lang BioNTech
Andrey Tovchigrechko BioNTech
Uğur Şahin BioNTech
Karim Beguir InstaDeep
Nicolas Lopez Carranza InstaDeep

Abstract

The peptide-major histocompatibility complex (pMHC) is a crucial protein in cell-mediated immune recognition and response. Accurate structure prediction is potentially beneficial for protein interaction prediction and therefore helps immunotherapy design. However, predicting these structures is challenging due to the sequential and structural variability. In addition, existing pre-trained models such as AlphaFold 2 require expensive computation thus inhibiting high throughput *in silico* peptide screening. In this study, we propose LightMHC: a lightweight model (2.2M parameters) equipped with attention mechanisms, graph neural networks, and convolutional neural networks. LightMHC predicts full-atom pMHC structures from amino-acid sequences alone, without template structures. The model achieved comparable or superior performance to AlphaFold 2 and ESMFold (93M and 15B parameters respectively), with five-fold acceleration (6.65 seconds/sample for LightMHC versus 36.82 seconds/sample for AlphaFold 2), potentially offering a valuable tool for immune protein structure prediction and immunotherapy design.

1 Introduction

In cellular immune responses, the peptide-major histocompatibility complex (pMHC) is critical for binding and presenting pathogen or tumour-derived peptides to T-cell receptors (TCRs) to initiate immune response [Chaplin, 2010]. Understanding their 3D structures could provide insights into TCR:pMHC recognition mechanisms [Crean et al., 2020], thus helping to identify potential peptide epitopes in targeted immunotherapy development [Kuhlman and Bradley, 2019, Lang et al., 2022]. However, resolving protein structures is expensive and time-consuming. In addition, TCR:pMHC complexes inherently have large structural and sequence variabilities. As a result, only a limited number of structures are available, hindering specific model training and validation [Berman et al., 2000]. While AlphaFold 2 [Jumper et al., 2021] and ESMFold [Lin et al., 2022] have demonstrated remarkable successes, these models cannot always capture accurately the conformational dynamics of flexible regions and remain computationally expensive, making them unsuitable for high-throughput inference. There also exist application-specific models that are limited to specific protein domains [Abanades et al., 2022, Delaunay et al., 2022] or backbone atoms [Delaunay et al., 2022, Cohen et al., 2022], or require pre-defined templates [Abanades et al., 2022], limiting their applicability to a broader range of immune recognition molecules. In this work, we focus on pMHC structure prediction task and present LightMHC: a light model combining attention-based graph neural networks and convolutional neural networks. The model inputs the target amino-acid se-

Machine Learning for Structural Biology Workshop, NeurIPS 2023.

quences and directly outputs the full-atom structure. The model does not require template structures as input, nor perform side-chain packing for post-processing, enabling efficient and scalable predictions. Notably, we conducted out-of-sample evaluations (Section 4.2) to assess the model's ability to generalise on unseen data, a step not taken in previous studies. This is particularly important given the limited availability of training data. LightMHC, with only 2.2M parameters, achieved comparable or superior performance and improved generalisability against AlphaFold 2 [Jumper et al., 2021] and ESMFold [Lin et al., 2022].

2 Related work

In immune protein structure prediction, progress covers pMHCs, TCRs and antibodies. MHCFold [Aronson et al., 2022] predicts pMHC backbones with limited evaluation on C_α RMSD. Delaunay et al. [2022] enhance performance with GNN but confine to 9-meric peptide backbone prediction, facing scalability issues with side-chain recovery. Pre-trained models like AlphaFold 2 [Jumper et al., 2021, Lin et al., 2022] tend to be overparametrised, while smaller specialised models match single-family performance [Delaunay et al., 2022]. Adaptations of AlphaFold 2 focus on pMHC but lack out-of-sample testing [Motmaen et al., 2022, Mikhaylov and Levine, 2023]. ImmuneBuilder [Abanades et al., 2023] inspired by AlphaFold 2, adopts untied weights with eight iterative refinement steps for TCR and antibodies. However, this extends training time and risks overfitting. For antibodies, DeepAb [Ruffolo et al., 2022] uses ResNet and LSTM to predict distance maps and torsion angles, yielding 3D coordinates. AbLooper [Ruffolo et al., 2022] employs equivariant graph neural networks for specific antibody segments, but requires the remaining part of the structure. IgFold [Ruffolo and Gray, 2022] combines AntiBERTy [Ruffolo et al., 2021] and AlphaFold 2's invariant point attention but lacks side-chain predictions and requires templates. Our proposed model addresses these limitations by integrating graph and convolutional neural networks, enabling full-atom pMHC structure prediction at low computational cost, while being rigorously assessed on out-of-sample distributions, thus demonstrating improved generalisability and biological consistency.

