Predicción de compuestos naturales inhibidores de la proteasa principal M^{pro} de los SARS-CoV usando modelos *Deep Learning*

Juan Romera de los Santos
Universidade da Coruña
A Coruña, España
juan.romera@udc.es

Cristian Robert Munteanu
RNASA-IMEDIR, CITIC, INIBIC
Universidade da Coruña
A Coruña, España
c.munteanu@udc.es

***Resumen*—Un nuevo coronavirus denominado SARS-CoV-2 ha alcanzado el estado de pandemia obligando a tomar medidas de seguridad que eviten su rápida expansión. Este patógeno causa síntomas parecidos a los de la neumonía, provocando en algunos pacientes síntomas más graves, llegando incluso en algunos casos a la muerte. Por estos motivos es necesario un tratamiento para los enfermos por COVID-19, la enfermedad que causa este virus. Una de las principales dianas para inhibir su desarrollo en el cuerpo humano es la proteasa M^{pro}, la cual presenta una alta similitud con la proteasa del SARS-CoV. Para reducir los costes y el tiempo de ensayos clínicos y de laboratorio en la búsqueda de un tratamiento efectivo se han empleado técnicas de cribado virtual, como los modelos de relación cuantitativa estructura-actividad. En el presente trabajo se desarrollaron este tipo de modelos empleando para ello algoritmos de *Deep Learning*. Para su entrenamiento se usaron datos de trabajos relacionados, intentando emplear un conjunto de datos con el mayor número de muestras posible de inhibidores de la proteasa M^{pro} de SARS-CoV. El mejor de los modelos presentó un 0.774 de AUROC. Con este modelo se predijeron posibles antivirales de una base de datos de compuestos naturales, hallando un gran número de ellos. Entre estos compuestos se encontraron algunos mencionados en trabajos anteriores.**

***Index Terms*—COVID-19, coronavirus, M^{pro}, *Deep Learning*, QSAR, SARS-CoV-2, compuestos naturales.**

I. Introducción

En diciembre de 2019 se detectó el primer caso de un nuevo coronavirus en la ciudad china de Wuhan [1] [2]. Debido a su rápida transmisión la organización mundial de la salud declaró unos meses más tarde la situación de emergencia global [3]. Este nuevo coronavirus se ha llamado coronavirus del síndrome respiratorio agudo grave dos (SARS-CoV-2 por sus siglas en inglés) y se transmite al estar en contacto estrecho con alguna persona que padezca la enfermedad. Entre los síntomas que produce se encuentran: fiebre, tos, pérdida del sentido del gusto u olfato, problemas al respirar y otro tipo de complicaciones [4] [5].

La familia *Coronaviridae* es responsable de causar enfermedades respiratorias con una sintomatología parecida a la neumonía, además de enfermedades gastrointestinales y neurológicas [2] [6]. El SARS-CoV-2 pertenece a esta familia, al igual que el SARS-CoV y el coronavirus del síndrome respiratorio de oriente medio (MERS-CoV por sus siglas en inglés). El SARS-CoV-2 ha demostrado tener una mayor capacidad de transmisión en comparación al SARS-CoV y al MERS-CoV [3]. También se ha comprobado que este nuevo virus comparte el 80 % de su genoma con otros virus tipo SARS [7]. Además, la evolución de los coronavirus ha demostrado que este no es un patógeno estable y que puede adaptarse a humanos llegando a ser más virulento [4] [8].

Los coronavirus son virus de ARN monocatenario de forma esférica con glicoproteínas proyectadas hacia el exterior [6] [9]. Existen tres dianas principales del SARS-CoV-2: la proteína de espiga, la ARN polimerasa dependiente de ARN y la proteasa M^{pro} [7].

La proteína de espiga media la entrada del virus en las células del huésped interactuando con el receptor ACE 2 [4]. Esta proteína se une al receptor con una mayor afinidad que el SARS-CoV debido a modificaciones en su estructura, lo cual explica en parte la alta tasa de transmisión del virus. La proteína de espiga es la diana preferida para el desarrollo de vacunas [6].

La proteasa M^{pro} se encarga de procesar dos poliproteínas (pp1a y pp1ab) para que puedan ser funcionales [8] [9]. Esta proteína muestra una identidad a nivel de secuencia con respecto a la proteasa de SARS-CoV del 96 % [2] [9] [10]. Esto es muy interesante ya que inhibidores de la enzima de SARS-CoV pueden tener también efecto sobre la enzima de este nuevo virus [11]. Al comparar los sitios activos de ambas enzimas se observa que no hay diferencias en los residuos del sitio activo pero su conformación es diferente, por ejemplo, en la posición 46 en SARS-CoV hay una alanina, mientras que en esta misma posición en SARS-CoV-2 hay una serina. Esto afecta tanto a la entrada como a la unión de ligandos en el sitio activo de la proteasa [4]. Se ha propuesto que una modificación en esta enzima, la variación Thr285Ala, puede contribuir a su alta infectividad [9]. Esta enzima es una diana interesante ya que está altamente conservada entre SARS-CoV y SARS-CoV-2, es imprescindible para la replicación del virus y no se encuentra en humanos [10]. Todo esto hace que M^{pro}, también llamada $3CL^{pro}$, sea la diana más empleada para llevar a cabo cribado virtual de medicamentos.

