EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Doctorado

Modelo predictivo sobre el emparejamiento de cadenas pesadas y ligeras de anticuerpos

presentada por

M.C. Manuel Erazo Valadez

como requisito para la obtención del grado de

Doctor en Ciencias de la Computación

Directora de tesis

Dra. María Yasmín Hernández Pérez

Codirectora de tesis

Dra. Elizabeth Ernestina Godoy Lozano

Cuernavaca, Morelos, México. Agosto de 2026.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

cenidet Centro Nacional de Investigación y Desarrollo Tecnológico	ACEPTACIÓN DE IMPRESIÓN DEL DOCUMENTO DE TESIS DOCTORAL	Código: CENIDET-AC-006-D20
		Revisión: O
	Referencia a la Norma ISO 9001:2008 7.1, 7.2.1, 7.5.1, 7.6, 8.1, 8.2.4	Página 1 de 1

Cuernavaca, Mor., a 03 de agosto de 2026

DR. CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
P R E S E N T E

AT'n: DRA. ANDREA MAGADÁN SALAZAR
PRESIDENTA DEL CLAUSTRO DOCTORAL DEL
DEPARTAMENTO DE CIENCIAS COMPUTACIONALES

Los abajo firmantes, miembros del Comité Tutorial del estudiante **M.C. Manuel Erazo Valadez** manifiestan que después de haber revisado el documento de tesis titulado **"Modelo predictivo sobre el emparejamiento de cadenas pesadas y ligeras de anticuerpos"**, realizado bajo la dirección de la **Dra. María Yasmín Hernández Pérez** y Codirigida por la **Dra. Elizabeth Ernestina Godoy Lozano,** el trabajo se ACEPTA para proceder a su impresión.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Conocimiento y Tecnología al Servicio de México"

DRA. MARÍA YASMÍN HERNÁNDEZ PÉREZ
TECNM/CENIDET

DRA. ELIZABETH ERNESTINA GODOY LOZANO
INSTITUTO NACIONAL DE SALUD PÚBLICA

DR. JAVIER ORTIZ HERNÁNDEZ
TECNM/CENIDET

DRA. ALICIA MARTÍNEZ REBOLLAR
TECNM/CENIDET

DR. JONATHAN VILLANUEVA TAVIRA
TECNM/CENIDET

DR. JUAN MAURICIO TÉLLEZ SOSA
INSTITUTO NACIONAL DE SALUD PÚBLICA

c.c.p: C Verónica Sotelo Boyas/ Jefa del Departamento de Servicios Escolares
c.c.p: C Nóe Alejandro Castro Sánchez / Jefe del Departamento de Ciencias Computacionales
c.c.p: Expediente

Educación
Secretaría de Educación Pública

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo Tecnológico
Subdirección Académica

Cuernavaca Mor, 07/agosto/2026
Oficio No. SAC/239/2026
Asunto: Autorización de impresión de tesis

MANUEL ERAZO VALADEZ
CANDIDATO AL GRADO DE DOCTOR EN CIENCIAS DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Modelo predictivo sobre el emparejamiento de cadenas pesadas y ligeras de anticuerpos"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Conocimiento y Tecnología al servicio de México"

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

c.c.p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/lmz

2026 año de Margarita Maza

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

Interior Internado Palmira S/N, Col. Palmira,
C.P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 4106,
e-mail: acad_cenidet@tecnm.mx | cenidet.tecnm.mx

Dedicatoria

A mi papá Oscar Manuel Erazo, que siempre será una de las personas más importantes en mi vida, que gracias a sus consejos y enseñanzas he llegado a cumplir cada uno de los objetivos que me he propuesto en la vida. Espero que desde el cielo pueda ver y este orgulloso de este objetivo que he cumplido. También este trabajo es dedicado para mi mamá y a mi hermano que son y siempre serán los pilares de mi vida y que sin ellos no hubiera alcanzado este objetivo.

Agradecimientos

Agradezco a la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), por otorgarme una beca durante todos los meses que tomó el desarrollo de este proyecto y así permitirme poder enfocarme solo en desarrollar esta tesis.

También quiero agradecer al Tecnológico Nacional de México (TecNM/CENIDET), por permitirme realizar un Doctorado en Ciencias de la computación y facilitarme en sus instalaciones todo lo necesario para poder desarrollar esta tesis.

A mis padres Oscar Erazo y Diana Valadez gracias por todo el apoyo incondicional que me han brindado a lo largo de mi vida, por impulsarme a realizar cada una de mis metas y por brindarme todo el amor que me han dado.

Muchas gracias, mamá por estar conmigo tanto en los momentos felices como en los momentos tristes. Este es un logro de los dos ya que aún recuerdo cada una de las desveladas hasta la madrugada que te quedabas conmigo para no me quedara solo haciendo mi tarea por eso y muchas cosas más gracias, mamá.

Agradezco a mi hermano Fernando Erazo por siempre estar a mi lado apoyándome y cuidándome en cada una de las decisiones que he tomado a lo largo de mi vida. Gracias por estar ahí siempre para mí y en ocasiones dejando de lado tus cosas para poder estar conmigo y apoyarme.

También le agradezco a dos personas muy importantes en mi vida, a mis abuelitos Rosalio Valadez y Areopajita Pineda por estar siempre apoyándome a mí y a mi familia. Gracias por todo abuelitos, sobre todo por estar siempre y brindarme su apoyo en las ocasiones que lo he necesitado.

A mis tías Pajis, Bere y a mi tío Ulises gracias por su apoyo.

Le agradezco a mi directora de tesis la Dra. María Yasmín Hernández Pérez y a mi codirectora de tesis Dra. Elizabeth Ernestina Godoy Lozano por todo el apoyo que me brindaron y por compartir sus conocimientos para la realización de esta tesis. Además, brindarnos su amistad incondicional y alentarme a seguir preparándome.

Resumen

Los anticuerpos son moléculas formadas por cadenas pesadas y cadenas ligeras cuyo emparejamiento correcto es esencial para la formación de un sitio de unión funcional. Sin embargo, en los métodos de secuenciación de nueva generación, ambas cadenas se obtienen de forma separada, lo que provoca la pérdida del emparejamiento natural y genera el problema de identificar, a partir de información de secuencia, qué combinaciones forman anticuerpos funcionales.

Para abordar este problema, se construyó un modelo computacional basado en una red neuronal siamesa asimétrica entrenada con la función de pérdida InfoNCE con temperatura τ=0.07, diseñada para aprender un espacio métrico en donde las cadenas que forman anticuerpos funcionales se encuentran cercanas en términos de distancia coseno, mientras que las combinaciones incompatibles se encuentran más alejadas. Las secuencias de aminoácidos se representaron mediante los factores Atchley, una codificación fisicoquímica que captura propiedades como polaridad, hidrofobicidad, carga, volumen molecular y tendencia estructural, y que demostró el mejor desempeño frente a representaciones alternativas evaluadas. La arquitectura cuenta con *encoders* especializados para cada tipo de cadena: el *encoder* de cadenas ligeras emplea activaciones GELU con normalización por lotes, y el *encoder* de cadenas pesadas utiliza activaciones ReLU, proyectando ambas cadenas a un espacio embebido de ocho dimensiones.

Una contribución central del trabajo es el uso de cadenas pesadas de ratón BALB/c como negativos verdaderos en el entrenamiento. Debido a las divergencias germinales, diferencias estructurales y ausencia de coevolución entre especies, estas cadenas no pueden formar anticuerpos funcionales con cadenas ligeras humanas, lo que proporciona una señal negativa con respaldo biológico sólido. Los resultados mostraron que esta decisión tuvo un impacto directo en la calidad del espacio métrico aprendido: la separación entre pares compatibles e incompatibles mejoró sustancialmente en comparación con el modelo entrenado sin negativos verdaderos.

El modelo fue evaluado sobre un conjunto de prueba independiente, obteniendo una exactitud de 0.97, F1-*Score* de 0.96, *Recall* de 0.94 y ROC-AUC de 0.97. En la evaluación de recuperación por ranking, se alcanzó un *Recall*@1 de 71.55%, *Recall*@5 de 91.77% y *Recall*@10 de 93.06%, lo que confirma la utilidad del modelo para reducir el espacio de cadenas candidatas en repertorios de gran escala. El análisis del espacio de *embeddings* mediante UMAP reveló una organización coherente con propiedades biológicas como la familia germinal V, el gen V específico y la longitud del CDR3, sin que estas propiedades fueran incluidas explícitamente en el entrenamiento.

Los resultados demuestran que es posible construir un modelo computacional con capacidad discriminativa real para el problema del emparejamiento de cadenas de anticuerpos, y que la calidad biológica de los negativos de entrenamiento es un factor determinante en el desempeño del modelo y que las representaciones fisicoquímicas compactas como los factores Atchley son una alternativa viable y efectiva para este tipo de tareas.

Abstract

Antibodies are molecules formed by heavy and light chains whose correct pairing is essential for the formation of a functional binding site. However, in next-generation sequencing methods, both chains are obtained separately, which causes the loss of the natural pairing and generates the problem of identifying, from sequence information, which combinations form functional antibodies.

To address this problem, a computational model was built based on an asymmetric siamese neural network trained with the InfoNCE loss function with temperature τ=0.07, designed to learn a metric space where chains forming functional antibodies are close in terms of cosine distance, while incompatible combinations are further apart. Amino acid sequences were represented using Atchley factors, a physicochemical encoding that captures properties such as polarity, hydrophobicity, charge, molecular volume and structural tendency, and which demonstrated the best performance against alternative representations evaluated. The architecture includes specialized encoders for each chain type: the light chain encoder uses GELU activations with batch normalization, and the heavy chain encoder uses ReLU activations, projecting both chains into an eight-dimensional embedding space.

A central contribution of this work is the use of BALB/c mouse heavy chains as true negatives during training. Due to germline divergences, structural differences and the absence of coevolution between species, these chains cannot form functional antibodies with human light chains, providing a biologically grounded negative signal. The results showed that this decision had a direct impact on the quality of the learned metric space: the separation between compatible and incompatible pairs improved substantially compared to the model trained without true negatives.

The model was evaluated on an independent test set, obtaining an accuracy of 0.97, F1-Score of 0.96, Recall of 0.94 and ROC-AUC of 0.97. In the ranking-based retrieval evaluation, a Recall@1 of 71.55%, Recall@5 of 91.77% and Recall@10 of 93.06% were achieved, confirming the utility of the model for reducing the candidate chain search space in large-scale repertoires. Analysis of the embedding space through UMAP revealed an organization consistent with biological properties such as the V germline family, the specific V gene and CDR3 length, without these properties being explicitly included during training.

The results demonstrate that it is possible to build a computational model with real discriminative capacity for the antibody chain pairing problem, and that the biological quality of the training negatives is a determining factor in model performance, and that compact physicochemical representations such as Atchley factors are a viable and effective alternative for this type of task.

Contenido

Lista de figuras

Lista de tablas

Capítulo 1.
Introducción

1. Introducción

Algunos organismos vivos poseen un sistema inmune adaptativo que les permite eliminar infecciones con una mayor eficacia que el sistema inmune innato. Este tipo de respuesta está presente exclusivamente en vertebrados y se basa en mecanismos de reconocimiento específico de antígenos mediados por receptores B y T. Los receptores de las células B, llamadas inmunoglobulinas (lg) en su forma soluble, son proteínas capaces de reconocer una amplia gama de antígenos. Los antígenos, a su vez, son cualquier sustancia extraña que pueda provocar una respuesta inmunológica (Janeway, 2000).

Cada anticuerpo es producido por linfocitos B y se caracterizan por su alta especificidad hacia un antígeno, así como por su capacidad de generar memoria inmunológica que protege al organismo en contra de futuras infecciones. Estas moléculas están formadas por dos cadenas pesadas (del inglés *heavy chain*, VH) y dos cadenas ligeras (del inglés *light chain*, VL), unidas mediante enlaces de disulfuro y estabilizadas por interacciones no covalentes. Las cadenas pesadas son de mayor tamaño y determinan la clase o isotipo del anticuerpo, mientras que las cadenas ligeras, de menor tamaño, se asocian a cada cadena pesada y contribuyen junto con ella a la formación del sitio de unión al antígeno. En humanos, aproximadamente el 60% de las cadenas ligeras son kappa (k) y el 40% restante son lambda (l). La correcta unión de las cadenas pesadas y ligeras es crucial para la formación de un sitio de unión funcional y eficiente, clave para la unión y posible neutralización de antígenos como toxinas o patógenos (Tonegawa, 1983).

La región variable de estas cadenas está compuesta por tres Regiones Determinantes de la Complementariedad (del inglés *Complementarity-Determining Region*, CDRs) y cuatro regiones estructurales denominadas *Frameworks* (FR1-FR4), esta región define la especificidad de unión al antígeno y mantiene la estructura necesaria para una interacción funcional. La complejidad estructural de estas regiones, junto con la diversidad generada por la recombinación somática, hipermutación y selección *clonal*, hacen que el repertorio de anticuerpos sea un sistema altamente dinámico, diverso y difícil de analizar de forma exhaustiva (Janeway, 2000; Schroeder & Cavacini, 2010).

Uno de los principales desafíos en el estudio computacional de anticuerpos es que, en la mayoría de los métodos de secuenciación de nueva generación (del inglés *Next Generation Sequencing,* NGS), las cadenas pesadas y ligeras se obtienen por separado. Esto provoca la pérdida del emparejamiento natural VH-VL (emparejamiento VH-VL, es decir, cadena pesada–cadena ligera), lo que genera el problema de identificar directamente qué cadenas forman anticuerpos funcionales. Esta falta de información limita la caracterización funcional de los repertorios inmunes, el descubrimiento de anticuerpos terapéuticos y la generación de modelos computacionales que permitan predecir combinaciones compatibles (Calis & Rosenberg, 2014).

En los últimos años, el aprendizaje automático ha adquirido una importancia significativa como rama de la inteligencia artificial con amplias aplicaciones en diversos campos de investigación. En el campo de la inmunología, se ha aprovechado el aprendizaje automático

para analizar la respuesta del sistema inmune adaptativo a las vacunas y las infecciones, así como para identificar moléculas con potencial terapéutico, incluidos los anticuerpos monoclonales (mAbs) (Greiff et al., 2020). Específicamente, se ha empleado para descubrir *in silico* nuevos mAbs, optimizar su unión con la molécula diana y comprender las propiedades biofísicas de la interacción antígeno-anticuerpo.

Al detectar este problema en el análisis de repertorios de anticuerpos, se desarrolló un modelo de aprendizaje automático basado en redes neuronales siamesas entrenadas con InfoNCE (*Information Noise-Contrastive Estimation*), diseñado para aprender un espacio métrico en donde las cadenas pesadas y ligeras que forman anticuerpos funcionales se encuentran cercanas en términos de distancia coseno, mientras que combinaciones que son incompatibles se encuentran más alejadas. Para lograrlo, las secuencias de aminoácidos de las cadenas pesadas y ligeras se representaron utilizando los factores Atchley, que es una codificación fisicoquímica que captura información relevante sobre la polaridad, hidrofobicidad, carga, volumen y tendencia estructural de los aminoácidos (Atchley et al., 2005).

La construcción del conjunto de datos es una actividad fundamental para la generación del modelo. Para representar las relaciones entre las cadenas, se utilizaron dos tipos de pares de cadenas pesadas y ligeras.

Los pares verdaderos humanos, se obtuvieron de anticuerpos completos con emparejamientos VH-VL conocidos.

Los negativos reales, generados a partir de anticuerpos de ratón BALB/c, cuyas cadenas pesadas no pueden formar anticuerpos funcionales con cadenas ligeras humanas debido a divergencias germinales, diferencias estructurales en las regiones variables y ausencia de coevolución entre especies.

Al utilizar pares de cadenas de humanos y ratón se obtiene un conjunto negativo biológicamente sólido que reduce la posibilidad de solapamiento entre pares funcionales y no funcionales, fortaleciendo el entrenamiento con InfoNCE y permitiendo una separación más clara en el espacio embebido.

Este trabajo presenta un modelo computacional diseñado para inferir la compatibilidad funcional entre cadenas pesadas y ligeras utilizando representaciones fisicoquímicas, aprendizaje métrico y un conjunto de datos estructurado que integra tanto pares humanos reales como negativos verdaderos de otra especie. Esta investigación contribuye al análisis computacional de repertorios inmunológicos y ofrece una herramienta viable para el diseño racional de anticuerpos terapéuticos.

1.1. Planteamiento del problema

Los anticuerpos son las moléculas encargadas de identificar y neutralizar agentes patógenos dentro del sistema inmune adaptativo, y su capacidad para hacerlo con especificidad depende de la interacción entre sus cadenas pesadas y ligeras al formar el sitio de unión al antígeno. Sin embargo, sigue siendo complicado predecir qué factores determinan el emparejamiento

correcto entre ambas cadenas, debido a la amplia variabilidad que existe en sus secuencias y características estructurales, así como al gran número de combinaciones posibles entre los genes que codifican sus regiones variables.

Desde el punto de vista computacional, resolver este problema es complicado debido a varias razones. Primero, el número de combinaciones posibles entre cadenas pesadas y ligeras depende del tamaño del repertorio de anticuerpos, lo que complica la evaluación de cada uno de los candidatos. Segundo, las secuencias de aminoácidos tienen una longitud variable y una alta variabilidad en las Regiones Determinantes de la Complementariedad (CDR), en especial en el CDR3, lo que complica su transformación en una representación numérica que conserve la información relevante para decidir la compatibilidad entre cadenas. Tercero, no existe una fuente natural de pares negativos verdaderos: al no contar con un repertorio confiable de combinaciones biológicamente imposibles, los negativos suelen generarse de forma aleatoria o artificial, lo que introduce ruido en el entrenamiento y limita la capacidad de cualquier modelo para predecir una frontera clara entre pares compatibles e incompatibles.

Tradicionalmente, se utilizan métodos heurísticos o reglas basadas en conocimiento experto para atacar este problema, como estrategias que generan emparejamientos de manera aleatoria o mediante combinaciones artificiales. Sin embargo, estas aproximaciones presentan limitaciones importantes: algunos emparejamientos considerados "negativos" podrían no estar biológicamente descartados, introduciendo ruido en el entrenamiento de los modelos y debilitando su capacidad de aprender fronteras claras entre combinaciones funcionales y no funcionales.

En este sentido, el aprendizaje automático y el reconocimiento de patrones ofrecen enfoques más prometedores al permitir extraer características relevantes de las secuencias de aminoácidos, identificar patrones complejos y aprender representaciones en espacios métricos en donde las parejas compatibles se encuentren cercanas y las incompatibles alejadas. Sin embargo, la calidad de los modelos depende de la construcción del conjunto de datos, específicamente de contar con positivos bien definidos y negativos reales con un respaldo biológico.

Una limitación es la ausencia de un conjunto de datos negativo que represente casos biológicamente imposibles de emparejarse. En este proyecto se propone utilizar anticuerpos de ratón BALB/c como fuente de verdaderos negativos, bajo la premisa de que sus cadenas pesadas no forman anticuerpos funcionales con cadenas ligeras humanas debido a divergencias germinales, diferencias estructurales y ausencia de coevolución entre especies. Integrar estos negativos en el entrenamiento de una red siamesa con InfoNCE busca fortalecer la separación en el espacio métrico entre pares humanos funcionales y combinaciones entre especies incompatibles.

1.2. Metodología de solución

La metodología propuesta se divide en cuatro fases principales: obtención de datos, preprocesamiento y construcción del conjunto de datos, construcción del modelo para el emparejamiento de cadenas pesadas y ligeras, y evaluación del modelo para el emparejamiento de cadenas pesadas y ligeras, la metodología está diseñada para abordar el problema del emparejamiento funcional entre cadenas pesadas y ligeras de anticuerpos. Esta secuencia metodológica integra técnicas de procesamiento bioinformático, representación numérica, aprendizaje automático y validación experimental. La figura 1.1, muestra las cuatro etapas de la metodología seguida durante esta investigación.

1.2.1. Fase 1: obtención de datos

En esta fase se realiza la recolección de secuencias de anticuerpos en repositorios públicos especializados, como OAS (*Observed Antibody Space*) y CoV-AbDab. El objetivo es construir un conjunto de datos que contenga tanto pares funcionales reales como negativos biológicamente fundamentados.

Se consideraron dos tipos de anticuerpos:

- Anticuerpos humanos: se obtienen las cadenas pesadas y ligeras con emparejamientos conocidos, que se utilizan como pares positivos funcionales, es decir combinaciones que si forman un anticuerpo funcional
- Anticuerpos de ratón BALB/c: de estos anticuerpos solo se extraen las cadenas pesadas, que sirven como verdaderos negativos al no ser compatibles con las cadenas ligeras de humano. Esta incompatibilidad refleja divergencias germinales, diferencias estructurales y ausencia de coevolución entre especies, por lo que forma un negativo biológicamente fundamentado y no generado de forma artificial

Todas las secuencias descargadas se almacenan en archivos FASTA, preservando la información necesaria para su anotación y codificación posterior.

1.2.2. Fase 2: Preprocesamiento y construcción del conjunto de datos

Esta fase considera las acciones necesarias para transformar las secuencias de aminoácidos en representaciones numéricas que puedan ser utilizadas por el modelo, mediante los siguientes procesos:

- Limpieza y filtrado de datos: se depuran los conjuntos de datos iniciales eliminando los registros que presentan información faltante, inconsistencias en las anotaciones o errores de formato, con el objetivo de garantizar la calidad y confiabilidad del conjunto de datos. Solo se conservan las cadenas pesadas y ligeras clasificadas como funcionales, descartando secuencias no productivas o con mutaciones que impidan la funcionalidad del anticuerpo.
- Anotación estructural (IMGT): las secuencias se procesan con la herramienta IMGT/HighV-QUEST de IMGT para identificar y extraer los CDRs (CDR1, CDR2, CDR3), las regiones que tienen una mayor variabilidad y que son determinante en el

reconocimiento del antígeno. También se identifican los Frameworks (FR1-FR4), estas son las regiones más conservadas, de soporte estructural. Esta anotación garantiza que el modelo trabaje con regiones comparables y estructuralmente relevantes.

- Codificación numérica: sobre las regiones anotadas se evaluaron y compararon tres esquemas de representación (bolsa de palabras TF-IDF, factores Atchley y ProtVec). Los factores Atchley resultaron ser la representación con mejor desempeño experimental: cada aminoácido de las regiones anotadas se transforma en un vector de características de 5 dimensiones que resumen sus propiedades fisicoquímicas (polaridad, carga, hidrofobicidad, volumen y tendencia estructural). Las secuencias codificadas se almacenan en archivos CSV separados para: cadenas ligeras humanas, cadenas pesadas humanas y cadenas pesadas de ratón.

1.2.3. Fase 3: Construcción del modelo para el emparejamiento de cadenas pesadas y ligeras de anticuerpos

En esta fase se aborda el diseño, la implementación y el entrenamiento del modelo computacional propuesto para la predicción del emparejamiento funcional entre cadenas pesadas y ligeras de anticuerpos. El modelo está basado en una red neuronal siamesa asimétrica entrenada con InfoNCE, arquitectura que permite proyectar los dos tipos de cadena en un espacio métrico compartido, en donde la compatibilidad funcional se refleja en la distancia coseno entre sus representaciones vectoriales.

A diferencia de los enfoques tradicionales que generan negativos de manera aleatoria, en esta investigación se incorporan cadenas pesadas de anticuerpos de ratón BALB/c como negativos biológicamente fundamentados, bajo la premisa de que no forman anticuerpos funcionales con cadenas ligeras de humanos debidos a divergencias germinales, diferencias estructurales y ausencia de coevolución entre especies. Adicionalmente, se utiliza la función de perdida InfoNCE, cuyo mecanismo de negativos in-batch agrega de manera implícita un comportamiento equivalente a la minería de negativos semi-hard, seleccionando automáticamente los ejemplos más informativos para el entrenamiento en cada iteración.

Esta etapa se divide en dos partes: primero se describe la construcción de las tripletas de entrenamiento y la estrategia de negativos, después se detalla la arquitectura de la red siamesa y su función de perdida.

Generación de tripletas:

Para el entrenamiento del modelo basado en una red neuronal siamesa asimétrica se construyeron tripletas con la estructura (ancla, positivo, negativo), donde cada elemento representa un tipo de cadena de anticuerpo con un rol especifico dentro de la función de perdida:

- Ancla (A): vector que representa una cadena ligera humana
- Positivo (P): vector que representa la cadena pesada humana que se empareja funcionalmente con el ancla, formando un anticuerpo real

- Negativo (N): vector que representa una cadena pesada que no forma un anticuerpo funcional con el ancla

El modelo recibe simultáneamente tres tipos de secuencias como entrada: las cadenas ligeras humanas, las cadenas pesadas humanas emparejadas y las cadenas pesadas de ratón BALB/c. Las cadenas ligeras actúan siempre como el ancla, las cadenas pesadas humanas emparejadas como positivos, y las cadenas pesadas de BALB/c como negativos verdaderos.

Red siamesa

La red propuesta es una arquitectura siamesa asimétrica, formada por dos encoders independientes —uno para cadenas ligeras y otro para cadenas pesadas—, dado que ambos tipos de cadena presentan distinta complejidad estructural:

- El EncoderLigera es una red neuronal densa compuesta por tres capas sucesivas, con activación GELU, normalización por lotes (BatchNorm1d) y regularización por Dropout
- El EncoderPesada utiliza activaciones ReLU y dos capas densas intermedias, con mayor profundidad para reflejar la complejidad estructural que tienen las cadenas pesadas en comparación con las cadenas ligeras

Cada encoder proyecta las secuencias codificadas en un espacio latente de baja dimensión (8 dimensiones), seguido de una normalización L2 que permite utilizar la similitud coseno como medida de comparación. Adicionalmente, se incorpora un módulo de cross-attention bidireccional entre las representaciones de cadena pesada y ligera, que permite que cada región (CDR/FR) de una cadena atienda a las regiones relevantes de la otra antes de la proyección final.

Durante el entrenamiento:

- Se minimiza la pérdida InfoNCE, obligando a que las combinaciones funcionales queden cercanas en el espacio embebido y las incompatibles queden alejadas.
- Los negativos más informativos se seleccionan de forma implícita mediante el mecanismo de negativos in-batch con temperatura $\tau=0.07$, equivalente a una minería semi-hard automática.

1.2.4. Fase 4: evaluación del modelo para el emparejamiento de cadenas pesadas y ligeras de anticuerpos

La evaluación del modelo se realizó sobre el conjunto de prueba, compuesto por cadenas pesadas y ligeras que no fueron utilizadas durante el entrenamiento. Para caracterizar su desempeño en distintos escenarios de uso, se utilizaron dos tipos de métricas complementarias: métricas de clasificación binaria, que evalúan la capacidad del modelo para distinguir pares compatibles a partir de un umbral de distancia, y métricas de recuperación (Recall@K), que evalúan su capacidad para encontrar el emparejamiento correcto dentro de un ranking de candidatos.

Métricas

- Accuracy, precisión, recall y F1-score, a partir de un umbral óptimo sobre la distancia coseno.
- Métricas de ranking como Recall@K (Top-K), que indican con qué frecuencia la cadena pesada correcta aparece entre las K cadenas más similares a una cadena ligera.
- Métricas globales como ROC AUC, para evaluar la separación entre pares compatibles e incompatibles.

Figura 1.1. Metodología de solución

1.3. Objetivos

En esta sección se presenta el objetivo general y los objetivos específicos de este trabajo de investigación

1.3.1. Objetivo general

Desarrollar un modelo computacional para el emparejamiento de cadenas pesadas y ligeras de anticuerpos, utilizando representaciones fisicoquímicas a nivel de aminoácidos, con el fin de inferir la compatibilidad funcional entre ambas cadenas.

1.3.2. Objetivos específicos

- Construir un conjunto de datos con información detallada sobre las características estructurales de las secuencias de las cadenas pesadas y ligeras de los anticuerpos, para facilitar el análisis computacional de su posible compatibilidad estructural.

- Analizar las propiedades de los anticuerpos, para identificar regiones clave involucradas en el emparejamiento entre cadenas pesadas y ligeras, por medio de técnicas estadísticas y visualización de datos.
- Construir modelos de aprendizaje automático, para evaluar la capacidad de distintas representaciones numéricas en tareas de predicción, utilizando las características derivadas de las secuencias de aminoácidos.
- Analizar los resultados obtenidos por los modelos, para identificar patrones y relaciones entre las propiedades estructurales de las proteínas y el emparejamiento entre cadenas pesadas y ligeras.
- Integrar las representaciones y configuraciones más efectivas en un modelo final, con el fin de proponer una solución optimizada para predecir el emparejamiento entre cadenas pesadas y ligeras de anticuerpos.

1.4. Hipótesis

Un modelo de aprendizaje que representa las cadenas pesadas y ligeras de anticuerpos a partir de sus propiedades fisicoquímicas, e incorpora ejemplos negativos con fundamento biológico durante su entrenamiento, es capaz de inferir la compatibilidad funcional entre ambas cadenas con un desempeño mayor al que se obtendría utilizando negativos generados de forma artificial, reflejado en las métricas de clasificación y recuperación con las que se evalúa el modelo.

1.5. Estructura del documento

En esta sección se describe la estructura de cómo está organizada esta investigación.

Capítulo 2. Marco teórico

Define los conceptos teóricos y técnicos fundamentales para el estudio, incluyendo la estructura de los anticuerpos, las regiones CDRs y FRs, y los fundamentos del aprendizaje automático y sus aplicaciones en bioinformática.

Capítulo 3. Análisis del estado del arte

Se presentan investigaciones relacionadas con el análisis computacional de anticuerpos y enfoques de aprendizaje automático para tareas similares a las de esta investigación.

Capítulo 4. Construcción de la representación de los aminoácidos

Describe el proceso de evaluación y comparación entre los diferentes métodos de representación numérica, incluyendo bolsa de palabras TF-IDF, factores Atchley y ProtVec.

Capítulo 5. Modelo para el emparejamiento de cadenas pesadas y ligeras

Describe el diseño, implementación y entrenamiento del modelo siamesa propuesto, basado en InfoNCE. Se explican sus componentes, la arquitectura y la lógica de los *encoders* asimétricos para cada tipo de cadena de los anticuerpos

Capítulo 6. Resultados

Presenta los hallazgos experimentales derivados de la evaluación del modelo. Se muestran las métricas que se obtuvieron, la calidad de los *embeddings* generados y la validez biológica de las predicciones.

Capítulo 7. Conclusiones y trabajos futuros

Se presentan las principales contribuciones del trabajo de investigación, mostrando los avances que se lograron en el modelado del emparejamiento de las cadenas de los anticuerpos. Se mencionan posibles líneas de investigación a futuro como el uso de datos negativos biológicamente más robustos o la extensión del modelo a otros repertorios inmunes.

Capítulo 2. Marco teórico

2. Marco teórico

En este capítulo se abordan los conceptos esenciales relacionados con la estructura de los anticuerpos, las estrategias de representación numérica de las secuencias de aminoácidos y los fundamentos del aprendizaje métrico mediante redes siamesas. Estos elementos son la base conceptual del modelo propuesto, ya que permiten entender cómo las propiedades fisicoquímicas de las cadenas pueden ser transformadas en vectores comparables y cómo un espacio métrico puede capturar la compatibilidad funcional de las cadenas pesadas y ligeras.

2.1. Sistema inmune

El sistema inmune es una red de células, tejidos y moléculas que han evolucionado para proteger a algunos organismos vivos en contra de agentes patógenos. Esta protección cuenta con dos funciones principales: el reconocimiento de los patógenos y la respuesta efectora en contra de ellos. El reconocimiento inmunológico puede distinguir entre los componentes que pertenecen al organismo y componentes extraños, como bacterias, virus, hongos o células anómalas.

Cuando el sistema inmune reconoce un patógeno, se activa una serie de mecanismos efectores destinados a eliminarlo. La respuesta inmune es adaptable y específica, permitiendo al organismo generar reacciones en contra de distintos tipos de patógenos. Algunas de estas respuestas generan memoria inmunológica, una característica que permite responder de forma más rápida e intensa en contra de futuras exposiciones al mismo patógeno. Este es el principio biológico de la vacunación, la cual entrena al sistema inmune para estar listo en contra de futuras infecciones.

El sistema inmune se divide en dos componentes: el sistema inmune innato y el sistema inmune adaptativo.

El Sistema inmune innato es la primera línea de defensa. Esta presente desde el nacimiento y se compone por mecanismos celulares y moleculares que actúan de forma inmediata en contra de los patógenos. Su principal función es prevenir la infección o eliminarla en las primeras horas después del contacto con el patógeno.

El sistema inmune adaptativo se desarrolla por la exposición inicial al patógeno. Se requiere de varios días para que se active, pero ofrece una respuesta altamente específica y duradera. Es el encargado de eliminar las amenazas que logran pasar del sistema inmune innato, y de generar células de memoria que permiten respuestas más eficientes en contra de futuras exposiciones con el mismo antígeno

Los dos sistemas trabajan en conjunto para mantener al organismo sano. La inmunidad innata no solo actúa como primera barrera, sino que también participa en la activación y modulación de la respuesta adaptativa, garantizando así una defensa inmunológica integrada y eficaz (Kindt et al., 2007).

2.2. Estructura de los anticuerpos

Los anticuerpos o inmunoglobulinas (lg), son producidas por los linfocitos B como parte de la respuesta inmune adaptativa. Su estructura está conformada por dos cadenas pesadas (IgH) idénticas y dos cadenas ligeras (IgL) idénticas, unidas por enlaces de disulfuro e interacciones no covalentes. Las cadenas ligeras pueden pertenecer a dos clases: kappa (k) y lambda (l) (Schroeder & Cavacini, 2010).

Funcionalmente los anticuerpos pueden dividirse en dos regiones principales:

La región variable (V), está ubicada en el extremo amino terminal de cada cadena, es responsable del reconocimiento y unión con el antígeno.

La región constante (C), es la que define el isotipo del anticuerpo y sus funciones efectoras en contra del antígeno, como la activación o la interacción con receptores celulares. Cuando el anticuerpo es el receptor de célula B (BCR), la región constante permanece anclada a la membrana y no ejecuta funciones efectoras (Schroeder & Cavacini, 2010).

A nivel proteico, la variabilidad de la región V no se distribuye de forma uniforme. Se concentra en tres segmentos conocidos como regiones determinantes de complementariedad (CDRs): CDR1, CDR2 y CDR3. Estas asas hipervariables son responsables del contacto directo con el antígeno. Cada cadena (pesada o ligera) aporta tres CDRs, y cuando ambas cadenas están emparejadas, sus CDRs forman un sitio hipervariable único, que constituye el sitio de unión al antígeno (Schroeder & Cavacini, 2010).

