

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

Estado del arte

Propuesta

Experimentos y Resultados

Base de datos

Modelos pre-entrenados

Resultados

Discusión

Conclusiones

Trabajos Futuros

Contribuciones y Publicaciones

Contenido



Contexto y Motivación

Estadísticas en Cáncer

Inmunoterapia del Cáncer

Vacunas Personalizadas

Problema y Objetivos

Estado del arte

Propuesta

Experimentos y Resultados

Base de datos

Modelos pre-entrenados

Resultados

Discusión

Conclusiones

Trabajos Futuros

Contribuciones y Publicaciones

Contexto y Motivación

3

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

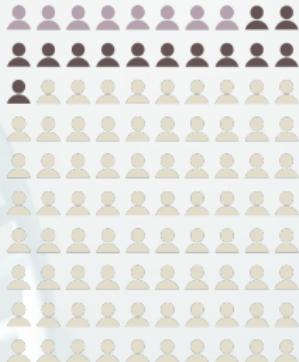
Contexto y Motivación

Porcentaje de casos y muertes



■ Developing cancer ■ Dying from cancer

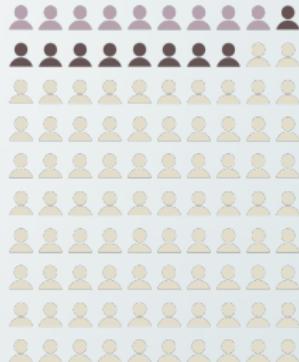
MALE



21% of males
worldwide develop cancer
during their lifetime

13% of males
worldwide die from the disease

FEMALE



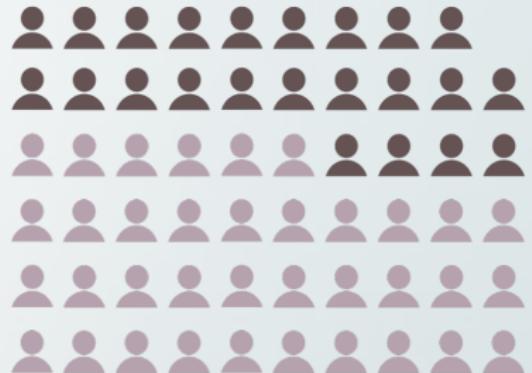
18% of females
worldwide develop cancer
during their lifetime

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].

Inmunoterapia del Cáncer



6

Es un tipo de tratamiento contra el Cáncer que estimula las defensas naturales del cuerpo para combatir el Cáncer [3].

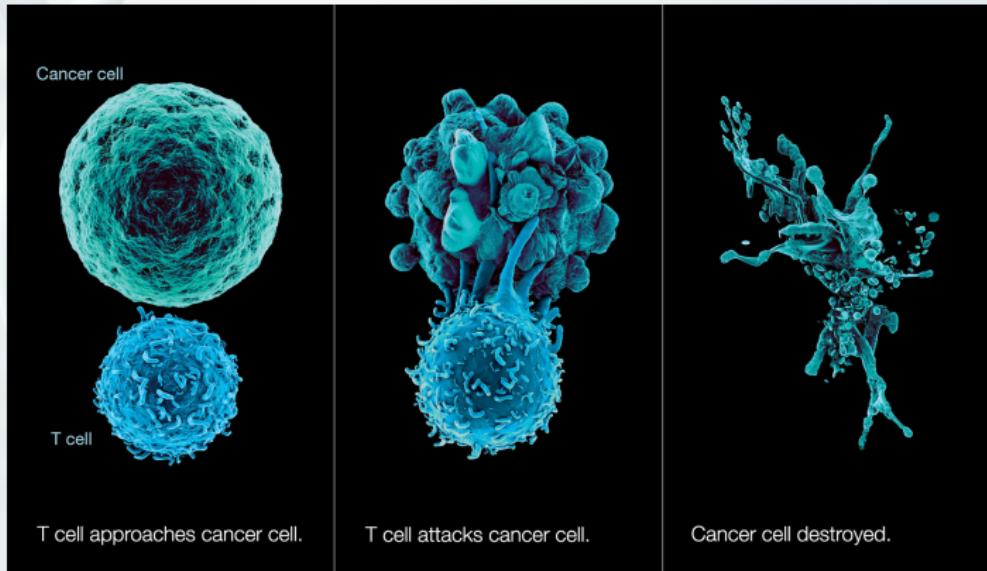


Figure: Ejemplo de como una célula T destruye células del cancer [4].

Contexto y Motivación

Inmunoterapia del Cáncer

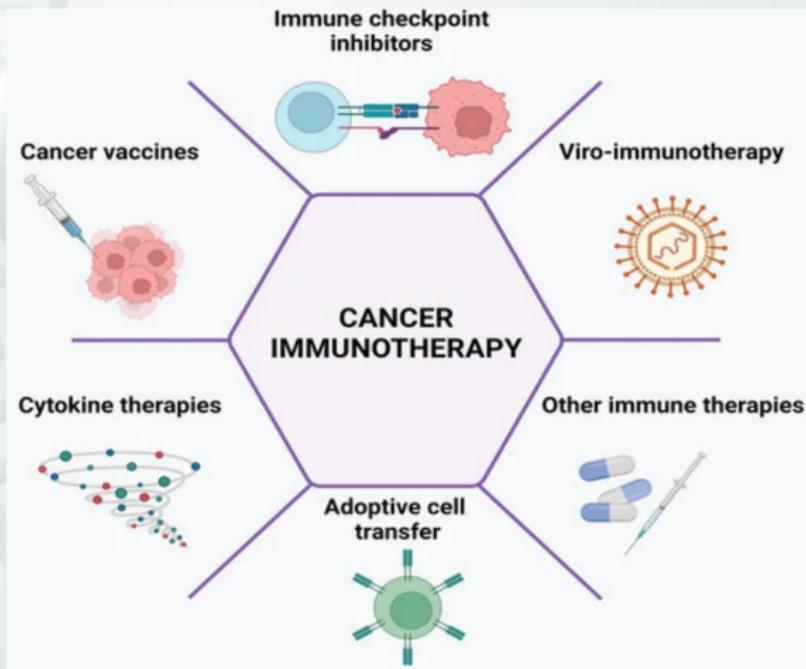


Figure: Tipos de tratamientos para la inmunoterapia del cáncer. Fuente: [5].

Contexto y Motivación

Neoantígenos



8

Es una **proteína** que se forma en las células de Cáncer cuando ocurre mutaciones en el DNA [6, 7].

