A Web Platform for Protein-Protein Interaction Prediction Using Transformers and Transfer Learning Applied to Peptide-MHC Bindings

Vicente Enrique Machaca Arceda
Protein Function Development (Maria-Jesus Martin)
Institut Pasteur

September 24, 2023

Abstract

In this work, we evaluated the use of transfer learning from TAPE, ProtBert-BFD, and ESM2, for PPI prediction of peptides and MHC-I. We will evaluate six pre-trained BERT architectures (four models from ESM-2) followed by BiLSTM layers. Furthermore, we will evaluate our results using Anthem dataset [1]. Finally, we will deploy a Web platform for the prediction of bindings between peptides and MHC-I.

Keywords: AI and machine learning, Bioinformatics, Software development, and Proteomics

1 Background, proposed project and its implementation

1.1 Introduction

Protein–protein interactions (PPI) are relevant mediator in biological processes; understanding them is beneficial as it enables us to comprehend the functions of proteins, the origins and progression of various illnesses, and can assist in the development of new drugs [2, 3]. Additionally, the human genome codes approximately 500000 proteins and 130000 to 650000 PPIs occurs in human body [2]. *In vivo* and *in vitro*, methods like biomolecular fluorescent complementary (BiFC), chromatography, and nuclear magnetic resonance (NMR) have been developed; however, they are time-consuming and labor-intensive [2, 4]. Consequently, *in silico* methods emerged as an alternative.

Moreover, in immunology, bindings between peptides and Major Histocompatibility Complex (MHC) represent a key factor for activating an immune respond. MHC class I (MHC-I) and MHC class II (MHC-II) present peptides at the cell surface to CD8+ and CD4+ T Cells, respectively [5, 6]. Lamentably, MHC proteins are encoded by highly polymorphic genes, called Human Leukocytes Antigens or (HLAs); the considerable polymorphic nature of MHC genes affords substantial variation in peptide binding, thereby influencing the set of peptides presented to T cells. [6]. In consequence, proposals methods are categorized as allele-specific or pan-specific. Allele-specific methods [7–13], train a model for each MHC allele; meanwhile, pan-specific methods [1, 14–24] train a global model taking peptides and MHC as inputs. Therefore, due to the highly polymorphic nature of MHC, pan-specific methods arise with high possibility of future applications. Additionally, immunotherapy is considered a promising approach to cancer treatment, especially since traditional methods based on surgeries, radiotherapies, and chemotherapies have low effectiveness [25, 26]. This strategy capitalizes on the observation that cancer cells generate distinctive neoepitopes recognized by the MHC [27]. Furthermore, these neoepitopes or neoantigens are considered the leading causes of an immune response [28–30].

Recently, the advent of Transformers has ushered in a new era in artificial intelligence, demonstrating significant success across various Natural Language Processing (NLP) tasks [31]. These models have also found application in neoantigen detection, particularly in predicting pMHC binding and presentation. For example, BERTMHC [32] is a pan-specific pMHC-II binding and presentation prediction

method that employs a BERT architecture and leverages transfer learning from the Tasks Assessing Protein Embeddings (TAPE) [33]. The methodology involves stacking an average pooling layer followed by a Fully Connected (FC) layer after the TAPE model. Empirical assessments have shown that BERTMHC outperforms both NetMHCIIpan3.2 and PUFFIN. Additionally, ImmunoBERT [34] utilizes transfer learning from TAPE but focuses on pMHC-I prediction. This approach involves stacking a classification token's vector after the TAPE model. Furthermore, MHCRoBERTa [35] and HLAB [24] also leverage transfer learning. MHCRoBERTa employs self-supervised training with data from UniProtKB and Swiss-Prot databases, followed by fine-tuning with data from the Immune Epitope Database (IEDB) [36]. MHCRoBERTa performs better than NetMHCpan4.0 and MHCflurry2.0 in terms of Spearman Rank Correlation Coefficient (SRCC). In contrast, HLAB leverages transfer learning from ProtBert-BFD [37] and incorporates a BiLSTM model in cascade. Notably, on the HLA-A*01:01 allele, HLAB demonstrates a slight performance advantage over state-of-the-art methods, including NetMHCpan4.1, with at least a 0.0230 improvement in Area Under the Curve (AUC) and a 0.0560 increase in accuracy.

In this work, we evaluated the use of transfer learning from TAPE, ProtBert-BFD, and ESM2, for PPI prediction of peptides and MHC-I. We will evaluate six pre-trained BERT architectures (four models from ESM-2) followed by BiLSTM layers. Furthermore, we will evaluate our results using Anthem dataset [1]. Finally, we will deploy a Web platform for the prediction of bindings between peptides and MHC-I.

