Universidad La Salle

Deep Learning and Transformers in MHC-peptide Binding and Presentation Towards Personalized Vaccines in Cancer Immunology: A Brief Review

MSc. Vicente Machaca Arceda

2023

Content

Content

Immunotherapy to Treat Cancer

Immunotherapy is a type of cancer treatment that helps your immune system fight cancer [1].

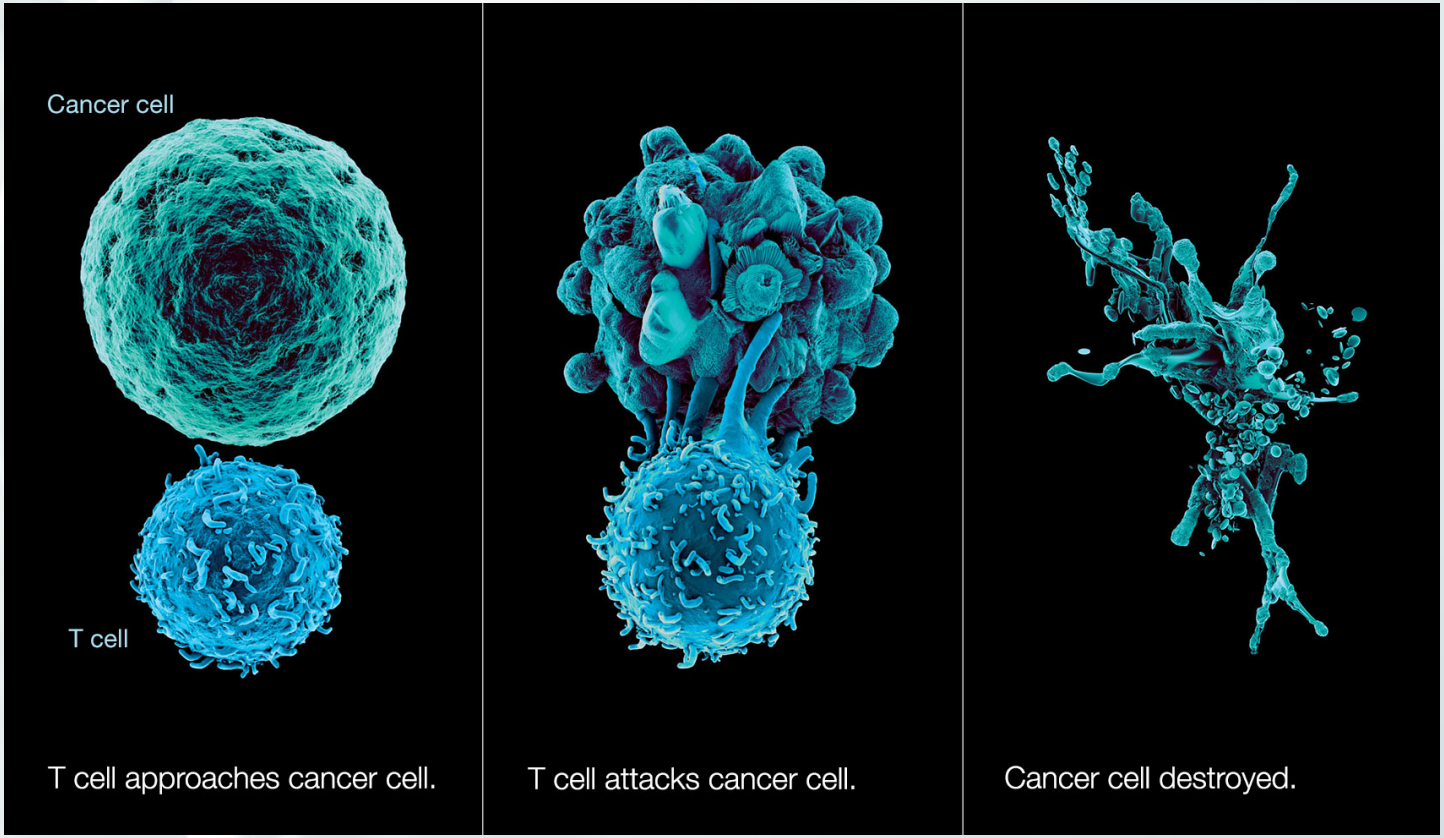

Figure: Example of how a T cell attack a cancer cell [2].