La ARN polimerasa dependiente de ARN es una enzima

que se encarga de la replicación del genoma viral y es, por lo tanto, una proteína importante en el ciclo de vida del virus. Esta enzima es también una de las dianas preferidas para el uso de medicamentos frente al coronavirus. Por ejemplo, el remdesivir actúa sobre esta diana [4].

Actualmente no existe un tratamiento efectivo contra la enfermedad que causa este nuevo virus, la enfermedad por coronavirus de 2019 (COVID-19 por sus siglas en inglés). Los medicamentos más prometedores en los que se está centrando la organización mundial de la salud son el remdesivir, la cloroquina e hidroxicloroquina, una combinación de lopinavir y ritonavir y una combinación de este último con interferón-β. Aunque recientemente se han detenido los ensayos clínicos con hidroxicloroquina ya que ha demostrado no proporcionar una mejora significativa en el tratamiento de la enfermedad [1]. Algunos suplementos alimenticios como la vitamina C y la D también se encuentran en fase de pruebas para comprobar si reducen la probabilidad de padecer la COVID-19 [4]. Recientemente en otro estudio se ha mostrado que la dexametasona reduce la tasa de mortalidad actuando como un anti-inflamatorio [3]. Otro estudio ha mostrado la alta eficacia de un derivado de productos naturales, la ivermectina, en la inhibición *in vitro* del SARS-CoV-2 [12].

Afortunadamente en el momento de la redacción de este trabajo se han comenzado a distribuir diferentes vacunas para evitar el contagio por coronavirus. La vacuna es la solución ideal, ya que evita padecer la enfermedad, pero aún así es de gran utilidad descubrir medicamentos eficaces frente a este nuevo virus ya que pueden ofrecer información acerca de este tipo de patógenos, además de tener tratamientos en caso de que la vacuna falle o surjan complicaciones de diversa índole.

Los métodos computacionales son muy importantes en la búsqueda de tratamientos para nuevas enfermedades debido a la necesidad de soluciones en poco tiempo. Es por ello que el cribado virtual de compuestos antivirales es de gran relevancia en la búsqueda de posibles medicamentos efectivos. Se ha mostrado que este desarrollo tecnológico permite reducir costes en el descubrimiento de nuevos tratamientos facilitando el proceso convencional [13]. Las técnicas más empleadas para ello son la relación cuantitativa estructura-actividad (QSAR por sus siglas en inglés) y el acoplamiento molecular [5]. La primera de ellas permite obtener información acerca de la actividad de una molécula a partir de su fórmula química, por ejemplo, conocer si una molécula es un posible inhibidor de una determinada enzima. La segunda técnica ofrece información acerca de la fuerza de unión entre dos moléculas, es decir, conocer la afinidad de unión. Cuanto mayor sea esta afinidad, más fuerte será la unión y, por ejemplo, para el caso del inhibidor, menor será la cantidad necesaria para inhibir la enzima y más eficaz será esta inhibición.

Para el desarrollo de modelos QSAR se suelen emplear técnicas de aprendizaje máquina. Estas herramientas permiten la obtención de modelos que aprendan patrones de un conjunto de datos y puedan crear predicciones en base a esos datos. Con técnicas más actuales como las de *Deep Learning* se pueden extraer características de forma automática de los datos [14]. Esto no solo ahorra trabajo con respecto a las técnicas de aprendizaje máquina, sino que se ha encontrado que esta extracción automática de características puede llevar a unos mejores resultados para determinadas tareas [6].

Los objetivos del presente trabajo consistieron en entrenar diferentes algoritmos de *Deep Learning* y emplearlos para predecir posibles antivirales que tengan eficacia frente a la COVID-19. Las moléculas en las que se centró este estudio son productos naturales, ya que algunos de estos compuestos han demostrado poseer propiedades antivirales. Por ejemplo xantoangelol E, rhoifolin, herbacetin y pectolinarin, entre otros, tienen actividad antiviral frente a la enzima M^{pro} de SARS-CoV [1]. Para ello se escogieron modelos de *Deep Learning* provistos por DeepChem [15], empleando datos de inhibidores de la proteasa M^{pro} de SARS-CoV para su entrenamiento. DeepChem es un proyecto que pretende ofrecer herramientas de libre acceso de alta calidad para el descubrimiento de fármacos. Los datos de inhibidores se extrajeron de trabajos anteriores llevando a cabo una labor de integración y limpieza de los mismos. Con el mejor de los modelos se consiguieron predecir más de 7000 compuestos naturales con posible actividad inhibitoria de la enzima M^{pro}, algunos de ellos mencionados en trabajos anteriores. En el resto del artículo se van a tratar los siguientes apartados: una revisión de los trabajos relacionados, la metodología que se siguió para la obtención de los modelos, una presentación de los resultados obtenidos, así como su comparación con trabajos parecidos y las conclusiones extraídas.