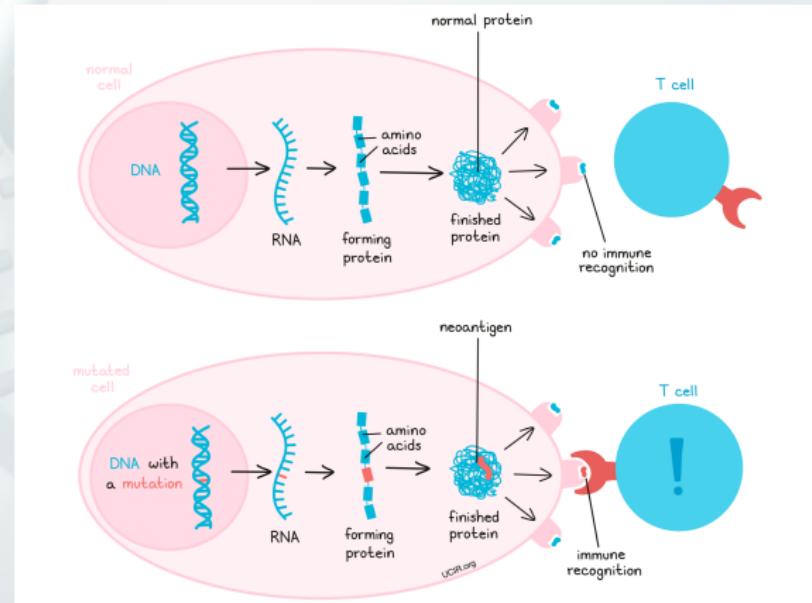


Figure: Neoantígenos y células T. Fuente: [8].

Contexto y Motivación

Vacunas personalizadas

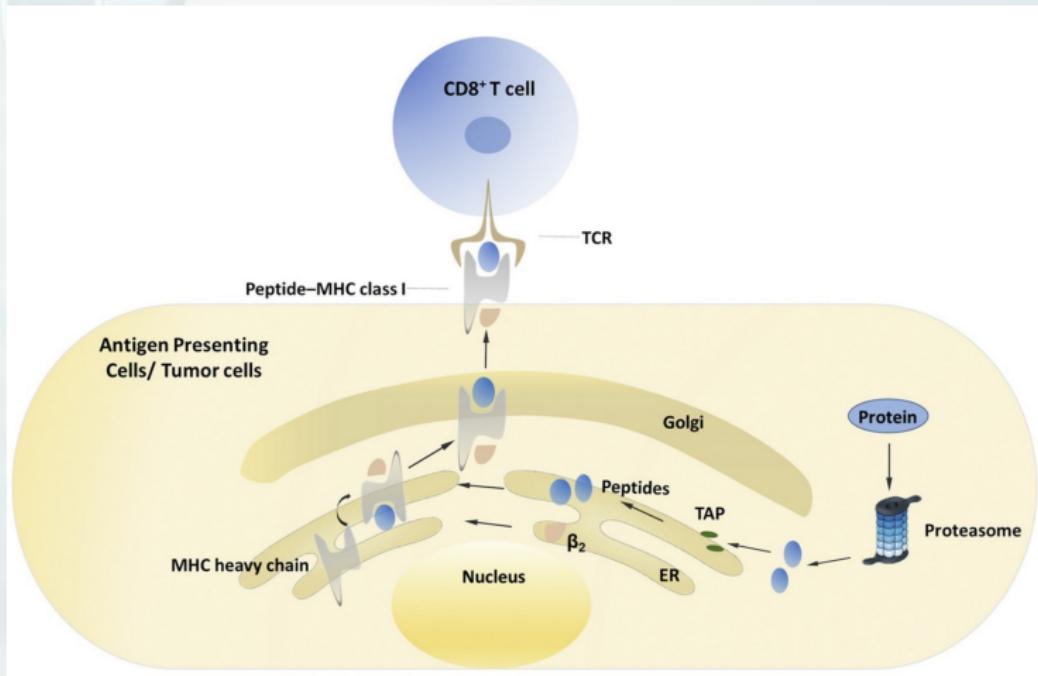


Figure: Presentación de antígenos por MHC-I. Fuente: [9]

Contexto y Motivación

Vacunas personalizadas



10

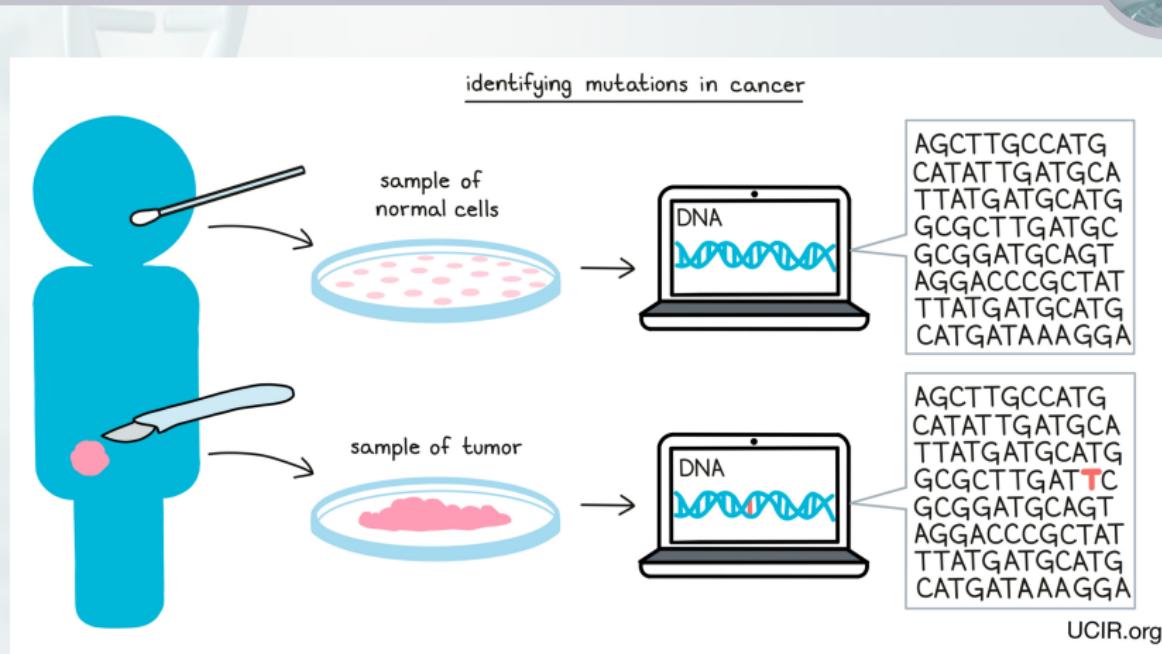


Figure: Proceso para la generación de vacunas contra el cáncer. Fuente: [8].

Contexto y Motivación

Vacunas personalizadas

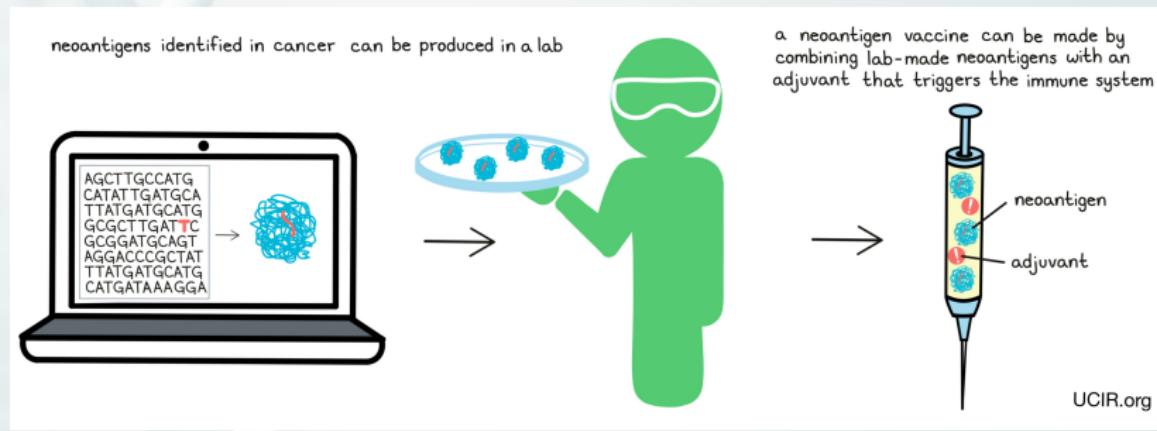


Figure: Proceso para la generación de vacunas contra el cáncer. Fuente: [8].

