ESTUDIO COMPARATIVO DE TÉCNICAS DE REDUCCIÓN DE DIMENSIÓN APLICADAS A LA PREDICCIÓN DE MAPAS DE CONTACTO DE PROTEÍNAS

DIEGO FERNANDO PEÑA UNIGARRO

Tesis de Maestría en Computación

Director: PhD. Néstor Díaz Mariño
Codirector: MsC. Ember Martínez Flor

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Sistemas
Grupo de Inteligencia Computacional - GICO
Áreas de Investigación: Bioinformática
Popayán, marzo de 2023

DIEGO FERNANDO PEÑA UNIGARRO

ESTUDIO COMPARATIVO DE TÉCNICAS DE REDUCCIÓN DE DIMENSIÓN APLICADAS A LA PREDICCIÓN DE MAPAS DE CONTACTO DE PROTEÍNAS

Tesis presentada a la Facultad de Ingeniería
Electrónica y Telecomunicaciones de la
Universidad del Cauca para la obtención del
Título de

Magíster en:
Computación

Director: PhD. Néstor Díaz Mariño
Codirector: MsC. Ember Martínez Flor

Popayán
2023

DEDICATORIA

A mi madre Sandra Unigarro, a mi Padre Edgar Peña, a mi hermano Andrés Peña, a mi novia Angela Pantoja, a mi primo Carlos Rodríguez y especialmente a mi abuelo Jesús Unigarro, quienes fueron una fuente de apoyo inagotable en los momentos más difíciles y me dieron la fuerza para seguir adelante.

Agradecimientos

Al Dr. Néstor Díaz, al Dr. Víctor Bucheli Guerrero y a MsC. Ember Martínez por su apoyo durante mi proyecto de investigación. Especialmente quiero agradecer al Dr. Néstor Díaz por todas las oportunidades y conocimientos brindados, siendo siempre su liderazgo y experiencia el eje fundamental de esta investigación y el soporte durante todo el programa de maestría. Fueron sus enseñanzas lo que me permitieron aprender y crecer como investigador, superando gracias a su asesoramiento todos los desafíos con confianza y determinación.

También quisiera agradecer a los docentes de la maestría en computación de la Universidad del Cauca por su papel formativo y académico en estos últimos años. Especialmente al grupo de investigación en inteligencia computacional GICO por su gestión y apoyo durante la publicación del artículo. Agradecer igualmente a la fundación CEIBA y el programa Bécate Nariño por darme la oportunidad de acceder a los estudios de maestría con su apoyo y financiación.

Resumen

La definición de la estructura nativa de la proteína a partir de su cadena de aminoácidos es una de los problemas más importantes y desafiantes de la bioinformática, debido a la gran variedad de aplicaciones y nuevos descubrimientos que su solución podría traer. Dada la complejidad de este problema se han intentado buscar soluciones desde diferentes enfoques siendo los computacionales uno de los más utilizados. Los avances actuales de los modelos de predicción de la estructura de las proteínas se han dado gracias al esfuerzo de resolver en primera instancia problemas intermedios de predicción, que aporten a la obtención de estructuras terciarias más precisas. La predicción de mapas de contacto como etapa de predicción intermedia utiliza la definición de contactos entre residuos para generar una representación bidimensional de la estructura de la proteína, que puede ser utilizada para definir el plegamiento tridimensional de la proteína. Si bien han existido avances significativos en los modelos de predicción de mapas de contacto los resultados obtenidos aún no han sido los suficientemente sólidos para ser utilizados en aplicaciones reales. En este proyecto de investigación se realiza un estudio comparativo de técnicas de reducción de dimensión, en donde se observa que el mapeo de las características de entrada a un espacio más compacto (menor número de características) mejora de manera estadísticamente significativa la detección de contactos reales de la proteína en la mayoría de rangos considerados. Esto teniendo en cuenta que la integración de dichos métodos se da con la implementación de un predictor de mapas de contacto que hace uso de una arquitectura presente en el estado del arte actual (basado en aprendizaje profundo).

Abstract

The definition of protein native structure from its amino acid chain information is one of the most challenging and important problems in bioinformatics, due to the great variety of applications and new discoveries that the solution of this issue would bring. It is because the complexity of this problem that different approaches have been tested included the computational ones, which are being widely used due to high storage and processing capabilities of current computer machines. One of the reasons new computational models advances on protein structure prediction are caused because they are focused on solve intermediate prediction problems that contribute to obtain more accurate tertiary structures. Contact map prediction as an intermediate prediction step uses the definition of contacts between residues to generate a two-dimensional representation of the protein structure that can be used for three-dimensional protein folding definition. Although there have been significant advances in contact map prediction models, the results obtained have not been solid enough yet to be used in real applications. In this research project a comparative study of dimensionality reduction techniques is performed, where it has been observed that the mapping of input features to a more compact space (a smaller number of features) generates statistically significantly improves in the detection of real protein contacts in most of the considered ranges. This taking into account that the integration of these methods is given with the implementation of a contact map predictor that makes use of an architecture present in the current state of the art (based on deep learning).

Tabla de Contenido

Índice de Tablas

Índice de Figuras

Índice de Anexos

Capítulo 1

1. Introducción

1.1. Planteamiento del problema

Las proteínas como bloques elementales para todos los organismos vivos, son capaces de ejecutar diversas funciones biológicas fundamentales como por ejemplo la creación de hormonas, la catálisis de reacciones bioquímicas, el transporte de nutrientes y la transmisión de señales biológicas [1], [2]. La gran variedad de estas biomoléculas y sus funciones está directamente relacionada con las posibles variaciones (secuencias) que se pueden presentar entre los 20 aminoácidos estándar y sobre todo al plegamiento que adoptan en el espacio (estructura 3D) [3]. Para el estudio de las propiedades funcionales y estructurales de las proteínas se han dividido los siguientes niveles en su organización: estructura primaria, secundaria, terciaria y cuaternaria [4].

La predicción de la estructura terciaria de las proteínas (polipéptido) a partir de su cadena de aminoácidos (estructura primaria) ha sido considerada como la problemática más desafiante en la bioinformática, puesto que tener plenamente identificada la conformación tridimensional de la proteína brinda información importante acerca de su función [1]. Si bien es cierto que existen métodos experimentales que permiten determinar las estructuras terciarias de las proteínas como, la cristalografía de rayos X [5], y la resonancia magnética nuclear de proteínas (RMN) [6], estas técnicas no pueden ser aplicada de manera generalizada a cualquier tipo de proteína, sin contar que suponen una inversión considerable de recursos [7]. Esta situación ha representado un gran problema debido a la brecha que se ha generado entre el número de secuencias de proteínas descubiertas y el número de estructuras 3D identificadas, dificultando en parte el avance de nuevas investigaciones

en áreas de estudio relacionadas con la biología, la bioquímica, la biofísica y medicina entre otras [1], [3]. Por este motivo, en situaciones donde se cuenta con información experimental limitada e insuficiente para crear conjeturas y entendimientos sólidos es necesario recurrir a otros enfoques como los modelos computacionales que pueden suplir la carencia de datos experimentales, brindando nuevas perspectivas que pueden ser corroboradas más fácilmente en el laboratorio [7].

Un aspecto ampliamente estudiado en el campo de la predicción de estructuras de proteínas se basa en el análisis de los contactos existentes entre los residuos de un polipéptido (secuencia de aminoácidos de una proteína), esto mayoritariamente debido a que dicha información forma una representación bidimensional de la estructura terciaria de la proteína, permitiendo así la posterior definición de la estructura nativa [8]. Al igual que la predicción de estructura terciaria, la definición de contactos entre residuos puede ser deducida de la cadena de aminoácidos, sin embargo, la predicción de mapas de contacto representa un problema intermedio de predicción más sencillo, el cual al ser utilizado como información de entrada posibilita la generación de estructuras nativas más precisas [4].

Si bien existen diversos trabajos en el estado del arte que abordan la predicción de mapas de contacto [9], la mayoría ellos solo se han enfocado en la selección y variación del modelo de predicción para obtener mejores resultados, dejando de lado otras etapas importantes como el procesamiento de los datos de entrada. Los métodos de reducción de dimensión (RD) como etapa de preprocesamiento pueden ayudar a mejorar tareas de predicción, clasificación y visualización debido a su capacidad de aprovechar el espacio de características original y mapearlo a un espacio de baja dimensión con características más compactas que contenga la información más importante [10].

En el caso específico de la predicción de mapas de contacto la integración de RD podría ser de gran relevancia puesto que podría no solo mejorar el desempeño de los resultados eliminando datos ruidosos, sino también potencialmente podría verse un ahorro de recursos como memoria de almacenamiento o tiempo de ejecución. Si bien, existen algunos trabajos relacionados que han utilizado el análisis de componentes principales (PCA) como etapa de extracción de características [11], [12], no fue posible

encontrar estudios comparativos que identifiquen los métodos de RD que puedan potencialmente ser integrados a modelos recientes de predicción de mapas de contacto.

El estudio comparativo como enfoque dentro del espectro de los métodos de investigación científica permite analizar, examinar y contrastar las reacciones de sistemas o modelos en condiciones similares, a perturbaciones en sus diferentes parámetros [13], [14]. Si bien el aprendizaje automático es una disciplina donde los aportes teóricos juegan un papel fundamental, los descubrimientos y avances empíricos también han demostrado ser de gran utilidad para la obtención de mejores modelos [15]. Para el caso específico de esta investigación se implementó un pipeline de predicción que permitió responder la siguiente pregunta de investigación:

> ¿Cuál es la representación óptima de los datos que las técnicas de RD pueden alcanzar cuando se integran a un modelo de predicción de mapas de contacto?

El pipeline de predicción de mapas de contacto integra varias etapas: En primer lugar, se tiene una etapa de procesamiento de características en donde se integran las técnicas de RD seleccionadas (revisión sistemática), para encontrar un espacio reducido de las características secuenciales de entrada. En segundo lugar, se implementó un modelo de predicción de mapas de contacto con una arquitectura basada en redes neuronales residuales (ResNet) [16], donde se establece una variación en la representación de uno de los conjuntos de características de entrada (características secuenciales) haciendo uso de métodos de RD basados en extracción de características. Finalmente se tiene una etapa de evaluación que se implementa siguiendo los lineamientos y métricas definidas en el CASP[1] para la predicción de mapas de contacto, de manera que se pueda contrastar experimentalmente el impacto de integrar dichas técnicas RD en el desempeño del modelo.

Con la ejecución de cada una de las etapas de construcción del modelo se obtuvieron varios resultados importantes. En primer lugar, vale la pena mencionar que el esquema de revisión sistemática realizado mediante el uso de *web scripting*, permitió

[1] https://predictioncenter.org/

automatizar en importante medida la búsqueda de trabajos relevantes, obteniendo así un espectro más amplio de trabajos relacionados (integración de RD con características secuenciales similares a las utilizadas). Una vez analizados los diferentes trabajos relevantes se pudo observar la importancia actual de los métodos de RD como etapa de preprocesamiento de datos, seleccionando aquellos comúnmente utilizados en escenarios donde se tiene un volumen de datos mayor al millón de puntos.

Una respuesta parcial a la pregunta de investigación se dio en el segundo resultado relevante en donde se puede observar un rango específico de dimensión en donde se pueden potencialmente obtener los mejores resultados. Lo interesante de dicho rango es que es similar en cada una de las técnicas de RD seleccionadas.

Finalmente, el último resultado que se obtuvo tiene que ver con las mejoras estadísticamente significativas conseguidas en algunas métricas de evaluación de predicción de mapas de contacto, cuando se utilizaron algunas técnicas de RD para mapear las características secuenciales de las proteínas a un espacio de baja dimensión.

1.2. Aportes del proyecto

Desde el punto de vista de investigación, las contribuciones de este proyecto se pueden ver reflejadas en la generación de nuevo conocimiento, debido principalmente a la novedad de esta propuesta de investigación, dada la carencia de trabajos académicos relacionados encontrados en las diferentes bases de datos electrónicas consultadas (*Scopus, ScienceDirect, Springerlink, IEEE Xplore*). El estudio experimental del desempeño de un modelo de predicción de mapas de contacto, cuando se cambia la representación de las características de entrada a un espacio de baja dimensión, obedece también a las necesidades actuales de procesar grandes volúmenes de información (Big Data), especialmente considerando la naturaleza de las características (alta dimensión) y el volumen de datos que pueden generar la representación de miles de cadenas de proteínas necesaria para entrenar un modelo basado en aprendizaje profundo.

Desde el punto de vista de innovación y desarrollo hay tres aportes principales: El primero relacionado con la revisión sistemática que recopila y analiza la información de múltiples artículos de la literatura, para brindar una referencia en cuanto a los métodos de RD más utilizados teniendo en cuenta criterios como: su aplicación en el en problema de predicción de mapas de contacto o en escenarios en donde se utilicen características secuenciales similares; y en su potencial integración al modelo de predicción utilizado. El segundo aporte está relacionado con el pipeline de predicción, el cual logra adaptar e implementar un modelo de predicción cuya arquitectura forma parte del estado de arte actual, además de integrar la etapa de RD, junto con la evaluación. Lo anterior es de gran relevancia puesto que tanto la selección de métodos de los RD, como su posterior aplicación al modelo se establecen en un escenario real en donde a partir de la caracterización de una cadena de proteína completa se obtiene una representación bidimensional.

Finalmente, es importante mencionar que la implementación del pipeline utilizado en esta investigación puede facilitar futuros procesos de investigación en el área de la predicción de mapas de contacto. Esto debido a que la codificación completa e integrada de cada una de las etapas dentro del pipeline está disponible en un repositorio[2] para su libre descarga. Además, la utilización del API de Keras, así como la codificación en Python 3 representa un aporte significativo con respecto a los trabajos originales los cuales hacen uso de Python 2 y Tensorflow de manera directa.

1.3. Objetivos

A continuación, se presentan los objetivos tal y como fueron aprobados en el documento de anteproyecto por parte del Consejo de Facultad de la Facultad de Ingeniería Electrónica y Telecomunicaciones de la Universidad del Cauca.

[2] https://github.com/diferpun/IntegratedModel

1.3.1. Objetivo general

Comparar experimentalmente el impacto de integrar técnicas de RD a un predictor de mapas de contacto, para identificar el espacio embebido con la menor dimensión que provea potenciales mejoras en el desempeño del modelo.

1.3.2. Objetivos específicos

1. Seleccionar las técnicas de RD que serán utilizadas, a través de una revisión sistemática para establecer criterios de discriminación relacionados con la capacidad de integración de estas técnicas con modelos utilizados en la predicción de mapas de contacto.

2. Adaptar un pipeline de predicción de mapas de contacto en donde se implementen las técnicas de reducción de dimensión seleccionadas y se integren a un modelo de aprendizaje de máquina para la discriminación de parejas de residuos en contacto.

3. Evaluar experimentalmente las técnicas de RD seleccionadas teniendo en cuenta el pipeline planteado y algunas métricas de desempeño existentes.

1.4. Resultados obtenidos

A continuación, se resumen los principales resultados de la presente investigación.

Monografía de la investigación. Corresponde al presente documento y resume el problema, los objetivos planteados para resolver dicho problema, los productos obtenidos (modelos desarrollados) así como la evaluación realizada. Finalmente, presenta las conclusiones junto con unas ideas de trabajo futuro que serían de interés para realizar nuevos aportes en esta área de investigación.

Modelos implementados. Este ítem corresponde al repositorio del código fuente desarrollado para los modelos y herramientas adicionales de procesamiento y evaluación.

Artículo de investigación. Como resultado de la investigación realizada se publicó un artículo titulado “Aplicación de métodos lineales de reducción de dimensión y redes neuronales convolucionales para la predicción de estructura secundaria de proteínas” que ha sido publicado en una revista categoría B del PUBLINDEX de Minciencias con la siguiente citación:

Peña, D. F., Díaz, N., & Bucheli, V. Aplicación de métodos lineales de reducción de dimensión y redes neuronales convolucionales para la predicción de estructura secundaria de proteínas. RISTI - Revista Ibérica de Sistemas y Tecnologías de la Información, E38, 199-213. diciembre 2020, ISSN: 1646-9895.

Este artículo representó un primer resultado en donde se hace una integración de técnicas de reducción de reducción de dimensión en una arquitectura de aprendizaje profundo para la predicción de la estructura secundaria de las proteínas (entrada al modelo de predicción de mapas de contacto), el cual es replicado dentro del pipeline planteado.

A continuación, se describe de manera general el contenido y organización de la presente monografía:

CAPÍTULO 1: Introducción: Hace referencia al presente capítulo que introduce el tema de investigación, presenta la pregunta de investigación que originó el trabajo, los aportes al problema, también los objetivos (general y específicos) definidos para el proyecto, un breve resumen de los resultados obtenidos y finalmente la organización de la monografía.

CAPÍTULO 2: Marco conceptual y estado del arte: Este capítulo presenta el estado del arte de trabajos relacionados con la predicción de mapas de contacto y los métodos de reducción de dimensión.

CAPÍTULO 3: Revisión Sistemática: Este capítulo abarca básicamente el desarrollo del primer objetivo específico que consiste en la elección de las técnicas de reducción de dimensión, teniendo en cuenta las características del conjunto de datos utilizado y

la capacidad de integración al modelo de predicción de mapas de contacto implementado.

CAPÍTULO 4: Modelado: En este capítulo se presenta el diseño general del pipeline planteado (objetivo específico número dos), que abarca de forma general los siguientes aspectos: descripción de los conjuntos de datos, descripción del modelo de predicción y la definición del esquema de evaluación.

CAPÍTULO 5: Resultados: Aquí se presentan los resultados de la evaluación experimental de las técnicas de reducción de dimensión integradas al modelo de predicción de mapas de contacto, teniendo en cuenta las métricas de evaluación previamente definidas.

CAPÍTULO 6: Conclusiones y trabajo futuro: En este capítulo se presentan las conclusiones obtenidas al finalizar el trabajo de grado, así como recomendaciones de trabajo futuro para algunas etapas del pipeline implementado.

CAPÍTULO 7: Bibliografía: Este último capítulo contiene las referencias bibliográficas de los artículos y libros consultados para la realización del proyecto.

Capítulo 2

2. Marco conceptual y estado del arte

2.1. Conceptos biológicos

La formación de cadenas de aminoácidos (residuos) unidos a través de enlaces covalentes se conoce como proteínas o polipéptidos y se consideran como las constituyentes primordiales de todos los organismos vivos [2]. Dichas macromoléculas son consideradas como las más abundantes e importantes, debido a que están inmersas en varios procesos biológicos [3]. De este modo, considerando la importancia de las proteínas en los seres vivos, la investigación de sus propiedades estructurales y funcionales siempre han sido una prioridad por parte de los científicos [4].

Como se indica en la Figura 1, la organización estructural de una proteína ha sido dividida en cuatro niveles fundamentales [4]. La estructura primaria establece la cantidad y los tipos de aminoácidos que componen a la proteína, así como el orden en que estos se disponen a lo largo de la secuencia [4], [17]. Cada polipéptido posee una constitución característica, debido a la variabilidad con que los aminoácidos pueden aparecer dentro de la proteína [17]. El segundo nivel estructural hace referencia a la formación de estructuras regulares repetitivas debidas a la relación espacial que se establece entre los aminoácidos de una proteína [18].

Dentro de la estructura secundaria se puede encontrar principalmente dos patrones bien definidos conocidos como: hélices alfa (α) donde los aminoácidos se enrollan en una configuración helicoidal asemejando la forma de un resorte; y laminas beta (β) donde los aminoácidos se configuran de manera extendida [4], [17]. El tercer nivel estructural expone la relación espacial que guardan entre si las diferentes zonas o

áreas de cada cadena polipeptídica que forman una proteína, describiendo así todos los aspectos de su plegamiento tridimensional (estructura terciaria o nativa) [17]. Por último, el cuarto nivel estructural hace referencia a la conformación en el espacio de varias cadenas polipeptídicas que generalmente aparecen en moléculas complejas [4].

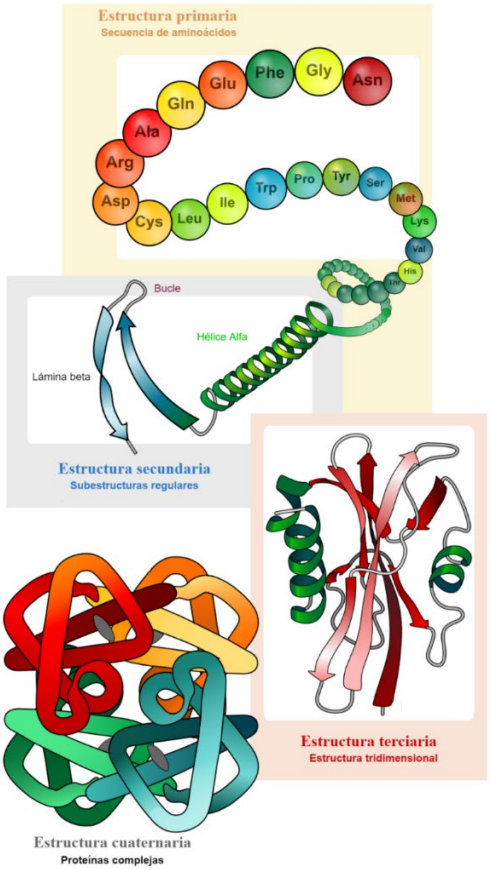


Figura 1. Niveles estructurales de una proteína**.** Fuente: https://byjus.com/chemistry.

Dentro del estudio estructural de las proteínas se ha presentado especial atención y esfuerzo en identificar la estructura terciaria (estructura nativa) de las proteínas, puesto que tener plenamente identificada su conformación tridimensional brinda información importante acerca de su función [1]. Esta información podría ser de gran ayuda para el avance de estudios científicos, así como el entendimiento de procesos biológicos, químicos y patológicos que podrían acelerar la implementación de nuevas aplicaciones y terapias médicas como vacunas [1], [3]. A pesar de la existencia de métodos experimentales que permiten determinar las estructuras terciarias de las proteínas como, por ejemplo, la cristalografía de rayos X [5], y la resonancia magnética nuclear

de proteínas (RMN) [6]. No todos los tipos de proteínas pueden ser caracterizados estructuralmente por estas técnicas, sin contar la considerable inversión de tiempo y recursos que dichas aproximaciones suponen [7].

Las repercusiones de la anterior situación pueden observarse en la Figura 2, donde existe una comparación de la cantidad de estructuras primarias y estructuras terciarias en función del tiempo en años. La gráfica de la izquierda ilustra el crecimiento anual del número nuevas estructuras de proteínas almacenadas en el PDB [19], junto con el método con el cual fue encontrada. La imagen de la derecha contiene el crecimiento anual de las secuencias de proteínas descubiertas y almacenadas en las bases de datos UniprotKB/TrEMBL [20], UniprotKB/SwissProt [20], junto con el crecimiento de las estructuras de proteínas en el PDB. Es en esta imagen en donde se resalta una diferencia significativa en el crecimiento de las estructuras de proteínas definidas y en las secuencias de proteínas descubiertas, siendo estas últimas las que mayor velocidad de adquisición e identificación poseen. Por este motivo surge la necesidad de utilizar métodos alternativos para el establecimiento de estructuras terciarias de proteínas que permitan complementar a los estudios experimentales [7].

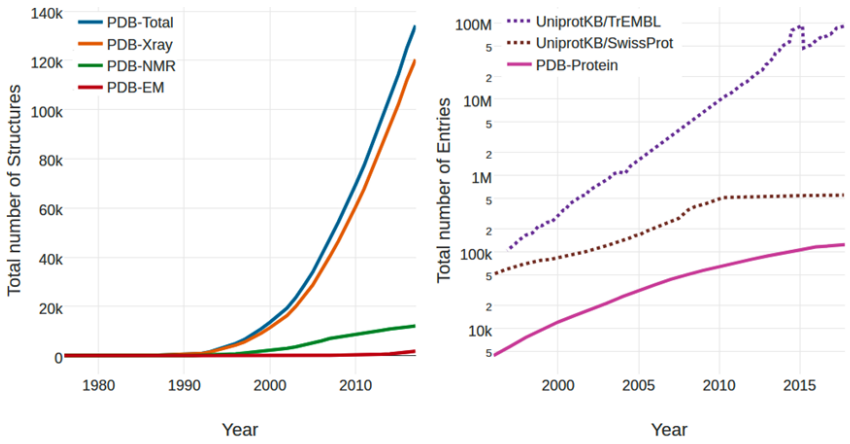


Figura 2. Comparación de la cantidad de estructuras primarias y estructuras terciarias descubiertas en función del tiempo en años. Fuente: [7].

En 1972, Anfinsen y sus colegas recibieron el premio nobel por su trabajo de investigación relacionado con el plegamiento de proteínas, obteniendo como resultado la postulación del dogma de Anfinsen que representa uno de los principios básicos en

la biología molecular [21]. El dogma de Anfinsen establece que la estructura nativa de una proteína está determinada únicamente por su secuencia de aminoácidos, lo cual demostró ser cierto al menos para las proteínas globulares [7]. La definición de este postulado fue de gran importancia puesto que motivó esfuerzos desde diferentes campos del conocimiento, para intentar encontrar una forma efectiva de definir el plegamiento tridimensional de un polipéptido utilizando la información de la cadena de aminoácidos [3].

Dentro de los enfoques para resolver el problema de predicción de estructura nativa de las proteínas se pueden destacar dos aproximaciones fundamentales [17]. La primera conocida como predicción basada en la secuencia, que agrupa el modelado ab initio y de novo, que buscan resolver el problema de predicción de estructura terciaria a partir del análisis de las propiedades físico-químicas de los aminoácidos, o con el estudio de información evolutiva de las proteínas respectivamente [22], [23]. El modelado por comparación representa el segundo enfoque y básicamente intenta encontrar nuevos plegamientos tridimensionales a través del análisis de estructuras de proteínas conocidas (plantillas) [24]. Este último enfoque se puede dividir en el modelado por homología o hilvanado [25]. El modelado por homología utiliza proteínas con estructura definida para encontrar estructuras terciarias desconocidas con la condición de similitud en la secuencia de aminoácidos [24]. El modelado por hilvanado consiste en plegar la proteína con estructura desconocida de todas las maneras empleadas por las proteínas con plegamientos tridimensionales previamente definidos y calcular cuál de ellas posee la conformación de menor energía [25].

Los modelos computacionales representan uno de los enfoques más utilizados para abordar el problema de predicción de proteínas, debido a la gran capacidad de procesamiento y almacenamiento de los sistemas informáticos actuales, así como el continuo crecimiento de la información disponible en bases de datos como el *Protein Data Bank* (PDB) [19], donde cada vez son acumuladas nuevas secuencias y estructuras resueltas de proteínas [7]. No obstante, a pesar del considerable avance de los modelos computacionales para la definición del plegamiento de las proteínas, únicamente los modelos por comparación mediante el uso de plantillas han permitido generar modelos lo suficientemente confiables para ser utilizados en aplicaciones reales [26], [27]. El modelado mediante técnicas ab initio y de novo por su parte, aún

no han sido capaces de generar estructuras de proteínas lo suficientemente aproximadas para ser usadas de manera regular en la práctica [7]. Buscando mejorar la calidad de las estructuras nativas de las proteínas definidas con métodos computacionales, los investigadores han intentado lograr un progreso resolviendo en primera instancia tareas intermedias de predicción, que aporten a la adquisición de estructuras nativas con un mayor grado de precisión [4].

Una tarea ampliamente estudiada en el campo de la predicción de estructuras de proteínas consiste en el análisis de los contactos existentes entre los residuos de un polipéptido, puesto que la identificación de contactos nativos restringe el plegamiento y caracterizan la estructura terciaria [8]. Además, considerando que esta representación puede ser deducida a partir de la estructura primaria, se ha dedicado mucho esfuerzo en resolver el problema de predicción de la estructura nativa a través de este enfoque conocido como predicción de mapas de contacto, el cual puede definirse como un puente entre la estructura primaria y la estructura terciaria [9].

2.2. Predicción de mapas de contacto

Un mapa de contacto denota una representación bidimensional (2D) de la estructura nativa de la proteína, donde se relaciona la interacción en el espacio tridimensional de los residuos i, j, en base a la distancia euclidiana que los separa [9]. Los contactos nativos presentes dentro de una proteína se representan a través de una matriz de tamaño $L \times L$, siendo L la longitud de la proteína o el número de aminoácidos que componen la estructura primaria. Cada uno de los valores $C(i,j)$ que conforman dicha matriz son calculados haciendo uso de la siguiente expresión:

$$C(i,j) = \begin{cases} 1, & si\ \Delta C_\beta < T \\ 0, & si\ \Delta C_\beta \geq T \end{cases} \quad (1),$$

donde ΔC_β denota la distancia euclidiana entre dos residuos medida desde los carbonos beta (C_β). El parámetro T es un umbral de distancia que establece un valor de discriminación para definir si los residuos i, j forman un contacto ($C(i,j) = 1$) o por el contrario describen la inexistencia de este ($C(i,j) = 0$) [7].

Con los valores binarios definidos para cada una de las parejas de aminoácidos presentes en una proteína se obtiene el patrón de la Figura 3, que presenta una fotografía bidimensional de la estructura terciaria de la proteína. Una de las principales ventajas que brinda el análisis de contactos entre los residuos, radica en su capacidad para reconstruir estructuras terciarias, inclusive con la presencia de errores en la definición de los contactos (falsos positivos, falsos negativos). Esto significa que se pueden obtener estructuras nativas de baja resolución haciendo uso de mapas de contactos ruidosos [8], o en casos extremos con un porcentaje de precisión de al menos 22% del total de los contactos (con sensibilidad del 100%) [28].

A pesar del grado de tolerancia que esta tarea de predicción permite, la tasa de discriminación de contactos y no contactos, especialmente los más informativos, es aún muy baja. Teniendo en cuenta los últimos resultados del CASP, el predictor que presentó uno de los mejores desempeños promedio considerando el balance entre precisión y sensibilidad fue tFold [29], con una precisión de 56.089 % y una sensibilidad del 23.676% (https://predictioncenter.org/casp14/rrc_avrg_results.cgi, visitado en junio de 2022). En consecuencia, la predicción de mapas de contacto puede considerarse aún como un problema abierto.

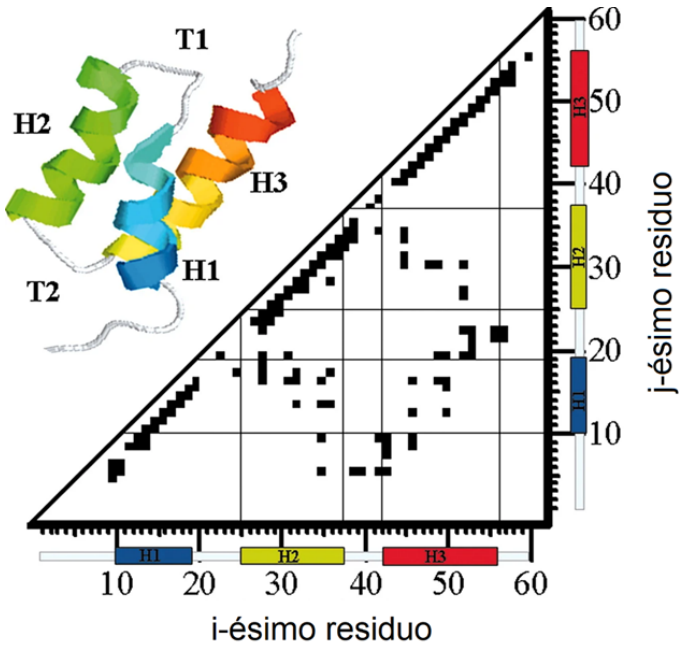


Figura 3. Estructura nativa y mapa de contacto del dominio B de la proteína 1BDD. Fuente: [30].

Como se explicó anteriormente no todos los contactos entre residuos son iguales, razón por la cual han sido divididos en categorías bien definidas basándose en su separación dentro de la cadena (Figura 4) [7], [31]. Los contactos de rango corto representan la primera clase y son aquellas interacciones entre residuos cuya separación dentro de la cadena es mayor a 6 y menor o igual a 12 ($6 < |i - j| \leq 12$) [31]. El segundo agrupamiento conocido como contactos de rango medio reúnen residuos en contacto que cuentan con una separación entre 12 y 24 aminoácidos ($12 < |i - j| \leq 24$) [31]. Finalmente, se dan los contactos de rango largo que tienen separaciones mayores a 24 aminoácidos ($|i - j| > 24$) [31]. Dentro de la anterior agrupación se suele distinguir el subconjunto denominado contactos de rango extralargo que cuentan con una separación superior a 50 aminoácidos ($|i - j| > 50$) [31]. Generalmente los contactos de rango largo y extralargo representan los más difíciles de predecir y los que más información aportan al momento de reconstruir la estructura nativa de una proteína [32].

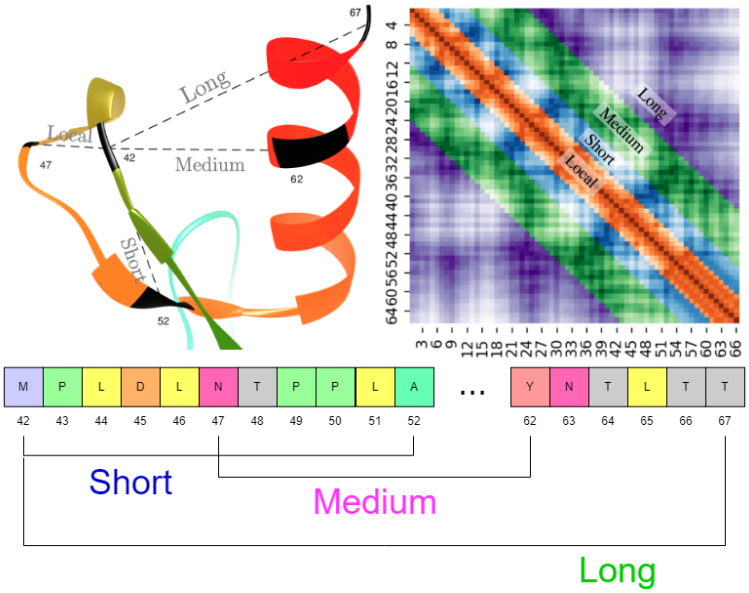


Figura 4. Clasificación de los tipos de contacto según su separación en la secuencia. Fuente: [33].

2.3. Modelos de predicción de mapas de contacto

La predicción de la estructura terciaria de proteínas en escenarios donde el número de plantillas estructurales disponibles es limitado (predicción ab initio), representa uno de

los grandes desafíos para la bioinformática actual [34]. Si bien inicialmente, herramientas como Rosetta [35] y FRAGFOLD [36] demostraron cierto grado de éxito en la predicción ab initio para segmentos específicos de una proteína, en años recientes el enfoque de predicción de estructura terciaria basado en la definición de los contactos entre residuos ha marcado un nuevo rumbo en esta disciplina [9]. Esto principalmente debido a los prometedores resultados y alcances que han sido obtenidos con dicha información [28].