3 Methods

3.1 Graph Representation

The protein structure is represented by a graph of amino acids from all chains. Nodes are labelled by amino acid type and chain. Edges remain constant for all structures during both training and testing phases. These edges are derived from structure 1AKJ, where nodes are connected by edges if their Cα distance is within 8 Å, indicating contact [Xia and Ku, 2021]. Edge type depends on bond nature and chain relation. The reference structure is fixed in training and inference, limiting accurate topology due to conformational changes or variable chain lengths. This ensures no data leakage and robust model generalisation. The approach extracts only a structure topology, differing from protein folding models with geometrical features based on closest dataset templates.

3.2 Neural Network Architecture

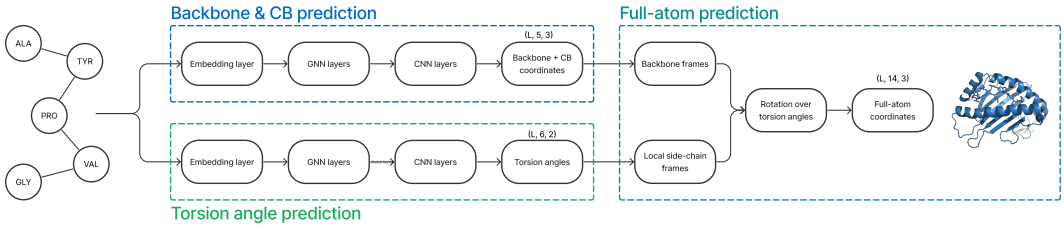


Figure 1: Neural Network Architecture. Each model stage (backbone atom prediction, torsion angle prediction, and full-atom prediction) is represented by a large rectangular box in blue, green and turquoise, respectively. Model layers, and their outputs are represented by small rectangular boxes.

LightMHC contains three stages, backbone & C$_\beta$ atom prediction, torsion angle prediction, and full-atom prediction, as illustrated in Figure 1. Backbone & C$_\beta$ atom and torsion angle prediction

share the same architecture with untied weights. Nodes are first encoded and then refined by GNN layers. After GNN processing, MHC and peptide node embeddings are concatenated, padded, and input to a CNN to predict backbone coordinates or torsion angles. Afterwards, these predictions are combined to derive full-atom structures similar to AlphaFold 2 [Jumper et al., 2021]. Specifically, GNN consists of four message-passing layers with a multi-head attention mechanism [Shi et al., 2021], following Delaunay et al. [2022]. The message-passing layer is wrapped into a Transformer architecture involving a ReLU-activated feed-forward network. The CNN has five ReLU-activated 1D-convolutional layers that combine node information to yield a coherent structure. Notably, one-hop GNN layers lack consideration of distant chain effects. Adding edges increases computational complexity, and more GNN layers would exacerbate the over-smoothing issue [Rusch et al., 2023]. Hence, CNN synthesises node-level information, compensating for GNN layers' sparse-attention limits. The full-atom prediction is adapted from the AlphaFold 2 algorithm 24 [Jumper et al., 2021]. Since LightMHC predicts backbone atom coordinates directly, Gram-Schmidt algorithm is used to derive backbone rigid frames (orientation and position of each residue's backbone) from predicted coordinates (Appendix A). Side chain coordinates are determined using derived backbone frames, idealised local frames, and predicted torsion angles.

3.3 Loss

We modify the loss in Delaunay et al. [2022] to adapt longer sequences while preserving geometrical and chemical consistency. Specifically, the loss is a sum of three terms, 1) $\mathcal{L}_{\text{Huber}}$ (Appendix B.1), a Huber loss on the backbone atom coordinates between ground truth and prediction, 2) $\mathcal{L}_{\text{Inter-bond}}$ (Appendix B.2), a Huber loss of distances on the bond between two neighbour residues, 3) $\mathcal{L}_{\text{Dihedral}}$ (Appendix B.3), an MSE of the trigonometric functions ($\cos$ and $\sin$) of dihedral angles.

3.4 Post-processing

Atom clashes are reduced by design due to the use of literature ideal local geometry as described in Appendix A and inter-residue bond loss (Appendix B.2). However, few structural violations may subsist and are removed using the PyRosetta IdealMover algorithm [Leaver-Fay et al., 2011] in the fast mode which is the most suited for minimal post-processing modifications. This provides clash-free structures while keeping the original shape of the predicted structures (the median full-atom RMSD between before and after applying IdealMover on our pMHC random partition test set is 0.2 Å). The mover is single-thread and paralleled. Analogous post-processing (AMBER99SB, [Hornak et al., 2006]) has also been used in AlphaFold 2.

