

UNIVERSIDAD NACIONAL DE SAN AGUSTÍN
ESCUELA DE POSGRADO
UNIDAD DE POSGRADO DE LA FACULTAD DE
INGENIERIA DE PRODUCCIÓN Y SERVICIOS



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

Tesis presentada por el Magister:
Vicente Enrique Machaca Arceda

Para optar el Grado de:
Doctor en Ciencia de la Computación

Asesor:
Prof. Dr. Cristian Lopez Del Alamo

Arequipa - Perú
2022

Declaración de autenticidad

I, Yo Vicente Machaca Arceda, declaro que la tesis titulada, ‘Detección de neo antígenos utilizando aprendizaje profundo en el marco del desarrollo de vacunas personalizadas en la inmunoterapia del Cáncer’ y el trabajo presentado en este son de mi propiedad intelectual y confirmo que:

- Este trabajo fue desarrollado durante mi candidatura a grado de doctor de esta universidad.
- Ninguna parte de esta tesis ha sido presentado para otro grado de esta universidad o cualquier otra institución.
- Cuando cito a otros autores, las fuentes has sido brindadas y con excepción de estas citas, mi trabajo es de mi autoría.
- He agradecido las principales fuentes de ayuda.
- En caso de que mi tesis haya sido desarrollado con un equipo de trabajo, yo he sido claro y he detallada la parte exacta de mi autoría.

Firma:

Fecha:

“Con fe, disciplina y desinteresada devoción al deber, no hay nada que merezca la pena que no puedas lograr.”

Muhammad Ali Jinnah

Dedico este trabajo a mi esposa Pamela Laguna Laura, quien me ha acompañado durante todo este proceso, me ha motivado y sobre todo me ha dado su amor, que me ha ayudado a prevalecer y siempre seguir adelante. De igual forma, a mis padres Vicente Machaca Chino y Victoria Arceda Arenas, de ellos he aprendido el valor de la disciplina, la fuerza por emprender y la importancia de los valores sin importar las circunstancias; gracias a ellos he logrado cumplir mis objetivos.

Resumen

La detección de neo antígenos, es la fase más importante para el desarrollo de vacunas personalizadas contra el cáncer. El proceso para identificar neo antígenos, es complejo y existen varias sub fases como: secuenciamiento, alineamiento, detección de mutaciones, identificación de péptidos, *peptide-MHC binding*, *peptide-MHC presentation* y la interacción pMHC-TCR. La mayoría de publicaciones, se ha centrado en el problema de *peptide-MHC binding*, y han logrado buenos resultados, pero menos del 5 % de los péptidos identificados, llegan a la membrana de las células y logran presentarse ante las células T. En este contexto, surge un nuevo problema llamado *peptide-MHC presentaion*, enfocado en predecir que péptidos logran enlazarse a la molécula MHC y permanecer unida a ellas hasta llegar a la membrana. Gracias a la tecnología de *Mass spectrometry*, se está secuenciando cada vez más muestras de compuestos pMHC de la membrana de las células; de esta forma se están construyendo nuevas bases de datos que puedan dar solución al problema de *peptide-MHC presentation*.

Las redes neuronales *Transformers* han revolucionado el campo de NLP, y se han abierto a muchas otras aplicaciones. Luego, las redes BERT, como una actualización a las *Transformer*, han sido aplicadas en problemas de interacción de proteínas. Pero, la interacción entre un péptido y la molécula, es una interacción entre proteínas; de esta forma han surgido trabajos que utilizan redes BERT para predecir la afinidad *peptide-MHC*. De esta forma, en esta tesis, se propone el uso de redes BERT para dar solución al problema de *peptide-MHC presentation*. Además en la propuesta se utilizó varias muestras de *Mass spectrometry*, recolectada de bases de datos públicas y trabajos similares. Finalmente, también se ha aplicado *transfer learning*, del modelo TAPE y ESM-1b (modelos entrenados con millones de secuencias de aminoácidos).

Esta tesis, presenta dos contribuciones: primero, se ha realizado una revisión sistemática de la literatura referente a la detección de neo antígenos y enfocada en estudiar los métodos basados en *deep learning*; segundo, se ha desarrollado un nuevo método basado en redes BERT y *transfer learning* para dar solución al problema de *peptide-MHC presentaion*.

Índice general

Declaración de autenticidad	I
Resumen	IV
Índice de figuras	VII
Índice de tablas	IX
Abreviaciones	X
1. Introducción	1
1.1. Contexto y Motivación	1
1.2. Problema	3
1.2.1. Formulación del problema	5
1.3. Objetivos	5
1.3.1. Objetivo General	5
1.3.2. Objetivos específicos	6
1.4. Contribuciones	6
1.5. Organización del Trabajo	7
2. Marco Conceptual	8
2.1. Bioinformática y Biología Molecular	8
2.1.1. Bioinformática	8
2.1.1.1. DNA, RNA y Proteínas	8
2.1.2. Mutaciones	11
2.2. Sistema inmunitario	12
2.2.1. Células T y APC	12
2.2.2. MHC I y II	13
2.2.3. Neo antígenos	14
2.3. <i>Machine Learning</i>	15
2.3.1. Algoritmos de aprendizaje	15
2.3.1.1. La tarea, T	16
2.3.1.2. El desempeño, P	17
2.3.1.3. La experiencia, E	17
2.3.2. Redes neuronales	18

2.4.	<i>Deep learning</i>	19
2.4.1.	<i>Deep Feedforward networks</i>	20
2.4.2.	<i>Convolutional Neural Networks</i>	20
2.4.3.	<i>Recurrent Neural Networks</i>	21
2.4.4.	<i>Transformers</i>	22
2.4.5.	<i>BERT</i>	23
3.	Estado del Arte	24
3.1.	Revisión Sistemática de la Literatura (RSL)	24
3.1.1.	Cadenas de búsqueda y bases de datos	24
3.1.2.	Selección de artículos	25
3.2.	Resultados de la RSL	27
3.2.1.	Peptide-MHC (pMHC) binding and presentation	27
3.2.1.1.	CNN	29
3.2.1.2.	RNN	29
3.2.1.3.	Transformers	33
3.2.2.	<i>Pipelines</i>	33
3.2.3.	<i>Bases de datos</i>	35
3.2.4.	<i>Reviews</i>	36
4.	Propuesta	44
4.1.	Detección de neo antígenos (<i>pipeline</i>)	44
4.2.	Predicción de la afinidad peptido-MHC (peptide-MHC binding)	47
5.	Resultados	49
6.	Conclusiones	51

Índice de figuras

1.1. Marco de desarrollo para la creación de vacunas personalizadas contra el cáncer basadas en neoantígenos. (a) proporciona una visión general de cada etapa (Han et al., 2020). (b) una visión general de cada fase con un énfasis en el desarrollo <i>in-silico</i>	3
2.1. Localización y estructura del DNA. Fuente: NCI (2022).	9
2.2. Transcripción y traducción. Fuente: NCI (2020).	10
2.3. <i>Alternative Splicing</i> . Fuente: NCI (2020).	10
2.4. Ejemplos de SNV en el DNA. Fuente: Socratic.org (2022)	11
2.5. Ejemplos de variaciones en el DNA. Fuente: PacBio (2021)	12
2.6. Presentación de antígenos por MHC-I. Fuente: Zhang et al. (2019)	13
2.7. Presentación de antígenos por MHC-II. Fuente: Zhang et al. (2019)	14
2.8. Proceso para la detección de neo antígenos y generación de vacunas personalizadas. Fuente: (Mattos et al., 2020)	15
2.9. Representación de una neurona. Fuente: Raff (2022).	18
2.10. Representación de una red neuronal.	19
2.11. Relación entre Inteligencia Artificial, <i>Machine Learning</i> y <i>Deep Learning</i> . Fuente: El Naqa and Murphy (2022).	19
2.12. Representación de un <i>Deep Feedforward Network</i> . Fuente: El Naqa and Murphy (2022).	20
2.13. Ejemplo de una convolución en procesamiento de imágenes. Fuente: Shuchen (2022).	21
2.14. Arquitectura de LeNet-5, una CNN para el reconocimiento de dígitos. Fuente: LeCun et al. (1998).	21
2.15. Ejemplo del procesamiento del <i>input gate</i> , <i>forget gate</i> y <i>output gate</i> de LSTM. Fuente: Zhang et al. (2021).	22
2.16. ejemplo del mecanismo de atención de una red <i>Transformer</i> . Fuente: Zhang et al. (2021).	23
3.1. The three phases and approaches that researches focus in the problem of neoantigen detection.	27
3.2. Peptide presentation by MHC class I. Source: Zhang et al. (2019)	28
4.1. Proceso general utilizado para la detección de neo antígenos a partir de secuencias de DNA. Fuente: Gopanenko et al. (2020).	46
4.2. Propuesta de <i>transfer learning</i> de ESM-1b y una red neuronal paralela para la predicción de la afinidad entre un péptido y MHC (peptide MHC binding).	48

5.1. Accuracy durante cada <i>epoch</i> , 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, B*44:02 y B*44:03.	50
--	----

Índice de tablas

3.1. Cadenas de búsqueda utilizadas en la RSL.	25
3.2. Bases de datos utilizadas en la RSL.	25
3.3. Cantidad de artículos encontrados y seleccionados según los criterios de inclusión y exclusión en la RSL.	26
3.4. Criterios de inclusión y exclusión de artículos utilizados en la RSL.	26
3.5. List of research since 2018 that uses CNNs for peptide-MHC binding and presentation.	30
3.6. List of research since 2018 that uses CNNs s with RNN or attention mechanisms for peptide-MHC binding and presentation.	31
3.7. List of research since 2018 that uses RNNs for peptide-MHC binding and presentation.	32
3.8. List of research since 2018 that uses Transformers (self-attention) for peptide-MHC binding and presentation.	34
3.9. Listado de <i>pipelines</i> desde el 2018, para la detección de neo antígenos. . .	35
3.10. Bases de datos públicas de <i>pMHC binding</i> , <i>pMHC presentation</i> , interacción pMHC-TCR y estructuras 3D de proteínas.	36
3.11. Listado de los <i>reviews</i> , que se enfocan en estudios de <i>Next-Generation Sequencing</i> para la detección de neo antígenos e inmunoterapia del Cáncer. . .	37
3.12. Listado de los <i>reviews</i> , que se enfocan en estudios de la interacción de péptidos y la molécula MHC, para la detección de neo antígenos.	38
3.13. Listado de los <i>reviews</i> , que se enfocan en estudios de <i>Mass spectrometry</i> para la detección de neo antígenos.	39
3.14. Listado de los <i>reviews</i> , que se enfocan en presentar en proceso general de detección de neo antígenos y vacunas personalizadas del año 2022 y 2021. . .	40
3.15. Listado de los <i>reviews</i> , que se enfocan en presentar en proceso general de detección de neo antígenos y vacunas personalizadas del año 2020 y 2019. . .	41
3.16. Listado de los <i>reviews</i> , que se enfocan en estudios que utilizan propiedades estructurales de los aminoácidos para la de detección de neo antígenos. . .	42
3.17. Listado de los <i>reviews</i> , que se enfocan en estudios de la interacción de compuestos pMHC con TCR.	42
3.18. Listado de los <i>reviews</i> , que se enfocan en presentar buenas prácticas en el proceso de detección de neo antígenos y generación de vacunas personalizadas,	43
5.1. Resultados obtenidos en cada base de datos.	49

Abreviaciones

ANN	Artificial Neural Network
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
bp	Base pair in DNA
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DNA	Deoxyribonucleic Acid
GNN	Graph Neural Netowrk
G-BERT	Graph Bidirectional Encoder Representations from Transformers
HLA	Human Leukocyte Antigens
MCC	Matthews Correlation Coefficient
MHC-I	Major Histocompatibility Complex Class I
MHC-II	Major Histocompatibility Complex Class II
MHC-III	Major Histocompatibility Complex Class III
mRNA	Messenger Ribonucleic Acid
NLP	Natural Language Processing
pMHC	Peptide-MHC ligand
pMHC-TCR	pMHC T-cell receptor ligand
RNA	Ribonucleic Acid
RoBERTa	Optimized BERT
RSL	Revisión Sistemática de la Literatura
tRNA	Transfer Ribonucleic Acid
TCR	T-cell receptor

Capítulo 1

Introducción

1.1. Contexto y Motivación

El cáncer representa el desafío de salud global más significativo ([Siegel et al., 2023](#)). Además, según el Instituto de Investigación del Cáncer del Reino Unido, se registraron más de 18 millones de nuevos casos y 10 millones de muertes en 2020 ([UK, 2023b](#)). Además, se predice que habrá alrededor de 28 millones de nuevos casos anualmente para alrededor de 2040 si la incidencia se mantiene estable y el crecimiento de la población y el envejecimiento continúan según las tendencias recientes ([UK, 2023a](#)). Esto representa un aumento del 54.9 % desde 2020, con un aumento esperado mayor en hombres (60.6 %) que en mujeres (48.8 %).

En este contexto, se sabe que los métodos tradicionales basados en cirugía, radioterapia y quimioterapia tienen baja eficacia y efectos secundarios adversos ([Peng et al., 2019](#)). Por lo tanto, ha surgido el desarrollo de la inmunoterapia contra el cáncer, con el objetivo de estimular el sistema inmunológico del paciente ([Borden et al., 2022](#)). Existen tratamientos como vacunas personalizadas, terapias con linfocitos T adoptivos e inhibidores de puntos de control inmunológico. De entre estos, las vacunas basadas en neoantígenos han mostrado un gran potencial al potenciar las respuestas de los linfocitos T y se consideran las más propensas a tener éxito ([Borden et al., 2022](#)). Además, los neoantígenos se utilizan en la terapia de bloqueo de puntos de control inmunológico. Los neoantígenos se consideran biomarcadores predictivos y objetivos para el tratamiento sinérgico en la inmunoterapia contra el cáncer ([Fang et al., 2022](#)).

El desarrollo de vacunas personalizadas contra el cáncer es un proceso largo que depende de la detección precisa de neoantígenos (ver Figura 1.1). Estos neoantígenos son péptidos que se encuentran exclusivamente en las células cancerosas. El objetivo de un tratamiento

basado en vacunas personalizadas es entrenar a los linfocitos (células T) del paciente para que reconozcan estos neoantígenos y activen el sistema inmunológico (Mattos et al., 2020; Peng et al., 2019). El proceso se resume en la Figura 1.1(b) y consta de las siguientes fases:

1. Obtener muestras de tejidos cancerosos y sanos. Ambos tejidos se secuencian para obtener ADN y/o ARN. Algunos enfoques incluyen información del *immunopeptidome* obtenida mediante *Mass Spectrometry* (MS).
2. En la etapa *in-silico*, se realiza el alineamiento de secuencias, se desarrolla un proceso de llamada de variantes para detectar variaciones y/o mutaciones, y se anotan estas variantes (posible detección de neoantígenos). Hay disponibles varias herramientas con buen rendimiento para esta etapa.
3. En esta etapa *in-silico*, se priorizan los neoantígenos. Este paso es crucial y ha recibido una atención significativa en la investigación en los últimos años debido a su complejidad y la baja efectividad de los enfoques actuales. Aquí, se evalúa la afinidad de los candidatos neoantígenos (péptidos) de la etapa anterior con el *Major Histocompatibility Complex* (MHC), conocido como la unión pMHC. Luego, se evalúa la afinidad de pMHC para unirse al *T-cell Receptor* (TCR). Al final de esta etapa, se obtienen los neoantígenos.
4. En la etapa *in-vitro*, en el laboratorio se inducen las células T del paciente para que reconozcan los neoantígenos. En este punto, se desarrollan las vacunas. Esta etapa la llevan a cabo biotecnólogos y biólogos.
5. Finalmente, el oncólogo realiza una evaluación clínica de la vacuna.

La detección *in-silico* de neoantígenos se basa en las etapas segunda y tercera representadas en la Figura 1.1(a). En este contexto, debido a la complejidad del proceso y la variedad de métodos disponibles, se han desarrollado herramientas de software y flujos de trabajo para agilizar el uso de estas herramientas. Además, los Transformers han marcado el comienzo de una nueva era en la inteligencia artificial, demostrando logros destacados en una variedad de tareas de procesamiento del lenguaje natural (Patwardhan et al., 2023). Estos modelos también han encontrado aplicación en la detección de neoantígenos, especialmente en la tercera etapa de la Figura 1.1(b). Se han propuesto modelos BERT y redes de aprendizaje profundo con mecanismos de atención para predecir la unión péptido-MHC y pMHC-TCR obteniendo resultados prometedores. Sin embargo, aún existe mucho camino por recorrer y con el incremento constante de muestras de ADN/proteínas, sumado a los nuevos mecanismos para entrenar modelos

Transformers con billones de parametros, se espera lograr avances significativos en este campo de estudio.

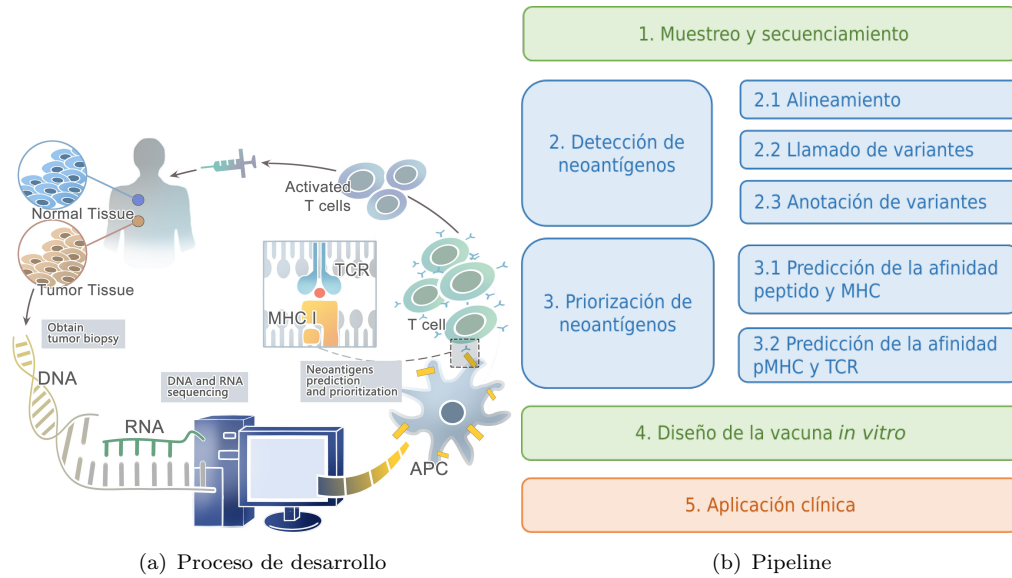


FIGURA 1.1: Marco de desarrollo para la creación de vacunas personalizadas contra el cáncer basadas en neoantígenos. (a) proporciona una visión general de cada etapa (Han et al., 2020). (b) una visión general de cada fase con un énfasis en el desarrollo *in-silico*.

1.2. Problema

Los neoantígenos son péptidos mutados específicos de tumores y son considerados los principales causantes de una respuesta inmune (Borden et al., 2022; Chen et al., 2021a; Gopanenko et al., 2020). Es así que surgen varios esfuerzos e investigación en la Inmunoterapia del cáncer, concentradas en el estudio y detección de neoantígenos. Es así, que el desarrollo de vacunas personalizadas basadas en neoantígenos es considerado uno de los métodos con mayor probabilidad de éxito (Borden et al., 2022). Incluso varias compañías como BioNTech, Genocoea Biosciences, Neon Therapeutics y Gritstone Oncology realizan investigación y ofrecen el servicio de generar vacunas personalizadas a pacientes de cáncer.

