

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



## Contexto y Motivación

- Estadísticas en Cáncer
- Inmunoterapia del Cáncer
- Vacunas Personalizadas

## Problema y Objetivos

- Problema
- Objetivos

## Estado del arte

## Propuesta

## Experimentos y Resultados

- Base de datos
- Modelos pre-entrenados
- Resultados

## Discusión y Conclusiones



## Contexto y Motivación

Estadísticas en Cáncer  
Inmunoterapia del Cáncer  
Vacunas Personalizadas

## Problema y Objetivos

Problema  
Objetivos

## Estado del arte

## Propuesta

## Experimentos y Resultados

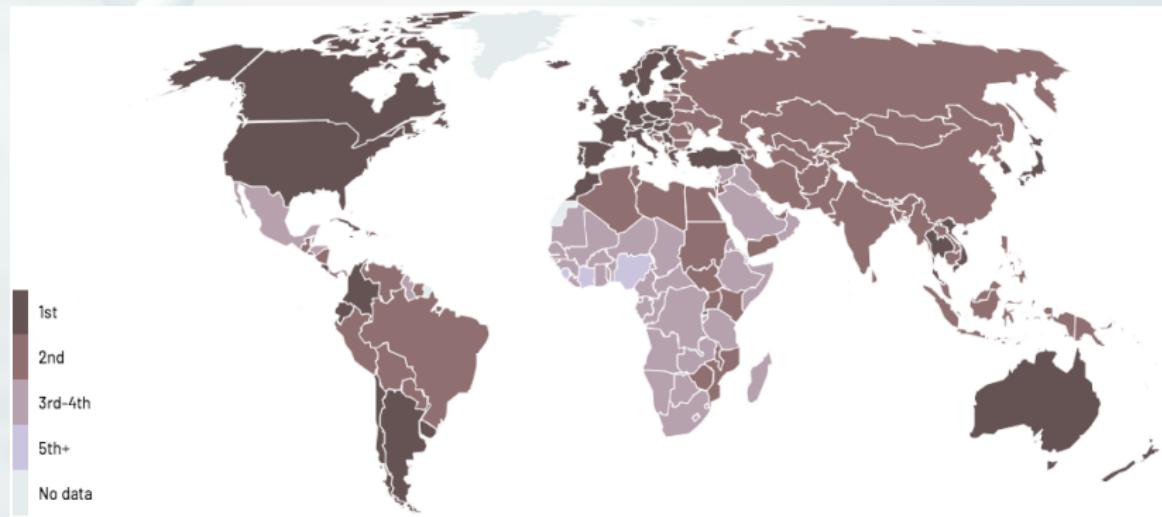
Base de datos  
Modelos pre-entrenados  
Resultados

## Discusión y Conclusiones

# Contexto y Motivación

3

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:** Atlas Cancer [2].

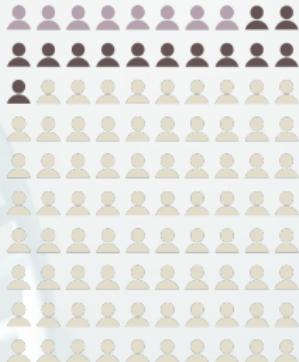
# Contexto y Motivación

## Porcentaje de casos y muertes



■ Developing cancer ■ Dying from cancer

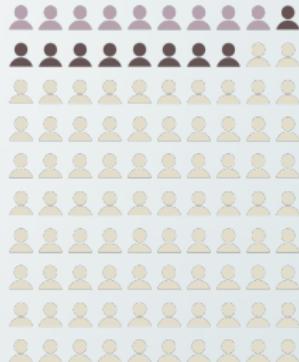
### MALE



**21% of males**  
worldwide develop cancer  
during their lifetime

**13% of males**  
worldwide die from the disease

### FEMALE



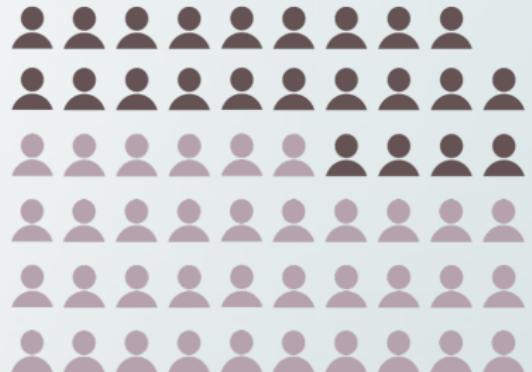
**18% of females**  
worldwide develop cancer  
during their lifetime

**9% of females**  
worldwide die from the disease

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

# Contexto y Motivación

Predicción de nuevos casos



New cases 2018       New cases 2040 (+demographic changes)

