Universidad Del Papaloapan

Campus Tuxtepec

ESTUDIO QSAR DE INHIBIDORES DE TROMBINA COMO FUNCIÓN DEL TIEMPO

TESIS

PARA OBTENER EL GRADO DE

LICENCIADA EN CIENCIAS QUÍMICAS

PRESENTA:

Narelle Montañez Godinez

Dr. Guillermo Ramírez Galicia

Director de Tesis

San Juan Bautista Tuxtepec, Oaxaca. 2018

UNIVERSIDAD DEL PAPALOAPAN

CAMPUS TUXTEPEC

San Juan Bautista Tuxtepec, Oax. a 03 de abril de 2018
Asunto: Autorización de impresión de tesis

M. E. YESENIA BARRIENTOS ARENAL
JEFA DEL DEPARTAMENTO DE SERVICIOS ESCOLARES
UNIVERSIDAD DEL PAPALOAPAN
P R E S E N T E

Sirva la presente para infórmale que los miembros de la Comisión Revisora del trabajo de tesis de la **C. NARELLE MONTAÑEZ GODÍNEZ** pasante de la carrera de la Licenciatura en Ciencias Químicas con número de matrícula 08060006, revisó y aprobó el trabajo de investigación denominado **"ESTUDIO QSAR DE INHIBIDORES DE TROMBINA COMO FUNCIÓN DEL TIEMPO"** mismo que será presentado como prueba escrita del acto de recepción profesional. para obtener el Título de Licenciada en Ciencias Químicas.

Por lo anterior y de acuerdo a los lineamientos institucionales, se le da trámite legal para que proceda a su impresión el trabajo presentado.

UNIVERSIDAD DEL PAPALOAPAN
3 ABR. 2018
DEPARTAMENTO DE SERVICIOS ESCOLARES CAMPUS TUXTEPEC

A t e n t a m e n t e
terra uberrima, mens aperta
Bǫ̈u Lo-tama, chí jí jú

Dra. Roxana Martínez Pascual
Jefa de Carrera
Licenciatura en Ciencias Químicas

UNPA OAXACA
LICENCIATURA EN CIENCIAS QUÍMICAS CAMPUS TUXTEPEC

C.c.p. M. en C. Héctor López Arjona. Vice-Rector Académico de la UNPA, para su conocimiento
C.c.p. Archivo de la jefatura de carrera

UNIVERSIDAD DEL PAPALOAPAN

CAMPUS TUXTEPEC

Tuxtepec, Oaxaca a 20 de marzo de 2018
ASUNTO: Designación de sinodales

NARELLE MONTAÑEZ GODÍNEZ
PASANTE DE LA LICENCIATURA EN CIENCIAS QUÍMICAS
PRESENTE

Por este medio le informo la propuesta de la jefatura de carrera de los Profesores-Investigadores que fungirán como revisores en su examen para obtener el Título de Licenciada en Ciencias Químicas.

Dr. Francisco Noé Mendoza Ambrosio	UNPA	Presidente
Dr. Óscar Abelardo Ramírez Marroquín	UNPA	Secretario
Dr. Guillermo Ramírez Galicia	UNPA	Vocal
Dr. Edgar García López	UNPA	1er Suplente
Dr. Ramón Garduño Juárez	UNAM	2° Suplente

UNIVERSIDAD DEL PAPALOAPAN
21 MAR. 2018
DEPARTAMENTO DE SERVICIOS ESCOLARES CAMPUS TUXTEPEC

Sin más por el momento. le envío cordiales saludos.

A t e n t a m e n t e
terra uberrima, mens aperta
Bẽu Lo-tama, chi ji jú

Dra. Roxana Martínez Pascual
Jefa de Carrera
Licenciatura en Ciencias Químicas

M. en C. Héctor López Arjona
Vice-rector Académico
Vo. Bo.

UNIVERSIDAD DEL PAPALOAPAN
UNPA
OAXACA
VICE-RECTORIA ACADEMICA

C.c.p. Dr. Guillermo Ramírez Galicia -Director de Tesis
C.c.p. Yesenia Barrientos Arenal -Jefa de Servicios Escolares
C.c.p. Archivo de la jefatura

A mis padres,

como testimonio de gran admiración y gratitud.

A Mario:

Por tu apoyo y comprensión

pero sobre todo por el amor.

Agradecimientos

*

Quiero agradecer a mis padres estoy segura de que he llegado hasta aquí gracias a todo el tiempo, consejos y amor que ustedes me han brindado; así que la culminación de este proyecto mas allá de ser un logro mío, es un logro suyo que al igual que a mí nos han costado muchas lágrimas, noches sin dormir y horas de angustia, así que ¡Enhorabuena para ustedes! y también va por ti querida hermana que siempre me has alentado. Gracias a ustedes mi familia por ser mi ejemplo, mi mayor inspiración, pero sobre todo por ser mi fortaleza.

*

A Mario por los innumerables consejos y por todo el apoyo que me has dado no tan solo para esta tesis sino para cada aspecto de mi vida, la cual ha cambiado desde que estas en ella, por creer en mi, incluso cuando ni yo misma lo hago y como Saint-Exupèry dijo "El amor no consiste en mirarse el uno al otro, sino en mirar juntos en la misma dirección". Gracias por estar en la misma dirección.

*

Quiero agradecer a la Universidad del Papaloapan por el apoyo recibido, así como a los integrantes de esta institución, profesores investigadores, al departamento de escolares sobre todo a la Licenciada Yessenia, a los encargados de laboratorio, biblioteca, sala y a quien sea que en algún momento llegue a darle lata, gracias.

*

Agradezco a mi tutor el Dr. Guillermo Ramírez Galicia por ayudarme a aterrizar mis ideas y mis conocimientos no tan solo en esta tesis sino en la vida misma, gracias por la paciencia, por la guía, por el buen humor y por ayudarme a salir de los obstáculos que se me presentaron aun cuando pensaba que no podía mas y ser mi compañero de anécdotas y por todos los "empujoncitos" que me ha dado. Gracias.

*

Agradezco a mis sinodales por ayudarme a darle forma a esta tesis, definitivamente sin ustedes sería un desastre. Gracias

*

También agradezco a mis amigos que han sido mis compañeros de vida por todas las palabras de animo que he recibido y las que no también; Brenz, Conchis, Mary, Jair, Vicky y compañía, Kryzz, Yessi, Heidy, química Lety, Luz, Karla, Chely gracias muchas gracias.

*

A todos aquellos que probablemente omití por la emoción y que de alguna manera contribuyeron e hicieron posible la culminación de este proyecto. Les doy mi más sincero agradecimiento, sé que me llevo de cada persona una experiencia y una lección de vida; gracias.

"Qué Dios nos dé la sabiduría para descubrir lo correcto, la voluntad para elegirlo y la fuerza para hacer que perdure"

Contenido

1. Introducción

Los seres vivos utilizan frecuentemente una serie de reacciones enzimáticas en cadena como respuesta a un estímulo externo y/o interno, un ejemplo de esto es la formación de coágulos sanguíneos en el interior de un vaso sanguíneo (trombosis), que dependiendo de su ubicación puede generar diferentes afecciones y complicaciones de salud. Actualmente existen diferentes medicamentos en el mercado para evitar y/o contrarrestar la formación de los trombos; sin embargo, presentan diferentes efectos secundarios. Una solución ha sido la investigación de compuestos alternativos, como los heterocíclicos no peptídicos derivados del benzimidazol que presentan efecto inhibitorio sobre la trombina.

La trombina presenta diferentes cavidades, conocidas por D- y P-cavidad, con tres principales sitios activos, S1 a S3, siendo el sitio S1 el principal, el cual se encuentra ubicado cercano al aminoácido Serina 195 [1]. Conocer el sitio de inhibición sobre esta enzima, la forma en que interactúa y los cambios conformacionales que sufre con el tiempo por medio de las diferentes herramientas computacionales han permitido predecir la actividad de nuevas moléculas; entre ellas el análisis QSAR por medio de regresiones multilineales, regresiones no lineales vía redes neuronales artificiales y simulaciones moleculares a partir del reconocimiento molecular (Docking), que describan las interacciones entre la trombina y sus inhibidores. Algunos de los cuales, presentan efectos no deseados como la falta de selectividad, baja bio-disponibilidad oral y que poseen grupos funcionales reactivos como aldehídos, cetonas y ácidos borónicos, entre otros [2].

En la presente tesis se realizó el análisis de 50 compuestos derivados del benzimidazol. Las conformaciones de compuestos fueron obtenidas previamente por medio de acoplamiento molecular, en once conformaciones de la trombina obtenidas por dinámica molecular [3].

2. Antecedentes

En el proceso de coagulación normal intervienen diferentes elementos celulares y plasmáticos que deben interactuar equilibradamente; de esta manera pueden participar plaquetas, endotelio, leucocitos, eritrocitos y factores plasmáticos. El proceso de coagulación, para su correcta función, utiliza sistemas de regulación, siendo la antitrombina y el sistema de proteínas C (Autotrombina II) y S (glicoproteína plasmática vitamina K-dependiente, que actúa como un cofactor de la proteína C), con el sistema fibrinolítico los más importantes. La activación de la coagulación puede llevarse a cabo de dos maneras: 1) por vía intrínseca ó 2) por vía extrínseca. La primera se observa cuando un vaso sanguíneo se rompe mientras que la segunda se presenta cuando el endotelio libera el factor tisular, durante este proceso se van generando pequeñas cantidades de trombina [4], que es la proteasa clave en la coagulación. De esta manera la trombina activa enzimas específicas que conducen a la generación de más enzimas, hasta llegar a construir un coágulo, el cual está constituido por fibrina; la parte final del proceso sucede cuando el fibronógeno se convierte en fibrina por acción de la trombina (Figura 1).

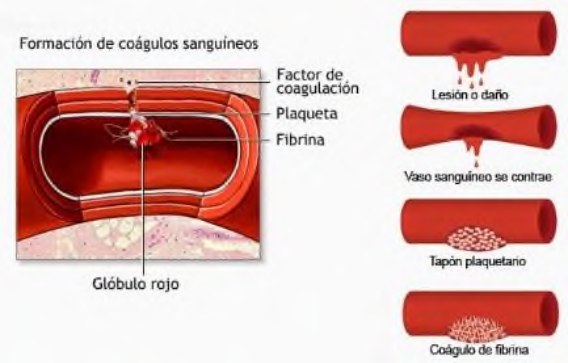


Figura 1: Formación de coágulos sanguíneos.

Así mismo se pueden formar coágulos insolubles con la activación de plaquetas y eritrocitos en exceso; la trombina cataliza la conversión del fibronógeno soluble a insoluble en la cascada de la coagulación y puede actuar sobre otros substratos como el factor V, VIII, XI y XIII [4-7].

Un trombo es un coágulo sanguíneo que se forma en un vaso y permanece allí, bloqueando el flujo sanguíneo e impidiendo la oxigenación y el flujo adecuado de

sangre hacia los tejidos, lo cual puede ocasionar diferentes daños y su gravedad depende de la vena, arteria o capilar obstruido, generando una trombosis provocando que los tejidos o células sufran isquemia; es decir, impedimento del flujo correcto del oxígeno, pudiendo producir una lesión celular y si esta se prolonga puede provocar necrosis o muerte celular conocida como infarto ya sea del miocardio, pulmonar o cerebral dependiendo del órgano afectado [6-8].

2.1 Trombosis y sus factores de riesgo

Las consecuencias de una trombosis incluyen diversas complicaciones, agudas o crónicas, siendo responsables de más de la mitad de los decesos en las sociedades desarrolladas como una de las principales causas de mortalidad mundial. La trombosis venosa es la tercera causa de muerte relacionada con problemas cardiovasculares solo superada por el infarto y el ataque al miocardio [4]. Esto se agrava al adquirir factores de riesgo propios de la vida moderna, como lo son la obesidad, infecciones, embarazo, inmovilización, ingesta de anticonceptivos orales y por la predisposición genética, provoca que la trombosis se convierta en una causa importante de morbi-mortalidad.

Debido a esto, el estudio sobre la inhibición de trombosis e inhibiendo la función plaquetaria con anticoagulantes de acción directa y de acción indirecta es importante (Tabla 1). Los anticoagulantes de acción directa son aquellos que inhiben por si solos la coagulación y los de acción indirecta inhiben mediante la acción de otras proteínas (Figura 2), siendo la inhibición de la trombosis una estrategia terapéutica clave, como ejemplo de inhibidor directo de trombina tenemos a la bivalirudina que está basado en la hirudina que es un anticoagulante producido por las sanguijuelas, la cual se une a la trombina generando un cambio conformacional impidiendo su acción [9].

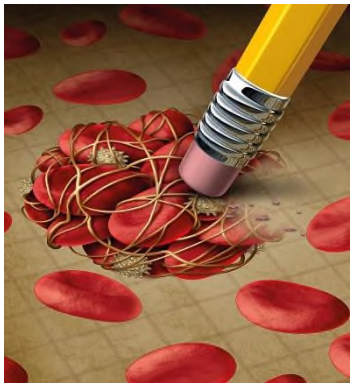


Figura 2: Ejemplificación de la acción de un anticuagulantes.

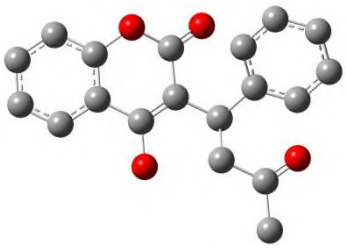
El uso de los medicamentos existentes en el mercado para la inhibición de la trombosis provoca diferentes efectos secundarios, por ejemplo, la heparina, presenta un alto efecto anticoagulante y es incapaz de inhibir a la trombina unido al coágulo, conocido como el efecto pro-agregante plaquetario, así mismo presenta un efecto rebote trombótico cuando se retira el tratamiento. Este es uno de los motivos por el cual se han estudiado ampliamente los inhibidores directos de trombina [9-12].

Efecto biológico	Anticoagulantes	Mecanismo de acción	Efectos adversos
Acción indirecta	Heparina no fraccionada	Antitrombina III	Sangrado Osteoporosis Trombocitopenia Necrosis cutánea Urticaria
Acción indirecta	Heparina de bajo peso molecular	Antitrombina III	Sangrado Osteoporosis Trombocitopenia Necrosis cutánea Urticaria
Acción indirecta	Warfarina y Acecumarol (derivados del Dicumarol)	Inhiben la interconversión de la vitamina K desde su forma oxidada hasta su forma reducida.	Alopecia Malformaciones congénitas Necrosis cutánea Sangrado
Acción Directa	Hirudina y Argatroban	Trombina	Hemorragia espontanea. Hematomas

Tabla 1: Detalles de los anticoagulantes existentes en el mercado.

El anticoagulante oral más usado en el mundo es la Warfarina ((RS)-4-hidroxi-3-(1-fenil-3-oxo-butil-cumarina)), la cual se absorbe fácilmente por el tubo digestivo

ayudada por su alta solubilidad en lípidos, y una vida media entre 36 y 42 horas, que actúa inhibiendo el ciclo de inter-conversión de la vitamina K desde su forma oxidada a su forma reducida, la vitamina K es esencial para desarrollar ciertos factores entre ellos los de coagulación y anticoagulación, dentro de sus efectos adversos se encuentran que pueden provocar alopecia (pérdida anormal del cabello) y mal formaciones genéticas [9-11].



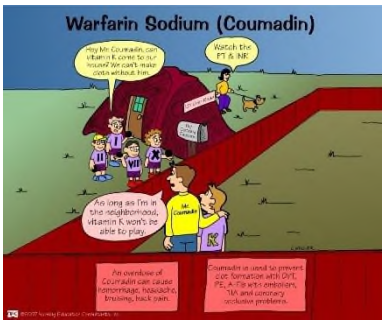


Figura 3: Representación de la Warfarina.

"*Mira el tiempo de protrombina y el radio internacional nacionalizado.
*Hígado
*Ey, Señor Cumadina, ¿puede vitamina k venir a nuestra casa? No podemos hacer coágulos sin él.
*Mientas yo esté en el vecindario, vitamina K no está disponible para jugar.
Una sobredosis de Cumadina puede causar hemorragia, dolor de cabeza hematomas y dolor de espalda.
La Cumadina es usada para prevenir la formación de coágulos en trombosis venosa profunda, embolismo pulmonar, fibrilación auricular con embolismo, ataque isquémico transitorio y problemas de oclusión coronaria."

2.2 Desarrollo de nuevos fármacos

El camino que debe seguir un fármaco comienza con diversos experimentos denominados estudios preclínicos, este tipo de estudios analizan si el compuesto candidato es seguro para iniciar su evaluación en seres humanos, por medio de cultivos celulares y animales como modelos, con el objetivo de predecir cómo actúa el organismo sobre el candidato a fármaco, es decir, se realiza un análisis farmacocinético. Por otra parte, los estudios farmacodinámicos determinan como el candidato a fármaco actúa sobre el organismo, como por ejemplo los efectos dosis-respuesta, los cambios bioquímicos y fisiológicos, de este modo podemos conocer si el candidato a fármaco es perjudicial o tóxico [13].

El desarrollo de nuevos medicamentos, después del descubrimiento de un fármaco, puede dividirse en diferentes fases en las cuales se estudia ampliamente la farmacodinámica y farmacocinética en humanos incluyendo sus efectos terapéuticos, adversos y sus riesgos. Cuando un compuesto satisface las pruebas farmacológicas, toxicológicas y bioquímicas durante la investigación preclínica se presenta ante las agencias regulatorias: en México es la Secretaria de Salud a través de la COFEPRIS (Comisión Federal para la Protección Contra Riesgos Sanitarios), la FDA (Food and Drug Administration) en el caso de Estados Unidos, la EMEA (European Medicines Agency) en la Comunidad Europea, la MHLW (Ministry of Health, Labour and Welfare) en Japón entre otros. De ser aprobado por la agencia regulatoria correspondiente, el nuevo fármaco se valorará, se le asigna un expediente con la aplicación que autoriza a una compañía farmacéutica a realizar los ensayos clínicos [13, 14].

La fase clínica está constituida por tres etapas: I, II y III. Durante los ensayos clínicos en humanos se observan los efectos adversos que ocasiona el nuevo fármaco que no son fácilmente reconocidos en los animales tales como la somnolencia o la confusión mental. La mitad de los efectos indeseables se identifican en esta etapa, debido entre otras cosas a la diferencia de tamaño y/o especie que repercuten en la farmacodinámica y farmacocinética que puede presentar el fármaco [14].

En la última fase se aplica la fármaco-vigilancia, cuando el medicamento es aprobado se mantiene una vigilancia continua de la seguridad del nuevo medicamento en las condiciones y en un número extenso de pacientes [14].

El proceso en general lleva un tiempo aproximado de 2 a 5 años para identificar la enfermedad y aislar la proteína involucrada, otros 2 a 5 años en escalar un fármaco efectivo, de 1 a 3 años en pruebas preclínicas, de 2 a 3 meses para realizar la formulación y escalamiento, de 2 a 10 años para realizar las pruebas clínicas en humanos y por último, en el caso de Estados Unidos, la aprobación de

la FDA que puede tardar de 2 a 3 años; en México la Secretaría de Salud a través de la COFEPRIS no tiene un tiempo fijo para dicha aprobación [14-16].

De manera general, los compuestos que son sintetizados pasan por un proceso de estudios muy amplio, aquellos que resultan prometedores se estudian más a fondo sus propiedades farmacocinéticas, metabolismo y su potencial toxicidad; sin embargo la demanda de información de los diferentes productos sintetizados con propiedades farmacocinéticas ha aumentado junto con la necesidad de exámenes más exhaustivos y con una mayor cantidad de información de los datos de Liberación, Absorción, Distribución, Metabolismo, Excreción, Toxicidad (LADMET), para establecer si alguna de estas estructuras cumplirá con las características necesarias para colocarse como un buen fármaco.

2.3 Análisis *in silico*

Los análisis de toxicología *in silico*, provienen de bases de datos con información ya existente, extrapolación formada por agrupamientos, relación estructura-actividad, y relación cuantitativa estructura-actividad, entre otros [17, 18].

Los programas de cómputo que hacen factible los análisis *in silico* hacen que los proyecto sean más accesibles para los analistas utilizando los sitios activos que pueden contener las proteínas, haciéndolos una herramienta poderosa en la predicción de las moléculas; estos programas necesitan contener información exacta de las moléculas a analizar sin despreciar factores importantes que expliquen las uniones entre el receptor y el ligando [17].

Comúnmente se utilizan técnicas de regresión lineal o multilineal para establecer los modelos matemáticos con correlaciones cuantitativas, definiendo de este modo los descriptores moleculares relevantes para el modelo y elegir el mejor de ellos, lo cual es un paso crucial para predecir la actividad biológica y mejorar modelos matemáticos [15, 19].

Los descriptores moleculares utilizados pueden contener información global de la molécula (características estructurales) o contener información enfocada en

características locales de la molécula (subestructuras). Estas propiedades pueden ser experimentales, calculadas o una combinación de ambas; por ejemplo, las propiedades fisicoquímicas y las similitudes o características compartidas con otras moléculas.

Los modelos *in silico* han sido utilizados para determinar, exitosamente, las propiedades ADMET y las propiedades que proporcionan información sobre el tamaño y la frecuencia de la dosis a utilizar, así como la absorción oral, la biodisponibilidad, penetración en el cerebro y su volumen de distribución [20].

2.4 Química computacional.

La química computacional es un área muy extensa en la que podemos encontrar diversos métodos y niveles de cálculo entre ellos puede involucrar el uso de modelos matemáticos para predecir las propiedades químicas, biológicas y físicas de los compuestos; para la investigación de átomos, moléculas y macromoléculas mediante modelado molecular, se divide en mecánica molecular y mecánica cuántica [21].

2.4.1 Mecánica molecular

En la mecánica molecular se aplican leyes de la física clásica; utilizando un modelo de una molécula compuesta por átomos que se mantienen unidos por enlaces considerando a la molécula como una colección de partículas. Utilizando constantes de fuerzas de tensión de enlace y de flexión que se comportan como osciladores armónicos, permitiendo las interacciones entre los átomos no enlazados; por lo tanto, el método construye una expresión de la energía potencial; es decir la unión de las posiciones atómicas, además utiliza un campo de fuerza en el cual los efectos electrónicos se encuentran implícitos [15, 21].

2.4.2 Mecánica cuántica

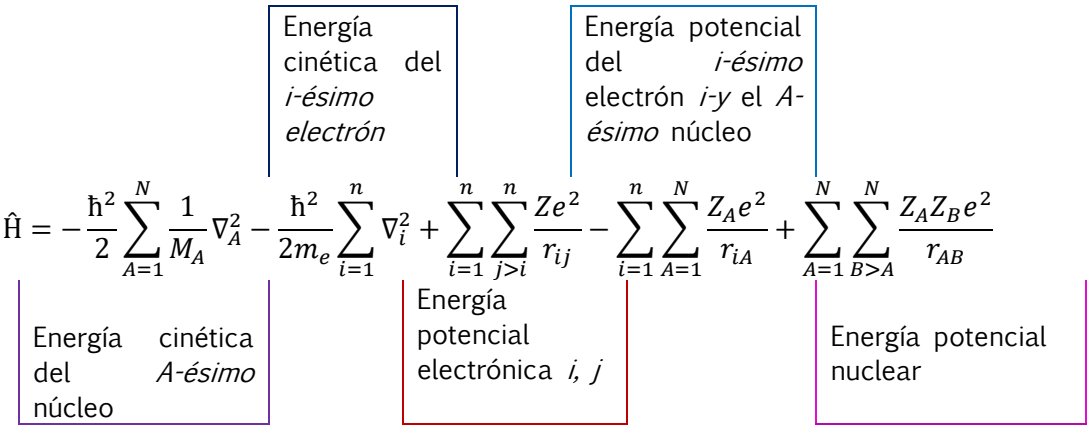
La mecánica cuántica [22] se basa en la resolución de la ecuación de Schrödinger (Ecuación 1), donde los distintos métodos de cálculo de la estructura electrónica se caracterizan por sus distintos niveles de aproximación a la solución exacta de la misma [15].

$$\hat{H}\psi = E\psi(x)$$

Ecuación 1. Ecuación de Schrödinger independiente del tiempo

La metodología utilizada en los métodos teórico-computacionales se realiza utilizando algoritmos basados en propiedades fisicoquímicas de las moléculas para predecir, simular y estudiar sus interacciones con otras moléculas basándose en parámetros obtenidos de manera experimental, estos métodos calculan la geometría molecular y las propiedades electrónicas para realizar los modelos [15].

Para lograr calcular propiedades moleculares como la geometría, la energía mínima, la energía libre de Gibbs y constantes de disociación se utilizan programas como el Gaussian03, el cual permite realizar las estructuras moleculares para su posterior optimización con los métodos *ab initio* y semiempíricos; los cuales se basan en el principio de que los núcleos y los electrones se distinguen unos de otros, las interacciones entre electrón-electrón y electrón-núcleo están dirigidas por el movimiento y la carga de los electrones, los métodos de mecánica cuántica se resuelven mediante aproximaciones de la ecuación de onda de Schrödinger en un sistema de un electrón y un núcleo, para describir la energía cinética y potencial de los electrones y los núcleos de un sistema se utiliza el operador Hamiltoniano $\hat{H}$ (Ecuación 2), la función de onda electrónica, Ψ (Ecuación 3), que describe los movimientos y la posición del electrón para los diferentes estados del electrón. La función de probabilidad la cual es la función de onda al cuadrado Ψ^2 y si se normaliza indica la probabilidad de encontrar un electrón en tal estado, la energía está en particular asociada al estado electrónico [15].