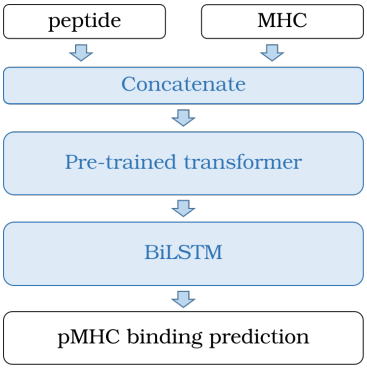
2 Proposal and methodology

We propose the development of a Web platform for PPI prediction of peptides and MHC. This project is based on previous works, where we performed a review of peptide-MHC interactions [38], and implemented a model using transformers and transfer learning for peptide-MHC bindings [39]. Consequently, we will use Transformers and transfer learning from BERT models pre-trained on large protein datasets. These pre-trained models are TAPE [33], ProtBert-BFD [37] and ESM-2 [40]; furthermore, in Table 1, we present the major difference between these models.

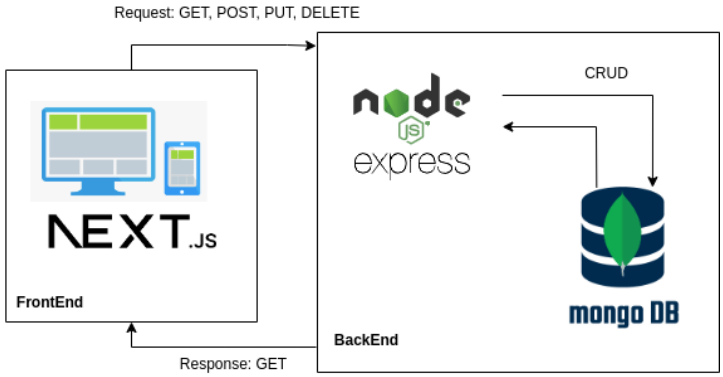
Table 1: Major differences between TAPE, ProtBert-DFB, and ESM-2.

Model	**Dataset**	**Samples**	**Layers**	**Hidden size**	**Att. heads**	**Params.**
TAPE	Pfam	30M	12	768	12	92M
ProtBert-BFD	BFD	2122M	30	1024	16	420M
ESM-2 (6 layers)	Uniref50	60M	6	320	20	8M
ESM-2 (12 layers)	Uniref50	60M	12	480	20	35M
ESM-2 (30 layers)	Uniref50	60M	30	640	20	150M
ESM-2 (33 layers)	Uniref50	60M	33	1280	20	650M

For fine-tuning, we will stack in cascade a BiLSTM at the end of the pre-trained model. The BiLSTM is based on HLAB [24] and has two layers with 768 units. In Figure. 1a, we present our proposal. This model takes the aminoacid sequences of a peptide and the MHC; then these sequences are concatenated and encoded using one-hot; then it feed-forward the pre-trained transformers and the BiLSTM model; finally, we will predict 1 for physical binding and 0 for no binding. Furthermore, we will use Anthem dataset [1] for fine-tuning.



(a) Transformer model.



(b) Web platform architecture.

Figure 1: (a) The proposed model for PPI prediction of peptides and MHC. (b) The Web architecture.

3 Infrastructure

4 Limitations

5 Expected results and their impact

6 Ethics

7 Gantt chart

References

[1] Shutao Mei, Fuyi Li, Dongxu Xiang, Rochelle Ayala, Pouya Faridi, Geoffrey I Webb, Patricia T Illing, Jamie Rossjohn, Tatsuya Akutsu, Nathan P Croft, et al., "Anthem: a user customised tool for fast and accurate prediction of binding between peptides and hla class i molecules," *Briefings in Bioinformatics*, vol. 22, no. 5, pp. bbaa415, 2021.

[2] Xiaotian Hu, Cong Feng, Tianyi Ling, and Ming Chen, "Deep learning frameworks for protein–protein interaction prediction," *Computational and Structural Biotechnology Journal*, vol. 20, pp. 3223–3233, 2022.

[3] Kanchan Jha, Sriparna Saha, and Sourav Karmakar, "Prediction of protein-protein interactions using vision transformer and language model," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2023.

[4] V Srinivasa Rao, K Srinivas, GN Sujini, and GN Kumar, "Protein-protein interaction detection: methods and analysis," *International journal of proteomics*, vol. 2014, 2014.

[5] Charles A Janeway Jr, "Immunobiology the immune system in health and disease," *Artes Medicas*, 1997.

[6] Esam T Abualrous, Jana Sticht, and Christian Freund, "Major histocompatibility complex (mhc) class i and class ii proteins: impact of polymorphism on antigen presentation," *Current Opinion in Immunology*, vol. 70, pp. 95–104, 2021.