4 Experimental Setting

4.1 Data

We focus on predicting the pMHC structures of MHC class I proteins due to the increased structural variability of the peptide in the binding groove compared to class II MHC [Jones, 1997]. For pMHC, we use a dataset of 749 crystal structures from the RCSB Protein Data Bank (PDB) [Berman et al., 2000]. After excluding structures with missing backbone information, we retained 665 structures. We aligned structures based on the MHC Cα atoms using a randomly selected reference structure (PDB code: 1AKJ). For benchmarking purposes, the MHC chain is restricted to its binding interface (α_1, α_2 domains) [Motmaen et al., 2022].

4.2 Benchmark

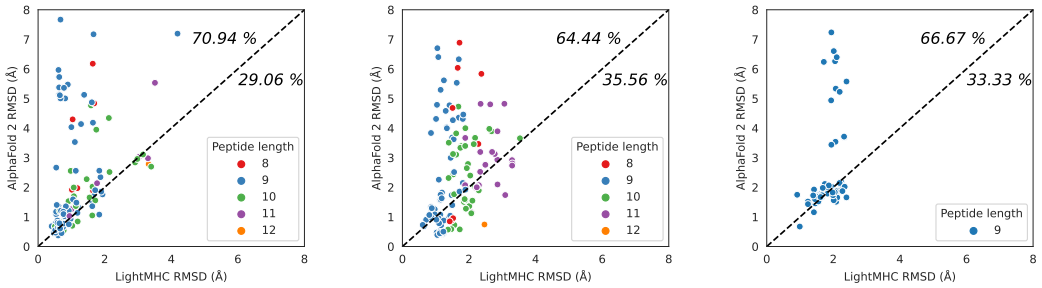
We conducted in-sample and out-of-sample evaluations, using distinct dataset partitions. Each partition consists of both a training and a test set, and our model was retrained and evaluated on each of these partitions. For in-sample assessment, we use a random partition, ensuring that no common peptides are shared between training and test sets. For out-of-sample evaluation, we use two additional partitions: a peptide sequence partition, where sequences are clustered to minimise similarity between train and test peptide sequences using the PAM30 scoring matrix embedding [Dayhoff, 1978], and a peptide structure partition, where the structures are clustered based on peptide backbone root mean square deviation (RMSD). Due to the significant influence of the peptide length on its conformation,

the peptide structure partition contains only 9-mer peptides to avoid any bias stemming from the peptide length. We evaluated the accuracy of our predictions by calculating the RMSD for both backbone and full-atom peptide structures. The RMSD was calculated for the entire structure and stratified by chain, distinguishing between the peptide and MHC chains. The model is benchmarked against AlphaFold 2 with Motmaen et al. [2022] methodology and ESMFold [Lin et al., 2022].

5 Results and Discussions

Table 1: Median RMSD (Å) on the pMHC dataset stratified per partition, chain and atoms considered. Statistical significance: p-values are computed via a signed Wilcoxon paired one-sided rank test between our model and AlphaFold 2 (*, **, *** denote significance levels at $p < 0.1$, 0.05, and 0.01).

Partition	Model	MHC (Å)		Peptide (Å)	
		Backbone	Full atom	Backbone	Full atom
Random	AlphaFold 2	0.60	1.17	1.40	2.08
	ESMFold	1.01	1.73	14.41	14.07
	LightMHC	**0.56**	**1.09*****	**0.98*****	**2.04*****
Peptide Sequence	AlphaFold 2	**0.54**	1.11	2.00	3.16
	ESMFold	1.02	1.78	8.27	9.62
	LightMHC	0.55	**1.05*****	**1.52*****	**2.78*****
Peptide Structure	AlphaFold 2	0.89	1.37	**1.81**	3.50
	ESMFold	1.00	1.71	14.58	15.32
	LightMHC	**0.55*****	**1.26*****	1.84	**3.37*****



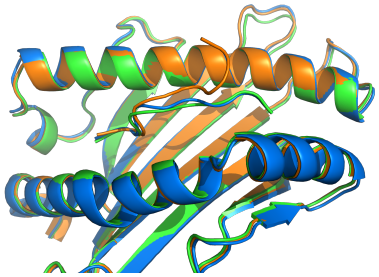
(a) Peptide Random partition (b) Peptide Sequence partition (c) Peptide Structure partition

Figure 2: Peptide RMSD for LightMHC vs AlphaFold 2. Each dot represents a structure from each test set. The numbers on each plots represent the proportion of samples above/below the x=y line. The legend displays the peptide lengths present in each test set.

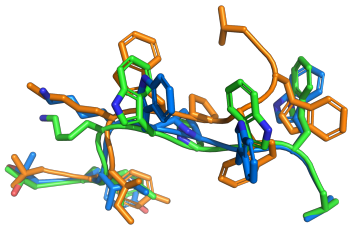
Table 2: Median peptide backbone RMSD (Å) stratified by peptide length on the peptide sequence partition (Appendix C). Statistical significance: see Table 1.

