

Abstract

The adaptive immune system relies on intracellular peptide presentation to T cells to detect infected or abnormal cells [1]. This process is mediated by the Major Histocompatibility Complex (MHC), molecules that display peptide fragments for immune surveillance [1]. In humans, MHC molecules are referred to as Human Leukocyte Antigens (HLA) [1]. Peptide-MHC prediction supports epitope identification and immunotherapy development [2], [3]. This study presents IncleMHC v2, a novel multi-allele transformer-based tool for predicting binding affinity (BA) and eluted-ligand (EL) presentation of longer peptides with MHC Class I alleles. IncleMHC v2 learns sequence length-dependent patterns across multiple MHC alleles and across peptide lengths. Five-fold cross-validation produced a mean PCC of 0.806 ± 0.005 and SCC of 0.746 ± 0.004 for BA predictions. Length-stratified analysis showed BA PCC values of 0.853, 0.801, 0.783, 0.770, and 0.713 for 8-, 9-, 10-, 11- and 12-mers. Performance noticeably decreased to PCCs of 0.443 and 0.540 for 13- and 14-mers. In contrast, EL prediction achieved an AUC of 0.994 across all peptide lengths. Per-allele EL analysis across 29 alleles produced a mean AUC of 0.995. Evaluation included 175 test alleles, including 19 not observed during training, supporting assessment across allele diversity. The dataset and five-fold cross-validation framework were adapted from NetMHCpan 4.2. These results suggest, but do not establish, stronger prediction performance for longer peptides (11–12-mers) than published NetMHCpan 4.2 results, while performance remained comparable across shorter lengths.

Key Terms

Longer peptide: > 9-mers long

MHC: a molecule that shows peptides on a cell's surface.

Binding (BA): how tightly a peptide sticks to MHC.

Presentation (EL): whether a cell actually displays the peptide.

Transformer: an AI that learns how parts of a sequence relate.

PCC and AUC: performance scores; higher is better.

IncleMHC v2: Predicting MHC *Binding* Affinity and Eluted Ligand *Presentation* of Longer Peptides Using a Novel Transformer-Based Framework

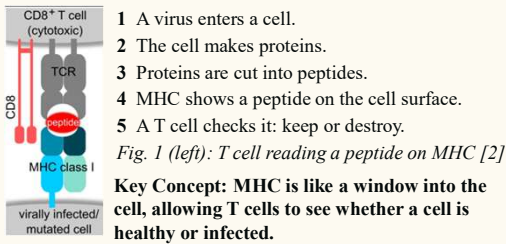
Gagan Girish | Camas High School , Grade 10

Introduction

Cells show small protein pieces, called peptides, on a molecule named MHC. T cells check these pieces to spot infected or abnormal cells [1]. A stronger, more stable bond with MHC gives a clearer signal and a better immune response. Lab tests are slow and costly, and current software is less accurate for peptides longer than 9 amino acids [3].

Research Question

Can a transformer predict MHC binding and presentation of longer peptides better than NetMHCpan and MHCflurry?



Methods

Data: About 1.8M labeled examples across many MHC types [4]: 214K for binding strength (BA) and 1.6M for presentation (EL).

Processing: Peptide and MHC sequences were turned into number grids (matrices) that the AI learns from [5].

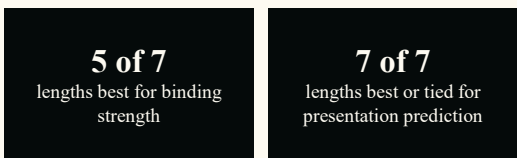
Model: A multi-allele transformer learns how peptide, MHC, binding, and presentation are linked.

Testing: 555 held-out examples, compared with NetMHCpan 4.2 [6] and MHCflurry [7].

Peptide	MHC	BA	EL
EVMPVSMK	HLA-A*03:01	0.431	1
EVMPVSMK	HLA-A*11:01	0.728	1
KTFPPTEPK	HLA-A*03:01	0.517	1

Table 1. Example training rows (BA: binding strength; EL: presentation, 1 = presented).

Results



IncleMHC v2 scored highest on longer peptides (10–14 amino acids). NetMHCpan 4.2 did better at lengths 8 and 9 [6].

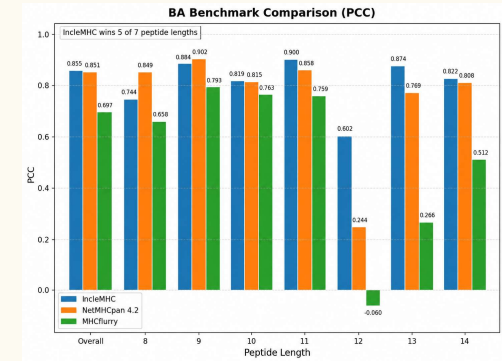


Fig. 2. Binding strength (PCC); higher is better.

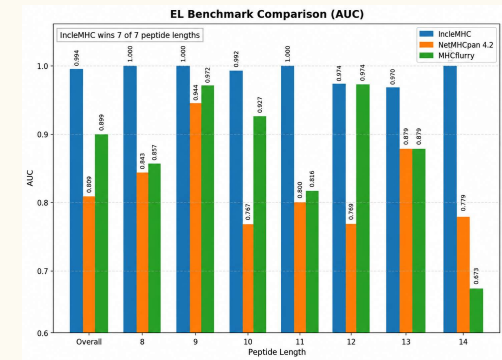


Fig. 3. Presentation (AUC); higher is better.

Discussion

IncleMHC v2 scored higher than both tools for longer peptides (10–14 amino acids), the gap it was built to fill. NetMHCpan 4.2 stayed ahead at lengths 8 and 9, which it predicts very well [6]. Limits: the test set is small (555 examples) and there is no real-world case study yet.

Length	Binding	Presentation
8, 9	NetMHCpan 4.2	IncleMHC v2
10, 11, 13, 14	IncleMHC v2	IncleMHC v2
12	IncleMHC v2	Tie (MHCflurry)

Table 2. Best tool at each peptide length.

Conclusion

Answer
Yes, on this test.

IncleMHC v2 scored higher than NetMHCpan and MHCflurry on binding for peptides of 10–14 amino acids, and was best or tied on presentation at every length.

Future Work: test against more tools, run a real-world case study, and extend IncleMHC to MHC Class II.

References

[1] J. Neefjes et al., "Towards a systems understanding of MHC class I and MHC class II antigen presentation," Nat. Rev. Immunol., vol. 11, no. 12, pp. 823–836, 2011.
[2] Assay Genie, MHC class I / T cell receptor illustration (Fig. 1) [add page title and URL].
[3] M. Nielsen and M. Andreatta, "NetMHCpan-3.0: improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets," Genome Med., vol. 8, no. 1, p. 33, 2016.
[4] A. Sette and J. Sidney, "Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism," Immunogenetics, vol. 50, no. 3–4, pp. 201–212, 1999.
[5] A. Elhaggar et al., "ProtTrans: Toward understanding the language of life through self-supervised learning," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 10, pp. 7112–7127, 2022.
[6] B. Reynisson et al., "NetMHCpan-4.1 and NetMHCpan-4.0: Improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data," Nucleic Acids Res., vol. 48, no. W1, pp. W449–W454, 2020.
[7] T. J. O'Donnell, A. Rubinsteyn, and U. Laserson, "MHCflurry 2.0: Improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing," Cell Syst., vol. 11, no. 1, pp. 42–48, 2020. Charts (Figs. 2–3) are student-produced.