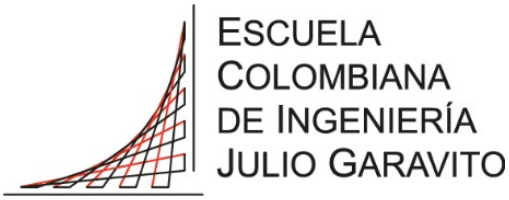
Estudio de modelos de aprendizaje automático para Tuberculosis en el proceso de *Drug Discovery*

Mateo Hueza Echeverri

Trabajo Dirigido

Tutores:
Álvaro David Orjuela Cañón PhD - Andrés Leonardo Jutinico Alarcón PhD



UNIVERSIDAD DEL ROSARIO
ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
PROGRAMA DE INGENIERÍA BIOMÉDICA
BOGOTÁ D.C
2022

Agradecimientos

Agradezco a mis padres Gloria Elsa Echeverri y Carlos Alberto Hueza, quienes me han apoyado y enseñado durante cada etapa, sin ellos esto no sería posible. A mi hermana Valentina por apoyarme incondicionalmente en cada momento. A mis compañeros y amigos por estar a mi lado en esta travesía. Especialmente a Daniela Muñoz, a Manuela Suárez y a Arturo Cordero, por su apoyo incondicional, por ayudarme a crecer como persona, amigo y como profesional. Es con ellos que viví y ahora culmino mi etapa universitaria.

Hago mención al Semillero de Procesamiento de Señales e Imágenes PROMISE y al semillero de Inteligencia Artificial en la Salud Semill-IAS por brindarme conocimientos adicionales en ámbitos inerdisciplinarios.

Por último, deseo agradecer La Universidad del Rosario y la Escuela Colombiana de Ingeniería Julio Garavito por permitirme la oportunidad de ser parte de su comunidad; y de igual manera agradezco a mis profesores por sus enseñanzas a lo largo de esta carrera, especialmente a mi tutor Álvaro Orjuela por su acompañamiento en este proyecto.

Resumen

La tuberculosis (TB) es catalogada como la segunda causa de mortalidad a nivel mundial[1]. A lo largo de los años han surgido variedad de medicamentos y fármacos cuyo objetivo es la eliminación de la especie bacteriana *Mycobaterium tuberculosis*, causante de la enfermedad. Clasificados como primera, segunda y tercera línea, estos grupos de fármacos se encargan de atacar a la bacteria de diferente manera y magnitud según la gravedad de la enfermedad; sin embargo, una problemática que crece de manera preocupante es el surgimiento de cepas que presentan resistencia a uno o varios fármacos existentes. Las resistencias se presentan por diferentes causas donde la principal responde a los deficiencias en tratamientos a la enfermedad, así como por transmisión de cepas resistentes[2].

En este documento se presenta un proyecto de investigación planteado desde el área de la bioinformática, en el que se busca dar una solución a la actual problemática de la drogorresistencia en tres proteínas presentes en cepas resistentes la especie bacteriana *Mycobaterium tuberculosis*. Se busca plantear y comparar modelos de predicción de valores de pIC50 (escala logarítmica del IC50), que hace referencia a la concentración necesaria del fármaco para disminuir la actividad de la proteína en un 50 %. Esta predicción servirá para nuevos fármacos, tomando como punto de partida la estructura molecular de compuestos químicos ya conocidos. Para ello se caracterizan con los datos provenientes de la base ChEMBL[3] y que tengan como proteína objetivo las proteínas *N-Acetiltransferasa* codificada por el gen *eis*, la *ATP sintentasa subunidad c* codificada por el gen *atpE* y por la *Subunidad beta de ARN polimerasa dirigida por ADN* codificada por el gen *rpoB*. Se esocogen estas proteínas dado que en ellas se presenta la resistencia a ciertos fármacos bactericidas de segunda y tercera línea. A cada compuesto se le calculan descriptores relacionados con la ley de lipinski: el peso molecular (MW), número de donadores de enlaces por puentes de hidrógeno (NumHDonnors), número de aceptores de enlaces por puentes de hidrógeno (NumHAcceptors) y el coeficinente de reparto octanol/agua (LogP); de igual manera se calcula una huella que cuenta con 881 descriptores que junto con los cuatro ya mencionados, se toman como la entrada de los modelos de regresión a plantear y los valores de pIC50 conocidos se toman como salida. Lo anterior corresponde al conjunto de entrenamiento con el que se generan diferentes modelos de regresión para esta predicción del pIC50. Para finalizar se comparan las características de funcionamiento de los modelos para así establecer los más adecuados para la problemática.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

INTRODUCCIÓN

La tuberculosis (TB) es una de las enfermedades más antiguas que se conoce en la humanidad, causada por la bacteria *Mycobacterium tuberculosis*, catalogada por la Organización Mundial de la Salud como la segunda causa de muertes en humanos[1]. A lo largo de la historia ha sido tema de discusión y foco central de investigación dadas las consecuencias que ocasiona, así como la cantidad de personas que la han padecido. Ha pasado de ser de ser una enfermedad letal a ser una enfermedad curable; sin embargo, en el año de 1981 a causa de la epidemia de VIH/SIDA, surgió una coinfección con TB, lo que llevo a un rebrote de esta y así mismo a una mutación drogorresistente, es decir, una mutación inmune a los fármacos existentes en la época [1]. Su comportamiento y funcionamiento varía, por ejemplo, el lugar en que se aloja la bacteria, ya que puede ser a nivel pulmonar o extrapulmonar. De la misma manera, los fármacos para el tratamiento se clasifican en primera línea como la isoniazida, la rifampicina, la pirazinamida y la estreptomicina, que poseen diferente intensidad de actividad bactericida. En segunda línea se tienen las quinolonas, capreomicina, protionamida, etionamida, kanamicina y rifabutina las cuales presentan actividad bactericida, mientras que los ácidos paraaminosalicílico, cicloserina, clofasimina y macrólidos poseen actividad bacteriostática. La clasificación en primero o segunda línea de fármacos responde a la eficacia y a la tolerabilidad de cada uno[4]. Sin embargo, en el momento en que las cepas empiezan a ser resistentes a fármacos antituberculosos es cuando se clasifican como drogorresistentes (DR-TB), y en casos que su resistencia llega a ser a varios fármacos son llamadas extensamente drogorresistentes (XDR-TB) o en el caso de no verse afectada por ningún fármaco, totalmente drogorresistentes. (TDR-TB). El problema yace en que en los últimos años estas cepas son más frecuentes y resistentes [5] [6].