Ecuación 2. Expresión matemática del operador Hamiltoniano para N-núcleos y n-electrones. En general se divide en una parte de energía cinética y una parte de energía potencial.

$$\Psi = a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_Nx_N = \sum_{i=1}^{N} a_ix_i$$

Ecuación 3. Aproximación de la función de onda como una combinación lineal de N-orbitales atómicos.

Uno de los principales inconvenientes de realizar cálculos *ab initio* es el tiempo de cómputo utilizado. Una buena opción son los métodos semiempíricos, los cuales surgieron debido a las dificultades que presentan los métodos *ab initio* al utilizarlos para las moléculas medianas y grandes y la necesidad de describir cualitativamente la molécula (orbitales moleculares, cargas atómicas, vibraciones, geometrías moleculares, momento dipolar, etc.) son ampliamente usados debido a que su base es reproducir datos experimentales. Estos métodos los podemos clasificar en tres tipos: 1) los métodos que tratan solamente los electrones π, 2) los que utilizan la aproximación ZDO (Métodos de Pople) y 3) los que utilizan la aproximación NDDO (Métodos de Dewar).

- Tratamientos OM semiempíricos de moléculas conjugadas planas

Los orbitales de una molécula orgánica no saturada plana se pueden dividir en Orbitales Moleculares σ y π. Los primeros métodos semiempíricos para compuestos orgánicos conjugados planos trataban los electrones π separadamente de los electrones σ, basado en la simetría de los orbitales y en la mayor polarizabilidad

de los electrones π que los hace más susceptibles a perturbaciones tales como las que ocurren en las reacciones químicas.

Las principales teorías OM π electrónicas son:

- Método OM del electrón libre

La teoría π-electrónica más simple desarrollada en 1950; en este método se ignoran las repulsiones interelectrónicas del tipo $1/r_{ij}$, y el efecto de los electrones σ se representan por la función de energía potencial de la partícula en cierta región, con V=0, o fuera de esta región, con V=∞, y la función de onda no tiene en cuenta el espín o el principio de Pauli [23].

Las transiciones π-electrónicas que dan la absorción electrónica de más baja frecuencia implican que un electrón vaya desde el HOMO al LUMO.

En este modelo el desprecio de las repulsiones electrónicas da a lugar a que los términos electrónicos "singulete" y "triplete" tengan la misma energía. La transición electrónica de mayor longitud de onda observada es una transición singulete-singulete, ya que las transiciones singulete-triplete están prohibidas [23].

- Método OM Hückel

Desarrollado en 1930; presenta una aproximación más simple, ya que incorpora los efectos de las repulsiones π-electrónicas como la suma de las energías de los orbitales y de una forma promediada, ignorando las repulsiones electrostáticas, este método es ampliamente usado para predecir propiedades y reactividades de compuestos conjugados [23].

- Método de Pariser-Parr-Pople

Esta teoría π-electrónica toma en cuenta la repulsión electrónica fue desarrollada en 1953; se utiliza un Hamiltoniano que incluye las repulsiones electrónicas y la función π-electrónica que se describe como producto antisimétrico de espin y orbitales π-electrónicos, este método realiza la aproximación de la interpenetración

cero (ZDO) que incluye el desprecio de las integrales de interpenetración cuando se evalúan las integrales de repulsión electrónica, de este modo este método ignora muchas pero no todas las integrales de repulsión electrónica; simplificando los cálculos [23].

La aproximación ZDO se puede justificar interpretando los orbitales atómicos usados para expresar los orbitales moleculares como ortogonalizados en lugar de orbitales atómicos ordinarios; cada orbital en una serie de orbitales ortogonalizados es una combinación lineal de orbitales atómicos ordinarios y los coeficientes se eligen de forma que los miembros de la serie sean mutuamente ortogonales [23].

Este método se utiliza poco en la actualidad ya que ha sido sustituido por métodos semiempíricos más generales, sin embargo, muchas de las aproximaciones de esta teoría para evaluar integrales se utilizan en las actuales teorías semiempíricas [23].

- Método OM Pople

Este método se aplica a todas las moléculas y tratan a todos los electrones de valencia; pertenecen a dos categorías las que utilizan el Hamiltoniano como la suma de términos de un electrón y las que usan un Hamiltoniano que incluye los términos de repulsión de dos electrones como términos de un electrón [23].

Se han desarrollado varios métodos semiempíricos de dos electrones que son aplicables tanto a moléculas planas como no planas, entre ellos podemos mencionar los métodos CNDO e INDO.

Método de desprecio completo la diferencial de interpenetración (CNDO) fue propuesto por Pople, Santry y Segal en 1965, mientras que el método de desprecio intermedio de la diferencial de interpenetración (INDO) fue propuesto por Pople, Beveridge y Dobosh en 1967. En ambos métodos solo tratan explícitamente a los electrones de valencia.

El método CNDO usa una base mínima de orbital atómico tipo Slater, donde la carga de los electrones de valencia (capa interna) es igual al número atómico del

átomo menos el número de los electrones de valencia, a su vez usa ZDO para todas las parejas de orbitales atómicos en las integrales de interpenetración y repulsión electrónica, hay varios orbitales atómicos de valencia sobre cada átomo y la aproximación ZDO desprecia las integrales de repulsión electrónica.

CNDO/1 obtiene la energía de ionización a partir de los niveles de energía atómica deducida con los datos de espectros atómicos; por lo tanto, con esta elección para la integral de atracción electrón-core dos átomos neutros o moléculas separadas por varios angstroms experimentan una atracción mutua.

Por otra parte, el método INDO es una mejora del CNDO este método se basa en que la diferencial de interpenetración entre los orbitales atómicos del mismo átomo no se desprecian en las integrales de repulsión electrónica de un centro, así que se desprecian menos integrales bioelectrónicas que en CNDO esto mejora los resultados CNDO, especialmente donde es importante la distribución electrónica.

Como resultado de todo esto los métodos CNDO e INDO dan buenas longitudes y ángulos de enlace, momentos dipolares algo erráticos y energías de disociación pobres.

Las versiones CNDO e INDO que fueron parametrizadas para predecir espectros electrónicos se denominan CNDO/S e INDO/S; estos incluyen alguna interacción de configuraciones, pese a que el estado fundamental de una molécula de capa cerrada esta generalmente representada por una función de onda de determinante simple.

Actualmente los métodos CNDO e INDO se usan poco debido a que se han vuelto obsoletos comparados con los métodos semiempíricos mejorados, la excepción es INDO/S que se usa ampliamente para realizar los cálculos de espectros electrónicos ya que provee buenos resultados para las energías de excitación vertical de moléculas grandes incluyendo los compuestos de metales de transición.

La aproximación del desprecio de la diferencial de interpenetración diatómica (NDDO) que fue sugerido en 1965 por Pople, Santry y Segal, este método es una mejora del método INDO en el que la diferencial de interpenetración se desprecia solamente entre orbitales atómicos centrados en diferentes átomos; el grado de desprecio de la interpenetración en NDDO es más justificable que en CNDO e INDO, este método satisface las condiciones de varianza rotacional y de hibridación sin necesidad de usar un valor común para las integrales de repulsión electrónica implicando los diferentes orbitales atómicos de valencia sobre un átomo dado [23].

- Método Dewar

El objetivo de J. Pople en los métodos CNDO e INDO era reproducir lo mejor posible los resultados de cálculos *ab initio* de orbitales moleculares utilizando una base mínima con teorías que requerían mucho menos tiempo de un ordenador. Los métodos basados en la filosofía Pople realizan bien la geometría molecular, pero fallan en las energías de enlace. M. Dewar y colaboradores propusieron varios métodos semiempíricos que tienen un gran parecido con los métodos NDDO e INDO; sin embargo, su propósito era diferente: obtener una teoría que pudiera proporcionar las energías de enlace molecular con precisión química y que a su vez pudiera usarse para moléculas grandes sin una cantidad de tiempo de cálculo prohibida. Para lograr esto Dewar basa su metodología en la parametrización de sus métodos por medio de una serie de elementos.

Las teorías tipo Dewar tratan solamente los electrones de valencia y la mayor parte de ellas usan bases mínimas de orbitales atómicos s y p tipo Slater para desarrollar los orbitales moleculares de los electrones de valencia [23].

En este método se ignora la energía vibracional en el punto cero (0 K) y los cambios de entalpia de 0 a 298 K, la justificación de esto es que estas cantidades son despreciables y se ajustan los parámetros para incluir valores experimentales de la entalpia a 298 K de muchos compuestos de forma que se incluyen correcciones para las cantidades despreciadas. La parametrización se basa en

reproducir las entalpias, geometría molecular y momento dipolar a través de datos experimentales, minimizando los errores en calores de formación, geometrías y momentos dipolares calculados. El proceso es parecido al de optimizar una geometría molecular, en la cual se varían las distancias de enlace, ángulos de enlace y ángulos diedros para optimizar la energía electrónica molecular, incluyendo la repulsión nuclear [23].

La primera teoría tipo Dewar útil fue el método MINDO/3 que fue parametrizado para compuestos que contiene C, H, O, N, B, F, Cl, Si, P y S; sin embargo para compuestos que contienen solo C, H, O y N, los errores absolutos medios MINDO/3 son grandes, produciendo errores en los calores de formación de compuestos con anillos pequeños, compuesto con triples enlaces, aromáticos, globulares compactas, compuestos de boro y moléculas con átomos con pares electrónicos solitarios [23].

Las teorías basadas en NDDO tipo Dewar dan mejores resultados que MINDO/3. Debido a esto se desarrolló el método MNDO (Modificación del desprecio de la diferencial de interpenetración diatómica), este método fue parametrizado para compuestos que contienen H, Li, Be, B, C, N, O, F, Al, Si, Ge, Sn, Pb, P, S, Cl, Br, I, Zn y Hg [23].

El método Austin 1 (AM1), es la tercera generación de un procedimiento semiempírico desarrollado por Dewar y colaboradores que toma un enfoque similar a MNDO en la aproximación de las integrales de dos electrones, pero usa una expresión modificada para la repulsión nuclear core-core asignándole a cada átomo un comportamiento de esfera, dejando los elementos de la matriz Fock son prácticamente los mismos, siendo la única modificación la función de la repulsión core. La ecuación modificada resulta en unas fuerzas atractiva que mimetizan las interacciones de van der Waals; la modificación también necesita una reparametrización del modelo, el cual es llevado a cabo con énfasis en el momento dipolar, potenciales de ionización y geometrías moleculares [24-26].

AM1 como PM3 fueron parametrizados para reproducir calores de formación, geometría, momento dipolar y la primera energía vertical de ionización [25].

La distribución de carga electrónica en una molécula se encuentra estrechamente relacionada con un gran número de propiedades o fenómenos observables entre ellos el momento dipolar μ. En general, las medidas del momento dipolar no se usan para obtener las longitudes de enlace, ni para conocer con exactitud la separación de cargas. Sin embargo, el conocimiento de los momentos dipolares de un compuesto resulta muy útil para determinar la conformación molecular e informar acerca de la posición atómica relativa en el espacio de una especie, obteniéndose su simetría, los momentos dipolares obtenidos con AM1 son bastantes confiables mostrando un error de 0.35 D [13].

En 1989 Stewart reparametrizó el método semiempírico AM1 para dar el método PM3. Los principales cambios se observan en las integrales de repulsión electrónica monocéntricas que se toman como parámetros a optimizar y la función de repulsión de "core" contiene solamente dos términos gausianos por átomo [23].

Posteriormente en 1993 Dewar y colaboradores modificaron el método semiempírico AM1 para obtener el método SAM1 (modelo semi *ab initio*). La diferencia más importante entre SAM1 y AM1 es que SAM1 evalúa las ERI usando una base STO-3G, además que es más lento que AM1 aunque es más rápido que otros cálculos *ab initio* por la aproximación NDDO [23].

La limitación más importante de las versiones originales de los métodos MNDO, AM1 y PM3 es el uso de bases de OA de valencia *s* y *p* solamente de manera que no eran útiles para metales de transición [23].

Thiel y Voityuk resolvieron este problema al extender y reparametrizar el MNDO para incluir métales del bloque d y cuando se forma un compuesto con C, H, O y N [23].

Los métodos semiempíricos están disponibles en diferentes programas, entre ellos Gaussian que incluye MNDO, AM1, PM3, MINDO/3, INDO Y CNDO [23].

El método semiempírico PM3 esta parametrizado para átomos que presentan las diferentes moléculas analizadas como C, H, N, O, Br y S; siendo los más importantes debido a que se encuentran en las estructuras analizadas pero puede aplicarse este mismo nivel de parametrización a B, F, Cl, Si, P, Li, Be, Al, Ge, Sn, Pb, I, Zn, Hg; una parametrización adecuada da la capacidad de tratar las moléculas y las propiedades calculadas utilizando información experimental o teórica para obtener los valores de los parámetros, de este modo se minimizan los errores en la longitud y ángulos del enlace por otro lado debido a que las constantes de fuerza de flexión y tensión de enlace no pueden medirse experimentalmente se describe una expresión aproximada para la energía potencial vibracional molecular; a partir de una suma cuadrática de la tensión de enlace, flexión de ángulos de enlace y la torsión [23, 27].

Tanto AM1 como PM3 son métodos implementados para utilizar integrales estándar; además aumenta la eficiencia y este cambio proporciona gradientes y frecuencias analíticas [28].

Para seleccionar el método adecuado se debe tomar en cuenta la disponibilidad de recursos computacionales, la calidad de las soluciones requeridas y la parametrización del método [15].

2.4.3 Optimización geométrica

La optimización geométrica comienza con el cálculo de la energía en un punto único, posteriormente se cambian las coordenadas para el conjunto de átomos y se recalcula un nuevo punto de energía, con el fin de determinar la energía de la nueva conformación, la primera o segunda derivada de la energía (grados de libertad geométricos) dependiendo del método, con respecto a las coordenadas atómicas que determinan cuanto y en qué dirección se debe cambiar el siguiente incremento de geometría. A continuación, se determina de nuevo la energía y sus derivadas y el proceso se repite y continúa hasta que se alcanza la mínima energía.

Este procedimiento puede presentar el inconveniente de conducir al sistema hacía mínimos de energía próximos a la posición de partida (mínimos locales), los cuales pueden no coincidir con el mínimo global correspondiente a la geometría óptima [15, 29].

Cuando se encuentra un punto estacionario en la Hipersuperfice de Energías Potenciales se localiza la conformación más estable de una molécula utilizando la optimización geométrica; los puntos críticos de una Hipersuperfice de Energías Potenciales son:

Mínimo Global: Energía más baja, indica la conformación más estable [15].

Mínimo local: Constituyen regiones donde un cambio en la geometría en cualquier dirección nos da una geometría de mayor energía [15].

Punto de silla: Punto entre dos energías extremas; definimos a silla como el punto en la Hipersuperfice de Energías Potenciales (Figura 4), en el cual la energía incrementa en todas direcciones excepto una, correspondiente a la pendiente primera derivada de la superficie que le otorga un valor de cero [15].

Obtenemos la minimización de la energía a través del mínimo de energía global y local; en el caso de la determinación del punto silla nos lleva a un estado de transición, la capacidad de una optimización geométrica para converger a un mínimo depende de la geometría de partida, de la función de energía de potencial usada, y de las condiciones impuestas para conseguir un gradiente mínimo aceptable [15].

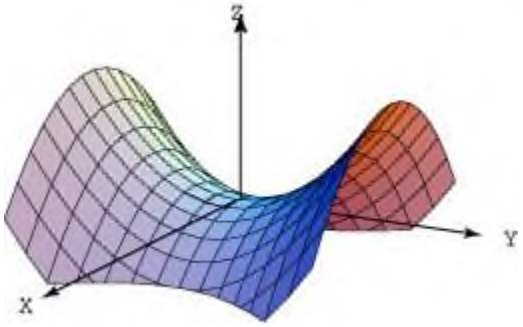


Figura 4: Representación de un punto de silla.

2.4.4 QSAR

Diferentes disciplinas de la ciencia se integran con el fin de complementarse, tal es el caso de la estadística que se integra con la inteligencia artificial (redes neuronales artificiales), con el objetivo de analizar los datos existentes de una manera óptima y de esta manera extraer específicamente la información o datos necesarios para el análisis de la relación cuantitativa entre la actividad biológica y la estructura química (QSAR por sus siglas en inglés) usando datos cuantitativos clasificados por algún proceso estadístico [30].

De este modo, la creación de descriptores químicos es una herramienta importante para el análisis QSAR, el cual permite predecir las propiedades tóxicas y características perjudiciales, así como las propiedades biológicas y farmacológicas de un conjunto de compuestos que pueden ser medidas experimentalmente, pero necesitarían mucha demanda en costos y tiempo [31, 32].

La base de los estudios QSAR es utilizar una serie de moléculas que presentan algún tipo de actividad biológica, bajo la premisa de que la actividad biológica que presenta un compuesto es función de una serie de parámetros fisicoquímicos, conocidos como descriptores.

Hansh y Fujita introdujeron la metodología QSAR en 1960, con la finalidad de construir un modelo matemático que prediga si un compuesto es activo o no, si es selectivo, si presentará propiedades farmacocinéticas ADMET o sí presentará toxicidad, todo esto para mejorar nuestro conocimiento sobre las relaciones estructura-actividad. Así mismo, es posible identificar los compuestos peligrosos a un bajo costo y una reducción del número de animales usados para las pruebas experimentales de toxicidad introduciéndose como diseño racional en el proceso de fármacos en los años 80's. Con el paso del tiempo, estos estudios se consideraron como una metodología basada en el uso de series de compuestos derivados de una estructura común, en los que se observan diferentes respuestas biológicas a partir de distintos sustituyentes. En el campo de los estudios QSAR y en la predicción de la toxicidad se han propuesto y desarrollado diferentes técnicas

de inteligencia artificial, tales como las redes neuronales artificiales, que permiten mejorar la predicción en las moléculas de interés [15, 19, 20, 31, 33, 34]

El análisis QSAR se realiza a través de diferentes pasos. El primero consiste en construir un modelo a partir de un conjunto de moléculas con actividad biológica conocida (también llamada conjunto de entrenamiento), posteriormente se realiza la optimización de dichas estructuras moleculares con el fin de obtener y seleccionar los descriptores moleculares que presenten la información relevante que describa a la actividad biológica de interés.

Aunque en principio cualquier descriptor puede ser utilizado para proponer un modelo, es necesario establecer una serie de técnicas para la selección de descriptores a partir de un conjunto de descriptores. Una técnica simple y más utilizada es el método de correlación que selecciona aquellos descriptores que presenten un alto coeficiente de correlación. Una vez seleccionados los descriptores se realiza un análisis de datos para encontrar la ecuación matemática que mejor describa la actividad biológica y por último se realiza la validación de compuestos no utilizados en la realización del modelo con actividad biológica conocida para validar y verificar que el modelo prediga de manera correcta la actividad biológica de interés. Esta técnica asegura que el modelo se puede utilizar para predecir actividades de un conjunto más extenso de moléculas.

La validación del modelo estadístico se puede realizar por dos métodos: validación interna y/o validación externa. En el método de validación interna se utilizan los datos a partir del cual se construyó el conjunto de entrenamiento, dentro de esta técnica la más utilizada es el método de validación cruzada. Por otra parte, la validación externa parte de la idea de generalizar la construcción correcta del modelo cuando se obtiene una cantidad amplia de datos y pueden ser divididos en un conjunto de entrenamiento y en un conjunto de ensayo o de prueba, donde se buscará un modelo de calibración para predecir las actividades de las moléculas en el conjunto de ensayo [31].

La ecuación de Hansch y Fujita (Ecuación 4) correlaciona la actividad biológica con las propiedades o descriptores fisicoquímicas los cuales constituyen las variables independientes del modelo.

$$\log\frac{1}{C} = -k(log\ P)^2 + k'(log\ P) + s + \cdots k''$$

Ecuación 4: Ecuación de Hansch y Fujita

donde *C* es la concentración a la que produce efecto la actividad biológica; *P* es el coeficiente de partición Octanol-Agua, s es la constante electrónica de Hammett y k, k', k'' son constantes cuyos valores se determinan por análisis de regresión u otros métodos estadísticos.

Los descriptores que sirven para describir las propiedades fisicoquímicas se pueden catalogar como descriptores constitucionales, topológicos, electrostático, geométrico o cuánticos; eligiendo de manera adecuada estos descriptores y el modelo se puede describir cuantitativamente estas variables o datos biológicos, el análisis de los datos biológicos y fisicoquímicos se pueden analizar de manera estadística de dos maneras: de forma conjunta o separada con métodos clasificatorios de reconocimiento de tendencias y utilizando técnicas de regresión lineal [15, 31, 34].

Los descriptores fisicoquímicos reflejan la interacción molecular con su receptor y sus propiedades globales (estéricas, electrónicas, hidrófobicas) por otro lado los descriptores cuánticos expresan las propiedades electrónicas y geométricas de las moléculas relacionadas con las interacciones estéricas, electrónicas e hidrofóbicas, a partir de los orbitales moleculares frontera HOMO-LUMO (HOMO, acrónimo en inglés para el orbital molecular ocupado de más alta energía, mientras que LUMO es el acrónimo en inglés para el orbital molecular desocupado de más baja energía).

Para obtener resultados exitosos en el desarrollo de un modelo QSAR debemos utilizar la selección adecuada de datos o propiedades biológicas que formaran los conjuntos de entrenamiento requeridos para la formación de modelos convenientes como por ejemplo las variables dependientes, de la misma manera a pesar del dominio que se presenta en el campo por usar análisis de regresión múltiple es

ampliamente recomendable el uso de redes neuronales artificiales las cuales son herramientas valoradas debido a su capacidad para superar problemas estadísticos; mejorando por mucho los resultados de los modelos creados [35].

2.4.4.1 Métodos estadísticos

Una regresión multilineal deriva de una regresión lineal simple, la regresión lineal simple se obtiene de manipular dos variables; una variable dependiente (la actividad biológica) y una variable independiente (algún descriptor). Para el caso de la regresión multilineal se utiliza un número de variables independientes para observar su influencia sobre la variable dependiente. Al igual que en la regresión lineal simple, se considera que los valores de la variable dependiente se generan por una combinación lineal de los valores de una o más variables (descriptores) moleculares [31, 36, 37].

Los coeficientes son elegidos de forma que la suma de los cuadrados de la diferencia de los valores experimentales y los valores pronosticados (residuos) sea mínima, de manera que se minimiza la varianza residual para determinar estos coeficientes [36, 37].

Este tipo de ecuación recibe el nombre de hiperplano, pues cuando tenemos más de dos variables independientes en vez de una recta de regresión tenemos un plano, si tenemos tres variables independientes tendríamos un espacio de cuatro dimensiones y así sucesivamente (Figura 5), un ejemplo de este tipo de ecuación es la siguiente (Ecuación 5):

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + u$$

Ecuación 5: Ejemplo de una regresión multilineal

donde u es un término de error que incluye los residuos, errores de media y varianza, b_i son coeficientes desconocidos y x_i son los descriptores utilizados.

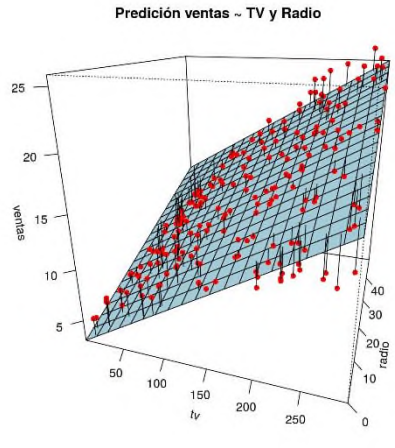


Figura 5: Regresión multilineal del peso de una persona en función de 2 variables independientes (tv y radio). En la imagen se observa un plano (en azul) que describe el comportamiento de la variable dependiente (Ventas, círculos en rojo) [37].

En la práctica se debe tener cuidado al elegir a los descriptores en el modelo; por lo tanto, se pueden tomar en cuenta los siguientes criterios:

- Tener Sentido numérico.
- No repetir descriptores.
- Los descriptores introducidos deben tener justificación teórica.
- Los descriptores deben tener relación con el modelo.
- Los descriptores independientes deben ser lineales con las variables dependientes.

El coeficiente de correlación lineal es la medida numérica de la relación lineal entre dos variables; refleja la consistencia del efecto que un cambio en una variable tiene sobre la otra; el coeficiente de correlación lineal siempre tiene un valor entre -1 y +1. Un valor de +1 indica una correlación positiva perfecta y un valor de -1 indica una correlación negativa perfecta. Si x aumenta y se presenta un aumento general de y será entonces una correlación positiva la línea de mínimos cuadrados tiene una pendiente hacia arriba; por otro lado si el valor de x aumenta pero el valor de y disminuye, la relación resulta negativa, la línea de cuadrados mínimos cuadrados tiene una pendiente hacia abajo [38].

