LaSalle Universidad
DIRECCIÓN DE INVESTIGACIÓN

Transformers meets neoantigen detection: A systematics literature review

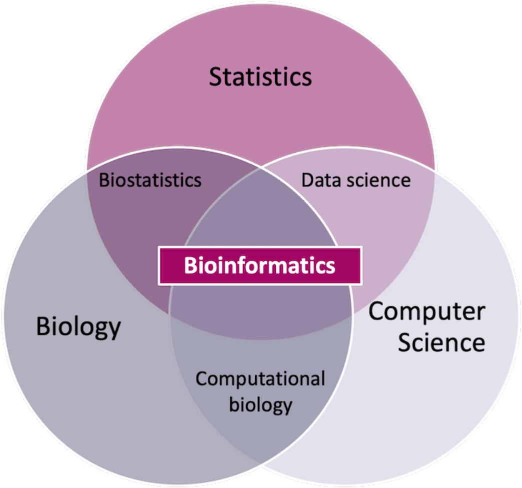
Jornadas de Investigación 2024

Ph.D. Vicente Machaca Arceda

Contenido

Content

Bioinformática



DNA
Localización

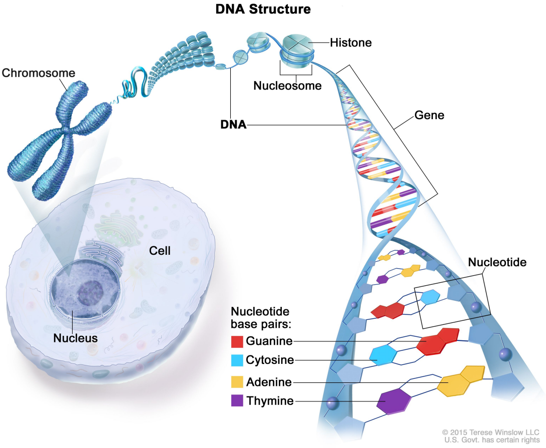


Figure: Where DNA is located [1].

```
>5341592 Part of the capsid of crAssphage
TTACCGTACAGTAAAGGTTTCGAGAGAAGGAGAGAAAGAGAAAG
AACGAATAGTCGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTG
TGTGTGTGTGTGTG
>5263167 [E. coli] [hypothetical protein]
TATGACCCATGCACCACTAGGGAGCTTAAATTCTGTTGGTGGTG
TTGCTACTGAAATTAACTCTGTAAACTATGTATCTCCTAGATCT
TGGTTAACATCATC
>5152774 Part of the capsid of crAssphage
ACTAACCGACTGACTGACGTGACTGACTGACTGACTGACTGACT
GACTGACTGACTGACTGACTGACTGACTGAATGACTGAC
```

DNA
De DNA a proteínas

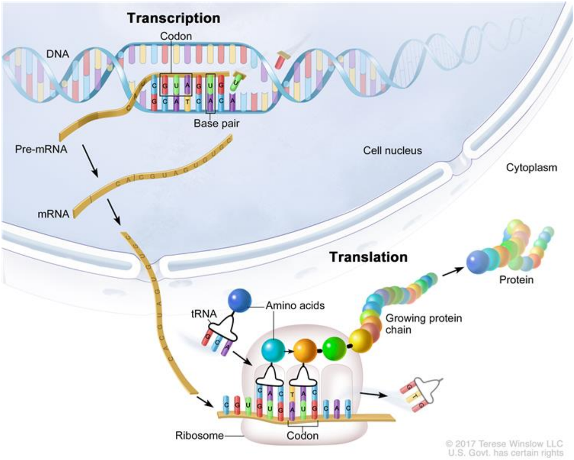


Figure: Transcription and translation [2].

Content

Variantes y Mutaciones

Tipos

- ***Single-Nucleotide Variant* (SNV)**, cambios a menos de 10 bases.
- ***Structural Variation* (SV)**, cambios a mas de 10 bases, incluso pueden llegar a aumentar la cantidad de cromosomas.