II. Trabajo relacionado

El cribado virtual de medicamentos se ha empleado con anterioridad a la situación provocada por la COVID-19, y como era de esperar han surgido diversos estudios aplicados al SARS-CoV-2. En esta sección se van a presentar trabajos relacionados con el cribado virtual de compuestos antivirales para el SARS-CoV-2, especialmente aquellos que se centran en la diana M^{pro}.

Gosh et al. [11] calcularon descriptores moleculares de inhibidores de M^{pro} con herramientas adicionales para entrenar algoritmos de aprendizaje máquina clásicos. Para este entrenamiento emplearon 88 moléculas. Consiguieron mejorar los resultados obtenidos inicialmente con una optimización basada en Monte Carlo, permitiendo escoger los descriptores más relevantes. El mejor de estos modelos mostró una exactitud del 94.44 % sobre una partición de los datos, obteniendo peores resultados para el conjunto de entrenamiento, lo cual podría indicar que la partición no era adecuada. Este modelo lo usaron posteriormente para predecir posibles antivirales en un conjunto de datos de compuestos naturales previamente mencionados como posibles inhibidores de la proteasa principal de SARS-CoV-2. Entre estas moléculas encontraron 13 con actividad inhibitoria frente a la enzima, todas con estructura muy similar. Por lo tanto, aunque consiguieron buenas medidas de rendimiento, parece que se podría estar infravalorando el error del modelo.

Alves et al. [3] calcularon diferentes descriptores moleculares de inhibidores de M^{pro} para entrenar un algoritmo de clasificación de tipo *Random Forest* con un conjunto de datos de 113 muestras. Este algoritmo lo emplearon posteriormente para predecir antivirales de una base de datos de medicamentos. Encontraron que el acoplamiento molecular falló en la discriminación de compuestos activos e inactivos. Este fallo de los estudios de acoplamiento molecular fue apoyado también por otros autores [1]. Los modelos conseguidos por estos investigadores alcanzaron el 71-83 % de exactitud, un 55-72 % de sensibilidad y 72-100 % de precisión.

Tejera et al. [16] construyeron un modelo QSAR para buscar antivirales de la proteasa M^{pro} empleando para ello datos de 229 moléculas, inhibidoras y no inhibidoras de la proteasa de SARS-CoV. Además llevaron a cabo un estudio de acoplamiento molecular con las moléculas más prometedoras que sirvió además como validación del modelo. Con esta validación comprobaron que el modelo se comportaba de forma adecuada, prediciendo los posibles antivirales. El modelo que emplearon es *GraphConvModel* de DeepChem, utilizando una sola partición aleatoria estratificada de los datos y obteniendo un 0.914 de área bajo la curva característica operativa del receptor (AUROC por sus siglas en inglés), 0.83 de precisión y 0.841 de exactitud. Por lo que el error real del modelo podría ser superior al observado.

Kumar y Roy [5] emplearon un modelo basado en regresión lineal múltiple para encontrar inhibidores de la proteasa M^{pro}. Este modelo fue entrenado con descriptores moleculares en dos dimensiones de 69 moléculas que inhiben la proteasa M^{pro} de SARS-CoV. También llevaron a cabo un estudio de acoplamiento molecular para conocer las interacciones entre las moléculas y la proteasa y validar sus resultados, los cuales encontraron que eran adecuados. Además emplearon el modelo entrenado para la predicción de moléculas en dos bases de datos de antivirales.

Tang et al. [10] emplearon inhibidores conocidos de la proteasa M^{pro} de SARS-CoV para entrenar un algoritmo de aprendizaje por refuerzo. Este algoritmo permite la creación virtual de moléculas que pueden tener acción antiviral frente al coronavirus. Con la ayuda de este modelo consiguieron predecir 47 moléculas, las cuales se sometieron también a un análisis de acoplamiento molecular. Este trabajo se centra en la generación de nuevos compuestos químicos frente al SARS-CoV-2.

Abdel-Basset et al. [14] desarrollaron un entorno de trabajo basado en *Deep Learning*. Este entorno está formado por cuatro bloques: el primero permite aprender la estructura de la proteína usando *Dense Net*, el segundo se encarga de aprender la topología de los medicamentos gracias a un grafo, el tercer módulo se encarga de aprender características de los medicamentos gracias a una estructura *ConvLSTM* y el último módulo calcula la afinidad entre las estructuras previamente caracterizadas. El modelo desarrollado presentaba un error cuadrático medio de 0.195 y 0.111 con los dos conjuntos de datos que lo probaron. Con este modelo predijeron medicamentos disponibles comercialmente que pudieran tener actividad frente al SARS-CoV-2. Se centraron en diversas dianas del virus, entre ellas la proteasa M^{pro} y la ARN polimerasa dependiente de ARN. Encontraron que algunos de los medicamentos predichos como antivirales ya estaban en fase de pruebas clínicas para comprobar su eficacia.

Ton et al. [17] emplearon modelos de *Deep Learning* de QSAR para desarrollar un sistema capaz de predecir el acoplamiento molecular de fármacos con la proteasa principal de SARS-CoV-2. El modelo se entrenó empleando datos de inhibidores conocidos de la proteasa de SARS-CoV. El principal aporte de estos autores fue el aumento en la velocidad de este tipo de ensayos, ya que los estudios de acoplamiento molecular son computacionalmente costosos. Con el modelo desarrollado predijeron inhibidores de SARS-CoV-2 en un conjunto de datos de compuestos químicos disponibles en el mercado.