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

Ecuación 6: Fórmula de cálculo para el coeficiente de correlación.

La función de correlación cruzada es la correlación entre las observaciones de dos series de tiempo x_t y y_t separadas por unidades de tiempo k (la correlación entre y_{t+k} y x_{t+k}. La correlación cruzada sirve para determinar si existe una relación entre dos series de tiempo. Para determinar si existe una relación entre las dos series, se debe buscar una correlación grande; normalmente una correlación se vuelve significativa cuando el valor absoluto es más grande que $\frac{2}{\sqrt{n-|k|}}$; donde n es el número de observaciones y k es el retraso (lag: es el número de periodos de tiempo que separan las dos series de tiempo, el número determinado de rezagos varía desde $-\sqrt{N}+10$ a $\sqrt{N}+10$). Este cálculo es una regla general basada en la aproximación normal de muestra grande. Si la correlación cruzada poblacional de retraso k es cero para k=1, 2… entonces para una n grande $r_{xy}(k)$, tendrá una distribución aproximadamente normal, con una media (μ) cero y una desviación estándar (σ) $\frac{1}{\sqrt{n-|k|}}$ [39].

Se puede expresar la correlación cruzada de dos señales en términos del coeficiente de correlación cruzada, eso se calcula normalizando la correlación cruzada ajustando m=0; el coeficiente de correlación se encuentra entre -1 y +1, el valor de cero indicaría que no existe correlación entre las señales [40].

La técnica de error de dejar uno afuera (LOO), es un caso especial de la clase general de métodos de estimación de error de validación cruzada. En la validación cruzada los casos se dividen aleatoriamente en *k* particiones de prueba igual de excluyentes de aproximadamente el mismo tamaño; los casos que no se encuentran en cada partición de prueba se utilizan de forma independiente para el entrenamiento y el clasificador resultante se prueba en la partición de prueba correspondiente. Las tasas de error promedio sobre todas las particiones *k* es la estimación de error de validación cruzada. Este método tiene un costo computacional alto por lo tanto es más utilizado en muestras pequeñas debido a

que para una muestra de tamaño n, un clasificador es generado utilizando n-1 casos y son analizados los casos remanentes; esto es repetido n veces. Cada caso es usado como un caso prueba y cada vez se acerca más a que todos los casos sean usados, el error estándar es el número de errores sobre los errores individuales de los casos divididos por n y la idea de separar la muestra en diferentes casos en entrenamiento y prueba es ayudar al clasificador con una tasa de error mínimo [41].

Un análisis de regresión genera una ecuación para describir la relación estadística entre uno o más descriptores y la variable de respuesta para predecir nuevas observaciones. La regresión lineal generalmente utiliza el método de estimación de mínimos cuadrados ordinarios, del cual se obtiene la ecuación a minimizar la suma de los residuos al cuadrado.

La regresión lineal múltiple examina las relaciones lineales entre una respuesta continua y de dos o más descriptores. Si el número de descriptores es grande, antes de ajustar un modelo de regresión con todos los descriptores, se deberían utilizar las técnicas de selección de modelo paso a paso o de los mejores subconjuntos para excluir los descriptores que no estén asociados con la respuesta [42].

Se utiliza una estandarización para centrar las variables, para escalar las variables o para ambas. Cuando se centran las variables, se reduce la multicolinealidad causada por los términos polinómicos y los términos de interacción, que mejoran la precisión de las estimaciones de los coeficientes; esto se puede realizar por el método: Restar la media y dividir entre la desviación estándar; este método centra y escala las variables. Cada coeficiente representa el cambio esperado en la respuesta ante el cambio de una desviación estándar en la variable y colocarlos en una escala comparable mejorando la interpretación del modelo [43].

En el contexto de los análisis de ajuste del modelo, los valores atípicos son observaciones con valores de descriptores mayores al promedio; es importante

identificar los valores atípicos, porque estos pueden afectar significativamente el modelo, proporcionando resultados potencialmente engañosos o incorrectos. Si se identifica un valor atípico en los datos, debe examinar la observación para determinar porque se trata de un valor poco común e identificar la solución.

El análisis por apalancamiento mide la distancia desde el valor de X de una observación hasta el promedio de los valores de X de todas las observaciones incluidas en un conjunto de datos, se utiliza para identificar observaciones que tengan valores predictores pocos comunes en comparación con los datos restantes.

Las observaciones con apalancamiento grande pueden tener un gran efecto sobre el valor ajustado y por lo tanto sobre el modelo de regresión [44].

La desviación estándar es la medida de dispersión más común que indica que tan dispersos están los datos con respecto a la media; mientras mayor sea la desviación estándar, mayor será la dispersión de los datos y se puede utilizar para establecer un valor de referencia para estimar la variación general de un proceso [45].

2.4.4.2 Redes neuronales

En 1947, el matemático John Von Neumann diseñó una estructura computacional que se basa en procesamiento riguroso secuencial de datos e instrucciones para llevar al cabo el almacenado de datos en la memoria. La arquitectura que utilizó Von Neumann se basa en la lógica de procedimientos que normalmente utilizamos para soluciones parciales a algún tipo de problema, que culminan a resolver el problema concreto con la solución final [46].

Las redes neuronales artificiales o sistemas neuronales artificiales son campos multidisciplinarios, debido a la amplia contribución de diferentes áreas tales como física, matemáticas, biología e ingeniería entre otras. Este tipo de sistema emula el comportamiento del cerebro humano específicamente las redes neuronales, esto hace que el sistema sea altamente complejo, la unidad análoga a la unidad biológica es el elemento procesador. De manera general este elemento tiene varias entradas y las combina normalmente como una suma básica. La suma de las entradas es

modificada por una función de transferencia y el valor de salida de esta función de transferencia se pasa directamente a la salida del elemento procesador por lo que los modelos de Redes Neuronales Artificiales son dirigidos a partir de los datos, es decir, son capaces de encontrar relaciones o patrones de forma inductiva por medio de algoritmos de aprendizaje basado en los datos [15, 47].

Una cualidad importante de las redes neuronales artificiales es su adaptabilidad dinámica o su capacidad para variar su comportamiento en diferentes situaciones y con diferentes problemas, esto se debe a que utiliza técnicas inspiradas en el aprendizaje, generalización o auto-organización utilizando las unidades elementales de procesamiento, su comportamiento en general determina su capacidad para ensayar hipótesis, detectar patrones estadísticos y regularidades, así como ajustar un modelo implícito que sea implementado en la arquitectura de la red, lo cual no depende de la suma de los potenciales de las neuronas. Las redes neuronales son modelos no lineales puesto que la función de activación es no lineal, son modelos paramétricos y los parámetros corresponden a los pesos de las conexiones entre neuronas, son modelos adaptativos, puesto que la aparición de nuevos datos permite el reaprendizaje de los parámetros adaptando los valores anteriores a los datos actuales, son tolerantes a los fallos ya que su comportamiento está distribuido entre todos los parámetros además tienen la capacidad de aproximar funciones con un grado preciso [16, 48].

De manera formal una red neuronal artificial puede explicarse a partir del concepto de grafo, este es un objeto integrado por un conjunto de nodos (vértices) y de conexiones (links) que puede ser dirigido cuando todas las conexiones tienen asignado un sentido y no dirigido cuando las conexiones son direccionales; los grafos densos son aquellos que tienen todos sus nodos conectados entre sí y los grafos dispersos tienen sus conexiones escasas; estos grafos pueden componerse de diferentes tipos de conexiones y nodos [49].

Una manera de representar los grafos es con círculos para los nodos y líneas o flechas para las conexiones como se muestra en la Figura 6; las redes neuronales artificiales normalmente cumplen con ciertas propiedades y características dentro de sus elementos principales que son la neurona artificial, la arquitectura de las redes neuronales artificiales y los modos de operación de las redes [49].

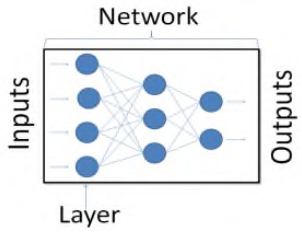


Figura 6: Representación de redes neuronales artificiales de tres capas.

Las redes neuronales artificiales, en general, se forman por un conjunto de procesadores elementales que se denominan neuronas artificiales, estás constituyen los dispositivos simples de cálculo que, a partir de información de otras neuronas o entrada de información del exterior, que proporcionan una única respuesta (salida). Las neuronas artificiales las podemos clasificar como neuronas artificiales de entrada (reciben la información del exterior), neuronas artificiales de salida (envían la información analizada al exterior) y neuronas artificiales ocultas las cuales responden o transmiten impulsos que lleva a actuar a otras neuronas presentando una ampliación de la transmisión entre neuronas cuando se repite la conexión entre ellas [48, 49].

El éxito en el uso de las redes neuronales para resolver conflictos en diferentes áreas se debe en gran sentido a la facilidad con la que una red neuronal artificial adquiere información o conocimiento de algún problema, a su vez la capacidad de almacenar la información ya adquirida y la accesibilidad rápida y eficiente a la información para su posterior uso y por supuesto su eficacia para resolver problemas incluso el de resolverlos a partir de subsoluciones, el conflicto surge cuando tratamos de comprender por qué da cierta soluciones a los problemas, la

capacidad para determinar cierto conjunto de condiciones y su contribución a mejorar todas las utilidades [50].

2.4.4.3 Métodologia QSAR utilizada

Nuestro grupo de trabajo ha publicado una serie de artículos [51, 52, 53, 54, 55, 56] en donde se ha establecido la metodología general de los artículos QSAR (QSPR) que se ha implementado en esta tesis, que se presentaran como antecedentes.

La metodología aplicada se muestra en la Figura 7. Inicialmente las moléculas bajo estudio fueron construidas como especies neutras, posteriormente se optimizan utilizando los métodos semiempiríco AM1 o PM3, se utilizan estos métodos de cálculo debido a la cantidad de compuestos estudiados. Una vez obtenido un mínimo de energía se realiza el análisis conformacional para cada compuesto con el objetivo de localizar su mínimo global. La naturaleza del punto estacionario se observa mediante los cálculos de los modos vibracionales.

Utilizando el mínimo global, se calculan los descriptores moleculares por medio del programa DRAGON [57], obteniendo un total de 1200 a 1500 descriptores en 19 tipos diferentes por cada molécula, incluyendo los descriptores cuánticos calculados con el método semiempiríco.

Una vez hecho esto, el grupo de moléculas se separaron en los diferentes tres conjuntos de predicción, validación y entrenamiento.

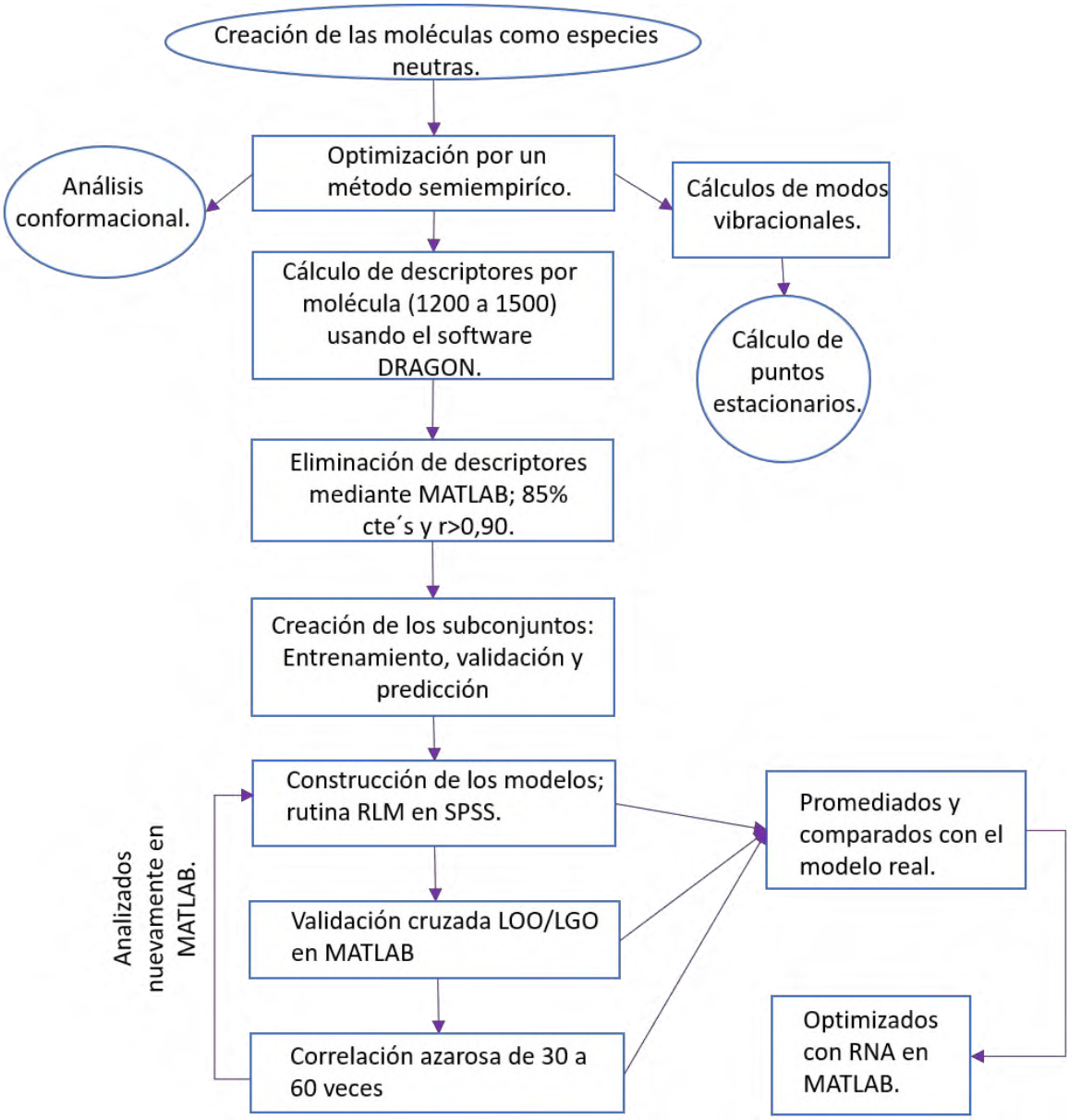


Figura 7. Metodología utilizada en el análisis QSAR de nuestro grupo de trabajo.

Utilizando una subrutina de MATLAB [58] se eliminaron aquellos descriptores que presentaron un porcentaje de 85% de constantes o variables constantes cercanas y una correlación alta $r > 0.90$.

De igual manera la construcción de los modelos de entrenamiento se realizó con el programa SPSS, utilizando la rutina RLM escalonada. Los modelos prometedores fueron sometidos a validación cruzada bajo el procedimiento LOO y LGO, ambas escritas con el programa MATLAB.

Los mejores modelos fueron sometidos a una prueba de correlación azarosa, en el cual los valores de la respuesta biológica fueron mezclados aleatoriamente y examinados nuevamente por el método de RLM para determinar la correlación. Este procedimiento se debe repetir varias veces, en este caso se repitió entre 30 a 60 veces. Esta metodología también se realizó con el programa MATLAB.

En todos los casos, los modelos RLM aceptados fueron optimizados bajo la metodología de Redes Neuronales Artificiales utilizando el programa MATLAB, bajo una estructura de *NX2NX1*, donde N es el número de descriptores obtenidos en el mejor modelo RLM.

3. Justificación

Con el fin de aprovechar los diferentes conocimientos en química teórica y los conocimientos obtenidos en la literatura de los efectos de diferentes compuestos heterocíclicos se pretende encontrar una buena actividad de estos compuestos heterocíclicos en la inhibición de trombina y lograr explicar con cálculos de simulación molecular, la interacción de estos compuestos heterocíclicos y el centro activo de la trombina de la manera más real posible y de esta manera proponer un mecanismo en la disminución de los efectos adversos a los organismos; esto debido a que el uso de estructuras cristalográficas en los cálculos realizados pueden generar resultados ilegítimos o falsos, ya que el entorno químico en estado sólido no coincide realmente con su comportamiento real y esto se ve reflejado en la estructura proteica y particularmente en la conformación [59]. En este análisis se pretende realizar cálculos que realmente representen el comportamiento de dicha enzima ante un conjunto de moléculas que hasta este punto se han comportado como inhibidoras en dicho estado cristalográfico.

4. Planteamiento del problema y discusión

Los inhibidores de trombina actuales presentan efectos adversos a la salud, por lo tanto se ha buscado una alternativa viable para estos fármacos, los cuales son

compuestos heterocíclicos inhibidores de trombina, pero a su vez estos inhibidores presentan inconvenientes en su biodisponibilidad oral ya que aunque se han encontrado buenos resultados de aplicación de manera *in vitro* son demasiado reactivos para considerarse un fármaco, esto podría interpretarse para llevar a cabo una posterior solución con un modelo QSAR adecuado. Estas adecuaciones se pueden realizar a partir de los estudios de dinámica molecular previos, en donde se muestran la evolución conformacional de la trombina en el tiempo, lo cual permitirá realizar cambios al modelo QSAR a partir de estas "mejores" conformaciones de los 50 inhibidores.

5. Hipótesis

El cambio conformacional de una enzima deja en disposición o bloquea el sitio activo, permitiendo la formación o no del complejo enzima-molécula activa. Los compuestos que poseen las mejores propiedades antitrombótica y biológicas para actuar como fármaco permitirán establecer cuáles son las interacciones entre el centro activo y la molécula heterocíclica como una función del tiempo a escala de nanosegundos, permitiendo de esta manera, escoger cual es el mejor compuesto biodisponible.

6. Objetivos

Encontrar las relaciones QSAR que reaccionen adecuadamente con el sitio activo de trombina en función del tiempo.

Establecer un modelo QSAR en función del tiempo para la interacción Enzima-molécula activa para inhibidores de la trombina.

7. Infraestructura

El presente estudio fue realizado en el Laboratorio de Química Teórica en la Universidad del Papaloapan Campus Tuxtepec.

Ordenador

Es una máquina capaz de procesar la información siguiendo las instrucciones que le han sido proporcionadas por el usuario. A este conjunto de instrucciones se les denomina programa, para realizar sus tareas los ordenadores necesitan de dos elementos el hardware, o elementos físicos tangibles y el software formado por los programas que lo hacen gestionar los elementos físicos en la realización de las tareas.

En este caso se utilizó un ordenador Dell Modelo: Optiplex 755, Evaluación: 3.3, Procesador: Intel (R) Pentium(R) Dual CPU E2180 @ 2.00 GHz 2.00 GHz, Memoria (RAM): 3.00 GB. Sistema operativo Windows Vista ™ Bussiness, Copyright © 2007 Microsoft Corporation. Service Pack 2, a 32 bits.

MATLAB

Este programa se utiliza en computación numérica y visualización de datos; se caracteriza por la simplicidad que se puede resolver problemas en matemática aplicada, física, química, ingeniería, finanzas entre otras, ya que se basa en un software de matrices para el análisis de ecuaciones [58].

En el presente proyecto se utilizó el programa MATLAB R2006a Versión 7.2.0.232 (R2006a).

DRAGON

Es un software que proporciona al usuario una variedad de descriptores moleculares, derivados de diferentes representaciones moleculares permitiendo al usuario escoger los descriptores moleculares que son más adecuados para su investigación específica [57].

Actualmente DRAGON incluye en sus cálculos 1664 descriptores y han sido desarrollados para trabajar tanto en Linux como en Windows. Hay dos versiones para Windows DRAGON profesional que solo puede funcionar en modo autónomo y DRAGON plus, el cual puede funcionar tanto en modo fondo independiente.

DRAGON no fue diseñado como software QSAR ya que solo contiene descriptores moleculares y no realiza análisis QSAR ni optimización geométrica [57].

En el presente proyecto utilizamos DRAGON versión 2.1, propuesto por R. Todeschini, V. Consonni y M. Pavan en 2002.

SPSS

Es un sistema comprensivo de análisis estadístico, puede adquirir datos de cualquier tipo de archivos y utilizarlos para generar informes tabulares, gráficos y diagramas de distribuciones y tendencias, estadísticos descriptivos y análisis estadísticos complejos, aunado a esto SPSS utiliza un lenguaje de comandos, algunas de estas funciones solo se pueden acceder a través de sintaxis de comandos. Este tipo de comandos detallados están disponibles de dos maneras: integrados en el sistema de ayuda global y como un documento independiente de formato PDF [60].

Se utilizó la versión 13.0 para Windows.

GAUSSIAN

Gaussian es un sistema conectado de programas que permite realizar una amplia variedad de cálculos semiempiricos y ab initio de orbitales moleculares.

Es un paquete de programas versátil, ampliamente usado que incluye todos los métodos ab initio comunes, tales como Hartree-Fock, CI, MCSCF, funcional de la densidad, MP o CC y también incluye muchos métodos semiempíricos; algunos paquetes incluyen los métodos de la mecánica molecular [28].

Optimiza la energía, calcula frecuencias vibracionales, propiedades termodinámicas y constantes de apantallamiento RMN, busca estados de transición, calcula MEP e incluye los efectos del disolvente; está disponible en versiones para supercomputador, estaciones de trabajo y computadores personales que trabajan bajo Windows [28].

En este caso se utilizaron las versiones GaussView 4.1.2 y Gaussian 03 para Windows.

OPEN BABEL

Es una caja de herramientas química que ha sido diseñada para poder "leer" casi todas las coordenadas de los compuestos químicos en diferentes tipos de formatos; esta caja de herramientas permite a cualquier persona buscar, convertir, analizar o almacenar datos de modelado molecular, química, materiales de estado sólido, bioquímica o áreas relacionadas. Permite la interconversión de datos químicos de un formato a otro e incluye diferentes tipos de formatos de archivos [61].

Se utilizó la versión Babel 1.3

8. Metas

Encontrar un modelo matemático que describa la interacción de las moléculas de interés con el sitio activo tomando en cuenta 11 conformaciones de la enzima calculada previamente en el intervalo de tiempo de 0.0 a 5.0 ns.

Establecer cuál es la mejor molécula que actúa como inhibidor de la trombina a través de la escala de tiempo

9. Metodología

Los modelos QSAR se realizaron a partir de las conformaciones de 50 derivados del benzimidazol, las cuales fueron obtenidas previamente [3] del análisis de acoplamiento molecular, utilizando el programa AutoDock 4.0.1 en diferentes conformaciones (de 0.0 a 5.0 ns) de la enzima trombina (PDB 1A4W) obtenidas utilizando el programa NAMD 2.6 [65] de dinámica molecular, el campo de fuerzas de Charmmss22 [62] para la proteína y el modelo TIP3 para las moléculas del agua.

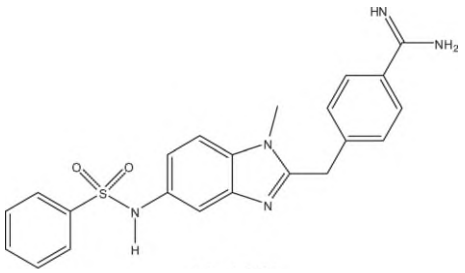


Figura 8. Estructura del benzimidazol

Se realizaron las siguientes actividades, para conocer el modelo que logrará describir de manera acertada el comportamiento, estructura y afinidad de 50 moléculas con respecto a su interacción con la trombina.

Inicialmente se analizó los cambios conformacionales que presentó la trombina durante una dinámica molecular durante 10 ns. En la Figura 9 se observa el cambio del radio de giro en función del tiempo.

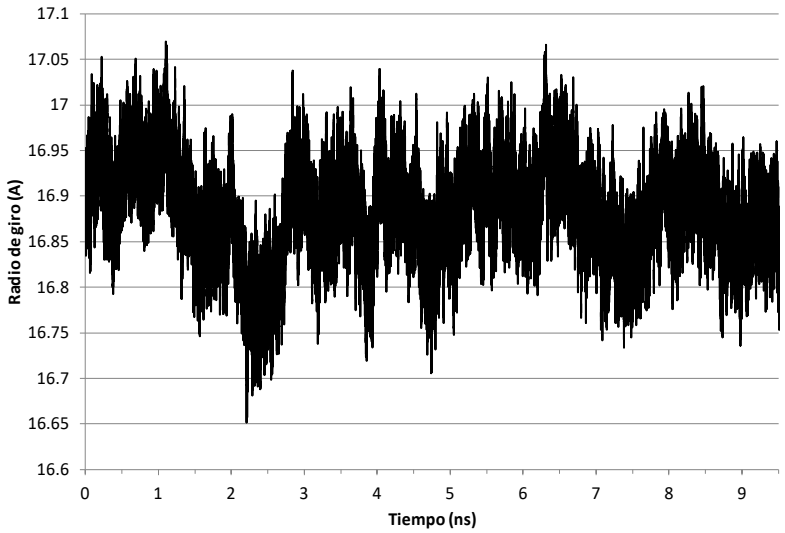


Figura 9. Radio de giro en función del tiempo para la trombina

En la Figura 9 se observan pequeñas variaciones conformacionales de la trombina en los primeros 5.0 ns, después de este intervalo de tiempo la molécula tiende a tener un comportamiento promedio así que podría decirse que llega al equilibrio, por lo que se decide realizar el análisis QSAR durante los primeros 5.0 ns. Para realizar este análisis, se tomaron diferentes conformaciones de la trombina cada 0.5 ns durante 0.0 a 5.0 ns, con estas estructuras se realizó el análisis de

acoplamiento molecular. Finalmente, con estas conformaciones se realizaron los modelos QSAR en función del tiempo.