Los CDRs se encuentran intercaladas con cuatro regiones más conservadas llamadas *Frameworks* (FR1, FR2, FR3 y FR4). Los *Frameworks* forman la estructura de la región variable y son esenciales para mantener la conformación tridimensional adecuada de los CDRs. Aunque no participan directamente en la unión al antígeno, los FRs aseguran la estabilidad del dominio variable y permiten que las CDRs se proyecten en la orientación correcta para interactuar eficazmente con el epítopo correspondiente (Schroeder & Cavacini, 2010).

En la figura 2.1, se muestra la estructura que tienen el anticuerpo y el emparejamiento de las cadenas pesadas y ligeras.

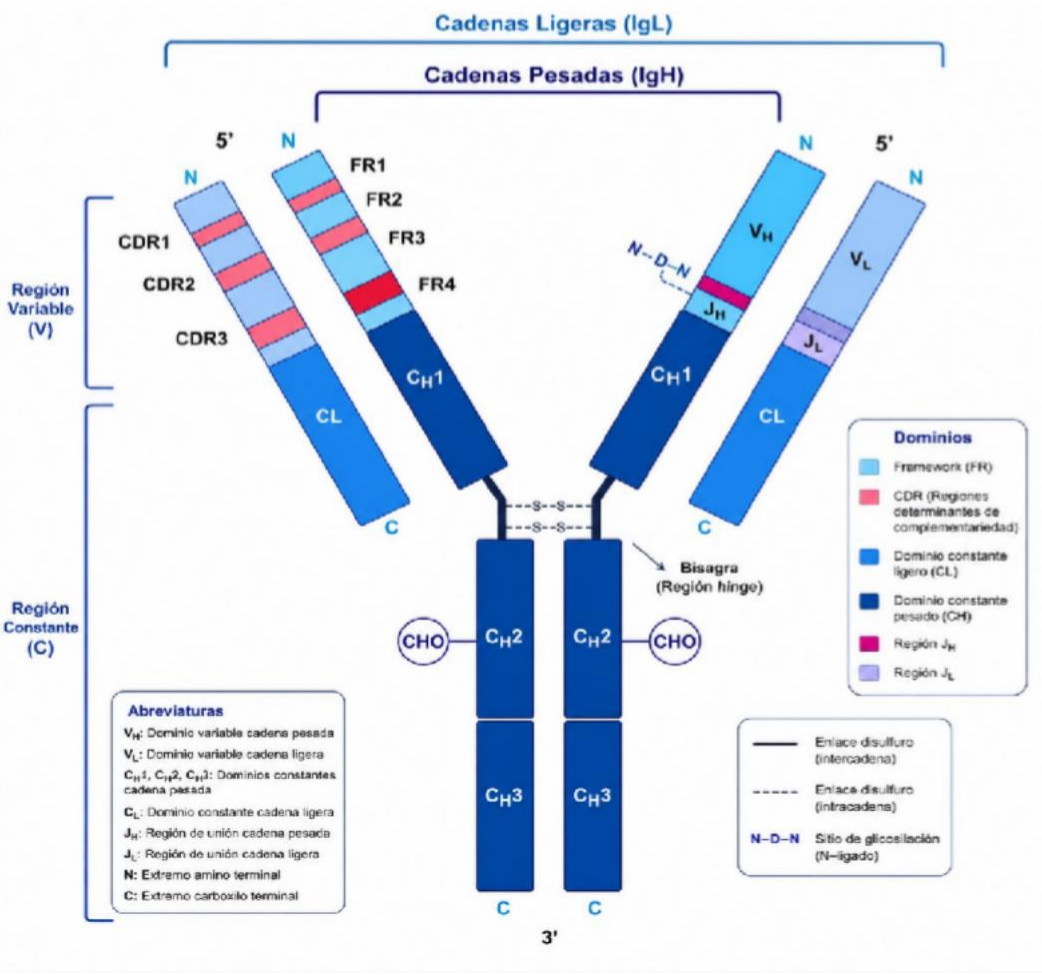

Figura 2.1. Estructura del anticuerpo

2.3. Bioinformática aplicada al análisis de anticuerpos

La bioinformática es la disciplina que utiliza técnicas computacionales al análisis de información biológica a gran escala, con el objetivo de extraer, organizar e interpretar datos que, por su volumen o complejidad, resultan difíciles de abordar mediante métodos experimentales tradicionales (Khan et al., 2024). En el caso de los anticuerpos, la bioinformática permite pasar de una secuencia de aminoácidos a una representación estructurada y analizable computacionalmente, sin perder la información funcional que esa secuencia contiene.

Un ejemplo directo de esto, ya utilizado en la metodología de este trabajo, es la anotación estructural mediante las herramientas de IMGT (Manso et al., 2022): estas herramientas bioinformáticas toman una secuencia de aminoácidos como entrada y la segmentan automáticamente en sus regiones funcionales (CDR1, CDR2, CDR3 y Frameworks FR1-FR4), es decir, traducen el conocimiento biológico sobre la estructura del anticuerpo —descrito en la sección anterior— en información delimitada y estandarizada que puede procesarse computacionalmente.

A partir de esta anotación, la bioinformática provee además los fundamentos para representar cada región de forma numérica, ya sea mediante propiedades fisicoquímicas de los aminoácidos, mediante métodos estadísticos de procesamiento de texto, o mediante modelos de lenguaje entrenados sobre secuencias biológicas. Estas representaciones son las que finalmente permiten aplicar técnicas de aprendizaje automático y aprendizaje profundo sobre datos de origen biológico, y son el tema de la siguiente sección.

2.4. Representación de las secuencias de aminoácidos

El análisis computacional de los anticuerpos requiere realizar una conversión de sus secuencias de aminoácidos en vectores numéricos que puedan ser utilizados por los algoritmos de aprendizaje automático. En esta fase del preprocesamiento y representación de las cadenas que conforman a los anticuerpos es de suma importancia, ya que de esta fase depende que se conserven las características estructurales, funcionales y bioquímicas de los anticuerpos. Una representación inadecuada puede provocar la pérdida de información importante para la investigación, puede disminuir la capacidad predictiva de los modelos.

Las cadenas pesadas y ligeras de los anticuerpos tienen una alta variabilidad en longitud y composición, especialmente en los CDRs, en donde se encuentra la especificidad del reconocimiento de los antígenos. En este trabajo de investigación se evaluaron tres enfoques para representar estas secuencias de aminoácidos.

2.4.1. TF-IDF

Es una técnica estadística de procesamiento de texto que convierte contenido textual en vectores numéricos, con el objetivo de establecer la importancia de una palabra dentro de un documento y su relevancia dentro del conjunto de documentos (corpus) (Jurafsky & Martin, 2021). Se utiliza comúnmente en tareas como categorización de documentos, extracción de palabras clave y recuperación de información.

En el contexto de esta investigación, los conceptos de TF-IDF se adaptan del dominio textual al dominio de las secuencias de aminoácidos: cada secuencia de aminoácidos (o región anotada de la cadena) se trata como un documento, cada aminoácido se trata como una palabra o término, y el conjunto de secuencias utilizado equivalente al corpus.

TF-IDF integra tres componentes clave para transformar datos textuales en vectores numéricos:

- Frecuencia del término (TF): refleja la frecuencia en la que aparece un término dentro de un documento.
- Frecuencia inversa (IDF): reduce el peso de los términos comunes en el conjunto de datos
- TF-IDF: producto de los anteriores, resalta los términos más relevantes dentro del corpus.

Con estos tres componentes, cada cadena de aminoácidos queda representada como un vector numérico ene que cada entrada indica la importancia de un aminoácido en específico, lo que

permite que los algoritmos de aprendizaje automático trabajen con datos numéricos en lugar de texto.

En esta investigación, la frecuencia de término (TF) de un aminoácido se calcula como la proporción entre su número de apariciones y el número total de aminoácidos en la secuencia. La frecuencia inversa de documento (IDF) se calcula con el logaritmo del cociente entre el número total de secuencias y el número de secuencias que contienen dicho aminoácido —añadiendo uno al denominador para evitar divisiones por cero—. Finalmente, el valor de TF-IDF se obtiene multiplicando TF por IDF (Jurafsky & Martin, 2021). En la ecuación 1 se presenta la fórmula para calcular TF-IDF.

$$TF - IDF\ (\alpha) = \frac{Número\ de\ veces\ que\ aparece\ (\alpha)\ en\ la\ secuencia}{Número\ total\ de\ aminoácidos\ en\ la\ secuencia} \times \log\left(\frac{N}{n_a + 1}\right) \quad \textbf{(1)}$$

A pesar de que TF-IDF permite representar numéricamente una secuencia de aminoácidos, esta representación es puramente estadística: refleja qué tan frecuente o distintivo es un aminoácido dentro del conjunto de secuencias, pero no incorpora ninguna información sobre sus propiedades fisicoquímicas (polaridad, carga, volumen), que son las que realmente determinan la compatibilidad estructural entre cadenas pesadas y ligeras. Esta limitación es la principal razón por la que, en este trabajo, se comparó su desempeño frente a los factores Atchley y ProtVec (capítulo 4), como parte de la selección de la representación numérica más adecuada para el modelo final.

2.4.2. Factores Atchley

Los factores Atchley son cinco valores numéricos asignados a cada aminoácido, obtenidos mediante un análisis de componentes principales (PCA) sobre un conjunto amplio de descriptores fisicoquímicos, que resumen las propiedades más relevantes de cada aminoácido en un espacio numérico reducido (Atchley et al., 2005).

Estos cinco valores codifican propiedades fisicoquímicas fundamentales: polaridad, preferencia por estructura secundaria, volumen molecular, composición relativa y carga electrostática. Esto los hace adecuados para reducir la dimensionalidad en aplicaciones de aprendizaje automático. Entre sus ventajas destacan la capacidad para capturar propiedades relevantes con bajo costo computacional y su flexibilidad para usarse en diferentes técnicas analíticas. En la tabla 2.1 se muestran los cinco factores para cada uno de los aminoácidos.

Tabla 2.1. Factores Atchley de los aminoácidos (Atchley et al., 2005).

Aminoácido	Símbolo	Factor I	Factor II	Factor III	Factor IV	Factor V
Alanina	A	-0.591	-1.302	-0.733	1.570	-0.146
Cisteína	C	-1.343	0.465	-0.862	-1.020	-0.255
Ácido aspártico	D	1.050	0.302	-3.656	-0.259	-3.242
Ácido glutámico	E	1.357	-1.453	1.477	0.113	-0.837
Fenilalanina	F	-1.006	-0.590	1.891	-0.397	0.412
Glicina	G	-0.384	1.652	1.330	1.045	2.064
Histidina	H	0.336	-0.417	-1.673	-1.474	-0.078
Isoleucina	I	-1.239	-0.547	2.131	0.393	0.816
Lisina	K	1.831	-0.561	0.533	-0.277	1.648
Leucina	L	-1.019	-0.987	-1.505	1.266	-0.912
Metionina	M	-0.663	-1.524	2.219	-1.005	1.212
Asparagina	N	0.945	0.828	1.299	-0.169	0.933
Prolina	P	0.189	2.081	-1.628	0.421	-1.392
Glutamina	Q	0.931	-0.179	-3.005	-0.503	-1.853
Arginina	R	1.538	-0.055	1.502	0.440	2.897
Serina	S	-0.228	1.399	-4.760	0.670	-2.647
Treonina	T	-0.032	0.326	2.213	0.908	1.313
Valina	V	-1.337	-0.279	-0.544	1.242	-1.262
Triptófano	W	-0.595	0.009	0.672	-2.128	-0.184
Tirosina	Y	0.260	0.830	3.097	-0.838	1.512

A diferencia de representaciones puramente estadísticas como TF-IDF, que capturan la frecuencia con la que aparece un aminoácido, pero no su naturaleza fisicoquímica, los factores Atchley incorporan directamente propiedades biológicamente relevantes para la interacción entre cadenas —como la carga electrostática y la polaridad, determinantes en la complementariedad estructural VH-VL descrita en la sección anterior—. Esta propiedad es la razón por la que, en este trabajo, se comparó su desempeño frente a TF-IDF y ProtVec (capítulo 4), resultando en la representación seleccionada para el modelo final.

2.4.3. *Word embeddings*

Los *word embeddings* son una técnica de representación numérica, ampliamente utilizada en procesamiento de lenguaje natural (NLP), que transforma palabras o documentos completos en vectores que codifican tanto el significado semántico como las dependencias contextuales entre términos, facilitando el desarrollo de tareas NLP mediante el análisis lingüístico computacional (Selva Birunda & Kanniga Devi, 2021).

Esta traducción de palabras a formato matemático mejora la capacidad para analizar y comprender datos textuales, incrementando así el rendimiento de múltiples técnicas de aprendizaje automático. En bioinformática, se emplean para representar secuencias de ADN, ARN y proteínas, permitiendo capturar patrones evolutivos y estructurales. Algunos modelos como ProtVec utilizan embeddings para representar proteínas a partir de subsecuencias de aminoácidos, permitiendo a los algoritmos de aprendizaje considerar similitudes funcionales y estructurales entre proteínas (Selva Birunda & Kanniga Devi, 2021).

A diferencia de TF-IDF —que solo captura frecuencia relativa de aminoácidos— y de los factores Atchley —que codifican propiedades fisicoquímicas fijas para cada aminoácido de forma individual—, los word embeddings buscan capturar relaciones contextuales entre subsecuencias, aprendidas a partir de grandes cantidades de secuencias de proteínas. En este trabajo, esta capacidad se evalúa específicamente a través de ProtVec (sección 2.3.4), como parte de la comparación de representaciones numéricas del capítulo 4, en la que se contrasta frente a TF-IDF y factores Atchley para determinar cuál conserva mejor la información relevante para el problema de emparejamiento VH-VL.

2.4.4. ProtVec

Es un modelo preentrenado basado en *embeddings* que representa numéricamente las secuencias de proteínas mediante técnicas de aprendizaje no supervisado. Este método divide las secuencias en tripéptidos superpuestos, codificándolos en vectores continuos que capturan patrones de secuencia y relaciones contextuales.

ProtVec es particularmente útil para tareas de clasificación de proteínas, anotación funcional y predicción estructural, ya que considera dependencias tanto locales como globales. Una de sus características destacadas es su capacidad de codificar relaciones bidireccionales entre aminoácidos a lo largo de la secuencia, lo cual mejora el desempeño predictivo de los modelos de aprendizaje automático en bioinformática (Asgari & Mofrad, 2015).

ProtVec se utiliza con secuencias de aminoácidos de anticuerpos obtenidas mediante un proceso de secuenciación y almacenadas en un conjunto de datos. Estos datos sirven como entrada para ProtVec, el cual genera vectores numéricos que pueden ser utilizados por algoritmos de aprendizaje automático. La Figura 2.2, muestra el proceso de uso de ProtVec.

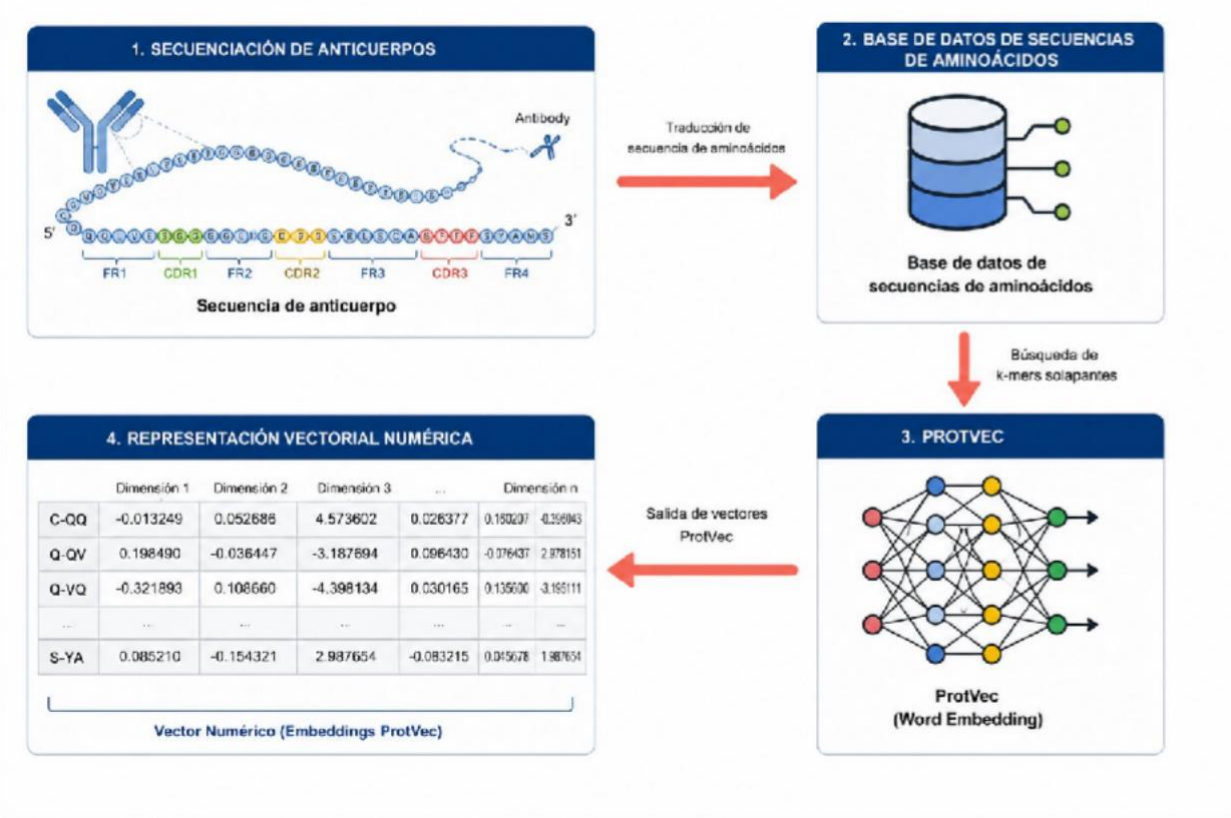

Figura 2.2. Ejemplo de cómo funciona ProtVec.

2.5. Técnicas de aprendizaje automático para la evaluación de representaciones de secuencias de aminoácidos

Con el objetivo de determinar qué representación numérica de las secuencias de aminoácidos captura de mejor manera la información relevante para la tarea de clasificación de anticuerpos, se emplearon distintas técnicas de aprendizaje automático como herramientas de evaluación comparativa. Estas técnicas permiten verificar la calidad de una representación a utilizando su desempeño en distintas tareas de clasificación y análisis exploratorio, proporcionando evidencia sobre la capacidad discriminativa de cada codificación evaluada.

Las técnicas utilizadas en esta investigación son árboles de decisión, regresión logística, máquinas de soporte vectorial (SVM), K-means, agrupamiento jerárquico y redes de Kohonen (SOM). Estas técnicas permiten evaluar las representaciones desde perspectivas complementarias: mientras los métodos de clasificación evalúan la forma de dividir las clases utilizando métricas como exactitud, precisión y F1, los métodos de agrupamiento revelan la estructura del espacio de representación y su coherencia biológica.

Es importante señalar que estas técnicas no constituyen el modelo propuesto en esta investigación, sino que se utilizan como línea base y como herramienta de análisis exploratorio previo al diseño del modelo final basado en una red neuronal siamesa asimétrica con pérdida InfoNCE. Los resultados obtenidos con estos métodos orientaron la selección

de la representación de factores Atchley como la codificación más adecuada para el problema de emparejamiento funcional entre cadenas pesadas y ligeras de anticuerpos.

2.5.1. Árboles de decisión

Los árboles de decisión son modelos predictivos de aprendizaje supervisado que representan un conjunto de reglas de clasificación en forma de estructura jerárquica ramificada, en la que cada nodo interno corresponde a una condición sobre un atributo, cada rama representa el resultado de dicha condición y cada nodo hoja corresponde a una clase de salida (Costa & Pedreira, 2023). Su amplio uso en aprendizaje automático se debe a su interpretabilidad, facilidad de implementación y capacidad para manejar tanto variables numéricas como categóricas sin requerir normalización previa de los datos.

Los principales algoritmos de inducción de árboles de decisión son ID3, C4.5 y CART. El algoritmo ID3 construye el árbol de manera recursiva seleccionando en cada nodo el atributo que maximiza la ganancia de información, calculada mediante la entropía de Shannon. C4.5 extiende a ID3 introduciendo la razón de ganancia como criterio de partición, lo que corrige el sesgo hacia atributos con muchos valores distintos, e incorpora mecanismos de poda para mejorar la generalización del modelo. Por su parte, CART utiliza el índice Gini como criterio de división y produce árboles binarios, siendo aplicable tanto a tareas de clasificación como de regresión (Mienye & Jere, 2024). Entre los hiperparámetros más relevantes se encuentra la profundidad máxima del árbol, cuyo control es fundamental para evitar el sobreajuste y favorecer la capacidad de generalización del modelo (Meng et al., 2023).

Para ejemplificar el funcionamiento de los árboles de decisión, considérese un conjunto de secuencias de aminoácidos representadas mediante sus propiedades fisicoquímicas. Un árbol de decisión, al que se le han enseñado estas representaciones, podría crear una regla en su primer nodo, como, por ejemplo: ¿El nivel promedio de hidrofobicidad de la secuencia es superior a 0.35? Las secuencias que cumplen esto irían por un camino, y las que no, por el otro. En los siguientes niveles, el árbol va añadiendo más criterios sobre otras propiedades, como la polaridad o la carga total, hasta que cada camino final asigna la secuencia a un grupo: un anticuerpo que identifica el SARS-CoV-2 o uno que reconoce otro tipo de agente infeccioso. Esta forma de dividir la información de manera repetitiva deja que el modelo capte las combinaciones de propiedades fisicoquímicas que definen a cada grupo, de una manera fácil de entender y de seguir gracias a las reglas que se van creando. En la figura 2.3, se muestra el árbol de decisión del ejemplo de clasificación de anticuerpos.

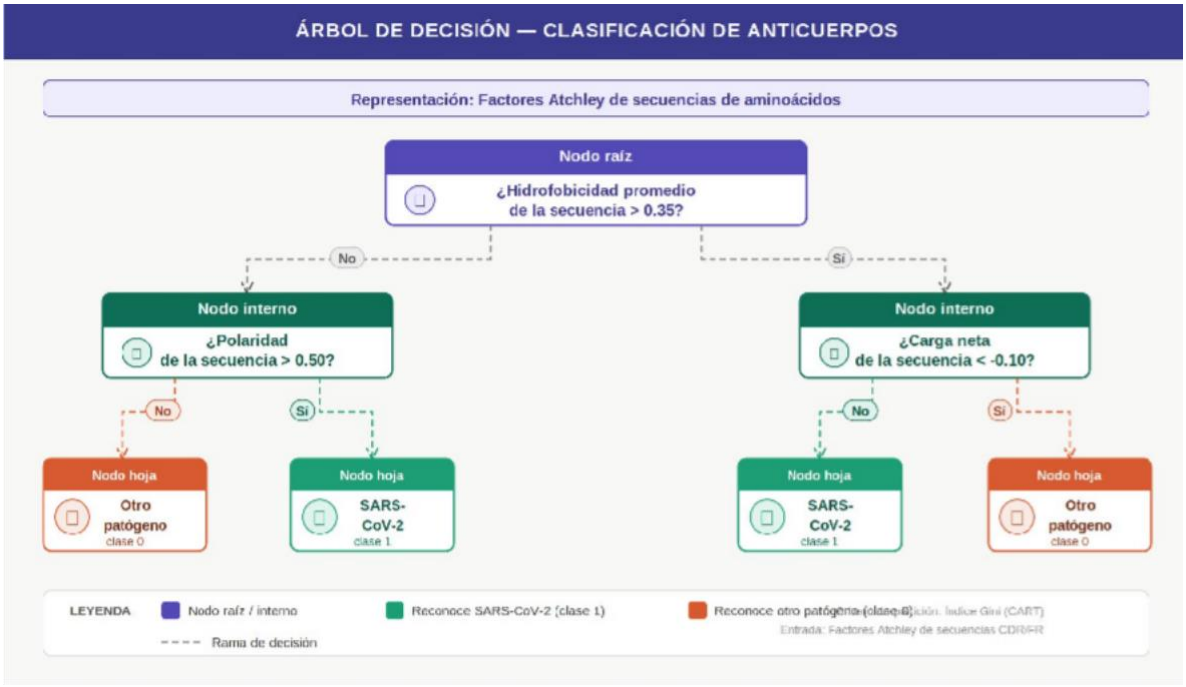

Figura 2.3. Árbol de decisión del ejemplo de clasificación de anticuerpos.

2.5.2. Regresión logística

La regresión logística es un modelo de clasificación supervisada que calcula la probabilidad de pertenencia de una instancia a una clase mediante una función logística aplicada a una combinación lineal de las variables de entrada (Suresh et al., 2024). Se trata de un algoritmo de clasificación y no de regresión como tal, es ampliamente utilizado en problemas de clasificación binaria por su simplicidad, interpretabilidad y sólida fundamentación estadística.

El modelo estima sus parámetros utilizando la maximización de la función de verosimilitud, optimizada a través de métodos como el descenso del gradiente o el método de newton-Raphson. La salida del modelo es una probabilidad entre 0 y 1, obtenida mediante la función sigmoide, que permite asignar cada instancia a la clase positiva o negativa según un umbral de decisión, típicamente establecido en 0.5 (Dey et al., 2025). La ecuación 2 muestra la función sigmoide se define como:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad \textbf{(2)}$$

Donde z representa la combinación lineal de las características de entrada con los pesos aprendidos por el modelo y e es la base del algoritmo natural, aproximadamente igual a 2.718. Cuando $\sigma(z) \geq 0.5$el modelo predice la clase positiva; en caso contrario, predice la clase negativa (Hu et al., 2025).

Para mostrar cómo funciona, tomemos una secuencia de aminoácidos codificada con los factores de Atchley, la cual genera un vector numérico de propiedades fisicoquímicas. La regresión logística luego toma una suma ponderada de esas propiedades, por ejemplo, dando

más importancia a la hidrofobicidad y la polaridad de las zonas CDR y le aplica la función sigmoide. Si la probabilidad que resulta es mayor que 0.5, la secuencia se considera una cadena pesada que encaja con una cadena ligera que reconoce el SARS-CoV-2; de lo contrario, se clasifica como negativa. Esta manera de operar linealmente hace que la regresión logística sea bastante útil como punto de partida para ver si una representación se puede separar de forma lineal en el espacio de características. En la figura 2.4 se muestra cómo funciona el algoritmo de regresión logística

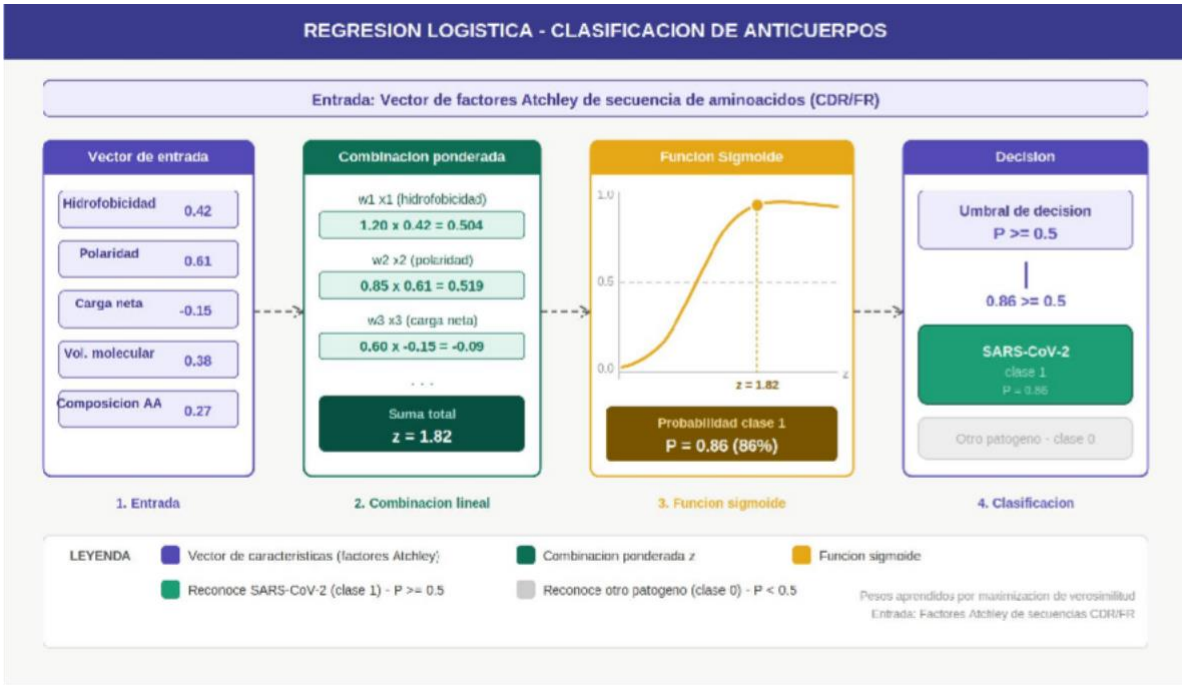

Figura 2.4. Ejemplo de regresión logística.

2.5.3. Máquinas de soporte vectorial

Las máquinas de soporte vectorial (SVM, por sus siglas en inglés) son algoritmos de aprendizaje supervisado que identifican el hiperplano de separación óptimo entre clases en un espacio de características, maximizando así el margen entre los puntos más cercanos de cada clase, conocidos como vectores de soporte (Cervantes et al., 2020). Las SVM sobresalen por su solidez en espacios de alta dimensión, su base teórica en la minimización del riesgo estructural y su aptitud para abordar problemas no linealmente separables a través del empleo de funciones de kernel.

El kernel permite mapear los datos de entrada a un espacio de mayor dimensión donde las clases se vuelven linealmente separables, sin requerir el cálculo de dicha transformación. Los kernels más empleados son el lineal, el polinomial y el de función de base radial (RBF); este último es el más frecuente en la práctica debido a su habilidad para modelar fronteras de decisión no lineales de manera eficiente (Rezvani et al., 2024). En escenarios de alta dimensionalidad, como el que exhiben las representaciones de secuencias de aminoácidos analizadas en este estudio, las SVM constituyen una alternativa particularmente apropiada,

dado que su complejidad de clasificación depende de la cantidad de vectores de soporte y no de la dimensionalidad del espacio de entrada (Kazemi, 2024).

Para mostrar cómo funciona, en un grupo de secuencias de aminoácidos que se representan con los factores Atchley en un plano. Las SVM encuentran la línea (hiperplano) que separa las secuencias de cadenas pesadas que encajan con los anticuerpos que detectan el SARS-CoV-2 de las que no, haciendo que la distancia entre los dos grupos sea lo más grande posible. Los puntos que están más cerca de esa línea, en cada grupo (los vectores de soporte), son los que definen dónde está y cómo está orientada esa línea divisoria. Si las representaciones no se pueden separar en línea recta, la función kernel RBF cambia el espacio de características a uno de más dimensiones donde sí se pueden separar, lo que deja que el modelo cree límites de decisión complicados sin que aumente mucho el esfuerzo computacional.

2.5.4. K-Means

K-Means es un algoritmo de clustering no supervisado que divide un conjunto de datos en k grupos, asignando cada instancia al clúster cuyo centroide esté más próximo, medido por distancia euclidiana (Ikotun et al., 2023). Al contrario que los métodos de clasificación supervisada, K-Means no necesita etiquetas de clase durante el entrenamiento, lo que lo hace particularmente útil para explorar la estructura inherente de los datos y determinar si las representaciones producen agrupaciones coherentes de manera natural.

El algoritmo funciona de forma iterativa en un par de pasos. En la primera fase, cada punto de datos se asigna al centroide más próximo; en la segunda, los centroides se recalculan como el promedio de todos los puntos asignados a cada grupo. Este procedimiento se repite hasta que las asignaciones se estabilizan o se llega a un límite superior de repeticiones (Zubair et al., 2024). Entre las principales restricciones del algoritmo figuran su susceptibilidad a la elección de los centroides iniciales y la exigencia de definir el número de grupos k antes de empezar. Con el fin de atenuar el primero de estos inconvenientes se emplea la versión K-Means++, que establece los centroides de manera distribuida para optimizar la convergencia. La identificación del valor más adecuado de k se respalda en mediciones como el coeficiente de silueta, el índice de Davies-Bouldin y el criterio del codo respecto a la inercia dentro del clúster (Ezugwu et al., 2022).

Para dar un ejemplo de cómo funciona, tomemos en cuenta un grupo de secuencias de cadenas pesadas de anticuerpos, codificadas con los factores de Atchley. El método K-Means organiza estas secuencias en k grupos basándose en su semejanza fisicoquímica, sin saber de antemano cuál es su clase real. Si la forma en que las representamos capta datos biológicamente importantes, las secuencias que encajan con las cadenas ligeras que identifican al SARS-CoV-2 se agruparán en clústeres diferentes a los de las secuencias que no encajan. Esto demuestra lo bien que la codificación que hemos usado puede diferenciar entre ellas. La figura 2.5, muestra el funcionamiento del algoritmo K-Means.

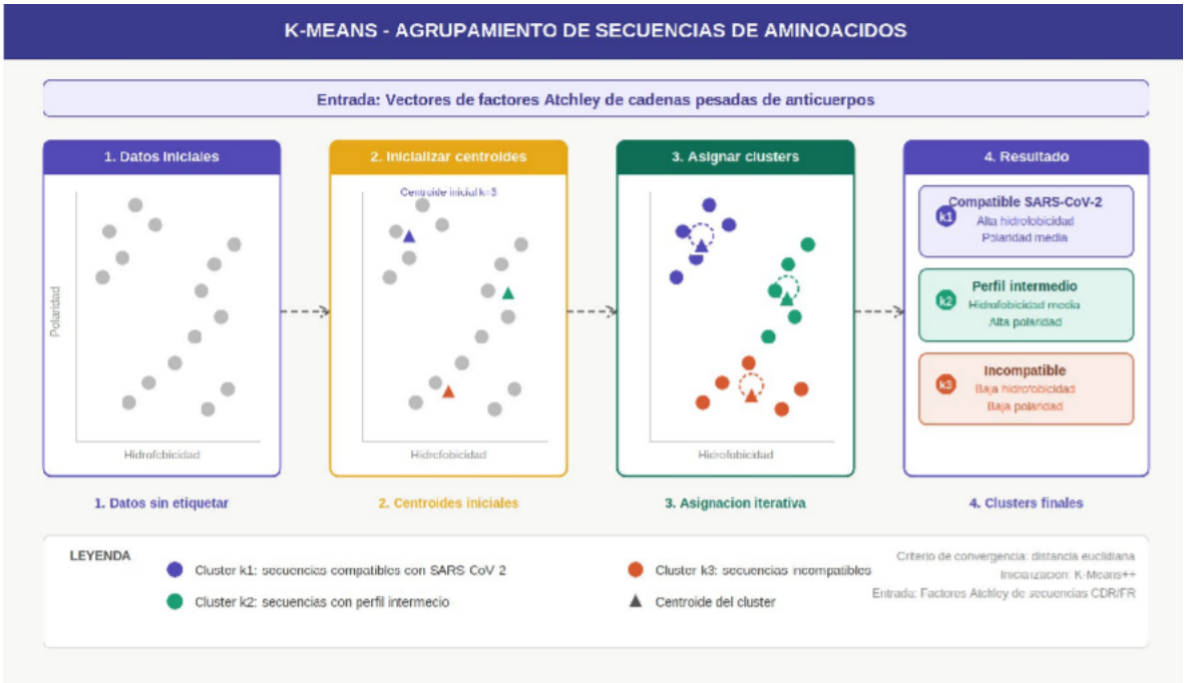

Figura 2.5. Ejemplo de K-Means.