En Colombia la presencia de la enfermedad ha ido creciendo. Es por ello por lo que, en 2018 el Ministerio de Salud y Protección, llevó a cabo talleres para brindar información y de igual manera junto con el Instituto de Nacional de Salud y la Organización Panamericana de la Salud se creó la Red Nacional Para la Tuberculosis, la cual se encarga de ayudar al avance en la investigación relacionada con esta enfermedad para adquirir mejor entendimiento [7].
La constante aparición de cepas drogorresistentes es una problemática que va en aumento y que gana importancia en la Red cada vez mayor ya que desde 2001 se encuentran en constante crecimiento el número de casos de personas con TB multi drogorresistente. Entre 2012 y 2013 se caracterizaron cepas en 29 de los 32 departamentos del territorio colombiano [8] y entre los años 2006 y 2016 se presentaron 51 casos de TB extremadamente drogorresistente [2]. Además

cabe resaltar que la población menor a 15 años también se ha visto afectada, teniendo 291 casos entre 2010 y 2015 de TB drogorresistente [9].

A nivel mundial el conocimiento acerca de la TB ha crecido, existen foros, plataformas, y todo tipo de herramientas que con ayuda del internet facilitan conocer más sobre la enfermedad. Existen bases de datos en las que se encuentran caracterizadas diferentes cepas de la bacteria, proteínas específicas, genes y demás datos que pueden tomarse como punto de partida de investigaciones para hacerle frente a la situación.

En la actualidad la problemática recae en el tratamiento de la enfermedad, pues existen varios factores para tener en cuenta, dado que al diferentes cepas dentro de la especie bacteriana *Mycobacterium tuberculosis* y de la misma manera diferentes resistencias, se hace evidente la necesidad de buscar otras metodologías para dar respuesta a la drogorresistencia y se vuelve una prioridad la búsqueda de nuevos fármacos para llevar a cabo del tratamiento partiendo de proteínas específicas relacionadas con la resistencia a compuestos específicos. Un grupo de soluciones son las investigaciones *In Silico*, es decir aquellas que se llevan a cabo de manera computacional[10]. Existen gran variedad de soluciones computacionales en varios campos, que incluyen bioquímica, biología química y estructural, química y bioinformática, química computacional, química física, síntesis orgánica entre otros. [10]. Con estas herramientas se pueden caracterizar cepas resistentes, y plantear fármacos especializados en proteínas específicas[11]. Se pueden escoger proteínas objetivo causantes de drogorresistencia aplicando elementos de la dinámica molecular para caracterizar mutaciones específicas que han llevado a la existencia a dichas variaciones de TB [12]. Una opción por utilizar son los modelos de Aprendizaje automático para la predicción de características de interés relacionadas a la drogorresistencia pues ya se han planteado soluciones para la TB mediante su uso; generalmente modelos para el diagnóstico de la enfermedad[13] [14] [15], así como modelos para predicción de características inhibitorias [16] y de visualización de las proteínas objetivo [17]. Es por esto que, se desea plantear un modelo capaz de predecir el valor de IC50 (concentración para inhibir la actividad de proteínas objetivo al 50 %) mediante el uso de descriptores para caracterizar cada proteína. Estos fundamentan los modelos de predicción en los que la salida será el valor pIC50.

1.1 Objetivos

1.1.1. General

Analizar modelos de aprendizaje automático en un proceso de ***Drug Discovery*** para Tuberculosis Drogorresistente.

1.1.2. Específicos

1. Determinar tres proteínas objetivo asociadas a la drogorresistencia de la bacteria *Mycobaterium Tuberculosis*.

2. Identificar compuestos químicos que tengan como proteína objetivo las tres proteínas asociadas, y generar descriptores para cada uno de ellos.

3. Estudiar tres modelos de aprendizaje automático en el proceso de *Drug Discovery* en el caso específico de la tuberculosis, comparando los resultados y comportamiento de cada modelo.

Capítulo 2

MARCO TEÓRICO

2.1 Tuberculosis y la drogorresistencia

La tuberculosis es una enfermedad generada por la especie bacteriana *Mycobacterium tuberculosis* y representa uno de los más grandes riesgos para la salud humana. Al ser una bacteria gram positiva y además ácido-alcohol resistente, presenta una capa de peptidoglicano la que es responsable del creciente grupo de bacterias resistentes [5], por ende, cada vez se presentan varias cepas con drogorresistencias en proteínas específicas codificadas desde genes específicos, como lo es el caso de los genes eis, rpoB y atpE. Estos hacen parte de un grupo de genes que se llevan estudiando hace más de 15 años.

El gen rpoB es codificante de la proteína "Subunidad beta de ARN polimerasa dirigida por ADN" y es en esta proteína en donde se ha estado generando la resistencia a la rifampicina, que hace de parte de la primera línea de fármacos antituberculosos. Ya que a pesar de no aumentar la sensibilidad de la bacteria al fármaco, sí aumenta por otros mecanismos la afinidad del núcleo a un precursor de la transcripción. Es decir que genera resistencia de manera indirecta activando mecanismos que funcionan en paralelo[18]. Es por ello por lo que de encontrar fármacos que puedan deshabilitarla, ayudaría a disminuir la resistencia a la rifampicina.

El gen atpE por su lado, es codificador de la proteína "ATP sintetasa subunidad c", proteína de función catalizadora en la reacción química en que se produce adenosín trifosfato ATP a partir del adenosín difosfato ADP. Es una proteína que de ser inhabilitada, generaría una pérdida a nivel energético en la bacteria. Es considerara una proteína objetivo importante a la hora de tratar *Mycobaterium tuberculosis* ya que comparte la misma vía que la proteína objetivo de la Isoniazida[19].

Por su parte, el gen eis que codifica la proteína "n-acetiltransferasa" que es una proteína, cuya función es atacar al sistema inmune del hospedador generando inflamaciones, autofagía en macrófagos y demás efecos negativos. Representa uno de los grandes retos del *Drug discovery*, ya que acetila grupos de los antibióticos aminoglucósidos llevando a inhibirlos y por ende generar resistencia[20].

2.2 Descriptores y regla de Lipinski

Los descriptores hacen referencia a propiedades fisicoquímicas calculadas a partir de la estructura tridimensional de un compuesto químico, y son una forma de describir de manera numérica dicho compuesto. En este caso se calculan tomando como punto de partida los datos de *The Chemistry Development Kit*[21] base de datos de compuestos químicos. Con ayuda de los descriptores se empiezan a cimentar las bases para crear modelos de Aprendizaje automático. Con los descriptores se pueden tener en cuenta parámetros relevantes a nivel farmacológico como lo es la Ley de cinco de Lipinski, en la que se estipula que existirá una absorción pobre del fármaco si se cumple más de una de 4 reglas[22]:

1. Existen más de 5 donadores de enlaces de hidrógeno,
2. Existen más de 10 aceptores de enlaces de hidrógeno,
3. El peso molecular es superior a 500 uma
4. El valor del LogP es mayor a 5.

