Universidad Nacional de San Agustín

Detección *in Silico* de Neoantígenos Utilizando Transformers y Transfer Learning en el Marco de Desarrollo de Vacunas Personalizadas para Tratar el Cáncer

MSc. Vicente Machaca Arceda

2023

Contenido

Contexto y Motivación
- Estadísticas en Cáncer
- Inmunoterapia del Cáncer
- Vacunas Personalizadas

Problema y Objetivos
- Motivación y Problema
- Objetivo

Revisión Sistemática de la Literatura (RSL)
- Metodología
- Resultados

Propuesta

Resultados

Conclusiones y Trabajos futuros

Contenido

Contexto y Motivación
- Estadísticas en Cáncer
- Inmunoterapia del Cáncer
- Vacunas Personalizadas

Problema y Objetivos
- Motivación y Problema
- Objetivo

Revisión Sistemática de la Literatura (RSL)
- Metodología
- Resultados

Propuesta

Resultados

Conclusiones y Trabajos futuros

Contexto y Motivación

An la actualidad, el cáncer representa el mayor problema de salud mundial [1].

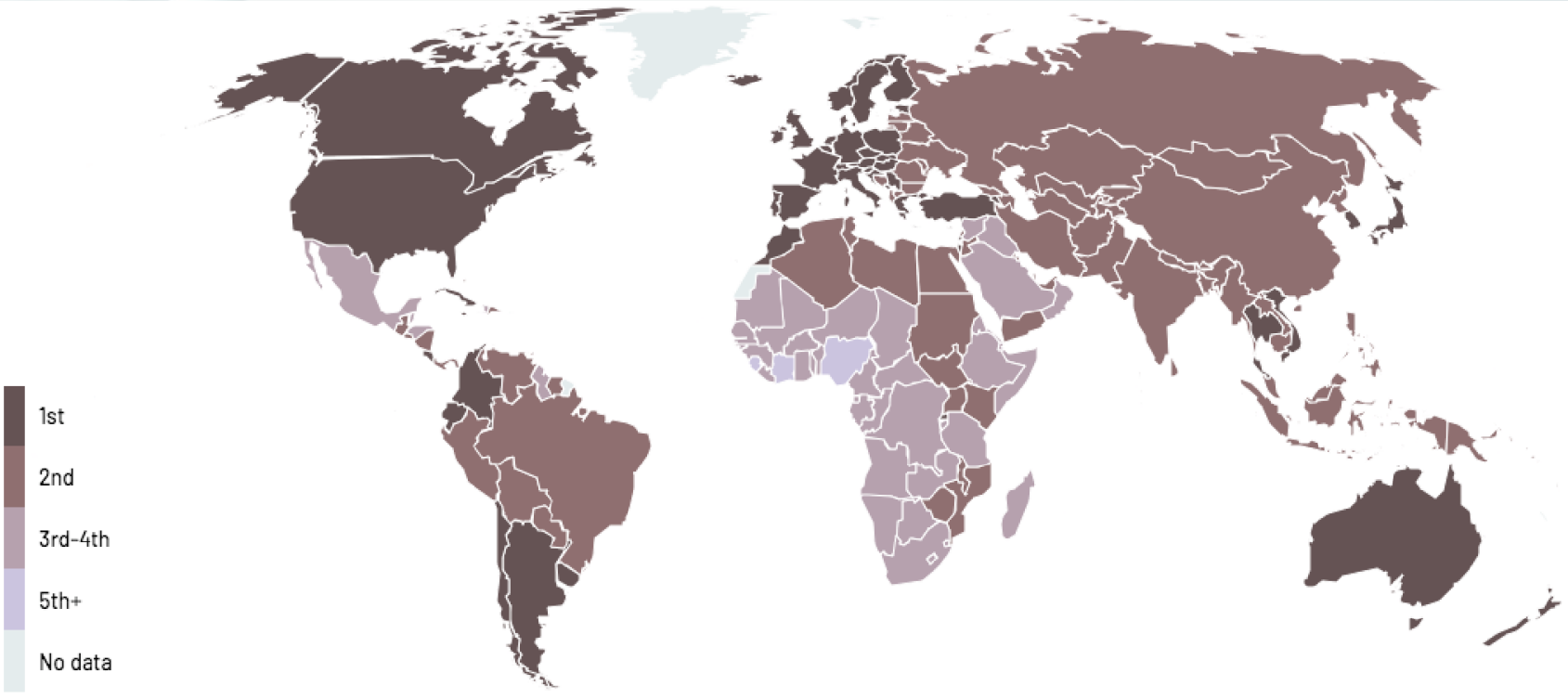


Figure: Ranking de las muertes por cáncer entre 30 y 69 años. **Fuente**: The Atlas Cancer [2].

Contexto y Motivación

Muertes por tipos de cáncer

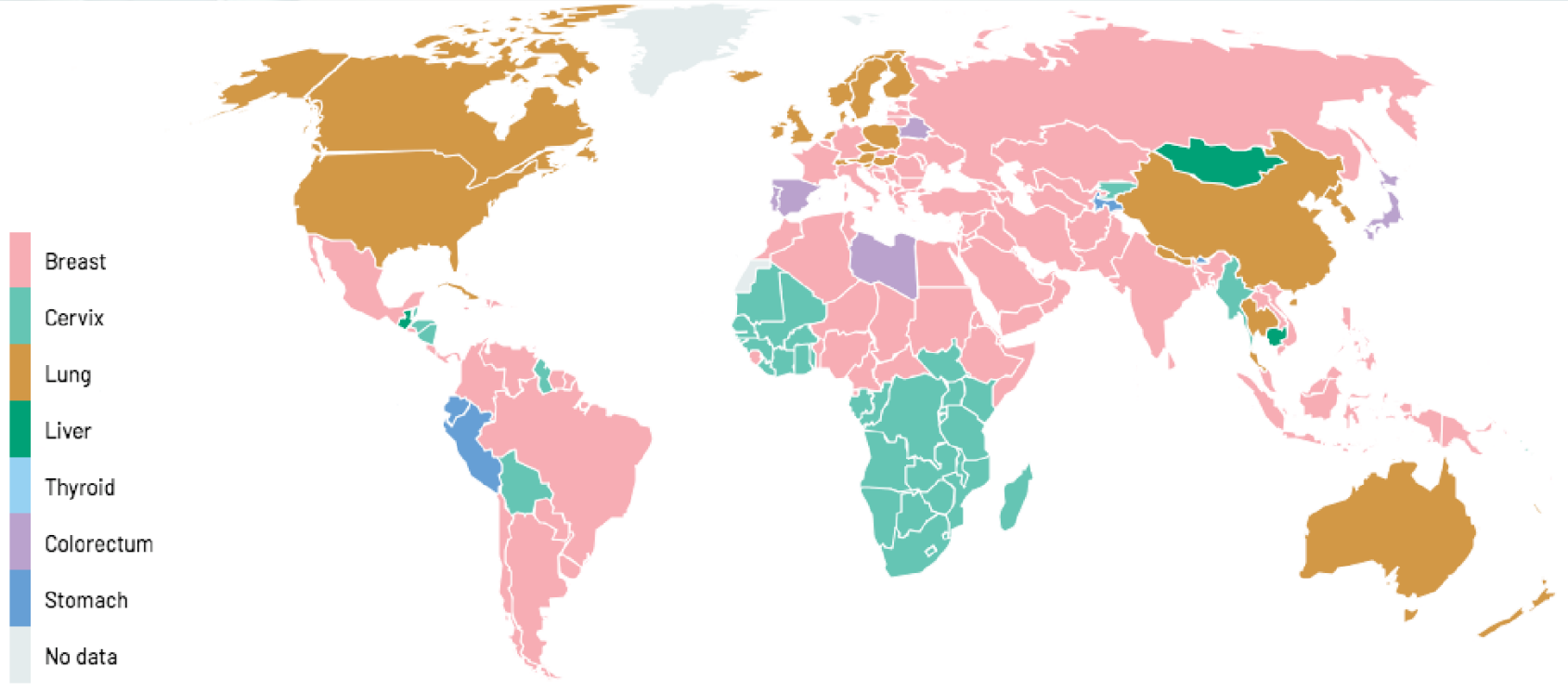


Figure: Ranking de las muertes por tipo de cáncer en mujeres. **Fuente**: The Atlas Cancer [2].