2.5.5. Agrupamiento Jerárquico

El agrupamiento jerárquico se trata de una técnica de clustering que elabora una jerarquía de grupos, la cual se presenta a través de una estructura arbórea conocida como dendrograma. Este dendrograma facilita la observación de las relaciones de semejanza entre las instancias en distintos niveles de detalle, sin tener que predefinir el número de clústeres (Ran et al., 2023). Se pueden distinguir dos enfoques fundamentales: el divisivo, que inicia con un solo grupo abarcando todos los datos para luego subdividirlo de manera continua, y el aglomerativo, que comienza con tantos grupos como instancias hay y los va uniendo progresivamente según un criterio de similitud.

Dentro del método aglomerativo, la opción de enlace de Ward es una de las más empleadas en el campo, dado que en cada fase une los dos clústeres cuya combinación resulta en la menor adición a la varianza total dentro de los clústeres. Esto propende a la generación de grupos que son densos y de dimensiones parecidas (Dogan & Birant, 2022). La elección del tipo de enlace ya sea simple, completo, promedio o Ward influye de manera notable en la conformación y la calidad de los clústeres que se obtienen. Por ello, su selección debe basarse en la estructura que se anticipa en los datos (Randriamihamison et al., 2021). A diferencia de K-Means, el agrupamiento jerárquico no exige que se fije el valor de k de antemano; la cantidad de clústeres se define al seccionar el dendrograma a una altura específica, ya sea mediante un límite de distancia o un criterio de validación interna.

Para mostrar cómo funciona, en un grupo de secuencias de cadenas pesadas codificadas con los factores Atchley. El algoritmo aglomerativo empieza considerando cada secuencia como un grupo aparte y, en cada paso, junta los dos grupos que más se parecen, usando el criterio de Ward. El gráfico que se forma, llamado dendrograma, deja ver a simple vista en qué punto

de similitud se separan las secuencias que concuerdan con anticuerpos que detectan el SARS-CoV-2 de las que no, lo que da pistas sobre cómo está organizado el espacio de representación sin que haga falta saber de antemano qué secuencia pertenece a qué grupo. En la figura 2.6, se muestra el funcionamiento del agrupamiento jerárquico.

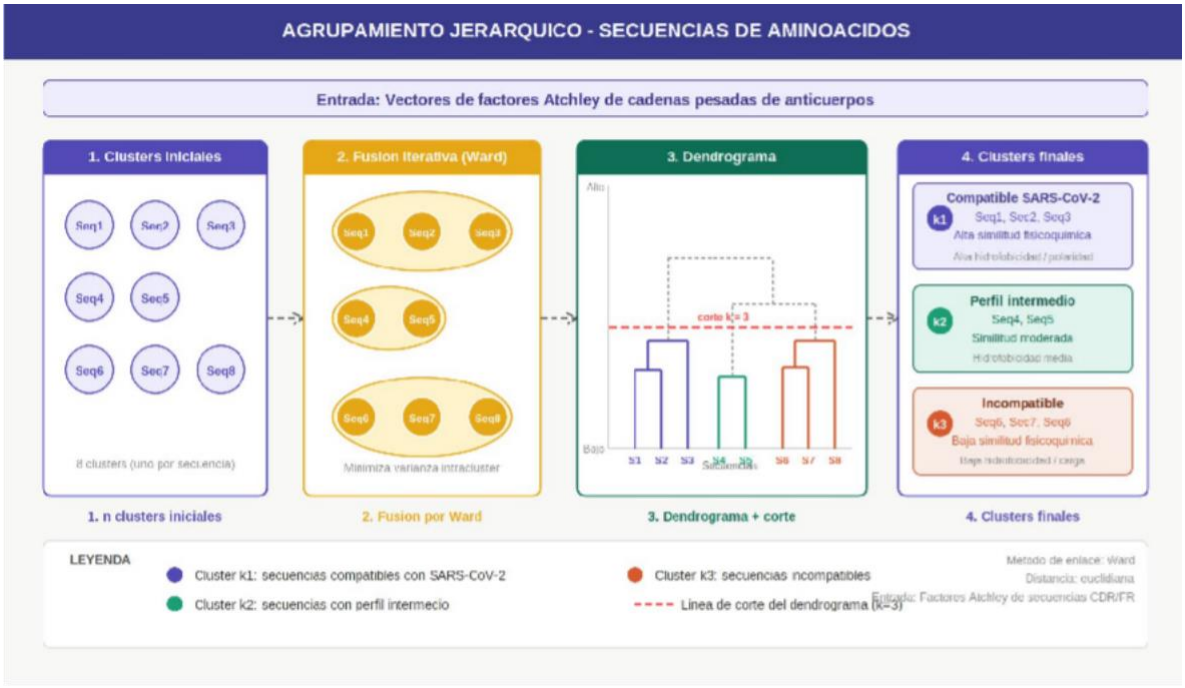

Figura 2.6. Ejemplo de agrupamiento jerárquico.

2.5.6. Redes de Kohonen

Las redes de Kohonen, también llamadas mapas autoorganizados (SOM, por sus siglas en inglés), constituyen un tipo de red neuronal artificial que aprende por medio de aprendizaje competitivo no supervisado para generar una representación discreta y de baja dimensión generalmente bidimensional del espacio de entrada, conservando las relaciones topológicas que se encuentran en los datos originales (Nogales et al., 2025). Esta cualidad de preservación topológica la diferencia de otros enfoques de reducción de dimensionalidad, puesto que las instancias parecidas en el espacio de alta dimensión son proyectadas a ubicaciones próximas en el mapa bidimensional.

El proceso de entrenamiento implica exponer repetidamente cada vector de entrada a la red, reconocer la neurona ganadora o unidad de mejor coincidencia (BMU, por sus siglas en inglés) como aquella cuyo vector de pesos es el más cercano al vector de entrada, y modificar los pesos de la BMU y de sus neuronas circundantes para aproximarlos al vector de entrada, disminuyendo gradualmente el radio de vecindad conforme avanza el entrenamiento (Tripathi, 2024). A diferencia de K-Means, el SOM no delimita clústeres de manera explícita, sino que organiza el espacio de entrada en áreas de activación, lo cual resulta particularmente beneficioso para detectar patrones ocultos y tener una visión de la distribución general de los datos en espacios de alta dimensión (Nogales et al., 2025).

Para ilustrar su funcionamiento, en un conjunto de secuencias de cadenas pesadas codificadas mediante los factores Atchley. El SOM proyecta estos vectores de alta dimensión sobre una cuadrícula bidimensional de neuronas, de modo que las secuencias con propiedades fisicoquímicas similares activan neuronas cercanas en el mapa. Las regiones del mapa que concentran secuencias compatibles con anticuerpos que reconocen el SARS-CoV-2 se distinguen visualmente de aquellas que agrupan secuencias incompatibles, permitiendo evaluar de forma intuitiva la capacidad discriminativa de la representación evaluada sin necesidad de entrenar un clasificador supervisado. La figura 2.7, muestra el funcionamiento de las redes de Kohonen.

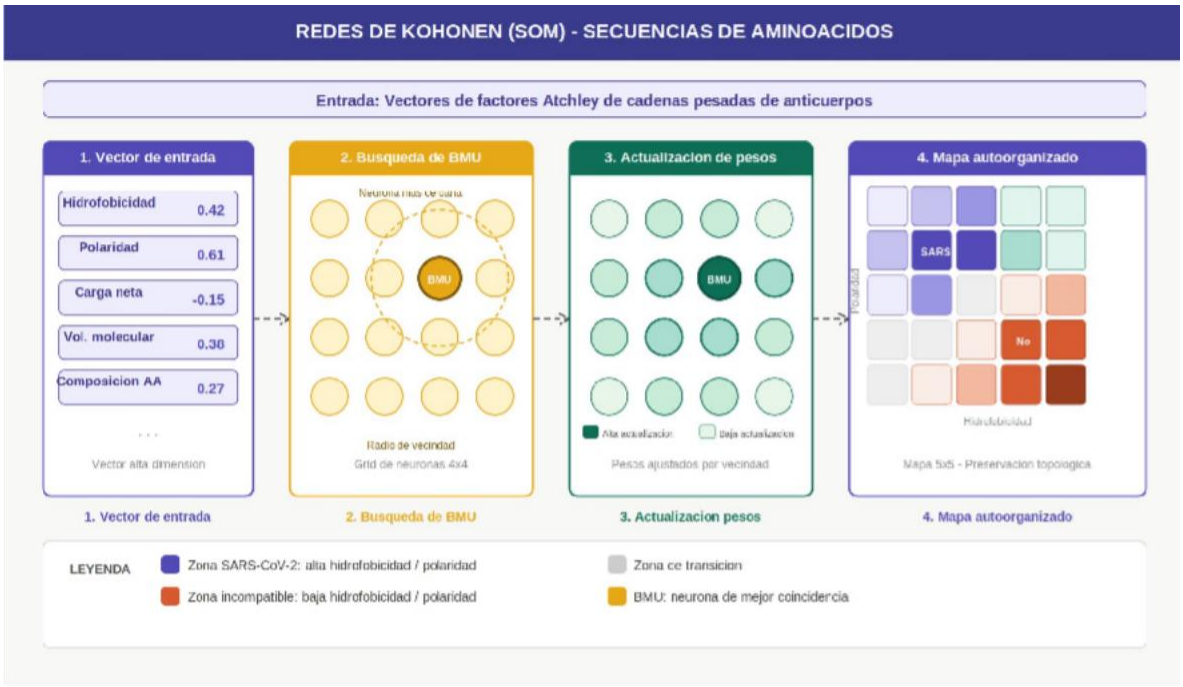

Figura 2.7. Ejemplo de redes de Kohonen.

2.6. Red siamesa asimétrica

Las redes siamesas son una arquitectura de aprendizaje profundo diseñada para comparar pares de entrada y aprender una representación que permita medir su similitud de manera eficaz. A diferencia de los modelos tradicionales que clasifican cada elemento por separado, una red siamesa aprende a medir relaciones entre pares de datos un espacio métrico común (Kaya & Bilge, 2019; Koch, Zemel, Salakhutdinov, et al., 2015).

Una red neuronal siamesa asimétrica es una extensión del modelo siamés en la que se utilizan dos *encoders* distintos en lugar de dos *encoders* idénticos. Mientras que una red siamesa clásica utiliza dos ramas con pesos compartidos para procesar entradas del mismo tipo, la versión asimétrica está diseñada para comparar dos elementos que provienen de dominios diferentes, permitiendo que cada rama aprenda representaciones especializadas para su propio tipo de dato (Kaya & Bilge, 2019).

En la red siamesa asimétrica, cada *encoder* transforma su entrada en un *embedding* numérico. Después, ambos *embeddings* se proyectan en un espacio métrico compartido en donde pueden ser comparados mediante una función de similitud, como la distancia coseno o la distancia euclidiana. El objetivo del entrenamiento es acercar las representaciones de pares compatibles y alejar las de pares incompatibles, lo que se logra mediante funciones de pérdida como *contrastive loss* o *triplet loss* (Jain et al., 2021).

Este tipo de arquitectura es especialmente útil cuando los dos elementos que se desean comparar no comparten la misma estructura o distribución, como ocurre en el caso del emparejamiento de cadenas pesadas y ligeras de anticuerpos, donde cada cadena posee propiedades secuenciales y fisicoquímicas distintas. La red siamesa dual permite que cada tipo de cadena sea codificada por un *encoder* adaptado a sus características, pero que ambos converjan en un espacio común donde la compatibilidad funcional pueda ser evaluada (Fierro et al., 2019).

2.7. Aprendizaje métrico

El aprendizaje métrico es un paradigma de aprendizaje automático en el que, en lugar de entrenar un modelo para predecir directamente una clase o una etiqueta, se entrena para aprender una función que transforme los datos de entrada en un espacio embebido (un espacio vectorial de menor dimensión) en el que la distancia entre dos representaciones refleje qué tan similares o compatibles son los elementos originales (Kaya & Bilge, 2019). Este enfoque es especialmente útil cuando el objetivo no es clasificar un elemento de forma aislada, sino comparar pares o conjuntos de elementos entre sí, como ocurre al evaluar la compatibilidad entre dos secuencias.

Para poder comparar dos representaciones dentro del espacio embebido, se utiliza una función de similitud. Las más comunes son la distancia coseno, que mide el ángulo entre dos vectores y es independiente de su magnitud, y la distancia euclidiana, que mide la distancia lineal entre ellos. La elección de la función de similitud determina cómo se interpreta la cercanía entre representaciones y, por lo tanto, afecta directamente la forma en que el modelo organiza el espacio embebido durante el entrenamiento (Kaya & Bilge, 2019).

2.7.1. Negativos verdaderos

Dentro del aprendizaje métrico, un negativo es cualquier ejemplo que se presenta al modelo como incompatible con el ancla durante el entrenamiento, y su calidad determina en gran medida qué tan bien aprende el modelo a discriminar (Robinson et al., 2021). En la literatura, los negativos se construyen comúnmente de dos formas:

- Negativos aleatorios o artificiales: se generan emparejando el ancla con cualquier otro elemento del conjunto de datos elegido al azar, sin verificar si esa combinación es realmente representativa de un caso incompatible en el dominio del problema. Por ejemplo, en tareas de recuperación de imágenes, un negativo aleatorio simplemente es cualquier imagen distinta a la que corresponde al ancla.

- Negativos verdaderos (o negativos biológicamente/semánticamente fundamentados): se seleccionan de forma que la incompatibilidad con el ancla esté respaldada por una razón concreta del dominio, y no solo por no ser el positivo. Esta distinción es relevante porque un negativo verdadero obliga al modelo a aprender fronteras de decisión más significativas, en lugar de aprender únicamente a distinguir el positivo de "cualquier otra cosa" (Robinson et al., 2021).

En esta investigación, los negativos verdaderos corresponden a las cadenas pesadas de anticuerpos de ratón BALB/c: a diferencia de un negativo aleatorio que podría ser, por ejemplo, cualquier otra cadena pesada humana no emparejada, una cadena de BALB/c es incompatible con una cadena ligera humana por razones biológicas concretas (divergencia en genes germinales, diferencias estructurales en el CDR3, ausencia de coevolución entre especies). El detalle completo de cómo se construyen y utilizan estos negativos se presenta en la sección de generación de tripletas del capítulo 5.

2.8. Métricas utilizadas para evaluar el modelo de emparejamiento de cadenas pesadas y ligeras de anticuerpos

La evaluación del desempeño del modelo siamés se basa en un conjunto de métricas para medir su capacidad discriminativa, su estabilidad de aprendizaje y su utilidad prácticas en tareas de emparejamiento. Dado que el modelo no solo debe de clasificar correctamente si un par de cadenas es compatible, sino también tiene que ordenar los candidatos preservando las distancias significativas en el espacio métrico, es necesario utilizar métricas que evalúen estos aspectos de rendimiento (Kaya & Bilge, 2019).

El entrenamiento de un modelo de aprendizaje métrico se guía mediante pérdidas contrastivas (contrastive losses), una familia de funciones de pérdida diseñadas específicamente para este paradigma: en lugar de penalizar una predicción incorrecta de clase, penalizan que representaciones de pares compatibles (positivos) queden lejos entre sí, o que representaciones de pares incompatibles (negativos) queden demasiado cerca. Dentro de esta familia se encuentran, entre otras, el *contrastive loss* —que funcionasobre pares individuales— y el *triplet loss* —que funciona sobre tripletas ancla-positivo-negativo, acercando el positivo al ancla y alejando el negativo simultáneamente— (Jain et al., 2021; Koch, Zemel, & Salakhutdinov, 2015).

En esta investigación, el modelo se entrena bajo este mismo paradigma: la red siamesa asimétrica descrita en la sección anterior actúa como la función que proyecta cada cadena al espacio embebido, la distancia coseno se utiliza como función de similitud entre cadenas pesadas y ligeras, y el entrenamiento se guía mediante InfoNCE —una pérdida contrastiva basada en tripletas, dentro de la familia de triplet loss, pero que en lugar de comparar contra un único negativo por iteración, contrasta el positivo contra todos los negativos disponibles en el lote de entrenamiento (negativos in-batch)—. Esta elección es la que permite que, en el espacio embebido final, las combinaciones de cadenas que forman un anticuerpo funcional queden cercanas, y las combinaciones biológicamente incompatibles —incluyendo los

negativos verdaderos de ratón BALB/c— queden alejadas. El detalle completo de esta implementación se presenta en el capítulo 5.

2.8.1. Métricas de clasificación

Estas métricas permiten evaluar cuantas veces el modelo clasifica correctamente los pares como compatibles o incompatibles.

El *Accuracy* representa la proporción de predicciones correctas. La ecuación 3, muestra cómo calcularlo.

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN} \quad \textbf{(3)}$$

La Sensibilidad (*Recall*) indica la capacidad del modelo para identificar correctamente los pares compatibles. La ecuación 4 muestra cómo calcularlo.

$$Recall = \frac{VP}{VP + FN} \quad \textbf{(4)}$$

El F1- *Score* combina la precisión y sensibilidad en una sola métrica equilibrada. La ecuación 5, muestra cómo calcularlo.

$$F1 - Score = 2 \times \frac{Precisión \times Sensibilidad}{Precisión + Sensibilidad} \quad \textbf{(5)}$$

La curva ROC (Receiver Operating Characteristic) y su área bajo la curva, ROC-AUC (Area Under the Curve), son una métrica global que evalúa la capacidad del modelo para separar la clase positiva de la negativa a lo largo de todos los posibles umbrales de decisión, y no solo en el umbral óptimo utilizado por Accuracy, Recall y F1-Score (Naidu et al., 2023). La curva ROC se construye graficando la tasa de verdaderos positivos (Recall) contra la tasa de falsos positivos para cada umbral posible; el área bajo esa curva resume, en un único valor entre 0 y 1, qué tan bien separa el modelo ambas clases —un valor de 1 indica separación perfecta, mientras que un valor de 0.5 equivale a una clasificación al azar—. La ecuación 6 muestra la forma de calcularlo.

$$FPR = \frac{FP}{FP + VN} \quad \textbf{(6)}$$

En esta investigación, el ROC-AUC es particularmente relevante porque, a diferencia de Accuracy o F1-Score, no depende de haber fijado previamente un único umbral de distancia coseno —el mismo umbral que se determina en la evaluación binaria (capítulo 6)—. Esto permite verificar que la separación entre pares compatibles e incompatibles en el espacio embebido es robusta en general, y no solo alrededor del punto de corte específico elegido para las demás métricas.

2.8.2. *Recall@*K

Esta métrica se utiliza para evaluar la capacidad del modelo de recuperación, como un modelo siamés, para encontrar el elemento correcto dentro de los primeros K resultados más similares según la distancia aprendida en el espacio métrico (Kaya & Bilge, 2019).

A diferencia de las métricas tradicionales de clasificación, *Recall@*K no evalúa si el modelo asigna una etiqueta correcta, sino su habilidad para ordenar candidatos según su similitud con un elemento dado (Wang et al., 2020).

Para cada ejemplo, el modelo genera una lista ordenada de los posibles emparejamientos, desde el más similar hasta el menos similar. *Recall@*K verifica si el emparejamiento aparece dentro de los primeros K lugares de esa lista. La ecuación 6 muestra cómo calcularlo.

$$Recall@K = \frac{1}{N} \sum_{i=1}^{N} 1\{y_i \in Top - K(i)\} \quad \textbf{(6)}$$

Donde:

- N: número total de ejemplos evaluados,
- y: emparejamiento correcto para el ejemplo i
- Top – K(i): los K candidatos más similares según el modelo
- $1\{\}$: función indicadora (vale 1 si la condición se cumple y 0 si no)

2.8.3. *Espacio de embeddings y UMAP*

Además de las métricas cuantitativas, en un modelo de aprendizaje métrico resulta útil inspeccionar de forma cualitativa el espacio de *embeddings* que el modelo aprendió durante el entrenamiento, para verificar si su organización tiene sentido más allá de lo que capturan las métricas numéricas. Sin embargo, este espacio suele tener una dimensionalidad demasiado alta para observarse directamente (por ejemplo, 8 dimensiones), por lo que se recurre a técnicas de reducción de dimensionalidad que proyecten los *embeddings* a un plano de 2 o 3 dimensiones, conservando en la medida de lo posible las relaciones de cercanía y lejanía del espacio original.

UMAP (del inglés, *Uniform Manifold Approximation and Projection*) es una de las técnicas más utilizadas para este fin. A diferencia de métodos lineales como PCA, UMAP preserva tanto la estructura local (qué puntos son vecinos cercanos entre sí) como buena parte de la estructura global del espacio original, lo que la hace especialmente adecuada para visualizar agrupamientos y patrones en datos biológicos de alta dimensionalidad (Diaz-Papkovich et al., 2021).

En esta investigación, UMAP se utiliza para proyectar en dos dimensiones el espacio de *embeddings* de 8 dimensiones aprendido por la red siamesa asimétrica, coloreando los puntos proyectados según propiedades biológicas conocidas (familia germinal, gen V, longitud del CDR3) y según su tipo de par (nativo o desplazado). Esto permite verificar, de forma visual, si el modelo organizó el espacio de una manera biológicamente coherente, sin que esas

propiedades hayan sido parte explícita del entrenamiento. El detalle de este análisis se presenta en el capítulo 6.

2.9. Conclusión

A lo largo de este capítulo se construyó el sustento teórico necesario para abordar el problema de investigación, conectando tres niveles: la biología del anticuerpo, es decir, qué determina la compatibilidad VH-VL; la bioinformática, es decir, cómo traducir esa biología en información procesable; y las herramientas computacionales capaces de aprender esa compatibilidad a partir de dicha información.

No todos los conceptos discutidos forman parte del modelo final: TF-IDF y los *word embeddings* (evaluados mediante ProtVec), junto con las técnicas de aprendizaje automático más clásicas (árboles de decisión, regresión logística, máquinas de soporte vectorial, K-Means, agrupamiento jerárquico y redes de Kohonen), se utilizaron únicamente durante la experimentación (capítulo 4) para comparar el desempeño de distintas representaciones numéricas, y quedaron descartados frente a los factores Atchley, que resultaron la representación con mejor desempeño y la seleccionada para el modelo final.

El modelo propuesto se construye, en cambio, a partir de la segunda mitad del capítulo: una red siamesa asimétrica con dos *encoders* independientes, enriquecida con un módulo de *cross-attention* bidireccional, entrenada bajo el paradigma de aprendizaje métrico mediante la función de pérdida InfoNCE y una estrategia de negativos verdaderos (cadenas pesadas de ratón BALB/c) como aportación central de este trabajo. Su desempeño se evalúa combinando métricas cuantitativas de clasificación y recuperación (*Accuracy*, *Recall*, F1-Score, ROC-AUC, Recall@K) con el análisis cualitativo del espacio de *embeddings* mediante UMAP, tal como se presenta en el capítulo 6.

Con este marco conceptual establecido, el siguiente capítulo revisa el estado del arte relacionado con el problema de emparejamiento VH-VL, antes de presentar el desarrollo metodológico completo del modelo propuesto.

Capítulo 3. Análisis del estado del arte

3. Análisis del estado del arte

Este capítulo presenta las investigaciones que contribuyeron al desarrollo del modelo predictivo para el emparejamiento de cadenas pesadas y ligeras de anticuerpo. Algunas de estas investigaciones se relacionan con métodos de representación numérica de secuencias de aminoácidos, mientras que otras se centran en el diseño de arquitecturas de aprendizaje profundo para tareas de similitud y emparejamiento en datos biológicos. Además, se identificaron estudios que se consideraron esenciales para un análisis más profundo, los cuales sirvieron como referencia metodológica para el diseño del modelo propuesto.

Con el objetivo de organizar las investigaciones revisadas, los artículos del estado del arte fueron clasificados según el aporte que tienen para la presente investigación. Esta clasificación permitió diferenciar entre los trabajos enfocados en la representación de las secuencias proteicas, investigaciones centradas en arquitecturas de aprendizaje métrico y modelos de similitud, y aquellos trabajos relacionados con el análisis computacional y el emparejamiento de secuencias de anticuerpos.

La tabla 3.1 muestra los tres clasificadores utilizados para dividir los trabajos del estado del arte.

Tabla 3.1. Criterios de clasificación de los artículos del estado del arte.

Clasificación	Criterio de inclusión	Propósito dentro de la investigación
Métodos de representación numérica de secuencias de aminoácidos	Estudios que proponen o evalúan métodos para transformar secuencias de proteínas en vectores numéricos, incluyendo enfoques estadísticos, fisicoquímicos y basados en aprendizaje profundo	Identificar las representaciones más adecuadas para capturar las propiedades estructurales y funcionales de las cadenas pesadas y ligeras de los anticuerpos
Arquitecturas de aprendizaje profundo para tareas de similitud y emparejamiento	Estudios que utilizan redes siamesas, aprendizaje métrico o modelos contrastivos para comparar pares de entidades en dominios biológicos o de otro tipo	Analizar arquitecturas, funciones de perdida y estrategias de entrenamiento aplicables al emparejamiento de cadenas de anticuerpos
Modelos computacionales aplicados al análisis y emparejamiento de cadenas de anticuerpos	Trabajos que tienen el objetivo de predecir, clasificar o evaluar la compatibilidad entre las cadenas pesadas y ligeras, o que desarrollan herramientas computacionales para el análisis de repertorios de anticuerpos	Servir como referencia para evaluar el desempeño del modelo propuesto e identifican las necesidades que esta investigación busca atender

3.1. Métodos de representación numérica de secuencias de aminoácidos

Esta clasificación permitió identificar los enfoques más relevantes para realizar la transformación de las secuencias proteicas en representaciones numéricas que puedan ser utilizadas por los modelos de aprendizaje automático. A partir de estas investigaciones se analizaron distintas estrategias utilizadas para codificar propiedades fisicoquímicas, estadísticas y contextuales de las cadenas de aminoácidos, así como sus ventajas y limitaciones en diferentes contextos de aplicación. Los artículos que integran esta clasificación se muestran en la Tabla 3.2.

Tabla 3.2. Artículos sobre métodos de representación numérica de secuencias de aminoácidos.

ID	Título	Autores
1	*Solving* the *protein sequence metric problem*	(Atchley et al., 2005)
2	Amino *Acid Encoding Methods* for *Protein Sequences*: A *Comprehensive Review* and *Assessment*	(Jing et al., 2020)
3	Amino *acid encoding* for *deep learning applications*	(ElAbd et al., 2020)
4	*Biological structure* and *function* emerge from *scaling unsupervised learning* to 250 *million protein sequences* (ESM-1b)	(Rives et al., 2021)
5	ProtTrans: *Toward understanding* the *language* of life *through* self-*supervised learning*	(Elnaggar et al., 2022)
6	*Deciphering* the *language* of *antibodies using* self-*supervised learning* (AntiBERTa)	(Leem et al., 2022)
7	AbLang: An *antibody language model* for *completing antibody sequences*	(Olsen et al., 2022)
8	*Evolutionary-scale prediction* of *atomic-level protein structure* with a *language model* (ESM-2)	(Lin et al., 2023)
9	IgLM: *Infilling language modeling* for *antibody sequence design*	(Shuai et al., 2023)

A continuación, se presenta una descripción y un análisis de estas investigaciones, organizadas en cuatro categorías metodológicas: representaciones fisicoquímicas clásicas, revisiones comparativas de métodos de codificación, modelos de lenguaje para proteínas y modelos de lenguaje específicos para anticuerpos.

3.1.1. Representaciones fisicoquímicas clásicas

Atchley et al. (2005) propusieron una representación numérica compacta de los aminoácidos basada en el análisis de componentes principales (PCA) aplicado sobre 500 descriptores fisicoquímicos y estructurales de los aminoácidos. Como resultado se obtuvieron cinco factores numéricos para cada uno de los aminoácidos que resumen de manera sintética sus propiedades fundamentales: polaridad, preferencia por estructura secundaria, volumen molecular, diversidad de codones y carga electrostática. A diferencia de enfoques como la

codificación one-hot, que trata a los aminoácidos categorías independientes sin relación entre ellos, los factores Atchley codifican similitudes y diferencias fisicoquímicas en un espacio continuo de baja dimensionalidad, con una correspondencia directa entre cada factor y una propiedad biológica medible e interpretable. Esta característica los hace especialmente adecuados para tareas donde la compatibilidad funcional entre regiones CDR depende de propiedades como la carga y la hidrofobicidad, como ocurre en el emparejamiento VH-VL (Atchley et al., 2005).

3.1.2. Revisiones comparativas de métodos de codificación

Jing et al. (2020) presentaron una revisión sistemática y comparativa exhaustiva de métodos de codificación de aminoácidos para tareas de predicción de propiedades proteicas. Los autores clasificaron 16 métodos representativos en cinco categorías: codificación binaria (one-hot), propiedades fisicoquímicas (entre ellas los factores Atchley), métodos evolutivos (como PSSM), métodos basados en estructura y métodos de aprendizaje automático. La evaluación experimental sobre predicción de estructura secundaria y reconocimiento de plegamientos mostró que los métodos evolutivos basados en alineamientos múltiples (PSSM) alcanzan el mejor desempeño general, aunque los métodos de aprendizaje automático muestran potencial creciente. Esta revisión es relevante para la presente investigación porque proporciona el marco comparativo que justifica la elección de una representación fisicoquímica como los factores Atchley en contextos donde no se dispone de datos evolutivos suficientes o donde la interpretabilidad es prioritaria (Jing et al., 2020).

ElAbd et al. (2020) llevaron a cabo una comparación empírica directa entre diversos esquemas clásicos de codificación de aminoácidos incluyendo one-hot, BLOSUM62, VHSE8 y *embeddings* que son aprendidos de manera end-to-end dentro del contexto de modelos de aprendizaje profundo para la predicción de afinidad de péptidos a proteínas HLA. Los hallazgos mostraron que un *embedding* aprendido que cuenta únicamente con 4 dimensiones alcanza un rendimiento similar al de una codificación clásica de 20 dimensiones, lo que demuestra que los *embeddings* aprendidos son más eficientes en términos de dimensionalidad. Sin embargo, los autores señalaron que los esquemas clásicos son capaces de captar conocimiento previo, lo cual puede ser positivo en situaciones donde los datos de entrenamiento son limitados, un descubrimiento que resulta directamente pertinente para la elección de la representación en conjuntos de datos de anticuerpos de tamaño moderado, como el utilizado en este estudio (ElAbd et al., 2020).

3.1.3. Modelos de lenguaje para proteínas

Rives et al. (2021) presentan ESM-1b, un modelo de lenguaje *transformer* con aproximadamente 650 millones de parámetros, entrenado de forma no supervisada sobre 250 millones de secuencias proteicas evolutivamente diversas. A diferencia de los modelos LSTM previos, ESM-1b procesa secuencias completas de manera bidireccional mediante el objetivo de modelado de lenguaje enmascarado, capturando dependencias de largo alcance entre aminoácidos. Un hallazgo clave es que las representaciones aprendidas organizan las proteínas a múltiples escalas, desde propiedades bioquímicas de aminoácidos individuales hasta homología distante entre proteínas. ESM-1b consigue predicciones de estructura

secundaria comparables a los perfiles HMM procedentes de alineamientos múltiples, pero sin necesidad de estos alineamientos durante la inferencia. Este trabajo estableció que los modelos de lenguaje *transformer* de gran escala son transferibles al dominio biológico y generalizables a regiones no vistas del espacio de secuencias (Rives et al., 2021).

Elnaggar et al. (2022) presentan ProtTrans, una familia de modelos de lenguaje proteico preentrenados con aprendizaje supervisado sin etiquetas a partir de datos UniRef y BFD, que contienen hasta 393 mil millones de aminoácidos. Los autores entrenaron seis arquitecturas distintas incluyendo variantes autorregresivas (*Transformer*-XL, XLNet) y autocodificadoras (BERT, Albert, Electra, T5) en supercomputadoras con hasta 5,616 GPUs. Los *embeddings* obtenidos sin necesidad de etiquetas sí capturan propiedades biofísicas de las secuencias, y el modelo ProtT5 alcanza resultados comparables al estado del arte en predicción de estructura secundaria, localización subcelular y anotación funcional, superando a ESM-1b en varias métricas gracias a una mayor diversidad y escala de datos. ProtTrans mostró que la arquitectura elegida influye notablemente en la calidad de las representaciones aprendidas (Elnaggar et al., 2022).

Lin et al. (2023) presentaron ESM-2, una familia de modelos *transformer* entrenados a escala evolutiva sobre secuencias de UniRef, con tamaños que van de 8 millones a 15 mil millones de parámetros. Un hallazgo central es que, a pesar de entrenarse exclusivamente sobre secuencias, los modelos desarrollan representaciones internas que codifican información sobre estructura tridimensional, la cual emerge de manera implícita conforme aumenta la escala del modelo. ESM-2 supera consistentemente a ESM-1b en igualdad de parámetros y sirve de base para ESMFold, que predice estructuras a nivel atómico directamente desde secuencias individuales sin necesidad de alineamientos múltiples. ESM-2 representa el estado del arte en representaciones generales de proteínas y es ampliamente empleado como *encoder* de referencia en tareas de bioinformática computacional, incluyendo el modelado de anticuerpos (Lin et al., 2023).

3.1.4. Modelos de lenguaje específicos de anticuerpos

Leem et al. (2022) desarrollaron AntiBERTa, un modelo de lenguaje *transformer* preentrenado con 12 capas específicamente en 57 millones de secuencias de receptores de células B (BCR) humanos 42 millones de cadenas pesadas y 15 millones de cadenas ligeras que fueron obtenidas de la base de datos *Observed Antibody Space* (OAS). Al contrario que modelos generales como ESM-1b o ProtTrans, AntiBERTa captura los patrones específicos generados por la recombinación V(D)J y la hipermutación somática que definen el repertorio inmune, propiedades que los modelos generales de proteínas no modelan explícitamente. Como caso de uso, los autores ajustaron el modelo para predecir posiciones del paratopo a partir de la secuencia, superando herramientas del estado del arte en múltiples métricas. AntiBERTa fue el primer modelo de lenguaje profundo específico de anticuerpos y sentó las bases para versiones más avanzadas (Leem et al., 2022).