Además, el MHC representa un factor clave, en la detección de neoantígenos, al ser el encargado de unirse al neoantígeno y presentarlo a la superficie de la célula. Debido a esto, este trabajo se enfoca en el desarrollo de un método para la predicción del enlace entre neoantígenos y MHC (*pMHC binding*), esto corresponde a la Fase 3.1 de la Figura 1.1(b) dentro del *pipeline* de detección de neoantígenos. Existen dos tipos: el MHC clase I (MHC-I) y MHC clase II (MHC-II), ambos presentan péptidos en la superficie celular a las células T CD8+ y CD4+, respectivamente (Janeway Jr, 1997;

Abualrous et al., 2021). En detalle, el ciclo de vida de los neoantígenos que se unen a MHC-I se puede resumir de la siguiente manera. Primero, una proteína cancerígena se degrada en péptidos en el citoplasma. Luego, los péptidos se unen al MHC (*pMHC binding*). Después, este compuesto sigue un camino hasta llegar a la membrana celular (*pMHC presentation*). Finalmente, el pMHC es reconocido por el TCR, desencadenando el sistema inmunológico (Janeway Jr, 1997; Wieczorek et al., 2017; Gasser et al., 2021). Por lo tanto, el *pMHC binding* es un paso muy importante para la inmunidad celular, y la predicción y comprensión de esta unión tienen un valioso potencial. Lamentablemente, la mayoría de los ligandos *pMHC* no llegan a la membrana celular (Mattos et al., 2020).

Adicionalmente, las proteínas MHC están codificadas por genes altamente polimórficos, llamados *Antígenos Leucocitarios Humanos* (HLA); la considerable naturaleza polimórfica de los genes MHC proporciona una variación sustancial en la unión con los neoantígenos, lo que influye en el conjunto de neoantígenos presentados a las células T (Abualrous et al., 2021). En consecuencia, los métodos propuestos se categorizan como *pan-specific* y *allele-specific*. Los métodos *allele-specific* (Rammensee et al., 1999; Reche et al., 2002; Kim et al., 2009; Nielsen and Andreatta, 2016; Vang and Xie, 2017; Shao et al., 2020; Bravi et al., 2021) entrenan un modelo para cada *allele* del MHC; mientras que los métodos *pan-specific* (Hu et al., 2019; Liu et al., 2019b; Wu et al., 2019; Phloyphisut et al., 2019; O'Donnell et al., 2018, 2020; Reynisson et al., 2020; Venkatesh et al., 2020; Ye et al., 2021; Mei et al., 2021; Chu et al., 2022; Zhang et al., 2022b; Mei et al., 2021; Hu et al., 2019; Gfeller et al., 2023) entrenan un modelo global que toma péptidos (neoantígenos) y MHC como entradas. Además, la naturaleza polimórfica del MHC eleva bastante la complejidad de este problema, se cree que existen las 10000 diferentes MHC *alleles* (Abelin et al., 2017), esto complica mucho la detección de neo antígenos. Por lo tanto, los métodos *pan-specific* surgen con una alta posibilidad de futuras aplicaciones.

Lamentablemente, a pesar de varios esfuerzos en el desarrollo de métodos para la detección de neoantígenos, menos del 5 % de neoantígenos detectados activan el sistema inmune (Mattos et al., 2020; Mill et al., 2022; Bulik-Sullivan et al., 2019; Bassani-Sternberg et al., 2015; Yadav et al., 2014). Según los autores de los métodos, las razones son:

1. La no inclusión en conjunto de varias fuentes de información como DNA-seq, RNA-seq, y datos de *Mass Spectrometry* (MS) (Kim et al., 2018). Por ejemplo, la mayoría de propuestas no utiliza datos de MS; en la actualidad, existe una creciente información de estos datos y se están aplicando a varios campos de la Bioinformática.
2. Uso de herramientas de bajo desempeño para la predicción del enlace péptido-MHC (pMHC) (etapa 3.1 de la Figura 1.1(b)). La mayoría de aplicaciones, se basa en el

uso de MHCFlurry (O'Donnell et al., 2020) y NetMHCpan4.1 (Reynisson et al., 2020). Sin embargo, actualmente, se cuenta con herramientas de mejor desempeño basado en *transformers* (Arceda, 2023). Esta tesis, se enfoca en resolver este problema.

3. Para la etapa 3.2 de la Figura 1.1(b), los autores no consideran la predicción del enlace pMHC al TCR (pMHC-TCR), varios autores consideran incluir esta tarea en trabajos futuros (Rubinsteyn et al., 2018).
4. Finalmente, no utilizar información de eventos de *alternative splicing*, variaciones estructurales en el ADN y las mutaciones de fusión de genes, esta información está fuertemente relacionada con varios tipos de cancer (Wood et al., 2020).

En conclusión, la detección de neoantígenos es un desafío que consta de múltiples etapas, y las herramientas actuales en el estado del arte presentan un rendimiento insuficiente. Uno de los factores clave detrás de este bajo rendimiento está relacionado con la predicción del enlace pMHC. Por esta razón, esta tesis se centra en abordar este problema mediante la propuesta de un método basado en Transformers para la predicción del enlace pMHC.

1.2.1. Formulación del problema

El presente estudio se centra en el problema de predicción del enlace pMHC-I (*pMHC binding prediction*). Esto representa un problema de clasificación binaria que toma como entrada la secuencia de aminoácidos de un péptido y el MHC. Un péptido podría representarse como: $p = \{A, \dots, Q\}$ y una representación similar para el MHC sería: $q = \{A, N, \dots, G\}$. Finalmente, necesitamos conocer la probabilidad de afinidad entre p y q . Si esta probabilidad es lo suficientemente alta, es posible que el péptido se enlace al MHC y por lo tanto, el péptido p en cuestión, sería un excelente candidato a neoantígeno.

1.3. Objetivos

1.3.1. Objetivo General

Proponer un método *in Silico* basado en *Transformers* y *Transfer Learning* para la detección de neo antígenos, enfocados en la predicción del enlace pMHC.

1.3.2. Objetivos específicos

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

1.4. Contribuciones

Las principales contribuciones de este trabajo son:

- (a) Se ha desarrollado una revisión sistemática de la literatura referente a los métodos basados en *Transformers* para la detección de neoantígenos. Esto ha generado dos publicaciones tituladas: “*Deep Learning and Transformers in MHC-Peptide Binding and Presentation Towards Personalized Vaccines in Cancer Immunology: A Brief Review*” (Machaca et al., 2023) y “*Transformers Meets Neoantigen Detection: A Systematic Literature Review*”.
- (b) Se ha implementado *fine-tuning* a seis modelos de Transformers para la predicción del enlace pMHC; además, se ha evaluado el uso de GAS y una metodología de *layer-freezing*. Los resultados fueron publicados en: “*Neoantigen Detection Using Transformers and Transfer Learning in the Cancer Immunology Context*” (Arceda, 2023) y “*Fine-tuning Transformers for Peptide-MHC Class I Binding Prediction*”.
- (c) Finalmente, se comparó los métodos propuestos con herramientas del estado del arte como: NetMHCpan4.1, MHCFlurry2.0, Anthem, ACME y MixMHCpred2.2. Los métodos propuestos obtuvieron los mejores resultados en *accuracy*, *Area Under the Curve* (AUC), *recall*, *f1-score* y *Matthews Correlation Coefficient* (MCC).

1.5. Organización del Trabajo

En el Capítulo 2 se presentan los conceptos básicos sobre Bioinformática e inmunoterapia del Cáncer, también son abordados los temas sobre *deep learning* y redes neuronales Transformers.

Luego, en el Capítulo 3 se describen los trabajos relacionados a la presente tesis. Debido a la gran cantidad de publicaciones, solo se ha considerado trabajos desde el 2018 y que hacen uso de Transformers o redes neuronales con mecanismos de atención.

El Capítulo 4, presenta la propuesta de la tesis. Esta se basa en un método para desarrollar *fine-tuning* a Transformers pre-entrenadas para diversas tareas de Proteómica.

Luego, en el Capítulo 5, se presentan los resultados de la investigación. Además, se presenta una comparación con los métodos del estado del arte.

Finalmente, en el Capítulo 6 son expuestos las conclusiones del presente trabajo así como también las direcciones para continuar con el mismo en la sección de trabajos futuros.

Capítulo 2

Marco Conceptual

El proyecto pertenece al área de Bioinformática y específicamente a la Inmunoinformática, en este contexto el marco teórico detalla conceptos de Biología Molecular (ADN, ARN y proteínas), Inmunología y Ciencias de la Computación.

2.1. Bioinformática y Biología Molecular

En esta sección, describiremos los principales conceptos referentes a Biología Molecular que serán considerados en la propuesta de la tesis.

2.1.1. Bioinformática

Según [Luscombe et al. \(2001\)](#), la Bioinformática involucra la tecnología que utiliza las computadoras para el almacenamiento, manipulación y distribución de información relacionada a la Biología Molecular como DNA, RNA y proteínas. También podemos considerar que la Bioinformática se enfoca al análisis de secuencias, estructuras y funciones de los genes y proteínas; algunas veces también puede ser llamado Computación Molecular Biológica ([Xiong, 2006](#)).

2.1.1.1. DNA, RNA y Proteínas

Deoxyribonucleic Acid (DNA) es una molécula dentro de las células que contiene información genética responsable del desarrollo y función del organismo ([NCI, 2022](#)). Gran parte del DNA se sitúa dentro del núcleo de las células (en organismos Eucariotes). Por ejemplo en la Figura [2.1](#), vemos como el DNA, forma parte de los cromosomas y estos

a su vez están en el núcleo. Luego, podemos notar, que los genes representan segmentos del DNA. Finalmente, en la Figura 2.1, notamos las bases nitrogenadas que componen el DNA: *Guanine*, *Cytosine*, *Adenine* y *Thymine*; normalmente, estas bases serán representadas por las letras: G, C, A, T respectivamente.

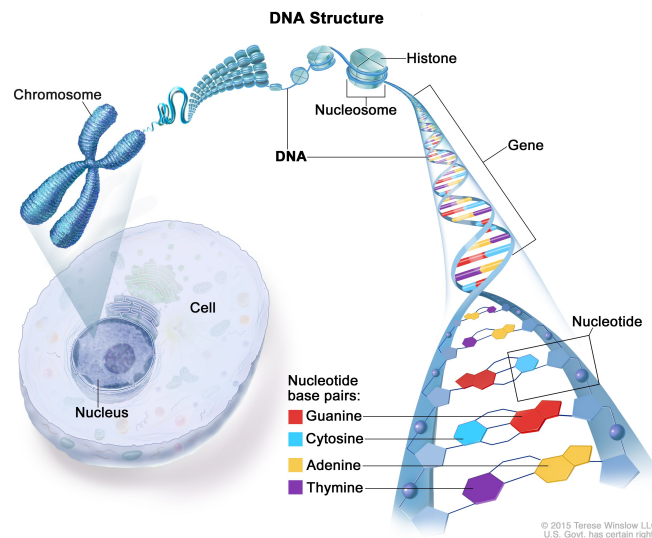


FIGURA 2.1: Localización y estructura del DNA. Fuente: [NCI \(2022\)](#).

Durante el ciclo de vida de la célula, ocurre un proceso llamado Transcripción (ver Figura 2.2), en este proceso se generan cadenas de *Ribonucleic Acid* (RNA) a partir de la cadena de DNA ([NCI, 2022](#)). Durante este proceso la base nitrogenada *Thymine* (T) es reemplazada por *Uracil* (U). El proceso mencionado, ocurre dentro del núcleo de la célula y en esta etapa el RNA es llamado *messenger RNA* (mRNA). Una vez el mRNA sale del núcleo, es transportado por *transfer RNA* (tRNA) hacia los Ribosomas (ver Figura 2.2). En esta, última etapa ocurre la Traducción, cada grupo de tres bases nitrogenadas (codones) se convierten en un aminoácido diferente, luego estos aminoácidos forman cadenas polipeptídicas y estas a su vez forman las proteínas; normalmente, cada gen genera una proteína ([Xiong, 2006](#); [NCI, 2022](#)).

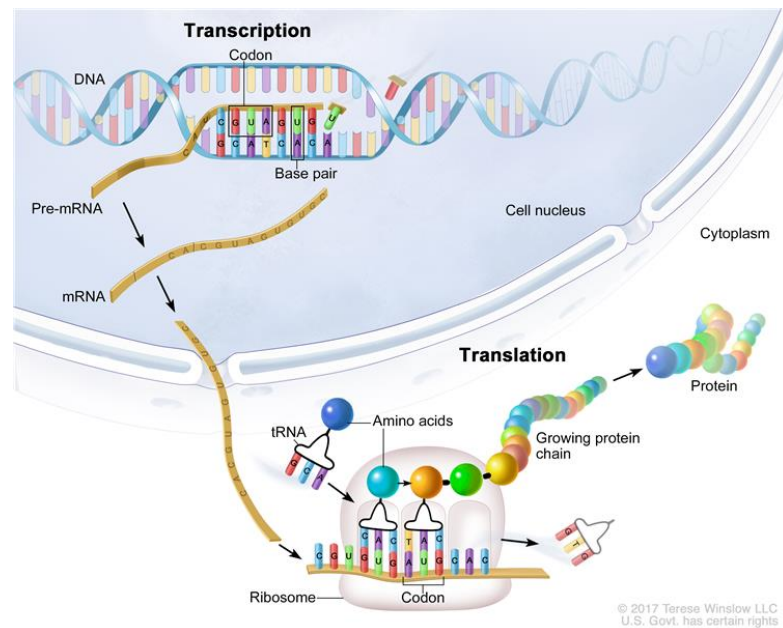
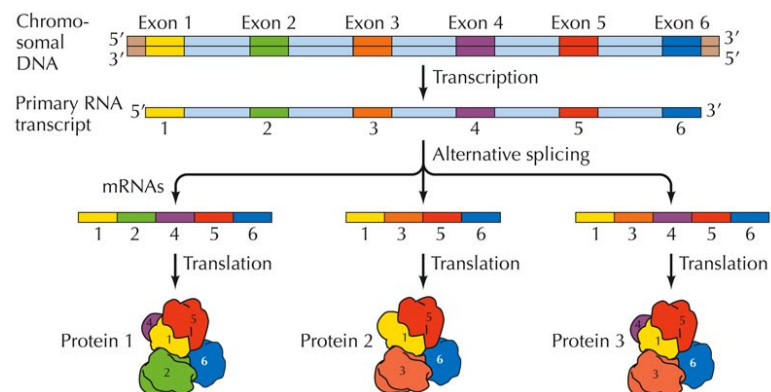


FIGURA 2.2: Transcripción y traducción. Fuente: NCI (2020).

Durante el proceso de Traducción, puede ocurrir un fenómeno llamado *Alternative Splicing*. Por ejemplo, en la Figura 2.3, notamos como un gen puede generar tres proteínas distintas, cada una con funciones distintas. Este fenómeno, complica bastante el análisis de DNA.



THE CELL, Fourth Edition, Figure 5.5 © 2006 ASM Press and Sinauer Associates, Inc.

FIGURA 2.3: *Alternative Splicing*. Fuente: NCI (2020).

2.1.2. Mutaciones

Las mutaciones también llamadas variaciones, representan cualquier cambio en la secuencia de DNA, estos pueden ocurrir durante la división celular o por la exposición a agentes químicos o radioactivos. Estas mutaciones pueden ser beneficiosas, dañinas (cuando afectan la generación de proteínas) o no tener algún efecto (NCI, 2022). Varios tipos de Cáncer son ocasionados por estas mutaciones (Borden et al., 2022; Chen et al., 2021a; Mattos et al., 2020).

Según el tipo de célula afectada, tenemos: mutaciones somáticas y mutaciones *germline* (una mutación en estas células puede ser heredada a la descendencia) (Clancy, 2008). Según (Xu, 2018), las variaciones genómicas pueden clasificarse en tres grupos: *Single-Nucleotide Variant* (SNV), inserciones y eliminaciones (INDELS) y *Structural Variation* (SV). Una mutación se considera SNV cuando las variaciones afectan a menos de 10 bases.

En la Figura 2.4, presentamos ejemplos de SNV. Por ejemplo, las sustituciones pueden afectar la generación de un aminoácido, pero las inserciones o eliminaciones pueden afectar en cadena la generación de varios aminoácidos, a este tipo de fenómeno se le conoce como *frameshit mutation* (Xu, 2018).

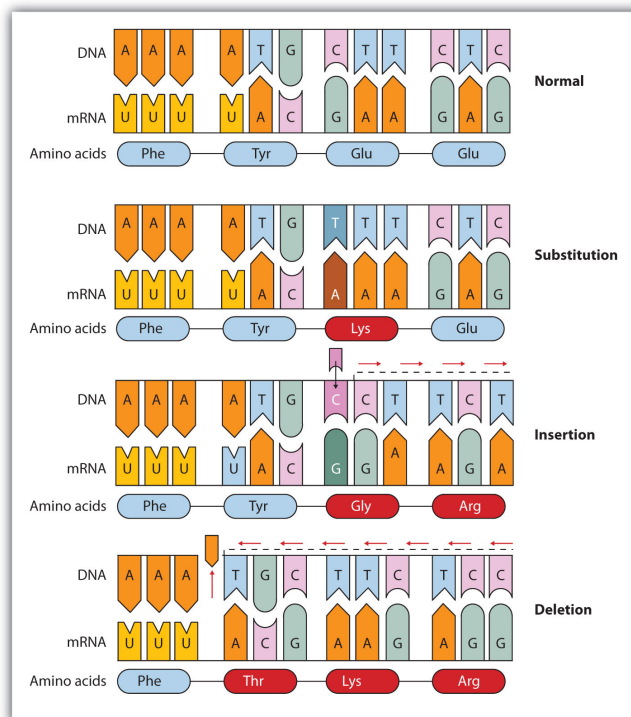


FIGURA 2.4: Ejemplos de SNV en el DNA. Fuente: [Socratic.org](https://www.socratic.org) (2022)

En la Figura 2.5, mostramos algunos tipos de SV. En este caso, también se pueden presentar INDELS, *Tanden duplication*, inversiones, traslocaciones y *Copy Number Variants* (CNV). Los CNVs, representan fuertes candidatos para ser biomarcadores de varios tipos de Cáncer (Pan et al., 2019; Lucito et al., 2007). Otra mutación importante, es referente a la fusión de genes, en estos casos dos o más genes se fusionan y forman una proteína completamente diferente, este tipo de mutación también está fuertemente relacionado a varios tipos de Cáncer (Kerbs et al., 2022; Kim and Zhou, 2019; Heyer and Blackburn, 2020).

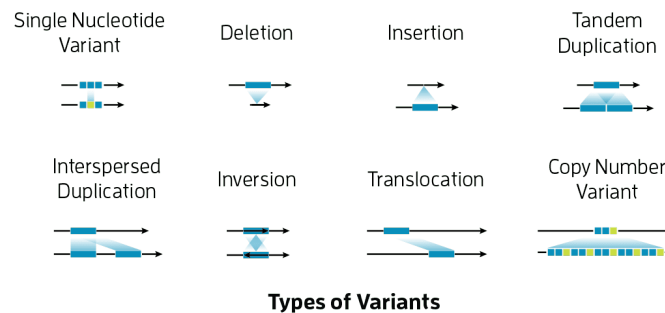


FIGURA 2.5: Ejemplos de variaciones en el DNA. Fuente: PacBio (2021)

2.2. Sistema inmunitario

El sistema inmunitario hace referencia al conjunto de células y procesos químicos que tiene como función protegernos de agentes extraños como: microbios, bacterias, células de Cáncer, toxinas, etc. Marshall et al. (2018). En esta sección, se explicará de forma breve el comportamiento del sistema inmunitario frente cuando un agente extraño (antígeno) ingresa al cuerpo humano.

2.2.1. Células T y APC

Las células T también llamadas linfocitos T, se forman a partir de la médula ósea y son los encargados de eliminar agentes extraños (antígenos) NCI (2022). Estas células están compuestas por un T-cell Receptor (TCR), que es el encargado de reconocer y enlazar a los antígenos. Luego, algunas células T, requieren de la acción de los *Antigen Presenting Cells* (APC), estas células APC son: células dentríticas, macrófagos, células B, fibroblastos y células epiteliales. Normalmente, los APC devoran los antígenos y luego los presentan a las células T para su eliminación (Marshall et al., 2018).

2.2.2. MHC I y II

Major Histocompatibility Complex (MHC) I y II, son proteínas que desempeñan un rol importante en el sistema inmunitario. Ambas proteínas tienen la función de presentar péptidos (antígenos) en la superficie de las células, para que sean reconocidas por la células T (Abualrous et al., 2021). MHC-I se encarga de la presentación de las células con núcleo, mientras que MHC-II, de las células APC.