Model	Peptide length				
	8	9	10	11	12
AlphaFold 2	4.07	1.33	2.44	2.78	**0.74**
ESMFold	8.79	12.73	8.69	6.36	16.10
LightMHC	**1.59***	**1.16*****	**1.89*****	**2.56**	2.46

The median RMSDs of LightMHC, AlphaFold 2, and ESMFold on pMHC under different partitions have been summarised in Table 1. All methods achieved low RMSD for MHC structure, as a result of high structural stability [Wilson and Fremont, 1993]. Regarding peptides, ESMFold often positions the peptide outside of the binding groove, resulting in large RMSD on peptides in all assessments. This



(a) pMHC backbone cartoon representation



(b) Peptide representation with side-chain sticks

Figure 3: Example of a predicted structure (PDB ID: 7EJM). Experimental, LightMHC, and AlphaFold 2 predicted structures are represented in green, blue, and orange respectively. ESMFold pMHC prediction is not shown as the peptide is outside the binding groove. Backbone atoms are represented with the cartoon representation. Peptide side-chain atom are represented as sticks on the right panel. Full-atom RMSD of LightMHC and AlphaFold 2 are 2.02 Å and 5.09Å respectively.

also holds true for AlphaFold 2 if templates were not provided, see ablation study in Appendix F.2, showing the importance of custom template structure guidance for pre-trained models. In contrast, with a fixed template that does not depend on the target pMHC, LightMHC outperformed template-dependent AlphaFold 2 on 70.94%, 64.44%, and 66.67% test cases for random, sequence, and structure partitions (see Figure 2). The differences are also statistically significant for random and sequence partitions. When stratifying the peptides by length (Table 2 and Appendix C), LightMHC remained significantly better across all lengths, with the exception of 12-mers where only one sample exists in the test set. Example predictions are shown in Figure 3 and Appendix D, highlighting the accuracy of LightMHC in generating full-atom predictions without side-chain packing. Particularly, AlphaFold 2 failed to anchor the peptide residues on N- and C- termini (Figure 3a).

Besides high accuracy, LightMHC's inference is more than five times faster (6.65 seconds/sample versus 36.82 seconds/sample for AlphaFold 2, inclusive of post-processing in both cases, see Appendix G), demonstrating high potential for large-scale applications. For instance, the state-of-the-art pMHC binding prediction model, NetMHCpan [Jurtz et al., 2017], leverages a sequence dataset with $\sim$ 850,000 experimental pMHC pairs. Predicting structures for all these pMHCs would require 362 days using AlphaFold 2 on a GPU, while LightMHC only needs 16 hours on a 100 CPU cores cluster.

An ablation study was conducted on model architecture, showing that both GNN and CNN layers are necessary for predicting accurate peptide structures Appendix F.1. Additionally, a model with identical architecture was trained and evaluated on a TCR dataset to study the performance of the proposed framework (Appendix H). However, no model consistently outperformed the best across different partitions. This suggests that TCR structure prediction is a more challenging task, potentially due to the higher sequential and structural variability.

6 Conclusion and Discussion

We introduced LightMHC: a lightweight model for predicting pMHC structures in immune proteins. Combining graph neural networks with attention and convolutional neural networks, our model predicts full-atom pMHC structures from sequences alone without the need for template protein structures. With only 2.2M parameters, LightMHC showed comparable or better performance, improved generalisability, and biological fidelity against large pre-trained models such as AlphaFold 2 and ESMFold. Importantly, LightMHC inference is more than five-folds faster than AlphaFold 2, enabling large-scale *in silico* peptide screening. Such a high throughput could potentially deepen our understanding of cell-mediated immunity and enhance immunotherapy design. Future research may improve the model on T-cell receptor predictions with pre-training, transfer learning, and ensembles.

References

Brennan Abanades, Guy Georges, Alexander Bujotzek, and Charlotte M. Deane. ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation. *Bioinformatics*, 38(7): 1877–1880, 2022.

Brennan Abanades, Wing Ki Wong, Fergus Boyles, Guy Georges, Alexander Bujotzek, and Charlotte M. Deane. ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. *Communications Biology*, 6(1):575, 2023.

Alon Aronson, Tanya Hochner, Tomer Cohen, and Dina Schneidman-Duhovny. Structure modeling and specificity of peptide-MHC class I interactions using geometric deep learning. *bioRxiv*, 2022. doi: 10.1101/2022.12.15.520566.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, January 2000.

David D. Chaplin. Overview of the immune response. *Journal of allergy and clinical immunology*, 125(2):S3–S23, 2010.

Tomer Cohen, Matan Halfon, and Dina Schneidman-Duhovny. NanoNet: Rapid and accurate end-to-end nanobody modeling by deep learning. *Frontiers in Immunology*, 13, 2022. ISSN 1664-3224. doi: 10.3389/fimmu.2022.958584.