Olsen et al. (2022) presentaron AbLang, un modelo de lenguaje específico para anticuerpos, entrenado por separado para cadenas pesadas y ligeras, con base en las secuencias de la base de datos OAS. Su aplicación más importante es la restauración de los residuos perdidos en

las secuencias de anticuerpos, problema habitual en la secuenciación de los repertorios BCR: más del 40% de las secuencias en OAS carecen de los primeros 15 residuos del extremo N-terminal. AbLang supera tanto a las líneas germinales IMGT como al modelo general ESM-1b en esta tarea, con precisiones de recuperación del 96% y 98% para los primeros 15 residuos de cadenas ligeras y pesadas respectivamente, y siete veces más rápido que ESM-1b. AbLang ilustra el valor de los modelos específicos de dominio frente a los generales cuando la tarea involucra características particulares del repertorio inmune, como la variabilidad en los extremos N-terminales producto de limitaciones experimentales de secuenciación (Olsen et al., 2022).

Shuai et al. (2023) presentaron IgLM (*Immunoglobulin Language Model*), un modelo generativo de lenguaje entrenado sobre 558 millones de secuencias variables de cadenas pesadas y ligeras de anticuerpos de OAS, condicionado al tipo de cadena y la especie de origen. A diferencia de los modelos anteriores orientados a generar representaciones fijas, IgLM formula el diseño de anticuerpos como una tarea de *infilling* de texto, lo que permite rediseñar tramos de longitud variable dentro de secuencias existentes utilizando contexto bidireccional. Las secuencias generadas exhiben propiedades biofísicas favorables y se parecen a anticuerpos humanos naturales en distribución de longitud y composición de CDRs. IgLM representa una extensión del paradigma de representación hacia el diseño activo de secuencias, ampliando el espectro de aplicaciones de los modelos de lenguaje de anticuerpos más allá de la predicción (Shuai et al., 2023).

En la Tabla 3.3 se resumen estas investigaciones, destacando el tipo de representación, la escala de datos, el dominio de aplicación y las tareas evaluadas.

Tabla 3.3. Comparación de métodos de representación numérica de secuencias de aminoácidos

Autores	Método	Tipo	Datos de entrenamiento	Dominio	Tarea evaluada
Atchley et al.	Factores Atchley	Fisicoquímica (PCA)	~500 descriptores de 20 aminoácidos	Proteínas generales	Codificación y análisis filogenético
Jing et al.	Revisión comparativa	16 métodos (5 categorías)	*Benchmarks* de estructura y plegamiento	Proteínas generales	Estructura secundaria
Dallago et al.	Comparación empírica	One-hot, BLOSUM62, VHSE8, *embeddings*	Péptidos HLA-DRB1	Péptidos y proteínas	Predicción afinidad péptido-HLA
Rives et al.	ESM-1b	*Transformer* no supervisado	250M secuencias (UniRef50)	Proteínas generales	Estructura secundaria

Autores	Método	Tipo	Datos de entrenamiento	Dominio	Tarea evaluada
Elnaggar et al.	ProtTrans	*Transformer* (BERT, T5, XLNet)	393 mil millones aa (UniRef + BFD)	Proteínas generales	Estructura secundaria
Leem et al.	AntiBERTa	*Transformer* BCR-específico	57M secuencias BCR humanas (OAS)	Anticuerpos	Predicción de paratopo
Olsen et al.	AbLang	*Transformer* cadena-específico	Secuencias OAS (HC y LC)	Anticuerpos	Restauración de residuos faltantes
Lin et al.	ESM-2	*Transformer* evolutivo	250M secuencias (UniRef)	Proteínas generales	Predicción de estructura atómica
Shuai et al.	IgLM	Modelo generativo autoregresivo	558M secuencias (OAS)	Anticuerpos	Generación y rediseño de CDRs
Barton, Galson & Leem	AntiBERTa2	*Transformer* + aprendizaje contrastivo	779.4M secuencias humanas (OAS)	Anticuerpos	Predicción de afinidad de unión

3.1.5. Análisis de los métodos de representación numérica

Los trabajos revisados muestran una evolución clara: desde representaciones fisicoquímicas ligeras e interpretables, pasando por revisiones comparativas que sistematizan decenas de esquemas de codificación, hasta modelos de lenguaje de gran escala entrenados sobre cientos de millones de secuencias, tanto generales de proteínas como específicos de anticuerpos. Esta línea de trabajo ha resuelto, en gran medida, el problema de obtener representaciones numéricas de alta calidad que capturan propiedades estructurales y evolutivas de las secuencias, incluso sin necesidad de alineamientos múltiples.

Sin embargo, persisten dos limitaciones relevantes para el problema de esta tesis. Primero, los modelos de lenguaje de mayor desempeño requieren volúmenes de datos y cómputo considerablemente mayores a los disponibles en este trabajo, y están optimizados como codificadores de propósito general, para tareas como predicción de estructura, restauración de residuos o generación de secuencias, no para modelar directamente la compatibilidad entre dos cadenas distintas, que es el problema central de este trabajo. Segundo, ninguno de los estudios revisados compara de forma sistemática representaciones ligeras interpretables

(como Atchley) contra representaciones aprendidas (*word embeddings*, modelos de lenguaje) específicamente en el contexto del emparejamiento VH-VL; la elección de representación en la literatura suele depender de la tarea original de cada modelo, no evaluarse empíricamente para este problema particular.

Esto justifica dos decisiones de este trabajo: primero, adoptar los factores Atchley como representación base, dado que las codificaciones clásicas retienen conocimiento previo valioso en escenarios de datos moderados, como el conjunto de anticuerpos utilizado aquí, evitando además el costo computacional de los modelos de lenguaje de gran escala; y segundo, no asumir esta elección de forma directa, sino compararla empíricamente contra TF-IDF y ProtVec (capítulo 4), aportando evidencia experimental específica para el problema de emparejamiento VH-VL que no se encuentra en la literatura revisada.

3.2. Arquitecturas de aprendizaje profundo para tareas de similitud y emparejamiento

Esta clasificación agrupa aquellas investigaciones que proponen o analizan arquitecturas diseñadas para aprender una función de similitud entre pares de entidades, así como las funciones de pérdida contrastivas que guían dicho aprendizaje. A partir de estos trabajos se identificaron los principios arquitectónicos y las estrategias de entrenamiento aplicables al problema de emparejamiento de cadenas de anticuerpos. Los artículos que pertenecen a esta clasificación se muestran en la Tabla 3.4.

Tabla 3.4. Artículos sobre arquitecturas de aprendizaje profundo para tareas de similitud y emparejamiento

ID	Título	Autores
11	Siamese neural *networks* for one-*shot image recognition*	(Koch, Zemel, & Salakhutdinov, 2015)
12	*Representation Learning* with *Contrastive Predictive Coding*	(van den Oord et al., 2018)
13	*Deep Metric Learning*: A *Survey*	(Kaya & Bilge, 2019)
14	SiameseCPP: a *sequence-based* Siamese *network* to *predict cell-penetrating peptides* by *contrastive learning*	Zhang et al. (Zhang et al., 2023)

A continuación, se describen y analizan estas investigaciones, organizadas en dos grupos: fundamentos de las arquitecturas siamesas y aprendizaje métrico, y funciones de pérdida contrastivas para el aprendizaje de espacios métricos.

3.2.1. Fundamentos de las arquitecturas siamesas y aprendizaje métrico profundo

Koch, Zemel y Salakhutdinov (2015) introdujeron las redes neuronales siamesas como arquitectura para el aprendizaje en escenarios de una sola muestra (one-*shot learning*), donde el modelo debe reconocer nuevas clases a partir de un único ejemplo por clase. La arquitectura está formada por dos redes gemelas que comparten pesos y parámetros, procesando cada una de las dos entradas del par; sus salidas se comparan mediante una función de energía que mide la distancia entre los vectores de representación aprendidos. El modelo se entrena para minimizar la distancia entre representaciones de pares que pertenecen a la misma clase y maximizarla entre pares de clases diferentes, utilizando una función de pérdida contrastiva Los autores demostraron que una red siamesa convolucional entrenada para verificación de pares generaliza con éxito a la clasificación one-*shot* de clases no vistas durante el entrenamiento, sin necesidad de reentrenamiento. Este trabajo estableció la arquitectura siamesa como el paradigma de referencia para el aprendizaje de similitud a partir de pares, con aplicaciones que se extienden desde reconocimiento de imágenes hasta verificación de secuencias biológicas (Koch, Zemel, & Salakhutdinov, 2015).

Kaya y Bilge (2019) realizaron una revisión sistemática y comparativa del campo del aprendizaje métrico profundo (*deep metric learning*), cubriendo las tres dimensiones que determinan su desempeño: estrategia de muestreo de pares, función de pérdida y arquitectura de la red. Los autores clasificaron los métodos en tres categorías principales: redes siamesas, redes triplete y variantes multimuestras, y analizaron sus respectivas funciones de pérdida contrastiva, triplete y N-*pair loss* destacando sus ventajas y limitaciones. Un hallazgo central es que la estrategia de muestreo de negativos es tan determinante para el rendimiento final como la elección de la función de pérdida: negativos demasiado fáciles generan gradientes débiles, mientras que negativos demasiado difíciles causan inestabilidad en el entrenamiento. Esta revisión es especialmente relevante para la presente investigación porque sistematiza las decisiones de diseño clave al construir un espacio métrico para el emparejamiento de cadenas de anticuerpos, incluyendo la selección de negativos biológicamente fundamentados como los negativos verdaderos derivados de cadenas pesadas de BALB/c (Kaya & Bilge, 2019).

3.2.2. Funciones de pérdida contrastivas para el aprendizaje de espacios métricos

Van den Oord, Li y Vinyals (2018) introdujeron el *Contrastive Predictive Coding* (CPC), un marco de aprendizaje no supervisado de representaciones que incluye la función de pérdida InfoNCE. A diferencia de la pérdida contrastiva binaria de las redes siamesas clásicas, que opera sobre un único par positivo y otro negativo, InfoNCE plantea el aprendizaje como un problema de clasificación multiclase: dado un contexto, el modelo debe reconocer la muestra positiva correcta entre N muestras, de las cuales N-1 son negativas. La pérdida se define como la entropía cruzada de esta clasificación, lo que equivale a maximizar una cota inferior de la información mutua entre las representaciones de las muestras positivas. Los autores demostraron que InfoNCE produce representaciones de alta calidad en cuatro dominios

distintos audio, imágenes, texto y aprendizaje por refuerzo sin necesidad de etiquetas. Esta función de pérdida es directamente relevante para la presente investigación, ya que fue adoptada como función de entrenamiento del modelo propuesto en sustitución de la pérdida triplete, por su mayor estabilidad numérica y su capacidad de aprovechar múltiples negativos por lote de entrenamiento (van den Oord et al., 2018).

Zhang et al. (2023) presentaron SiameseCPP, una red siamesa específicamente diseñada para la predicción de péptidos de penetración celular (*cell-penetrating peptides*, CPPs) a partir de secuencias, combinando aprendizaje contrastivo con representaciones de modelos de lenguaje proteicos. La arquitectura procesa secuencias mediante un *encoder transformer* bidireccional seguido de una capa BiGRU, y aplica aprendizaje contrastivo para maximizar el acuerdo entre representaciones de péptidos similares y minimizarlo entre péptidos disimilares. Una característica distintiva del modelo es que fusiona las representaciones aprendidas por el aprendizaje contrastivo con los *embeddings* directos del modelo de lenguaje preentrenado, combinando el conocimiento general de la secuencia con las relaciones de similitud específicas de la tarea. SiameseCPP supera a métodos previos del estado del arte en predicción de CPPs y representa un precedente metodológico directo para la arquitectura propuesta en esta investigación, donde una red siamesa asimétrica aplica aprendizaje contrastivo sobre representaciones de secuencias de cadenas pesadas y ligeras de anticuerpos.

Tabla 3.5. Comparación de arquitecturas de aprendizaje profundo para tareas de similitud y emparejamiento

Autores	Método	Tipo de arquitectura	Función de pérdida	Dominio	Tarea evaluada
Koch et al.	Red siamesa convolucional	Siamesa simétrica (pesos compartidos)	Contrastiva binaria	Imágenes	Clasificación one-*shot*
van den Oord et al.	CPC + InfoNCE	*Encoder* + modelo autoregresivo	InfoNCE (multiclase)	Audio, imágenes, texto	Aprendizaje de representaciones no supervisado
Kaya & Bilge	Revisión *deep metric learning*	Siamesa, triplete, multimuestras	Contrastiva, triplete, N-*pair*	Múltiples dominios	Clasificación, verificación, recuperación
Zhang et al.	SiameseCPP	Siamesa simétrica (*Transformer* + BiGRU)	Contrastiva + *embeddings* PLM	Péptidos biológicos	Predicción de péptidos de penetración celular

3.2.3. Análisis de las arquitecturas de aprendizaje profundo para tareas de similitud y emparejamiento

Los trabajos revisados establecen los fundamentos sobre los que se construye la arquitectura propuesta: las redes siamesas como paradigma para aprender similitud a partir de pares, la sistematización de las decisiones de diseño que determinan su desempeño (muestreo de negativos, función de pérdida, arquitectura), la función de pérdida InfoNCE como alternativa más estable y eficiente que la pérdida triplete clásica, y un precedente directo de red siamesa con aprendizaje contrastivo aplicada a secuencias biológicas. En conjunto, esta línea de trabajo ha resuelto el problema de establecer principios de diseño generales y robustos para el aprendizaje de espacios métricos a partir de pares.

Sin embargo, ninguno de los trabajos revisados atiende dos particularidades del problema de esta tesis. Primero, todas las arquitecturas siamesas descritas son simétricas, es decir, comparan dos entradas del mismo tipo mediante ramas con pesos compartidos; ninguna aborda el caso en que las dos entradas provienen de dominios distintos con propiedades estructurales diferentes entre sí, como ocurre entre una cadena pesada y una cadena ligera de anticuerpo. Segundo, la discusión sobre estrategias de muestreo de negativos se centra en negativos generados dentro del propio conjunto de entrenamiento mediante criterios de dificultad, sin considerar negativos seleccionados por un criterio biológico externo, como una especie distinta con incompatibilidad estructural conocida.

Esto justifica dos decisiones arquitectónicas centrales de este trabajo: primero, el uso de una red siamesa asimétrica, con dos encoders independientes, uno por tipo de cadena, en lugar de la variante simétrica dominante en la literatura revisada; y segundo, la combinación de negativos in-batch semi-hard con negativos verdaderos biológicamente fundamentados (cadenas BALB/c), una estrategia de muestreo que no se identificó en ninguno de los trabajos de esta clasificación.

3.3. Modelos computacionales aplicados al análisis y emparejamiento de cadenas de anticuerpos

Esta clasificación reúne las investigaciones que abordan directamente el análisis computacional de repertorios de anticuerpos, la predicción de propiedades estructurales y funcionales de las cadenas VH y VL, y la construcción de modelos para el emparejamiento de cadenas pesadas y ligeras. A partir de estos trabajos se identificaron los problemas metodológicos que motivan la propuesta de esta investigación y se establecen las referencias de comparación para la evaluación del modelo desarrollado. En la tabla 3.6 se muestran los artículos que pertenecen a esta clasificación.

Tabla 3.6. Artículos sobre modelos computacionales aplicados al análisis y emparejamiento de cadenas de anticuerpos

ID	Título	Autores
15	Germline VH/VL *pairing* in *antibodies*	(Jayaram et al., 2012)
16	Five *computational* developability *guidelines* for *therapeutic antibody profiling*	(Raybould et al., 2019)
17	*Antibody structure prediction using* interpretable *deep learning* (DeepAb)	(Ruffolo et al., 2022)
18	*Improving antibody language models* with *native pairing* (BALM)	(Burbach & Briney, 2024)
19	*Large scale paired antibody language models* (IgBert & IgT5)	(Kenlay et al., 2024)
20	ImmunoMatch *learns* and *predicts cognate pairing* of *heavy* and light *immunoglobulin chains*	(Guo et al., 2026)

A continuación, se describen y analizan estas investigaciones, organizadas en tres grupos: análisis estructural y germinal de la interfaz VH-VL, modelos de predicción estructural y desarrollabilidad de anticuerpos, y modelos de lenguaje y aprendizaje profundo para el emparejamiento de cadenas.

3.3.1. Análisis estructural y germinal de la interfaz VH-VL

Jayaram, Bhowmick y Martin (2012) llevaron a cabo un análisis sistemático y a gran escala de las preferencias de emparejamiento entre genes germinales VH y VL en anticuerpos humanos y de ratón. A partir de una base de datos de anticuerpos con emparejamiento conocido, los autores identificaron combinaciones preferenciales de genes V que se repiten más de lo esperado por azar, lo que sugiere que hay restricciones evolutivas que favorecen ciertos emparejamientos VH-VL sobre otros. El análisis mostró que las preferencias germinales son parcialmente independientes de la especificidad antigénica, lo que sugiere que la compatibilidad entre cadenas está al menos parcialmente codificada en las secuencias de las regiones marco (framework) más que exclusivamente en las CDRs. Este trabajo es fundamental para la presente investigación porque establece la evidencia biológica de que el emparejamiento VH-VL no es aleatorio y puede ser modelado computacionalmente, y porque las preferencias germinales identificadas constituyen el fundamento del uso de cadenas pesadas de ratón BALB/c como negativos verdaderos en el entrenamiento del modelo propuesto (Jayaram et al., 2012).

3.3.2. Modelos de predicción estructural y desarrollabilidad de anticuerpos

El conjunto de cinco directrices computacionales, denominado *Therapeutic Antibody* Profiler (TAP), se propuso para evaluar la desarrollabilidad de anticuerpos terapéuticos en etapas tempranas del diseño (Raybould et al., 2019). A través del análisis de 137 anticuerpos clínicos en fase post-I y su comparación con un repertorio humano de referencia, los autores identificaron cinco propiedades fisicoquímicas asociadas con problemas de desarrollabilidad: longitud total de las CDRs, hidrofobicidad en la vecindad de las CDRs (PSH), parches de carga positiva (PPC), parches de carga negativa (PNC) y asimetría de carga entre cadenas pesada y ligera (SFvCSP). Esta última métrica es especialmente relevante para la presente investigación porque cuantifica explícitamente la complementariedad electrostática entre VH y VL como indicador de compatibilidad funcional, un principio que subyace al diseño del espacio métrico aprendido por el modelo propuesto. TAP demostró que las directrices computacionales permiten anticipar problemas de agregación y baja expresión antes de etapas experimentales costosas, posicionando el análisis in silico de la interfaz VH-VL como una herramienta de valor clínico (Raybould et al., 2019).

Ruffolo, Sulam y Gray (2022) presentaron DeepAb, un modelo de aprendizaje profundo que puede predecir estructuras de anticuerpos únicamente a partir de las secuencias de sus cadenas pesada y ligera. La arquitectura utiliza una red neuronal convolucional residual 2D junto con *embeddings* de un modelo LSTM preentrenado sobre repertorios de inmunoglobulinas, lo que le permite generalizar más allá de estructuras conocidas sin necesidad de injertos de fragmentos estructurales. DeepAb es interpretable mediante mecanismos de atención criss-cross que identifican los pares de residuos inter-cadena que más influyen en las predicciones estructurales, especialmente en la región CDR-H3. Los resultados en *benchmarks* estándar muestran que DeepAb supera a métodos previos como RosettaAntibody en la predicción de la orientación VH-VL y la geometría de los bucles CDR. Este trabajo es relevante para la presente investigación porque confirma que la información de compatibilidad entre cadenas pesada y ligera está codificada en sus secuencias y puede extraerse mediante modelos de aprendizaje profundo, y porque el mecanismo de atención criss-cross es conceptualmente análogo al módulo de cross-*attention* bidireccional implementado en el modelo propuesto (Ruffolo et al., 2022).

3.3.3. Modelos de lenguaje y aprendizaje profundo para el emparejamiento de cadenas

Burbach y Briney (2024) introdujeron BALM (*Baseline Antibody Language Models*), un trabajo que evalúa de forma sistemática cómo el entrenamiento con secuencias de anticuerpos nativas genuinamente emparejadas mejoras el rendimiento de los modelos de lenguaje de anticuerpos comparado con el entrenamiento con secuencias aleatoriamente emparejadas o no emparejadas. Los autores entrenaron tres versiones diferentes del modelo: BALM emparejado (entrenado con ~1.6 millones de pares nativos), BALM barajado (pares aleatorios) y BALM sin emparejar, así como una versión de ESM-2 ajustada con pares nativos. Los resultados mostraron que el entrenamiento con pares nativos permite al modelo aprender características inmunológicamente relevantes que abarcan ambas cadenas

simultáneamente, incluyendo señales de compatibilidad inter-cadenas que no pueden ser simuladas con pares aleatorios. Este hallazgo es directamente relevante para la presente investigación porque valida el principio central del modelo propuesto: la compatibilidad VH-VL es una señal aprendible desde secuencia, y la calidad de los ejemplos negativos de entrenamiento determina la capacidad del modelo para discriminar pares compatibles de incompatibles (Burbach & Briney, 2024).

Kenlay et al. (2024) presentaron IgBert e IgT5, modelos de lenguaje de anticuerpos entrenados a gran escala que pueden procesar simultáneamente secuencias pareadas e impareadas de cadenas pesadas y ligeras. Los modelos fueron entrenados sobre más de dos mil millones de secuencias impareadas y dos millones de secuencias pareadas de OAS, iniciando desde los pesos de ProtBERT y ProtT5 respectivamente. A diferencia de modelos anteriores entrenados exclusivamente sobre secuencias individuales, IgBert e IgT5 codifican las dos cadenas de forma conjunta separadas por un *token* especial, permitiendo que el modelo capture dependencias cruzadas entre VH y VL. Estos modelos superan a AntiBERTa, AbLang y ESM-2 en una diversa gama de tareas de diseño y regresión relevantes para ingeniería de anticuerpos. IgBert e IgT5 representan el estado del arte en modelos de lenguaje que procesan la naturaleza heterodimérica del anticuerpo, y establecen el contexto tecnológico en el que se inscribe el modelo propuesto en esta investigación (Kenlay et al., 2024).

Guo et al. (2026) introducen ImmunoMatch, el primer conjunto de modelos de aprendizaje profundo diseñados específicamente para la clasificación de pares cognativos de cadenas VH y VL de anticuerpos. Los modelos, llamados ImmunoMatch-κ e ImmunoMatch-λ para las cadenas ligeras kappa y lambda, se desarrollaron mediante fine-*tuning* del modelo AntiBERTa2 en secuencias de longitud completa del dominio variable recuperado de los repertorios de células B de donantes humanos. El entrenamiento empleó pares positivos procedentes de la secuenciación de célula única y pares pseudo-negativos generados mediante mezcla controlada de secuencias, reproduciendo la distribución de los genes V y J observada in vivo. ImmunoMatch fue validado en múltiples contextos: discriminación de pares cognativos en repertorios de células B *naive*, de memoria y de centros germinales, aplicación a datos de transcriptómica espacial, y evaluación de posiciones clave de la interfaz VH-VL en anticuerpos terapéuticos (Guo et al., 2026).

Tabla 3.7. Comparación de modelos computacionales aplicados al análisis y emparejamiento de cadenas de anticuerpos

Autores	Modelo	Enfoque	Representación	Tarea principal	Datos
Jayaram et al.	Análisis germinal	Estadístico	Genes V/J	Preferencias VH-VL germinales	Anticuerpos humanos y de ratón
Raybould et al.	TAP	Fisicoquímico	Propiedades de secuencia	Evaluación de desarrollabilidad	137 anticuerpos clínicos
Ruffolo et al.	DeepAb	LSTM + CNN 2D	*Embeddings* de repertorio	Predicción estructural VH-VL	SAbDab
Burbach & Briney	BALM	Modelo de lenguaje (BERT)	Secuencias pareadas	Clasificación por especificidad	1.6M pares nativos (OAS)
Kenlay et al.	IgBert / IgT5	Modelo de lenguaje pareado	Secuencias VH+VL conjuntas	Diseño y regresión de anticuerpos	2B+ secuencias OAS
Guo et al.	ImmunoMatch	Fine-*tuning* AntiBERTa2	*Embeddings* contextuales	Clasificaciones pares cognativos VH-VL	Repertorios células B humanas

3.3.4. Análisis de los modelos computacionales aplicados al análisis y emparejamiento de cadenas de anticuerpos

Los trabajos revisados en esta clasificación abordan el problema de emparejamiento VH-VL de forma directa y constituyen la comparación más cercana al modelo propuesto. En conjunto, ya han resuelto varios aspectos fundamentales del problema: establecieron evidencia de que el emparejamiento VH-VL no es aleatorio y está parcialmente codificado en las regiones marco; demostraron que propiedades fisicoquímicas de la interfaz VH-VL, como la asimetría de carga entre cadenas, se relacionan con la compatibilidad funcional; confirmaron que esa compatibilidad puede extraerse de la secuencia mediante mecanismos de atención inter-cadena; y validaron que entrenar con pares nativos genuinamente emparejados produce señales de compatibilidad que no pueden simularse con pares aleatorios. El trabajo más reciente de esta clasificación llega incluso a formular el mismo problema que aborda esta tesis: clasificar si una cadena pesada y una ligera forman un par cognado.

Sin embargo, persisten limitaciones incluso en los trabajos más cercanos. El modelo de clasificación de pares cognados y los modelos de lenguaje pareados dependen de hacer *fine-tuning* sobre modelos de lenguaje preentrenados a gran escala, lo que implica un costo computacional considerable y una menor interpretabilidad respecto al origen de cada decisión del modelo. Además, los negativos utilizados por el modelo de clasificación de pares cognados son pseudo-negativos generados mediante mezcla controlada de secuencias dentro de la misma especie, reproduciendo la distribución de genes V y J observada in vivo, un enfoque que no aprovecha explícitamente la evidencia de incompatibilidad entre especies ya documentada previamente en esta misma clasificación. Por su parte, el mecanismo de atención inter-cadena del modelo de predicción estructural se utiliza para predecir estructura tridimensional, no para construir un espacio métrico de compatibilidad explícitamente comparable entre pares.

3.4. Conclusiones del estado del arte

El análisis de las tres líneas de investigación presentadas en este capítulo (métodos de representación numérica, arquitecturas de aprendizaje profundo para similitud, y modelos computacionales aplicados al emparejamiento VH-VL) permite consolidar cuatro brechas metodológicas que motivan la propuesta de esta investigación:

- Ningún trabajo, salvo el más reciente de esta clasificación, aborda el emparejamiento VH-VL como un problema de aprendizaje profundo supervisado sobre secuencias.
- Las arquitecturas de similitud revisadas son simétricas y no explotan la asimetría funcional entre cadena pesada y cadena ligera mediante *encoders* especializados.
- Ningún trabajo evalúa negativos con fundamento biológico entre especies (como cadenas pesadas de organismos con repertorios divergentes) como estrategia de entrenamiento.
- La mayoría de los modelos existentes opera como clasificador binario, sin producir un espacio métrico continuo que permita recuperación eficiente por similitud a escala de repertorio.

El modelo propuesto en esta investigación busca atender estas cuatro brechas mediante una red siamesa asimétrica entrenada con la función de pérdida *InfoNCE*, representaciones fisicoquímicas interpretables basadas en factores Atchley, y negativos verdaderos derivados de cadenas pesadas de ratón BALB/c, produciendo un espacio métrico continuo que habilita la recuperación eficiente de cadenas compatibles mediante búsqueda por similitud.

En la Tabla 3.8 se consolidan los trabajos revisados en las tres líneas de investigación, destacando para cada uno el método propuesto, la representación utilizada, la tarea principal abordada y la brecha metodológica identificada con respecto al problema de emparejamiento funcional VH-VL que motiva la presente investigación.

Tabla 3.8. Comparación integral de trabajos relacionados en el análisis computacional de anticuerpos.

Autores	Línea	Método / Modelo	Representación	Tarea principal	Brecha identificada
Atchley et al.	Representación	Factores Atchley	Fisicoquímica (PCA)	Codificación de aminoácidos	No diseñada para emparejamiento VH-VL
Jing et al.	Representación	Revisión comparativa	16 métodos (5 categorías)	Comparación de representaciones	Sin aplicación directa a compatibilidad inter-cadena
Koch et al.	Similitud	Red siamesa	Embeddings de imagen	Aprendizaje one-shot	No aborda asimetría VH-VL
Van den Oord et al.	Similitud	InfoNCE	Representaciones continuas	Aprendizaje contrastivo	No aplicado a secuencias proteicas
Zhang et al.	Similitud	SiameseCPP	Secuencias proteicas	Similitud de péptidos	Sin especialización en anticuerpos
Jayaram et al.	Emparejamiento	Análisis germinal	Genes V/J	Preferencias VH-VL germinales	Solo nivel germinal, no secuencia completa
Raybould et al.	Emparejamiento	TAP	Propiedades de secuencia	Evaluación de desarrollabilidad	No predice compatibilidad VH-VL
Ruffolo et al.	Emparejamiento	DeepAb	Embeddings de repertorio	Predicción estructural VH-VL	No produce espacio métrico continuo
Burbach & Briney	Emparejamiento	BALM	Secuencias pareadas	Clasificación por especificidad	Clasificador binario, no escalable
Guo et al.	Emparejamiento	ImmunoMatch	Embeddings contextuales	Clasificación de pares cognitivos VH-VL	Cross-encoder, costoso a escala de repertorio

Capítulo 4.

Construcción de la representación de los aminoácidos

4. Construcción de la representación de los aminoácidos

El análisis computacional de los anticuerpos requiere transformar las secuencias de las cadenas pesadas y ligeras que los conforman en representaciones numéricas que pueden ser utilizadas por los modelos de aprendizaje automático. Estas representaciones deben de tomar en cuenta las características funcionales y bioquímicas de las secuencias de aminoácidos, si perder generalidad y sin introducir ruido innecesario a los datos. Tomando en cuenta lo anterior, la elección de la representación de las cadenas de aminoácidos es una tarea de suma importancia porque las distintas representaciones pueden causar variaciones en el desempeño de los modelos de predicción.

En el caso de los anticuerpos, la especificidad de unión con el antígeno depende de la conformación de los CDRs, una representación adecuada tiene la posibilidad de generar una diferencia entre un modelo útil y uno ineficaz. Por estas razones, en este capítulo se realiza una comparación de distintas representaciones de cadenas de aminoácidos, evaluando su capacidad para describir propiedades importantes de los CDRs y de los FRs.

En la comparación se incluyen representaciones basadas en propiedades fisicoquímicas, como los son los factores Atchley, también enfoques inspirados en el procesamiento de lenguaje natural, como TF-IDF y ProtVec.

4.1. Obtención de datos de anticuerpos

Para construir el conjunto de datos, se recopilaron secuencias de anticuerpos humanos con el fin de analizar las regiones variables responsables del reconocimiento de antígenos. Estas regiones variables están compuestas por tres regiones determinantes de complementariedad (CDRs) y cuatro regiones estructurales denominadas *Frameworks* (FR1, FR2, FR3 y FR4). Mientras que los CDRs son los principales responsables del contacto directo con el antígeno, los *Frameworks* proporcionan la estructura que permite mantener la conformación adecuada del sitio de unión.

4.1.1. Anticuerpos anti-SARS-CoV-2

El repositorio *The Coronavirus Antibody Database* (CoV-AbDab), contiene registros exclusivamente de estructuras de anticuerpos anti-SARS-CoV-2. El repositorio contiene registros de anticuerpos de diversas especies, incluyendo humano (10,779 registros), ratón (223 registros), alpaca (673 registros) y de primates no humanos (7 registros). Dado que el enfoque del proyecto de investigación se centra en los anticuerpos humanos, se llevó a cabo un filtrado de los registros para conservar únicamente las estructuras de los anticuerpos obtenidos de seres humanos. En la tabla 4.1, un ejemplo de registro de los anticuerpos de este repositorio.

Tabla 4.1. Ejemplo de algunos atributos de un anticuerpo obtenido de CoV-AbDab.

Nombre	Tipo de anticuerpo	Cadena pesada	Cadena ligera	Gen V cadena pesada	Gen J cadena pesada
Zhang_P4J15	Anticuerpo (Ab)	QVQIQQWGAGLL KPSETLSVYDE..	DIQMTQSPSSLSAS..	IGHV4-34	IGHJ4

De la estructura completa de cada uno de los anticuerpos se extrajeron las secuencias de los tres CDRs y los cuatro *Frameworks* (FR1, FR2, FR3 y FR4) de las cadenas pesadas y ligeras, ya que, en conjunto, estas regiones conforman la parte variable del anticuerpo. Mientras que los CDRs son los responsables directos de la unión al antígeno, los *Frameworks* proveen soporte estructural esencial para mantener la conformación adecuada del sitio de unión.

4.1.2. Anticuerpos que atacan a otro antígeno

Con el fin de incorporar datos negativos a las tareas de clasificación, se obtuvieron datos de anticuerpos que identifiquen antígenos diferentes al SARS-Co-V-2. Los datos de anticuerpos completos que no reconocen el virus SARS-CoV-2 del repositorio *Observed Antibody Space* (OAS, por sus siglas en inglés). Para la búsqueda de estos datos, se establecieron los siguientes criterios: las muestras debían provenir de personas que no estuvieran enfermas al momento de la recolección, y estas debían haber sido obtenidas antes del descubrimiento del SARS-CoV-2, con el objetivo de garantizar que los anticuerpos no reconocieran a este virus.

Como resultado de la búsqueda se obtuvo un total de 24 archivos CSV, que contenían secuencias de cadenas pesadas y ligeras que conforman anticuerpos completos. En total se obtuvieron 350,209 anticuerpos completos. Al igual que con los datos de CoV-AbDab solo se extrajeron las secuencias de los CDRs en aminoácidos. En la tabla 4.2 se muestran algunos atributos que contiene el repositorio de OAS.

Tabla 4.2. Ejemplo de algunos atributos de un anticuerpo obtenido de OAS.

ID	Locus	Secuencia completa	FR1	CDR1
OAS_H_0001842	IGH	QVQLVQSGAEVKKPGSS VKVSCKASGGTF….	QVQLVQSGA..	GGTFSSYA

4.2. Representaciones de las cadenas de aminoácidos

Para utilizar las secuencias de cadenas pesadas y ligeras de los anticuerpos (cadenas de caracteres) como entrada en los modelos de aprendizaje automático, es necesario

transformarlas en representaciones numéricas que preserven sus propiedades estructurales y funcionales.

En este trabajo, se consideraron de forma integral las regiones variables de las cadenas, compuestas por tres regiones determinantes de complementariedad (CDRs) y cuatro regiones estructurales denominadas *Frameworks* (FR1, FR2, FR3 y FR4). Las CDRs son responsables del reconocimiento directo del antígeno, mientras que los *Frameworks* mantienen la conformación tridimensional que permite una interacción efectiva y estable. Por tanto, ambas regiones son fundamentales para una representación biológicamente significativa.