Contexto y Motivación

Muertes por tipos de cáncer

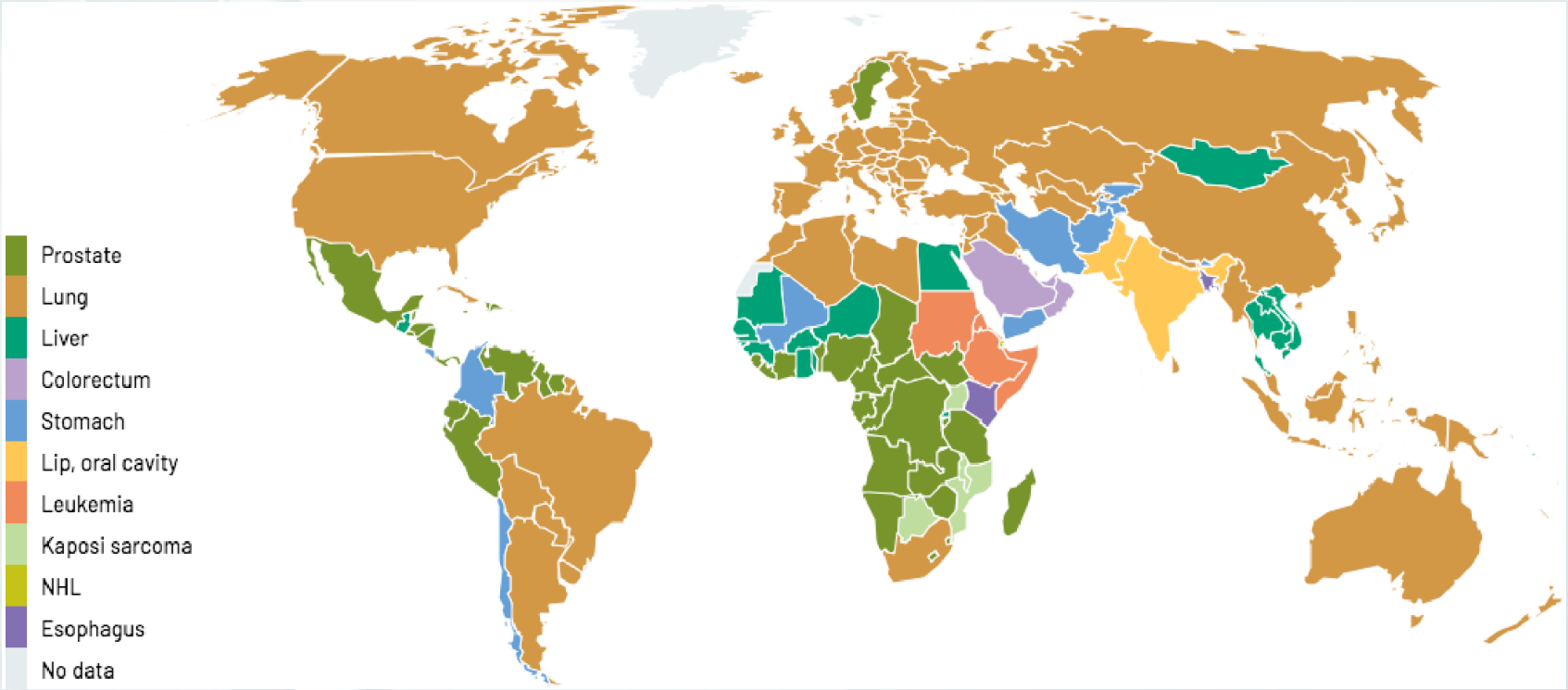


Figure: Ranking de las muertes por tipo de cáncer en hombres. **Fuente**: The Atlas Cancer [2].

Contexto y Motivación
Porcentaje de casos y muertes

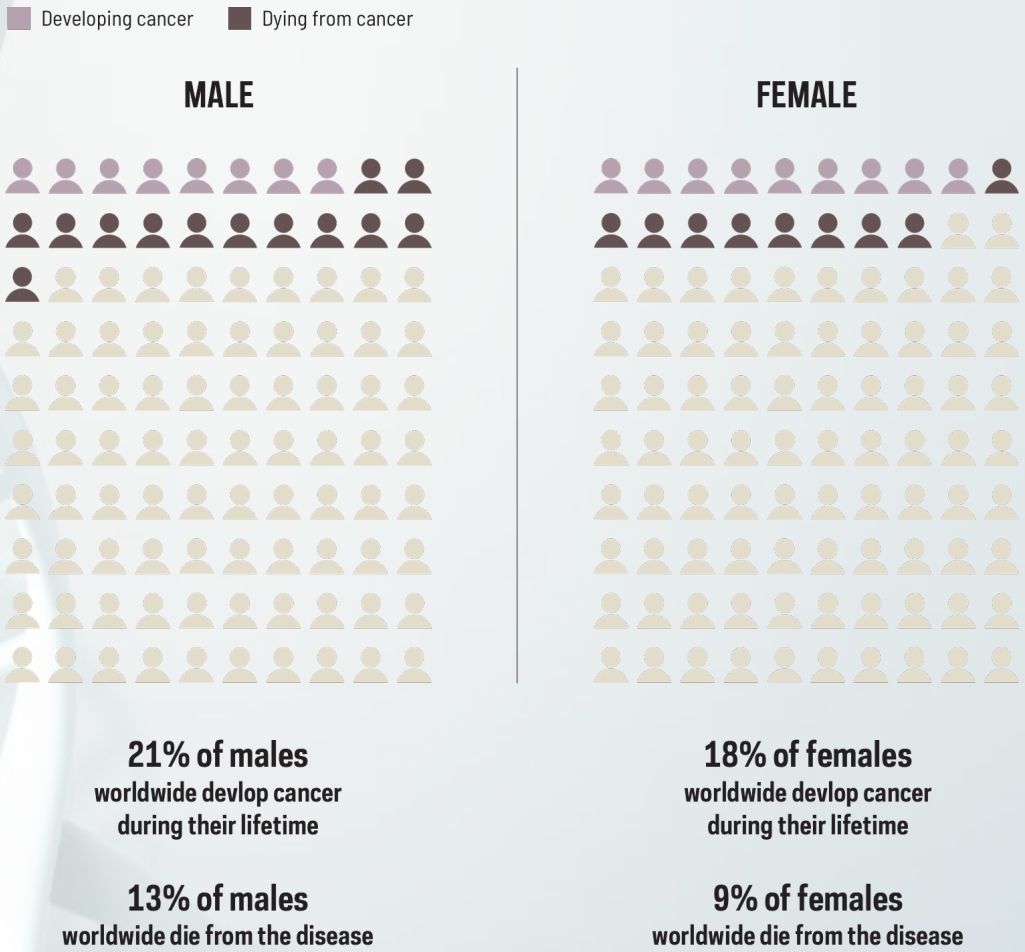


Figure: Porcentaje de casos y muertes por sexo. **Fuente** The Atlas Cancer [2].

Contexto y Motivación

Predicción de nuevos casos

Figure: Predicción de nuevos casos para el 2040. **Fuente** The Atlas Cancer [2].

Contenido

Contexto y Motivación
- Estadísticas en Cáncer
- Inmunoterapia del Cáncer
- Vacunas Personalizadas

Problema y Objetivos
- Motivación y Problema
- Objetivo

Revisión Sistemática de la Literatura (RSL)
- Metodología
- Resultados

Propuesta

Resultados

Conclusiones y Trabajos futuros

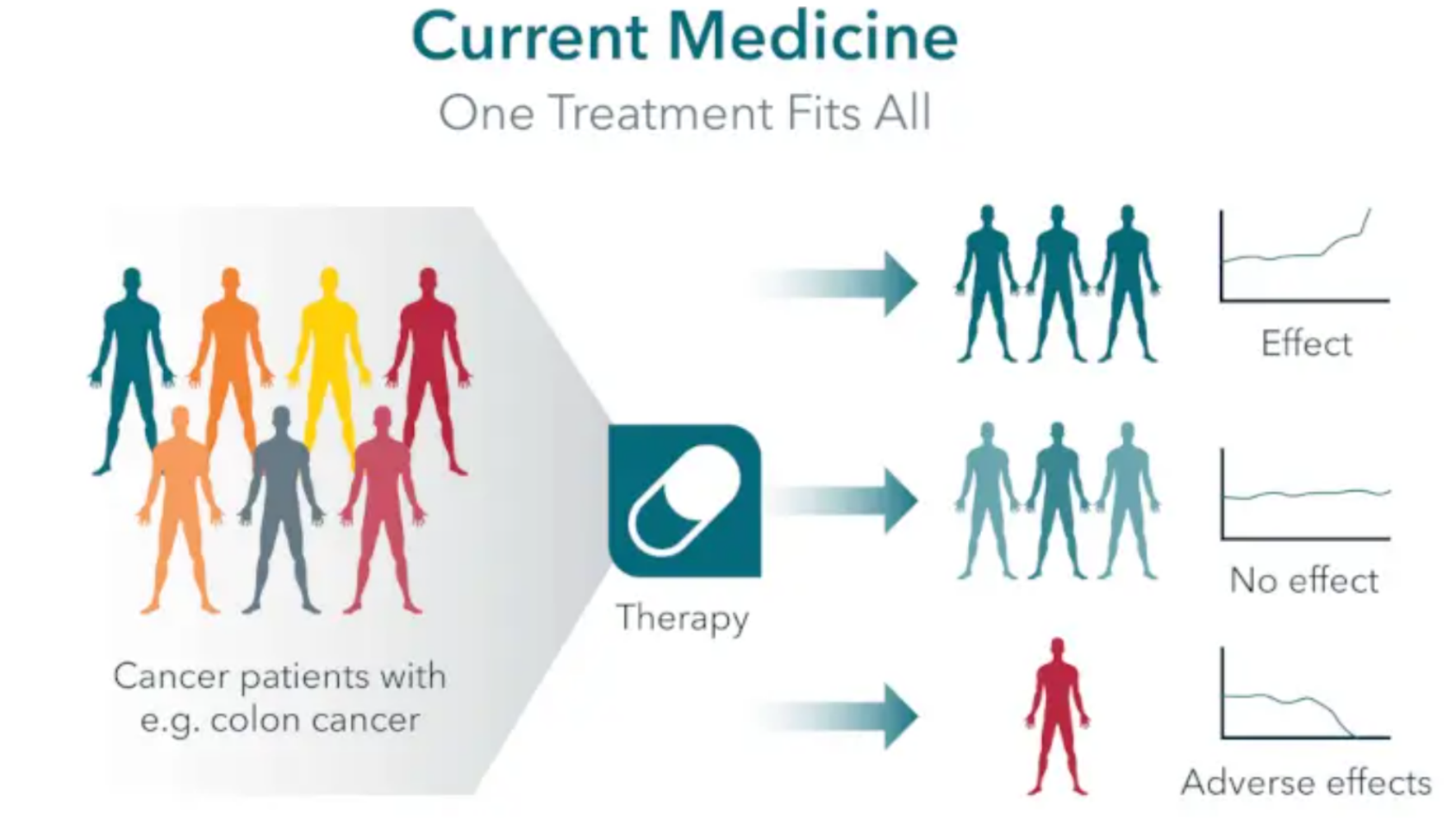


Figure: Pacientes con el mismo tipo de cáncer pueden reaccionar de forma disitinta a los mismos tratamientos. **Fuente** The Atlas Cancer [3].