Contexto y Motivación

Vacunas personalizadas

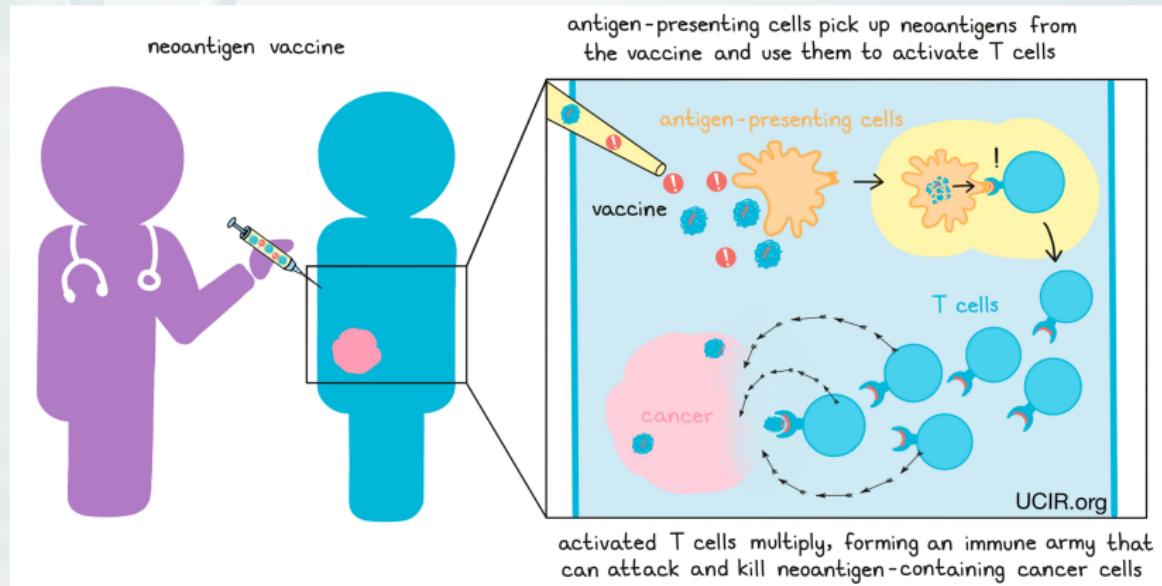


Figure: Proceso para la generación de vacunas contra el cáncer. Fuente: [8].

Contexto y Motivación

Vacunas personalizadas



13

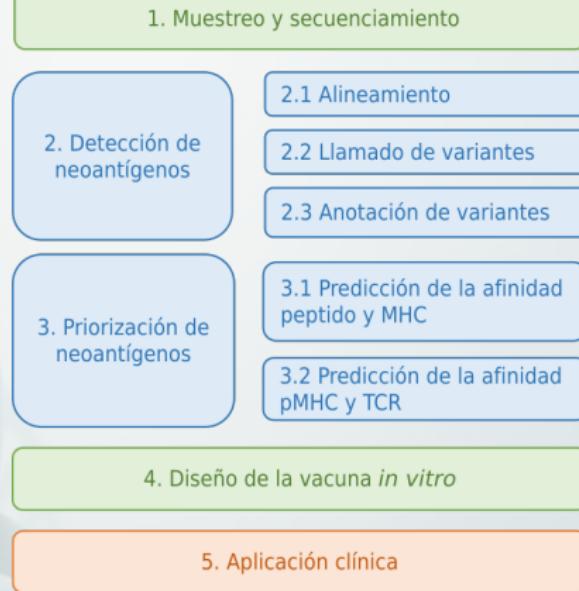


Figure: Resumen del proceso de generación de vacunas contra el cáncer.

Contenido



Contexto y Motivación

Estadísticas en Cáncer

Inmunoterapia del Cáncer

Vacunas Personalizadas

Problema y Objetivos

Estado del arte

Propuesta

Experimentos y Resultados

Base de datos

Modelos pre-entrenados

Resultados

Discusión

Conclusiones

Trabajos Futuros

Contribuciones y Publicaciones

Problema

Menos del 5% de neoantígenos detectados activan el sistema inmune [10, 11, 12, 13, 14].

Problema

Menos del 5% de neoantígenos detectados activan el sistema inmune [10, 11, 12, 13, 14].

- ▶ La no inclusión en conjunto de varias fuentes de información como DNA-seq, RNA-seq, y datos de MS [15].

Problema

Menos del 5% de neoantígenos detectados activan el sistema inmune [10, 11, 12, 13, 14].

- ▶ La no inclusión en conjunto de varias fuentes de información como DNA-seq, RNA-seq, y datos de MS [15].
- ▶ Uso herramientas de bajo desempeño para la predicción del enlace péptido-MHC (pMHC). La mayoría de aplicaciones, se basa en el uso de MHCFlurry [16] y NetMHCpan4.1 [17].

Problema

Menos del 5% de neoantígenos detectados activan el sistema inmune [10, 11, 12, 13, 14].

- ▶ La no inclusión en conjunto de varias fuentes de información como DNA-seq, RNA-seq, y datos de MS [15].
- ▶ Uso herramientas de bajo desempeño para la predicción del enlace péptido-MHC (pMHC). La mayoría de aplicaciones, se basa en el uso de MHCFlurry [16] y NetMHCpan4.1 [17].
- ▶ No consideran la predicción del enlace pMHC-TCR [18].

Problema

Menos del 5% de neoantígenos detectados activan el sistema inmune [10, 11, 12, 13, 14].

- ▶ La no inclusión en conjunto de varias fuentes de información como DNA-seq, RNA-seq, y datos de MS [15].
- ▶ Uso herramientas de bajo desempeño para la predicción del enlace péptido-MHC (pMHC). La mayoría de aplicaciones, se basa en el uso de MHCFlurry [16] y NetMHCpan4.1 [17].
- ▶ No consideran la predicción del enlace pMHC-TCR [18].
- ▶ No utilizar información de eventos de *alternative splicing*, variaciones estructurales y fusión de genes [19].