La elección de una representación adecuada es un paso crucial, ya que determina qué características se conservan del sistema biológico original. Una representación inadecuada puede distorsionar o perder información clave, afectando drásticamente el rendimiento del modelo. Por el contrario, cuando la representación es adecuada, el modelo puede generalizar a secuencias nunca vistas, lo cual es fundamental en un entorno alta variabilidad como los anticuerpos.

Las regiones variables de los anticuerpos están compuestas por tres CDRs y cuatro *Frameworks* (FRs). Para representar numéricamente estas secuencias, se consideraron ambas partes debido a su relevancia estructural: los CDRs por su interacción directa con el antígeno, y los FRs por su función de soporte y estabilidad conformacional.

A continuación, se describen las representaciones evaluadas durante este proyecto: TF-IDF, factores de Atchley y ProtVec.

4.2.1. Representación utilizando bolsa de palabras TF-IDF

La bolsa de palabras TF-IDF (*Term Frequency – Inverse Document Frequency*) es una representación numérica de documentos que pertenecen a un conjunto de documentos, comúnmente es utilizada en el campo de procesamiento de lenguaje natural y minería de texto. Esta representación toma en cuenta la frecuencia de palabras en un documento y la rareza de esas palabras en el conjunto de documentos completos.

En esta investigación, cada sección de las cadenas (CDRs y Frameworks) se considera como un documento, mientras que cada aminoácido, representado por su código de una letra (por ejemplo, A para alanina, C para cisteína, D para ácido aspártico), se considera como una palabra individual. De esta forma, el diccionario de palabras consta de 20 letras correspondientes a los 20 aminoácidos.

En la tabla 4.3 se muestra un ejemplo de la representación TF-IDF de las cadenas pesadas completas: cada fila corresponde a uno de los 20 aminoácidos (identificados por su código de una letra), y cada columna corresponde a una cadena pesada específica, identificada por el nombre del anticuerpo al que pertenece (por ejemplo, BA7054). El valor en cada celda es el peso TF-IDF de ese aminoácido dentro de esa cadena particular: valores más altos indican que el aminoácido aparece con mayor frecuencia relativa en esa cadena que en el resto del conjunto de datos.

Tabla 4.3: Ejemplo de vectores TF-IDF cadenas pesadas

Aminoácido	BA7054	BA7125	BA7208	6-2C	Jaiswal	IS-9A	CC68.109	CC99.103	SD1.040
A	0.28	0.26	0.25	0.27	0.3	0.29	0.31	0.32	0.33
C	0.05	0.06	0.05	0.05	0.06	0.06	0.07	0.06	0.06
D	0.18	0.17	0.19	0.16	0.18	0.19	0.2	0.21	0.19
E	0.1	0.12	0.11	0.12	0.13	0.12	0.13	0.11	0.1
F	0.08	0.09	0.07	0.1	0.11	0.09	0.1	0.11	0.12
G	0.42	0.4	0.43	0.41	0.42	0.44	0.45	0.43	0.41
H	0.03	0.02	0.04	0.02	0.03	0.04	0.05	0.03	0.02
I	0.07	0.08	0.06	0.07	0.09	0.08	0.07	0.08	0.07
K	0.12	0.11	0.13	0.1	0.12	0.13	0.14	0.13	0.12
L	0.14	0.15	0.13	0.14	0.16	0.15	0.17	0.15	0.14
M	0.11	0.1	0.12	0.11	0.13	0.12	0.13	0.14	0.11
N	0.16	0.15	0.17	0.14	0.15	0.18	0.16	0.17	0.15
P	0.22	0.2	0.21	0.23	0.24	0.22	0.23	0.25	0.21
Q	0.18	0.17	0.19	0.16	0.17	0.2	0.21	0.19	0.18
R	0.2	0.21	0.18	0.22	0.24	0.23	0.22	0.21	0.23
S	0.3	0.28	0.29	0.31	0.32	0.3	0.31	0.33	0.29
T	0.33	0.31	0.32	0.34	0.35	0.33	0.34	0.36	0.32
V	0.26	0.25	0.27	0.24	0.26	0.28	0.27	0.26	0.25
W	0.08	0.09	0.07	0.1	0.11	0.09	0.1	0.11	0.12
Y	0.1	0.11	0.09	0.12	0.13	0.11	0.12	0.13	0.14

En la figura 4.1, se muestran los 9 vectores de la tabla 4.3 que fueron generados utilizando TF-IDF, en donde se puede apreciar que los picos más altos pertenecen a los aminoácidos con una mayor relevancia en las cadenas pesadas.

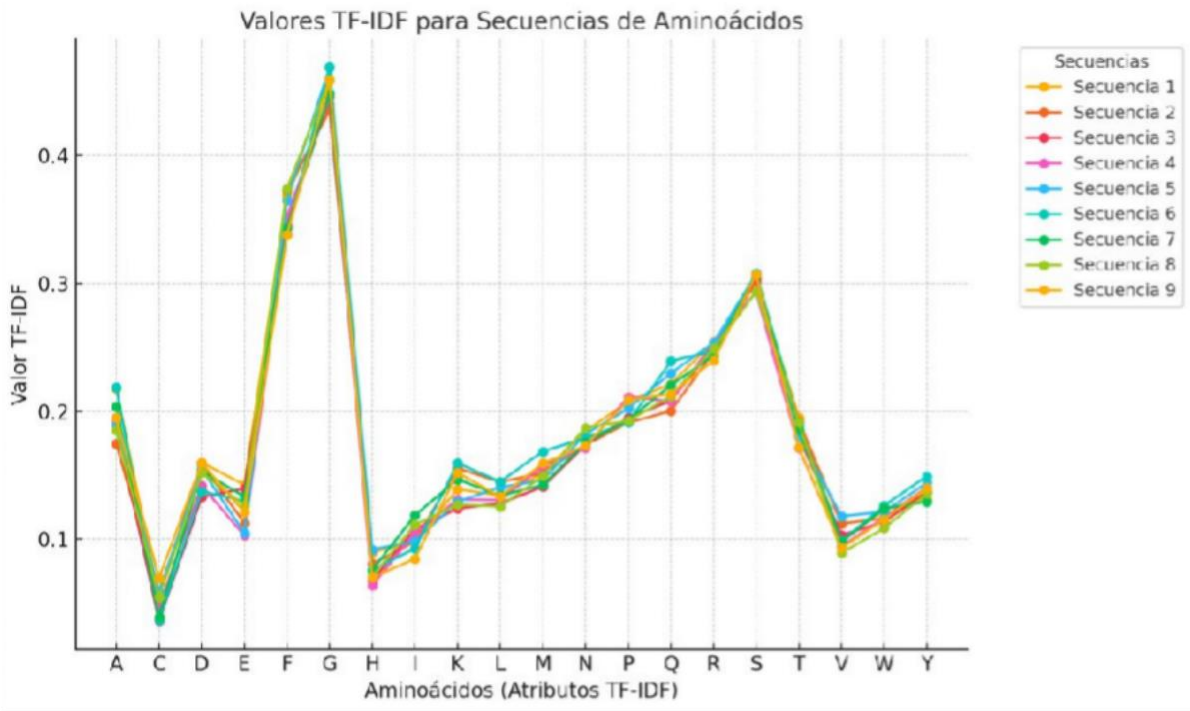

Figura 4.1. Vectores generados con TF – IDF

El siguiente paso fue generar los vectores que representen las secciones de las cadenas pesadas y ligeras, en donde también se puede observar que los picos más altos son los aminoácidos con una mayor relevancia en estas secciones. En la figura 4.2, se muestran algunos vectores generados de los CDRs de las cadenas pesadas.

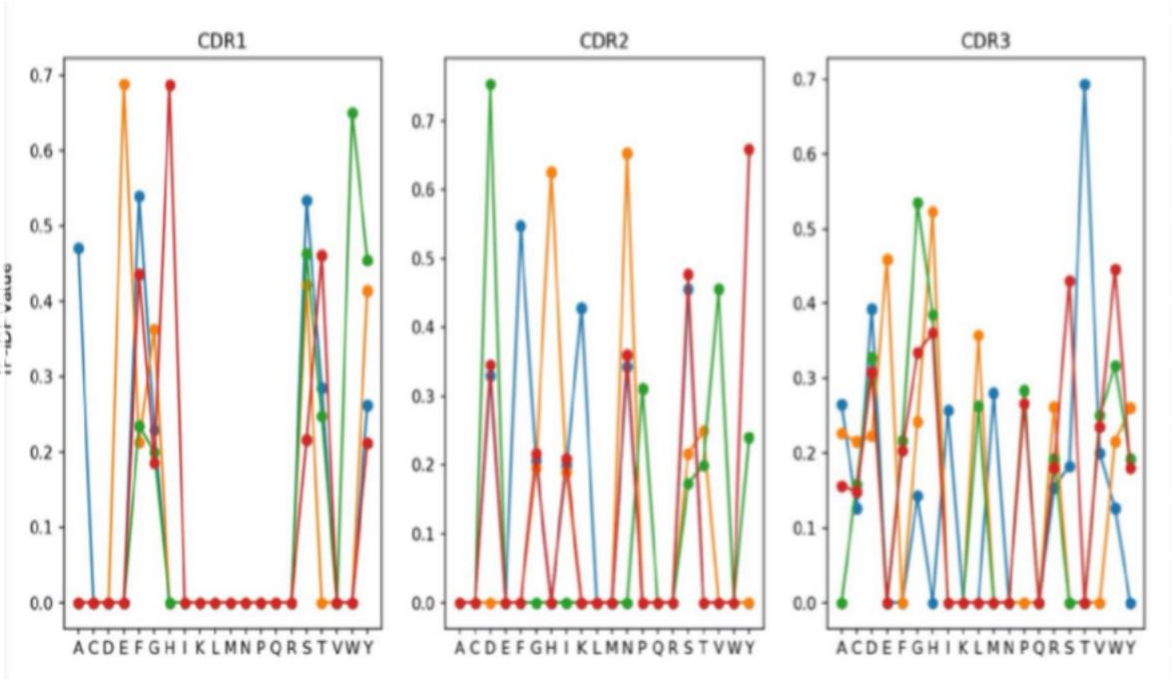

Figura 4.2. Vectores TF-IDF de los CDRs de cadenas pesadas

4.2.2. Representación de cadenas pesadas y ligeras por sus factores Atchley

Los factores de Atchley constituyen una codificación numérica basada en propiedades fisicoquímicas derivadas del análisis de componentes principales (PCA) aplicado a una amplia gama de características de los aminoácidos. Cada aminoácido es representado como un vector de 5 dimensiones que refleja su:

- Polaridad
- Hidrofobicidad
- Tamaño molecular
- Carga eléctrica
- Tendencia estructural secundaria

A diferencia de trabajar con los 20 aminoácidos y sus múltiples propiedades individuales, los factores Atchley proporcionan una representación compacta de 5 dimensiones de cada cadena de aminoácidos. Este enfoque reduce la dimensionalidad y al mismo tiempo conserva las propiedades bioquímicas y estructurales esenciales.

Los vectores finales de esta representación se obtienen calculando el promedio de cada factor de los aminoácidos que conforman las secuencias de aminoácidos. La ecuación 7, muestra la fórmula utilizada para para calcular cada factor Atchley para las secuencias.

$$F_i^{CDR} = \frac{1}{n}\sum_{j=1}^{n} F_i^{(j)} \tag{7}$$

Donde:

F_i^{CDR} es el valor del factor de Atchley para toda la secuencia.

$F_i^{(j)}$ es el valor del factor de Atchley para el j-ésimo aminoácido en la secuencia.

n es el número total de aminoácidos en la secuencia.

Para esta representación, el conjunto de datos contiene 35 atributos, correspondientes a los 5 factores Atchley calculados para cada uno de los tres CDRs (CDR1, CDR2 y CDR3) y los cuatro Frameworks (FR1, FR2, FR3 y FR4). Cada columna corresponde a uno de estos 35 atributos, nombrado como FnCDRm o FnFRm (por ejemplo, F1CDR1 es el primer factor Atchley del CDR1, F3FR2 sería el tercer factor Atchley del FR2), y cada fila corresponde a una cadena pesada distinta del conjunto de datos, identificada por el anticuerpo al que pertenece. La tabla 4.4 muestra un ejemplo de esta representación; por espacio, solo se incluyen las columnas correspondientes a CDR1 y CDR3, omitiendo CDR2 y los cuatro Frameworks.

Tabla 4.4. Representación de los tres CDRs de las cadenas pesadas con factores de Atchley

F1CDR1	F2CDR1	F3CDR1	F4CDR1	F5CDR1	……..	F1CDR3	F2CDR3	F3CDR3	F4CDR3	F5CDR3
-0.401	0.390	0.021	0.403	0.034	……..	-0.514	0.690	0.221	0.403	0.334
-0.044	0.714	0.337	0.183	0.179	……..	0.195	0.514	0.537	0.383	0.379
-0.244	0.731	0.347	-0.113	0.166	……..	0.331	0.631	0.547	-0.313	0.366
-0.261	0.367	0.775	0.053	0.537	……..	0.423	0.367	0.775	0.253	0.737
-0.401	0.390	0.021	0.403	0.034	……..	0.214	0.490	0.321	0.603	0.234
-0.097	0.660	1.104	-0.218	0.614	……..	0.297	0.560	1.204	-0.418	0.614
0.119	0.628	2.054	0.056	1.249	……..	0.419	0.628	2.254	0.256	1.149
-0.014	0.634	0.678	0.149	0.279	……..	0.254	0.634	0.878	0.349	0.479
-0.271	0.522	1.371	0.132	0.736	……..	0.371	0.522	1.471	0.332	0.736

4.2.3. Representación utilizando ProtVec

ProtVec es una técnica utilizada para representar numéricamente secuencias de anticuerpos, utilizando incrustaciones de palabras para transformar secuencias de aminoácidos en vectores de alta dimensión. El número de dimensiones es un factor crucial, ya que depende de la tarea específica que se vaya a realizar y tiene un impacto directo en el rendimiento de los modelos de aprendizaje automático.

ProtVec normalmente produce incrustaciones con dimensiones que oscilan entre 50 y 300, dependiendo de la configuración específica del modelo y de los requisitos de la tarea. Para trabajos de clasificación se recomienda un rango de 50 – 100 dimensiones. Mientras que, para aplicaciones más complejas como la predicción estructural o el análisis evolutivo, se recomienda un rango de 200 – 300 dimensiones. El valor inicial fue de 100 dimensione, por lo que el conjunto de datos que se genera con esta representación cuenta con 700 dimensiones ya que cada sección de las cadenas se representa con un vector de 100 dimensiones.

ProtVec genera las representaciones numéricas mediante el modelo Skip-gram de Word2Vec, descrito en la sección 2.3.3, aplicado sobre trigramas de aminoácidos extraídos de las secuencias. El modelo está entrenado con datos de proteínas a gran escala, capturando patrones estadísticos en secuencias de aminoácidos y codificándolos en vectores de 100 dimensiones. Sin embargo, los valores numéricos de estas representaciones dependen del proceso de entrenamiento, lo que dificulta determinar con precisión cómo contribuye cada aminoácido o secuencia al vector resultante.

En la tabla 4.5 se muestra un ejemplo de esta representación aplicada al CDR1 de la cadena ligera: cada fila corresponde a un anticuerpo distinto, identificado en la columna ID, y cada columna (CDR1Light1 a CDR1Light100) corresponde a una de las 100 dimensiones del vector ProtVec generado para esa región. Por espacio, solo se muestran las primeras cinco dimensiones y la última (CDR1Light100).

Tabla 4.5. Representación de CDR1 con ProtVec

ID	CDR1Light1	CDR1Light2	CDR1Light3	CDR1Light4	CDR1Light5	…	CDR1Light100
BA7054	-0.011314	-0.081608	-0.373200	-0.012507	-0.104019	…	0.332234
BA7125	-0.195655	-0.027455	-0.310991	0.096563	0.059360	…	0.502038
BA7208	-0.221828	0.132836	-0.309813	0.092073	0.125076	…	0.412048
6-2C	0.046774	-0.083658	0.150353	-0.298076	-0.252442	…	-0.039804
Jaiswal	-0.229373	0.056932	-0.328333	0.105182	0.106505	…	0.474955
IS-9A	0.133214	-0.084762	0.318797	-0.499609	-0.383408	…	-0.203989
CC68.109	0.116688	-0.199207	-0.001692	-0.065427	-0.097210	…	0.026315
CC99.103	0.027772	-0.127451	0.126028	-0.122607	-0.076175	…	0.030137
SD1.040	0.150177	-0.585391	0.350960	-0.307501	-0.282573	…	-0.001659

Después de generar cada una de las representaciones de las secuencias de aminoácidos, se obtuvieron tres conjuntos de datos distintos, uno para cada representación. Estos conjuntos de datos muestran diferentes características, ya que cada método de representación da como resultado un número distinto de atributos, lo que genera las variaciones en la estructura y dimensionalidad de los conjuntos de datos. En la tabla 4.6, se muestra un resumen de las características de cada conjunto de datos generado contemplando la representación de las cadenas pesadas y ligeras (cada cadena tiene 3 CDRs y 4 *Frameworks*).

Tabla 4.4. Conjuntos de datos

Conjunto de datos	Representación	Atributos por CDR y Framework	Total de atributos
DS1	TF-IDF	20	280
DS2	Factores Atchley	5	70
DS3	ProtVec	100	1,400

4.3. Evaluación de los conjuntos de datos

Una vez completado el procesamiento de los datos, y generado las representaciones de las cadenas de aminoácidos se evaluaron utilizando algoritmos de clasificación supervisada. El caso de estudio se centró en la clasificación de anticuerpos, diferenciando los anticuerpos capaces de reconocer el SARS-CoV-2 de aquellos que identifican a otros patógenos.

Los resultados permiten evaluar la efectividad de las representaciones de las cadenas pesadas y ligeras de anticuerpos y su impacto en la clasificación de anticuerpos. Este análisis es esencial para la creación de modelos de aprendizaje automático útiles en la investigación biomédica, en donde la identificación de anticuerpos es importante para comprender la respuesta inmune y así desarrollar nuevas terapias en contra de enfermedades virales.

4.3.1. Árboles de decisión

Se utilizaron árboles de decisión como primer enfoque de clasificación. La configuración de los hiperparámetros influye en el rendimiento y la capacidad de generalización de los modelos de aprendizaje automático. Algunos hiperparámetros se definieron manualmente para evitar el sobreajuste y optimizar la precisión de la clasificación, mientras que otros se mantuvieron en sus valores predeterminados. En la tabla 4.7, se presentan los hiperparámetros utilizados para cada parámetro del modelo.

Tabla 4.5. Hiperparámetros para árboles de decisión

Hiperparámetro	Valor	Descripción
criterio	Gini	Función para medir la calidad de una división
max_*depth*	5	Profundidad máxima del árbol para controlar el sobreajuste
random_state	42	Valor de inicialización para la generación de números aleatorios para garantizar la reproducibilidad
Divisor	*best*	Estrategia para elegir la división en cada nodo
min_*samples_split*	2	Número mínimo de muestras necesarias para dividir un nodo interno
min_*samples*_leaf	1	Número mínimo de muestras requeridas para estar en un nodo hoja
min_*weight_fraction*_leaf	0	Fracción mínima ponderada del total de muestras requeridas en un nodo hoja
max_*features*	none	Número de características a tener en cuenta a la hora de buscar la mejor división
max_leaf_*nodes*	none	Número máximo de nodos hoja
min_*impurity_decrease*	0	Un nodo se dividirá solo si la disminución de impurezas es mayor que este valor
class_weight	none	Pesos asociados a las clases
ccp_alpha	0	Parámetro de complejidad utilizado para la poda

Utilizando los hiperparámetros anteriormente mencionados, se realizaron experimentos con cada uno de los conjuntos de datos de las representaciones para evaluar el rendimiento del modelo con cada uno de ellos. En la tabla 4.8, se presentan los resultados para cada representación. Los factores de Atchley mostraron el mejor desempeño, alcanzando una precisión (*accuracy*) del 94% y un AUC de 0.96.

Tabla 4.6. Resultados de representaciones con árboles de decisión

Conjunto de datos	*Accuracy*	*Recall*	F1-*Score*	ROC AUC
DS1	0.86	0.86	0.86	0.88
DS2	0.94	0.93	0.93	0.96
DS3	0.80	0.81	0.82	0.84

4.3.2. Regresión logística

La regresión logística se seleccionó como un modelo base, por su interpretabilidad y eficiencia. Los hiperparámetros se configuraron para optimizar la convergencia y el rendimiento del modelo. En la tabla 4.9, se presentan los hiperparámetros utilizados.

Tabla 4.7. Hiperparámetros para regresión logística

Hiperparámetro	Valor	Descripción
pena	l2	Tipo de regularización
solucionador	lbfgs	Algoritmo utilizado para la optimización
max_iter	200	Número máximo de iteraciones para la convergencia
random_state	42	Valor de inicialización para la generación de números aleatorios para garantizar la reproducibilidad
C	1	Inverso de la fuerza de regularización
tol	0.0001	Detener la tolerancia para la convergencia de optimización
fit_*intercept*	True	Si se va a calcular la intersección del modelo
intercept_scaling	1	Factor de escala para el término de intercepción
warm_start	False	Reutilice la solución anterior para el inicio en caliente

Las representaciones de las secuencias de aminoácidos se utilizaron para comparar el rendimiento de la clasificación de anticuerpos. Nuevamente, los factores de Atchley superaron a las otras representaciones, con métricas ligeramente superiores a TF-IDF y sustancialmente mejores que ProtVec (Tabla 4.10).

Tabla 4.8.Resultados de regresión logística

Conjunto de datos	*Accuracy*	*Recall*	F1-*Score*	ROC AUC
DS1	0.85	0.84	0.84	0.87
DS2	0.86	0.85	0.85	0.88
DS3	0.80	0.79	0.78	0.81

4.3.3. Máquinas de soporte vectorial

Dado que las SVM son especialmente efectivas en espacios de alta dimensión, se utilizó este modelo para evaluar si representaciones como ProtVec podrían mejorar su desempeño (Tabla 4.11). Aunque todas las representaciones obtuvieron buenos resultados, los factores de Atchley nuevamente fueron superiores en todas las métricas (Tabla 4.12).

Tabla 4.9. Hiperparámetros utilizados en el modelo SVM

Hiperparámetro	Valor	Descripción
C	1	Parámetro de regularización: los valores más altos reducen la regularización
núcleo	RBF	Tipo de *kernel* utilizado en el algoritmo
grado	3	Grado de la función *kernel* polinómica
gamma	escam a	Coeficiente de *kernel*
coef0	0	Término independiente en la función del *kernel*
Reducción	True	Si se va a utilizar la heurística de contracción para la optimización
probabilidad	False	Permite estimaciones de probabilidad
tol	0.001	Criterio de tolerancia para la parada
cache *size*	200	Tamaño de la caché del *kernel* en MB

Tabla 4.10. Diferentes representaciones utilizando SVM

Conjunto de datos	*Accuracy*	*Recall*	F1-*Score*	ROC AUC
DS1	0.87	0.86	0.86	0.89
DS2	0.88	0.87	0.87	0.90
DS3	0.82	0.80	0.79	0.83

4.3.4. Resultados de la evaluación de los conjuntos de datos

La representación basada en los factores Atchley mostró un mejor desempeño en comparación con las otras dos técnicas evaluadas. Esto se debe a que considera las propiedades fisicoquímicas de cada aminoácido presente en la estructura de los anticuerpos, y además ofrece una baja dimensionalidad, lo cual contribuye a mejorar el rendimiento del modelo

Por su parte, la representación TF-IDF también obtuvo resultados positivos, sin embargo, esta técnica no conserva la información contextual ni estructural de las cadenas de aminoácidos, lo que puede limitar al modelo su capacidad de generalizar.

Al analizar las secuencias de las cadenas que componen los anticuerpos, se observó que la viabilidad de la longitud de las secuencias de aminoácidos, específicamente en las regiones determinantes de complementariedad (CDRs), representa un reto importante a la hora de representar las secuencias. Como era de esperarse, CDR3 presentó la mayor diversidad en longitud y en la frecuencia de aparición de cada uno de los aminoácidos. Esto demuestra que es la región la que juega un papel importante en la afinidad de los anticuerpos.

En la tabla 4.13, se muestra la comparación de los resultados que obtuvieron cada una de las representaciones.

Tabla 4.11. Comparación de representación de secuencias de aminoácidos

Representación	Modelos de ML	*Accuracy*	*Recall*	F1-*Score*	ROC AUC
Factores Atchley	Árboles de decisión	0.94	0.93	0.93	0.96
Factores Atchley	SVM	0.88	0.87	0.87	0.9
TF-IDF	SVM	0.87	0.86	0.86	0.89
TF-IDF	Árboles de decisión	0.86	0.86	0.86	0.88
Factores Atchley	Regresión logística	0.86	0.85	0.85	0.88
TF-IDF	Regresión logística	0.85	0.84	0.84	0.87
ProtVec	Árboles de decisión	0.80	0.81	0.82	0.84
ProtVec	SVM	0.82	0.80	0.79	0.83
ProtVec	Regresión logística	0.80	0.79	0.78	0.81

Capítulo 5.

Modelo para el emparejamiento de cadenas pesadas y ligeras

5. Modelo para el emparejamiento de cadenas pesadas y ligeras

En este capítulo se describe el diseño, implementación y entrenamiento del modelo computacional propuesto para la predicción del emparejamiento funcional entre cadenas pesadas y ligeras de anticuerpos. El modelo está basado en una red neuronal siamesa asimétrica entrenada con InfoNCE, está arquitectura permite proyectar los dos tipos de cadena en un espacio métrico compartido en donde la compatibilidad funcional se refleja en la distancia coseno entre sus representaciones vectoriales.

A diferencia de los enfoques tradicionales que generan negativos de manera aleatoria, en este trabajo se incorporan cadenas pesadas de anticuerpos de ratón BALB/c como negativos biológicamente fundamentados, bajo la premisa de que no forman anticuerpos funcionales con cadenas ligeras humanas debido a divergencias germinales, diferencias estructurales y ausencia de coevolución entre especies. Adicionalmente, se adopta la función de pérdida InfoNCE, cuyo mecanismo de negativos in-batch introduce de forma implícita un comportamiento equivalente a la minería de negativos semi-*hard*, seleccionando automáticamente los ejemplos más informativos para el entrenamiento en cada iteración.

El capítulo se organiza de la siguiente manera: primero se describe la construcción de las tripletas de entrenamiento y la estrategia de negativos; posteriormente se detalla la arquitectura del modelo y su función de pérdida; y finalmente se presenta el proceso de entrenamiento con la experimentación de distintas configuraciones de hiperparámetros.

5.1. Generación de tripletas para el entrenamiento

Para el entrenamiento del modelo siamés asimétrico se construyeron tripletas con la estructura (A, P, N), en donde cada elemento representa un tipo de cadena de anticuerpo con un rol específico dentro de la función de pérdida.

- Ancla (A): Vector que representa a una cadena ligera humana
- Positivo (P): Vector que representa a la cadena pesada humana que se empareja funcionalmente con el ancla, formando un anticuerpo real
- Negativo (N): Vector que representa una cadena pesada que no forma un anticuerpo funcional con el ancla.

El modelo recibe simultáneamente tres tipos de secuencias como entrada: las cadenas ligeras humanas, las cadenas pesadas humanas emparejadas y las cadenas pesadas de ratón BALB/c. A partir de estas tres fuentes se construyen las tripletas de entrenamiento. Las cadenas ligeras actúan siempre como ancla, las cadenas pesadas humanas emparejadas actúan como positivos, y las cadenas pesadas BALB/c actúan como negativos verdaderos. También, durante el entrenamiento la función de pérdida InfoNCE genera de forma automática negativos adicionales a partir de los demás pares humanos disponibles en cada batch, lo que

proporciona un comportamiento equivalente a la minería de negativos semi-*hard* sin necesidad de un proceso de selección explícito.

5.1.1. Negativos verdaderos: cadenas pesadas de ratón BALB/c

La primera estrategia de construcción de negativos consiste en agregar cadenas pesadas de anticuerpos de ratón BALB/c. Estas cadenas son biológicamente incompatibles con las cadenas ligeras humanas por tres razones principales.

- Divergencia en genes germinales VH: Los genes que codifican las regiones variables de las cadenas pesadas de ratón pertenecen a familias germinales distintas a las humanas. Esta divergencia evolutiva se traduce en diferencias en la composición y longitud de las regiones como los Frameworks y los CDR's, lo que provoca que la superficie de contacto entre las cadenas sea incompatible.
- Diferencias estructurales en el CDR3: la región CDR3 de las cadenas pesadas es la que tiene una mayor variabilidad y la que tiene una mayor importancia en la especificidad de unión con el antígeno. En el caso del ratón BALB/c, esta región presenta patrones de longitud y composición de aminoácidos distintos a los que han sido observados en anticuerpos humanos, lo que no permite que se forme un sitio de unión funcional al emparejarse con una cadena ligera humana.
- Ausencia de coevolución entre especies: en organismos vivos, las cadenas pesadas y ligeras de un mismo anticuerpo coevolucionan dentro del mismo individuo, desarrollando complementariedad estructural y electrostática entre sus interfaces. Al combinar cadenas de especies distintas, esta complementariedad no existe, lo que genera combinaciones que no pueden formar anticuerpos funcionales

El uso de estos negativos representa una contribución central de este trabajo respecto a enfoques previos, ya que provee ejemplos negativos con un fundamento biológico sólido, reduciendo la ambigüedad que introduce el uso de negativos aleatorios o artificiales en el entrenamiento.

5.1.2. Negativos semi-hard mediante InfoNCE in-batch

La segunda estrategia de selección de negativos se establece de manera implícita a través del funcionamiento propio de la función de pérdida InfoNCE. En cada iteración del entrenamiento, el modelo procesa un lote compuesto por 256 tripletas. Para cada ancla dentro del lote, el positivo corresponde a su cadena pesada emparejada, mientras que los negativos se generan automáticamente: son todas las demás cadenas pesadas incluidas en ese mismo lote. Este procedimiento, denominado negativos in-batch, tiene las siguientes propiedades:

- El comportamiento semi-hard implícito se deriva de la temperatura $\tau = 0.07$ utilizada en la función de pérdida InfoNCE. Este ajuste provoca que el gradiente de aprendizaje se concentre en los negativos más difíciles dentro del lote, es decir, aquellos cuya representación se encuentra más próxima al positivo en el espacio embebido. Este mecanismo resulta equivalente a una estrategia explícita de minería semi-hard: el modelo deja de aprender de negativos trivialmente fáciles y se enfoca en aquellos ubicados en la región de mayor ambigüedad. Como consecuencia, se ve obligado a refinar las fronteras

de decisión en el espacio métrico, fortaleciendo así la capacidad discriminativa del modelo.

- La actualización dinámica se produce porque el lote cambia en cada iteración y los embeddings se modifican continuamente con cada paso del optimizador. Como resultado, los negativos seleccionados de manera implícita también cambian en cada iteración, lo que provoca que el modelo trabaje con una amplia diversidad de casos difíciles a lo largo del entrenamiento. Esta variabilidad contribuye a enriquecer el proceso de aprendizaje y a fortalecer la capacidad del modelo para generalizar en escenarios complejos.
- Escala eficiente: con un batch de 256 muestras, cada ancla tiene acceso a 255 negativos in-batch en cada iteración, lo que proporciona una señal de aprendizaje mucho más rica que el uso de un único negativo fijo por tripleta.

La combinación de negativos BALB/c y negativos in-batch semi-*hard* permite que el modelo aprenda fronteras de decisión en dos niveles: los negativos BALB/c aportan una separación clara y biológicamente fundamentada entre especies, mientras que los negativos in-batch obligan al modelo a discriminar entre cadenas humanas compatibles e incompatibles con mayor precisión.

5.1.3. Composición del conjunto de tripletas

Las tripletas se construyeron a partir de dos conjuntos de datos independientes: uno conformado por cadenas ligeras humanas y otro por cadenas pesadas humanas, ambos codificados mediante factores de Atchley. Posteriormente, los dos conjuntos se integraron utilizando el identificador del anticuerpo, lo que garantizó que cada par ancla-positivo correspondiera a un anticuerpo real con emparejamiento VH-VL conocido. Se utilizaron las cadenas pesadas de ratón BALB/c obtenidas del repositorio OAS y procesadas con la codificación de los factores Atchley, las cuales se utilizaron como negativos verdaderos en proporciones equivalentes a los pares humanos. De esta manera, se mantuvo un balance entre positivos y negativos en el conjunto de entrenamiento. Finalmente, todos los vectores fueron normalizados mediante StandardScaler antes de ser utilizados por el modelo.

Cada tripleta está formada por tres elementos: el ancla, que es siempre una cadena ligera humana; el positivo, que es la cadena pesada humana que forma un anticuerpo funcional con esa cadena ligera; y el negativo verdadero, que es una cadena pesada de ratón BALB/c biológicamente incompatible con el ancla. Adicionalmente, durante el entrenamiento InfoNCE genera negativos semi-*hard* a partir de los demás pares humanos disponibles en cada batch: con un tamaño de batch de 256 muestras, cada ancla tiene acceso a hasta 255 negativos in-batch por iteración, seleccionando automáticamente los más informativos tomando en cuenta la temperatura $\tau = 0.07$. El uso de dos tipos de negativos permite cubrir dos niveles de dificultad distintos: los negativos BALB/c garantizan una separación biológica clara e inequívoca entre especies, mientras que los negativos semi-*hard* in-batch obligan al modelo a aprender fronteras más finas entre cadenas del mismo dominio humano. La tabla 5.1 muestra la distribución del conjunto de tripletas utilizadas para entrenamiento y para prueba.

Tabla 5.1. Distribución del conjunto de tripletas

Etapa	Tripletas	Porcentaje
Entrenamiento	20,000	80%
Prueba	5,000	20%
Total	25,000	100%

5.2. Análisis exploratorio de los datos

Con el objetivo de identificar patrones latentes en los datos y evaluar la estructura de las representaciones numéricas de los factores Atchley de las cadenas pesadas y ligeras de los anticuerpos, se realizó un análisis exploratorio basado en técnicas de clustering no supervisado. Este análisis tuvo como finalidad determinar la capacidad de las representaciones para distinguir subgrupos estructural y funcionalmente relevantes, así como explorar su potencial para sugerir unión entre las cadenas.

Explorar si las secuencias representadas con los factores Atchley permiten una agrupación coherente y biológicamente significativa de las cadenas pesadas y ligeras. Se evaluó si los algoritmos de clustering eran capaces de distinguir subgrupos funcionales dentro de los datos sin necesidad de etiquetas.

La elección del número de clústeres (k) es una fase importante en la aplicación de algoritmos de clustering, ya que determina la granularidad con la que se agrupan los datos. Para estimar este valor, se aplicó una estrategia combinada que incluye métricas estadísticas internas y validación biológica basada en el comportamiento de las cadenas de anticuerpos.