El proceso de presentación de los antígenos por MHC-I es el siguiente (Figura 2.6): la proteína foránea es degradado por el proteasoma y se producen péptidos (posibles antígenos), luego estos péptidos son transportados al Endoplasmic Reticulum (ER) con la ayuda de *Transporter associated Antigen Processing* (TAP), luego es migrado al aparato de Golgi para ser presentado en la superficie de la célula y es enlazado a la proteína MHC-I, una vez en la superficie, el antígeno puede ser reconocido por las células CD8+T (Zhang et al., 2019).

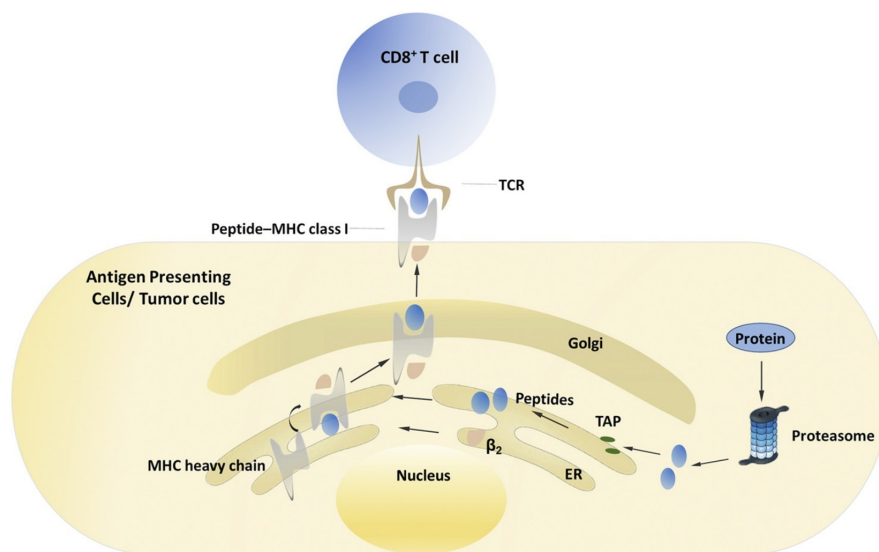


FIGURA 2.6: Presentación de antígenos por MHC-I. Fuente: Zhang et al. (2019)

Para el caso de MHC-II, es un proceso similar (Figura 2.7): primero, los patógenos son devorados por fagocitosis, los péptidos asociados a MHC-II son producidos en el Endoplasmic Reticulum (ER), para luego ser trasladados al aparato de Golgi, y luego ser transportados a la superficie de las células una vez enlazadas con MHC-II, finalmente, son reconocidas por las células CD4+T (Zhang et al., 2019).

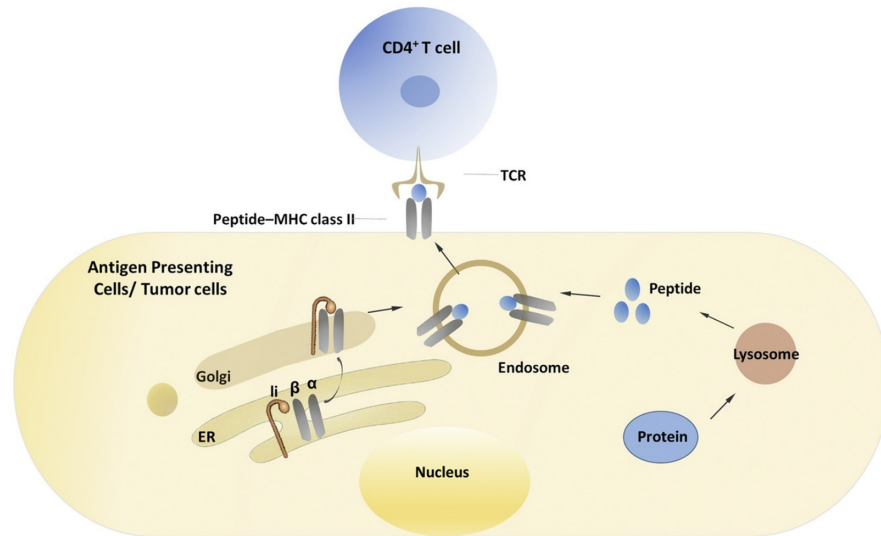


FIGURA 2.7: Presentación de antígenos por MHC-II. Fuente: [Zhang et al. \(2019\)](#)

2.2.3. Neo antígenos

Es una proteína que se forma en las células de Cáncer cuando ocurre mutaciones en el DNA. Los neo antígenos cumplen un rol importante al estimular una respuesta inmune en contra de células de Cáncer. En la actualidad, se estudia su uso en el desarrollo de vacunas contra el Cáncer [NCI \(2022\)](#). Una característica importante de los neo antígenos, es que solo están presentes en células tumorales y no en células sanas, debido a eso son considerados factores clave en la inmunoterapia del Cáncer [Borden et al. \(2022\)](#). En la actualidad hay varios métodos para detectar a predecir neo antígenos, pero solo una pequeña porción de ellos logran estimular al sistema inmune [Chen et al. \(2021a\)](#); [Hao et al. \(2021\)](#).

Este proceso para la detección de neo antígenos, generalmente consiste en: (1) extracción del tejido tumoral, (2) identificación de mutaciones, (3) detección de neo antígenos y predicción de inmunogenicidad, (4) desarrollo de experimentos in vitro y (5) desarrollo de la vacuna ([Mattos et al., 2020](#); [Peng et al., 2019](#)) (ver Figura 2.8).

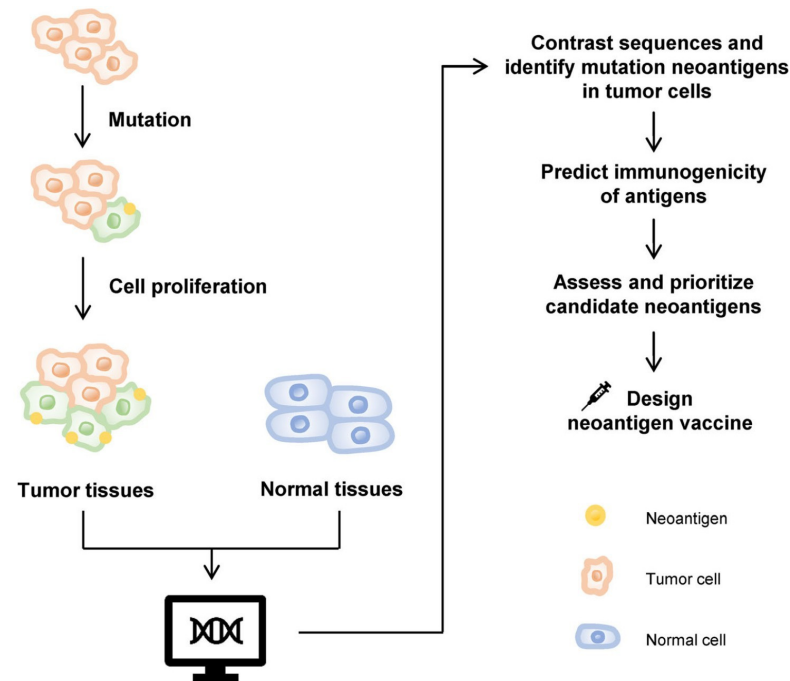


FIGURA 2.8: Proceso para la detección de neo antígenos y generación de vacunas personalizadas. Fuente: (Mattos et al., 2020)

2.3. Machine Learning

Machine Learning (ML) es una categoría de algoritmos computacionales capaces de emular algunas acciones inteligentes. Es el resultado de varias disciplinas como: inteligencia artificial, probabilidad, estadística, ciencia de la computación, teoría de la computación, psicología y filosofía (El Naqa and Murphy, 2022). *Machine Learning* tiene varias definiciones, pero una de las mas acertadas, según Samuel (1967): “Campo de estudio que brinda a las computadoras la habilidad de aprender sin haber sido explícitamente programado”.

2.3.1. Algoritmos de aprendizaje

Un algoritmo de aprendizaje o *machine learning algorithm*, es aquel algoritmo que no debe ser programado explícitamente, este aprende de la experiencia, a partir de datos (Goodfellow et al., 2016). Según Mitchell (1997): “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E ”. La traducción a español indicaría: “Un programa de computadora puede aprender de una experiencia E , para una tarea T y con una métrica de desempeño P , si el desempeño de

la tarea T , medido con P , mejorar con la experiencia E'' . Esto, nos da a entender que un programa de computadora puede aprender si mejora su desempeño según aumente su experiencia o datos.

2.3.1.1. La tarea, T

La tarea T de ML, puede ser descrito como de la forma en que el sistema de ML procesa una muestra o ejemplo. Según [Goodfellow et al. \(2016\)](#) las tareas más comunes de ML son:

- **Clasificación.** En este caso, el algoritmo de ML debe predecir la clase a la que pertenece la muestra. Entonces, al algoritmo debe producir una función: $f : \mathbb{R}^n \rightarrow \{1, \dots, k\}$. También puede escribirse como: $y = f(x)$, aquí x representa la entrada y la función f determinará la clase a la que pertenece.
- **Regresión.** El algoritmo debe producir una función: $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Es decir, dada como entrada un vector x de reales, el algoritmo de ML debe predecir un valor en los números reales.
- **Transcripción.** En este caso, dada como entrada datos no estructurados, el algoritmo de ML debe generar información de forma textual. Por ejemplo: dada una imagen como entrada, la salida sería el texto encontrado en la imagen.
- **Maquinas de traducción.** Como el nombre indica, la entrada es un texto en un lenguaje y la salida es un texto en otro lenguaje.
- **Salida estructurada.** En este caso la salida es un vector o alguna estructura de datos de varios valores. El procesamiento natural de lenguaje es un buen ejemplo, la entrada es un texto y la salida es un árbol que denota la estructura gramatical y semántica de la entrada.
- **Detección de anomalías.** En este tipo de problemas el algoritmo de ML, busca detectar eventos anómalos, es decir muestras que no corresponden a la distribución normal de los datos. Un ejemplo, es la detección de transacciones fraudulentas.
- **Síntesis y muestreo.** En este caso, el algoritmo de ML debe generar nuevas muestras a partir de un conjunto de entrenamiento. Esto se aplica en los videojuegos, para la generación automática de texturas para objetos de gran tamaño.

2.3.1.2. El desempeño, P

Es muy importante medir el desempeño de un algoritmo de ML, usualmente la métrica utilizada puede variar según la tarea T . Para tareas de clasificación, usualmente se suele aplicar *Precision* y *Recall*, estos están detallados en las Ecuaciones 2.1 y 2.2 respectivamente (Dalianis, 2018).

$$Precision : P = \frac{tp}{tp + fp} \quad (2.1)$$

$$Recall : R = \frac{tp}{tp + fn} \quad (2.2)$$

tp , hace referencia a la cantidad de muestras que eran verdaderas y han sido reconocidas como verdaderas; fp , son las muestras que eran falsas, pero fueron reconocidas como verdaderas; fn , son las muestras que eran negativas y fueron reconocidas como negativas. Otra métrica importante es el F -score, este puede ser definido como el peso promedio de *Precision* y *Recall* (Dalianis, 2018). En la Ecuación 2.3, presentamos la definición.

$$F - score : F_{\beta} = (1 + \beta^2) * \frac{P * R}{\beta^2 * P + R} \quad (2.3)$$

Cuando $\beta = 1$:

$$F - score : F_1 = 2 * \frac{P * R}{P + R} \quad (2.4)$$

Finalmente otra métrica, aunque no muy recomendada para datos no balanceados es el *accuracy*. Este representa el porcentaje de muestras reconocidas correctamente.

$$Accuracy : acc = \frac{tp + tn}{tp + tn + fp + fn} \quad (2.5)$$

Para otro tipo de problemas, como regresión se puede aplicar el *error rate*, esta es una medida en los números reales y nos indica que tan diferente es la predicción realizada por un algoritmo de ML Goodfellow et al. (2016).

2.3.1.3. La experiencia, E

Según el tipo de experiencia que realizan los algoritmos de ML, se pueden clasificar en: Aprendizaje supervisado y Aprendizaje no supervisado Goodfellow et al. (2016).

- **Aprendizaje supervisado.** En este caso, cada muestra par el entrenamiento tiene los datos de entrada x y una etiqueta l . La idea es que el algoritmo de ML, pueda aprender de estos datos y luego realizar predicción de la etiqueta j tomando como entrada sólo los datos x .
- **Aprendizaje no supervisado.** En este caso, solo se cuenta con muestras no etiquetadas. Entonces el algoritmo de ML, debe agrupar los datos en *clusters*. Un ejemplo de estos problemas es la segmentación de clientes, segmentación de noticias, etc.

2.3.2. Redes neuronales

Uno de los modelos mas representativos de ML son la redes neuronales. Estas se basan en unidades llamadas neuronas (perceptron). En la Figura 2.9, se muestra esta representación, donde x_i , representa un atributo, w_i es el peso que se asigna al atributo x_i , de esta forma la neurona representa el resultado de multiplicar un peso a un atributo: $\sum_{i=1}^d x_i \cdot w_i$, una representación vectorial sería: $\mathbf{x}^T \mathbf{w}$ (Nielsen, 2015). Luego, a dicho resultado se aplica una función de activación, la función mas utilizada es la función sigmoidea (Equación 2.6 y 2.7).

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (2.6)$$

, donde $z = \sum_i w_i \cdot x_i - b$.

$$\frac{1}{1 + e^{-\sum_i w_i \cdot x_i - b}} \quad (2.7)$$

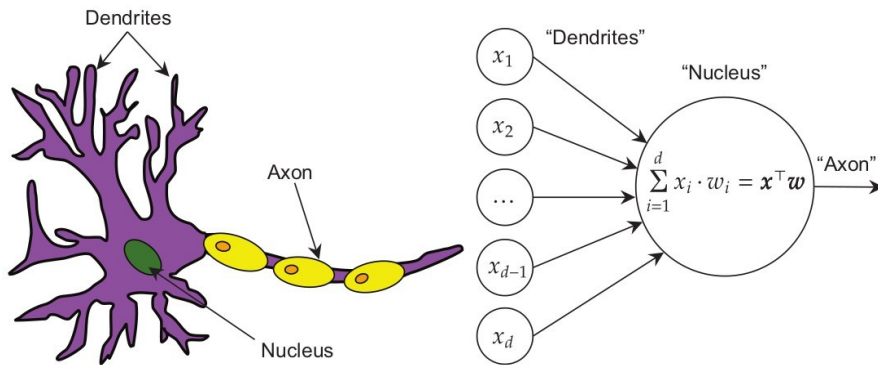


FIGURA 2.9: Representación de una neurona. Fuente: Raff (2022).

El perceptron, es capaz de solucionar varios problemas, pero para casos complejos puede formar una red, como se presenta en la Figura 2.10.

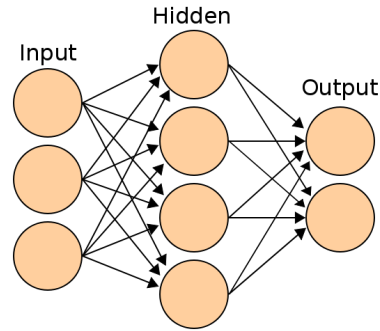


FIGURA 2.10: Representación de una red neuronal.

2.4. *Deep learning*

Deep learning (DL) es una subcategoría de *Machine Learning*, a diferencia de los algoritmos tradicionales de ML, usualmente DL trata con señales sin pre-procesamiento, los modelos (basados en redes neuronales) son mucho mas complejos tanto en dimensión como en el método de aprendizaje (El Naqa and Murphy, 2022). Por ejemplo, en la Figura 2.11, presentamos la relación entre inteligencia artificial, ML y DL, de ahí podemos concluir que ML es parte de la IA y DL es parte de ML (El Naqa and Murphy, 2022).

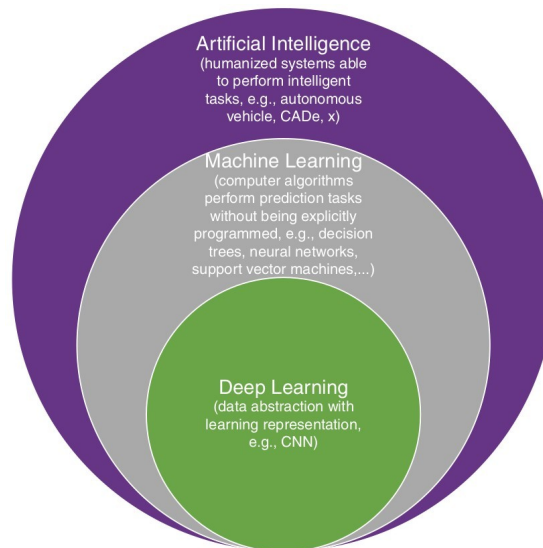


FIGURA 2.11: Relación entre Inteligencia Artificial, *Machine Learning* y *Deep Learning*.
Fuente: El Naqa and Murphy (2022).

2.4.1. *Deep Feedforward networks*

Deep Feedforward networks son perceptrones multicapa o *multilayer perceptrons* (MLP). Su objetivo es aproximar una función f^* , para el caso de clasificación, podría modelarse como $y = f^*(x)$. Luego, un *feedforward network*, define un mapeo $y = f(x; \theta)$ y aprende los valores de los parámetros θ [Goodfellow et al. \(2016\)](#). Entonces un *Deep Feedforward networks*, es una red neuronal tradicional pero con un número grande de neuronas y capas (Figura 2.12). Existen muchos tipos de *Deep Feedforward networks*, estas serán detalladas en los siguientes apartados.

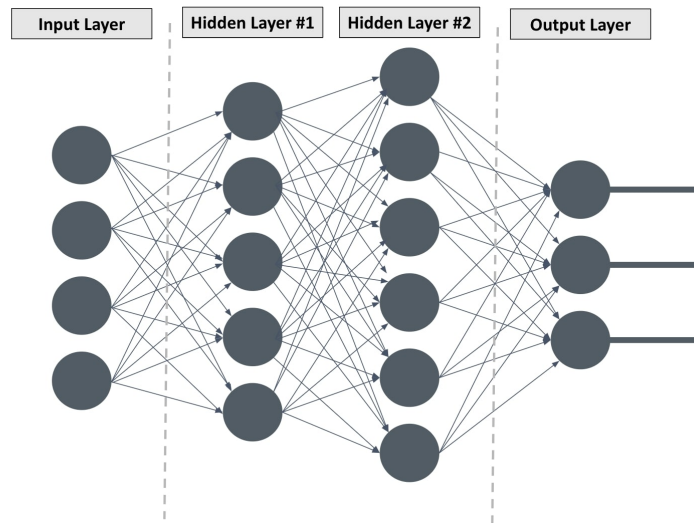


FIGURA 2.12: Representación de un *Deep Feedforward Network*. Fuente: [El Naqa and Murphy \(2022\)](#).

2.4.2. *Convolutional Neural Networks*

Una *Convolutional Neural Networks* (CNN), es una red neuronal basada en la operación de convoluciones (utilizada en procesamiento de imágenes). Generalmente estas redes neuronales se aplican a problemas de visión computacional ([Zhang et al., 2021](#)). La operación básica es la convolución, esta se presenta en la Figura 2.13. Se toman pequeñas ventanas de una imagen y se realiza el producto punto con un *kernel* ya establecido. Según los diferentes valores del *kernel*, se pueden obtener diferentes resultados en la imagen de salida como: detección de bordes, suavizados, dilatación, etc.