En cuanto a los métodos de predicción de mapas de contacto, con las diferentes clasificaciones mencionadas en [34], se puede obtener un agrupamiento simplificado de cuatro enfoques que están categorizados de acuerdo al tipo de información que utilizan. En una primera categoría se encuentran los modelos que hacen uso únicamente de información derivada de procesos de coevolución (*direct coupling analysis*) [37]. Luego se tiene el grupo de enfoques que se basan en modelos de aprendizaje automático con la extracción de características de la secuencia [9]. Por último, se definen los métodos basados en plantillas estructurales [38], y el grupo de métodos híbridos que hacen uso de varios de los enfoques anteriormente mencionados [32]. No obstante, cabe resaltar que los trabajos más actuales han integrado casi de manera general la información de coevolución junto con los modelos de aprendizaje automático debido a que este tipo de métodos presenta mejores tasas de discriminación entre contactos y no contactos [32], [39].

Los enfoques basados en información evolutiva de las proteínas [37], [40], aprovechan la extracción de información presente en el alineamiento múltiple de secuencias (MSA) para encontrar las posibles parejas de aminoácidos que están en contacto. Estos modelos principalmente se sustentan en la base, que la existencia de un contacto crítico para mantener el plegamiento limitará las propiedades fisicoquímicas de los aminoácidos involucrados [37]. Por lo tanto, si un residuo en un contacto dado muta y potencialmente perturba las propiedades del contacto, entonces es más probable que su pareja mute a un residuo fisicoquímicamente complementario, para asegurar que el pliegue nativo de la proteína permanezca estabilizado [37], [40]. Sin embargo, a pesar de que estos enfoques por sí solos, no han brindado los resultados más sobresalientes, se han convertido en una fuente importante de características que son aprovechadas por los modelos de aprendizaje automático que han empezado a ver

notables mejorías al combinar este tipo de información evolutiva en sus procesos de aprendizaje [34].

Dentro de los métodos de predicción de mapas de contacto basados en algoritmos clásicos de aprendizaje automático se pueden destacar, modelos basados en redes neuronales [28], [41], [42], modelos ocultos de Markov [43], máquinas de soporte vectorial [44], algoritmos genéticos [45], bosques aleatorios [46], entre otros. Estos métodos generalmente hacen uso de vectores de entrenamiento formados por características que pueden incluir: el tipo de estructura secundaria, la accesibilidad al solvente, el tipo de residuo (polaridad y propiedades fisicoquímicas), la longitud de separación de secuencia entre los residuos en consideración [34]. La principal ventaja de emplear este tipo de métodos reside en su capacidad de ser implementados en escenarios donde existen proteínas que no cuentan con muchas secuencias homólogas conocidas [9]. No obstante, a pesar de que en un principio los enfoques anteriormente descritos obtuvieron los resultados más relevantes en la predicción de mapas de contacto, la aparición de arquitecturas de aprendizaje profundo en los últimos años ha generado un nuevo marco de referencia en el estado del arte debido a las mejorías en los índices de desempeño que han alcanzado [47].

El tercer tipo de método de predicción de contactos está basado en plantillas donde se aprovechan la información de estructuras homólogas conocidas, para predecir contactos de secuencias desconocidas [48]. La principal ventaja de este tipo de métodos radica en el buen desempeño que pueden tener si se cuenta con un buen conjunto de estructuras de plantilla para guiar la predicción [38]. No obstante, esta técnica está limitada por la disponibilidad y calidad de las estructuras homólogas conocidas, de modo que no es recomendada cuando no se tiene esta información [34]. Finalmente, la última clasificación comprende modelos híbridos que hacen uso de varias fuentes de información para la predicción de mapas de contacto. Por ejemplo, el trabajo desarrollado en [38], combina información de estructuras homólogas y redes neuronales para la predicción de mapas de contacto. Igualmente, en [49], se hace uso de información fisicoquímica de la proteína junto con información evolutiva para crear una discriminación de contactos y no contactos entre residuos. Los trabajos presentados en [32], [39], emplean características obtenidas a partir de información evolutiva y de la secuencia de aminoácidos para alimentar modelos de predicción. Los

dos últimos trabajos mencionados representan una tendencia actual muy fuerte, provocando que sean catalogados dentro de los métodos basados en aprendizaje de máquina y no como híbridos [34].

Las arquitecturas de aprendizaje profundo conforman el estado del arte más reciente puesto que los recursos computacionales y las fuentes de información disponibles actualmente, han permitido sea posible recolectar los volúmenes de datos necesarios para entrenar este tipo de modelos en tiempos de entrenamiento aceptables [47]. Además, como se explicó anteriormente los datos de entrenamiento para este tipo de arquitecturas han usado casi como regla general una combinación de características extraídas de la secuencia de aminoácidos e información evolutiva.

Si bien en años recientes se han implementado diversos modelos de aprendizaje automático, es relevante destacar aquellas propuestas que han ocupado un lugar importante en las últimas versiones del CASP, en la categoría de predicción de mapas de contacto. MetaPSICOV [32], es un modelo de predicción que hace uso de información evolutiva y redes neuronales feed-forward para discriminar contactos entre los residuos de una proteína. A pesar de que esta arquitectura no emplea aprendizaje profundo es importante resaltarla debido a que es una de las mejores implementaciones del evento CASP 11, y uno de los primeros enfoques en demostrar la potencialidad de combinar información evolutiva con modelos de aprendizaje automático. Por otro lado, el modelo descrito en [39], conocido como RaptorX-Contact y presentado en las versiones 12 y 13 del CASP, está constituido por dos etapas de redes neuronales recurrentes profundas para la predicción de mapas de contacto y significó un gran progreso dados los niveles de precisión alcanzados con la adopción de este tipo de arquitecturas, posicionándolo como una de las mejores propuestas en los dos eventos.

En el CASP 13 del año 2018 se pueden señalar varios modelos, los cuales están basados en redes neuronales profundas. El modelo conocido como TripletRes [50] hace uso de una arquitectura basada en ResNet que es alimentada con información evolutiva mejorada obtenida del MSA. AlphaFold es un modelo propuesto en [51], y básicamente hace uso del análisis coevolutivo, para mapear la covariación de residuos en la secuencia, al contacto físico en la estructura de la proteína. Lo anterior se lleva

a cabo con la aplicación de redes neuronales profundas, las cuales identifican de manera robusta patrones en la información de entrada, que definirán el mapa de contacto predicho. DeepMetaPSICOV propuesto en [52], combina algunas características de los modelos MetaPSICOV [32], y DeepCov [53], para posteriormente entrenar una arquitectura basada en redes convolucionales residuales. DNCON2 [54] es un predictor conformado por dos niveles de redes neuronales convolucionales, que en primera instancia realizan una clasificación preliminar con varios umbrales de distancia para obtener probabilidades de contacto. Estas probabilidades formarán parte de los datos de entrada del segundo nivel de discriminación que permitirá obtener el mapa de contacto final. Por último, se puede resaltar el predictor Deepcontact que está conformado por 9 capas convolucionales que toman como entrada características de la secuencia junto con información coevolutiva para brindar probabilidades de contacto [55]. Finalmente, según [56] en la versión 14 del CASP la evaluación de los resultados mostró que las predicciones de tFold [29], TripletRes [50], y DeepPotential fueron las más precisas en la categoría de predicción de mapas de contacto, aunque se resalta que no existieron diferencias significativas con el CASP13 [56].

Adicionalmente se pueden destacar publicaciones que exploran la adopción de otras arquitecturas de aprendizaje profundo. En [32] y [45], se hace uso de redes neuronales recurrentes tipo Long Short-Term Memory (LSTM) para encontrar las parejas de residuos que presentan una mayor probabilidad de interacción. Por otro lado, el modelo GANcon propuesto en [57], representa un trabajo interesante que utiliza Generative Adversarial Networks (GAN) para la predicción de mapas de contacto. Donde básicamente se emplea una arquitectura de Encoder-Decoder para capturar la información de contacto subyacente en las características de la proteína, y de esta forma generar mapas de contacto sintéticos. Estos finalmente serán la entrada a una red de discriminación que aprende a distinguir entre mapas de contacto reales y artificiales, forzando al generador a producir contactos con un mayor grado de exactitud con respecto a los que efectivamente están en la proteína. Otros modelos adoptan metodologías basadas en sistemas de atención [58] y en el aprovechamiento de capas previamente entrenadas con la utilización de Transfer Learning [59].

2.4. Métodos de reducción de dimensión

Los métodos de reducción de dimensión (RD) tienen por objetivo construir un espacio de características reducido con la menor pérdida de información posible, donde se conserven ciertas propiedades estructurales del espacio original [60]. Estos métodos proporcionan una herramienta para el preprocesamiento de los datos, cuyo uso puede verse reflejado en una reducción del espacio de almacenamiento y en la posible supresión de datos redundantes, irrelevantes o ruidosos [10], [60]. Como se muestra en la Figura 5, los métodos de RD generalmente se clasifican en dos clases: selección de características y extracción de características [60]. El primer grupo se basa en elegir una porción de las características de un conjunto de datos que represente la información más relevante teniendo en cuenta la tarea que se desea abordar [60]. En contraste, la extracción de características encuentra una representación reducida haciendo uso de mapeos o transformaciones del espacio de alta dimensión [10].

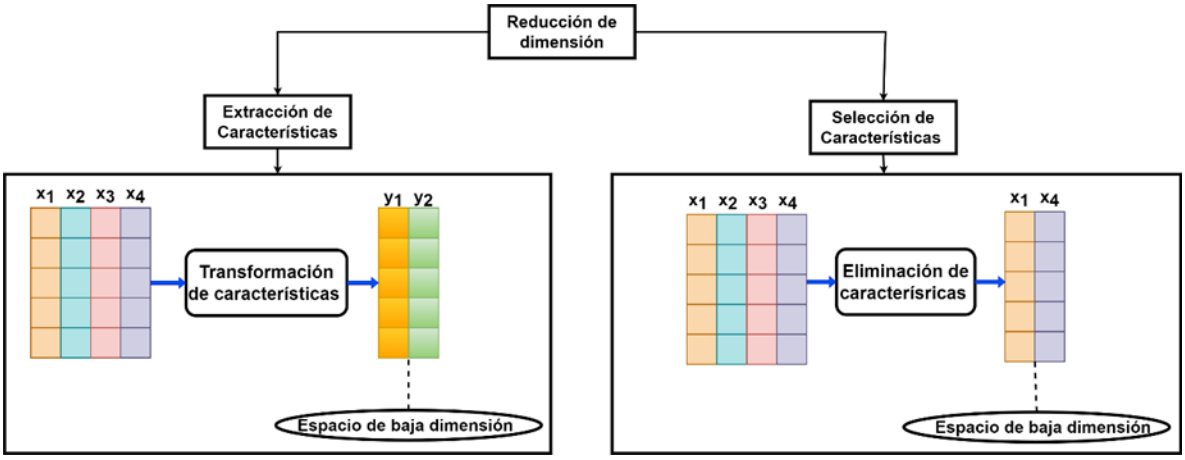


Figura 5. Tipos de métodos de reducción de dimensión.

Según [61] los métodos de RD basados en la selección de características pueden ser divididos en cuatro categorías. La primera categoría está conformada por las técnicas basadas en filtros que hacen uso de información estadística para eliminar atributos irrelevantes o ruidosos que no aportan a la tarea de aprendizaje [62]. Dentro de esta agrupación se pueden destacar las técnicas basadas en criterios de correlación, e información mutua [63]. En la segunda categoría se encuentran las técnicas de

envoltura, las cuales utilizan medidas de desempeño de un algoritmo de aprendizaje como referencia para la selección heurística de un subconjunto de características que denoten mejoras en los resultados [62]. Las técnicas de envoltura se pueden presentar en dos enfoques: los que están basados en algoritmos de selección secuenciales; y en algoritmos heurísticos de búsqueda [63]. Las técnicas de selección incrustada hacen parte de la tercera categoría donde se integra la etapa de selección de características como parte del proceso de entrenamiento del modelo [62]. La última categoría hace referencia a técnicas híbridas de selección de características, las cuales hacen uso simultáneo de varios de los enfoques presentes en las anteriores categorías [64].

En la literatura no existe una taxonomía absoluta de las técnicas de extracción de características puesto que pueden ser divididas según varios criterios como, por ejemplo: la clase de mapeos que realizan (lineal o no lineal) [10], la forma como preservan propiedades intrínsecas de los datos de alta dimensión (local o global), inclusive se realiza una categorización según la naturaleza del problema de optimización que resuelven (convexo o no convexo) [65]. Para esta investigación se tendrá en cuenta el paradigma de clasificación basado en mapeos lineales y no lineales puesto que brindan una visión general de las técnicas de RD disponibles.

Entre los enfoques que hacen uso de transformaciones lineales para obtener el espacio embebido se pueden mencionar los siguientes: *Principal Component Analysis* (PCA) es una técnica de RD no supervisada, la cual se basa en un problema de optimización convexo que busca la preservación de la varianza de las características originales a través de la proyección a un espacio de componentes principales [66]; *Classical multidimensional scaling* (CMDS) es una técnica estrechamente relacionada con PCA que busca encontrar un espacio de baja dimensión informativo, maximizando la dispersión de la proyección del conjunto de datos de entrada [67]; *Random projections* (RP) es una técnica lineal de reducción de dimensión que genera nuevas características a través de la proyección de cada uno de los datos a lo largo de direcciones aleatorias, que mejor preserven la información del conjunto de datos original [68]; *Singular Value Decomposition* (SVD) representa una técnica de RD que en términos de interpretación es similar a PCA, sin embargo, este método no trabaja directamente sobre la matriz de covarianza y ha mostrado buenos resultados al

momento de reducir la dimensión de matrices dispersas [69]; *Independent component analysis* (ICA) reduce el número de características a través de la búsqueda de componentes que se asumen como la mezcla lineal de un conjunto de fuentes independientes, separadas bajo la independencia estadística medida con información mutua [70]; *Linear discriminant analysis* (LDA) es una técnica de RD supervisada que busca proyectar el conjunto de datos a un espacio de baja dimensión, a través de una combinación lineal de las características de entrada que establezca la máxima separación entre las clases y la mínima separación entre elementos de una misma clase [71].

En el grupo de técnicas de extracción de características que obtienen los espacios embebidos a partir de transformaciones no lineales, se destacan: *Kernel Principal Component Analysis* (KPCA), el cual es una extensión de PCA que, en lugar de calcular la matriz de covarianza, calcula los vectores propios de una matriz kernel, lo cual produce un espacio embebido a través de un mapeo no lineal [72]; *Multidimensional scaling* (MDS) es una técnica de RD no supervisada que tiene por objetivo preservar distancias entre parejas de instancias de datos [67]; ISOMAP busca una representación de baja dimensión que preserve la distancia geodésica entre los puntos de datos [73]; *Locally linear embedding* (LLE) utiliza los vecinos más cercanos de cada punto para calcular un conjunto de pesos que serán usados para resolver un problema de optimización que finalmente encontrará un espacio embebido donde cada punto es descrito como una combinación lineal de sus vecinos [74]; *Laplacian Eigenmaps* (LE) es una técnica no supervisada que hace uso de descomposiciones espectrales y representaciones basadas en grafos para la obtención de espacios de baja dimensión [75]. Finalmente, dentro del grupo de técnicas de RD no lineales basadas en redes neuronales se pueden destacar los siguientes enfoques: *Curvilinear component análisis* (CCA) [76], *Curvilinear distance análisis* (CDA) [76], *Self-organizing map* (SOFM) [77], *Isotop* [78], y *Autoencoders* [79].

Al analizar algunos trabajos dentro del estado del arte se encuentra que, en [61] se demuestra de manera experimental que una buena selección de características en los datos dentro del área de la informática puede verse reflejado en una mejoría significativa del desempeño de clasificación en modelos de aprendizaje profundo como: redes neuronales convolucionales o redes neuronales recurrentes. Por otra

parte, en [53] se hace una búsqueda del número mínimo de características necesarias para alcanzar el desempeño de los modelos más representativos en la literatura, demostrando así que algunas características no aportan información valiosa en la predicción de mapas de contacto con redes neuronales profundas. Adicionalmente, existen trabajos que denotan la importancia de la calidad en el MSA, implementando filtros que eliminan secuencias ruidosas, las cuales pueden afectar el desempeño del clasificador [80], [81]. Inclusive uno de los mejores algoritmos presentados en el CASP 13 realiza un aporte en la adquisición de información coevolutiva de mayor calidad y resalta la importancia de encontrar mejores características para obtener modelos de predicción más precisos [50]. Un trabajo importante con los lineamientos planteados en este proyecto de investigación es el trabajo desarrollado en [82], donde se hace uso de características de Fisher junto con autoencoders para encontrar representaciones vectoriales más complejas que mejoren el modelo de discriminación de contactos y no contactos entre residuos. Igualmente, cabe resaltar los trabajos de [83], [84], los cuales trabajan con el conjunto de datos EDBDL'14 con aproximadamente 32 millones de parejas de aminoácidos, donde se resalta que las etapas de selección de características y submuestreo tienen una gran relevancia al momento de trabajar con un gran volumen de datos. Por lo tanto, la tarea de buscar espacios de características compactos que tengan la información más importante y discriminatoria es fundamental para la obtención de modelos de predicción de mayor calidad [61].

2.5. Caracterización de secuencias de aminoácidos

El conjunto de datos procesado y utilizado para las diferentes pruebas realizadas en esta investigación es el mismo empleado por los trabajos [39] y [85], donde las secuencias de aminoácidos son caracterizadas con dos tipos de información como se indica en la Figura 6. La primera fuente de información corresponde al conjunto de características secuenciales que agrupan atributos como: Matriz de puntuación de posición específica (PSSM), Matriz de puntuación de frecuencia específica (PSFM), tres tipos de estructura secundaria (SS3) y Accesibilidad al solvente (ACC). La segunda fuente de información, integra características de coevolución, información mutua (mutual information) y potencial de contacto (pairwise contact potential).

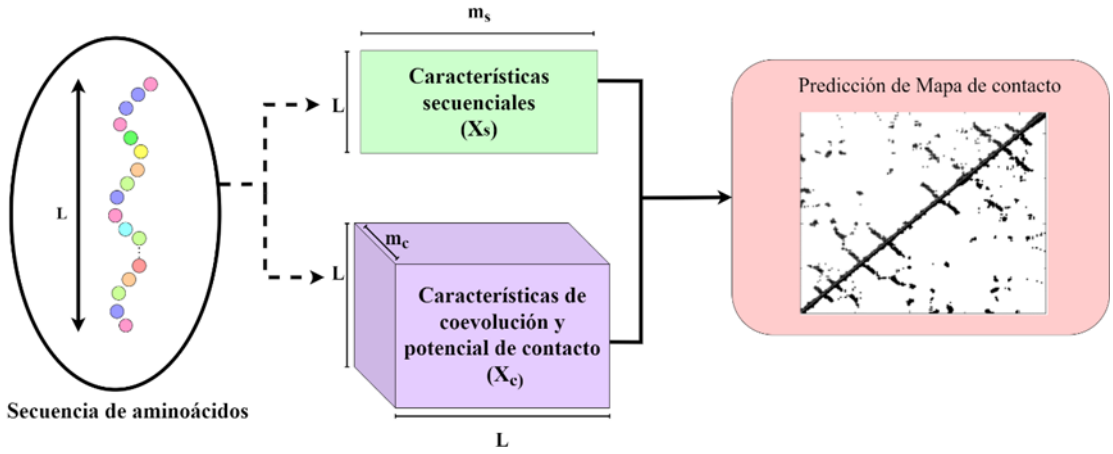


Figura 6. Conjunto de características utilizado para la obtención de un mapa de contacto predicho.

2.5.1. Características secuenciales

Este tipo de información establece atributos para cada uno de los residuos dentro de una secuencia de proteína, lo cual producirá finalmente una matriz de características X_s para cada una de las cadenas de proteínas objetivo que conforman el conjunto de datos. Dicha matriz será de tamaño $L \times D$, donde L representa la longitud de la secuencia y D el número de atributos que en este caso vendrá dado por las características secuenciales que se mencionan a continuación.

- **Matriz de puntuación de frecuencia especifica (PSFM).**

La PSFM representa la probabilidad de encontrar uno de los 20 aminoácidos estándar en una posición específica dentro de la estructura primaria de la proteína [86]. Esta probabilidad que se indica en la Tabla 2 es obtenida básicamente con el conteo normalizado de coincidencias dentro de todas las secuencias homólogas de un determinado aminoácido estándar, en una posición dentro de la cadena. Por ejemplo, si analizamos la primera posición de las secuencias de la Tabla 1 se puede observar que el número de apariciones de Alanina (A) es igual a 19, si se divide dicho número entre la cantidad total de secuencias de la tabla (25) se tendrá un valor igual a 0.76 que es la probabilidad de encontrar A en la primera posición dentro de la secuencia.

Tabla 1. Lista o colección de sitios de unión para una proteína hipotética. Fuente: [86].

ATGACATCAT	ATTCGCTAAT	ATTGCGAGAT	GTGTGATCAT	ATGTTGCCAG
ATGCGACAAT	GCTAGCTCAG	ATGCTGATAT	GTACTGACAT	ATGAGATTAT
ATGCTGCCAA	TAGCTAGCAT	TTGTGATGAT	ATGCATTCAG	ATCAGACCAT
ATGCGATAGG	ATCGCGCCAT	TTAGCATGCC	ATGAATACTT	ATGACAGCAT
ATCGACGTAC	ATCGCTACAT	ATTGCATCAG	ATGGACCCCT	ATGATGACTT

Tabla 2. Matriz de puntuación de frecuencia específica para las secuencias de la Tabla 1. Fuente: [86].

	1	2	3	4	5	6	7	8	9	10
A	0.76	0.04	0.08	0.28	0.12	0.44	0.24	0.12	0.80	0.04
C	0.00	0.04	0.12	0.32	0.28	0.12	0.28	0.68	0.08	0.04
T	0.12	0.92	0.16	0.16	0.28	0.12	0.40	0.08	0.08	0.68
G	0.12	0.00	0.64	0.24	0.32	0.32	0.08	0.12	0.04	0.24

Matemáticamente los valores de la Tabla 2 pueden obtenerse a través de la siguiente expresión:

$$P_{i,j} = \frac{1}{N}\sum_{k=1}^{N} I(a_{k,i} = j) \qquad (2),$$

donde $P_{i,j}$ denota la probabilidad de encontrar el aminoácido estándar j en la posición i, con $i \in \{1, \dots, L\}$ y $j \in \{1, \dots, 20\}$. Por otro lado, $k \in \{1, \dots, N\}$, denota el número de secuencias homólogas dentro del alineamiento, en este sentido $a_{k,i}$ representa el aminoácido en la posición i de la secuencia k. Por último, la función I retorna un valor binario teniendo en cuenta la coincidencia con uno de los 20 aminoácidos estándar como se describe en la ecuación (3).

$$I(P_{k,i} = j) = \begin{cases} 1, & si\ a_{k,i} = j \\ 0, & si\ a_{k,i} \neq j \end{cases} \qquad (3)$$

Una vez se tiene los valores de probabilidad de cada posición dentro de la secuencia se forma la siguiente matriz:

$$\text{PSFM} = \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,20} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ P_{L,1} & P_{L,2} & \cdots & P_{L,20} \end{bmatrix} \quad (4),$$

donde las filas de la matriz (L) representan las posiciones de la secuencia y las columnas de la matriz denotan la probabilidad de aparición de cualquiera de los veinte aminoácidos estándar [87].

- **Matriz de puntuación de posición específica (PSSM).**

La matriz de posición específica es deducida a partir de la PSFM donde cada uno de los valores se obtiene con la siguiente expresión

$$W_{ij} = log_2\left(\frac{P_{i,j}}{b_k}\right) \quad (5),$$

definiendo a b_k como la frecuencia con que una letra puede aparecer en el conjunto de datos. El modelo más simple asume que cada letra aparece con igual frecuencia, por lo cual se puede aproximar $b_k = 0.25$ para nucleótidos y $b_k = 0.05$ para aminoácidos. Finalmente, una vez se realizan los cálculos se obtiene la matriz de la ecuación (6) de dimensión $L \times 20$ [87].

$$\text{PSSM} = \begin{bmatrix} W_{1,1} & W_{1,2} & \cdots & W_{1,20} \\ W_{2,1} & W_{2,2} & \cdots & W_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ W_{L,1} & W_{L,2} & \cdots & W_{L,20} \end{bmatrix} \quad (6)$$

La estructura secundaria hace referencia a segmentos tridimensionales locales de una proteína que se forman después de que los residuos se juntan en una secuencia y antes de que la proteína se pliegue en la estructura terciaria [88]. En la estructura secundaria intervienen enlaces de hidrógenos a lo largo de la columna vertebral de la cadena polipeptídica que causan formas locales de plegamiento [18], [88]. La principal relevancia de identificar la estructura secundaria de una proteína radica en el

importante rol que desempeña en el plegamiento de la proteína [18]. Por este motivo, la clasificación de la estructura secundaria es utilizada como características de entrada para modelos de predicción de estructuras de proteínas y mapas de contacto [18], [39], [85]. Según el diccionario de estructura secundaria propuesto en [89], estas pueden ser agrupadas en ocho o tres categorías, para el caso específico de esta investigación será considerada una clasificación basada en los tres tipos de estructura secundaria que se presentan en la Figura 7. En consecuencia, el modelo de predicción de mapas de contacto utilizará como características de entrada tres valores que establecen la probabilidad con que un aminoácido dentro de la secuencia forma parte de una hélice-α (H), una lámina-β (E) o un giro (C) [39].

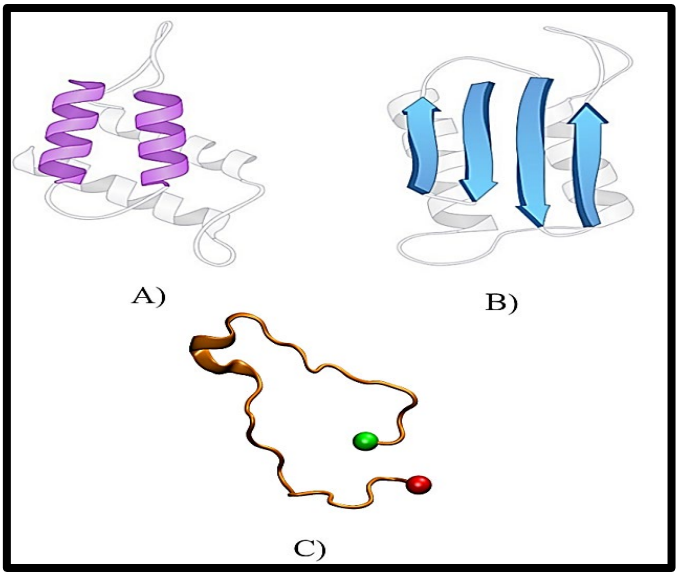


Figura 7. Clasificación de estructura secundaria teniendo en cuenta el enfoque de tres estados: A) hélice-α, B) lámina-β y C) giro.

- **Accesibilidad al solvente (ACC).**

El concepto de accesibilidad al solvente de los residuos en proteínas globulares fue introducido por primera vez en 1971 gracias a Lee y Richards [90], definiendo el termino área de superficie accesible al solvente (SASA) como la medida en que los átomos en la superficie de una proteína pueden formar contactos con el solvente. Se ha observado que la accesibilidad al solvente de un residuo juega un papel importante en la disposición espacial y el empaquetado de la proteína, lo cual ha provocado que

se preste especial atención a la identificación de esta característica para utilizarla como información complementaria al momento de definir la estructura terciaria de una proteína [91]. Al igual que la estructura secundaria esta métrica caracteriza a cada uno de los residuos dentro de una cadena de aminoácidos con tres valores de probabilidad que denotan que tan probable un determinado aminoácido tiene una accesibilidad al solvente expuesta, intermedia o enterrada (*exposed, intermediate, buried*) [91].

2.5.2. Características de coevolución y potencial de contacto

A diferencia de las características secuenciales este conjunto de atributos no caracterizará residuos aislados, si no parejas de estos. Lo anterior significa que cada una de estas características estarán definidas por una matriz 2D que al ser integradas formarán una matriz tridimensional X_c de dimensión $L \times L \times 4$, siendo L la longitud de la proteína. Las matrices bidimensionales utilizadas para entrenar el modelo son: Matriz de precisión normalizada generada por CCMpred [92], dos matrices de información mutua [93], y finalmente una matriz que refleja el potencial de contacto entre parejas de residuos [94].

2.6. Redes neuronales residuales

La aplicación de redes neuronales convolucionales profundas ha representado un avance en los modelos de aprendizaje automático, dada su capacidad para generar características en varios niveles de abstracción, lo que ha generado un nuevo marco de referencia que ha brindado mejores índices de precisión en diversas tareas de predicción [95]. Una vez confirmado el potencial de este tipo de arquitecturas se creó una tendencia por parte de los investigadores en implementar cada vez redes neuronales más profundas (más capas ocultas), para intentar abordar tareas complejas de predicción [16], [96]. Sin embargo, se ha podido observar que a medida que se incrementan las capas en una red neuronal se vuelve más difícil entrenarlas. Esto debido a la saturación de la precisión y a su posterior degradación en lo que se conoce como el problema del desvanecimiento/explosión del gradiente [96].

Las redes neuronales residuales o ResNets por sus siglas en inglés, intentan resolver el problema del desvanecimiento/explosión del gradiente agregando bloques

residuales, los cuales crean atajos o “conexiones de salto” que permiten “saltarse” capas intermedias para referenciar directamente entradas. Matemáticamente un bloque residual se puede definir con la siguiente expresión:

$$y = \mathcal{F}(x, \{W_i\}) + W_s x \quad (7),$$

donde x y y denotan la entrada y la salida de una capa en específica respectivamente. La función $\mathcal{F}(x, \{W_i\})$ representa el mapeo residual que se entrena, es decir $\mathcal{F} = W_{i+1}\sigma(W_i x)$, siendo σ la función de activación de ReLu [97]. El término $W_s x$ forma el atajo que conecta la capa de entrada con la capa de salida. Los pesos W_s son parámetros entrenables, cuya función es ajustar la dimensión de x, para que su adición con $\mathcal{F}(x, \{W_i\})$ sea posible, aunque esto también puede ser logrado concatenando a la entrada vectores o matrices de ceros [16].

La arquitectura del modelo de predicción de mapas de contacto que se implementa en este proyecto de investigación está basada en los trabajos de [39], [85], donde se utiliza el bloque residual de la Figura 8, en su forma unidimensional para las características secuenciales, y en su forma bidimensional para procesar la unión de las características de coevolución y el potencial de contacto. Cabe resaltar que a diferencia de trabajos como: [32], [37], [83], [84], que realizan predicciones individuales de contactos, la salida final del modelo utilizado será una matriz de tamaño $L \times L$ que representará la predicción completa del mapa de contacto. Lo anterior es de vital importancia puesto que establecerá un criterio de discriminación para la selección de métodos de RD.

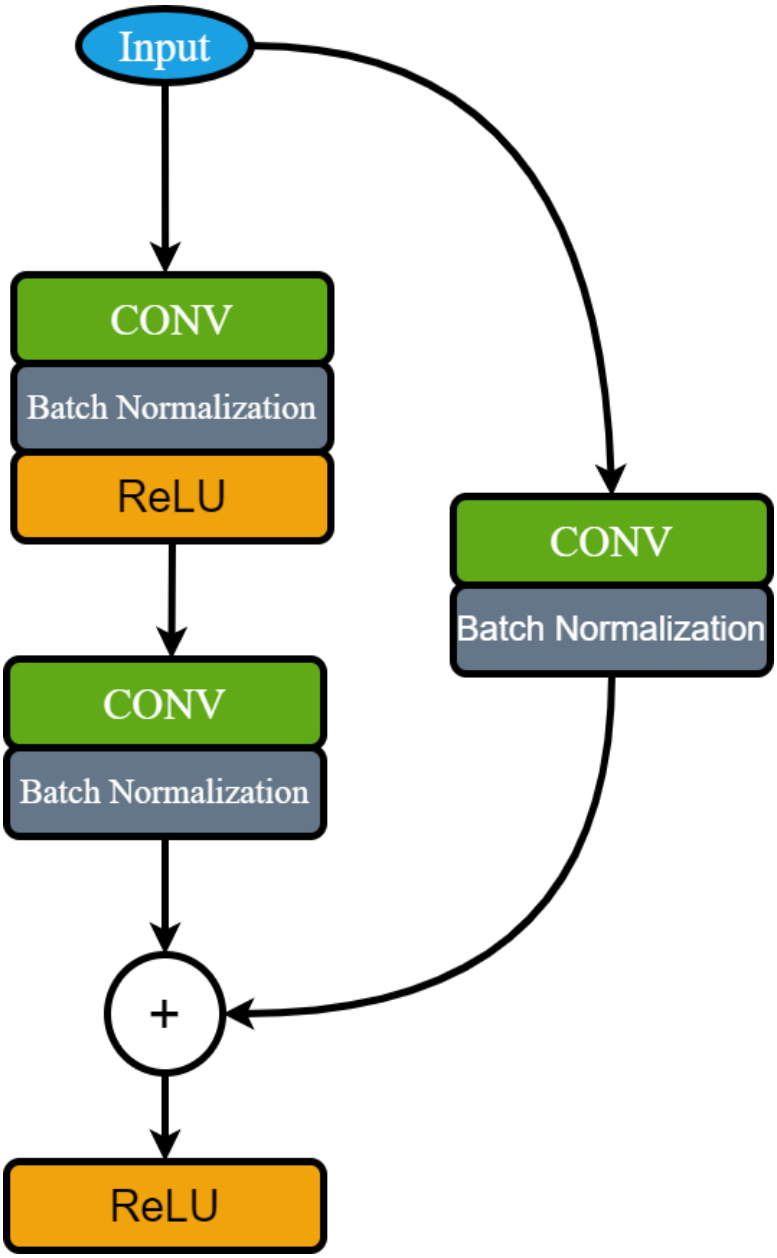


Figura 8. Arquitectura del bloque residual utilizado en el modelo de predicción de mapas de contacto.

Capítulo 3

3. Revisión Sistemática

En este capítulo se describe el proceso llevado a cabo para la realización de la revisión sistemática que permitió la selección de los métodos de reducción de dimensión que serán implementados e integrados al modelo de predicción de mapas de contacto.

3.1. Planeación de la revisión

3.1.1. Protocolo de investigación

La revisión sistemática como metodología de investigación es utilizada en el presente trabajo para abarcar y analizar de manera ordenada el material científico relacionado con la aplicación de métodos de reducción de dimensión en la etapa de preprocesamiento. La información recolectada durante esta etapa del proyecto será considerada para tomar decisiones con respecto a los métodos y técnicas que serán adaptadas finalmente al modelo de predicción.

3.1.2. Etapa de planificación

Antes de empezar a estructurar las etapas de la revisión sistemática y armar la cadena de búsqueda fue necesario llevar a cabo una inspección de unos primeros trabajos relacionados, que ayudaron a definir aspectos iniciales dentro de la planificación de esta revisión. Para el caso específico de esta investigación inicialmente se tuvieron en cuenta los siguientes aspectos: el estudio de las características del conjunto de datos utilizado; la revisión de los trabajos [39], [85], con sus respectivos repositorios; y el análisis de literatura gris basada en los artículos tipo revisión [98]–[100].