Utilizando el método del codo se graficó la inercia intra-clúster, es decir, la suma de las distancias cuadradas entre los puntos y el centroide de su clúster para valores de k entre 2 y 10. A medida que aumenta k, la inercia disminuye, pero llega un punto donde la mejora es marginal. Este punto de inflexión, donde la curva comienza a estabilizarse, se identificó visualmente entre $k = 3$ y $k = 4$, lo cual sugiere que estos valores representan una división eficiente de los datos. En la figura 5.1, se muestra la gráfica de la curva del codo.

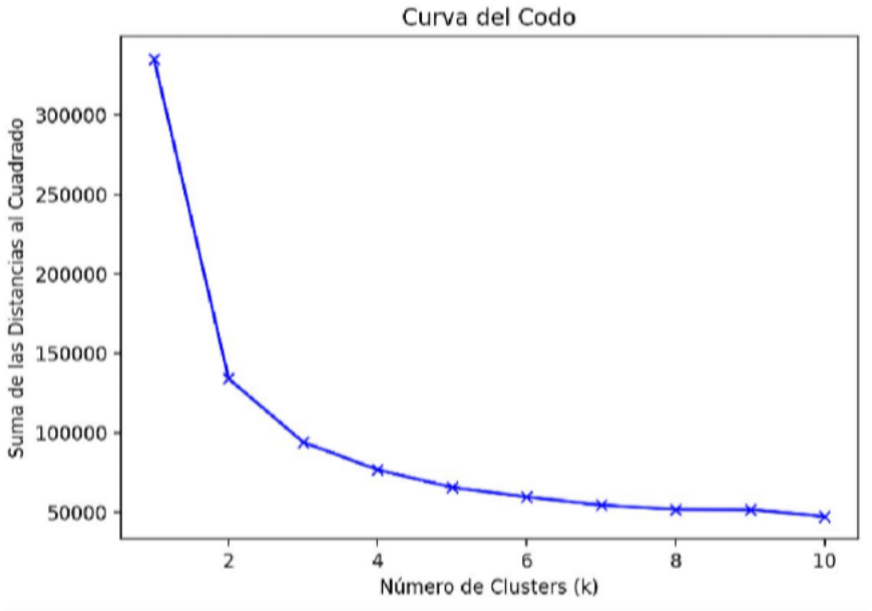

Figura 5.1. Curva generada por el método del codo

Por medio del Índice de *Silhouette*, se evalúa qué tan bien se separa cada punto de su clúster con respecto a los demás. El valor del índice varía entre -1 y 1, donde valores cercanos a 1 indican que los clústeres están bien definidos y separados. En este caso, el mayor promedio del *Silhouette score* se obtuvo en $k = 3$, lo que refleja buena cohesión interna y alta separación entre grupos.

Otra métrica utilizada fue Davies-Bouldin índex, que mide la relación entre la dispersión interna de los clústeres y su separación mutua. A diferencia de otras métricas, valores más bajos son mejores, ya que indican clústeres compactos y bien separados. El valor mínimo se alcanzó con $k = 3$, lo que indica una estructura de clústeres definida y con mínima superposición.

Por último, se calculó Calinski-Harabasz índex, este índice mide la razón entre la varianza inter-clúster y la intra-clúster. Un valor más alto sugiere que los clústeres están mejor definidos. En los resultados, este índice se maximiza para $k = 3$ y también presenta un valor alto en $k = 4$, lo que respalda estas configuraciones como las más apropiadas para representar la estructura de los datos. En la tabla 5.2, se muestra una comparación entre las técnicas utilizadas para determinar el valor óptimo de k.

Tabla 5.2. Métricas para determinar el valor de k

Métrica	Valor de *k*
Índice de *Silhouette*	$k = 3$ (mayor cohesión y separación)
Davies-Bouldin *index*	$k = 3$ (menor superposición entre grupos)
Calinski-Harabasz *index*	$k = 3$ y $k = 4$ (clústeres bien definidos)
Método del codo	$k = 3$ y $k = 4$ (punto de inflexión)

5.2.1. KMeans

El algoritmo KMeans fue el principal enfoque de agrupamiento utilizado, debido a su simplicidad, eficiencia computacional y capacidad para formar clústeres bien definidos en espacios vectoriales como los generados con la representación basada en los factores Atchley.

KMeans agrupa las secuencias de cadenas de anticuerpos en k clústeres utilizando un proceso iterativo que minimiza la suma de distancias cuadráticas entre los vectores de cada grupo y si centroide. Al trabajar con la representación de los factores Atchley, esta técnica permitió detectar agrupamientos basados en similitud fisicoquímica.

Los resultados que se obtuvieron fueron:

Para $k = 2$ se obtuvo una separación primaria entre cadenas pesadas y ligeras. Aunque estadísticamente aceptable, esta partición fue considerada demasiado general, ya que no distinguía variaciones internas relevantes dentro de cada tipo de cadena. Esta experimentación permite observar que la representación permite distinguir entre tipo de cadena (pesada y ligera). En la figura 5.2, se muestra gráficamente la división entre las cadenas.

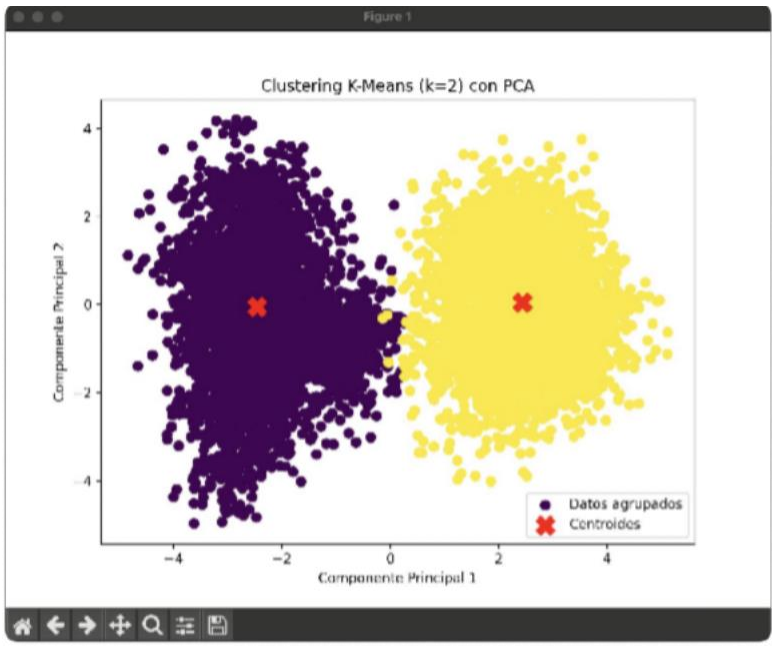

Figura 5.2. KMeans k = 2

Aunque se produce una clara división entre las cadenas pesadas y ligeras, las métricas internas reflejan una baja calidad en el agrupamiento. En la tabla 5.3, se muestran las métricas de evaluación de los clústeres.

Tabla 5.3. Métricas de evaluación de clústeres

Métrica	Resultado
Silhouette	0.293: baja cohesión / separación entre registros
Davies-Bouldin *index*	1.43: alta superposición entre grupos
Calinski-Harabasz *index*	4565.9: elevado por la gran distancia entre los dos tipos de cadenas, pero sin granularidad interna

Para $k = 3$ se formaron tres grupos, en donde de observa la separación entre cadenas pesadas y ligeras, pero también la posible existencia de subgrupos funcionales dentro de los tipos de cadenas. La distribución de los clústeres es la siguiente:

- Clúster 1: agrupó mayoritariamente cadenas pesadas, mostrando consistencia estructural interna y características bioquímicas similares.
- Clúster 0 y clúster 2: ambas agrupan cadenas ligeras, pero con diferencias en su estructura.

Esta segmentación logra una separación biológicamente entre los tipos de cadenas, lo que muestra que la representación basada en factores Atchley es capaz de capturar información estructural relevante.

La diferencia entre los clústeres que están conformados por las cadenas ligeras se encuentra en la región CDR1. Estas diferencias se reflejan en propiedades como:

- Longitud de CDR1
- Distribución de carga y volumen molecular (factores F3 y F5)
- Variabilidad en distintos sectores de las cadenas ligeras.

Al utilizar el valor de k = 3 se tienen la capacidad de subdividir las cadenas ligeras en función de propiedades fisicoquímicas.

La figura 5.3, muestra la división entre las cadenas, en donde se observa la diferencia entre las cadenas pesadas y los subgrupos de las cadenas ligeras.

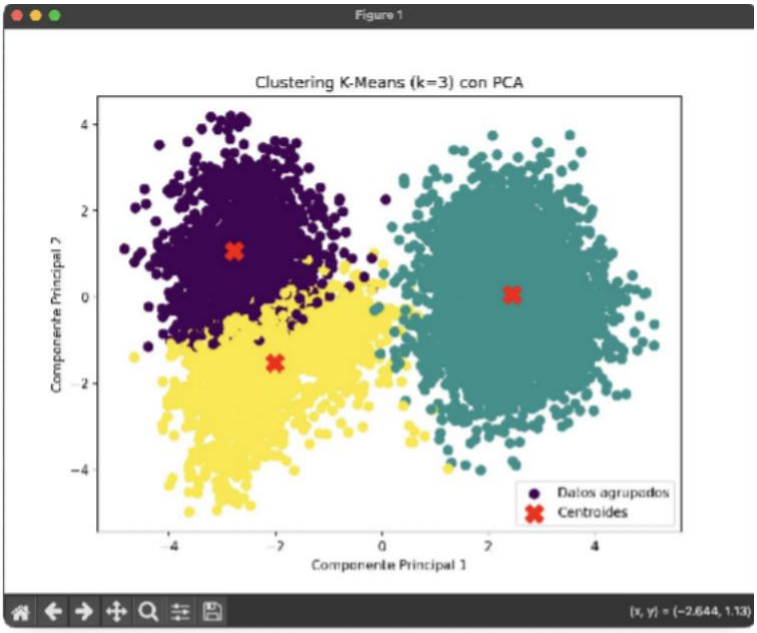

Figura 5.3. KMeans k = 3

Para $k = 4$ mostró una separación de las cadenas pesadas y ligeras, generando subgrupos estructuralmente diferenciados dentro de cada tipo. La distribución de los clústeres es la siguiente:

- Clústeres 0 y 2: cadenas ligeras, con diferencias en las frameworks especialmente en el FR3
- Clústeres 1 y 3: cadenas pesadas, con diferencias en propiedades del CDR3 (longitud) y CDR1 (carga)

Con esta separación entre clústeres parecía mostrar la posibilidad de inferir relaciones de unión entre clústeres, lo que tendría un gran valor para el complimiento del objetivo del análisis. En la figura 5.4, se muestra la división de los datos entre clústeres.

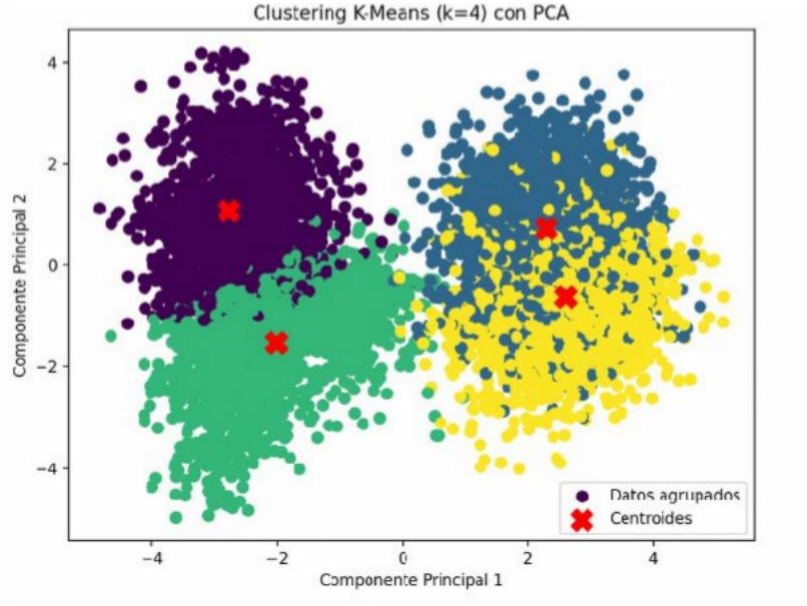

Figura 5.4. KMeans k = 4

Inicialmente, se planteó que, al existir clústeres definidos por el tipo de cadena, se podrían relacionar los clústeres de cadenas pesadas con los clústeres de cadenas ligeras como posibles pares funcionales. Por ejemplo, si una cadena ligera pertenece al clúster 0 y tiene relación con las cadenas del clúster 1 (en cuanto a propiedades complementarias), podría inferirse una compatibilidad.

Al obtener el número de cadenas que ligeras que se relacionan con los clústeres de cadenas pesadas se puede observar que cada uno de los clústeres de cadena ligera se relaciona con todos los clústeres de cadena pesada. Por lo que no se encontró un patrón importante entre los clústeres de los tipos de cadenas. La figura 5.5, muestra el mapa de calor que representa las relaciones de pares funcionales entre clústeres de cada tipo de cadena.

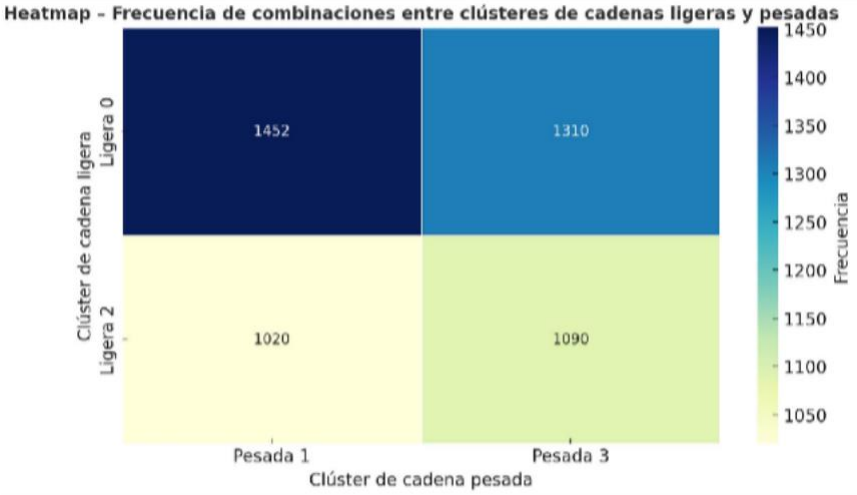

Figura 5.5. Relaciones de pares funcionales entre clústeres de tipos de cadenas

5.2.2. Agrupamiento jerárquico

El clustering jerárquico aglomerativo se utilizó como una técnica complementaria para explorar la organización de las cadenas pesadas y ligeras utilizando la representación basada en los factores Atchley. En comparación con KMeans, este método no requiere definir un número fijo de clústeres para construir una jerarquía de agrupamiento basada en la similitud de los vectores. Este enfoque es útil como herramienta exploratoria, ya que permite visualizar cómo se agrupan progresivamente las secuencias y detectar patrones globales de organización.

La figura 5.6, muestra el dendrograma que genera el modelo, en donde se observan dos agrupamientos principales, uno en color naranja y otro en color verde, que dividen el conjunto de datos en dos ramas principales.

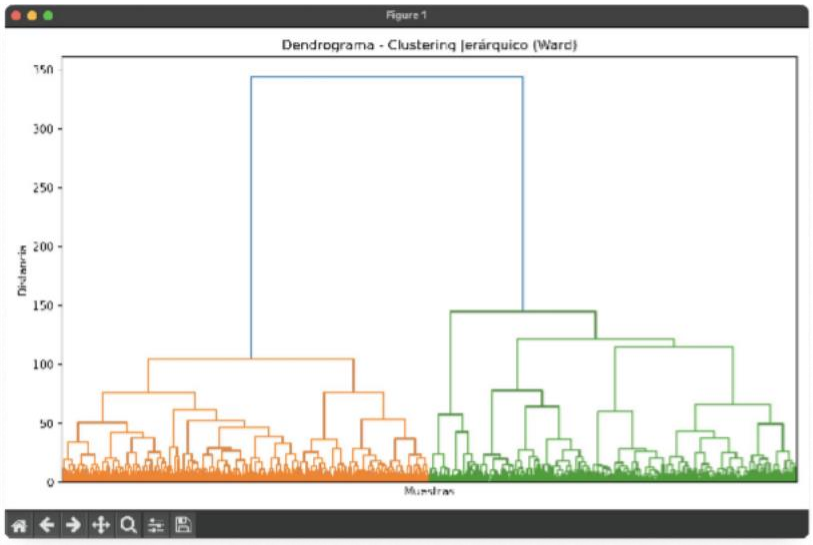

Figura 5.6. Dendrograma clustering jerárquico

La división principal de los datos corresponde a la división entre cadenas pesadas y ligeras. Este resultado confirma que los factores Atchley permiten capturar diferencias estructurales entre los tipos de cadenas. Sin embargo, al descender en el dendrograma y analizar las ramificaciones internas, no se identifican subgrupos funcionales consistentes basados en las regiones relevantes. Las ramificaciones son numerosas, poco compactas y no contienen patrones estructurales evidentes.

No se generan agrupamientos simétricos entre ramas de las cadenas ligeras y ramas de las cadenas pesadas que pudieran generar relaciones funcionales o emparejamientos de las cadenas. La figura 5.7, muestra el diagrama de calor de las relaciones encontradas entre clústeres.

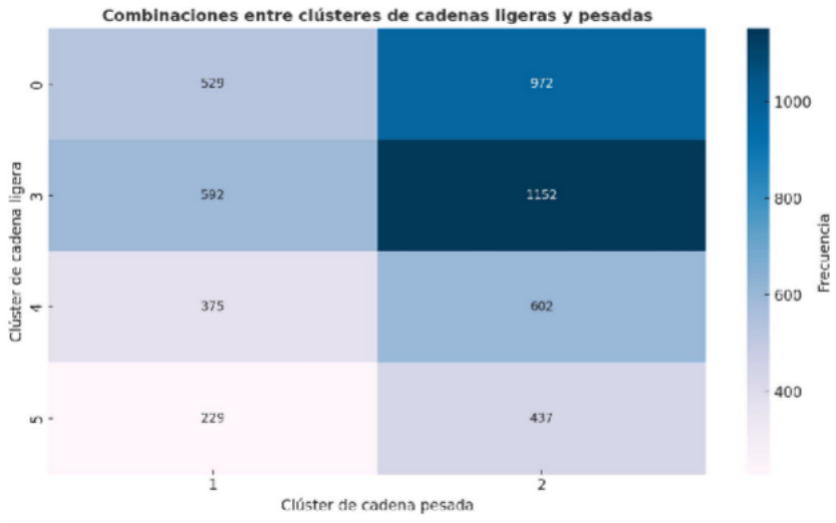

Figura 5.7. Relaciones de pares funcionales entre clústeres de tipos de cadenas clustering jerárquico

El clustering jerárquico aglomerativo demostró ser útil para confirmar una separación global por el tipo de cadena, pero no permite obtener una segmentación funcional interna confiable. La estructura del dendrograma revela agrupamientos extensos y dispersos, sin evidencia de complementariedad o correlación entre cadenas ligeras y pesadas.

Este método debe considerarse una herramienta exploratoria, útil para observar relaciones generales de similitud, pero insuficiente para tareas analíticas avanzadas como la predicción de compatibilidades funcionales entre cadenas.

5.2.3. Redes de Kohonen

Como parte del análisis exploratorio de similitud estructural entre las cadenas de anticuerpos, se implementaron redes autoorganizadas de Kohonen (Self-*Organizing Maps*, SOM) con el objetivo de visualizar la distribución de las secuencias representadas con los factores Atchley. Estas redes permiten proyectar los vectores de alta dimensión en una cuadricula bidimensional, preservando relaciones entre muestras.

A diferencia de métodos como KMeans, este método no define clústeres de forma explícita, sino que estructura el espacio de entrada en zonas de activación, lo cual es útil para identificar patrones.

La figura 5.8, muestra la matriz de distancias promedio entre nodos. Las regiones oscuras delimitan zonas de transición topológica, mientras que las zonas claras representan los conjuntos de vectores similares. Esto sirvió para confirmar que los factores Atchley permiten organizar las cadenas pesadas y ligeras con cierta coherencia.

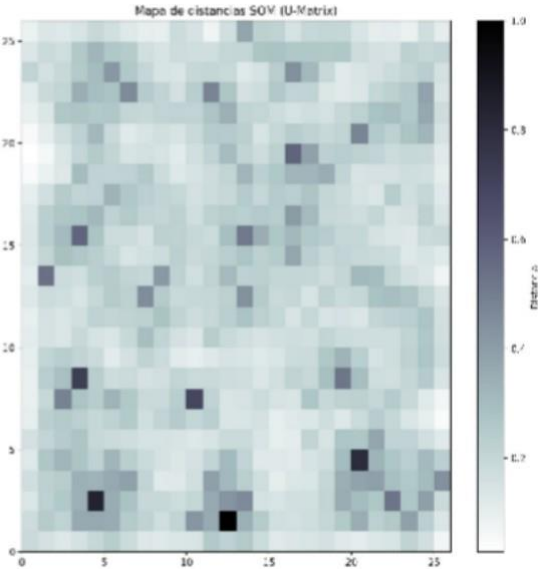

Figura 5.8. Mapa de distancias SOM

Después se aplicó KMeans con k = 4 sobre las ubicaciones neuronales que se clasifican como ganadoras. El análisis reveló agrupamientos definidos para cada clúster, con regiones predominantes para cada clúster. Esta segmentación se visualiza utilizando mapas de calor, en donde cada clúster mostró zonas específicas de alta densidad, aunque existe cierto solapamiento. En la figura 5.9, se muestran los mapas de calor para cada uno de los clústeres generados.

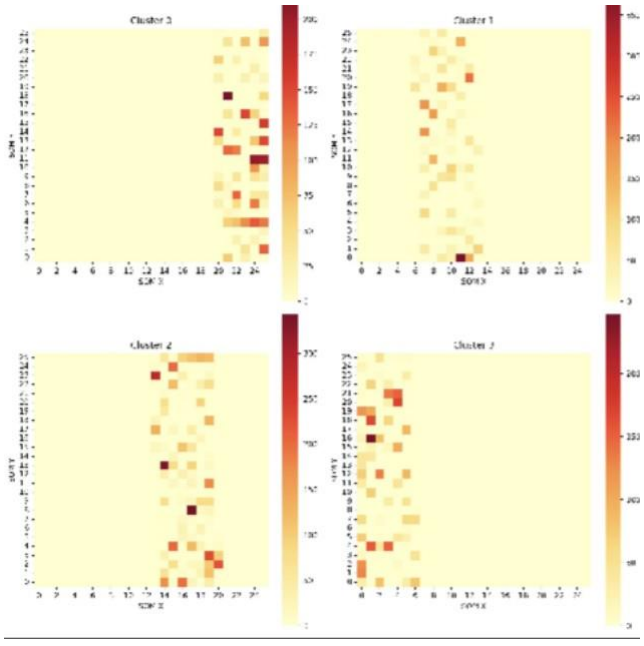

Figura 5.9. Mapas de calor SOM

Para investigar si los clústeres generados estaban organizando por tipo de cadena, se analizó cada clúster para observar cómo se distribuyen los datos. Los resultados mostraron que:

- Ningún clúster contiene exclusivamente un tipo de cadena
- El clúster 2 de las cadenas pesadas fue el más frecuente entre pares de cadenas que se unen
- La distribución es relativamente balanceada, lo cual indica que el SOM no separó explícitamente por tipo de cadena, sino por similitud estructural compartida.

El enfoque SOM + KMeans es una herramienta exploratoria robusta, pero no es suficiente para inferir compatibilidades funcionales entre cadenas. Para ello, se requiere un modelo supervisado específico, como las redes siamesas con pérdida de tripleta.

5.2.4. Conclusiones del análisis exploratorio

El análisis exploratorio utilizando algoritmos de clustering no supervisado proporcionó una perspectiva sobre la organización interna de las cadenas pesadas y ligeras de los anticuerpos representadas utilizando los factores Atchley. Esta representación que está basada en propiedades fisicoquímicas y estructurales permitió utilizar técnicas de agrupamiento para identificar patrones en los datos sin etiquetas.

El algoritmo KMeans permitió identificar estructuras diferenciadas en los datos. Con $k = 3$, se obtuvo una división entre las cadenas pesadas y ligeras, mientras que con $k = 4$, se observaron subdivisiones funcionales, con variaciones estructurales en regiones específicas como CDR1 y FR3. El intento de inferir relaciones funcionales entre clústeres de cadenas pesadas y ligeras no tuvo éxito, ya que los emparejamientos no presentaron patrones consistentes entre los tipos de cadenas.

El clustering jerárquico confirmó la separación global entre tipos de cadena, evidenciada en las ramas principales del dendrograma. Sin embargo, las divisiones internas de los tipos de cadenas tuvieron una baja cohesión funcional y no se generaron agrupamientos claros con los que se pudiera inferir emparejamientos entre las cadenas.

Las redes SOM implementadas con KMeans, facilitan la visualización del espacio estructural de las cadenas. Mientras que los mapas de calor facilitan identificar las zonas de alta densidad para los clústeres, pero al analizar los pares de cadenas se encontró que no existe una relación exclusiva entre clústeres opuestos. A pesar de una mayor recurrencia del clúster 2 de cadenas pesadas, la asociación fue difusa y no determinante.

En conjunto, los resultados indican que los algoritmos de clustering aplicados sobre la representación de factores de Atchley logran identificar patrones internos de similitud y subdivisiones estructurales relevantes, pero no permiten inferir compatibilidades funcionales entre cadenas.

Este análisis exploratorio, cumple un papel esencial al delimitar zonas estructuralmente coherentes, las cuales pueden ser aprovechadas en etapas posteriores mediante modelos de aprendizaje supervisado, como las redes siamesas con pérdida de tripleta. Dichos modelos tienen el potencial de aprender patrones de emparejamiento funcional directamente, utilizando como base las agrupaciones y características reveladas durante esta fase inicial.

5.3. Diseño del modelo

El modelo construido es una red neuronal siamesa asimétrica, diseñada para la predicción del emparejamiento de cadenas pesadas y ligeras de anticuerpos. Este tipo de arquitectura permite comparar dos entradas independientes utilizando un espacio vectorial compartido, en donde la similitud entre las representaciones se hace por medio de la distancia coseno entre sus *embeddings*.

A diferencia de una red siamesa simétrica que utiliza los mismos pesos para ambas ramas de entrada, en este caso se utilizan pesos distintos en cada rama. Esta decisión se tomó considerando que las cadenas pesadas y ligeras presentan diferencias estructurales significativas, observadas durante el análisis exploratorio de los datos, por lo que procesarlas con *encoders* separados permite que cada rama pueda aprender representaciones especializadas y adaptadas a las características propias de cada tipo de cadena.

El modelo fue entrenado utilizando la función de perdida InfoNCE, que permite aprovechar todos los pares del batch de entrenamiento como negativos de forma eficiente, aportando un comportamiento equivalente a la minería de negativos semi-*hard* mediante el parámetro de temperatura τ=0.07. La figura 5.10, muestra la arquitectura completa del modelo.

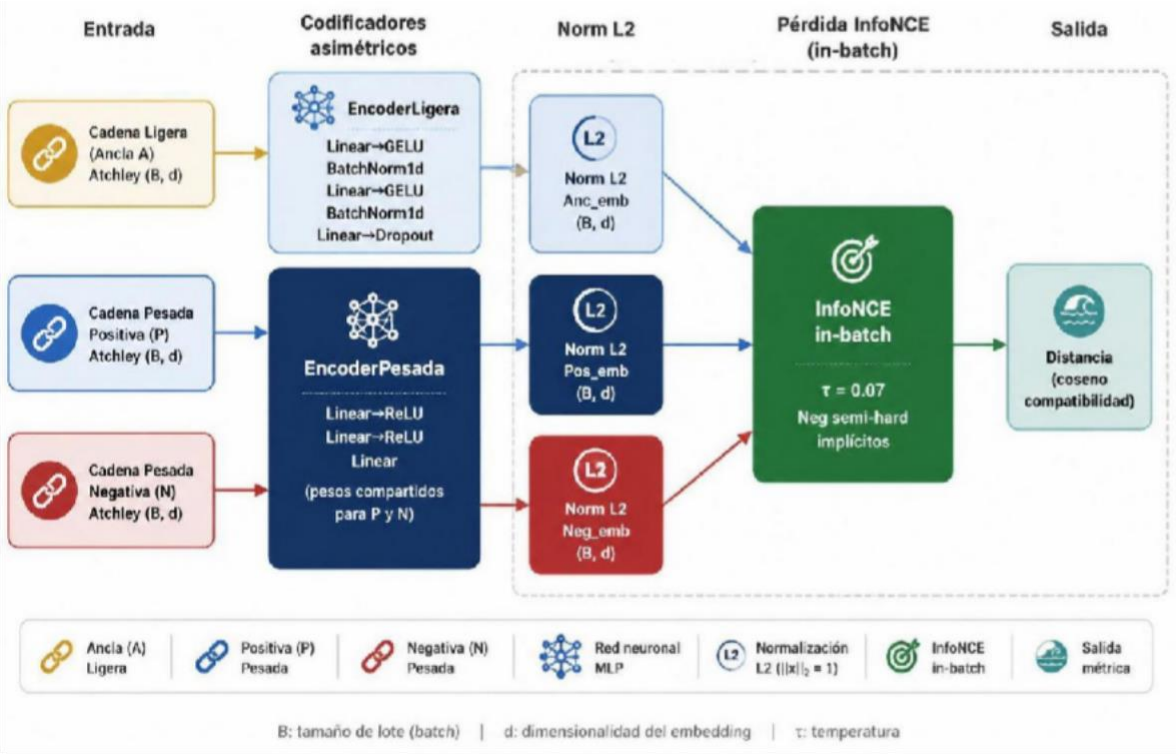

Figura 5.10. Arquitectura de la red siamesa asimétrica con InfoNCE in-batch

5.3.1. Estructura general del modelo

La estructura siamesa asimétrica se conforma de dos *encoders* independientes:

- Encoder para cadenas ligeras (EncoderLigera): red neuronal densa que se compone de tres capas sucesivas, utilizando GELU, con normalización por lotes (BatchNorm1d) y regularización por Dropout
- Encoder para cadenas pesadas (EncoderPesada): arquitectura distinta con activaciones ReLU y dos capas densas intermedias. Tiene mayor profundidad para reflejar la complejidad estructural que tienen las cadenas pesadas en comparación con las cadenas ligeras
- Red siamesa: se compone por ambos encoders, recibe como entrada una tripleta (ancla, positiva, negativa) y produce los embeddings correspondientes para calcular la pérdida.

Los dos *encoders* proyectan sus entradas en un espacio de 8 dimensiones, lo que permite realizar comparaciones por la distancia coseno. Las secuencias de los anticuerpos se representaron utilizando los factores Atchley y se normalizaron con StandardScaler. Los dos conjuntos de datos se unificaron utilizando el identificador de la secuencia para asegurar que cada para sea funcional.

5.3.2. Encoder de cadenas ligeras

El EncoderLigera es una red neuronal densa compuesta por tres capas sucesivas. Su arquitectura incorpora normalización por lotes (BatchNorm1d) después de cada activación, lo que estabiliza el entrenamiento al normalizar las activaciones de cada capa y reduce la dependencia de la inicialización de pesos. Se utiliza la función de activación GELU (Gaussian Error Linear *Unit*), que a diferencia de ReLU aplica una transformación suavizada que considera la probabilidad de cada activación, lo que ha demostrado mejor desempeño en redes profundas con datos de alta variabilidad como las secuencias de aminoácidos. Adicionalmente se aplica regularización por Dropout para reducir el sobreajuste. La figura 5.11, muestra la estructura interna de este *encoder*.

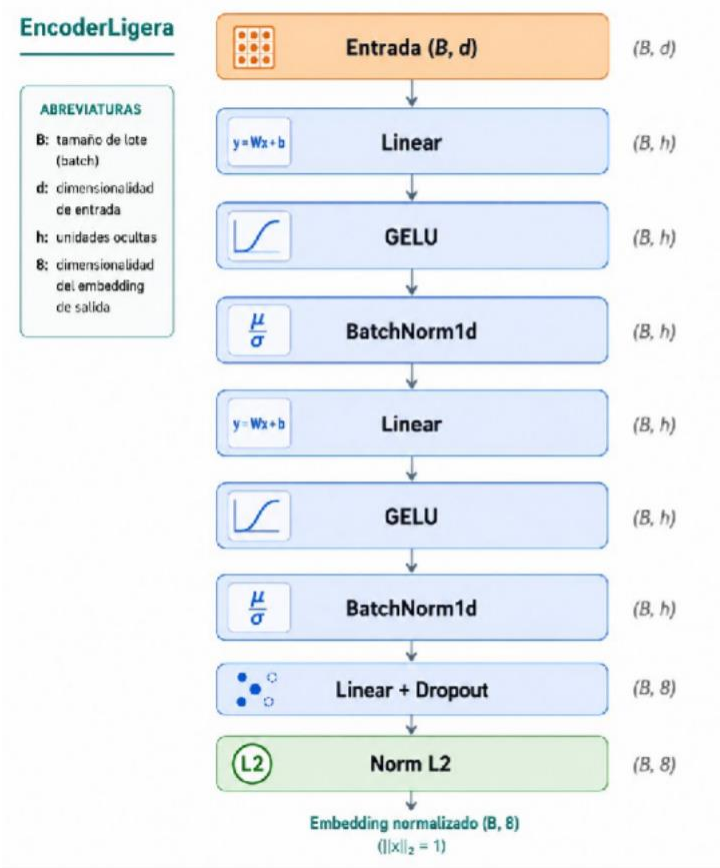

Figura 5.11. Estructura del EncoderLigera

5.3.3. Encoder de cadenas pesadas

El EncoderPesada tiene una arquitectura distinta al EncoderLigera, con activaciones ReLU y dos capas densas intermedias. Esta diferencia refleja la mayor complejidad estructural observada en las cadenas pesadas durante el análisis exploratorio: al tener mayor variabilidad en sus CDRs, particularmente en CDR3, se optó por una arquitectura más profunda que permita capturar patrones más complejos. Los pesos de este *encoder* son compartidos al procesar tanto las cadenas positivas como las negativas dentro de cada tripleta, garantizando que ambas sean proyectadas al mismo espacio con el mismo criterio. La figura 5.12 muestra su estructura.