Neoantigen

A new protein that forms on cancer cells when certain mutations occur in tumor DNA. Neoantigens used in vaccines and other types of immunotherapy are being studied in the treatment of many types of cancer [3, 4].

Currently, there is a lot of methods to detect neoantigens; however, only a small number of them manage to stimulate the immune system [5, 6].

Immunotherapy for Cancer
Personalized Vaccines

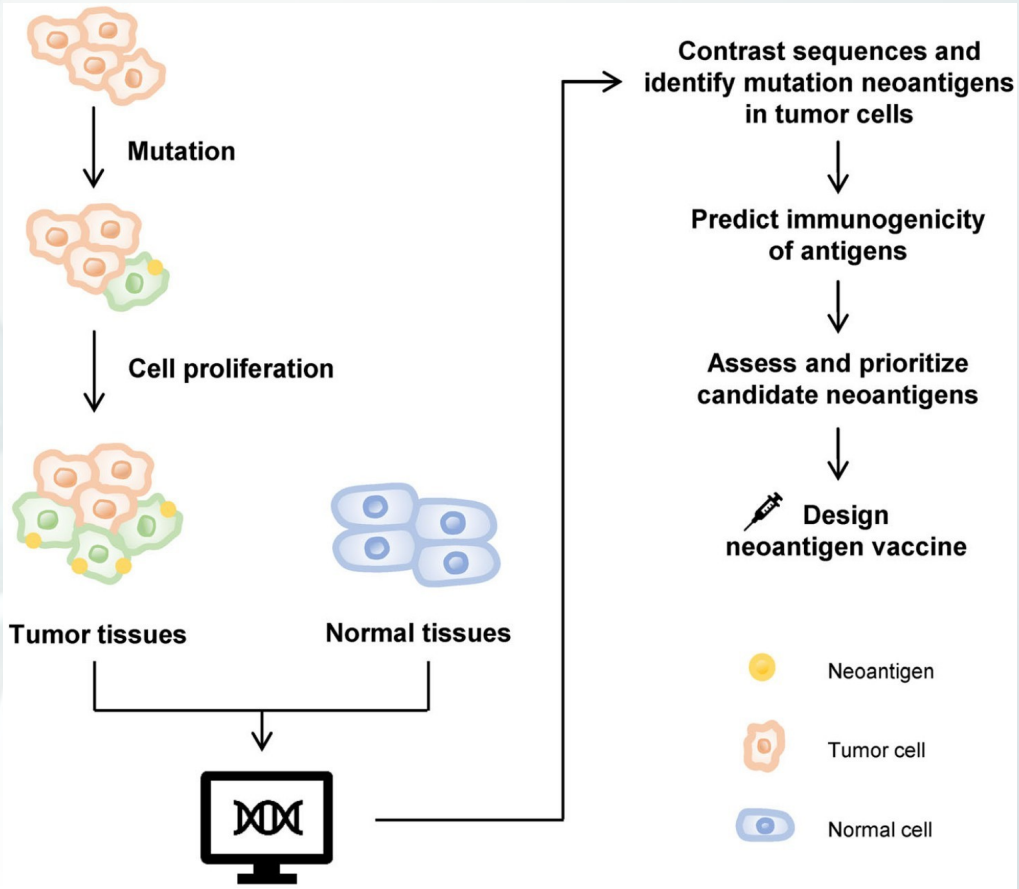

Figure: Personalized vaccines process for Cancer [7].