Variantes y Mutaciones
Ejemplo

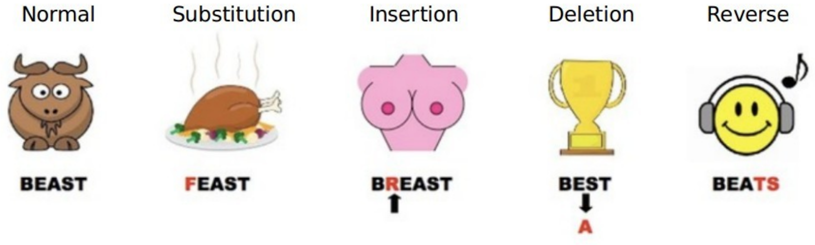


Figure: Overview of the Different Types of Point Mutations.

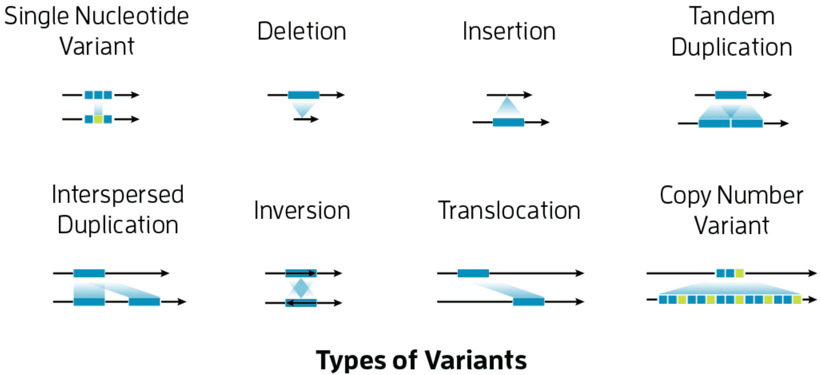


Figure: Example of structural variants. Source: [3]

Variaciones a nivel de cromosomas

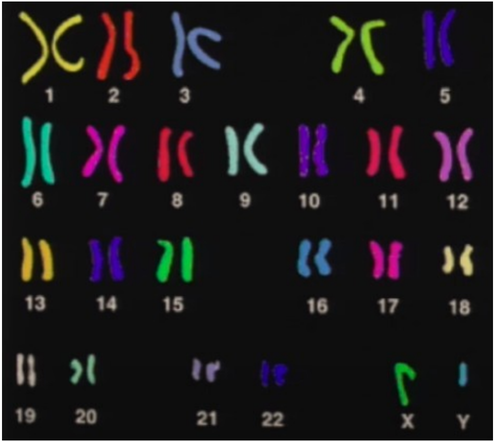


Figure: Los 46 cromosomas presentes en una célula.

Variaciones a nivel de cromosomas

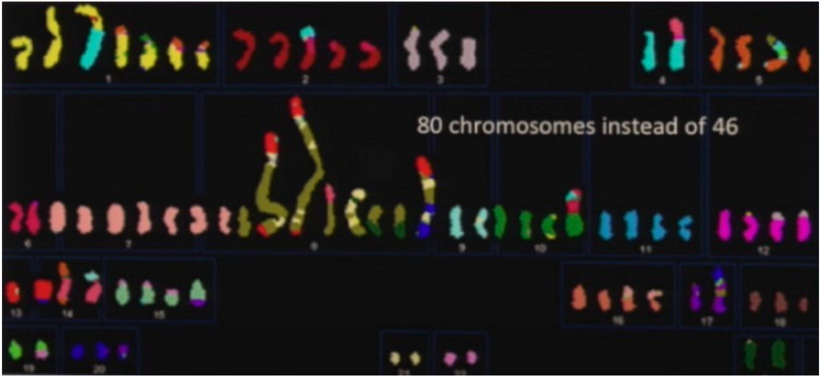


Figure: Cromosomas de una mujer con Cáncer de mama (1971).

Content

Inmunoterapia del Cáncer

Es un tipo de tratamiento contra el Cáncer que estimula las defensas naturales del cuerpo para combatir el Cáncer [4].