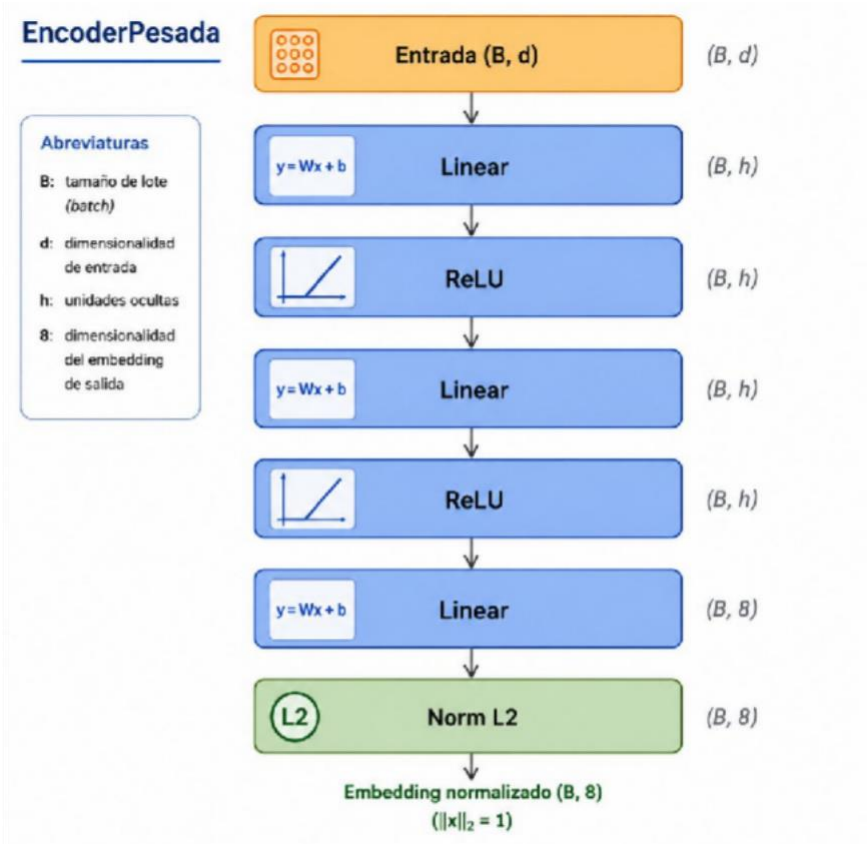

Figura 5.12. Estructura del EncoderPesada.

5.3.4. Espacio métrico compartido

Ambos *encoders* proyectan sus salidas mediante normalización L2, garantizando que todos los vectores tengan norma unitaria y se ubiquen sobre la hiperesfera unitaria. Esta normalización es fundamental para el uso de la distancia coseno como métrica de similitud: al estar todos los *embeddings* normalizados, la distancia coseno es equivalente al producto punto negativo, lo que simplifica el cálculo y estabiliza el entrenamiento.

Adicionalmente, el modelo incorpora un módulo de atención cruzada bidireccional (cross-*attention*) que opera antes de la proyección final al espacio métrico. Este módulo permite que la representación de cada cadena incorpore información de la otra durante el proceso de codificación: la cadena ligera atiende a las regiones de la cadena pesada y viceversa. El mecanismo utiliza las representaciones intermedias de ambos *encoders* como *queries* y *keys*, generando *embeddings* enriquecidos con contexto inter-cadena antes de la normalización L2. Esto permite al modelo capturar dependencias entre regiones específicas de ambas cadenas, añadiendo interpretabilidad biológica a las predicciones.

En el espacio métrico aprendido por el modelo, los *embeddings* de cadenas pesadas y ligeras que forman anticuerpos funcionales quedan cercanos, mientras que los *embeddings* de cadenas incompatibles, como los negativos BALB/c, quedan alejados. La figura 5.13 muestra esta separación en una proyección 2D del espacio embebido de 8 dimensiones.

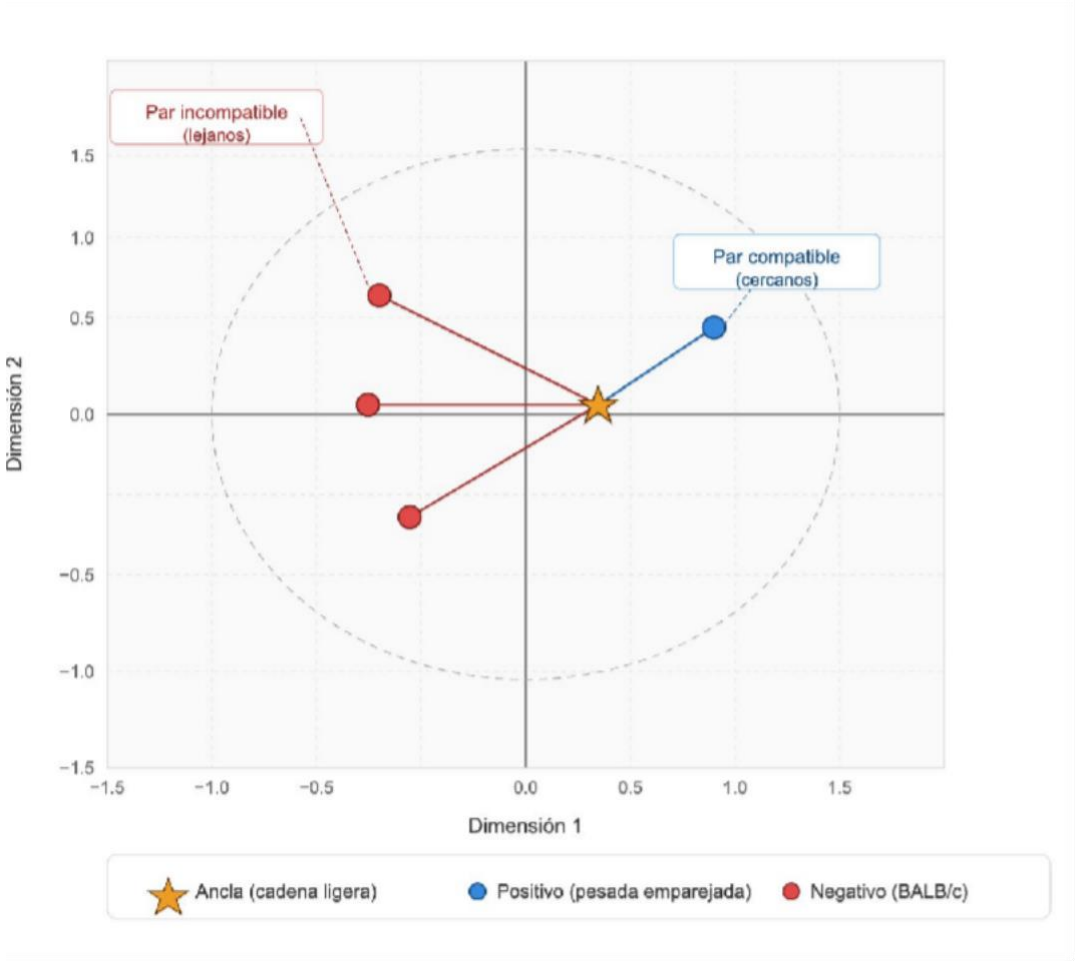

Figura 5.13. Representación del espacio métrico aprendido

En la tabla 5.4 se muestran los parámetros principales de la arquitectura del modelo.

Tabla 5.4. Parámetros de la arquitectura del modelo

Componente	Parámetro	Valor
EncoderLigera	Número de capas densas	3
EncoderLigera	Función de activación	GELU
EncoderLigera	Normalización	BatchNorm1d
EncoderLigera	Regularización	Dropout 10%
EncoderPesada	Número de capas densas	2
EncoderPesada	Función de activación	ReLU

Componente	Parámetro	Valor
EncoderPesada	Regularización	Dropout 20%
Ambos *encoders*	Dimensión del *embedding*	8
Ambos *encoders*	Normalización de salida	L2 (norma unitaria)
Espacio métrico	Métrica de similitud	Distancia coseno
Función de pérdida	Tipo	InfoNCE in-batch
Función de pérdida	Temperatura τ	0.07

5.3.5. Entrenamiento previo de los encoders

Los *encoders* fueron preentrenados como autoencoders completos, con el objetivo de que aprendieran representaciones comprimidas que conservan la mayor cantidad de información relevante sobre los dos tipos de cadenas. Este preentrenamiento garantiza que los *encoders* estén inicializados con los pesos que ya capturan la estructura interna de las secuencias, lo que acelera la convergencia y mejora la calidad del espacio métrico aprendido posteriormente.

El autoencoder para cadenas ligeras utiliza la misma arquitectura de tres capas con activaciones GELU, normalización por lotes y Dropout del 10%. La pérdida de reconstrucción se evaluó utilizando el error cuadrático medio (MSE), mostrando una curva de convergencia estable durante 100 épocas. Los resultados mostraron que los vectores reconstruidos conservan la forma general de los originales.

El autoencoder para cadenas pesadas emplea la arquitectura más profunda con activaciones ReLU y Dropout del 20%. La pérdida inicial fue alta al comienzo, pero disminuye de manera constante. Las reconstrucciones individuales reflejan alta fidelidad estructural.

La figura 5.14, muestra la evolución de la pérdida de reconstrucción (MSE *Loss*) durante el entrenamiento de los autoencoders para cadenas ligeras (línea azul) y pesadas (línea naranja) durante las 100 épocas de entrenamiento.

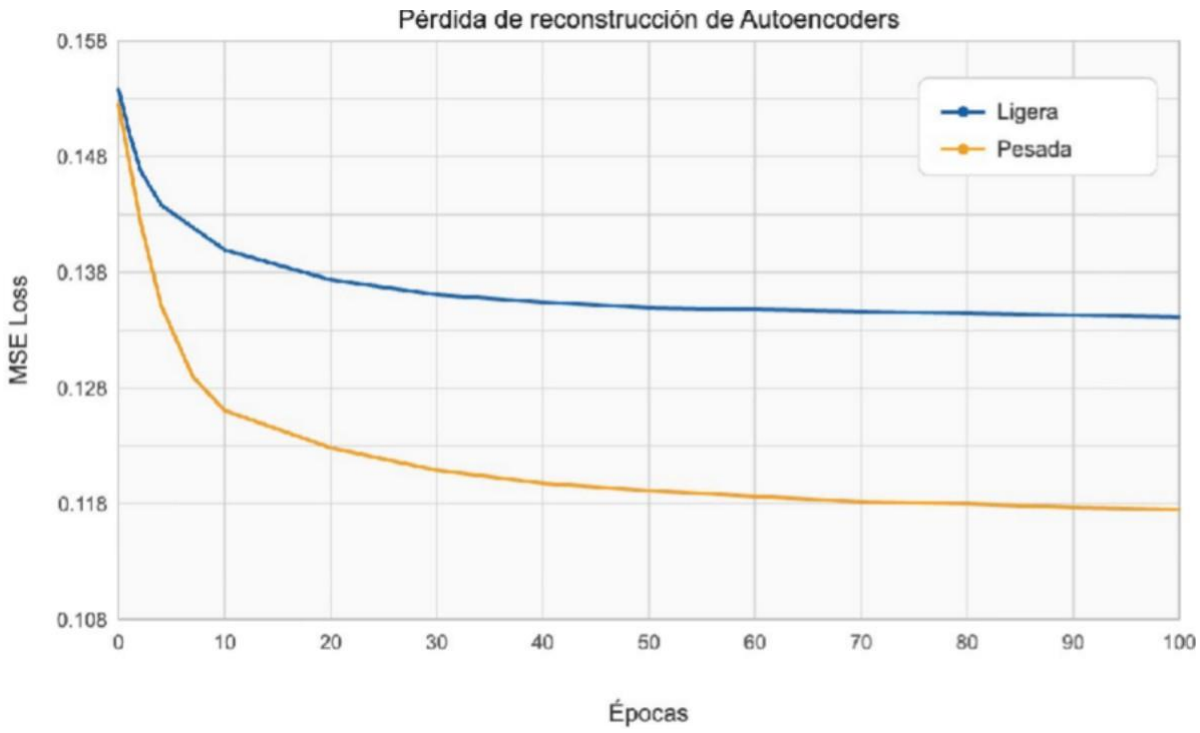

Figura 5.14. Pérdida de reconstrucción (MSE Loss) durante el preentrenamiento

En el caso de los dos tipos de cadenas, la pérdida disminuye progresivamente en las primeras épocas, lo que indica una rápida adaptación inicial a los datos. El autoencoder de las cadenas pesadas alcanza valores más bajos y se estabiliza más rápido que el de las cadenas ligeras. Como se observó en el análisis con algoritmos de clustering, el conjunto de cadenas pesadas tiene una menor variabilidad en comparación con el grupo de cadenas ligeras, lo que provoca que el autoencoder tenga un menor rendimiento al aprender los patrones de las cadenas ligeras.

La figura 5.15, muestra la reconstrucción de cinco vectores representativos de cadenas ligeras, comparando el vector original (línea azul continua) con el reconstruido por el autoencoder (línea naranja punteada).

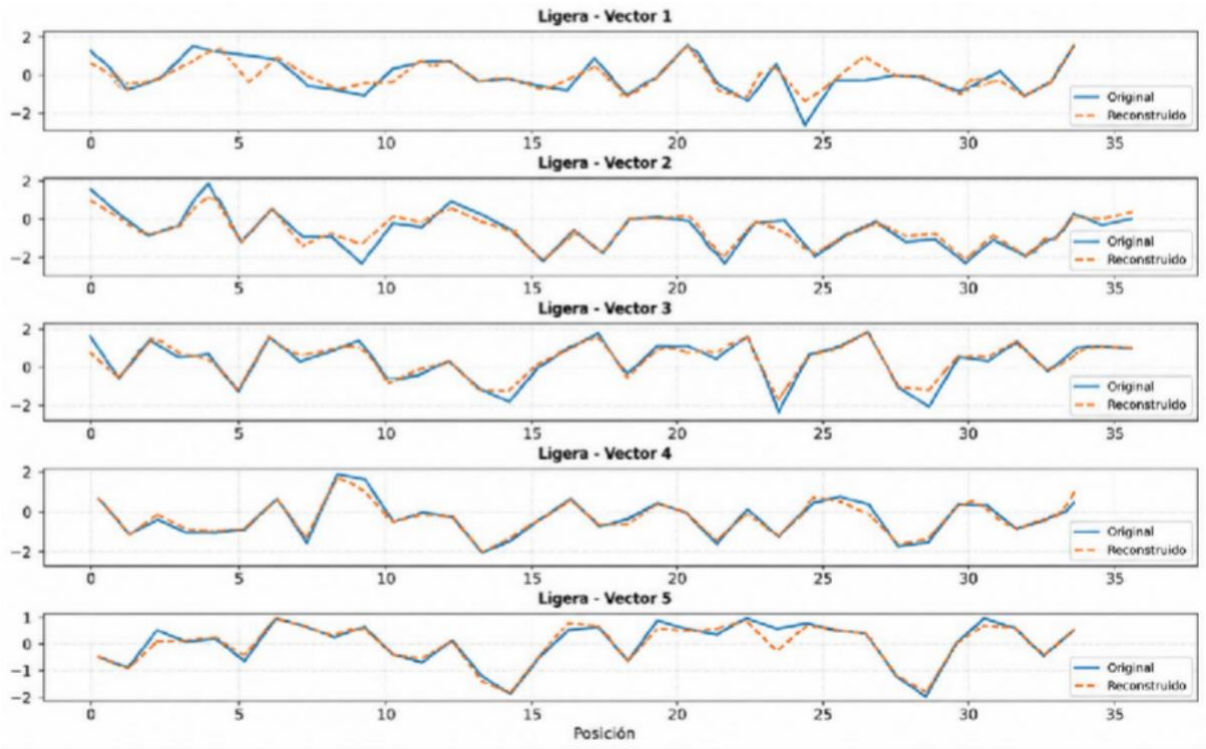

Figura 5.15. Comparación de vector original y reconstruido por el autoencoder de cadenas ligeras

La figura 5.16, muestra la comparación del vector original (línea azul continua) y el reconstruido (línea naranja punteada) por el autoencoder de cadenas pesadas. Al igual que el autoencoder de cadenas ligeras, el autoencoder de cadenas pesadas también obtiene un buen resultado al comparar ambos vectores.

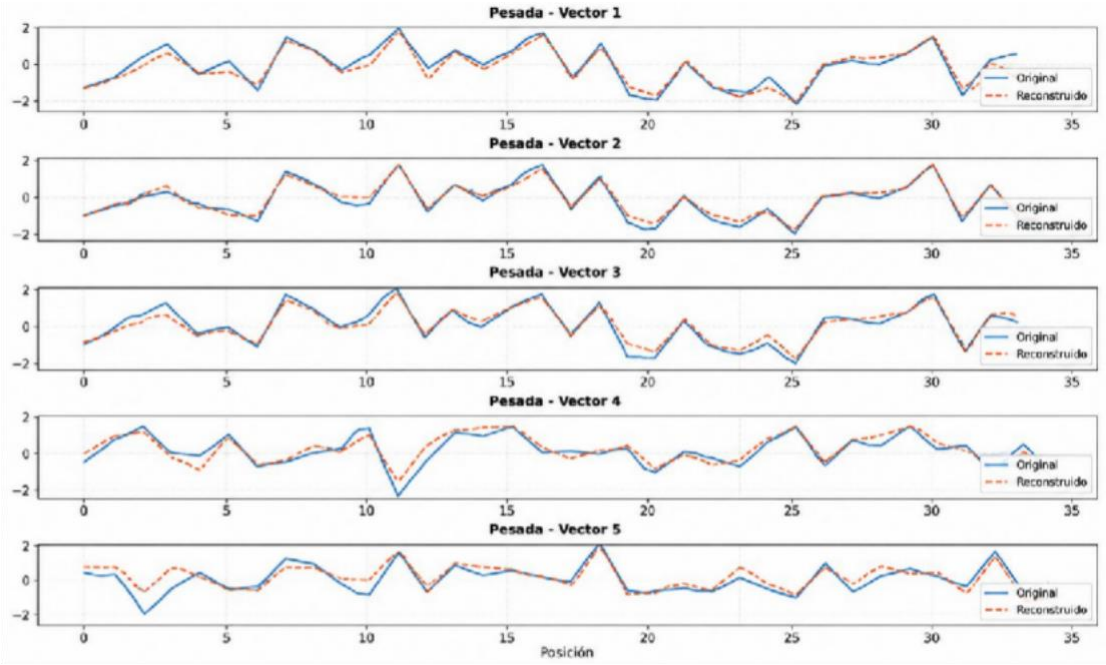

Figura 5.16. Comparación de vector original y reconstruido por el autoencoder de cadenas pesadas

5.4. Proceso de entrenamiento

El objetivo de esta etapa fue identificar la configuración que tuviera el mejor rendimiento del modelo siamés asimétrico para predecir la compatibilidad funcional entre las cadenas pesadas y ligeras de los anticuerpos. Para lograrlo, se realizaron múltiples experimentaciones utilizando distintos valores de los hiperparámetros más relevantes del modelo y analizando su impacto sobre las métricas de clasificación.

5.4.1. Entrenamiento con diferentes configuraciones

Durante el proceso experimental se ajustaron los siguientes hiperparámetros:

- Dimensión del embedding: se probaron diferentes dimensionalidades de 4, 8, 26, 36 y 64 dimensiones para evaluar el equilibrio entre la capacidad de representar las cadenas y el sobreajuste del modelo
- Activación en el encoder de cadena ligera: se compararon ReLU y GELU, seleccionando esta última por su comportamiento más estable en redes profundas
- Dropout en el encoder de cadenas pesadas: se evaluaron valores de 0.2 y 0.3 como medida de regularización
- Número de épocas: se realizaron pruebas con 50, 70 y 100 épocas para observar la evolución de la pérdida y la mejora marginal en el desempeño
- Tasa de aprendizaje (Learning Rate, LR): se utilizaron valores de 0.001 y 0.0005
- Tamaño del batch: se compararon tamaños de 26 y 32 muestras por iteración.

5.4.2. Evaluación y registro de métricas

Después del entrenamiento de cada configuración, se evaluó el modelo sobre un conjunto de validación. Se generaron las distancias coseno entre los *embeddings* de pares reales y los pares negativos, se aplicó el umbral óptimo que maximizó el F1-*Score*. Para cada experimentación se calcularon las siguientes métricas:

- Accuracy: proporción total de predicciones correctas
- Precision: proporción de pares predichos como unión que realmente se unen
- Recall: proporción de pares verdaderamente compatibles correctamente identificados
- F1-Sacore. Medida combinada de precisión y recall

En la tabla 5.5, se muestran las configuraciones de hiperparámetros y los resultados de cada una de las configuraciones con las que se experimentó.

Tabla 5.5. Configuraciones y métricas de las experimentaciones

Embedding	Activación Ligera	Dropout Pesada	Épocas	LR	Batch	*Accuracy*	*Precision*	*Recall*	F1 *Score*
4	ReLU	0.2	50	0.001	16	0.7032	0.6691	0.8012	0.7284
8	**GELU**	**0.2**	**70**	**0.001**	**16**	**0.97**	**0.98**	**0.94**	**0.96**
8	GELU	0.3	70	0.0005	32	0.7350	0.6823	0.8381	0.7486
8	GELU	0.2	100	0.001	16	0.7284	0.6902	0.8210	0.7494
8	GELU	0.2	50	0.0005	32	0.7205	0.6675	0.8042	0.7296
26	GELU	0.2	70	0.001	16	0.7345	0.6884	0.8320	0.7491
26	GELU	0.3	70	0.0005	32	0.7292	0.6827	0.8188	0.7441

Embedding	Activación Ligera	Dropout Pesada	Épocas	LR	Batch	*Accuracy*	*Precision*	*Recall*	F1 *Score*
36	GELU	0.2	70	0.001	16	0.7267	0.6811	0.8173	0.7400
36	GELU	0.3	100	0.001	16	0.7235	0.6752	0.8091	0.7368
64	GELU	0.2	70	0.001	16	0.7189	0.6703	0.8087	0.7325
64	GELU	0.3	100	0.0005	32	0.7152	0.6648	0.8054	0.7276

5.4.3. Conclusiones de las experimentaciones

La configuración con *embedding* de 8 dimensiones, activación GELU, Dropout de 0.2, 70 épocas, tasa de aprendizaje de 0.001 y batch de 16 fue la que obtuvo el mejor rendimiento, alcanzando un F1-*Score* de 0.96, el valor más alto registrado en todas las experimentaciones.

Las dimensiones más altas como 26, 36 y 64 muestran una mejora en la capacidad representativa, pero también presentan mayor variabilidad en la convergencia y métricas de precisión más bajas. Esto sugiere que el espacio embebido de 8 dimensiones ofrece el mejor balance entre capacidad de representación y generalización para el tamaño del conjunto de datos disponible.

Las experimentaciones en las que se utilizó la activación GELU con Dropout de 0.2 mostraron una consistencia general en obtener los mejores resultados a través de las distintas configuraciones evaluadas. En conjunto, las experimentaciones permitieron seleccionar la configuración final del modelo con un rendimiento robusto y balanceado, que será utilizada en la evaluación presentada en el capítulo de resultados.

Capítulo 6.
Resultados

6. Resultados

En este capítulo se presentan los resultados que se obtuvieron durante la evaluación del modelo siamés asimétrico para la predicción del emparejamiento funcional entre cadenas pesadas y ligeras de anticuerpos. Los resultados se organizan en tres secciones: la evaluación cuantitativa del modelo utilizando métricas de clasificación y recuperación, el análisis del espacio de *embeddings* aprendido y el análisis del mecanismo de atención cruzada del modelo.

La figura 6.1 muestra el flujo general de uso del modelo: a partir de dos archivos independientes con las representaciones Atchley de las cadenas pesadas y ligeras, el modelo proyecta los dos tipos de cadenas en el espacio métrico compartido y genera como salida una lista de posibles emparejamientos ordenados por su compatibilidad funcional. Este flujo es la aplicación del modelo en el análisis de repertorios inmunes, donde se parte de conjuntos de cadenas sin información de emparejamiento y se busca identificar las combinaciones con mayor probabilidad de generar anticuerpos funcionales.

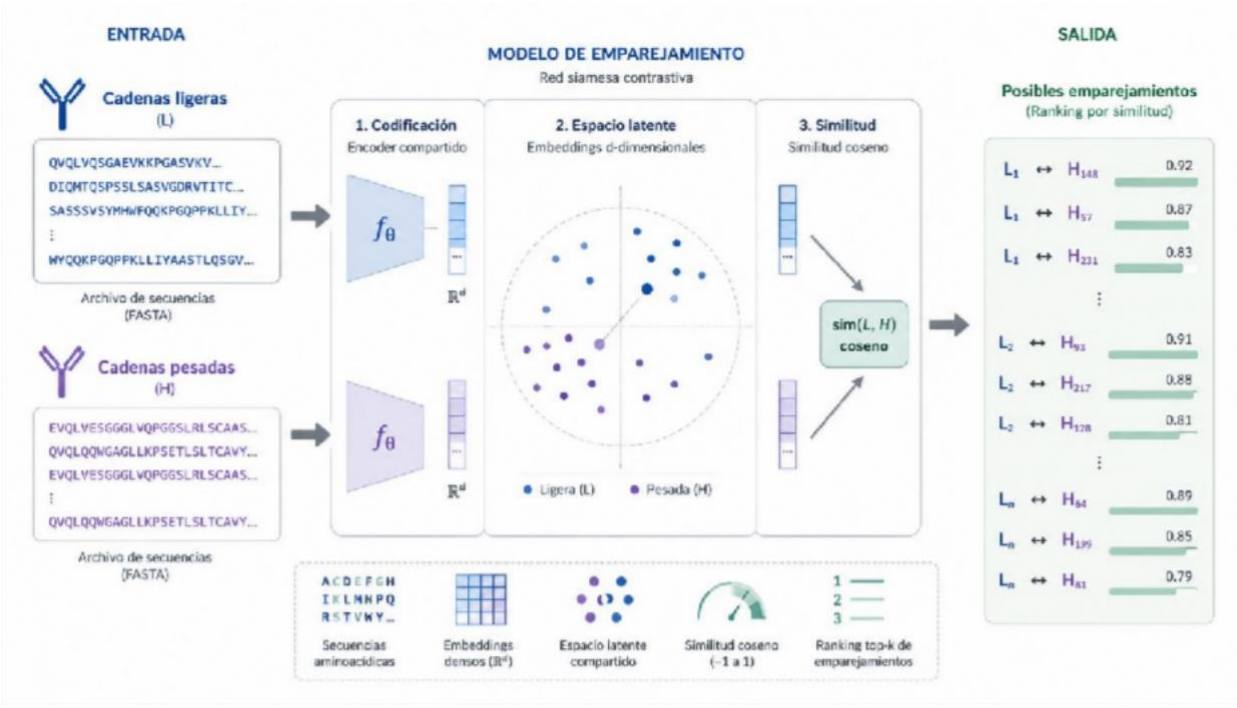

Figura 6.1. Flujo de entrada y salida del modelo

6.1. Evaluación del modelo

La evaluación del modelo se realizó sobre el conjunto de prueba que se compone de cadenas pesadas y ligeras que no fueron utilizadas durante el entrenamiento. Para caracterizar el desempeño del modelo en distintos escenarios de uso, se utilizaron dos tipos de métricas complementarias: métricas de clasificación binaria, que evalúan la capacidad del modelo para distinguir pares compatibles a partir de un umbral de distancia y métricas de

recuperación (*Recall*@K), que evalúan su capacidad para encontrar el emparejamiento correcto dentro de un ranking de candidatos.

6.1.1. Evaluación binaria con el umbral de distancia

Para la evaluación binaria, se determinó el umbral de distancia coseno que maximiza el F1-*Score* sobre el conjunto de validación. Los pares con una distancia coseno menor al umbral se clasifican como compatibles y los pares con distancia mayor son incompatibles.

Con el objetivo de mostrar visualmente el impacto de agregar los negativos verdaderos, la figura 6.2 muestra la distribución de las distancias coseno comparado con el modelo que no contempla los verdaderos negativos. En la figura se puede observar cómo evolucionó la calidad del espacio métrico aprendido en ambas versiones del modelo.

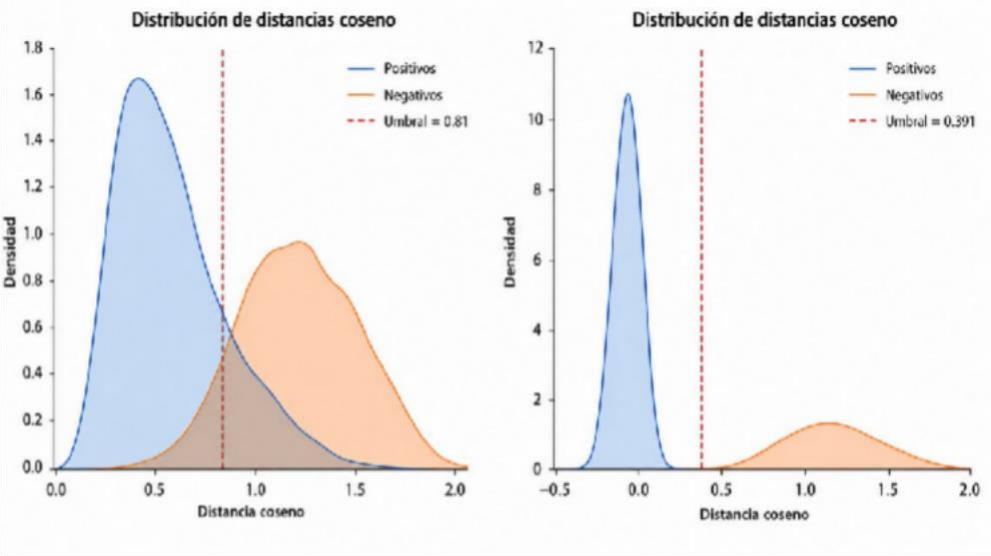

Figura 6.2. Distribución de distancias coseno entre pares positivos y negativos

En el modelo sin negativos verdaderos se observa un solapamiento considerable entre las distribuciones de pares positivos y negativos, lo que hace que el umbral se coloque en una zona con alta ambigüedad en donde se producen los falso positivos y los falsos negativos. El modelo con los negativos verdaderos muestra una separación notablemente más clara: los pares positivos se concentran con valores de distancia cercanos a cero mientras que los negativos se distribuyen en valores más altos, reduciendo la zona de solapamiento. Esta mejora es el resultado directo del uso de las cadenas pesadas de ratón BALB/c como negativos biológicamente correctos y la función de pérdida InfoNCE con temperatura τ=0.07, que juntos obligan al modelo a aprender un espacio métrico más discriminativo.

La tabla 6.1 presenta las métricas de clasificación que se obtuvieron con el umbral del modelo con negativos verdaderos. Los resultados muestran un desempeño alto en todas las métricas evaluadas, lo que confirma que el espacio métrico aprendido permite distinguir con precisión entre pares funcionales y no funcionales.

Tabla 6.1. Métricas de clasificación del modelo con los verdaderos negativos

Métrica	Valor
Accuracy	0.97
F1-*Score*	0.96
Recall	0.94
ROC-AUC	0.97
PR-AUC	0.96

6.1.2. Evaluación Top-K

La evaluación Top-K complementa las métricas de clasificación al medir la capacidad del modelo para recuperar el emparejamiento correcto dentro de los K candidatos más similares a una cadena ligera dada. A diferencia de la evaluación binaria, que requiere definir un umbral fijo, el *Recall*@K evalúa directamente la calidad del ordenamiento producido por el modelo, lo que lo convierte en la métrica más relevante para escenarios de uso práctico donde se busca generar una lista reducida de cadenas pesadas candidatas para validación experimental posterior.

La figura 6.3, muestra el comportamiento del *Recall*@K para los tres valores de K evaluados. Se puede observar que el modelo muestra una rápida convergencia conforme se aumenta el valor de K, lo que indica que el emparejamiento correcto tiende a ubicarse consistentemente entre los primeros candidatos del ranking generado por el modelo con negativos verdaderos.

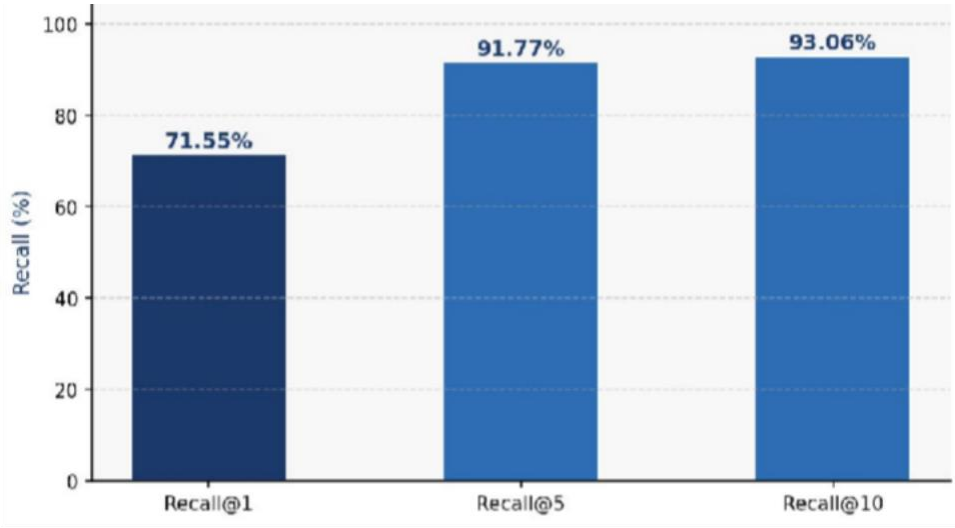

Figura 6.3. Recall@K del modelo

La tabla 6.2 presenta los valores específicos de *Recall*@K obtenidos en el conjunto de prueba. La diferencia entre el valor en K=1 y los valores en K=5 y K=10 refleja que, aunque no siempre el emparejamiento correcto es el más cercano, el modelo con negativos verdaderos lo ubica sistemáticamente en posiciones muy altas del ranking, lo que lo hace útil para reducir el espacio de búsqueda en repertorios de gran escala.

Tabla 6.2. Resultados de Recall@K

Métrica	Valor
Recall@1	71.55%
Recall@5	91.77%
Recall@10	93.06%

El comportamiento del *Recall*@K es especialmente adecuado para el problema del emparejamiento VH-VL en el contexto de secuenciación de nueva generación, donde el objetivo no es siempre identificar el único emparejamiento correcto sino reducir el espacio de cadenas candidatas a un conjunto manejable sobre el cual realizar validaciones experimentales. En este sentido, los resultados confirman que el modelo con negativos verdaderos es una herramienta viable para el análisis de repertorios inmunes de gran escala.

6.2. Análisis del espacio de *embeddings*

Para evaluar la calidad y coherencia biológica del espacio métrico aprendido por el modelo con negativos verdaderos, se realizó un análisis de los *embeddings* generados utilizando reducción de dimensionalidad UMAP. Este análisis permite verificar si las representaciones aprendidas reflejan propiedades biológicas conocidas de las cadenas de anticuerpos, añadiendo interpretabilidad a las predicciones más allá de las métricas cuantitativas.