Una vez finalizado este proceso preliminar de apropiación del conocimiento, se procedió a estructurar el desarrollo de la revisión sistemática teniendo en cuenta los conceptos obtenidos y el proceso de revisión descrito en [101], [102].

En la Figura 9 se indica el esquema general de tres fases que define el procedimiento que se llevó a cabo para el desarrollo de la revisión sistemática. En la primera fase principalmente se abordan aspectos como: la construcción de las preguntas de investigación, la selección de fuentes de información y la organización de palabras clave en una cadena de búsqueda. A continuación, en la segunda fase se realiza una identificación e interpretación de estudios relevantes para posteriormente seleccionar los artículos primarios. Finalmente, en la fase tres, con los artículos primarios seleccionados, se lleva a cabo un proceso de extracción, análisis y evaluación de información para escoger los métodos de reducción de dimensión que formarán parte del estudio comparativo. Adicionalmente, es importante destacar que, la planeación y la ejecución de la revisión sistemática representan etapas iterativas puesto que es posible volver a definir sus procesos, si los resultados no concuerdan con los objetivos y los resultados que se quieren obtener.

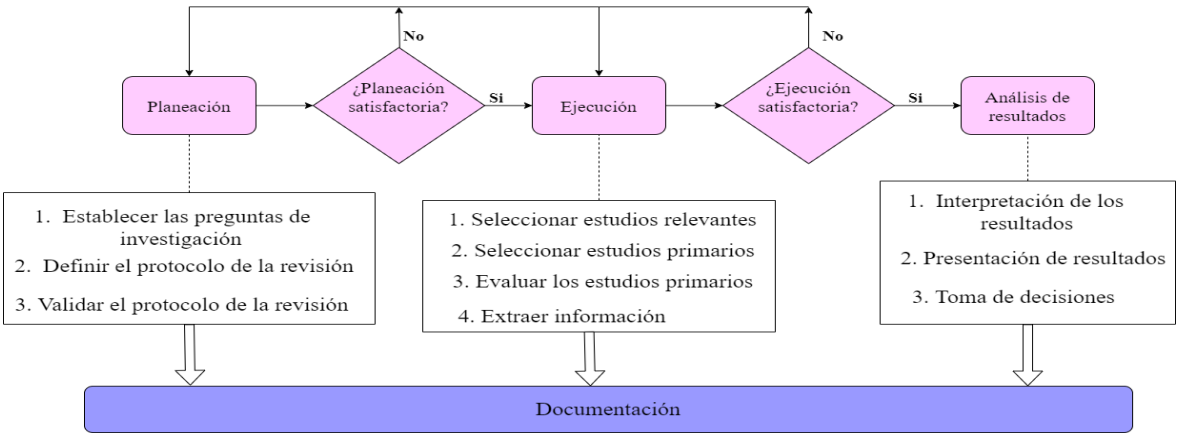


Figura 9. Esquema general de las fases definidas para el desarrollo la revisión sistemática. Fuente: [101], [102].

3.1.3. Preguntas de investigación

Las técnicas de reducción de dimensión tienen por objetivo construir un espacio de características reducido con la menor perdida de información posible, donde se conserven ciertas propiedades estructurales del espacio original [99]. Estas técnicas proporcionan una herramienta para el preprocesamiento de los datos, cuyo uso puede verse reflejado en una reducción del espacio de almacenamiento y en la supresión de

datos redundantes, irrelevantes o ruidosos. Sin embargo, no todos los métodos de reducción de dimensión disponibles pueden ser utilizados en procesos de clasificación o predicción, puesto que algunos de ellos se enfocan en tareas de visualización o el cálculo del espacio embebido no se ajusta al proceso de predicción o clasificación que se quiere llevar a cabo [100].

Como se explicó en el Capítulo 2: Marco Conceptual y Estado del Arte, el modelo de predicción requiere de dos fuentes de información: La primera basada en características secuenciales (matriz bidimensional); la segunda compuesta por el conjunto de características de coevolución (matriz tridimensional). Siendo en las primeras donde se aplicarán los métodos de reducción de dimensión seleccionados para el desarrollo del estudio comparativo. Por este motivo, el foco de la revisión sistemática estará centrado en la búsqueda de trabajos de investigación que integren métodos de reducción de dimensión como etapa de preprocesamiento a conjuntos de características secuenciales similares a las utilizadas para alimentar el modelo de predicción, y más importante aún que puedan ser adaptados a la arquitectura del modelo de predicción implementado.

Mediante el desarrollo de la revisión sistemática se pretende identificar tendencias en las propuestas y trabajos existentes que utilicen métodos de reducción de dimensión para comprimir un conjunto de características basadas en atributos obtenidos con información de la secuencia de la proteína. Esto permitirá obtener un panorama más amplio que permita identificar los métodos de reducción de dimensión más utilizados, para posteriormente realizar una selección final basada en la adaptabilidad de método a la arquitectura del modelo implementado.

Teniendo en cuenta lo expuesto anteriormente y el resultado que se quiere obtener con la revisión sistemática se formulan las preguntas de investigación que se presentan en la Tabla 3.

Tabla 3. Preguntas de investigación de la revisión sistemática.

Pregunta de investigación	Objetivo
Q1. ¿qué métodos de reducción de dimensión han sido utilizados para procesar características secuenciales de proteínas?	Identificar los métodos de reducción de dimensión que han sido utilizados para procesar características secuenciales dentro de procesos de clasificación o predicción.
Q2. ¿qué métodos de reducción de dimensión han permitido una mejora en el desempeño de modelos de predicción o clasificación?	Determinar los métodos de reducción de dimensión aplicados a características secuenciales que han mejorado el desempeño de modelos de clasificación o predicción
Q3. ¿qué métodos de reducción de dimensión son capaces de procesar un número considerable de elementos?	Determinar los métodos de reducción de dimensión que han sido utilizados en conjuntos de datos con un elevado número de elementos.
Q4. ¿qué métodos de reducción de dimensión han sido utilizados para reducir características secuenciales en la predicción de mapas de contacto?	Identificar que métodos de reducción de dimensión han sido utilizados en el proceso de predicción de mapas de contacto como etapa de preprocesamiento de características.
Q5. ¿cuáles son los métodos de reducción de dimensión que pueden ser adaptados a la arquitectura del modelo de predicción de mapas de contacto implementado?	Determinar los métodos de reducción de dimensión que pueden ser integrados al modelo de predicción implementado.

3.1.4. Selección de las fuentes

Una parte fundamental para el cumplimiento de los objetivos y la obtención de los resultados esperados en la revisión sistemática es la selección de las fuentes de literatura, puesto que proveen la población de trabajos y artículos académicos que serán analizados. Las fuentes de información fueron seleccionadas principalmente debido a que proveen bases de datos electrónicas relacionadas con la temática a tratar y cuentan con motores de búsqueda habilitados vía internet. Asimismo, cabe resaltar que las fuentes utilizadas tienen a su disposición gran variedad de artículos científicos que han sido publicados en revistas de alto impacto y que han sido base de estudio otros trabajos (elevado número de citaciones).

En total fueron cinco las fuentes electrónicas seleccionadas con el fin de obtener un espectro más extenso que permita encontrar la mayor cantidad de artículos relevantes.

Adicionalmente, al aplicar cadenas de búsquedas similares en diferentes fuentes permite no solo encontrar posibles aciertos o inconsistencias en los resultados, si no también puede ayudar a ajustar la cadena de búsqueda que se va a emplear. Dentro de las fuentes seleccionadas para desarrollar la revisión sistemática se encuentran: Google Scholar, Scopus, ScienceDirect, IEEE Xplore y SpringerLink.

3.1.5. Diseño de la búsqueda

Considerando el número de trabajos y la variedad de áreas que implementan métodos de reducción de dimensión en la etapa de preprocesamiento, se decidió enfocar la búsqueda en trabajos que apliquen dichos métodos, y en donde intervengan características secuenciales como: estructura secundaria, accesibilidad al solvente, matriz de puntuación de posición específica (PSSM) y matriz de frecuencia de posición específica (PSFM), las cuales son empleadas para entrenar el modelo de predicción de mapas de contacto.

En la Tabla 4 se muestran las cadenas de búsqueda definidas, las cuales combinan los conectores lógicos “AND” y “OR” con algunas palabras clave identificadas a partir del estudio de las características del conjunto de datos y de la literatura gris. El proceso de aplicación de las cadenas de búsqueda se realizó de manera iterativa, afinando la forma de la cadena de búsqueda y los elementos que la componen cuando los artículos presentados no eran coherentes con lo que deseaba encontrar. Adicionalmente, se estableció una ventana temporal de cinco años con el fin de recolectar trabajos recientes y observar las tendencias actuales relacionadas con la aplicación de métodos de reducción de dimensión.

En la mayoría de las fuentes de información seleccionadas se utilizó la misma cadena de búsqueda, exceptuando por IEEE Xplore donde fue necesario dividir la búsqueda en tres cadenas separadas, puesto que la inclusión de todas las palabras clave inicialmente consideradas generaban resultados que no estaban relacionados con los resultados obtenidos en los otros motores de búsqueda.

Tabla 4. Cadenas de búsqueda realizadas el 28 de abril del 2022.

Buscador	Cadena de búsqueda	Número de publicaciones
Google Scholar	("dimension reduction" OR "dimensionality reduction" OR "feature reduction") AND ("classification" OR "prediction") AND ("secondary structure" OR "solvent accessibility" OR "PSSM" OR "PSFM")	2.470
Scopus	("dimension reduction" OR "dimensionality reduction" OR "feature reduction") AND ("classification" OR "prediction") AND ("secondary structure" OR "solvent accessibility" OR "PSSM" OR "PSFM")	19
Science Direct	("dimension reduction" OR "dimensionality reduction" OR "feature reduction") AND ("classification" OR "prediction") AND ("secondary structure" OR "solvent accessibility" OR "PSSM" OR "PSFM")	270
Springer Link	("dimension reduction" OR "dimensionality reduction" OR "feature reduction") AND ("classification" OR "prediction") AND ("secondary structure" OR "solvent accessibility" OR "PSSM" OR "PSFM")	109
IEEE Xplore	"protein" AND "dimensionality reduction"	29
	"protein" AND "dimension reduction"	
	"protein" AND "feature reduction"	

3.1.6. Criterios de evaluación para artículos primarios

Una parte fundamental dentro de las actividades que intervienen en el desarrollo de la revisión sistemática es la evaluación de los artículos primarios que serán utilizados como referencia para este proyecto de investigación. La principal razón de llevar a cabo este proceso de cuantificación es el de establecer la calidad de la información que contiene cada uno de los trabajos seleccionados y su potencial aporte a este proyecto de investigación.

En esta revisión se planteó evaluar cada uno de los estudios primarios teniendo en cuenta los puntos observados en la Tabla 5, los cuales están directamente relacionados con las preguntas de investigación que se plantearon anteriormente. Cada uno de los artículos seleccionados tendrán asignado un valor que podrá estar entre 0 y 13, siendo este ultimo la máxima calificación que se puede obtener. El primer ítem busca evaluar el impacto de los resultados a través de la medición del cuartil de la revista donde se publicó el trabajo. El Segundo ítem considerado evalúa la similitud de las características secuenciales utilizadas con el conjunto de datos que alimenta el

modelo de predicción de mapas de contacto. En el tercer ítem de naturaleza binaria, mide si el trabajo aborda el problema de predicción de mapas de contacto, aplicando una etapa de preprocesamiento basada en reducción de dimensión. El cuarto y último ítem está pensado para evaluar el volumen de datos que el algoritmo de reducción de dimensión tiene que manejar.

Tabla 5. Ítems a evaluar dentro de los estudios primarios seleccionados.

Ítem a evaluar	Puntuación posible
El artículo fue publicado en una revista de alto impacto	0-4
Grado de similitud de las características utilizadas en el artículo, con las definidas para entrenar el modelo de predicción de mapas de contacto.	0-4
El problema que aborda el trabajo está centrado en la predicción de mapas de contacto.	0 o 1
El volumen de elementos que la etapa de preprocesamiento basada en reducción de dimensión maneja.	0-4

3.2. Ejecución de la revisión

La ejecución de la revisión sistemática tiene como objetivo obtener dos tipos de información que permitan la selección de los métodos de reducción de dimensión que serán utilizados. El primer tipo de información obtenida de la selección de estudios relevantes busca obtener un panorama global que identifique los métodos de reducción de dimensión que son utilizados y su frecuencia. Por su parte el segundo tipo de información a partir de los estudios primarios busca obtener una perspectiva más específica reflejada en las comparaciones de espacios de características reducidos y su influencia en el desempeño del modelo.

3.2.1. Selección de artículos relevantes

Una vez se aplica la cadena definida en la Tabla 1 se procede inicialmente a realizar un filtro que permita identificar y registrar estudios relevantes. Para este fin, se hace el análisis del título, *abstract*, palabras clave y de la sección en donde se presenta el preprocesamiento de las características. De este modo, se obtiene unos primeros criterios de discriminación basados, en la similitud de la tarea que aborda el artículo (problema y característica secuenciales que utiliza) con la predicción de mapas de

contacto, y en la detección de una etapa de preprocesamiento basada en métodos de reducción de dimensión. No obstante, dado el caso que no se obtenga la información necesaria de estos apartados, se hará un análisis de otras partes del documento para definir si cumple con los objetivos planteados en esta revisión.

Para la selección de estudios relevantes se tienen esencialmente los siguientes criterios de inclusión.

Criterios de inclusión

- La inclusión de un artículo relevante se llevará a cabo si y solo si se cumplen los siguientes aspectos: primero en la similitud del problema que el estudio aborda con la predicción de mapas de contacto; segundo que utilice características secuenciales similares; tercero y más importante que en el trabajo exista una etapa de preprocesamiento de características basada en reducción de dimensión.

Criterios de exclusión

- Que no se integre en el estudio una etapa de preprocesamiento basada en reducción de dimensión.
- Que el método de reducción de dimensión esté integrado internamente al modelo de predicción o clasificación y no como una etapa independiente de preprocesamiento
- Que el idioma del articulo no sea idioma inglés.
- Que el articulo no pueda ser descargado.
- El área de estudio sea diferente a las ciencias de la computación.
- Estudios que hayan sido encontrados previamente en otras de las fuentes seleccionadas.

Con la extracción de artículos relevantes se busca obtener una perspectiva global que forme parte de la información de referencia para la posterior selección de métodos de reducción de dimensión. Con el análisis superficial de dichos trabajos se quiere adquirir tres datos: el primero relacionado con el espectro de los métodos de reducción de dimensión que han sido utilizados; el segundo define las tasas de artículos que utilizan métodos de reducción de dimensión basados en selección de características, extracción de características; el tercero establece la frecuencia por separado del

conjunto de algoritmos basados en selección de características y el conjunto de algoritmos basados en extracción de características.

3.2.1.1. Ejecución del protocolo de búsqueda en Google Scholar

La primera fuente electrónica en donde se aplicó la cadena de búsqueda fue en Google Scholar, puesto que arroja el mayor número de resultados, además de permitir recopilar y seleccionar la información mediante el uso de web scraping [103].

El proceso llevado a cabo para recolectar, organizar y almacenar la información que provee Google Scholar es descrito en la Figura 10, donde la cadena de búsqueda previamente definida es utilizada para generar una URL que permitirá la recolección automatizada de metadatos relacionados con los 2.470 artículos. La ejecución del algoritmo de búsqueda implementado obtiene la información de cada una de las páginas resultantes, las cuales agrupan los artículos en conjuntos de diez elementos. No obstante, el programa de web scraping ejecutado únicamente iteraba y recopilaba información hasta la página número cien (1000 resultados). Al intentar obtener información de páginas superiores se pudo observar que los metadatos obtenidos contenían valores vacíos. Para comprobar posibles errores en la implementación del algoritmo se ingresó a Google escolar y se verifico manualmente que efectivamente este motor de búsqueda no retornaba resultados más allá de la página cien, lo cual concuerda con la salida del programa.

En total de los 2.470 resultados que reporta Google Scholar, se recuperaron 920 artículos en forma de metadatos. Cada uno de los 920 artículos son extraídos en forma de diccionarios, cuyas claves representan atributos como: titulo, sitio web de la publicación, autores, año de publicación, abstract y editorial. Con esta información disponible se empleó una etapa de filtrado que consistió en descartar aquellos trabajos en donde no se encontraran coincidencias entre los términos que componen tanto el título, como el abstract, y las expresiones, "dimension reduction", "dimensionality reduction" y "feature reduction". Con la aplicación de este filtro se reduce el número de resultados a 696.

Cada uno de los artículos obtenidos en esta instancia fueron almacenados en un archivo csv, el cual está compuesto por elementos caracterizados con los siguientes metadatos: título, autores, editorial y año. Si bien en la mayoría de los casos la información obtenida por el algoritmo estaba completa, en algunos artículos fue necesario completar manualmente los campos que estaban vacíos. Finalmente, con el archivo csv completo se procedió a ejecutar el protocolo de inclusión y exclusión de estudios relevantes, obteniendo así un número de 56 artículos.

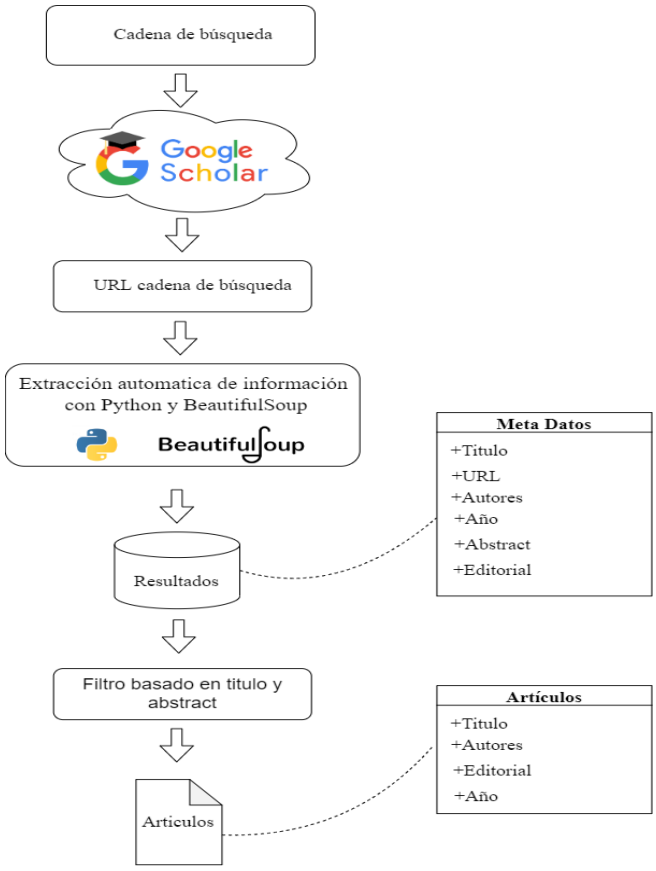


Figura 10. Esquema de *web scraping* llevado a cabo en la fuente *Google Scholar*

3.2.1.2. Ejecución del protocolo de búsqueda en otras fuentes

Con la recolección de unos primeros estudios relevantes se procedió a aplicar las cadenas de búsqueda en las fuentes electrónicas restantes. En Scopus se obtuvieron un total de 19 trabajos, sin embargo, solo uno de los artículos fue agregado a los

estudios relevantes, puesto que en su gran mayoría todos los fueron identificados previamente. Cuando se realizó el proceso manual de revisión en Google Scholar se pudo observar que una gran cantidad de trabajos estaban disponibles en Science Direct y Springer Link, lo cual concuerda con los resultados preliminares de la Tabla 4. Por este motivo de los 270 resultados encontrados en Science Direct únicamente dos fueron agregados como estudios relevantes, por su parte en Springer Link se obtuvieron 109 resultados donde solo uno fue seleccionado. En IEEE Xplore la cadena de búsqueda tuvo que ser modificada para poder obtener artículos coherentes con esta revisión sistemática. En total para esta fuente, se encontraron 29 artículos de los cuales se clasificó uno como artículo relevante.

3.2.1.3 Extracción de información de artículos relevantes

Con la finalización de la etapa de selección de estudios relevantes se obtuvieron los resultados que se muestran en la Tabla 6, donde se observa un total de 61 artículos. Para facilitar el análisis de cada uno de los trabajos se les agregó campos adicionales para describir aspectos como: La clase de documento (artículo de revista, articulo de conferencia o tesis); el tipo de método de reducción de dimensión (selección de características o extracción de características); el algoritmo especifico que aplican (PCA, LLE, Autoencoders, Lasso etc.); el número de citaciones y el impacto de la revista donde el trabajo fue publicado (cuartil).

Tabla 6. Número de estudios relevantes seleccionados en cada una de las fuentes.

Fuente	Número de estudios Relevantes
Google Scholar	56
Scopus	1
Science Direct	2
Springer Link	1
IEEE Xplore	1
Total	61

En la Figura 11 se muestra una primera contextualización que permite observar cual es el método de reducción de dimensión más utilizado entre la selección de características, la extracción de características y entre trabajos que mezclan estos dos tipos de métodos en la etapa de preprocesamiento. Como se puede observar el

número de trabajos seleccionados que hacen uso de métodos reducción de dimensión basados en extracción de características es mayor a los demás. No obstante, la diferencia con los métodos de selección de características no es muy elevada.

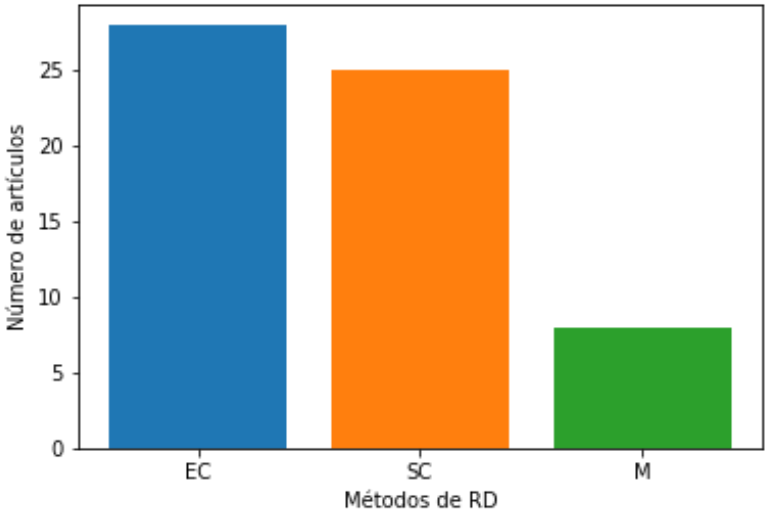


Figura 11. Distribución del número de artículos en función del método de reducción que utilizan: Extracción de características (EC), selección de características (SC), una mezcla de las dos (M).

Posteriormente se analizó por separado la frecuencia de las técnicas de extracción y selección de características como se indican en las Figura 12 y 13. Lo primero que se puede observar en los diagramas es el amplio espectro de técnicas que se han utilizado en los trabajos que están dentro de los criterios de selección de estudios relevantes. Lo anterior denota no solo la vigencia de dichas técnicas a pesar del auge de las arquitecturas basadas en aprendizaje profundo, si no su potencial aporte en el desempeño de los modelos actuales.

La muestra la gran variedad de algoritmos de extracción de características que han sido utilizados como etapa de preprocesamiento, en donde se pueden encontrar desde técnicas lineales como: Principal Component Analysis (PCA), Lineal Discriminant Analysis (LDA), Local Fisher Discriminant Análisis (LDA), Random projections RP; hasta técnicas no lineales como: Autoencoders (AE), Kernel Principal Component Analysis (KPCA), Locally Linear Embedding (LLE), Uniform Manifold Approximation and Projection (UMAP) y Locality Preserving Projections (LPP). No obstante, En la Figura 13 también se muestra una amplia gama de algoritmos de selección de

características basados en: métodos integrados como: Lasso y Elastic Net (EN); métodos de envoltura como: Recursive Feature Elimination (RFE); y métodos de filtro como: the Max-Relevance-Max-Distance (MRMD), Maximum-Relevance-Minimum-Redundancy (MRMR), XGBoost, RelieF y LightGBM.

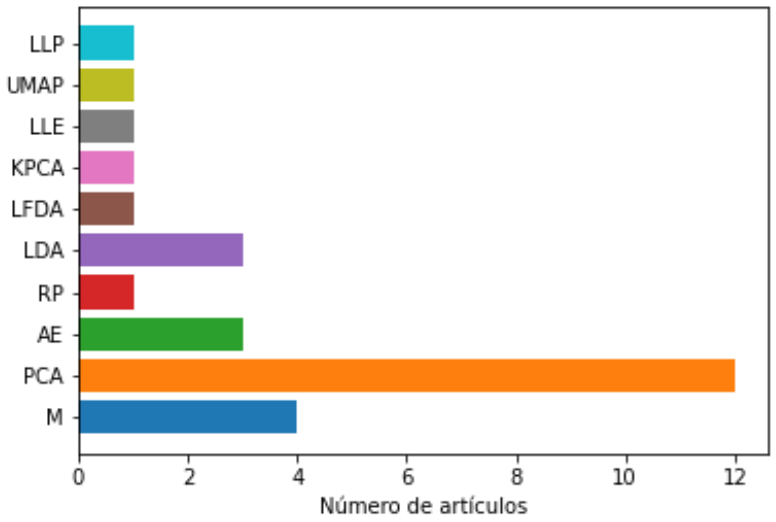


Figura 12. Distribución del número de artículos en función del algoritmo de extracción de características que implementan.

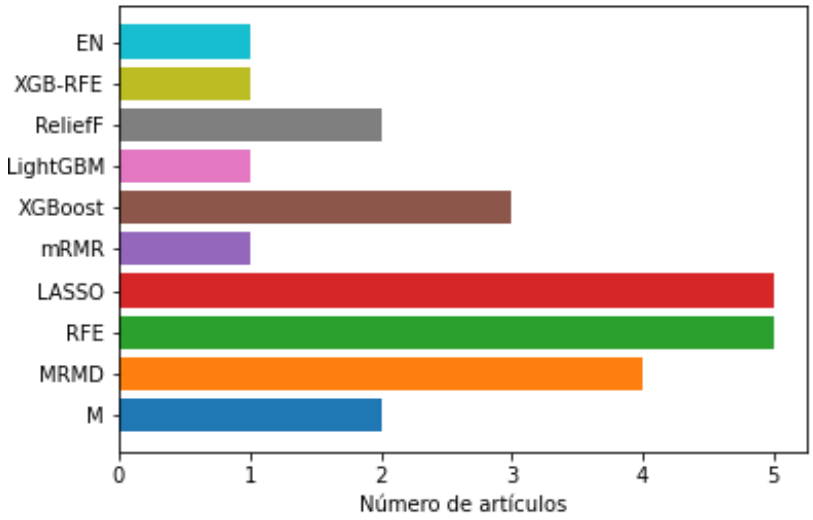


Figura 13. Distribución del número de artículos en función del algoritmo de selección de características que implementan.

3.2.2. Selección de artículos primarios

Con los estudios relevantes establecidos se procede a realizar un análisis más detallado de los artículos, para abarcar secciones adicionales del documento que permitan verificar los criterios de inclusión y exclusión que se definen para los artículos primarios. Especialmente se presta atención a las secciones que describen la etapa de preprocesamiento y los resultados, para detectar contrastes en el desempeño del modelo cuando se comparan varios métodos de reducción de dimensión.

Criterios de inclusión

- El criterio de inclusión de los estudios primarios se basa en detectar aquellos trabajos dentro de los estudios relevantes, que además de integrar métodos de reducción de dimensión en la etapa de preprocesamiento, establecen una comparación de desempeño entre las características reducidas, el espacio original, u otros métodos de reducción de dimensión.

- Que el método de reducción de dimensión que usa como etapa de preprocesamiento se integre con un modelo basado en Deep Learning.

- Que el método de reducción de dimensión sea integrado específicamente a un modelo de predicción de mapas de contacto.

Criterios de Exclusión

- Los criterios de exclusión se basan en descartar aquellos trabajos que cumplan con las siguientes condiciones: que no exista un contraste de desempeño del método de reducción de dimensión utilizado, bien sea con el espacio original o con otros espacios embebidos; que se aplique un método de reducción de dimensión, pero no sea integrado a un modelo de Deep learning; que el preprocesamiento basado en reducción de dimensión no sea integrado explícitamente con un modelo de predicción de mapas de contacto.

La selección y el análisis de los artículos primarios a diferencia de la información brindada por los estudios relevantes busca obtener una visión más específica de los métodos de reducción de dimensión. Esto quiere decir que además del estudio de las frecuencias dentro de los estudios primarios, se buscará establecer un punto de comparación basado en el aporte que los algoritmos de selección o extracción de características ofrecen al desempeño del modelo de predicción o clasificación.

La selección y el análisis de los artículos primarios a diferencia de la información brindada por los estudios relevantes busca obtener una visión más específica de los métodos de reducción de dimensión. Esto quiere decir que además del estudio de las frecuencias dentro de los estudios primarios, se buscará establecer un punto de comparación basado en el aporte que los algoritmos de selección o extracción de características ofrecen al desempeño del modelo de predicción o clasificación.

3.2.2.1. Extracción de la información

La extracción de información de cada uno de los estudios primarios comienza con la revisión y análisis de secciones específicas dentro del artículo. En una primera instancia se estudia la adquisición y caracterización del conjunto de datos utilizado para obtener parámetros, como: el tipo de características secuenciales que se extraen, la dimensión del vector de características final y el número de elementos que conforma el conjunto de datos de entrenamiento. En segundo lugar, se estudian las secciones del artículo que estén relacionadas con la etapa de preprocesamiento mediante métodos de reducción de dimensión. En esta parte se presta especial atención por encontrar, contrastes con otras técnicas de reducción de dimensión y el procedimiento para determinar el mejor conjunto de características reducidas. Finalmente se analiza detenidamente la sección de resultados para conseguir un punto de referencia que permita establecer cuáles son los métodos de reducción que permitieron mejorar el desempeño del modelo.

Con la aplicación de los criterios de inclusión y selección al conjunto de artículos relevantes se obtienen los estudios primarios que se presentan en la Tabla 7. Al comparar el número de artículos primarios con la fuente de donde se observa que a excepción de un trabajo obtenido de Scopus, todos los estudios primarios fueron

extraídos de Google Scholar. La aparición de esta tendencia se debe mayoritariamente a que una gran cantidad de artículos inicialmente recuperados a través de Google Scholar, están disponibles Science Direct y Springer Link.

Tabla 7. Número de estudios primarios seleccionados en cada una de las fuentes.

Fuente	Número de estudios primarios
Google Scholar	18
Scopus	1
Science Direct	0
Springer Link	0
IEEE Xplore	0
Total	19

Las figuras 14-18 resumen la información extraída de los estudios primarios. Lo primero que se puede observar es que a pesar de que los trabajos utilizan características secuenciales similares e inclusive algunos abordan un mismo problema existe gran variabilidad en los métodos de reducción de dimensión que han brindado los mejores resultados. En consecuencia, el análisis de frecuencias tanto de los estudios relevantes como de los estudios primarios no son suficientes para brindar un criterio claro de selección para dichos métodos. Por este motivo, se utilizará información adicional teniendo en cuenta las respuestas a las preguntas de investigación planteadas en esta revisión, la evaluación de cada uno de los artículos tomando como referencia la Tabla 5, y las especificaciones del conjunto de datos y el modelo de predicción de mapas de contacto utilizado.

La Figura 14 relaciona el tipo de método de reducción de dimensión (extracción de características y selección de características) que se utiliza dentro de los estudios primarios seleccionados. Como se puede observar a diferencia de la Figura 11 donde se relaciona los estudios relevantes, los algoritmos basados en selección de características son los más frecuentes. No obstante, la diferencia aún continúa siendo bastante estrecha.

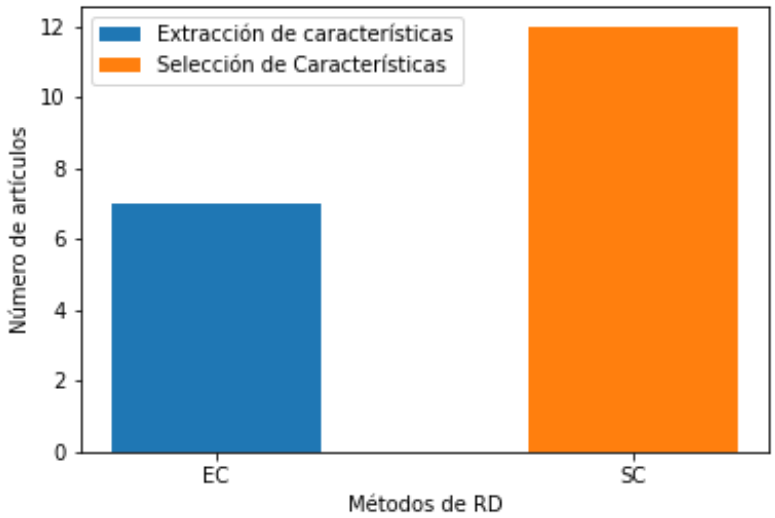


Figura 14. Distribución del número de estudios primarios en función del método de reducción de dimensión que denota el mejor desempeño.

Con la inclusión de trabajos que establecen un contraste de métodos de reducción de dimensión, se extrajeron los tres primeros algoritmos que según el estudio primario han mejorado el desempeño del modelo de predicción o clasificación. En la Figura 15 se observan los algoritmos de reducción de dimensión que finalmente fueron integrados a los modelos y que aportaron al incremento en la calidad de las métricas de evaluación. Las técnicas de selección de características Lasso, XGboost y LigthtGBM sobresalen en esta primera gráfica por ser utilizadas en más de un estudio primario. En la Figura 16 se muestran aquellos algoritmos de reducción de dimensión que tuvieron el segundo mejor desempeño al ser integrados en el modelo, lo que más resalta en este grafico es la frecuencia con que PCA es utilizado, superando incluso la longitud de cualquiera de las barras de la Figura 15. Los algoritmos que presentaron el tercer mejor rendimiento se exponen en Figura 17 donde predominan técnicas de extracción de características como: PCA, LLE y KPCA.

Uno de los aspectos que también se buscó evaluar dentro del artículo fue el número de elementos N que conformaban el conjunto de datos de entrenamiento, y en donde los métodos de reducción de dimensión son aplicados. En la Figura 18 se relaciona en escala logarítmica el número de elementos que usan para entrenan a los modelos, con el estudio primario y el método de reducción de dimensión utilizado (método con el mejor desempeño). Los artículos A11 y A15 representan los más destacable dentro

del diagrama de barras, puesto que manejan un elevado número de elementos, con la aplicación de los algoritmos de selección de características Lasso y XGboost respectivamente. Del mismo modo, se pueden encontrar dos algoritmos de extracción de características PCA y RP, que procesan una cantidad considerable de información en los trabajos A2 y A18.