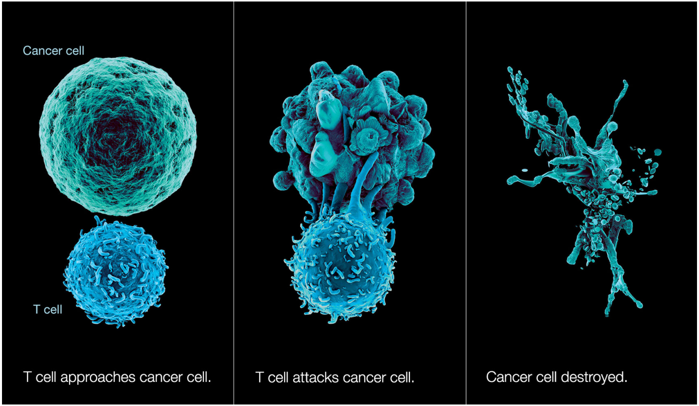


Figure: Ejemplo de como una célula T destruye células del cancer [5].

Es una **proteína** que se forma en las células de Cáncer cuando ocurre mutaciones en el DNA, cumplen un rol importante al **estimular una respuesta inmune** [1, 6].

En la actualidad hay varios métodos para detectar a predecir neo antígenos, pero **solo una pequeña cantidad de ellos** logran estimular al sistema inmune [7, 8].

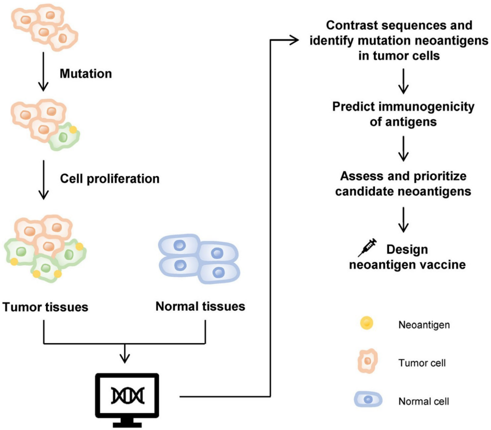
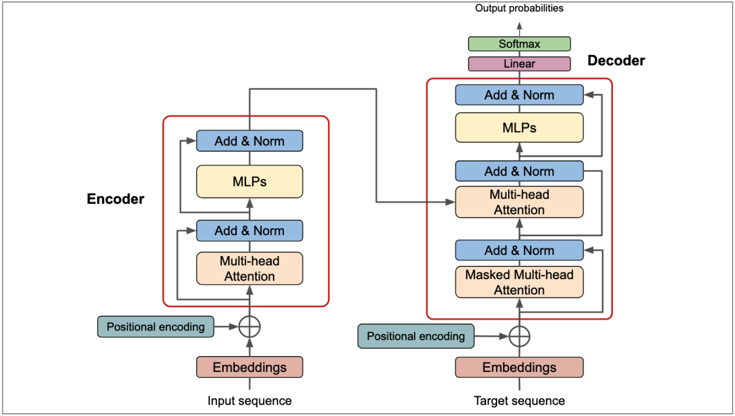


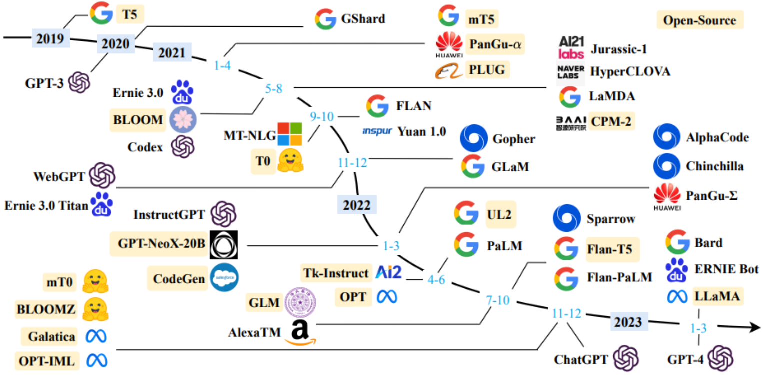
Figure: Proceso para la generación de vacunas personalizadas [9].