pMHC binding and presentation prediction

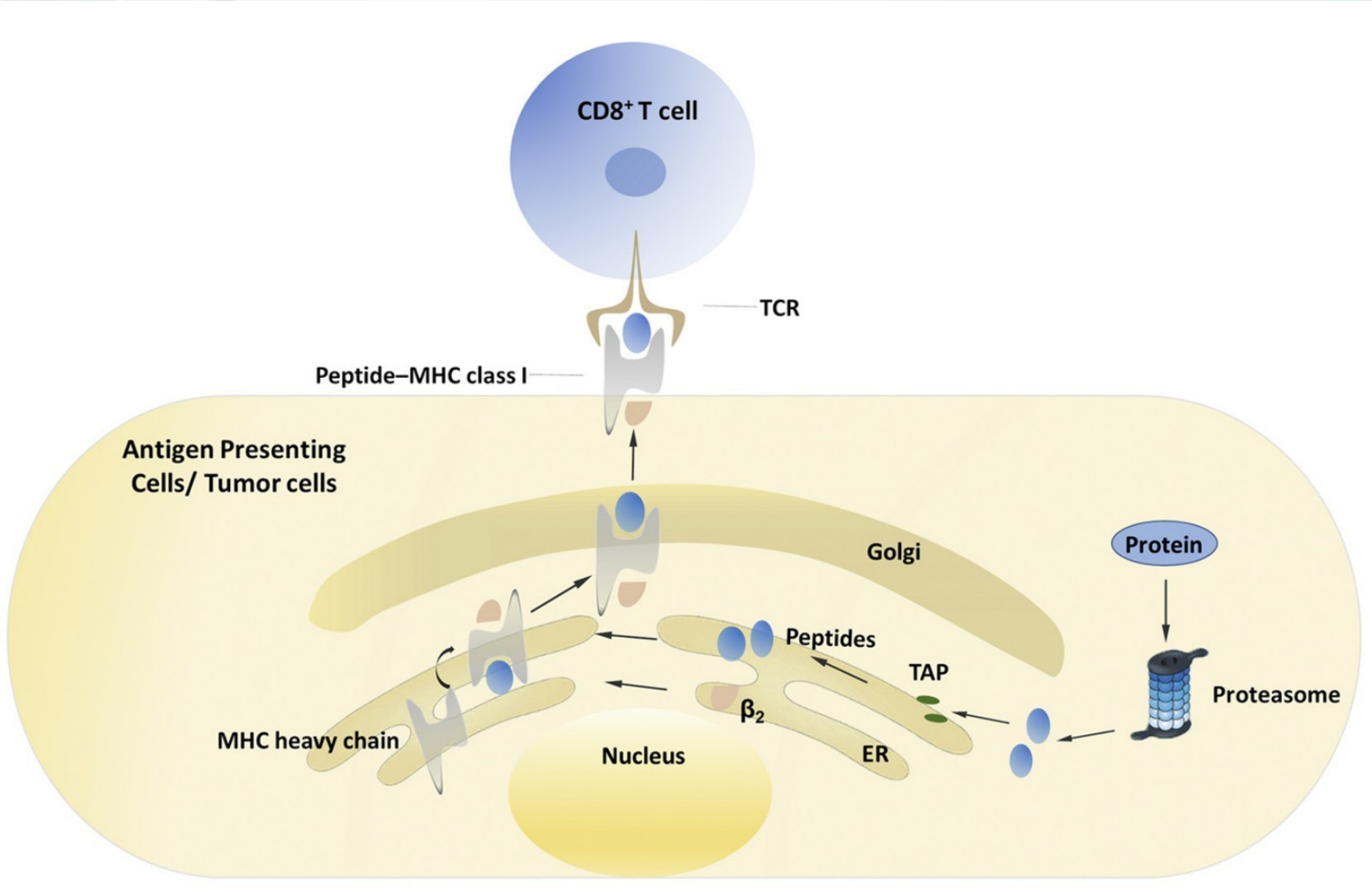

Figure: pMHC presentation process in MHC class I [8].