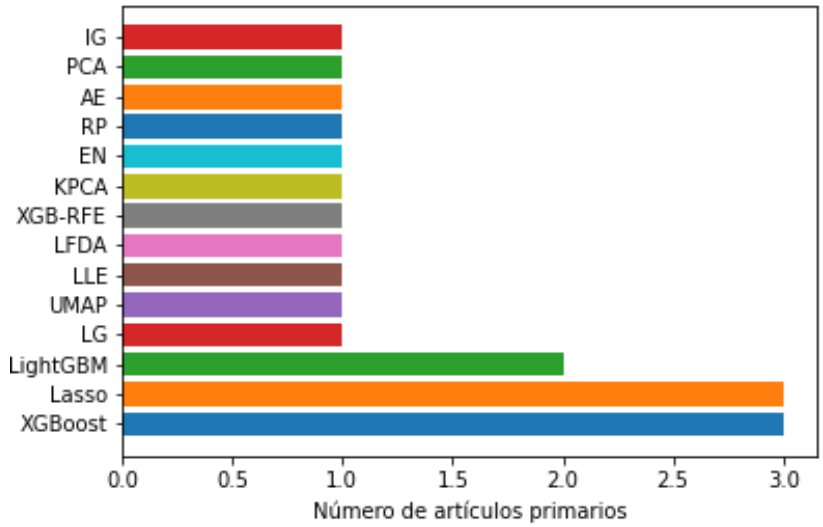


Figura 15. Frecuencia de los métodos de reducción de dimensión que presentaron el mejor desempeño cuando se integraron al modelo de predicción o clasificación.

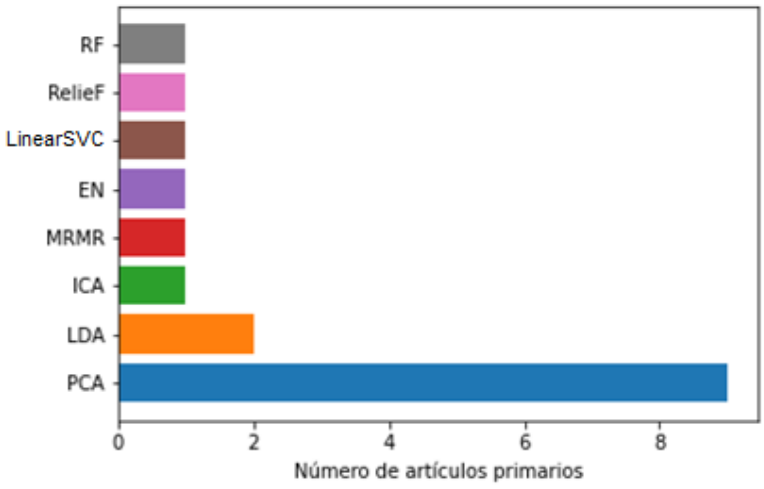


Figura 16. Frecuencia de los métodos de reducción de dimensión que presentaron el segundo mejor desempeño cuando se integraron al modelo de predicción o clasificación.

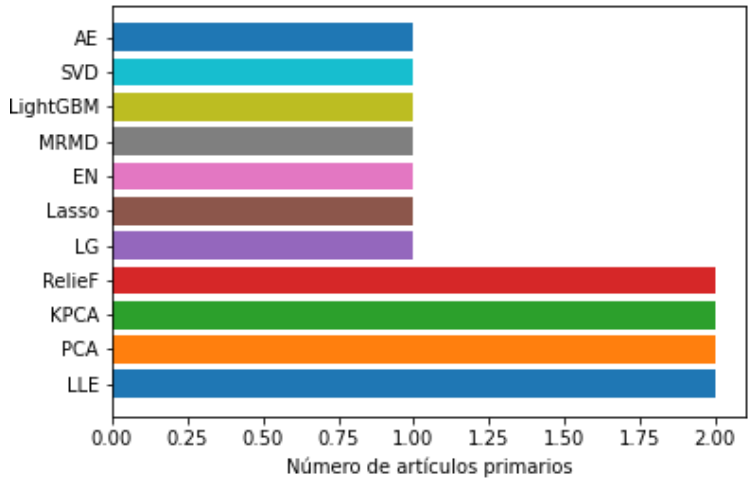


Figura 17. Frecuencia de los métodos de reducción de dimensión que presentaron el tercer mejor desempeño cuando se integraron al modelo de predicción o clasificación.

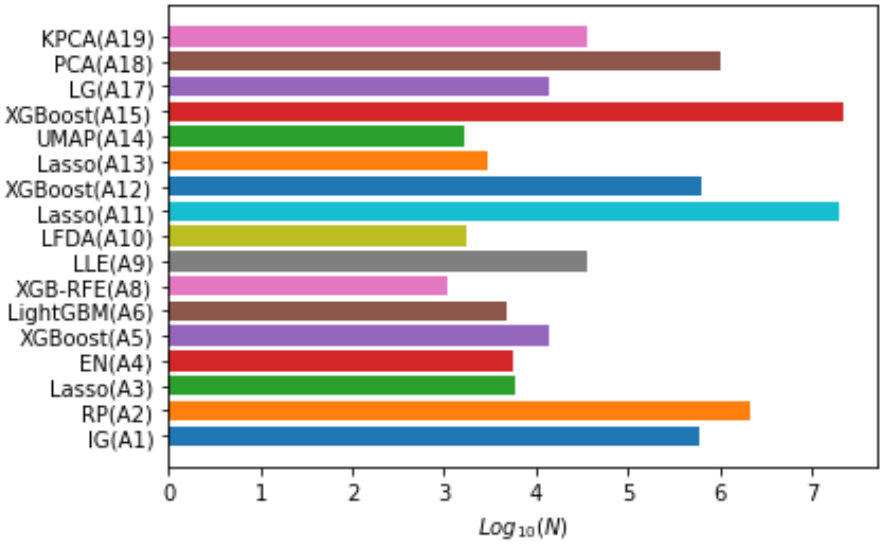


Figura 18. Estudios primarios y el método de reducción de dimensión que implementan en función de la representación logarítmica del número de elementos N que conforman el conjunto de datos de entrenamiento.

3.2.2.2. Evaluación de artículos primarios

Para obtener un punto de referencia adicional a la información extraída en las secciones anteriores, se procede a evaluar cada uno de los estudios primarios

tomando como referencia los ítems establecidos en la Tabla 5, los cuales incluyen aspectos como: el tipo de trabajo (revista, conferencia o tesis); el año de publicación, el cuartil de Scimago; el score de evaluación alcanzado, además de su relación con las preguntas de investigación planteadas en esta revisión.

Tabla 8. Evaluación de estudios primarios seleccionados en esta investigación.

Titulo	Tipo	Año	Scimago Cuartil	Evaluación
A1. Dimensionality reduction for protein secondary structure and solvent accessibility prediction	Revista	2018	4	5
A2. Dimensionality reduction based multi-kernel framework for drug-target interaction prediction	Revista	2021	2	8
A3. Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure	Revista	2019	1	7
A4. Prediction of protein-protein interactions based on elastic net and deep forest	Revista	2021	1	6
A5. Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier	Revista	2020	1	6
A6. Structural protein fold recognition based on secondary structure and evolutionary information using machine learning algorithms	Revista	2021	3	5
A7. DeepStack-DTIs: Predicting Drug-Target Interactions Using LightGBM Feature Selection and Deep-Stacked Ensemble Classifier	Revista	2021	3	5
A8. StackPDB: predicting DNA-binding proteins based on XGB-RFE feature optimization and stacked ensemble classifier	Revista	2021	1	6
A9. A deep learning model for plant lncRNA-protein interaction prediction with graph attention	Revista	2020	2	5
A10. Prediction of human phosphorylated proteins by extracting multi-perspective discriminative features from the evolutionary profile and physicochemical properties through LFDA	Revista	2021	2	5
A11. DeepACTION: A deep learning-based method for predicting novel drug-target interactions	Revista	2020	3	5

A12. DNN-DTIs: Improved drug-target interactions prediction using XGBoost feature selection and deep neural network	Revista	2021	1	7
A13. Fertility-lightgbm: A fertility-related protein prediction model by multi-information fusion and light gradient boosting machine	Revista	2021	2	5
A14. UMAP-DBP: an improved DNA-binding proteins prediction method based on uniform manifold approximation and projection	Revista	2021	0	2
A15. EPSILON-CP: using deep learning to combine information from multiple sources for protein contact prediction	Revista	2017	1	9
A16. Representing Amino Acid Contacts In Protein Interfaces	Tesis	2020	0	3
A17. GTB-PPI: Predict Protein–protein Interactions Based on L1-regularized Logistic Regression and Gradient Tree Boosting	Revista	2020	1	6
A18. Application of linear dimensionality reduction techniques and convolutional neural networks for protein secondary structure prediction	Revista	2020	4	6
A19. Prediction of protein-protein interaction sites through eXtreme gradient boosting with kernel principal component analysis	Revista	2021	1	7

3.3. Análisis de resultados y selección de métodos de reducción de dimensión

En esta última sección se recopila la información extraída de las diferentes etapas del protocolo de búsqueda, para dar respuesta a las preguntas de investigación y finalmente seleccionar los métodos de reducción de dimensión que serán integrados al modelo de predicción de mapas de contacto.

La respuesta a la pregunta de investigación **Q1** se puede obtener estudiando las figuras 11-13, donde se pueden obtener detalles importantes inmersos dentro de los estudios relevantes. Si se analiza la Figura 11 se puede observar que no existe una tendencia definida en la aplicación de métodos de reducción de dimensión basados en extracción de características o selección de características, debido a que son utilizados aproximadamente en igual proporción. Las figuras 11-13 denotan una gran variedad

de algoritmos utilizados para reducir la dimensión del conjunto de datos, donde existe un uso recurrente de PCA, LDA, AE para las técnicas de extracción de características, y de Lasso, RFE, XGBoost, MRMD para las técnicas de selección de características.

Otro importante enfoque de discriminación se basa en la pregunta **Q2** donde se busca identificar los métodos de reducción de dimensión que han alcanzado los mejores resultados cuando son integrados al modelo de predicción o clasificación. Con el fin de resolver esta incógnita se estudian los resultados y las conclusiones de los artículos primarios para obtener los tres métodos de reducción de dimensión que presentan los mejores resultados. Del mismo modo que con los estudios relevantes en la Figura 14, se calcula la frecuencia de dichos métodos agrupados en enfoques de extracción y selección de características. Al analizar sus frecuencias tampoco se encuentran diferencias significativas que permitan establecer un criterio de selección que beneficie a cualquiera de los dos conjuntos. Para intentar generar una discriminación más efectiva se procede a evaluar de manera individual la aparición de los métodos de selección y extracción de características que finalmente son integrados al modelo por su efecto positivo en los resultados. Como se observa en la Figura 15 dentro de los estudios primarios tampoco existe una inclinación marcada en los algoritmos utilizados, los únicos que cuentan con una recurrencia mayor a uno son: Lasso (A3, A11, A13), XGBoost (A5, A12, A15), LightGBM (A6, A7).

Para dar respuesta a la pregunta **Q3** y teniendo en cuenta que las características secuenciales que se utilizan para entrenar el modelo superan el millón de puntos, se buscó en los estudios primarios el número de elementos que componen el conjunto de datos de entrenamiento. Los estudios primarios A2, A11, A15 y A18, como se observa en la Figura 18 implementan la etapa de preprocesamiento basada en reducción de dimensión en un conjunto de datos igual o superior al millón de elementos. El articulo A2 establece un estudio comparativo con tres técnicas de extracción de características, RP, PCA y SVD, donde RP presenta los mejores resultados seguido de PCA. En A11 se contrastan tres algoritmos de selección de características y un algoritmo de extracción de características. En este trabajo los tres mejores resultados se obtienen con Lasso, PCA y ReliefF. El trabajo presentado en A15 contiene el conjunto de entrenamiento más grande de todos los estudios primarios donde se realiza un análisis de importancia de características con el algoritmo XGBoost.

Finalmente, en el artículo A18 se realiza un estudio comparativo entre PCA y LDA, siendo la aplicación del primero el que mejor desempeño alcanza.

Durante el desarrollo de esta revisión sistemática únicamente los trabajos A15 y A16 responden la pregunta de investigación **Q4** puesto que abordan de manera específica la predicción de mapas de contacto. Como se mencionó anteriormente A15 utiliza XGBoost para ranquear las características, mientras que en A16 se realiza un estudio comparativo entre dos técnicas de extracción de características AE y PCA, donde AE obtiene los mejores resultados.

La respuesta de **Q5** se basa en la posibilidad de integración de los métodos que se encontraron en esta revisión, con el modelo de predicción de mapas de contacto implementado. Al considerar la arquitectura del predictor y la naturaleza de las características utilizadas para su entrenamiento, se concluye que la mejor opción para aplicar reducción de dimensión a las características secuenciales radica en la utilización de métodos no supervisados. En consecuencia, únicamente serán consideradas las técnicas de extracción de características puesto que, todos los algoritmos de selección de características detectados en la revisión son de naturaleza supervisada. Cabe resaltar que la aplicación de XGBoost en A15 es posible puesto que en este artículo se predicen parejas de aminoácidos individuales y no el mapa de contacto completo como se trabaja en el modelo de aprendizaje profundo implementado.

Las tres primeras técnicas de reducción de dimensión seleccionadas se encuentran en A2 puesto que denotan algoritmos no supervisados con la capacidad de procesar considerables volúmenes de información, además de estar entre las técnicas de reducción de dimensión con los mejores desempeños. En primer lugar, se tiene a PCA que representa por mucho la técnica de reducción de dimensión no supervisada más frecuente, siendo utilizada en A18 donde demostró mejorar el proceso de clasificación. En segundo lugar, se tiene el algoritmo RP que demuestra tener el mejor desempeño en A2. La tercera técnica es SVD que cuenta con un desempeño aceptable en A2 y es utilizado de manera recurrente en los artículos A4, A5, A13. El algoritmo AE es seleccionada debido a su aplicación especifica en la predicción de mapas de contacto dentro del articulo A16 y en el hecho de que dentro del mismo trabajo se haya

reportado una mejoría en los resultados. La cuarta técnica de reducción de dimensión seleccionada es LLE debido a su aparición recurrente dentro de los artículos primarios A9, A10, A13, donde destaca por generar espacios embebidos sobresalientes. Adicionalmente, en A9 el algoritmo LLE es integrado en un modelo de aprendizaje profundo, lo cual es beneficiosos teniendo en cuenta la arquitectura del modelo de predicción de mapas de contacto que se implementa. El ultimo método de reducción de dimensión considerado es KPCA que aparece en A19 como la técnica de reducción de dimensión que genera una mejoría en el desempeño y se encuentra de manera recurrente dentro de los trabajos A3, A5, A4 y A17, de métodos de reducción de dimensión probados.

Capítulo 4

4. Modelado

El pipeline de predicción implementado, junto con la integración de las técnicas de RD puede ser observado en la Figura 19. El trabajo definido en la primera etapa se basa en el proceso de obtención, selección y procesamiento de las cadenas de proteínas que conforman el conjunto de datos de entrenamiento, validación y prueba. En la segunda etapa se describe la implementación e integración de las técnicas de RD al modelo de predicción, con la definición de espacios embebidos generados a través de las características secuenciales. Posteriormente, en la tercera etapa se describe el funcionamiento del modelo implementado teniendo en cuenta los diferentes procedimientos implementados para llevar a cabo el entrenamiento y la predicción de mapas de contacto. En una cuarta etapa se describe el protocolo de evaluación llevado a cabo para evaluar el desempeño de las predicciones obtenidas. Finalmente, en la última etapa se describirán los pasos llevados a cabo para obtener los mejores espacios embebidos y modelos.

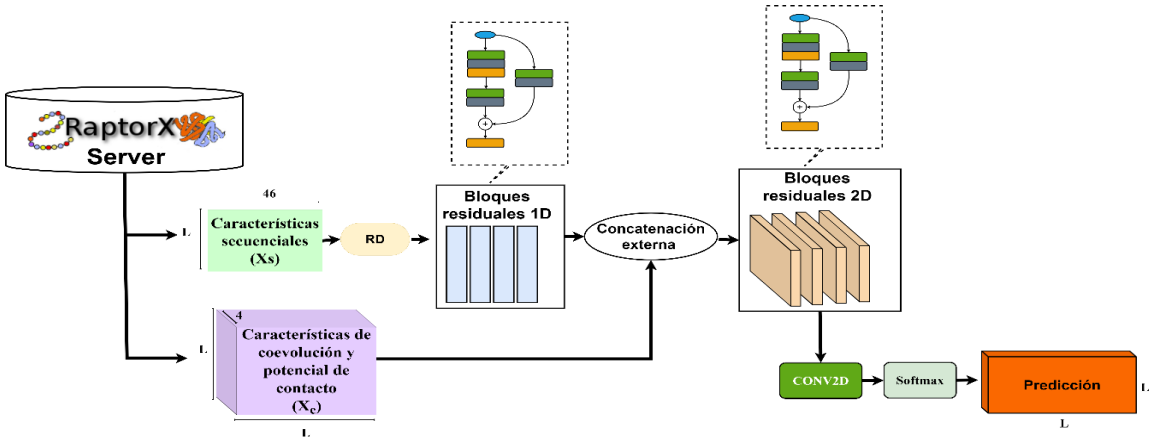


Figura 19. Pipeline de predicción de mapas de contacto implementado en esta investigación.

4.1. Conjunto de datos

El conjunto de datos utilizado para el entrenamiento del modelo de predicción de mapas de contacto fue descargado del *RaptorX Server*[3], donde es posible obtener tres archivos en formato *pickle*[4]. Dichos archivos contienen tres agrupaciones de cadenas de proteínas que se utilizan como conjuntos de datos de entrenamiento, validación y prueba. La información dentro de los archivos *pickle* está organizada en forma de lista de diccionarios donde sus claves y valores denotan los atributos que caracterizan las diferentes cadenas de proteínas seleccionadas. Una vez se tienen los archivos *pickle* descargados se procede a extraer y procesar la información contenida en ellos. Sin embargo, la implementación del pipeline de predicción establecido en *Raptor X Contact* [39], fue trabajado en su mayoría en la versión 2.7 de Python. Por lo tanto, la información presente en los archivos tiene que ser inicialmente adaptada para que pueda ser leída y utilizada en Python 3.8.

En la Tabla 9 se detalla la información que integra cada uno de los diccionarios almacenados en los diferentes archivos *pickle*. No obstante, es importante resaltar que únicamente algunos atributos son tenidos en cuenta para la construcción del conjunto de datos y el entrenamiento del modelo. La clave "*name*" contiene un identificador compuesto por caracteres alfanuméricos que especifican el código PDB de la proteína (primeros cuatro caracteres) [19], y la cadena a la que pertenece dicha secuencia (último carácter). Otro aspecto relevante para identificar un polipéptido es la secuencia de aminoácidos (orden y número de residuos), la cual se almacena como una cadena de caracteres dentro de la clave "*sequence*". Como se observa en la Figura 19, el modelo de predicción utiliza dos fuentes de información: la primera obtenida a partir de la secuencia de aminoácidos (naturaleza unidimensional), y la segunda de naturaleza bidimensional formada por la caracterización de parejas de residuos. El conjunto de atributos que agrupan las características secuenciales utilizadas es: los tres estados de estructura terciaria ("*SS3*"), accesibilidad al solvente ("*ACC*"), la matriz de puntuación de posición específica ("*PSSM*") y Matriz de puntuación de frecuencia específica ("*PSFM*"), para un total de 46 atributos. Los atributos bidimensionales se obtienen a partir de las características de coevolución y del potencial de contacto

[3] http://raptorx.uchicago.edu/download/ZGZlcGVuYUB1bmljYXVjYS5lZHUuY28=/

[4] https://docs.python.org/es/3/library/pickle.html

disponibles en las claves "*ccmpredZ*" y "*OtherPairs*", que forman un total de cuatro arreglos bidimensionales de tamaño $L \times L$. Para definir las características secuenciales y las características bidimensionales a utilizar se tuvo en cuenta los trabajos [39], [85], junto con la revisión de sus respectivos repositorios[5,6]. El mapa de contacto está almacenado en forma de matriz de tamaño $L \times L$, dentro de la llave "*contactMatrix*". Cada uno de los elementos de dicha matriz tiene únicamente tres valores posibles: cero para la ausencia de contacto, uno para la existencia de contacto y menos uno para interacciones desconocidas entre contactos [39].

4.1.1. Obtención de cadenas de proteínas

Como se explica en [39], [85], la lista de cadenas de proteínas descargadas para el entrenamiento del modelo está basado en el subconjunto de proteínas PDB25, el cual fue extraído del Protein Data Bank (PDB) [19]. Para evitar sesgos durante el entrenamiento y conseguir una mejor capacidad de generalización, cada una de las parejas de secuencias en el conjunto de datos comparten una identidad menor al 25 por ciento. Además, se establecen algunos criterios de exclusión a las cadenas de proteínas que serán utilizadas. En primer lugar, se fija un límite a la longitud de la secuencia (L), eliminando aquellas cadenas de proteínas con longitud menor a 26 y mayor a 700. El segundo filtro hace referencia a la exclusión de estructuras de proteínas cuya resolución sea menor a 2,5 Å. En tercer lugar, no se tienen en cuenta dominios formados por múltiples cadenas proteicas. Finalmente, se excluyen las proteínas con información incompleta o inconsistente en bases de datos como PDB, DSSP y ASTRAL. Una vez finaliza el proceso de selección se obtiene un conjunto de datos compuesto por: 6367 cadenas de proteínas de entrenamiento, 400 cadenas de proteínas de validación y 500 cadenas de proteínas de prueba.

[5] https://github.com/j3xugit/RaptorX-Contact

[6] https://github.com/lnyile/Protein-Contact-Map-Rse_Dense

Tabla 9. Claves y valores que integran el conjunto de datos de entrenamiento, validación y prueba.

Llave	Valor
"ACC"	Tres clases de accesibilidad al solvente (expuesta, intermedia o enterrada).
"PSFM"	Matriz de puntuación de frecuencia específica (PSFM).
"name"	Identificador del target que concatena el código PDB más la cadena a la que pertenece.
"DISO"	Región desordenada.
"sequence"	Estructura primaria o cadena de aminoácidos.
"ccmpredZ"	Matriz de precisión normalizada generada por CCMpred.
"contactMatrix"	Matriz de tamaño $L \times L$ que representan el mapa de contacto.
"PSSM"	Matriz de puntuación de posición específica.
"SS3"	Tres clases de estructura secundaria hélice-α (H), una lámina-β (E) y un giro (C).
"psicovZ"	Matriz de precisión normalizada generada con PSICOV.
"*OtherPairs*"	Matriz tridimensional que contiene dos matrices información mutua (normalizada y no normalizada) y una matriz de potencial de contacto.
"SS8"	Ocho tipos de estructura secundaria: tres clases de hélices (G para hélice, H para hélices-α, I para hélices-π), dos tipos de Lámina (E para lamina-β extendida y B para puente-β) y tres tipos de giro (T giro con enlace de hidrogeno, S para laso de alta curvatura y L para giro irregular).

4.1.2. Obtención de conjuntos de prueba 76CAMEO y MEMS400

Además del conjunto de datos de prueba obtenido de PDB25, se tienen dos conjuntos de cadenas independientes conocidos como 76CAMEO [104], y MEMS400 [105], los cuales están conformados por 76 y 400 cadenas de proteínas respectivamente. Si bien estas cadenas ya contaban con las características secuenciales, de coevolución y de potencial de contacto, no contenían el mapa de contacto real. Por este motivo fue necesario implementar un bloque de código adicional que permita la obtención y el

procesamiento automático de mapas de contacto reales a partir del código PDB de la proteína y del carácter que define la cadena (figuras 20-21).

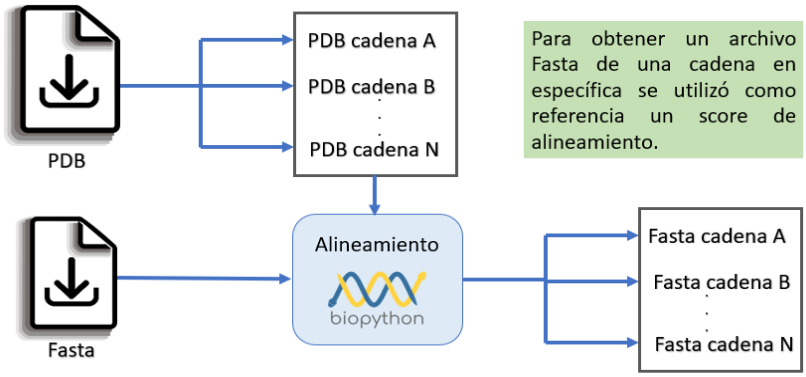


Figura 20. Obtención archivos PDB y Fasta de cadenas de proteínas individuales.

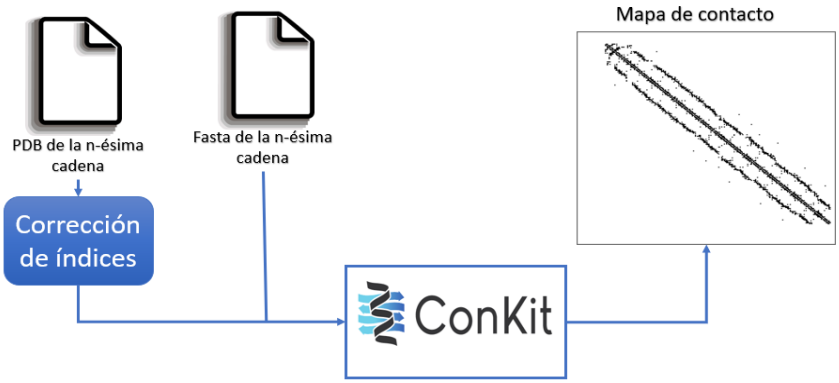


Figura 21. Procesamiento de archivos PDB y Fasta de cadenas individuales para la obtención de mapa de contacto.

La primera parte de la obtención de los mapas de contacto se ilustra en la Figura 20, donde básicamente se desarrolla la descarga de los archivos PDB y Fasta de la proteína a partir de los cuatro caracteres alfanuméricos que conforman el código PDB de la proteína. Posteriormente se procesan dichos archivos para obtener la estructura y la secuencia de una cadena de interés. Para asegurar que la secuencia de la cadena de interés sea correctamente extraída se toma como referencia un score de alineamiento entre la secuencia contenida en el PDB de la cadena y todas las cadenas

contenidas en el archivo Fasta. De este modo, se almacenará la secuencia con el más alto score, en un archivo fasta individual.

En la última parte del procesamiento que se muestra en la Figura 21, se describe el proceso que finalmente obtendrá el mapa de contacto real a partir de los archivos PDB y Fasta de una cadena de proteína en especifica. Una vez se tienen los archivos PDB individuales de cada cadena, se necesita hacer unos ajustes en los índices de los residuos que contienen. De modo que exista concordancia entre la posición real de cada aminoácido dentro de la cadena. Una vez se tienen todas estas correcciones se extrae el mapa de contacto en archivos de texto en formato RR con la aplicación de ConKit [106]. Finalmente, con los archivos de texto plano se genera su representación matricial que será utilizada para evaluar la predicción que se obtiene con el modelo implementado.

4.1.3. Análisis de longitud de las proteínas

El espectro de las longitudes presentes en las cadenas de proteínas dentro del conjunto de datos de entrenamiento (PDB25) puede ser observado en la Figura 22, donde es posible denotar que no existe una distribución uniforme de longitudes. Por esta razón se estableció un criterio adicional de discriminación seleccionando en los conjuntos de entrenamiento validación y prueba, cadenas de proteínas que cuenten con una longitud mayor o igual a 26 y menor o igual a 430 residuos ($26 \leq L \leq 430$). Lo anterior no solo permitirá realizar un mayor número de entrenamientos en menor tiempo, si no también existirá una reducción del consumo de memoria en la unidad de procesamiento gráfico (GPU) utilizada para el entrenamiento del modelo, evitando así futuros errores debidos a la falta de memoria. Con este nuevo filtrado se tienen un número de 5.813 cadenas de proteínas que reflejan un 91% del conjunto de entrenamiento original (6.367), obteniendo así una muestra significativa para establecer experimentos y resultados confiables, adicionalmente se tendrán 372 cadenas de proteínas de validación y 460 cadenas de proteína de prueba. Para los conjuntos independientes de prueba al aplicar la anterior restricción se tendrán 70 cadenas de proteínas para 76CAMEO y 371 proteínas para MEMS400.

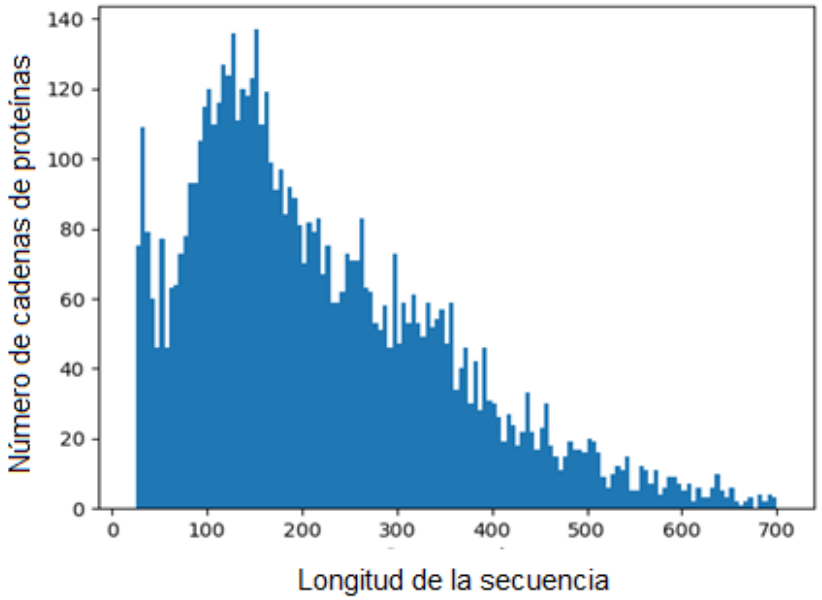


Figura 22. Distribución de la longitud de las cadenas de proteínas que conforman el conjunto de datos de entrenamiento.

4.1.4. Matriz de características secuenciales unificadas

Como se mencionó anteriormente, el estudio comparativo propuesto en este proyecto de investigación, será desarrollado con la integración de RD para encontrar espacios embebidos de menor dimensión que mejor representen las características secuenciales (entrada unidimensional del modelo). Para adaptar las técnicas de reducción de dimensión seleccionadas es necesario extraer una matriz unificada $U \in R^{N\times 46}$, donde N representa el número de residuos presentes en todas las cadenas de proteínas de un conjunto de datos en específico (entrenamiento, validación o prueba). El valor de N se calcula con la ecuación (8) en donde se suman las longitudes de las secuencias de aminoácidos (L_i).

$$N = \sum_{i=1}^{n} L_i \tag{8}$$

El procedimiento para obtener la matriz U, que establece la representación matricial de las características secuenciales, se describe en la Figura 23, y comprende las siguientes etapas: 1) Extracción de las características secuenciales de cada uno de

los targets, 2) representación vectorial de cada aminoácido con la concatenación de las características secuenciales (matriz de tamaño $\mathrm{L} \times 46$), 3) unificación de matrices de características secuenciales de todas las cadenas de proteínas que conforman el conjunto de datos. Para todo el conjunto de datos de entrenamiento compuesto por las 6.367 proteínas el valor de N es igual a 1.402.007 residuos.

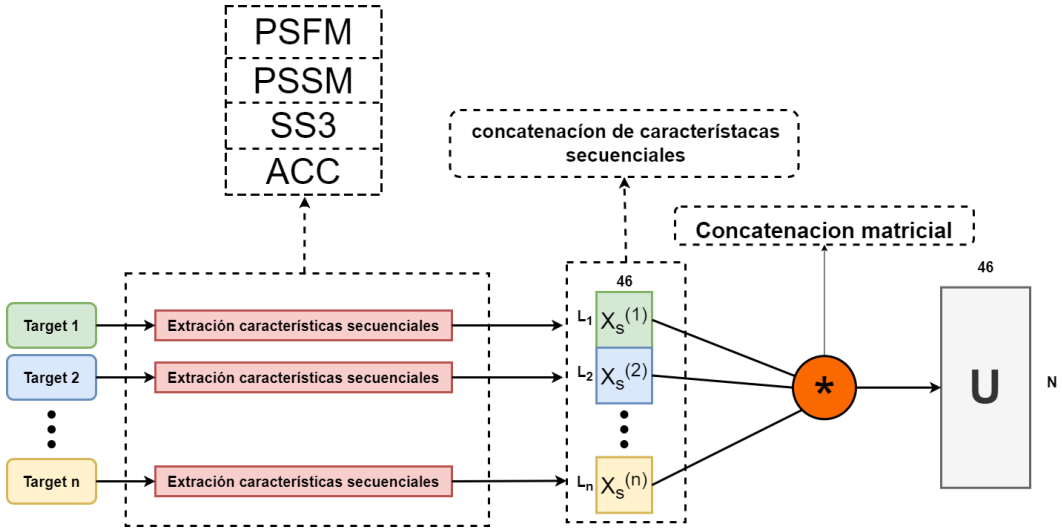


Figura 23. Esquema general del procesamiento de datos para la obtención de la matriz unificada U.

4.2. Implementación de técnicas de reducción de dimensión

Los métodos de RD como se muestra en la Figura 24, tienen como objetivo encontrar a partir de una matriz $U \in R^{N \times D}$, un espacio embebido $Y \in R^{N \times d}$, con $d < D$, que preserve la estructura o propiedades de U tanto como sea posible bajo un criterio establecido [107]. La importancia de obtener la matriz U radica en su papel para generar el mapeo o proyección que permitirá efectuar la reducción de dimensión a las diferentes características secuenciales. Es relevante resaltar que para aplicar reducción de dimensión la matriz unificada tiene que ser obtenida para el conjunto de datos de entrenamiento, validación y prueba. No obstante, el mapeo o proyección para nuevos puntos (conjuntos de datos de validación y prueba) de una técnica de RD en específica se define únicamente con la matriz unificada debida al conjunto de datos de entrenamiento.

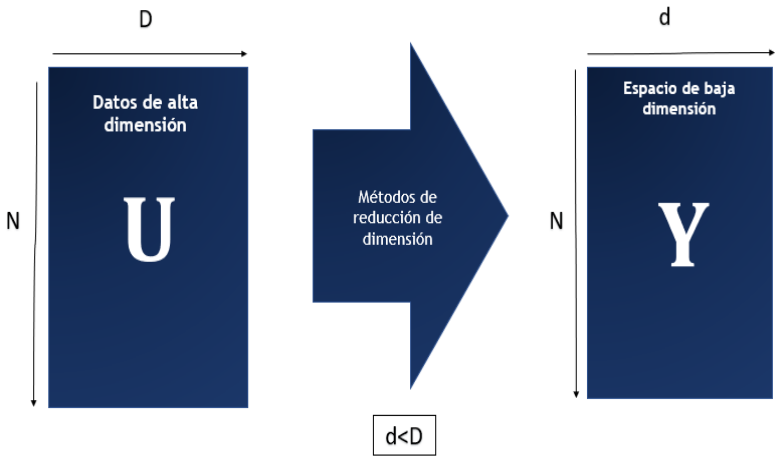


Figura 24. Esquema general de un método de reducción de dimensión aplicado a la matriz unificada U.