Problema

Formulación del problema



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

Estado del arte

Propuesta

Experimentos y Resultados

Base de datos

Modelos pre-entrenados

Resultados

Discusión

Conclusiones

Trabajos Futuros

Contribuciones y Publicaciones

Estado del arte

Publicaciones sobre estudios de Neoantígenos

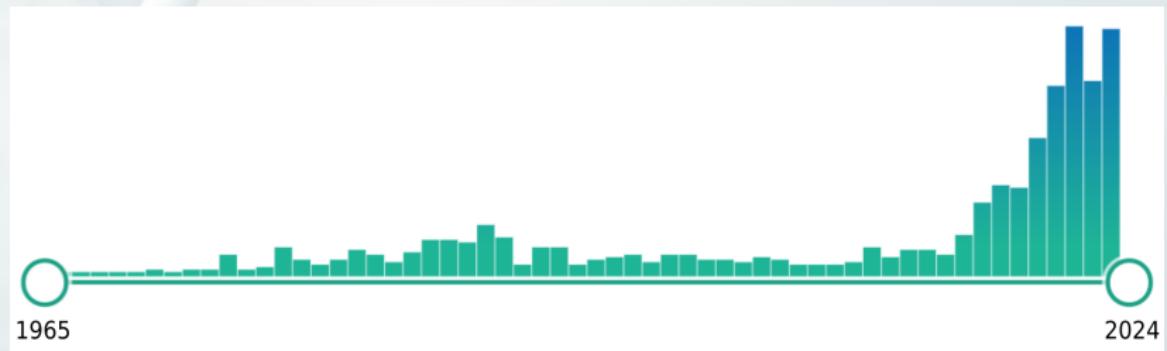


Figure: Publicaciones sobre estudios de Neoantígenos. Fuente: PubMed.



Table: Preguntas de investigación

Preguntas de investigación

- Q1.** ¿Como se aplican los modelos *Transformers* para la detección de neoantígenos?
- Q2.** ¿Que problemas y limitaciones hacen frente los modelos *Transformers* en la detección de neoantígenos?
- Q3.** ¿Que *pipelines* se han desarrollado para la detección de neoantígenos?
- Q4.** ¿Que pruebas clínicas, de vacunas personalizadas de neoantígenos, han sido aplicadas?

Estado del arte

Cadena de búsqueda

Table: Cadenas de búsqueda utilizadas para cada fase de detección de neoantígenos.

Categoría	Cadena de búsqueda
Priorización de neoantígenos	(mhc OR hla) AND (peptide OR epitope OR antigen) AND (specificity OR immunogenicity OR binding OR affinity OR predict* OR detection OR presentation OR classification) AND (transformer* OR bert* OR attention OR 'transfer learning' OR method* OR predict*), (tcr OR 't cell' OR t-cell) AND (mhc OR peptide OR epitope OR antigen) AND (specificity OR immunogenicity OR binding OR affinity OR predict* OR detection OR presentation OR classification) AND (transformer* OR bert* OR attention OR 'transfer learning' OR method* OR predict*)
Pipelines	(pipeline OR toolkit) AND (tcr OR 't cell' OR t-cell OR mhc OR hla OR peptide OR epitope OR antigen* OR neoantigen*) (pipeline OR tool* OR workflow OR application OR web*) AND (peptide OR epitope OR antigen* OR neoantigen* OR neoepito*) AND (immunotherapy OR detection OR identify* OR predict* OR presentation*)
Ensayos clínicos	(neoantigen OR neoepitope OR dendritic cell) AND (vaccines OR immunology)



Table: Bases de datos utilizadas en la RSL.

Bases de datos

PubMed

Scopus

Google scholar

Web of Science

Estado del arte

Criterios de inclusión



Table: Criterios de inclusión y exclusión.

Criterio de inclusión	Criterio de exclusión
Artículos en revistas Q1 o Q2 según Scimago o Conferencias ERA (A o B).	Artículos de Conferencias de bajo nivel.
Artículos publicados desde el 2018.	Publicaciones que no son del área de computación o Bioinformática.
Artículos que hagan uso de <i>Transformers</i> o <i>Deep Learning</i> con mecanismos de atención.	



Se realizo un filtro por título, obteniendo un total de **151 artículos**. Luego, se seleccionó un subconjunto en función de los criterios de inclusión, y finalmente, se revisó el *abstract* de cada uno para llegar a un número de **79 artículos**.

Estado del arte

Resumen



Table: Redes *Transformers* utilizadas en la predicción del enlace pMHC.

Year	Name	Model
2022[20]	HLAB	BERT de ProtBert en cascada con BiLSTM.
2022[21]	MHC RoBERTa	RoBERTa pre-entrenado y seguido de 12 <i>multi-head</i> SA y una capa lineal.
2022[22]	TransPHLA	Usa SA basado en cuatro bloques.
2021[23]	ImmunoBERT	BERT TAPE pre-entrenado y seguido de una capa lineal. Se enfocó en pMHC-I.
2021[24]	BERTMHC	BERT TAPE pre-entrenado y seguido de una capa lineal. Se enfocó en pMHC-II.

Contenido



Contexto y Motivación

Estadísticas en Cáncer

Inmunoterapia del Cáncer

Vacunas Personalizadas

Problema y Objetivos

Estado del arte

Propuesta

Experimentos y Resultados

Base de datos

Modelos pre-entrenados

Resultados

Discusión

Conclusiones

Trabajos Futuros

Contribuciones y Publicaciones

Propuesta



Figure: Propuesta para la predicción del enlace pMHC.

Propuesta

Fine-tuning

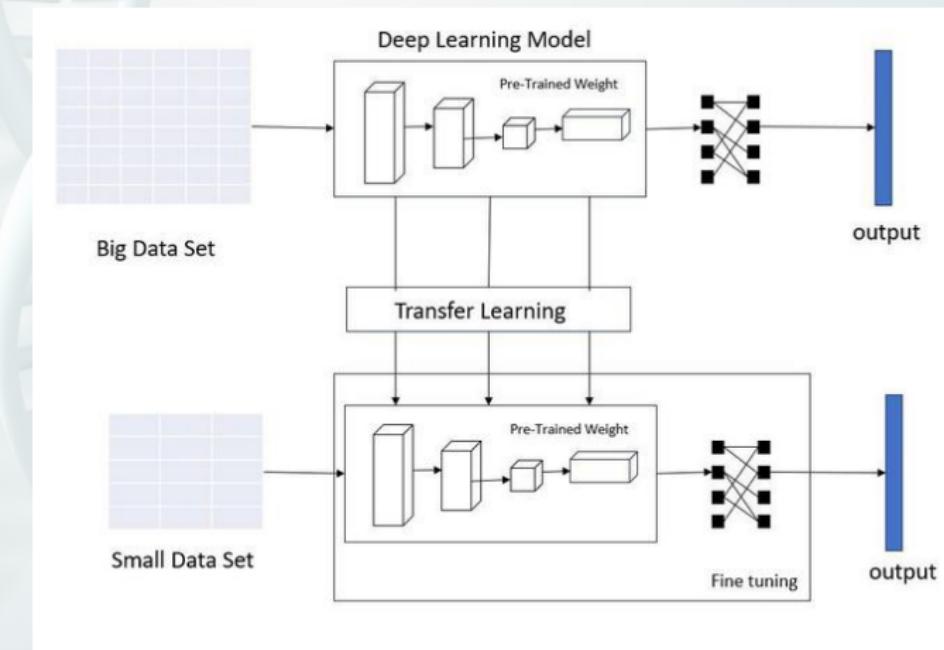


Figure: Ejemplo de *Fine-tuning*. Fuente: [25].



La arquitectura BERT, precede a un bloque BiLSTM. Este está compuesto por **dos capas con 768 unidades** (inspirados en HLAB [20]).

Se utilizaron los siguientes hiperparámetros:

- ▶ $lr = 5e^{-5}$.
- ▶ *weight decay* = 0.0001 (regularización).
- ▶ *linear warm-up steps* de 1000.
- ▶ Optimizador ADAM: ($\beta_1 = 0.9$, $\beta_2 = 0.999$).
- ▶ *Early stopping*.

