Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Química Biológica

Desarrollo y aplicaciones de estrategias avanzadas en docking molecular: impacto de la flexibilidad del receptor, búsqueda de sustratos de los BacCYPS y búsqueda virtual de antimicrobianos

Trabajo de tesis presentado para optar por el título de doctor
de la Universidad de Buenos Aires en el área Química Biológica

Lic. Juan Manuel Prieto

Director: Marcelo A. Martí
Lugar de Trabajo: Departamento de Química Biológica, FCEN-UBA e INQUIBICEN-UBA/CONICET
Fecha de Defensa: 3 de Julio de 2025
Ciudad Autónoma de Buenos Aires

Resumen

Esta tesis doctoral se explora el uso de métodos computacionales, principalmente docking molecular, para estudiar las interacciones proteína-ligando y, así como para facilitar la búsqueda virtual de nuevos ligandos con potencial actividad biológica. La precisión del docking se ve afectada por múltiples factores, entre ellos la correcta representación del entorno del sitio de unión y la flexibilidad del receptor. Para abordar estas limitaciones, el estudio incorpora estrategias avanzadas como la inclusión de "sitios de solvente" y Bias Docking, junto con el análisis de la flexibilidad conformacional del receptor.

Uno de los principales desafíos en el docking molecular es describir y predecir cómo un receptor proteico acomoda su estructura al interactuar con un ligando. En muchos casos, estas interacciones inducen cambios conformacionales que pueden variar desde pequeños ajustes en las cadenas laterales del sitio activo hasta modificaciones estructurales globales que alteran significativamente la conformación de la proteína. Mientras que algunos receptores presentan estructuras casi idénticas en sus formas apo (sin ligando) y holo (complejo con ligando), otros requieren un ajuste estructural más profundo para permitir la unión del ligando. Los métodos de docking tradicionales están optimizados para modelos de receptor rígido, lo que limita su precisión en escenarios donde la flexibilidad del receptor es un factor clave. Para superar estas deficiencias, en esta tesis se evalúan estrategias computacionales que integran información conformacional dinámica con el objetivo de mejorar la predicción de interacciones y la selección de ligandos en estudios de cribado virtual.

Las metodologías propuestas se aplican al estudio de dos sistemas proteicos de relevancia biológica: el complejo transportador de lipoproteínas (LolCDE) y los citocromos P450 bacterianos (BacCYPs), ambos implicados en procesos clave de transporte y metabolismo celular. En el primer caso los desarrollos de docking se aplican en un esquema de búsqueda virtual de nuevos inhibidores, como punto de partida para el desarrollo de antibióticos. En el segundo caso, el docking es integrado con otros análisis bioinformáticos con el objetivo de predecir el sustrato de un BacCYP a partir de solamente el conocimiento de su secuencia proteica.

Los resultados obtenidos destacan la efectividad de estrategias avanzadas de docking molecular para mejorar la predicción de interacciones proteína-ligando. En LolCDE, la identificación de sitios de solvente y el uso de Bias Docking permitieron optimizar la selección de ligandos en virtual screening, mientras que en BacCYPs, la combinación de análisis filogenético, modelado estructural y docking molecular demostró ser una herramienta valiosa para la predicción de sustratos. Además, el estudio de la flexibilidad del receptor evidenció la importancia de considerar ajustes conformacionales en los métodos de docking. En conjunto, estas metodologías proporcionan un marco computacional robusto, aplicable al estudio de otras proteínas de interés y al diseño racional de nuevos fármacos, resaltando el valor de herramientas de código abierto como AutoDock y RDKit en la investigación biomolecular.

Palabras clave: Docking, Bias Docking, Bioinformática, Flexibilidad del receptor, Sitios de solvente,Predicción de sustratos, Virtual Screening, LolCDE, BacCYPs.

Development and Applications of Advanced Strategies in Molecular Docking: Impact of Receptor Flexibility, Substrate Search for BacCYPs, and Virtual Screening of Antimicrobials

Abstract

This doctoral thesis explores the use of computational methods, primarily molecular docking, to study protein-ligand interactions and facilitate virtual screening for new ligands with potential biological activity. Docking accuracy is influenced by multiple factors, including the correct representation of the binding site environment and receptor flexibility. To address these limitations, this study incorporates advanced strategies such as the inclusion of "solvent sites" and Bias Docking, along with the analysis of receptor conformational flexibility.

One of the main challenges in molecular docking is describing and predicting how a protein receptor accommodates its structure when interacting with a ligand. In many cases, these interactions induce conformational changes that can range from minor adjustments in the side chains of the active site to global structural modifications that significantly alter the protein's conformation. While some receptors exhibit nearly identical structures in their apo (ligand-free) and holo (ligand-bound) forms, others require a deeper structural adjustment to allow ligand binding. Traditional docking methods are optimized for rigid receptor models, which limits their accuracy in scenarios where receptor flexibility is a key factor. To overcome these shortcomings, this thesis evaluates computational strategies that integrate dynamic conformational information to improve interaction predictions and ligand selection in virtual screening studies.

The proposed methodologies are applied to the study of two biologically relevant protein systems: the lipoprotein transport complex LolCDE and bacterial cytochromes P450 (BacCYPs), both involved in key cellular transport and metabolic processes. In the first case, docking developments are applied in a virtual screening framework to identify new inhibitors, as a starting point for antibiotic development. In the second case, docking is integrated with other bioinformatics analyses to predict the substrate of a BacCYP solely from its protein sequence.

The results obtained highlight the effectiveness of advanced molecular docking strategies in improving the prediction of protein-ligand interactions. In LolCDE, the identification of solvent sites and the use of Bias Docking optimized ligand selection in virtual screening, while in BacCYPs, the combination of phylogenetic analysis, structural modeling, and molecular docking proved to be a valuable tool for substrate prediction. Additionally, the study of receptor flexibility demonstrated the importance of considering conformational adjustments in docking methods. Together, these methodologies provide a robust computational framework applicable to the study of other proteins of interest and the rational design of new drugs, emphasizing the value of open-source tools such as AutoDock and RDKit in biomolecular research.

***Keywords*:** Docking, Bias Docking, Bioinformatics, Receptor Flexibility, Solvent Sites, Substrate Prediction, Virtual Screening, LolCDE, BacCYPs

Introducción

1.1 Introducción General

Las interacciones proteína-ligando juegan un papel fundamental en la regulación de procesos biológicos, siendo clave en funciones como la señalización celular, la regulación enzimática y la unión de sustratos. En este trabajo definimos a un ligando como a una molécula que se une a una proteína, usualmente en una cavidad específica llamada sitio de unión, lo que puede desencadenar o inhibir una respuesta biológica. Estas interacciones son determinantes en la actividad de numerosas rutas metabólicas y procesos celulares, influyendo en mecanismos esenciales como la transmisión de señales, la catálisis enzimática y la modulación de receptores de membrana.

Comprender la naturaleza y dinámica de estas interacciones es crucial en diversas áreas científicas, incluyendo la bioquímica, la biología estructural y el diseño racional de fármacos. En este último campo, la identificación de compuestos capaces de modular la actividad de una proteína permite el desarrollo de nuevas estrategias terapéuticas dirigidas a enfermedades específicas. La caracterización de los sitios de unión y la predicción de interacciones proteína-ligando han sido impulsadas por avances en modelado molecular y simulaciones computacionales, que permiten explorar la flexibilidad estructural y la energía de unión, que determina la afinidad, de estos complejos.

En este contexto, el desarrollo de herramientas computacionales como el *docking* molecular y la dinámica molecular han revolucionado la forma en que se estudian estas interacciones, facilitando la identificación de ligandos con alta afinidad y selectividad. Sin embargo, la precisión de estos métodos depende de múltiples factores, incluyendo la flexibilidad del receptor, la presencia de solventes y la inclusión de restricciones informadas mediante enfoques como el *bias docking*. Así, la combinación de estas técnicas computacionales con datos experimentales ofrece un marco robusto para mejorar la predicción de interacciones y optimizar el diseño de fármacos

1.2 Los métodos computacionales en el desarrollo de fármacos.

El desarrollo de nuevos fármacos es un campo de gran relevancia que se sitúa en la intersección entre la ciencia básica y la investigación aplicada. El primer fármaco diseñado de manera racional basado en la estructura de su receptor fue el captopril, un inhibidor de la enzima convertidora de angiotensina, lanzado al mercado en la década de 1980 [2] [1]. En este ámbito, la química biológica y la bioinformática estructural desempeñan un papel fundamental, ya que proporcionan las bases necesarias para comprender las interacciones moleculares y los mecanismos de acción de los posibles compuestos terapéuticos (fármacos, o también llamados

drogas, al derivarse del término americano “drugs”). El modelado molecular ha emergido como una herramienta clave en el diseño de fármacos, permitiendo predecir y optimizar interacciones proteína-ligando de manera eficiente. Métodos como el docking molecular, tema central de la presente tesis, la dinámica molecular y el cribado (o búsqueda) virtual de ligandos, han revolucionado la identificación de compuestos con potencial terapéutico, reduciendo costos y tiempos en comparación con enfoques experimentales tradicionales.

La combinación de estas técnicas computacionales mejora la precisión en la selección de candidatos, facilitando el desarrollo de nuevos fármacos con mayor afinidad y especificidad hacia sus blancos moleculares[4]. Las metodologías computacionales, en particular los métodos de docking molecular, han emergido como herramientas indispensables en el proceso de diseño y descubrimiento de fármacos. Estos métodos facilitan tanto la identificación de compuestos candidatos como su optimización de manera más eficiente, rápida y económica en comparación con los enfoques experimentales tradicionales[3].

El docking molecular, al predecir la interacción, y el modo de unión, entre moléculas pequeñas (ligandos) y macromoléculas biológicas (receptores), permite explorar un vasto espacio químico y seleccionar aquellos compuestos con mayor potencial para convertirse en fármacos. Además, los resultados obtenidos mediante docking pueden guiar modificaciones estructurales específicas, mejorando la afinidad y selectividad de los compuestos de interés.

No obstante, a pesar de los avances alcanzados y del éxito relativo de estas técnicas, los métodos de docking aún presentan limitaciones que necesitan ser superadas para alcanzar un mayor nivel de precisión y aplicabilidad.Algunos de estos desafíos incluyen la mejora en la predicción de la flexibilidad conformacional de los receptores, la inclusión más precisa de efectos de solvatación, y una mayor comprensión y precisión del cálculo de de las interacciones ligandos y sus blancos. Por tanto, el perfeccionamiento continuo de estas técnicas es esencial para continuar optimizando el diseño racional de fármacos y acelerar el desarrollo de terapias más efectivas.

1.3 Modelado Molecular y Dinámica Molecular (MD)

El **modelado molecular** es una disciplina computacional que busca reproducir, comprender y predecir el comportamiento de sistemas moleculares a partir de los principios de la fisicoquímica. Debido a la complejidad y el tamaño de muchas biomoléculas, como las proteínas y ácidos nucleicos, el modelado molecular suele basarse en **métodos de mecánica molecular**, en lugar de cálculos mecano-cuánticos completos. Estos métodos emplean **campos de fuerzas**, que simplifican las interacciones atómicas mediante ecuaciones clásicas, prescindiendo del movimiento electrónico y evaluando la energía en función de las posiciones nucleares, de acuerdo con la conocida aproximación de Born-Oppenheimer. [5]
Entre las estrategias más utilizadas en modelado molecular se encuentra la **dinámica molecular (MD)**, una técnica que simula la evolución temporal de un sistema biomolecular a lo largo del tiempo. A través de MD es posible analizar el comportamiento dinámico de proteínas y otras macromoléculas, proporcionando información clave sobre su estructura, estabilidad y propiedades termodinámicas. Dado que la función de una proteína depende no solo de su

estructura tridimensional sino también de su **dinámica interna y sus posibles cambios conformacionales**, el uso de simulaciones de dinámica molecular es fundamental para entender sus mecanismos de acción.
En estos estudios, el solvente juega un papel crucial, ya que influye en la estabilidad conformacional de las biomoléculas, y en la formación de interacciones específicas. En entornos biológicos, el **agua** es el solvente predominante y determina la organización espacial de proteínas y ligandos, facilitando interacciones como enlaces de hidrógeno y efectos hidrofóbicos.

1.3.1 Cosolventes

Además del agua, se pueden emplear **cosolventes**, que son moléculas pequeñas agregadas al sistema con el objetivo de modificar las propiedades del medio y revelar posibles sitios de interacción en la proteína. Estos cosolventes pueden imitar grupos funcionales de ligandos y, de este modo, localizarse en regiones estratégicas de la superficie proteica, formando lo que se conoce como *hot spots* o sitios de alta afinidad. La identificación de estos sitios mediante MD en solventes mixtos es una estrategia valiosa para el diseño racional de fármacos, ya que permite mapear interacciones clave antes de realizar estudios de docking molecular o *virtual screening*. La integración de datos de solventes y cosolventes en estudios de dinámica molecular mejora la precisión en la identificación de sitios de unión relevantes, optimizando la predicción de interacciones proteína-ligando y facilitando el desarrollo de compuestos con alta afinidad y selectividad.[6]

1.3.2 Detección de Sitios

Una aplicación relevante de la dinámica molecular es la identificación de los denominados "*sitios de solvente"*, regiones en la superficie de la proteína donde ciertas moléculas de solvente tienden a localizarse con mayor frecuencia durante la simulación. Estos sitios proporcionan información sobre los puntos de interacción preferidos entre la proteína y potenciales ligandos. Por ejemplo, mediante simulaciones en mezclas de agua y solventes orgánicos, es posible mapear regiones hidrofóbicas, polares o iónicas dentro del sitio activo de la proteína. Estos datos pueden mejorar los modelos de *docking*, ya que permiten definir interacciones clave que deben priorizarse al diseñar o seleccionar ligandos con alta afinidad. En este contexto, y en función de los trabajos previos del grupo, en la presente tesis utilizaremos los "sitios de solvente" como una de las estrategias para implementar y mejorar la eficacia de los métodos de docking.

1.4 Docking Molecular y sus Aplicaciones

Los métodos de **docking molecular** permiten predecir la estructura de complejos proteína-ligando y estimar su afinidad a partir de la estructura de la proteína y el ligando por separado, de manera eficiente. En este enfoque, la proteína generalmente se mantiene en una

conformación fija (modelo rígido), mientras que el ligando es tratado como una molécula flexible, explorando múltiples orientaciones y conformaciones. De estas, se selecciona aquella que presenta la mayor afinidad de unión.

Para llevar a cabo un estudio de **docking**, es fundamental contar con la estructura tridimensional del receptor o, al menos, del sitio de unión a nivel atómico. Estas estructuras pueden obtenerse mediante técnicas experimentales como **difracción de rayos X** (si la resolución es inferior a 2,5 Å), **resonancia magnética nuclear (RMN)** o mediante **modelado por homología**, siempre que exista una alta identidad con una estructura conocida. Actualmente, incluso se están comenzando a utilizar estructuras generadas por inteligencia artificial (IA).

En la mayoría de los casos, el **docking proteína-ligando** no se realiza sobre toda la superficie de la proteína, sino que se restringe a una región específica de interés. Esta selección se basa en información previa obtenida a partir de complejos co-cristalizados, predicciones computacionales de ***"pockets"*** o sitios de unión potencialmente drogables, entre otras estrategias.

1.4.1 Métodos de Docking: Precisión, Sensibilidad y especificidad.

Como ya mencionamos, el docking es una técnica computacional fundamental para el diseño de fármacos y la investigación en biología estructural, la misma permite predecir cómo dos moléculas, típicamente una proteína (el receptor) y un ligando (usualmente una molécula orgánica pequeña), interactúan entre sí, formando el complejo correspondiente, y estimar la energía libre de unión (ΔGu) a partir de las estructuras individuales de las moléculas involucradas[7].

El proceso de docking busca determinar a partir de la estructura del receptor proteico, y el ligando, por separado, la orientación y posición óptimas de un ligando dentro del sitio de unión de su receptor, así como también estimar la afinidad de esta unión, que está determinada por la energía libre de unión entre ambas moléculas. Los softwares para docking, tales como Autodock, (Autodock)Vina, GLIDE, GOLD o DOCK, están diseñados para ofrecer resultados ante una amplia variedad de compuestos con diferentes propiedades fisicoquímicas. Sin embargo, este enfoque generalista implica que el rendimiento de estas herramientas no siempre sea óptimo para sistemas específicos.

Los métodos de docking molecular constan de dos componentes esenciales. El primer componente lo comprenden la estrategia y el algoritmo de ***búsqueda conformacional***. Entre los métodos más utilizados se encuentran los algoritmos estocásticos, como el **Simulated Annealing** [12] [11] (basado en Monte Carlo), y los **algoritmos genéticos** [10] [9] [8], los cuales operan introduciendo cambios aleatorios en los grados de libertad conformacionales del ligando, como traslaciones, rotaciones y torsiones de enlaces. Cada pose generada es evaluada mediante una función de puntaje predefinida (el 2do componente), que determina si la conformación se mantiene como una opción viable o se descarta. Para simplificar la búsqueda,

las distancias y ángulos de enlace generalmente se mantienen fijos, incluso en anillos alifáticos. Este enfoque permite optimizar el proceso de docking, reduciendo la complejidad computacional y facilitando la identificación de poses biológicamente relevantes en un tiempo acotado.

El segundo es el método de estimación de la afinidad de unión (denotado como ΔG_u), conocido también como función de puntaje. Esta función permite evaluar la estabilidad del complejo proteína-ligando predicho, y clasificar las poses generadas de acuerdo con su energía libre de unión. Estas funciones se construyen considerando diversos factores que contribuyen a la interacción molecular, tales como **las interacciones electrostáticas**,**la energía de enlace rotacional**, **las interacciones de van der Waals**, entre otras.

Es importante mencionar, que para evaluar la precisión y eficacia de los métodos de docking, se emplean métricas estadísticas típicas como sensibilidad, especificidad, precisión y otras medidas de rendimiento. Cuando el objetivo del docking es identificar el modo de unión correcto (predicción de pose), el "verdadero positivo" (VP) se define como una pose correcta cuya ΔG_u sea más favorable que cualquier otra pose generada. Por el contrario, un "falso positivo" (FP) corresponde a poses incorrectas que, a pesar de no alojar al ligando en el sitio activo de manera adecuada, son predichas con valores de ΔG_u altamente favorables. Los "verdaderos negativos" (VN) son poses incorrectas con valores de ΔG_u poco favorables, mientras que los "falsos negativos" (FN) representan poses correctas que no logran una buena puntuación y, por ende, son clasificadas erróneamente con valores de ΔG_u poco favorables. Como en cualquier método de clasificación el objetivo es maximizar la precisión (calculada como VP/(VP+FP)) y la sensibilidad (VP/(VP+FN)) , aunque en muchos métodos de docking es un reto lograr una precisión alta, debido a la existencia de numerosos FP, es decir, poses incorrectas con altos puntajes.[13]

1.4.2 Bias Docking

El *bias docking (o sea docking sesgado)* es una variante del *docking* molecular que incorpora información adicional para guiar la predicción de interacciones proteína-ligando. A diferencia del *docking* tradicional, donde la búsqueda de poses se realiza de manera general, el *bias docking* introduce restricciones o preferencias basadas en datos experimentales, conocimiento previo del sistema, o hipótesis estructurales que sesgan el espacio de búsqueda conformacional del ligando. En la práctica, éstos sesgos favorecen configuraciones específicas del ligando dentro del sitio de unión, enfocándose, en nuestra implementación, en interacciones biológicamente relevantes.

AutoDock Bias es una herramienta basada en AutoDock4 y AutoDockTools que implementa este enfoque mediante la modificación de los mapas de energía de los tipos de átomos y el archivo de parámetros de *docking* (DPF), para incluir sesgos energéticos sobre átomos específicos del ligando. La técnica se basa en la introducción de pozos de energía en los mapas de AutoDock4, representados como términos gaussianos invertidos que favorecen la ubicación de ciertos átomos del ligando en posiciones definidas dentro de la cuadrícula de cálculo. Esta modificación permite priorizar interacciones específicas, como enlaces de

hidrógeno, contactos aromáticos, u otros criterios definidos por el usuario. El *bias docking* ha demostrado que puede mejorar la precisión en la selección de la conformación correcta del complejo, optimizando la identificación de la estructura *holo* más relevante dentro de un conjunto de conformaciones alternativas. Este desarrollo, fue realizado por el grupo de trabajo de la presente tesis, y por lo tanto será una de las estrategias centrales de la misma. [14]

1.4.3 Virtual Screening

El Virtual Screening (VS) es una estrategia computacional ampliamente utilizada en el descubrimiento de fármacos basada en el docking, y que busca identificar compuestos con alta afinidad de unión a un blanco biológico deseado. A diferencia de la *pose prediction* en *docking*, donde se busca determinar cómo un ligando específico interactúa con su proteína objetivo, el VS no requiere conocimiento previo de un ligando que pueda unirse a la proteína. En su lugar, el objetivo es identificar nuevos ligandos a partir de una gran biblioteca química virtual de posibles ligandos, seleccionando aquellos que tienen mayor probabilidad de interacción con la proteína de interés.

Este enfoque ha revolucionado el diseño de fármacos al permitir la evaluación rápida de millones de compuestos en busca de candidatos potenciales, reduciendo el tiempo y los costos asociados a los métodos experimentales tradicionales. El VS se emplea tanto para identificar nuevos *hit compounds,* como para optimizar candidatos existentes, buscando análogos con mayor potencia, selectividad o mejores propiedades farmacocinéticas [16] [15].

El VS ha demostrado ser una herramienta clave en la identificación de nuevos fármacos, con numerosos casos de éxito en la literatura. Por ejemplo, la identificación del inhibidor de la tirosina quinasa *Imatinib* (*Gleevec*) y la optimización de antivirales como *Oseltamivir* (*Tamiflu*) han sido facilitadas por enfoques computacionales de tipo VS. A medida que las técnicas de VS evolucionan, se integran cada vez más con inteligencia artificial y las simulaciones de dinámica molecular, mejorando la precisión en la predicción de interacciones proteína-ligando y reduciendo la tasa de falsos positivos en la selección de compuestos. VS [17].

1.4.5 Evaluación de la Búsqueda Virtual y sus Desafíos

Para evaluar la efectividad de los métodos de docking en un esquema de búsqueda virtual (BV), también se emplean métricas estadśiticas mencionadas par ala evaluación de predicción de pose, pero en este caso se realiza una clasificación diferente de los verdaderos positivos (VP), falsos positivos (FP), verdaderos negativos (VN) y falsos negativos (FN). Los VP son aquellos compuestos que el *docking* predice como ligandos activos y que posteriormente muestran afinidad y actividad experimentalmente. Los FP son compuestos predichos como activos, pero que no presentan actividad en ensayos experimentales. Los VN son correctamente identificados como no ligandos, mientras que los FN representan compuestos que sí interactúan con la proteína, pero no fueron identificados por el método computacional.

Los métodos de *docking*, debido a limitaciones inherentes como la predicción inexacta del ΔG_u, y el tratamiento insuficiente de la flexibilidad del receptor, suelen mostrar un rendimiento limitado, con precisión y sensibilidad moderadas o bajas. Esta falta de precisión y sensibilidad afecta tanto la predicción de poses como la identificación de nuevos ligandos, siendo uno de los principales desafíos del VS. La alta incidencia de FP es especialmente problemática, ya que genera predicciones incorrectas de afinidad o actividad del ligando en el sitio de unión[18].

En este contexto, ***el desarrollo de estrategias que mejoren la precisión de los cálculos de ΔG_u sin incurrir en costos computacionales elevados resulta fundamental. La presente tesis aborda esta problemática, explorando enfoques que optimicen la selección de ligandos y mejoren la tasa de éxito de estudios experimentales mediante métodos computacionales avanzados.***

1.4.5 El problema del cálculo de la energía libre de unión (ΔGu).

La determinación precisa del ΔG_u en interacciones proteína-ligando o proteína-proteína, es un aspecto crucial en bioinformática estructural, ya que este valor es el indicador principal de la afinidad entre el ligando y su receptor. Así, el ΔG_u se convierte en un parámetro central en gran parte de los métodos computacionales utilizados para predecir interacciones moleculares. Existen diversas estrategias para calcular ΔG_u, las cuales pueden clasificarse en función de su rapidez y precisión. Los métodos de docking, por ejemplo, son extremadamente rápidos, pero adolecen de baja precisión debido a las aproximaciones simplificadas que utilizan para el cálculo de afinidad[22]. En contraste, existen métodos de tiempo de cálculo moderado, que emplean el post-procesamiento de trayectorias de Dinámica Molecular de los complejos para obtener una estimación más precisa de ΔG_u ,a estos se los conoce como métodos de punto final [21]. Finalmente, hay una serie de métodos aún más precisos que requieren un muestreo extenso, como la Dinámica Molecular Replica-Exchange[20] o la Metadinámica[19], que permiten calcular la energía libre de unión absoluta o relativa entre diferentes fármacos y receptores. Estas técnicas, aunque altamente precisas, son de alto costo computacional y por tanto, suelen reservarse para etapas avanzadas en un proyecto de búsqueda virtual y/o para evaluar casos específicos debido a su complejidad de implementación.

1.5 La flexibilidad del Receptor.

Uno de los mayores desafíos para los métodos de docking es la capacidad de describir y predecir cómo el receptor proteico ajusta su estructura al unirse al ligando. Cuando las proteínas interactúan con un ligando (o sustrato), experimentan un ajuste conformacional (ajuste inducido) que puede variar desde casi nulo—cuando las estructuras de la proteína sin ligando (apo) son prácticamente indistinguibles de las del complejo (holo)—hasta cambios significativos, considerados como verdaderos cambios conformacionales. Entre estos extremos, se pueden observar movimientos en las posiciones de las cadenas laterales, de los residuos del sitio activo, así como variaciones en el tamaño y la forma de dicho sitio [24] [23].

En una primera aproximación, los métodos de docking tradicionales suelen tratar al receptor como una estructura rígida, lo que resulta óptimo para casos en los que el receptor mantiene una conformación estable, pero tiende a ser ineficaz en situaciones donde se requiere una mayor flexibilidad del receptor. En estos casos, los resultados de docking pueden ser poco precisos, especialmente cuando el ajuste conformacional inducido es significativo y se requiere una representación más dinámica de la estructura del receptor.

En la presente tesis, nos proponemos **determinar la capacidad del Bias-docking para seleccionar la estructura correcta (holo)** a partir de un conjunto de conformaciones del receptor que reflejan distintos estados del sitio activo. Para ello, desarrollaremos una metodología sistemática para caracterizar la variabilidad conformacional del sitio activo, particularmente para la transición entre estructuras **Apo y Holo**. A través de esta caracterización, buscamos establecer una relación entre la variabilidad estructural del sitio activo y la precisión del Bias-docking en la identificación de la conformación funcionalmente relevante. La implementación de esta metodología permitirá mejorar la interpretación de los resultados de docking, y optimizar la selección de estructuras para estudios de interacción proteína-ligando.

1.6 Similitud Química

A pesar de que pequeñas variaciones estructurales pueden resultar en grandes diferencias en la actividad biológica, compuestos y/o proteínas similares tienden a comportarse de manera similar, desde una perspectiva de la unión proteína-ligando. Los métodos actuales, como las huellas dactilares moleculares (finger prints) y los índices de similitud química (como el de Tanimoto), ofrecen enfoques para medir esta similitud entre los ligandos. Sin embargo, estos métodos presentan desafíos significativos, incluyendo una adecuada y detallada representación molecular que derive en una manera óptima para predecir la correlación entre similitud estructural y actividad biológica. Es conocido que a menudo, compuestos que son altamente similares pueden mostrar diferentes actividades biológicas, mientras que compuestos estructuralmente diversos pueden presentar actividades similares. Por otro lado, la vastedad del espacio químico, con un número inmenso de moléculas posibles, complica la identificación de relaciones de similitud significativas dentro de bases de datos amplias.

Para abordar estos retos, se han propuesto diferentes enfoques integrales, como pipelines que integran datos de actividad biológica y métricas de similitud química, que permiten optimizar la selección de compuestos en las etapas iniciales del descubrimiento de fármacos [25]. En este contexto, podemos mencionar el índice de Tanimoto y la herramienta LigQ, desarrollada en el grupo de investigación donde se realiza esta tesis, que han sido utilizadas para mejorar la clasificación y selección de compuestos con alta probabilidad de interacción con la proteína objetivo y el armado de las bases de datos. En esta tesis, también utilizaremos esta herramientas de comparación de ligandos, para ordenarlos, clasificarlos y filtrarlos en las etapas previas a la realización del docking.

1.7 Herramientas bioinformáticas - Interacciones

Como ya mencionamos antes, una de las claves en el proceso de formación del complejo proteína-ligando es la formación de interacciones moleculares específicas entre ambos. Durante el desarrollo de esta tesis, se implementaron y programaron varios *scripts* con el objetivo de automatizar y optimizar distintos aspectos del análisis computacional de las interacciones proteína-ligando. En particular, se desarrolló un *script* específico para evaluar interacciones, que permite analizar y cuantificar los contactos entre los ligandos y los residuos del sitio activo de las proteínas estudiadas. Este script evalúa interacciones como puentes de hidrógeno y aromáticas entre otras funciones, que describiremos brevemente a continuación.