La molécula de benzimidazol y de un análogo (Figura 10) fueron sustituidas en diferentes puntos; siendo estos sustituyentes grupos funcionales como quinolinas, quinoxalinas, sulfonamidas, carboxamidas, anillos aromáticos y aminas. [51] (ver Anexo 1).

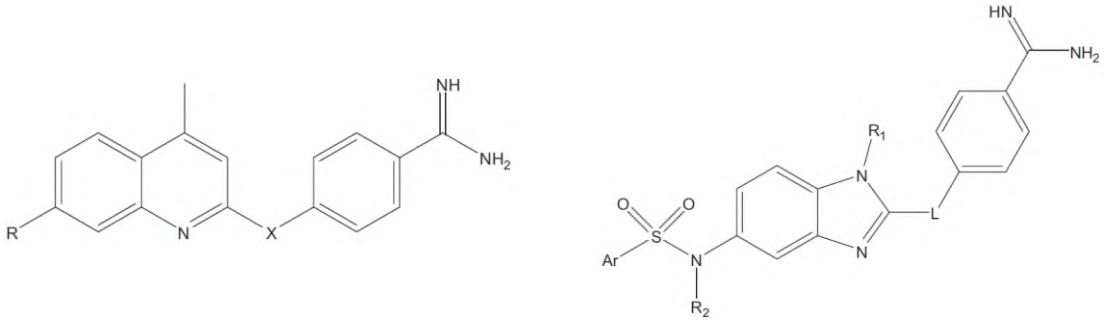


Figura 10. Estructuras básicas utilizadas en el análisis QSAR.

A partir de los archivos generados por el análisis de acoplamiento molecular (*.PDBQ) por cada 0.5 ns, se convirtieron a archivos tipo *.PDB para poder visualizarlos y analizarlos con GausView 4.1.2.

Posteriormente, las estructuras con formato *.PDB se transformaron al formato *.GJF verificando la valencia con la adición de átomos de hidrógeno, a estos archivos se les realizó el cálculo único, utilizando el método semiempírico PM3 que se encuentra en el programa Gaussian 03 obteniendo los resultantes en archivos tipo *.OUT (Figura 11A). Esto permite conservan las características geométricas obtenidas en el estudio de la dinámica molecular previo.

A los archivos *.OUT obtenidos se les realizó dos cálculos diferentes e independientes, el primero es la optimización geométrica para obtener la geometría molecular más estable con menor energía (Figura 11B) y el segundo fue el cálculo de modos vibracionales, Posteriormente a los cálculos de optimización geométrica; también se calcularon los modos vibracionales [29].

El tiempo que tomó realizar cada uno de estos cálculos fue entre 1 a 5 minutos dependiendo la molécula y el tipo de cálculo realizado siendo los cálculos de

frecuencia los más extensos. Estos tiempos se deben al uso del método semiempírico PM3, el cual presenta como una característica y ventaja un corto gasto de tiempo de cálculo.

Para cada molécula se crearon los archivos de entrada dependiendo del tipo de cálculo a realizar. La plantilla del archivo de entrada presenta 4 líneas donde se plantean la información necesaria para realizar los diferentes cálculos (Figura 11); en la primera línea se coloca el controlador chekpoint (*.chk), este controlador crea un archivo donde realizan el cálculo de todas las integrales necesarias en el proceso; en la segunda línea se coloca el tamaño máximo del almacenaje que se ocupará en los cálculos; en la tercera línea se incluyen los controladores sobre el tipo de cálculo, el método y otras opciones de cálculo, finalmente en la cuarta línea se presenta la opción de colocar información de la molécula.

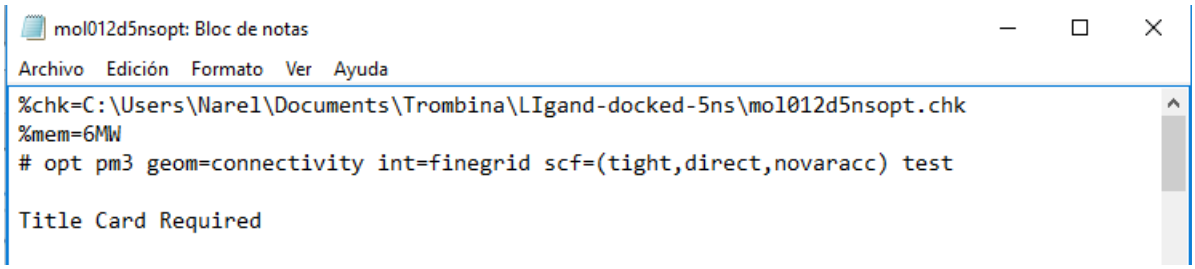


Figura 11: Información en un archivo de entrada.

La cantidad de memoria dinámica para realizar el cálculo se establece por medio del comando %Mem seguido de la cantidad de mega palabras que se desee utilizar; para este caso se utilizaron 6MW [28].

La sección de una ruta de trabajo en Gaussian se inicia con # como el primer carácter de la línea seguido con la información de trabajo; los tipos de cálculos que en este caso se utilizan son energía, frecuencia y optimización.

El controlador OPT realiza la optimización geométrica, por medio de la localización de un punto estacionario, este es el algoritmo por defecto para la optimización a un mínimo local, para optimizaciones a estados de transición y para puntos de silla de orden superior [28].

El controlador FREQ calcula las constantes de fuerza de la molécula y como resultados presenta sus modos de vibración; las constantes de fuerza son calculadas por una sola diferenciación numérica para los métodos en los que solo la primera derivada está disponible y por doble diferenciación numérica para aquellos métodos que solo disponen de energías. Los modos vibracionales son calculados por la segunda derivada de la energía con respecto coordenadas nucleares cartesianas y luego transformada a coordenadas ponderadas en la masa, este tipo de cálculo solo es viable solo para un punto estacionario [28].

Los controladores agregados a los archivos son los siguientes:

- *geom*: Determina las opciones relacionadas con la coordinación del archivo de entrada y la geometría del archivo de salida y controla la información que se imprime de la molécula y el uso de controles internos en la matriz Z; este controlador trabaja en conjunto con otros controladores, en este caso se utiliza el controlador *connectivity* [28, 63].
    - *connectivity*: Especifica explícitamente el enlace de los átomos a través de una sección de entrada adicional respetando a la especificación de la geometría; este controlador requiere una línea de entrada por átomo específicamente ordenado a la molécula y especificar al átomo enlazado y al orden de enlace [28].
- *int:* El controlador integral modifica el método de cálculo con el uso de integrales y sus derivadas de dos electrones, debe utilizarse la misma red para aquellos cálculos que deseen ser comparados [28].
    - *finegrid*: Es utilizada por defecto en el controlador Integral; es una red que ha sido optimizada para usar un mínimo de puntos requeridos para determinado punto de precisión; tiene 7000 puntos por átomo [28].
- *scf*: Controla el funcionamiento del procedimiento del SCF (Self Consistent Field) [28].
    - *tight:* Es el controlador utilizado por defecto para el sistema *scf;* revela el uso de convergencia utilizada [28].

- *Direct:* Opción de almacenamiento de integrales; realiza un cálculo directo SCF, en el que las integrales de dos electrones se recalculan según sea lo necesario; este es un controlador por defecto de gaussian y es factible para diferentes métodos excepto orbitales complejos [28].
- *novaracc:* Comienza utilizando integrales modestas para posteriormente utilizar integrales más exactas en SCF [28].

test: Suprime la creación automática de un archivo de entrada [28].

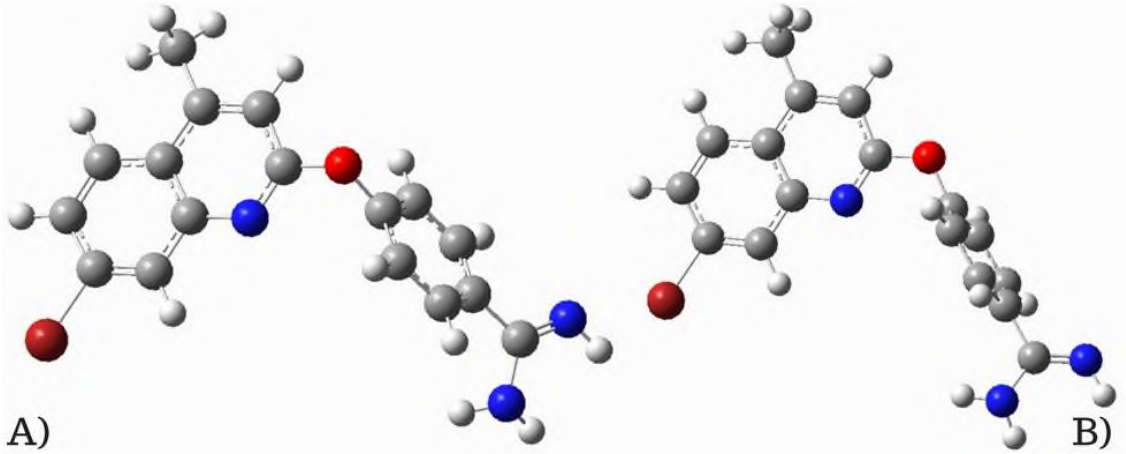


Figura 12. Estructura molecular de 2-[-4-Oxibenzamidina]-7-Bromo-4-Metilquinolina (código mol121) en 0.0 ns en las diferentes etapas del cálculo: A) Cálculo único y B) Cálculo de optimización total. Se observa un cambio conformacional (giro) en anillo aromático cuando se realiza la optimización.

Los archivos optimizados fueron almacenados como *.GJF para convertirlos a coordenadas *.MOL utilizando el programa DRAGON, esto se realiza para lograr obtener los descriptores moleculares que determinen mejor los compuestos moleculares.

Para cada conjunto de moléculas en cada 0.5 ns se obtuvieron alrededor de 1200 descriptores. Además, de estos descriptores se adicionó el logaritmo de la actividad biológica (log IC_{50}) y los descriptores cuánticos obtenidos de los cálculos realizados por el método semiempírico PM3, los cuales están relacionados con la reactividad: la energía mínima, la energía del LUMO, la energía del HOMO, el momento dipolar en las direcciones X, Y, Z y el momento dipolar total.

Continuando con la metodología observada en los antecedentes, se realizó un filtro manual de los descriptores, descartando aquellos descriptores que presentaban valores nulos o presentaban valores constantes en un porcentaje del 20% y seleccionando aquellos descriptores que contenían información relevante relacionada con la estructura, geometría, distribución de carga y las propiedades fisicoquímicas y moleculares, reduciendo la cantidad de descriptores. Este proceso de análisis se realizó las veces que fueran necesarias hasta obtener descriptores que realmente proporcionaran información utilizando el programa SPSS.

Hecho esto se guardó la información de los coeficientes para la construcción del primer modelo mediante la aproximación de algoritmos; los archivos obtenidos en SPSS se guardaron como *.SAV para los datos y *.SPO para el archivo de resultados, se realizó una regresión lineal múltiple con los descriptores seleccionados en la cual los descriptores fueron tomados como variables independientes, mientras la actividad biológica log IC_{50} fue tomada como variable dependiente esto se realizó tanto para el primer modelo como para el modelo en el cual se tomó el conjunto de entrenamiento como modelo matemático basándose en la evolución de una población de modelos a fin de encontrar el más óptimo. Esto se realizó con la finalidad de encontrar los descriptores que tengan una relación directa con el log IC_{50} y que describan el comportamiento de las 50 moléculas al interactuar con la trombina. Los modelos se obtienen a partir de un análisis de correlación y regresión lineal múltiple utilizando mínimos cuadrados ordinarios.

Al realizar esto se almacenaron los datos de los coeficientes para la creación de un modelo preliminar con una cantidad de descriptores mucho menor; estos datos fueron utilizados para crear conjuntos de predicción, validación y entrenamiento utilizando un algoritmo de números aleatorio programado en MATLAB, que indica las moléculas que se integraran a cada conjunto; la cantidad de moléculas para crear cada subconjunto fue respetada del análisis anterior realizado por Ramírez-Galicia y colaboradores [51]. Esto para poder realizar la metodología de investigación

del análisis cuantitativo de la relación estructura actividad biológica y poder obtener un modelo predictivo.

Los datos del conjunto de entrenamiento es el punto inicial, se considera que este conjunto cuenta con la diversidad estructural de todos los compuestos, esto se realiza para que el conjunto de entrenamiento sea eficiente, pues a partir de este conjunto se utilizó para generar un nuevo modelo. Con base de los datos de entrenamiento se vuelve a realizar un modelo con los datos obtenidos en SPSS; así como el uso de Redes Neuronales Artificiales.

Los descriptores encontrados en el conjunto de entrenamiento son analizados por dos metodologías de aplicabilidad de dominio: por estandarización (SA) y por apalancamiento (LA), esto se realizó para cada 0.5 nanosegundo estudiado. En el caso del análisis LA, se utiliza el programa escrito en MATLAB. En este programa se necesitan los descriptores y su número en cada conjunto; esto para obtener los valores de apalancamiento en el entrenamiento (he), predicción (hp) y validación (hv).

Para el análisis de Redes Neuronales Artificiales se utilizó un archivo por nanosegundo que contiene a los descriptores encontrados en la RLM y log IC_{50} correspondientes a cada modelo, para poder ingresarlo a MATLAB; ingresando diferentes algoritmos y datos sobre la secuencia de conteo de los conjuntos de entrenamiento, predicción y validación; realizando los gráficos para visualizar el mejor modelo de los diferentes nanosegundos. Estos datos son utilizados para realizar cálculos sobre la pendiente, intersección con el eje, correlación lineal, desviación estándar y para obtener la R^2.

Posteriormente, se realizaron las regresiones lineales de los modelos obtenidos utilizando como variable independiente los valores de la actividad biológica conocida y los valores obtenidos a partir de redes neuronales artificiales para el conjunto de predicción; los gráficos se realizaron; para verificar que realmente el modelo desarrollado presenta una mejora, comparando los nuevos datos con los revisados

en la bibliografía de Ramírez-Galicia y colaboradores [51]. Los diferentes términos calculados son: R^2, q^2_{LOO}, SA, LA, σ y R^2_{CC}.

10. Resultados y discusión

Se analizaron un total de 550 moléculas las cuales se pueden visualizar en el Anexo 1, 50 moléculas por cada 0.5 ns calculado partiendo desde 0.0 hasta 5.0 ns, usando el método semiempírico PM3 para obtener la estructura molecular más estable de cada una de las estructuras en cada tiempo [28].

Para verificar que tan factible y por lo tanto estables serían las conformaciones tomadas en la dinámica molecular realizada previamente [51] se realizó la comparación entre estructuras no optimizadas y estructuras optimizadas; Las estructuras provenientes del análisis de acoplamiento molecular, en teoría, no deberían variar a pesar de los cambios en el tiempo, o variar muy poco (Figura 13).

En la Figura 13 podemos observar que el grupo metilo en la posición 4 de la quinolina presenta la misma conformación tanto en el cálculo de optimización (Figuras 13B, D, F y H) en los 4 tiempos mostrados; por otra parte, si observamos el oxígeno de la posición 2 de la quinolina presenta el ángulo formado por el oxígeno y los dos carbones a los que se une; sin embargo podemos observar que la forma compacta de la molécula se respeta en los primeros 2 tiempos (0.5 y 1.5 ns) y existe un cambio conformacional completamente en 2.5 ns haciendo ver a la molécula con una apariencia lineal para posteriormente en 3.5 ns regresar a la formación compacta; esto lo podemos corroborar con el valor del ángulo diedro el cual tiene un valor muy cercano a 180° en los 0.5, 1.5 y 3.5 ns y un valor muy cercano a cero en 2.5 ns.

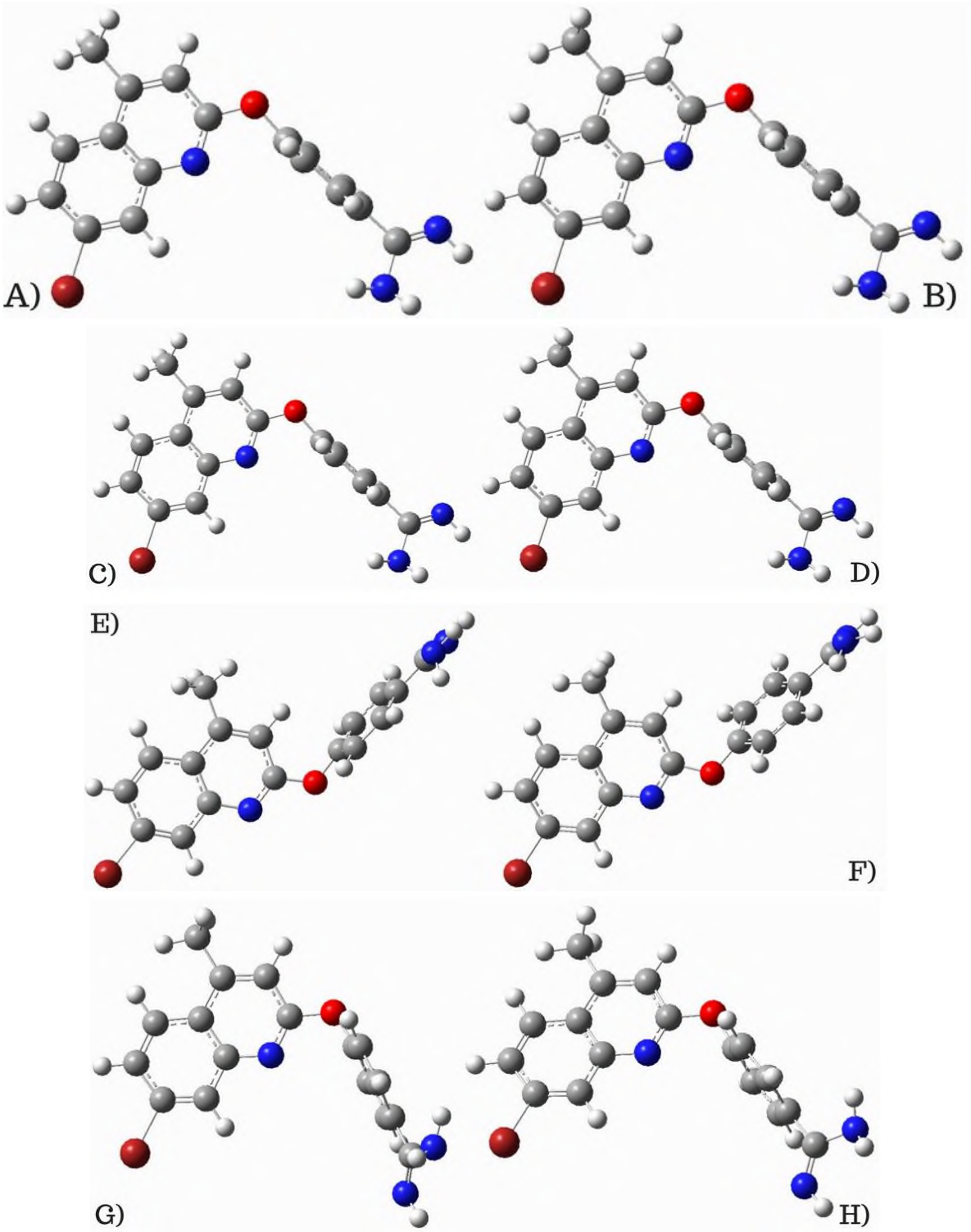


Figura 13: Comparación de la molécula 121 (2-[-4-Oxibenzamidina]-7-Bromo-4-Metilquinolina) en 4 tiempos diferentes. Cálculo único en A) 0.5, C) 1.5, E) en 2.5 y G) en 3.5 ns. Cálculo de optimización en B) 0.5, D) 1.5, F) 2.5 y H) 3.5 ns.

Si observamos el grupo amidino podemos visualizar independientemente el cambio que se observa en 2.5 ns; el giro de este grupo en 3.5 ns comparado con los 0.5 y 1.5 ns observando como el doble enlace del C=N se encuentra en una posición inversa y en 3.5 ns está en el inferior, la medida del ángulo diedro es igual solo cambia el sentido de orientación; es decir el ángulo diedro en 0.5 y 1.5 ns es muy cercano a 180º y en 2.5 y 3.5 ns es muy cercano a -180º. Esto nos indica que el centro activo está cambiando en el tiempo. El mismo análisis se realizó para todas las moléculas en los diferentes tiempos (Ver anexo 1).

Utilizando los descriptores moleculares se creó el modelo utilizando el programa SPSS. Se observó que los descriptores que se repitieron a lo largo del tiempo fueron; Mor16m; seguido de Mor14e, RDF130m, R1m+, dipolo total cuántico, G1u, MATS8m, R1m, IDDE y GATS8e (Ver Tabla 2).

Descriptor	Definición	Ponderado por	Grupo descriptor
Mor16m	3D-MoRSE señal 16	Masas atómicas	3D-MoRSE
Mor14e	3D-MoRSE señal 14	Electronegtividades atómicas de Sanderson	3D-MoRSE
RDF130m	Función de distribución Radial señal 13.0	Volumen atómico de Van Der Waals	RDF
R1m+	Autocorrelación máxima R de bandera 1	Masas atómicas	GETAWAY
D_{Total}	Dipolo Total	Sin ponderación	Cuántico
G1u	Primer componente índice de simetría direccional	Sin ponderación	WHIM
MATS8m	Autocorrelación de Moran bandera 8	Masas atómicas	Autocorrelación 2D
R1M	Autocorrelación R de bandera 1	Masas atómicas	GETAWAY
IDDE	Información sobre las distancias	Sin ponderación	Índices de información
GATS8e	Autocorrelación de Geary bandera 8	Electronegatividad atómica de Sanderson	Autocorrelación 2D

Tabla 2: Resumen de los descriptores predominantes en los modelos.

En el caso del descriptor Mor16m (ver Ecuación 7) se trata de un descriptor ponderado en masa atómica de señal 16 perteneciente a los descriptores 3-DMorSE,

este descriptor es derivado desde el espectro infrarrojo utilizando una función de dispersión generalizada. Este tipo de descriptores pueden calcular las propiedades atómicas en una suma por diferentes funciones de dispersión otorgando un buen arreglo de los átomos de la molécula es un diseño tridimensional.

El descriptor Mor14e también pertenece a los descriptores 3-DMorSE pero ponderado por la electronegatividad de Sanderson en la señal 14.

$$Mor(s,v) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \frac{w_i w_j \sin(sr_{ij})}{sr_{ij}}$$

Ecuación 7: Estructura matemática de los descriptores Mor.

El descriptor RDF130m (Ecuación 8) es un descriptor de funciones radiales (RDF Radial Function Descriptor; Descriptor de Función Radial) ponderado por la masa atómica, este descriptor molecular es obtenido por una función de base radial centrado sobre diferentes distancias interatómicas que van desde el centro hasta los 0.5 a 15.5 Å, este tipo de descriptores parece relacionar la presencia de átomos electronegativos en la atmosfera interna en un volumen esférico de radio R. Entre compuestos similares, además proporciona información sobre las distancias interatómicas en toda la molécula e información valiosa sobre distancias de enlace.

Formalmente un ensamble de átomos A puede ser interpretado como la probabilidad de distribución radial de encontrar un átomo en un volumen esférico de radio R. La forma general del código de Función de distribución radial está representada por la siguiente ecuación:

$$g(R) = f \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} w_i \cdot w_j \cdot e^{-\beta \cdot \left(R - r_{ij}^2\right)}$$

Ecuación 8: Estructura matemática de los descriptores RDF.

donde *f* es el factor de escala, *w* son las propiedades características de los átomos *i* y *j*, r_{ij} son las distancias interatómicas entre el *i*-ésimo y el *j*-ésimo átomo y A es el número de átomos; el término exponencial contiene la distancia interatómica r_{ij} y el término suavizado β la cual define la probabilidad de distribución de las distancias interatómicas individuales; β puede ser interpretado como un factor de

temperatura que define el movimiento de los átomos. Siendo $g(R)$ calculado generalmente a un número de puntos discretos con intervalos definidos [64].

Para incluir las propiedades atómicas definidas w de los átomos i y j puede ser usando el código RDF en diferentes tareas de forma que los requerimientos de la información sean representados. Estas propiedades atómicas permiten la discriminación de los átomos de una molécula.

El descriptor R1m es un descriptor de autocorrelación de atraso 1 ponderado por la masa atómica, mientras que el descriptor R1m+ es de autocorrelación de atraso 1 máxima ponderado también por la masa atómica. Ambos descriptores corresponden al conjunto de descriptores GATEWAY que son calculados a partir de la matriz de apalancamiento obtenida por las coordenadas atómicas centradas. Los descriptores R1m y R1m+ son análogos obtenidos desde la matriz palanca/geometría.