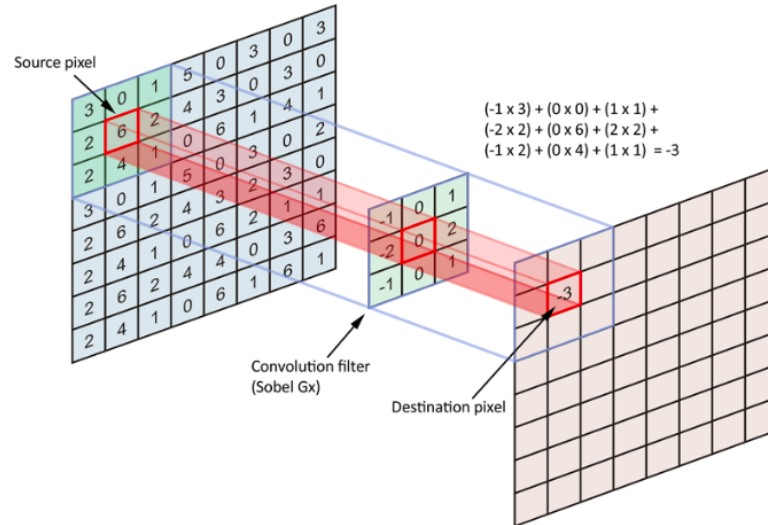


FIGURA 2.13: Ejemplo de una convolución en procesamiento de imágenes. Fuente: Shuchen (2022).

Con inspiración en la operación de convolución, se plantean las CNN por primera vez por LeCun et al. (1998). En la Figura 2.14, se presenta la LeNet-5, planteado por los autores. Luego, surgen diversas propuestas como AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan and Zisserman, 2014), GoogleNet (Szegedy et al., 2015) y ResNet (He et al., 2016).

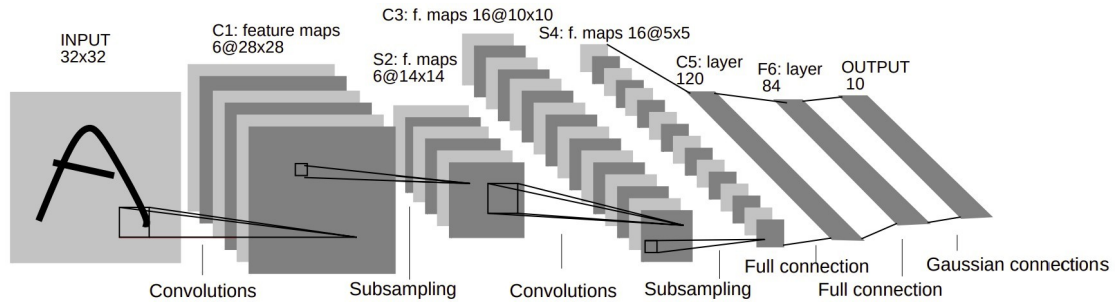


FIGURA 2.14: Arquitectura de LeNet-5, una CNN para el reconocimiento de dígitos. Fuente: LeCun et al. (1998).

2.4.3. Recurrent Neural Networks

Mientras que las CNN están especializadas para manejar información espacial, las *Recurrent Neural Networks* (RNN), se especializan en información secuencial (Zhang et al., 2021). En este campo, se habla del tiempo como una variable y se tratan problemas de series temporales por ejemplo.

El término RNN, aparece por primera vez en los trabajos de Rumelhart et al. (1985) y Jordan (1997). Algunos autores, comentan también que el inicio de las RNN fue con las

redes de Hopfield (Hopfield, 1982). En general estas RNN, tienen dos entradas: estado actual y estado anterior; luego la RNN predice el siguiente estado. El problema de estas redes neuronales surge por una falta de memoria, es decir cuando tenemos varios estados, el estado inicial va a influenciar cada vez menos a los estados futuros.

Como alternativa de solución al problema mencionado anteriormente, surgen Long Short-Term Memory, propuesta por Hochreiter and Schmidhuber (1997). Una red neuronal LSTM, es capaz de recordar un dato relevante de una secuencia y almacenarlo varios instantes de tiempo. En la Figura 2.15, explicamos brevemente el funcionamiento de LSTM, los datos que ingresan a una compuerta (*gate*), son los datos de entrada en un tiempo específico y el estado oculto anterior. Luego, es procesado por tres capas totalmente conectadas: *input gate*, *forget gate* y *output gate* (Zhang et al., 2021).

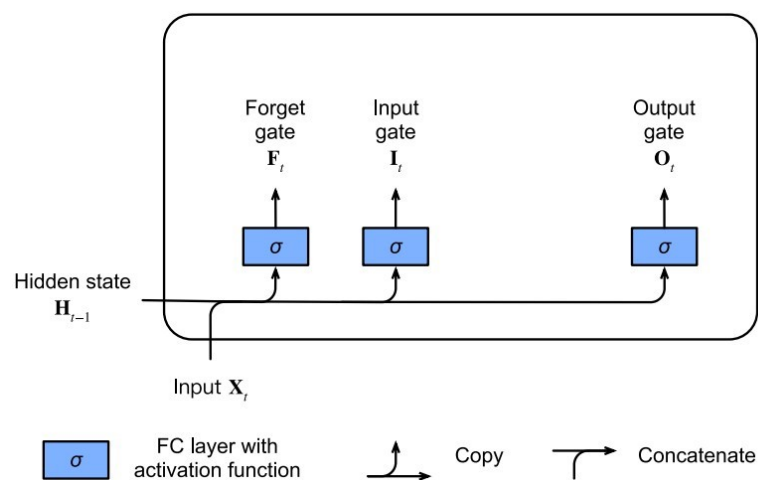


FIGURA 2.15: Ejemplo del procesamiento del *input gate*, *forget gate* y *output gate* de LSTM. Fuente: Zhang et al. (2021).

2.4.4. Transformers

Los *Transformers* son propuestas por Vaswani et al. (2017), para dar solución al problema de *long-range dependency*. Por ejemplo el autor comenta: “The Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution”. Del enunciado anterior, *transduction* hace referencia a la conversión secuencias de entrada hacia otro formato. Otro termino interesante es *self-attention* (Figura 2.16), este permite al modelo mirar hacia otras palabras en la secuencia de entrada para tener un mejor entendimiento de cierta palabra en la secuencia (Kelvin, 2022).

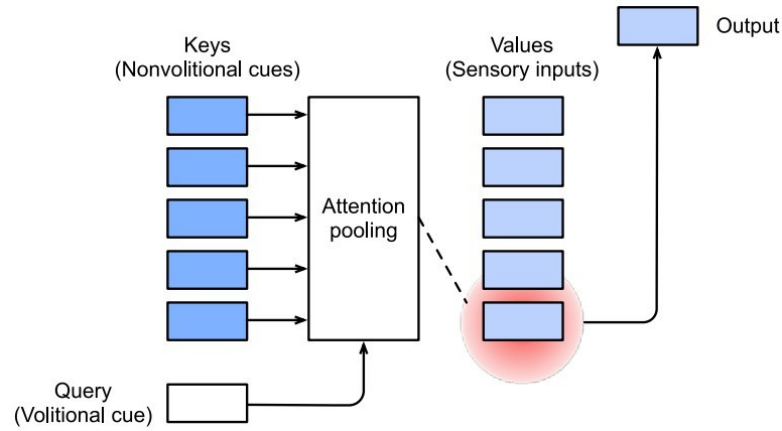


FIGURA 2.16: ejemplo del mecanismo de atención de una red *Transformer*. Fuente: [Zhang et al. \(2021\)](#).

2.4.5. *BERT*

Bidirectional Encoder Representations from Transformers (BERT), propuesta por [Devlin et al. \(2018\)](#), está inspirada por la red *Transformer* y su mecanismo de atención, la cuál entiende la relación contextual entre diferentes palabras. A diferencia de una RNN, BERT no tiene dirección, es decir lee la secuencia entera. Esta característica, le permite al modelo aprender información contextual de una palabra con respecto a las otras ([Kelvin, 2022](#)).

Capítulo 3

Estado del Arte

En este capítulo presentaremos los resultados de la Revisión Sistemática de la Literatura (RSL) referente a los métodos de detección de neo antígenos con técnicas de *deep learning* y desde una perspectiva en las ciencias de la computación.

3.1. Revisión Sistemática de la Literatura (RSL)

Con el objetivo de mapear las principales técnicas de detección de neo antígenos, se planteó desarrollar una Revisión Sistemática de la Literatura (RSL). La RSL, se enfocó en los métodos basados en *deep learning* y desde una perspectiva de las ciencias de la computación. Se definió este objetivo, porque en la literatura ya existían varios otros *reviews*, enfocados en el proceso general de vacunas personalizadas, y detección de neo antígenos. En esta sección, se describe el proceso que se llevó a cabo y sus resultados.

3.1.1. Cadenas de búsqueda y bases de datos

En la Tabla 3.1, se presentan las cadenas de búsqueda utilizadas para la RSL. Generalmente los términos sinónimos a *neoantigen* utilizados en la literatura son *peptide* y *epitope*. Luego, algunos trabajos se enfocan en predecir el enlace entre un péptido y la molécula MHC, pero para células humanas la molécula MHC tiene el nombre de HLA. Además, hay varias clases como MHC-I y MHC-II. Debido a eso, se tenía que considerar todos esos sinónimos de MHC. También, otra diferencia existe en el término “enlace”, del enlace péptido con MHC, algunos trabajos se refieren a él con los términos: *binding*, *presentation*, *prediction* y *detection*. Finalmente, algunos trabajos se enfocan en otra fase de la detección de neo antígenos, esta consiste en predecir el enlace entre el compuesto pMHC y T-cell Receptor (TCR) de las células T.

Luego, se utilizó Google Scholar y Mendeley como motores de búsqueda al ser estos unos motores que indexan casi la totalidad de artículos científicos. Utilizando estas herramientas, se obtuvo artículos de las bases de datos descritas en la Tabla 3.2.

TABLA 3.1: Cadenas de búsqueda utilizadas en la RSL.

Cadena de búsqueda

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

TABLA 3.2: Bases de datos utilizadas en la RSL.

Bases de datos

IEEE Xplore

Science Direct

Springer

ACM Digital Library

PubMed

BioRxiv

3.1.2. Selección de artículos

Con las cadenas de búsqueda y considerando solo los artículos desde el 2018, se analizó el título de cada artículo encontrado por los motores de búsqueda y se seleccionaron 334 artículos. En la Tabla 3.3, se presenta la cantidad de artículos publicados por año. Para el caso del 2022, solo se tienen 57 artículos porque esta tesis se redactó a mediados del año 2022.

Del total de artículos encontrados (342), se seleccionó un subconjunto basado en los criterios de inclusión y exclusión presentados de la Tabla 3.4. Estos criterios incluían que el artículo pertenezca a un *conference* o *journal* reconocido, que tenga una metodología detallada y que pertenezca al área de ciencia de la computación. Luego, en la Tabla 3.4, se

puede ver que hay un puntaje según cada criterio de inclusión, se utilizó este puntaje para calificar cada artículo y luego se seleccionaron los artículos que tenían un puntaje mayor a 4. En este proceso, se analizó el *abstract* de los artículos y ciertas partes importantes según era necesario para asignar el puntaje. Al finalizar esta etapa, se obtuvieron 253 artículos, estos son los trabajos que se han analizado en la RSL. Adicionalmente, a los artículos seleccionados, se han considerado otros trabajos importantes que proponían bases de datos, *pipelines* y *reviews*.

TABLA 3.3: Cantidad de artículos encontrados y seleccionados según los criterios de inclusión y exclusión en la RSL.

Año	Artículos encontrados	Artículos seleccionados
2018	53	42
2019	79	52
2020	81	67
2021	64	51
2022	57	41
Total	342	253

TABLA 3.4: Criterios de inclusión y exclusión de artículos utilizados en la RSL.

Criterios de inclusión	Criterios de exclusión	Puntaje
Artículos con categoría ERA (A, B o C) si son conferencias y Journals Q1, Q2 o Q3.	No considerar los trabajos de baja calidad, que no esten rankeados.	3
Trabajos que se basen en <i>deep learning</i> para la detección de neo antígenos.	Trabajos que se basan en el uso de alguna herramienta (investigaciones realizadas por científicos de otras areas).	2
La metodología es detallada.		2
Tiene resultados clínicos		2
Tiene repositorio de código fuente.		1
Comparte la base de datos utilizada.		1

3.2. Resultados de la RSL

El proceso para la detección de neo antígenos, es complejo, y generalmente consiste en: (1) extracción del tejido tumoral y secuenciamiento, (2) identificación de mutaciones, (3) detección de péptidos como resultado de alineamiento con muestras sanas, (4) predicción de *peptide-MHC binding* (*pMHC*), (5) predicción de *pMHC presentation* y (6) predicción del enlace pMHC-TCR (Mattos et al., 2020; Peng et al., 2019). De este proceso, la mayoría de investigaciones se centra en el problema de *peptide-MHC binding*, *peptide-MHC presentation* y predicción del enlace pMHC-TCR. Entonces, se va a reportar los trabajos relacionados según esta clasificación. Tambien, se van a incluir en otra clasificación, los pipelines que integran varias herramientas para todo el proceso de detección de neo antígenos; Investigaciones que presentan bases de datos; y finalmente *reviews* relacionados a la tesis.

Neoantigen detection is an extensive process, and it can be divided into three phases or approaches. For example, in Figure 3.1, we show these three phases: first, from DNA-seq or RNA-seq as inputs, there are tools to elude possible peptides; second, it is necessary to determine which peptides bind to MHC (pMHC); finally, some works study the interaction between pMHC and TCR. The first stage (eluted peptides), depends on several tools and is almost resolved, so the current research article on neoantigen detection focus on pMHC binding and pMHC-TCR interaction. In this context, in this review, we focus only on deep learning techniques used in pMHC binding and pMHC-TCR interaction.

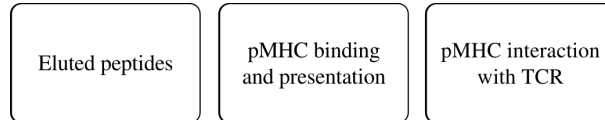


FIGURA 3.1: The three phases and approaches that researches focus in the problem of neoantigen detection.

3.2.1. Peptide-MHC (pMHC) binding and presentation

Not all peptides eluted from tumor mutations are neoantigens; it depends on which peptides could bind to MHC. So, several works focus on peptide-MHC binding (pMHC). Moreover, there are three types of MHC: MHC class I, class II and class III. In human cells, they are named Human Leukocyte Antigen (HLA); for MHC class I, we have HLA-A, HLA-B, and HLA-C; and for MHC class II, HLA-DR, HLA-DQ, and HLA-DP Neefjes et al. (2011).

Most works focus in pMHC class I, because there is a lot of data, and the problem is not too complicated. The pMHC class I, represent a problem where a 9-mer peptide could bind to a MHC class I. In the other hand, the prediction of peptide MHC class II binding is challenging because the peptides are larger (normally 15-mer) and it binds to two amino acid chains in MHC class II.

The prediction of peptide-MHC binding usually refers to when a peptide bind to the MHC inside a nucleic cell. However, peptide-MHC presentation refers to the peptides that are presented by MHC in the membrane of cells. For instance, the process for peptide MHC binding and presentation is: first, antigens are degraded by the proteasome; then, the resulting peptides are translocated via a transporter associated with antigen presentation (TAP) into the endoplasmic reticulum (ER) and loaded onto MHC (peptide MHC binding); then, peptide-MHC class I complexes are released from the ER and transported via the Golgi to the plasma membrane (peptide MHC presentation); finally, the peptide-MHC is presented to TCR [Neefjes et al. \(2011\)](#) (see Figure 3.2). So, peptide-MHC binding (pMHC) and peptide-MHC presentation are two problems because not all pMHC are presented in the membrane [Mill et al. \(2022\)](#); [Mattos et al. \(2020\)](#); [Mill et al. \(2022\)](#); [Bulik-Sullivan et al. \(2019\)](#); [Bassani-Sternberg et al. \(2015\)](#); [Yadav et al. \(2014\)](#).

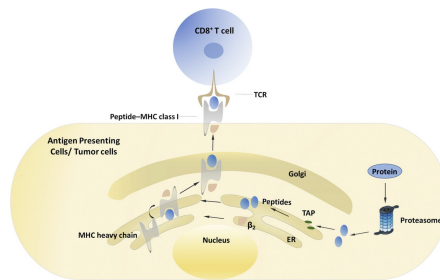


FIGURA 3.2: Peptide presentation by MHC class I. Source: [Zhang et al. \(2019\)](#)

In this section, we analyzed all research that used deep learning techniques for peptide-MHC binding and peptide-MHC presentation. Moreover, we divided the deep learning techniques in Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Transformers, and classic machine learning methods.

3.2.1.1. CNN

Convolutional Neural Networks (CNN) have been used in a range of different fields like computer vision, speech recognition, face recognition, object detection, etc. [Alzubaidi et al. \(2021\)](#). Moreover, CNNs, are helpful because they notably reduce the number of parameters and enhance generalization. CNNs are regularly used in computer vision because they extract local spatial features and combine them with higher-order features [Marais et al. \(2020\)](#).

Several works apply CNNs for peptide-MHC binding and presentation. They differ in how they compute the inputs (amino acid sequences). Typically, they use one-hot encoding or BLOSUM to vectorize each amino acid, then join each vector and build a matrix that could be managed like an image. In Table 3.5, we present a list of the most recent works that use CNNs for neoantigen detection; we include the method used for encoding amino acids, what type of MHC class they focus and a link to the repository and database used for the experiments. Also, we distinguish if the research is used for pMHC binding or pMHC presentation.

Furthermore, CNNs could be used with RNNs to improve generalization, or even we could include attention mechanisms. In this context, in Table 3.5, we present a list of works that used modified CNNs architectures. For instance, the first three rows detail the proposals with CNNs and attention mechanisms: DeepNetBim [Yang et al. \(2021\)](#), DeepAttentionPan [Jin et al. \(2021\)](#) and ACME [Hu et al. \(2019\)](#). Then, in the last row, we describes MHCherryPan [Xie et al. \(2020\)](#), which combines CNNs and RNNs.

3.2.1.2. RNN

In Table 3.7, we present a list of most recent works that uses RNNs for peptide-MHC binding and presentation prediction. For instance, in this case there works that used RNN with attentions mechanisms: MATHLA [Ye et al. \(2021\)](#), DeepSeqPanII [Liu et al. \(2021\)](#) and DeepHLApan [Wu et al. \(2019\)](#). There is also a works that uses standalone RNN, like a GRU model [Heng et al. \(2021\)](#), MHCnuggets which used a LSTM, and BVLSTM-MHC which used a bilateral and variable long short-term memory (BVLSTM).

TABLA 3.5: List of research since 2018 that uses CNNs for peptide-MHC binding and presentation.

Ref.	Approach Name		MHC class	Encoding	Description
You et al. (2022)	peptide-MHC binding	DeepMHCII	MHC class-II	Binding core and Peptides Flanking Regions (PFR), sequences and interactions between a peptide and an MHC-II molecule	Uses a Binding Interaction Convolutional Layer (BICL) in order to get the representation of the interaction between peptides and MHC-II.
Li et al. (2021)	peptide-MHC binding	DeepImmuno	MHC class-I	AAindex1 database	The authors evaluated several machine learning methods, CNN got the best results.
Lang et al. (2021)	peptide-MHC presentation	APPM	MHC class-I	One-hot encoding	Three parallel CNN with MS data.
Lee et al. (2021)	peptide-MHC presentation	MHCfovea	MHC class-I	One-hot encoding	Clusters N-terminal and C-terminal sub-motifs on observed and unobserved alleles, then MHCfovea calculated the hyper-motifs in order to disclose the relation between binding motifs and MHC-I
Junet and Daura (2021)	peptide-MHC binding	CNN-PepPred	MHC class-II	BLOSUM	CNN is used to know motifs in peptide-MHC complexes.
Pei and Hsu (2020)	peptide-MHC binding	IConMHC	MHC class-I	PCA of Aminoacid interaction from AAindex3 database	Learns from physical and chemical interaction properties between pairwise amino acids from the two molecules.
Saxena et al. (2020)	peptide-MHC binding	OnionMHC	MHC class-I	BLOSUM and structural features	Utilized complex structure and the peptide sequence features to predict the binding affinity of peptides to HLA-A*02:01 -
Ng et al. (2020)	peptide-MHC presentation	MINERVA	MHC class-I	physicochemical properties	The model learned the determinants of peptideMHC presentation,

TABLA 3.6: List of research since 2018 that uses CNNs s with RNN or attention mechanisms for peptide-MHC binding and presentation.