Rory M. Crean, Bruce J. MacLachlan, Florian Madura, Thomas Whalley, Pierre J. Rizkallah, Christopher J. Holland, Catriona McMurran, Stephen Harper, Andrew Godkin, Andrew K. Sewell, et al. Molecular rules underpinning enhanced affinity binding of human T cell receptors engineered for immunotherapy. *Molecular Therapy-Oncolytics*, 18:443–456, 2020.

Margaret Dayhoff. *Atlas of protein sequence and structure*. National Biomedical Research Foundation, 1978.

Antoine P. Delaunay, Yunguan Fu, Alberto Bégué, Robert McHardy, Bachir A. Djermani, Michael Rooney, Andrey Tovchigrechko, Liviu Copoiu, Marcin J. Skwark, Nicolas Lopez Carranza, Maren Lang, Karim Beguir, and Uğur Şahin. Peptide-MHC Structure Prediction With Mixed Residue and Atom Graph Neural Network. *bioRxiv*, 2022. doi: 10.1101/2022.11.23.517618.

R. A. Engh and R. Huber. *Structure quality and target parameters*. Springer Netherlands, 2006.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, April 2019.

Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3):712–725, 2006.

E Yvonne Jones. MHC class I and class II structures. *Current Opinion in Immunology*, 9(1):75–79, 1997. ISSN 0952-7915.

John Jumper et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589, August 2021. Number: 7873.

Vanessa Jurtz, Sinu Paul, Massimo Andreatta, Paolo Marcatili, Bjoern Peters, and Morten Nielsen. NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *The Journal of Immunology*, 199(9):3360–3368, 11 2017. ISSN 0022-1767.

Brian Kuhlman and Philip Bradley. Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology*, 20(11):681–697, 2019.

Franziska Lang, Barbara Schrörs, Martin Löwer, Özlem Türeci, and Ugur Sahin. Identification of neoantigens for individualized therapeutic cancer vaccines. *Nature reviews Drug discovery*, 21(4): 261–282, 2022.

Andrew Leaver-Fay, Michael Tyka, and Steven M. et al. Lewis. Rosetta3. In *Methods in Enzymology*, volume 487, pages 545–574. Elsevier, 2011.

Jinwoo Leem, Saulo H P. de Oliveira, Konrad Krawczyk, and Charlotte M. Deane. STCRDab: the structural T-cell receptor database. *Nucleic Acids Research*, 46(D1):D406–D412, 10 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx971. URL `https://doi.org/10.1093/nar/gkx971`.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, and Alexander Rives. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Victor Mikhaylov and Arnold J. Levine. Accurate modeling of peptide-MHC structures with AlphaFold. *bioRxiv*, 2023. doi: 10.1101/2023.03.06.531396.

Amir Motmaen, Justas Dauparas, Minkyung Baek, Mohamad H. Abedi, David Baker, and Philip Bradley. Peptide binding specificity prediction using fine-tuned protein structure prediction networks. preprint, Bioinformatics, July 2022.

Jeffrey A. Ruffolo and Jeffrey J Gray. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Biophysical Journal*, 121(3):155a–156a, 2022.

Jeffrey A. Ruffolo, Jeffrey J Gray, and Jeremias Sulam. Deciphering antibody affinity maturation with language models and weakly supervised learning. *arXiv preprint arXiv:2112.07782*, 2021.

Jeffrey A Ruffolo, Jeremias Sulam, and Jeffrey J Gray. Antibody structure prediction using interpretable deep learning. *Patterns*, 3(2), 2022.

T. Konstantin Rusch, Michael M. Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks. *arXiv preprint arXiv:2303.10993*, 2023.

Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification, May 2021. Number: arXiv:2009.03509.

Ian A. Wilson and Daved H. Fremont. Structural analysis of MHC class I molecules with bound peptide antigens. *Seminars in Immunology*, 5(2):75–80, April 1993.

Tian Xia and Wei-Shinn Ku. Geometric Graph Representation Learning on Protein Structure Prediction. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1873–1883, Virtual Event Singapore, August 2021. ACM.

Appendices

A Backbone coordinate generation

LightMHC's backbone coordinate prediction is described as follows, in comparison to AlphaFold 2 [Jumper et al., 2021]. Local ideal residue geometry [Engh and Huber, 2006] defines optimal relative positions of a residue type's backbone atoms. AlphaFold 2 predicts a backbone frame (rotation matrix and translation vector) for each residue and derives full-atom coordinates from these frames. In contrast, LightMHC outputs backbone atom coordinates, from which a frame is extracted using the Gram-Schmidt algorithm. These frames are combined with local ideal residue geometry [Engh and Huber, 2006] to refine predicted backbone coordinates. Local ideal geometry ensures biological consistency and avoids atom clashes. We empirically found this modified approach yielded better results on our model than AlphaFold 2's original method.