Estos descriptores ayudan a conocer información de la solubilidad, velocidad de difusión, la concentración máxima a nivel gastrointestinal y a su vez la velocidad de eliminación[23]. Se calculan dado que son una manera muy eficiente y utilizada a la hora de caracterizar compuestos químicos en procesos de *Drug Discovery*, así como en proyectos donde se lleva a cabo el análisis de liberación farmacológica[24] [25] [26].

2.3 Aprendizaje automático

El aprendizaje automático, *Machine Learning* en inglés, hace referencia a un tipo de estudio en que se permite a un sistema computacional aprender a realizar una operación o proceso al proporcionarle varios ejemplos de cómo se realiza. Este tipo de estudio permite que el sistema a su vez pueda mejorar gracias a sus conocimientos de aprendizajes pasados. De manera que puede mejorar entre más aprendizajes realice.Se puede presentar en diferentes formas como lo son los modelos predictivos, modelos basados en neuronas, entre otros[27].

Existen varios modelos de aprendizaje automático que sirven para generar regresiones de valores, se destacan los árboles aleatorios (*Random forest*), y el resgresor de k-vecinos cercanos (*k-neighbors*). Esto debido a que son capaces de generalizar y son muy utilizados en el Aprendizaje automático en general[28].

2.3.1. Random forest

Los regresores *Random Forest* hacen referencia a ensambles o combinaciones de árboles de decisión. Dichos árboles son estructuras de clasificación en las que cada característica que se ingresa genera un nodo, o punto de separación. Con ello según una característica específica se pueden generar varias clasificaciones. El *Random Forest* es un tipo de aprendizaje automático supervisado, es decir que se le dice cuáles son los datos y características, y cuál es el valor que se desea predecir. Es un modelo que ayuda a la interpretabilidad, ya que funciona mediante decisiones basadas en preguntas sencillas. Uno de sus hipeparámetros principales es la función que se desea optimizar a la hora de establecer los nodos.[28].

2.3.2. K-Neighbors

Es un tipo de Aprendizaje Automático no supervisado y a diferencia de la mayoría de los modelos de este tipo, no presenta una función con la que se realiza la predicción. Este funciona aprendiendo el conjunto de datos. Su funcionamiento es sencillo, se le da el valor de k vecinos que se desea y una distancia máxima. El modelo se encarga agrupar los datos en grupos según la clasificación que se otorga, y el valor de salida será dictado por el valor que presentan sus vecinos cercanos. Este modelo tiene como ventaja, que se adecua muy bien a cambios en los datos[28].

Capítulo 3

METODOLOGÍA

En el proceso de *Drug discovery* para la generación de modelos de predicción de pIC50, se lleva a cabo la metodología mostrada en la Figura 3.1. Donde se comienza como la búsqueda de las proteínas objetivo a trabajar, se adquieren los datos esenciales de cada una, se calculan los descriptores y finalmente se generan los modelos de regresión.

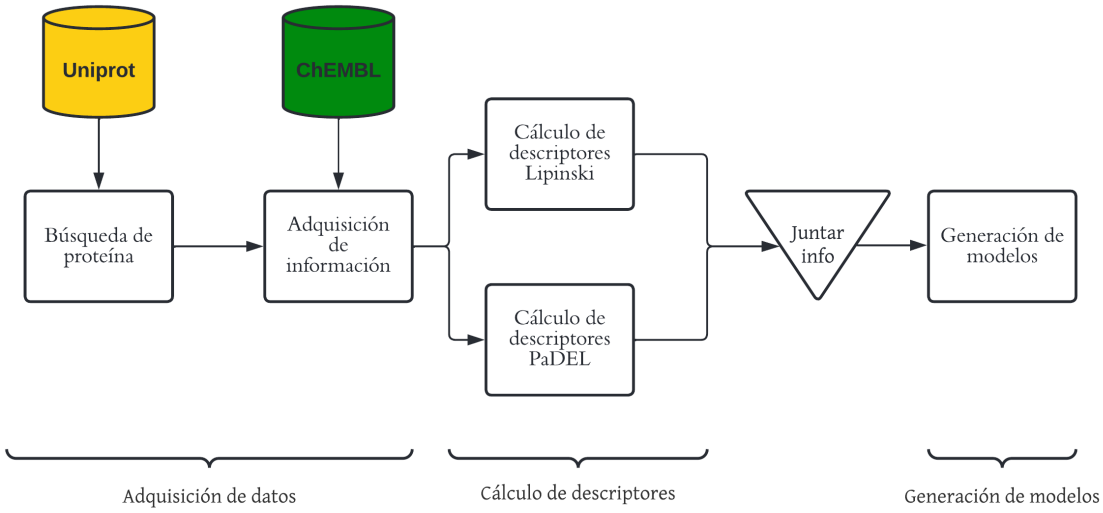


Figura 3.1: Esquemático de la metodología

3.1 Adquisición de datos

Dado que el fin del proyecto es generar modelos de predicción para el valor pIC50 de compuestos relacionados a cada proteína objetivo se debe tener como punto de partida una base de datos con información característica de cada una. Sin embargo, una de las limitantes corresponde a que no hay una diferenciación de proteínas que se presenten drogorresistencias, por ello se deben generar los datos de manera manual.
Luego de una búsqueda se establecieron tres genes relacionados con drogorresistencia y con-

sigo las proteínas que codifican, serán las proteínas objetivo del proyecto. En la tabla 3.1 se presentan los genes seleccionados, las proteínas a las que corresponden y los fármacos a los que exhiben resistencia.

Gen	Proteína	Fármaco
rpoB	Subunidad beta de ARN polimerasa dirigida por ADN	Rifabutina
atpE	ATP sintentasa subunidad c	Bedaquilina
eis	N-acetiltransferasa Eis	Amikacina, kanamicina, capreomicina y viomicina

Tabla 3.1: Genes y proteínas a trabajar

Para cada gen a trabajar junto con la proteína que codifica se consulta su correspondiente código de identificación en la base datos Uniprot. En esta se confirma que corresponda a una proteína de la bacteria *Mycobacterium tuberculosis*. Con este código de identificación se tiene acceso a información general acerca de la proteína como funciones, proteínas relacionadas y códigos de acceso en bases de datos químicas. Es entonces que se toma el código de identificación para la base de datos ChEMBl, el *ChEMBL ID*. Este último es el punto de partida para extraer los datos de los correspondientes compuestos relacionados a la proteína codificada por el gen.

Trabajando desde *Python* se hace uso de la base de datos ChEMBL a manera de cliente, de manera que con el *ChEMBL ID* encontrado previamente se encuentran una serie de compuestos químicos que tienen como proteína objetivo la proteína de interés consultada. Con ello la información adquirida corresponde un DataFrame en el que se encuentra de cada compuesto su respectivo código de identifiación *ChEMBL ID*, el valor de IC50 relacionado con la proteína y la descripción de su estructura química.