El dipolo total es un descriptor cuántico, que se obtiene por la distribución de distribución de cargas en la molécula.

El descriptor G1u pertenece al grupo de descriptores WHIM, representa la direccionalidad del primer componente WHIM, en este caso no se encuentra ponderado. Este descriptor se obtiene a partir de índices estadísticos de los átomos proyectados sobre los tres componentes principales obtenidos desde matrices de covarianza ponderadas en coordinadas atómicos.

El descriptor MATS8m corresponde a la serie de descriptores de autocorrelación de Moran, en este caso de atraso 8 y ponderado por la masa atómica, perteneciente a los descriptores autocorrelación 2D al igual que el descriptor GATS8e pero corresponde a la serie de descriptores de autocorrelación de Geary de atraso 8 ponderado la electronegatividad de Sanderson.

El descriptor IDDE pertenece al grupo de los índices de información que presenta sobre el contenido de la igualdad del grado de distancia, los descriptores de autocorrelación 2D son descriptores calculados desde gráficos moleculares de la suma de los pesos atómicos de los átomos terminales; todos estos descriptores fueron generados por un conjunto de entrenamiento de 26 moléculas [65].

Utilizando el programa SPSS se seleccionaron los descriptores y sus valores que mejor describe a cada serie de moléculas (Ecuación 9). Para verificar la fiabilidad de este modelo se compararon los resultados con la actividad biológica experimental mediante el cálculo del residuo, un valor muy cercano a cero indicaba que los valores del modelo son aceptables, un ejemplo puede visualizarse en la Tabla 3.

$$MODELO = Cte\ B + Descriptor_1 * CteB_1 + \cdots + Descriptor_n * Cte_n$$

Ecuación 9: Estructura matemática de los modelos QSAR.

Nombre	log IC50 $_{exp}$	log IC5O $_{Calc}$	logIC50$_{ANN}$	Residuo$_{MLR}$	Residuo$_{ANN}$
m012d4o	-2.2218	-1.7507	-1.7384	0.4711	0.4834
m015d4o	-1.7447	-0.7250	-0.8127	1.0197	0.9320
m038d4o	-1.7959	-1.3192	-1.4687	0.4766	0.3272
m062d4o	-2.2218	-1.0646	-1.9040	1.1572	0.3178
m076d4o	-2.0000	-2.0549	-1.6752	-0.0549	0.3248
m088d4o	0.5798	0.0261	0.2034	-0.5536	-0.3764
m105d4o	0.1139	-1.5184	-1.5136	-1.6324	-1.6275
m121d4o	-1.1804	-1.3387	-1.5622	-0.1582	-0.3817
m186d4o	-2.0000	-1.4271	-1.8262	0.5729	0.1738
m188d4o	-1.5686	-1.8513	-1.6479	-0.2826	-0.0793
m224d4o	-1.9208	-1.9222	-1.9192	-0.0014	0.0016
m234d4o	-1.1675	-2.0986	-1.6157	-0.9311	-0.4482
m244d4o	-1.8860	-2.0947	-1.6150	-0.2086	0.2710
m289d4o	-1.8239	-1.8168	-1.5693	0.0071	0.2546
m298d4o	1.2787	0.9750	0.7793	-0.3037	-0.4994
m316d4o	0.6020	0.2530	0.6966	-0.3490	0.0945
m325d4o	0.1139	-0.1940	0.1723	-0.3079	0.0583
m336d4o	0.2041	0.5664	0.7429	0.3623	0.5388
m367d4o	-0.2218	-0.4738	-0.2743	-0.2519	-0.0524
m505d4o	-1.2676	-1.0850	-1.2788	0.1825	-0.0112
m517d4o	-0.4815	-0.8926	-0.8626	-0.4111	-0.3811
m533d4o	-2.0000	-1.5690	-1.8477	0.4310	0.1523
m556d4o	-2.2676	-2.4931	-2.3248	-0.2255	-0.0572
m570d4o	-2.0000	-1.4796	-2.0050	0.5204	-0.0050
m592d4o	-2.0315	-1.5602	-1.9685	0.4712	0.0630

Tabla 3: Comparación de la actividad biológica experimental contra la calculada por MLR y ANN y sus residuos respectivos para el grupo de entrenamiento en 4.0 y 1.5 ns respectivamente; este grupo no presento valores atípicos.

La separación de las 50 moléculas en los tres diferentes subconjuntos se realizaron de forma azarosa por medio de una subrutina programada en MATLAB, respetando el número de moléculas propuesto por Ramírez-Galicia y colaboradores [51], tomando 27 moléculas para el conjunto de entrenamiento, 18 para el conjunto de predicción y 5 moléculas para el conjunto de validación.

Se realizaron los modelos de Regresión Lineal Múltiple por cada 0.5 ns, todos los modelos realizados posteriormente fueron refinados con Redes Neuronales Artificiales. La visualización se realizó por medio de gráficos de cada modelo y en cada refinamiento. Para determinar si los modelos habían mejorado se comparó el valor dado de r^2 y la presencia de valores atípicos.

En la Figura 14 se muestran las regresiones lineales (Figura 14A y 14C) y las regresiones no lineales por ANN (Figura 14B y 14D), incluyendo, en rojo, a los 4 valores atípicos encontrados en 0.5 ns. Los valores del coeficiente de correlación, $R^{2,}$ observan una ligera mejora entre el caso A) y B) para el conjunto de predicción, pasando de R^2_{MLR}=0.6695 a R^2_{RNA}= 0.6953 y una mejora considerable para el conjunto de entrenamiento, pasando de R^2_{MLR}=0.8568 a R^2_{RNA}= 0.9654; para el grupo de validación se presentaron valores R^2_{MLR}=0.873 y R^2_{RNA}=0.7622 (Tabla 4).

Para el caso de las Figuras 14C y 14D, se descartó el valor atípico del segundo cuadrante, sin embargo, se observó que el valor de la R^2 para el grupo de validación disminuyó con respecto a los modelos que conservaron sus valores atípicos siendo de R^2_{MLR}=0.5627 y R^2_{RNA}=0.3899 (Tabla 4); para el caso de los grupos entrenamiento y predicción se mantuvieron iguales debido a que los valores atípicos que fueron descartados pertenecían al grupo validación.

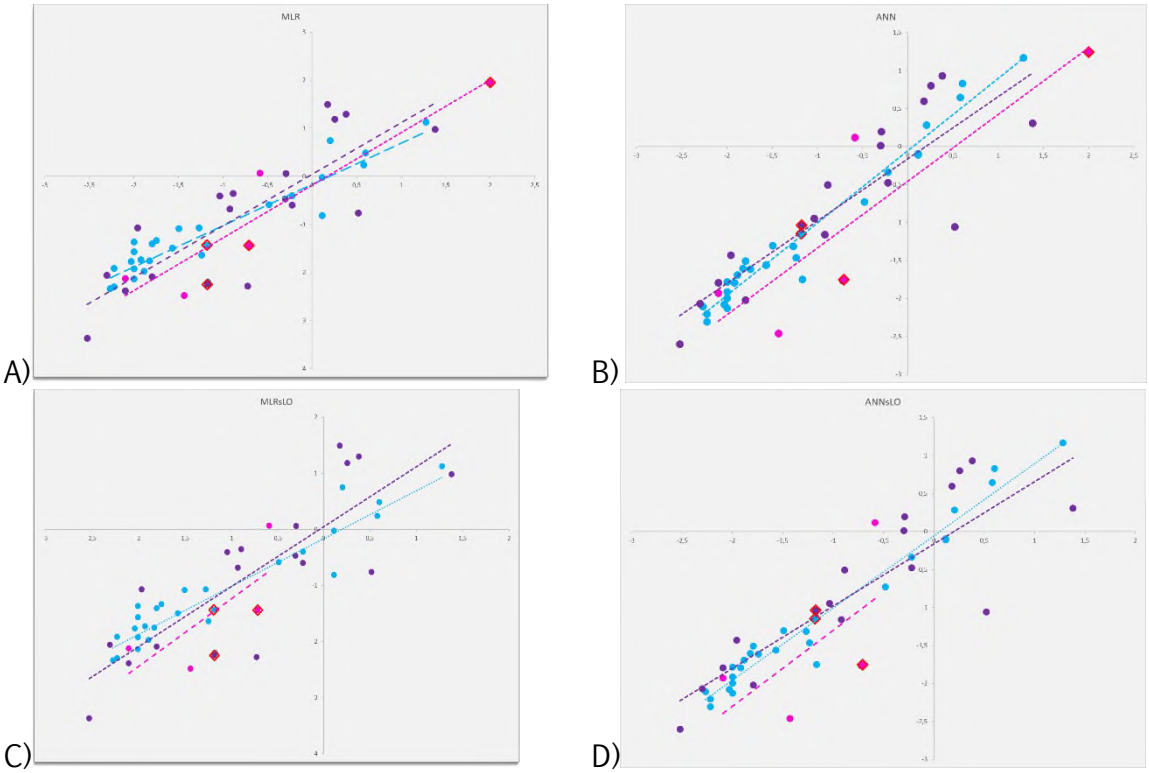
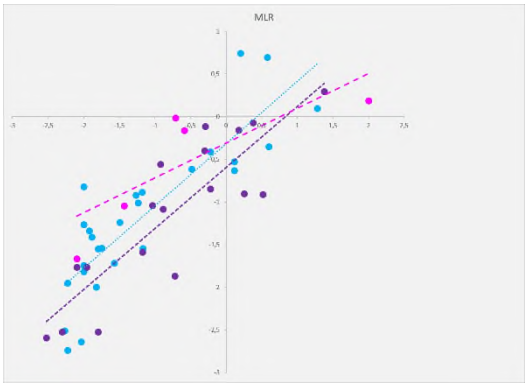


Figura 14: Comparación de los gráficos de los modelos RLM y RNA en 0.5 ns; A) modelo RLM incluyendo valores atípicos. B) modelo RNA incluyendo valores atípicos. C) modelo RLM sin valores atípicos y D) modelo RNA sin valores atípicos. Los puntos azules corresponden a entrenamiento, morados a predicción y en rosa validación; así como sus líneas correspondientes. Los puntos enmarcados en rojo corresponden a los valores atípicos.

En la Figura 15 se observa la comparación para los modelos RML y RNA en 1.0 ns. El coeficiente de correlación, R^2, incrementa en el conjunto de entrenamiento pasando de 0.7295 en el modelo RML a 0.9071 en el modelo RNA; Sin embargo, para el conjunto de predicción, la R^2 disminuye de 0.7516 en el modelo RML a 0.6264 en el modelo RNA y para el caso de validación tenemos un valor de R^2 de 0.6498 para el modelo MLR y R^2 de 0.8324 para RML mejorando de manera considerable el uso de Redes Neuronales Artificiales para el caso de validación; el modelo 1.0 ns no presentó valores atípicos en sus gráficos (Tabla 4).

A) 

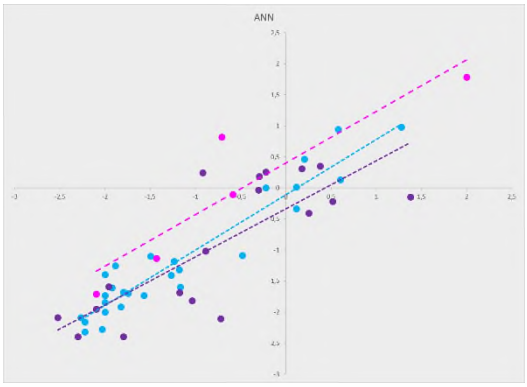
B) 

Figura 15: Comparación de los gráficos de los modelos A) RLM y B) RNA en 1.0 ns. Los puntos azules corresponden a entrenamiento, morados a predicción y en rosa validación; así como sus líneas correspondientes.

En la Figura 16 se observa la comparación del modelo correspondiente a 1.5 ns, en el cual podemos observar como el coeficiente de correlación mejoró con la aplicación de RNA comparado con RLM para el caso de entrenamiento siendo de 0.9354 para RLM y 0.9922 para RNA, para el caso de predicción el valor del coeficiente de correlación disminuyó siendo 0.8274 para RLM y 0.6550 para RNA; este grupo presentó un valor atípico; el grupo de validación también presentó una disminución en el valor del coeficiente de correlación al aplicar RNA siendo de 0.9888 para RLM y 0.9638 para RNA; este grupo también presentó un valor atípico (Tabla 4).

En la Figura 17 se observan la comparación de los modelos QSAR en 3.0 ns incluyendo los valores atípicos (A y B) y excluyendo los valores atípicos (C y D). En la Figura 17A se observan dos valores atípicos, los cuales se muestran suavizados, en el caso de la Figura 17B con el uso de Redes Neuronales Artificiales. El coeficiente de correlación incrementa en el conjunto de entrenamiento de 0.8170 en el modelo RLM a 0.9391 en el modelo RNA. En este caso también se observa el incremento de coeficiente de correlación en el conjunto de predicción, pasando de 0.5854 en el modelo RLM a 0.6280 en el modelo RNA. Y para el caso de validación 0.9276 en RLM y 0.8308 para RNA (Tabla 4).

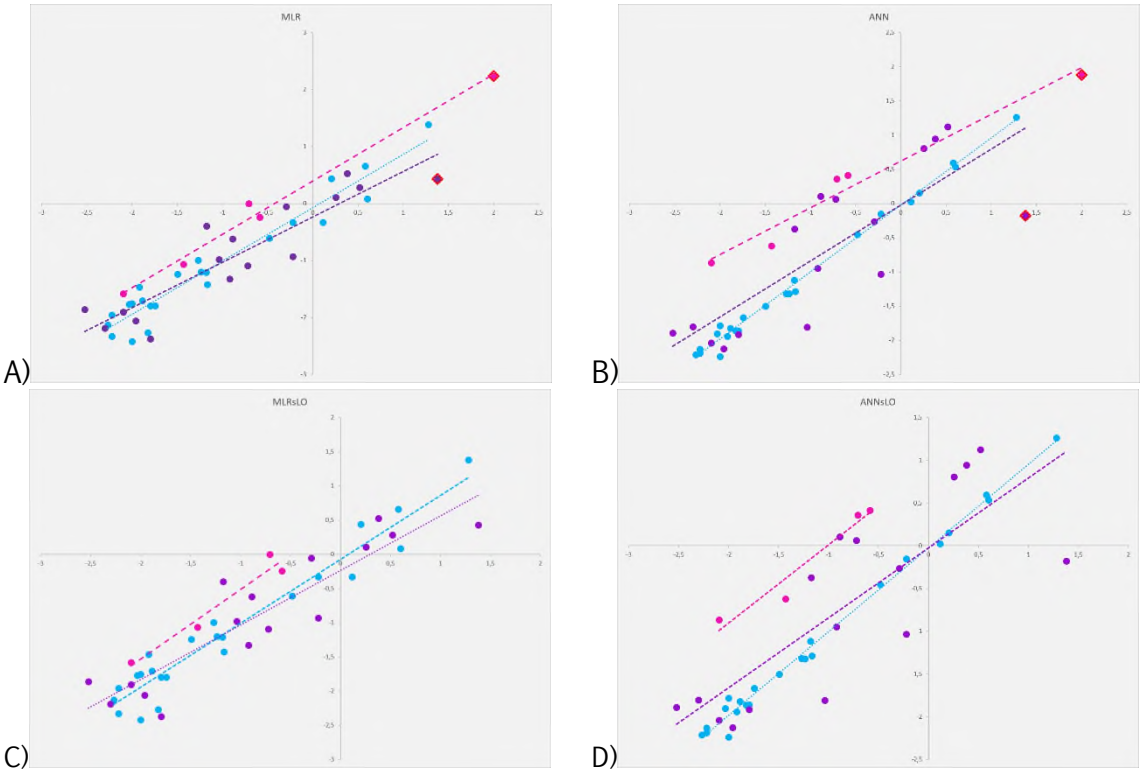


Figura 16: Comparación de los gráficos de los modelos RLM y RNA en 1.5 ns; A) modelo RLM incluyendo valores atípicos. B) modelo RNA incluyendo valores atípicos. C) modelo RLM sin valor atípico y D) modelo RNA sin valor atípico. Los puntos azules corresponden a entrenamiento, morados a predicción y en rosa validación; así como sus líneas correspondientes. Los puntos enmarcados en rojo corresponden a los valores atípicos.

En la Figura 17 vemos los gráficos correspondientes a RLM y RNA sin el valor atípico perteneciente al grupo de validación; para ambos casos RLM y RNA disminuyo el coeficiente de correlación siendo 0.6048 para el caso de MLR y 0.3567 para RNA en el caso del grupo de validación; los grupos de entrenamiento y predicción no presentaron cambio en su coeficiente de correlación (Tabla 4).

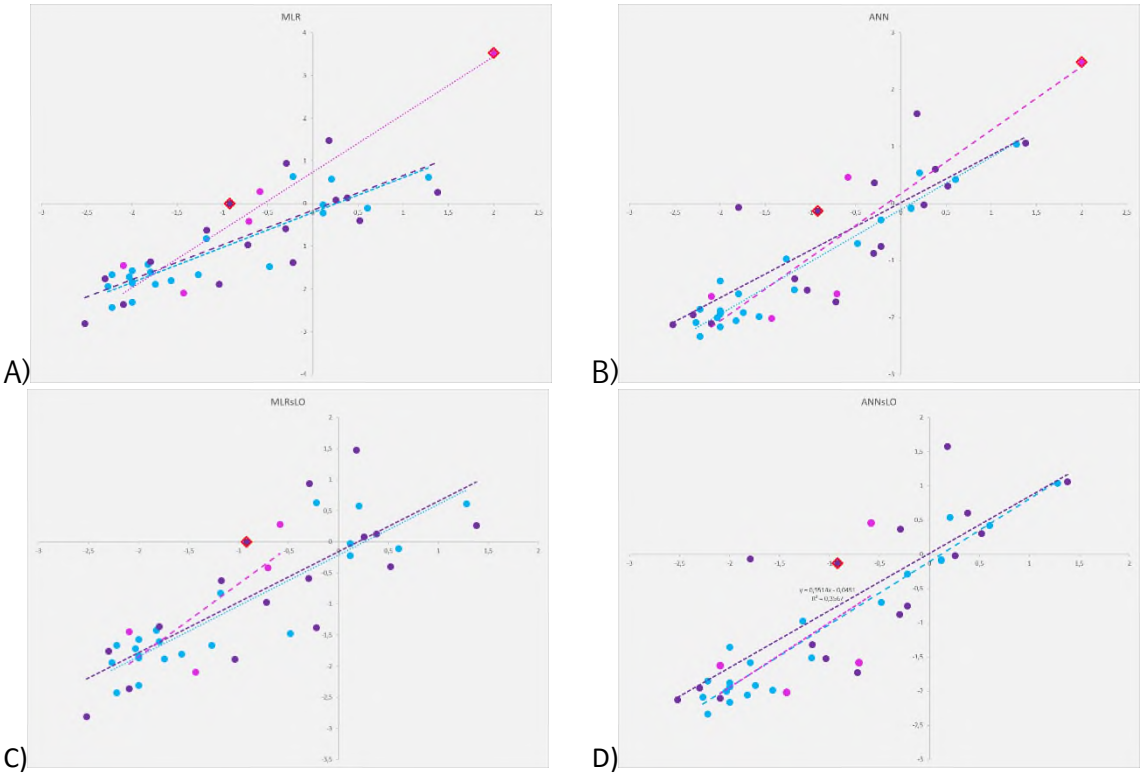


Figura 17: Comparación de los gráficos de los modelos en 3.0 ns A) RLM con valores atípicos, B) RNA con valores atípicos, C) RLM sin valores atípicos y D) RNA sin valores atípicos. Los puntos azules corresponden a entrenamiento, morados a predicción y en rosa validación; así como sus líneas correspondientes. Los puntos enmarcados en rojo corresponden a los valores atípicos.

	MLR	ANN	MLRsLO	ANNsLO
0.5 ns				
Entrenamiento	y=0.8568x-0.1678 R^2=0.8568	y=0.9512x-0.0554 R^2=0.9654	y=0.8568x-0.1678 R^2=0.8568	y=0.9512x-0.0554 R^2=0.9654
Predicción	y=1.0719x+0.0407 R^2=0.6695	y=0.8192x-0.1621 R^2=0.6953	y=1.0719x+0.040 R^2=0.6695	y=0.8192x-0.1621 R^2=0.6953
Validación	y=1.0948x-0.1879 R^2=0.8733	y=0.8792x-0.4591 R^2=.07622	y=1.204x-0.0437 R^2=0.5627	y=0.9998x-0.2997 R^2=0.389
1.0 ns				
Entrenamiento	y=0.7295x-0.317 R^2=0.7295	y=0.8912x-0.1107 R^2=0.9071	-	-
Predicción	y=0.7151x-0.5969 R^2=0.7516	y=0.7731x-0.3356 R^2=0.6264	-	-
Validación	y=0.4071x-0.3104 R^2=0.6498	y=0.8311x+0.4007 R^2=0.8324	-	-
1.5 ns				
Entrenamiento	y=0.9354x-0.0736 R^2=0.9354	y=0.9801x-0.0238 R^2=0.9922	y=0.9354x-0.0736 R^2=0.9354	y=0.9801x-0.0238 R^2=0.9922
Predicción	y=0.7974x-0.2359 R^2=0.8274	y=0.8179x-0.0244 R^2=0.655	y=0.7974x-0.2359 R^2=0.8274	y=0.8179x-0.0244 R^2=0.655
Validación	y=0.9376x+0.3971 R^2=0.9888	y=0.6839x+0.62 R^2=0.9638	y=1.0168x+0.5017 R^2=0.947	y=0.9117x+0.9209 R^2=0.9389
3.0 ns				
Entrenamiento	y=0.817x-0.2136 R^2=0.817	y=0.919x-1071 R^2=0.9391	y=0.817x-0.2136 R^2=0.817	y=0.919x-1071 R^2=0.9391
Predicción	y=0.8122x-0.1599 R^2=0.5854	y=0.8349x+0.0149 R^2=0.628	y=0.8122x-0.1599 R^2=0.5854	y=0.8349x+0.0149 R^2=0.628
Validación	y=1.3536x+0.7343 R^2=0.9276	y=1.1164x+0.1728 R^2=0.8308	y=1.1733x+0.4961 R^2=0.6048	y=0.9514x-0.0451 R^2=0.3567

Tabla 4: Ecuaciones y valores de R^2 para cada conjunto de los modelos 0.5 ns, 1.0 ns, 1.5 ns y 3.0 ns; correspondientes a las Figuras 14, 15, 16 y 17.

Los modelos que presentaron mejor coeficiente de correlación presentan al menos un descriptor GETEWAY (Tabla 5); este tipo de descriptores se caracterizan por intentar armonizar la geometría molecular 3D proporcionada por la matriz de influencia molecular y la relación de la topología atómica con diferente información química como peso atómico, polarizabilidad y electronegatividad, volumen de van der Waals.

Modelo	0.5 ns	1.0 ns	1.5 ns	3.0 ns
Descriptor	SIC0	GATS8e	MATS8m	R1m
	R8v+	E1s	Mor16m	HATS7u
	Mor11p	R3u+	R1e+	Pw5
	H8u	MATS4m	X	
	HATS6p		X1A	
			G1u	

Tabla 5: Descriptores encontrados en los modelos que presentaron mejores coeficientes de correlación.

Todos los modelos se analizaron por medio de los métodos de estandarización y apalancamiento con el fin de establecer la aplicabilidad de dominio; la molécula que resultó con la mayor cantidad de valores atípicos (Ver Tabla 6) fue la 6-[N-3,5 -Bis- (Trifluorometil) fenil sulfonamida -2- (4- Metil- Benzamidina)]- 1H-3-Metil-Benzimidazol (molécula 356), la cual es la única molécula que contiene átomos de Flúor (seis) en su estructura (Figura 18) y sus valores de actividad biológica distaban de los demás pertenecientes al mismo grupo (Ver Tabla 6); esta molécula fue removida de algunos modelos para intentar mejorar la tendencia de su regresión y el valor de los coeficientes de correlación, R^2.

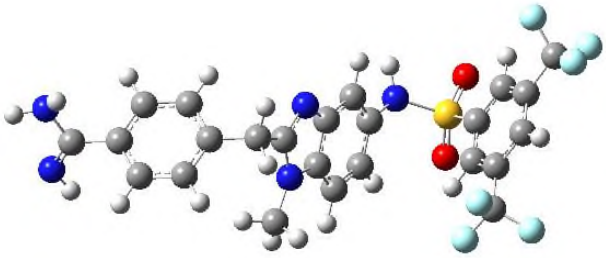


Figura 18: Estructura de la 6-[N-3,5-Bis-(Trifluorometil)fenil sulfonamida-2-(4-Metil-Benzamidina)]-1H-3-Metil-Benzimidazol (Molécula 356).