1.7.1 Puente de Hidrógeno

Un **puente de hidrógeno** es un tipo de interacción no covalente que ocurre cuando un átomo de hidrógeno, que está unido covalentemente a un átomo electronegativo (como oxígeno, nitrógeno o flúor), interactúa con otro átomo electronegativo de una molécula cercana. Los puentes de hidrógeno presentan una variabilidad que depende de varios factores, como los átomos involucrados, la geometría de la interacción y el entorno químico en el que se forman. En particular, cuando los átomos de nitrógeno (N), hidrógeno (H) y oxígeno (O) participan en un puente de hidrógeno, tienden a adoptar una orientación casi lineal, lo que maximiza la energía de la interacción. La distancia entre el nitrógeno y el oxígeno (N···O) suele encontrarse en un rango de 2.8 a 3.2 Å, lo que es típico para enlaces de hidrógeno fuertes. Además, el ángulo N–H···O juega un papel crucial en la estabilidad y la fuerza de la interacción. Este ángulo es generalmente superior a los 150°, lo que refuerza la naturaleza lineal del puente de hidrógeno y contribuye a una mayor estabilidad y energía de interacción. Las interacciones más lineales tienden a ser más fuertes.

1.7.2 Aromática (Π - Π)

Las interacciones **pi-pi** son interacciones no covalentes [26], que ocurren entre anillos aromáticos, debido a la atracción entre las nubes de electrones deslocalizados (orbital pi) de los sistemas aromáticos. Estas interacciones desempeñan un papel fundamental en la estabilización de complejos moleculares en sistemas biológicos, como en las estructuras de proteínas, la doble hélice de ADN y en el reconocimiento molecular entre proteínas y ligandos.

Desde una perspectiva estructural, las interacciones pi-pi pueden clasificarse principalmente en tres configuraciones:

1. **Stacking Paralelo (Face-to-Face)**: En este tipo de interacción, los anillos aromáticos están orientados de forma casi paralela, con una separación centroide-centroide de entre ***3.3 y 4.4 Å***.

2. **Stacking Desplazado (Offset Stacked)**: En esta variación del stacking paralelo, los anillos aromáticos permanecen alineados, pero con un desplazamiento lateral.
3. **Interacción en T (T-shaped)**: En este tipo de interacción, un anillo se orienta de manera perpendicular al otro, formando un ángulo cercano a 90°.

1.8 Bases de datos

El desarrollo de esta tesis se apoya en el uso de bases de datos bioinformáticas que contienen información estructural y funcional de proteínas y ligandos. Estas bases de datos permiten acceder a datos experimentales de estructuras proteicas, y a colecciones de compuestos químicos que pueden ser evaluados mediante métodos computacionales.

La bioinformática ha emergido como una disciplina fundamental en la era genómica, permitiendo el análisis, almacenamiento y gestión de grandes volúmenes de datos biológicos. Las bases de datos bioinformáticas juegan un papel crucial en este contexto, ya que sirven como repositorios estructurados que almacenan información biológica diversa, como secuencias de ADN, proteínas, estructuras moleculares, interacciones genéticas y datos de expresión génica entre otras. Estas herramientas, no solo facilitan el acceso a la información, sino que también permiten la integración y el análisis de datos para generar conocimiento biológico relevante.

1.9 El Protein Data Bank (PDB)

Uno de los recursos clave utilizados en este trabajo es el **Protein Data Bank (PDB)**, una de las bases de datos más importantes en el campo de la bioinformática y la biología estructural. Fundada en 1971, el PDB alberga información tridimensional detallada sobre estructuras de macromoléculas biológicas, como proteínas, ácidos nucleicos (ADN y ARN) y otros complejos moleculares.

Las estructuras registradas en el PDB han sido determinadas experimentalmente, principalmente mediante técnicas de difracción de rayos X, seguidas por modelos generados a través de resonancia magnética nuclear (RMN). En años recientes, la microscopía electrónica también ha ganado relevancia como método alternativo para la obtención de estructuras. Actualmente, el proceso para determinar la estructura tridimensional de una proteína mediante cristalografía de rayos X requiere un tiempo significativamente menor, reduciéndose en al menos tres órdenes de magnitud en comparación con el pasado. Además, el volumen de datos en las bases de datos de estructuras ha experimentado un crecimiento exponencial, conteniendo hoy en día 300 veces más entradas que hace 25 años[27].

Un avance reciente y transformador en el campo ha sido el desarrollo de AlphaFold, un sistema de inteligencia artificial creado por DeepMind, capaz de predecir estructuras proteicas a partir de la secuencia, con una precisión comparable a los métodos experimentales. AlphaFold ha permitido la predicción de estructuras para millones de proteínas, incluyendo muchas que no habían sido resueltas experimentalmente, lo que ha ampliado significativamente nuestro conocimiento sobre el plegamiento y la función de las proteínas [28].

Este hito no sólo complementa los datos experimentales, sino que también acelera la investigación en biología estructural y abre nuevas vías para el estudio de enfermedades y el diseño de fármacos.

1.10 Estructuras de proteínas utilizadas en estas tesis

En el desarrollo de esta tesis, se trabajó con diversas estructuras proteicas, incluyendo las de los **citocromos P-450 bacterianos (BacCYPs),** y el complejo **LolCDE**, con el fin de evaluar la aplicabilidad de las metodologías desarrolladas en diferentes sistemas biológicos de relevancia.

1.10.1 Citocromos P450

Los citocromos P450 (también conocidos como P450s o BACYPs) son enzimas naturales de gran eficiencia, capaces de llevar a cabo reacciones químicas que aún representan un desafío para la química moderna. Su descubrimiento se remonta a la década de 1950, cuando se investigaban pigmentos en células hepáticas, denominados citocromos (del griego *kytos*, célula, y *chroma*, color). Durante esos estudios, el científico alemán Martin Klingenberg identificó un pigmento que, al unirse al monóxido de carbono, presentaba una absorbancia máxima a 450 nm, lo que sugería que se trataba de una hemoproteína distinta a las conocidas hasta entonces [30]. Más adelante, en 1964, Omura y Sato confirmaron su naturaleza hemoproteica en microsomas hepáticos de mamíferos [29]. Al ser reducido por NADPH, este pigmento se unía al CO, generando un pico de absorbancia característico a 450 nm, lo que llevó a llamarlo citocromo P-450 (donde "P" hace referencia a "pigmento" y "450" a su absorbancia en el espectro UV).

Puntualmente en las bacterias, los CYPs cumplen roles esenciales en el metabolismo y la adaptación a diversos ambientes. Estas enzimas están involucradas en la biosíntesis de metabolitos secundarios, como antibióticos y toxinas, que son cruciales para la supervivencia y competencia bacteriana [32]. A diferencia de los CYPs eucariotas, en las bacterias los citocromos P450 no están asociados a membranas y se solubilizan con facilidad [31]. Estas proteínas son extremadamente ubicuas, y la cantidad de secuencias de CYPs conocidas actualmente es asombrosa. Una simple búsqueda en la base de datos RefSeq revela que actualmente se han identificado más de 28,000 CYPs diferentes, lo que refleja su diversidad y amplia distribución en la naturaleza.

1.11.2 Estructura P-450

Las P450s presentan un plegamiento característico que es exclusivo de esta familia de enzimas. Su estructura secundaria está compuesta principalmente por hélices α, y alberga en su sitio activo un cofactor hemo, el cual se encuentra encajado entre dos dominios: uno mayormente α-helicoidal (denominado dominio alfa, con hélices designadas de A a L) y otro más pequeño, rico en hojas β (conocido como dominio beta, con láminas designadas de 1 a 4). La región central que rodea al cofactor hemo proporciona la estructura necesaria para la

activación del oxígeno por parte de las P450s. En su estado de reposo, el hierro del grupo hemo se encuentra coordinado axialmente por una cisteína altamente conservada como ligando proximal y, típicamente, por una molécula de agua como ligando distal.

Para llevar a cabo su función catalítica, las CYPs deben fijar sus sustratos dentro del bolsillo del sitio activo, ubicado en la parte superior del sitio distal del hemo, donde también se une el oxígeno. La comparación de estructuras cristalográficas de diferentes CYPs revela un plegamiento altamente conservado, pero con una variabilidad significativa en la forma y tamaño de los sitios o bolsillos de unión a sustratos. Estas diferencias son clave para determinar la especificidad de cada CYP tanto en la selección del sustrato como en el tipo de reacción que realiza.

El sitio catalítico en torno al grupo hemo contiene regiones específicas denominadas sitios de reconocimiento de sustratos (SRS), los cuales participan en la unión y catálisis de los sustratos. La secuencia de aminoácidos de estos SRSs varía considerablemente entre diferentes CYPs, lo que explica su capacidad para procesar una amplia gama de compuestos [33].

Los citocromos P450 (CYPs) tienen aplicaciones amplias y diversas, que incluyen la producción de fármacos, metabolitos y el diseño de biosensores. Debido a su versatilidad, estas enzimas son cruciales en áreas como la síntesis de compuestos orgánicos y la eliminación de contaminantes ambientales, convirtiéndolas en una familia de proteínas de gran interés para la investigación científica y aplicaciones tecnológicas [34]. Comprender en profundidad sus propiedades estructurales y funcionales es esencial para explorar su potencial y aplicarlo en diversos campos científicos e industriales.

1.11.3 Complejo LolCDE

El complejo LolCDE (***Figura 1***) es un complejo proteico de la familia de transportadores ABC responsable de la extracción de lipoproteínas de la membrana interna en bacterias Gram-negativas, y su transferencia a la proteína LolA, una chaperona periplásmica. Está compuesto por las proteínas transmembrana LolC y LolE, que forman un heterodímero, y por la proteína citoplasmática LolD, que actúa como homodímero y contiene el dominio de unión a nucleótidos responsable de la hidrólisis de ATP, que provee la energía libre necesaria para el proceso de extracción lipídica. En *E. coli*, LolC y LolE desempeñan funciones diferenciadas: mientras que LolC interactúa con la chaperona LolA, LolE se une directamente a las lipoproteínas. Esta especificidad funcional, ausente en humanos, convierte a LolCDE en un blancoatractivo desde el punto de vista del diseño de inhibidores que interfieran con el transporte lipoproteico bacteriano. [35].

La elección de LolCDE como sistema de estudio en esta tesis responde a su relevancia biológica, y a su potencial como blanco terapéutico en el contexto del desarrollo de nuevas estrategias antimicrobianas. Esta proteína fue seleccionada como uno de los dos blancos prioritarios dentro de un proyecto colaborativo entre el la organización internacional GARDP (Global Antibiotic Research and Development Partnership) [36] y el el grupo donde se realizó

la presente tesis, cuyo objetivo es fomentar la investigación y desarrollo de nuevos tratamientos frente a bacterias resistentes, especialmente en contextos de salud pública desatendidos.

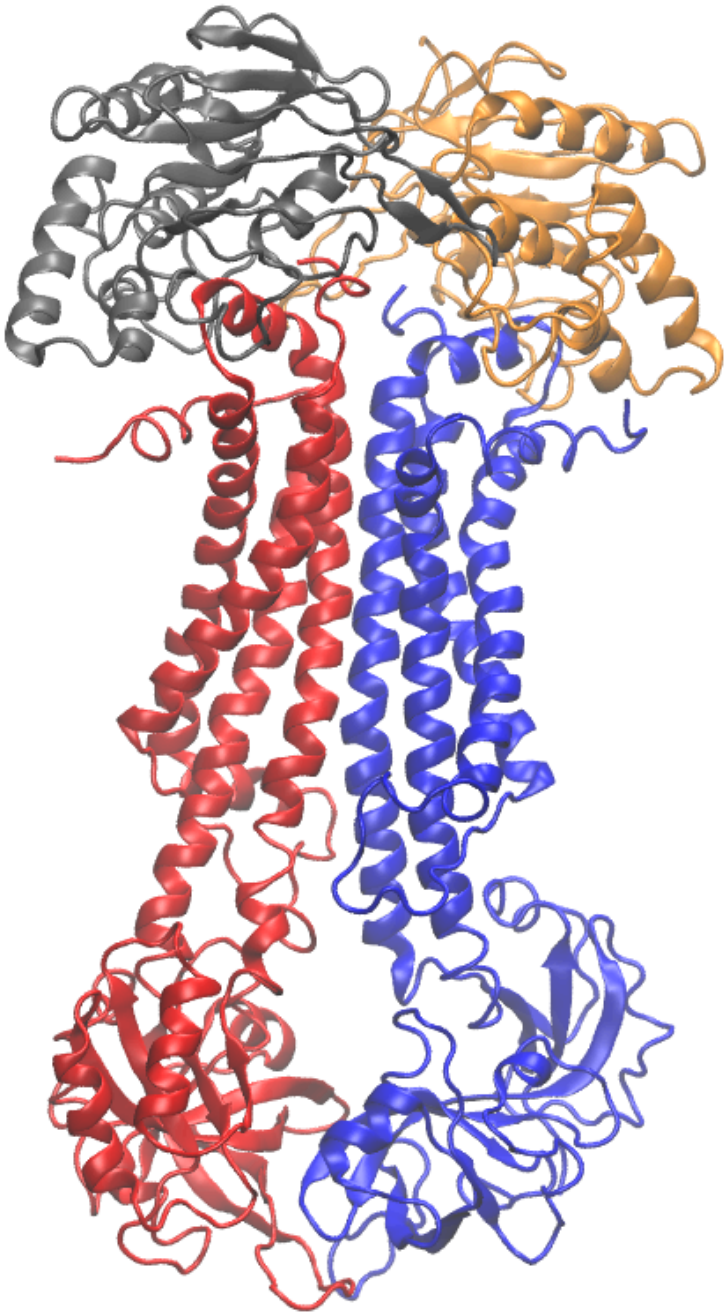


Figura 1 Estructura tridimensional del complejo transportador **LolCDE de *Escherichia coli***, obtenida a partir del modelo cristalográfico depositado en el PDB (ID: 7ARM) Se observan tres subunidades del complejo: **LolC** en rojo, **LolE** en azul y **LolD** en gris y naranja. Este complejo ABC (ATP-binding cassette) participa en el transporte de lipoproteínas desde la membrana interna hacia el periplasma en bacterias Gram-negativas.

En ese marco, se propuso un proyecto de colaboración que busca la aplicación de métodos de screening virtual, docking y simulaciones moleculares para identificar nuevas moléculas capaces de unirse a blancos con potencial antimicrobiano no explorados previamente. En el caso particular de LolCDE, se busca contribuir al descubrimiento de inhibidores que afecten su función esencial en el transporte de lipoproteínas, lo cual podría representar una vía innovadora para el desarrollo de antibióticos efectivos frente a patógenos multirresistentes, como *Klebsiella pneumoniae (Kp)*.

En resumen, el conjunto de avances recientes en modelado molecular, docking sesgado y simulaciones de dinámica molecular, combinados con herramientas bioinformáticas y modelos estructurales derivados de inteligencia artificial, abre nuevas posibilidades para abordar los desafíos en estudio y la predicción de las interacciones proteína-ligando. Esta tesis se inscribe en este marco conceptual, proponiendo el desarrollo y aplicación de metodologías computacionales integradas para mejorar la precisión en la selección de ligandos, considerar de forma más realista la flexibilidad del receptor, y explorar nuevos blancos terapéuticos. A continuación, se presentan los objetivos específicos que orientan este trabajo.

1.12 Objetivos de la Tesis

El objetivo principal de esta tesis es desarrollar, optimizar y aplicar estrategias computacionales avanzadas que mejoren la predicción de las interacciones proteína-ligando en estudios de docking molecular, abordando especialmente los desafíos asociados a la flexibilidad del receptor, la identificación de sustratos en enzimas bacterianas, y la búsqueda virtual de inhibidores para un blanco con potencial antimicrobiano.

El trabajo integra herramientas de docking convencional y sesgado, análisis estructurales basados en dinámica molecular, modelado por inteligencia artificial (AlphaFold) y técnicas de bioinformática y quimioinformática, organizadas en tres líneas principales:

1) Optimizar la performance del Virtual Screening mediante la incorporación de información derivada de sitios de solvente y la aplicación sistemática de Bias Docking. Como modelo biológico, se estudió, en este caso, se utilizó el complejo transportador LolCDE, evaluando la capacidad de estas estrategias para mejorar la selección de ligandos candidatos como potenciales inhibidores antimicrobianos.

2) Desarrollar un pipeline de predicción funcional de Citocromos P450 bacterianos (BacCYPs) a partir de su secuencia proteica, combinando modelado estructural mediante AlphaFold, análisis filogenético, estudio de diversidad de sustratos y docking sesgado. Este enfoque busca inferir preferencias de sustrato y funciones catalíticas de BacCYPs aún no caracterizados experimentalmente, contribuyendo al aprovechamiento biotecnológico de esta familia en potenciales aplicaciones industriales y farmacológicas.

3) Analizar de manera sistemática el impacto de la flexibilidad conformacional del sitio activo en la predicción de poses de docking, caracterizando las diferencias entre estructuras Apo y Holo, en el contexto de la teoría del ajuste inducido. Se estudiaron cambios en ángulos χ de residuos clave, variaciones en volumen y geometría del pocket mediante técnicas de Convex Hull, y se propuso una estrategia para introducir flexibilidad dirigida que mejore la predicción de unión partiendo de estructuras Apo. Este eje constituye además la base conceptual para futuros desarrollos que integren modelos multiestado derivados de AlphaFold y técnicas de aprendizaje automático para optimizar protocolos de docking flexibles.

Metodología

En este capítulo se ofrecerá una visión general de los métodos empleados a lo largo de la presente tesis, los cuales se dividen en dos grandes categorías:

1. Métodos computacionales
2. Métodos bioinformáticos.

Estos enfoques complementarios constituyen las herramientas fundamentales para abordar las diversas problemáticas planteadas en el trabajo de tesis. A lo largo de la misma, cada uno de estos métodos será aplicado de manera específica para resolver cuestiones particulares, adaptándose a las necesidades de cada capítulo. Los detalles técnicos y la implementación de estos métodos en cada contexto y caso particular serán descritos de forma exhaustiva en los capítulos correspondientes.

2.1 Métodos Computacionales

2.1.1 Modelado Estructural

Se utilizó AlphaFold v2.3.2 (DeepMind) instalado en un entorno Ubuntu 20.04 con soporte GPU (NVIDIA CUDA 11.4). Los modelos se generaron utilizando la base de datos de secuencias UniProt y PDB actualizada al año 2024. Se emplearon cinco ejecuciones por proteína, seleccionando el modelo con mayor score de predicción (pLDDT), y verificando su calidad con la herramienta MolProbity. La alineación de secuencias se realizó con HHblits y MMseqs2, configurados para tres iteraciones y un umbral de e-value de 0.001. Las estructuras finales fueron comparadas con modelos experimentales cuando estaban disponibles y seleccionadas según su estabilidad estructural, evaluada mediante simulaciones cortas de dinámica molecular. Se justificó el uso de AlphaFold por su alta precisión en modelado de proteínas difíciles, como los BacCYPs, esenciales para la presente tesis.

2.1.2 Docking Proteína - Ligando

El software más utilizado en esta tesis es el software de Docking **AutoDock-GPU**[37], una herramienta de código abierto ampliamente adoptada por la comunidad científica,por lo cual muchos aspectos metodológicos están basados en dicho programa.

En el docking, normalmente se considera la proteína como un cuerpo rígido, limitando la flexibilidad al ligando. Para un ligando flexible en una proteína fija, el docking involucra seis grados de libertad: tres traslacionales y tres rotacionales, más los grados de libertad internos del ligando, que se limitan a las torsiones alrededor de enlaces simples (torsiones o ángulos diedros), manteniendo fijos los enlaces covalentes y los ángulos, para explorar estos espacios conformacionales se utiliza un algoritmo de búsqueda. Estas variables son utilizadas por dicho algoritmo para encontrar la pose óptima (pose prediction).

2.1.2.1 Algoritmo de búsqueda

En **AutoDock-GPU**, se utiliza un algoritmo genético, optimizado para aprovechar la capacidad de las GPU, lo que permite realizar cálculos más rápidos y eficientes en comparación con AutoDock4 (la versión previa), facilitando la búsqueda conformacional del ligando sin comprometer la precisión

El **algoritmo genético** se emplea para explorar el **espacio conformacional del ligando**, considerando sus posibles posiciones y orientaciones dentro del sitio de unión de la proteína. Cada conformación del ligando se trata como un **individuo en una población**, y su calidad se evalúa mediante una **función de aptitud**, que refleja la afinidad de unión con el receptor (es decir, estima el ΔG_u mediante la función de puntaje que describo más adelante).

El proceso evolutivo del algoritmo consta de tres etapas clave:

1. **Selección**: Se eligen las conformaciones con mejor afinidad de unión para la siguiente generación.
2. **Reproducción (Cruce genético)**: Se combinan características de los ligandos seleccionados para generar nuevas conformaciones.
3. **Mutación**: Se introducen cambios aleatorios en los **ángulos de torsión** de los enlaces flexibles del ligando, permitiendo la exploración de nuevas configuraciones espaciales.

Este ciclo se repite hasta que se alcanza un criterio de convergencia. En nuestro caso un número máximo de iteraciones. De este modo, el algoritmo genético permite identificar la conformación más estable y favorable del ligando dentro del sitio activo.

2.1.2.2 Función de puntaje

El objetivo de la función de puntaje (scoring function) es cuantificar qué tan favorable es una conformación específica del complejo proteína-ligando, para ello usa una función matemática que evalúa la afinidad de unión (***energía libre de unión***) entre un ligando y una proteína. Esta función es clave para predecir qué tan bien se acomoda un ligando en el sitio de unión del receptor y qué tan fuerte es esta interacción.

La función de puntaje del Autodock está formada por un campo de fuerzas que utiliza un enfoque mixto, conocido como semiempírico, que integra un potencial basado en principios fisicoquímicos de la mecánica molecular, junto con parámetros de solvatación y factores de ponderación para diferentes componentes de la función, los cuales fueron calibrados con datos experimentales para calcular la energía libre de unión.

La **función de puntaje semiempírica** que se observa en la ***Ecuación 1*** y considera varios factores importantes:

- **Interacciones de van der Waals**: Evalúa las interacciones atractivas y repulsivas entre átomos no enlazados a distancias cortas. Estas interacciones son importantes para

describir cómo los átomos del ligando y el receptor se atraen o se repelen a medida que se acercan.

- **Interacciones enlace de hidrógeno**: Evalúa los enlaces puente de hidrógeno que se forman entre el ligando y el receptor , la dirección viene dada el ángulo (θ) y la distancia.
- **Interacciones electrostáticas**: AutoDock4 considera las cargas parciales de los átomos y cómo afectan a las interacciones entre los átomos cargados del ligando y del receptor. Evalúa las interacciones electrostáticas mediante un potencial de Coulomb apantallado.
- **Término de solvatación**: Calcula los efectos de solvatación o la interacción del ligando y el receptor con el solvente (generalmente agua). Este término refleja cómo el desplazamiento del agua en el sitio de unión afecta la afinidad del ligando,basado en el volumen atómico V de los átomos alrededor del átomo bajo análisis, un parámetro atómico de solvatación S y una función Gaussiana dependiente de la distancia entre átomo

$$V = W_{\mathrm{vdW}} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \quad \} \text{ Lennard-Jones}$$

$$+ W_{\mathrm{EH}} \sum_{i,j} E(\theta) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + \quad \} \text{ Enlace de Hidrógeno}$$

$$+ W_{\mathrm{elec}} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} + \quad \} \text{ Coulomb}$$

$$+ W_{\mathrm{sv}} \sum_{i,j} (S_i V_j + S_j V_i) e^{-r_{ij}^2 / 2\sigma^2} \quad \} \text{ Desolvatación}$$

Ecuación 1: *Función de puntaje utilizada en AutoDock-GPU la misma representa la energía libre de unión entre el ligando y la proteína, calculada mediante la combinación de diferentes términos energético*

Para estimar la energía libre de unión, la función de puntaje incluye un término adicional ***Ecuación 2,*** que representa la pérdida de entropía del ligando al restringir sus grados de libertad conformacionales durante el proceso de unión (∆**Sconf**). Este término penaliza la flexibilidad del ligando al calcular la energía necesaria para modificar los ángulos de torsión de sus enlaces rotacionales. Cuanto más flexible es un ligando (es decir, con más enlaces rotables), mayor es la energía requerida para adoptar una conformación óptima. Esto se refleja en el parámetro N_tor, ya que ∆**Sconf** es directamente proporcional al número de torsiones. Además, W_conf representa un peso constante que define cuánta energía se asigna a esta penalización, calibrado en AutoDock a partir de datos experimentales para reflejar la contribución energética de cada torsión restringida en el complejo proteína-ligando.

$$\Delta Sconf = W\_conf * N\_tor$$

$$\Delta Gu = V + \Delta Sconf$$

Ecuación 2 *Corrección entropía conformacional en AutoDock-GPU (ΔS_conf). Este término penaliza la flexibilidad del ligando, ya que al fijarse en el sitio activo pierde grados de libertad*

En resumen, la función de puntaje de AutoDock combina dos contribuciones principales para estimar la energía libre de unión (ΔGu) entre un ligando y un receptor. Por un lado, utiliza un potencial que modela la interacción ligando-proteína e incorpora la solvatación de manera implícita mediante tres enfoques: (i) una constante dieléctrica variable que atenúa las cargas, (ii) la variación en la superficie accesible al solvente, y (iii) parámetros dependientes de la carga para estimar la interacción con el agua. Por otro lado, incluye un término de entropía conformacional basado en la flexibilidad del ligando. Cuanto menor sea el puntaje (más negativo), mayor es la afinidad de unión predicha, lo que indica una interacción más fuerte entre el ligando y el receptor.

2.1.2.3 Preparación de las estructuras de ligandos y proteínas

En el caso de las proteínas las estructuras fueron obtenidas de PDB y preparadas mediante una verificación estructural y la protonación de sus residuos, llevada a cabo con las herramientas de AutoDock. Para generar los correspondientes archivos PDBQT del receptor, se empleó el script *prepare_receptor4.py*, para calcular las cargas atómicas, se utilizó el método rápido de Gasteiger PEOE49, el cual asigna las cargas en función de las diferencias de electronegatividad entre los átomos.

Los átomos de hidrógeno capaces de formar enlaces de hidrógeno, es decir, los que están unidos a O, N o S, se representaron de manera explícita. El resto de los hidrógenos, los no polares, se trataron de manera implícita, fusionándose con el átomo pesado al que estaban covalentemente unidos, y los parámetros de van der Waals y las cargas se ajustaron en consecuencia.

En el caso de los ligandos, algunos fueron obtenidos directamente de PDB (junto con su receptor), en el resto de los casos fueron generados a partir de las cadenas SMILES correspondientes obtenidas de bases de datos como PubChem y convertidos al formato MOL utilizando la librería RDKit [39]. Posteriormente, se generaron las estructuras tridimensionales (3D) de los ligandos, y sus energías fueron minimizadas utilizando OpenBabel [38], garantizando así conformaciones energéticamente favorables para los experimentos de docking.

2.1.2.4 Cálculo de la energía en grillas

La **grilla de energía en AutoDock** es una estructura tridimensional (cubo o un prisma rectangular) de puntos que define el espacio de búsqueda en el que se evalúan las posibles interacciones entre el ligando y el receptor durante el proceso de docking. Esta grilla permite

que AutoDock optimice el cálculo de la energía de interacción, al pre computar las contribuciones energéticas de cada tipo de átomo del ligando en cada punto de la grilla, lo que reduce considerablemente el tiempo de cálculo ***Figura 2***.

La generación de la grilla se realiza mediante el programa **AutoGrid**, que es parte del conjunto de herramientas de AutoDock. AutoGrid crea un mapa de la energía de interacción entre un átomo del ligando (o un grupo funcional específico) y cada uno de los diferentes puntos de la grilla. La grilla abarca el área de interés en la proteína (generalmente el sitio activo o un dominio específico), y su tamaño y resolución pueden ajustarse de acuerdo con la geometría del sistema y el tamaño del ligando.