Contexto y Motivación

Reacciones distintas para cada paciente

Figure: Cada paciente necesita un tratamiento personalizado. **Fuente** The Atlas Cancer [3].

Inmunoterapia del Cáncer

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

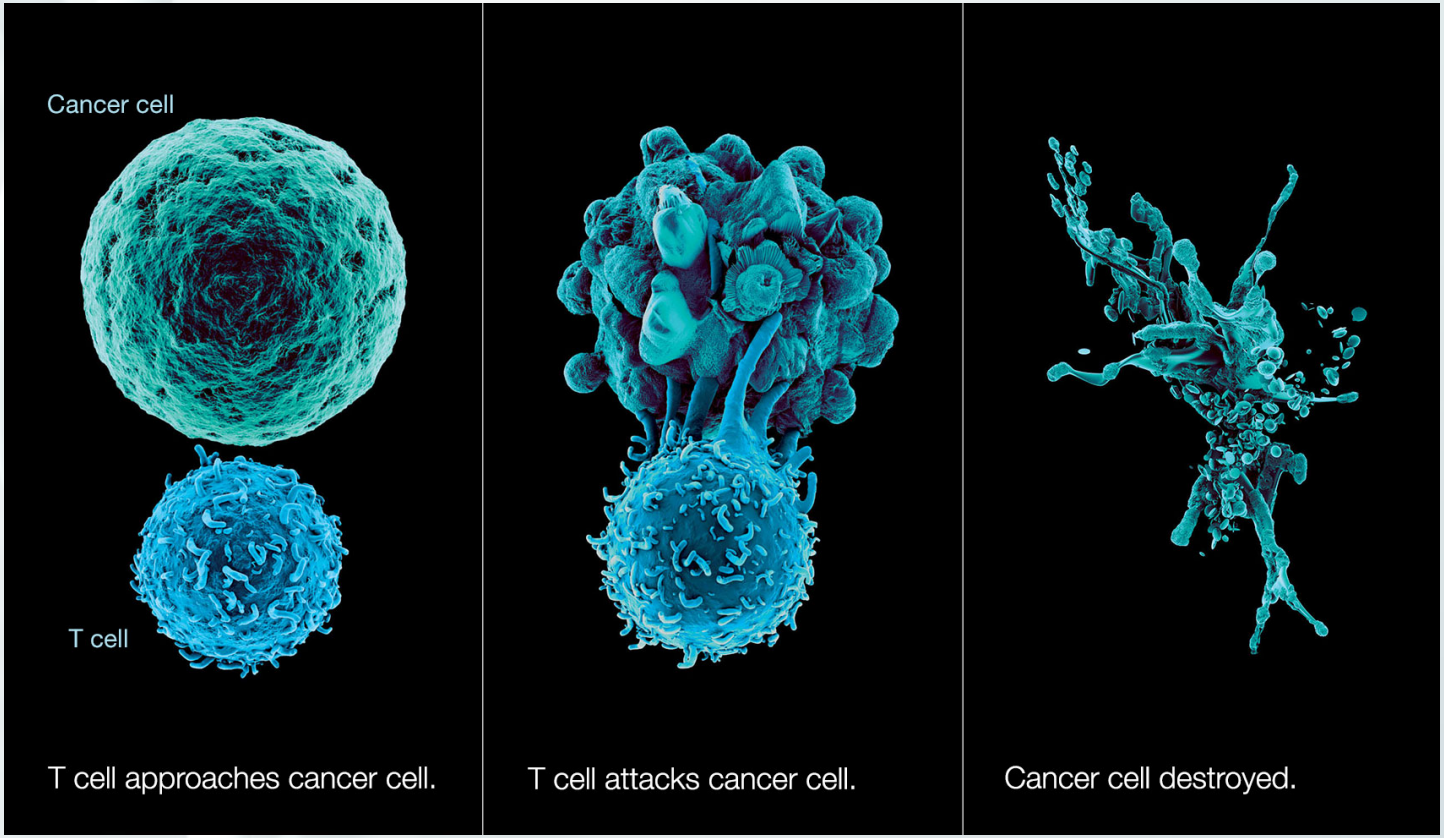


Figure: Ejemplo de como una célula T destruye células del cancer [5].

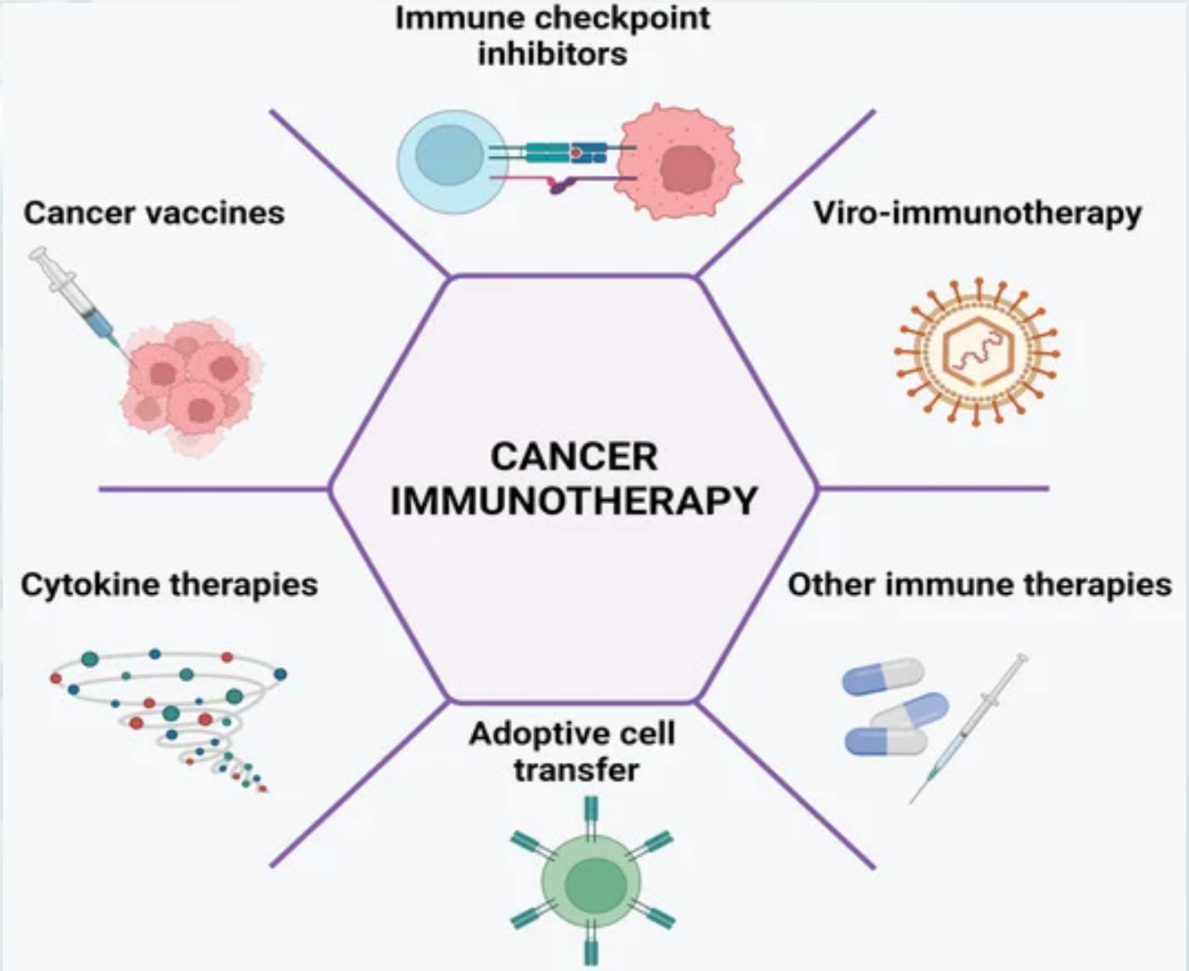


Figure: Tipos de tratamientos para la inmunoterapia del cáncer. **Fuente**: [6].

Contenido

Contexto y Motivación
- Estadísticas en Cáncer
- Inmunoterapia del Cáncer
- Vacunas Personalizadas

Problema y Objetivos
- Motivación y Problema
- Objetivo

Revisión Sistemática de la Literatura (RSL)
- Metodología
- Resultados

Propuesta

Resultados

Conclusiones y Trabajos futuros

Es una **proteína** que se forma en las células de Cáncer cuando ocurre mutaciones en el DNA [7, 8].