B Loss

B.1 Huber Loss

The role of the Huber loss is to give the correct overall shape and position of the amino acids. Let $y_a^{(j)}$ and $\hat{y}_a^{(j)}$ denote the true and predicted coordinates of the j^{th} atom in the a^{th} amino acid, respectively. L is the length of the sequence and n_a is the number of atoms of the a^{th} amino-acid. The formula of this loss is given by:

$$\mathcal{L}_{\text{Huber}} = \frac{1}{L}\sum_{a=1}^{L}\sum_{j=1}^{n_a}\frac{1}{n_a}l(y_a^{(j)}, \hat{y}_a^{(j)}). \tag{1}$$

$$\text{with } l(y_a^{(j)}, \hat{y}_a^{(j)}) = \begin{cases} ||y_a^{(j)} - \hat{y}_a^{(j)}||^2 & \text{if } |y_a^{(j)} - \hat{y}_a^{(j)}| \leq 1.0 \\ |y_a^{(j)} - \hat{y}_a^{(j)}| & \text{else.} \end{cases}$$

We observed empirically that Huber loss led to more stable training trajectories than MSE. We also attempted to use AlphaFold 2 FAPE loss [Jumper et al., 2021] on backbone frames extracted from predicted backbone coordinates, but obtained a slower convergence and less accurate structures.

B.2 Inter-bond Loss

We adapt AlphaFold 2 inter-residue bond loss [Jumper et al., 2021]. The inter-residue bond connects the carbonyl carbon atom of the a^{th} amino-acid with the nitrogen atom of the $a + 1^{th}$ amino-acid. The predicted (resp. true) bond length between the a^{th} and $a + 1^{th}$ amino-acids are denoted as $\hat{d}_a$ and d_aThe inter-residue bond loss is defined as:

$$\mathcal{L}_{\text{Inter-bond}} = \frac{1}{L-1}\sum_{a=1}^{L-1}||d_a - \hat{d}_a||^2. \tag{2}$$

B.3 Dihedral Loss

Let $\theta_a^{(j)}$ the j^{th} dihedral angle of the a^{th} amino-acid. Following Xia and Ku [2021], the dihedral loss is defined as:

$$\mathcal{L}_{\text{Dihedral}} = \frac{1}{L}\sum_{a=1}^{L}\sum_{j=1}^{N_{angles}}\frac{1}{N_{angles}}(\sin(\theta_a^{(j)}) - \sin(\hat{\theta}_a^{(j)}))^2 + (\cos(\theta_a^{(j)}) - \cos(\hat{\theta}_a^{(j)})^2 \tag{3}$$

C Dataset peptide lengths

We provide in Table 3 the number of PDB per peptide length, stratified by dataset partition and subset. In addition to Table 2 in the main text, Table 4 provides supplementary stratification results per peptide length on the random partition.

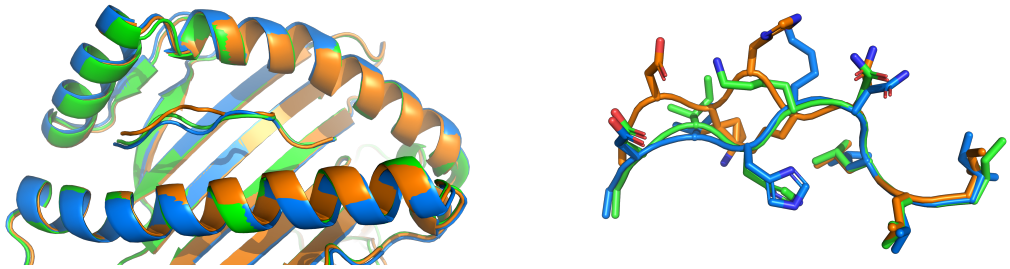
Table 3: Peptide length for each partition and subset.

Partition	Set	8	9	10	11	12
Random	Train	27	382	106	28	3
	Test	9	74	27	6	1
Sequence	Train	28	387	100	12	3
	Test	8	71	33	22	1
Structure	Train	0	387	0	0	0
	Test	0	71	0	0	0

Table 4: Median peptide backbone RMSD (Å) stratified by peptide length on the random partition. Statistical significance: see Table 1.

Model	Peptide length				
	8	9	10	11	12
AlphaFold 2	1.92	1.25	1.66	2.56	**2.78**
ESMFold	14.90	14.41	14.34	11.90	5.64
LightMHC	**1.11*****	**0.77*****	**1.46****	**2.40**	3.32

D Additional predicted structures visualisation



(a) pMHC backbone cartoon representation

(b) Peptide representation with side-chain sticks

Figure 4: Example of a predicted structure (PDB ID: 7KGO). Experimental, LightMHC, and AlphaFold 2 predicted structures are represented in green, blue, and orange respectively. ESMFold prediction is not shown as the peptide is outside the binding groove. Backbone atoms are represented with the cartoon representation. Peptide side-chain atom are represented as sticks on the right panel. Full-atom RMSD of LightMHC and AlphaFold 2 are 1.61 Å and 4.05Å respectively.