4.2.1. Implementación de técnicas de reducción de dimensión

Teniendo en cuenta la revisión sistemática desarrollada en el Capítulo 3, donde se seleccionaron las cinco técnicas de RD, se procedió a realizar su implementación para obtener el modelo de mapeo o proyección generado con los datos de entrenamiento. El parámetro N que define el número de instancias en la matriz unificada de entrenamiento (1.402.007 instancias), tiene gran relevancia por su relación directa con el costo computacional de las técnicas de RD seleccionadas, determinando su aplicación práctica en la predicción de mapas de contacto. La adopción de PCA, RP y SVD no presentó ningún problema durante su implementación, debido a que representan algoritmos lineales de RD [68], [108], que son capaces de manejar volúmenes de datos con un millón de puntos. Por otra parte, la complejidad de las técnicas no lineales de RD se ve ampliamente afectada por la magnitud de N, dado que un valor superior al millón de puntos puede generar errores por falta de memoria. Este escenario fue el caso de KPCA y LLE, que necesitan el cálculo de matrices de gran tamaño ($N \times N$ para el caso de KPCA) para realizar reducción de dimensión. Esto principalmente porque se necesita información de distancias entre parejas de puntos dentro de la matriz U [107]. De modo que, para el caso específico de esta investigación, no es práctico aplicar dichas técnicas de RD, dada la cantidad de memoria que se necesita y la ausencia de implementaciones escalables que aborden dicho problema. De este modo, dentro del grupo de técnicas de RD no lineales seleccionadas en la revisión sistemática, únicamente Autoencoders fue implementado.

4.2.2. Implementación de análisis de componentes principales (PCA)

Análisis de componentes principales o PCA por sus siglas en inglés es un enfoque clásico de reducción de dimensión ampliamente utilizado en el análisis exploratorio de los datos por su capacidad de reducir el número de observaciones, mediciones o variables correlacionadas [109]. Como se observa en el contraste que se presenta en la Figura 25 para PCA, la correlación es importante puesto que en un caso ideal donde se tenga una correlación perfecta (Figura 25a), es posible realizar una proyección con una mínima perdida de información (proyección ideal) [110]. Sin embargo, en conjuntos de datos reales los puntos de datos presentan una proyección no tan ideal como se muestra en la Figura 25b. PCA se agrupa dentro de los enfoques no supervisados y puede potencialmente mejorar el proceso de aprendizaje (desempeño y memoria) en modelos de predicción y clasificación [111]. Para aplicar PCA a un conjunto de datos de alta dimensión es necesario realizar las siguientes suposiciones: El mapeo realizado al momento de cambiar de base es de naturaleza lineal; La varianza más grande contiene la información más relevante; Los componentes principales del nuevo espacio de características son ortogonales [110].

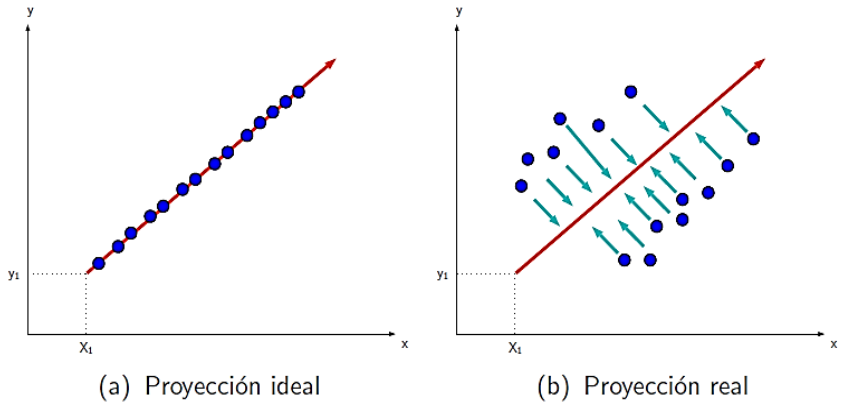


Figura 25. Contraste entre dos tipos de proyecciones: proyección ideal (a) y proyección real (b).

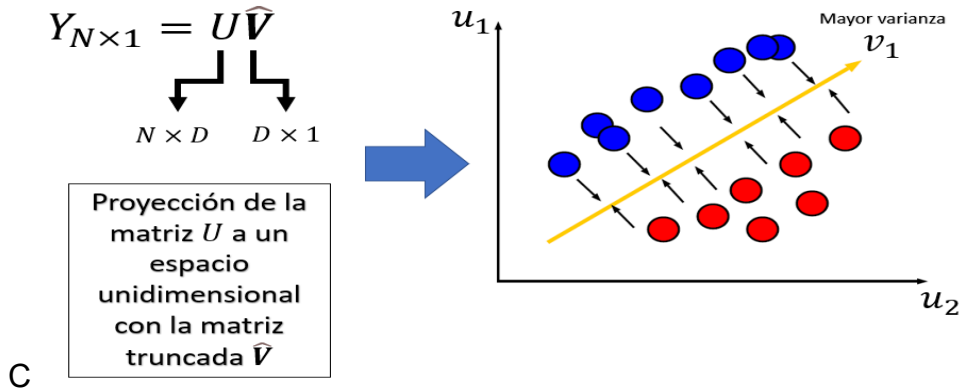
Dada una matriz de datos $U \in R^{N \times D}$ y una matriz de proyección $V \in R^{D \times D}$ que contiene los vectores propios de la matriz de covarianza $C \in R^{D \times D}$, entonces se puede encontrar un subespacio de Y con dimensión menor que el espacio original, haciendo

uso de una matriz truncada $\hat{V}$ (Figura 26). Dicha matriz truncada contiene los primeros d vectores propios asociados a los valores propios λ_i con mayor magnitud [109]. Cada uno de los vectores propios conservan un porcentaje de varianza de los datos originales, siendo v_1 el vector que más varianza contiene [110]. Esto permite lograr una representación aproximada que conserve la información más relevante del espacio original. La definición matemática de la proyección se define con la ecuación (9).

$$Y_{N\times d} = U_{N\times D}\hat{V}_{D\times d} \tag{9}$$

Adicionalmente, es importante definir la matriz de covarianza que puede ser calculada con la ecuación (10), siempre y cuando la matriz unificada $U_{N\times D}$ este centrada, es decir con media cero y desviación estándar igual a uno.

$$C_{D\times D} = \frac{1}{m} U^T_{D\times N} U_{N\times D} \tag{10}$$



C

Figura 26. Representación gráfica de una proyección lineal de un espacio bidimensional a un espacio unidimensional.

Con el fin de obtener datos menos redundantes y darles la misma varianza se procede a aplicar PCA *whitening* que es una variación del algoritmo original, y consiste

principalmente dividir cada uno de los componentes principales (y_i) por la raíz del valor propio asociado, como se menciona en [109], y se muestra en la ecuación (11).

$$y_{wi} = \frac{y_i}{\sqrt{\lambda_i}} \tag{11}$$

4.2.3. Implementación de proyecciones aleatorias (RP)

RP es una técnica de RD utilizada para reducir datos de alta dimensión, que se presenta como una alternativa a PCA. En RP la dimensión original del conjunto de datos D se "proyecta" a un subespacio de dimensión d, con $d \ll D$. Este procedimiento se lleva a cabo usando una matriz aleatoria $R \in R^{D \times d}$, con columnas formadas por vectores unitarios (la norma de cada columna es igual a uno) [68]. Al analizar la ecuación (12) que define RP, se puede observar que existe cierta semejanza con una proyección lineal, sin embargo, a diferencia de la matriz de proyección V utilizada en PCA, la matriz R es aproximadamente ortonormal por lo cual en el sentido estricto RP no puede considerarse como una proyección lineal [112].

$$Y_{N \times D} = U_{N \times D} R_{D \times d} \tag{12}$$

La idea clave de RP surge a partir del lema de Johnson-Lindenstrauss [113], que establece que, si los puntos dentro de un espacio vectorial son proyectados dentro de un subespacio seleccionado aleatoriamente de una dimensión adecuada, entonces las distancias entre los puntos tienen alta probabilidad de ser preservadas de manera aproximada (Figura 27).

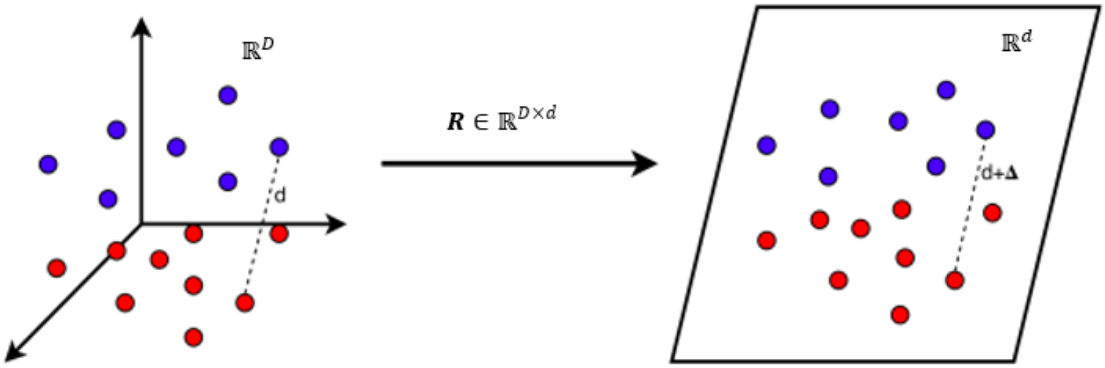


Figura 27. Reducción de dimensión basada en proyecciones aleatorias. Fuente: https://www.groundai.com/project/random-projections-of-mel-spectrograms-as-low-level-features-for-automatic-music-genre-classification/1

Teniendo en cuenta como se calcula la matriz R se pueden encontrar dos tipos de RP [114]. El primero conocido como RP Gaussiana, donde la matriz R se construye eligiendo elementos aleatoriamente de una distribución Gaussiana. El segundo definido como RP disperso (*sparse*), es el que se utiliza en este proyecto de investigación puesto que posee un costo computacional más bajo debido a su simplicidad [114]. Esto convierte a esta técnica de RD en una buena opción para obtener un espacio de baja dimensión puesto que el costo computacional para el cálculo de la "proyección" es menor con respecto a PCA y SVD, lo cual podría ser determinante al momento de trabajar con un considerable volumen de puntos. La matriz aleatoria, utilizada para reducir la dimensión se calcula con la siguiente ecuación:

$$\mathrm{R}_{ij} = \begin{cases} \sqrt{\frac{s}{d}} & con\ probabilidad \quad \frac{1}{2s} \\ 0 & con\ probabilidad \quad 1 - \frac{1}{s} \\ \sqrt{\frac{s}{d}} & con\ probabilidad \quad \frac{1}{2s} \end{cases} \qquad (13),$$

definiendo a d como la dimensión del subespacio y a s como un factor de densidad que generalmente es establecido como $\frac{1}{\sqrt{D}}$ [114].

4.2.4. Implementación de Descomposición de Valores Singulares (SVD)

Una técnica de RD estrechamente relacionada con PCA es la versión truncada de SVD [115]. Esta técnica se basa en el hecho de que una matriz puede ser factorizada a través de la multiplicación de tres matrices como se indica en la Figura 28.

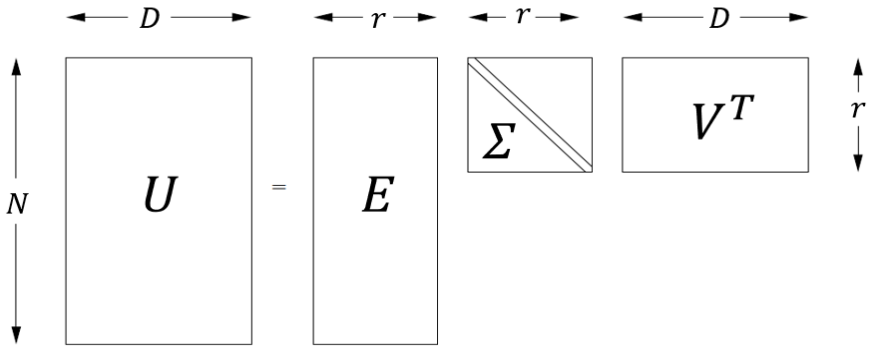


Figura 28. Versión truncada de SVD. Fuente: [115].

Las matrices $E \in R^{N \times r}$ y $V^T \in R^{D \times r}$ son ortogonales con $r < min\,(N, D)$, que contienen los vectores singulares izquierdos y derechos de la matriz U respectivamente [115]. El termino $\Sigma \in R^{r \times r}$ representa una matriz diagonal donde se encuentran los valores singulares de U. La reducción de dimensión se realiza a través de la definición de un valor $d < r$, con el fin de obtener una factorización truncada que reconstruye de manera aproximada a la matriz U, como se muestra en la ecuación (14).

$$U_{N \times D} \approx E_{N \times d} \Sigma_{d \times d} V^T_{d \times D} \quad (14)$$

Como se explica en [69], la reducción de dimensión llevada a cabo por esta técnica se basa en la ecuación (15), en donde se obtiene un espacio embebido al seleccionar los vectores singulares la matriz E asociados a los valores singulares Σ más grandes. En cuanto al costo computacional este procedimiento es similar al empleado para calcular las componentes principales [69].

$$Y_{N \times d} = E_{N \times d} \Sigma_{d \times d} \quad (15),$$

4.2.5. Implementación de AutoEncoders

A diferencia de las anteriores técnicas de RD, como se observa en la Figura 29, la aplicación de Autoencoders es capaz de generar espacios embebidos a través de mapeos no lineales que contengan la información más relevante de un conjunto de datos de entrada [116].

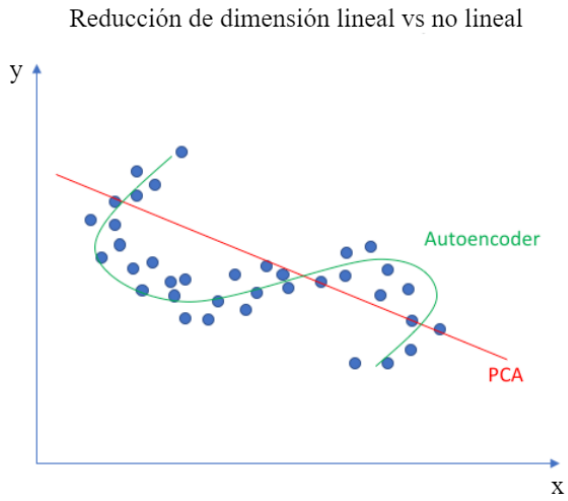


Figura 29. Comparación entre el mapeo lineal llevado a cabo por PCA y el mapeo no lineal establecido por Autoencoders. Fuente: [117].

Un autoencoder es una red neuronal que se compone de dos partes: un encoder y un decoder, cada una de las cuales tiene su propio conjunto de parámetros ha ser entrenados [116]. La función de un encoder es la de tomar un conjunto de datos de entrada (matriz de características, imágenes, audio, video o texto) y mapearlo dentro de un espacio latente codificado a través de la capa conocida como *bottleneck* [117]. Posteriormente, dentro del decoder se recoge la información del espacio latente para armar una copia aproximada del conjunto de datos de entrada. De esta manera, la arquitectura de AE busca básicamente la reconstrucción de la entrada minimizando una función de loss que generalmente es el error cuadrático medio (MSE) [116]. Para esta implementación se hace uso del MSE que se define con la siguiente ecuación:

$$Loss = \frac{1}{N}\sum(x - \hat{x})^2 \qquad (16),$$

donde en este caso x representa un vector de entrada y $\hat{x}$ representa la copia generada por la red neuronal [117]. Si bien un modelo que intente la reconstrucción de la entrada no parece de utilidad, en este tipo de arquitecturas se presta mayor atención en el espacio latente que se genera en el *bottleneck*. Esto con el fin de generar un nuevo espacio con características: menos correlacionadas, menos redundantes y por lo tanto con una dimensión menor a la del espacio original [116].

De acuerdo con su arquitectura este tipo de modelos se ha clasificado en los siguientes tipos: *undercomplete autoencoder*, *denoising autoencoder*, *sparse autoencoder y adversarial autoencoder* [116].

Teniendo en cuenta las características de la matriz unificada y el objetivo de reducir la dimensión se implementó la técnica de RD basada en *undercomplete autoencoder.* Esta implementación representa el enfoque más sencillo de RD capaz de obtener características útiles con el estableciendo un espacio latente definido por el número de neuronas dentro del *bottleneck* [116]*.* Al ser una red neuronal esta técnica es capaz de aprender una mejor generalización que PCA especialmente cuando existen relaciones no lineales entre las características de entrada [116].

La arquitectura del autoencoder utilizada como técnica de RD en esta investigación se muestra en la Figura 30. Como se menciona en [116], generalmente el número de neuronas dentro de las capas ocultas del encoder y el decoder son menores a la dimensión del espacio original. Por este motivo y para reducir el espacio de búsqueda de hiperparámetros se definieron dos capas ocultas para los bloques del encoder y del decoder con un número de neuronas menores al del espacio original. Adicionalmente, se estableció un número de neuronas simétricas, es decir el mismo número de neuronas para las capas uno y cinco, y para las capas dos y cuatro. Un parámetro importante es el número de neuronas para la capa conocida como bottleneck puesto que determina la dimensión del espacio embebido. Adicionalmente, a las capas ocultas mencionadas se añadieron capas de normalización del batch y un factor de regularización L2.

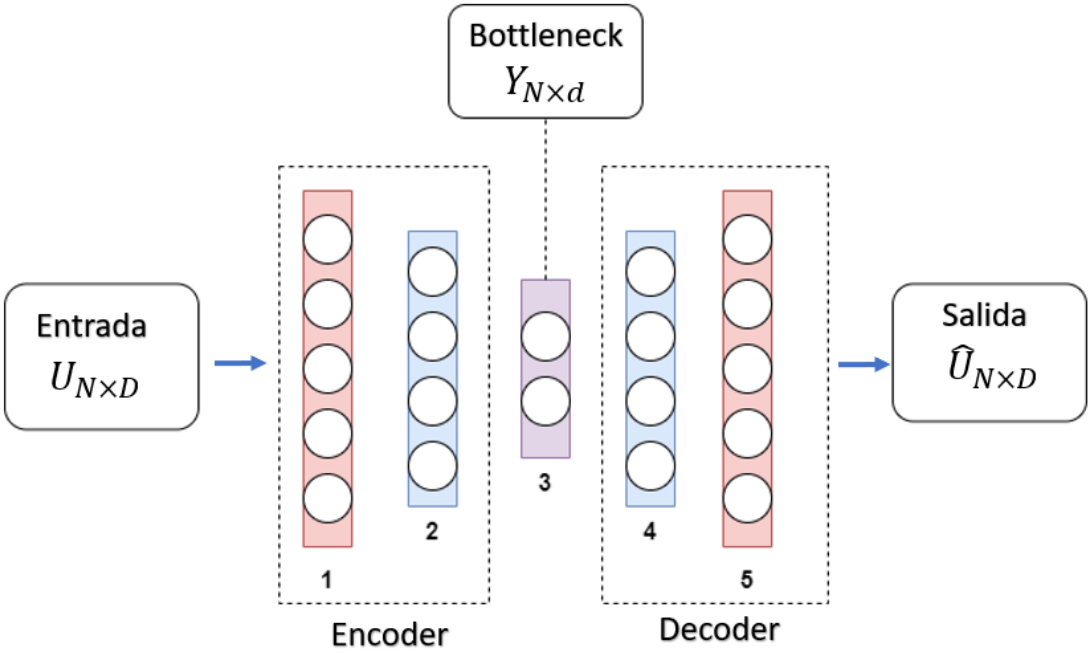


Figura 30. Arquitectura del Autoencoder utilizado en esta investigación, el cual cuenta con cinco capas ocultas en total.

4.3. Modelo de predicción de mapas de contacto

Como se ha mencionado anteriormente, el modelo de predicción implementado está basado en la arquitectura propuesta en los trabajos [39], [85], donde se hace uso de dos fuentes de información distintas para entrenar un modelo basado en redes neuronales residuales (ResNet). Un primer paso para la implementación de dicho modelo fue el análisis del código fuente de los trabajos mencionados anteriormente. Esto con el fin de encontrar una implementación que se adecúe a las necesidades del proyecto de investigación. Analizando los repositorios de ambos trabajos se pudo observar que en el trabajo en [39] esta implementado con la librería Theano[7], mientras que en [85] se implementa con Tensorflow[8], ninguno de los cuales utiliza alguna API que facilite su implementación y ejecución. Debido a las múltiples variaciones que se deben llevar a cabo en la entrada unidimensional, en este proyecto se implementó un modelo de predicción de mapas de contacto funcional mediante el API de Keras[9], con

[7] https://theano-pymc.readthedocs.io/en/latest/
[8] https://www.tensorflow.org/
[9] https://keras.io/

Tensorflow como backend. Esto significó un importante avance para el proyecto debido a que permite: una mayor facilidad de implementación de los bloques residuales, la integración dinámica de las técnicas de RD, y la construcción de un modelo integrado.

4.3.1. Arquitectura del modelo de predicción

La arquitectura del modelo de predicción de mapas de contacto utilizado se describe en la Figura 19 y se basa mayoritariamente como se ha mencionado anteriormente en los trabajos [39], [85]. Dentro de la arquitectura del modelo de predicción de mapas de contacto existen dos tipos de unidades residuales, agrupadas de acuerdo con la dimensión del conjunto de características de entrada que reciben. Sin embargo, cada una de las unidades residuales tanto para las características unidimensionales y bidimensionales están compuestos por las mismas capas.

El camino principal de la unidad residual (main path) está compuesto por: una capa convolucional, seguido de una capa de normalización (batch normalización) y de la función de activación (ReLu); las siguientes capas son una capa de convolución y de normalización; para finalizar el camino principal se encuentra una capa de sumatoria que se encarga de unir el atajo (shortcut path) con la entrada, para posteriormente finalizar con una función de activación (ReLu). El atajo al igual que los anteriores bloques está compuesto por: una capa convolucional (con un kernel de 1×1) y una capa de normalización, la principal función de estas capas es ajustar la dimensión de la entrada de la unidad residual para poder posteriormente tener la posibilidad de sumarla a la salida.

4.3.2. Formato de las características de entrada

Para alimentar el modelo de predicción se utilizan dos tipos de características: unidimensionales (características secuenciales), y características bidimensionales (información de coevolución y potencial de contacto). El proceso de adaptación de la anterior información como entrada al modelo de predicción, se basa en la extracción y concatenación de las diferentes características disponibles en los diccionarios. Como se mencionó anteriormente en total son 46 los atributos secuenciales concatenados por cada residuo, lo cual termina en una matriz $X_s^i \in R^{L\times 46}$ para el i-ésimo target. Los

cuatro tipos de características bidimensionales de tamaño $L \times L$, para el target i se concatenan en una matriz tridimensional $X_c^i \in R^{L \times L \times 4}$.

En las características secuenciales es donde se van a integrar las técnicas de RD a través de la matriz unificada U**,** la cual contiene la caracterización de todos los residuos que componen todas las cadenas de un conjunto de datos en específico (entrenamiento, validación y prueba). No obstante, tanto la matriz U de características original, como los espacios embebidos generados Y, no pueden ser ingresados al modelo directamente, debido a que el formato de entrada solo recibe información de una única cadena de proteína. En consecuencia, tanto las características secuenciales originales, como los espacios embebidos calculados tienen que ser reorganizados nuevamente en matrices separadas para cada una de las cadenas. Del mismo modo el mapa de contacto real (label) fue reorganizado en una matriz tridimensional de tamaño $L \times L \times 3$, usando el formato *one hot encoding,* donde se representan los tres tipos de interacciones: no contacto, contacto y desconocido respectivamente.

Para una mayor facilidad en el entrenamiento y la evaluación se reorganizó la información de cada uno de los targets en listas de diccionarios similares a los que se mostraron en la Tabla 1. Sin embargo, las claves de los diccionarios solo almacenan los siguientes datos: El identificador del target, que es el código PDB más la cadena a la que pertenece; una cadena de caracteres con la secuencia de aminoácidos; la matriz de características secuenciales X_s^i o el espacio Y^i generado para cada proteína; y la matriz de características bidimensionales X_c^i.

4.3.3. Entrenamiento del modelo de predicción

Una vez se procesan las características de entrada se procede a entrenar el modelo de predicción. A diferencia de los modelos clásicos de predicción, el entrenamiento requiere etapas adicionales de procesamiento que deben abordar los siguientes aspectos: Un módulo intermedio que permita concatenar las características unidimensionales y bidimensionales; Entrenar el modelo de predicción implementado con cadenas de longitud variable; Implementar un protocolo de evaluación para la predicción de mapas de contacto; Correr el modelo de predicción de mapas de contacto haciendo uso de máquinas virtuales de Google Cloud.

4.3.3.1. Concatenación externa

En el pipeline de la Figura 19 se mostró que existe una etapa intermedia de procesamiento que tiene la función de concatenar las salidas de los bloques residuales 1D, con las características bidimensionales. Esto para posteriormente ingresar a los bloques residuales 2D y obtener el mapa de contacto predicho. En los trabajos [39], [85] se utiliza una operación similar al producto externo entre dos vectores, para la transformación de las características secuenciales, en características basadas en parejas de aminoácidos.

Sea $A = \{a_1, a_2, a_3, .., a_L\}$, con $a_i \in R^n$ y L como la longitud de la cadena de aminoácidos, la salida del primer módulo compuesto por los bloques residuales 1D para una cadena de proteínas. Cada uno de los vectores a_i es una representación que la red neuronal convolucional ha creado para caracterizar los residuos que componen una cadena (n características). Al considerar todas las posibles parejas de aminoácidos y concatenar los vectores de características es posible armar una matriz de tamaño $L \times L \times 3n$, cuyos elementos componen la concatenación de a_i, $a_{(i+j)/2}$, a_j, que denotan la pareja de residuos i, j más la información del residuo que queda en medio de ellos $(i + j)/2$. Finalmente, a dicha matriz se le concatenan las características de coevolución X_c^i sumando cuatro dimensiones de profundidad ($L \times L \times 3n + 4$). La implementación del anterior procedimiento de concatenación externa en el API de Keras, fue posible gracias a la integración de funciones lambda dentro de las capas del modelo

4.3.3.2. Entrenamiento de proteínas de diferentes longitudes

Para adaptar la implementación propuesta en [39] y ser utilizadas con el API de Keras, se tuvo que solucionar el problema relacionado con la longitud variable de las cadenas de proteínas. Esto representó un gran problema puesto, que con Keras se tiende a utilizar conjuntos de datos con tamaños de muestras y características constantes, lo cual difiere con la naturaleza de las secuencias de proteínas que en la mayoría de los casos son de longitudes variables. Para abordar esta problemática se utiliza el enfoque

del Descenso del Gradiente Estocástico (SGD) [118], donde los parámetros que conforman las diferentes capas se actualizan con cada cadena de proteína de entrada [39], [85]. El uso de las funcionalidades: "*train_on_batch*", "*test_on_batch*", y "*predict_on_batch*" permitieron no solo aplicar SGD, si no también solucionar el problema de longitudes variables, al poder entrenar el modelo con un tamaño del batch (longitud de la secuencia) diferente. En este sentido, el proceso de entrenamiento del modelo que se realiza en cada batch, está relacionado con la información (características unidimensionales y bidimensionales) de una única cadena proteica.

Si bien el anterior proceso descrito permitió el entrenamiento del modelo, existieron algunos problemas de memoria debidas a la variabilidad de los *batches* de entrenamiento. Para solucionar este problema dos estrategias fueron consideradas y utilizadas. El primer enfoque de solución consistió en llamar a las funciones *garbage collector* para en cierta medida limpiar la memoria cada vez que termina una época. El segundo enfoque consistió en reducir el número de tensores que se debe crear y almacenar en memoria, a través de la creación de una etapa *padding* (Figura 31) que restringe el número de valores que puede tomar L, los cuales fueron definidos a múltiplos de diez. Esto también ayudo a agilizar el proceso de entrenamiento de cada uno de los batches.

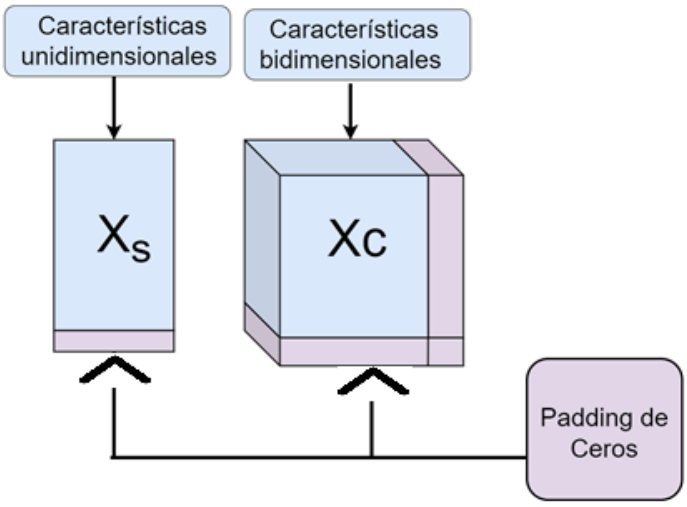


Figura 31. *Padding* llevado a cabo en el conjunto de datos de entrenamiento.

4.3.3.3. Función de Loss

El entrenamiento del modelo se lleva a cabo mediante el método *maximum-likelihood* que básicamente maximiza la probabilidad de ocurrencia de los contactos nativos (1), las parejas que no están en contacto (0) y la clase extra que representa interacciones desconocidas (-1). Dado que existen más de dos clases, la función de Loss seleccionada fue *categorical crossentropy* promediada para las parejas de residuos que conforman el conjunto de datos de entrenamiento. La función de Loss utilizada se define con la siguiente ecuación:

$$Loss = -\sum_{i=1}^{m} y_i log \hat{y}_i \qquad (17),$$

donde y_i denota la clase real, $\hat{y}_i$ la predicción del modelo y m el número de clases que en este caso será igual a tres.

4.3.3.4. Entrenamiento

Debido al considerable volumen de información que se maneja, al tiempo que tarda cada modelo en ser entrenado, y al número de entrenamientos que se deben realizar, se implementó el procedimiento de la Figura 32**,** en donde se utilizan servicios en la nube para llevar a cabo las diferentes pruebas. En primer lugar, para facilitar el control de versiones del modelo, se decidió recurrir a Github de manera que se puedan actualizar los cambios en el modelo en la máquina virtual de Google Cloud, desde el computador local. En segundo lugar, teniendo en cuenta el considerable volumen de la información dentro de los conjuntos de datos de entrenamiento, validación, prueba y de la matriz unificada de entrenamiento se almacenaron estos archivos en Google Drive para ser posteriormente cargados directamente desde la máquina virtual. Finalmente, se hace uso de una de las GPUs disponibles en esta plataforma para agilizar el tiempo en que el modelo finaliza su entrenamiento.

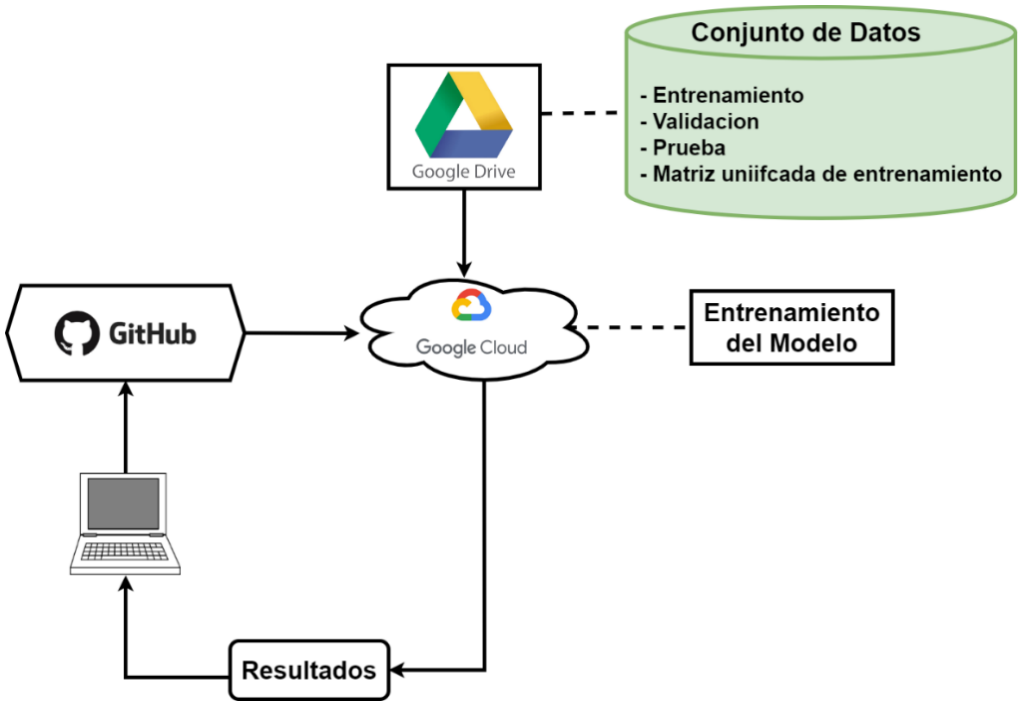


Figura 32. Esquema general del proceso de entrenamiento del modelo de predicción de mapas de contacto.

4.3.3.5. Protocolos de evaluación

Con el fin de encontrar diferencias significativas entre el modelo base (sin técnicas de reducción de dimensión) y los modelos que integran el módulo de RD como etapa de preprocesamiento, se implementó el protocolo de evaluación que se muestra en la Figura 33.

Las métricas de evaluación fueron codificadas con ayuda de la librería Numpy[10] de Python, teniendo en cuenta la representación matricial de la predicción (matriz de tamaño $L \times L$) como parámetro de entrada, y siguiendo los lineamientos que se especifican en el CASP[11]. Una vez implementadas las métricas de evaluación se procede a contrastar los valores obtenidos con los resultados disponibles en el CASP 13 para las predicciones del target T0950-D1[12]. Esto con el fin de encontrar errores

[10] https://numpy.org/
[11] https://predictioncenter.org/casp14/doc/rr_help.html
[12] https://predictioncenter.org/casp13/rrc_results.cgi

que puedan generar mediciones erróneas. Adicionalmente, se realizó una prueba en donde se evaluaron dos mapas de contacto idénticos, de modo que se observó, como era de esperarse, resultados con los valores más altos para cada una de las métricas. Con las medidas de desempeño de lista reducida y lista completa para cada una de las cadenas disponibles dentro del conjunto de datos de prueba, se procede a guardar los valores en archivos CSV por medio de la librería Pandas[13]. Finalmente, con los archivos CSV se hace uso de la herramienta KEEL [119], para llevar a cabo test de Friedman [120], y con esta información realizar el análisis de Nemenyi [121], como post hoc test para determinar los modelos con mejor desempeño.