 0.5M people

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

# Contexto y Motivación

Reacciones distintas para cada paciente



6

## Current Medicine

One Treatment Fits All



Cancer patients with  
e.g. colon cancer



Therapy



Effect



No effect



Adverse effects

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

# Contexto y Motivación

Reacciones distintas para cada paciente



7

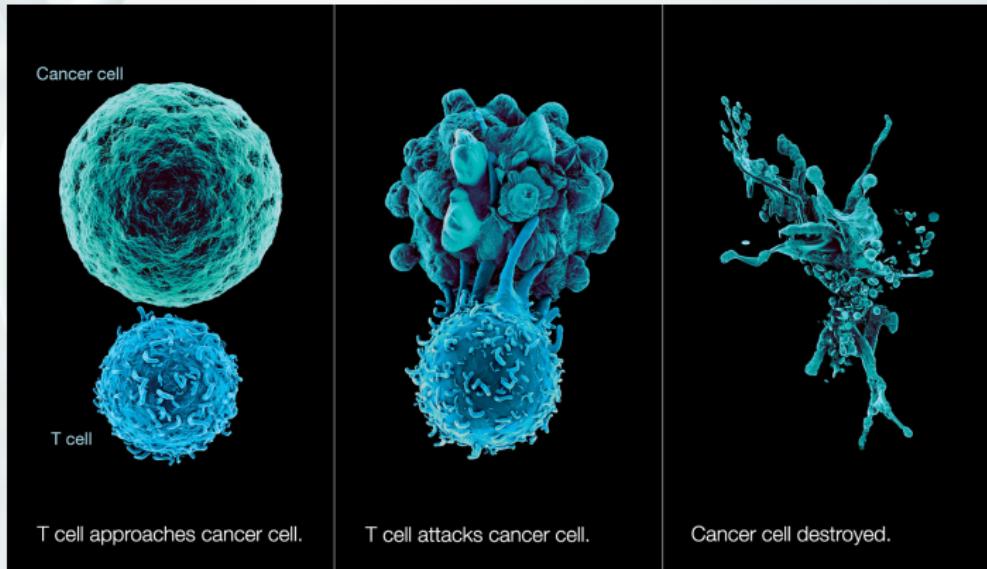


**Figure:** Cada paciente necesita un tratamiento personalizado. **Fuente** The Atlas Cancer [3].

# Inmunoterapia del Cáncer

8

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].

# Contexto y Motivación

## Inmunoterapia del Cáncer

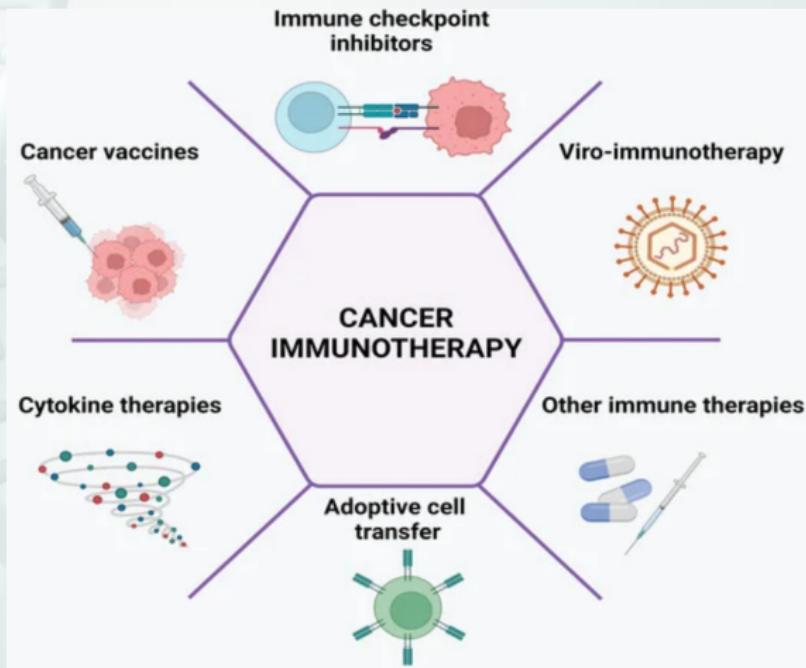


Figure: Tipos de tratamientos para la inmunoterapia del cáncer. Fuente: [6].

# Contexto y Motivación

## Neoantígenos



10

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

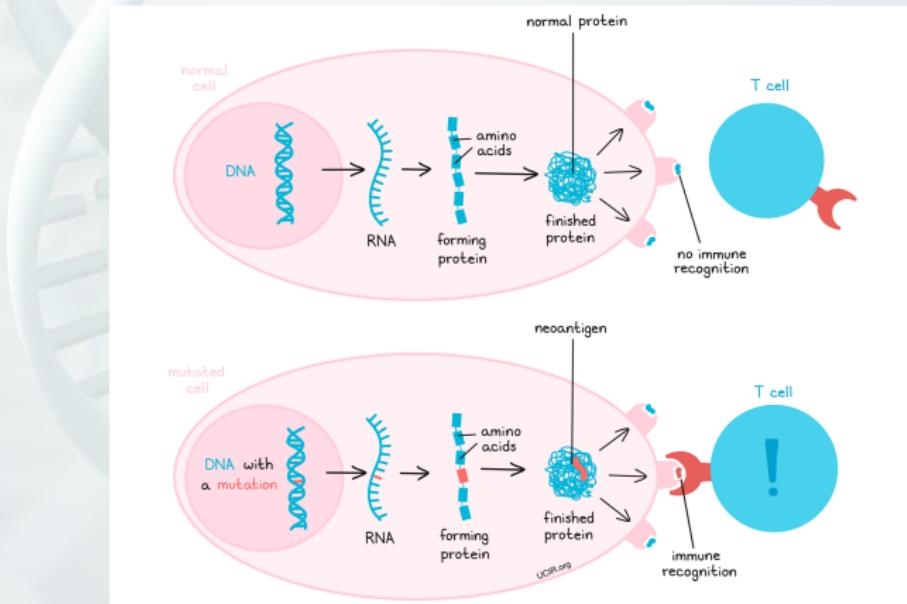


Figure: Neoantígenos y células T. Fuente: [9].