El proceso de generación de la grilla incluye los siguientes pasos:

1. **Definición del centro y tamaño de la grilla**: Se especifican las coordenadas centrales del sitio de unión y las dimensiones de la grilla, lo que determina el área en la que se explorarán las posibles posiciones del ligando.
2. **Precalculo de los mapas de energía**: AutoGrid calcula previamente los mapas de interacción para diferentes tipos de átomos del ligando (como hidrógenos, carbonos, oxígenos, etc.) con todos los átomos del receptor. Para este proceso, se coloca secuencialmente un átomo de prueba (para cada uno de los tipos de átomo del ligando) en cada punto de la grilla, y se calcula la energía de interacción entre el átomo de prueba y los átomos de la proteína, siguiendo los términos independientes de la carga de la ecuación
3. **Resolución de la grilla**: La resolución determina el número de puntos que cubren el espacio tridimensional. Una resolución más alta (más puntos) proporciona mayor precisión pero aumenta el tiempo de cálculo. En este trabajo el espacio se utilizó un espaciado entre puntos de 0,375 Å,

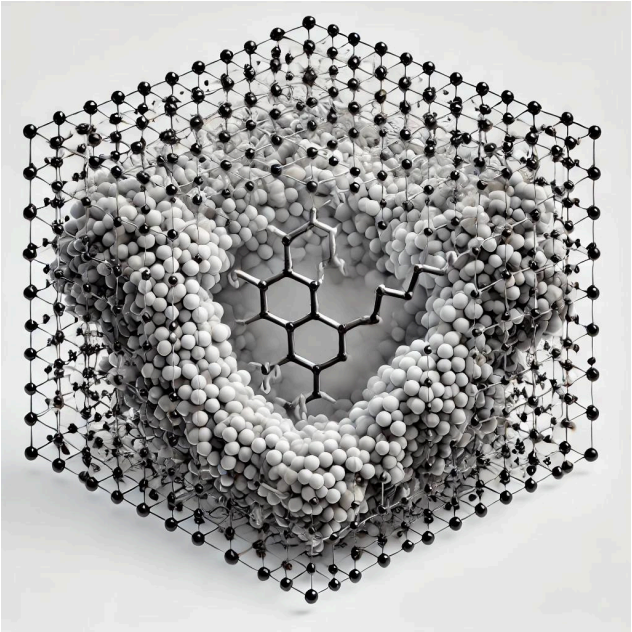


Figura 2 *Representación de la grilla tridimensional utilizada en el proceso de docking molecular, que cubre el sitio activo de la proteína, definida por un conjunto de puntos equidistantes (0,375 Å en este trabajo).*

2.1.2.5 Docking convencional

Para realizar el docking convencional, como ya mencionamos, se utilizó AutoDock-GPU (versión OpenCL y CUDA acelerada de AutoDock 4.2.6) con parámetros por defecto. Las estructuras de los ligandos y la proteína se prepararon siguiendo lo descrito en la sección anterior. Para la proteína rígida, se generaron mapas de energía a pares para cada tipo de átomo de los ligandos en el entorno del sitio de unión. La grilla se definió a partir de las coordenadas del ligando de referencia, centrada en su centro geométrico, y su tamaño se ajustó para que cada lado fuera el doble de la distancia máxima entre dos átomos del ligando, garantizando que cubriera adecuadamente el sitio activo. El espaciado entre puntos de la grilla se mantuvo en el valor predeterminado (0,375 Å). Se realizaron 100 simulaciones independientes de docking, y las poses generadas fueron agrupadas en clústeres utilizando un umbral de RMSD de 2 Å entre átomos pesados, permitiendo identificar conformaciones equivalentes y facilitar el análisis de los resultados.

2.1.2.6 Bias Docking (Sesgado)

El método de bias docking, desarrollado previamente en el grupo donde se llevó adelante esta tesis, es una modificación de AutoDock, que consiste en modificar los mapas de energía de AutoDock, aprovechando que los grupos polares de los ligandos suelen ocupar las mismas posiciones que los sitios hidrofílicos de agua o etanol observados en simulaciones de DM. Además, los sitios hidrofóbicos del etanol tienden a replicar las interacciones hidrofóbicas de los ligandos, en particular las que involucran anillos aromáticos.

El efecto del sesgo se logra agregando un término de energía adicional a la función original de AutoDock, por ejemplo, para cada átomo pesado del ligando capaz de formar enlaces de hidrógeno (como los átomos de tipo OA, NA y N unido a HD), de acuerdo a la ***Ecuación 3***

$$\Delta G_{\mathrm{SSBD}} = \Delta G_{\mathrm{AutoDock}} - RT \sum_{i=1}^{N} [\ln(\mathrm{PFP}_i)] e^{-\frac{\sqrt{(x-x_i)^2+(y-y_i)^2+(z-z_i)^2}}{R_{90,i}}}$$

Ecuación 3 *Modificación de la función de energía de AutoDock para incorporar el sesgo hidrofílico e hidrofóbico en el método de Bias Docking. La ecuación introduce un término adicional basado en la probabilidad de ocupación de solvente (PFP) y la distancia entre el sitio del solvente y los átomos pesados del ligando. R90 representa el radio medio del sitio de solvente, mientras que la profundidad y el ancho del pozo energético dependen de la PFP y R90, respectivamente.*

En esta ecuación, ΔG_{SSBD} representa la función de puntaje modificada, mientras que $\Delta G_{AutoDock}$ es la función de puntaje original de AutoDock4. Los términos R y T corresponden a la constante de los gases y la temperatura (298 K), respectivamente. La suma se realiza sobre el número total N de sitios de agua (o etanol-OH) que interactúan con el sitio de unión de la proteína. La *probe finding probability* (PFP) del sitio de solvente y las coordenadas (x, y, z) son los puntos de la grilla, mientras que (x_i, y_i, z_i) son las coordenadas del sitio de solvente. R90, es el radio medio del sitio de solvente

Por tanto, para cada sitio de agua o etanol-OH identificado, se genera un pozo de energía en los mapas correspondientes a OA, NA y N en la posición del sitio de solvente. La profundidad del pozo, que representa la magnitud de la recompensa energética, aumenta en función de la PFP del sitio de solvente. El ancho del pozo, que indica su extensión espacial al alejarse del centro del sitio de solvente, se incrementa con el valor de R90.

Para incluir el bias hidrofóbico, el método es similar. Para los anillos aromáticos el sistema crea para cada anillo aromático del ligando un nuevo átomo (*dummy atom*) artificial localizado en el centro del anillo (sin carga ni capacidad de interacción según la función de puntaje de AutoDock). Luego se crea un mapa de energía nuevo para el tipo de átomo artificial cuya energía se calculó usando la Ecuación anterior, con $\Delta G_{AutoDock} = 0$ y usando los sitios hidrofóbicos determinados a partir del CH3 del etanol. El método sesgado fue empleado del mismo modo que el convencional, pero con los mapas de energía modificados.

2.1.3 Dinámica Molecular

El modelado molecular es una disciplina que busca simular, entender y predecir el comportamiento de sistemas moleculares, principalmente mediante el uso de herramientas computacionales. Dado que muchos sistemas son demasiado grandes para ser tratados con métodos de mecánica cuántica, como es el caso de macromoléculas (proteínas o ácidos nucleicos), se emplean métodos de mecánica molecular. Estos métodos se basan en campos de fuerzas clásicos (similares la función de puntaje de Autodock) que modelan las interacciones entre los átomos mediante ecuaciones simplificadas, omitiendo el movimiento de los electrones y calculando la energía únicamente en función de las posiciones de los núcleos, de acuerdo con la conocida aproximación de Born-Oppenheimer.

En esta tesis trabajamos con las simulaciones de dinámica molecular se realizaron utilizando el paquete **AMBER 22** [40]. Los sistemas fueron preparados con **LEaP**, la dinámica se llevó a cabo con **pmemd.cuda**, y el análisis posterior se realizó con **cpptraj.**

La DM clásica es una técnica de simulación computacional basada en la mecánica molecular, que permite estudiar los movimientos de átomos y moléculas que interactúan entre sí a lo largo del tiempo. Este método permite el cálculo de diversas propiedades dinámicas, estructurales y termodinámicas del sistema, por ejemplo, las proteínas. Dado que la función de las proteínas está estrechamente vinculada a su dinámica estructural, en particular a sus

movimientos internos y cambios conformacionales, el estudio de la dinámica de las proteínas adquiere una importancia fundamental.

Las propiedades dinámicas y termodinámicas de cualquier sistema de partículas dependen tanto de la posición como del momento de todas las partículas que lo componen. Para determinar estas propiedades, es necesario conocer la probabilidad de que el sistema se encuentre en cada uno de sus posibles estados microscópicos, lo que implica evaluar un gran número de réplicas del sistema (un "ensamble") y calcular el promedio de esas réplicas para obtener la propiedad termodinámica deseada. Sin embargo, en sistemas de partículas como biomoléculas en solución (proteínas, fragmentos de ADN, oligosacáridos, etc.), simular la cantidad de réplicas requeridas para representar un experimento real (que sería del orden de billones) es infinitamente costoso en términos computacionales.

La DM utiliza un enfoque alternativo, en lugar de simular múltiples réplicas simultáneamente, el método sigue la evolución temporal de un único estado microscópico del sistema, o sea una réplica, y a partir de esa trayectoria en el tiempo, se calculan las propiedades de interés, promediando los valores obtenidos a lo largo de la simulación. La idea clave detrás de esta estrategia es que, dado un tiempo de simulación lo suficientemente largo, el sistema pasará por todos los microestados accesibles, permitiendo obtener una representación adecuada de su comportamiento. Este enfoque se denomina hipótesis ergódica, y es un principio fundamental de la Mecánica Estadística. Este enfoque permite, por ejemplo, estudiar la dinámica de biomoléculas en entornos complejos, como proteínas en solución, y estimar propiedades termodinámicas o de estabilidad estructural a partir de una única trayectoria de simulación de DM.

En las simulaciones de DM, los microestados sucesivos del sistema se generan mediante la integración temporal de las ecuaciones de movimiento de Newton. A través de este proceso, se obtiene una trayectoria que describe cómo varían las posiciones y velocidades de las partículas a lo largo del tiempo de acuerdo a la ***Ecuación 4***. La fuerza que actúa sobre cada partícula se calcula derivando la función de energía potencial que define el campo de fuerzas.

$$\frac{d^2 x_i}{dt^2} = \frac{F_{x_i}}{m_i}$$

***Ecuación 4**: x_i es una coordenada dada (x,y o z), t es el tiempo, mi es la masa de la partícula y Fx_i es la fuerza sobre la partícula en la coordenada x_i*

2.1.3.1 Campos de fuerza

Los **campos de fuerza** son modelos matemáticos que describen las interacciones entre las partículas (átomos o moléculas) en las simulaciones de DM. Estos campos permiten calcular las fuerzas que actúan sobre cada partícula en función de su posición relativa respecto a las demás. En esencia, un campo de fuerza define cómo se comportan las moléculas en un entorno simulado, proporcionando una aproximación de la energía potencial total del sistema.

$$V(\mathbf{r}^N) = \sum_{enlaces} k_r(r - r_0)^2 \; + \; + \sum_{ángulos} k_\theta(\theta - \theta_0)^2 \; + \; + \sum_{diedros} \frac{V_n}{2}[1 + \cos(n\omega - \gamma)] \; + \; + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right)$$

Ecuación 5 Expresión general del potencial de energía molecular empleado en simulaciones clásicas. La ecuación incluye términos para las interacciones internas (enlaces, ángulos y diédricos) y no enlazantes (interacciones electrostáticas y de van der Waals, mediante el potencial de Lennard-Jones).

Esta ecuación (***Ecuación 5***) representa entonces la **energía potencial total** de un sistema de partículas. Cada término de la ecuación describe una contribución específica a la energía potencial del sistema, y en conjunto permiten modelar cómo los átomos interactúan entre sí. Esta ecuación, de manera similar a la función de puntaje de Autodock, tiene las siguientes contribuciones o términos:

Término de enlaces:

Este término representa la energía de los enlaces covalentes entre átomos adyacentes en una molécula. Utiliza un potencial armónico, donde:

- ***r*** es la distancia actual entre dos átomos enlazados.
- r_0 es la distancia ideal de enlace (o longitud de equilibrio).
- k_r es una constante de fuerza que mide la rigidez del enlace.

Este término aumenta a medida que la distancia entre los átomos se aleja de su valor de equilibrio r_0, es decir, a medida que se estira o comprime el enlace.

Término de ángulos:

Este término modela la energía asociada a los ángulos formados entre tres átomos conectados por dos enlaces covalentes. Al igual que el término de enlaces, utiliza un potencial armónico:

- $\boldsymbol{\theta}$ es el ángulo actual formado entre los tres átomos.
- $\boldsymbol{\theta_0}$ es el ángulo de equilibrio o ideal.
- $\boldsymbol{k_\theta}$ es una constante de fuerza que indica la resistencia a cambiar el ángulo.

Este término aumenta si el ángulo entre los átomos se desvía del ángulo ideal θ_0

Término de diedros:

Este término describe la energía asociada a la torsión o rotación alrededor de un enlace entre dos átomos. Involucra cuatro átomos consecutivos (formando un ángulo diedro), y su energía depende del ángulo de torsión:

- ω es el ángulo diedro actual.
- V_n es una constante que determina la magnitud de la barrera rotacional.
- n es el número de posibles rotaciones (n veces en 360 grados).
- γ es un desplazamiento de fase.

Este término se utiliza para modelar las barreras rotacionales que afectan la conformación de moléculas flexibles.

Término de electrostáticos:

Este término modela las interacciones electrostáticas entre todas las partículas del sistema:

- q_i y q_j son las cargas de los átomos i y j.
- r_{ij} es la distancia entre los átomos i y j.
- ϵ_0 es la permitividad eléctrica del vacío.

Este término sigue la ley de Coulomb, que describe la atracción o repulsión entre partículas cargadas. La interacción es inversamente proporcional a la distancia entre las partículas: cuanto más cerca están, más fuerte es la interacción.

Interacciones de van der Waals (Lennard-Jones):

Este término describe las interacciones de van der Waals entre átomos que no están directamente enlazados:

- El primer término $\frac{A_{ij}}{r_{ij}^{12}}$ describe la repulsión a muy corta distancia entre los átomos, modelando el principio de exclusión de Pauli, que evita que los electrones ocupen el mismo espacio.
- El segundo término $\frac{B_{ij}}{r_{ij}^{6}}$ representa la atracción de van der Waals, que domina distancias más grandes, modelando las fuerzas de dispersión de London.

El potencial de Lennard-Jones combina ambos efectos y presenta un equilibrio entre la atracción y la repulsión entre átomos.

Con esta función podemos entonces determinar la energía del sistema en cada microestado conformacional, y mediante su derivada analítica obtenemos la fuerza que actúa sobre cada átomo lo que permite predecir su evolución en el tiempo generando la DM.

2.1.3.2 Condiciones Generales

Después de definir el campo de fuerzas y el algoritmo que propaga esas fuerzas a lo largo del tiempo, es fundamental mencionar de manera breve los algoritmos auxiliares que permiten iniciar y estabilizar la simulación, asegurando que se reproduzcan correctamente las características del ensamble molecular a las temperaturas y presiones deseadas (en este caso condiciones normales de presion i.e 1 atm y temperatura i.e 25C). Para un análisis más detallado de cualquiera de estos temas, se puede consultar el manual de usuario de AMBER y las referencias proporcionadas aquí.

Las simulaciones computacionales de macromoléculas pueden realizarse bajo tres condiciones: (i) en vacío, (ii) en un entorno dieléctrico que simula implícitamente el apantallamiento de cargas debido al solvente, o (iii) rodeando la macromolécula con moléculas de agua explícitas y utilizando un modelo de solvente. Hay una amplia gama de modelos de agua disponibles para simular el solvente de manera explícita alrededor de una macromolécula.

El modelo de agua utilizado en esta tesis fue el TIP3P (Transferable Intermolecular Potential 3-Point) [41], modela la molécula de agua con tres cargas puntuales colocadas en los átomos de oxígeno y de hidrógeno, junto con parámetros de potencial Lennard-Jones aplicados en el oxígeno.

La asignación de las velocidades iniciales a los átomos se hace de manera aleatoria a partir de una distribución de velocidades de Maxwell-Boltzmann ***Ecuación 6***, que da la probabilidad p de que un átomo i de masa m_i tenga velocidad v_{ix} en la dirección x a la temperatura de interés T y donde k_B es la constante de Boltzmann.

$$p(v_{ix}) = \sqrt{\frac{m_i}{2\pi k_B T}} e^{-\frac{m_i v_{ix}^2}{2k_B T}}$$

Ecuación 6 Distribución de velocidades de **Maxwell-Boltzmann** para un átomo i de masa m_i, que describe la probabilidad de que dicho átomo tenga una velocidad v_{ix} en la dirección x a una temperatura T. Esta distribución se utiliza para asignar velocidades iniciales a los átomos en simulaciones de dinámica molecular.

Las simulaciones de esta tesis se llevaron a cabo en un ensamble isotérmico-isobárico (o sea a T y Presión constantes), que es comúnmente utilizado debido a su similitud con condiciones experimentales típicas. La temperatura en las simulaciones se controla ajustando la velocidad de las partículas, según el teorema de equipartición de la energía, que relaciona la temperatura con el promedio temporal de la energía cinética. Este ajuste se realiza mediante un algoritmo de tipo termostato. Para mantener la temperatura constante en esta tesis, se empleó un termostato estocástico basado en la dinámica de Langevin.

Para evitar efectos de borde en las simulaciones de sistemas que se realizan con un número finito de partículas, se usa lo que se conoce como “condiciones periódicas de contorno”, es decir la celda de simulación se repite en las tres direcciones de modo de generar un arreglo periódico infinito, en el que las partículas experimentan fuerzas como si estuvieran en el seno del fluido.

2.1.3.3 Sitios de solvente

El método usado para detectar los sitios de interacción proteína-solvente proviene de trabajos previos de nuestro grupo, en los cuales se desarrollaron diferentes modos de obtención de primero los sitios de hidratación (sitios de agua o water sites, WS). Los WS se definen como regiones del espacio adyacentes a la superficie proteica donde la probabilidad de hallar un átomo de solvente (agua en este caso) es mayor que la de encontrarlo en el seno de la solución. La estrategia para identificarlos implica la obtención de un conjunto de estructuras solvatadas de la proteína, al que se le aplica un algoritmo de clustering jerárquico. Este conjunto de estructuras solvatadas puede obtenerse de una simulación de DM o de apo-estructuras derivadas de diversos experimentos de cristalografía de rayos-X.

Los algoritmos de clustering jerárquico se basan en la creación de una matriz de distancias entre todos los elementos que se desean analizar y requieren dos parámetros. El primer parámetro, denominado “ξ”, que representa la distancia mínima que debe existir entre dos puntos para que se incluyan en el mismo clúster. El segundo parámetro, que llamaremos “μ”, se refiere al número mínimo de puntos necesarios para definir un clúster. El algoritmo inicia tomando la posición del primer punto y localiza todos los puntos que se encuentran a una distancia menor que ξ. A continuación, todos los puntos encontrados se agrupan para formar un clúster. Si el número de puntos que pertenecen al clúster es igual o mayor que “μ”, se define este clúster como un sitio de solvente, y se calcula el centro de masa de todos los puntos que lo integran, lo que establece su posición en el espacio.

Si extendemos el mismo análisis a otros tipos de solventes, por ejemplo el grupo CH3 del etanol, o el centro del anillo el Fenol, podemos definir los sitios de solvente cómo a las regiones en la superficie de las macromoléculas donde las moléculas del cosolvente, presentan alta ocupación. Los mismos se calculan utilizando el mismo algoritmos de clustering que para los WS, pero a partir de DM en solventes mixtos agua/etanol o agua/fenol. Estos sitios son cruciales para comprender interacciones biomoleculares y procesos de solvatación.

2.2 Métodos Bioinformáticos

2.2.1 Análisis Filogenético

Los árboles filogenéticos son herramientas fundamentales en biología, ya que permiten visualizar y comprender las relaciones evolutivas entre diferentes organismos y/o sus componentes moleculares. A través de estas representaciones gráficas, es posible observar cómo las especies han divergido a lo largo del tiempo, y reconocer su ascendencia común. Estos árboles pueden aplicarse a distintos niveles de organización, desde el ADN pasando por las proteínas y hasta los organismos enteros, facilitando así una comprensión clara de su trayectoria evolutiva a lo largo del tiempo [42].

Para la construcción de los árboles filogenéticos de proteínas, se parte de un alineamiento múltiple de secuencias, una herramienta que permite evaluar el grado de similitud o divergencia entre un conjunto de proteínas. Para realizar el alineamiento se le asigna a cada nucleótido (o aminoácido) una letra y se forma una matriz donde cada columna representa una posición en la secuencia y cada fila una de las secuencias a comparar (ver el ejemplo en la ***Figura 3***).

Protein 1	N	G	H	P	W	I/	L	A	A	Q
Protein 2	N	–	H	P	W	I	L	A	L	Q
Protein 3	N	G	–	I	L	V	A	A	A	Q
Protein 4	N	–	V	V	A	A	A	L	L	Q
Protein 5	N	I	L	A	A	A	A	L	A	Q

Figura 3 Matriz de alineamiento múltiple*, donde cada fila representa una secuencia distinta (Proteína 1 a 5) y cada columna corresponde a una posición alineada entre los residuos. El alineamiento permite identificar coincidencias, conservaciones y variaciones entre las secuencias, asignando puntuaciones que optimizan la correspondencia estructural y funcional entre los residuos.*

Durante el proceso de alineamiento, a cada coincidencia se le asigna un puntaje, al igual que a cada cambio o sustitución. Si es necesario, se introducen espacios en blanco (gaps) para lograr una mejor correspondencia entre las secuencias o sus segmentos. En la ***Figura 4*** se puede observar el resultado de un alineamiento como ejemplo. Para optimizar la

calidad del análisis, se eliminan las regiones menos informativas, reduciendo el impacto de inserciones, deleciones, y de zonas altamente variables que podrían generar ruido y afectar la precisión de los resultados. [43].

Figura 4 *Alineamiento de secuencias de aminoácidos*

BMGE (Block Mapping and Gathering with Entropy) [44] es una herramienta muy utilizada para filtrar alineamientos de secuencias, eliminando posiciones ambiguas mediante un análisis de la entropía de shannon** en cada posición, que permite conservar sólo las regiones evolutivamente relevantes. Este proceso reduce el ruido y mejora la precisión del árbol filogenético. Los parámetros clave incluyen:

- **-b**: tamaño mínimo de bloque conservado.
- **-w**: tamaño de ventana para el cálculo de entropía.
- **-h**: umbral máximo de variabilidad aceptada.
- **-g**: proporción máxima de gaps permitidos.

La construcción de los árboles filogenéticos en este trabajo se realizó utilizando el programa PhyML [46], un método de máxima verosimilitud que permite optimizar la topología del árbol, las longitudes de las ramas y los parámetros del modelo evolutivo. Este enfoque garantiza una representación precisa y robusta de las relaciones evolutivas, especialmente en conjuntos de datos complejos como los BacCYPs. Para seleccionar el modelo evolutivo más adecuado, se utilizó el módulo SMS (Smart Model Selection) integrado en PhyML [45]. Este módulo evalúa de forma automática distintos modelos evolutivos y selecciona el más apropiado utilizando criterios de información, como el AIC (Criterio de Información de Akaike), y el BIC (Criterio de Información Bayesiano). Este proceso, basado en el principio de máxima verosimilitud, permite una inferencia más eficiente y precisa, minimizando el riesgo de errores en la interpretación de las relaciones filogenéticas entre las secuencias del MSA.

2.2.2 Similitud Química y Tratamiento computacional de moléculas orgánicas pequeñas

En esta tesis utilizamos el índice de Tanimoto como principal descriptor de la similitud química entre compuestos (o ligandos).

2.2.2.1 Smiles

El sistema SMILES (Simplified Molecular Input Line Entry System) [47] es una forma sencilla de representar compuestos químicos utilizando una anotación en línea. Este método tipográfico emplea solo caracteres para describir las moléculas, especificando los átomos que las componen y cómo están conectados entre sí mediante enlaces. Una de sus principales ventajas es que cada SMILES es único para cada estructura molecular, lo que facilita su almacenamiento de manera comprimida y estandarizada. Esto permite realizar búsquedas eficientes en bases de datos y almacenar información molecular de forma compacta.

En un SMILES ***Figura 5***, los átomos se representan con su símbolo químico (por ejemplo, "C" para carbono y "Cl" para cloro), mientras que los enlaces simples no se especifican, ya que se sobreentiende la presencia de los mismos en la secuencia de caracteres. Los enlaces dobles y triples se indican con los símbolos "=" y "#", respectivamente. Los ciclos o anillos de átomos se denotan mediante la numeración de un átomo, permitiendo "cerrar" el anillo al reconectar este átomo con otro punto de la secuencia (o sea otro átomo del SMILE). Los hidrógenos generalmente no se incluyen a menos que sea necesario para aclarar propiedades específicas. También se pueden manejar estados de protonación, quiralidad e isótopos mediante símbolos adicionales.

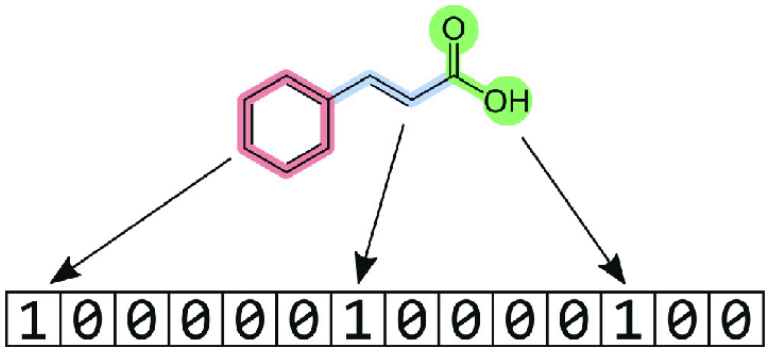


Figura 5 *Ejemplo de cómo una molécula se traduce en una huella digital. La presencia de subestructuras específicas se indica con un "1" en una posición determinada del vector* [48].

Para agilizar el proceso de filtrado de moléculas utilizando similitud química, se genera entonces un "*fingerprint*" a partir del SMILES de cada una. Este *fingerprint*, o huella digital molecular, es una representación abstracta que captura ciertas características clave de la estructura de la molécula, lo que permite comparar de manera eficiente grandes cantidades de compuestos. Las fingerprints (**huellas dactilares moleculares**) son representaciones binarias

o vectoriales que codifican la información estructural de moléculas, permitiendo la comparación y análisis de la similitud entre compuestos químicos. Estos vectores representan la presencia o ausencia de ciertas características estructurales de una molécula, como grupos funcionales, enlaces, o patrones de sustitución. Cada característica se codifica como un bit (0 o 1), donde:

- **1** indica la presencia de la característica.
- **0** indica la ausencia de la característica.

En esta, hemos generado un vector de 2048 bits para cada molécula utilizando RDKit, una biblioteca popular para la química computacional en Python.

2.2.2.2 Índice de Tanimoto

El **índice de Tanimoto** [49] es una métrica ampliamente utilizada en química y biología computacional para evaluar la similitud estructural entre moléculas pequeñas. Se basa en la comparación de huellas dactilares moleculares, que son representaciones binarizadas de las características estructurales de los compuestos. El índice, se calcula como la relación entre el número de características que comparten dos moléculas, y el número total de características únicas entre ambas. Su valor varía entre 0 y 1, donde 1 indica que las moléculas son idénticas, mientras que 0 sugiere que no comparten ningún elemento estructural en común, o sea son completamente diferentes.

Además del índice de Tanimoto, existen otros índices de similitud, como el **índice de Dice** y el **índice de Coseno**, que también se utilizan para comparar estructuras moleculares, cada uno tiene sus propias particularidades y áreas de aplicación. La elección del índice a utilizar depende del contexto y de las características específicas de los datos a analizar. El índice de Tanimoto es particularmente valioso para seleccionar compuestos similares químicamente que podrían actuar como ligandos efectivos de la misma proteína objetivo, y es por ello que es el elegido en la presente tesis.

Específicamente el índice se define con la ***Ecuación 7***. Donde $|\boldsymbol{A} \cap \boldsymbol{B}|$ representa el número de características compartidas entre las dos moléculas (es decir, los bits que están activados en ambas fingerprints), $|\boldsymbol{A} \cup \boldsymbol{B}|$ corresponde al número total de características únicas presentes en ambas moléculas. $\boldsymbol{A} \cdot \boldsymbol{B}$ es el producto escalar de los vectores binarios de las fingerprints, que indica la cantidad de posiciones en las que ambos tienen un bit activado, valor de 1, $|A|$ y $|B|$ corresponden al número total de bits activados en las fingerprints de las moléculas A y B, respectivamente.

$$Tanimoto(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{A \cdot B}{|A| + |B| - A \cdot B}$$

Ecuación 7 : Ecuación de tanimoto que permite evaluar cuán simil o disímiles son 2 ligandos.

A modo de ejemplo en la ***Figura 6***, se generaron dos moléculas que, a pesar de su similitud visual, presentaron un índice de Tanimoto moderado, cercano a 0.50. Este resultado indica que solo el 50% de sus fingerprints coinciden, lo que sugiere que, aunque las estructuras pueden parecer similares a simple vista, existen diferencias significativas en sus características químicas subyacentes.

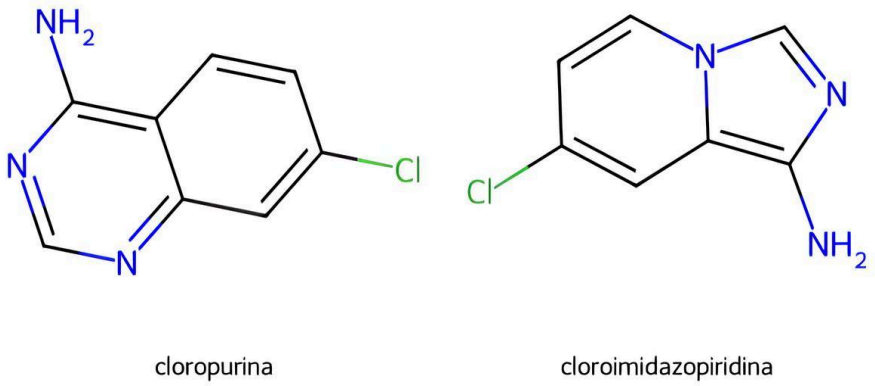


Figura 6 : Estructuras 2D de dos moléculas con un índice de similitud de Tanimoto cercano a 0.5, calculado a partir de sus fingerprints moleculares. Se observa la presencia de un anillo aromático clorado en ambas, con diferencias en la disposición de los átomos nitrogenados.