6.2.1. Organización biológica del espacio aprendido

Una forma de validar que un modelo de aprendizaje métrico aprendió representaciones biológicamente significativas es verificar si el espacio embebido organiza las secuencias de acuerdo con propiedades biológicas conocidas, sin que estas propiedades fueran incluidas explícitamente como información de entrenamiento. Para este análisis se proyectaron los *embeddings* generados por el modelo con negativos verdaderos en dos dimensiones mediante UMAP y se colorearon con tres propiedades biológicas independientes: la familia germinal V de la cadena pesada, el gen V específico y la longitud del CDR3-HC.

La figura 6.4 muestra el resultado de esta proyección. Si el espacio métrico aprendido captura propiedades biológicas relevantes, se esperaría observar agrupaciones coherentes con cada una de las propiedades de color, lo que indicaría que el modelo las infirió implícitamente a partir de las secuencias de aminoácidos durante el entrenamiento con tripletas.

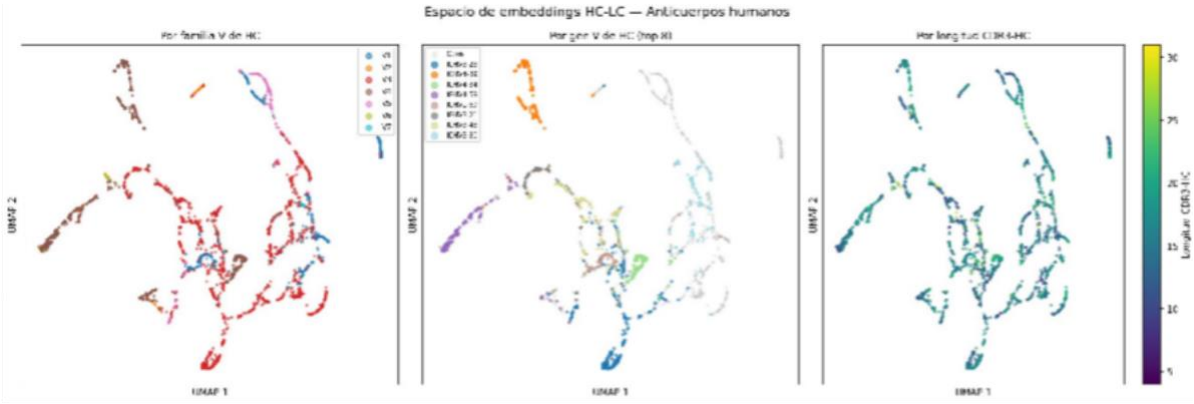

Figura 6.4. Proyección UMAP del espacio de embeddings HC-LC

Los tres paneles muestran una organización coherente del espacio embebido con respecto a las propiedades biológicas utilizadas para el coloreado. En el panel izquierdo se observa que las cadenas de la misma familia V tienden a agruparse en regiones específicas del espacio, lo que indica que el modelo aprendió a capturar las diferencias estructurales entre familias germinales presentes en los datos de entrenamiento y las proyectó de forma organizada en el espacio métrico. El panel central revela que genes V específicos forman agrupaciones diferenciadas, consistente con las diferencias en la composición de aminoácidos de sus CDRs, lo que refleja que el modelo aprendió representaciones que preservan la especificidad génica de las cadenas pesadas. El panel derecho muestra un gradiente de longitud de CDR3-HC a lo largo del espacio embebido, lo que indica que el modelo incorporó la diversidad de longitud de esta región, la de mayor variabilidad estructural en los anticuerpos, como un factor relevante en la organización del espacio métrico aprendido.

Estos resultados son especialmente relevantes cuando se contrastan con el análisis exploratorio del capítulo 5. En ese análisis, los algoritmos de clustering no supervisado lograron separar cadenas pesadas de ligeras, pero no pudieron organizar las cadenas por propiedades biológicas específicas de forma consistente. El espacio métrico aprendido por el modelo con negativos verdaderos, en cambio, logra una organización que refleja propiedades inmunológicamente significativas, lo que confirma que el aprendizaje supervisado con tripletas captura información estructural que el clustering no supervisado no pudo extraer.

6.2.2. Separación entre pares nativos y pares desplazados

Un aspecto fundamental para validar el espacio métrico aprendido es verificar que los pares que forman anticuerpos reales, denominados paras nativos, sean proyectados de forma diferente en comparación a las combinaciones de cadenas de anticuerpos distintos, denominados pares desplazados. Esta distinción es precisamente la tarea que el modelo con negativos verdaderos aprendió durante el entrenamiento con tripletas, por lo que su visualización en el espacio UMAP permite evaluar de forma cualitativa si el espacio embebido refleja este aprendizaje.

La figura 6.5 muestra la proyección UMAP distinguiendo entre pares nativos y pares desplazados. Se busca observar si los pares nativos tienden a ubicarse en posiciones específicas dentro del espacio embebido que los distingan de las combinaciones artificiales.

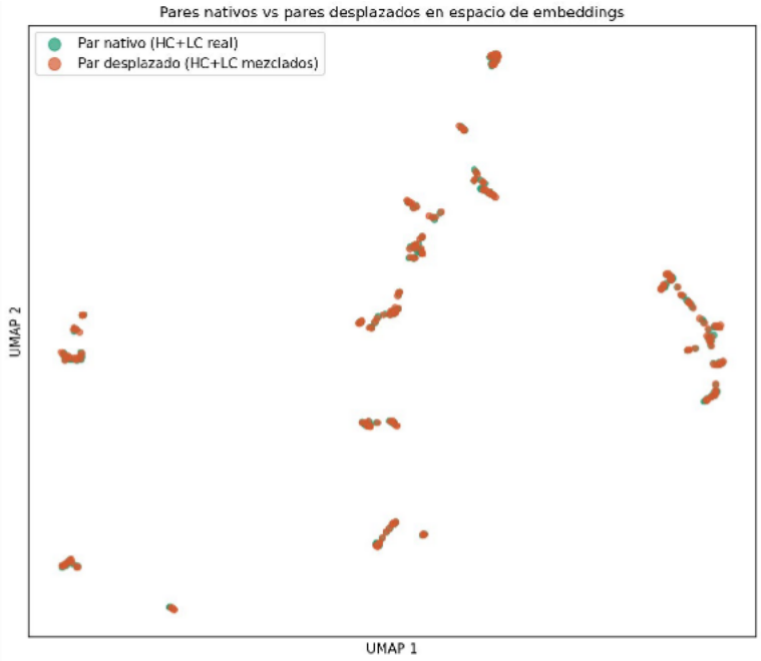

Figura 6.5. Proyección UMAP de pares nativos (verde) y pares desplazados (naranja) en el espacio de embeddings

La figura muestra que los pares nativos y desplazados forman agrupaciones locales en el espacio embebido, con los pares nativos ubicados en posiciones consistentemente específicas dentro de cada clúster. La proximidad entre ambos tipos de pares en la proyección UMAP refleja que el modelo no los separa en regiones completamente distintas del espacio global, sino que aprende distancias relativas entre ellos que permiten la clasificación mediante el umbral de distancia coseno. Este comportamiento es consistente con el objetivo del entrenamiento con InfoNCE: aprender distancias relativas informativas dentro de cada batch, más que separaciones absolutas en el espacio completo.

6.3. Análisis del mecanismo de atención cruzada

El modelo con negativos verdaderos incorpora un módulo de cross-*attention* bidireccional que permite que la representación de cada cadena incorpore información de la otra durante el proceso de codificación. El análisis de los pesos de atención aprendidos permite identificar qué regiones de las cadenas pesadas y ligeras el modelo consideró más relevantes para determinar la compatibilidad funcional, añadiendo interpretabilidad biológica a las predicciones más allá de las métricas cuantitativas.

6.3.1. Importancia de regiones en el cross-attention

Para entender que regiones de las cadenas pesadas y ligeras tienen una mayor importancia en el proceso de evaluación de compatibilidad aprendido por el modelo con negativos

verdaderos, se calcularon los pesos de atención acumulados por región para cada dirección del cross-*attention*. Esta información permite identificar si el modelo aprendió a enfocarse en regiones con relevancia biológica conocida, como los CDR's responsables del contacto directo con el antígeno, o si distribuyo la atención de forma entre todas las regiones.

La figura 6.6, muestra los pesos de atención acumulados separados por región: la atención recibida por cada región de la cadena ligera desde la cadena pesada, y la atención enviada por cada región de la cadena pesada hacia la cadena ligera. La comparación entre ambos paneles permite entender la asimetría en el flujo de información entre las dos cadenas.

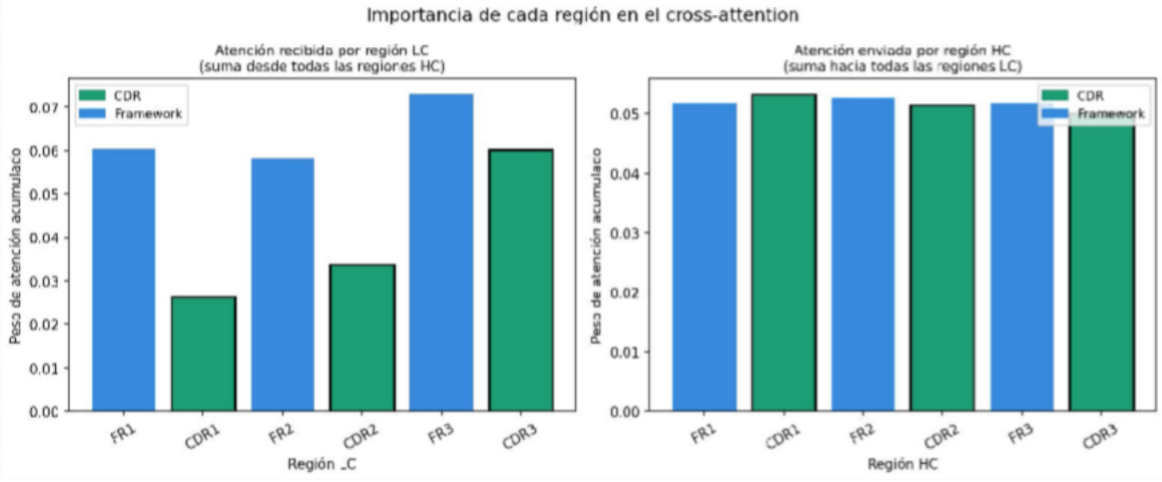

Figura 6.6. Pesos de atención acumulados por región en el cross-attention del modelo

El panel izquierdo revela que no todas las regiones de la cadena ligera reciben la misma atención desde la cadena pesada. Las diferencias entre los pesos de los CDRs y los *Frameworks* de LC reflejan que el modelo aprendió a priorizar ciertas regiones sobre otras al evaluar la compatibilidad, lo que es biológicamente coherente dado que distintas regiones tienen distintos roles en la interfaz VH-VL. El panel derecho muestra que la atención enviada por las regiones de HC es más uniforme, lo que sugiere que todas las regiones de la cadena pesada contribuyen de forma relativamente equilibrada al proceso de evaluación de compatibilidad con la cadena ligera.

6.3.2. Patrones de atención HC ↔ LC por región IMGT

El análisis de importancia por región muestra el peso total de cada región, pero no revela cómo se distribuye la atención entre cada par específico de regiones HC-LC. Para obtener una visión más detallada de los patrones de interacción aprendidos por el modelo con negativos verdaderos, se calcularon los mapas de calor de los pesos de cross-*attention* promedio para cada combinación de región HC y región LC, en ambas direcciones.

La figura 6.7 presenta estos mapas de calor. Cada celda representa el peso de atención promedio entre un par específico de regiones, lo que permite identificar qué regiones de HC atienden preferentemente a qué regiones de LC y viceversa. Si el modelo aprendió patrones biológicamente coherentes, se esperaría que las regiones con mayor relevancia en la interfaz VH-VL mostraran pesos de atención sistemáticamente más altos.

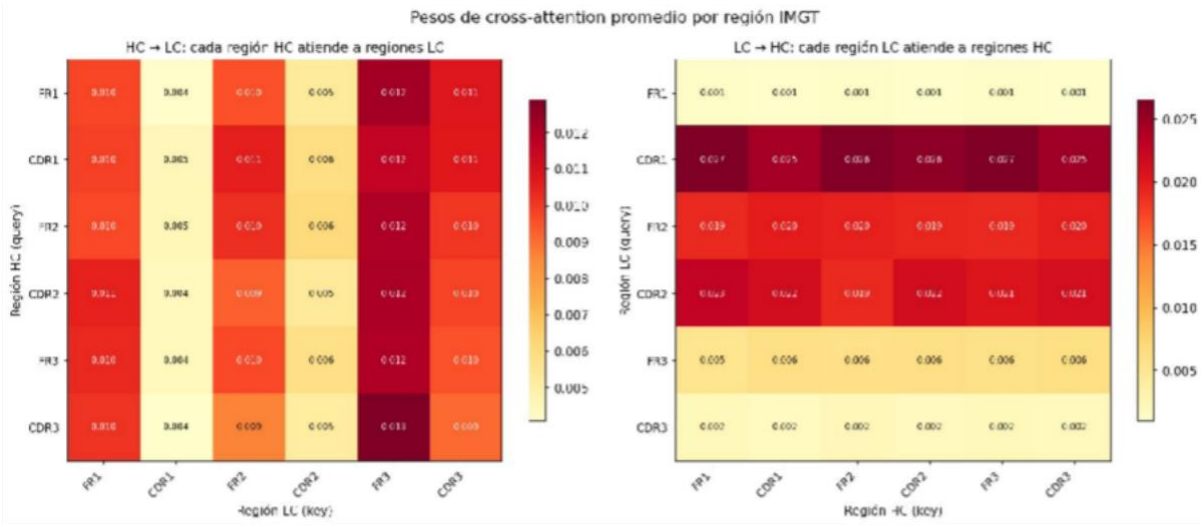

Figura 6.7. Mapas de calor de pesos de cross-attention promedio por par de regiones IMGT

El mapa de calor HC→LC muestra que algunas regiones de LC concentran consistentemente más atención desde todas las regiones de HC, independientemente de la región origen. Este patrón uniforme entre filas sugiere que FR3-LC actúa como región receptora central en la compatibilidad VH-VL, siendo consultadas de forma prioritaria por toda la cadena pesada al evaluar la compatibilidad. En particular, FR3-LC recibe la mayor atención acumulada desde HC (~0.072), seguida de FR1-LC y CDR3-LC (~0.060), mientras que CDR1-LC es la región menos consultada (~0.027), lo que sugiere que los *frameworks* estructurales de la cadena ligera juegan un papel relevante en la interfaz VH-VL.

El mapa de calor LC→HC muestra un patrón distinto con mayor variabilidad entre regiones. CDR1-LC y CDR2-LC son las regiones que envían notablemente más atención hacia HC (valores promedio de ~0.026 y ~0.023 respectivamente), mientras que FR1-LC y CDR3-LC envían valores mínimos (~0.001 y ~0.002). Esto indica que el modelo aprendió una asimetría funcional: CDR1-LC y CDR2-LC actúan principalmente como emisoras de información hacia HC, mientras que FR3-LC actúa principalmente como receptora.

En conjunto, el análisis del mecanismo de atención cruzada revela que el modelo con negativos verdaderos aprendió patrones de interacción biológicamente coherentes entre las regiones de las cadenas pesadas y ligeras. La asimetría observada entre los patrones HC→LC y LC→HC es consistente con las diferencias estructurales entre ambos tipos de cadenas descritas en el capítulo 2, y valida que la arquitectura asimétrica con *encoders* independientes y cross-*attention* bidireccional fue una elección adecuada para capturar estas interacciones.

6.4. Discusión

Los resultados mostrados en este capítulo confirman la hipótesis planteada al inicio de la investigación. La incorporación de cadenas pesadas de ratón BALB/c como negativos biológicamente fundamentados demostró tener un impacto positivo y medible en la calidad del espacio métrico aprendido por el modelo, produciendo una separación más clara entre

pares compatibles e incompatibles en comparación con el modelo sin negativos verdaderos. Este resultado valida la premisa central del trabajo: que la calidad biológica de los negativos de entrenamiento determina directamente la capacidad discriminativa del modelo.

Una de las contribuciones más relevantes de este trabajo es demostrar que la elección de los negativos de entrenamiento no es una decisión trivial en el contexto del emparejamiento VH-VL. Los enfoques basados en negativos aleatorios o artificiales introducen ambigüedad porque no garantizan que los pares considerados negativos sean verdaderamente incompatibles desde el punto de vista biológico. Al utilizar cadenas pesadas de una especie distinta con incompatibilidad documentada, se establece una frontera clara y fundamentada en el espacio métrico que el modelo puede aprender de forma más efectiva.

Además, el estudio del espacio de *embeddings* demuestra que el modelo no solamente obtiene la capacidad de distinguir pares compatibles de incompatibles, sino también de estructurar el espacio de representaciones de forma coherente con rasgos inmunológicos conocidos como la longitud del CDR3, el gen V y la familia germinal. Este nivel de organización, en conjunto con los patrones de atención cruzada que tienen coherencia biológica y que se identifican en la sección 6.3, indica que el modelo ha aprendido principios estructurales del emparejamiento VH-VL más allá de solamente recordar los pares del grupo de entrenamiento.

Capítulo 7.
Conclusiones y trabajo futuro

7. Conclusiones y trabajo futuro

En este capítulo se presentan las conclusiones del trabajo de investigación y los trabajos futuros que se generan a partir de los resultados. Las conclusiones resumen las principales contribuciones del trabajo, comparan los resultados con la hipótesis que se planteó al inicio de la investigación y se destacan los hallazgos sobre el emparejamiento funcional de cadenas pesadas y ligeras de anticuerpos. Los trabajos futuros destacan las líneas de investigación que se pueden seguir para continuar con el desarrollo y fortalecer el modelo propuesto.

7.1. Conclusiones

Este trabajo presentó el diseño, implementación y evaluación de un modelo computacional para la predicción del emparejamiento funcional de cadenas pesadas y ligeras de anticuerpos, basado en una red neuronal siamesa asimétrica entrenada con la función de pérdida InfoNCE y representaciones fisicoquímicas utilizando los factores Atchley.

La comparación entre tipos de representación numérica de secuencias de aminoácidos demostró que los factores Atchley ofrecen el mejor balance entre capacidad descriptiva y dimensionalidad para usarse en métodos computacionales. Su capacidad de capturar propiedades fisicoquímicas relevantes de cada uno de los aminoácidos en un espacio compacto contribuyó a la eficiencia computacional del modelo y a su desempeño en las métricas de clasificación. Este resultado es consistente con la literatura y refuerza la importancia de elegir cuidadosamente la representación de entrada antes de diseñar el modelo de aprendizaje profundo.

Una de las principales contribuciones de este trabajo fue la incorporación de cadenas pesadas de anticuerpos de ratón BALB/c como negativos verdaderos en el entrenamiento del modelo. A diferencia de los enfoques tradicionales que utilizan negativos aleatorios, este criterio está fundamentado en la incompatibilidad biológica documentada entre las cadenas de especies distintas, derivada de divergencias germinales, diferencias estructurales y la ausencia de coevolución. Los resultados demostraron que esta decisión tuvo un impacto directo y medible en la calidad del espacio métrico aprendido: la separación entre pares compatibles e incompatibles mejoró sustancialmente en comparación con el modelo entrenado sin negativos verdaderos, lo que se reflejó en mejores métricas de clasificación y en una distribución de distancias coseno con menor solapamiento entre clases.

La arquitectura siamesa asimétrica, con *encoders* independientes para cadenas pesadas y ligeras y un módulo de cross-*attention* bidireccional, demostró ser adecuada para capturar las diferencias estructurales entre ambos tipos de cadenas y las interacciones entre sus regiones. El análisis del espacio de *embeddings* mostró que el modelo organiza las representaciones de forma coherente con propiedades inmunológicas conocidas, lo que sugiere que aprendió principios estructurales del emparejamiento de las cadenas pesadas y ligeras más allá de memorizar los pares del conjunto de entrenamiento. El análisis del mecanismo de atención cruzada reveló patrones de interacción entre regiones biológicamente

coherentes con el conocimiento existente sobre el emparejamiento de las cadenas, añadiendo interpretabilidad a las predicciones del modelo.

En conjunto, los resultados de esta investigación demuestran que es posible construir un modelo computacional con capacidad discriminativa real para el problema del emparejamiento de las cadenas de anticuerpos, que la calidad biológica de los negativos de entrenamiento es un factor determinante en el desempeño del modelo, y que las representaciones fisicoquímicas compactas como los factores Atchley son una alternativa viable y efectiva frente a representaciones de mayor dimensionalidad.

7.2. Trabajos futuros

Los resultados mostrados en este trabajo de investigación abren líneas de investigación que pueden fortalecer el modelo propuesto.

Una línea de mejora es el conjunto de datos de entrenamiento, explorando fuentes adicionales de negativos biológicamente fundamentados y ampliando el número de pares positivos con anticuerpos de distintos contextos inmunológicos. Un conjunto de datos con mayor diversidad y representatividad del repertorio inmune humano ayudaría a mejorar la capacidad de generalización del modelo a casos no conocidos durante el entrenamiento.

Una segunda línea de investigación sería la exploración de representaciones que pueden aportar más información de las secuencias de aminoácidos. Por ejemplo, integrar información posicional, contextual o estructural en la representación de entrada podría permitirle al modelo capturar patrones de compatibilidad que las representaciones fisicoquímicas promediadas no logran modelar por completo.

Una tercera línea es la extensión del modelo a otros contextos inmunológicos, incluyendo repertorios de enfermedades infecciosas, condiciones autoinmunes o poblaciones con características genéticas distintas. Esta extensión permitiría evaluar la generalidad de los principios aprendidos y ampliar el alcance de las aplicaciones del modelo.

La validación experimental de las predicciones es otra línea fundamental. Llevar los emparejamientos propuestos por el modelo a validación in vitro permitiría verificar su coherencia biológica real y generar retroalimentación que guíe mejoras iterativas del modelo.

Referencias

Asgari, E., & Mofrad, M. R. K. (2015). Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLOS ONE*, *10*(11), e0141287. https://doi.org/10.1371/journal.pone.0141287

Atchley, W. R., Zhao, J., Fernandes, A. D., & Drüke, T. (2005). Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences*, *102*(18), 6395–6400. https://doi.org/10.1073/pnas.0408677102

Barton, J., Galson, J. D., & Leem, J. (2024). Enhancing Antibody Language Models with Structural Information. *bioRxiv*. https://doi.org/10.1101/2023.12.12.569610

Burbach, S. M., & Briney, B. (2024). Improving antibody language models with native pairing. *Patterns*, *5*(4), 100967. https://doi.org/10.1016/j.patter.2024.100967

Calis, J. J. A., & Rosenberg, B. R. (2014). Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends in Immunology*, *35*(12), 581–590. https://doi.org/10.1016/j.it.2014.09.004

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215. https://doi.org/10.1016/j.neucom.2019.10.118

Costa, V. G., & Pedreira, C. E. (2023). Recent advances in decision trees: an updated survey. *Artificial Intelligence Review*, *56*(5), 4765–4800. https://doi.org/10.1007/s10462-022-10275-5

Dey, D., Haque, Md. S., Islam, Md. M., Aishi, U. I., Shammy, S. S., Mayen, Md. S. A., Noor, S. T. A., & Uddin, Md. J. (2025). The proper application of logistic regression model in complex survey data: a systematic review. *BMC Medical Research Methodology*, *25*(1), 15. https://doi.org/10.1186/s12874-024-02454-5

Diaz-Papkovich, A., Anderson-Trocmé, L., & Gravel, S. (2021). A review of UMAP in population genetics. *Journal of Human Genetics*, *66*(1), 85–91. https://doi.org/10.1038/s10038-020-00851-4

Dogan, A., & Birant, D. (2022). K-centroid link: a novel hierarchical clustering linkage method. *Applied Intelligence*, *52*(5), 5537–5560. https://doi.org/10.1007/s10489-021-02624-8

ElAbd, H., Bromberg, Y., Hoarfrost, A., Lenz, T., Franke, A., & Wendorff, M. (2020). Amino acid encoding for deep learning applications. *BMC Bioinformatics*, *21*(1), 235. https://doi.org/10.1186/s12859-020-03546-x

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., & Rost, B. (2022). ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, *110*, 104743. https://doi.org/10.1016/j.engappai.2022.104743

Fierro, A. N., Nakano, M., Yanai, K., & Pérez, H. M. (2019). Siamese and triplet convolutional neural networks for the retrieval of images with similar content. *Informacion Tecnologica*, *30*(6), 243–254. https://doi.org/10.4067/S0718-07642019000600243

Greiff, V., Yaari, G., & Cowell, L. G. (2020). Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Current Opinion in Systems Biology*, *24*, 109–119. https://doi.org/10.1016/j.coisb.2020.10.010

Guo, D., Dunn-Walters, D. K., Fraternali, F., & Ng, J. C. F. (2026). ImmunoMatch learns and predicts cognate pairing of heavy and light immunoglobulin chains. *Nature Methods*, *23*(1), 106–117. https://doi.org/10.1038/s41592-025-02913-x

Hu, Y., Zhang, X., Slavin, V., Belsti, Y., Tiruneh, S. A., Callander, E., & Enticott, J. (2025). Beyond Comparing Machine Learning and Logistic Regression in Clinical Prediction Modelling: Shifting from Model Debate to Data Quality. *Journal of Medical Internet Research*, *27*, e77721. https://doi.org/10.2196/77721

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Jain, H., Harit, G., & Sharma, A. (2021). Action Quality Assessment Using Siamese Network-Based Deep Metric Learning. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(6), 2260–2273. https://doi.org/10.1109/TCSVT.2020.3017727

Janeway, C. A. (2000). How the immune system works to protect the host from infection: A personal view. En *National Academy of Sciences*. https://www.pnas.org

Jayaram, N., Bhowmick, P., & Martin, A. C. R. (2012). Germline VH/VL pairing in antibodies. *Protein Engineering, Design and Selection*, *25*(10), 523–529. https://doi.org/10.1093/protein/gzs043

Jing, X., Dong, Q., Hong, D., & Lu, R. (2020). Amino Acid Encoding Methods for Protein Sequences: A Comprehensive Review and Assessment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *17*(6), 1918–1931. https://doi.org/10.1109/TCBB.2019.2911677

Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing* (3a ed.). Pearson.

Kaya, M., & Bilge, H. Ş. (2019). Deep Metric Learning: A Survey. *Symmetry*, *11*(9), 1066. https://doi.org/10.3390/sym11091066

Kazemi, M. (2024). Support vector machine in ultrahigh-dimensional feature space. *Journal of*

Statistical Computation and Simulation, *94*(3), 517–535. https://doi.org/10.1080/00949655.2023.2263128

Kenlay, H., Dreyer, F. A., Kovaltsuk, A., Miketa, D., Pires, D., & Deane, C. M. (2024). Large scale paired antibody language models. *PLOS Computational Biology*, *20*(12), e1012646. https://doi.org/10.1371/journal.pcbi.1012646

Khan, A. M., Singh, H., Ranganathan, S., Gojobori, T., & Gao, X. (2024). Editorial: 21st International Conference on Bioinformatics (InCoB 2022)—accelerating innovation to meet biological challenges: the role of bioinformatics. *Frontiers in Genetics*, *15*, 1365223. https://doi.org/10.3389/fgene.2024.1365223

Kindt, T. J., Goldsby, R. A., & Osborne, B. A. (2007). *INMUNOLOGÍA de Kuby* (6a ed.). McGraw Hill.

Koch, G., Zemel, R., & Salakhutdinov, R. (2015). Siamese Neural Networks for One-Shot Image Recognition. *ICML Deep Learning Workshop*, *2*. https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf

Koch, G., Zemel, R., Salakhutdinov, R., & others. (2015). Siamese neural networks for one-shot image recognition. *ICML deep learning workshop*, *2*(1), 1–30.

Leem, J., Mitchell, L. S., Farmery, J. H. R., Barton, J., & Galson, J. D. (2022). Deciphering the language of antibodies using self-supervised learning. *Patterns*, *3*(7), 100513. https://doi.org/10.1016/j.patter.2022.100513

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123–1130. https://doi.org/10.1126/science.ade2574

Manso, T., Folch, G., Giudicelli, V., Jabado-Michaloud, J., Kushwaha, A., Nguefack Ngoune, V., Georga, M., Papadaki, A., Debbagh, C., Pégorier, P., Bertignac, M., Hadi-Saljoqi, S., Chentli, I., Cherouali, K., Aouinti, S., El Hamwi, A., Albani, A., Elazami Elhassani, M., Viart, B., … Kossida, S. (2022). IMGT® databases, related tools and web resources through three main axes of research and development. *Nucleic Acids Research*, *50*(D1), D1262–D1272. https://doi.org/10.1093/nar/gkab1136

Meng, L., Bai, B., Zhang, W., Liu, L., & Zhang, C. (2023). Research on a Decision Tree Classification Algorithm Based on Granular Matrices. *Electronics*, *12*(21), 4470. https://doi.org/10.3390/electronics12214470

Mienye, I. D., & Jere, N. (2024). A Survey of Decision Trees: Concepts, Algorithms, and Applications. *IEEE Access*, *12*, 86716–86727. https://doi.org/10.1109/ACCESS.2024.3416838

Naidu, G., Zuva, T., & Sibanda, E. M. (2023). *A Review of Evaluation Metrics in Machine Learning Algorithms* (pp. 15–25). https://doi.org/10.1007/978-3-031-35314-7_2

Nogales, A., Guadalupe, D., & García-Tejedor, Á. J. (2025). Self-organizing maps to evaluate optimal strategies for balancing binary class distributions: a methodological approach. *Journal of Big Data*, *12*(1), 141. https://doi.org/10.1186/s40537-025-01188-5

Olsen, T. H., Moal, I. H., & Deane, C. M. (2022). AbLang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, *2*(1), vbac046. https://doi.org/10.1093/bioadv/vbac046

Ran, X., Xi, Y., Lu, Y., Wang, X., & Lu, Z. (2023). Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review*, *56*(8), 8219–8264. https://doi.org/10.1007/s10462-022-10366-3

Randriamihamison, N., Vialaneix, N., & Neuvial, P. (2021). Applicability and Interpretability of Ward's Hierarchical Agglomerative Clustering With or Without Contiguity Constraints. *Journal of Classification*, *38*(2), 363–389. https://doi.org/10.1007/s00357-020-09377-y

Raybould, M. I. J., Marks, C., Krawczyk, K., Taddese, B., Nowak, J., Lewis, A. P., Bujotzek, A., Shi, J., & Deane, C. M. (2019). Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(10), 4025–4030. https://doi.org/10.1073/pnas.1810576116

Rezvani, S., Pourpanah, F., Lim, C. P., & Wu, Q. M. J. (2024). Methods for class-imbalanced learning with support vector machines: a review and an empirical evaluation. *Soft Computing*, *28*(20), 11873–11894. https://doi.org/10.1007/s00500-024-09931-5

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, *118*(15), e2016239118. https://doi.org/10.1073/pnas.2016239118

Robinson, J., Chuang, C.-Y., Sra, S., & Jegelka, S. (2021). *Contrastive Learning with Hard Negative Samples*.

Ruffolo, J. A., Sulam, J., & Gray, J. J. (2022). Antibody structure prediction using interpretable deep learning. *Patterns*, *3*(2), 100406. https://doi.org/10.1016/j.patter.2021.100406

Schroeder, H. W., & Cavacini, L. (2010). Structure and function of immunoglobulins. *Journal of Allergy and Clinical Immunology*, *125*(2), S41–S52. https://doi.org/10.1016/j.jaci.2009.09.046

Selva Birunda, S., & Kanniga Devi, R. (2021). *Innovative Data Communication Technologies and Application* (J. S. Raj, A. M. Iliyasu, R. Bestak, & Z. A. Baig, Eds.; Vol. 59). Springer Singapore. https://doi.org/10.1007/978-981-15-9651-3

Shuai, R. W., Ruffolo, J. A., & Gray, J. J. (2023). IgLM: Infilling language modeling for antibody sequence design. *Cell Systems*, *14*(11), 979–989.e4. https://doi.org/10.1016/j.cels.2023.10.001

Suresh, N., Thomas, B., & Joseph, J. (2024). Bibliometric Analysis and Visualization of

Scientific Literature on Heart Disease Classification Using a Logistic Regression Model. *Cureus*. https://doi.org/10.7759/cureus.63337

Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature*, *302*(5909), 575–581. https://doi.org/10.1038/302575a0

Tripathi, K. (2024). The novel hierarchical clustering approach using self-organizing map with optimum dimension selection. *Health Care Science*, *3*(2), 88–100. https://doi.org/10.1002/hcs2.90

van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*. https://doi.org/10.48550/arXiv.1807.03748

Wang, X., Han, X., Huang, W., Dong, D., & Scott, M. R. (2020). *Multi-Similarity Loss with General Pair Weighting for Deep Metric Learning*. http://arxiv.org/abs/1904.06627

Zhang, X., Wei, L., Ye, X., Zhang, K., Teng, S., Li, Z., Jin, J., Kim, M. J., Sakurai, T., Cui, L., Manavalan, B., & Wei, L. (2023). SiameseCPP: a sequence-based Siamese network to predict cell-penetrating peptides by contrastive learning. *Briefings in Bioinformatics*, *24*(1). https://doi.org/10.1093/bib/bbac545

Zubair, Md., Iqbal, MD. A., Shil, A., Chowdhury, M. J. M., Moni, M. A., & Sarker, I. H. (2024). An Improved K-means Clustering Algorithm Towards an Efficient Data-Driven Modeling. *Annals of Data Science*, *11*(5), 1525–1544. https://doi.org/10.1007/s40745-022-00428-2

Anexo A

En este anexo se presentan los productos académicos generados a partir del trabajo de investigación desarrollado durante el doctorado. Estos productos reflejan las contribuciones científicas derivadas de la investigación sobre el emparejamiento funcional de cadenas pesadas y ligeras de anticuerpos, y han sido sometidos a procesos de revisión por pares en revistas científicas indexadas.

Artículo publicado: Erazo-Valadez, M., Hernández, Y., Godoy-Lozano, E. E., Ortiz-Hernández, J., Téllez-Sosa, J., Flores-Sedano, J. J., & Cuevas-Chavez, A. (2026). *Preprocessing of amino acid chains of antibody structure for machine learning analysis*. *International Journal of Combinatorial Optimization Problems and Informatics*, 17(1), 196–214. https://doi.org/10.61467/2007.1558.2026.v17i1.1030

Artículo aceptado: Erazo-Valadez, M., Hernández, Y., Godoy-Lozano, E. E., Ortiz-Hernández, J., & Téllez-Sosa, J. (en prensa). *Functional pairing prediction of antibody heavy and light chains using Atchley-based representations and an asymmetric Siamese network. Programming and Computer Software*. Springer. ISSN 0361-7688. https://link.springer.com/journal/11086