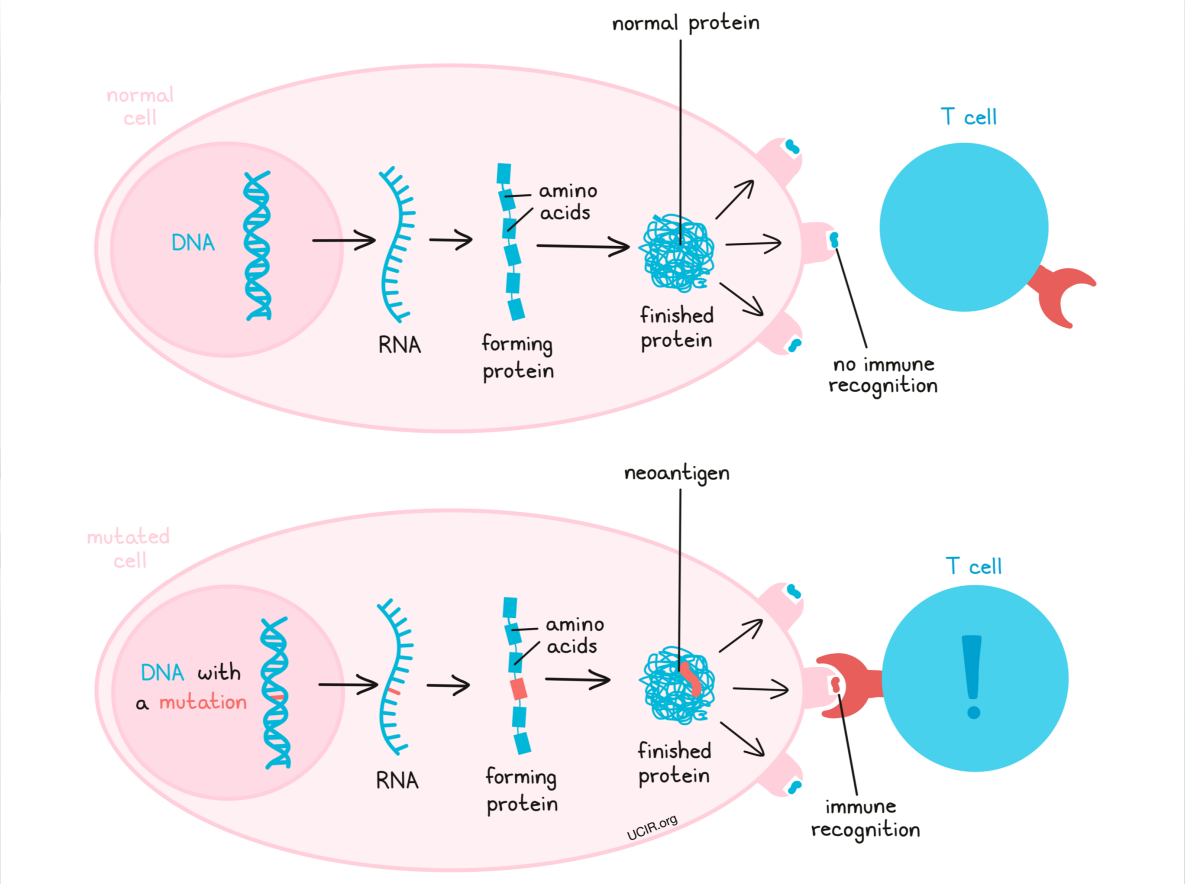


Figure: Neoantígenos y células T. **Fuente**: [9].

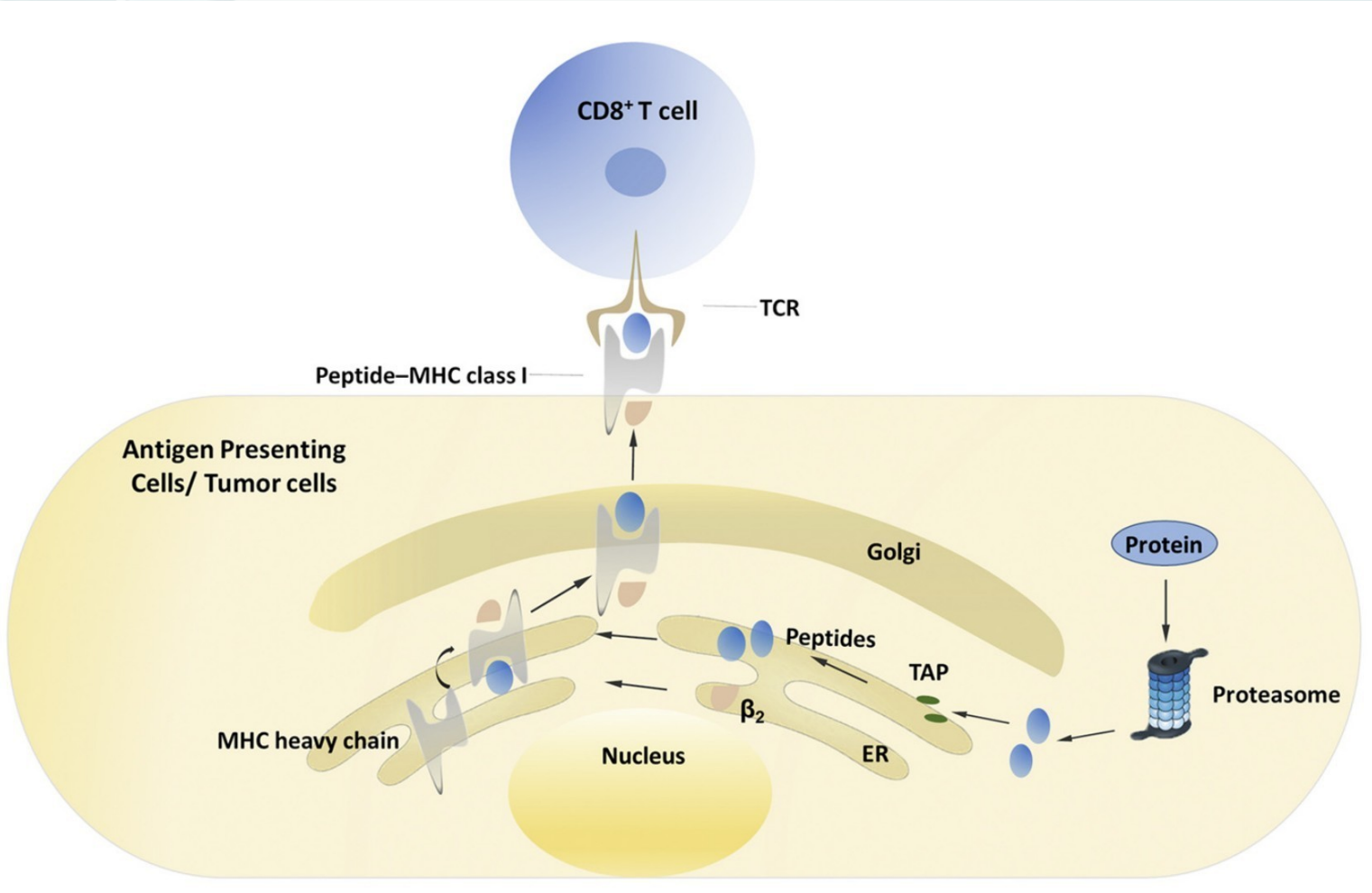


Figure: Presentación de antígenos por MHC-I. Fuente: [10]

Contexto y Motivación
Vacunas personalizadas

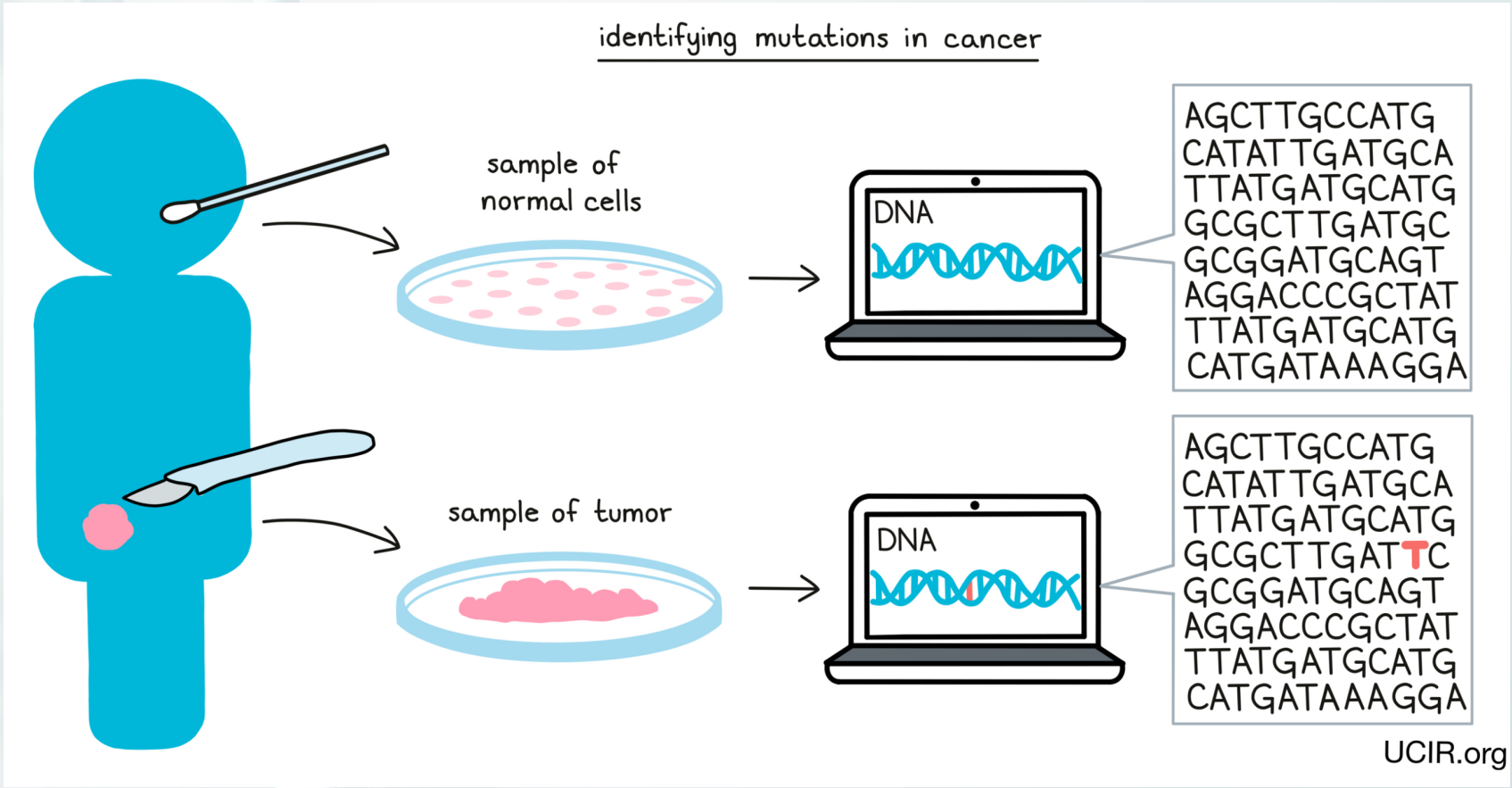


Figure: Proceso para la generación de vacunas contra el cáncer. **Fuente**: [9].