Content

Research Questions

Table: Research questions used in SLR.

Research questions
Q1. How deep learning and transformers are used in MHC-peptide binding and presentation prediction?
Q2. What type of input data and pre-processing methods are used?
Q3. Which are the most promising methods?

Content

Search Strings

We searched in IEEE Xplore, Science Direct, Springer, ACM Digital Library, PubMed, and Scopus.

Table: Search string used in SLR.

Search Strings
(MHC-I OR MHC-II OR MHC OR HLA) AND (peptide OR epitope) AND (binding OR affinity OR prediction OR detection OR presentation) OR (neoantigen detection)

Content

Inclusion Criteria

Conference papers with ERA category (A or B) or journal articles Q1/Q2. Moreover, papers published since 2018.

Results

We analyzed papers' titles, and then a small subset of **54 papers were selected**.

Table: Number of papers found in databases according to search string.

Year	Research papers
2018	46
2019	72
2020	86
2021	61
2022	58
Total	**323**

Content

Input Encoding

This is a **binary classification problem**. A peptide could be represented like: $p = \{A, ..., Q\}$ and a MHC like: $q = \{A, N, ..., Q, E\}$. Finally, we need to know the probability of affinity between p and q (pMHC)

Input Encoding
One-hot

A	R	N	...	V
1	0	0		0
0	1	0		0
0	0	1		0
0	0	0		0
0	0	0		0
0	0	0		0
.	.	.		.
.	.	.		.
.	.	.		.
0	0	0		1

Figure: pMHC presentation process in MHC class I [8].

Input Encoding
BLOSUM

Table: BLOSUM62 matrix. Normally, it is used to represent amino acids numerically. Each amino acid is represented by a row.

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	0																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	-1	0	-2	-2	0	6													
D	-3	1	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	0	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	-1	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	0	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-1	-4	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1-	-3	-2	-3	1	2	11

Input Encoding
Another Alternatives

Also, there are another alternatives like: universal **Google encoder** [10], **AAindex** [11, 12] (a database of numerical indices representing physicochemical and biochemical properties of amino acids), **3D amino acid coordinates** [13], and **physicochemical properties** of each amino acid [14, 15, 16].

Content

DANN, CNN, and RNN

Currently, the state-of-art method is **NetMHCPan4.1** [17], it is a DANN of 40 assembled ANNs with 60 and 70 neurons. Additionally, it uses NN_alignMA [18] to handle poly-specific datasets of Mass Spectrometry (MS)-eluted ligands.

DANN, CNN, and RNN

Convolutional Neural Networks

Table: List of research since 2018 that uses CNNs for peptide-MHC binding and presentation.

Year	Ref.	Approach	Name	MHC	Encoding
2022	[19]	pMHC(b)	DeepMHCII	II	PFR
2021	[12]	pMHC(b)	DeepImmuno	I	AAindex1
2021	[20]	pMHC(p)	APPM	I	One-hot
2021	[21]	pMHC(p)	MHCfovea	I	One-hot
2021	[22]	pMHC(b)	CNN-PepPred	II	BLOSUM
2020	[23]	pMHC(b)	IConMHC	I	PCA and AAindex3
2020	[24]	pMHC(b)	OnionMHC	I	BLOSUM and structural features
2020	[25]	pMHC(p)	MINERVA	I	Physicochemical properties
2019	[26]	pMHC(b)	CNN-NF	I	Sequence, Hydropathy, Polarity, Length
2019	[27]	pMHC(b)	DeepSeqPan	I	One-hot
2018	[28]	pMHC(b)	ConvMHC	I	Contact side HLA.peptide

DANN, CNN, and RNN
Recurrent Neural Networks

Table: List of research since 2018 that uses RNNs for peptide-MHC binding and presentation. MATHLA, DeepSeqPanII and DeepHLApan uses RNN with attention mechanims, meanwhile the other focus on GRU and LSTM.