# Contexto y Motivación

## Vacunas personalizadas

11

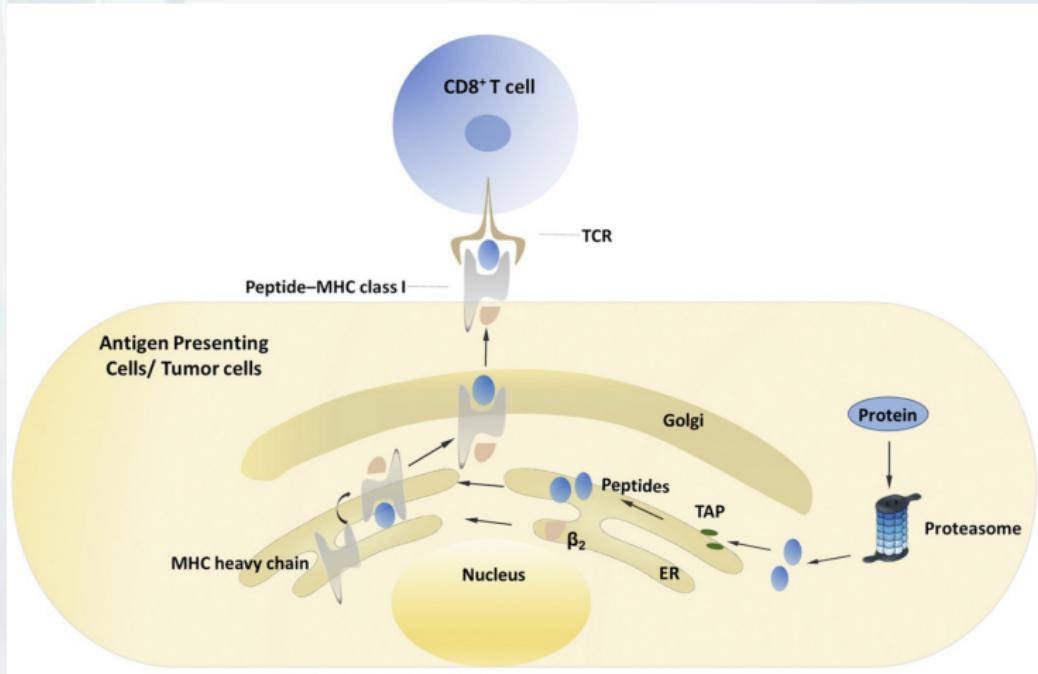


Figure: Presentación de antígenos por MHC-I. Fuente: [10]

# Contexto y Motivación

## Vacunas personalizadas

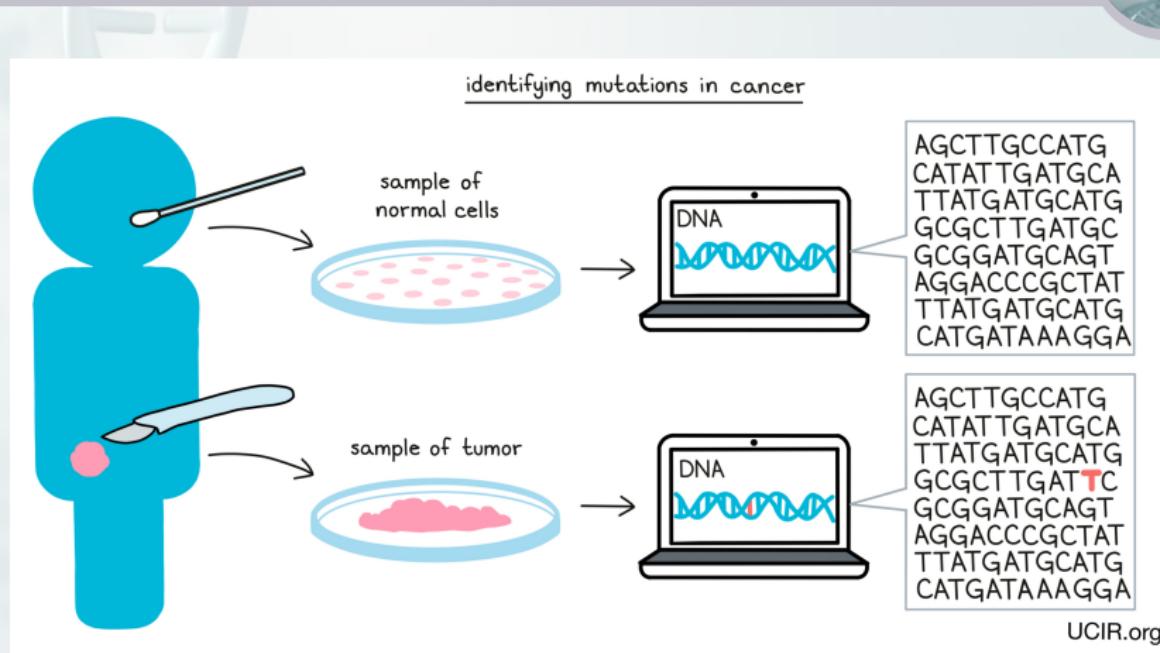


Figure: Proceso para la generación de vacunas contra el cáncer. Fuente: [9].

# Contexto y Motivación

## Vacunas personalizadas



13

neoantigens identified in cancer can be produced in a lab

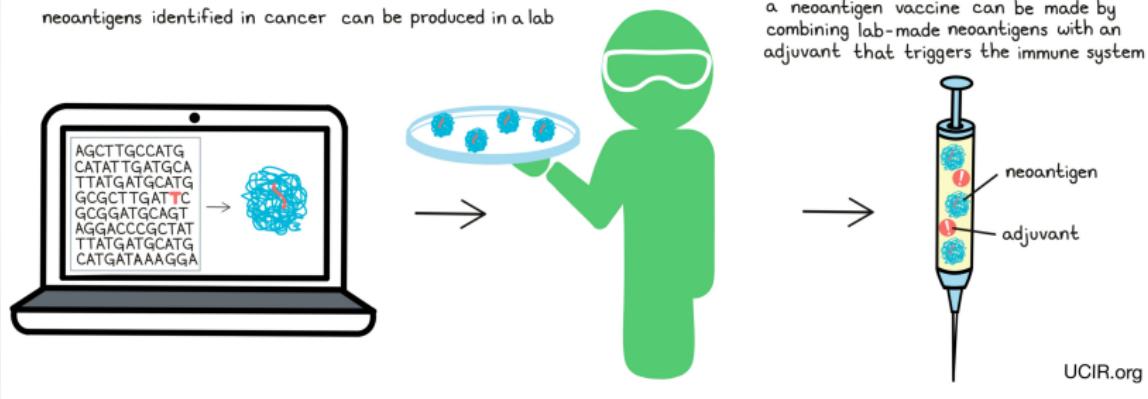


Figure: Proceso para la generación de vacunas contra el cáncer. Fuente: [9].