Para un posterior análisis diferenciado entre los componentes activos de los inactivos, se realiza una clasificación en función del valor del IC50, los que presentan un valor menor a 1000nM se clasifican como *active*, con valores mayores a 10000nM como *inactive*, y los restantes en *intermediate*. La clasificación se adjunta a los datos de cada componente, generando una nueva característica.

En los datos organizados se tiene cada compuesto con su respectiva información: ChEMBL ID, estructura en formato SMILES, el valor del IC50 en nM, y por último la clasificación de su bioactividad.

3.2 Cálculo de descriptores

El cálculo de los descriptores se lleva a cabo para caracterizar los compuestos relacionados a cada proteína objetivo. Se hallan descriptores mediante funciones que al recibir como entrada la estructura tridimensional encuentra los diferentes descriptores, a estructura se entrega en formato *SMILES*. Se encuentran de los descriptores relacionados con la ley Lipinski, es

decir el peso molecular (MW), número de donadores de enlaces por puentes de hidrógeno (NumHDonnors), número de aceptores de enlaces por puentes de hidrógeno (NumHAcceptors) y el coeficinente de reparto octanol/agua (LogP). De igual manera se tiene en cuenta el valor de IC50, el cual se cambia a la escala pIC50. Para esto, se convierte el valor de nM a M, y se calcula su el negativo de su antilogaritmo (Ver ecuación 3.1). Con ello los valores quedan por debajo de 10 y su interpretación es más sencilla.

$$_{P}IC50 = -log(IC50[M]) \tag{3.1}$$

Para tener más descriptores, se calculan las otra serie de descriptores también basados en la estructura química, dichos descriptores en conjunto forman una "*footprint*" o "huella"; que se asemeja al concepto de huella digital, de manera que cada compuesto tiene una huella propia que lo caracteriza en su totalidad. Se utiliza el Software *PaDEL*, encargado de crear huellas tanto bidimensionales como tridimensionales. En este caso, se calcula la huella *PubChem*, que corresponde a 881 características en las que se da la información de manera binaria. Es decir, si presenta la característica se denota con un 1, de lo contrario un 0, teniendo en cuenta caracterísitcas tan sencillas como la presencia de elementos químicos específicos hasta la presencia de grupos químicos complejos como grupos aromáricos o ácidos orgánicos.

3.3 Modelos de regresión

Para preparar los datos que serán usados en el modelo de regresión de los valores de pIC50 se toma la totalidad de los descriptores obtenidos para cada compuesto. Al ser una gran cantidad de características, se procede a rechazar las que tienen baja varianza estadística, en este caso al tratarse de valores booleanos en los descriptores, se eliminan en las que más del 80 % de los datos, presentan el mismo valor (1 ó 0) [29]. Realizada la disminusión de carcaterísticas, se tiene lista la entrada del modelo, que corresponde a las toda la matriz de valores exceptuando el pIC50, pues está será la salida o valor que se desea predecir para cada estructura.

Mediante el uso de la librería *ScikitLearn* para crear los modelos se preparan tres regresores, los escogidos fueron *RandomForestRegressor*, un *ExtraTreeRegressor* y un *KNeighborsRegressor*. Estos fueron escogidos luego de hacer un experimento con varios regresores tomando los de menor error medio, menor error cuadrático medio y menor tiempo de cómputo. Sin embargo, una de las limitaciones es el tamaño de los datos de trabajo. Por consiguiente a la hora de realizar el entrenamiento de los modelos se realiza bajo la metodología de validación cruzada *LeaveoneOut*, que se utiliza como una métrica que ayuda a escoger modelos de aprendizaje automático[30]. Esta metodología consiste en tomar cada uno de los datos tanto para entrenamiento como para validación, de esta manera se realizan tantas iteraciones como instancias exista y en cada iteración, se escoge un elemento del conjunto de datos para ser un elemento de prueba mientras el modelo toma como entrenamiento los datos restantes. Para finalizar se revisan los parámetros de funcionamiento de error medio (ME) y de error cuadrático medio (RMSE) de cada uno de los modelos para compararlos y establecer cuál de ellos se comporta de manera más adecuada.

Capítulo 4

RESULTADOS Y DISCUSIÓN

4.1 Proteína 1: Subunidad beta de ARN polimerasa dirigida por ADN

ChEMBL ID: CHEMBL3430898 - UniProt ID: P9WG45.

Esta proteína corresponde a la expresión del gen rpoB y hace parte de las proteínas en que se presenta la resistencia a la Rifabutina, fármaco de tercera línea. La proteína que pertenece a la sepa de *Mycobaterium Tuberculosis* ATCC 25618 / H37Rv. Se encarga de catalizar la transcripsción de ADN en ARN. Por consiguiente de ser inhabilitada, detiene el proceso de transcripción.

Para el caso del gen rpoB, se encontraron 10 componentes relacionados. De los cuales se clasificaron 10 como *active* y 1 como *inactive*. Es evidente que no hay un buen balance, pues la relación es 10 a 1.

Luego del cálculo del pIC50 a partir del IC50 de cada uno de los compuestos se encuentran los valores en la Figura 4.1, es evidente que hay una variación significativa. El valor de pIC50 de los componentes está en un promedio de $x = 6{,}96$ con una desviación estándar std=1,09, es decir que los componentes encontrados requieren en promedio una concentración de 1x10-7M o 100nM para disminuir al 50 % la actividad de la proteína. Además. varios de los compuestos se encuentran por debajo de 6 en escala de pIC50, y al ser una escala logarítmica, implica que el valor de IC50 es mayor y por consiguiente se requiere una concentración muy alta en comparación al resto. Es decir que en un caso de tener que escoger entre los compuestos que se tienen, serían los menos indicados porque no ayudan a optimizar.

En la Tabla 4.1 se presentan los valores de los descriptores para analizar la Ley de Lipinski para los compuestos que tienen como proteína objetivo la codificada por el gen rpoB. Y de esta tabla, es evidente que ninguno de los compuestos de llegar a ser aceptado, podría administrarse de manera oral, pues no cumple con la regla cinco de Lipinski, presentan más de uno de los descriptores por fuera de los valores requeridos. Todos estos presentan un valor de peso molecular (MW) muy alto, es decir que son compuestos estructuralmente grandes, igualmente los valores de los aceptores son muy altos y por ende podría tener interacciones no favorables a la hora de ser administrado. Ya que no tendría buena solubilidad.