Cozac et al. [13] entrenaron algoritmos de aprendizaje máquina clásicos, basados en grafos como los que se usan en este trabajo y agrupamientos de modelos para predecir inhibidores de la ARN polimerasa dependiente de ARN. Para ello emplearon datos de inhibidores de esta enzima de los virus hepatitis C, polio, dengue e influenza. De esta forma consiguieron predecir algunos de los compuestos que se encuentran en fase clínica como el remdesivir. Además llevaron a cabo un estudio de acoplamiento molecular de los compuestos más prometedores. Uno de los modelos que mostró mejores resultados para esta tarea fue el *GraphConvModel* de DeepChem.

La mayoría de los trabajos mencionados anteriormente emplearon un conjunto de datos bastante reducido. En este trabajo se integraron datos empleados con anterioridad para conseguir un conjunto de datos lo más extenso posible, dotando así a los algoritmos de una mayor capacidad de generalización. Además, se emplearon algoritmos de *Deep Learning* que aún no han sido probados por otros autores para el cribado virtual de antivirales en el SARS-CoV-2 para intentar obtener medidas de rendimiento superiores a las del estado del arte.

III. Materiales y métodos

Para predecir los compuestos naturales inhibidores de M^{pro} fue necesario un conjunto de datos lo más grande posible y de alta calidad de inhibidores. En este caso se emplearon los de la proteasa M^{pro} de SARS-CoV, ya que como se ha mencionado presenta un alto parecido con la de SARS-CoV-2. También fue necesaria una estimación del error de los modelos entrenados para conocer su rendimiento en esta tarea. El problema se modeló como uno de clasificación compuesto por dos clases: si el compuesto es capaz de inhibir la enzima o no. Con el mejor modelo se predijeron moléculas de una base de datos de compuestos naturales y se realizó una comparación de las moléculas predichas con un trabajo anterior.

El código y los datos están disponibles en la siguiente dirección: https://github.com/juanromeradelossantos/SARS-CoV-antivirals-DeepChem.

III-A. Recolección de datos

Los datos para el entrenamiento de los algoritmos tienen el formato especificación de introducción lineal molecular

simplificada (SMILES por sus siglas en inglés). Para establecer las dos clases de compuestos se emplearon los valores de concentración inhibitoria semimáxima (IC50). Un compuesto fue etiquetado como inhibidor de la proteasa si tiene una IC50 inferior a 10 μmolar y como no inhibidor en caso contrario. El motivo de emplear este umbral es para reconocer solamente los inhibidores más potentes de la enzima. Estos datos se extrajeron de trabajos realizados con anterioridad [10] [11] [16]. Para poder fusionar los datos se convirtieron los SMILES a SMILES canónicos, en caso de que no tuvieran este formato. Este tipo de representación es única para cada molécula, por lo tanto así se evitan duplicados en los datos. Se tuvieron que eliminar todas la entradas que no contaban con la IC50 ya que es el valor de referencia para establecer las clases. Tras este proceso de integración y limpieza de datos se obtuvo un conjunto formado por 333 muestras, 121 etiquetadas como inhibidores de la proteasa M^{pro} y 212 como no inhibidores.

Como el conjunto de datos seguía siendo reducido se probó a aumentarlo incorporando datos sintéticos generados a partir de los originales. Para ello se aprovechó la idea de que cada molécula puede tener más de una representación SMILES [18]. Primero se convirtieron los SMILES canónicos del paso anterior a moléculas y posteriormente se transformaron esas moléculas en SMILES aleatorios. Se generaron diez SMILES para cada molécula, exceptuando once de ellas que mostraron errores, y se juntaron con los SMILES canónicos iniciales. Tras este aumento de datos se consiguieron obtener 3542 muestras, siendo 1243 inhibidores y 2299 no inhibidores.

Para la predicción de compuestos antivirales se empleó una base de datos de 401624 compuestos naturales [19]. Estos compuestos naturales pueden definirse como productos químicos producidos por organismos vivos. La base de datos empleada era la más grande de libre acceso en el momento de desarrollar este trabajo.

III-B. Modelos entrenados

Los modelos se escogieron de la API de DeepChem, un proyecto que pretende proporcionar modelos de *Deep Learning* para el descubrimiento de nuevos compuestos químicos, entre otras tareas. Se empleó la versión estable ejecutada en *Google Colab* con aceleración GPU. Se probó a modificar diferentes hiperparámetros de cada uno de los modelos, pero se encontró que funcionaban mejor con los parámetros por defecto a excepción del tamaño de los lotes, el porcentaje de neuronas que se desactivan aleatoriamente en cada capa para evitar el sobreajuste y el número de veces que se pasa el conjunto de datos a la red neuronal para su entrenamiento. Los modelos escogidos fueron: *GraphConvModel*, *WeaveModel*, *DAGModel*, *MPNNModel* y *ChemCeption*.