# Contexto y Motivación

## Vacunas personalizadas

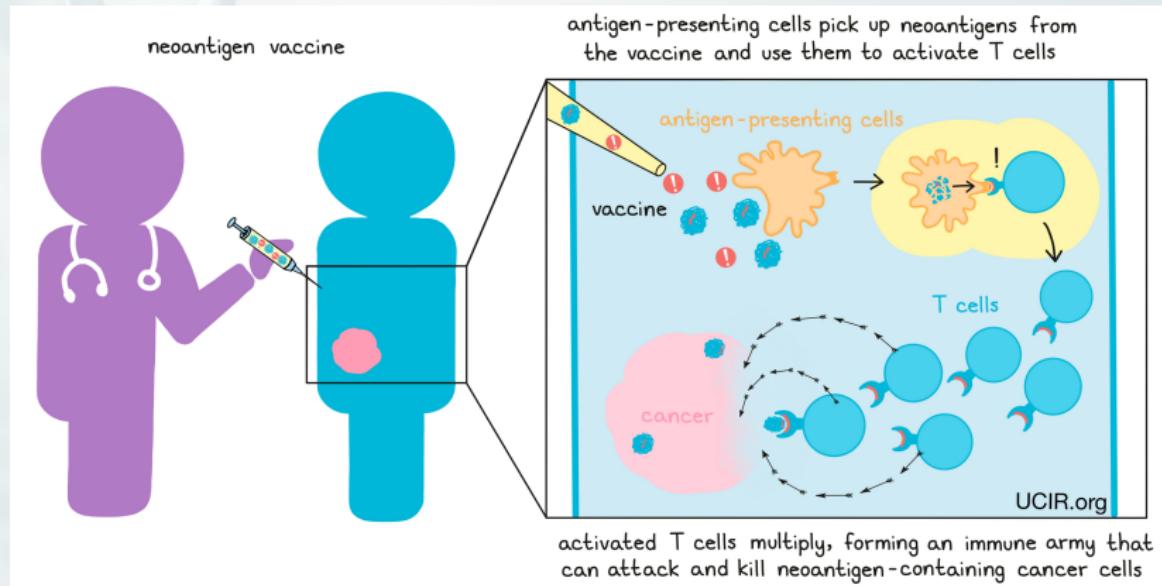
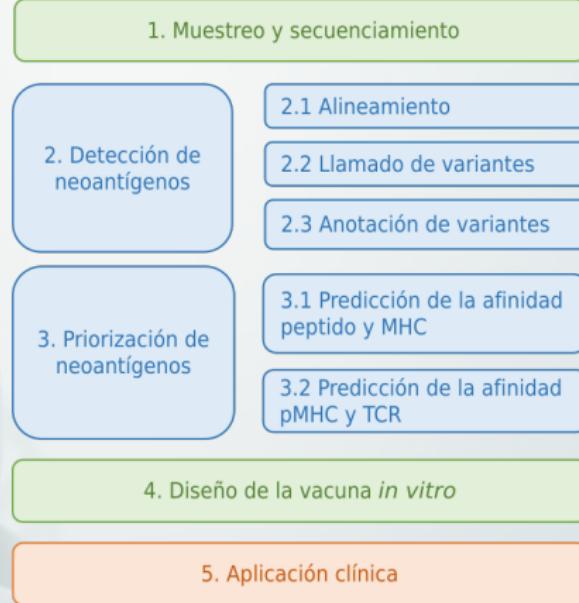


Figure: Proceso para la generación de vacunas contra el cáncer. Fuente: [9].

# 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

- Problema
- Objetivos

## Estado del arte

## Propuesta

## Experimentos y Resultados

- Base de datos
- Modelos pre-entrenados
- Resultados

## Discusión y Conclusiones

# Problema



**Menos del 5% de neoantígenos detectados activan el sistema inmune [11, 12, 13, 14, 15].**

# Problema

**Menos del 5% de neoantígenos detectados activan el sistema inmune [11, 12, 13, 14, 15].**

- ▶ La no inclusión en conjunto de varias fuentes de información como DNA-seq, RNA-seq, y datos de MS [16].

# Problema

**Menos del 5% de neoantígenos detectados activan el sistema inmune [11, 12, 13, 14, 15].**

- ▶ La no inclusión en conjunto de varias fuentes de información como DNA-seq, RNA-seq, y datos de MS [16].
- ▶ Uso herramientas de bajo desempeño para la predicción del enlace péptido-MHC (pMHC). La mayoría de aplicaciones, se basa en el uso de MHCFlurry [17] y NetMHCpan4.1 [18].

# Problema

**Menos del 5% de neoantígenos detectados activan el sistema inmune [11, 12, 13, 14, 15].**

- ▶ La no inclusión en conjunto de varias fuentes de información como DNA-seq, RNA-seq, y datos de MS [16].
- ▶ Uso herramientas de bajo desempeño para la predicción del enlace péptido-MHC (pMHC). La mayoría de aplicaciones, se basa en el uso de MHCFlurry [17] y NetMHCpan4.1 [18].
- ▶ No consideran la predicción del enlace pMHC-TCR [19].

# Problema

**Menos del 5% de neoantígenos detectados activan el sistema inmune [11, 12, 13, 14, 15].**

- ▶ La no inclusión en conjunto de varias fuentes de información como DNA-seq, RNA-seq, y datos de MS [16].
- ▶ Uso herramientas de bajo desempeño para la predicción del enlace péptido-MHC (pMHC). La mayoría de aplicaciones, se basa en el uso de MHCFlurry [17] y NetMHCpan4.1 [18].
- ▶ No consideran la predicción del enlace pMHC-TCR [19].
- ▶ No utilizar información de eventos de *alternative splicing*, variaciones estructurales y fusión de genes [20].

# Problema

## Formulación del problema



18

Es un problema de clasificación binaria que toma como entrada la secuencia de aminoácidos de un péptido ( $p = \{A, \dots, Q\}$ ) y el MHC ( $q = \{A, N, \dots, G\}$ ). Finalmente, necesitamos conocer la probabilidad de afinidad entre  $p$  y  $q$ .



**Figure:** Problema de predicción del enlace pMHC.



## Objetivo general

Implementar un método *in silico* basado en *Transformers* y *Transfer Learning* para la detección de neoantígenos, enfocados en la predicción de la unión pMHC.