Content

Transformers



LLMs


2019
2020
2021
2022
2023
1-4
5-8
9-10
11-12
1-3
4-6
7-10
11-12
1-3
T5
GPT-3
GShard
mT5
PanGu-α
PLUG
Open-Source
Jurassic-1
HyperCLOVA
LaMDA
CPM-2
Ernie 3.0
BLOOM
Codex
MT-NLG
T0
FLAN
Yuan 1.0
Gopher
GLaM
AlphaCode
Chinchilla
PanGu-Σ
WebGPT
Ernie 3.0 Titan
InstructGPT
GPT-NeoX-20B
CodeGen
UL2
PaLM
Sparrow
Flan-T5
Flan-PaLM
Bard
ERNIE Bot
LLaMA
Tk-Instruct
OPT
GLM
AlexaTM
mT0
BLOOMZ
Galatica
OPT-IML
ChatGPT
GPT-4

Inmunoterapia del Cáncer
Pipeline

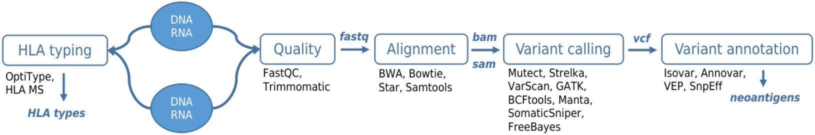
1. Sampling and sequencing

2. Neoantigen candidates detection
   - 2.1 Alignment
   - 2.2 Variant calling
   - 2.3 Variant annotation

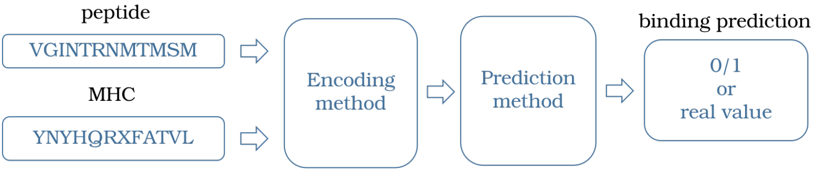
3. Neoantigen prioritization
   - 3.1 Peptide-MHC binding prediction
   - 3.2 pMHC-TCR binding prediction

4. Vaccine development in-vitro

5. Clinical trials



HLA typing
OptiType,
HLA MS
HLA types
DNA
RNA
DNA
RNA
Quality
FastQC,
Trimmomatic
fastq
Alignment
BWA, Bowtie,
Star, Samtools
bam
sam
Variant calling
Mutect, Strelka,
VarScan, GATK,
BCFtools, Manta,
SomaticSniper,
FreeBayes
vcf
Variant annotation
Isovar, Annovar,
VEP, SnpEff
neoantigens


peptide
VGINTRNMTMSM
MHC
YNYHQRXFATVL
Encoding method
Prediction method
binding prediction
0/1
or
real value

Protein Language Models

Table: Pre-trainned BERT models for several protein tasks: TAPE, ProtBert, ESM1, 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
ProtT5-XL	Uniref50, BFD	2122M	24	1024	32	3B
ProtT5-XXL	Uniref50, BFD	2122M	24	1024	128	11B
ESM-1 (6 layers)	Uniref50	60M	6	768	12	43M
ESM-1 (12 layers)	Uniref50	60M	12	768	12	85M
ESM-1 (34 layers)	Uniref50	60M	34	1280	20	670M
ESM-1b	Uniref50	60M	34	1280	20	650M
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
ESM-2 (36 layers)	Uniref50	60M	36	2560	20	3B
ESM-2 (48 layers)	Uniref50	60M	48	5120	20	15B

Fin-tuning
pMHC

Transformer models used in pMHC

Table: Transformers and deep learning methods with attention mechanism used for pMHC binding prediction.

Year	Name	Input	Model
2023[10]	ESM-GAT	One-hot	BERT with transfer learning from ESM1b and ESM2 fine-tuned with a Graph Attention Network (GAT) at the end. It outperformed NetMHCpan4.1.
2023[11]	CapsNet-MHC	BLOSUM62	Capsule Neural Network, it outperformed state-of-art tools for small peptides of 8 to 11-mer.
2023[12]	STMHCpan	One-hot	A Star-Transformer model, it use usefull for anylenght peptides and could extended for predicting T-cell responses.
2023[13]	DapNet-HLA	Fused word embedding	Combined the advantages of CNN, SENet (for pooling), and LSTM with attention.
2022[14]	HLAB	One-hot	BERT from ProtBert pre-trained model followed by a BiLSTM with attention mechanism.
2022[15]	MHC RoBERTa	One-hot	RoBERTa pre-trained and followed by 12 multi-head SA and a FC layers, it outperformed NetMHCPan 3.0.
2022[16]	TransPHLA	One-hot	It used SA mechanism based on four blocks, it slightly outperformed NetMHCpan4.1 and is faster making predictions.
2021[17]	CapTransformer	One-hot	Transformer with cross attention pooling to capture local and global information.
2021[18]	ImmunoBERT	One-hot	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.