Los cuatro primeros se basan en grafos convolucionales. Este tipo de modelos emplean descriptores atómicos para calcular descriptores moleculares. Están basados en redes neuronales convolucionales cuya entrada es un grafo que representa la molécula, siendo los nodos los átomos y las aristas los enlaces. Al final se emplean una o varias capas completamente conectadas de neuronas para la clasificación. Se ha demostrado que estos descriptores moleculares pueden prestar un mayor rendimiento que los calculados con herramientas adicionales para determinadas tareas [20]. El primer modelo es una implementación de grafos convolucionales, el segundo son grafos convolucionales tipo *Weave* que calculan características de los enlaces de forma explícita, el tercero emplea una representación de la molécula como grafos acíclicos dirigidos en vez de grafos no dirigidos como los anteriores, el cuarto trata las operaciones de convolución como un problema más general de paso de mensajes entre los nodos del grafo. El quinto modelo emplea una representación de las moléculas como imágenes y se emplea una red neuronal convolucional para el tratamiento de dichas imágenes.

Antes de entrenar cada modelo fue necesario un paso de transformación de los datos de partida. Esta transformación consistió en convertir los SMILES a grafos moleculares específicos para cada modelo, o imágenes en el caso del modelo *ChemCeption*.

III-C. Estimación del error

Para la estimación del error de los modelos entrenados se emplearon como medidas de rendimiento el AUROC, la exactitud, la sensibilidad y la precisión. Para el entrenamiento de los modelos se escogió una validación cruzada con tres paquetes, siendo uno de ellos el conjunto de test y el resto el conjunto de entrenamiento. Como el número de muestras de cada clase no es parecido, se escogió un reparto de los datos que asegurara la misma proporción de muestras de cada clase en cada conjunto de datos. El motivo de no emplear una validación cruzada con más paquetes es debido al elevado tiempo de entrenamiento de los modelos.

Para el cálculo de las medidas de rendimiento es necesario obtener las tasas de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Para ello fue necesario escoger un umbral para determinar la clase a la cual pertenecía cada molécula, ya que la salida al realizar predicciones con los modelos era continua. Por lo tanto, para cada clase se obtuvo un valor entre cero y uno indicando la probabilidad de pertenencia a cada clase. El valor escogido para esta umbralización fue de 0.5.

IV. Resultados y discusión

El objetivo principal de este trabajo era encontrar posibles inhibidores de la proteasa M^{pro}, centrándose en compuestos naturales. Para ello se recolectaron datos de inhibidores de la proteasa de SARS-CoV de trabajos anteriores y se emplearon para entrenar diferentes modelos de *Deep Learning*. A continuación se presentan los principales resultados obtenidos, los modelos entrenados y las predicciones obtenidas de posibles antivirales, junto a una discusión de los mismos.

En trabajos previos se ha observado que es común realizar un estudio de acoplamiento molecular junto a los modelos QSAR, sin embargo, existen algunos trabajos recientes que mencionan la baja correlación de este tipo de estudios con la realidad debido a la gran flexibilidad que presenta la enzima en su sitio activo [1] [3]. Es por este motivo que en este trabajo

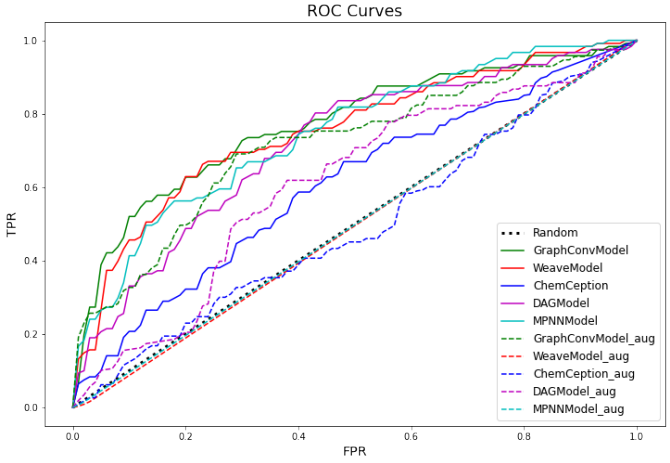

Figura 1. Curvas ROC para cada uno de los modelos entrenados, empleando los datos originales y los aumentados, para el conjunto de test.

no se ha presentado un estudio de acoplamiento molecular a modo de validación de los resultados.

IV-A. Modelos obtenidos

La mayoría de modelos que se obtuvieron mostraron un claro sobreajuste, ya que las medidas obtenidas para el conjunto de entrenamiento eran superiores a las del conjunto de test. En la tabla I pueden verse las medidas que se calcularon para cada modelo y conjunto de datos, tanto para los datos originales como los aumentados y el conjunto de entrenamiento y de test. Aunque para algunos de los modelos se empleó la estrategia de desactivar un porcentaje de neuronas de forma aleatoria con la finalidad de reducir este sobreajuste, no se consiguió reducirlo a valores normales. Este problema puede ser debido a la baja cantidad de datos, sumado a la elevada complejidad de los modelos empleados.