Estos valores fueron utilizados por BERTMHC [24] después de buscar los mejores parámetros utilizando *grid search*.



Librerías

Se utilizó **PyTorch 1.13** y **Python 3.9** para definir los modelos de *deep learning*.

GPU

Se utilizo dos GPU:

- ▶ GPU propia: RTX3070 (8GB).
- ▶ GPU de la plataforma Paperspace.

Propuesta

Gradient Accumulation Steps

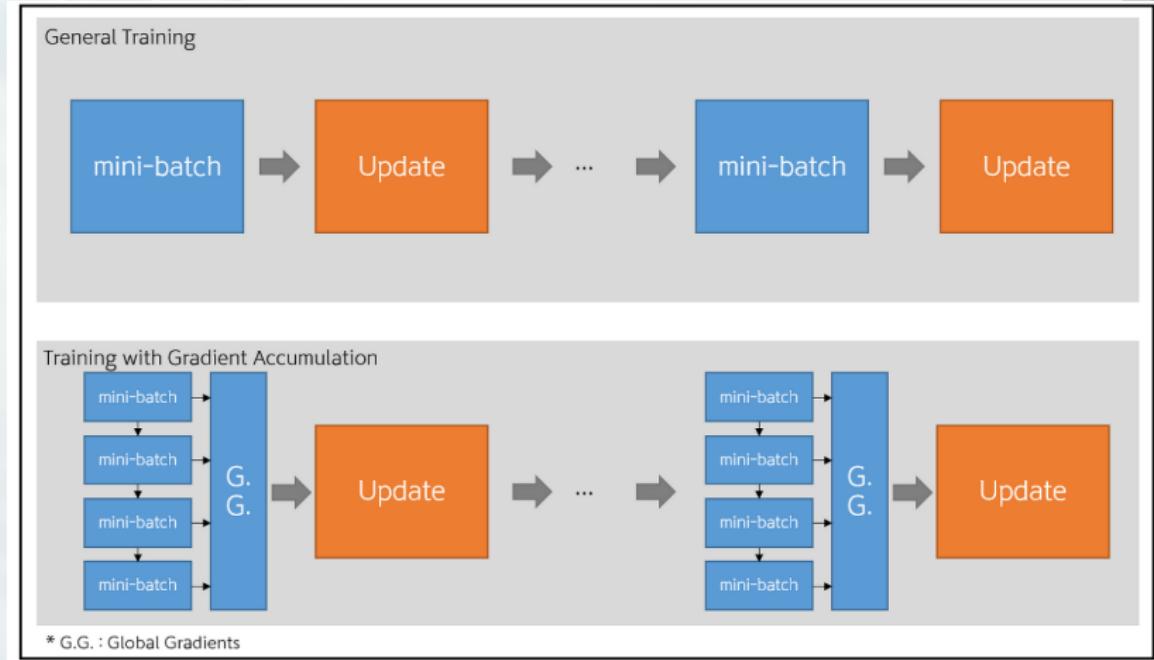


Figure: Ejemplo de GAS. Fuente: [25].

Propuesta

Gradient Accumulation Steps



$$x = x - \alpha \cdot$$

DO THE PARAMETER

UPDATE AFTER ACCUMULATION



Figure: Ejemplo de GAS. Fuente: [26].

Propuesta

Congelamiento de capas

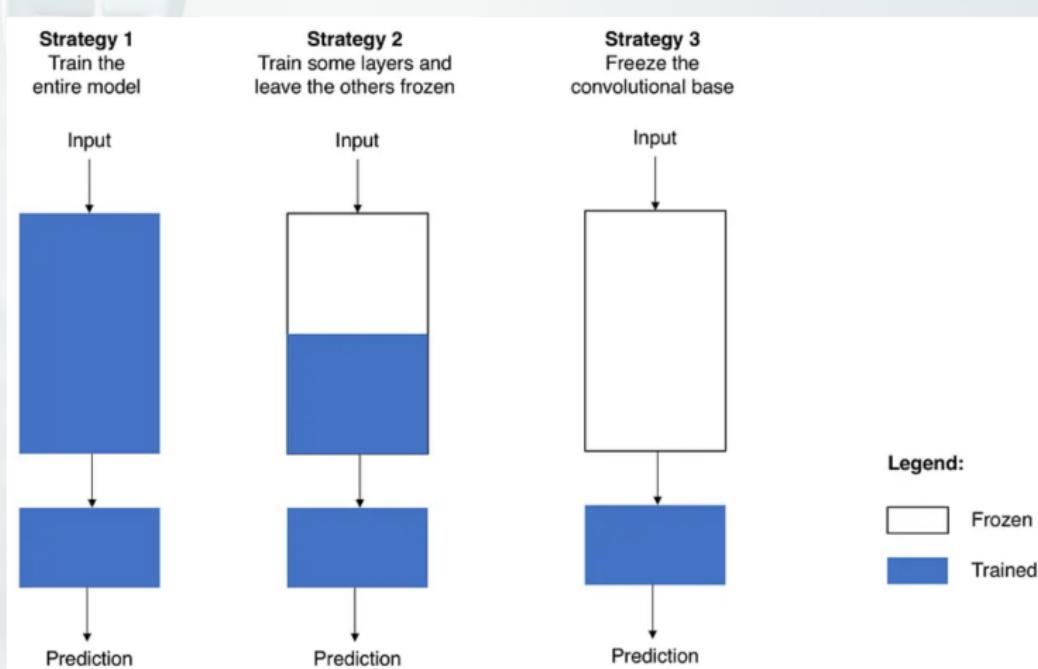


Figure: Ejemplo de congelamiento de capas. Fuente: [27].

Contenido



Contexto y Motivación

Estadísticas en Cáncer

Inmunoterapia del Cáncer

Vacunas Personalizadas

Problema y Objetivos

Estado del arte

Propuesta

Experimentos y Resultados

Base de datos

Modelos pre-entrenados

Resultados

Discusión

Conclusiones

Trabajos Futuros

Contribuciones y Publicaciones

Base de datos



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

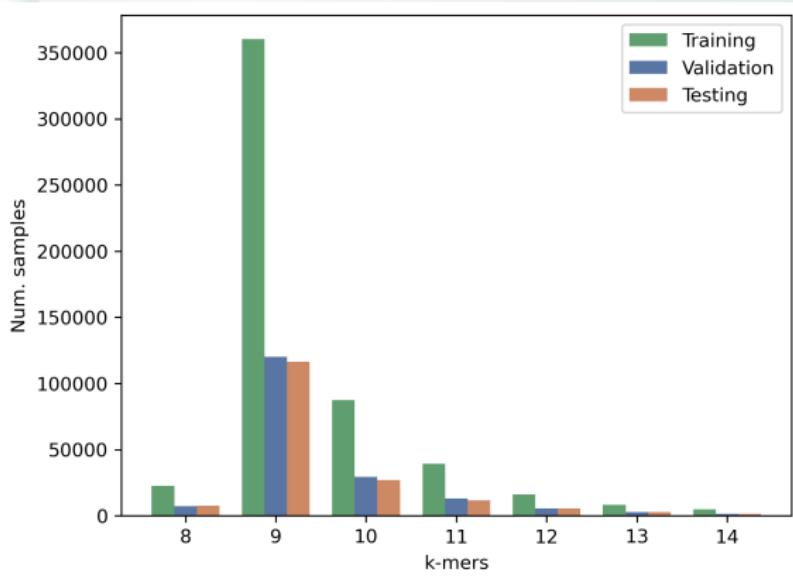


Figure: Número de muestras por k -mer.