Los modelos con mejor predicción son los que presentan valores de residuos pequeños. En este caso de MLR con valores de residuos pequeños fue el modelo en 4.0 ns, mientras que para RNA correspondiente en 1.5 ns, ambos para el grupo de entrenamiento (Tabla 3).

Para el modelo en 0.5 ns (Ver Figura 14), se observa la presencia de 4 valores atípicos, entre ellos la molécula 356. Aunque el modelo de MLR también fue refinado por la metodología de Redes Neuronales Artificiales y se observa que la R^2 mejora,

la confiabilidad del resultado no se puede aceptar debido a que presentan comportamientos diferentes a la mayoría de las moléculas; los valores atípicos se pueden observar más cerca de los demás valores sobre el eje Y; por ejemplo, el valor atípico del segundo cuadrante (Figura 14A) presenta un valor de 2.0 para ambos ejes, mientras que en la Figura 14B se observa que aunque conserva el valor de 2.0 en el eje X, el valor del eje Y se encuentra entre 1.5 y 2.0 unidades. Después de realizar los modelos sin la presencia de los valores atípicos, se observó que la molécula 356 influía en la tendencia del modelo. Claramente esta molécula presenta un mecanismo diferente que la mayoría de los compuestos en el conjunto de entrenamiento como se puede observar en la Figura 16C y 16D y la Figura 17C y 17D.

En el caso del modelo en 3.0 ns (Figura 17) se observaron dos valores atípicos en la gráfica de Regresión Multilíneal los cuales fueron suavizados posteriormente en la aplicación de Redes Neuronales Artificiales, siendo más evidente este refinamiento que el hecho al valor atípico de la molécula 356 la cual se localiza en el segundo cuadrante el cual pasa de estar posicionado sobre 3 y 4 para el eje Y a 2 y 3, permaneciendo igual para el eje X (Figura 17A y la Figura 17B) observando una mejora significativa tanto para el grupo de entrenamiento como para el grupo de predicción. Una vez realizada la aplicabilidad del dominio se descartó el valor atípico del segundo cuadrante (ver Figuras 17C y 17D) correspondiente a la molécula 356 observando que el valor de la R^2 para el grupo de validación disminuye; sin embargo, se observa una mejora en la tendencia de la pendiente calculada con Redes Neuronales.

1.5 ns				1.0 ns			
Nombre	**logIC50$_{exp}$**	**logIC50$_{calc}$**	**logIC50$_{ANN}$**	**Nombre**	**logIC50$_{exp}$**	**logIC50$_{calc}$**	**logIC50$_{ANN}$**
m023d15o	-1.4318	-1.0683	-0.6215	m023d1o	-1.4318	-1.0460	-1.1333
m132d15o	-0.7077	0.0005	0.3609	m132d1o	-0.7077	-0.0141	0.8214
m212d15o	-2.0969	-1.5844	-0.8651	m212d1o	-2.0969	-1.6663	-1.7090
m339d15o	-0.5850	-0.2438	0.4134	m339d1o	-0.5850	-0.1622	-0.1042
m356d15o	2.0000	2.2357	1.8827	m356d1o	2.0000	0.1878	1.7836

Tabla 6: Comparación de la actividad biológica y la calculada en MLR y ANN para el grupo de validación; para 1.5 ns se observan un valor atípico (molécula 356 en rojo); en el caso del modelo 1.0 ns, no se observaron valores atípicos.

Para verificar y corroborar todos los valores obtenidos se compararon todos los modelos obtenidos en los diferentes tiempos (Figura 19). En esta figura se observa que las tendencias de las pendientes son similares, incluso para los modelos en 2.0 y 3.0 ns se empalman, observando algo similar para los modelos en 4.5 y 5.0 ns. Todos los gráficos se cruzan en aproximadamente en -1.25 unidades para ambos ejes (X, Y), si ubicamos estos valores de Log IC50 en los datos de la actividad biológica experimental y calculada podrían corresponder a las moléculas mostradas en la Tabla 7.

Molécula	Log IC50	1.5ns	2.0ns	2.5ns	5.0ns
m505d0o	-1,2676	-0,9989	-0,424	-1,531	-1,2786
m466d0o	-1,2366	-1,2023	-1,0741	-1,2116	-0,8995
m121d0o	-1,1804	-1,2089	-1,2531	-0,5944	0,2142
m234d0o	-1,1675	-1,425	-1,8269	-1,671	-1,7118

Tabla 7: Valores de la actividad biológica en MLR; los valores marcados corresponden al valor aproximado a -1.25 que es el punto aproximado de cruza de los gráficos.

El modelo MLR en 1.5 ns presenta su ordenada al origen cercana a cero (b=-0.0735), mientras que el valor de su pendiente es muy cercano a uno (m=0.9354). (Ver Figura 21 y 22). Podemos agrupar dos tendencias, los modelos en 2.5, 4.5 y 5.0 ns presentan un comportamiento similar, mientras que los modelos 1.0, 2.0, 3.5 y 4.0 ns presentan otro comportamiento.

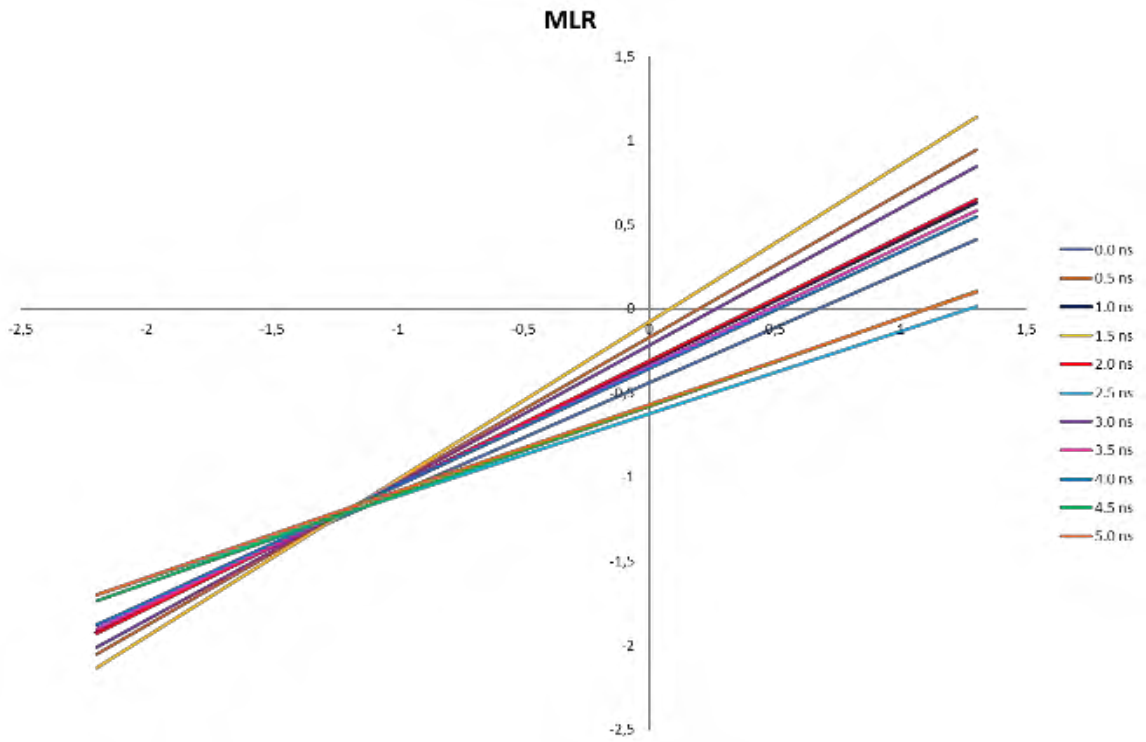


Figura 19: Gráfico correspondiente a la pendiente de los diferentes modelos de los diferentes tiempos bajo estudio; para el grupo entrenamiento.

Para el caso de la Figura 20 se utilizaron los datos de los modelos refinados con RNA. En este caso se puede observar que el comportamiento de las pendientes es más homogéneo; es decir que presentan un patrón de comportamiento muy similar. Al comprar estos resultados con las pendientes obtenidas por RLM se observa que el modelo en 2.5 ns se encuentra fuera de esta tendencia. Por otra parte, los modelos en 0.5 y 1.5 ns se empalman entre ellos, además se nota un comportamiento similar entre los modelos en 1.0 y 3.0 ns. Los modelos obtenidos por RNA mejoran debido a que aumenta el número de modelos con la ordenada al origen cercano a cero. También se observan 4 gráficos mejorados en RNA comparado con uno del gráfico de MLR.

Los modelos mejorados en RNA son en 0.0, 1.0, 2.0 y 4.0 ns; todos los modelos se interceptan entre sí en aproximadamente Log IC_{50} entre -1.2 y -1.3. Este valor corresponde a las moléculas mostradas en la Tabla 8. Se observó que cuatro modelos aumentan el valor de sus pendientes, siendo más cercanos a 1 (Ver Figura

22), mientras que para las MLR tan solo fue un modelo. Los modelos que presentaron estos valores fueron: el modelo 3.0 ns, seguido de los modelos en 1.0, 0.5 y 1.5 ns, los últimos dos presentan valores muy similares.

Molécula	LogIC50	0,5	1	1,5	2	2,5	4	5
m505d0o	-1,2676	-1,3131	-1,4058	-1,3197	-1,1484	-1,594	-1,2788	-1,1048
m466d0o	-1,2366	-1,4617	-1,1841	-1,3216	-0,9114	-1,3256		-1,2337
m121d0o	-1,1804	-1,1496	-1,3178	-1,1185	-1,555	-1,302	-1,5622	0,1887
m234d0o	-1,1675	-1,7505	-1,5949	-1,2889	-1,7608	-1,7089	-1,6157	-1,5342

Tabla 8: Valores de la actividad biológica en ANN; los valores marcados corresponden al valor aproximado entre -1.2 y -1.3 que es el punto aproximado de cruza de los gráficos.

Los valores de la intersección con el eje más cercano a cero fue el modelo en 4.0 ns seguido de los modelos en 0.0 y 2.0 ns con el mismo valor, pero con signo diferente (Figura 21). Podemos concluir que los mejores modelos son los calculados en 3.0, 1.0, 0.5 y 1.5 ns. El modelo en 3.0 ns presenta la pendiente de 1.00 y una intersección en el eje con un valor de 0.14.

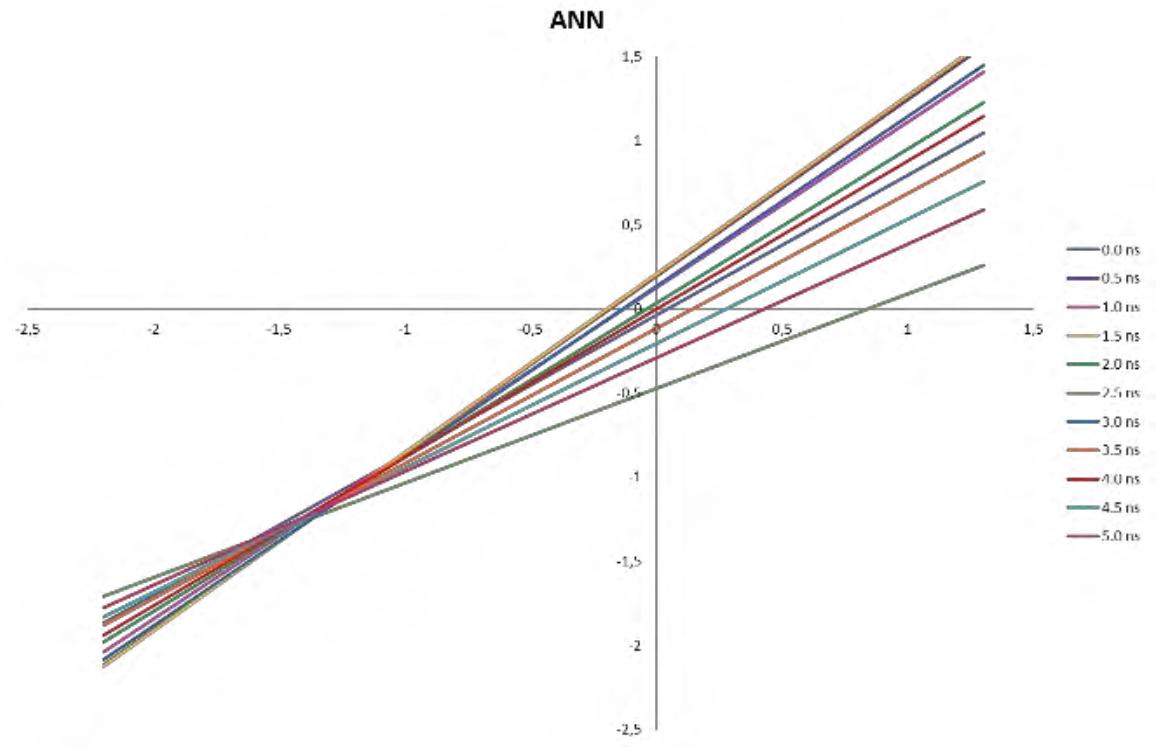


Figura 20: Gráfico correspondiente a la pendiente de los diferentes modelos que han sido refinados por Redes Neuronales Artificiales de los diferentes tiempos estudiados; para el grupo entrenamiento.

Para observar mejor el comportamiento de la pendiente y de la intersección con el eje de los modelos se realizó un gráfico de estos valores; los cuales siguen la misma tendencia entre los análisis de MLR y ANN correspondiente a la pendiente o la intersección; para el caso del gráfico de la pendiente los valores correspondientes a MLR son inferiores a los de ANN al igual que para el caso de la intersección con el eje, verificando de este modo el mejoramiento de los modelos con ANN contra MLR.

Para el caso de la pendiente buscamos que el valor sea lo más cercano a 1 para poder decir que tenemos un buen modelo; los valores que tomo la pendiente para MLR son inferiores a uno en todos los modelos, siendo el más cercano el modelo 1.5 ns con un valor de 0.94, por otro lado, en ANN el gráfico presenta valores más altos y lo vemos reflejado en que tenemos un valor de 1 el correspondiente al modelo 3.0 ns, un valor muy cercano a 1.0000 (0.9800), para el modelo en 1.0 ns y dos valores aceptables de los modelos 0.5 y 1.5 ns con un valor de 1.05 y 1.06 respectivamente, mejorando notablemente los valores de los modelo con el uso de ANN.

En la Figura 21, se espera que el valor ideal de la intersección con el origen sea cero; los valores que tomaron las intersecciones con el eje son mayores para el caso de ANN comparados con los valores que toma para MLR. Para el caso de MLR, los valores son muy inferiores a cero incluso el mejor valor dista de cero con un valor de -0.07, caso contrario que se observa en ANN donde tenemos para el caso del modelo 4.0 ns con un valor de cero y un valor bastante aceptable para los modelos en 0.0 y 2.0 ns con un valor de 0.04 con signo negativo para el modelo en 0.0 ns y con signo positivo para el modelo en 2.0 ns; observando de igual manera que para la pendiente una mejora de los modelos al aplicar las ANN.

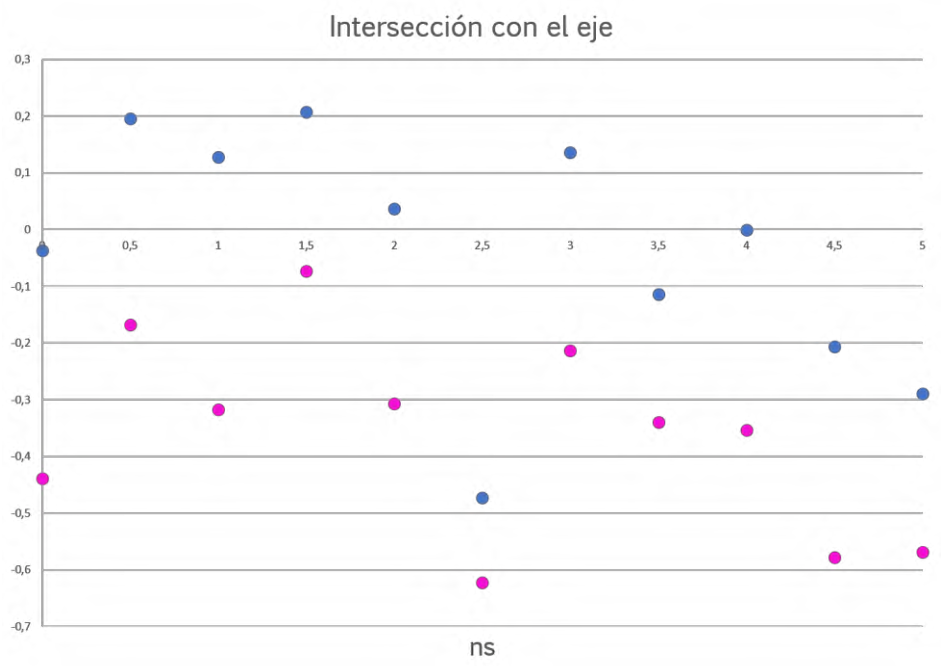


Figura 21: Intersección del eje para los modelos de RLM con el gráfico en color rosa y RNA con el gráfico en color azul.

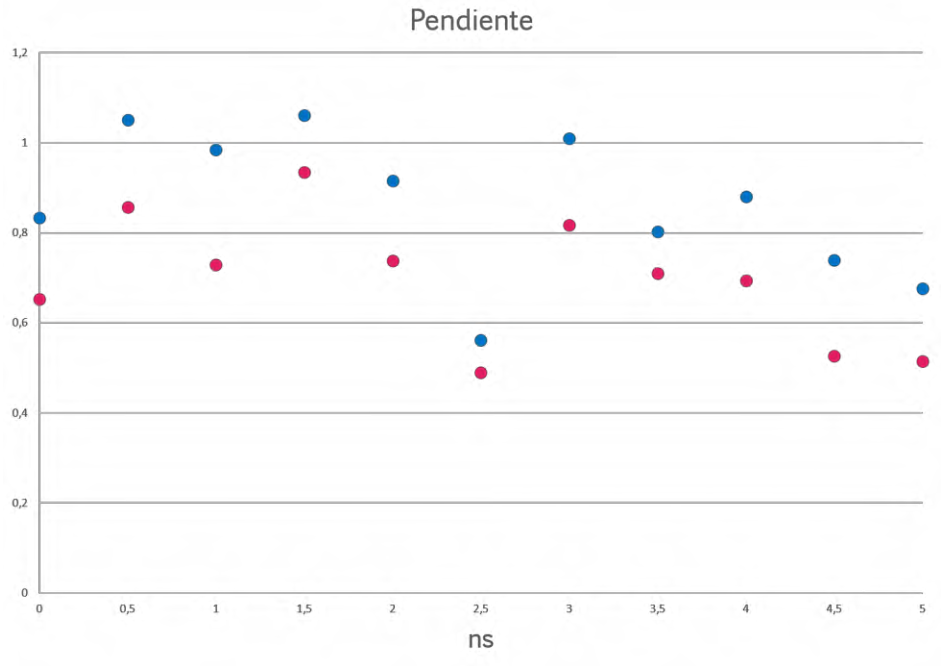


Figura 22: Comportamiento de la pendiente en los modelos MLR con color rosa y ANN mostradas con color azul.

En este gráfico de la Figura 23 podemos observar que existe una relación entre el radio de giro, la pendiente y la intersección con el eje ya que presentan un comportamiento similar. El radio de giro representa la elasticidad que tiene la

proteína en función al tiempo. Al observar las gráficas podemos decir que la proteína presento menor elasticidad en el modelo 2.5 ns y también en este modelo se presentó el valor menor de la pendiente y de la intersección con el eje; siendo el modelo más flexible el 3.5 ns. Podemos decir que si el modelo presenta un radio de giro con valores menores es indicación de que la enzima es más rígida, lo cual la hace reducir las interacciones con los ligandos. Debido a esto mismo es que este modelo presentó un comportamiento errático en la mayoría de nuestros análisis siendo el modelo que más presentó valores atípicos.

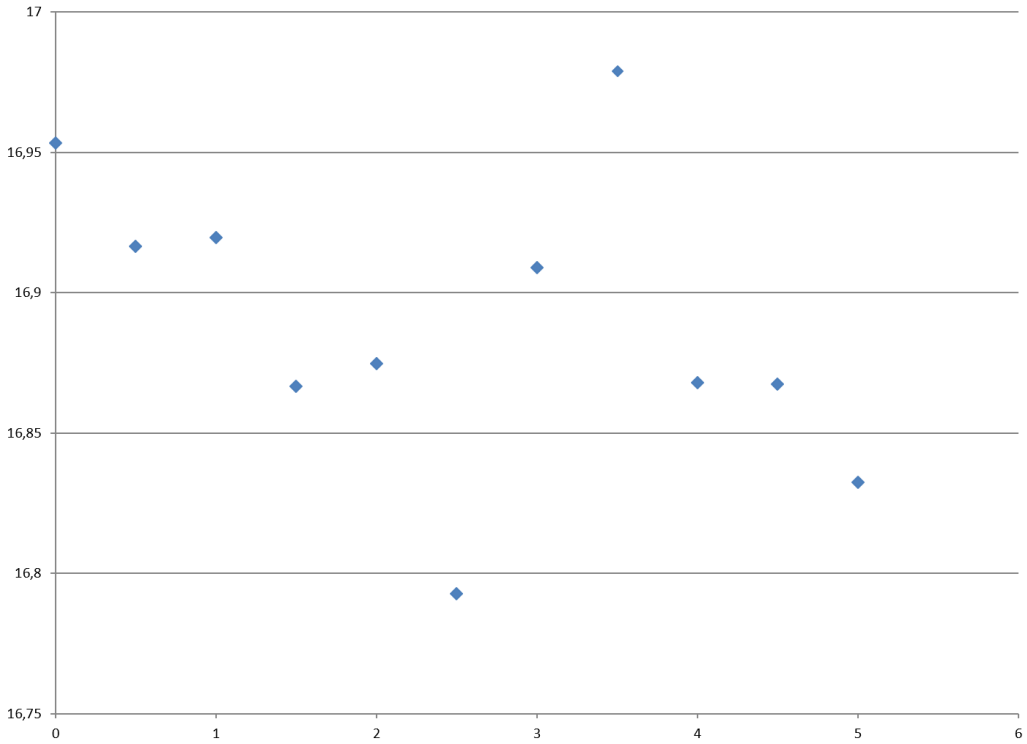
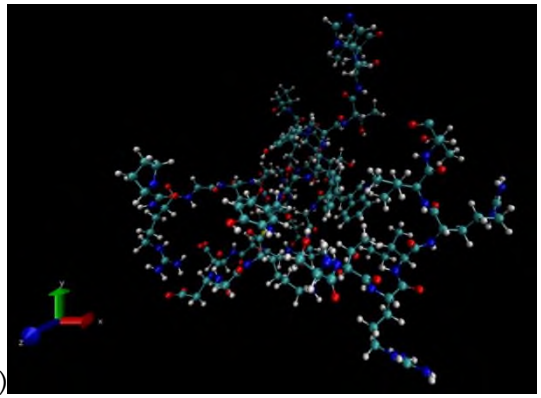
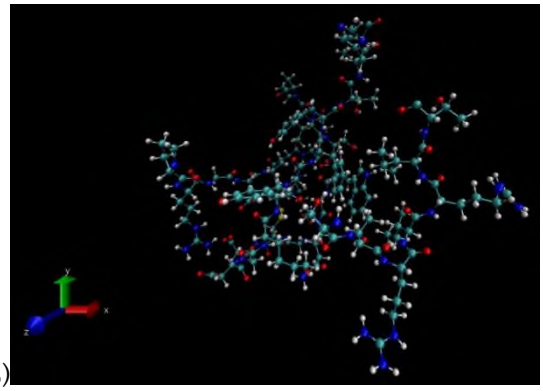


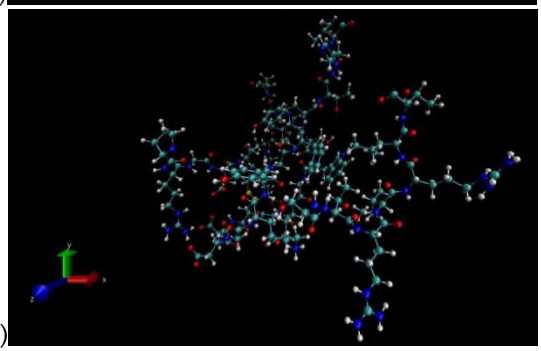
Figura 23: Resumen de los valores de radio de giro obtenidos del análisis realizado por Galicia y colaboradores [51].