Year	Ref.	Approach	Name	MHC	Encoding
2021	[29]	pMHC(b)	MATHLA	I	BLOSUM
2021	[30]	pMHC(b)	DeepSeqPanII	II	One-hot and BLOSUM
2021	[31]	pMHC(b)	GRU-based RNN	II	Embeding layer
2021	[32]	pMHC(b)	BVLSTM-MHC	I	One-hot and BLOSUM
2020	[33]	pMHC(b)	MHCnuggets	I, II	One-hot
2019	[34]	pMHC(b)	DeepHLApan	I	One-hot

Content

Transformers

The **Transformer Neural Network** is a novel architecture that aims to solve sequence-to-sequence tasks while handling long-range dependencies with ease. It was proposed in the paper “Attention Is All You Need” [35].

Bidirectional Encoder Representations from Transformers (BERT) [36] is a recent paper published by researchers at Google AI Language. It has caused a stir in the Machine Learning community by presenting state-of-the-art results in a wide variety of NLP tasks,

Table: List of pre-trained BERT models.

Model	Parameters	Layers
TAPE	92M	12
ProtBert	420M	30
ESM1	43M, 85M y 670M	6, 12, and 34
ESM1-b	650M	33
ESM2	8M, 35M, 150M, 650M, 3B, 15B	6, 12, 30, 33, 36, and 48

Transformers

Table: Transformers used for pMHC binding and presentation prediction.

Year	Name	Model
2022[37]	**HLAB**	BERT from ProtBert pre-trained model followed by a BiLSTM with attention mechanism.
2022[38]	MHC RoBERTa	RoBERTa pre-trained and followed by 12 multi-head SA and a FC layers, it outperformed NetMHCPan 3.0.
2022[39]	**TransPHLA**	It used SA mechanism based on four blocks, it slightly outperformed NetMHCpan4.1 and is faster making predictions.
2021[40]	ImmunoBERT	BERT from TAPE pre-trained followed by a linear layer. Authors claimed that N-terminal and C-terminals are highly relevant after analysis with SHAP and LIME.
2021[41]	BERTMHC	BERT from TAPE pre-trained followed by a linear layer. It outperformed NetMHCIIpan3.2 and PUFFIN.

Content

State of art methods

Although **NetMHCpan4.1** is the state-of-art pan method, the transformers overcome exciting results. These BERT models [41, 40, 38], used transfer learning from TAPE [42], and ProtBert [43], which are models self-supervised trained with Pfam, UniRef50, UniRef100, UniProtKB, Swiss-prot and BFD datasets.

Despite, NetMHCpan4.1 [17] is considered the state-of-the-art pan-specific method; **HLAB** [37] and **TransPHLA** [39] slightly outperformed it on allele-specific testing.

Content

Limitations

They ignored Posttranslational modifications (PTMs) such as phosphorylation, glycosylation, and deamidation, which influence the specificity of MHC binding and presentation and several aspects of the biology underlying pMHC presentation are poorly understood. Furthermore, to get accurate results for neoantigen detection, we need to integrate pMHC-TCR studies.

Another limitations are related to high computing requirements for training BERT architectures. For instance, the biggest ESM2 model has 15 billion parameters.

Content

Future works

Future work could include the use of transfer learning from ESM1-b [44] and ESM2 [45].

Moreover, there is pHLA3D, a dataset of 3D structures of the alpha/beta chains and peptides of MHC-I proteins; it opens new perspectives for studying pMHC prediction.

References I

[1] Cancer.net,
"Qué es la inmunoterapia," 2022.

[2] NortShore,
"Immunotherapy," 2022.

[3] NCI,
"National cancer institute dictionary," 2022.

[4] 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.

References II

[5] 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.