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

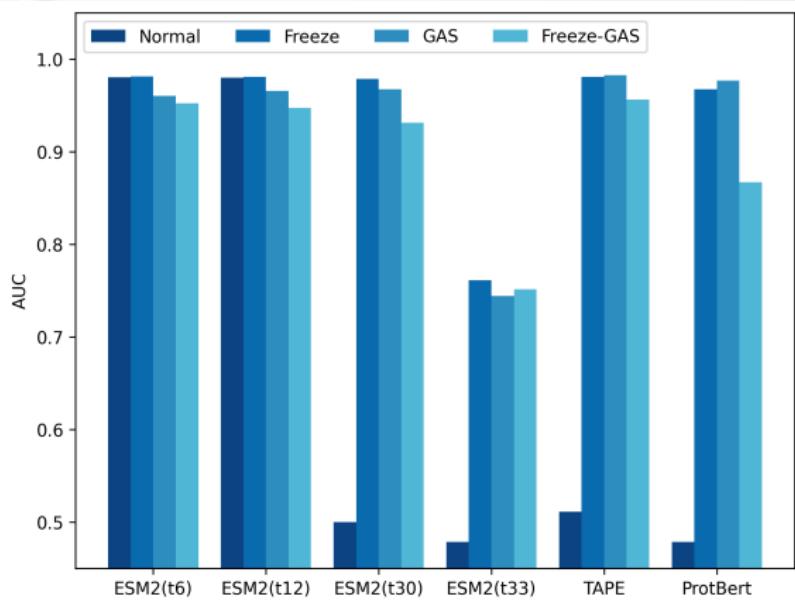
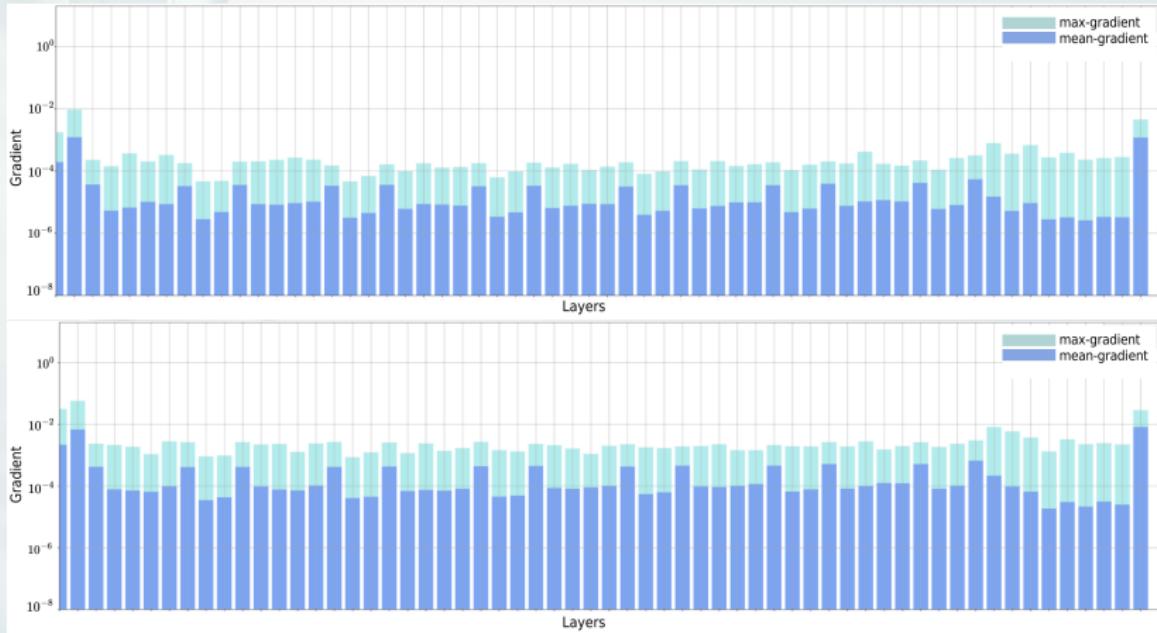


Figure: Comparación del AUC por modelo y metodología de entrenamiento.

Resultados

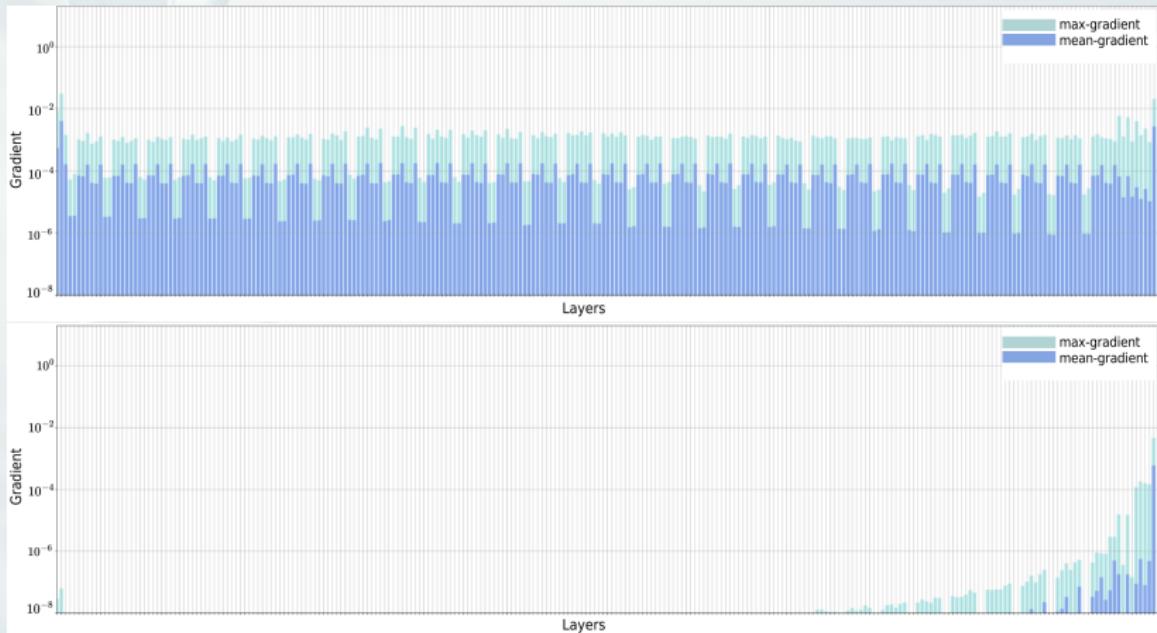
Problema de *vanish gradient* para ESM2(t6)

39



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	0.9453	0.9392	0.9797	0.8780
ESM2(t6)-Freeze	0.9401	0.9398	0.9402	0.9400	0.9830	0.8802
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	0.9393	0.9304	0.9493	0.9397	0.9787	0.8787
ESM2(t30)-GAS	0.9346	0.9337	0.9352	0.9345	0.9808	0.8691
ESM2(t30)-Freeze-GAS	0.9363	0.9319	0.9411	0.9365	0.9818	0.8726
TAPE-Normal	-	-	-	-	-	-
TAPE-Freeze	0.9395	0.9404	0.9382	0.9393	0.9815	0.8790
TAPE-GAS	0.9415	0.9352	0.9484	0.9418	0.9841	0.8831
TAPE-Freeze-GAS	0.9359	0.9297	0.9428	0.9362	0.9820	0.8719