E Experiments Parameters

Our model is implemented with PyTorch and PyTorch Geometric [Fey and Lenssen, 2019] and trained on a Nvidia A100 40GB GPU. The model can be run on a single CPU at inference time. We use Adam with an initial learning rate of 1×10^{-3} as the optimiser. Hyper-parameters were empirically set based on the training set results without extensive tuning. The official implementations of AlphaFold 2 [Jumper et al., 2021] (with Motmaen et al. [2022] 200 residue gap trick added to the peptide sequence) and ESMFold [Lin et al., 2022] were used for benchmarking. Model and training hyper-parameters are defined in Table 5 and Table 6 respectively. CNN layers have a kernel of size 25, a dilation of 1 and a stride of 1.

Table 5: Model hyper-parameters

Block	Number of layers	Input size	Output size
GNN blocks	2	64	128
CNN blocks	5	128	12

Table 6: Training hyper-parameters

Parameter	Value	Parameter	Value
Max epochs	500 (pMHC) / 1,000 (TCR)	MHC maximum length (α_1, α_2 domains)	180
Batch size	16	Peptide maximum length	13
Output shape	(L, 14, 3)	Embedding dimension d_{emb}	128

F Ablation studies

F.1 Model parts ablation

To study the impact of each part of our model, we ran it on the randomly partition pMHC dataset removing either the GNN layers or the CNN layers. In the latter case, the CNN layers are replaced by a single feed-forward layer to reduce GNN 128-dim output to the corresponding dimension (backbone / torsion angles). The results are reported in Table 7.

Table 7: Median RMSD on the pMHC randomly partition dataset for different kept layers.

Layers	MHC (Å)		Peptide (Å)	
	Backbone	Full atom	Backbone	Full atom
GNN + CNN	0.56	1.09	**0.98**	**2.04**
CNN	**0.44**	**0.96**	1.19	2.10
GNN	0.47	1.01	1.30	2.20

F.2 AlphaFold 2 input ablation

AlphaFold 2 takes as input Multiple Sequence Alignment (MSA) information as well as templates. The input derived from the templates consists notably of geometrical and distance features [Jumper et al., 2021]. We run AlphaFold 2 on the pMHC randomly partitioned dataset using MSA only and MSA + templates. The results are reported in Table 8.

Table 8: AlphaFold 2 median RMSD on the pMHC randomly partitioned dataset for different input features.

Input	MHC (Å)		Peptide (Å)	
	Backbone	Full atom	Backbone	Full atom
MSA + Template	**0.60**	**1.17**	**1.40**	**2.08**
MSA	2.16	2.53	16.69	17.33

G Inference time

For AlphaFold 2 and ESMFold, We assess the running time on one A100 40GB GPU and 8 e2-standard 64 GB CPUs. For our model, we report the inference time on a single e2-standard 64 GB given that PyRosetta IdealMover [Leaver-Fay et al., 2011] is not suitable for GPU inference. Note that AlphaFold 2 inference time does not include MSA and template search. The times reported are meant to provide an order of magnitude of the approximative relative performance of each model when they are run in a real-world setting, rather than a thorough analysis.

H Extension of the model to T-cell receptors (TCR) proteins

A T-cell receptor is a protein found on the surface of T-cells that recognises and binds to specific pMHCs, initiating an immune response against foreign substances in the body. The TCR structures were aligned based on the TCR Cα atoms using a randomly selected reference structure (PDB code: 1AO7). TCR is composed of two different protein chains: alpha and beta chains, each comprising constant and variable regions. The variable regions are responsible for antigen recognition and are further divided into complementarity-determining regions (CDRs) and framework regions (FRs). CDR loops are short stretches of amino acids within the variable regions of the TCR chains that are directly involved in recognising and binding to antigens. There are three CDR loops in both the alpha and beta chains of the TCR, labelled CDR1, CDR2, and CDR3. The CDR3 loop, in particular, is the most diverse and crucial for antigen recognition.

H.1 Data and benchmark

For the TCR analysis, similarly to pMHC, we used a curated dataset of 531 crystal structures from the STCRDab [Leem et al., 2017].