Como podemos ver en la gráfica de la figura 1 el mejor de los modelos fue *GraphConvModel* entrenado con el conjunto de datos originales, confirmado gracias a las medidas de la tabla I. En la gráfica podemos ver una media de las curvas ROC obtenidas para cada iteración de la validación cruzada con todos los modelos, solamente para el conjunto de test. Podemos apreciar que el uso del conjunto de datos aumentado no proporciona una mejoría notable para el problema que se está tratando.

El mejor de los modelos era capaz de distinguir bien las dos clases, ya que presentaba una medida de AUROC del 77.4 %. Reconocer la clase positiva, en este caso los inhibidores de la proteasa M^{pro}, lo conseguía con un porcentaje de aciertos aceptable, aunque ligeramente bajo, como nos indica la medida de sensibilidad. Esto supondría que del total de posibles inhibidores se estarían reconociendo un 57.0 % aproximadamente. Las moléculas clasificadas como inhibidores se obtenían también con cierta seguridad, aunque también reducida, como nos indica la precisión. Esto supondría que de todos los inhibidores predichos el 64.7 % son efectivamente inhibidores, por lo que se estarían cometiendo algunos errores.

En otro estudio anterior [3] obtuvieron una precisión en sus modelos de entre el 72-100 % y una exactitud del 71-83 %. En otro trabajo [11] se consiguieron valores del 100 % de sensibilidad y 94.44 % de exactitud. Aunque el inconveniente de ambos trabajos es que empleaban un conjunto de datos reducido, por lo que sus modelos tienen una menor capacidad de generalización en comparación al presentado en este estudio. Además, las medidas finales obtenidas en el segundo trabajo eran solamente sobre una partición de los datos, la cual parece no ser adecuada ya que se obtienen mejores medidas para el conjunto de test que para el de entrenamiento. Por lo tanto podría existir una infravaloración del error cometido por el modelo. En otro trabajo [16] alcanzaron un AUROC de 0.914 con un modelo basado en grafos convolucionales con un número de muestras mayor, pero estos resultados los obtuvieron con una sola partición aleatoria de los datos. Por lo tanto el error podría ser superior al esperado. En comparación con estos datos las medidas obtenidas con el mejor de los modelos del presente trabajo son inferiores. Aunque cabe destacar que en este caso las medidas proporcionadas se obtienen a partir de más de una partición de los datos, por lo que los valores obtenidos de rendimiento son más cercanos al real.

Este trabajo empleó una mayor cantidad de datos para el entrenamiento de los algoritmos por lo que su capacidad de generalización es mayor que los modelos de los trabajos mencionados anteriormente. El inconveniente es que al existir una diversidad mayor de moléculas era más difícil que los modelos pudieran aprender a reconocer los compuestos de forma adecuada. Este problema debido a la falta de datos se intentó reducir empleando datos creados de forma artificial a partir de los originales, aunque los resultados no mejoraron significativamente. También sería interesante llevar a cabo un análisis de las moléculas que se emplearon para el entrenamiento con la finalidad de identificar si hay alguna que tenga una estructura y propiedades muy diferentes del resto para eliminarla de los datos y de esta forma obtener un modelo más robusto.

IV-B. Predicción de inhibidores

Con el mejor de los modelos se predijeron los posibles antivirales de la base de datos de compuestos naturales y se obtuvieron 7690 moléculas con una probabilidad superior a 0.99 de ser inhibidores de la proteasa M^{pro}. Este número de compuestos sigue siendo bastante elevado por lo que sería interesante tener en cuenta más técnicas con la finalidad de reducir el número de antivirales finales.

Comparando los resultados con un trabajo anterior que también se centró en compuestos naturales [11], dos de los trece compuestos que ellos encontraron como activos también se hallaron en este trabajo como activos (Myricitrin, Oolonghomobisflavan-A), mientras que nueve de ellos el modelo los marcó como inactivos (Rutin, Quercetin 3-vicianoside, Kouitchenside I, Neohesperidin, Baicalin, Cyanidin, Hespiridine, Theasinensin-D, 22-Hydroxyhopan-3-one) y el resto no se encontraron en la base de datos.

Tabla I
MEDIDAS DE RENDIMIENTO OBTENIDAS CON CADA MODELO, EMPLEANDO LOS DATOS ORIGINALES Y AUMENTADOS PARA LOS CONJUNTOS DE ENTRENAMIENTO Y TEST.