Conclusiones

Con el auge de los Transformers, cada vez se desarrollan mas propuestas de fine-tuning para diversas areas en Proteómica.

Se ha mejorado mucho el acierto para la detección de neo antígenos; sin embargo la gran variedad de tipos de Cancer aún es una tarea muy compleja.

Entrenar estos modelos implica un alto costo computacional lo cual dificulta la investigación de laboratorios perqueños.

References I

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

[2] NCI,
"Nci dictionary of cancer terms,"
https://www.cancer.gov/publications/dictionaries/cancer-terms/def/transcription, 2020,
Accessed: 2020-03-20.

[3] PacBio,
"Two review articles assess structural variation in human genomes," https://www.pacb.com/blog/two-review-articles-assess-structural-variation-in-human-ge
2021,
Accessed: 2021-05-07.

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

References II

[5] NortShore,
"Immunotherapy," 2022.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Nasser Hashemi, Boran Hao, Mikhail Ignatov, Ioannis Ch Paschalidis, Pirooz Vakili, Sandor Vajda, and Dima Kozakov,
"Improved prediction of mhc-peptide binding using protein language models,"
Frontiers in Bioinformatics, vol. 3, 2023.

References IV

[11] Mahmood Kalemati, Saeid Darvishi, and Somayyeh Koohi, "Capsnet-mhc predicts peptide-mhc class i binding based on capsule neural networks," *Communications Biology*, vol. 6, no. 1, pp. 492, 2023.

[12] Zheng Ye, Shaohao Li, Xue Mi, Baoyi Shao, Zhu Dai, Bo Ding, Songwei Feng, Bo Sun, Yang Shen, and Zhongdang Xiao, "Stmhcpan, an accurate star-transformer-based extensible framework for predicting mhc i allele binding peptides," *Briefings in Bioinformatics*, vol. 24, no. 3, pp. bbad164, 2023.

[13] Yuanyuan Jing, Shengli Zhang, and Houqiang Wang, "Dapnet-hla: Adaptive dual-attention mechanism network based on deep learning to predict non-classical hla binding sites," *Analytical Biochemistry*, vol. 666, pp. 115075, 2023.

References V

[14] 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.

[15] 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 VI

[16] 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.

[17] Cheng Chen, Zongzhao Qiu, Zhenghe Yang, Bin Yu, and Xuefeng Cui,
"Jointly learning to align and aggregate with cross attention pooling for peptide-mhc class i binding prediction,"
in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 18–23.

References VII

[18] 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.

[19] 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.

References VIII

[20] 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.

[21] 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.

[22] Xiaoyun Yang, Liyuan Zhao, Fang Wei, and Jing Li, "Deepnetbim: deep learning model for predicting hla-epitope interactions based on network analysis by harnessing binding and immunogenicity information," *BMC bioinformatics*, vol. 22, no. 1, pp. 1–16, 2021.

[23] Jing Jin, Zhonghao Liu, Alireza Nasiri, Yuxin Cui, Stephen-Yves Louis, Ansi Zhang, Yong Zhao, and Jianjun Hu, "Deep learning pan-specific model for interpretable mhc-i peptide binding prediction with improved attention mechanism," *Proteins: Structure, Function, and Bioinformatics*, vol. 89, no. 7, pp. 866–883, 2021.

[24] Ziqi Chen, Martin Renqiang Min, and Xia Ning, "Ranking-based convolutional neural network models for peptide-mhc class i binding prediction," *Frontiers in Molecular Biosciences*, vol. 8, pp. 634836, 2021.

References X

[25] 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.

[26] 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.

References XI

[27] 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.

Questions?