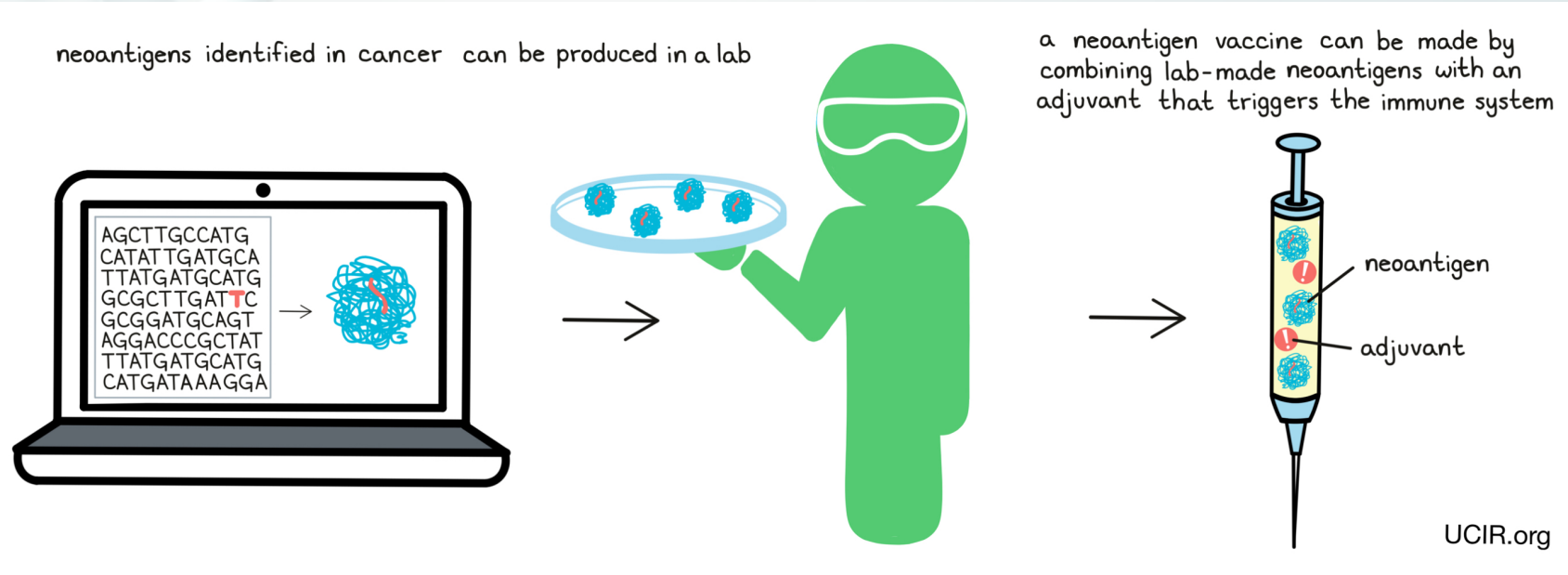


Figure: Proceso para la generación de vacunas contra el cáncer. **Fuente**: [9].

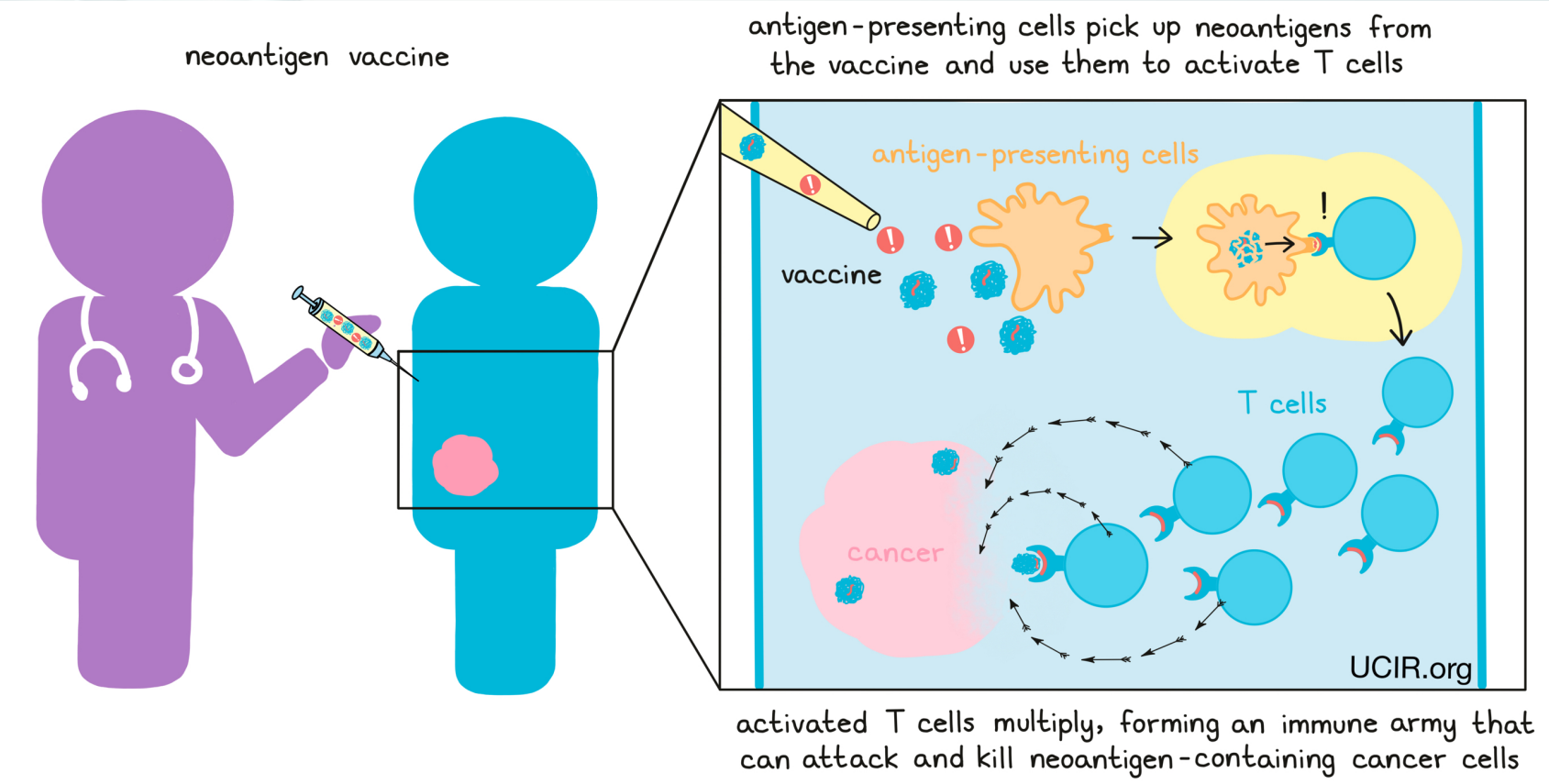


Figure: Proceso para la generación de vacunas contra el cáncer. **Fuente**: [9].

Contexto y Motivación
Vacunas personalizadas

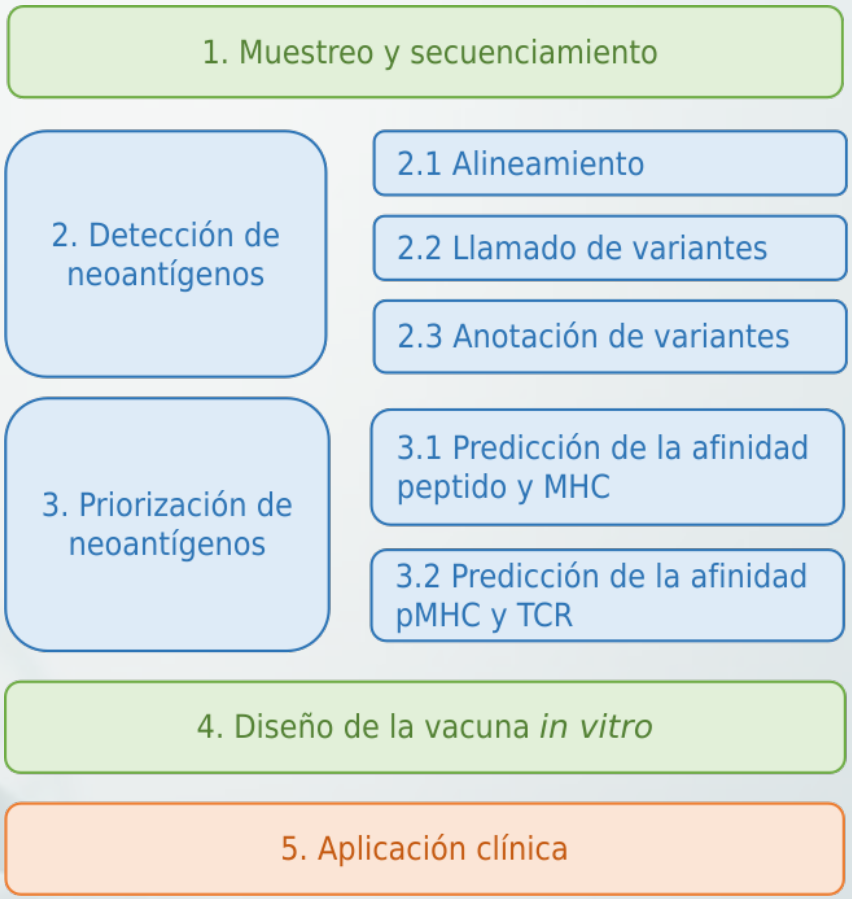


Figure: Resumen del proceso de generación de vacunas contra el cáncer.