Para este gen, se consiguieron 10 compuestos, es decir que para el modelo son 10 instancias

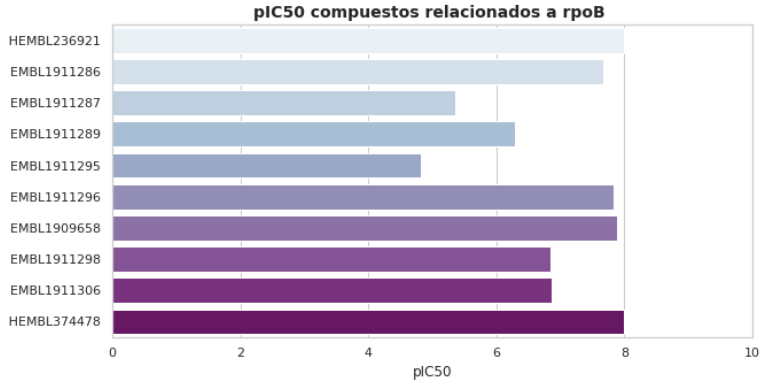


Figura 4.1: Valores de pIC50 de compuestos relacionados al gen rpoB

molecule_chembl_id	MW	LogP	NumHDonors	NumHAcceptors
CHEMBL236921	695.76	3.63	4.0	12.0
CHEMBL1911286	849.95	5.0	3.0	14.0
CHEMBL1911287	694.78	3.51	4.0	12.0
CHEMBL1911289	722.83	3.99	3.0	12.0
CHEMBL1911295	790.95	6.06	4.0	12.0
CHEMBL1911296	725.86	4.65	3.0	12.0
CHEMBL1909658	820.94	3.23	4.0	15.0
CHEMBL1911298	961.1	4.3	3.0	17.0
CHEMBL1911306	861.01	3.14	4.0	16.0
CHEMBL374478	822.95	4.34	6.0	15.0

Tabla 4.1: Descriptores de los compuestos relacionados al gen rpoB

con los cuatro descriptores de Lipinski sumados con las 881 de la huella. Luego de realizar la eliminación de los de baja varianza, se dejan solo 23 características de la huella. Con estos datos, se hace la implementación de los modelos de regresión obteniendo los datos valores presentados en la tabla 4.2.

Regresor	MAE	RMSE
Random forest	1.0118±0,83	1.0058±0,41
Extra tree	1.0046±1,05	1.0023±0,50
K Neighbors	1.0774±0,68	1.0380±0,35

Tabla 4.2: Resultados de los regresores trabajados para el gen rpoB

Teniendo en cuenta los parámetros calculados con la validación cruzada, en este caso el que presenta menores valores y por ende mejor funcionamiento, es el *Extra tree*. Sin embargo se debe caer en cuenta que la diferencia no es tan acentuada. Sobre todo con el Random Forest, ya que el Extra Tree comparte muchas similitudes con este.

4.2 Proteína 2: ATP sintetasa subunidad C

ChEMBL ID: CHEMBL2364166 - UniProt ID: P9WPS1.

Esta proteína corresponde a la expresión del gen atpE y hace parte de las proteínas en que se presenta la resistencia a la bedaquilina, fármaco de tercera línea. La proteína que pertenece a la cepa de *Mycobaterium Tuberculosis* ATCC 25618 / H37Rv. Su función es la producción de adenosín trifosfato (ATP) a partir del adenosín difosfato (ADP), es una proteína de membrana.

Para el caso del gen apE, se encontraron 16 componentes relacionados. De los cuales se clasificaron en su totalidad como *active*. Esto hace que los modelos de regresión sean específicamente para compuestos activos.

De igual manera en la Figura 4.2, se presentan los valores pIC50 de cada compuesto respectivamente, En este caso se tiene un promedio de 7,13 con desviación de 0.73, es decir un valor de IC50 promedio de 713nM. En este caso, los valores del pIC50 tienen a ser mucho mayores, es decir que el valor el IC50 es menor, lo que trae consigo beneficio, en cuanto a que al requerir menos concentración, se puede optimizar más en caso de tener que ser producido.

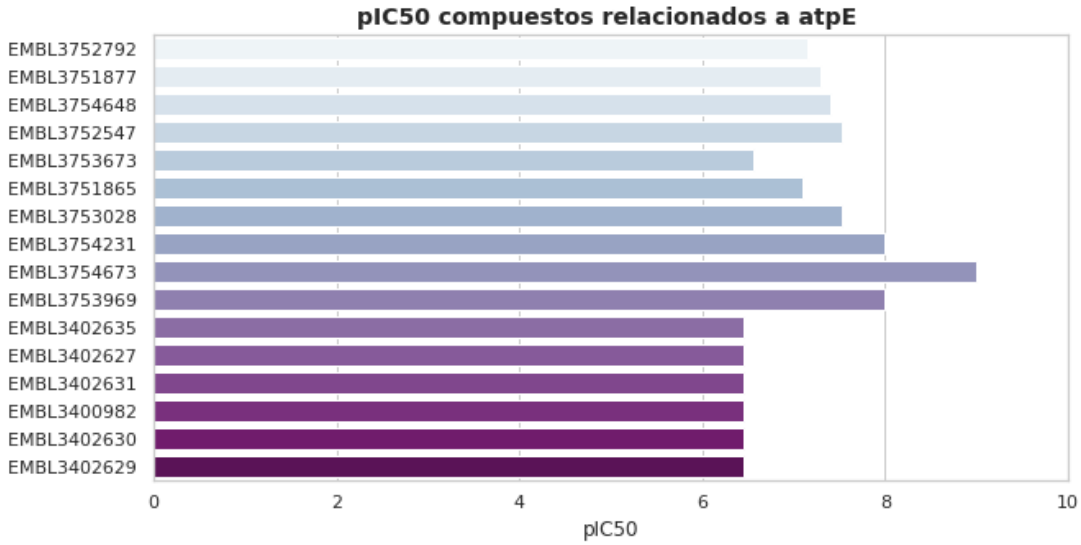


Figura 4.2: Valores de pIC50 de compuestos relacionados al gen atpE

En la tabla 4.3 se presentan los valores de los descriptores para analizar la Ley de Lipinski para los compuestos que tienen como proteína objetivo la codificada por el gen atpE. Y de esta tabla, al igual que en el caso del rpoB, ninguno es administrable vía oral. Sin embargo, la cantidad tanto como de aceptores como de donadores de enlaces por puentes de hidrógeno, es significativamente menor. Para este grupo de conjunto de compuestos, a pesar de que los aceptores y donadores estén en valores adecuados, el índice LogP está muy alto y por ende se ve afectada su solubilidad y adicional a ello, el valor del peso molecular también tiende a ser muy alto, no hay ningún compuesto quw alcance a estar por debajo de 500 uma.