[6] Qing Hao, Ping Wei, Yang Shu, Yi-Guan Zhang, Heng Xu, and Jun-Ning Zhao,
"Improvement of neoantigen identification through convolution neural network,"
Frontiers in immunology, vol. 12, 2021.

[7] 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.

References III

[8] Xiaomei Zhang, Yue Qi, Qi Zhang, and Wei Liu,
"Application of mass spectrometry-based mhc immunopeptidome profiling in neoantigen identification for tumor immunotherapy,"
Biomedicine & Pharmacotherapy, vol. 120, pp. 109542, 2019.

[9] L Mattos, M Vazquez, F Finotello, R Lepore, E Porta, J Hundal, P Amengual-Rigo, CKY Ng, A Valencia, J Carrillo, et al.,
"Neoantigen prediction and computational perspectives towards clinical benefit: recommendations from the esmo precision medicine working group,"
Annals of oncology, vol. 31, no. 8, pp. 978–990, 2020.

[10] Norwin Kubick and Michel Edwar Mickael,
"Predicting epitopes based on tcr sequence using an embedding deep neural network artificial intelligence approach,"
bioRxiv, 2021.

References IV

[11] Shuichi Kawashima and Minoru Kanehisa,
"Aaindex: amino acid index database,"
Nucleic acids research, vol. 28, no. 1, pp. 374–374, 2000.

[12] Guangyuan Li, Balaji Iyer, VB Surya Prasath, Yizhao Ni, and Nathan Salomonis,
"Deepimmuno: deep learning-empowered prediction and generation of immunogenic peptides for t-cell immunity,"
Briefings in bioinformatics, vol. 22, no. 6, pp. bbab160, 2021.

[13] Yi Shi, Zehua Guo, Xianbin Su, Luming Meng, Mingxuan Zhang, Jing Sun, Chao Wu, Minhua Zheng, Xueyin Shang, Xin Zou, et al.,
"Deepantigen: a novel method for neoantigen prioritization via 3d genome and deep sparse learning,"
Bioinformatics, vol. 36, no. 19, pp. 4894–4901, 2020.

References V

[14] Pieter Moris, Joey De Pauw, Anna Postovskaya, Sofie Gielis, Nicolas De Neuter, Wout Bittremieux, Benson Ogunjimi, Kris Laukens, and Pieter Meysman,
"Current challenges for unseen-epitope tcr interaction prediction and a new perspective derived from image classification,"
Briefings in Bioinformatics, vol. 22, no. 4, pp. bbaa318, 2021.

[15] 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.,
"Nettcr-2.0 enables accurate prediction of tcr-peptide binding by using paired tcrα and β sequence data,"
Communications biology, vol. 4, no. 1, pp. 1–13, 2021.

References VI

[16] Alan M Luu, Jacob R Leistico, Tim Miller, Somang Kim, and Jun S Song,
"Predicting tcr-epitope binding specificity using deep metric learning and multimodal learning,"
Genes, vol. 12, no. 4, pp. 572, 2021.

[17] 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.

References VII

[18] Bruno Alvarez, Birkir Reynisson, Carolina Barra, Søren Buus, Nicola Ternette, Tim Connelley, Massimo Andreatta, and Morten Nielsen,
"Nnalign_ma; mhc peptidome deconvolution for accurate mhc binding motif characterization and improved t-cell epitope predictions,"
Molecular & Cellular Proteomics, vol. 18, no. 12, pp. 2459–2477, 2019.

[19] Ronghui You, Wei Qu, Hiroshi Mamitsuka, and Shanfeng Zhu,
"Deepmhcii: a novel binding core-aware deep interaction model for accurate mhc-ii peptide binding affinity prediction,"
Bioinformatics, vol. 38, no. Supplement_1, pp. i220–i228, 2022.