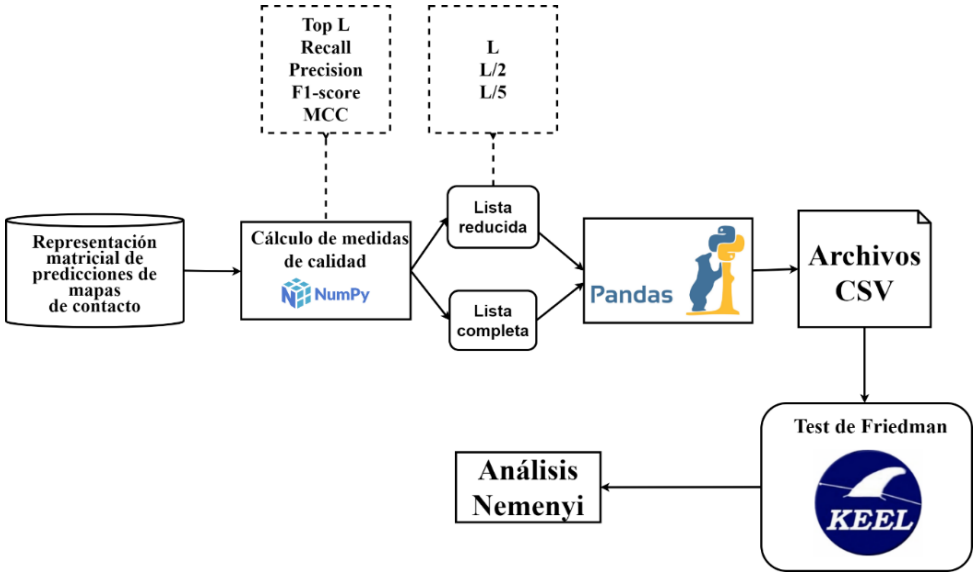


Figura 33. Protocolo de evaluación para predicciones de mapas de contacto.

4.3.4. Métricas de evaluación

Principalmente, fueron 4 las métricas de desempeño utilizadas para evaluar las predicciones sobre las cadenas de proteínas que conforman el conjunto de entrenamiento. La mayoría de ellas se utilizan tanto en la lista completa, como en la lista reducida de predicción. Dentro de cada tipo de lista los resultados son analizados de acuerdo con la separación de la pareja de residuos en contacto dentro de la

[13] https://pandas.pydata.org/

secuencia de la proteína, agrupando las predicciones de contactos en varias categorías. Para la lista reducida la evaluación se realiza en las siguientes categorías de predicciones: contactos de rango extra-largo; contactos de rango largo; contactos de rango largo y medio; y contactos de rango largo, medio y corto. Para la lista completa las predicciones se agrupan en: contactos de rango extra-largo; contactos de rango largo, contactos de rango largo y medio, y finalmente contactos de rango largo, medio y corto.

4.3.4.1. Sensibilidad

La sensibilidad es una métrica de evaluación que establece para este caso el número de contactos nativos verdaderos que fueron detectados por el modelo. El valor de este parámetro según el CASP[14] se calcula con la siguiente ecuación:

$$Sensibilidad = \frac{TP}{N_p} \quad (18),$$

donde TP denota la tasa de verdaderos positivos, es decir predicciones de contactos del modelo que realmente son contactos y N_p es el número total de contactos presentes en la estructura teniendo en cuenta un determinado rango (largo, medio, corto, etc).

4.3.4.2. Precisión y Top L

La precisión mide la cantidad de predicciones que el modelo definió como contactos y que efectivamente son contactos nativos de la estructura. El valor de esta métrica se calcula con la siguiente ecuación:

$$precisión = \frac{TP}{TP + FP} \quad (19),$$

[14] https://predictioncenter.org/casp14/doc/rr_help.html

con TP (Verdaderos positivos) como la tasa de predicciones de contactos que realmente eran contactos y FP (Falsos negativos) el número de predicciones erróneas de contactos. El cálculo de la Top L se obtiene con la expresión (19) sin embargo para la evaluación se toman los L/k contactos más probables de ser contactos según el modelo.

4.3.4.3. F1-score

Esta métrica se define como el promedio armónico entre dos factores: precisión y sensibilidad [122]. Para obtener el valor de esta medida de desempeño se hace uso de la ecuación (20).

$$F1\ score = \frac{2 \times precisión \times Sensibilidad}{precisión + Sensibilidad} \quad (20),$$

4.3.4.4. Coeficiente de Correlación de Matthews (MCC)

El coeficiente de correlación de Matthews (MCC), es una herramienta para la evaluación de modelos que mide las diferencias entre los valores reales y los valores predichos [122]. Esta métrica de evaluación se calcula únicamente para la lista completa de contactos.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (21),$$

4.3.5. Test de Friedman

El test de Friedman es una prueba no paramétrica que se utiliza para encontrar diferencias significativas, estableciendo una comparación de tres o más mediciones dentro un mismo grupo [120]. Para el caso específico de esta investigación se utilizará dicha prueba para encontrar variaciones significativas al momento de aplicar métodos de RD a las características secuenciales de entrada. El primer paso para llevar a cabo

el test de Friedman es realizar un ranqueo de mayor a menor de cada una de las columnas (bloques) que componen la matriz formada por n mediciones y p grupos, como se observa en la Figura 34. Las mediciones (x_{ij}) representan las diferentes métricas de calidad definidas en subsecciones anteriores, teniendo en cuenta las diferentes listas (completa y reducida) y los tipos de contactos (corto, medio, largo), esto definido para cada una de las estructuras definidas en el conjunto de datos de prueba. Una vez se obtiene el ranqueo (r_{ij}) por separado de cada una de las métricas, se procede a calcular las diferencias de los promedios de los rankings para cada grupo R_j [120]. De este modo se puede observar la diferencia en los rankings de tal forma que se pueda determinar un p-value y a partir de este refutar o no la hipótesis nula, que para este caso en especial sería "los métodos de RD aplicados a las características secuenciales no tienen efecto en los resultados del predictor".

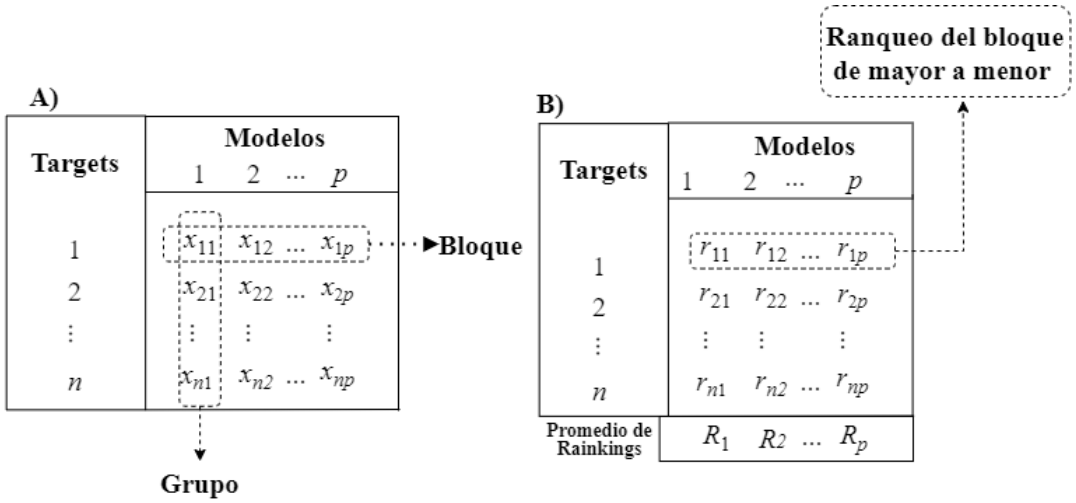


Figura 34. Cálculo de los rainkings en el test de Friedman. Fuente: [123].

Para obtener el p-value es necesario conocer dos valores, el chi-cuadrado de los rankings (X_r^2) y los grados de libertad (df). Según [120], los grados de libertad se obtienen con el número de modelos (grupos) menos uno ($p-1$), para el chi-cuadrado se utiliza la siguiente ecuación:

$$X_r^2 = \frac{12}{np(p+1)}\sum_{j=1}^{p}(R_j)^2 - 3n(p+1) \qquad (22),$$

Donde n representa el número de targets, p el número de modelos y R la suma de los rankings para cada uno de los modelos, es decir:

$$R_j = \frac{1}{n}\sum_{i=1}^{n} r_i \quad (23),$$

4.3.6. Análisis de Nemenyi

Una vez se aplica el test de Friedman y se obtiene la refutación de la hipótesis nula, se procede a realizar un estudio post hoc, el cual será desarrollado haciendo uso del Análisis de Nemenyi [121]. Dicho análisis, se basa en el cálculo de una distancia critica (DC) entre los promedios de los rankings de cada modelo, para obtener un criterio que ayude a definir la implementación que presenta los mejores resultados. El cálculo de DC se realiza mediante la siguiente ecuación:

$$CD = q_\alpha \sqrt{\frac{p(p+1)}{6n}} \quad (24),$$

donde p es el número de modelos y n es el número de experimentos (targets). Los valores de q_α se obtienen teniendo en cuenta el nivel de confianza α y el número de modelos que se consideran. En [121], se presenta la Tabla 10 en donde brindan dichos valores hasta para un número de diez modelos.

Tabla 10. Valores de q_α para la prueba de Nemenyi en función del valor de confianza y el número de modelos. Fuente [121].

	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
$\boldsymbol{q_{0.05}}$	1.960	2.343	2.569	2.728	2.850	2.949	1.960	2.343	3.164
$\boldsymbol{q_{0.10}}$	1.645	2.052	2.291	2.459	2.589	2.693	2.780	2.855	2.920

4.4. Modelado

Con las fases de procesamiento de datos finalizada se procede a llevar a cabo la etapa de modelado, la cual fue realizada en su totalidad con el conjunto de datos de entrenamiento, validación y prueba obtenidos del PDB25. En esta etapa buscó

principalmente definir dos aspectos: el primero es el establecimiento de la dimensión del espacio embebido para cada una de las técnicas de RD seleccionadas; y el segundo que consiste en seleccionar un conjunto de hiperparámetros para el modelo base sin RD y para los modelos entrenados con los espacios de baja dimensión que hayan presentado el mejor desempeño en la primera parte.

4.4.1. Establecimiento de la dimensión para los espacios embebidos

Para llevar a cabo esta tarea se definieron diferentes criterios para cada una de las técnicas de RD, de manera que se puedan encontrar las características que mejor desempeño tengan, tomando como referencia un modelo base y las medidas de calidad aplicadas al conjunto de datos de prueba. Los hiperparámetros de la Tabla 11, son los que se definen dentro del modelo base, el cual será utilizado como referencia para establecer las dimensiones de los espacios embebidos generados a partir de las técnicas de RD seleccionadas. Estos hiperparámetros fueron seleccionados teniendo en cuenta los recursos computacionales disponibles y las arquitecturas presentadas en los trabajos [39], [85], [124], [125], los cuales hacen uso de bloques residuales para la predicción de mapas de contacto.

Tabla 11. Definición de hiperparámetros para el modelo base.

Hiperparámetros en el bloque residual unidimensional	
Número de unidades residuales	2
Número de filtros en cada capa convolucional	20
Tamaño de los filtros en cada capa convolucional	17
Función de activación	ReLu
Hiperparámetros en el bloque residual bidimensional	
Número de unidades residuales	8
Número de filtros en cada capa convolucional	64
Tamaño de los filtros en cada capa convolucional	(3,3)
Función de activación	ReLu
Hiperparámetros generales	
Número de filtros en la capa convolucional de salida	3
Tamaño de los filtros en la capa convolucional de salida	(3,3)
Función de activación de salida	SoftMax
Función de loss	*Categorical Crossentropy*

Optimizador	Adam
Tasa de aprendizaje	0.01
Épocas	50

4.4.1.1. Establecimiento de dimensión para PCA y SVD

Dada la estrecha relación existente entre PCA y SVD se aplicaron en ambos dos enfoques para determinar el número de componentes que van a ser utilizados. El primer enfoque llevado a cabo se relaciona con el análisis del aporte de cada uno de los componentes principales para PCA y componentes singulares para SVD. De este modo, se podrá obtener un acercamiento inicial de cuál es la cantidad de dimensiones que se deben considerar para cada técnica. El segundo procedimiento fue el de iterar las dimensiones de los espacios embebidos para integrarlos directamente en la etapa de entrenamiento del modelo, haciendo uso únicamente de una muestra significativa de los conjuntos de datos en entrenamiento (600 cadenas), validación (60 cadenas) y prueba (60 cadenas), para posteriormente ranquear con test de Friedman y encontrar los modelos que presenten un mejor F1-score para los contactos de rango medio y largo de la lista completa. Es importante resaltar que la prueba anteriormente descrita se repitió tres veces con el fin de observar patrones que puedan ser analizados junto con el aporte de varianza de cada componente (dimensión). Finalmente se hace un contraste de los resultados obtenidos para realizar pruebas con el conjunto de entrenamiento completo.

Al examinar la Figura 36 se puede observar que, con un número de componentes principales de 24 y 33, todos los experimentos aparecen de manera más frecuente entre los cinco mejores rankeados según el test de Friedman. Al contrastar la anterior información con la Figura 35 se identifica una congruencia con los resultados de la Figura 36, puesto que después del componente principal número 20 la cantidad de información (varianza), que poseen los demás componentes se reduce considerablemente. Por este motivo, se establecieron las dimensiones 24, 30 y 33 como candidatas para ser probadas e integradas en pruebas con los conjuntos de datos completos.

De manera similar que el caso anterior, en SVD se puede observar en la Figura 37 que el aporte de cada uno de los componentes singulares va decreciendo especialmente en componentes singulares mayores a 20 en donde se destaca los pequeños aportes al momento de reconstruir la matriz original de entrenamiento. Al comparar los niveles de varianza representativa con los resultados de los experimentos que se presentan en la Figura 38 se puede ver que existe una tendencia marcada en un número de componentes singulares iguales a 21, 27 y 33, siendo estos valores los candidatos seleccionados para ser usadas en posteriores etapas de modelado.

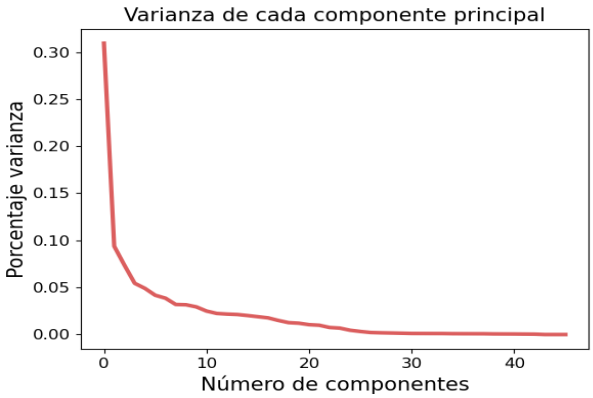


Figura 35. En este grafico se evalúa el aporte (varianza) de cada uno de los 46 componentes principales.

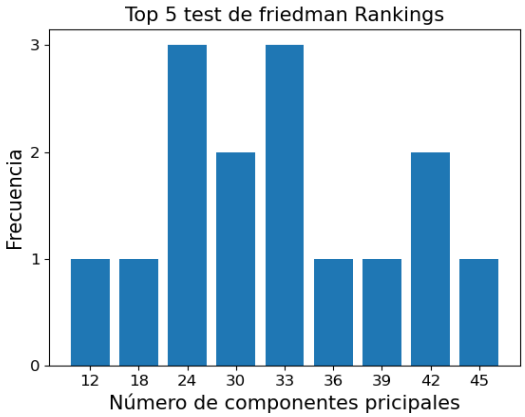


Figura 36. Frecuencia con que un número de componentes principales en específico apareció dentro de los cinco modelos mejor rankeados según el test de Friedman y la métrica F1-score, para los contactos de rango largo y medio de la lista completa.

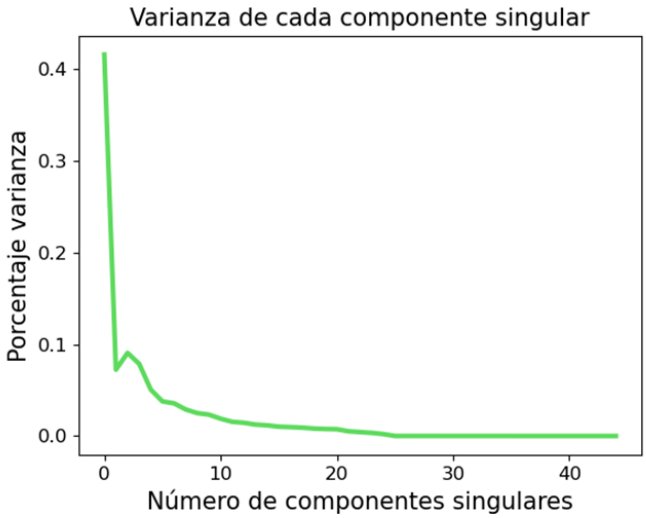


Figura 37. En este grafico se evalúa el aporte (varianza) de cada uno de los 45 componentes singulares.

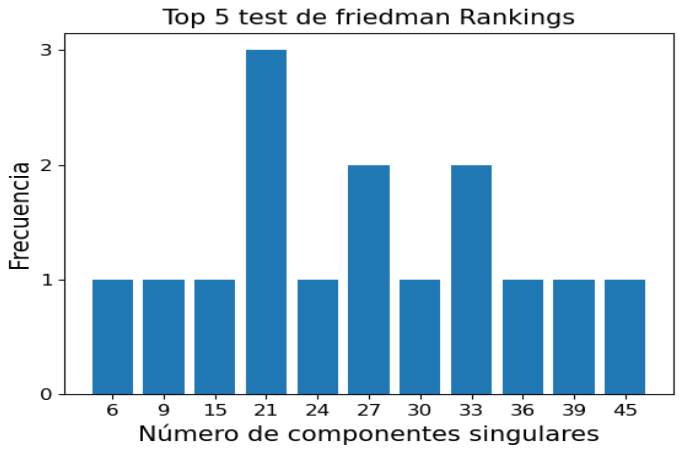


Figura 38. Frecuencia con que un número de componentes singulares en específico apareció dentro de los cinco modelos mejor rankeados según el test de Friedman y la métrica F1-score, para los contactos de rango largo y medio de la lista completa.

4.4.1.2. Establecimiento de dimensión para RP

Esta técnica de reducción de dimensión representa un caso particular, puesto que debido a su implementación no existe una forma directa de representar el contenido de información que cada una de las dimensiones posee. Por este motivo, para encontrar las dimensiones candidatas se tomaron como referencia únicamente los

resultados obtenidos en la predicción de contactos de rango medio y largo, en la misma muestra significativa establecida para SVD y PCA. Al igual que en el análisis anterior se hace uso del F1-score como métrica de referencia para comparar el potencial aporte que tendrán cada una de las dimensiones dentro del espacio de componentes aleatorias. En la Figura 39, se puede observar que existe una recurrencia en las dimensiones 15, 27, 30, 39 y 45, las cuales se encuentran en un rango de dimensión similar a los observados en PCA y SVD. Puesto que las dimensiones 39 y 45 no representan una reducción de dimensión significativa con respecto al espacio de datos original, se definieron las dimensiones 15, 27 y 30 como candidatas para ser probadas en las fases posteriores de modelado.

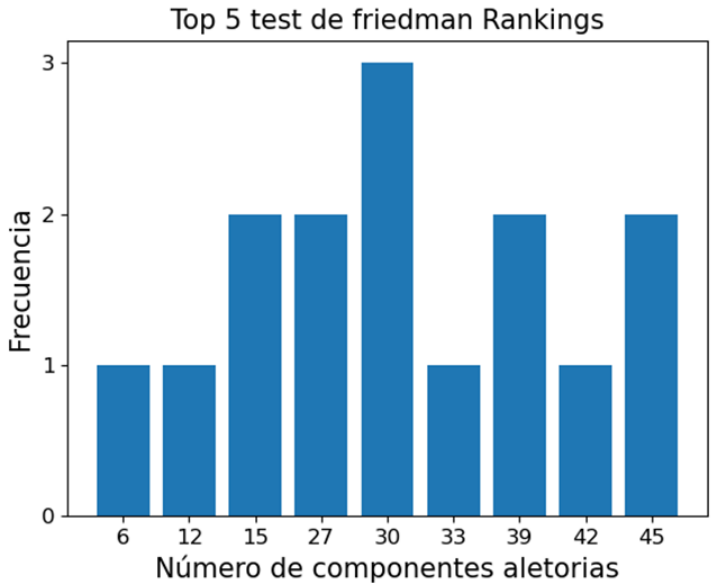


Figura 39. Frecuencia con que un número de componentes aleatorias en específico apareció dentro de los cinco modelos mejor rankeados según el test de Friedman y la métrica F1-score, para los contactos de rango largo y medio de la lista completa.

4.4.1.3. Establecimiento de dimensión para AE

A diferencia de las anteriores técnicas de RD, en donde los espacios embebidos de varias dimensiones podían ser generados dinámicamente, AE necesita ser entrenado cada vez que se necesita generar un espacio embebido de una dimensión determinada. Lo anterior, debido a que cada vez que se quiere cambiar la dimensión

se necesita cambiar el número de neuronas dentro de la capa conocida como "*bottleneck*", modificando la arquitectura del modelo (Figura 30) y por ende realizar un nuevo entrenamiento. En consecuencia, al momento de integrar esta técnica de RD con el modelo de predicción, no sería práctico realizar el entrenamiento de dos modelos y esto sin considerar los otros hiperparámetros definidos que conforman el AE. Para abordar el anterior problema se utilizó la matriz unificada U de todo el conjunto de datos de entrenamiento y validación, para de este modo evaluar el espacio latente generado, a partir de su capacidad para reconstruir el espacio original del conjunto de datos de validación. De modo que, para la definición de la dimensión del espacio embebido que finalmente se integra al modelo de predicción se realizará mediante un proceso de búsqueda de hiperparámetros que brinden el mejor resultado en cuanto a la exactitud (*accuracy*) con que se reconstruye el conjunto de datos de validación.

En la Tabla 12 se establece el espacio de búsqueda considerado para encontrar el modelo que presente la mejor exactitud, al momento de reconstruir el espacio de validación. Para buscar la mejor configuración del AE se utilizó el modelo *hyperband* que permite ajustar una serie de hiperparámetros a través de un método lógico de parada temprana teniendo en cuenta un número definido de configuraciones [126]. La principal razón por la cual se seleccionó este modelo de búsqueda de hiperparámetros fue debido a que es de 5 a 30 veces más rápido que otras técnicas típicas de optimización como se menciona en [126].

Tabla 12. Espacio de búsqueda de hiperparámetros para encontrar el mejor desempeño en la reconstrucción del conjunto de datos de validación.

Hiperparámetro	Rango de valores
Factor de regularización (L2)	[0.1, 1e-2, 1e-3, 1e-4,1e-5]
Tasa de aprendizaje	[0.1, 1e-2, 1e-3,1e-4,1e-5]
Número de neuronas capas uno y cinco	Valores entre [2,46]
Número de neuronas capas dos y cuatro	Valores entre [2,46]
Número de neuronas de la capa "bottleneck"	Valores entre [2,40]
Función de activación de la capa de salida	[*softmax, tanh*]
Épocas	Máximo 50 épocas

Tabla 13. Hiperparámetros que lograron obtener un mejor desempeño del conjunto de datos de validación, el cual alcanzó un valor de exactitud máximo de 0.9067 entre todos los modelos probados.

Hiperparámetros	Valores
Factor de regularización (L2)	0.001
Tasa de aprendizaje	0.0001
Número de neuronas capas uno y cinco	28
Número de neuronas capas dos y cuatro	24
Número de neuronas de la capa “bottleneck”	24
Función de activación de la capa de salida	"tanh"
Épocas	50

Posteriormente con el espacio de búsqueda definido se procede a ejecutar *hyperband* en donde se obtiene los hiperparámetros de la Tabla 13, los cuales alcanzaron una exactitud del 0.90 al momento de reconstruir el conjunto de datos de validación. El número de épocas es un hiperparámetro cuyo valor no es determinado por el modelo de *hyperband*, por lo cual se decidió entrenar el mejor modelo nuevamente con un número de épocas igual a 500. Teniendo en cuenta las variaciones de la función de loss y la exactitud presentada en las figuras 40-41, se decidió definir un número de épocas igual a 50 puesto que fue este valor que mejor exactitud alcanzó cuando se llevó a cabo la búsqueda basada en *hyperband*.

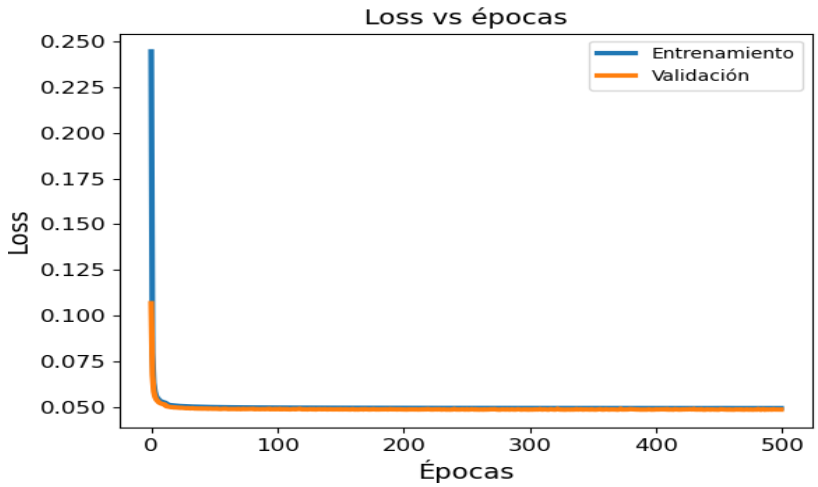


Figura 40. Valores de la función de Loss en función de las épocas para el conjunto de datos de entrenamiento y validación.

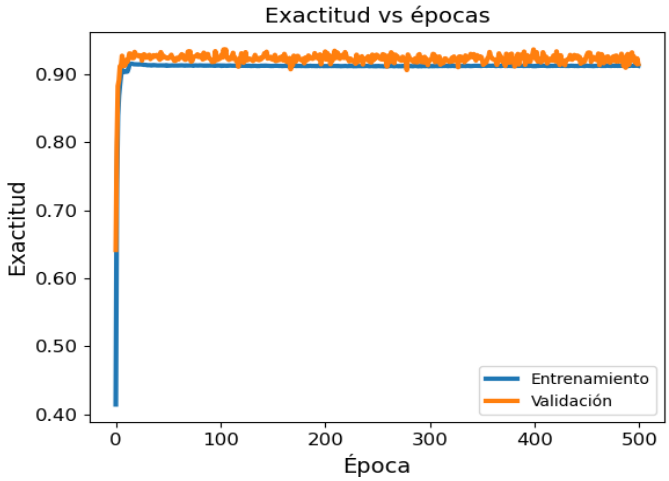


Figura 41. Precisión del conjunto de datos de entrenamiento y validación en función de las épocas.

4.4.1.4. Análisis y selección de espacios embebidos

Con los resultados de las pruebas realizadas anteriormente, se procedió a entrenar el modelo de predicción de mapas de contacto con el conjunto de datos PDB25 completo, para evaluar el desempeño, tanto del modelo base (sin RD) y los modelos que integran los espacios embebidos seleccionados. Posteriormente, se hace un estudio estadístico a través del test de Friedman y análisis de Nemenyi. con el fin de discriminar a los candidatos que pasarán a la etapa de afinación de hiperparámetros. La selección tendrá como referencia el F1-score para los contactos de rango medio y largo en la lista completa (ver tablas 14-15).

Tabla 14. En esta tabla se presentan los mejores modelos organizados de mayor a menor según el escalafón definido por el test de Friedman.

Técnica de RD	Dimensión	Ranking
PCA	24	6,0033
Base	46	6,5533
SVD	27	6,638
PCA	33	6,6457
AE	24	7,6576
RP	15	8,3413
SVD	21	8,3989
SVD	33	8,4033
RP	30	8,4924
RP	27	8,7054

Tabla 15. Contraste entre los cinco modelos mejor ranqueados con test de Friedman y los modelos restantes, haciendo uso del análisis de Nemenyi en donde se establece una diferencia crítica de 3,391 y un p-value de 1.80e-10.

Modelo	PCA 24	Base 46	SVD 27	PCA 33	AE1 24	RP 15	SVD 21	SVD 33	PCA 30	RP 30	RP 27
PCA 24	0	0	0	0	**1**	1	1	1	1	1	1
Base 46	0	0	0	0	**1**	1	1	1	1	1	1
SVD 27	0	0	0	0	**1**	1	1	1	1	1	1
PCA 33	0	0	0	0	**1**	1	1	1	1	1	1
AE 24	0	0	0	0	0	0	0	0	0	0	1

Con los resultados parciales mostrados en la Tabla 15, fue posible distinguir cinco candidatos que formarán parte de la etapa de afinación de hiperparámetros. Dichos candidatos según el análisis de Nemenyi, han superado en el valor de F1-score, al menos a uno de los modelos que integran las diferentes técnicas de RD seleccionadas, tanto para contactos de rango largo y medio de la lista completa. Lo interesante de la información presentada en la Tabla 15 es el hecho que en el entrenamiento no parece existir diferencias significativas, entre el modelo base y los modelos PCA, SVD y AE.

4.4.2. Selección de hiperparámetros

La búsqueda aleatoria (*random search*), a diferencia de la búsqueda por rejilla (*grid search*) que prueba todas las combinaciones posibles en un espacio de búsqueda de hiperparámetros, sólo prueba algunas combinaciones que son seleccionadas aleatoriamente (Figura 42). Si bien, es improbable conseguir la mejor combinación con una selección aleatoria, este enfoque permite encontrar picos de desempeño que, aunque no representan necesariamente el mejor conjunto posible de hiperparámetros, puede proporcionar un modelo que se acerque al modelo ideal [127].

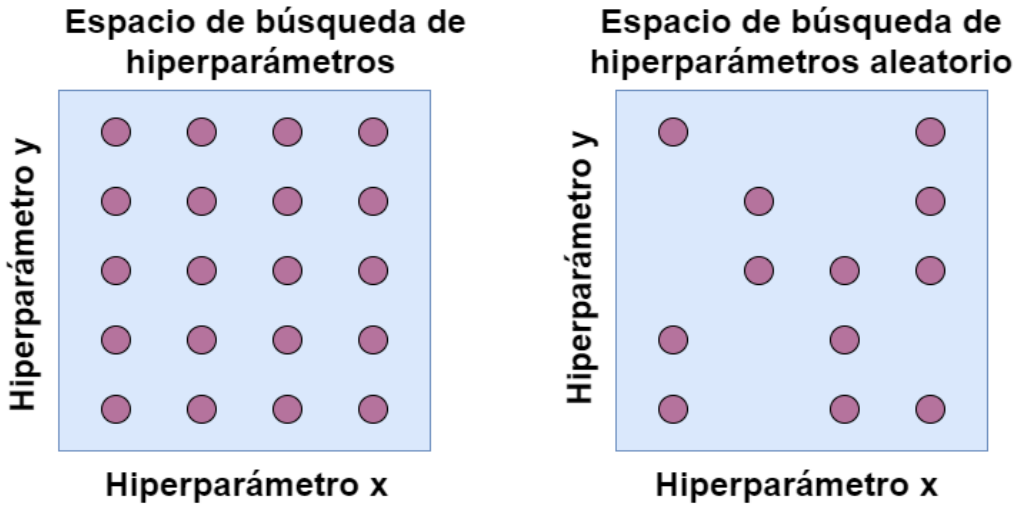


Figura 42. Esquema general de la búsqueda por rejilla y la búsqueda aleatoria para la selección de hiperparámetros.

Como se ha explicado anteriormente, la arquitectura del modelo implementado difiere en algunos aspectos de los problemas clásicos de predicción. Dichos aspectos se basan principalmente en: el costo computacional del entrenamiento; cadenas de longitudes variables; un proceso de unificación de características (concatenación externa); la naturaleza bidimensional de la salida del modelo; el proceso de evaluación que contiene diferentes tipos de contacto; y el volumen de información de entrada. Por este motivo, para generar una búsqueda de hiperparámetros efectiva (considerando los recursos computacionales disponibles), se optó por realizar una búsqueda de hiperparámetros aleatoria. Esto debido a que en términos de rendimiento este enfoque cuenta con una mayor eficiencia puesto que se tiene un costo computacional más bajo al limitar el número de modelos a ser entrenados [127].

La rejilla donde se procede a realizar la búsqueda aleatoria está conformada por los valores que se muestran en la Tabla 16, donde existe un total de 150 combinaciones posibles. Por cuestiones de costos computacionales y tiempo de entrenamiento se seleccionaron veinte combinaciones de hiperparámetros por modelo, para finalmente seleccionar los mejores resultados teniendo en cuenta el F1-score de la lista completa del rango de contactos medio y largo del conjunto de datos PDB25. Cabe resaltar que los hiperparámetros de los modelos iniciales de la Tabla 11, también cuentan como una combinación adicional. En total se entrenaron 100 diferentes combinaciones

estableciendo un número de épocas igual a veinte. Con el análisis de test de Friedman se seleccionaron los candidatos que presentaron un F1-score mayor.

Tabla 16. Conjunto de hiperparámetros y el rango de valores tenidos en cuenta para el afinamiento de hiperparámetros aleatorio.

Hiperparámetro	Rango de valores
Optimizador	Adamax, Nadam, Adam
Tasa de aprendizaje	0.1, 1e-2, 1e-3,1e-4,1e-5
Regularización	L1, L2
Factor de regularización	0.1, 1e-2, 1e-3, 1e-4,1e-5

Con las técnicas de RD definidas y la ejecución de la búsqueda por rejilla para el rango de hiperparámetros de la Tabla 16, se pudieron establecer los modelos de predicción de mapas de contactos finales, cuyos hiperparámetros están definidos en la Tabla 17. El establecimiento de dichos hiperparámetros al igual que en casos anteriores estuvo basado en el análisis de los resultados obtenidos al aplicar test de Friedman y el análisis de Nemenyi entre los modelos de la Tabla 15 (ver Anexo 2.2-2.7), pudiendo detectar una combinación de hiperparámetros dominante para cada uno de los casos. Cabe resaltar que tanto para el modelo base como para el modelo que integra SVD sobresale la configuración de hiperparámetros inicial de la Tabla 11.

Tabla 17. Conjunto de hiperparámetros seleccionados teniendo en cuenta la búsqueda por rejilla, el test de Friedman y el análisis de Nemenyi.