## Objetivos específicos

- ▶ Analizar los métodos que utilizan *Transformers* para la predicción del enlace pMHC en el contexto de detección de neoantígenos.



## Objetivos específicos

- ▶ Analizar los métodos que utilizan *Transformers* para la predicción del enlace pMHC en el contexto de detección de neoantígenos.
- ▶ Analizar los modelos basados en *Transformers* TAPE, ProtBert-BFD, y EMS2 pre-entredados para diversas tareas en Proteómica y de los cuáles se puede aplicar *Transfer Learning*.



## Objetivos específicos

- ▶ Analizar los métodos que utilizan *Transformers* para la predicción del enlace pMHC en el contexto de detección de neoantígenos.
- ▶ Analizar los modelos basados en *Transformers* TAPE, ProtBert-BFD, y EMS2 pre-entredados para diversas tareas en Proteómica y de los cuáles se puede aplicar *Transfer Learning*.
- ▶ Implementar *fine-tuning* a los modelos TAPE, ProtBert-BFD, y EMS2 para la tarea de predicción del enlace pMHC, aplicando *Gradient Accumulation Steps* (GAS) y una metodología de congelamiento de capas.



## Objetivos específicos

- ▶ Analizar los métodos que utilizan *Transformers* para la predicción del enlace pMHC en el contexto de detección de neoantígenos.
- ▶ Analizar los modelos basados en *Transformers* TAPE, ProtBert-BFD, y EMS2 pre-entredados para diversas tareas en Proteómica y de los cuáles se puede aplicar *Transfer Learning*.
- ▶ Implementar *fine-tuning* a los modelos TAPE, ProtBert-BFD, y EMS2 para la tarea de predicción del enlace pMHC, aplicando *Gradient Accumulation Steps* (GAS) y una metodología de congelamiento de capas.
- ▶ Comparar los modelos de mejor desempeño con las herramientas del estado del arte como: NetMHCpan4.1, MHCFlurry2.0, Anthem, ACME y MixMHCpred2.2.

# Contenido



## Contexto y Motivación

Estadísticas en Cáncer

Inmunoterapia del Cáncer

Vacunas Personalizadas

## Problema y Objetivos

Problema

Objetivos

## Estado del arte

## Propuesta

## Experimentos y Resultados

Base de datos

Modelos pre-entrenados

Resultados

## Discusión y Conclusiones

# Estado del arte



# Contenido



## Contexto y Motivación

- Estadísticas en Cáncer
- Inmunoterapia del Cáncer
- Vacunas Personalizadas

## Problema y Objetivos

- Problema
- Objetivos

## Estado del arte

## Propuesta

### Experimentos y Resultados

- Base de datos
- Modelos pre-entrenados
- Resultados

## Discusión y Conclusiones

# Propuesta



**Figure:** Propuesta para la predicción del enlace pMHC.

# Propuesta

*Fine-tuning*



# Propuesta

*Gradient Accumulation Steps*



# Propuesta

*Congelamiento de capas*



27

# Contenido



## Contexto y Motivación

- Estadísticas en Cáncer
- Inmunoterapia del Cáncer
- Vacunas Personalizadas

## Problema y Objetivos

- Problema
- Objetivos

## Estado del arte

## Propuesta

## Experimentos y Resultados

- Base de datos
- Modelos pre-entrenados
- Resultados

## Discusión y Conclusiones

# Base de datos

*Training: 539,019; Validation: 179,673; y Testing: 172,580.*

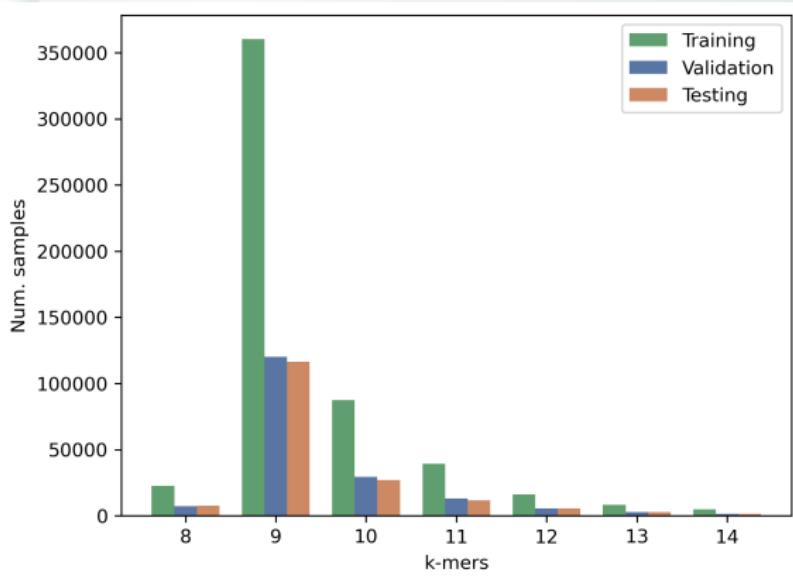


Figure: Número de muestras por  $k$ -mer.

# Modelos pre-entrenados

**Table:** Diferencias entre TAPE, ProtBert-DFB, y ESM2. HS: *Hidden size*; AH: *Attention heads*.

Modelo	BD	Muestras	Capas	HS	AH	Params.
TAPE	Pfam	30M	12	768	12	92M
ProtBert-BFD	BFD	2122M	30	1024	16	420M
ESM2(t6)	Uniref50	60M	6	320	20	8M
ESM2(t12)	Uniref50	60M	12	480	20	35M
ESM2(t30)	Uniref50	60M	30	640	20	150M
ESM2(t33)	Uniref50	60M	33	1280	20	650M