2.2.2.3 Utilización de Indices de Similitud Química en un esquema de VS (el método LigQ)

El Virtual Screening, cómo ya mencionamos en la introducción, es una metodología computacional ampliamente utilizada para identificar nuevos inhibidores de pequeñas moléculas dirigidos a un objetivo molecular (una proteína) específico. Este proceso implica la evaluación sistemática de miles o millones de compuestos provenientes de grandes bases de datos, con el objetivo de seleccionar un subconjunto de ligandos potencialmente activos que serán posteriormente analizados en el laboratorio. Dado que el costo asociado a la evaluación experimental de cada compuesto es elevado, es fundamental optimizar la selección virtual (o bioinformática) de ligandos, para maximizar la cantidad de inhibidores reales (activos) minimizando, al mismo tiempo, el número total de compuestos a testear, ya que de este modo se disminuye el costo total del proyecto.

En el grupo de investigación donde se desarrolla la presente tesis, previamente se desarrolló e implementó una herramienta denominada **LigQ**, diseñada para facilitar y optimizar el proceso de Virtual Screening, que se centra, en su paso crítico, en la utilización de índices de similitud química, como el de Tanimoto.

En su conjunto la herramienta LigQ permite:

1. **Identificar la mejor estructura y el sitio de unión** de una proteína específica, utilizando criterios de accesibilidad y propiedades de unión.
2. **Detectar ligandos conocidos y potenciales** que interactúan con dominios proteicos similares, a partir de bases de datos como PDB y ChEMBL.
3. **Seleccionar un conjunto optimizado de compuestos comerciales** enriquecidos en potenciales ligandos, utilizando como punto central los criterios de similitud química.
4. **Preparar las estructuras moleculares** de estos compuestos para su posterior evaluación mediante docking molecular.

LigQ fue utilizado en esta tesis para el enriquecimiento del conjunto de ligandos destinados al Virtual Screening contra LolE. La aplicación de LigQ facilitó la identificación y preparación de ligandos potenciales, optimizando el proceso de VS y aumentando la eficiencia del análisis experimental posterior. Además, algunas de las funcionalidades y/o herramientas de LigQ fueron utilizados para la identificación de sustratos de los CYPs.

2.2.3 Análisis estructural de la interacción proteína ligando

Como mencionamos en la introducción, las interacciones entre una proteína y su ligando pueden estar mediadas por varios tipos de fuerzas no covalentes, como enlaces de hidrógeno, interacciones hidrofóbicas, interacciones iónicas y fuerzas de Van der Waals. Cada uno de estos tipos de interacción contribuye a la estabilidad del complejo proteína-ligando y a su afinidad, determinando en última instancia la eficacia (o fuerza) de la unión. Para identificar y analizar las interacciones moleculares entre proteínas y ligandos de manera sistemática, se desarrolló un script en Python 3 que automatiza el proceso de detección de sitios activos, aceptores y donores de puentes de hidrógeno, así como interacciones aromáticas. Este script facilita el análisis de simulaciones de dinámica molecular y modelos estructurales generados mediante herramientas de predicción como AlphaFold.

2.2.3.1 Script búsqueda de interacciones

El script utiliza bibliotecas como RDKit para el manejo de moléculas y cálculo de coordenadas atómicas, Bio.PDB para la manipulación de estructuras de proteínas en formato PDB, y pandas para la gestión y análisis de datos. Se definieron patrones SMARTS para la detección de aceptores y donores de puentes de hidrógeno, además de algoritmos para el cálculo del centro de masa de las estructuras y la identificación de anillos aromáticos.

Para cada proteína y ligando analizado, el script extrae las coordenadas atómicas, calcula el centro de masa del ligando y determina los residuos del sitio activo de la proteína en un radio predefinido. Posteriormente, identifica y almacena las coordenadas de los átomos involucrados en interacciones potenciales, verificando las distancias y ángulos para determinar la validez de cada interacción. Las interacciones confirmadas se almacenan en archivos CSV para su análisis posterior, y se generan representaciones visuales en VMD mediante la creación automática de scripts en TCL.

El script también genera representaciones 2D de los ligandos ***(Figura 7)***, resaltando los sitios activos importantes. Esta representación permite visualizar de manera clara las características estructurales de los ligandos y su relación con las interacciones que forma con el receptor, lo que es crucial para comprender la mecánica de las interacciones a nivel molecular.

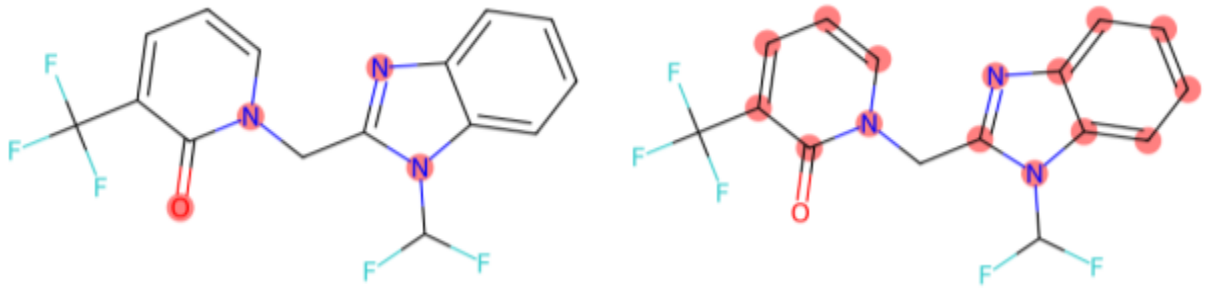


Figura 7*: Salida del script de interacciones. En el panel izquierdo se resaltan las interacciones tipo aceptor de puentes de hidrógeno, mientras que en el panel derecho se destacan las interacciones dadoras y aromáticas.*

Los parámetros utilizados para la detección de interacciones incluyen umbrales de distancia y ángulo específicos para cada tipo de interacción. Para los puentes de hidrógeno, se utilizó en este caso, un umbral de distancia de 3.5 Å y ángulos entre 100° y 200°. Las interacciones aromáticas se consideraron válidas cuando la distancia entre centros de anillos fue inferior a 5.5 Å y el ángulo entre planos aromáticos estuvo entre 0°-30° o 85°-95°. Estos valores fueron definidos en un archivo YAML de configuración, lo que permitió ajustar dinámicamente los parámetros sin modificar el código fuente.

Este enfoque automatizado del análisis de interacciones permitió analizar de manera eficiente un gran número de complejos proteínas y ligandos, contribuyendo significativamente a la comprensión de los mecanismos de unión y la especificidad de sustrato en diversas enzimas, un ejemplo de visualización de interacciones se puede observar en la ***Figura 8***

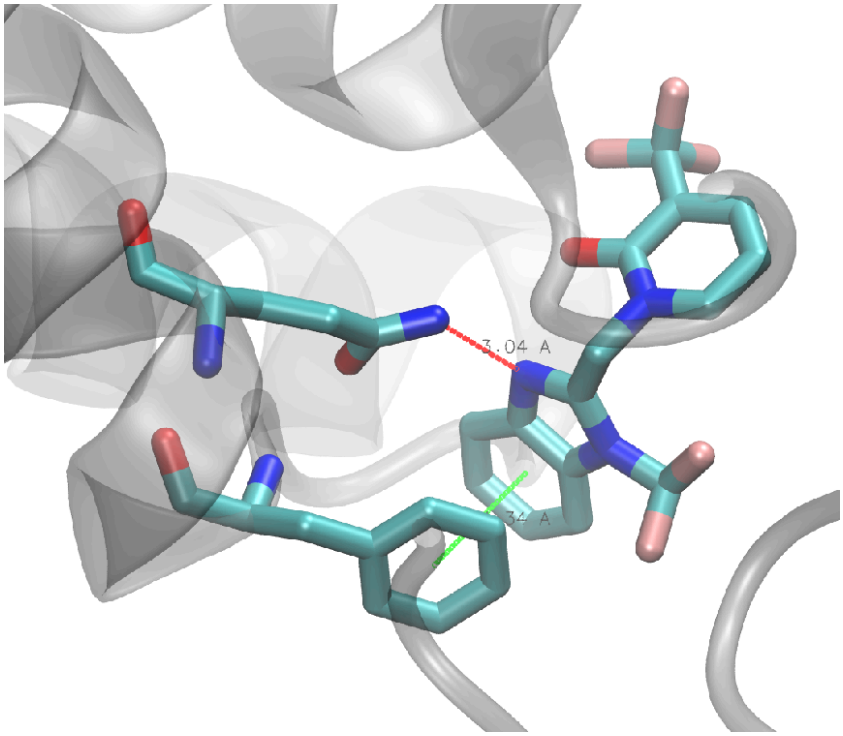


Figura 8 *Salida del script de interacciones generado para VMD. El script genera un archivo* ***TCL****, el cual es utilizado para generar la imagen con las interacciones y sus respectivas distancias.*

2.2.3.2 Determinación del volumen de una cavidad (*Convex Hull*).

La envolvente convexa (*convex hull*) es una herramienta geométrica que permite calcular el volumen de un espacio tridimensional, al determinar el poliedro convexo más pequeño que contiene un conjunto de puntos. Para estimar el volumen del sitio activo de una proteína, implementamos un método que construye la envolvente convexa utilizando como puntos de referencia los átomos más cercanos al centro de masa del ligando, seleccionando únicamente aquellos ubicados a una distancia máxima de hasta 8 Å. De esta manera, el volumen obtenido representa una aproximación confiable del espacio accesible al ligando dentro del sitio activo, proporcionando información valiosa sobre su capacidad de alojamiento. Este procedimiento ha sido implementado en un script con el objetivo de evaluar de manera eficiente el volumen del sitio activo, obteniendo un valor aproximado y una representación gráfica que facilita su análisis visual. A modo de ejemplo en la ***Figura 9***

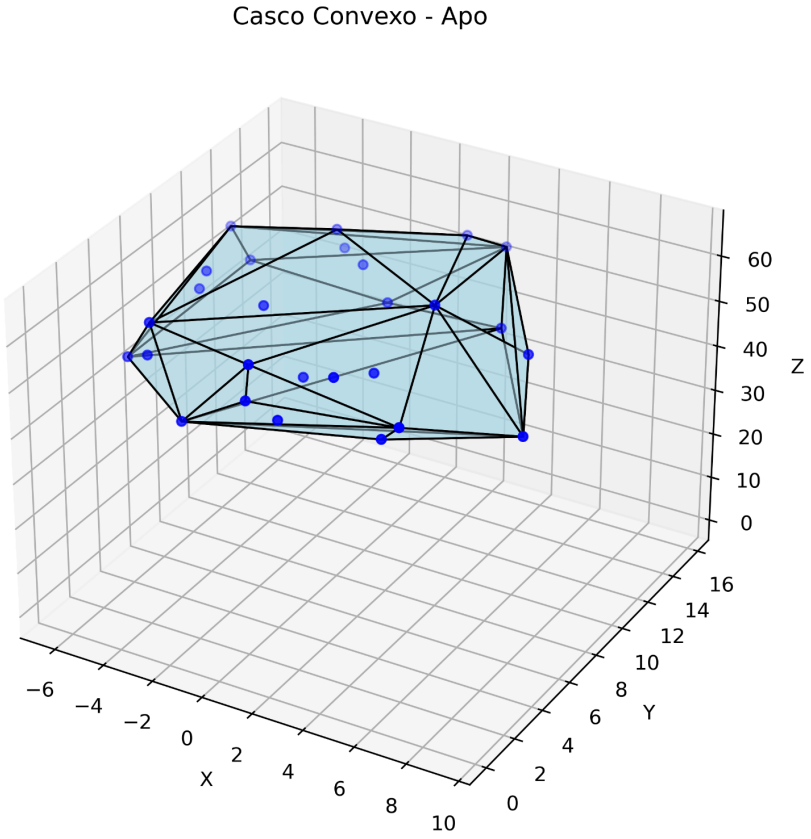


Figura 9 Se observa el volumen estimado para el sitio activo de 1R1W , utilizando el método convex hull, siendo el conjunto de puntos el centro de masa de cada residuo dentro del radio de 8Å generado por el centro de masa del ligando

2.2.4 Análisis estadísticos de los resultados de Docking y Virtual Screening

Los resultados del docking molecular pueden evaluarse a través de dos parámetros fundamentales: la energía libre de unión, y la población relativa de las conformaciones obtenidas. Estos valores proporcionan información clave sobre la estabilidad y la frecuencia con la que una determinada conformación (o pose) aparece entre las soluciones generadas por el algoritmo. Para facilitar su análisis e interpretación, estos datos pueden representarse en un gráfico 2D, como se muestra en la ***Figura 10***, donde el eje horizontal corresponde a la energía libre de unión, y el eje vertical a la población. Este tipo de representación visual permite identificar de manera clara las conformaciones más estables y predominantes, que son las que se ubican en el extremo superior (alta población) e izquierda (baja energía) del gráfico.

Sin embargo, uno de los desafíos que surge en estos análisis es la dificultad para comparar poblaciones que pueden presentar valores de energía distintos, dependiendo de las características específicas de cada ligando y cada proteína. Esta variabilidad puede dificultar la interpretación directa de los resultados, y por lo tanto la comparación entre diferentes sistemas (como veremos en algunas secciones de la presente tesis).

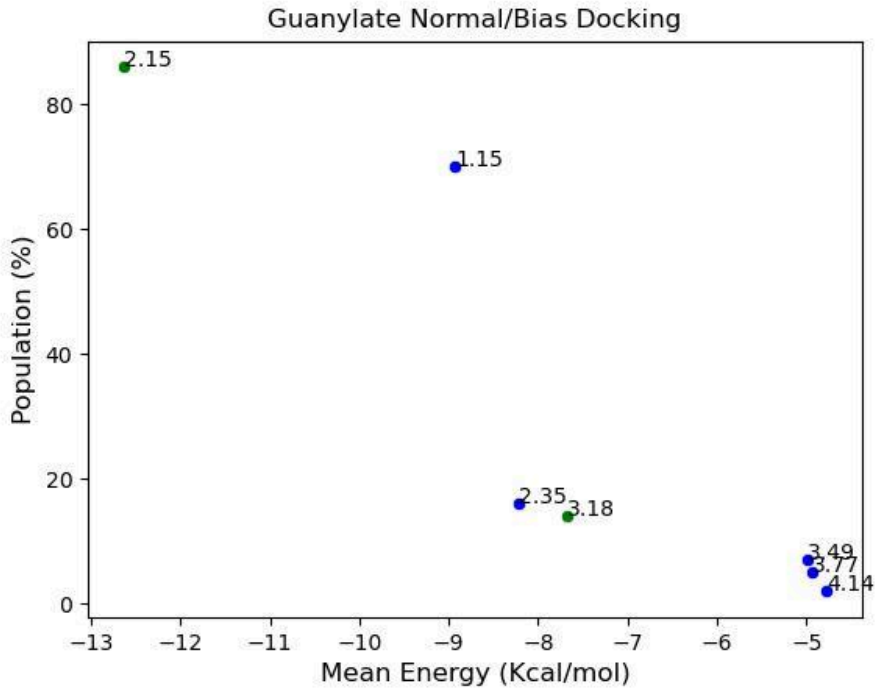


Figura 10: Gráfico 2D resumiendo los resultados obtenidos de un docking de la familia guanylate

Una forma de abordar este problema es normalizando los datos, lo que permite ajustar las poblaciones y las energías (de diferentes ligandos y/o receptores) en una escala común. La normalización facilita la comparación entre distintos complejos, proporcionando una visión más clara de las tendencias generales y permitiendo identificar patrones que podrían pasar desapercibidos en una evaluación directa de los datos crudos.

2.2.4.1 Normalización (Zeta Score)

Para facilitar la comparación entre diferentes resultados de los experimentos de docking, se aplicó la normalización utilizando el Zeta Score (Z-score). Este estadístico cuantifica la diferencia entre un valor observado Xi y un parámetro (media , mediana, etc) *(**Ecuación 8**)*

$$Z = \frac{X_i - \mu}{\sigma}$$

Ecuación 8 Definición de Z-score de cada para receptores - ligando. Donde X_i es el valor a normalizar, μ es la media y σ representa la desviación estándar, ambos de la población.

Esta normalización no sólo se utilizó para comparar las energías de unión de los ligandos en los ensayos de docking, sino que también fue fundamental para calcular los clústeres de poblaciones en cada uno. Al aplicar el Z score, se obtiene una medida estandarizada que permite identificar patrones y tendencias dentro de los datos, facilitando la agrupación de ligandos con características similares en función de sus energías de unión y poblaciones normalizadas. De esta manera se obtienen 2 valores de Z Scores , Z_{energy} y $Z_{cluster}$. Estos Z-scores fueron luego combinados para rankear (u ordenar) los ligandos y las poses en las diferentes aplicaciones de los ensayos de docking.

Docking-based virtual screening en LoICDE

3.1 Docking-based virtual screening en LolCDE

El objetivo de este capítulo es llevar a cabo una campaña de **Búsqueda Virtual (BV)** basada en docking contra el complejo **LolCDE de Kp**, una proteína esencial en bacterias Gram-negativas que participa en el transporte de lipoproteínas desde la membrana interna hacia el periplasma. Dada su relevancia funcional y su potencial como blanco terapéutico, LolCDE fue seleccionada como uno de los sistemas modelo en el marco de un proyecto colaborativo con GARDP, orientado a la identificación de nuevos inhibidores con potencial antimicrobiano mediante enfoques computacionales.

Este trabajo se enmarca dentro de una estrategia racional que busca explorar la capacidad de unión de pequeñas moléculas a través de la caracterización detallada del **sitio activo** y la identificación de **farmacóforos clave**. Se estima que, mediante esta campaña, podrían seleccionarse alrededor de **100 compuestos** con potencial capacidad de unión, derivados de una librería inicial construida a partir de ligandos reportados en literatura, y enriquecida mediante criterios fisicoquímicos y de diversidad estructural.

En las siguientes secciones se describen los métodos utilizados para definir el pocket de unión, generar los modelos de docking, aplicar filtrados bioquímicos y realizar el análisis de resultados, en el contexto de una búsqueda sistemática de **ligandos candidatos a inhibidores de LolCDE**.

3.2 Métodos

Para llevar a cabo la BV, se tomó como punto de partida la estructura cristalográfica de la proteína LolCE (obtenida de ***Escherichia coli***), obtenida previamente y depositada en el Protein Data Bank (PDB) bajo el código 7ARM. La definición del sitio de unión se estableció mediante una cuadrícula (grid) con dimensiones de 50 x 50 x 50 Å, y un espaciado entre puntos de 0,375 Å, centrada en el bolsillo de unión identificado en las estructuras cristalográficas dada la presencia en la estructura del sustrato natural.

Los experimentos de docking se realizaron utilizando AutoDock GPU y la versión Bias, considerando el receptor como una estructura rígida mientras que los ligandos fueron tratados con flexibilidad conformacional. Para cada ligando, se llevaron a cabo 100 ejecuciones independientes con el objetivo de explorar exhaustivamente los posibles modos de unión (o poses).

Las estructuras de los ligandos se obtuvieron a partir de sus respectivos SMILES, se convirtieron al formato **mol** utilizando RDKit [51]. Una vez en formato mol, fueron convertidos en estructuras 3D y sus energías se minimizaron utilizando OpenBabel[50].Este proceso garantizó que los ligandos estuvieran en conformaciones energéticamente favorables antes de ser sometidos al procedimiento de docking. El conjunto de ligandos fue obtenido de Zinc y pre-filtrado por diversidad .

Además del docking, se realizaron simulaciones de dinámica molecular para evaluar la estabilidad de los complejos formados entre LolCDE y los ligandos seleccionados. Se llevaron a cabo tres tipos de simulaciones bajo diferentes condiciones:

1. **Simulación 1:** Duración de 20 ns con una constante de restricción de 10 kcal/mol.
2. **Simulación 2:** Duración de 20 ns con una constante de restricción de 5 kcal/mol.
3. **Simulación 3:** Duración extendida de 50 ns con una restricción más suave de 1 kcal/mol.

Las restricciones fueron utilizadas para evitar que la proteína se despliegue, al menos de modo parcial, al ser extraída de la membrana y simulada en un entorno acuoso. Estas simulaciones permitieron evaluar la estabilidad de los complejos formados y refinar la selección de ligandos con mayor potencial de unión. La combinación del enfoque de docking con las simulaciones de dinámica molecular permitió obtener una caracterización más robusta de las interacciones entre LolCDE y los ligandos propuestos.

Además se realizaron MD de LolCDE en presencia de mezclas de solvente para la determinación de los sitios de solvente o hot-spots de interacción.

3.3 Resultados: LolCDE - búsqueda de hotspots

Se realizaron simulaciones de MD con solventes mixtos específicamente en mezclas agua-fenol y agua-etanol. El objetivo de estas simulaciones fue identificar "Hot spots" de interacción entre los cosolventes y la proteína. La lipoproteína fue eliminada de la estructura 7ARM antes de la solvatación.

Se realizaron tres MD, cada una con diferentes tiempos de simulación y diversas restricciones aplicadas a los átomos del esqueleto de los residuos transmembrana, estas últimas buscan emular el hecho de que la proteína se encuentra embebida y restringida dentro de la membrana plasmática. El mismo procedimiento se llevó a cabo en las simulaciones en agua-etanol, donde no se observaron sitios de solvatación.

Estas condiciones variables se diseñaron para estudiar el efecto de la flexibilidad del esqueleto proteico en las interacciones con cosolventes a diferentes escalas de tiempo y niveles de restricción. Uno de los primeros resultados obtenidos de las dinámicas fue la identificación de sitios específicos de interacción del cosolvente, en este caso fenol, con la proteína o "Solvent-Protein Interaction Sites" (SPF).

A lo largo de las simulaciones, se observó que los cosolventes interactúan de manera constante con las mismas regiones clave de la proteína. Específicamente, los sitios de fenol que denominaremos S1, S2 y S3 fueron comunes en todas las simulaciones de DM, independientemente de la fuerza de las restricciones o la duración de la misma. Sin embargo, un cuarto sitio (S4) de interacción apareció únicamente en la última simulación, en la cual las

restricciones eran más débiles (1 kcal/mol) y la simulación se extendió a 50 ns. Esto sugiere que simulaciones más largas con restricciones reducidas permitieron identificar nuevas regiones de unión para los cosolventes (***Tabla 1***). Los sitios se pueden observar en el contexto de la estructura de la proteína en la ***Figura 11***.

Simulación #	S1 - SPF	S2 - SPF	S3 - SPF	S4 - SPF
1	0.90	2.54	4.34	-
2	2..10	1.05	4.34	-
3	0.66	1.38	0.84	2.87

Tabla 1 *El SPF de los Sitios de Solvente en cada simulación representa la frecuencia o intensidad de interacción entre el cosolvente y la proteína en sitios específicos a lo largo de las simulaciones de dinámica molecular. Un valor mayor de SPF indica un contacto más frecuente y sostenido del cosolvente con el sitio correspondiente.*

Estos sitios de interacción con los solventes se encontraron cerca de los residuos que se conoce interactúan con las lipoproteínas, los cuales han demostrado ser cruciales para la función de LolCDE en estudios previos de mutagénesis [52]. La identificación de estos sitios de unión de solventes coincide con la relevancia funcional de dichos residuos, lo que sugiere un posible papel en la modulación de las interacciones con lipoproteínas.

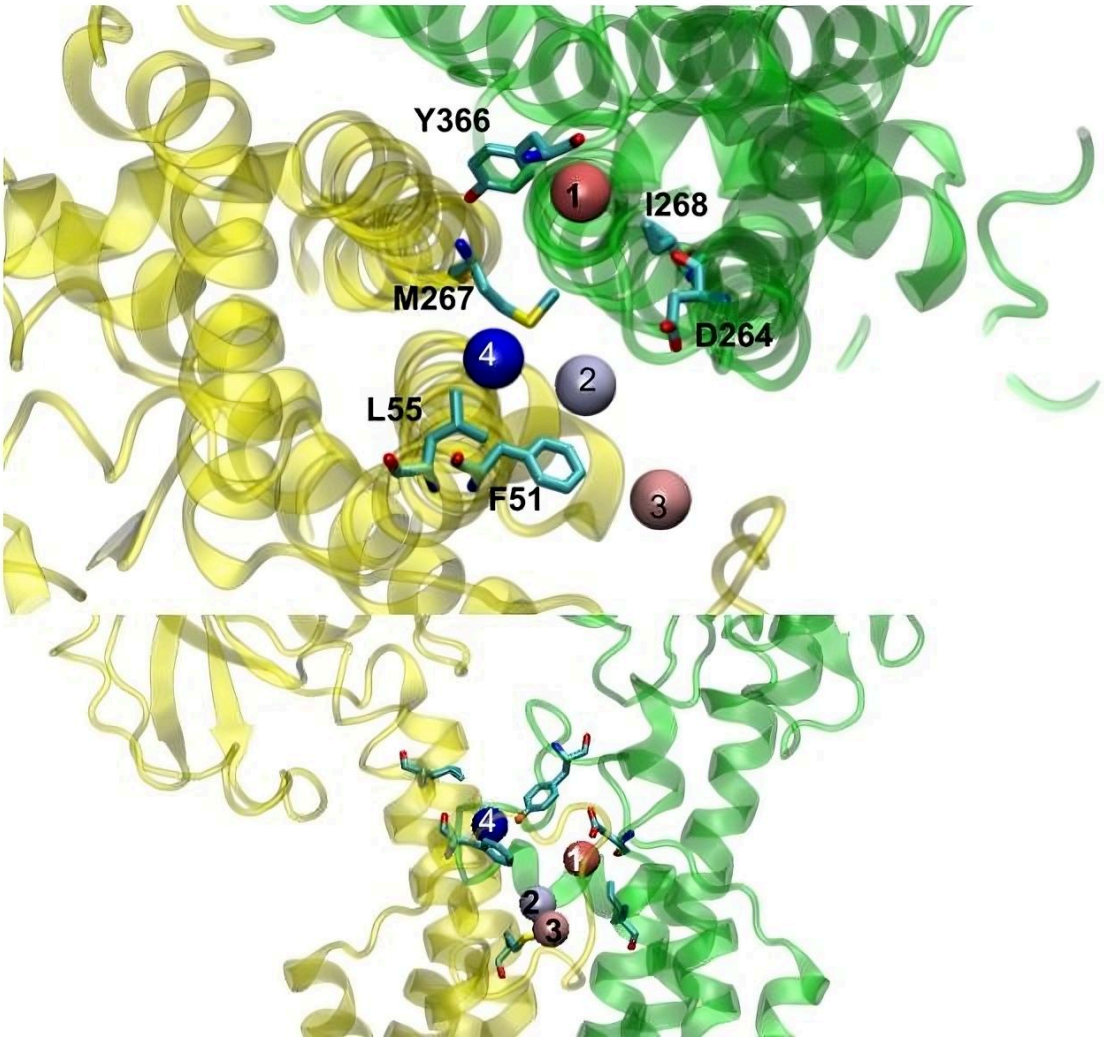


Figura 11 *Distribución espacial de los sitios de solvente 1a 4 y los residuos de LolC y LolE involucrados en las interacciones, formando el farmacóforo. LolE (verde) y LolC (amarillo). En formato "Licorice" se muestran los residuos orientados hacia los sitios de solvente. Panel superior desde una vista superior, Panel inferior vista lateral.*

3.4 Docking Ligandos Conocidos LolCDE

Se realizaron dockings con compuestos de referencia para validar tantos los parámetros del docking, como el modo de unión de los inhibidores conocidos, caracterizar el pocket, y determinar mejor el farmacóforo. Los compuestos de referencia, que han mostrado ser activos frente a LolCDE, fueron derivados de literatura, en un análisis previo realizado en el marco de la colaboración con GARDP , y se encuentran resumidos en la ***Tabla 2***.