Contenido

Contexto y Motivación
- Estadísticas en Cáncer
- Inmunoterapia del Cáncer
- Vacunas Personalizadas

Problema y Objetivos
- Motivación y Problema
- Objetivo

Revisión Sistemática de la Literatura (RSL)
- Metodología
- Resultados

Propuesta

Resultados

Conclusiones y Trabajos futuros

Motivación

El cáncer representa el mayor problema de salud mundial, pero lamentablemente los métodos basados en cirugías, radioterapias, quimioterapias tienen baja efectividad [11].

La inmunoterapia del cáncer es una alternativa para el desarrollo de vacunas personalizadas, pero este proceso depende de una correcta detección de neo antígenos [12, 11].

Problema

Menos del **5% de péptidos** detectados en *pMHC binding*, llegan a la membrana de la células. Para *peptide-MHC presentation*, propuestas recientes solo llegan a **0.6 de *presicion* y 0.4 de *recall*** [13].

En este contexto, la tesis se enfoca en el problema de *pMHC presentation*, considerándolo como un problema de clasificación binaria, y tomando como entrada la secuencia de aminoácidos del péptido y la secuencia de aminoácidos de la proteína MHC.

Contenido

Contexto y Motivación
- Estadísticas en Cáncer
- Inmunoterapia del Cáncer
- Vacunas Personalizadas

Problema y Objetivos
- Motivación y Problema
- Objetivo

Revisión Sistemática de la Literatura (RSL)
- Metodología
- Resultados

Propuesta

Resultados

Conclusiones y Trabajos futuros

Objetivo general

Proponer un método basado en *deep learning* para la detección de neo antígenos, enfocados en el problema de *peptide-MHC presentation*.

Contenido

Contexto y Motivación
- Estadísticas en Cáncer
- Inmunoterapia del Cáncer
- Vacunas Personalizadas

Problema y Objetivos
- Motivación y Problema
- Objetivo

Revisión Sistemática de la Literatura (RSL)
- Metodología
- Resultados

Propuesta

Resultados

Conclusiones y Trabajos futuros

Table: Cadenas de busqueda utilizadas en la RSL.

Cadena de busqueda
neoantigen AND (detection OR pipeline) AND deep learning
(MHC OR HLA) AND binding AND deep learning
(MHC-I OR MHC-II OR MHC OR HLA) AND (peptide OR epitope) AND (binding OR affinity OR prediction OR detection OR presentation)
TCR interaction prediction

Table: Bases de datos utilizadas en la RSL.

Bases de datos
IEEE Xplore
Science Direct
Springer
ACM Digital Library
PubMed
BioRxiv

Table: Criterios de inclusión y exclusión de artículos utilizados en la RSL.

Criterios de inclusión	Criterios de exclusión
Artículos con categoría ERA (A, B o C) si son conferencias y Journals Q1, Q2 o Q3. Sobre *deep learning* La metodología es detallada. Tiene repositorio de código fuente y base de datos (deseable).	Trabajos de baja calidad, que no esten rankeados.

Table: Cantidad de artículos encontrados y seleccionados según los criterios de inclusión y exclusión en la RSL.

Año	Artículos encontrados	Artículos seleccionados
2018	57	21
2019	72	31
2020	86	29
2021	61	34
2022	58	19
Total	**334**	**134**

Contenido

Contexto y Motivación
- Estadísticas en Cáncer
- Inmunoterapia del Cáncer
- Vacunas Personalizadas

Problema y Objetivos
- Motivación y Problema
- Objetivo

Revisión Sistemática de la Literatura (RSL)
- Metodología
- Resultados

Propuesta

Resultados

Conclusiones y Trabajos futuros

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

Year	Ref.	Approach	Name	MHC	Encoding
2022	[14]	pMHC(b)	DeepMHCII	II	PFR
2021	[15]	pMHC(b)	DeepImmuno	I	AAindex1
2021	[16]	pMHC(p)	APPM	I	One-hot
2021	[17]	pMHC(p)	MHCfovea	I	One-hot
2021	[18]	pMHC(b)	CNN-PepPred	II	BLOSUM
2020	[19]	pMHC(b)	IConMHC	I	PCA and AAindex3
2020	[20]	pMHC(b)	OnionMHC	I	BLOSUM and structural features
2020	[21]	pMHC(p)	MINERVA	I	Physicochemical properties
2019	[22]	pMHC(b)	CNN-NF	I	Sequence, Hydropathy, Polarity, Length
2019	[23]	pMHC(b)	DeepSeqPan	I	One-hot
2018	[24]	pMHC(b)	ConvMHC	I	Contact side HLA.peptide

Table: List of research since 2018 that uses CNNs s with RNN or attention mechanisms for peptide-MHC binding and presentation. MHCherryPan uses CNN with RNN, the other uses CNN witn Attention mechanims.

Year	Ref.	Approach	Name	MHC	Encoding
2021	[25]	pMHC(b)	DeepNetBim	I	BLOSUM
2021	[26]	pMHC(b)	Deep Attention Pan	I	BLOSUM
2019	[27]	pMHC(b)	ACME	I	BLOSUM
2020	[28]	pMHC(b)	MHCherryPan	I	BLOSUM

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

Table: List of research since 2018 that uses Transformers (self-attention) for peptide-MHC binding and presentation.

Year	Ref.	Approach	Name	MHC	Encoding
2022	[35]	pMHC(b)	MHCRoBERTa	I	Tokenized from a pre-trained model
2022	[36]	pMHC(b)	TransPHLA	I	Character embedding model
2021	[37]	pMHC(b)	BERTMHC	II	Embeding layer
2021	[38]	pMHC(p)	ImmunoBERT	I	Embeding layer

Table: Public databases of *pMHC binding*, *pMHC presentation*, pMHC-TCR interaction, and 3D structures of proteins.

Name	Year ref.	Description
VDJdb	2018 [39]	TCR binding to pMHC, contains 5491 samples.
IEDB	2018 [40]	The bigger database, contains information *T-cell epitopes*
TSNAdb	2018 [41]	It contains 7748 samples of mutations and HLA of 16 types of cancer.
NeoPeptide	2019 [42]	It contains samples of neoantigens resulting from somatic mutations and related items. 1818137 epitopes of more than 36000 neoantigens.
pHLA3D	2019 [43]	Presents 106 3D structures of the alpha, *beta2M* chains, and peptides of HLA-I molecules
dbPepNeo	2020 [44]	It has validated samples of the *peptide-MHC* bond, from MS. It contains 407794 low-quality samples, 247 medium-quality, and 295 high-quality samples.
dbPepNeo2.0	2022 [45]	It gathers a list of neoantigens and HLA molecules. It presents 801 high-quality and 842,289 poor-quality HLAs. Also, 55 class II neoantigens and 630 TCR-bound neo antigens.
IntroSpect	2022 [46]	Tool for building databases on *peptide-MHC binding*. It uses data from *Mass Spectrometry*.
IPD-IMGT/HLA	2022 [47]	With 25000 MHC molecules and 45 alleles.

Table: List of *pipelines* since 2018 for the detection of neoantigens.

Name	Year ref.	Input	Output
Neopepsee	2018 [48]	RNA-seq, somatic mutations (VCF), HLA type (optional)	Neoantigens and gene expression levels
PGV Pipeline	2018 [49]	DNA-seq	Neoantigens
ScanNeo	2019 [50]	RNA-seq	Neoantigens
NeoPredPipe	2019 [51]	Mutations (VCF) y HLA type	Neoantigens and variant annotation
pVACtools	2020 [52]	Mutations (VCF)	Neoantigens
ProGeo-neo	2020 [53]	RNA-seq y somatic mutations (VCF)	Neoantigens
Neoepiscope	2020 [54]	Somatic mutations (VCF) and BAM files	Neoantigens and mutations
NeoANT-HILL	2020 [55]	RNA-seq y somatic mutations (VCF)	Neoantigens and gene expression levels
NAP-CNB	2021 [56]	RNA-seq	Neoantigens
PEPPRMINT	2021 [57]	DNA-seq	Neoantigens
Valid-NEO	2022 [58]	Somatic mutations (VCF), HLA type (optional)	Neoantigens

Propuesta

La propuesta se basa el los modelos BERTMHC [37] y APPM [59]. Tambien, se utilizará *transfer learning* de ESM-1b [60], esta red neuronal fue entrenada con 250 millones de proteínas a diferencia de TAPE (utilizada por BERTMHC), que fue entrenada con 30 milllones de proteínas.

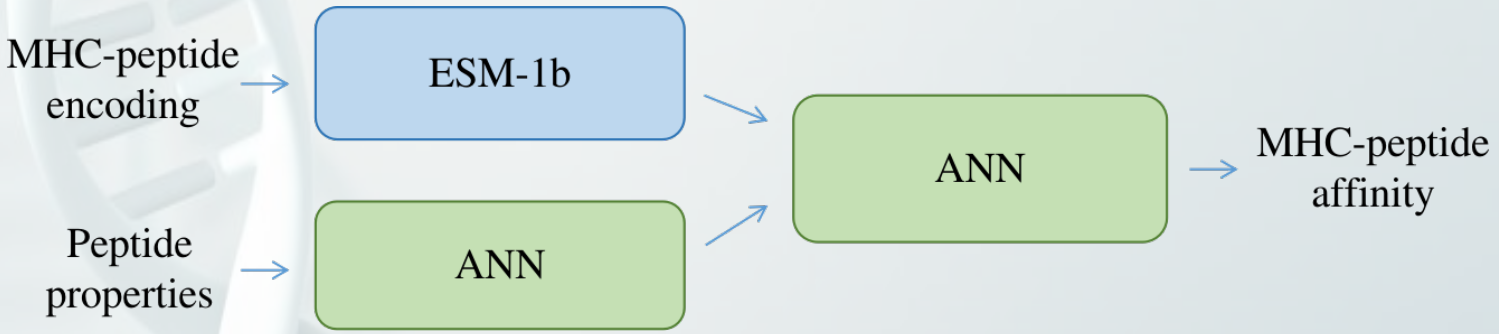


Figure: Proceso general utilizado para la detección de neo antígenos a partir de secuencias de DNA. Fuente: [61].