# Resultados

Entrenamiento por 3 epochs

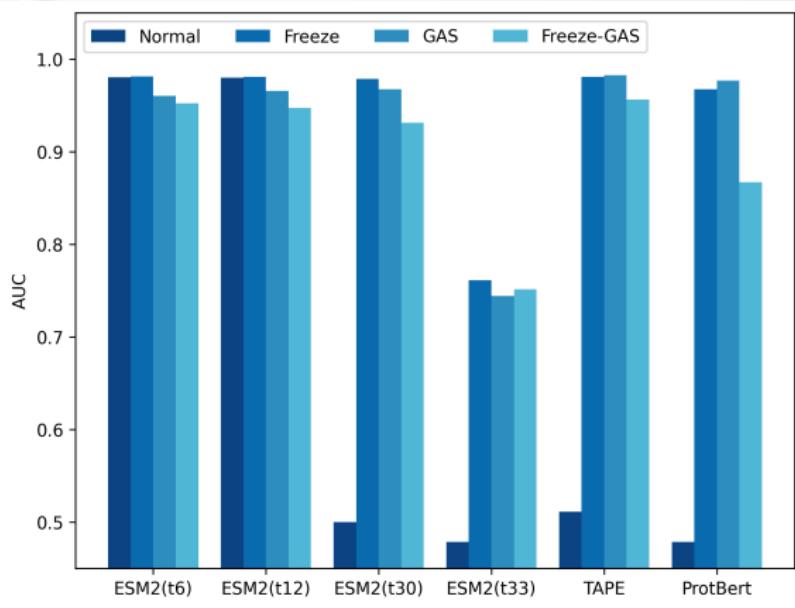
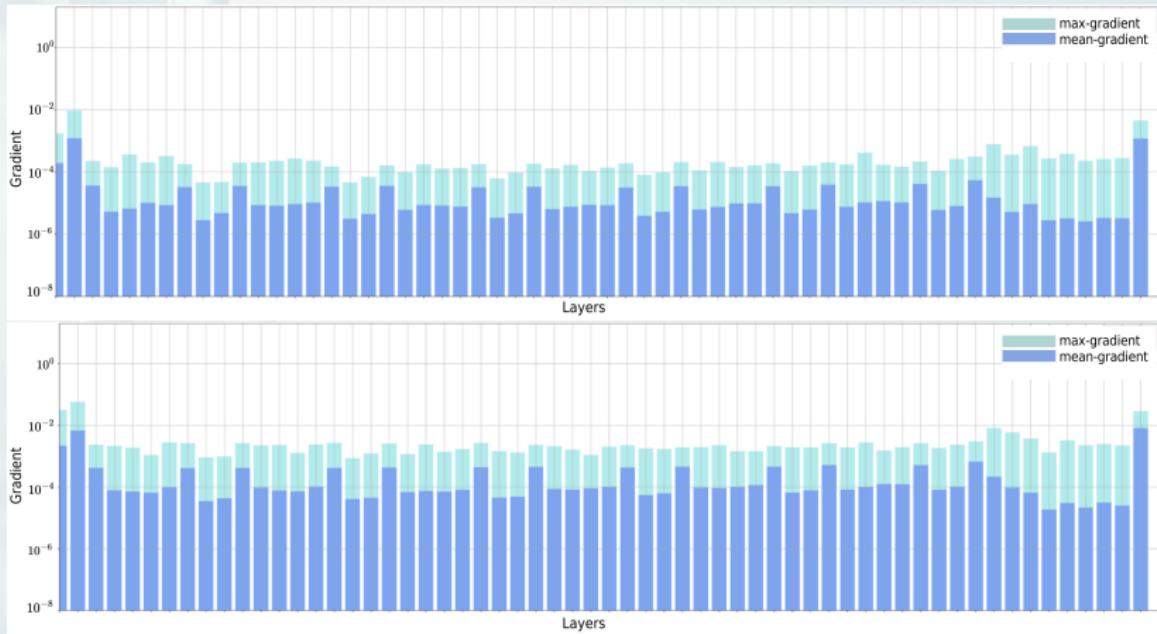


Figure: Comparación del AUC por modelo y metodología de entrenamiento.

# Resultados

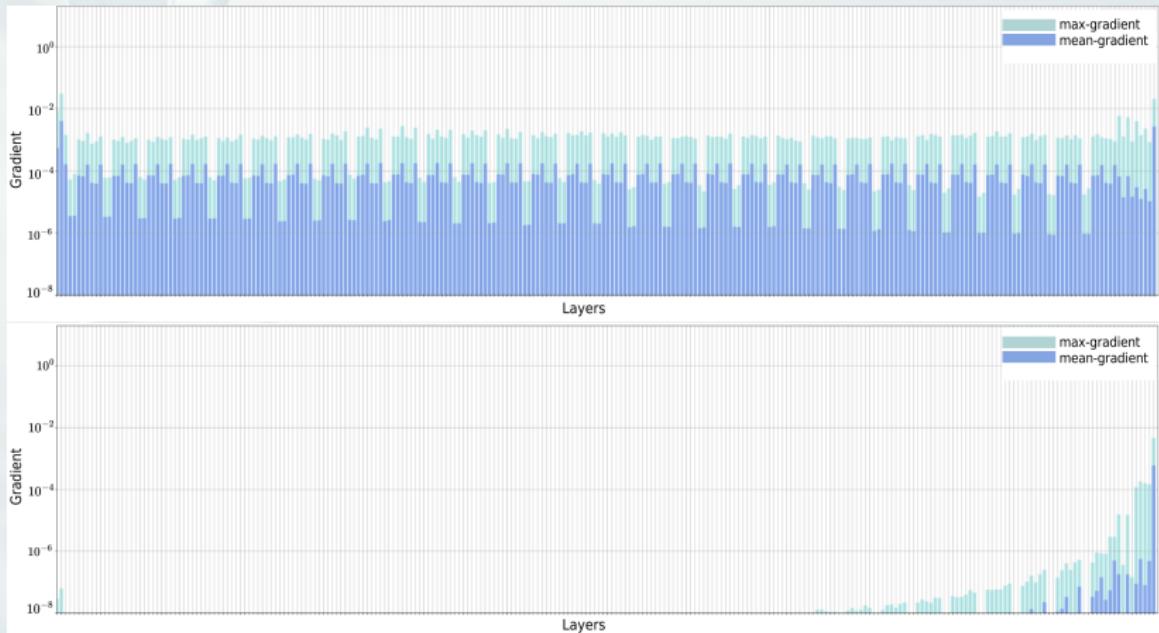
Problema de *vanish gradient* para ESM2(t6)