molecule_chembl_id	MW	LogP	NumHDonors	NumHAcceptors
CHEMBL3752792	507.63	5.77	1.0	6.0
CHEMBL3751877	586.53	6.53	1.0	6.0
CHEMBL3754648	559.71	6.88	1.0	6.0
CHEMBL3752547	638.61	7.65	1.0	6.0
CHEMBL3753673	638.61	7.65	1.0	6.0
CHEMBL3751865	666.66	8.43	1.0	6.0
CHEMBL3753028	557.69	6.92	1.0	6.0
CHEMBL3754231	586.53	6.53	1.0	6.0
CHEMBL3754673	557.69	6.92	1.0	6.0
CHEMBL3753969	638.61	7.62	1.0	6.0
CHEMBL3402635	538.11	6.89	1.0	5.0
CHEMBL3402627	544.5	6.92	1.0	5.0
CHEMBL3402631	578.95	7.58	1.0	5.0
CHEMBL3400982	550.53	6.99	1.0	6.0
CHEMBL3402630	578.95	7.58	1.0	5.0
CHEMBL3402629	555.06	6.18	1.0	7.0

Tabla 4.3: Descriptores de los compuestos relacionados al gen atpE

Para este gen, se consiguieron 16 compuestos, de igual manera para el modelo son 16 instancias con los cuatro descriptores de Lipinski sumados con las 881 de la huella. Posterior a la remoción de los de baja varianza, se dejan 78 características de la huella. En este caso la matriz es mayor, tanto en instancias como en características a trabajar, por lo que esperan mejore valores de funcionamiento. Se realiza la implementación de los modelos de regresión utilizando los mismos parámetros trabajados en el caso anterior, obteniendo los valores presentados en la Tabla 4.4 confirmando así mejores resultados que en el caso del gen rpoB; pero, en este caso el mejor modelo de regresión es el de *K Neighbors*. Esto se puede explicar, porque al no presentarse la diferenciación entre grupos por actividad, todos pertenecen a una misma clase y por ende no existe tanta diferencia entre ellos. Es lo equivalente a eliminar un valor atípico que sería un compuesto perteneciente a la otra clase.

Regresor	MAE	RMSE
Random forest	0.3735±0,42	0.6112±0,39
K Neighbors	0.3253±0,42	0.5737±0,40
Extra tree	0.3842±0,48	0.6199±0,42

Tabla 4.4: Resultados de los regresores trabajados para el gen atpE

4.3 Proteína 3: N-acetiltransferasa eis

ChEMBL ID: CHEMBL3879870 - UniProt ID: P9WFK7.

Esta proteína corresponde a la expresión del gen eis y es una proteína en la que se genera resistencia a varios fármacos de segunda línea como lo son la amikacina, la kanamicina, la capreomicina y la viomicina. La proteína que pertenece a la cepa de *Mycobaterium Tuberculosis* ATCC 25618 / H37Rv. Es una proteína agresiva, ya que se encarga de acetilar varios fármacos antituberculosos, es por ello que genera drogorresistencia. Para el caso del gen eis, se encontraron 27 componentes relacionados. De los cuales se clasificaron 10 como *active* y 12 *inactive*. Y para este caso la muestra de de mayor tamaño está mejor balanceada.

En la Tabla 4.5 se presentan los valores de los descriptores para analizar la Ley de Lipinski para los compuestos que tienen como proteína objetivo la codificada por el gen eis. Para el caso todos los compuestos relacionados cumplen con la "ley de cinco" de Lipinski y por consiguiente, todos podrían ser administrados vía oral. Esto simplifica su administración y por ende el tratamiento del paciente.

molecule_chembl_id	MW	LogP	NumHDonors	NumHAcceptors
CHEMBL3958709	283.28	2.9	0.0	5.0
CHEMBL3926253	301.27	3.04	0.0	5.0
CHEMBL3950034	301.27	3.04	0.0	5.0
CHEMBL3976122	317.73	3.56	0.0	5.0
CHEMBL3905343	362.18	3.67	0.0	5.0
CHEMBL3891229	313.31	2.91	0.0	6.0
CHEMBL3960292	301.27	3.04	0.0	5.0
CHEMBL3980494	317.73	3.56	0.0	5.0
CHEMBL3964177	362.18	3.67	0.0	5.0
CHEMBL3909271	297.31	3.21	0.0	5.0
CHEMBL3979506	333.34	4.06	0.0	5.0
CHEMBL3900249	235.24	2.0	0.0	5.0
CHEMBL3971387	263.29	2.64	0.0	5.0
CHEMBL3927142	309.32	2.12	0.0	4.0
CHEMBL3899105	327.31	2.25	0.0	4.0
CHEMBL3903088	327.31	2.25	0.0	4.0
CHEMBL3931070	343.77	2.77	0.0	4.0
CHEMBL3927499	388.22	2.88	0.0	4.0
CHEMBL3940094	339.35	2.12	0.0	5.0
CHEMBL1410785	327.31	2.25	0.0	4.0
CHEMBL1612166	343.77	2.77	0.0	4.0
CHEMBL1538832	388.22	2.88	0.0	4.0
CHEMBL3890216	323.35	2.42	0.0	4.0
CHEMBL3952477	359.38	3.27	0.0	4.0
CHEMBL3976851	261.28	1.21	0.0	4.0
CHEMBL1702313	289.33	1.85	0.0	4.0
CHEMBL3915019	354.32	2.02	0.0	6.0

Tabla 4.5: Descriptores de los compuestos relacionados al gen eis

En el caso del gen eis, es evidente que los valores del pIC50 son menores que en los anteriores dos genes tal como se ve en la Figura 4.3. El promedio se reduce a un valor de 5.22 con una desviación de 1.3. Por consiguiente, los compuestos relacionados la proteína del gen eis, tienen en promedio un valor de 100uM. Lo que la hace la concentración más alta, pero

sigue siendo a la única con que es posible administrar los compuestos de manera oral. Todos lo compuestos cumplen con la Ley de cinco, pero requieren concentraciones altas.