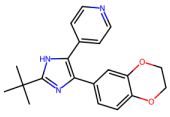
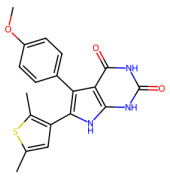
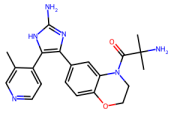
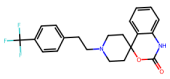
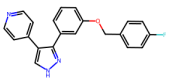
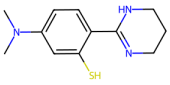
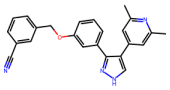
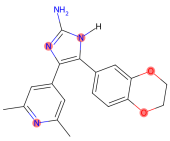
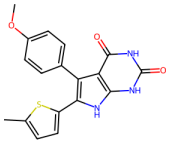
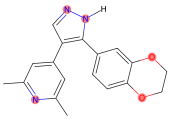
Ligando	Nombre	Estructura	Ligando	Nombre	Estructura
Ligand 1	Compund-1		Ligand 6	G0793	
Ligand 2	SMT-738		Ligand 7	Aubacin	
Ligand 3	Compund-2		Ligand 8	CCT-00432	
Ligand 4	Lolamycin		Ligand 9		
Ligand 5	G0507		Ligand 10		

Tabla 2 *Ligandos considerados positivos , de los cuales se tiene conocimiento que unen a LolCDE.*

Se realizaron simulaciones de docking estándar para los 10 ligandos positivos, y para cada caso se calculó la población y la energía de unión. La mayoría de los ligandos exhibieron resultados favorables en los análisis de docking estándar, aunque el 30% de los casos fallaron, mostrando poblaciones menores al 50%. El resto de los casos presentaron buenos resultados, con poblaciones entre el 50% y el 90%, aunque en ningún caso se observaron poblaciones superiores al 90%. A modo de ejemplo se puede observar el caso del ligando 8 (CTT 00432) , siendo un resultado de docking que consideramos correcto, mientras que en el caso del ligando 4, que consideramos fallo, ya que se observa un gran número de poblaciones todas de baja población. Ambos ejemplos se observan en la ***Figura 12***

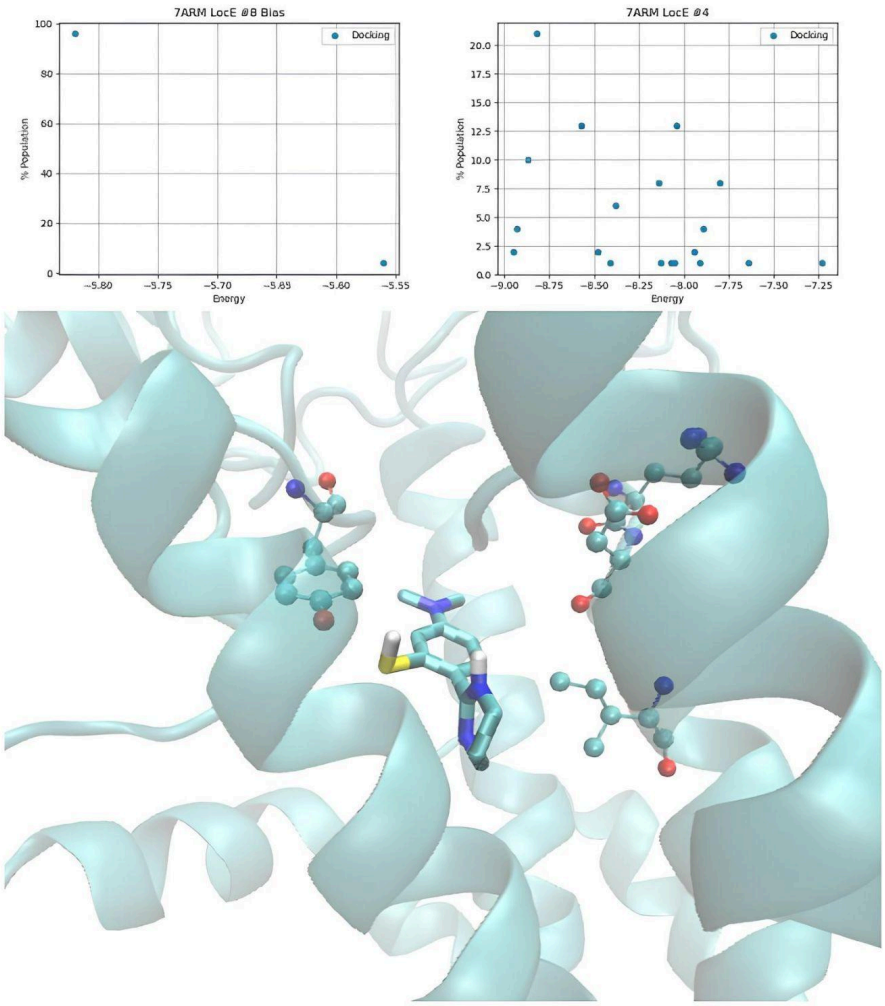


Figura 12 *El panel izquierdo muestra un el resultado del docking del ligando 8 (CCT-00432) mediante un gráfico de energía libre de unión (kcal/mol), en cual se observa una población mayor al 90%, considerado un docking correcto. El panel derecho presenta el docking con el ligando 4 no habiendo poblaciones mayoritarias, siendo considerado un mal docking . En el panel inferior se visualiza, mediante VMD, la pose de menor energía, destacando interacciones clave como enlaces de hidrógeno y contactos hidrofóbicos.*

En la comparación de las poses con los sitios identificados previamente, los resultados mostraron consistentemente una preferencia por el sitio 1, independientemente de si la población superaba el 90% (o si se acoplaron múltiples poses, como por ejemplo, para el ligando 4), como se muestra en la ***Figura 13***. Esta preferencia fue consistente en todas las poses, a pesar de la diversidad de los resultados. Además, dentro del sitio activo se identificó una marcada afinidad por los residuos TYR 366 y ASP 264 de LolE. Sobre el sitio 2, en cambio, se encontraron muy pocos ligandos y en general en poses de baja población.

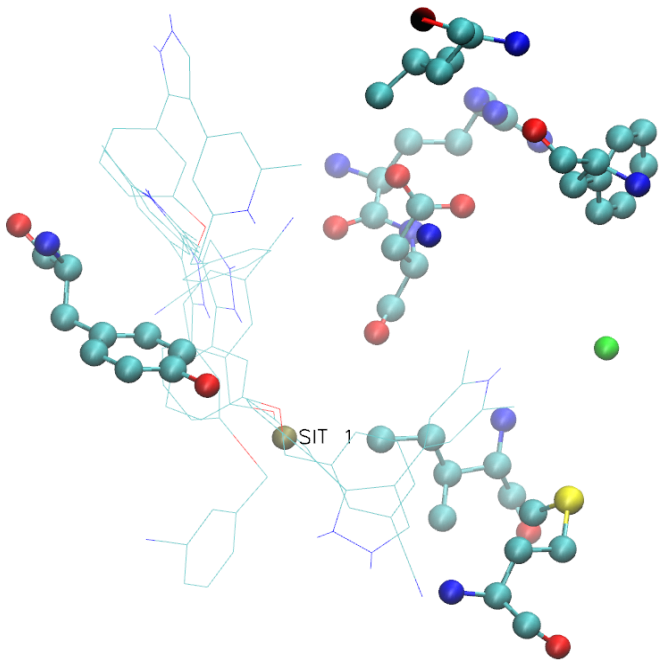


Figura 13. *Sitio de unión detectado para el ligando 4 se observa que a pesar de las múltiples poses todas oscilan alrededor del sitio denominado 1.*

Llamativamente, para el Sitio 3, a pesar de haber obtenido el mayor valor de SPF, no se encontraron ligandos que interactúan en su zona. Probablemente esto se deba a que se encuentra en una región más profunda de la proteìna donde es posible se dificulte el acceso de los ligandos, El sitio 4 se encuentra más cerca de la cadena LolC y sólo en proximidad a LolE-TYR 366 en una zona menos activa que a priori debe tener un alta restricción por estar estar en la zona embebida en la membrana (***Figura 14***).

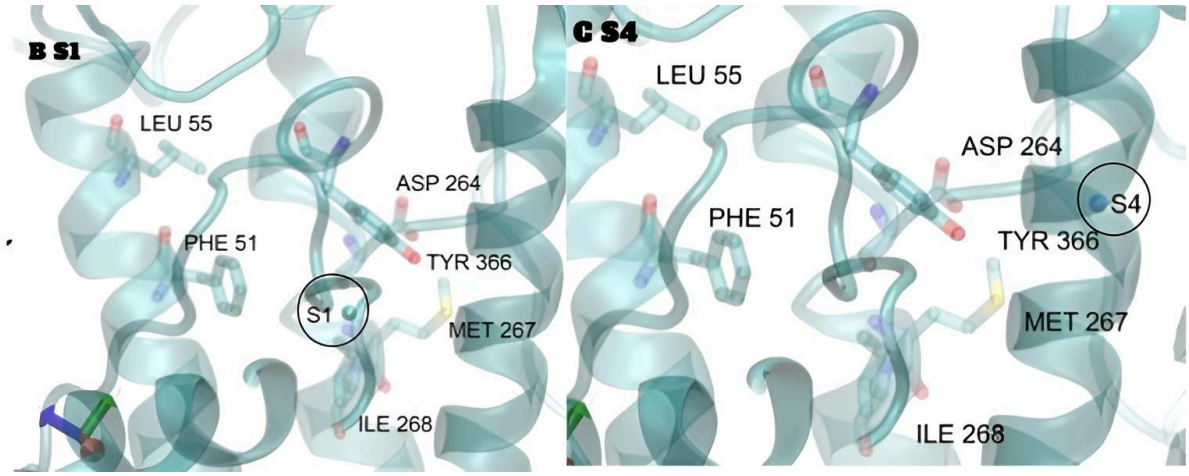


Figura 14 Ubicación del Sitio 1 (Panel Izquierdo) y del Sitio 4 (Panel Derecho), en el mismo se observa el desplazamiento del S4 hacia una zona más externa del sitio activo y con menos residuos con los que interactuar el ligando.

En base a los datos obtenidos en esta etapa de análisis, se determinó que el Sitio 1 presenta las características más favorables para su selección como punto de bias para los dockings subsecuentes. Esta decisión se fundamenta en la consistencia observada en las interacciones de los ligandos con este sitio, independientemente de las condiciones de simulación o la diversidad de poses de acoplamiento generadas. Además, el Sitio 1 mostró una alta frecuencia de contactos y una afinidad preferencial hacia los residuos clave, lo que refuerza su relevancia como un punto crítico de unión para futuros estudios enfocados en explorar su potencial funcional y farmacológico.

La siguiente tabla (***Tabla 3***) resume los resultados de los análisis de docking realizados, tanto en condiciones normales como con bias (utilizando el Sitio 1 como punto de referencia), para los 10 ligandos activos de LolCE. La escala utilizada para evaluar los resultados se basa en porcentajes de población y se define de la siguiente manera: **Excelente (valores mayores al 90%)**, **Muy Bueno (comprendidos entre 90 % y 80%)**, **Bueno (en la franja del 80% al 70%)**, **Aceptable (70% - 60%)** y **Malo (50% o menos)**.

Ligand	Standard Docking	Bias Docking
Ligand 1	Bueno	Excelente
Ligand 2	Aceptable	Excelente
Ligand 3	Aceptable	Excelente
Ligand 4	Aceptable	Muy Bueno
Ligand 5	Bueno	Muy Bueno
Ligand 6	Bueno	Muy Bueno
Ligand 7	Bueno	Muy Bueno
Ligand 8	Bueno	Excelente
Ligand 9	Bueno	Excelente
Ligand 10	Muy Bueno	Excelente

Tabla 3 *Resultados de los docking con y sin bias para los 10 ligandos activos de LolCE.*

La ***Figura 15*** muestra un análisis más pormenorizado de los resultados de docking utilizando la pose mayoritaria de los casos con bias. En el mismo se observa que en todos los casos un grupo fenilo del ligando se encuentra asentado sobre el sitio de bias en casi todas las

poses principales, lo que sugiere una interacción consistente y significativa con esta región de la proteína, la cual está definida por el residuo TYR 366. Esta interacción podría influir en la orientación y estabilidad de los ligandos durante el docking, destacando la importancia del sitio de bias en la unión molecular. La recurrencia de este patrón, sugiere que el grupo fenilo desempeña un papel clave en la afinidad y selectividad de los ligandos hacia el sitio activo de LolCDE.

Es importante destacar que aunque existe cierta flexibilidad en la orientación del grupo fenilo, este siempre permanece cerca del sitio de bias. Esto indica que, a pesar de ajustes espaciales, su proximidad constante contribuye a la estabilización de los complejos ligando-proteína. En conjunto, esto resalta al sitio de bias, definido por TYR 366, como un punto focal crucial en la interacción molecular.

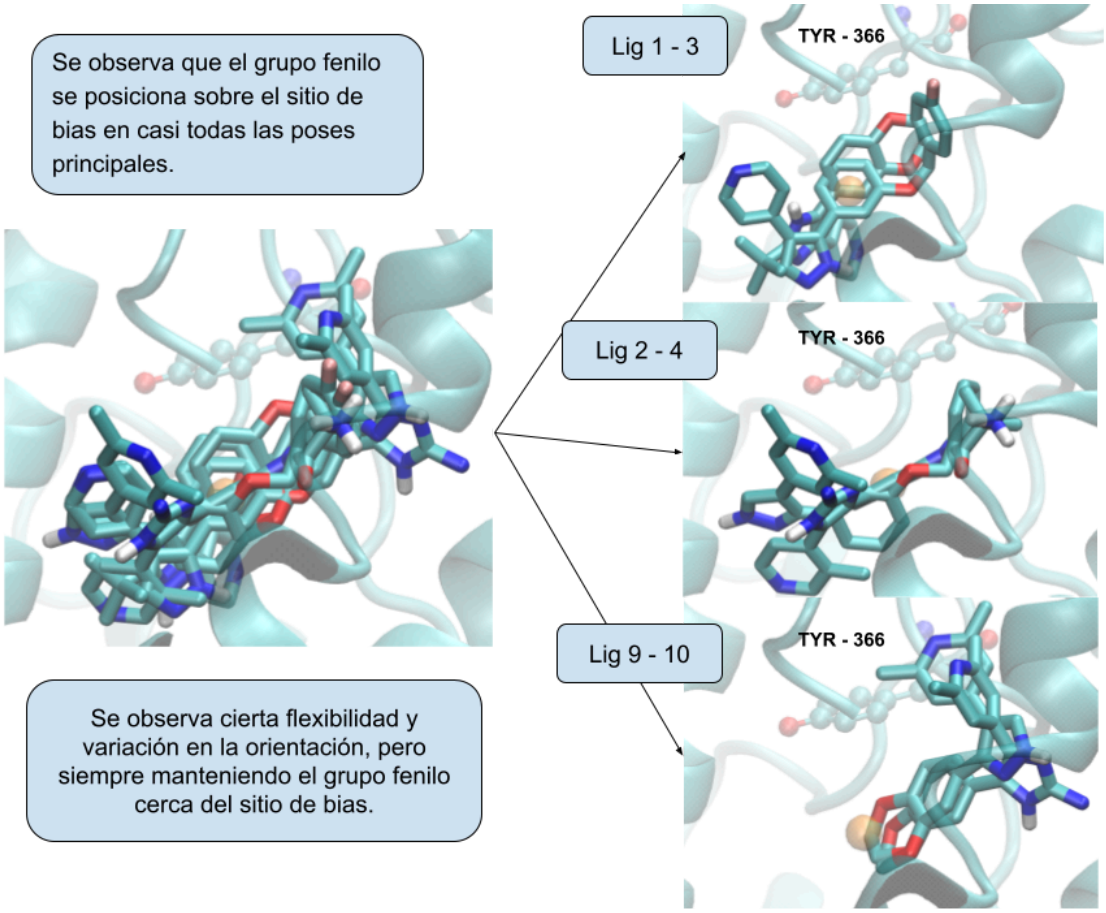


Figura 15 *Poses de docking de los ligandos 1-3, 2-4, and 9-10 de LolCE. . En todos los casos, el grupo fenilo de los ligandos establece una interacción constante con el residuo TYR366 en el sitio de unión..*

Con los resultados obtenidos del docking, procedió a analizar las interacciones que se encuentran entre los ligandos y el sitio activo en la ***Tabla 4*** se resumen los resultados obtenidos. En los mismos se observa que el residuo TYR 366 juega un papel recurrente como

un punto clave de interacción, específicamente mediante interacciones tipo π-π. El residuo forma interacciones consistentes con los ligandos 1, 3, 5 y 8, confirmando su relevancia como parte del sitio de bias. Esto refuerza la hipótesis de que TYR 366 es un punto focal en la orientación y estabilización de los ligandos dentro del sitio activo.

Además, se destacan otros residuos, como TYR 260 y PHE 51, que también participan en interacciones relevantes, ya sea mediante enlaces de hidrógeno o interacciones tipo π-π, lo que sugiere que estas regiones complementan el papel estabilizador del residuo TYR 366. En particular, PHE 51 aparece como un segundo punto recurrente en las interacciones con los ligandos, reforzando su posible contribución al proceso de unión molecular.

Finalmente, los residuos ILE 365, MET 261 y PHE 367 también intervienen en algunas interacciones, aunque de manera más variable. Esto podría indicar que estos residuos aportan flexibilidad estructural al sitio activo, permitiendo acomodar diferentes ligandos y ajustarse a sus disposiciones espaciales.

Ligando	Posición	Residuo	Interacción
Ligand 1	260	TYR	H.B.
	260	TYR	Pi-Pi
	366	TYR	Pi-Pi
Ligand 2	51	PHE	Pi-Pi
Ligand 3	366	TYR	Pi-Pi
Ligand 4	51	PHE	Pi-Pi
Ligand 5	260	TYR	Pi-Pi
	366	TYR	Pi-Pi
	51	PHE	Pi-Pi
	367	PHE	Pi-Pi
Ligand 6	260	TYR	Pi-Pi
	51	PHE	Pi-Pi

	365	ILE	H.B.
	367	PHE	Pi-Pi
Ligand 7	51	PHE	Pi-Pi
	261	MET	H.B.
	367	PHE	Pi-Pi
Ligand 8	366	TYR	Pi-Pi
	365	ILE	H.B.
Ligand 9	260	TYR	Pi-Pi
Ligand 10	51	PHE	Pi-Pi
	261	MET	H.B.

Tabla 4 Resumen de las interacciones encontradas entre las poses principales de los resultados de docking bias.

En la ***Tabla 5*** se presenta un resumen complementario de las interacciones observadas, reorganizado desde la perspectiva de los **residuos participantes**. Este formato permite identificar con mayor claridad cuáles son los aminoácidos que contribuyen con más frecuencia a la estabilización de los ligandos en el sitio activo de la proteína. Entre ellos, destacan **PHE 51**, que participa en interacciones tipo π-π en el **60% de los casos analizados**, y **TYR 260**, con un **50% de aparición**, distribuidos entre interacciones aromáticas (π-π) y puentes de hidrógeno (40% y 10%, respectivamente).

A partir de los resultados obtenidos, se ha definido entonces un parámetro claro para identificar y caracterizar un buen resultado de docking. El siguiente paso en este proceso implica la preparación de un conjunto de datos robusto y representativo, que servirá como base para realizar un virtual screening efectivo y dirigido, para buscar potenciar nuevos inhibidores de LolE.

ID / Residue	TYR 260 (π-π)	TYR 260 (H.B.)	TYR 366 (π-π)	PHE 51 (π-π)	MET 261 (H.B.)	ILE 365 (H.B.)	PHE 367 (π-π)	Total
Ligando 1	SI	SI	SI	NO	NO	NO	NO	3
Ligando 2	NO	NO	NO	SI	NO	NO	NO	1
Ligando 3	NO	NO	SI	NO	NO	NO	NO	1
Ligando 4	NO	NO	NO	SI	NO	NO	NO	1
Ligando 5	SI	NO	SI	SI	NO	SI	NO	4
Ligando 6	SI	NO	NO	SI	NO	SI	SI	4
Ligando 7	NO	NO	NO	SI	SI	NO	SI	3
Ligando 8	NO	NO	SI	NO	NO	SI	NO	2
Ligando 9	SI	NO	NO	NO	NO	NO	NO	1
Ligando 10	NO	NO	NO	SI	SI	NO	NO	2
Total	4	1	4	6	2	3	2	
% Aparición	40	10	40	60	20	30	20	

Tabla 5. *Resumen de las interacciones observadas se muestran los residuos involucrados en interacciones específicas con los ligandos, indicando el tipo de interacción: Pi-Pi stacking (interacciones aromáticas) o puentes de hidrógeno (H.B.). Cada fila representa un ligando e indica el total de interacciones que contribuyen a su estabilización , mientras que las columnas indican con cuantos ligandos interacciona cada residuo. Los residuos más recurrentes en la estabilización de ligandos fueron PHE 51 y TYR 260 (π-π), con un 60% y 40% de presencia, respectivamente.*

3.5 Dinámica Molecular de Complejos Proteína - Ligando Seleccionados

El objetivo de este sección consistió en llevar adelante simulaciones de dinámica molecular para un grupo reducido de aproximadamente 5 complejos ligando-proteína, específicamente LolCE y los inhibidores conocidos, seleccionados en el paso anterior, con el fin de estudiar su comportamiento dinámico y validar las interacciones de unión identificadas en estudios previos.

Durante las simulaciones, se analizó la movilidad de los ligandos en el sitio activo y las interacciones presentes en la misma. Los resultados se encuentran resumidos en la ***Tabla 6*** y un ejemplo de los mismos, se puede observar en la ***Figura 16*** donde se observa para el caso del ligando 2. Al analizar su RMSD (Cuadrante A) se aprecia un leve movimiento del ligando respecto al cuadro inicial. También se puede ver la formación de un puente de hidrógeno (en el Cuadrante C) entre MET 261 y el ligando , donde la distancia entre ambos está a lo largo de la distancia de enlace (2.3 A). Todos estos datos se analizaron para los 10 ligandos, revelando que las interacciones aromáticas son las más predominantes en todos los casos estudiados. Estas interacciones (pi-stacking) entre los anillos aromáticos de los ligandos y los residuos TYR, PHE o TRP de la proteína, desempeñan un papel fundamental en la unión molecular. En particular, los residuos LolE-TYR 366 y LolE-TYR 260 emergen como actores clave, ya que participan activamente en la estabilización de los ligandos dentro del sitio activo. También, se identificaron otras interacciones de menor frecuencia, como enlaces de hidrógeno e interacciones hidrofóbicas, que podrían contribuir de manera complementaria a la estabilidad general del complejo ligando-proteína. Estas interacciones secundarias, aunque menos frecuentes (en relación al tiempo simulado), refuerzan la unión y podrían ser relevantes para el diseño de ligandos con mayor afinidad, y/o especificidad.

Ligando #	**RMSD SD (Å)**	**RMSD Average (Å)**	**Hydrogen Bond** (% presence of the interaction)	**Aromatic** (% presence of the interaction
Ligando 1	0.224	0.657	GLU 263 34% ARG 263 29% MET 261 10%	TYR 366 100%
Ligando 2	0.253	1.510	ASP 264 16% SER 3 14%	TYR 260 100%
Ligando 3	0.198	0.597	MET 261 57%	TYR 366 99%
Ligando 4	0.516	1.314	ARG 263 41% MET 261 23% GLU 263 11%	TYR 366 98.6%
Ligando 5	0.109	0.388	GLU 263 42%	TYR 366 100%
Ligando 6	0.205	0.580	GLU 263 8%	TYR 366 99.7%
Ligando 7	0.251	0.962	ASN 4 75%	TYR 366 38.5% TYR 260 100%
Ligando 8	0.262	0.564	ILE 365 39%	TYR 366 99% TYR 260 100%
Ligando 9	0.273	0.759	ARG 263 39%	TYR 366 89%

			MET 261 27%	
Ligando 10	0.199	0.597	MET 261 43%	TYR 366 95%

***Tabla 6. Detalles de las interacciones de los ligandos**, incluyendo los valores de **RMSD** (desviación cuadrática media), la presencia de **puentes de hidrógeno** y **interacciones aromáticas**.**RMSD SD (Å)** representa la **desviación estándar** del RMSD, mientras que **RMSD promedio (Å)** indica el valor medio del mismo.Se indica el **porcentaje de presencia** de cada tipo de interacción para cada ligando, destacándose las interacciones aromáticas predominantes, en particular aquellas que involucran a los residuos **LoIE–TYR 366** y **LoIE–TYR 260**.*

El comportamiento dinámico observado confirma que los ligandos tienden a mantener su posición en el sitio activo, interactuando de manera consistente y estable con los residuos clave de la proteína LoIE identificados previamente. Estos hallazgos subrayan el papel crítico de las interacciones aromáticas en la unión molecular, y proporcionan una base sólida para futuros esfuerzos de optimización de ligandos, orientados a mejorar su eficacia terapéutica.

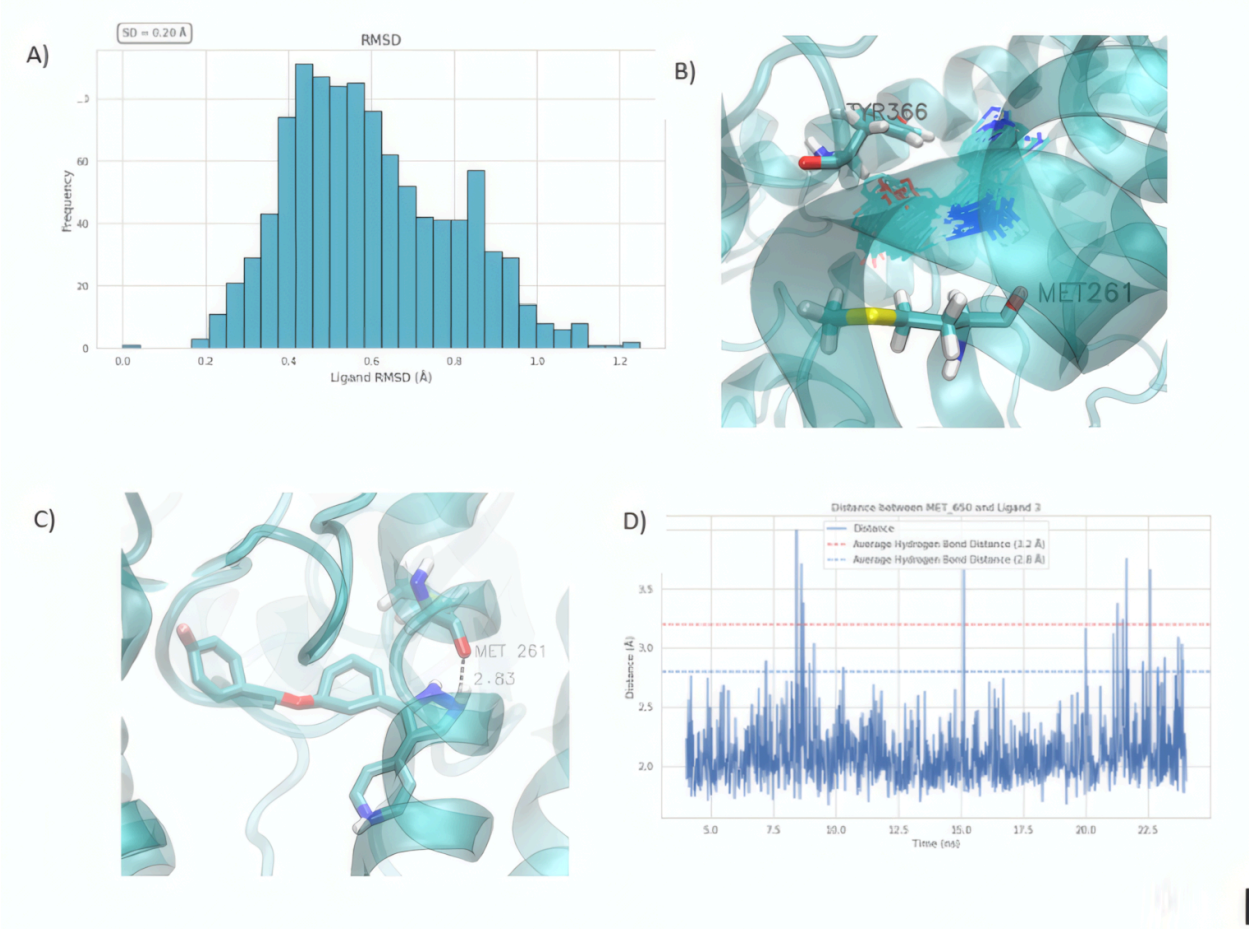


Figura 16: *El análisis de la movilidad del ligando 3 se llevó a cabo midiendo el Desvío Cuadrático Medio (RMSD) con respecto a su pose inicial.* ***Cuadrante A*** *El histograma muestra la distribución de los valores de RMSD del ligando a lo largo de la simulación.****Cuadrante B*** *Representación estructural de las interacciones clave entre el ligando y los residuos del sitio de unión, como LolE-MET261 y LolE-TYR366.* ***Cuadrante C*** *Se destaca la interacción de enlace de hidrógeno entre LolE-MET261 y el ligando.****Cuadrante D*** *La distancia entre LolE-MET261 y el ligando se monitorea durante toda la dinámica, mostrando una distancia consistente que respalda la formación estable del enlace.*

Las simulaciones de dinámica molecular iniciales identificaron a LolE-TYR366 como el principal sitio de interacción en el bolsillo de unión de fenol, lo que justificó su selección como punto de sesgo en los futuros estudios de docking. Sin embargo, los análisis posteriores con ligandos activos revelaron que LolE-PHE51 también desempeña un papel significativo en la unión, complementando la interacción con LolE-TYR366. Además, se observó la contribución de otros residuos, como MET261, en la estabilización del ligando. Esto sugiere que el sitio activo de LolCDE involucra la participación cooperativa de varios residuos aromáticos-hidrofóbicos, cuya interacción con el ligando depende, en parte, de su tamaño y disposición espacial. Para resumir esta información gráficamente, en la ***Figura 17***, en el panel izquierdo muestra el sitio activo, incluyendo el punto de sesgo y los residuos participantes, mientras que el panel derecho presenta el total de interacciones observadas en las 10 simulaciones dinámicas realizadas.