Propuesta
BERTMHC

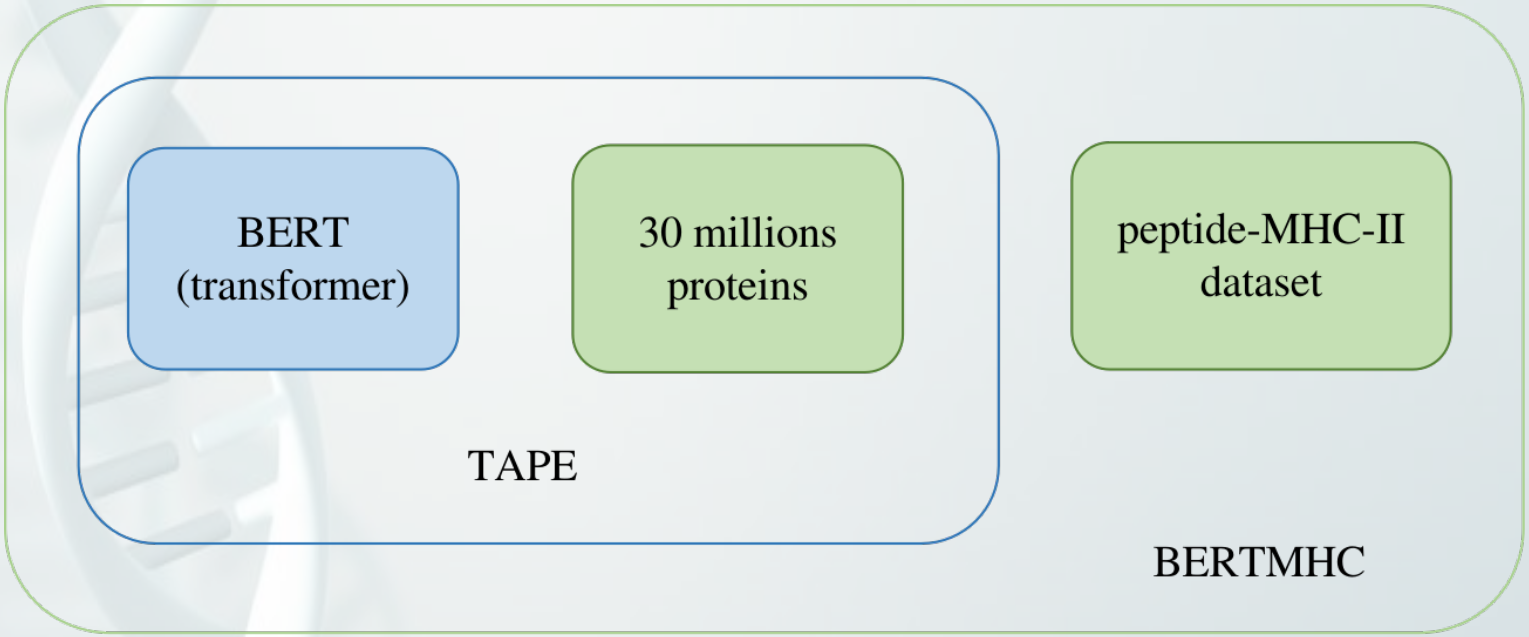


Figure: BERTMHC.

Propuesta
APPM

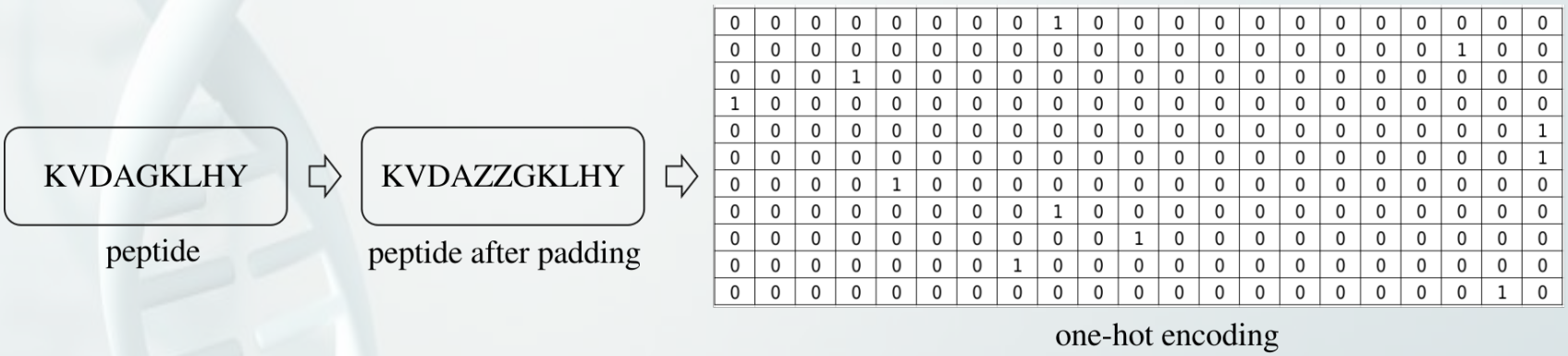


Figure: Proceso para obtener una matriz (imagen) a partir de un péptido (APPM).

MHC alleles utilizados

Table: Cantidad de muestras por tipo de *allele*.

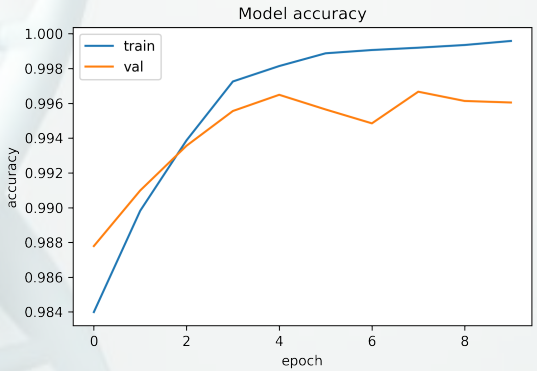
Alleles	**Label = 1**	**Label = 0**	**Train**	**Test**
A*01:01	3398	48700	45498	6600
A*02:01	6779	165342	160921	11200
A*02:03	1780	116299	107879	10200
A*31:01	1879	45918	41597	6200
B*44:02	1525	44760	40085	6200
B*44:03	1487	39482	34769	6200
MHC-II alleles	1917	496	1533	384

Resultados

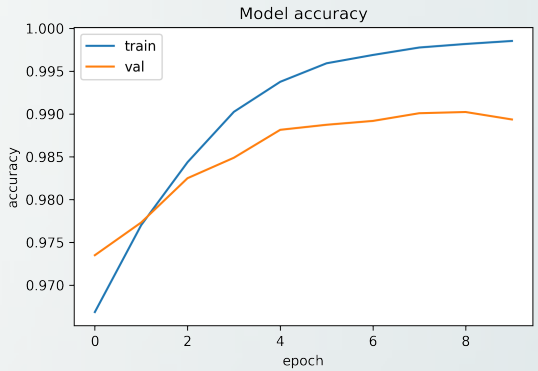
Table: Resultados obtenidos en cada base de datos.

Allele	*Accuracy*	*F1 score*	*Precision*	*Recall*
A*01:01	0.978	0.917	0.982	0.887
A*0201	0.962	0.956	0.965	0.948
A*02:03	0.992	0.979	0.994	0.969
A*31:01	0.980	0.968	0.989	0.951
B*44:02	0.991	0.981	0.968	0.997
B*44:03	0.992	0.987	0.995	0.980

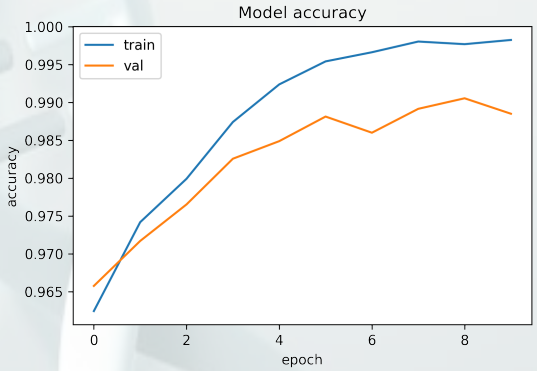
Resultados



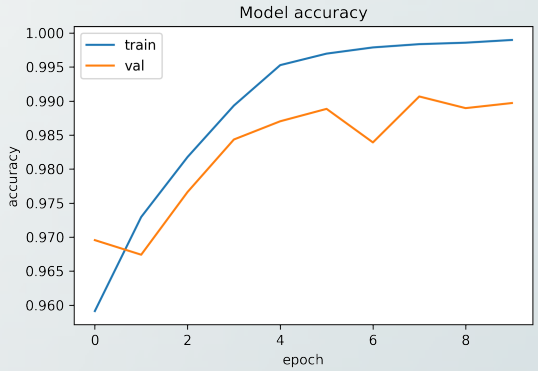
(a) A*01:01



(b) A*02:01



(c) A*02:03



(d) A*31:01

Figure: *Accuracy* durante cada *epoch*, para cada base de datos. Las bases de datos representan las células HLA A*01:01, A*02:01, A*02:03, A*31:01.

Conclusiones

Se ha desarrollado una RSL, sobre los métodos de detección de neoantígenos utilizando *deep learning*. Esto ha logrado identificar las tendencias, retos y problemas del tema de interes.

Se ha realizado experimentos preliminares, sobre el uso de CNNs para el problema de peptide-MHC presentation. Se ha utilizado muestras de MS con un enfoque *single allele* (se entrena varios modelos para cada tipo de MHC).

Trabajos futuros

Recientemente un trabajo [62] tambien propone el uso de *transfer learning* pero de un modelo pre-entrenado con 250 millones de proteínas. Entonces, se plantea utilizar la misma red, aumentar la cantidad de muestras y evaluar los resultados.