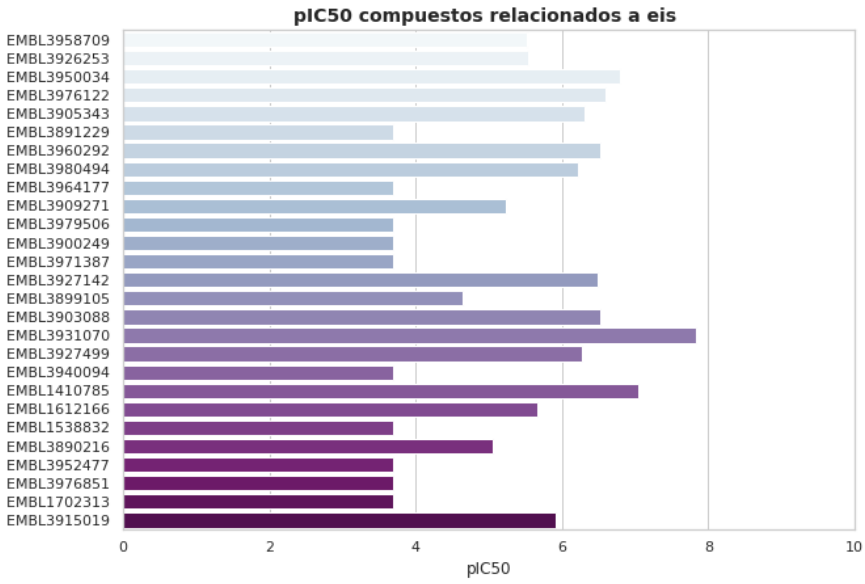


Figura 4.3: Valores de pIC50 de compuestos relacionados al gen eis

Para este gen, se consiguieron 27 compuestos, que implican 27 instancias para generar modelos. Con la remoción de características de baja varianza, se dejan 81 provenientes de la huella. Para este gen tenemos la matriz de mayor tamaño. Luego de la implementación de los modelos, se presentan los resultados en la tabla 4.6.

Regresor	MAE	RMSE
Random forest	1.2796±0,74	1.1312±0,36
K Neighbors	1.0640±0,75	1.0315±0,38
Extra tree	1.0117±0,81	1.0058±0,51

Tabla 4.6: Resultados de los regresores trabajados para el gen eis

A pesar de tener la muestra más grande, los resultados no son los mejores dentro de las tres proteínas, además, también cabe resaltar que es el mejor de los tres modelos para este caso es el *Extra tree*. En este caso, a diferencia de los otros dos, el modelo *Random Forest* es el que presenta el peor funcionamiento. En el caso de *K neighbors* es un funcionamiento adecuado, pero no el mejor. Se puede inferir que, si la finalidad fuese predecir la clasificación de bioactividad, este sería el mejor modelo porque podría agrupar las clases de manera adecuada.

4.4 Modelos de regresión

En todos los casos se logran implementar los modelos de regresión, en donde los hiperparámetros fueron adecuados de la manera más óptima. Algunos hiperparámetros no pudieron ser valores grandes, a causa del tamaño de las muestras con que se entrenaron. A pesar de que se utilizó la validación cruzada para el entrenamiento, el tamaño de las muestras sigue siendo inadecuado.

En la tabla 4.7 se presentan los hiperparámetros de cada modelo.

Regresor	Hiperparámetros
Random forest	n.estimators=10, criterion=squared, n.jobs=-1
K Neighbors	n.neighbors=5, weights=uniform, algorithm=auto, n.jobs=-1
Extra tree	n.estimators=10, criterion=squared, n.jobs=-1

Tabla 4.7: Hiperparámetros de los modelos de regresión

Cada uno de los modelos se trabajó con el método de validación cruzada utilizando 5 Folds. En el caso del *Random Forest*, se estableció para trabajar con 10 estimadores, y con criterio de error cuadrártico medio. El *Extra Tree* se trabajó con el parámetro de mejor decisión aleatoria, y con un número aleatorio de características. Para finalizar, el *K Neighbors Regressor* se trabajó con un valor de 5 vecinos cercanos y con una detección automática del algoritmo de detección de valores cercanos. Los parámetros de los modelos fueron variados hasta encontrar los mejores, en el caso del Random Forest y del Extra Tree, se varió el número de estimadores de 1 a 10.

La falta de datos es un impedimento para poder genera modelos adecuados. Si bien se logra hacer una descripción adecuada de cada una de las proteínas, no es posible llegar a un resultado que presente una validez significativa. No existe criterio para establecer si los modelos funcionan de manera adecuada.

Capítulo 5

CONCLUSIONES

La caracterización de cada compuesto relacionado a las proteínas objetivo seleccionadas se realizó adecuadamente, utilizando los descriptores de la ley de Lipinski así como los correspondientes a la huella calculada mediante Padel.

El proceso del *Drug Discovery* trabajado es completamente dependiente de la base de datos inicial, pues si bien se pudo llegar a modelos de regresión, una muestra de mayor tamaño hubiera sido más adecuada y se hubiesen presentado mejores resultados.

En general se evidenció que en cuanto a los modelos generados, el que mejor se adecúa a la problemática es el *Extra tree*. Pues mostró los valores de error tanto cuadrático medio como error medio más bajos. Sin embargo, los hiperparámetros no se pudieron ajustar de manera tan libre. Dado que, al tener un conjunto de datos tan pequeños, es muy posible llegar sobreajustes en los modelos.

Capítulo 6

RECOMENDACIONES Y TRABAJOS FUTUROS

Como trabajo futuro a corto plazo se tiene el cambiar la base de datos escogida o complementar la misma con más compuestos relacionados, generando así un conjunto de entrenamiento más robusto y por ende generando mejores predicciones. A medio plazo se plantea implementar diferentes modelos de aprendizaje automático, dado que al tener un conjunto de datos mayor se pueden tomar otro tipo de modelos diferentes a los regresores; de igual manera, establecer más proteínas objetivo para comparar la metodología trabajada en más casos. Para finalizar un trabajo futuro a largo plazo es generar una base de datos con los resultados encontrados, almacenando los los compuestos relacionados a cada proteína junto con los valores de pIC50 calculados respectivamente.

Como recomendación para futuros trabajos similares propone aplicar la temática a otro tipo de enfermedades, de manera que se pueda realizar una comparación para establecer la importancia de una base de datos robusta.

Bibliografía

[1] A. Natarajan, P. Beena, A. V. Devnikar y S. Mali, «A systemic review on tuberculosis,» *Indian Journal of Tuberculosis*, vol. 67, n.º 3, págs. 295-311, 2020, Cited By :14.

[2] A. Zabaleta y C. Llerena, «Extensively resistant tuberculosis, Colombia, 2006-2016,» *Biomedica*, vol. 39, n.º 4, págs. 707-714, 2019, Cited By :3.

[3] D. Mendez, A. Gaulton, A. P. Bento et al., «ChEMBL: towards direct deposition of bioassay data,» *Nucleic Acids Research*, vol. 47, n.º D1, págs. D930-D940, nov. de 2018.

[4] C. Silva, V. Bermúdez, N. Arraiz, F. Bermúdez, M. Rodríguez y L. V. E. Leal., «Fármacos de primera línea utilizados en el tratamiento de la tuberculosis,» *Archivos Venezolanos de Farmacología y Terapéutica*, 2007.

[5] N. A. Shaik, K. R. Hakeem, B. Banaganapalli y R. Elango, *Essentials of Bioinformatics Volume II, In Silico Life Sciences: Medicine*. Springer, 2019.