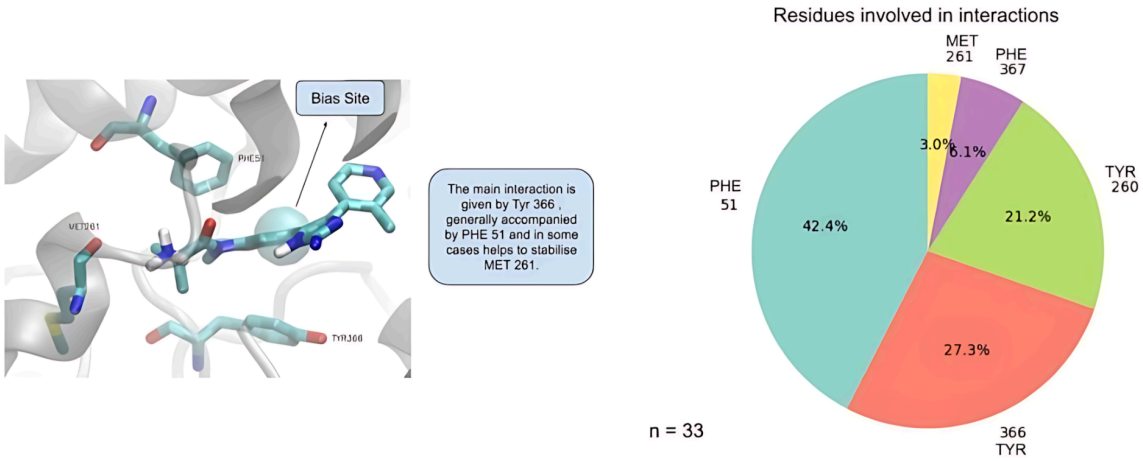


Figura 17. Panel Izquierdo *Sitio activo de LolCDE con los residuos participantes de las Interacciones clave en el sitio de unión.* ***Panel Derecho:*** *El porcentaje indica la proporción del total de interacciones en las que participa el residuo a lo largo de todos los casos.*

En resumen, los resultados de las simulaciones de dinámica molecular no solo **validan** las interacciones clave identificadas en los estudios de docking de la sección anterior (especialmente las π-π con TYR 366 y TYR 260), sino que también **enriquecen** el entendimiento al revelar cómo estas interacciones se mantienen o ajustan en un escenario dinámico, y más realista. Esta sinergia entre métodos estáticos y dinámicos refuerza la hipótesis de que las interacciones aromáticas son fundamentales para el diseño de inhibidores de LolCE, y proporciona una base sólida para optimizar ligandos con mayor afinidad y especificidad.

3.6 Preparación del conjunto de compuestos para el VS.

Tomando como punto de partida los resultados de docking para los ligandos conocidos de LolCE, se seleccionaron aquellos candidatos que dieran un buen resultado de docking y mantuvieran buenas interacciones en la MD, para utilizarlos como moléculas **semilla** en la construcción de nuevos conjuntos de compuestos. Los compuestos seleccionados fueron los ligandos 1 , 5 ,6 y 7. El conjunto de **semillas** se expandió evaluando la similitud química mediante el Índice de Tanimoto, utilizando compuestos provenientes de la base de datos ChemBL (2.5 millones de compuestos) [54] y la base de datos de compuestos adquiribles ZINC (12 millones de compuestos) [53]. Los rangos de similitud de Tanimoto utilizados variaron entre 0.6 y 0.8, dependiendo de la base de datos. El protocolo seguido se resume en la ***Figura 18.***

Los *fingerprints* se calcularon previamente para cada compuesto en las bases de datos ChemBL y ZINC utilizando SMILES [56] y MACCS *fingerprints* [55]. Luego, se calculó el valor de similitud de Tanimoto para cada compuesto al compararlo con los ligandos semilla, y se graficó su distribución. A partir de esta distribución se utilizaron los valores de tanimoto mayores dentro del rango de 0.6 a 0.8, hasta obtener cerca de 1000 ligandos candidatos para el VS. Una vez que alcanzado el número objetivo de compuestos, se calculó la intersección de ligandos presentes en los conjuntos filtrados de todas las bases de datos, con el fin de eliminar compuestos redundantes.

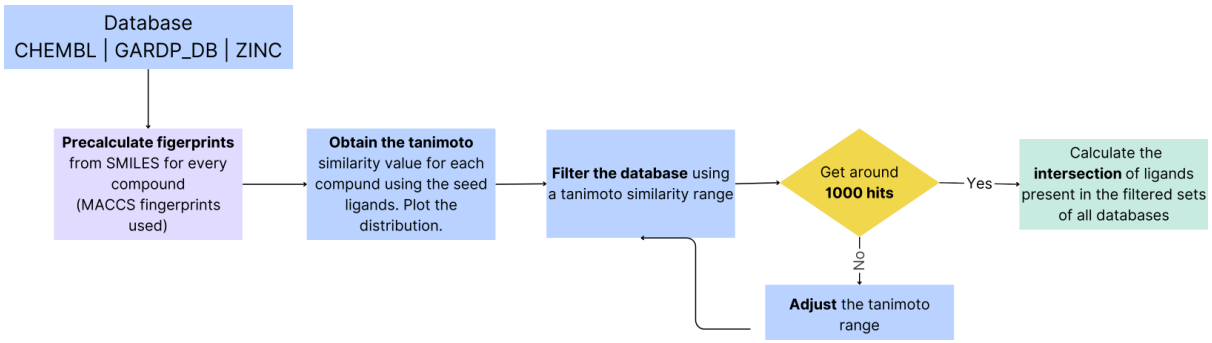


Figura 18. *Protocolo general que describe los pasos para identificar compuestos similares en bases de datos químicas utilizando la similitud de Tanimoto, incluyendo el cálculo de huellas dactilares (fingerprints), el filtrado y la intersección de ligandos.*

Siguiendo este protocolo, se obtuvo una biblioteca de ***3382*** compuestos únicos para LolCE (***Figura 19***). El origen de estos compuestos en las bases de datos fue el siguiente: 1032 compuestos provienen de ZINC (índice de Tanimoto 0.78–0.80), 1274 de ChemBL (índice de Tanimoto 0.75–0.80), de los cuales 299 se encontraron simultáneamente en ZINC y ChemBLDe esta manera queda establecido la base de datos para el virtual screening.

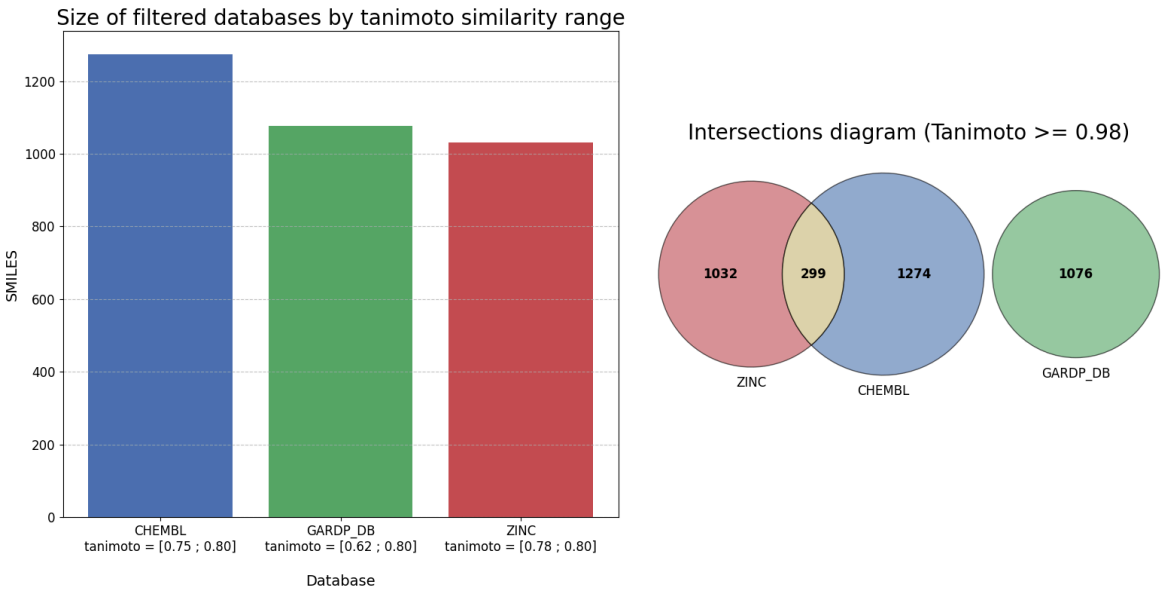


Figura 19. *Panel Izquierdo: Gráfico de barras que muestra el número de ligandos recuperados de cada base de datos en función de rangos específicos de similitud de Tanimoto para la biblioteca de compuestos de LolCE. Panel Derecho: Diagrama de Venn que representa la intersección de ligandos entre las bases de datos, filtrados por similitud. El tamaño de los círculos refleja las proporciones de los conjuntos de datos.*

3.7 High-Throughput Docking

Se realizó entonces el docking de una biblioteca de 3382 ligandos, tanto con bias como docking convencional contra la estructura de LolCDE, para identificar cuáles de ellos pueden ser relevantes como inhibidores de la proteína. Para analizar los resultados se calculó el Z-Score para los parámetros obtenidos del docking, la energía libre de unión y la población del clúster (es decir, la frecuencia con la que un ligando aparecía en un clúster específico a lo largo de 100 ejecuciones de docking), como se detalla en la sección de métodos.

Con los datos obtenidos, se realizó el flujo de trabajo que se muestra en la ***Figura 20*** el cual establece un método sistemático para la selección y priorización de los ligandos a partir de los resultados iniciales del Docking. Este proceso combina análisis estadísticos, geométricos y químicos para identificar ligandos con el mayor potencial de afinidad y estabilidad en el sitio activo de la proteína objetivo. El análisis comienza con el cálculo de dos Z-scores: Zenergy, que mide la desviación de las energías de interacción con respecto a la media, y Zcluster, que evalúa la densidad o agrupamiento de las poses de ligandos. Tras este filtrado inicial, los ligandos restantes son ordenados en función de Zenergy (de mayor afinidad energética a menor) y Zcluster (de mayor a menor densidad), otorgandole prioridad a la energía. Posteriormente, se calcula la distancia de cada ligando al punto de bias, una región clave en el sitio activo, como el residuo TYR 366 en LolCE. Los ligandos que excedan una distancia de 7.5 Å (o sea se encuentran muy lejos del mismo) son descartados. Este criterio geométrico permite

concentrarse en los ligandos que interaccionan directamente con las regiones críticas del sitio activo identificadas previamente en los dockings de los controles positivos y la MD.

De esta manera se obtiene un primer subset de 1000 complejos proteína-ligandos, a las cuales se somete a un análisis más fino de interacciones moleculares, para evaluar la cantidad y calidad de las interacciones entre ellos y la proteína. Este análisis final prioriza ligandos en función de cuatro parámetros: Zenergy, Zcluster, distancia al punto de bias, y el total de interacciones. De esta manera, se selecciona un conjunto final de 100 ligandos optimizados para su unión. En una siguiente instancia, estos ligandos candidatos que representan las opciones más prometedoras, deberían ser adquiridos y evaluados experimentalmente en el laboratorio para confirmar su actividad sobre LolCDE.

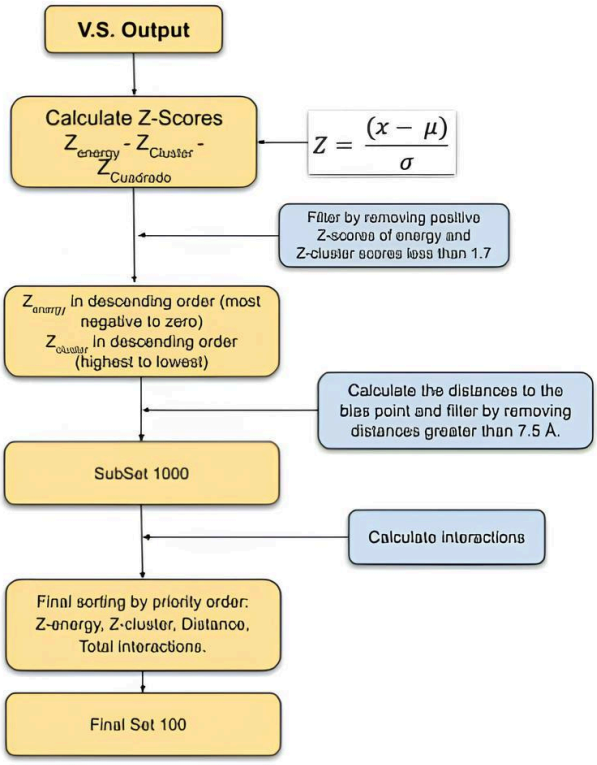


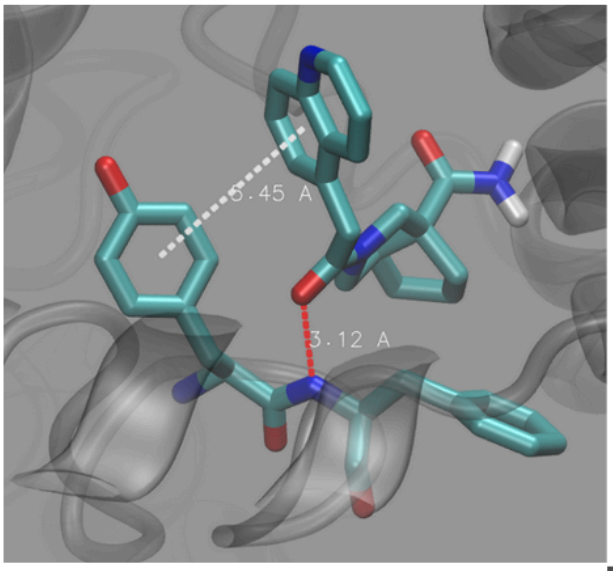
Figura 20 *Diagrama de flujo del protocolo para la selección de los 1000 compuestos principales.*

3.7 Análisis Compuestos Candidatos

Una vez obtenido el conjunto de los **1000 compuestos mejor rankeados**, se procedió a evaluar visualmente los primeros candidatos con el objetivo de confirmar si ocupaban el **sitio activo, si se encontraban en una conformación coherente,** y si presentaban **interacciones clave similares a las observadas en los controles positivos**. Para ello, se analizaron en detalle las **interacciones no covalentes** más relevantes (principalmente interacciones

aromáticas del tipo π–π y enlaces por puente de hidrógeno), identificando los residuos del sitio activo que participan con mayor frecuencia en el reconocimiento de ligandos.

A continuación, se presentan algunos ejemplos representativos de ligandos seleccionados, junto con un resumen general de los residuos más frecuentes en las interacciones

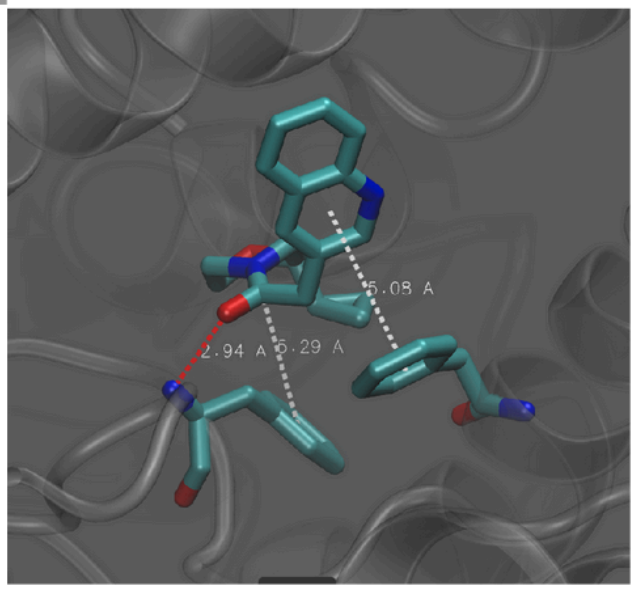


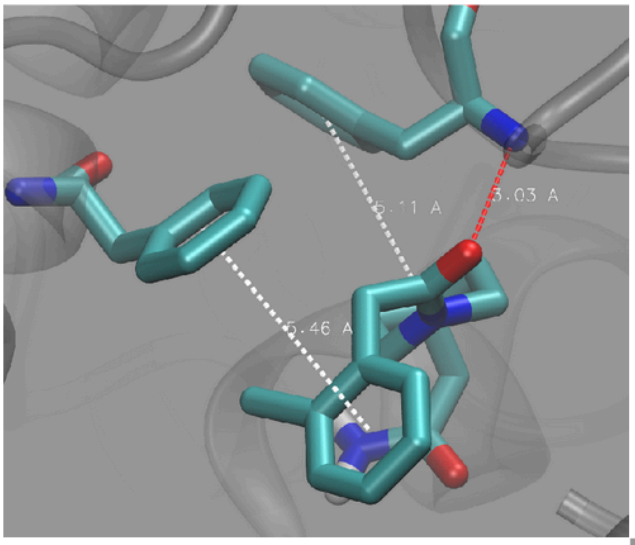
Ligando ***GARDP0240973***

Se identificaron dos interacciones aromáticas: una con **TYR 366**, y otra con **PHE 367**. La interacción con TYR 366 es particularmente relevante, ya que este residuo aparece en el 40 % de los casos positivos analizados, siendo una de las interacciones más frecuentes. La interacción con PHE 367 también contribuye a estabilizar la unión del ligando.

Ligando ***GARDP0238222***

Se observaron **tres interacciones, 1 π-π con PHE 51** y **dos interacciones adicionales con PHE 367 , siendo una de ellas π-π y la otra P.H.**. PHE 51 representa el residuo más recurrente entre todos los positivos analizados (60 % de aparición), lo que resalta su rol clave en el anclaje del ligando. Estas múltiples interacciones sugieren una alta afinidad en esta región del sitio activo.



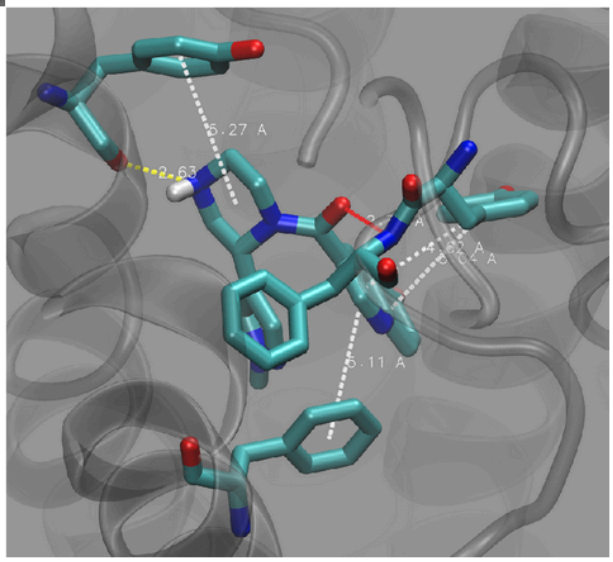


Ligando ***GARDP0240889***

Este ligando presenta **interacciones aromáticas múltiples**, principalmente con **PHE 367** y **PHE 51**, ambas observadas en el 20 % y 60 % de los casos respectivamente (Tabla 5). La conformación muestra un posicionamiento favorable que permite la formación simultánea de estas interacciones, lo cual podría contribuir significativamente a la estabilidad del complejo.

Ligando ***GARDP0245777***

se observaron **seis interacciones** distribuidas entre **cuatro residuos aromáticos claves**: **PHE 367**, **TYR 260**, **PHE 51** y **TYR 366**. Todos estos residuos han sido identificados como frecuentes en la Tabla 5, siendo PHE 51 y TYR 366 los más comunes (60 % y 40 %, respectivamente). La participación simultánea de estos cuatro puntos de anclaje sugiere una **afinidad elevada y una ocupación óptima del sitio activo**, destacando a este ligando como uno de los más prometedores del set analizado.



El análisis de las interacciones presentes en los candidatos seleccionados revela una recurrencia marcada de ciertos residuos del sitio activo, en particular la **PHE 51**, **TYR 366** y **PHE 367**, con apariciones en el 60 %, 40 % y 20 % de los casos respectivamente. Estas interacciones, mayormente del tipo **π–π stacking**, coinciden con los contactos observados en los **ligandos del grupo utilizados como control positivo**, reforzando su relevancia funcional. La reiterada participación de estos residuos sugiere que los mismos representan **elementos estructurales clave para el reconocimiento molecular**, y por tanto, constituyen **blancos prioritarios en el diseño racional de nuevos inhibidores** dirigidos contra este sitio activo.

3.8 Conclusión

En este capítulo se ha llevado a cabo una campaña de búsqueda virtual basada en docking, para identificar posibles ligandos de LolCDE, combinando enfoques de docking y dinámica molecular.

El sitio activo, se definió, a partir de la estructura cristalográfica en presencia del sustrato (en E Coli), y en función del docking realizado con controles positivos sobre un modelo de la estructura de la especie blanco donde se identificaron puntos clave de interacción, en particular el residuo TYR 366 como un sitio de unión recurrente.

El análisis de docking confirmó que la mayoría de los ligandos positivos evaluados mostraron afinidad por LolCDE, con diferencias en la estabilidad y frecuencia de interacción. La posterior validación mediante simulaciones DM permitió observar la persistencia de estas interacciones en un entorno dinámico, resaltando la importancia de las interacciones aromáticas en la estabilización de los ligandos dentro del sitio activo.

A partir de estos resultados, se estableció un protocolo para la pre-selección de compuestos basados en la similitud química, dando lugar a una biblioteca de 3,382 compuestos. Se implementó luego un proceso de selección por docking de alto rendimiento, utilizando criterios de afinidad energética, agrupamiento conformacional y proximidad al sitio de unión clave, lo que permitió reducir el conjunto a 100 candidatos óptimos para estudios futuros.

Estos hallazgos proporcionan una base sólida para la optimización y validación experimental de nuevos inhibidores de LolCDE, representando un avance significativo en la identificación de posibles fármacos dirigidos contra esta proteína y permitió establecer un diagrama de flujo de trabajo, el cual puede ser aplicado al estudio de otras proteínas, facilitando la identificación y validación de ligandos en diversos contextos biomoleculares. Este protocolo representa una herramienta valiosa para acelerar el descubrimiento de compuestos bioactivos en el campo de la biología estructural y el diseño de fármacos.

Integración de estrategias bioinformáticas para la predicción de sustratos de citocromos P450 bacterianos (BacCYPs)

4.1 Predicción de sustratos de citocromos P450 bacterianos

El objetivo del presente capítulo consistió en utilizar la metodología de docking como criterio para seleccionar sustratos potenciales para proteínas blancos, en este caso en particular trabajamos con los **citocromos bacterianos P450** (BacCYP) las misma son proteínas que contienen un grupo hemo, similares a sus contrapartes eucariotas, y realizan reacciones de oxidación en sustratos químicos complejos. Este proceso requiere oxígeno molecular y dos electrones. A diferencia de las eucariotas, las BacCYPs son proteínas solubles y participan en diversas rutas bioquímicas. Para alcanzar este objetivo, se implementó una estrategia de 3 etapas que integra diversos enfoques bioinformáticos:

1. **Análisis filogenético**: Se construyó un árbol filogenético basado en el análisis de secuencias de BacCYPs, lo que permitió clasificar estas enzimas en diferentes grupos y evaluar la diversidad de sustratos dentro de cada uno de estos grupos.
2. **Modelado estructural**: Se determinaron modelos de homología y predicciones generadas por AlphaFold para estimar la estructura tridimensional de las BacCYPs de estructura desconocida. AlphaFold resultó especialmente útil para modelar proteínas con baja identidad de secuencia respecto a estructuras conocidas.
3. **Docking molecular**: Se empleó para analizar las interacciones entre las BacCYPs y sus posibles sustratos.

4.2 Métodos

El presente estudio integró múltiples enfoques bioinformáticos para abordar la predicción de sustratos de citocromos P450 bacterianos (BacCYPs), combinando análisis de tipo filogenetico-evolutivo, modelado estructural y docking molecular.

- **Análisis filogenético y alineamiento múltiple de secuencias (MSA):** Se recopilaron más de 1800 secuencias de BacCYPs, incluyendo todas las estructuras cristalizadas disponibles hasta octubre de 2023. A partir de las mismas, se realizó un MSA utilizando **MAFFT-DASH**, incorporando información estructural para mejorar la alineación de homólogos lejanos. A partir de este alineamiento, se construyó un árbol filogenético y se generaron **Modelos Ocultos de Markov (HMM , *Hidden Markov Model*)** para identificar motivos conservados y clasificar las enzimas en 14 grupos monofiléticos.

- **Análisis de diversidad de sustratos:** Se recolectaron todos los sustratos conocidos de BacCYPs y se calculó la similitud química entre ellos utilizando el **índice de Tanimoto** (ver métodos generales). Con esta información, se construyó un dendrograma que permitió agrupar los sustratos en 10 grupos principales, revelando una alta diversidad

estructural.

- **Modelado estructural:** Se generaron modelos tridimensionales de diversos BacCYPs empleando dos estrategias. Por un lado, se aplicó **modelado por homología** clásico con plantillas (templates) de identidad variable para evaluar su impacto en la calidad del modelo, utilizando Modeller, y cuantificada mediante **RMSD-Cα** y **QMEAN**. Por otro lado, se utilizaron predicciones de **AlphaFold v2.0**, particularmente en proteínas sin estructura conocida, demostrando este método una mayor precisión general.

- **Docking molecular y bias docking:** Se realizaron más de 500 simulaciones de docking utilizando **AutoDockGPU**. Las grillas se generaron con **AutoGrid**, y se aplicaron bias en aquellos casos donde existía información previa de interacciones críticas (por ejemplo con el Hierro del grupo hemo), incorporando estos puntos como restricciones al algoritmo de docking.

- **Re-docking y validación:** Se evaluó la capacidad de los modelos para reproducir las poses cristalinas originales mediante re-docking. Los resultados se validaron considerando valores de **energía de unión**, **población de clúster** y **RMSD** entre la pose predicha y la experimental utilizada como referencia.

- **Filtrado y análisis cuantitativo:** Se descartaron poses con energías positivas o poblaciones bajas (<30%). Se aplicaron **Z-scores** para normalizar los valores de energía y población, permitiendo una comparación más robusta entre diferentes compuestos y receptores. Este análisis facilitó la identificación de ligandos positivos y permitió estudiar la **selectividad o promiscuidad** de cada BacCYP.

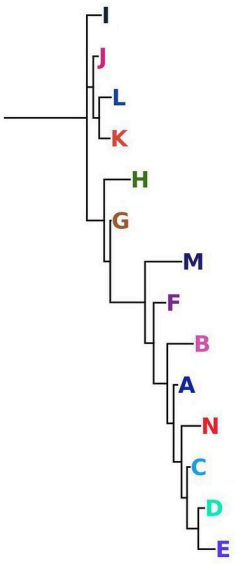
4.3 Diversidad de secuencias y Análisis filogenético

El análisis filogenético es una herramienta clave en biología evolutiva que permite estudiar las relaciones entre diferentes organismos o secuencias. En este estudio, se realizó un análisis exhaustivo de la diversidad de secuencias de BacCYPs mediante un Alineamiento Múltiple de Secuencias (MSA), abarcando un conjunto representativo de 1809 secuencias, denominado conjunto 1. Este conjunto incluye todo el espectro conocido de BacCYPs, e incorpora las 202 estructuras cristalizadas hasta la fecha de Octubre 2023.

Para realizar el MSA, se utilizó **MAFFT** [58], una herramienta ampliamente reconocida por su precisión y eficiencia. MAFFT emplea guías jerárquicas y refinamiento iterativo, lo que permite manejar grandes volúmenes de datos con robustez. Además, su funcionalidad opcional para integrar información estructural mejora la alineación en el caso de homólogos lejanos. Una extensión de esta herramienta, **MAFFT-DASH** [57], fue utilizada en el estudio para incorporar restricciones estructurales derivadas de la base de datos DASH, logrando un alineamiento más preciso mediante la combinación de información estructural y secuencial, especialmente en secuencias con baja similitud.

Como resultado del MSA, las BacCYPs se clasificaron en 14 grupos monofiléticos, de los cuales 11 cuentan con al menos una estructura cristalizada representativa. Posteriormente, se generaron **Modelos Ocultos de Markov (HMM, por sus siglas en inglés)** para cada uno de estos grupos. Los HMM son herramientas probabilísticas derivadas de alineamientos múltiples que asignan estados representativos y probabilidades específicas a cada posición del alineamiento[59] resultan particularmente útiles para identificar dominios funcionales en nuevas secuencias y predecir funcionalidades basándose únicamente en su composición aminoacídica, incluso en ausencia de información adicional.Todo este análisis y sus resultados se resumen en la ***Figura 21*** panel A. En relación con la disponibilidad de estructuras cristalizadas de referencia, el análisis del árbol filogenético muestra que la mayoría de los principales grupos de BacCYPs tienen al menos una estructura cristalizada representativa. El MSA y los HMMs evidencian aquellos motivos de secuencia que están conservados en todos los BacCYPs (ejemplos representativos se muestran en la ***Figura 21*** panel B, así como los sitios y residuos específicos de cada grupo.

A



	Secuencias de BacCyps (conjunto 1)	BacCyps con estructura cristalizada	BacCyps con sustratos anotados
I	210 (11.6 %)	30	18
J	126 (7.0 %)	41	43
L	23 (1.3 %)	7	16
K	36 (2 %)	18	15
H	38 (2.1 %)	7	6
G	113 (6.2 %)	5	3
M	50 (2.8 %)	5	3
F	139 (7.7 %)	4	1
B	87 (4.8 %)	4	3
A	218 (12.0 %)	3	5
N	63 (3.5 %)	0	0
C	213 (11.8 %)	0	0
D	168 (9.3 %)	3	36
E	127 (7.0 %)	0	3
	198 (10.9 %)	6	5
	1809	133	157

B

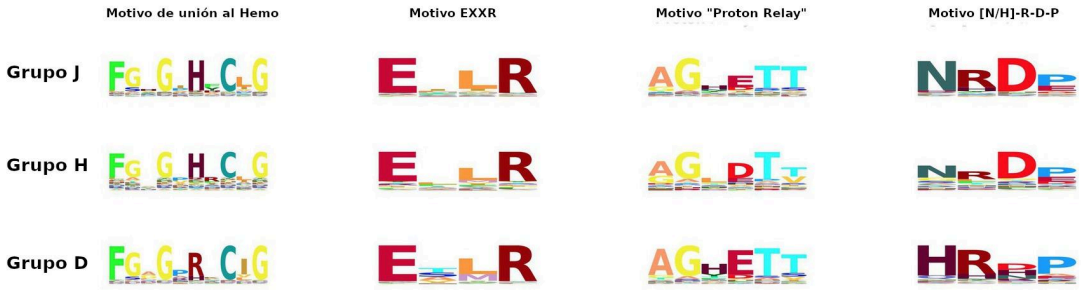


Figura 21 *Panel A: Árbol filogenético de los principales grupos de BacCYPs, mostrando el número de proteínas con estructuras cristalizadas y sustratos anotados. Panel B: Logos HMM de los grupos J, H y D, destacando variaciones clave en los motivos de unión al hemo, EXXR, "Proton Relay Motif" y [N-H]-R-D-P.*

El dominio BacCYPs exhibe una notable variación en su longitud, con un promedio de 403 residuos. Dentro de este dominio, se identifican 33 posiciones conservadas que ofrecen un alto valor informativo. Entre estas destacan:

1. El motivo de unión al hemo (posiciones 364-373) con la secuencia F-[G/S]-X-G-X-[H/R]-X-C-X-G, donde la última posición (G) se sustituye por A en el grupo N.
2. El motivo EXXR localizado en las posiciones 292-295.
3. El "Proton Relay Motif" ubicado entre las posiciones 236-241, con la forma [A/G]-G-X-[D/E]-T-[T/S] [60]
4. El motivo [N/H]-R-D-P en las posiciones 332-335, que distingue a los grupos A, B, C, D, E, F, M y N (principalmente con histidina) de otros grupos, donde prevalece una asparagina.