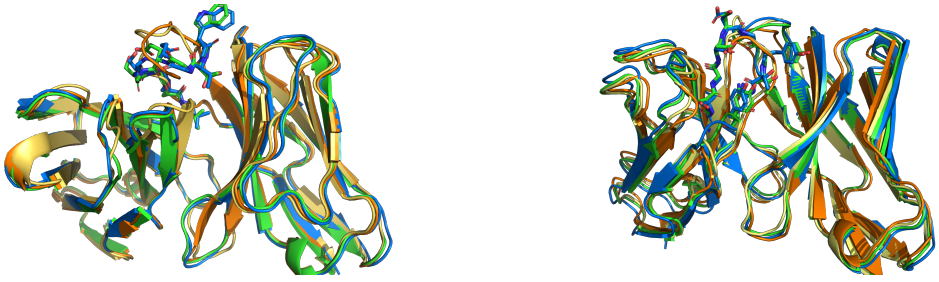
In the in-sample assessment, we performed a random partition to ensure that no common TCR sequences were shared between the training and test sets. For the out-of-sample evaluation, we employed a sequence similarity partition based on the two TCR sequences similarly to the pMHC case.

Performing a structure partition for TCR proved to be challenging due to the significant variability in loop lengths. However, by considering the results on the CDR3α and CDR3β loops, which are the most variable and critical parts of TCR structures in terms of binding affinity and specificity to the pMHC complex, we were able to assess the performance of our model in capturing their structural characteristics.

H.2 Metrics

Similarly to pMHC complexes we report RMSD stratified by chain (TCR α and β chains). Additionally, we computed the RMSD specifically for each CDR loop as they are known for their challenging modelling and critical biological functions.

H.3 Results



(a) PDB ID: 4NQC (b) PDB ID: 7M72

Figure 5: Examples of predicted TCR structures. Experimental, our model, AlphaFold 2 and ESMFold predicted structures are represented in green, blue, orange, and yellow respectively. CDR3β loop atom sticks are represented for experimental and our model structures (others omitted for legibility).

Table 9: Median RMSD (Å) on the TCR dataset stratified per partition, chain and atoms considered.

Partition	Model	TCRα (Å)		TCRβ (Å)		CDRs α + β (Å)	
		Backbone	Full atom	Backbone	Full atom	Backbone	Full atom
Random	AlphaFold 2	**1.06**	1.62	**1.28**	**1.80**	**1.46**	2.08
	ESMFold	1.45	2.00	1.35	1.92	1.65	2.28
	LightMHC	1.15	**1.60**	1.43	1.88	1.48	**2.01**
Sequence	AlphaFold 2	**1.53**	1.88	1.50	1.81	2.04	2.41
	ESMFold	1.76	2.01	**1.20**	**1.66**	2.14	2.72
	LightMHC	1.60	**1.83**	1.59	2.03	**1.89**	**2.35**

Partition	Model	CDR1α (Å)		CDR1β (Å)		CDR1 α + β (Å)	
		Backbone	Full atom	Backbone	Full atom	Backbone	Full atom
Random	AlphaFold 2	1.06	**1.25**	1.29	1.68	**1.25**	**1.63**
	ESMFold	1.37	1.87	**1.09**	**1.38**	1.33	1.74
	LightMHC	**0.86**	1.28	1.36	1.83	1.34	1.73
Sequence	AlphaFold 2	1.65	1.91	1.58	1.76	1.68	2.06
	ESMFold	1.77	2.33	**1.44**	**1.71**	**1.63**	**2.02**
	LightMHC	**1.50**	**1.88**	2.15	2.85	1.99	2.44

Partition	Model	CDR2α (Å)		CDR2β (Å)		CDR2 α + β (Å)	
		Backbone	Full atom	Backbone	Full atom	Backbone	Full atom
Random	AlphaFold 2	**0.98**	**1.40**	**0.85**	**1.32**	**0.99**	**1.39**
	ESMFold	1.22	1.76	1.11	1.52	1.26	1.79
	LightMHC	1.15	1.55	1.30	1.75	1.30	1.76
Sequence	AlphaFold 2	1.66	1.96	1.60	1.83	1.93	1.98
	ESMFold	**1.41**	1.88	**1.30**	**1.27**	**1.38**	**1.64**
	LightMHC	1.68	**1.67**	1.49	1.92	1.60	1.83

Partition	Model	CDR3α (Å)		CDR3β (Å)		CDR3 α + β (Å)	
		Backbone	Full atom	Backbone	Full atom	Backbone	Full atom
Random	AlphaFold 2	**1.51**	**2.18**	1.81	**2.20**	1.79	2.62
	ESMFold	1.88	3.05	1.89	2.80	1.99	2.84
	LightMHC	1.61	**2.18**	**1.55**	2.36	**1.78**	**2.34**
Sequence	AlphaFold 2	2.56	2.96	2.09	2.37	2.39	2.77
	ESMFold	3.07	4.17	**1.53**	**2.25**	2.88	3.56
	LightMHC	**2.30**	**2.81**	1.86	2.85	**2.05**	**2.80**