Ref.	Approach Name		MHC class	Encoding	Description
Yang et al. (2021)	peptide-MHC binding	DeepNetBim	MHC class-I	BLOSUM	Stands for binding and immunogenecity prediction with a CNN and an attention mechanism to make the final prediction. Encode data using BLOSUM50
Jin et al. (2021)	peptide-MHC binding	Deep Attention Pan	MHC class-I	BLOSUM	Uses CNN with a attention mechanism and uses 20 trained netowrks to stabilize prediction performance. Moreover, they used BLOSUM62 to encode input data.
Hu et al. (2019)	peptide-MHC binding	ACME	MHC class-I	BLOSUM	They used CNN with attention mechanism to extract interpretable binding patterns. For instance, this module assigned weights to the feature vectors (residue positions) and then computed their weighted average to facilitate prediction Moreover, each amino acid is decoded using BLOSUM50.
Xie et al. (2020)	peptide-MHC binding	MHCherryPan	MHC class-I	BLOSUM	Combines CNN and RNN in order to deal with any length peptide sequences. it can be used in RNA-protein, DNA-protein, and RNA-DNA binding prediction.

TABLA 3.7: List of research since 2018 that uses RNNs for peptide-MHC binding and presentation.

Ref.	Approach Name	MHC class	Encoding	Description
Ye et al. (2021)	peptide-MHC binding	MATHLA MHC class-I	BLOSUM	Integrates bi-directional long short-term memory network and multiple head attention mechanism. The model got good results for large ligand sequences (11 to 15 amino acids)
Liu et al. (2021)	peptide-MHC binding	DeepSeqPanIIMHC class-II	One-hot encoding and BLOSUM	It is a LSTM with attention mechanism, it have three parts: sequence encoders, binding context extractor, and affinity predictor.
Heng et al. (2021)	peptide-MHC binding	GRU-based RNN MHC class-II	Embedding layer	The authors uses a GRU model and present represent each sequence using an embedding layer which outputs 128 dim vectors.
Jiang et al. (2021)	peptide-MHC binding	BVLSTM-MHC MHC class-I	One-hot encoding and BLOSUM	Review of current methods for predicting peptide-MHC-I binding problem. Moreover, the authors developed a bilateral and variable long short-term memory (BVLSTM).
Shao et al. (2020)	peptide-MHC binding	MHCnuggets MHC class-I and II	Onehot encoding	It uses a a long short-term memory (LSTM) model. Additionally, the authors processed 26.3 million allele-peptide comparisons yielding 101,326 unique predicted immunogenic missense mutations (IMM)
Wu et al. (2019)	peptide-MHC binding	DeepHLApan MHC class-I	Onehot encoding	Uses a Bidirectional Gated Recurrent Unit (Bi-GRO) with attention mechanism. The model got good results on unseen HLA alleles.

3.2.1.3. Transformers

The attention mechanism was first proposed by Bahdanau in 2014 [Bahdanau et al. \(2014\)](#), in order to resolve the bottleneck problem using a fixed-length encoding vector. With this new approach, the authors got comparable state-of-art results for English to French translation. Then, this attention mechanism is used for natural language inference [Parikh et al. \(2016\)](#), and a structured attention networks is proposed [Kim et al. \(2017\)](#). Nevertheless, these attentions modules are commonly used in conjunction with a recurrent network. Then, in 2017, the paper: “Attention Is All You Need” [Vaswani et al. \(2017\)](#), proposed a new network architecture (Transformer) base solely on attention mechanisms.

In Table 3.8, we present a list of four works that recently used a transformer network in peptide-MHC binding and presentation. The majority of them uses a BERT architecture [Devlin et al. \(2018\)](#), which stands for a bidirectional transformer for language representation. For instance, we have

3.2.2. Pipelines

Debido a la complejidad del proceso y la gran cantidad de métodos desarrollados, se ha desarrollado software y *pipelines* que pretenden facilitar el uso de estas herramientas. Entre los *pipelines* más conocidas antes del 2018 tenemos: Somaticseq ([Fang et al., 2015](#)), CloudNeo ([Bais et al., 2017](#)), MuPeXI ([Bjerregaard et al., 2017](#)), NeoepitopePred ([Tran et al., 2015](#)), y NeoFuse ([Gros et al., 2016](#)). Estas herramientas en su mayoría toman como entrada archivos Variant Calling Files (VCF) y archivos de alineamiento BAM, para la detección de mutaciones (inserciones, eliminaciones y fusión de genes) y posibles neo antígenos. Luego, también hemos detallado, un conjunto de herramientas a partir del 2018, en la Tabla 3.9.

TABLA 3.8: List of research since 2018 that uses Transformers (self-attention) for peptide-MHC binding and presentation.

Ref.	Approach Name	MHC class	Encoding	Description
Wang et al. (2022)	peptide-MHC binding	MHCroBERTa MHC class-I	Tokenized from a pre-trained model	The problem is resolved like a NLP approach. Moreover, they used transfer learning. This work outperformed NetMHCpan 3.0.
Chu et al. (2022)	peptide-MHC binding	TransPhLA MHC class-I	Character embedding model	The authors used a character embedding model to vectorize each amino acid. They used self-attention mechanism based in four blocks: the embedding block, encoder block, feature optimization block, encoder block.
Cheng et al. (2021)	peptide-MHC binding	BERTMHC MHC class-II	Embedding layer	The authors used the pre-trained model TAPE to embed each aminoacid into a 768-dimensional vector. The self-attention mechanism learns the interaction of all possible amino acid pairs in the input sequence
Gasser et al. (2021)	peptide-MHC presentation	ImmunoBERTMHC class-I	Embedding layer	They used TAPE as the embedding layer. Also, The authors used SHAP and LIME to analyze the results and they concluded that N- and C-terminals are highly relevant.

TABLA 3.9: Listado de *pipelines* desde el 2018, para la detección de neo antígenos.

Nombre	Autor-año	Entrada	Salida
Neopepsee	Kim et al. (2018)	RNA-seq, somatic mutations (VCF), tipo de HLA (opcional)	Neo antígenos y niveles de expresión de los genes
PGV Pipeline	Rubinsteyn et al. (2018)	DNA-seq	Neo antígenos
ScanNeo	Wang et al. (2019)	RNA-seq	Neo antígenos
NeoPredPipe	Schenck et al. (2019)	Mutaciones (VCF) y tipo de HLA	Neo antígenos y anotación de variantes
pVACtools	Hundal et al. (2020)	Mutaciones (VCF)	Neo antígenos
ProGeo-neo	Li et al. (2020)	RNA-seq y somatic mutations (VCF)	Neo antígenos
neoepiscope	Wood et al. (2020)	Somatic mutations (VCF) y archivos BAM	Neo antígenos y mutaciones
neoANT-HILL	Coelho et al. (2020)	RNA-seq y somatic mutations (VCF)	Neo antígenos, y niveles de expresión de los genes
NAP-CNB	Wert-Carvajal et al. (2021)	RNA-seq	Neo antígenos
Valid-NEO	Teraï et al. (2022)	Somatic mutations (VCF), tipo de HLA (opcional)	Neo antígenos

3.2.3. Bases de datos

En la Tabla 3.10, presentamos una lista de bases de datos públicas. Estas bases de datos se centran en la interacción *peptide-MHC* ([Wu et al., 2018](#); [Zhou et al., 2019](#); [Tan et al., 2020](#); [Lu et al., 2022](#)) y pMHC con TCR ([Shugay et al., 2018](#); [Bagaev et al., 2020](#)). También, hay un trabajo que presenta las estructuras 3D de las péptidos y HLA abriendo una nueva rama de investigación desde otro enfoque. Finalmente, la base de datos por excelencia IEDB ([Vita et al., 2018](#)).

TABLA 3.10: Bases de datos públicas de *pMHC binding*, *pMHC presentation*, interacción pMHC-TCR y estructuras 3D de proteínas.

Nombre	Autor-año	Descripción
VDJdb	Shugay et al. (2018) y Bagaev et al. (2020)	Base de datos del enlace TCR con pMHC, cuenta con 5491 muestras
IEDB	Vita et al. (2018)	La base de datos mas grande, contiene información <i>T-cell epitopes</i> de humanos y otros organismos.
TSNAdb	Wu et al. (2018)	Contiene 7748 muestras de mutaciones y HLA de 16 tipos de Cáncer.
NeoPeptide	Zhou et al. (2019)	Contiene muestras de neo antígenos, resultado de mutaciones somáticas y artículos relacionados. Contiene 1818137 epitopes de ms de 36000 neo antígenos.
pHLA3D	Oliveira et al. (2019)	Presenta 106 estructuras 3D de las cadenas alpha, β_2M y peptidos de las moléculas HLA-I
dbPepNeo	Tan et al. (2020)	Tiene muestras validadas del enlace <i>peptide-MHC</i> , a partir de MS. Contiene 407794 muestras de baja calidad, 247 de mediana calidad y 295 muestras de alta calidad.
dbPepNeo2. 0	Lu et al. (2022)	Recolecta una lista de neo antígenos y moléculas HLA. Presenta 801 muestras de alta calidad y 842289 de mala calidad de HLAs. Tambien, 55 neo antígenos de clase II y 630 neo antígenos enlazados a TCR.
IntroSpect	Zhang et al. (2022a)	Herramienta para la construcción de bases de datos sobre <i>peptide-MHC binding</i> . Utiliza datos de <i>Mass Spectrometry</i>
IPD-IMGT/HLA	Robinson et al. (2020)	Tiene secuencias de moléculas MHC, 25000 alleles de 45 genes.

3.2.4. Reviews

La detección de neo antígenos es un problema interdisciplinar y esto ha originado varios *reviews* desde diferentes perspectivas. Entonces se ha planteado la siguiente clasificación: basados en *Next-Generation Sequencing*, *Mass Spectrometry*, interacción *peptide-MHC*,

basados en información estructural, enfocados en TCR, buenas prácticas y los enfocados en el proceso completo de generación de vacunas personalizadas.

Primero, presentamos los trabajos que se enfocan en estudios de *Next-Generation Sequencing* (Tabla 3.11), para la detección de neo antígenos e inmunoterapia del Cáncer. Estos trabajos principalmente utilizan información secuencial de *DNA* y gracias a las tecnologías modernas ahora se pueden considerar las secuencias de *RNASeq*. Las tecnologías de *RNASeq*, proveen información mas precisa de la transcripción e identificación de isoformas que otros métodos (Wang et al., 2009). Mayormente, estas tecnologías se limitan a algoritmos alineamiento con genomas de referencia (Groisberg et al., 2018).

TABLA 3.11: Listado de los *reviews*, que se enfocan en estudios de *Next-Generation Sequencing* para la detección de neo antígenos e inmunoterapia del Cáncer.

Autor-año	Título
Zhou et al. (2022)	A Comprehensive Survey of Genomic Mutations in Breast Cancer Reveals Recurrent Neoantigens as Potential Therapeutic Targets
Battaglia (2020)	Neoantigen prediction from genomic and transcriptomic data
Mirandola et al. (2020)	The Quest for the Next-Generation of Tumor Targets: Discovery and Prioritization in the Genomics Era
Groisberg et al. (2018)	Immunotherapy and next-generation sequencing guided therapy for precision oncology: what have we learnt and what does the future hold?

Algunos trabajos son más específicos, y se enfocan en la interacción de un péptido y la molécula MHC. Esta interacción es un factor clave, porque si se forma el enlace pMHC y luego este compuesto es presentado a las células T, es posible activar el sistema inmune. En la Tabla 3.12, se presenta estos *reviews*. La mayoría de estos trabajos, se centran en la molécula MHC-I (Mateo et al., 2020; Mei et al., 2020; Schmidt and Lill, 2019; Mei et al., 2020) , molécula MHC-II (Jensen et al., 2018) y todos los tipos de MHC en general (Nielsen et al., 2020; Liu et al., 2020,?). También, hay trabajos que estudian la complejidad de esta molécula y todos sus *alleles* (Radwan et al., 2020).

TABLA 3.12: Listado de los *reviews*, que se enfocan en estudios de la interacción de péptidos y la molécula MHC, para la detección de neo antígenos.

Autor-año	Título
Mateo et al. (2020)	Comparison of machine learning models for the prediction of cancer cells using MHC class I complexes
Mei et al. (2020)	A comprehensive review and performance evaluation of bio-informatics tools for HLA class I peptide-binding prediction
Nielsen et al. (2020)	Immunoinformatics: predicting peptide-MHC binding
Liu et al. (2020)	A review on the methods of peptide-MHC binding prediction
Paul et al. (2020b)	Major histocompatibility complex binding, eluted ligands, and immunogenicity: benchmark testing and predictions
Radwan et al. (2020)	Advances in the Evolutionary Understanding of MHC Polymorphism
Schmidt and Lill (2019)	MHC class I presented antigens from malignancies: A perspective on analytical characterization & immunogenicity
Jensen et al. (2018)	Improved methods for predicting peptide binding affinity to MHC class II molecules
Mei et al. (2020)	A comprehensive review and performance evaluation of bio-informatics tools for HLA class I peptide-binding prediction
Liu et al. (2020)	A review on the methods of peptide-MHC binding prediction
Paul et al. (2020b)	Major histocompatibility complex binding, eluted ligands, and immunogenicity: benchmark testing and predictions
Radwan et al. (2020)	Advances in the Evolutionary Understanding of MHC Polymorphism
Schmidt and Lill (2019)	MHC class I presented antigens from malignancies: A perspective on analytical characterization & immunogenicity
Jensen et al. (2018)	Improved methods for predicting peptide binding affinity to MHC class II molecules

La mayoría de *reviews* estudian las técnicas basadas en secuencias de DNA y RNA, pero recientemente se está utilizando *Mass spectrometry*, para secuenciar los péptidos y moléculas MHC ya enlazados y presentes en las membranas de las células. Este avance ha impulsado la creación de nuevas bases de datos y métodos para el problema de *peptide-MHC presentation*. En este contexto, en la Tabla 3.13, se presentan todos los *reviews*, enfocados en estudiar *Mass spectrometry* para la detección de neo antígenos.

TABLA 3.13: Listado de los *reviews*, que se enfocan en estudios de *Mass spectrometry* para la detección de neo antígenos.

Autor-año	Título
Kote et al. (2020)	Mass spectrometry-based identification of MHC-associated peptides
Kote et al. (2020)	Mass spectrometry-based identification of MHC-associated peptides
Zhang et al. (2019)	Application of mass spectrometry-based MHC immunopeptidome profiling in neoantigen identification for tumor immunotherapyA
Chen et al. (2021b)	Identification of MHC peptides using mass spectrometry for neoantigen discovery and cancer vaccine development
Creech et al. (2018)	The role of mass spectrometry and proteogenomics in the advancement of HLA epitope prediction
Zhang et al. (2019)	Application of mass spectrometry-based MHC immunopeptidome profiling in neoantigen identification for tumor immunotherapyA
Creech et al. (2018)	The role of mass spectrometry and proteogenomics in the advancement of HLA epitope prediction

En si la detección de neo antígenos, es un proceso muy largo e integra métodos de secuenciamiento, alineamiento, detección de mutaciones, identificación de péptidos, predicción de la interacción *peptide-MHC*, y finalmente el trabajo biotecnológico para la generación de vacunas. Entonces, en la Tabla 3.14 y 3.15, se presenta el lista de *reviews*, que explican el problema de generación de vacunas pero desde una vista panoramica incluyendo todo el proceso completo. Algunos trabajos se enfocan en demostrar la posibilidad de crear vacunas personalizadas contra en Cáncer ([Lang et al., 2022](#); [Richard et al., 2022](#); [Pao et al., 2022](#); [Reynolds et al., 2022](#); [McCaffrey, 2022](#); [Fritsch et al., 2020](#)) y otros trabajos, priorizan la importancia de los neo antígenos ([Okada et al., 2022](#); [Zheng et al., 2022](#); [Wang et al., 2021b](#); [Pearlman et al., 2021](#); [Arnaud et al., 2020](#); [Han et al., 2020](#)).

TABLA 3.14: Listado de los *reviews*, que se enfocan en presentar en proceso general de detección de neo antígenos y vacunas personalizadas del año 2022 y 2021.

Autor-año	Título
Tran et al. (2022)	A tale of solving two computational challenges in protein science: neoantigen prediction and protein structure prediction
Lang et al. (2022)	Identification of neoantigens for individualized therapeutic cancer vaccines
Okada et al. (2022)	Identification of Neoantigens in Cancer Cells as Targets for Immunotherapy
Bollineni et al. (2022)	Chasing neoantigens; invite naïve T cells to the party
Richard et al. (2022)	Neoantigen-based personalized cancer vaccines: the emergence of precision cancer immunotherapy
Pao et al. (2022)	Therapeutic Vaccines Targeting Neoantigens to Induce T-Cell Immunity against Cancers
Fang et al. (2022)	Neoantigens and their potential applications in tumor immunotherapy
Zheng et al. (2022)	Neoantigen: A Promising Target for the Immunotherapy of Colorectal Cancer
Redwood et al. (2022)	What's next in cancer immunotherapy?-The promise and challenges of neoantigen vaccination
Reynolds et al. (2022)	Neoantigen Cancer Vaccines: Generation, Optimization, and Therapeutic Targeting Strategies
Roesler and Anderson (2022)	Beyond Sequencing: Prioritizing and Delivering Neoantigens for Cancer Vaccines
McCaffrey (2022)	Bioinformatic Techniques for Vaccine Development: Epitope Prediction and Structural Vaccinology
Fotakis et al. (2021)	Computational cancer neoantigen prediction: current status and recent advances
Wang et al. (2021a)	Beyond tumor mutation burden: tumor neoantigen burden as a biomarker for immunotherapy and other types of therapy
Ferreira et al. (2021)	Glycoproteogenomics: Setting the Course for Next-generation Cancer Neoantigen Discovery for Cancer Vaccines
Blass and Ott (2021)	Advances in the development of personalized neoantigen-based therapeutic cancer vaccines
Wang et al. (2021b)	Gene fusion neoantigens: Emerging targets for cancer immunotherapy
Pearlman et al. (2021)	Targeting public neoantigens for cancer immunotherapy

TABLA 3.15: Listado de los *reviews*, que se enfocan en presentar en proceso general de detección de neo antígenos y vacunas personalizadas del año 2020 y 2019.

Autor-año	Título
Arnaud et al. (2020)	Biotechnologies to tackle the challenge of neoantigen identification
Fritsch et al. (2020)	Personal neoantigen cancer vaccines: a road not fully paved
Holtsträter et al. (2020)	Bioinformatics for cancer immunotherapy
Roudko et al. (2020)	Computational prediction and validation of tumor-associated neoantigens
Esprit et al. (2020)	Neo-antigen mRNA vaccines
Chen et al. (2020)	Personalized neoantigen vaccination with synthetic long peptides: recent advances and future perspectives
Londhe and Date (2020)	Personalized neoantigen vaccines: A glimmer of hope for glioblastoma
Han et al. (2020)	Progress in neoantigen targeted cancer immunotherapies
Keshavarzi Arshadi and Salem (2020)	AI and Immunoinformatics
Jiang et al. (2019)	Tumor neoantigens: from basic research to clinical applications
Mardis (2019)	Neoantigens and genome instability: impact on immunogenic phenotypes and immunotherapy response
de Miranda and Trajanoski (2019)	Advancing cancer immunotherapy: a vision for the field
Li et al. (2018)	Recent updates in cancer immunotherapy: a comprehensive review and perspective of the 2018 China Cancer Immunotherapy Workshop in Beijing
Sidhom et al. (2018)	Applications of Artificial Intelligence & Machine Learning in Cancer Immunology
Doytchinova and Flower (2018)	In silico prediction of cancer immunogens: current state of the art

Gracias a *Next-generation Sequencing* y *Mass spectrometry*, se ha logrado muchos avances en la Bioinformática, pero a veces es necesario tener información adicional como

las propiedades estructurales de los aminoácidos. Debido a esto, han surgido varias investigaciones y los *reviews* de la Tabla 3.16, que explican como se pueden utilizar este tipo de propiedades para predecir la interacción pMHC. Lamentablemente, solo se ha identificado dos trabajos (Perez et al., 2022; Antunes et al., 2018), porque no se cuenta con muchas muestras de este problema.