[7] H-G Rammensee, Jutta Bachmann, Niels Philipp Nikolaus Emmerich, Oskar Alexander Bachor, and SSYFPEITHI Stevanović, "Syfpeithi: database for mhc ligands and peptide motifs," *Immunogenetics*, vol. 50, pp. 213–219, 1999.

[8] Pedro A Reche, John-Paul Glutting, and Ellis L Reinherz, "Prediction of mhc class i binding peptides using profile motifs," *Human immunology*, vol. 63, no. 9, pp. 701–709, 2002.

[9] Yohan Kim, John Sidney, Clemencia Pinilla, Alessandro Sette, and Bjoern Peters, "Derivation of an amino acid similarity matrix for peptide: Mhc binding and its application as a bayesian prior," *BMC bioinformatics*, vol. 10, pp. 1–11, 2009.

[10] Morten Nielsen and Massimo Andreatta, "Netmhcpan-3.0; improved prediction of binding to mhc class i molecules integrating information from multiple receptor and peptide length datasets," *Genome medicine*, vol. 8, no. 1, pp. 1–9, 2016.

[11] Yeeleng S Vang and Xiaohui Xie, "Hla class i binding prediction via convolutional neural networks," *Bioinformatics*, vol. 33, no. 17, pp. 2658–2665, 2017.

[12] Xiaoshan M Shao, Rohit Bhattacharya, Justin Huang, IK Sivakumar, Collin Tokheim, Lily Zheng, Dylan Hirsch, Benjamin Kaminow, Ashton Omdahl, Maria Bonsack, et al., "High-throughput prediction of mhc class i and ii neoantigens with mhcnuggetshigh-throughput prediction of neoantigens with mhcnuggets," *Cancer immunology research*, vol. 8, no. 3, pp. 396–408, 2020.

[13] Barbara Bravi, Jérôme Tubiana, Simona Cocco, Remi Monasson, Thierry Mora, and Aleksandra M Walczak, "Rbm-mhc: a semi-supervised machine-learning method for sample-specific prediction of antigen presentation by hla-i alleles," *Cell systems*, vol. 12, no. 2, pp. 195–202, 2021.

[14] Yan Hu, Ziqiang Wang, Hailin Hu, Fangping Wan, Lin Chen, Yuanpeng Xiong, Xiaoxia Wang, Dan Zhao, Weiren Huang, and Jianyang Zeng, "Acme: pan-specific peptide–mhc class i binding prediction through attention-based deep neural networks," *Bioinformatics*, vol. 35, no. 23, pp. 4946–4954, 2019.

[15] Zhonghao Liu, Yuxin Cui, Zheng Xiong, Alierza Nasiri, Ansi Zhang, and Jianjun Hu, "Deepseqpan, a novel deep convolutional neural network model for pan-specific class i hla-peptide binding affinity prediction," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[16] Jingcheng Wu, Wenzhe Wang, Jiucheng Zhang, Binbin Zhou, Wenyi Zhao, Zhixi Su, Xun Gu, Jian Wu, Zhan Zhou, and Shuqing Chen, "Deephlapan: a deep learning approach for neoantigen prediction considering both hla-peptide binding and immunogenicity," *Frontiers in Immunology*, p. 2559, 2019.

[17] Poomarin Phloyphisut, Natapol Pornputtapong, Sira Sriswasdi, and Ekapol Chuangsuwanich, "Mhcseqnet: a deep neural network model for universal mhc binding prediction," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–10, 2019.

[18] Timothy J O'Donnell, Alex Rubinsteyn, Maria Bonsack, Angelika B Riemer, Uri Laserson, and Jeff Hammerbacher, "Mhcflurry: open-source class i mhc binding affinity prediction," *Cell systems*, vol. 7, no. 1, pp. 129–132, 2018.

[19] Timothy J Donnell, Alex Rubinsteyn, and Uri Laserson, "Mhcflurry 2.0: improved pan-allele prediction of mhc class i-presented peptides by incorporating antigen processing," *Cell systems*, vol. 11, no. 1, pp. 42–48, 2020.

[20] Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen, "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*, vol. 48, no. W1, pp. W449–W454, 2020.

[21] Gopalakrishnan Venkatesh, Aayush Grover, G Srinivasaraghavan, and Shrisha Rao, "Mhcattnnet: predicting mhc-peptide bindings for mhc alleles classes i and ii using an attention-based deep neural model," *Bioinformatics*, vol. 36, no. Supplement_1, pp. i399–i406, 2020.

[22] Yilin Ye, Jian Wang, Yunwan Xu, Yi Wang, Youdong Pan, Qi Song, Xing Liu, and Ji Wan, "Mathla: a robust framework for hla-peptide binding prediction integrating bidirectional lstm and multiple head attention mechanism," *BMC bioinformatics*, vol. 22, no. 1, pp. 1–12, 2021.