Modelo	Dimensión	Optimizador	Tasa de aprendizaje	Regularización	Fator de regularización
Base	46	Adam	0.001	Sin regularización	0
PCA	24	Nadam	0.0001	L2	0.0001
AE	24	Nadam	0.0001	L2	1e-05
SVD	27	Adam	0.001	Sin regularización	0

Capítulo 5

5. Resultados

Una vez finalizó el proceso de modelado descrito en el anterior capítulo, se procede a analizar los resultados obtenidos con cada uno de los modelos y configuraciones seleccionadas, los cuales presentaron los mejores desempeños dentro del pipeline de predicción planteado en este proyecto de investigación. El análisis de los resultados será dividido teniendo en cuenta tres aspectos importantes: En el primero se analizan las diferencias obtenidas en el proceso de entrenamiento y en la selección del umbral del logit; finalmente en la segunda y tercera parte se realiza un estudio de las diferencias encontradas en el análisis de Nemenyi de la lista completa y reducida de contactos para los datos de prueba utilizados (PDB25, 76CAMEO y MEMS400).

Como se explicó anteriormente los conjuntos de datos de pruebas utilizados para este estudio cuentan con las siguientes características: el conjunto de prueba PDB25 está conformado por 460 cadenas de proteínas; el conjunto 76CAMEO cuenta con 70 proteínas; mientras que MEMS400 esa conformado por 470 cadenas de proteínas. Adicionalmente, para los tres conjuntos de datos mencionados se tiene que, todas las cadenas que los conforman cuentan entre si con una similitud entre secuencias menor al 25%. Otro aspecto importante a destacar con los resultados que se presentan, es que los conjuntos de datos 76CAMEO y MEMS400 como se menciona en [85], representan *hard targets datasets*, lo cual permite tener un panorama más representativo del efecto de integrar técnicas de RD en el desempeño del modelo de predicción de mapas de contacto.

5.1. Análisis del umbral del logit

Un parámetro importante dentro de la obtención de la predicción de mapas de contacto es el establecimiento de un valor umbral del logit para la definición de contacto o no contacto. Lo anterior es importante, especialmente al momento de la evaluación de la lista completa de contactos, en donde todos los contactos dentro de un determinado rango son tenidos en cuenta. La definición de este parámetro fue realizada por medio del valor promedio del F1-score para todas las cadenas de proteínas dentro del conjunto de datos de prueba en función del umbral, el cual fue establecido en un rango de 0.05 a 0.91 en pasos de 0.01. Los resultados del procedimiento pueden ser observados en la Figura 43. Lo que destaca dentro de esta figura es el valor promedio máximo que alcanzan los modelos que tienen etapa de preprocesamiento basada en técnicas de RD, los cuales superan al modelo base. Especialmente se puede resaltar los modelos que integran PCA y AE, los cuales también sobresalen en la evaluación mediante valores promedios, test de Friedman junto con el análisis de Nemenyi.

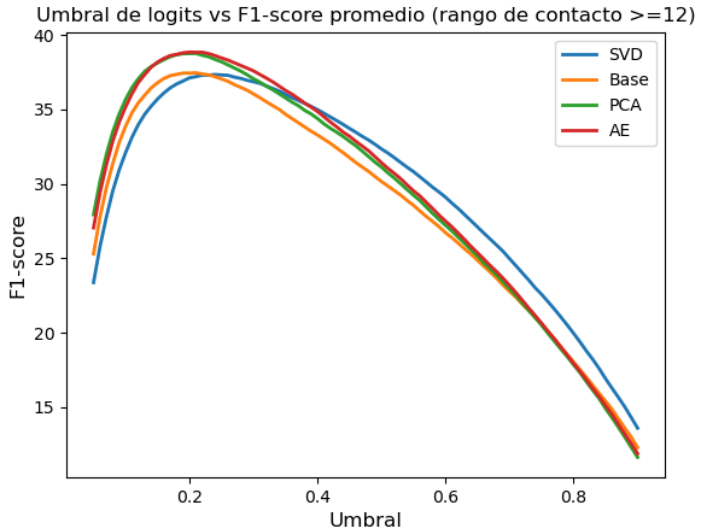


Figura 43. En el diagrama se relaciona el valor del F1-score en función del umbral del logit teniendo en cuenta los contactos de rango medio y largo para los cuatro modelos considerados.

En la Tabla 18 se definen el umbral del logit de cada uno de los modelos seleccionados, los cuales establecieron un pico en la métrica F1-Score para los contactos de rango medio y largo. Los valores que se presentan en esta tabla son

importantes puesto que son los utilizados para realizar el análisis final basado en el test de Friedman y en el análisis de Nemenyi.

Tabla 18. Se relaciona el valor umbral del logit que alcanza el máximo promedio de F1-score para contactos de rango medio y largo.

Modelo	Umbral	Máximo F1-Score
AE-24	0.2	38.85
PCA-24	0.21	38.76
RAW-46	0.21	37.46
SVD-27	0.24	37.34

5.2. Análisis de resultados

Uno de los aspectos más complicados al momento de realizar el análisis fue el manejo de la gran cantidad de tablas comparativas donde se condensan los rankings del Test de Friedman y el análisis de Nemenyi, puesto que se deben analizar los siguientes aspectos: dos tipos de listas de contacto, varias métricas de calidad, y para el caso de la lista reducida de contactos se tienen en cuenta los tops L/k contactos más probables (donde k puede tomar el valor de 1, 5 y 2). En total un número de 183 tablas (ANEXO 2.1) tuvieron que ser creadas y estudiadas, por este motivo fue necesario llevar a cabo el siguiente protocolo de condensación de resultados. En primer lugar, con el fin de evitar el error humano se automatizó la creación de las 183 tablas tomando de manera directa los datos de los archivos xlsx que contienen los cálculos para el análisis de Nemenyi. En segundo y último lugar se ideó un sistema de comparación resumido que compacta la información de todas las tablas para comparar el modelo base con los modelos que integran técnicas de RD, codificando lo resultados en los siguientes valores: El valor de uno denota una diferencia significativa entre el modelo base y la técnica de reducción de dimensión seleccionada; el valor de cero indica la ausencia de una diferencia significativa entre el modelo base y las técnicas de reducción de dimensión; por otra parte, el valor de menos uno denota que el método de reducción de dimensión seleccionado fue superado por el modelo sin ninguna clase de preprocesamiento.

5.2.1. Resultados para la lista completa de contactos

5.2.1.1. Resultados lista completa para el conjunto de datos PDB25

Al analizar la Tabla 26 se observa que para el rango extra-largo existe una diferencia estadísticamente significativa en tres de las cuatro métricas en favor de los modelos PCA y AE, siendo en la métrica de precisión en donde no existe diferencia significativa con el modelo base. En cuanto a los valores promedio que se muestran en la Tabla 25, PCA y AE superan al modelo base en cada una de las métricas mostrando así un mejor desempeño. Entre AE y PCA no existe un modelo definitivo puesto que AE posee un mejor MCC, F1-score y una mejor sensibilidad, mientras que PCA cuenta con una mayor precisión. Al estudiar las anteriores métricas se podría decir que AE tiene una capacidad más alta en detectar contactos extralargos reales. Por su parte SVD no cuenta con diferencias significativas con respecto al modelo base, pero también no se ve superado por este e incluso en términos de valores promedio supera al modelo base en la métrica de sensibilidad.

Dentro de los resultados obtenidos en el rango largo de contactos es importante destacar la superioridad de los modelos AE y PCA frente al modelo base en todas las métricas, lo cual sugiere la superioridad de estos dos modelos al momento de distinguir contactos de rango largo. En cuanto a los valores promedios los dos anteriores modelos mencionados tienen los valores más grandes para todas las métricas. El modelo AE cuenta con un mejor F1 score y sensibilidad, mientras que PCA tienen una mejor precisión, en cuanto al MCC se puede decir que en valores promedios existe un empate. El desempeño de SVD es similar al caso anterior en donde no existe diferencias significativas con el modelo base, excepto con la precisión donde se ve superado, sin embargo, en este caso podemos observar que supera al modelo base (diferencia estadísticamente significativa) en la métrica de sensibilidad.

El caso de los rangos medio y largo es muy similar a los resultados obtenidos en el extra-largo en donde AE y PCA superan al modelo base en todas las métricas exceptuando la precisión. No obstante, AE posee un desempeño inferior al modelo base. En cuanto a los valores promedio AE y PCA cuentan con los valores más grandes en comparación con los demás modelos en las métricas de MCC, F1-score y

sensibilidad. Para el caso de SVD es interesante denotar que posee una mayor sensibilidad.

El comportamiento observado en las predicciones en los rangos medio, largo y corto es similar al anterior, sin embargo, la principal diferencia radica en que PCA si logra vencer al modelo base de forma significativa en todas las métricas denotando así un mejor desempeño. AE por su parte vence en tres de las cuatro medidas, sin embargo, del mismo modo que en el caso anterior no cuenta con una mejor precisión con respecto al modelo base. En cuanto a los valores promedio se repite el mismo escenario del caso anterior.

Para este conjunto de datos se pudo observar una clara superioridad de los modelos que integran reducción de dimensión. De igual forma que en casos anteriores, tanto AE como PCA presentan una superioridad soportada en valores promedio y en el test de Friedman y análisis de Nemenyi. Un patrón interesante que resalta en esta base de datos es la diferencia significativa y los valores superiores de sensibilidad alcanzados en cada uno de los rangos, mostrando que los espacios embebidos generados por AE, PCA y SVD tuvieron un efecto positivo para la detección de contactos reales.

Tabla 19. Valores promedio de las métricas: coeficiente de correlación de Matthews (MCC), F1-score (F1), precisión (P) y sensibilidad (S), para la lista completa de contactos del conjunto de datos de prueba PDB25.

Extra-largo				
Modelo	**MCC**	**F1**	**P**	**S**
BASE-46	0.280	27.118	35.168	25.887
AE-24	**0.291**	**28.251**	35.761	**28.112**
PCA-24	0.287	27.774	**36.566**	26.534
SVD-27	0.275	26.784	33.635	26.268
Largo				
Modelo	**MCC**	**F1**	**P**	**S**
BASE-46	0.341	33.856	40.003	32.904
AE-24	0.353	**35.053**	40.159	**35.270**
PCA-24	**0.354**	34.918	**42.200**	33.838
SVD-27	0.339	33.737	38.964	33.450
Medio-Largo				
Modelo	**MCC**	**F1**	**P**	**S**
BASE-46	0.376	37.464	43.891	35.946

AE-24	0.389	**38.851**	43.342	**38.854**
PCA-24	**0.390**	38.764	**45.554**	37.410
SVD-27	0.373	37.346	41.979	36.912
Medio-Largo-Corto				
Modelo	**MCC**	**F1**	**P**	**S**
BASE-46	0.399	40.184	44.531	39.333
AE-24	0.410	**41.314**	44.255	**42.010**
PCA-24	**0.412**	41.213	**46.765**	40.092
SVD-27	0.397	39.975	43.334	40.087

Tabla 20. Esquema comparativo del desempeño en lista completa de los modelos que integran RD versus el modelo base para el conjunto de datos de prueba PDB25.

Extra-largo				
Modelo	**MCC**	**F1**	**P**	**S**
AE	1	1	0	1
PCA	1	1	0	1
SVD	0	0	-1	0
Largo				
Modelo	**MCC**	**F1**	**P**	**S**
AE	1	1	1	1
PCA	1	1	1	1
SVD	0	0	-1	1
Medio-Largo				
Modelo	MCC	F1	P	S
AE	1	1	-1	1
PCA	1	1	0	1
SVD	0	0	-1	1
Medio-Largo-Corto				
Modelo	**MCC**	**F1**	**P**	**S**
AE	1	1	-1	1
PCA	1	1	1	1
SVD	0	0	-1	1

5.2.1.2. Resultados lista completa para el conjunto de datos 76CAMEO

En este caso no se observan muchos escenarios en donde existan diferencias significativas con respecto al modelo base, sin embargo, este tampoco puede superar a los modelos que integran reducción de dimensión.

En el rango de contactos extra-largo no existe diferencias significativas en el desempeño teniendo en cuenta las cuatro métricas de evaluación. En cuanto a los

valores promedio se observa, que PCA tiene los valores más grandes, exceptuando por la métrica de sensibilidad que es en donde AE gana. El valor de precisión al igual que en casos anteriores es más grande con respecto a AE. Los valores promedios más bajos corresponden al modelo de SVD, sin ser vencido por el modelo base de manera significativa.

Al igual que el caso anterior en la mayoría de las métricas en el rango largo de contactos, no existe diferencia significativa. No obstante, AE consigue obtener un mayor desempeño con respecto al modelo base en la métrica de sensibilidad. En cuanto a los valores promedios PCA obtiene los índices más altos en todas las métricas exceptuando sensibilidad que es en donde AE obtiene el mayor valor. AE supera en los valores medios al modelo base exceptuando la métrica de sensibilidad. Es interesante mencionar que para este rango los valores promedio de F1-score y sensibilidad que presenta el modelo SVD superan al modelo base. El patrón descrito anteriormente se repite en los resultados para el rango de contacto medio y largo.

Tabla 21. Valores promedio de las métricas: coeficiente de correlación de Matthews (MCC), F1-score (F1), precisión (P) y sensibilidad (S), para la lista completa de contactos del conjunto de datos de prueba CAMEO76.

Extra-largo				
Modelo	**MCC**	**F1**	**P**	**S**
BASE-46	0.164	15.186	21.568	17.135
AE-24	0.173	16.199	21.491	**18.293**
PCA-24	**0.179**	**16.479**	**25.647**	17.805
SVD-27	0.159	14.847	20.023	17.016
Largo				
Modelo	**MCC**	**F1**	**P**	**S**
BASE-46	0.211	19.839	27.685	21.924
AE-24	0.229	21.815	28.045	**23.790**
PCA-24	**0.233**	**22.030**	**30.132**	23.217
SVD-27	0.217	20.521	27.320	22.714
Medio-Largo				
Modelo	**MCC**	**F1**	**P**	**S**
BASE-46	0.253	23.960	33.837	25.366
AE-24	0.269	25.777	33.586	**27.701**
PCA-24	**0.271**	**25.903**	**34.773**	26.873
SVD-27	0.259	24.725	32.790	26.797
Medio-Largo-Corto				

Modelo	MCC	F1	P	S
BASE-46	0.290	28.504	35.489	29.628
AE-24	0.301	29.472	36.029	31.132
PCA-24	**0.305**	**29.489**	**39.078**	**29.922**
SVD-27	0.300	28.944	38.549	30.478

Es en el rango medio-largo-corto en donde existen más diferencias significativas con respecto al modelo base, PCA obtiene un mejor MCC y una mejor precisión mientras que AE al igual que en casos anteriores cuenta con una mayor sensibilidad, del mismo modo esto se ve reflejado en los valores promedios. AE y PCA tiene mejores valores promedios que el modelo base, por su parte SVD en valores promedios tiene un mejor F1-score y sensibilidad que el modelo base.

Los resultados de este conjunto de datos no cuentan con una amplia diferencia significativa, no obstante, se puede observar al igual que con PDB25 una mejoría en la sensibilidad. Los mejores valores promedios se obtuvieron los métodos de RD, donde incluso SVD tuvo mejores valores promedio en F1 score y sensibilidad.

Tabla 22. Esquema comparativo del desempeño en lista completa de los modelos que integran RD versus el modelo base para el conjunto de datos de prueba CAMEO76.

Extra-largo				
Modelo	**MCC**	**F1**	**P**	**S**
AE-24	0	0	0	0
PCA-24	0	0	0	0
SVD-27	0	0	0	0
Largo				
Modelo	**MCC**	**F1**	**P**	**S**
AE-24	0	0	0	1
PCA-24	0	0	0	0
SVD-27	0	0	0	0
Medio-Largo				
Modelo	MCC	F1	P	S
AE-24	0	0	0	1
PCA-24	0	0	0	0
SVD-27	0	0	0	0
Medio-Largo-Corto				
Modelo	**MCC**	**F1**	**P**	**S**
AE-24	0	0	0	1

PCA-24	1	0	1	0
SVD-27	0	0	0	0

5.2.1.3. Resultados lista completa para el conjunto de datos MEMS400

Un patrón recurrente que se encuentra en los tres conjuntos de datos de prueba es la tendencia de los modelos que integran reducción de dimensión a obtener una mejoría en la métrica de sensibilidad, lo que demuestra que los espacios embebidos proveen una mejora en la detección de contactos de residuos reales, teniendo en cuenta que en la mayoría de los casos el modelo base no cuenta con diferencias estadísticamente significativas en las demás métricas.

Para este conjunto de datos en el rango extra-largo AE y PCA tienen una mejor sensibilidad con respecto al modelo base, pero no poseen diferencia significativa en las métricas MCC y F1-score. El modelo base del mismo modo que en casos anteriores cuenta con una mejor precisión, En cuanto a los valores promedio AE y PCA superan en todas las métricas exceptuando precisión. SVD se ve superado en todos los ámbitos por el modelo base.

Para el caso de rango largo de contactos AE no cuenta con diferencias significativas con respecto al modelo base en todas las métricas, por su parte PCA vence al modelo base en sensibilidad, pero pierde en precisión. El modelo PCA no tiene diferencias significativas en las métricas de MCC y F1-score. En cuanto a los valores promedio AE cuenta con los mejores valores para MCC y F1-score, mientras que PCA cuenta con el mayor valor de sensibilidad al igual que el caso anterior el modelo base cuenta con la mayor precisión. SVD se ve superado en cada uno de los escenarios por el modelo base.

El comportamiento de los modelos es el mismo para los rangos medio-largo y Medio-largo-corto, en donde AE no cuenta con diferencias significativas con respecto al modelo base en todas las métricas, pero no se ve superado en ningún escenario por el modelo base. Por otro lado, PCA vence al modelo base en sensibilidad, PCA no tiene diferencias significativas en las métricas de MCC y F1 score, y es vencido por el modelo base en la métrica de precisión. En cuanto a los valores promedio AE cuenta

con los mejores valores para MCC y F1-score, mientras que PCA cuenta con el mayor valor de sensibilidad, al igual que el caso anterior el modelo base cuenta con la mayor precisión. SVD se ve superado en cada uno de los escenarios por el modelo base.

Los modelos que integran reducción de dimensión demostraron que una representación más compacta de las características secuenciales igualó y superó en varios escenarios al modelo base, lo que verifica que una representación reducida de los datos puede mejorar los resultados incluso en arquitecturas actuales como lo son las basadas en aprendizaje profundo.

Tabla 23. Valores promedio de las métricas: coeficiente de correlación de Matthews (MCC), F1-score (F1), precisión (P) y sensibilidad (S), para la lista completa de contactos del conjunto de datos de prueba MEMS400.

Extra-largo				
Modelo	**MCC**	**F1**	**P**	**S**
BASE-46	0.200	18.976	**24.040**	21.043
AE-24	**0.210**	**20.327**	22.295	**24.388**
PCA-24	0.201	19.534	20.576	24.124
SVD-27	0.157	14.938	18.886	17.386
Largo				
Modelo	**MCC**	**F1**	**P**	**S**
BASE-46	0.267	26.082	**26.695**	31.109
AE-24	**0.270**	**26.553**	26.177	32.135
PCA-24	0.267	26.150	24.250	**33.688**
SVD-27	0.226	22.003	24.061	25.689
Medio-Largo				
Modelo	MCC	F1	P	S
BASE-46	0.292	28.675	**29.139**	33.482
AE-24	**0.296**	**29.323**	28.536	34.670
PCA-24	0.292	28.841	26.698	**36.163**
SVD-27	0.249	24.443	26.211	27.937
Medio-Largo-Corto				
Modelo	**MCC**	**F1**	**P**	**S**
BASE-46	0.301	29.851	**29.452**	34.883
AE-26	**0.308**	**30.647**	29.273	36.127
PCA-26	0.300	29.753	27.678	**36.705**
SVD-27	0.262	25.810	27.329	29.377

Tabla 24. Esquema comparativo del desempeño en lista completa de los modelos que integran RD versus el modelo base para el conjunto de datos de prueba MEMS400.

Extra-largo				
Modelo	**MCC**	**F1**	**P**	**S**
AE-24	0	0	-1	1
PCA-24	0	0	-1	1
SVD-27	-1	-1	-1	-1
Largo				
Modelo	**MCC**	**F1**	**P**	**S**
AE-24	0	0	0	0
PCA-24	0	0	-1	1
SVD-27	-1	-1	-1	-1
Medio-Largo				
Modelo	MCC	F1	P	S
AE-24	0	0	0	0
PCA-24	0	0	-1	1
SVD-27	-1	-1	-1	-1
Medio-Largo-Corto				
Modelo	**MCC**	**F1**	**P**	**S**
AE-24	0	0	0	0
PCA-24	0	0	-1	0
SVD-27	-1	-1	-1	-1

5.2.2. Resultados para la lista reducida de contactos

5.2.2.1. Resultados lista reducida para el conjunto de datos PDB25

Como se muestra en la Tabla 19, para el rango extralargo de contactos existe una clara superioridad del modelo AE en los valores promedio de las métricas utilizadas en relación con los demás modelos. Adicionalmente, al contrastar la anterior información con la Tabla 20 se puede observar que para este rango los modelos de AE y PCA presentan una diferencia significativa con respecto al modelo base en todas las métricas (F1-score, precisión y sensibilidad). Con respecto a SVD se observa que según el análisis de Nemenyi no existe una diferencia significativa con el modelo base. Lo que demuestra que al menos para este rango la reducción de dimensión no genera una disminución del desempeño, si no una mejoría.

Al igual que el caso anterior, dentro del rango largo de contactos tanto AE como PCA demuestran los mejores índices promedio en todas las métricas de evaluación, junto con una diferencia estadísticamente significativa en relación con el modelo base. Igualmente, SVD no presenta una diferencia estadísticamente significativa con respecto al modelo base, comprobando de manera experimental que para este rango y el anterior si existe una mejoría en el desempeño, considerando además que los contactos de rango extralargo y largo son los más difíciles de identificar y los que poseen más información acerca de la estructura de la proteína.

Cuando se evalúan los contactos de rango medio y largo se observa también que los modelos de AE y PCA presentan los mejores valores promedio para cada una de las métricas de evaluación, sin embargo, no existe una completa superioridad de AE en todas las métricas, puesto que PCA tiene un valor promedio mayor en las métricas que evalúan los $L/2$ contactos más probables. Al igual que en los casos anteriores AutoEncoders y PCA tienen diferencias significativas con respecto al modelo base, mientras que SVD no es superado, ni es mejor que el modelo base.

Al evaluar los contactos de rango medio el modelo PCA es el que cuenta con un mayor valor promedio para cada una de las métricas de rango medio, sin embargo, AE tiene un mayor valor promedio que el modelo base. Al llevar a cabo análisis de Nemenyi (Tabla 20) se muestra la existencia de diferencia estadísticamente significativa con respecto al modelo base para AE y PCA, por el contrario, SVD no mejora al modelo base, pero tampoco es superado por este.

Para el rango de contactos corto, AE contiene los mejores índices de valores en las métricas de evaluación entre todos los modelos, y es el único modelo que posee una mejora con respecto al modelo base con diferencia estadísticamente significativa para todas las métricas. PCA por su parte si posee mejores valores promedios, sin embargo, no tiene una diferencia estadísticamente significativa al momento de evaluar los $L\ y\ L/5$ contactos más probables. Por su parte SVD si cuenta con diferencia estadísticamente significativa para los $L/5$ contactos más probables, con un mayor valor promedio en el F1-score, precisión y en la sensibilidad.

Tabla 25. Valores promedio de las métricas: F1-score (F1), precisión (P) y sensibilidad (S), para la lista reducida de contactos del conjunto de datos de prueba PDB25.

Extra-largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
BASE-46	26.718	23.290	37.044	28.523	34.610	27.937	22.507	48.757	16.501
AE-24	**28.141**	**24.518**	**39.610**	**29.844**	**36.091**	**29.489**	**23.348**	**50.564**	**16.900**
PCA-24	27.696	24.109	38.866	29.519	35.719	29.086	23.059	49.972	16.816
SVD-27	26.546	23.134	37.149	28.003	33.973	27.418	22.250	48.313	15.875
Largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
BASE-46	35.325	40.327	33.479	30.512	53.413	22.365	18.631	65.821	11.147
AE-24	**36.683**	**41.734**	**34.953**	**31.719**	**55.182**	**23.416**	**19.470**	**67.903**	**11.729**
PCA-24	36.529	41.593	34.712	31.522	54.967	23.163	19.199	67.378	11.522
SVD-27	35.246	40.167	33.529	30.292	52.942	22.254	18.620	65.626	11.162
Medio-largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
BASE-46	37.517	49.862	31.320	29.405	62.398	19.806	16.464	73.705	9.419
AE-24	**38.983**	**51.696**	**32.639**	30.483	64.556	20.555	**17.163**	**76.220**	**9.837**
PCA-24	38.831	51.510	32.470	**30.578**	**64.699**	**20.621**	17.071	76.179	9.770
SVD-27	37.370	49.609	31.199	29.373	62.335	19.740	16.577	74.100	9.478
Medio									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
BASE-46	35.891	26.575	67.153	42.759	40.385	52.543	39.003	58.630	32.272
AE-24	36.850	27.243	69.094	43.780	41.284	53.836	39.827	59.773	32.919
PCA-24	**36.919**	**27.321**	**69.158**	**43.893**	**41.403**	**54.218**	**40.210**	**60.263**	**33.351**
SVD-27	35.985	26.569	67.381	42.987	40.429	53.047	39.257	58.905	32.360
Corto									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
BASE-46	34.499	22.781	79.437	45.417	36.924	64.555	46.397	57.627	41.280
AE-24	**35.020**	**23.115**	**80.858**	**46.376**	**37.627**	**66.180**	**47.401**	**58.708**	**42.293**
PCA-24	34.688	22.897	80.190	45.908	37.221	65.655	47.061	58.207	42.000
SVD-27	34.493	22.781	79.328	45.902	37.312	65.198	46.866	58.227	41.652

Tabla 26. Esquema comparativo del desempeño en lista reducida de los modelos que integran RD versus el modelo base para el conjunto de datos prueba PDB25.

Extra-largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
AE-24	1	1	1	1	1	1	1	1	1
PCA-24	1	1	1	1	1	1	1	1	1
SVD-27	0	0	0	0	0	0	0	0	0

Largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
AE-24	1	1	1	1	1	1	1	1	1
PCA-24	1	1	1	1	1	1	1	1	1
SVD-27	0	0	0	0	0	0	0	0	0
Medio-largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
AE-24	1	1	1	1	1	1	1	1	1
PCA-24	1	1	1	1	1	1	1	1	1
SVD-27	0	0	0	0	0	0	0	0	0
Medio									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
AE-24	1	1	1	1	1	1	1	1	1
PCA-24	1	1	1	1	1	1	1	1	1
SVD-27	0	0	0	0	0	0	0	0	0
Corto									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
AE-24	1	1	1	1	1	1	1	1	1
PCA-24	0	0	0	1	1	1	0	0	0
SVD-27	0	0	0	0	0	0	1	1	1

5.2.2.2. Resultados lista reducida para el conjunto de datos 76CAMEO

A diferencia del caso anterior en donde existían diferencias significativas entre los modelos, para este conjunto de datos de prueba parece existir un desempeño similar entre todos modelos. Por ejemplo, para el caso de rango extra-largo la Tabla 22 muestra que según test de Friedman y análisis de Nemenyi ningún modelo sobresale. No obstante, al analizar los valores promedios se puede observar que los mejores valores son obtenidos con AE y PCA, lo cual denota una mejoría en este aspecto una mejoría en el desempeño.

Tabla 27. Valores promedio de las métricas: F1-score (F1), precisión (P) y sensibilidad (S), para la lista reducida de contactos del conjunto de datos de prueba CAMEO76.

Extra-largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
BASE-46	18.594	14.921	29.565	19.923	21.451	22.030	16.392	30.791	12.488
AE-24	**19.865**	**15.980**	31.190	21.203	23.070	23.037	17.883	**33.180**	14.238
PCA-24	19.836	15.937	**31.500**	**21.571**	**23.317**	**23.793**	**18.018**	32.884	**14.479**
SVD-27	17.979	14.478	28.238	18.970	20.687	20.567	15.713	29.576	12.457

Largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
BASE-46	25.978	26.538	30.258	23.901	35.561	20.989	16.285	45.129	11.596
AE-24	**27.963**	**28.615**	**32.037**	**25.548**	**38.357**	**22.176**	16.753	47.654	11.334
PCA-24	27.936	28.531	31.990	25.087	37.804	21.068	**17.294**	**49.108**	**11.617**
SVD-27	26.287	26.805	30.298	24.643	36.583	21.560	17.170	47.250	12.170
Medio-largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
BASE-46	29.019	34.891	27.790	24.267	44.504	18.324	14.936	53.823	9.247
AE-24	**31.706**	**37.905**	**30.640**	25.690	47.094	19.197	15.628	56.462	9.581
PCA-24	31.601	37.833	30.343	**25.977**	**47.581**	**19.522**	**15.872**	**56.955**	**9.740**
SVD-27	30.461	36.321	29.489	25.463	46.149	19.319	15.018	54.257	9.233
Medio									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
BASE-46	30.234	22.264	59.232	34.929	32.953	45.335	31.383	47.067	28.193
AE-24	**31.479**	**23.304**	**60.980**	35.618	33.849	46.076	30.749	46.833	26.983
PCA-24	31.243	23.054	60.923	**36.390**	**34.267**	**47.184**	**32.081**	**47.970**	**28.492**
SVD-26	30.299	22.345	59.266	35.654	33.738	46.069	31.356	47.142	27.620
Corto									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
BASE-46	31.284	20.845	72.342	38.840	32.141	56.035	37.867	48.220	34.639
AE-24	31.911	21.276	73.988	40.074	33.082	58.248	37.871	48.400	34.592
PCA-24	**31.991**	**21.283**	**74.244**	**40.357**	**33.269**	**57.755**	**38.408**	**48.694**	**35.092**
SVD-26	31.843	21.233	73.409	39.553	32.648	56.903	38.953	49.483	35.510

Para el caso de contactos de rango largo en contraste del caso anterior, si existe una diferencia significativa con respecto al modelo base, tanto AE como PCA tienen un mejor desempeño en todas las métricas que evalúan los top L contactos más probables, adicionalmente PCA también tiene un mejor desempeño en las predicciones que conforman los top $L/5$ contactos más probables. En cuanto a los valores promedio PCA y AutoEncoders poseen los valores más altos entre todos los modelos, específicamente AE gana en los top L y $L/2$, mientras que PCA resalta en las métricas de los top $L/5$. Algo importante a resaltar de SVD para este rango de contactos es que posee un valor promedio más grande que el modelo base. Lo cual denotaría una mejoría en el desempeño con la integración de técnicas de RD.

Para este rango conformado por los contactos de rango medio y largo la aplicación de técnicas de reducción de dimensión parece tener un efecto positivo en el desempeño de la predicción teniendo en cuenta diferencias estadísticamente significativas en dos de los tres escenarios que se presentan dentro de la lista reducida (top L y $L/2$). Para

la lista reducida top L tanto PCA como AE presentan mejor desempeño que el modelo base, lo cual también se ve reflejado en los valores promedio que son más grandes que el modelo base. Al igual que el caso anterior SVD tiene mejores valores promedios que el modelo base en todas las métricas consideradas excepto en la sensibilidad para los tops $L/5$ contactos. Sin embargo, no cuenta con diferencia estadísticamente significativa cuando se contrasta con el modelo base.

Para los rangos medio y corto no existe diferencia significativa entre los diferentes modelos. Sin embargo, para el caso de contactos de rango medio parece existir una superioridad en el valor promedio de todos los modelos que integran reducción de dimensión (AE, PCA, SVD), el único valor promedio en donde SVD es superado por el modelo base es en la métrica de F1-score y la sensibilidad para los $L/5$, contactos más probables.

En este sentido analizando los resultados para este conjunto de datos se pudo observar que, a pesar de no encontrar diferencias significativas en la mayoría de los escenarios con el modelo base, este no supero a ninguno de los modelos que integran reducción de dimensión. La mejoría en este caso estuvo centralizada en algunos aspectos sobre todo en los valores promedios, con lo cual se podría afirmar que para este conjunto de datos la aplicación de técnicas de RD también mejora el desempeño de las predicciones.

Tabla 28. Esquema comparativo del desempeño en lista reducida de los modelos que integran RD versus el modelo base para el conjunto de prueba 76CAMEO.

Extra-largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
AE-24	0	0	0	0	0	0	0	0	0
PCA-24	0	0	0	0	0	0	0	0	0
SVD-27	0	0	0	0	0	0	0	0	0
Largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
AE-24	1	1	1	0	0	0	0	0	0
PCA-24	1	1	1	0	0	0	1	1	1
SVD-27	0	0	0	0	0	0	0	0	0
Medio-largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
AE-24	1	1	1	0	0	0	0	0	0

PCA-24	1	1	1	1	1	1	0	0	0
SVD-27	0	0	0	0	0	0	0	0	0
Medio									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
AE-24	0	0	0	0	0	0	0	0	0
PCA-24	0	0	0	0	0	0	0	0	0
SVD-27	0	0	0	0	0	0	0	0	0
Corto									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
AE-24	0	0	0	0	0	0	0	0	0
PCA-24	0	0	0	0	0	0	0	0	0
AE-24	0	0	0	0	0	0	0	0	0

5.2.2.3. Resultados lista reducida para el conjunto de datos MEMS400

Analizando la Tabla 24 para el rango de contactos extralargo se observa que AE es el único modelo que pudo superar al original puesto que presenta diferencias significativas, además de obtener mejores valores promedios entre todos los demás modelos (Tabla 23). Del mismo modo cabe resaltar que PCA obtiene un mejor desempeño que el modelo base en todas las métricas relacionadas con los top $L/5$ de los valores promedios. El espacio embebido SVD para este caso en específico se vio completamente superado por el modelo base.

Tabla 29. Valores promedio de las métricas: F1-score (F1), precisión (P) y sensibilidad (S), para la lista reducida de contactos del conjunto de datos de prueba MEMS400.