Modelo	Conjunto de datos		AUROC	Exactitud	Sensibilidad	Precisión
GraphConvModel	Original	Entrenamiento	0.987 ± 0.003	0.939 ± 0.009	0.880 ± 0.020	0.949 ± 0.019
		Test	0.774 ± 0.049	0.729 ± 0.025	0.570 ± 0.048	0.647 ± 0.043
	Aumentado	Entrenamiento	0.999 ± 0.000	0.998 ± 0.001	0.997 ± 0.002	0.996 ± 0.004
		Test	0.715 ± 0.030	0.664 ± 0.001	0.494 ± 0.136	0.523 ± 0.003
WeaveModel	Original	Entrenamiento	0.973 ± 0.012	0.884 ± 0.023	0.788 ± 0.133	0.897 ± 0.109
		Test	0.752 ± 0.042	0.714 ± 0.044	0.472 ± 0.176	0.664 ± 0.109
	Aumentado	Entrenamiento	0.499 ± 0.000	0.647 ± 0.003	0.000 ± 0.000	0.000 ± 0.000
		Test	0.495 ± 0.006	0.647 ± 0.003	0.000 ± 0.000	0.000 ± 0.000
ChemCeption	Original	Entrenamiento	0.999 ± 0.000	0.990 ± 0.007	0.991 ± 0.011	0.984 ± 0.022
		Test	0.606 ± 0.058	0.612 ± 0.066	0.388 ± 0.035	0.483 ± 0.101
	Aumentado	Entrenamiento	0.999 ± 0.000	0.988 ± 0.003	0.978 ± 0.003	0.988 ± 0.007
		Test	0.495 ± 0.021	0.540 ± 0.025	0.362 ± 0.084	0.350 ± 0.035
DAGModel	Original	Entrenamiento	1.000 ± 0.000	0.998 ± 0.002	1.000 ± 0.000	0.995 ± 0.004
		Test	0.716 ± 0.042	0.669 ± 0.023	0.569 ± 0.028	0.546 ± 0.035
	Aumentado	Entrenamiento	0.999 ± 0.000	0.998 ± 0.002	1.000 ± 0.000	0.994 ± 0.007
		Test	0.608 ± 0.019	0.614 ± 0.021	0.529 ± 0.103	0.457 ± 0.034
MPNNModel	Original	Entrenamiento	0.990 ± 0.006	0.951 ± 0.017	0.933 ± 0.025	0.937 ± 0.029
		Test	0.746 ± 0.023	0.708 ± 0.008	0.554 ± 0.096	0.615 ± 0.029
	Aumentado	Entrenamiento	0.500 ± 0.000	0.649 ± 0.003	0.000 ± 0.000	0.000 ± 0.000
		Test	0.497 ± 0.003	0.649 ± 0.003	0.000 ± 0.000	0.000 ± 0.000

Algunos de esos 7690 compuestos son: Pestalpolyol A y D, Tautomycetin, Hydroxypentafuhalol A Tetradecaacetate, Decafuhalol A Hexacosaacetate, Nonafuhalol A Tricosaacetate, Hexafuhalol A Hexadecaacetate y Deshydroxyoctafuhalol C Eicosaacetate.

Pestalpolyol A y D son compuestos que se pueden extraer de productos de fermentación de *Pestalotiopsis* sp. y han demostrado actividad antitumoral [21]. Tautomycetin puede extraerse de *Streptomyces griseochromogenes* sp. y ha demostrado actividad antibacteriana frente a *Sclerotinia sclerotiorum*, además de actividad antifúngica y cambiar la morfología de células K562 de leucemia humanas [22]. Hydroxypentafuhalol A Tetradecaacetate, Decafuhalol A Hexacosaacetate, Nonafuhalol A Tricosaacetate, Hexafuhalol A Hexadecaacetate y Deshydroxyoctafuhalol C Eicosaacetate pueden extraerse del alga marrón *Carpophyllum angustifolium*. Este alga es rica en compuestos que presentan actividad tóxica, antibiótica, antifúngica y algicida [23].

A la vista de las funciones de algunos de los compuestos predichos como activos, es probable que alguno de ellos pueda presentar actividad inhibitoria de la enzima M^{pro} en la realidad. Por lo tanto, sería interesante intentar reducir el número de antivirales con otras técnicas y, si fuera posible, probar los más prometedores en el laboratorio.

V. CONCLUSIONES

Este trabajo empleó algoritmos de *Deep Learning* e inhibidores de la proteasa M^{pro} de SARS-CoV para predecir compuestos naturales que pudieran tener actividad inhibitoria frente a la proteasa de SARS-CoV-2. Los modelos entrenados mostraron un rendimiento mejorable, teniendo el mejor de ellos un AUROC de 0.774. Este modelo se empleó para predecir posibles antivirales de una base de datos de compuestos naturales. Comparando resultados con un trabajo previo se encontraron varias coincidencias. Como trabajo futuro se podría intentar reducir el número de compuestos activos, ya que este fue bastante grande, y comprobar la eficacia de estos inhibidores finales en pruebas de laboratorio y, si fuera pertinente, en pruebas clínicas.

REFERENCIAS

[1] S. Verma, D. Twilley, T. Esmear, C. B. Oosthuizen, A.-M. Reid, M. Nel, and N. Lall, "Anti-sars-cov natural products with the potential to inhibit sars-cov-2 (covid-19)," *Frontiers in Pharmacology*, vol. 11, p. 1514, 2020.

[2] X. Liu and X.-J. Wang, "Potential inhibitors against 2019-ncov coronavirus m protease from clinically approved medicines," *Journal of Genetics and Genomics*, vol. 47, no. 2, p. 119, 2020.

[3] V. M. Alves, T. Bobrowski, C. C. Melo-Filho, D. Korn, S. Auerbach, C. Schmitt, E. N. Muratov, and A. Tropsha, "Qsar modeling of sars-cov mpro inhibitors identifies sufugolix, cenicriviroc, proglumetacin, and other drugs as candidates for repurposing against sars-cov-2," *Molecular Informatics*, 2020.