TABLA 3.16: Listado de los *reviews*, que se enfocan en estudios que utilizan propiedades estructurales de los aminoácidos para la de detección de neo antígenos.

Autor-año	Título
Perez et al. (2022)	Structural Prediction of Peptide–MHC Binding Modes
Antunes et al. (2018)	Structure-based methods for binding mode and binding affinity prediction for peptide-MHC complexes

Generalmente, con la predicción del enlace pMHC, podría terminar el trabajo Bioinformático, para luego proceder a los trabajos *in vitro* e *in vivo*. Pero, algunos trabajos, también buscan entender que hace que un compuesto pMHC se enlace al TCR y así se genere una respuesta inmune. Esto también ha generado bastantes *reviews* presentados en la Tabla 3.17.

TABLA 3.17: Listado de los *reviews*, que se enfocan en estudios de la interacción de compuestos pMHC con TCR.

Autor-año	Título
Kast et al. (2021)	Advances in identification and selection of personalized neoantigen/T-cell pairs for autologous adoptive T cell therapies
Schaap-Johansen et al. (2021)	T Cell Epitope Prediction and Its Application to Immunotherapy
Zvyagin et al. (2020)	An overview of immunoinformatics approaches and databases linking T cell receptor repertoires to their antigen specificity
Sidney et al. (2020)	Epitope prediction and identification- adaptive T cell responses in humans
Zvyagin et al. (2020)	An overview of immunoinformatics approaches and databases linking T cell receptor repertoires to their antigen specificity
Spear et al. (2019)	Understanding TCR affinity, antigen specificity, and cross-reactivity to improve TCR gene-modified T cells for cancer immunotherapy

Finalmente, se han desarrollado *reviews* que detallan los principales desafíos, buenas prácticas y perspectivas futuras en la detección de neo antígenos (Tabla 3.18). De estos trabajos, el *review* de [Gopanenko et al. \(2020\)](#) y [Borden et al. \(2022\)](#), explican detalladamente, todos los métodos de cada fase para la detección de neo antígenos; adicionalmente, explican las ventajas de cada método y los problemas actuales. También, resaltamos el trabajo de [Richters et al. \(2019\)](#), que resalta las buenas prácticas de este campo de estudio.

TABLA 3.18: Listado de los *reviews*, que se enfocan en presentar buenas prácticas en el proceso de detección de neo antígenos y generación de vacunas personalizadas,

Autor-año	Título
Borden et al. (2022)	Cancer Neoantigens: Challenges and Future Directions for Prediction, Prioritization, and Validation
Chen et al. (2021a)	Challenges targeting cancer neoantigens in 2021: a systematic literature review
Gopanenko et al. (2020)	Main strategies for the identification of neoantigens
Mattos et al. (2020)	Neoantigen prediction and computational perspectives towards clinical benefit: recommendations from the ESMO Precision Medicine Working Group
Richters et al. (2019)	Best practices for bioinformatic characterization of neoantigens for clinical utility
Garcia-Garijo et al. (2019)	Determinants for Neoantigen Identification
Auricchio et al. (2018)	The perfect personalized cancer therapy: cancer vaccines against neoantigens
Barros et al. (2018)	Immunological-based approaches for cancer therapy
Türeci et al. (2018)	Challenges towards the realization of individualized cancer vaccines
Villani et al. (2018)	Systems immunology: Learning the rules of the immune system
Richters et al. (2019)	Best practices for bioinformatic characterization of neoantigens for clinical utility
Garcia-Garijo et al. (2019)	Determinants for Neoantigen Identification
Barros et al. (2018)	Immunological-based approaches for cancer therapy

Capítulo 4

Propuesta

En este capítulo presentaremos la propuesta y como se relaciona con los métodos tradicionales de detección de neo antígenos.

4.1. Detección de neo antígenos (*pipeline*)

Según [Gopanenko et al. \(2020\)](#), la detección de neo antígenos podría clasificarse en tres grupos: (1) basados en genómica, (2) basados en *Mass Spectrometry* (MS) y (3) basados en estructura.

La detección de neo antígenos basada en genómica sigue un proceso muy largo e involucra muchas herramientas, debido a esto se han propuesto bastantes *pipelines*. El proceso general consta de varias etapas presentadas en la Figura 4.1, a continuación detallaremos cada una de ellas y explicaremos en qué fase se ubica la propuesta de esta tesis:

1. **Secuenciamiento.** La primera fase consiste en el secuenciamiento de DNA, en este caso se toman muestras de sangre al tener menos riesgo de no ser contaminadas por un tumor ([Borden et al., 2022](#)). Para la secuenciación, se puede optar por *Whole Genome Sequencing* (WGS) o *Whole Exome Sequencing* (WES), la primera tiene la ventaja de tener mucha más información de mutaciones pero es muy costoso. Esta fase, también puede retroalimentarse con secuenciamiento de RNA (seqRNA). Una tendencia reciente fomenta el uso de *RiboSeq*, este tiene la ventaja de tener más información de las proteínas formadas en los Ribosomas, lamentablemente no se tienen muchas muestras ([Borden et al., 2022](#)).

2. **Alineamiento y procesamiento.** En esta fase, se evalúa la calidad del secuenciamiento, se elimina el ruido y se realiza un alineamiento con un genoma base. Como resultado se obtienen archivos BAM (resultado del alineamiento) y FastQC (calidad de cada secuenciación).
3. **Identificación de neo antígenos.** En esta fase se analiza las mutaciones de la secuencia, generalmente se obtienen *Variant Calling Files* (VCF). En esta etapa, es importante secuenciar las proteínas *Human Leukocyte Antigens* (HLA), estas representan las proteínas MHC mencionadas anteriormente. Luego con información del tipo de HLA y mutaciones, se puede identificar los posibles neo antígenos. Esta fase puede ser retroalimentada de *RiboSeq* y datos de MS.

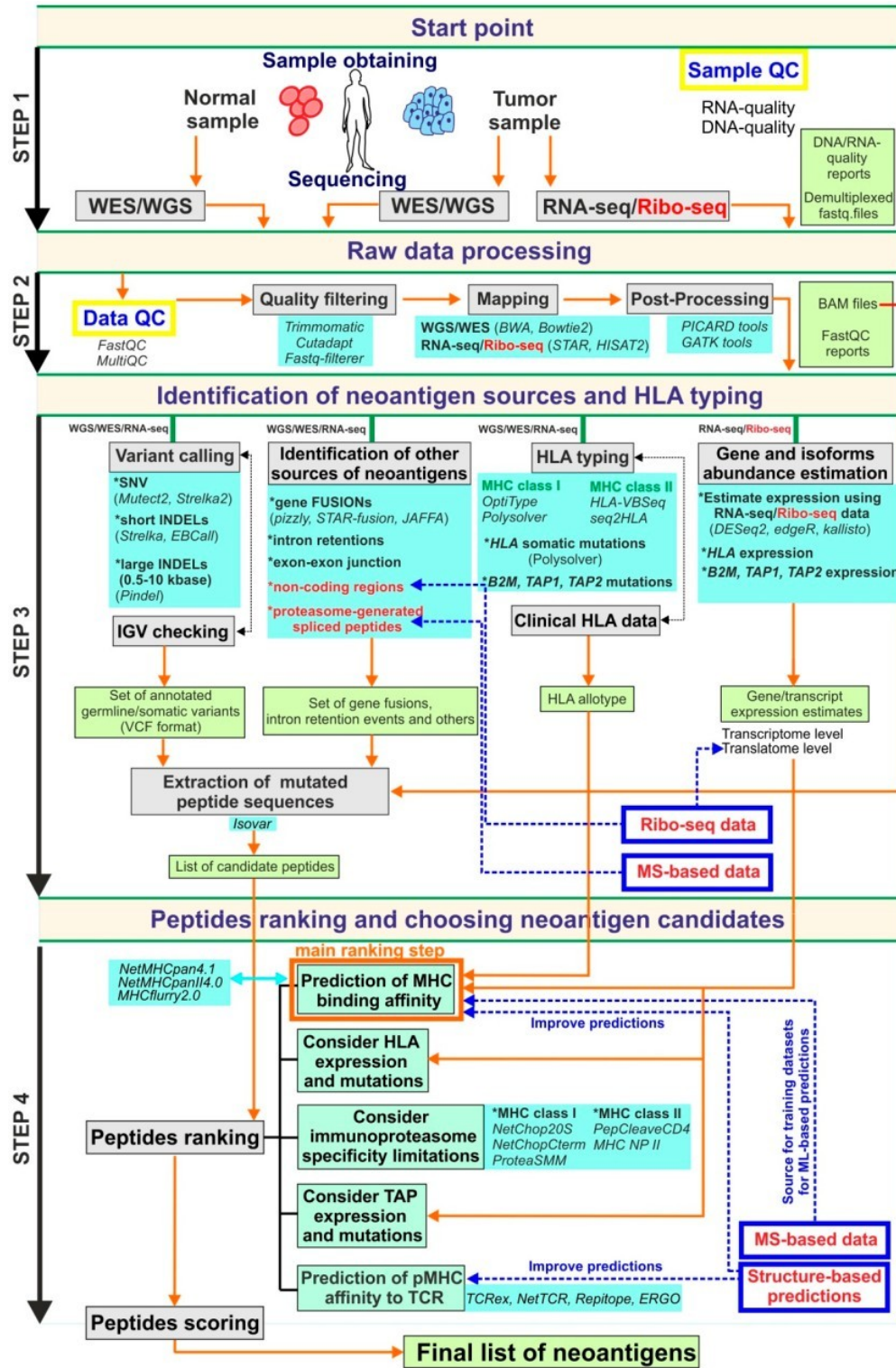


FIGURA 4.1: Proceso general utilizado para la detección de neo antígenos a partir de secuencias de DNA. Fuente: [Gopanenko et al. \(2020\)](#).

4. **Priorización de neo antígenos.** En esta fase se filtran los neo antígenos identificados anteriormente. Este problema es conocido mayormente como: *MHC-peptide binding*, en este caso se predice el enlace entre el neo antígeno y la proteína MHC

(la propuesta de la tesis se enfoca en esta etapa). Las herramientas con mejor desempeño son *NetMHCpan4.1* y *MHCflurry2.0* según varios *benchmarks* (Bonsack et al., 2019; Zhao and Sher, 2018; Paul et al., 2020a; Trolle et al., 2015). Recientemente una nueva propuesta ha superado a *NetMHCpan4.1*, esta propuesta obtuvo buenos resultados utilizando *protein language models* (Hashemi et al., 2022). Finalmente, se predice la afinidad de T-Cell Receptor (TCR) con pMHC (peptide-MHC binding).

Recientemente, se está utilizando otros enfoques para mejorar la detección de neo antígenos, por ejemplo, se puede utilizar datos MS para mejorar la identificación de neo antígenos. Luego, el enfoque basado en estructura que utiliza información de propiedades químicas y físicas de los péptidos puede ser utilizada para mejorar la predicción de afinidad TCR y pMHC (Borden et al., 2022; Gopanenko et al., 2020).

4.2. Predicción de la afinidad péptido-MHC (peptide-MHC binding)

La propuesta se inspira en los trabajos de Cheng et al. (2021) y Hashemi et al. (2022). Ambos proponen el uso de *transfer learning* a partir de los modelos pre-entrenados BERT (Devlin et al., 2018) y ESM-1b (Rives et al., 2021) respectivamente.

El modelo *Bidirectional Encoder Representations from Transformers*. (BERT), fue diseñado para el pre-entrenamiento de representaciones bidireccionales de textos no etiquetados. Este modelo fue diseñado inicialmente para el procesamiento natural del lenguaje, pero en el trabajo de Rao et al. (2019), se planteó su uso para secuencias de aminoácidos. Es así que Rao et al. (2019) entrenan BERT con 31 millones de secuencias de proteínas y llaman a su propuesta *Tasks Assessing Protein Embeddings* (TAPE).

Recientemente, Facebook desarrolla el modelo ESM-1b (Rives et al., 2021). La propuesta se basa en el modelo RoBERTa (Liu et al., 2019a), la cuál es una optimización de BERT. Luego, ESM-1b fue entrenado con la base de datos Uniref50 (Suzek et al., 2015), esta base de datos cuenta con aproximadamente 250 millones de secuencias de proteínas. En este caso, se realizó un entrenamiento no supervisado, se ocultaron las etiquetas referentes a la estructura o función de las proteínas.

Entonces, la propuesta de la tesis se basa en utilizar *transfer learning* del modelo pre-entrenado ESM-1b, luego se va a utilizar otra red neuronal paralela que se alimente de datos físico-químicos de los aminoácidos. Se propone utilizar las propiedades físico-químicas de los aminoácidos, porque en varios ensayos clínicos se ha comprobado que influyen en la predicción *peptide-MHC binding* y *pMHC-TCR presentation* (Gopanenko et al., 2020; Borden et al., 2022). Luego, las dos redes neuronales paralelas se unirán en una red neuronal totalmente conectada (ver Figura 4.2). El objetivo, es aprovechar las propiedades físico-químicas de los aminoácidos para mejorar la afinidad *peptide-MHC*.

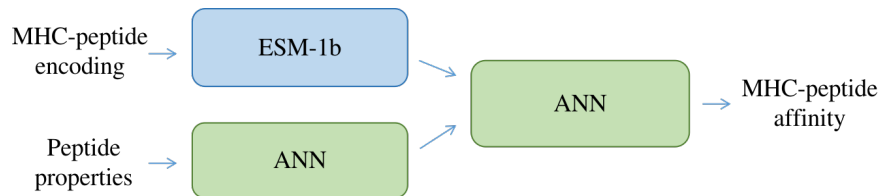


FIGURA 4.2: Propuesta de *transfer learning* de ESM-1b y una red neuronal paralela para la predicción de la afinidad entre un péptido y MHC (peptide MHC binding).

Para los entrenamientos y experimentos se utilizará la base de datos HLA3D (Li et al., 2022), esta contiene información de 1296 aminoácidos. Luego, también utilizaremos las muestras recolectadas de Hashemi et al. (2022).

Capítulo 5

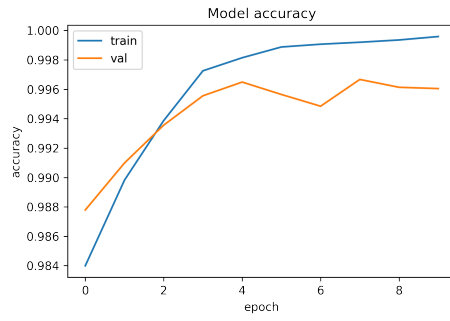
Resultados

En la Tabla 5.1, presentamos el *accuracy*, *f1 score*, *precision* y *recall* de cada base de datos (*allele*). Como podemos ver, en todos los casos superamos el 0.9 de *accuracy*, esto valida la propuesta y da origen a seguir trabajando en mejorar la propuesta.

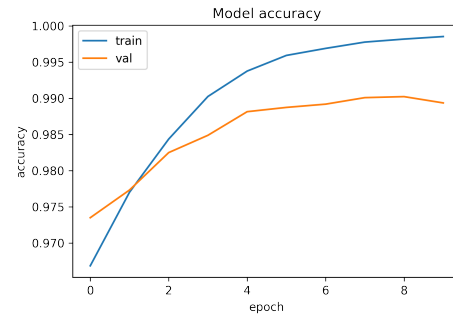
Luego, en la Figura 5.1, presentamos el *accuracy* obtenido durante el entrenamiento de cada base de datos con el conjunto de muestras de entrenamiento y validación. En este caso, utilizamos el 20 % de las muestras de entrenamiento como validación. Como podemos ver, con solo 10 *epochs*, se lograron buenos resultados. Tambien se evaluao con mas *epochs*, pero los resultados no mejoraron.

TABLA 5.1: Resultados obtenidos en cada base de datos.

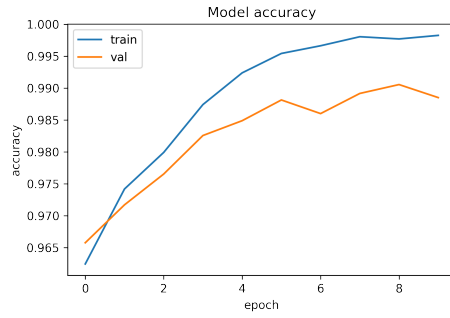
<i>Allele</i>	<i>Accuracy</i>	<i>F1 score</i>	<i>Precision</i>	<i>Recall</i>
A*01:01	0.978	0.917	0.982	0.887
A*0201	0.962	0.956	0.965	0.948
A*02:03	0.992	0.979	0.994	0.969
A*31:01	0.980	0.968	0.989	0.951
B*44:02	0.991	0.981	0.968	0.997
B*44:03	0.992	0.987	0.995	0.980



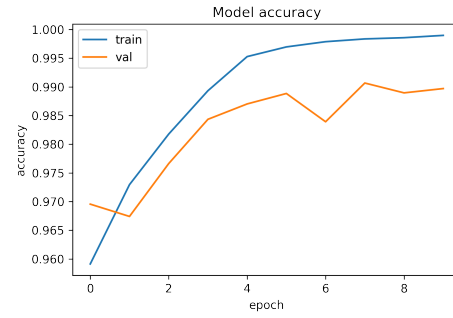
(a) A*01:01



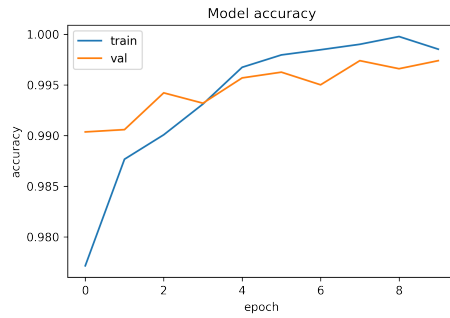
(b) A*02:01



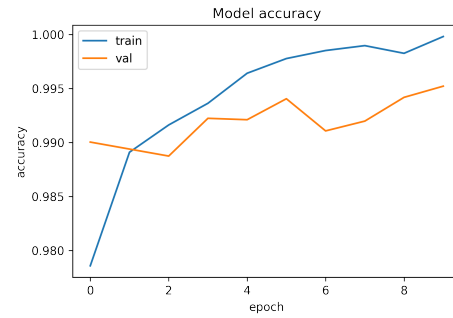
(c) A*02:03



(d) A*31:01



(e) B*44:02



(f) B*44:03

FIGURA 5.1: *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, B*44:02 y B*44:03.

Capítulo 6

Conclusiones

Primera: Se ha realizado una búsqueda sistemática de la literatura sobre los principales métodos basados en *deep learning*, utilizados para la detección de neo antígenos. Estos métodos involucran las *Shallow Neural Networks*, redes neuronales convolucionales, redes neuronales recurrentes y recientemente las redes *Transformers* y BERT.

Segunda: Se ha presentado un nuevo método basado en redes neuronales BERT y con *transfer learning*, de los modelos pre entrenados TAPE y ESMb-1. El método propuesto ha sido evaluado con colección de varias muestras tomadas de bases de datos públicas y trabajos similares.

Bibliografía

- Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G. L., Eisenhaure, T. M., et al. (2017). Mass spectrometry profiling of hla-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity*, 46(2):315–326.
- Abualrous, E. T., Sticht, J., and Freund, C. (2021). Major histocompatibility complex (mhc) class i and class ii proteins: impact of polymorphism on antigen presentation. *Current Opinion in Immunology*, 70:95–104.
- Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L. (2021). Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8(1):1–74.
- Antunes, D. A., Abella, J. R., Devaurs, D., Rigo, M. M., and Kavraki, L. E. (2018). Structure-based methods for binding mode and binding affinity prediction for peptide-mhc complexes. *Current topics in medicinal chemistry*, 18(26):2239–2255.
- Arceda, V. E. M. (2023). Neoantigen detection using transformers and transfer learning in the cancer immunology context. In *International Conference on Practical Applications of Computational Biology & Bioinformatics*, pages 97–102. Springer.
- Arnaud, M., Duchamp, M., Bobisse, S., Renaud, P., Coukos, G., and Harari, A. (2020). Biotechnologies to tackle the challenge of neoantigen identification. *Current opinion in biotechnology*, 65:52–59.
- Aurisicchio, L., Pallocca, M., Ciliberto, G., and Palombo, F. (2018). The perfect personalized cancer therapy: cancer vaccines against neoantigens. *Journal of Experimental & Clinical Cancer Research*, 37(1):1–10.
- Bagaev, D. V., Vroomans, R. M., Samir, J., Stervbo, U., Rius, C., Dolton, G., Greenshields-Watson, A., Attaf, M., Egorov, E. S., Zvyagin, I. V., et al. (2020). Vdjdb in 2019: database extension, new analysis infrastructure and a t-cell receptor motif compendium. *Nucleic Acids Research*, 48(D1):D1057–D1062.

- Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Bais, P., Namburi, S., Gatti, D. M., Zhang, X., and Chuang, J. H. (2017). Cloudneo: a cloud pipeline for identifying patient-specific tumor neoantigens. *Bioinformatics*, 33(19):3110–3112.
- Barros, L., Pretti, M. A., Chicaybam, L., Abdo, L., Boroni, M., and Bonamino, M. H. (2018). Immunological-based approaches for cancer therapy. *Clinics*, 73.
- Bassani-Sternberg, M., Platscher-Frankild, S., Jensen, L. J., and Mann, M. (2015). Mass spectrometry of human leukocyte antigen class i peptidomes reveals strong effects of protein abundance and turnover on antigen presentation*[s]. *Molecular & Cellular Proteomics*, 14(3):658–673.
- Battaglia, S. (2020). Neoantigen prediction from genomic and transcriptomic data. *Methods in Enzymology*, 635:267–281.
- Bjerregaard, A.-M., Nielsen, M., Hadrup, S. R., Szallasi, Z., and Eklund, A. C. (2017). Mupexi: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunology, Immunotherapy*, 66(9):1123–1130.
- Blass, E. and Ott, P. A. (2021). Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. *Nature Reviews Clinical Oncology*, 18(4):215–229.
- Bollineni, R. C., Tran, T. T., Lund-Johansen, F., and Olweus, J. (2022). Chasing neoantigens; invite naïve t cells to the party. *Current Opinion in Immunology*, 75:102172.
- Bonsack, M., Hoppe, S., Winter, J., Tichy, D., Zeller, C., Küpper, M. D., Schitter, E. C., Blatnik, R., and Riemer, A. B. (2019). Performance evaluation of mhc class-i binding prediction tools based on an experimentally validated mhc-peptide binding data set. *Cancer immunology research*, 7(5):719–736.
- Borden, E. S., Buetow, K. H., Wilson, M. A., and Hastings, K. T. (2022). Cancer neoantigens: Challenges and future directions for prediction, prioritization, and validation. *Frontiers in Oncology*, 12.
- Bravi, B., Tubiana, J., Cocco, S., Monasson, R., Mora, T., and Walczak, A. M. (2021). Rbm-mhc: a semi-supervised machine-learning method for sample-specific prediction of antigen presentation by hla-i alleles. *Cell systems*, 12(2):195–202.
- Bulik-Sullivan, B., Busby, J., Palmer, C. D., Davis, M. J., Murphy, T., Clark, A., Busby, M., Duke, F., Yang, A., Young, L., et al. (2019). Deep learning using tumor hla peptide mass spectrometry datasets improves neoantigen identification. *Nature biotechnology*, 37(1):55–63.

- Chen, I., Chen, M., Goedegebuure, P., and Gillanders, W. (2021a). Challenges targeting cancer neoantigens in 2021: a systematic literature review. *Expert Review of Vaccines*, 20(7):827–837.
- Chen, R., Fulton, K. M., Twine, S. M., and Li, J. (2021b). Identification of mhc peptides using mass spectrometry for neoantigen discovery and cancer vaccine development. *Mass spectrometry reviews*, 40(2):110–125.
- Chen, X., Yang, J., Wang, L., and Liu, B. (2020). Personalized neoantigen vaccination with synthetic long peptides: recent advances and future perspectives. *Theranostics*, 10(13):6011.
- Cheng, J., Bendjama, K., Rittner, K., and Malone, B. (2021). Bertmhc: improved mhc–peptide class ii interaction prediction with transformer and multiple instance learning. *Bioinformatics*, 37(22):4172–4179.
- Chu, Y., Zhang, Y., Wang, Q., Zhang, L., Wang, X., Wang, Y., Salahub, D. R., Xu, Q., Wang, J., Jiang, X., et al. (2022). A transformer-based model to predict peptide–hla class i binding and optimize mutated peptides for vaccine design. *Nature Machine Intelligence*, 4(3):300–311.
- Clancy, S. (2008). Genetic mutation. *Nature Education*, 1(1):187.
- Coelho, A. C. M., Fonseca, A. L., Martins, D. L., Lins, P. B., da Cunha, L. M., and de Souza, S. J. (2020). neoant-hill: an integrated tool for identification of potential neoantigens. *BMC Medical Genomics*, 13(1):1–8.
- Creech, A. L., Ting, Y. S., Goulding, S. P., Sauld, J. F., Barthelme, D., Rooney, M. S., Addona, T. A., and Abelin, J. G. (2018). The role of mass spectrometry and proteogenomics in the advancement of hla epitope prediction. *Proteomics*, 18(12):1700259.
- Dalianis, H. (2018). Evaluation metrics and evaluation. In *Clinical text mining*, pages 45–53. Springer.
- de Miranda, N. F. and Trajanoski, Z. (2019). Advancing cancer immunotherapy: a vision for the field.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Doytchinova, I. A. and Flower, D. R. (2018). In silico prediction of cancer immunogens: current state of the art. *BMC immunology*, 19(1):1–9.

- El Naqa, I. and Murphy, M. J. (2022). Machine and deep learning in oncology, medical physics and radiology.
- Esprit, A., de Mey, W., Bahadur Shahi, R., Thielemans, K., Franceschini, L., and Breckpot, K. (2020). Neo-antigen mrna vaccines. *Vaccines*, 8(4):776.
- Fang, L. T., Afshar, P. T., Chhibber, A., Mohiyuddin, M., Fan, Y., Mu, J. C., Gibeling, G., Barr, S., Asadi, N. B., Gerstein, M. B., et al. (2015). An ensemble approach to accurately detect somatic mutations using somaticseq. *Genome biology*, 16(1):1–13.
- Fang, X., Guo, Z., Liang, J., Wen, J., Liu, Y., Guan, X., and Li, H. (2022). Neoantigens and their potential applications in tumor immunotherapy. *Oncology Letters*, 23(3):1–9.
- Ferreira, J. A., Relvas-Santos, M., Peixoto, A., Silva, A. M., and Santos, L. L. (2021). Glycoproteogenomics: setting the course for next-generation cancer neoantigen discovery for cancer vaccines. *Genomics, Proteomics & Bioinformatics*, 19(1):25–43.
- Fotakis, G., Trajanoski, Z., and Rieder, D. (2021). Computational cancer neoantigen prediction: current status and recent advances. *Immuno-Oncology and Technology*, 12:100052.
- Fritsch, E. F., Burkhardt, U. E., Hacohen, N., and Wu, C. J. (2020). Personal neoantigen cancer vaccines: a road not fully paved. *Cancer immunology research*, 8(12):1465–1469.
- Garcia-Garijo, A., Fajardo, C. A., and Gros, A. (2019). Determinants for neoantigen identification. *Frontiers in immunology*, 10:1392.
- Gasser, H.-C., Bedran, G., Ren, B., Goodlett, D., Alfaro, J., and Rajan, A. (2021). Interpreting bert architecture predictions for peptide presentation by mhc class i proteins. *arXiv preprint arXiv:2111.07137*.
- Gfeller, D., Schmidt, J., Croce, G., Guillaume, P., Bobisse, S., Genolet, R., Queiroz, L., Cesbron, J., Racle, J., and Harari, A. (2023). Improved predictions of antigen presentation and tcr recognition with mixmhcpred2. 2 and prime2. 0 reveal potent sars-cov-2 cd8+ t-cell epitopes. *Cell Systems*, 14(1):72–83.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>.
- Gopanenko, A. V., Kosobokova, E. N., and Kosorukov, V. S. (2020). Main strategies for the identification of neoantigens. *Cancers*, 12(10):2879.
- Groisberg, R., Maymani, H., and Subbiah, V. (2018). Immunotherapy and next-generation sequencing guided therapy for precision oncology: what have we learnt and what does the future hold? *Expert review of precision medicine and drug development*, 3(3):205–213.

- Gros, A., Parkhurst, M. R., Tran, E., Pasetto, A., Robbins, P. F., Ilyas, S., Prickett, T. D., Gartner, J. J., Crystal, J. S., Roberts, I. M., et al. (2016). Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. *Nature medicine*, 22(4):433–438.
- Han, X.-J., Ma, X.-l., Yang, L., Wei, Y.-q., Peng, Y., and Wei, X.-w. (2020). Progress in neoantigen targeted cancer immunotherapies. *Frontiers in Cell and Developmental Biology*, 8:728.
- Han, Y. (2018). Deep convolutional neural networks for peptide-mhc binding predictions.
- Hao, Q., Wei, P., Shu, Y., Zhang, Y.-G., Xu, H., and Zhao, J.-N. (2021). Improvement of neoantigen identification through convolution neural network. *Frontiers in immunology*, 12.
- Hashemi, N., Hao, B., Ignatov, M., Paschalidis, I., Vakili, P., Vajda, S., and Kozakov, D. (2022). Improved predictions of mhc-peptide binding using protein language models. *bioRxiv*.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Heng, Y., Kuang, Z., Xie, W., Lan, H., Huang, S., Chen, L., Shi, T., Xu, L., Pan, X., and Mei, H. (2021). A simple pan-specific rnn model for predicting hla-ii binding peptides. *Molecular Immunology*, 139:177–183.
- Heyer, E. E. and Blackburn, J. (2020). Sequencing strategies for fusion gene detection. *BioEssays*, 42(7):2000016.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Holtsträter, C., Schrörs, B., Bukur, T., and Löwer, M. (2020). Bioinformatics for cancer immunotherapy. *Bioinformatics for Cancer Immunotherapy*, pages 1–9.
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558.
- Hu, Y., Wang, Z., Hu, H., Wan, F., Chen, L., Xiong, Y., Wang, X., Zhao, D., Huang, W., and Zeng, J. (2019). Acme: pan-specific peptide-mhc class i binding prediction through attention-based deep neural networks. *Bioinformatics*, 35(23):4946–4954.

- Hundal, J., Kiwala, S., McMichael, J., Miller, C. A., Xia, H., Wollam, A. T., Liu, C. J., Zhao, S., Feng, Y.-Y., Graubert, A. P., et al. (2020). pvactools: a computational toolkit to identify and visualize cancer neoantigens. *Cancer immunology research*, 8(3):409–420.
- Janeway Jr, C. A. (1997). Immunobiology the immune system in health and disease. *Artes Medicas*.
- Jensen, K. K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J. A., Yan, Z., Sette, A., Peters, B., and Nielsen, M. (2018). Improved methods for predicting peptide binding affinity to mhc class ii molecules. *Immunology*, 154(3):394–406.
- Jiang, L., Yu, H., Li, J., Tang, J., Guo, Y., and Guo, F. (2021). Predicting mhc class i binder: existing approaches and a novel recurrent neural network solution. *Briefings in Bioinformatics*, 22(6):bbab216.
- Jiang, T., Shi, T., Zhang, H., Hu, J., Song, Y., Wei, J., Ren, S., and Zhou, C. (2019). Tumor neoantigens: from basic research to clinical applications. *Journal of hematology & oncology*, 12(1):1–13.
- Jin, J., Liu, Z., Nasiri, A., Cui, Y., Louis, S.-Y., Zhang, A., Zhao, Y., and Hu, J. (2021). Deep learning pan-specific model for interpretable mhc-i peptide binding prediction with improved attention mechanism. *Proteins: Structure, Function, and Bioinformatics*, 89(7):866–883.
- Jordan, M. I. (1997). Serial order: A parallel distributed processing approach. In *Advances in psychology*, volume 121, pages 471–495. Elsevier.
- Junet, V. and Daura, X. (2021). 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*, 37(23):4567–4568.
- Kast, F., Klein, C., Umaña, P., Gros, A., and Gasser, S. (2021). Advances in identification and selection of personalized neoantigen/t-cell pairs for autologous adoptive t cell therapies. *OncoImmunology*, 10(1):1869389.
- Kelvin, J. (2022). Rnns, lstms, cnns, transformers and bert.
- Kerbs, P., Vosberg, S., Krebs, S., Graf, A., Blum, H., Swoboda, A., Batcha, A. M., Mansmann, U., Metzler, D., Heckman, C. A., et al. (2022). Fusion gene detection by rna-sequencing complements diagnostics of acute myeloid leukemia and identifies recurring nrip1-mir99ahg rearrangements. *haematologica*, 107(1):100.
- Keshavarzi Arshadi, A. and Salem, M. (2020). Ai and immunoinformatics. *Artificial Intelligence in Medicine*, pages 1–9.

- Kim, P. and Zhou, X. (2019). Fusiondb: fusion gene annotation database. *Nucleic acids research*, 47(D1):D994–D1004.
- Kim, S., Kim, H. S., Kim, E., Lee, M., Shin, E.-C., and Paik, S. (2018). Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Annals of Oncology*, 29(4):1030–1036.
- Kim, Y., Denton, C., Hoang, L., and Rush, A. M. (2017). Structured attention networks. *arXiv preprint arXiv:1702.00887*.
- Kim, Y., Sidney, J., Pinilla, C., Sette, A., and Peters, B. (2009). Derivation of an amino acid similarity matrix for peptide: Mhc binding and its application as a bayesian prior. *BMC bioinformatics*, 10:1–11.
- Kote, S., Pirog, A., Bedran, G., Alfaro, J., and Dapic, I. (2020). Mass spectrometry-based identification of mhc-associated peptides. *Cancers*, 12(3):535.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- Lang, F., Riesgo-Ferreiro, P., Löwer, M., Sahin, U., and Schrörs, B. (2021). Neofox: annotating neoantigen candidates with neoantigen features. *Bioinformatics*, 37(22):4246–4247.
- Lang, F., Schrörs, B., Löwer, M., Türeci, Ö., and Sahin, U. (2022). Identification of neoantigens for individualized therapeutic cancer vaccines. *Nature Reviews Drug Discovery*, 21(4):261–282.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- Lee, K.-H., Chang, Y.-C., Chen, T.-F., Juan, H.-F., Tsai, H.-K., and Chen, C.-Y. (2021). Connecting mhc-i-binding motifs with hla alleles via deep learning. *Communications Biology*, 4(1):1–12.
- Li, G., Iyer, B., Prasath, V. S., Ni, Y., and Salomonis, N. (2021). Deepimmuno: deep learning-empowered prediction and generation of immunogenic peptides for t-cell immunity. *Briefings in bioinformatics*, 22(6):bbab160.
- Li, X., Lin, X., Mei, X., Chen, P., Liu, A., Liang, W., Chang, S., and Li, J. (2022). Hla3d: an integrated structure-based computational toolkit for immunotherapy. *Briefings in bioinformatics*, 23(3):bbac076.
- Li, Y., Wang, G., Tan, X., Ouyang, J., Zhang, M., Song, X., Liu, Q., Leng, Q., Chen, L., and Xie, L. (2020). Progeo-neo: a customized proteogenomic workflow for neoantigen prediction and selection. *BMC medical genomics*, 13(5):1–11.

- Li, Z., Song, W., Rubinstein, M., and Liu, D. (2018). Recent updates in cancer immunotherapy: a comprehensive review and perspective of the 2018 china cancer immunotherapy workshop in beijing. *Journal of Hematology & Oncology*, 11(1):1–15.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019a). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Liu, Y., Ouyang, X.-h., Xiao, Z.-X., Zhang, L., and Cao, Y. (2020). A review on the methods of peptide-mhc binding prediction. *Current Bioinformatics*, 15(8):878–888.
- Liu, Z., Cui, Y., Xiong, Z., Nasiri, A., Zhang, A., and Hu, J. (2019b). Deepseqpan, a novel deep convolutional neural network model for pan-specific class i hla-peptide binding affinity prediction. *Scientific reports*, 9(1):1–10.
- Liu, Z., Jin, J., Cui, Y., Xiong, Z., Nasiri, A., Zhao, Y., and Hu, J. (2021). 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*.
- Londhe, V. Y. and Date, V. (2020). Personalized neoantigen vaccines: A glimmer of hope for glioblastoma. *Expert Review of Vaccines*, 19(5):407–417.
- Lu, M., Xu, L., Jian, X., Tan, X., Zhao, J., Liu, Z., Zhang, Y., Liu, C., Chen, L., Lin, Y., et al. (2022). dbpepneo2. 0: A database for human tumor neoantigen peptides from mass spectrometry and tcr recognition. *Frontiers in immunology*, page 1583.
- Lucito, R., Suresh, S., Walter, K., Pandey, A., Lakshmi, B., Krasnitz, A., Sebat, J., Wiggler, M., Klein, A. P., Brune, K., et al. (2007). Copy-number variants in patients with a strong family history of pancreatic cancer. *Cancer biology & therapy*, 6(10):1592–1599.
- Luscombe, N. M., Greenbaum, D., and Gerstein, M. (2001). What is bioinformatics? a proposed definition and overview of the field. *Methods of information in medicine*, 40(04):346–358.
- Machaca, V. E., Goyzueta, V., Cruz, M., and Tupac, Y. (2023). Deep learning and transformers in mhc-peptide binding and presentation towards personalized vaccines in cancer immunology: A brief review. In *International Conference on Practical Applications of Computational Biology & Bioinformatics*, pages 14–23. Springer.
- Marais, W. J., Holz, R. E., Reid, J. S., and Willett, R. M. (2020). Leveraging spatial textures, through machine learning, to identify aerosols and distinct cloud types from multispectral observations. *Atmospheric Measurement Techniques*, 13(10):5459–5480.

- Mardis, E. R. (2019). Neoantigens and genome instability: Impact on immunogenomic phenotypes and immunotherapy response. *Genome medicine*, 11(1):1–12.
- Marshall, J. S., Warrington, R., Watson, W., and Kim, H. L. (2018). An introduction to immunology and immunopathology. *Allergy, Asthma & Clinical Immunology*, 14(2):1–10.
- Mateo, N., Canon, Á. D. O., and Charry, O. J. P. (2020). Comparison of machine learning models for the prediction of cancer cells using mhc class i complexes. In *16th International Symposium on Medical Information Processing and Analysis*, volume 11583, pages 180–187. SPIE.
- Mattos, L., Vazquez, M., Finotello, F., Lepore, R., Porta, E., Hundal, J., Amengual-Rigo, P., Ng, C., Valencia, A., Carrillo, J., et al. (2020). Neoantigen prediction and computational perspectives towards clinical benefit: recommendations from the esmo precision medicine working group. *Annals of oncology*, 31(8):978–990.
- McCaffrey, P. (2022). Bioinformatic techniques for vaccine development: Epitope prediction and structural vaccinology. In *Vaccine Design*, pages 413–423. Springer.
- Mei, S., Li, F., Leier, A., Marquez-Lago, T. T., Giam, K., Croft, N. P., Akutsu, T., Smith, A. I., Li, J., Rossjohn, J., et al. (2020). A comprehensive review and performance evaluation of bioinformatics tools for hla class i peptide-binding prediction. *Briefings in bioinformatics*, 21(4):1119–1135.
- Mei, S., Li, F., Xiang, D., Ayala, R., Faridi, P., Webb, G. I., Illing, P. T., Rossjohn, J., Akutsu, T., Croft, N. P., et al. (2021). Anthem: a user customised tool for fast and accurate prediction of binding between peptides and hla class i molecules. *Briefings in Bioinformatics*, 22(5):bbaa415.
- Mill, N. A., Bogaert, C., van Crielinge, W., and Fant, B. (2022). neoms: Attention-based prediction of mhc-i epitope presentation. *bioRxiv*.
- Mirandola, L., Marincola, F., Rotino, G., Figueroa, J. A., Grizzi, F., Bresalier, R., and Chiriva-Internati, M. (2020). The quest for the next-generation of tumor targets: Discovery and prioritization in the genomics era. In *Immuno-Oncology*, pages 239–253. Springer.
- Mitchell, T. M. (1997). *Machine learning*, volume 1. McGraw-hill New York.
- NCI (2020). Nci dictionary of cancer terms. <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/transcription>. Accessed: 2020-03-20.
- NCI (2022). National cancer institute dictionary.