32



# Resultados

Problema de *vanish gradient* para ESM2(t30)



# Resultados

Entrenamiento por 30 epochs



	Accuracy	Precision	Recall	F1-score	AUC	MCC
ESM2(t6)-Normal	0.9390	0.9333	<b>0.9453</b>	0.9392	0.9797	0.8780
ESM2(t6)-Freeze	<b>0.9401</b>	<b>0.9398</b>	0.9402	<b>0.9400</b>	<b>0.9830</b>	<b>0.8802</b>
ESM2(t6)-GAS	0.9366	0.9322	0.9413	0.9368	0.9818	0.8732
ESM2(t6)-Freeze-GAS	0.9354	0.9326	0.9383	0.9355	0.9813	0.8708
ESM2(t30)-Normal	-	-	-	-	-	-
ESM2(t30)-Freeze	<b>0.9393</b>	0.9304	<b>0.9493</b>	<b>0.9397</b>	0.9787	<b>0.8787</b>
ESM2(t30)-GAS	0.9346	<b>0.9337</b>	0.9352	0.9345	0.9808	0.8691
ESM2(t30)-Freeze-GAS	0.9363	0.9319	0.9411	0.9365	<b>0.9818</b>	0.8726
TAPE-Normal	-	-	-	-	-	-
TAPE-Freeze	0.9395	<b>0.9404</b>	0.9382	0.9393	0.9815	0.8790
TAPE-GAS	<b>0.9415</b>	0.9352	<b>0.9484</b>	<b>0.9418</b>	<b>0.9841</b>	<b>0.8831</b>
TAPE-Freeze-GAS	0.9359	0.9297	0.9428	0.9362	0.9820	0.8719

# Resultados

Comparación con los métodos *state-of-art*

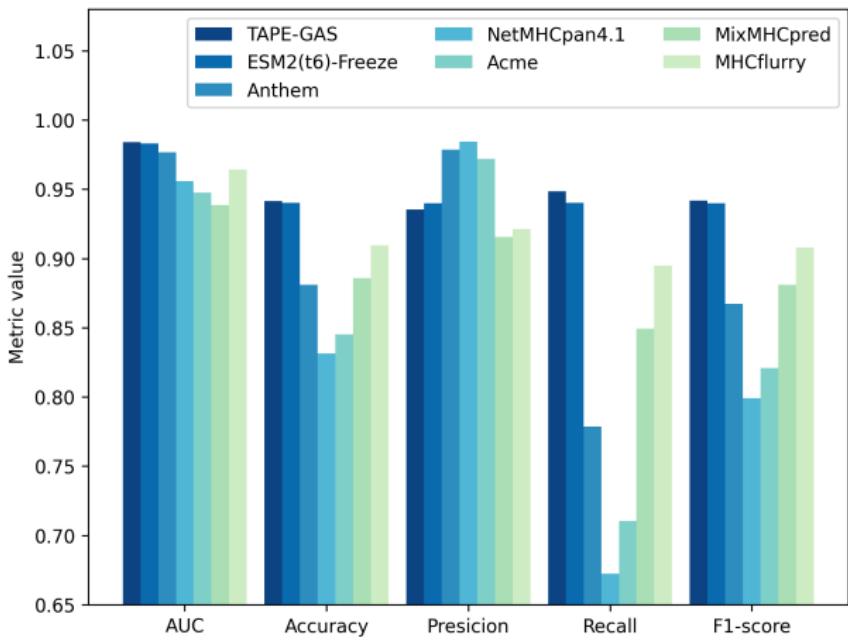
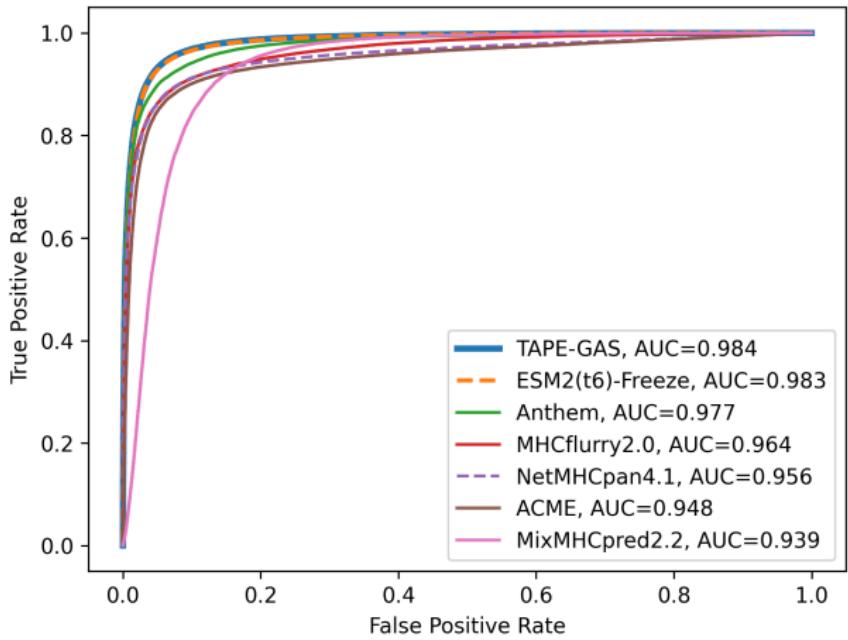


Figure: Comparación de TAPE-GAS y ESM2(t6) contra los mejores métodos del estado del arte.

# Resultados

Comparación con los métodos *state-of-art*



**Figure:** Comparación de TAPE-GAS y ESM2(t6) contra los mejores métodos del estado del arte.

# Resultados

Comparación con los métodos *state-of-art*



**Table:** Desempeño de TAPE-GAS y ESM2(t6)-Freeze, entrenados por 30 epochs, contra Anthem, NetMHCpan4.1, ACME, MixMHCpred2.2, y MhcFlurry2.0.