[4] D. Kumar, G. Chauhan, S. Kalra, B. Kumar, and M. S. Gill, "A perspective on potential target proteins of covid-19: Comparison with sars-cov for designing new small molecules," *Bioorganic chemistry*, p. 104326, 2020.

[5] V. Kumar and K. Roy, "Development of a simple, interpretable and easily transferable qsar model for quick screening antiviral databases in search of novel 3c-like protease (3clpro) enzyme inhibitors against sars-cov diseases," *SAR and QSAR in Environmental Research*, vol. 31, no. 7, pp. 511–526, 2020.

[6] A. Keshavarzi Arshadi, J. Webb, M. Salem, E. Cruz, S. Calad-Thomson, N. Ghadirian, J. Collins, E. Diez-Cecilia, B. Kelly, H. Goodarzi *et al.*, "Artificial intelligence for covid-19 drug discovery and vaccine development. front," *Artif. Intell*, vol. 3, p. 65, 2020.

[7] M. U. Mirza and M. Froeyen, "Structural elucidation of sars-cov-2 vital proteins: Computational methods reveal potential drug candidates against main protease, nsp12 polymerase and nsp13 helicase," *Journal of Pharmaceutical Analysis*, 2020.

[8] Y. W. Chen, C.-P. B. Yiu, and K.-Y. Wong, "Prediction of the sars-cov-2 (2019-ncov) 3c-like protease (3cl pro) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates," *F1000Research*, vol. 9, 2020.

[9] J. Song, "2019-ncov 3c-like protease carries an activity-enhancing t285/a variation which may contribute to its high infectivity," 2020.

[10] B. Tang, F. He, D. Liu, M. Fang, Z. Wu, and D. Xu, "Ai-aided design of novel targeted covalent inhibitors against sars-cov-2," *bioRxiv*, 2020.

[11] K. Ghosh, S. A. Amin, S. Gayen, and T. Jha, "Chemical-informatics approach to covid-19 drug discovery: Exploration of important fragments and data mining based prediction of some hits from natural origins as main protease (mpro) inhibitors," *Journal of molecular structure*, vol. 1224, p. 129026, 2020.

[12] L. Caly, J. D. Druce, M. G. Catton, D. A. Jans, and K. M. Wagstaff, "The fda-approved drug ivermectin inhibits the replication of sars-cov-2 in vitro," *Antiviral research*, p. 104787, 2020.

[13] R. Cozac, N. Medzhidov, and S. Yuki, "Predicting inhibitors for sars-cov-2 rna-dependent rna polymerase using machine learning and virtual screening," *arXiv preprint arXiv:2006.06523*, 2020.

[14] M. Abdel-Basset, H. Hawash, M. Elhoseny, R. K. Chakrabortty, and M. Ryan, "Deeph-dta: Deep learning for predicting drug-target interactions: A case study of covid-19 drug repurposing," *IEEE Access*, vol. 8, pp. 170 433–170 451, 2020.

[15] B. Ramsundar, P. Eastman, P. Walters, V. Pande, K. Leswing, and Z. Wu, *Deep Learning for the Life Sciences*. O'Reilly Media, 2019, https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.

[16] E. Tejera, C. R. Munteanu, A. López-Cortés, A. Cabrera-Andrade, and Y. Pérez-Castillo, "Drugs repurposing using qsar, docking and molecular dynamics for possible inhibitors of the sars-cov-2 mpro protease," *Molecules*, vol. 25, no. 21, p. 5172, 2020.

[17] A.-T. Ton, F. Gentile, M. Hsing, F. Ban, and A. Cherkasov, "Rapid identification of potential inhibitors of sars-cov-2 main protease by deep docking of 1.3 billion compounds," *Molecular informatics*, 2020.

[18] E. J. Bjerrum, "Smiles enumeration as data augmentation for neural network modeling of molecules," *arXiv preprint arXiv:1703.07076*, 2017.

[19] M. Sorokina and C. Steinbeck, "Review on natural products databases: where to find data in 2020," *Journal of cheminformatics*, vol. 12, pp. 1–51, 2020.

[20] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," *arXiv preprint arXiv:1509.09292*, 2015.

[21] J. Li, J. Xie, Y.-H. Yang, X.-L. Li, Y. Zeng, and P.-J. Zhao, "Pestalpolyols a–d, cytotoxic polyketides from pestalotiopsis sp. cr013," *Planta medica*, vol. 81, no. 14, pp. 1285–1289, 2015.

[22] X.-C. Cheng, M. Ubukata, and K. Isono, "The structure of tautomycetin, a dialkylmaleic anhydride antibiotic," *The Journal of antibiotics*, vol. 43, no. 7, pp. 890–896, 1990.

[23] K.-W. Glombitza and A. Schmidt, "Nonhalogenated and halogenated phlorotannins from the brown alga carpophyllum angustifolium," *Journal of natural products*, vol. 62, no. 9, pp. 1238–1240, 1999.