Extra-largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
BASE-46	18.489	15.646	29.036	20.052	23.182	21.547	17.261	34.572	13.095
AE-24	**19.531**	**16.487**	**31.275**	**21.447**	**24.650**	**23.447**	**18.711**	**37.067**	**14.385**
PCA-24	19.072	16.127	30.197	20.923	24.073	22.604	18.099	36.119	13.653
SVD-27	15.030	12.619	23.825	16.140	18.579	17.292	13.999	28.128	10.638
Largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
BASE-46	27.931	28.790	30.694	26.543	40.429	21.708	18.287	53.209	11.744
AE-24	28.613	29.543	31.360	**27.356**	**41.755**	**22.275**	18.882	55.064	12.028
PCA-24	**28.673**	**29.447**	**31.968**	27.287	41.564	22.215	**19.090**	**55.430**	**12.261**
SVD-27	24.061	24.980	26.347	22.875	35.164	18.743	16.023	47.162	10.401
Medio-largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**

BASE-46	30.112	35.645	29.269	26.197	46.879	19.828	16.514	57.782	10.135
AE-24	30.979	36.672	30.316	26.840	48.101	20.420	17.125	59.728	10.508
PCA-24	**31.095**	**36.675**	**30.352**	**27.088**	**48.379**	**20.432**	**17.343**	**60.289**	**10.685**
SVD-27	26.270	31.606	25.236	22.805	41.501	17.178	14.427	51.731	8.841
Medio									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
BASE-46	26.018	18.557	65.741	31.712	27.520	51.964	32.750	40.256	34.181
AE-24	26.057	18.588	65.663	**32.420**	**28.019**	**53.500**	33.880	41.536	35.234
PCA-24	**26.178**	**18.648**	**66.491**	32.344	27.926	53.381	**34.543**	**41.993**	**36.309**
SVD-27	23.945	17.298	58.691	28.796	25.490	45.940	28.873	36.670	29.322
Corto									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
BASE-46	22.605	14.112	71.305	30.855	22.925	57.471	34.175	36.211	36.999
AE-24	**23.119**	**14.417**	**73.557**	**31.569**	**23.392**	**59.192**	**35.203**	**37.071**	**38.239**
PCA-24	21.665	13.526	68.242	28.603	21.210	53.051	32.685	34.511	35.689
SVD-27	21.254	13.326	65.689	28.414	21.316	51.419	31.655	34.048	33.900

Los resultados de la predicción de contactos de rango largo se puede observar una superioridad estadísticamente significativa con respecto al modelo base, para los modelos AE y PCA, lo cual se ve reflejado también en los valores promedios más altos en todas las métricas de estos dos modelos. Con respecto a los valores promedios presentados en la Tabla 23 no existe una dominancia clara entre AE y PCA, puesto que AE domina en los top $L/2$, y PCA en los top L y $L/2$. A diferencia de los conjuntos de datos anteriores SVD se ve superado completamente por el modelo base con diferencias estadísticamente significativas.

El rango de contactos largo y medio tiene un comportamiento muy similar al analizado en el rango anterior puesto que se puede observar una superioridad significativa con respecto al modelo base en todas las métricas para los modelos de PCA y AE, lo cual se refleja en los valores promedios más altos de estos dos modelos, en donde PCA domina con respecto a los demás. Sin embargo, en este escenario AE también supera al modelo base. Como se muestra con el análisis de test de Friedman y Nemenyi, SVD se ve superado tanto por el modelo base y además obtiene los más bajos valores promedio para cada una de las métricas.

En el rango medio no existe diferencia estadísticamente significativa en la mayoría de las métricas con respecto al modelo base, AE y PCA superan al modelo base en el conjunto de métricas que conforman el top $L/5$ de predicciones. En cuanto a los

valores promedio se puede observar que el modelo PCA cuenta con las mejores métricas de desempeño, seguido por AE, el cual también tiene unos valores superiores con respecto al modelo base. SVD al igual que en casos anteriores se ve superado por el modelo base, junto con los menores valores promedio.

El rango de contactos corto presenta un resultado es interesante puesto que diferencia los resultados anteriores, AE según el test de Friedman y análisis de Nemenyi presenta una diferencia significativa con respecto al modelo base en todas las métricas de evaluación, dominando de igual forma en los valores promedios que se observan en la Tabla 23. Para este caso en particular tanto PCA como SVD no pueden superar al modelo base.

Tabla 30. Esquema comparativo del desempeño en lista reducida de los modelos que integran RD versus el modelo base para el conjunto de datos de prueba MEMS400.

Extra-largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
AE-24	1	1	1	1	1	1	1	1	1
PCA-24	0	0	0	0	0	0	1	1	1
SVD-27	-1	-1	-1	-1	-1	-1	-1	-1	-1
Largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
AE-24	1	1	1	1	1	1	1	1	1
PCA-24	1	1	1	1	1	1	1	1	1
SVD-27	-1	-1	-1	-1	-1	-1	-1	-1	-1
Medio-largo									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
AE-24	1	1	1	1	1	1	1	1	1
PCA-24	1	1	1	1	1	1	1	1	1
SVD-27	-1	-1	-1	-1	-1	-1	-1	-1	-1
Medio									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
AE-24	0	0	0	0	0	0	1	1	1
PCA-24	0	0	0	0	0	0	1	1	1
SVD-27	-1	-1	-1	-1	-1	-1	-1	-1	-1
Corto									
Modelo	**F1-L**	**P-L**	**S-L**	**F1-L/2**	**P-L/2**	**S-L/2**	**F1-L/5**	**P-L/5**	**S-L/5**
AE-24	1	1	1	1	1	1	1	1	1
PCA-24	-1	-1	-1	-1	-1	-1	-1	-1	-1
SVD-27	-1	-1	-1	-1	-1	-1	-1	-1	-1

Capítulo 6

6. conclusiones

Dada la gran cantidad de trabajos encontrados durante la revisión sistemática es importante destacar la vigencia de los métodos de RD en la actualidad, puesto que aún son utilizados como etapas de preprocesamiento válidas capaces de incrementar el desempeño de modelos de clasificación o predicción en el área de la bioinformática. Si bien se analizaron trabajos en donde se utilizaban características relacionadas con las características secuenciales de las proteínas, no se encontró un patrón definido en cuanto a la mejor técnica de RD a utilizar puesto que varió de acuerdo al problema en específico, denotando así la importancia de realizar estudios específicos para cada problema.

Un aspecto importante a resaltar es la capacidad de algunas técnicas de RD para manejar grandes volúmenes datos, siendo los basados en proyecciones lineales los más eficientes y los más utilizados en este trabajo de investigación, dada su capacidad para procesar conjuntos de datos que superan el millón de muestras. Si bien durante la revisión sistemática se encontró la utilización de técnicas de extracción de características no lineales basadas en Manifold learning estas no fueron muy útiles, debido a que el cálculo del espacio embebido requiere cantidades de memoria que generalmente ocasionan errores cuando el número de puntos es mayor o igual al millón. Por lo tanto, a menos que se implemente una versión escalable o una aproximación con un costo computacional más reducido de dichas técnicas, su implementación en aplicaciones reales de *big data* no es práctica.

En cuanto al modelo de predicción utilizado es importante resaltar que al ser una arquitectura bastante cercana al estado del arte actual permitió establecer un escenario de predicción más realista, en donde se procesa la proteína de forma

completa y no aminoácidos individuales, brindando mayor valor a la selección, integración e implementación de las técnicas de RD que finalmente fueron utilizadas. Como se pudo observar en las diferentes pruebas la integración de técnicas de RD tiene un efecto en el desempeño del modelo de predicción de mapas de contacto, lo cual denota una clara dependencia con la calidad de las características de entrada. Sin embargo, se pudieron observar escenarios en donde no existía diferencia estadísticamente significativa entre el modelo base (sin ninguna clase de preprocesamiento) y los diferentes modelos entrenados haciendo uso de espacios embebidos. Lo cual potencialmente podría servir cuando el volumen de datos sea considerablemente mayor, si bien no se detectaron variaciones significativas en el tiempo de entrenamiento, un espacio de baja dimensión podría potencialmente ahorrar el espacio en memoria de almacenamiento, sin sacrificar el rendimiento del algoritmo de predicción de mapas de contacto, cómo se pudo observar en los resultados presentados.

Durante el desarrollo de este trabajo de investigación se pudieron detectar varias oportunidades de trabajos futuros tanto en la predicción de mapas de contacto como en la exploración de técnicas de RD alternativas. Un primer aspecto a resaltar es la complejidad de esta tarea de predicción puesto que difiere metodológicamente de los problemas clásicos de predicción, sobre todo en las arquitecturas que se utilizan, la naturaleza de las características de entrada y la forma cómo se evalúan los resultados. Esto deja la necesidad de implementar pipelines integrados que permiten encapsular el entrenamiento y la evaluación en un solo script. Si bien en este proyecto de investigación se avanzó en gran medida en este problema con la adaptación de la arquitectura utilizada con el API de Keras, y en la implementación de funciones en Python para el cálculo rápido y directo de las métricas según los lineamientos del CASP, no se encontró un método, metodología o implementación definida para el ajuste de hiperparámetros. Esto debido a que la evaluación del modelo no se observa directamente con una única métrica, si no en la calidad de la predicción en las diferentes listas y clases de contacto que existen. Un segundo aspecto es la carencia de herramientas para extraer mapas de contacto reales a través de los archivos Fasta y PDB de manera que se puedan expandir fácilmente los conjuntos de datos de entrenamiento, prueba y validación. Este módulo también fue abordado en este proyecto de investigación haciendo uso de la herramienta Conkit, junto con un script de corrección de índices de aminoácidos o residuos que permitió obtener mapas de contacto sin desfases.

Capítulo 7

7. Bibliografía

[1] K. A. Dill and J. L. MacCallum, “The Protein-Folding Problem, 50 Years On,” *Science (80-.).*, vol. 338, no. 6110, pp. 1042–1046, Nov. 2012, doi: 10.1126/science.1219021.

[2] N. D. Jana, S. Das, and J. Sil, “Backgrounds on Protein Structure Prediction and Metaheuristics,” 2018, pp. 1–28.

[3] U. H. E. Hansmann, “Protein folding in silico: an overview,” *Comput. Sci. Eng.*, vol. 5, no. 1, pp. 64–69, Jan. 2003, doi: 10.1109/MCISE.2003.1166554.

[4] S.-H. Feng, J.-Y. Xu, and H.-B. Shen, *Artificial intelligence in bioinformatics*. Elsevier Inc., 2020.

[5] D. Whitford, *Proteins: Structure and Function*, 1st ed. London, Uk, 2013.

[6] J. M. Tyszka, S. E. Fraser, and R. E. Jacobs, “Magnetic resonance microscopy: Recent advances and applications,” *Curr. Opin. Biotechnol.*, vol. 16, no. 1 SPEC. ISS., pp. 93–99, 2005, doi: 10.1016/j.copbio.2004.11.004.

[7] S. Vorberg, “Bayesian statistical approach for protein residue-residue contact prediction,” lmu, 2017.

[8] M. Vendruscolo, E. Kussell, and E. Domany, “Recovery of protein structure from contact maps,” *Fold. Des.*, vol. 2, no. 5, pp. 295–306, 1997, doi: 10.1016/S1359-0278(97)00041-2.

[9] J. Xie, W. Ding, L. Chen, Q. Guo, and W. Zhang, “Advances in Protein Contact Map Prediction Based on Machine Learning,” *Med. Chem. (Los. Angeles).*, vol. 11, no. 3, pp. 265–270, 2015, doi: 10.2174/1573406411666141230095427.

[10] S. Ayesha, M. K. Hanif, and R. Talib, “Overview and comparative study of dimensionality reduction techniques for high dimensional data,” *Inf. Fusion*, vol. 59, no. January, pp. 44–58, 2020, doi: 10.1016/j.inffus.2020.01.005.

[11] P. Chen, C. Liu, L. Burge, M. Mohammad, B. Southerland, and C. Gloster, “Prediction of inter-residue contact clusters from hydrophobie cores,” *Proc. - 7th Int. Conf. Mach. Learn. Appl. ICMLA 2008*, no. May 2014, pp. 703–708, 2008, doi: 10.1109/ICMLA.2008.74.

[12] J. P. dos S. Pires, “Representing Amino Acid Contacts In Protein Interfaces,” 2020.

[13] A. D. Keromytis, “Comparative Analysis,” *SpringerBriefs Comput. Sci.*, vol. 1, pp. 57–60, 2011, doi: 10.1007/978-1-4419-9866-8_5.

[14] S. A. H. Bukhari, “What is Comparative Study,” *SSRN Electron. J.*, 2012, doi:

10.2139/ssrn.1962328.
[15] C. Drummond, “Machine learning as an experimental science,” *AAAI Work. - Tech. Rep.*, vol. WS-06-06, pp. 1–5, 2006, doi: 10.1007/bf00115008.
[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.
[17] C. E. S. Toca, *Predicción de mapas de contactos de proteínas mediante multiclasificadores*. Editorial Universitaria, 2014.
[18] Y. Y. Ji and Y. Q. Li, “The role of secondary structure in protein structure selection,” *Eur. Phys. J. E*, vol. 32, no. 1, pp. 103–107, 2010, doi: 10.1140/epje/i2010-10591-5.
[19] H. M. Berman *et al.*, “The Protein Data Bank and the challenge of structural genomics,” *Nat. Struct. Biol.*, vol. 7, no. SUPPL., pp. 957–959, 2000, doi: 10.1038/80734.
[20] A. Bateman *et al.*, “UniProt: The universal protein knowledgebase,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D158–D169, 2017, doi: 10.1093/nar/gkw1099.
[21] C. B. Anfinsen and S. Moore, “The Thermodynamic Hypothesis of Protein Folding: the Work of Christian Anfinsen,” *J. Biol. Chem.*, vol. 281, no. 14, pp. e11–e11, 2006.
[22] J. Lee, P. L. Freddolino, and Y. Zhang, “Ab initio protein structure prediction,” in *From protein structure to function with bioinformatics*, Springer, 2017, pp. 3–35.
[23] B. Kuhlman and P. Bradley, “Advances in protein structure prediction and design,” *Nat. Rev. Mol. Cell Biol.*, vol. 20, no. 11, pp. 681–697, 2019, doi: 10.1038/s41580-019-0163-x.
[24] X. Qu, R. Swanson, R. Day, and J. Tsai, “A Guide to Template Based Structure Prediction,” *Curr. Protein Pept. Sci.*, vol. 10, no. 3, pp. 270–285, 2009, doi: 10.2174/138920309788452182.
[25] J. Xu, F. Jiao, and L. Yu, “Protein Structure Prediction Using Threading,” in *Protein Structure Prediction*, M. J. Zaki and C. Bystroff, Eds. Totowa, NJ: Humana Press, 2008, pp. 91–121.
[26] T. Schwede, “Protein modeling: What happened to the ‘protein structure gap’?,” *Structure*, vol. 21, no. 9, pp. 1531–1540, 2013, doi: 10.1016/j.str.2013.08.007.
[27] S. Ornes, “Let the structural symphony begin,” *Nature*, vol. 536, no. 7616, pp. 361–363, Aug. 2016, doi: 10.1038/536361a.
[28] B. He, S. M. Mortuza, Y. Wang, H. Bin Shen, and Y. Zhang, “NeBcon: Protein contact map prediction using neural network training coupled with naïve Bayes classifiers,” *Bioinformatics*, vol. 33, no. 15, pp. 2296–2306, 2017, doi: 10.1093/bioinformatics/btx164.
[29] T. Shen *et al.*, “When homologous sequences meet structural decoys: Accurate contact prediction by tFold in CASP14—(tFold for CASP14 contact prediction),” *Proteins Struct. Funct. Bioinforma.*, vol. 89, no. 12, pp. 1901–1910, 2021, doi: 10.1002/prot.26232.
[30] K. Itoh and M. Sasai, “Flexibly varying folding mechanism of a nearly symmetrical protein: B domain of protein A,” *Proc. Natl. Acad. Sci. U. S. A.*, vol.

103, no. 19, pp. 7298–7303, 2006, doi: 10.1073/pnas.0510324103.
[31] R. Shrestha *et al.*, “Assessing the accuracy of contact predictions in CASP13,” *Proteins Struct. Funct. Bioinforma.*, vol. 87, no. 12, pp. 1058–1068, 2019, doi: 10.1002/prot.25819.
[32] D. T. Jones, T. Singh, T. Kosciolek, and S. Tetchner, “MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins,” *Bioinformatics*, vol. 31, no. 7, pp. 999–1006, 2015, doi: 10.1093/bioinformatics/btu791.
[33] B. Adhikari, “A fully open-source framework for deep learning protein real-valued distances,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020, doi: 10.1038/s41598-020-70181-0.
[34] B. Adhikari and J. Cheng, “Protein Residue Contacts and Prediction Methods,” vol. 1415, no. 2, 2016, pp. 463–476.
[35] S. Ovchinnikov, H. Park, D. E. Kim, F. DiMaio, and D. Baker, “Protein structure prediction using Rosetta in CASP12,” *Proteins Struct. Funct. Bioinforma.*, vol. 86, pp. 113–121, 2018, doi: 10.1002/prot.25390.
[36] T. Kosciolek and D. T. Jones, “De novo structure prediction of globular proteins aided by sequence variation-derived contacts,” *PLoS One*, vol. 9, no. 3, 2014, doi: 10.1371/journal.pone.0092197.
[37] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil, “PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments,” *Bioinformatics*, vol. 28, no. 2, pp. 184–190, 2012, doi: 10.1093/bioinformatics/btr638.
[38] I. Walsh, D. Baù, A. J. Martin, C. Mooney, A. Vullo, and G. Pollastri, “Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks,” *BMC Struct. Biol.*, vol. 9, pp. 1–20, 2009, doi: 10.1186/1472-6807-9-5.
[39] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, “Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model,” *PLoS Comput. Biol.*, vol. 13, no. 1, pp. 1–34, Jan. 2017, doi: 10.1371/journal.pcbi.1005324.
[40] P. J. Kundrotas and E. G. Alexov, “Predicting residue contacts using pragmatic correlated mutations method: Reducing the false positives,” *BMC Bioinformatics*, vol. 7, pp. 1–9, 2006, doi: 10.1186/1471-2105-7-503.
[41] W. Chen, J. Sun, and C. Gao, “Improving Residue-Residue Contacts Prediction from Protein Sequences Using RNN-Based LSTM Network,” *Proc. - Int. Conf. Mach. Learn. Cybern.*, vol. 2019-July, pp. 1–7, 2019, doi: 10.1109/ICMLC48188.2019.8949207.
[42] W. Ding, J. Xie, D. Dai, H. Zhang, H. Xie, and W. Zhang, “CNNcon: Improved Protein Contact Maps Prediction Using Cascaded Neural Networks,” *PLoS One*, vol. 8, no. 4, 2013, doi: 10.1371/journal.pone.0061533.
[43] P. Björkholm, P. Daniluk, A. Kryshtafovych, K. Fidelis, R. Andersson, and T. R. Hvidsten, “Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue-residue contacts,” *Bioinformatics*, vol. 25, no. 10, pp. 1264–1270, 2009, doi:

10.1093/bioinformatics/btp149.
[44] W. H. Chan and M. S. Mohamad, “Prediction of protein residue contact using support vector machine,” *Commun. Comput. Inf. Sci.*, vol. 295 CCIS, pp. 323–332, 2012, doi: 10.1007/978-3-642-32826-8_33.
[45] P. Chen and J. Li, “Prediction of protein long-range contacts using an ensemble of genetic algorithm classifiers with sequence profile centers,” *BMC Struct. Biol.*, vol. 10, no. SUPPL. 1, pp. 1–13, 2010, doi: 10.1186/1472-6807-10-S1-S2.
[46] Y. Li, Y. Fang, and J. Fang, “Predicting residue-residue contacts using random forest models,” *Bioinformatics*, vol. 27, no. 24, pp. 3379–3384, 2011, doi: 10.1093/bioinformatics/btr579.
[47] S. M. Kandathil, J. G. Greener, and D. T. Jones, “Recent developments in deep learning applied to protein structure prediction,” *Proteins Struct. Funct. Bioinforma.*, vol. 87, no. 12, pp. 1179–1189, 2019, doi: 10.1002/prot.25824.
[48] Y. Shao and C. Bystroff, “Predicting interresidue contacts using templates and pathways,” *Proteins Struct. Funct. Genet.*, vol. 53, no. S6, pp. 497–502, 2003, doi: 10.1002/prot.10539.
[49] M. Schneider and O. Brock, “Combining physicochemical and evolutionary information for protein contact prediction,” *PLoS One*, vol. 9, no. 10, 2014, doi: 10.1371/journal.pone.0108438.
[50] Y. Li, C. Zhang, E. W. Bell, D. J. Yu, and Y. Zhang, “Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13,” *Proteins Struct. Funct. Bioinforma.*, vol. 87, no. 12, pp. 1082–1091, 2019, doi: 10.1002/prot.25798.
[51] A. W. Senior *et al.*, “Improved protein structure prediction using potentials from deep learning,” *Nature*, vol. 577, no. 7792, pp. 706–710, 2020, doi: 10.1038/s41586-019-1923-7.
[52] S. M. Kandathil, J. G. Greener, and D. T. Jones, “Prediction of interresidue contacts with DeepMetaPSICOV in CASP13,” *Proteins Struct. Funct. Bioinforma.*, vol. 87, no. 12, pp. 1092–1099, 2019, doi: 10.1002/prot.25779.
[53] D. T. Jones and S. M. Kandathil, “High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features,” *Bioinformatics*, vol. 34, no. 19, pp. 3308–3315, 2018, doi: 10.1093/bioinformatics/bty341.
[54] B. Adhikari, J. Hou, and J. Cheng, “DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks,” *Bioinformatics*, vol. 34, no. 9, pp. 1466–1472, 2018, doi: 10.1093/bioinformatics/btx781.
[55] Y. Liu, P. Palmedo, Q. Ye, B. Berger, and J. Peng, “Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks,” *Cell Syst.*, vol. 6, no. 1, pp. 65-74.e3, 2018, doi: 10.1016/j.cels.2017.11.014.
[56] V. Ruiz-Serra, C. Pontes, E. Milanetti, A. Kryshtafovych, R. Lepore, and A. Valencia, “Assessing the accuracy of contact and distance predictions in CASP14,” *Proteins Struct. Funct. Bioinforma.*, vol. 89, no. 12, pp. 1888–1900, 2021, doi: 10.1002/prot.26248.
[57] H. Yang, M. Wang, Z. Yu, X. M. Zhao, and A. Li, “GANcon: Protein Contact

Map Prediction with Deep Generative Adversarial Network," *IEEE Access*, vol. 8, pp. 80899–80907, 2020, doi: 10.1109/ACCESS.2020.2991605.

[58] J. Liu and X. Gong, "Attention mechanism enhanced LSTM with residual architecture and its application for protein-protein interaction residue pairs prediction," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–11, 2019, doi: 10.1186/s12859-019-3199-1.

[59] Z. Li, S. Wang, Y. Yu, and J. Xu, "Predicting membrane protein contacts from non-membrane proteins by deep transfer learning," 2017, [Online]. Available: http://arxiv.org/abs/1704.07207.

[60] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction," *J. Appl. Sci. Technol. Trends*, vol. 1, no. 2, pp. 56–70, 2020, doi: 10.38094/jastt1224.

[61] Z. Chen *et al.*, "Feature selection may improve deep neural networks for the bioinformatics problems," *Bioinformatics*, vol. 36, no. 5, pp. 1542–1552, 2020, doi: 10.1093/bioinformatics/btz763.

[62] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[63] N. El Aboudi and L. Benhlima, "Review on wrapper feature selection approaches," *Proc. - 2016 Int. Conf. Eng. MIS, ICEMIS 2016*, 2016, doi: 10.1109/ICEMIS.2016.7745366.

[64] K. L. Chiew, C. L. Tan, K. S. Wong, K. S. C. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," *Inf. Sci. (Ny).*, vol. 484, pp. 153–166, 2019, doi: 10.1016/j.ins.2019.01.064.

[65] A. Gracia, S. González, V. Robles, and E. Menasalvas, "A methodology to compare Dimensionality Reduction algorithms in terms of loss of quality," *Inf. Sci. (Ny).*, vol. 270, pp. 1–27, 2014, doi: 10.1016/j.ins.2014.02.068.

[66] S. Wold, K. Esbensen, and P. Geladi, "Principal Component Analysis," *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1–3, pp. 37–52, 1987, [Online]. Available: http://files.isec.pt/DOCUMENTOS/SERVICOS/BIBLIO/Documentos de acesso remoto/Principal components analysis.pdf.

[67] A. R. Tagilayev, *Modern Multidimensional Scaling*, vol. 36, no. 15. New York, NY: Springer New York, 2005.

[68] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: Applications to image and text data," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01*, 2001, pp. 245–250, doi: 10.1145/502512.502546.

[69] T. F. Abidin, B. Yusuf, and M. Umran, "Singular Value Decomposition for dimensionality reduction in unsupervised text learning problems," *ICETC 2010 - 2010 2nd Int. Conf. Educ. Technol. Comput.*, vol. 4, pp. 422–426, 2010, doi: 10.1109/ICETC.2010.5529649.

[70] J. Wang and C. I. Chang, "Independent component analysis-based

dimensionality reduction with applications in hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1586–1600, 2006, doi: 10.1109/TGRS.2005.863297.

[71] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," *AI Commun.*, vol. 30, no. 2, pp. 169–190, May 2017, doi: 10.3233/AIC-170729.

[72] B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998, doi: 10.1162/089976698300017467.

[73] J. B. Tenenbaum, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science (80-.).*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000, doi: 10.1126/science.290.5500.2319.

[74] S. T. Roweis, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science (80-.).*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000, doi: 10.1126/science.290.5500.2323.

[75] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Adv. Neural Inf. Process. Syst.*, 2002, doi: 10.7551/mitpress/1120.003.0080.

[76] P. Demartines and J. Hérault, "Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets," *IEEE Trans. Neural Networks*, vol. 8, no. 1, pp. 148–154, 1997, doi: 10.1109/72.554199.

[77] P. Campoy, "Dimensionality Reduction by self organizing maps that preserve distances in output space," *Proc. Int. Jt. Conf. Neural Networks*, no. July 2009, pp. 432–438, 2009, doi: 10.1109/IJCNN.2009.5179009.

[78] J. A. Lee and M. Verleysen, "Nonlinear projection with the isotop method," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2415 LNCS, no. August, pp. 933–938, 2002, doi: 10.1007/3-540-46084-5_151.

[79] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, no. November, pp. 232–242, 2016, doi: 10.1016/j.neucom.2015.08.104.

[80] H. Fukuda and K. Tomii, "DeepECA: An end-to-end learning framework for protein contact prediction from a multiple sequence alignment," *BMC Bioinformatics*, vol. 21, no. 1, pp. 2–4, 2020, doi: 10.1186/s12859-019-3190-x.

[81] C. Zhang, W. Zheng, S. M. Mortuza, Y. Li, Y. Zhang, and A. Valencia, "DeepMSA: Constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins," *Bioinformatics*, vol. 36, no. 7, pp. 2105–2112, 2020, doi: 10.1093/bioinformatics/btz863.

[82] T. Du, L. Liao, C. H. Wu, and B. Sun, "Prediction of residue-residue contact matrix for protein-protein interaction with Fisher score features and deep learning," *Methods*, vol. 110, pp. 97–105, 2016, doi: 10.1016/j.ymeth.2016.06.001.

[83] I. Triguero, S. Del Río, V. López, J. Bacardit, J. M. Benítez, and F. Herrera, "ROSEFW-RF: The winner algorithm for the ECBDL'14 big data competition:

An extremely imbalanced big data bioinformatics problem," *Knowledge-Based Syst.*, vol. 87, pp. 69–79, 2015, doi: 10.1016/j.knosys.2015.05.027.

[84] T. Hasanin, T. M. Khoshgoftaar, J. Leevy, and N. Seliya, "Investigating random undersampling and feature selection on bioinformatics big data," *Proc. - 5th IEEE Int. Conf. Big Data Serv. Appl. BigDataService 2019, Work. Big Data Water Resour. Environ. Hydraul. Eng. Work. Medical, Heal. Using Big Data Technol.*, pp. 346–356, 2019, doi: 10.1109/BigDataService.2019.00063.

[85] Z. Li, Y. Lin, A. Elofsson, and Y. Yao, "Protein Contact Map Prediction Based on ResNet and DenseNet," *Biomed Res. Int.*, vol. 2020, pp. 1–12, Apr. 2020, doi: 10.1155/2020/7584968.

[86] I. Erill, "Information theory and biological sequences: insights from an evolutionary perspective," *Inf. Theory New Res. New York Nov. Sci. Publ.*, pp. 1–28, 2012.

[87] Y. Ma, R. Liu, H. Lv, J. Han, D. Zhong, and X. Zhang, "A computational method for prediction of matrix proteins in endogenous retroviruses," *PLoS One*, vol. 12, no. 5, p. e0176909, May 2017, doi: 10.1371/journal.pone.0176909.

[88] W. Pirovano and J. Heringa, "Protein secondary structure prediction," *Data Min. Tech. Life Sci.*, pp. 327–348, 2010.

[89] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, Dec. 1983, doi: 10.1002/bip.360221211.

[90] B. Lee and F. M. Richards, "The interpretation of protein structures: estimation of static accessibility," *J. Mol. Biol.*, vol. 55, no. 3, pp. 379--IN4, 1971.

[91] D. T. H. Chang, H. Y. Huang, Y. T. Syu, and C. P. Wu, "Real value prediction of protein solvent accessibility using enhanced PSSM features," *BMC Bioinformatics*, vol. 9, no. SUPPL. 12, pp. 1–12, 2008, doi: 10.1186/1471-2105-9-S12-S12.

[92] S. Seemayer, M. Gruber, and J. Söding, "CCMpred - Fast and precise prediction of protein residue-residue contacts from correlated mutations," *Bioinformatics*, vol. 30, no. 21, pp. 3128–3130, 2014, doi: 10.1093/bioinformatics/btu500.

[93] S. D. Dunn, L. M. Wahl, and G. B. Gloor, "Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction," *Bioinformatics*, vol. 24, no. 3, pp. 333–340, 2008, doi: 10.1093/bioinformatics/btm604.

[94] M. R. Betancourt and D. Thirumalai, "Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes," *Protein Sci.*, vol. 8, no. 2, pp. 361–369, 2008, doi: 10.1110/ps.8.2.361.

[95] A. Arsenov, I. Ruban, K. Smelyakov, and A. Chupryna, "Evolution of convolutional neural network architecture in image classification problems," *CEUR Workshop Proc.*, vol. 2318, pp. 35–45, 2018.

[96] R. Grosse, "Lecture 15: Exploding and vanishing gradients," *Univ. Toronto Comput. Sci.*, 2017.

[97] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” 2010.
[98] F. Anowar, S. Sadaoui, and B. Selim, “Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE),” *Comput. Sci. Rev.*, vol. 40, p. 100378, 2021, doi: 10.1016/j.cosrev.2021.100378.
[99] V. S. Sumithra and S. Surendran, “A Review of Various Linear and Non Linear Dimensionality Reduction Techniques,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 3, pp. 2354–2360, 2015, [Online]. Available: www.ijcsit.com.
[100] A. S. Nsang and A. Ralescu, “A review of dimensionality reduction methods and their applications,” *Proc. 20th MAICS 2009 - Midwest Artif. Intell. Cogn. Sci. Conf.*, no. January 2009, pp. 118–123, 2009.
[101] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, “Lessons from applying the systematic literature review process within the software engineering domain,” *J. Syst. Softw.*, vol. 80, no. 4, pp. 571–583, 2007, doi: 10.1016/j.jss.2006.07.009.
[102] J. Biolchini, P. G. Mian, A. C. C. Natali, and G. H. Travassos, “Systematic review in software engineering,” *Syst. Eng. Comput. Sci. Dep. COPPE/UFRJ, Tech. Rep. ES*, vol. 679, no. 05, p. 45, 2005.
[103] D. M. Thomas and S. Mathur, “Data analysis by web scraping using python,” in *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2019, pp. 450–454.
[104] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks, “Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing,” *Cell*, vol. 149, no. 7, pp. 1607–1621, Jun. 2012, doi: 10.1016/j.cell.2012.04.012.
[105] N. V. Grishin, “Membrane protein structure predictions for exploration,” *Cell*, vol. 149, no. 7, pp. 1424–1425, 2012, doi: 10.1016/j.cell.2012.06.004.
[106] F. Simkovic, J. M. H. Thomas, and D. J. Rigden, “ConKit: A python interface to contact predictions,” *Bioinformatics*, vol. 33, no. 14, pp. 2209–2211, 2017, doi: 10.1093/bioinformatics/btx148.
[107] J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction*, 1st ed. New York, NY: Springer New York, 2007.
[108] S. Tanwar, T. Ramani, and S. Tyagi, “Dimensionality Reduction Using PCA and SVD in Big Data: A Comparative Case Study,” in *Future Internet Technologies and Trends*, 2018, pp. 116–125.
[109] A. Hyvärinen, J. Hurri, and P. O. Hoyer, “Principal Components and Whitening,” in *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*, London: Springer London, 2009, pp. 93–130.
[110] J. Shlens, “A Tutorial on Principal Component Analysis,” 2014, [Online]. Available: http://arxiv.org/abs/1404.1100.
[111] I. T. Jollife and J. Cadima, “Principal component analysis: A review and recent developments,” *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, 2016, doi: 10.1098/rsta.2015.0202.

[112] M. Nabil, “Random Projection and Its Applications,” 2017, [Online]. Available: http://arxiv.org/abs/1710.03163.
[113] W. B. Johnson, “Extensions of Lipschitz mappings into a Hilbert space,” *Contemp. Math.*, vol. 26, pp. 189–206, 1984.
[114] K. K. Vu, “Random projection for high-dimensional optimization,” Université Paris-Saclay (ComUE), 2016.
[115] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of massive data sets*. Cambridge university press, 2020.
[116] R. Garcia-Dias and A. Mechelli, “Machine Learning: Methods and Applications to Brain Disorders,” *Mach. Learn. Methods Appl. to Brain Disord.*, p. 193, 2019.
[117] H. Nugroho, M. Susanty, A. Irawan, M. Koyimatu, and A. Yunita, “Fully Convolutional Variational Autoencoder For Feature Extraction Of Fire Detection System,” *J. Ilmu Komput. dan Inf.*, vol. 13, no. 1, p. 9, 2020, doi: 10.21609/jiki.v13i1.761.
[118] C. Der Fuh and I. Hu, “Bayesian stochastic estimation of the maximum of a regression function,” *Random Walk, Seq. Anal. Relat. Top. A Festschrift Honor Yuan-Shih Chow*, pp. 269–280, 2006, doi: 10.1142/9789812772558_0018.
[119] J. Alcalá-Fdez *et al.*, “Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework.,” *J. Mult. Log. \& Soft Comput.*, vol. 17, 2011.
[120] M. Friedman, “The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance,” *J. Am. Stat. Assoc.*, vol. 32, no. 200, pp. 675–701, Dec. 1937, doi: 10.1080/01621459.1937.10503522.
[121] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.
[122] A. Tharwat, “Classification assessment methods,” *Appl. Comput. Informatics*, 2018, doi: 10.1016/j.aci.2018.08.003.
[123] R. Eisinga, T. Heskes, B. Pelzer, and M. Te Grotenhuis, “Exact p -values for pairwise comparison of Friedman rank sums , with application to comparing classifiers,” *BMC Bioinformatics*, pp. 1–18, 2017, doi: 10.1186/s12859-017-1486-2.
[124] M. Gao, H. Zhou, and J. Skolnick, “DESTINI: A deep-learning approach to contact-driven protein structure prediction,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–13, 2019, doi: 10.1038/s41598-019-40314-1.
[125] Y. Li, J. Hu, C. Zhang, D. J. Yu, and Y. Zhang, “ResPRE: High-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks,” *Bioinformatics*, vol. 35, no. 22, pp. 4647–4655, 2019, doi: 10.1093/bioinformatics/btz291.
[126] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, “Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization,” *J. Mach. Learn. Res.*, vol. 18, no. 185, pp. 1–52, 2018, [Online]. Available: http://jmlr.org/papers/v18/16-558.html.
[127] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012.