[23] Yanyi Chu, Yan Zhang, Qiankun Wang, Lingfeng Zhang, Xuhong Wang, Yanjing Wang, Dennis Russell Salahub, Qin Xu, Jianmin Wang, Xue Jiang, et al., "A transformer-based model to predict peptide–hla class i binding and optimize mutated peptides for vaccine design," *Nature Machine Intelligence*, vol. 4, no. 3, pp. 300–311, 2022.

[24] Yaqi Zhang, Gancheng Zhu, Kewei Li, Fei Li, Lan Huang, Meiyu Duan, and Fengfeng Zhou, "Hlab: learning the bilstm features from the protbert-encoded proteins for the class i hla-peptide binding prediction," *Briefings in Bioinformatics*, 2022.

[25] Miao Peng, Yongzhen Mo, Yian Wang, Pan Wu, Yijie Zhang, Fang Xiong, Can Guo, Xu Wu, Yong Li, Xiaoling Li, et al., "Neoantigen vaccine: an emerging tumor immunotherapy," *Molecular cancer*, vol. 18, no. 1, pp. 1–14, 2019.

[26] Akshita Thakur, Akanksha Sharma, Hema K Alajangi, Pradeep Kumar Jaiswal, Yong-beom Lim, Gurpal Singh, and Ravi Pratap Barnwal, "In pursuit of next-generation therapeutics: Antimicrobial peptides against superbugs, their sources, mechanism of action, nanotechnology-based delivery, and clinical applications," *International Journal of Biological Macromolecules*, 2022.

[27] Aurélie Durgeau, Yasemin Virk, Stéphanie Corgnac, and Fathia Mami-Chouaib, "Recent advances in targeting cd8 t-cell immunity for more effective cancer immunotherapy," *Frontiers in immunology*, vol. 9, pp. 14, 2018.

[28] Elizabeth S Borden, Kenneth H Buetow, Melissa A Wilson, and Karen Taraszka Hastings, "Cancer neoantigens: Challenges and future directions for prediction, prioritization, and validation," *Frontiers in Oncology*, vol. 12, 2022.

[29] Ina Chen, Michael Chen, Peter Goedegebuure, and William Gillanders, "Challenges targeting cancer neoantigens in 2021: a systematic literature review," *Expert Review of Vaccines*, vol. 20, no. 7, pp. 827–837, 2021.

[30] Alexander V Gopanenko, Ekaterina N Kosobokova, and Vyacheslav S Kosorukov, "Main strategies for the identification of neoantigens," *Cancers*, vol. 12, no. 10, pp. 2879, 2020.

[31] Narendra Patwardhan, Stefano Marrone, and Carlo Sansone, "Transformers in the real world: A survey on nlp applications," *Information*, vol. 14, no. 4, pp. 242, 2023.

[32] Jun Cheng, Kaïdre Bendjama, Karola Rittner, and Brandon Malone, "Bertmhc: improved mhc–peptide class ii interaction prediction with transformer and multiple instance learning," *Bioinformatics*, vol. 37, no. 22, pp. 4172–4179, 2021.

[33] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song, "Evaluating protein transfer learning with tape," *Advances in neural information processing systems*, vol. 32, 2019.

[34] Hans-Christof Gasser, Georges Bedran, Bo Ren, David Goodlett, Javier Alfaro, and Ajitha Rajan, "Interpreting bert architecture predictions for peptide presentation by mhc class i proteins," *arXiv preprint arXiv:2111.07137*, 2021.

[35] Fuxu Wang, Haoyan Wang, Lizhuang Wang, Haoyu Lu, Shizheng Qiu, Tianyi Zang, Xinjun Zhang, and Yang Hu, "Mhcroberta: pan-specific peptide–mhc class i binding prediction through transfer learning with label-agnostic protein sequences," *Briefings in Bioinformatics*, vol. 23, no. 3, pp. bbab595, 2022.

[36] Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters, "The immune epitope database (iedb): 2018 update," *Nucleic acids research*, vol. 47, no. D1, pp. D339–D343, 2018.

[37] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al., "Prottrans: Toward understanding the language of life through self-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 10, pp. 7112–7127, 2021.

[38] Vicente Enrique Machaca, Valeria Goyzueta, Maria Cruz, and Yvan Tupac, "Deep learning and transformers in mhc-peptide binding and presentation towards personalized vaccines in cancer immunology: A brief review," in *International Conference on Practical Applications of Computational Biology & Bioinformatics*. Springer, 2023, pp. 14–23.

[39] Vicente Enrique Machaca Arceda, "Neoantigen detection using transformers and transfer learning in the cancer immunology context," in *International Conference on Practical Applications of Computational Biology & Bioinformatics*. Springer, 2023, pp. 97–102.

[40] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al., "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.