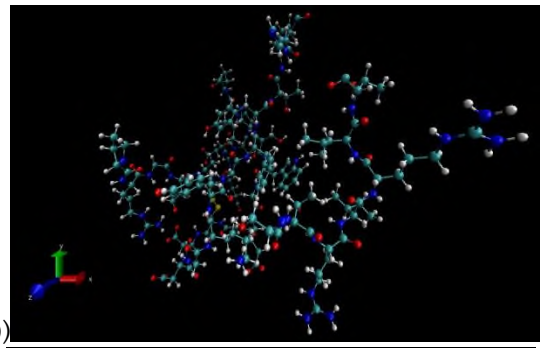
En la Figura 24 se puede apreciar la conformación del sitio de unión de la proteína en los diferentes tiempos observando en cada nanosegundo. El sitio activo toma una conformación espacial diferente. Si observamos el sitio en 0.0 ns y 3.5 ns la distribución del sitio es más amplia observando “huecos”, caso contrario se observa en el sitio en el tiempo 2.5 ns en el cual el sitio se ve más compactado disminuyendo el espacio de los “huecos”; por otro lado, el sitio en el tiempo 1.5

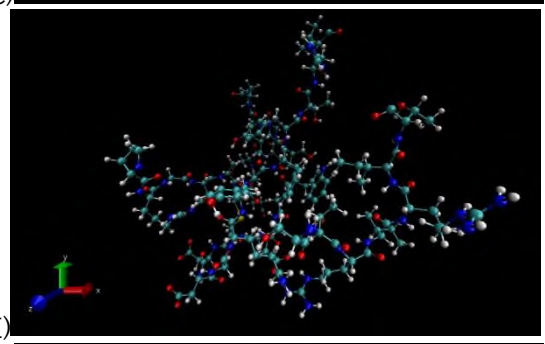
ns parece estar más “desordenado”, la distribución del sitio se observa de manera diferente que en los otros sitios.

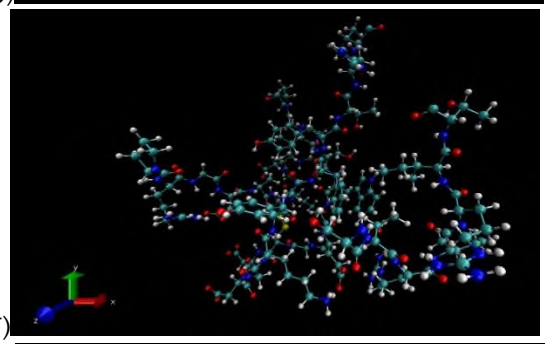
A) 

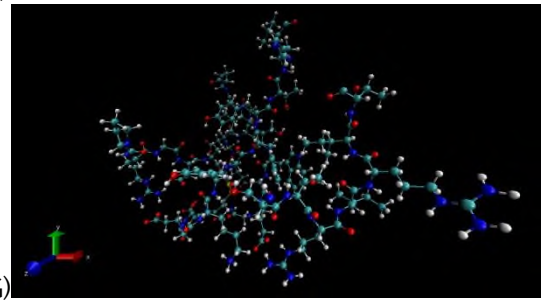
B) 

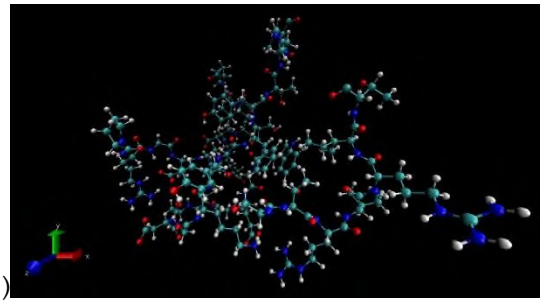
C) 

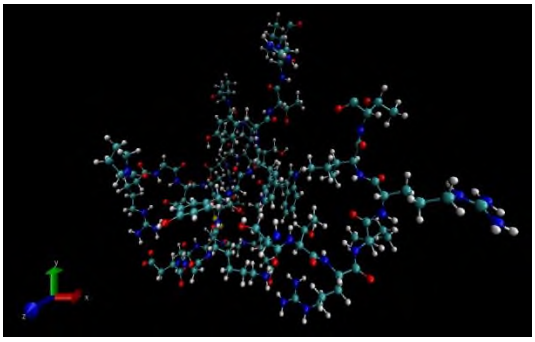
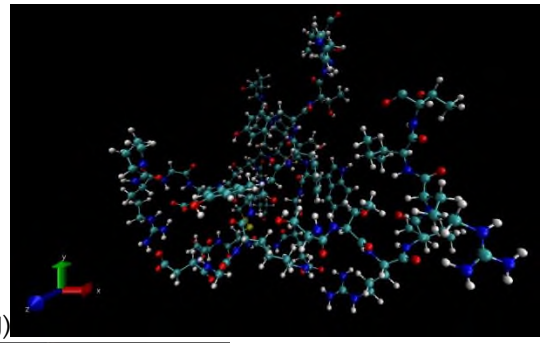
D) 

E) 

F) 

G) 

H) 

I)  J) 

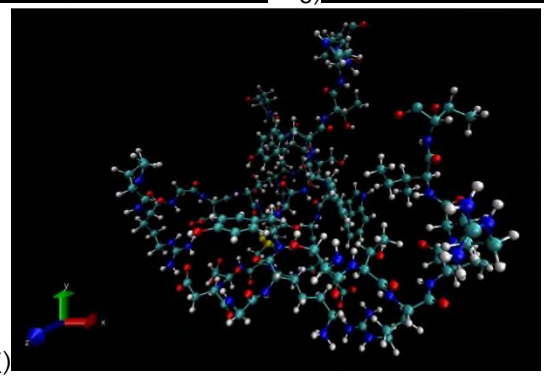
K) 

Figura 24: Sitios de unión de la trombina en los diferentes tiempos A) 0.0 ns B) 0.5 ns C) 1.0 ns D) 1.5 ns E) 2.0 ns F) 2.5 ns G) 3.0 ns H) 3.5 ns I) 4.0 ns J) 4.5 ns K) 5.0 ns.

En la Figura 25 podemos observar los valores del error absoluto medio de los modelos comparada con la R^2 calculada para RLM y de las RNA. En esta figura se observa que para el caso del modelo en 1.5 ns los puntos se encuentran empalmados, esto es debido a que los valores presentados son muy similares entre ellos. Este modelo presentó el valor de R^2 más alto tanto para la MLR y la RNA, siendo de 0.9354 y de 0.9922 respectivamente. Para el caso del R^2RNA calculado y el MAE se realizó una línea para seguir su tendencia ambos valores fueron más altos para el modelo 1.5 ns siendo 0.9961 para el caso de R^2RNA y 0.9977 para el MAE; al observar la tendencia que presentan estos valores podemos observar que presentan un comportamiento similar en los primeros nanosegundos y el punto donde difieren en valores pero no en comportamiento porque ambos presentan un punto bajo es en el modelo 2.5 ns, volviendo a presentar valores similares en el modelo en 3.0 ns y después una serie de altibajos pero con valores que difieren bastante entre uno y otro terminando con una diferencia bastante obvia en el modelo en 5.0 ns. El modelo ideal sería aquel con una R^2 alta muy cercana a uno y un valor de MAE muy bajo muy cercano a cero; por lo tanto, aunque el modelo

1.5 ns presenta una R^2 muy buena su valor de MAE es muy elevado esto indicaría que el error del modelo dista mucho del valor medio de todos los modelos. Esta diferencia puede deberse a que no hay presencia de una tendencia específica del error de los modelos; es decir cada modelo presenta un error variable comparado con los demás modelos.

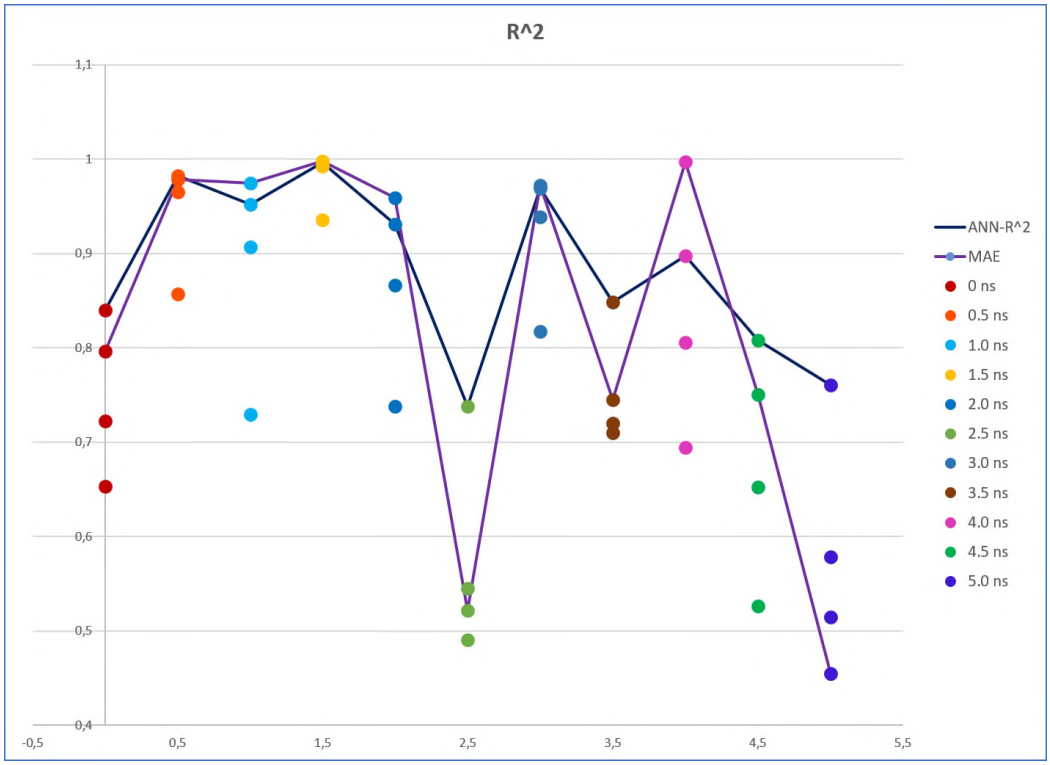


Figura 25: Valores de los modelos de la R^2 y el valor de error absoluto medio.

En la Figura 26 podemos observar el comportamiento del análisis por estandarización de las moléculas a través del tiempo, al realizar este análisis los valores presentes en los modelos se centralizan, así que de este modo aquellos modelos que indiquen un cambio considerable en el modelo serán fácilmente identificables, esto se realiza con un límite superior que determina el área idónea, estos valores que sobrepasan el límite son analizadas posteriormente; los primeros dos puntos que se observan (color rosa) corresponden al modelo 2.0 ns y corresponden a las moléculas m012 y m289; después se observan 3 puntos que son valores atípicos de la molécula m121, una corresponde al modelo 3.5 ns (punto lavanda), y los otros dos son puntos empalmados que corresponden a los modelos

0.5 ns (azul oscuro) y 4.5 ns (gris); por último el punto rojo corresponde a la molécula 298 del modelo en 2.5 ns. Estos valores que se encuentran en la parte superior del límite se comportan como valores atípicos en los modelos; sin embargo, no todos los valores atípicos encontrados en los modelos aparecen de esta manera en el gráfico de análisis de estandarización. Los modelos que tuvieron un mejor comportamiento, es decir; no presentaron valores atípicos fueron los modelo 0.0 ns, 1.0 ns, 1.5 ns, 3.0 ns, 4.0 ns y 5.0 ns.

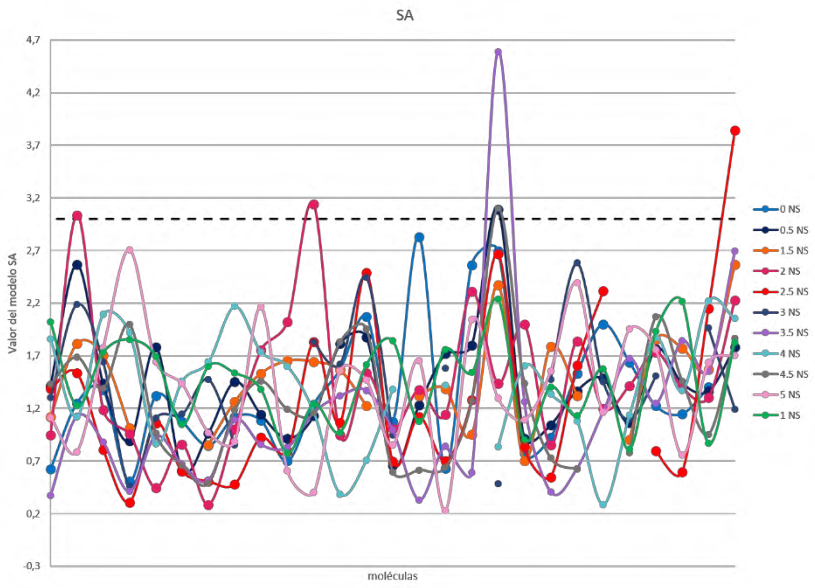


Figura 26: Valores de los modelos obtenidos por Análisis de estandarización SA.

Al analizar el modelo en 0.5 ns, se observa que su valor atípico correspondiente a la molécula 121 (2-[-4-Oxibenzamidina]-7-Bromo-4-Metilquinolina), la cual se presentó como valor atípico en 3 modelos diferentes: en 0.5, 3.5 y 4.5ns.

En el caso del modelo en 0.5 ns se utilizaron 5 descriptores: SIC0, Rv8+, Mor11p, H8u y HATS6p. Al verificar los datos se observó que para el caso de la molécula mol121 los valores correspondientes a los descriptores SIC0 y HATS6p fueron los más altos de todo el modelo para que esta molécula dejara de presentarse como valor atípico debía reducir su valor del descriptor SIC0 un mínimo de 0.007 unidades, y para HATS6p un mínimo de 0.004 unidades, para los descriptores restantes la molécula presento los valores promedio en los descriptores; además el descriptor

HATS6p solo se encontró en este modelo lo cual lo hace diferir de los demás modelos.

El descriptor SIC0 es un descriptor topológico que toma en cuenta todos los átomos y considera el contenido de información proporcionado por varias clases de átomos en función de sus átomos vecinos y el descriptor HATS6p es un descriptor GETAWAY ponderado por la polarizabilidad de retraso 6; ambos descriptores presentan información interna de la molécula y la influencia de un átomo sobre otro átomo, esto puede deberse a la presencia del átomo de Br en su estructura, ya que las tres moléculas que presentan el átomo de Br en su estructura presentaron valores atípicos en los subgrupos de predicción y validación; al solo existir tres moléculas con átomos de bromo y estar cada una repartida en un diferente subgrupo, estas moléculas no son representativas del resto del modelo; este comportamiento ocurre de manera similar en los modelos en 3.5 y 4.5 ns.

En el caso del modelo en 3.5 ns se observaron tres descriptores R1m+, HATS7p y GATS8m, para la misma molécula mol121 se presentó como valor atípico, para descifrar los valores con los que se dejaría de comportar como valor atípico se observó que al modificar los valores aparecían nuevos valores atípicos, sin embargo se puede observar que R1m+ es un descriptor GETAWAY que también se presenta en el modelo en 2.5 ns que presentó un valor atípico y que HATS7p al igual que HATS6p es un descriptor GETAWAY que también presentó valores atípicos en el modelo en 0.5 ns, el último descriptor de este modelo es un descriptor de autocorrelación 2D que solo se presentó en este modelo. Las demás moléculas con átomos de Br presente son valores atípicos en los demás subgrupos.

De igual manera la molécula 121 se presentó como valor atípico en el modelo en 4.5 ns en el cual se obtuvieron dos descriptores el descriptor HATS7v y el descriptor SIC0 que al igual que en el modelo en 0.5 ns presentó el valor más alto del grupo y con el mismo valor que el modelo 0.5 ns pero en este caso para que el valor dejara de comportarse como atípico en lugar de restarle los 0.007 para que entrara

en el rango había que adicionárselos; de igual manera las demás moléculas con átomos de Br presentaron valores atípicos.

En el caso del modelo en 2.0 ns se presentaron dos moléculas que se comportaba como atípicos en el gráfico uno corresponde a la molécula mol012 y el otro a la molécula 289; para este modelo se utilizaron tres descriptores R8p+, RDF030e y Mor22m; al analizar los valores de estos descriptores para la molécula 012 se observó que en el descriptor R8p+ presentaba el valor más alto entre todas las moléculas siendo para los otros dos descriptores valores dentro del promedio, para que el punto de la molécula 012 deje presentarse como atípico el valor para el descriptor R8p+ debe ser reducido en 0.001; este descriptor pertenece a los descriptores GETAWAY es un descriptor de autocorrelación máxima y esta ponderado por la polarizabilidad atómica.

Para el caso de la molécula 289 presentó valores dentro del promedio para los descriptores R8p+ y Mor22m, para el caso del descriptor RDF030e presentó valores muy altos comparado con las demás moléculas, para que este valor deje de tener un comportamiento atípico debe ser menor en 1.082 unidades, este descriptor pertenece a los descriptores RDF, este tipo de descriptores relaciona la presencia de átomos electronegativos en un radio interno, este descriptor esta ponderado por la electronegatividad atómica de Sanderson.

En el caso del modelo en 2.5 ns se utilizaron dos descriptores, R1m+ y HATS7v, para el caso del descriptor HATS7v los valores están dentro del promedio para la molécula 298, sin embargo, para el descriptor R1m+ el valor es muy alto comparado con los valores de las demás moléculas del modelo, para que este deje de presentarse como valor atípico debe reducir su valor 0.092 unidades, este descriptor pertenece a los descriptores GETAWAY y esta ponderado por masas atómicas con una autocorrelación máxima, esto podría deberse por la presencia del átomo de azufre que se encuentra en la estructura ya que en el subgrupo de validación se presentaron moléculas con estructura similar y con presencia de azufre que a su

vez presentaron valores atípicos; que son valores que se encuentran afuera del límite superior.

Así que para el análisis de estandarización el mejor modelo fue el modelo 1.0 ns ya que no presentó valores atípicos ni valores descentralizados.

Con el análisis de apalancamiento podemos evaluar que tan poco comunes son los valores de los descriptores de los modelos; comparando sus valores con la media de estos.

Al realizar el análisis de apalancamiento se observó que algunos modelos compartían entre si el valor de h*, como se puede observar en la Figura 19; este valor depende del número de variables obtenidas y del número de moléculas analizadas, h* es inversamente proporcional al número de moléculas analizadas (3p/n; p: descriptores n: moléculas). Las moléculas que comparten el valor de h* fueron 2.0 y 3.5 ns con una h* de 0.444, también los modelos en 2.5 y 4.5 ns presentan el mismo valor de h* con 0.346; esto se debe a que comparten la cantidad de variables y moléculas a analizar. Así mismo, los modelos en 2.0 y 3.5 ns se obtuvieron utilizando 3 variables y 27 moléculas, por lo cual comparten el valor de h*.

Este análisis busca comparar el valor de h de cada molécula con el valor de h* utilizando todos los descriptores en el modelo; por lo tanto, los valores que superan el límite superior son aquellos que presentan una gran diferencia de los valores y son denominados como valores atípicos.

Los modelos en 2.5 y 3.5 ns presentaron valores atípicos. Para el caso del modelo en 2.5 ns el valor atípico corresponde a la molécula 298, el cual efectivamente es por mucho el valor más alto en el modelo elevando el promedio del modelo. Esta molécula contiene un átomo de azufre en su estructura, lo cual probablemente eleva su valor. En un análisis más general, se encontró que algunas otras moléculas

de otros subgrupos presentan valores similares. El valor de h* del modelo en 2.5 ns fue de 0.35 y el valor h para la molécula 298 fue 0.50.

De igual manera la molécula 298 se presentó como valor atípico en el modelo en 3.5 ns, presentando el valor más alto de su modelo 0.49 siendo el valor de h* 0.44; este modelo presenta un comportamiento similar al caso anterior.

Para este análisis el mejor modelo sería aquel en el que los componentes del modelo no presenten altas diferencias entre su h* y su valor de cada molécula; por lo tanto, se determinó que el mejor modelo es el correspondiente al modelo en 1.0 ns el cual presenta menos valores altos lo cual indica baja diferencia entre sus valores y h*.

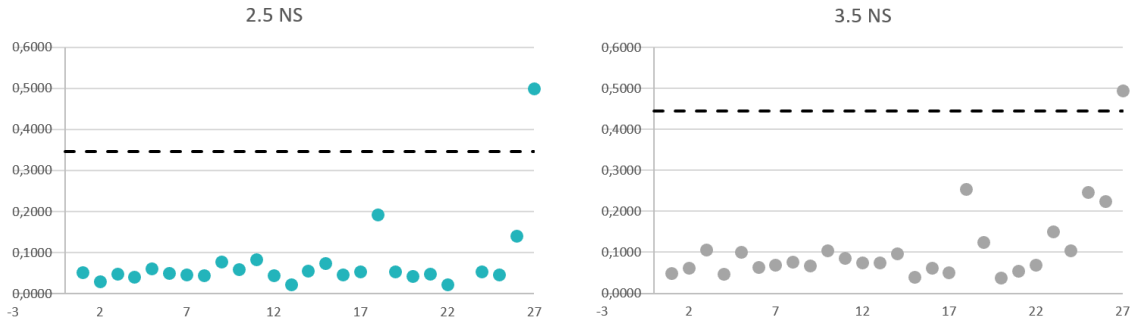


Figura 27: Comportamiento de los modelos 2.5 ns con un valor de h*=0.3462 y 3.5 ns con un valor de h*=0.4444 al realizar el análisis de apalancamiento (LA).

Para probar si los modelos calculados presentan correlación al azar, revisamos estos por medio de un programa escrito en MATLAB. Entre menor sean los valores del coeficiente de correlación, menor será la correlación azarosa. Los modelos con la correlación más baja corresponden a los modelos en 2.5, 4.5 y 5 ns.

Observando la Figura 28 podemos visualizar que el valor más alto se presentó en el modelo en 1.5 ns para ambos casos entrenamiento y predicción; para el caso de predicción se obtuvo un valor de 0.30 y para el caso de entrenamiento 0.28. Los valores más bajos para el caso de entrenamiento corresponden a los modelos en 4.5 y 5.0 ns con un valor de 0.075 y para el caso de predicción es el modelo en 2.5 ns con un valor de 0.09.

Por otro lado, en el caso de los modelos 0.0 ns los valores tanto para el caso de entrenamiento y predicción se empalman con un valor de 0.13.

Los valores de los coeficientes de correlación en la Figura 28 son el promedio de repetir el procedimiento 30 veces de forma azarosa.

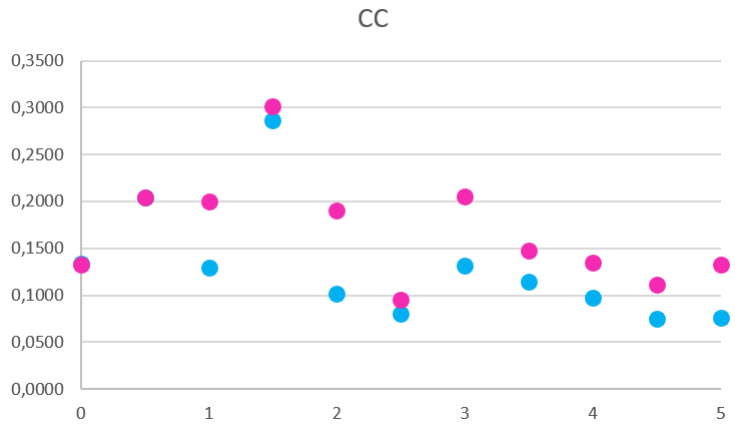


Figura 28: Resultados obtenidos del análisis de correlación azarosa para todos los modelos tanto en los conjuntos de entrenamiento en color azul y de predicción en color rosa.

Para calcular la capacidad predictiva de los modelos se utiliza el análisis de validación cruzada una manera de realizarlo es con el método de dejar uno afuera (LOO), este método recalcula dejando una observación fuera a la vez, es importante mencionar que este análisis solo se realiza para los resultados obtenidos con la RLM.

El modelo ideal presentaría un valor de 1; en este caso ningún modelo presentó este valor, siendo nuestro mejor modelo el correspondiente a 0.0 ns con un valor de validación cruzada de 0.55, lo cual indica que es el modelo con más alta capacidad predictiva que tenemos y el modelo con menor capacidad predictiva es el modelo en 4.5 ns con el valor 0.0 unidades lo cual podría interpretarse como capacidad predictiva nula.

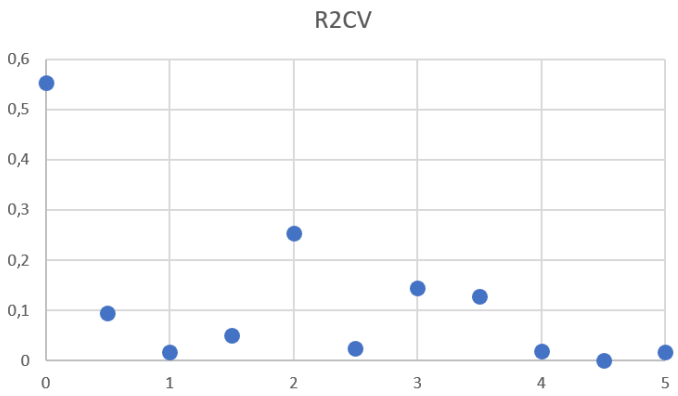


Figura 29: Coeficiente de validación cruzada para el grupo de entrenamiento con el método de LOO.