El sitio activo de BacCYP (consideramos que todos los aminoácidos de la cadena que están involucrados en interacciones con ligandos en cristales de PDB corresponden al sitio activo) contiene aproximadamente 24 residuos. Una vez caracterizada la variabilidad del sitio activo se procedió a analizar la diversidad en los sustratos (o ligandos) de los BacCYP.

4.4 Diversidad de sustratos

El siguiente paso consistió en analizar la diversidad de los sustratos conocidos para los BacCYPs, para ellos se construyó un dendrograma basado en la similitud química determinada con el índice de tanimoto (explicado en métodos) utilizando todos los sustratos conocidos, los cuales se pudieron agrupar en 10 grupos, como se muestra en la ***Figura 22***

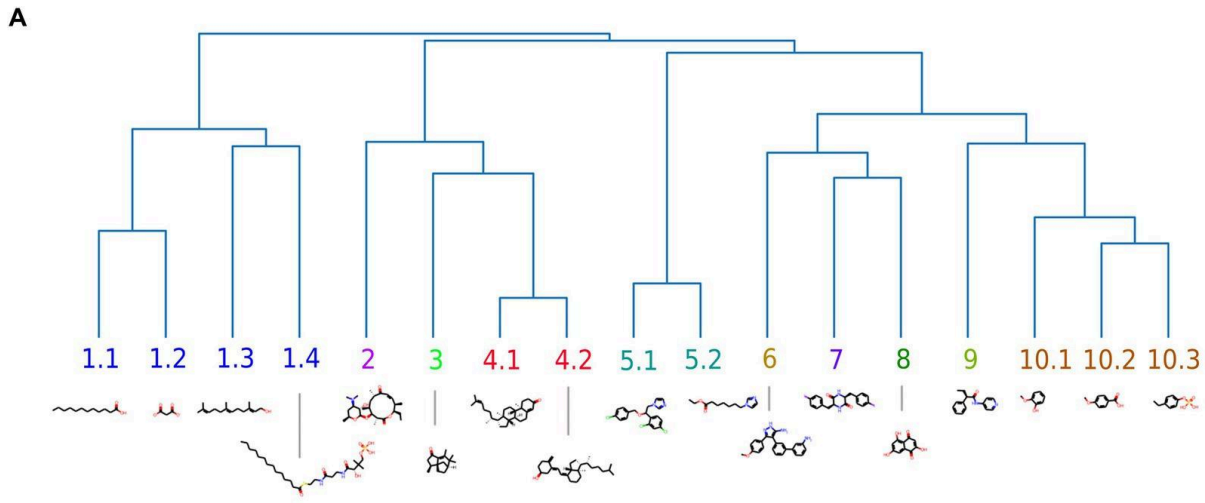


Figura 22 *Dendrograma que representa los 10 grupos principales de compuestos que interactúan con las BacCYPs, junto con sus subgrupos correspondientes, incluyendo un ejemplo ilustrativo de cada grupo para visualizar las relaciones de similitud fisicoquímica.*

Entre los grupos principales, se distinguen claramente a:

- Compuestos tipo esteroles (grupo 4).
- Ácidos grasos (grupo 1).
- Moléculas relacionadas con el alcanfor (grupo 3).
- Macrólidos (grupo 2).

Por otro lado, otros grupos (5, 6, 7 y 9) contienen estructuras más complejas, caracterizadas por la presencia de múltiples anillos aromáticos. Mientras que el grupo 10 parece unir sustratos más pequeños.

Una vez obtenidos tanto el conjunto de estructuras como el de sustratos, se evaluó si existía alguna relación entre los grupos de citocromos y los sustratos definidos por los respectivos árboles/dendrogramas, es decir, si algún grupo filogenético de BacCYPs mostraba preferencia por un grupo específico de ligandos. Para ello, se calculó la distribución de Tanimoto para ***todos los pares de ligandos*** de la base de datos (Set Todas las Proteínas) y para aquellos dentro de ***cada grupo filogenético*** (Set Por Grupo filogenético). Los resultados en la ***Figura 23*** muestran que no existen diferencias significativas entre ambos grupos, lo que sugiere que no hay una preferencia particular de los receptores por ningún grupo específico de ligandos.

La notable similitud entre ambas distribuciones indica que un análisis filogenético amplio, que abarque todo el universo de BacCYPs, no es suficiente por sí solo para ofrecer información consistente que permita predecir la especificidad de los sustratos hacia proteínas de un grupo específico.

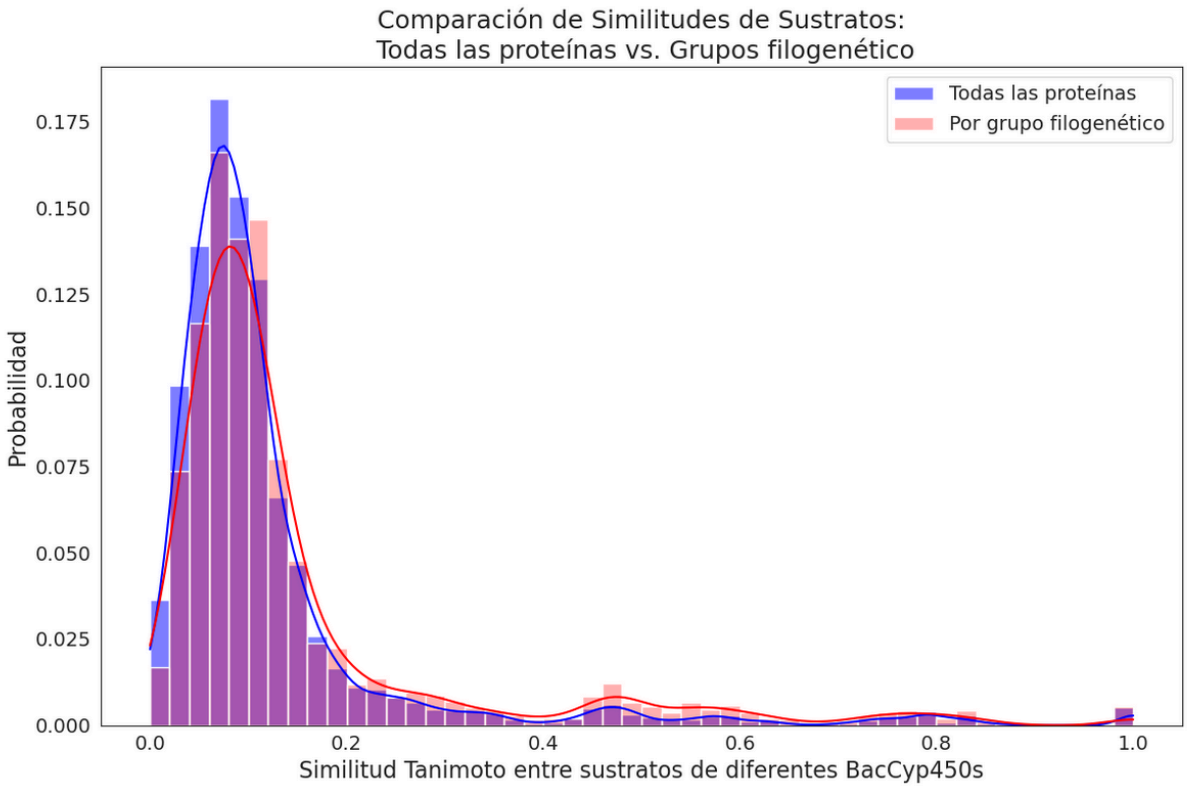


Figura 23 Gráficos de barras que muestran la similitud basada en el Índice de Tanimoto entre los SPIs que interactúan con distintos subconjuntos de proteínas: en azul, aquellos que se unen a todos los BacCYPs anotados; en rojo, los que se asocian con BacCYPs de un mismo grupo filogenético.

No obstante, al analizar los sustratos de los BacCYPs con una identidad de secuencia superior al 70%, se observan picos característicos en los valores del Índice de Tanimoto (TI) de 0.5, 0.8 y cercanos a 1, superiores a los que se observan en la Figura 23. Estos hallazgos indican que, aunque no es sencillo predecir el sustrato de un BacCYP basándose únicamente en su grupo filogenético, la identificación de BacCYPs con alta identidad de secuencia y sustratos conocidos puede ofrecer información valiosa sobre los sustratos de proteínas aún no caracterizadas.

4.5 Modelado Estructural

En esta sección nos propusimos evaluar distintos enfoques para la generación de modelos estructurales de proteínas. En particular, se buscó analizar la calidad y precisión de los modelos como una función de la diversidad de secuencias. Este estudio es crucial para determinar cuán confiables son las estructuras generadas, especialmente al considerar aplicaciones posteriores como en este caso el docking de sustratos

Para lograr este objetivo, se emplearon dos enfoques principales: el modelado basado en homología y el uso de AlphaFold. En el caso del modelado por homología, seleccionamos 12 estructuras como casos de prueba. Para cada una, se construyeron 9 modelos utilizando plantillas con diferentes niveles de identidad de secuencia: alta identidad (plantillas de la misma subclase), identidad moderada (plantillas del mismo grupo), y baja identidad (plantillas de grupos distintos). Estos modelos fueron evaluados comparando las estructuras resultantes con las correspondientes estructuras reales, considerando métricas como el CA-RMSD y el puntaje QMEAN. El CA-RMSD (Root Mean Square Deviation de átomos de carbono alfa) mide la desviación promedio entre los átomos equivalentes de las estructuras modeladas y experimentales, proporcionando una indicación directa de la similitud estructural. Por su parte, el puntaje QMEAN evalúa la calidad global de un modelo estructural en comparación con estructuras experimentales conocidas, combinando diferentes parámetros estadísticos. Estas métricas son esenciales para cuantificar la precisión y confiabilidad de los modelos generados.

Por otro lado, AlphaFold fue empleado para generar 20 estructuras adicionales. Para garantizar la imparcialidad de la evaluación, seleccionamos estructuras depositadas en el Protein Data Bank (PDB) después de la fecha límite utilizada para el entrenamiento del modelo AlphaFold. Esto aseguró que las estructuras analizadas no estuvieran incluidas en los datos de entrenamiento del modelo, permitiendo una evaluación objetiva de su capacidad predictiva.

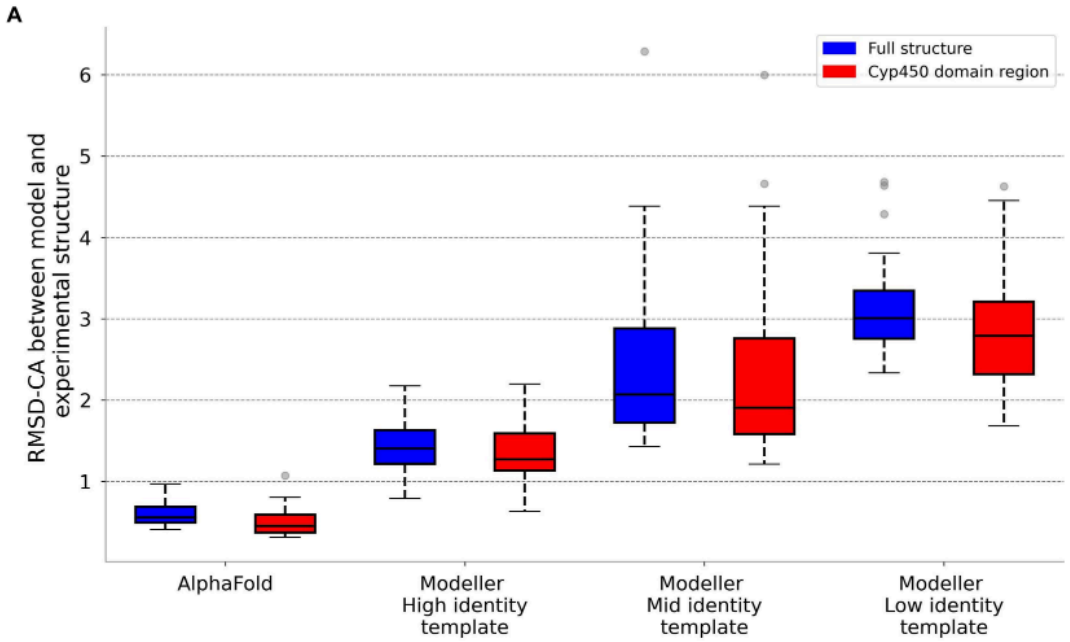


Figura 24 *Evaluación comparativa de la calidad entre estructuras de referencia y aquellas modeladas mediante AlphaFold o Modeller. Se presentan los valores de RMSD-Cα, calculados tanto para la proteína completa como para la región específica que corresponde al dominio BacCYP.*

La ***Figura 24*** ilustra la comparación de la calidad de los modelos estructurales generados mediante dos enfoques: **AlphaFold** y el **modelado basado en homología** utilizando plantillas con distintos niveles de identidad de secuencia, los resultados obtenidos. En la mismas se observa que los modelos generados por AlphaFold destacan por su notable superioridad frente a los obtenidos mediante modelado por homología, incluso superando a los que emplean plantillas de alta identidad.

Respecto a los modelos por homología, se evidencia que los modelos construidos con plantillas de alta identidad logran, en más del 90% de los casos, un RMSD-Cα inferior a 2 Å respecto a la estructura experimental, lo que indica una predicción de alta calidad [61]. Sin embargo, cuando se utilizan plantillas de identidad moderada, la precisión de los modelos disminuye, alcanzando un RMSD-Cα promedio cercano a 2.5 Å en comparación con las estructuras cristalizadas. Por otro lado, los modelos basados en plantillas de baja identidad presentan valores de RMSD-Cα más altos, lo que los hace menos confiables para análisis posteriores.

Estos hallazgos destacan el potencial de AlphaFold como una herramienta robusta para la predicción de estructuras proteicas y sientan las bases para explorar su aplicación en tareas avanzadas como el Docking Molecular.

4.6 Docking : Construcción del set de datos

Para analizar si el docking molecular podría emplearse para identificar los posibles sustratos de un BacCYP específico, se evaluó el rendimiento en la predicción de poses de unión, y la selectividad del Autodock-bias utilizando un conjunto representativo de BacCYPs frente a compuestos pertenecientes a los diversos grupos de ligandos.

Para ello primero se construyó primero un set de datos de validación. Se seleccionaron 15 estructuras de proteínas de BacCYP, cada una de las cuales se une a un ligando de un grupo diferente, denominadas Proteínas Unidas al Grupo de Ligandos Representativos (RLP). Es importante notar que no se obtuvieron cristales adecuados para los subgrupos 1.4 y 6, impidiendo su evaluación, todo esto se encuentra descrito en la ***Tabla 7***.

Proteína Uniprot ID	Estructura representativa PDB	Grupo filogenético Proteína	Grupo de ligando que une la proteína
O31440	1IZO	F	1.1
Q8VQF6	3BDZ	J	1.2
P9WPP3	6T0J	J	1.3
Q9KIZ4	1QD5	J	2

P00183**	1AKD	H	3
Q06069	5IKI	J	4.1
P18326	3CV9	K	4.2
P00183**	1PHA	K	5.1
P9WPN8	5LI8	J	5.2
P9WPP7	3G5H	J	7
Q9KZF5	2NZ5	J	8
P9WPP9	2CI0	E	9
A0A076MY51	5OMU	I	10.1
Q6N8N2	4DNJ	J	10.2
Q8NSW2	5GWE	J	10.3

Tabla 7 *Proteínas seleccionadas para las evaluaciones de docking molecular. También se indica el subgrupo específico de ligandos con el que interactúan, y la estructura cristalográfica representativa registrada en el PDB.*
***Se aclara que la proteína P00183 se une tanto al grupo 3 como al grupo 5.1.*

Para evaluar el método se armó un set de 34 ligandos de pruebas, dos por cada grupo de sustratos excepto el grupo 5.1 con tres ligandos, y el 1.4 con solo uno)(***Tabla 8***). Los sustratos implicados en las actividades catalíticas de las BacCYPs se obtuvieron de la base de datos Rhea [64]. Los ligandos unidos a BacCYPs en estructuras cristalizadas se recopilaron del Protein Data Bank (PDB) y se filtraron para conservar únicamente aquellos biológicamente relevantes utilizando la base de datos MOAD. Además, los compuestos con actividad demostrada en ensayos de unión se extrajeron de ChEMBL (versión 32) [63], seleccionando exclusivamente ligandos con un valor Pchembl superior a 6, como medida de su afinidad [62].

Ligando	Grupo	Ligando	Grupo
9FL	2	CHEMBL_91	5.1
CHEMBL_16089	2	FJQ	5.1
CAH	3	PFZ	5.1
CAM	3	JZ3	10.1
CHEMBL_3804971	6	V55	10.1
CHEMBL_3804159	6	MLA	1.2
CHEMBL_4467901	6	MLI	1.2
YTT	7	7ZU	4.2
FLV	8	VDX	4.2
NQ	8	6XD	5.2
1CM	9	M65	5.2
CII	9	FIV	10.2
CHEMBL_30807	1.1	TWO	10.2
VGJ	1.1	CHEMBL_83951	1.3
CHEMBL_112570	4.1	DXJ	1.3
CHEMBL_2048331	4.1	88L	10.3
ZMO	1.4	FW6	10.3

Tabla 8 *Set ligandos de prueba, el mismo fue armado de manera de representar equilibradamente cada grupo*

Este enfoque integral garantiza una selección robusta y relevante de sustratos para el estudio de las BacCYPs

4.7 Re-Dockings

Una vez armado el set de datos, como primer paso se evaluaron casos de receptores (BAC CYPs específicos) con ligandos cristalizados, para evaluar la viabilidad de los dockings y analizar qué valores podemos esperar de energía y población. De los 15 grupos de citocromos analizados en 13 encontramos energías bajas con un valor promedio de -7.7 kcal/mol y poblaciones mayores al 87%, lo que a priori sugiere estamos hablando de buenos resultados de Docking (***Figura 25***).

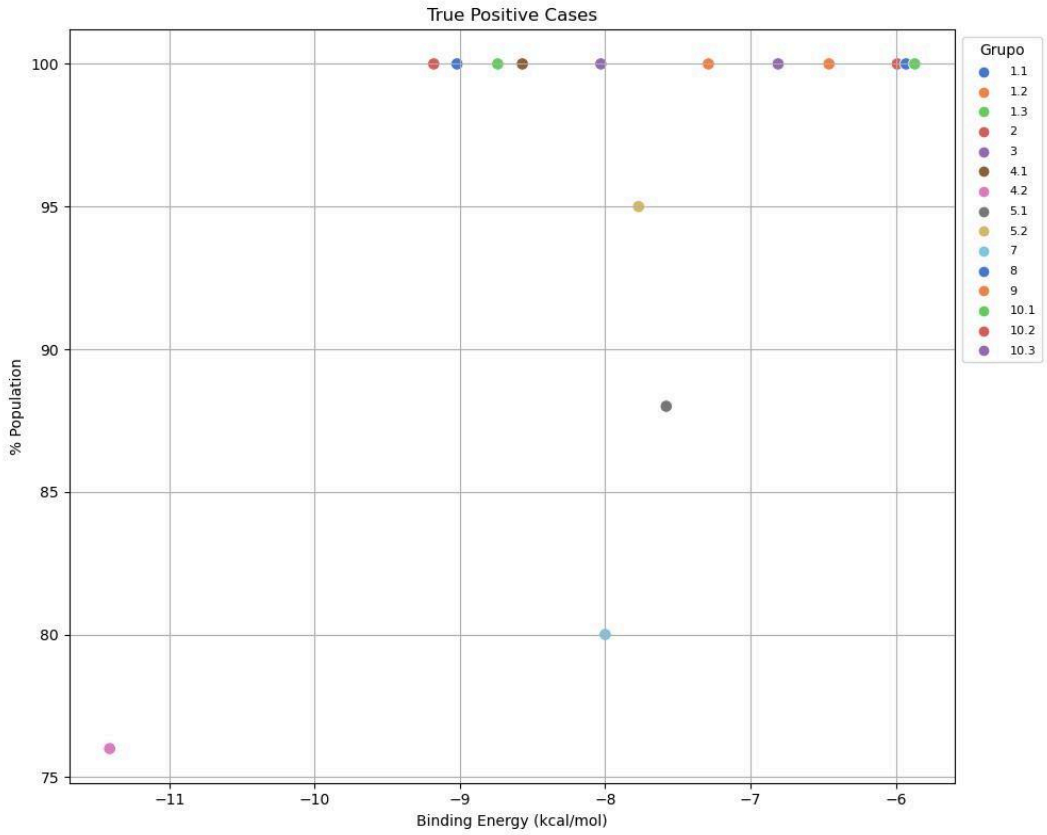


Figura 25 *Gráfico de población en función de la energía de unión (binding energy) , en el mismo se observa que en la mayoría de los grupos los re-docking son correctos dado que presentan alta población y energías bajas.*

Se utilizó como criterio adicional para analizar el resultado de la predicción de pose, el valor de rmsd del sustrato respecto al cristal siendo los siguientes valores:

- RMSD < 2 Å: Un valor de RMSD inferior a 2 Å se considera generalmente como una buena superposición entre la pose predicha y la conformación observada experimentalmente
- RMSD entre 2 y 4 Å: Indica una superposición moderada, donde la molécula se coloca en el sitio de unión correcto, pero su orientación y conformación pueden no ser totalmente precisas
- RMSD > 4 Å: Se considera una mala superposición, lo que sugiere que el algoritmo de docking no ha logrado predecir con precisión el modo de unión.

En la ***Tabla 9*** se encuentra el resumen de los resultados de RMSD de los dockings. En líneas generales se obtuvieron buenos resultados (con un promedio de 2.227), a excepción del grupo 7 y 8. En el caso del ligando del grupo 7, se encuentra la pose pero en un cluster de menor población, mientras que en el caso del ligando del grupo 8, este es muy pequeño y está en la misma zona que el cristal pero rotado.

Proteína	Grupo	RMSD	Proteína	Grupo	RMSD
1AKD	3	2.18	1IZO	1.1	2.62
1PHA	5.1	3.45	1Q5D	2	0.59
2CI0	9	3.05	2NZ6	8	4.60
3DBZ	1.2	1.85	3CV9	4.2	1.36
3G5H	7	4.75	4DNJ	10.2	0.90
5GWE	10.3	0.35	5IKI	4.1	2.63
5LIB	5.2	3.21	5MOU	10.1	0.57
6TOJ	1.3	1.29			

Tabla 9 *Resumen de los valores de RMSD obtenidos para los re-dockings de cada ligando con su receptor.*

4.8 Docking y Bias Docking

Una vez validados los sistemas, se procedió a realizar los ensayos de docking cruzado. Para ello se seleccionó un conjunto de ligandos que incluyó, como mínimo, dos compuestos de cada grupo de sustratos (*a excepción del grupo 1.4*). De cada grupo de receptores, se eligió un caso representativo, correspondiente a una proteína cristalizada del PDB con capacidad conocida para unirse al grupo, y se llevaron a cabo dockings con todo el conjunto de ligandos (***Tabla 10***).

Las poses predichas de los 510 dockings se caracterizaron considerando tanto su energía de unión estimada como su población, normalizadas mediante Z-scores, lo que permitió un análisis cuantitativo detallado de las interacciones entre las proteínas y los posibles ligandos.

Grupo Ligando	Receptor	Ligandos	Grupo Ligando	Receptor	Ligandos
1.1	1IZO	VGJ CHEBI_30807 PAM	5.1	1PHA	CHEMBL_91 FJQ PFZ
1.2	3DBZ	MLI MLA	5.2	5LI7	M65 6XD
1.3	6TOJ	CHEBI_83951 DXJ RWZZ	7	3G5H	YTT CHEMBL_44 7901
1.4	-	ZMP ZMO	8	2NZ5	NQ FLV 226
2	1Q5E	EPB CHEBI_16089 9LF	9	2CI0	1CM CII
3	1AKD	CAM CAH	10.1	5OMU	V55 JZ3 3DM
4.1	5IKI	CHEMBL_112570 CHEMBLI_204833 A9H	10.2	4DNJ	ANN FIV TWO

Tabla 10 *Ligandos representativos de cada grupo , para realizar dockings todos contra todos*

Este procedimiento permitió evaluar de manera sistemática las interacciones proteína-ligando, considerando tanto la diversidad estructural de los compuestos, como las propiedades específicas de las proteínas representativas.

Con respecto al bias, para determinar los sitios más adecuados para introducir estos sesgos, se analizaron interacciones críticas entre BacCYPs y sus sustratos en complejos conocidos. En particular, se implementó un sesgo para los ligandos que coordinan el grupo hemo, junto con hasta tres sesgos adicionales basados en las interacciones observadas, para esto se utilizó el script de interacciones desarrollado en el transcurso de esta tesis. Este protocolo permitió mejorar la precisión de las predicciones del docking al incorporar información previa sobre la química y geometría de las interacciones clave (***Tabla 11***).

Proteína Uniprot ID	Estructura	Cadena	Residuo	Interacción
O31440	2ZQJ	A	ARG 242	Aceptor
Q8VQF6	3BDZ	-	-	-
P9WPP3	6T0J	-	-	-
Q9KIZ4	1Q5D	A	PHE 96	AROMÁTICA
		A	ALA 180	ACEPTOR
		A	GLY 304	ACEPTOR
P00183	1PHC	A	PHE 87	AROMÁTICA
		A	TYR 96	DONOR
Q06069	5XNT	A	THR 89	DONOR
P18326	3CV9	-	-	-
P9WPP7	3CXX	A	GLN 385	ACEPTOR
		A	THR 77	DONOR
P9WPN8	5LI8	-	-	-
Q9KZF5	2NZA	A	HIS 290	DONOR
		A	ARG 291	ACEPTOR
P9WPP9	2BZ9	A	TYR 76	AROMÁTICA
		A	MET 433	ACEPTOR
A0A076MY51	5OMU	-	-	-
Q6N8N2	4DNJ	A	SER 97	DONOR
		A	SER 247	ACEPTOR
Q8NSW2	5GWE	A	SER 136	DONOR

A	ARG 221	ACEPTOR
A	SER 288	DONOR

Tabla 11 *Puntos de bias seleccionado para cada caso , en los casos que no se encontraron interacciones de interés se realizó docking normal.*

4.9 Filtrado y análisis de los datos de docking

Tras completar los 525 *dockings* realizados con y sin sesgo (*bias*), se procedió al análisis detallado de los resultados. Cada pose predicha fue evaluada en función de su energía de unión estimada y su población, ambas normalizadas mediante puntajes Z (*Z-scores*). De acuerdo con trabajos previos de nuestro grupo, el análisis combinado de estos parámetros para cada par proteína-ligando permite identificar como positivos (es decir, posibles ligandos) aquellas poses consideradas atípicas. Estas se localizan típicamente en el cuadrante superior izquierdo de los gráficos 2D, donde destacan por su energía de unión negativa y una población significativamente elevada (***Figura 26***). Como era de esperarse, en esa zona se encuentran las poses predichas para nuestras muestras de referencia (casos de Verdaderos Positivos). Además, con los valores bajos de RMSD, cuyo promedio es de 2.227 (como se mencionó anteriormente), se puede concluir que estas predicciones corresponden a una pose potencialmente correcta.