Resultados

Comparación con los métodos *state-of-art*

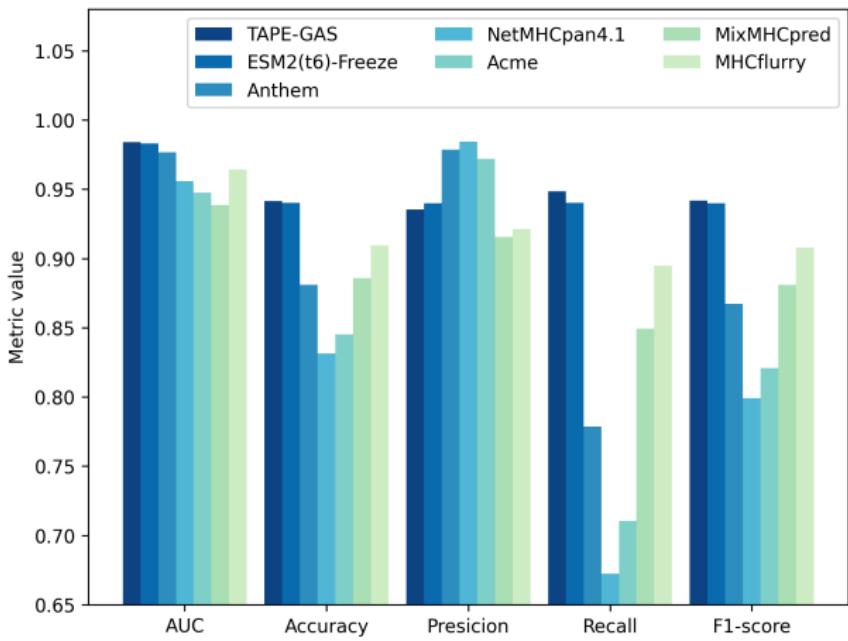


Figure: Comparación de TAPE-GAS y ESM2(t6) contra los mejores métodos del estado del arte.

Resultados

Comparación con los métodos *state-of-art*

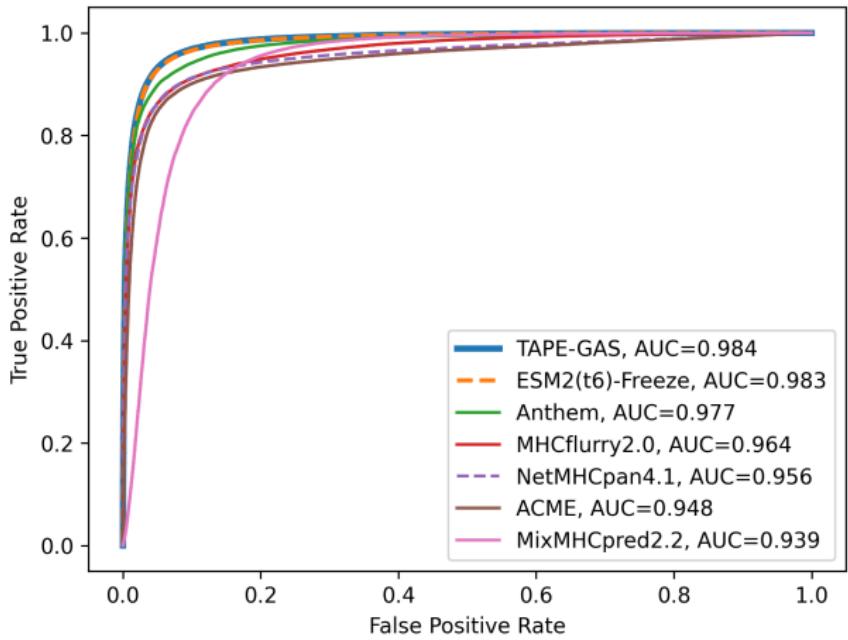


Figure: Comparación de TAPE-GAS y ESM2(t6) contra los mejores métodos del estado del arte.

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	0.9415	0.9352	0.9484	0.9418	0.9841	0.8831
ESM2(t6)-Freeze	0.9401	0.9398	0.9402	0.9400	0.9830	0.8802
Anthem	0.8811	0.9786	0.7787	0.8673	0.9768	0.7785
NetMHCpan4.1	0.8312	0.9844	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



45

Contexto y Motivación

Estadísticas en Cáncer

Inmunoterapia del Cáncer

Vacunas Personalizadas

Problema y Objetivos

Estado del arte

Propuesta

Experimentos y Resultados

Base de datos

Modelos pre-entrenados

Resultados

Discusión

Conclusiones

Trabajos Futuros

Contribuciones y Publicaciones

Discusión

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



46

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 [28].



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 [29]
- ▶ Los modelos *Transformer* grandes requieren más datos para el entrenamiento [29], 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 [30].

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.

Contenido



Contexto y Motivación

Estadísticas en Cáncer

Inmunoterapia del Cáncer

Vacunas Personalizadas

Problema y Objetivos

Estado del arte

Propuesta

Experimentos y Resultados

Base de datos

Modelos pre-entrenados

Resultados

Discusión

Conclusiones

Trabajos Futuros

Contribuciones y Publicaciones

Conclusiones

PRIMERA



PRIMERA

Se revisó y analizó los métodos que utilizan *Transformers* para la predicción del enlace pMHC. Este análisis ha demostrado que existen varios problemas como:

Conclusiones

PRIMERA



PRIMERA

Se revisó y analizó los métodos que utilizan *Transformers* para la predicción del enlace pMHC. Este análisis ha demostrado que existen varios problemas como:

- ▶ No inclusión de datos de MS.

Conclusiones

PRIMERA



PRIMERA

Se revisó y analizó los métodos que utilizan *Transformers* para la predicción del enlace pMHC. Este análisis ha demostrado que existen varios problemas como:

- ▶ No inclusión de datos de MS.
- ▶ Bajo rendimiento en las herramientas de predicción pMHC y PMHC-TCR.

Conclusiones

PRIMERA



PRIMERA

Se revisó y analizó los métodos que utilizan *Transformers* para la predicción del enlace pMHC. Este análisis ha demostrado que existen varios problemas como:

- ▶ No inclusión de datos de MS.
- ▶ Bajo rendimiento en las herramientas de predicción pMHC y PMHC-TCR.
- ▶ Falta de métodos para la detección de fusión de genes y *alternative splicing*.

Conclusiones

PRIMERA



PRIMERA

Se revisó y analizó los métodos que utilizan *Transformers* para la predicción del enlace pMHC. Este análisis ha demostrado que existen varios problemas como:

- ▶ No inclusión de datos de MS.
- ▶ Bajo rendimiento en las herramientas de predicción pMHC y PMHC-TCR.
- ▶ Falta de métodos para la detección de fusión de genes y *alternative splicing*.

Sin embargo, a pesar de estas limitaciones, ya se han realizado ensayos clínicos con resultados motivadores.