- Neefjes, J., Jongsma, M. L., Paul, P., and Bakke, O. (2011). Towards a systems understanding of mhc class i and mhc class ii antigen presentation. *Nature reviews immunology*, 11(12):823–836.
- Ng, F. S., Vandenberghe, M., Portella, G., Cayatte, C., Qu, X., Hanabuchi, S., Landry, A., Chaerkady, R., Yu, W., Collepardo-Guevara, R., et al. (2020). Minerva: Learning the rules of hla class i peptide presentation in tumors with convolutional neural networks and transfer learning. *Available at SSRN 3704016*.
- Nielsen, M. and Andreatta, M. (2016). Netmhcpa-3.0; improved prediction of binding to mhc class i molecules integrating information from multiple receptor and peptide length datasets. *Genome medicine*, 8(1):1–9.
- Nielsen, M., Andreatta, M., Peters, B., and Buus, S. (2020). Immunoinformatics: predicting peptide–mhc binding. *Annual Review*.
- Nielsen, M. A. (2015). *Neural networks and deep learning*, volume 25. Determination press San Francisco, CA, USA.
- O'Donnell, T. J., Rubinsteyn, A., Bonsack, M., Riemer, A. B., Laserson, U., and Hammerbacher, J. (2018). Mhcflurry: open-source class i mhc binding affinity prediction. *Cell systems*, 7(1):129–132.
- O'Donnell, T. J., Rubinsteyn, A., and Laserson, U. (2020). Mhcflurry 2.0: improved pan-allele prediction of mhc class i-presented peptides by incorporating antigen processing. *Cell systems*, 11(1):42–48.
- Okada, M., Shimizu, K., and Fujii, S.-i. (2022). Identification of neoantigens in cancer cells as targets for immunotherapy. *International Journal of Molecular Sciences*, 23(5):2594.
- Oliveira, D. M. T., de Serpa Brandão, R. M. S., da Mata Sousa, L. C. D., Lima, F. d. C. A., do Monte, S. J. H., Marroquim, M. S. C., de Sousa Lima, A. V., Coelho, A. G. B., Costa, J. M. S., Ramos, R. M., et al. (2019). phla3d: An online database of predicted three-dimensional structures of hla molecules. *Human Immunology*, 80(10):834–841.
- PacBio (2021). Two review articles assess structural variation in human genomes. <https://www.pacb.com/blog/two-review-articles-assess-structural-variation-in-human-genomes/>. Accessed: 2021-05-07.
- Pan, X., Hu, X., Zhang, Y.-H., Chen, L., Zhu, L., Wan, S., Huang, T., and Cai, Y.-D. (2019). Identification of the copy number variant biomarkers for breast cancer subtypes. *Molecular Genetics and Genomics*, 294(1):95–110.

- Pao, S.-C., Chu, M.-T., and Hung, S.-I. (2022). Therapeutic vaccines targeting neoantigens to induce t-cell immunity against cancers. *Pharmaceutics*, 14(4):867.
- Parikh, A. P., Täckström, O., Das, D., and Uszkoreit, J. (2016). A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.
- Patwardhan, N., Marrone, S., and Sansone, C. (2023). Transformers in the real world: A survey on nlp applications. *Information*, 14(4):242.
- Paul, S., Croft, N. P., Purcell, A. W., Tschärke, D. C., Sette, A., Nielsen, M., and Peters, B. (2020a). Benchmarking predictions of mhc class i restricted t cell epitopes in a comprehensively studied model system. *PLoS computational biology*, 16(5):e1007757.
- Paul, S., Grifoni, A., Peters, B., and Sette, A. (2020b). Major histocompatibility complex binding, eluted ligands, and immunogenicity: benchmark testing and predictions. *Frontiers in immunology*, 10:3151.
- Pearlman, A. H., Hwang, M. S., König, M. F., Hsiue, E. H.-C., Douglass, J., DiNapoli, S. R., Mog, B. J., Bettegowda, C., Pardoll, D. M., Gabelli, S. B., et al. (2021). Targeting public neoantigens for cancer immunotherapy. *Nature cancer*, 2(5):487–497.
- Pei, B. and Hsu, Y.-H. (2020). Iconmhc: a deep learning convolutional neural network model to predict peptide and mhc-i binding affinity. *Immunogenetics*, 72(5):295–304.
- Peng, M., Mo, Y., Wang, Y., Wu, P., Zhang, Y., Xiong, F., Guo, C., Wu, X., Li, Y., Li, X., et al. (2019). Neoantigen vaccine: an emerging tumor immunotherapy. *Molecular cancer*, 18(1):1–14.
- Perez, M. A., Cuendet, M. A., Röhrig, U. F., Michielin, O., and Zoete, V. (2022). Structural prediction of peptide–mhc binding modes. In *Computational Peptide Science*, pages 245–282. Springer.
- Phloyphisut, P., Pornputtapong, N., Sriswasdi, S., and Chuangsuwanich, E. (2019). Mhc-seqnet: a deep neural network model for universal mhc binding prediction. *BMC bioinformatics*, 20(1):1–10.
- Radwan, J., Babik, W., Kaufman, J., Lenz, T. L., and Winternitz, J. (2020). Advances in the evolutionary understanding of mhc polymorphism. *Trends in Genetics*, 36(4):298–311.
- Raff, E. (2022). *Inside Deep Learning*. Manning Publications Co.
- Rammensee, H.-G., Bachmann, J., Emmerich, N. P. N., Bachor, O. A., and Stevanović, S. (1999). Syfpeithi: database for mhc ligands and peptide motifs. *Immunogenetics*, 50:213–219.

- Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. (2019). Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32.
- Reche, P. A., Glutting, J.-P., and Reinherz, E. L. (2002). Prediction of mhc class i binding peptides using profile motifs. *Human immunology*, 63(9):701–709.
- Redwood, A. J., Dick, I. M., Creaney, J., and Robinson, B. W. (2022). What’s next in cancer immunotherapy?-the promise and challenges of neoantigen vaccination. *Onco-immunology*, 11(1):2038403.
- Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020). Netmhcpa-4.1 and netmhciipa-4.0: improved predictions of mhc antigen presentation by concurrent motif deconvolution and integration of ms mhc eluted ligand data. *Nucleic acids research*, 48(W1):W449–W454.
- Reynolds, C. R., Tran, S., Jain, M., and Narendran, A. (2022). Neoantigen cancer vaccines: Generation, optimization, and therapeutic targeting strategies. *Vaccines*, 10(2):196.
- Richard, G., Princiotta, M. F., Bridon, D., Martin, W. D., Steinberg, G. D., and De Groot, A. S. (2022). Neoantigen-based personalized cancer vaccines: the emergence of precision cancer immunotherapy. *Expert Review of Vaccines*, 21(2):173–184.
- Richters, M. M., Xia, H., Campbell, K. M., Gillanders, W. E., Griffith, O. L., and Griffith, M. (2019). Best practices for bioinformatic characterization of neoantigens for clinical utility. *Genome medicine*, 11(1):1–21.
- Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15).
- Robinson, J., Barker, D. J., Georgiou, X., Cooper, M. A., Flicek, P., and Marsh, S. G. (2020). Ipd-imgt/hla database. *Nucleic acids research*, 48(D1):D948–D955.
- Roesler, A. S. and Anderson, K. S. (2022). Beyond sequencing: Prioritizing and delivering neoantigens for cancer vaccines. *Vaccine Design*, pages 649–670.
- Roudko, V., Greenbaum, B., and Bhardwaj, N. (2020). Computational prediction and validation of tumor-associated neoantigens. *Frontiers in Immunology*, 11:27.
- Rubinsteyn, A., Kodysh, J., Hodes, I., Mondet, S., Aksoy, B. A., Finnigan, J. P., Bhardwaj, N., and Hammerbacher, J. (2018). Computational pipeline for the pgv-001 neoantigen vaccine trial. *Frontiers in immunology*, 8:1807.

- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.
- Samuel, A. L. (1967). Some studies in machine learning using the game of checkers. ii—recent progress. *IBM Journal of research and development*, 11(6):601–617.
- Saxena, S., Animesh, S., Fullwood, M. J., and Mu, Y. (2020). Onionmhc: A deep learning model for peptide—hla-a* 02: 01 binding predictions using both structure and sequence feature sets. *Journal of Micromechanics and Molecular Physics*, 5(03):2050009.
- Schaap-Johansen, A.-L., Vujović, M., Borch, A., Hadrup, S. R., and Marcatili, P. (2021). T cell epitope prediction and its application to immunotherapy. *Frontiers in Immunology*, 12.
- Schenck, R. O., Lakatos, E., Gatenbee, C., Graham, T. A., and @miscNCIdictionary2022, author = NCI, title = National Cancer Institute Dictionary, year = 2022, url = <https://www.cancer.gov/publications/dictionaries/genetics-dictionary>, urldate = 2022-03-20 Anderson, A. R. (2019). Neopredpipe: high-throughput neoantigen prediction and recognition potential pipeline. *BMC bioinformatics*, 20(1):1–6.
- Schmidt, M. and Lill, J. R. (2019). Mhc class i presented antigens from malignancies: A perspective on analytical characterization & immunogenicity. *Journal of proteomics*, 191:48–57.
- Shao, X. M., Bhattacharya, R., Huang, J., Sivakumar, I., Tokheim, C., Zheng, L., Hirsch, D., Kaminow, B., Omdahl, A., Bonsack, M., et al. (2020). High-throughput prediction of mhc class i and ii neoantigens with mhcnugetshigh-throughput prediction of neoantigens with mhcnugets. *Cancer immunology research*, 8(3):396–408.
- Shuchen, D. (2022). Understanding deep self-attention mechanism in convolution neural networks.
- Shugay, M., Bagaev, D. V., Zvyagin, I. V., Vroomans, R. M., Crawford, J. C., Dolton, G., Komech, E. A., Sycheva, A. L., Koneva, A. E., Egorov, E. S., et al. (2018). Vdjdb: a curated database of t-cell receptor sequences with known antigen specificity. *Nucleic acids research*, 46(D1):D419–D427.
- Sidhom, J.-W. et al. (2018). *Applications of Artificial Intelligence & Machine Learning in Cancer Immunology*. PhD thesis, Johns Hopkins University.
- Sidney, J., Peters, B., and Sette, A. (2020). Epitope prediction and identification-adaptive t cell responses in humans. In *Seminars in immunology*, volume 50, page 101418. Elsevier.

- Siegel, R. L., Miller, K. D., Wagle, N. S., and Jemal, A. (2023). Cancer statistics, 2023. *Ca Cancer J Clin*, 73(1):17–48.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Socratic.org (2022). How does a deletion mutation differ from a substitution mutation?
- Spear, T. T., Evavold, B. D., Baker, B. M., and Nishimura, M. I. (2019). Understanding tcr affinity, antigen specificity, and cross-reactivity to improve tcr gene-modified t cells for cancer immunotherapy. *Cancer Immunology, Immunotherapy*, 68(11):1881–1889.
- Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and Consortium, U. (2015). Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.
- Tan, X., Li, D., Huang, P., Jian, X., Wan, H., Wang, G., Li, Y., Ouyang, J., Lin, Y., and Xie, L. (2020). dbpepneo: a manually curated database for human tumor neoantigen peptides. *Database*, 2020.
- Terai, Y. L., Huang, C., Wang, B., Kang, X., Han, J., Douglass, J., Hsiue, E. H.-C., Zhang, M., Purohit, R., deSilva, T., et al. (2022). Valid-neo: A multi-omics platform for neoantigen detection and quantification from limited clinical samples. *Cancers*, 14(5):1243.
- Tran, E., Ahmadzadeh, M., Lu, Y.-C., Gros, A., Turcotte, S., Robbins, P. F., Gartner, J. J., Zheng, Z., Li, Y. F., Ray, S., et al. (2015). Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science*, 350(6266):1387–1390.
- Tran, N. H., Xu, J., and Li, M. (2022). A tale of solving two computational challenges in protein science: neoantigen prediction and protein structure prediction. *Briefings in bioinformatics*, 23(1):bbab493.
- Trolle, T., Metushi, I. G., Greenbaum, J. A., Kim, Y., Sidney, J., Lund, O., Sette, A., Peters, B., and Nielsen, M. (2015). Automated benchmarking of peptide-mhc class i binding predictions. *Bioinformatics*, 31(13):2174–2181.
- Türeci, Ö., Löwer, M., Schrörs, B., Lang, M., Tadmor, A., and Sahin, U. (2018). Challenges towards the realization of individualized cancer vaccines. *Nature Biomedical Engineering*, 2(8):566–569.

- UK, C. R. (2023a). Worldwide cancer incidence statistics.
- UK, C. R. (2023b). Worldwide cancer statistics.
- Vang, Y. S. and Xie, X. (2017). Hla class i binding prediction via convolutional neural networks. *Bioinformatics*, 33(17):2658–2665.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Venkatesh, G., Grover, A., Srinivasaraghavan, G., and Rao, S. (2020). Mhcatttnnet: predicting mhc-peptide bindings for mhc alleles classes i and ii using an attention-based deep neural model. *Bioinformatics*, 36(Supplement_1):i399–i406.
- Villani, A.-C., Sarkizova, S., and Hacohen, N. (2018). Systems immunology: learning the rules of the immune system. *Annual review of immunology*, 36:813.
- Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., and Peters, B. (2018). The immune epitope database (iedb): 2018 update. *Nucleic acids research*, 47(D1):D339–D343.
- Wang, F., Wang, H., Wang, L., Lu, H., Qiu, S., Zang, T., Zhang, X., and Hu, Y. (2022). Mhcroberta: pan-specific peptide–mhc class i binding prediction through transfer learning with label-agnostic protein sequences. *Briefings in Bioinformatics*, 23(3):bbab595.
- Wang, P., Chen, Y., and Wang, C. (2021a). Beyond tumor mutation burden: tumor neoantigen burden as a biomarker for immunotherapy and other types of therapy. *Frontiers in Oncology*, 11:672677.
- Wang, T.-Y., Wang, L., Alam, S. K., Hoepfner, L. H., and Yang, R. (2019). Scanneo: identifying indel-derived neoantigens using rna-seq data. *Bioinformatics*, 35(20):4159–4161.
- Wang, Y., Shi, T., Song, X., Liu, B., and Wei, J. (2021b). Gene fusion neoantigens: Emerging targets for cancer immunotherapy. *Cancer Letters*, 506:45–54.
- Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63.
- Wert-Carvajal, C., Sánchez-García, R., Macías, J. R., Sanz-Pamplona, R., Pérez, A. M., Alemany, R., Veiga, E., Sorzano, C. Ó. S., and Muñoz-Barrutia, A. (2021). Predicting mhc i restricted t cell epitopes in mice with nap-cnb, a novel online tool. *Scientific reports*, 11(1):1–10.

- Wieczorek, M., Abualrous, E. T., Sticht, J., Álvaro-Benito, M., Stolzenberg, S., Noé, F., and Freund, C. (2017). Major histocompatibility complex (mhc) class i and mhc class ii proteins: conformational plasticity in antigen presentation. *Frontiers in immunology*, 8:292.
- Wood, M. A., Nguyen, A., Struck, A. J., Ellrott, K., Nellore, A., and Thompson, R. F. (2020). Neoepiscopes improves neoepitope prediction with multivariant phasing. *Bioinformatics*, 36(3):713–720.
- Wu, J., Wang, W., Zhang, J., Zhou, B., Zhao, W., Su, Z., Gu, X., Wu, J., Zhou, Z., and Chen, S. (2019). Deephlapan: a deep learning approach for neoantigen prediction considering both hla-peptide binding and immunogenicity. *Frontiers in Immunology*, page 2559.
- Wu, J., Zhao, W., Zhou, B., Su, Z., Gu, X., Zhou, Z., and Chen, S. (2018). Tsnadb: a database for tumor-specific neoantigens from immunogenomics data analysis. *Genomics, proteomics & bioinformatics*, 16(4):276–282.
- Xie, X., Han, Y., and Zhang, K. (2020). Mhcherrypan: a novel pan-specific model for binding affinity prediction of class i hla-peptide. *International Journal of Data Mining and Bioinformatics*, 24(3):201–219.
- Xiong, J. (2006). *Essential bioinformatics*. Cambridge University Press.
- Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and structural biotechnology journal*, 16:15–24.
- Yadav, M., Jhunjhunwala, S., Phung, Q. T., Lupardus, P., Tanguay, J., Bumbaca, S., Franci, C., Cheung, T. K., Fritsche, J., Weinschenk, T., et al. (2014). Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature*, 515(7528):572–576.
- Yang, X., Zhao, L., Wei, F., and Li, J. (2021). Deepnetbim: deep learning model for predicting hla-epitope interactions based on network analysis by harnessing binding and immunogenicity information. *BMC bioinformatics*, 22(1):1–16.
- Ye, Y., Wang, J., Xu, Y., Wang, Y., Pan, Y., Song, Q., Liu, X., and Wan, J. (2021). Mathla: a robust framework for hla-peptide binding prediction integrating bidirectional lstm and multiple head attention mechanism. *BMC bioinformatics*, 22(1):1–12.
- You, R., Qu, W., Mamitsuka, H., and Zhu, S. (2022). Deepmhcii: a novel binding core-aware deep interaction model for accurate mhc-ii peptide binding affinity prediction. *Bioinformatics*, 38(Supplement_1):i220–i228.

- Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J. (2021). Dive into deep learning. *arXiv preprint arXiv:2106.11342*.
- Zhang, L., Liu, G., Hou, G., Xiang, H., Zhang, X., Huang, Y., Zhang, X., Li, B., and Lee, L. J. (2022a). Introspect: Motif-guided immunopeptidome database building tool to improve the sensitivity of hla i binding peptide identification by mass spectrometry. *Biomolecules*, 12(4):579.
- Zhang, X., Qi, Y., Zhang, Q., and Liu, W. (2019). Application of mass spectrometry-based mhc immunopeptidome profiling in neoantigen identification for tumor immunotherapy. *Biomedicine & Pharmacotherapy*, 120:109542.
- Zhang, Y., Zhu, G., Li, K., Li, F., Huang, L., Duan, M., and Zhou, F. (2022b). Hlab: learning the bilstm features from the protbert-encoded proteins for the class i hla-peptide binding prediction. *Briefings in Bioinformatics*.
- Zhao, T., Cheng, L., Zang, T., and Hu, Y. (2019). Peptide-major histocompatibility complex class i binding prediction based on deep learning with novel feature. *Frontiers in Genetics*, 10:1191.
- Zhao, W. and Sher, X. (2018). Systematically benchmarking peptide-mhc binding predictors: From synthetic to naturally processed epitopes. *PLoS computational biology*, 14(11):e1006457.
- Zheng, Y., Fu, Y., Wang, P.-P., and Ding, Z.-Y. (2022). Neoantigen: A promising target for the immunotherapy of colorectal cancer. *Disease Markers*, 2022.
- Zhou, S., Liu, S., Zhao, L., and Sun, H.-X. (2022). A comprehensive survey of genomic mutations in breast cancer reveals recurrent neoantigens as potential therapeutic targets. *Frontiers in oncology*, 12.
- Zhou, W.-J., Qu, Z., Song, C.-Y., Sun, Y., Lai, A.-L., Luo, M.-Y., Ying, Y.-Z., Meng, H., Liang, Z., He, Y.-J., et al. (2019). Neopeptide: an immunoinformatic database of t-cell-defined neoantigens. *Database*, 2019.
- Zvyagin, I. V., Tsvetkov, V. O., Chudakov, D. M., and Shugay, M. (2020). An overview of immunoinformatics approaches and databases linking t cell receptor repertoires to their antigen specificity. *Immunogenetics*, 72(1):77–84.