Actualmente se cuenta con una base de datos de proteínas MHC [43], entonces utilizando AlphaFold de Google, se plantea predecir la estructura de varios péptidos y analizar el enlace péptido-MHC desde un punto de vista de la computación gráfica.

References I

[1] Rebecca L Siegel, Kimberly D Miller, Nikita Sandeep Wagle, and Ahmedin Jemal,
"Cancer statistics, 2023,"
Ca Cancer J Clin, vol. 73, no. 1, pp. 17–48, 2023.

[2] Cancer Atlas,
"Cancer atlas - the burden," 2023.

[3] Personalized Medicine,
"Pdx and personalized medicine," 2023.

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

[5] NortShore,
"Immunotherapy," 2022.

References II

[6] Mateusz Kciuk, Esam Bashir Yahya, Montaha Mohamed Ibrahim Mohamed, Summya Rashid, Muhammad Omer Iqbal, Renata Kontek, Muhanad A Abdulsamad, and Abdulmutalib A Allaq,
"Recent advances in molecular mechanisms of cancer immunotherapy,"
Cancers, vol. 15, no. 10, pp. 2721, 2023.

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

[8] 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 III

[9] UCIR,
"Neoantige-based therapy," 2023.

[10] 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.

[11] 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 IV

[12] 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.

[13] Nil Adell Mill, Cedric Bogaert, Wim van Criekinge, and Bruno Fant, "neoms: Attention-based prediction of mhc-i epitope presentation," *bioRxiv*, 2022.

[14] 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 V

[15] 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.

[16] 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.

[17] 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.

References VI

[18] 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.

[19] 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.

[20] 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,"

References VII

Journal of Micromechanics and Molecular Physics, vol. 5, no. 03, pp. 2050009, 2020.

[21] 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.

[22] 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 VIII

[23] 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.

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

[25] 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.

References IX

[26] 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.

[27] 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 X

[28] Xuezhi Xie, Yuanyuan Han, and Kaizhong Zhang,
"Mhcherrypan: a novel pan-specific model for binding affinity prediction of class i hla-peptide,"
International Journal of Data Mining and Bioinformatics, vol. 24, no. 3, pp. 201–219, 2020.

[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 XI

[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 XII

[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 XIII

[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] 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 XIV

[36] 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.

[37] 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 XV

[38] 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.

[39] Mikhail Shugay, Dmitriy V Bagaev, Ivan V Zvyagin, Renske M Vroomans, Jeremy Chase Crawford, Garry Dolton, Ekaterina A Komech, Anastasiya L Sycheva, Anna E Koneva, Evgeniy S Egorov, et al.,
"Vdjdb: a curated database of t-cell receptor sequences with known antigen specificity,"
Nucleic acids research, vol. 46, no. D1, pp. D419–D427, 2018.

References XVI

[40] 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.

[41] Jingcheng Wu, Wenyi Zhao, Binbin Zhou, Zhixi Su, Xun Gu, Zhan Zhou, and Shuqing Chen, "Tsnadb: a database for tumor-specific neoantigens from immunogenomics data analysis," *Genomics, proteomics & bioinformatics*, vol. 16, no. 4, pp. 276–282, 2018.

References XVII

[42] Wei-Jun Zhou, Zhi Qu, Chao-Yang Song, Yang Sun, An-Li Lai, Ma-Yao Luo, Yu-Zhe Ying, Hu Meng, Zhao Liang, Yan-Jie He, et al.,
"Neopeptide: an immunoinformatic database of t-cell-defined neoantigens,"
Database, vol. 2019, 2019.

[43] Deylane Menezes Teles Oliveira, Rafael Melo Santos de Serpa Brandão, Luiz Claudio Demes da Mata Sousa, Francisco das Chagas Alves Lima, Semiramis Jamil Hadad do Monte, Mário Sérgio Coelho Marroquim, Antonio Vanildo de Sousa Lima, Antonio Gilberto Borges Coelho, Jhonatan Matheus Sousa Costa, Ricardo Martins Ramos, et al.,
"phla3d: An online database of predicted three-dimensional structures of hla molecules,"
Human Immunology, vol. 80, no. 10, pp. 834–841, 2019.

References XVIII

[44] Xiaoxiu Tan, Daixi Li, Pengjie Huang, Xingxing Jian, Huihui Wan, Guangzhi Wang, Yuyu Li, Jian Ouyang, Yong Lin, and Lu Xie, "dbpepneo: a manually curated database for human tumor neoantigen peptides," *Database*, vol. 2020, 2020.

[45] Manman Lu, Linfeng Xu, Xingxing Jian, Xiaoxiu Tan, Jingjing Zhao, Zhenhao Liu, Yu Zhang, Chunyu Liu, Lanming Chen, Yong Lin, et al., "dbpepneo2. 0: A database for human tumor neoantigen peptides from mass spectrometry and tcr recognition," *Frontiers in immunology*, p. 1583, 2022.

References XIX

[46] Le Zhang, Geng Liu, Guixue Hou, Haitao Xiang, Xi Zhang, Ying Huang, Xiuqing Zhang, Bo Li, and Leo J Lee,
"Introspect: Motif-guided immunopeptidome database building tool to improve the sensitivity of hla i binding peptide identification by mass spectrometry,"
Biomolecules, vol. 12, no. 4, pp. 579, 2022.

[47] James Robinson, Dominic J Barker, Xenia Georgiou, Michael A Cooper, Paul Flicek, and Steven GE Marsh,
"Ipd-imgt/hla database,"
Nucleic acids research, vol. 48, no. D1, pp. D948–D955, 2020.

References XX

[48] Sora Kim, Han Sang Kim, Eunyoung Kim, MG Lee, E-C Shin, and S Paik,
"Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information,"
Annals of Oncology, vol. 29, no. 4, pp. 1030–1036, 2018.

[49] Alex Rubinsteyn, Julia Kodysh, Isaac Hodes, Sebastien Mondet, Bulent Arman Aksoy, John P Finnigan, Nina Bhardwaj, and Jeffrey Hammerbacher,
"Computational pipeline for the pgv-001 neoantigen vaccine trial,"
Frontiers in immunology, vol. 8, pp. 1807, 2018.

References XXI

[50] Ting-You Wang, Li Wang, Sk Kayum Alam, Luke H Hoeppner, and Rendong Yang,
"Scanneo: identifying indel-derived neoantigens using rna-seq data,"
Bioinformatics, vol. 35, no. 20, pp. 4159–4161, 2019.

[51] Ryan O Schenck, Eszter Lakatos, Chandler Gatenbee, Trevor A Graham, and Alexander RA @miscNCIdictionary2022, author = NCI, title = National Cancer Institute Dictionary, year = 2022, url = https://www.cancer.gov/publications/dictionaries/genetics-dictionary, urldate = 2022-03-20 Anderson,
"Neopredpipe: high-throughput neoantigen prediction and recognition potential pipeline,"
BMC bioinformatics, vol. 20, no. 1, pp. 1–6, 2019.

References XXII

[52] Jasreet Hundal, Susanna Kiwala, Joshua McMichael, Christopher A Miller, Huiming Xia, Alexander T Wollam, Connor J Liu, Sidi Zhao, Yang-Yang Feng, Aaron P Graubert, et al.,
"pvactools: a computational toolkit to identify and visualize cancer neoantigens,"
Cancer immunology research, vol. 8, no. 3, pp. 409–420, 2020.

[53] Yuyu Li, Guangzhi Wang, Xiaoxiu Tan, Jian Ouyang, Menghuan Zhang, Xiaofeng Song, Qi Liu, Qibin Leng, Lanming Chen, and Lu Xie,
"Progeo-neo: a customized proteogenomic workflow for neoantigen prediction and selection,"
BMC medical genomics, vol. 13, no. 5, pp. 1–11, 2020.

References XXIII

[54] Mary A Wood, Austin Nguyen, Adam J Struck, Kyle Ellrott, Abhinav Nellore, and Reid F Thompson,
"Neoepiscope improves neoepitope prediction with multivariant phasing,"
Bioinformatics, vol. 36, no. 3, pp. 713–720, 2020.

[55] Ana Carolina MF Coelho, André L Fonseca, Danilo L Martins, Paulo BR Lins, Lucas M da Cunha, and Sandro J de Souza,
"neoant-hill: an integrated tool for identification of potential neoantigens,"
BMC Medical Genomics, vol. 13, no. 1, pp. 1–8, 2020.

References XXIV

[56] Carlos Wert-Carvajal, Rubén Sánchez-García, José R Macías, Rebeca Sanz-Pamplona, Almudena Méndez Pérez, Ramon Alemany, Esteban Veiga, Carlos Óscar S Sorzano, and Arrate Muñoz-Barrutia,
"Predicting mhc i restricted t cell epitopes in mice with nap-cnb, a novel online tool,"
Scientific reports, vol. 11, no. 1, pp. 1–10, 2021.

[57] Laura Y Zhou, Fei Zou, and Wei Sun,
"Prioritizing candidate peptides for cancer vaccines by pepprmint: a statistical model to predict peptide presentation by hla-i proteins,"
bioRxiv, 2021.

References XXV

[58] Yuri Laguna Terai, Chun Huang, Baoli Wang, Xiaonan Kang, Jing Han, Jacqueline Douglass, Emily Han-Chung Hsiue, Ming Zhang, Raj Purohit, Taylor deSilva, et al.,
"Valid-neo: A multi-omics platform for neoantigen detection and quantification from limited clinical samples,"
Cancers, vol. 14, no. 5, pp. 1243, 2022.

[59] 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.

References XXVI

[60] 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.

[61] 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.

References XXVII

[62] Nasser Hashemi, Boran Hao, Mikhail Ignatov, Ioannis Paschalidis, Pirooz Vakili, Sandor Vajda, and Dima Kozakov, "Improved predictions of mhc-peptide binding using protein language models," *bioRxiv*, 2022.