Conclusiones

SEGUNDA



SEGUNDA

Se analizó los modelos *Transformers* pre-entrenados como: TAPE, ProtBert-BFD, y la familia de modelos de EMS2, para tareas de Proteómica.

Conclusiones

SEGUNDA



SEGUNDA

Se analizó los modelos *Transformers* pre-entrenados como: TAPE, ProtBert-BFD, y la familia de modelos de EMS2, para tareas de Proteómica.

- ▶ ProtBert-BFD fue entrenado con la base de datos mas grande proteínas (2122M).

Conclusiones

SEGUNDA



SEGUNDA

Se analizó los modelos *Transformers* pre-entrenados como: TAPE, ProtBert-BFD, y la familia de modelos de EMS2, para tareas de Proteómica.

- ▶ ProtBert-BFD fue entrenado con la base de datos mas grande proteínas (2122M).
- ▶ TAPE se entreno con la base de datos mas pequeña (30M).

Conclusiones

SEGUNDA



SEGUNDA

Se analizó los modelos *Transformers* pre-entrenados como: TAPE, ProtBert-BFD, y la familia de modelos de ESM2, para tareas de Proteómica.

- ▶ ProtBert-BFD fue entrenado con la base de datos mas grande proteínas (2122M).
- ▶ TAPE se entreno con la base de datos mas pequeña (30M).
- ▶ Familia de modelos ESM2, recientemente publicada (60M).



TERCERA

Se aplicó *fine-tuning* a los seis modelos pre-entrenados agregando un bloque BiLSTM. Adicionalmente, se evaluó el uso de GAS y una metodología congelamiento de capas.



TERCERA

Se aplicó *fine-tuning* a los seis modelos pre-entrenados agregando un bloque BiLSTM. Adicionalmente, se evaluó el uso de GAS y una metodología congelamiento de capas.

- ▶ Los modelos ESM2 mejoran al aplicar el congelamiento de capas.



TERCERA

Se aplicó *fine-tuning* a los seis modelos pre-entrenados agregando un bloque BiLSTM. Adicionalmente, se evaluó el uso de GAS y una metodología congelamiento de capas.

- ▶ Los modelos ESM2 mejoran al aplicar el congelamiento de capas.
- ▶ GAS ofreció una mitigación menor del problema de *vanish gradientes*, lo que permitió el entrenamiento efectivo de modelos más grandes.
- ▶ ProtBert, era uno de los modelos mas grandes y obtuvo un desempeño pobre.



TERCERA

Se aplicó *fine-tuning* a los seis modelos pre-entrenados agregando un bloque BiLSTM. Adicionalmente, se evaluó el uso de GAS y una metodología congelamiento de capas.

- ▶ Los modelos ESM2 mejoran al aplicar el congelamiento de capas.
- ▶ GAS ofreció una mitigación menor del problema de *vanish gradientes*, lo que permitió el entrenamiento efectivo de modelos más grandes.
- ▶ ProtBert, era uno de los modelos mas grandes y obtuvo un desempeño pobre.
- ▶ TAPE obtuvo los mejores resultados, el cual se debe a la calidad de las muestras utilizadas en su pre-entrenamiento.

Conclusiones

CUARTA



CUARTA

Se volvió a entrenar los modelos de mejor desempeño ESM2(t6)-Freeze y TAPE-GAS, por 30 *epochs*, y luego se realizó una comparación con: NetMHCpan4.1, MHCflurry2.0, Anthem, ACME y MixMHCpred2.2.



CUARTA

Se volvió a entrenar los modelos de mejor desempeño ESM2(t6)-Freeze y TAPE-GAS, por 30 *epochs*, y luego se realizó una comparación con: NetMHCpan4.1, MHCflurry2.0, Anthem, ACME y MixMHCpred2.2.

- ▶ Los modelos propuestos ESM2(t6)-Freeze y TAPE-GAS superaron a los demás métodos del estado del arte en *AUC*, *accuracy*, *recall*, *f1-score* y *MCC*.

Conclusiones

QUINTA



QUINTA

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

- ▶ El método implementado obtuvo el mejor desempeño en términos de AUC, *accuracy*, *recall*, *f1-score* y MCC. Está publicado en: <https://github.com/arceda/pmh>

Contenido



Contexto y Motivación

Estadísticas en Cáncer

Inmunoterapia del Cáncer

Vacunas Personalizadas

Problema y Objetivos

Estado del arte

Propuesta

Experimentos y Resultados

Base de datos

Modelos pre-entrenados

Resultados

Discusión

Conclusiones

Trabajos Futuros

Contribuciones y Publicaciones

- ▶ Evaluar modelos grandes como: ProtT5-XL y ProtT5-XXL, ESM2(t36) y ESM2(t48) en otras bases de datos.
- ▶ Evaluar otras estrategias de entrenamiento como DistilBERT y LoRA.
- ▶ Evaluar el uso arquitecturas basadas en grafos en el proceso de *fine-tuning*.
- ▶ Desarrollar un *pipeline* para la detección de neoantígenos.

Contenido



Contexto y Motivación

Estadísticas en Cáncer

Inmunoterapia del Cáncer

Vacunas Personalizadas

Problema y Objetivos

Estado del arte

Propuesta

Experimentos y Resultados

Base de datos

Modelos pre-entrenados

Resultados

Discusión

Conclusiones

Trabajos Futuros

Contribuciones y Publicaciones



Conference Proceeding

- ▶ “Deep Learning and Transformers in MHC-Peptide Binding and Presentation Towards Personalized Vaccines in Cancer Immunology: A Brief Review” [?].
- ▶ “Neoantigen Detection Using Transformers and Transfer Learning in the Cancer Immunology Context” [?].

Journals

- ▶ “Transformers Meets Neoantigen Detection: A Systematic Literature Review” (Aceptado Q2).
- ▶ “Fine-tuning Transformers for Peptide-MHC Class I Binding Prediction” (En evaluación Q1).



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] Cancer.net,
“Qué es la inmunoterapia,” 2022.
- [4] NortShore,
“Immunotherapy,” 2022.

References II



- [5] 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.
- [6] NCI,
“National cancer institute dictionary,” 2022.
- [7] 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

- [8] UCIR,
“Neoantige-based therapy,” 2023.
- [9] 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.
- [10] 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



- [11] Nil Adell Mill, Cedric Bogaert, Wim van Crielinge, and Bruno Fant,
“neoms: Attention-based prediction of mhc-i epitope presentation,”
bioRxiv, 2022.
- [12] 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



- [13] 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.
- [14] 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



- [15] 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.
- [16] 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



65

- [17] 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.
- [18] 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



- [19] 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.
- [20] 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.
- [21] 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 IX



- [22] 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.
- [23] 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 X



- [24] Jun Cheng, Kaïdre Bendjama, Karola Rittner, and Brandon Malone,
“Bertmhc: improved mh_c–peptide class ii interaction prediction with transformer and multiple instance learning,”
Bioinformatics, vol. 37, no. 22, pp. 4172–4179, 2021.
- [25] Simon JD Prince,
UNDERSTANDING DEEP LEARNING.,
MIT PRESS, 2023.
- [26] Medium,
“Gradient accumulation steps,” 2023.
- [27] Towards data science,
“Transfer learning from pre-trained models,” 2023.

References XI



69

- [28] 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.
- [29] 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.

References XII



- [30] 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.