	Accuracy	Precision	Recall	F1-score	AUC	MCC
TAPE-GAS	<b>0.9415</b>	0.9352	<b>0.9484</b>	<b>0.9418</b>	<b>0.9841</b>	<b>0.8831</b>
ESM2(t6)-Freeze	<b>0.9401</b>	0.9398	<b>0.9402</b>	<b>0.9400</b>	<b>0.9830</b>	<b>0.8802</b>
Anthem	0.8811	<b>0.9786</b>	0.7787	0.8673	0.9768	0.7785
NetMHCpan4.1	0.8312	<b>0.9844</b>	0.6724	0.7991	0.9557	0.6982
ACME	0.8452	0.9717	0.7105	0.8208	0.9476	0.7165
MixMHCpred2.2	0.8857	0.9155	0.8493	0.8811	0.9386	0.7733
MhcFlurry2.0	0.9093	0.9211	0.8948	0.9078	0.9642	0.8189

# Contenido



## Contexto y Motivación

- Estadísticas en Cáncer
- Inmunoterapia del Cáncer
- Vacunas Personalizadas

## Problema y Objetivos

- Problema
- Objetivos

## Estado del arte

## Propuesta

## Experimentos y Resultados

- Base de datos
- Modelos pre-entrenados
- Resultados

## Discusión y Conclusiones

# Discusión

¿Porque el modelo mas pequeño de la familia ESM2 es el mejor?



El modelo más pequeño, ESM2(t6) supero a los demás. Las causas de este fenómeno pueden ser:

# Discusión

¿Porque el modelo mas pequeño de la familia ESM2 es el mejor?



El modelo más pequeño, ESM2(t6) supero a los demás. Las causas de este fenómeno pueden ser:

- ▶ En conjunto de datos que consta de 559,019 muestras, que no consideramos lo suficientemente grande para ESM2(t33), un modelo que cuenta con 650 millones de parámetros.

# Discusión

¿Porque el modelo mas pequeño de la familia ESM2 es el mejor?



El modelo más pequeño, ESM2(t6) supero a los demás. Las causas de este fenómeno pueden ser:

- ▶ En conjunto de datos que consta de 559,019 muestras, que no consideramos lo suficientemente grande para ESM2(t33), un modelo que cuenta con 650 millones de parámetros.
- ▶ El uso de *Rotary Position Embedding* (RoPE) en lugar de la codificación posicional absoluta. Si bien RoPE puede llevar a un ligero aumento en el costo de entrenamiento, se ha observado que mejora la calidad de los resultados, especialmente para modelos más pequeños [21].

# Discusión

## Congelamiento de capas y GAS



### Congelamiento de capas

Para los modelos ESM2, esta metodología arrojó los mejores resultados, mientras que para TAPE y ProtBert-BFD, produjo los resultados esperados



### Congelamiento de capas

Para los modelos ESM2, esta metodología arrojó los mejores resultados, mientras que para TAPE y ProtBert-BFD, produjo los resultados esperados

### GAS

Esta técnica alivia ligeramente el problema de *vanish gradients* acumulando las gradientes durante las iteraciones. En si, este técnica extiende la cantidad de iteraciones de entrenamiento que se pueden realizar antes de que el modelo posiblemente vuelva a enfrentar el problema de *vanish gradient*.



### ProtBert-BFD

ProtBert-BFD (420M parámetros) obtuvo el peor resultado a pesar de que este modelo fue **pre-entrenado con el conjunto de datos más grande (2122 millones de muestras)**. Las causas son:

- ▶ Ruido en las muestras y a los errores en las secuencias en el conjunto de datos BFD [22]
- ▶ Los modelos *Transformer* grandes requieren más datos para el entrenamiento [22], y en nuestro caso este modelo se entreno con 559,019 muestras.



### TAPE

TAPE logró los mejores resultados. Este solo fue pre-entrenado con 30 millones de muestras; sin embargo, secuencias pertenecen a *Reference Proteomes* en lugar de abarcar toda la base de datos de UniProtKB [23].

### ESM2

ESM2(t6) logró resultados que compiten estrechamente con el desempeño de TAPE; sin embargo ESM2(t6) tiene 8M versus los 92M de parámetros de TAPE.

# 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] 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.

# References IV



- [12] Nil Adell Mill, Cedric Bogaert, Wim van Crielinge, and Bruno Fant,  
“neoms: Attention-based prediction of mhc-i epitope presentation,”  
*bioRxiv*, 2022.
- [13] Brendan Bulik-Sullivan, Jennifer Busby, Christine D Palmer, Matthew J Davis, Tyler Murphy, Andrew Clark, Michele Busby, Fujiko Duke, Aaron Yang, Lauren Young, et al.,  
“Deep learning using tumor hla peptide mass spectrometry datasets improves neoantigen identification,”  
*Nature biotechnology*, vol. 37, no. 1, pp. 55–63, 2019.

# References V



- [14] Michal Bassani-Sternberg, Sune Pletscher-Frankild, Lars Juhl Jensen, and Matthias Mann,  
“Mass spectrometry of human leukocyte antigen class i peptidomes reveals strong effects of protein abundance and turnover on antigen presentation\*[s],”  
*Molecular & Cellular Proteomics*, vol. 14, no. 3, pp. 658–673, 2015.
- [15] Mahesh Yadav, Suchit Jhunjhunwala, Qui T Phung, Patrick Lupardus, Joshua Tanguay, Stephanie Bumbaca, Christian Franci, Tommy K Cheung, Jens Fritzsche, Toni Weinschenk, et al.,  
“Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing,”  
*Nature*, vol. 515, no. 7528, pp. 572–576, 2014.

# References VI



- [16] 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.
- [17] Timothy J O'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.

## References VII



49

- [18] Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen,  
“Netmhcpant-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.
- [19] 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 VIII



- [20] 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.
- [21] 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.

# References IX



- [22] 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.
- [23] Robert D Finn, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Jaina Mistry, Alex L Mitchell, Simon C Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, et al.,  
“The pfam protein families database: towards a more sustainable future,”  
*Nucleic acids research*, vol. 44, no. D1, pp. D279–D285, 2016.