[6] G. Singh, P. Kesharwani y A. K. Srivastava, «Tuberculosis treated by multiple drugs: An overview,» *Current Drug Delivery*, vol. 15, n.º 3, págs. 312-320, 2018, Cited By :5.

[7] C. E. Rincón-Torres, V. Rubio, C. Castro et al., «National Network for Knowledge Management, Research, and Innovation in Tuberculosis in Colombia,» *Revista Panamericana de Salud Publica/Pan American Journal of Public Health*, vol. 45, 2021.

[8] J. G. Rodríguez-Castillo, C. Llerena, L. Argoty-Chamorro et al., «Population structure of multidrug-resistant Mycobacterium tuberculosis clinical isolates in Colombia,» *Tuberculosis*, vol. 125, 2020, Cited By :2.

[9] A. Zabaleta, C. Llerena y A. Valbuena, «Tuberculosis resistente en menores de 15 años, Colombia 2010-2015,» *Biomedica*, vol. 39, n.º 2, págs. 330-338, 2019, Cited By :3.

[10] S. Macalino, J. B. Billones, V. G. Organo y Carrillo, «In Silico Strategies in Tuberculosis Drug Discovery.,» *Basel, Switzerland*, vol. 25, n.º 3, 2020. DOI: 10.3390/molecules25030665.

[11] S. Swaminathan, J. C. Sundaramurthi, A. N. Palaniappan y S. Narayanan, «Recent developments in genomics, bioinformatics and drug discovery to combat emerging drug-resistant tuberculosis,» *Tuberculosis*, vol. 101, págs. 31-40, 2016.

[12] G. Mugumbate, B. Nyathi, A. Zindoga y G. Munyuki, «Application of Computational Methods in Understanding Mutations in Mycobacterium tuberculosis Drug Resistance,» *Frontiers in Molecular Biosciences*, vol. 8, 2021.

[13] X. Chen, J. Sa, M. Li e Y. Zhou, «Combined prediction model of tuberculosis based on generalized regression neural network,» en *Proceedings of 2020 IEEE International Conference on Artificial Intelligence and Computer Applications, ICAICA 2020*, Cited By :2, 2020, págs. 577-581.

[14] K. Ghazvini, M. Yousefi, F. Firoozeh y S. Mansouri, «Predictors of tuberculosis: Application of a logistic regression model,» *Gene Reports*, vol. 17, 2019, Cited By :12.

[15] R. S. Wallis, C. Wang, D. Meyer y N. Thomas, «Month 2 Culture Status and Treatment Duration as Predictors of Tuberculosis Relapse Risk in a Meta-Regression Model,» *PLoS ONE*, vol. 8, n.º 8, 2013.

[16] M. Hassam, J. A. Shamsi, A. Khan, A. Al-Harrasi y R. Uddin, «Prediction of inhibitory activities of small molecules against Pantothenate synthetase from Mycobacterium tuberculosis using Machine Learning models,» *Computers in biology and medicine*, vol. 145, 2022.

[17] T. R. Lane, F. Urbina, L. Rank et al., «Machine Learning Models for Mycobacterium tuberculosis in Vitro Activity: Prediction and Target Visualization,» *Molecular Pharmaceutics*, vol. 19, n.º 2, págs. 674-689, 2022.

[18] Y. Hu, Z. Morichaud, S. Chen, J.-P. Leonetti y K. Brodolin, «Mycobacterium tuberculosis RBPA protein is a new type of transcriptional activator that stabilizes the A -containing RNA polymerase holoenzyme,» *Nucleic Acids Research*, vol. 40, n.º 14, págs. 6547-6557, 2012. DOI: 10.1093/nar/gks346.

[19] M. A. Isa, M. B. Abubakar, M. M. Mohammed, M. M. Ibrahim y F. A. Gubio, «Identification of potent inhibitors of ATP synthase subunit C (AtpE) from mycobacterium tuberculosis using in silico approach,» *Heliyon*, vol. 7, n.º 12, 2021. DOI: 10.1016/j.heliyon.2021.e08482.

[20] J. L. Houghton, T. Biswas, W. Chen, O. V. Tsodikov y S. Garneau-Tsodikova, «Chemical and structural insights into the regioversatility of the aminoglycoside acetyltransferase Eis,» *European journal of chemical biology*, vol. 7, n.º 14, 2013. DOI: 10.1002/cbic.201300359.

[21] C. W. Yap, «Padel-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints,» *Journal of Computational Chemistry*, vol. 32, n.º 7, 2010. DOI: 10.1002/jcc.21707.

[22] C. A. Lipinski, F. Lombardo, B. W. Dominy y P. J. Feeney, «Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings,» *Advanced Drug Delivery Reviews*, vol. 46, n.º 1, págs. 3-26, 2001, ISSN: 0169-409X. DOI: https://doi.org/10.1016/S0169-409X(00)00129-0.

[23] P. R. Duchowicz, C. Talevi A.and Bellera, E. A. Castrp y L. E. Buno-Lanch, «Estudio QSPR Sobre Solubilidades Acuosas de Sustancias Orgánicas,» *XXVI Congreso Argentino de Química Dr. Ángel del Carmen Devia*, 2006.

[24] C. Rauch, «Toward a mechanical control of drug delivery. On the relationship between Lipinski's 2nd rule and cytosolic pH changes in doxorubicin resistance levels in cancer cells: A comparison to published data,» *European Biophysics Journal*, vol. 38, n.º 7, págs. 829-846, 2009, Cited By :38.

[25] D. S. Diningrat, A. N. Sari, N. S. Harahap y Kusdianti, «IN SILICO STUDY OF THE TOXICITY AND ANTIVIRAL ACTIVITY PREDICTION OF JAMBLANG (Syzygium cumini) LEAVES ESSENTIAL OIL AS ACE2 INHIBITOR,» *Pharmacologyonline*, vol. 3, págs. 1334-1351, 2021.

[26] S. Hosseini, S. Ketabi y G. Hasheminasab, «QSAR study of antituberculosis activity of oxadiazole derivatives using DFT calculations,» *Journal of Receptors and Signal Transduction*, 2022.

[27] S. Singh, K. R. Ramkumar y A. Kukkar, «Machine Learning Techniques and Implementation of Different ML Algorithms,» en *2021 2nd Global Conference for Advancement in Technology, GCAT 2021*, 2021.

[28] S. Singh, K. R. Ramkumar y A. Kukkar, *Python Machine Learning: Unlock deeper insights into Machine Leaning with this vital guide to cutting-edge predictive analytics.* 2015.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort et al., «Scikit-learn: Machine Learning in Python,» *Journal of Machine Learning Research*, vol. 12, págs. 2825-2830, 2011.

[30] J. Shao, «Linear model selection by cross-validation,» *Journal of the American Statistical Association*, vol. 88, n.º 422, págs. 486-494, 1993.