References VIII

[20] Franziska Lang, Pablo Riesgo-Ferreiro, Martin L¨ower, Ugur Sahin, and Barbara Schr¨ors,
"Neofox: annotating neoantigen candidates with neoantigen features,"
Bioinformatics, vol. 37, no. 22, pp. 4246–4247, 2021.

[21] Ko-Han Lee, Yu-Chuan Chang, Ting-Fu Chen, Hsueh-Fen Juan, Huai-Kuang Tsai, and Chien-Yu Chen,
"Connecting mhc-i-binding motifs with hla alleles via deep learning,"
Communications Biology, vol. 4, no. 1, pp. 1–12, 2021.

[22] Valentin Junet and Xavier Daura,
"Cnn-peppred: an open-source tool to create convolutional nn models for the discovery of patterns in peptide sets—application to peptide–mhc class ii binding prediction,"
Bioinformatics, vol. 37, no. 23, pp. 4567–4568, 2021.

References IX

[23] Baikang Pei and Yi-Hsiang Hsu,
"Iconmhc: a deep learning convolutional neural network model to predict peptide and mhc-i binding affinity,"
Immunogenetics, vol. 72, no. 5, pp. 295–304, 2020.

[24] Shikhar Saxena, Sambhavi Animesh, Melissa J Fullwood, and Yuguang Mu,
"Onionmhc: A deep learning model for peptide—hla-a* 02: 01 binding predictions using both structure and sequence feature sets,"
Journal of Micromechanics and Molecular Physics, vol. 5, no. 03, pp. 2050009, 2020.

References X

[25] Felicia SL Ng, Michel Vandenberghe, Guillem Portella, Corinne Cayatte, Xiaotao Qu, Shino Hanabuchi, Aimee Landry, Raghothama Chaerkady, Wen Yu, Rosana Collepardo-Guevara, et al.,
"Minerva: Learning the rules of hla class i peptide presentation in tumors with convolutional neural networks and transfer learning,"
Available at SSRN 3704016, 2020.

[26] Tianyi Zhao, Liang Cheng, Tianyi Zang, and Yang Hu,
"Peptide-major histocompatibility complex class i binding prediction based on deep learning with novel feature,"
Frontiers in Genetics, vol. 10, pp. 1191, 2019.

References XI

[27] 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.

[28] Youngmahn Han,
"Deep convolutional neural networks for peptide-mhc binding predictions," 2018.

[29] 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.

References XII

[30] Zhonghao Liu, Jing Jin, Yuxin Cui, Zheng Xiong, Alireza Nasiri, Yong Zhao, and Jianjun Hu,
"Deepseqpanii: an interpretable recurrent neural network model with attention mechanism for peptide-hla class ii binding prediction,"
IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021.

[31] Yu Heng, Zuyin Kuang, Wenzhao Xie, Haoqi Lan, Shuheng Huang, Linxin Chen, Tingting Shi, Lei Xu, Xianchao Pan, and Hu Mei,
"A simple pan-specific rnn model for predicting hla-ii binding peptides,"
Molecular Immunology, vol. 139, pp. 177–183, 2021.

References XIII

[32] Limin Jiang, Hui Yu, Jiawei Li, Jijun Tang, Yan Guo, and Fei Guo, "Predicting mhc class i binder: existing approaches and a novel recurrent neural network solution," *Briefings in Bioinformatics*, vol. 22, no. 6, pp. bbab216, 2021.

[33] 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.

References XIV

[34] 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.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin,
"Attention is all you need,"
Advances in neural information processing systems, vol. 30, 2017.

References XV

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova,
"Bert: Pre-training of deep bidirectional transformers for language understanding,"
arXiv preprint arXiv:1810.04805, 2018.

[37] 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.

[38] 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.

References XVI

[39] 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.

[40] 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.

References XVII

[41] 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.

[42] 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.

References XVIII

[43] 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.

[44] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al.,
"Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences,"
Proceedings of the National Academy of Sciences, vol. 118, no. 15, 2021.

References XIX

[45] 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.