En el caso del coeficiente de la correlación cruzada un valor de 1 nos indicaría que los modelos presentan un 100% de correlación por lo tanto si observamos la Figura 30, podemos determinar que el conjunto que presentó mejores valores fue el de entrenamiento y el conjunto con los peores valores es el de validación. El modelo con mejores valores pertenece al grupo de entrenamiento y es el modelo en 1.5 ns con un valor de 0.87, lo cual es un valor bastante aceptable, seguido de los modelos en 0.5 ns del grupo de entrenamiento y el modelo en 4.0 ns del grupo de validación ambos con un valor de 0.76. El modelo con el valor más bajo para el coeficiente de correlación pertenece al grupo de validación, siendo el modelo 1.5 ns con un valor de 0.00 que indicaría que no existe correlación alguna, los demás modelos presentaron una correlación baja. No es de extrañarse que los modelos con coeficientes más bajos pertenezcan al grupo de validación ya que este grupo en particular esta creado con un número de moléculas reducido y cada molécula contiene átomos diferentes entre ellas lo que ocasionar que estas moléculas no tengan correlación alta entre ellas o no tengan.

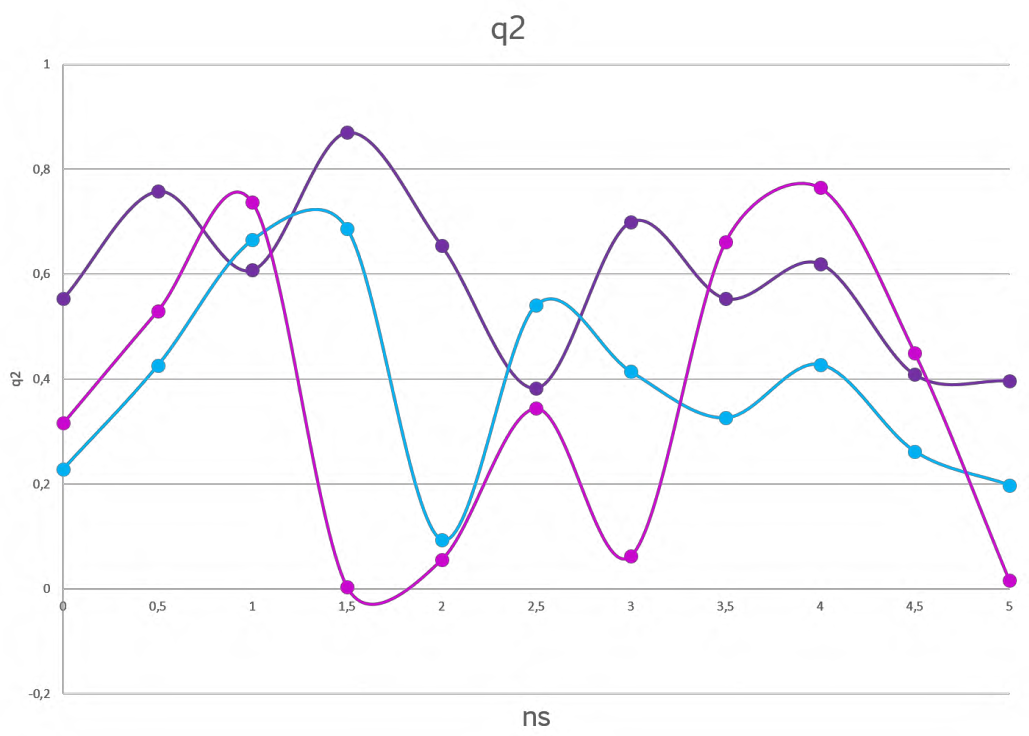


Figura 30: Coeficiente de validación cruzada de los grupos de entrenamiento representados en color morado, predicción en color azul y validación con color rosa.

El cálculo de la raíz cuadrada del error cuadrático medio mide el grado de variación entre los datos calculando la diferencia entre el valor real de cada punto y el valor de ajuste; por lo tanto, el valor más bajo de los modelos será aquel que esté más cerca del valor ideal debido a que indicará una variación mínima y por lo tanto podríamos decir que es un buen modelo. En este caso el valor que presenta su valor más cercano a cero corresponde al modelo en 0.0. ns con un valor de 0.63, lo cual lo convierte en nuestro mejor modelo, seguido del modelo en 2.5 ns con un valor de 1.07 unidades. En caso opuesto los modelos que obtuvieron valores altos son los que mayor diferencia presentan una mayor diferencia con el valor ideal; el más alto corresponde al modelo en 1.5 ns con un valor de 1.30 unidades, seguido de los modelos en 0.5 y 3.0 ns ambos con un valor de 1.23 unidades.

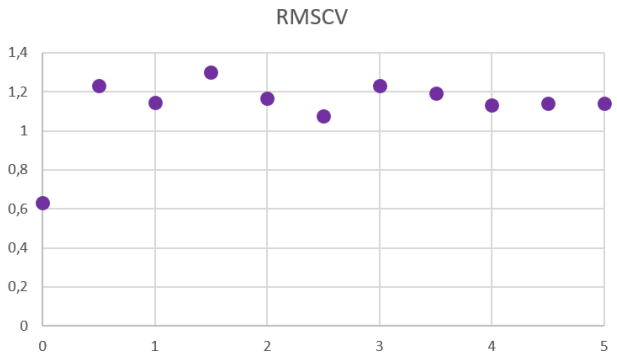


Figura 31: Resultados de la Raíz del Error cuadrático de la media en los diferentes modelos.

En la Tabla 9 podemos observar los modelos que se obtuvieron del análisis QSAR en los diferentes tiempos analizados, los mejores modelos obtenidos corresponden a los modelos 0.5ns, 1.0 ns, 1.5 ns y 3.0 ns; estos modelos comparten la presencia de los descriptores GETAWAY, en general la mayoría de los descriptores encontrados en los modelos se relacionan con la estructura interna de la molécula y su distribución, ya que los descriptores están ponderados por electronegatividades, polarizabilidad, forma y estructura. En la referencia Ramírez-Galicia y colaboradores [51]; mencionaron que la interacción ligando-proteína se debía a las interacciones $\pi - \pi$, interacciones electrostáticas, $\pi -$ catión e hidrofóbicas; siendo las interacciones electrónicas las más destacables en estos modelos.

Modelo	Ecuación.
Ecuación Nativa	$\log IC_{50}^{Calc}$=7.003+22.653SIC_0+0.697MAXDN+0.544$G_{internal}$-9.876R4m$^+$-6.584RBF
0.0 ns	$\log IC_{50}^{Calc}$=-10.824+4.851R1m-12.672R7m+1.672IDDE
0.5 ns	$\log IC_{50}^{Calc}$=-5.588+29.519SIC_0-63.517R8v$^+$+1.468Mor11p-1.605H8u-17.588HATS6p
1.0 ns	$\log IC_{50}^{Calc}$=-42.888-1.299GATS8e+4.069E1s-24.782R3u$^+$+45.449MATS4m
1.5 ns	$\log IC_{50}^{Calc}$=-47.674+62.97MATS8m+0.635Mor16m+11.806R1e$^+$+0.451DX-58.034X1A+39.15G1u
2.0 ns	$\log IC_{50}^{Calc}$=4.029-120.39R8p$^+$-0.076RDF030e-0.835Mor22m
2.5 ns	$\log IC_{50}^{Calc}$=0.616+6.707R1m$^+$-35.499HATS7v
3.0 ns	$\log IC_{50}^{Calc}$=0.214+6.173R1m-7.566HATS7u-65.465PW5
3.5 ns	$\log IC_{50}^{Calc}$=1.395+7.452R1m$^+$-24.591HATS7p-173.075GATS8m
4.0 ns	$\log IC_{50}^{Calc}$=-6.299+1.943Mor16m+21.864JGI1
4.5 ns	$\log IC_{50}^{Calc}$=-9.602+41.454SIC_0-32.971HATS7v
5.0 ns	$\log IC_{50}^{Calc}$=-5.111+3.821R1m+0.33DY

Tabla 9: Ecuaciones que relacionan la actividad biológica con los descriptores en cada modelo.

11. Conclusión

El realizar el análisis QSAR permitió determinar el mejor modelo de RLM que representa mejor el comportamiento e interacción de las 50 moléculas derivadas de quinolina y quinoxalina en el sitio inhibidor de la trombina en el transcurso del tiempo; el mejor modelo corresponde en 1.5 ns, el cual presenta descriptores importantes, tales como 3D MoRSE, WHIM y GETAWAY. El modelo RLM en 1.5 ns está caracterizado por la influencia entre sus átomos en la distribución molecular; así mismo se observa la presencia de pares atómicos electronegativos y polarizables. La actividad biológica está altamente relacionada con la estructura tridimensional de la molécula y las propiedades electrónicas de sitios específicos de unión del ligando y el receptor; así que la presencia de los descriptores mencionados representa los efectos fisicoquímicos necesarios para proporcionar la información necesaria y aportar una contribución significativa.

Los modelos fueron mejorados con la aplicación de las Redes Neuronales Artificiales obteniendo en este caso 4 mejores modelos, en 0.5 ns, 1.0, 1.5 y 3.0 ns; todos estos modelos presentaron en común al menos un descriptor GETAWAY. Estos modelos están correlacionados por la geometría y por los átomos del ligando y con el sitio receptor; los cuales han sido altamente evaluados para modelos QSAR.

12. Perspectivas

- Es necesario analizar los 50 inhibidores de la trombina en un tiempo mayor al realizado en la presente tesis; para evaluar el comportamiento de las moléculas con la trombina de manera amplia.
- Evaluar nuevas moléculas reportadas con posible actividad biológica en la inhibición de trombina, para comprobar los modelos presentados en esta tesis.
- Diseñar a partir de los modelos obtenidos en esta tesis, una molécula con posible actividad biológica en la inhibición de la trombina, y volver a evaluar los modelos para verificar la eficacia de estos.

13. Bibliografía

[1]. R. Frédérick, S. Robert, C. Charlier, J. Wouters, B. Masereel, L. Pochet. Mechanism-based trombin inhibitors: design, synthesis, and molecular docking of a new selective 2-oxo-2H-1-benzopyran derivative. J. Med. Chem. 2007, 50, 3645-3650.

[2]. MR. Wiley, MJ. Fisher. Small molecules direct trombin inhibitors. Exp. Opin. Ther. Pat. 1997, 7, 1265-1282.

[3]. Datos (no publicados) proporcionados por el Dr. José Correa Basurto. Laboratorio de Modelado Molecular y Bioinformática, Escuela Superior de Medicina, Instituto Politécnico Nacional.

[4]. K. Mena-Ulecia, W. Tiznado, J. Caballero. Study of the differential activity of thrombin inhibitors using docking, QSAR, molecular dynamics and MM-GBSA. PLoS One 2015, 10, e0142774.

[5]. Formación de trombos <https://www.anticoagulante.es> (Pagina consultada en Noviembre del 2013)

[6]. C. Martínez Murillo, S. Quintana Gonzales. Factores de riesgo para trombosis. Rev. Hematol. Mex. 2005, Vol. 6, pp 1-8.

[7]. J. Mark Berg, L. Stryer, JL. Tymoczko, Bioquímica, Reverte, 2008, pp 296, 297.

[8]. A. Majluf-Cruz, F. Espinosa-Larrañaga. Fisiopatología de la Trombosis. Gac Méd Méx. 2007, 143, pp 11-14.

[9]. R. Carrillo-Esper, CR. Arias-Delgadillo, D. Sánchez-Ríos. Inhibidores directos de trombina. Med Int Mex 2011, 27, 38-51.

[10]. Tratamiento con anticoagulantes y antiagregantes plaquetarios de la enfermedad cerebrovascular < https://www.svneurologia.org/congreso/vascular-1.html > (Pagina consultada en Enero del 2014)

[11]. C. Trejo. Anticoagulantes: Farmacología, mecanismos de acción y usos clínicos. Cuad. Cir. 2004, 18, 83-90.

[12]. AR. Gennaro, Remington Farmacia Volumen II, Ed. Médica Panamericana, 2003, Capitulo 67 pp. 1471.

[13]. MI. Nicolas Vazquez, E. Marín Chiñas, FM. Castro Martínez, R. Miranda Ruvalcaba. Algunos aspectos básicos de la Química Computacional.. Editor UNAM. ISBN

9703233074, 9789703233076, Abril 2006, Facultad de estudios superiores Cuautitlán

[14]. GA. Magos Guerrero, M. Lorenzana-Jiménez: Las fases en el desarrollo de nuevos medicamentos. Rev Fac Med UNAM 2009,Vol 52, No. 6, pp 260-264.

[15]. H. Martínez Pacheco. Evaluación in sílico -Docking y QSAR- de derivados del ácido valproico sobre la HDAC8 con potencial uso antineoplásico. Universidad del Papaloapan, Campus Tuxtepec. 2012.

[16]. H. Waterbeemd, E. Gifford. ADMET in silico modeling: towards prediction paradise? Nature Publishing Group. 2003, 2, 192-204.

[17]. JD. Durrant, JA. McCammon. AutoClickChem: Click Chemistry in Silico. PLoS Comput Biol 2012, 8, e1002397.

[18]. EK. Brockmeier, G. Hodges, TH. Hutchinson, E. Butler, M. Hecker, KE. Tollefsen, N. Garcia-Reyero, P. Kille, D. Becker, K. Chipman, J. Colbourne, TW. Collette, A. Cossins, M. Cronin, P. Graystock, S. Gutsell, D. Knapen, I. Katsiadaki, A. Lange, S. Marshall, SF. Owen, EJ. Perkins, S. Plaistow, A. Schroeder, D. Taylor, M. Viant, G. Ankley, F. Falciani. The Role of Omics in the Application of Adverse Outcome Pathways for Chemical Risk Assessment. Tox. Sci. 2017, 158, 252-262.

[19]. QSAR Relación estructura-Actividad cuantitativa. Educación virtual. Facultad de ciencia bioquímicas <http://www.fbioyf.unr.edu.ar/evirtual/pluginfile.php/103521/mod_resource/content/1/QSAR.pdf> (Pagina consultada en Diciembre del 2013)

[20]. Investigación de un fenómeno natural: ¿Estudios *in vivo, in vitro* o *in silico.* Brenda Lorena Fina, Mercedes Lombarte, Alfredo Rigalli. Laboratorio de Biología Ósea, Facultad de Ciencias Médicas. Universidad Nacional de Rosario, Santa Fe 3100, Rosario, Argentina. Actual Osteol 2013 Vol 9(3), 2013 pp 283-288.

[21]. AC. Martínez Olguín. Estudio teórico de la reactividad de nuevos nanocúmulos de Al-B con potencial uso como acumuladores de Hidrógeno. Universidad del Papaloapan, Campus Tuxtepec. 2016

[22]. En el apartado se presentan los conceptos básicos sobre mecánica cuántica. Para mayor información se recomienda consultar: Química Cuántica, Ira N. Levine, Quinta edición, Pearson Educación S.A., Madrid, 2001.

[23]. Química Cuántica, Ira N. Levine, Quinta edición, Pearson Educación S.A., Madrid, 2001, Capitulo 16.

[24]. LEl. Bailey Chapman, MD. Troitiño Nuñez. Química Cuántica. La química cuántica en 100 problemas. Universidad Nacional de Educación a Distancia. Madrid 2015.

[25]. EG. Lewars. Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics. Third Edition. Springer International Publishing Switzerland 2016. Chapter 6: Semiempirical Calculations.

[26]. KI. Ramachandran, G. Deepa, K. Namboori. Computational Chemistry and Molecular Modeling: Principles and Applications. Spinger Verlag Berlin Heidelberg 2008. Chapter 7: Semiempirical Methods.

[27]. Cursos de Química Computacional. <http://www.agalano.com/Cursos/QC/P10_Semiempiricos.pdf> (Pagina consultada en Mayo del 2017).

[28]. GaussView 3.0, Gaussian, Inc. Carnegie Office Park-Building 6, Pittsburgh PA 15106 USA. Copyright (c) 2000-2003. Semichem, Inc.

[29]. Pérez-Nueno VI. Herramientas de cribado virtual aplicadas a inhibidores de entrada del VIH. Diseño de nuevos compuestos anti-VIH. Tesis de Doctorado, Departamento Química Orgánica. Escola Técnica Superior IQS: Barcelona; 2009.

[30]. G. Gini, MV. Craciun, Ch. König, E. Benfenati Combining Unsupervised and Supervised Artificial Neuronal Networks to Predict Aquatic Toxicity. J. Chem. Inf. Comput. Sci. 2004, 44, 1897-1902.

[31]. O. Deeb, H. Martínez.-Pacheco, G. Ramírez-Galicia, R. Garduño-Juárez. Applied Case Studies and Solutions in Molecular Docking-Based Drug Design; Chapter 2: Application of docking Methodologies in QSAR-Based Studies. Published by Medical Information Science Reference (an imprint of IGI Global) ISSN: 2327-9354, 2016.

[32]. R. Todeschini, V. Consonni. Molecular Descriptor for Chemoinformatics,. Second, Revised and Enlarged Edition. Published by Wiley-VCH 2009. ISBN 978-3-527-31852-0. QSAR/QSPR Modeling.

[33]. Enfermedades cardiovasculares; Nota descriptiva enero 2015 Organización Mundial de la Salud <http://www.who.int/mediacentre/factsheets/fs317/es/> (Pagina consultada en Agosto del 2016).

[34]. S. Andrade Ochoa. Estudio QSAR de compuestos encontrados en diversos aceites esenciales con actividad antiparasitaria, antifúngica y antimicrobiana. Universidad Autónoma de Chihuahua, Facultad de Ciencias Químicas. 2014.

[35]. J. Devillers, AT. Balaban. Topological Indices and Related Descriptors in QSAR and QSPR. Gordon and Breach Science Publishers 1999. Chapter 1: No-Free-Lunch Molecular Descriptors is QSAR and QSPR.

[36]. H. Pedrosa, L. Dicovsky. Sistema de Análisis Estadístico con SPSS. Instituto Nicagüarense de Teconología Agropecuaria. Managua, Nicaragua 2006.

[37]. Introducción a la Regresión Lineal Múltiple https://rpubs.com/Joaquin_AR/226291 (Consultada en febrero 2018)

[38]. R. Johnson, P. Kuby. Estadística elemental: Lo esencial; Décima Edición, Cengage Learning. 2008. Capítulo 3: Análisis descriptivo y presentación de datos bivariados. Sección 3.3: Correlación lineal.

[39]. Correlación cruzada <https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/time-series/how-to/cross-correlation/interpret-the-results/all-statistics-and-graphs/> (Pagina consultada en Octubre 2017).

[40]. "EEE305", "EEE801 Part A": Digital Signal Processing; Chapter 6: Describing Random Sequences; University of Newcastle upon Tyne. https://www.staff.ncl.ac.uk/oliver.hinton/eee305/Chapter6.pdf (Pagina consultada en Noviembre 2017).

[41]. JW. Shavlik, TG. Dietterich. Readings in machine learning. Morgan Kaufman Publishers, Inc. Chapter 2: Inductive learning from preclassified training examples; 2.2.2 A theory and methodology of inductive learning pp 78.

[42]. Análisis de regresión <https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/types-of-regression-analyses/> (Pagina consultada en Octubre del 2017)

[43]. Ajustar modelo de regresión <https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/regression/how-to/fit-regression-model/interpret-the-results/all-statistics-and-graphs/method-table/> (Pagina consultada en Octubre del 2017).

[44]. Maneras de identificar valores atípicos en regresión y ANOVA <https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-

statistics/regression/supporting-topics/model-assumptions/ways-to-identify-outliers/> (Pagina consultada en Noviembre del 2017).

[45]. Desviación estándar <https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/supporting-topics/data-concepts/what-is-the-standard-deviation/> (Pagina consultada en Noviembre del 2017).

[46]. Aplicaciones de las redes de neuronas en supervisión, diagnosis y control de procesos. María Jesús de la Fuente Aparicio; Eloisa Susana Gonzalez Palenzuela; Jesús María Zamarreño Cosme; Teodoro Calonge Cano. Ediciones de la Universidad Simón Bolívar. Capítulo I-1 Concepto generales – Modelos y arquitectura, pp 11.

[47]. JC. Brégains. Análisis síntesis y control de diagramas de radiación generados por distribuciones continuas y discretas utilizando algoritmos estadísticos y redes neuronales artificiales aplicaciones. Facultad de Física, Universidad de Santiago de Compostela, 2007.

[48]. R. Pino Diez, A. Gómez Gómez, N. de Abajo Martínez. Introducción a la inteligencia artificial: Sistemas expertos, redes neuronales artificiales y computación evolutiva. Universidad de Oviedo, Servicio de Publicaciones, 2001. Capítulo 3: Redes Neuronales Artificiales.

[49]. R. Flórez López, JM. Fernández Fernández. Las redes neuronales artificiales Fundamentos teóricos y aplicaciones prácticas, Gesbiblo 2008. Capítulo 1: Fundamentos Biológicos de las Redes Neuronales Artificiales, Capítulo 2: Las redes neuronales artificiales: Aspectos generales.

[50]. R. Andrews, J. Diederich, AB. Tickle. Survey and critique of techniques for extracting rules from trained artificial neuronal networks. Knowledge-Based Systems Volume 8 Number 6 December 1995. Pp 373-389.

[51]. G. Ramírez-Galicia, R. GarduñoJuárez., J. Correa-Basurto., O. Deeb. Exploring QSAR for inhibitory effect of a set of heterocyclic thrombin inhibitors by multilinear regression refined by artificial neuronal network and molecular docking simulations. J. Enz. Inhib. Med. Chem. 2012; 27, 174–186.

[52]. G. Ramírez-Galicia, R. Garduño-Juárez, B. Hemmateenejad, O. Deeb, M. Deciga-Campos, JC. Moctezuma-Eugenio. QSAR Study on the Antinociceptive Activity of Some Morphinans. *Chem. Biol. Drug Des.* 2007, 70, 53-64.

[53]. G. Ramírez-Galicia, R. Garduño-Juárez, B. Hemmateenejad, O. Deeb, S. Estrada-Soto. QSAR Study on the Relaxant Agents from Some Mexican Medicinal Plants and Synthetic Related Organic Compounds. Chem. Biol. Drug Des. 2007, 70, 143-53.

[54]. G. Ramírez-Galicia, R. Garduño-Juárez, B. Hemmateenejad, O. Deeb. MLR-ANN and RTO Approach to -Opioid Receptor Binding Affinity. Pooling Data from Different Sources. *Chem. Biol. Drug Des.* 2008, 71, 260-70)

[55]. G. Ramírez-Galicia, H. Martínez-Pacheco, R. Garduño-Juárez, O. Deeb. Exploring QSAR of Antiamoebic Agents of Isolated Natural Products by MLR, ANN and RTO Models. *Med. Chem. Res.*2012, 21, 2501-2516.

[56]. M.C. Contreras-Romo, M. Martínez-Archundia, O. Deeb, M.J. Sluzar, G. Ramírez-Salinas, R. Garduño-Juárez, A. Quintanar-Stephano, *G. Ramírez-Galicia*, J. Correa-Basurto Exploring the ligand recognition properties of the human vasopressin V1a receptor using QSAR and molecular modeling studies. *Chem. Biol. Drug Disg.* 2014, 83, 207-223.

[57]. A. Mauri, V. Consonni, M. Pavan, R. Todeschini. DRAGON software: An easy approach to molecular descriptor calculations match. Commun. Math. Comput. Chem. 2006, 56, 237-248.

[58]. Tutorial MATLAB <https://la.mathworks.com/support/learn-with-matlab-tutorials.html> (Pagina consultada en Enero 2014).

[59]. I. Putinhon Caruso, JM. Barbosa Filho, A. Suman de Araújo, F. Pereira de Souza, MA. Fossey, M. Lopes Cornélio. An integrated approach with experimental and computational tools outlining the cooperative binding between 2-phenylchromone and human serum albumin. Food Chem., 2016, 196, 935-942.

[60]. C. Helma, S. Kramer. SPSS Statistics Base 17.0 Useŕs Guide: Web site at http://www.spss.com35 A survey of the Predictive Toxicology Challenge 2000-2001. Bioinformatics. 2003, 19, 1179-1182

[61]. GR Hutchison, Ch. Morley Craig, JCh. Swain, H. De Winter, T. Vandermeersch, NM. O'Boyle (Ed.) Open babel documentation. Diciembre 5, 2011 <http://openbabel.org/docs/current/OpenBabel.pdf>

[62]. JC. Phillips, R. Braun, W. Wang, J. Gumbart, E. Taikhorshid, E. Villa, C. Chipot, RD. Skeel, L. Kale, K. Schulten. Scalable molecular dynamics with NAMD. J. Comput. Chem. 2005, 26, 1781-1802.

[63]. Manual de Gaussian 09 <http://users.etown.edu/h/hoffmang/Manuals/g09ur/k_geom.htm> (Pagina consultada en Agosto del 2017).

[64]. R. Todeschini, V. Consonni. Descriptor from Molecular Geometry; <http://michem.disat.unimib.it/chm/download/materiale/geometrical_descriptors.pdf> (Pagina consultada en Septiembre del 2017).

[65]. VCCLAB, Virtual Computational Chemistry Laboratory, http://www.vcclab.org, 2005. (Pagina consultada en Septiembre del 2017).