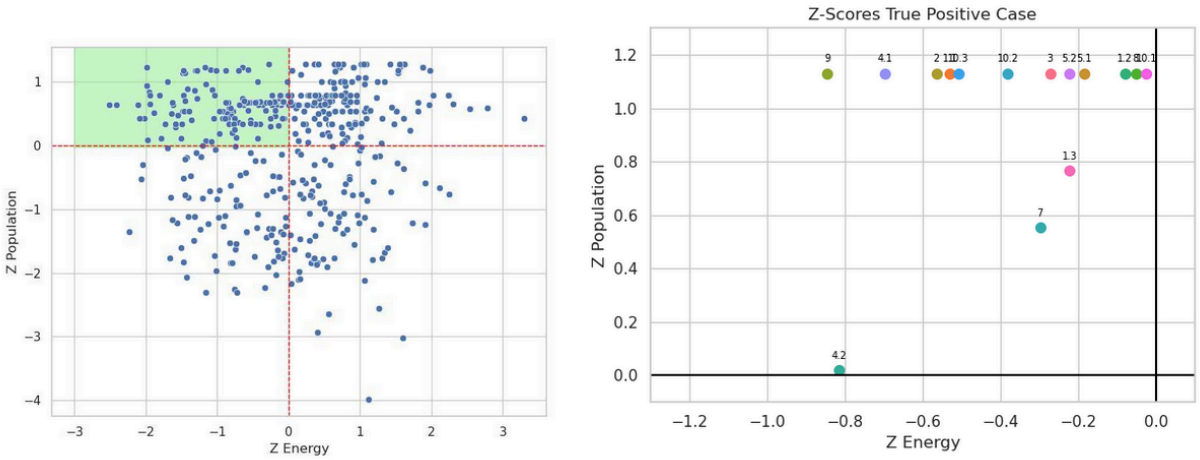


Figura 26: Gráfico de Población frente a Energía de Unión (puntajes Z) para el *docking* de ligandos. La zona verde representa el área que define los posibles ligandos.

Inicialmente, se realizaron varios filtros para depurar los datos obtenidos. En primer lugar, se eliminaron todas las poses con energías positivas, ya que estas incrementan significativamente el promedio, lo que provocaba que los valores negativos quedarán muy agrupados al calcular los puntajes Z (*Z-scores*). Además, las energías positivas carecen de sentido biológico al no representar interacciones favorables y suelen estar relacionadas con sobreposición de átomos en las estructuras correspondientes.

Posteriormente, para reducir el ruido generado por poses de baja calidad, se descartaron los clústeres con menos del 30% de la población total. Esto se justificó al observar casos con energías muy bajas pero también poblaciones pequeñas, lo que introducía variabilidad no representativa en los resultados. Luego de estos filtros, se probaron distintas combinaciones de Z score (ver en métodos) para analizar su impacto en los gráficos, y evaluar cuál de ellas permitía una mejor separación de los ligandos. De las distintas combinaciones de Z-scores pudimos observar que la que mejor separaba los casos en líneas generales es el Z Poblacion vs Z de energía , el Z-score normalizado puede mejorar en algunos casos puntuales cuando el ligando es muy pequeño y con muchos puntos de interacción , como se observa en los plots de la ***Figura 27***. Con estos resultados decidimos continuar con la clasificación de los Z-Scores “normales”, dado que la mayoría de los ligandos utilizados son de tamaños similares (un promedio de 30 átomos por ligando)

Con el fin de evaluar la selectividad de los BacCYP, consideramos como ligandos positivos a aquellos cuyo puntaje Z es superior a 0.9 del *Z-score* de los verdaderos positivos. En la ***Figura 28*** se ilustra la cantidad de grupos de ligandos que se predice que cada BacCYP puede unir. Cada columna indica cuántos grupos diferentes de ligandos puede unir un dado BacCYP. Los resultados indican que algunos receptores presentan mayor promiscuidad que otros. Tres de ellos se consideran ***altamente selectivos***, ya que se predice que solo se unen a los ligandos de prueba del mismo grupo que su compuesto de “verdadero” (K 4.2 , E 9 y J 10.3). En contraste, otros cuatro grupos de receptores muestran que los ligandos positivos provienen de 2 grupos distintos. Por otro lado, tres tipos de receptores se espera que sean considerablemente promiscuos, ya que presentan resultados favorables de *docking* con ligandos de prueba provenientes de hasta seis grupos diferentes.

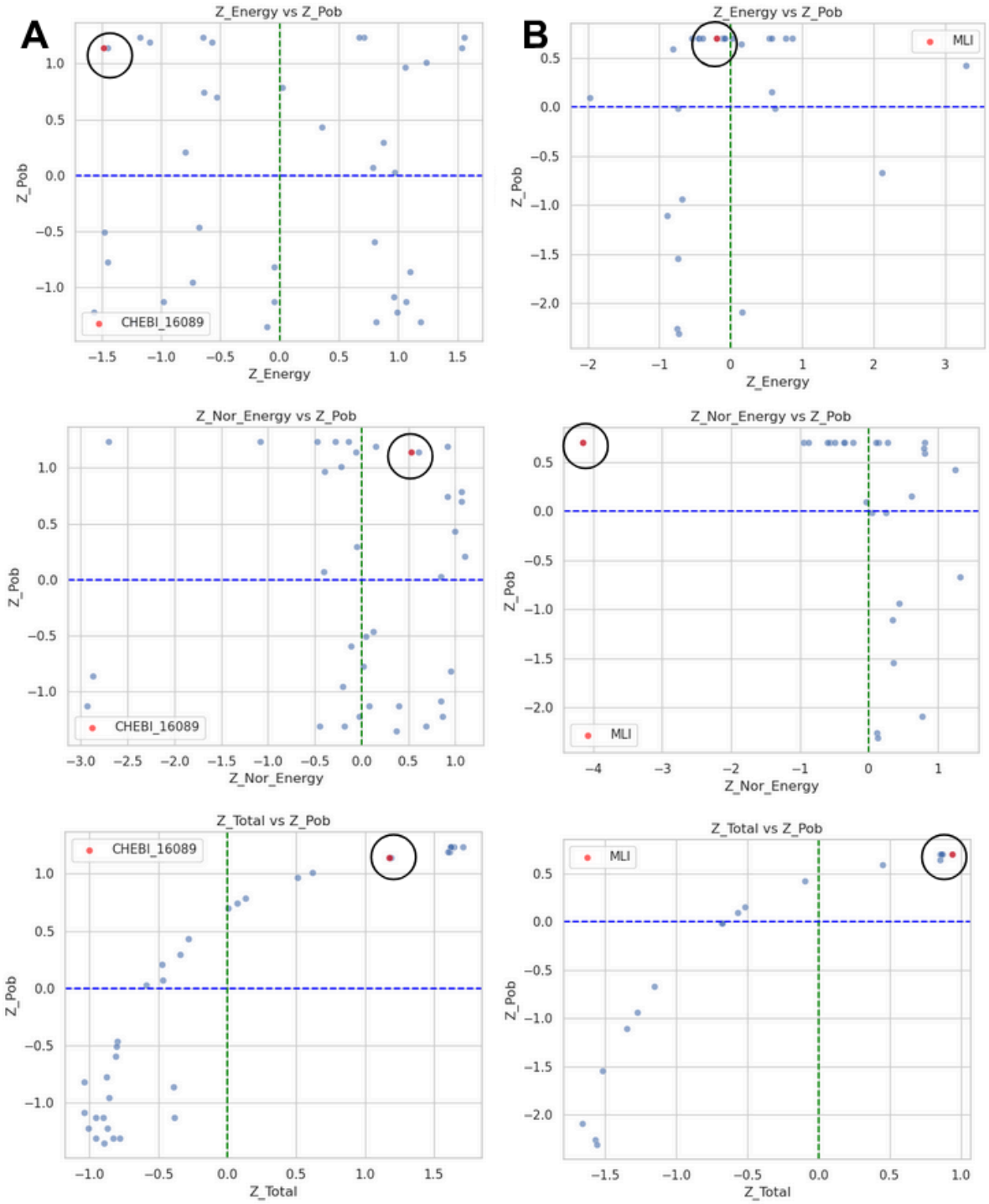


Figura 27 Se observan distintas combinaciones de los valores de Z para 2 proteínas distintas, la zona encerrada en un círculo negro corresponde al caso positivo (TP) de la proteína analizada. En el panel A (Izquierda) un ejemplo donde mejor separa el caso TP de los Z sin ningún tipo de modificación, en el panel B (Derecha) uno donde se observa una mejor separación en los casos donde se normaliza en función del tamaño del ligando, para la mayoría de los casos evaluados las distribuciones corresponden al panel A, por lo que se continuó el análisis usando el valor de Z_{energy} vs Z_{pob}.

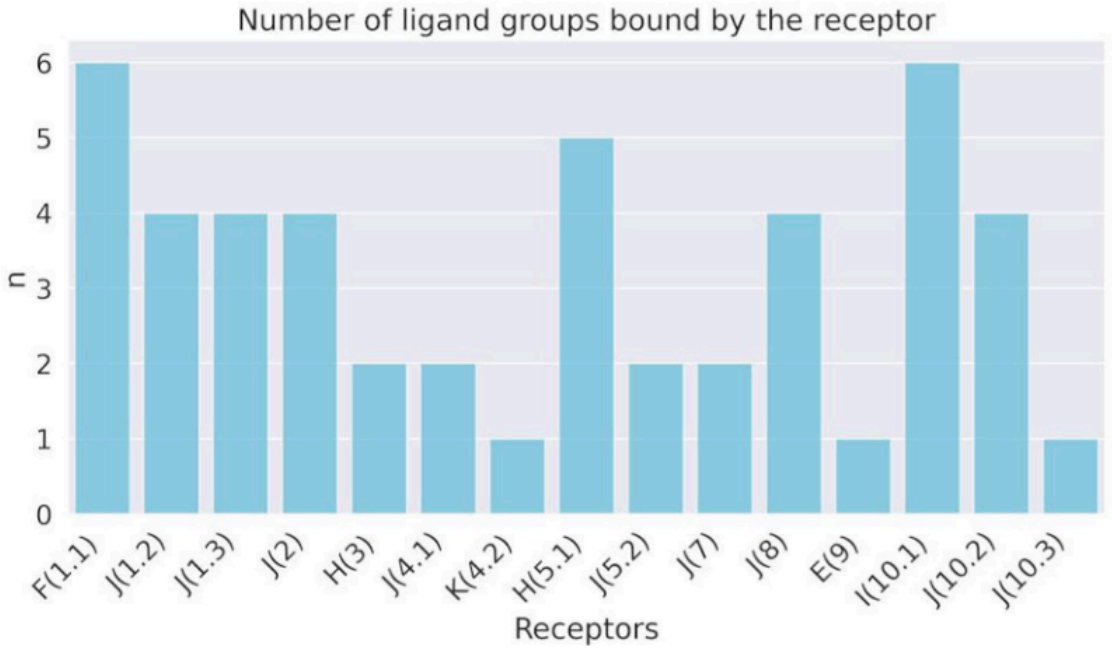


Figura 28 Distribución del número de 17 tipos diferentes de ligandos predichos que se unen a cada tipo definido filogenéticamente de BacCYP. Cada Columna indica el número de grupos que une el receptor.

El análisis detallado ***Figura 29*** revela, por ejemplo, que los ***RLPs*** promiscuos de los grupos filogenéticos F y H se predice que se unen a ligandos de prueba considerablemente grandes, como un ácido graso de cadena larga de 62 átomos, o un compuesto con tres anillos aromáticos. En el caso del RLP del grupo I, cuyo ligando de unión real (grupo 10.1) es un ligando con un anillo aromático y sustituyentes polares, los ensayos de *docking* sugieren una preferencia por ligandos de grupos cercanos.

Desde la perspectiva centrada en los ligandos, aquellos considerados promiscuos provienen de los grupos 3, 8 y 9. Los dos primeros son ligandos de tamaño mediano a pequeño, con uno o dos anillos aromáticos y grupos funcionales polares. En conjunto, nuestros resultados de *docking* indican que, si está disponible un sustrato potencial, el *docking* podría determinar si puede unirse o no, y, en ausencia de otra información, el docking puede proporcionar un filtro de selectividad que ayude a reducir las posibles opciones de sustratos.

Receptors \ Ligands	1.1	1.2	1.3	1.4	2	3	4.1	4.2	5.1	5.2	6	7	8	9	10.1	10.2	10.3	Total
F(1.1)	1	0	0	0	0	1	0	0	1	2	1	0	0	2	0	0	0	8
J(1.2)	0	1	1	0	0	2	0	0	0	0	0	0	1	0	0	0	0	5
J(1.3)	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	5
J(2)	0	0	0	0	1	1	0	1	0	0	0	0	0	2	0	0	0	5
H(3)	0	0	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	3
J(4.1)	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	2
K(4.2)	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
H(5.1)	0	0	1	0	0	1	0	0	1	0	0	1	2	0	0	0	0	6
J(5.2)	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	3
J(7)	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	2
J(8)	0	0	1	0	1	1	0	0	0	0	0	0	0	1	0	0	0	4
E(9)	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2
I(10.1)	0	0	1	0	0	1	0	0	0	0	0	0	1	0	2	1	1	7
J(10.2)	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	2	5
J(10.3)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
Total	2	2	4	0	2	9	3	4	2	5	1	2	6	7	2	4	5	

Figura 29 *Matriz de resultados de binding. Las filas corresponden a proteínas representativas unidas a grupos de ligandos, indicando entre paréntesis el grupo de ligando conocido que se une a la proteína correspondiente. Las columnas representan los diferentes grupos de tipos de ligandos. Los números indican la cantidad de ligandos de cada grupo que se predice que se unen al BacCYP correspondiente según los resultados de docking.*

Finalmente, la ***Figura 30*** presenta un ejemplo visual de la especificidad de BacCYP. En el panel izquierdo se muestra un BacCYP del grupo H (PDB ID 1AKD) con su ligando natural (CAM del grupo 3) en azul. Este ligando es relativamente pequeño y forma un fuerte enlace de hidrógeno con TYR 96. En contraste, un ligando más grande, VDX (grupo 10.3), se muestra en rojo y claramente no encaja en el sitio activo, chocando con TYR 96. En el panel derecho se ofrece una comparación visual directa, donde se muestra el receptor natural de VDX del grupo J, con ambos ligandos representados con el mismo código de colores. Para unirse a VDX, este BacCYP (5GWE) tiene una cavidad más grande y puede establecer dos enlaces de hidrógeno, uno con la ARG 193 y otro con la THR 81, en ambos extremos del ligando. Este BacCYP no puede unirse a CAM, ya que carece de TYR 96 y su cavidad es demasiado grande (los resultados de *docking* muestran muchas poses diferentes con poblaciones bajas). De manera clara, y como se esperaba, la especificidad de BacCYP está estrechamente relacionada con el tamaño de su sitio activo y la presencia de residuos específicos capaces de interactuar con los ligandos.

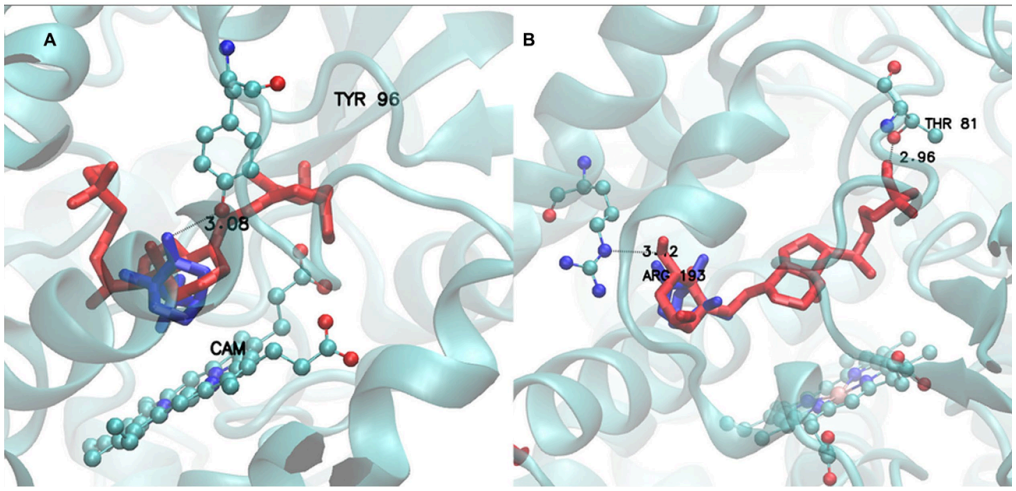


Figura 30 *Sitio activo del receptor del Grupo H (PDB 1AKD) con el ligando CAM (Grupo 3) en azul y el ligando VDX en rojo (A). Sitio activo del receptor del Grupo J (PDB 5GWE) con el ligando CAM (Grupo 3) en azul y el ligando VDX en rojo (B).*

4.10 Capacidad predictiva

Para demostrar la capacidad predictiva de la estrategia propuesta, llevamos a cabo todos los pasos descritos previamente en dos BacCYPs con sustratos ya conocidos (o sea controles positivos). La ***Figura 31*** ilustra de manera general este enfoque, que puede resumirse así: i) primero se toma la secuencia de un BacCYP nuevo y se clasifica en uno de los 14 grupos filogenéticos, evaluando la promiscuidad del grupo correspondiente. ii) Luego, se realiza una búsqueda en el PDB para identificar estructuras relevantes; si no se encuentra una, se genera un modelo utilizando AlphaFold o, en caso de existir una estructura muy similar, se construye un modelo de homología, por ejemplo utilizando modeller.

Una vez clasificada filogenéticamente la secuencia y determinada la estructura o modelo de la proteína, iii) se determinan los posibles los candidatos sustratos, productos e inhibidores (SPIs), que se identifican mediante los dos enfoques complementarios: a) el análisis de contexto genético y b) la búsqueda por correlación filogenética e identidad de secuencias. La búsqueda por identidad de secuencia se basa en la identificación de ligandos asociados a proteínas con secuencias similares, como se describió previamente. Por otro lado, el análisis de contexto genético, aunque excede los alcances principales de esta tesis, parte del principio de que, en bacterias, los genes involucrados en vías metabólicas suelen organizarse en operones. En este contexto, el producto de un gen puede actuar como sustrato de la proteína codificada por un gen cercano. Para este análisis, se utilizó la base de datos KEGG para estudiar la relación entre los sustratos y productos de los genes vecinos, y de este modo evaluar su similitud química mediante el índice de Tanimoto.

Después de aplicar ambos enfoques, se verifica la presencia de posibles SPIs. Si no se identificaron compuestos candidatos, iv) se procede con el docking molecular utilizando un conjunto representativo de los ligandos conocidos que unen BacCYPs. Si se identificaron compuestos candidatos a ser SPIs, se realiza el docking molecular en el dominio funcional de la proteína con estas moléculas. Este enfoque permite priorizar, entre todos los candidatos, aquellos con mayor probabilidad de ser verdaderos SPIs del BacCYP. Finalmente, como resultado de este flujo de trabajo, se obtiene una selección de compuestos aptos para poder realizar ensayos experimentales.

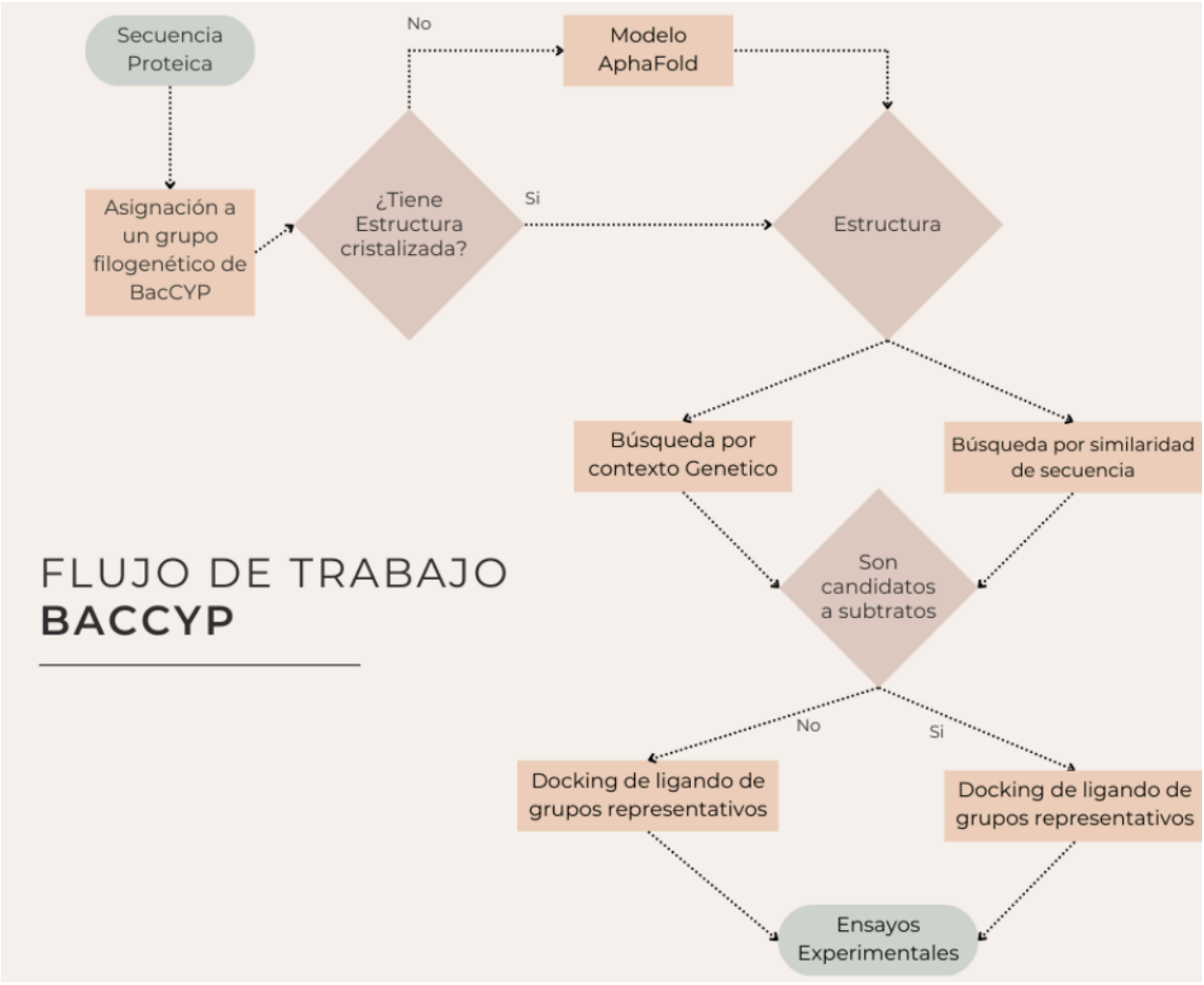


Figura 31 *Diagrama de flujo de la metodología propuesta para identificar sustratos de BacCYP*

El primer caso de estudio corresponde a CYP121 de *Mycobacterium tuberculosis* (cepa ATCC 25618 / H37Rv, Uniprot ID: P9WPP7), una proteína involucrada en la síntesis de micociclosina, cuyo sustrato es el compuesto cyclo(L-tirosil-L-tirosil), perteneciente al grupo de ligandos de tipo 7. El análisis de secuencia clasifica a MtCYP121 dentro del grupo filogenético J, que incluye ligandos de varios grupos (1.1, 1.2, 1.3, 2, 4.1, 4.2, 5.1, 5.2, 6, 7 y 8).

Por otro lado, el análisis del contexto genético mostró que los genes cercanos a MtCYP121 son P9WPF9 (aguas arriba) y P9WLF1 (aguas abajo). El gen P9WPF9 codifica para Cyclo(L-tirosil-L-tirosil) sintasa, una enzima que produce el sustrato de MtCYP121 a partir de L-tirosil-tRNA(Tyr), liberando la parte de tRNA(Tyr) durante la reacción. Además, la simulación de *docking* confirmó que cyclo(L-tirosil-L-tirosil) se une de manera efectiva a MtCYP121, lo que valida su identificación como el sustrato real (***Figura 32***).

Cabe destacar que al realizar *docking* con el conjunto completo de ligandos, realizado a modo de prueba adicional, se identificaron otros dos posibles sustratos (de los grupos 4.1 y 10.2

respectivamente). Este hallazgo resalta la ventaja de combinar el análisis del contexto genético con el *docking* para una asignación más precisa de sustratos, en lugar de emplear cada método por separado.

En cuanto a la estructura, el BacCYP más similar a MtCYP121 es MycG de *Micromonospora griseorubida* (Uniprot ID: Q59523), que comparte un 50% de identidad y está asociado a ligandos del grupo 2. Para modelar la estructura de MtCYP121, se utilizó el programa Modeller y la estructura PDB 2YGX. Aunque AlphaFold es una herramienta recomendada, en este caso se prefirió Modeller para evitar posibles sesgos relacionados con el uso de la estructura experimental de la proteína en el entrenamiento de AlphaFold v2.0.

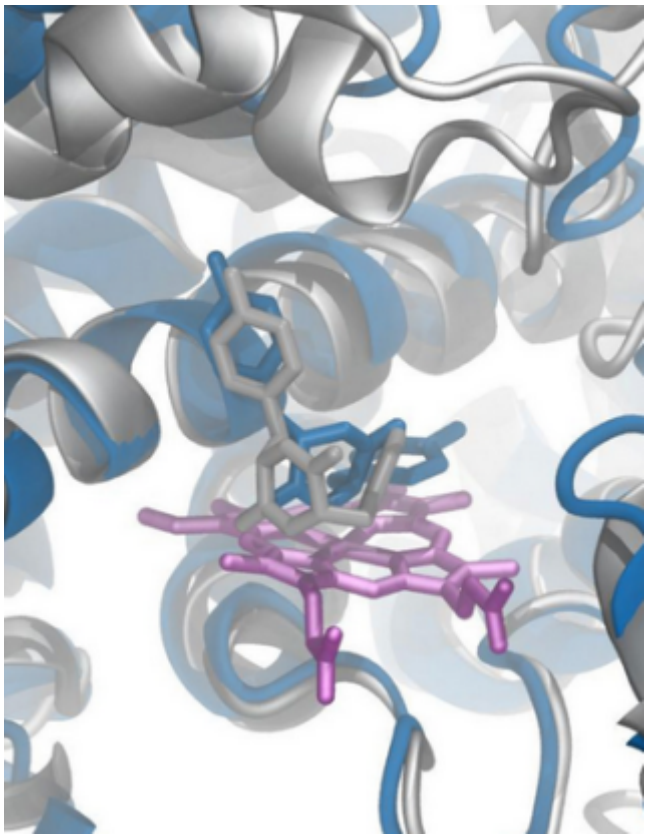


Figura 32 Comparación entre la estructura obtenida por rayos X (plateado) y la conformación obtenida por docking (azul) para el ligando natural Cyclo(L-tirosil-L-tirosil).

El segundo caso de ejemplo corresponde a la **Aromatic O-demethylase, subunidad del citocromo P450** (Uniprot ID **P0DPQ7**) de *Amycolatopsis sp.* ATCC 39116. Este caso resulta particularmente desafiante, ya que **no se dispone de información sobre los sustratos o productos de los genes vecinos**, y la primera estructura cristalizada de la proteína fue publicada en el **PDB el 4 de julio de 2018**, después de la fecha límite utilizada para el entrenamiento de la red neuronal de AlphaFold.

La proteína pertenece al **grupo I**, que se caracteriza por su promiscuidad, es decir, tiene la capacidad de unir sustratos de **nueve grupos químicos distintos**, entre ellos **guaiacol, 3-metoxicatacol y guaetol** (grupos 10.1 y 10.2). Para estudiar su estructura, se generó un modelo con **AlphaFold versión 2.0**, que resultó ser de **alta fidelidad**, con un **RMSD-CA de 0.726 Å** en comparación con la estructura experimental. Este hecho resalta la capacidad y precisión de AlphaFold para generar estructuras a partir de la secuencia, particularmente cómo en este caso en familias (o dominios) de proteínas donde existen diversas estructuras disponibles.

Los resultados del **docking molecular** mostraron que los sustratos del **grupo 10.1** fueron los segundos mejores posicionados en términos de afinidad, justo detrás del **grupo 8**, mientras que los del **grupo 10.2** ocuparon el quinto lugar. **Aunque no se pudo identificar un sustrato específico**, los resultados de nuestra estrategia muestran, que la combinación de modelado estructural y docking permitió **restringir la posible gama de sustratos** a unos pocos grupos (2 o 3), entre los cuales se encuentran aquellos ya conocidos.

4.11 Conclusión

Los citocromos P450 bacterianos (BacCYPs) desempeñan un papel clave en diversas rutas metabólicas, catalizando reacciones de oxidación sobre una amplia variedad de sustratos. Sin embargo, la identificación de los sustratos específicos de cada una de estas enzimas sigue siendo un gran desafío debido a la gran diversidad estructural y funcional de los BacCYPs.

En este estudio, se implementaron tres enfoques complementarios para la **predicción de los sustratos de BacCYPs**: el análisis filogenético, el modelado estructural, y el docking molecular. Se demostró que **la clasificación filogenética por sí sola no es suficiente** para inferir con precisión los sustratos, ya que proteínas filogenéticamente cercanas pueden presentar sustratos y especificidades significativamente diferentes. No obstante, cuando se combina esta información con modelado estructural y docking molecular, es posible **restringir el rango de posibles sustratos**, mejorando ampliamente la capacidad predictiva del método.

Los resultados del modelado estructural con **AlphaFold** mostraron además que este enfoque proporciona modelos de alta precisión, incluso superiores a aquellos obtenidos mediante modelado por homología con plantillas de identidad moderada. Esta precisión estructural permitió realizar simulaciones de docking molecular confiables, lo que facilitó la identificación de sustratos potenciales con base en su afinidad y ajuste en el sitio activo de los BacCYPs estudiados.

El análisis de docking molecular reveló que, si bien algunos BacCYPs presentan una alta especificidad por cierto grupo de sustratos, otros exhiben una promiscuidad significativa, pudiendo unir compuestos químicamente muy diferentes. En particular, el caso de **Aromatic O-demethylase (P0DPQ7)** evidenció cómo, incluso en ausencia de información filogenética o de contexto genómico, el modelado estructural y el docking pueden **acotar el rango de sustratos posibles**, proporcionando información útil para estudios experimentales futuros.

En conclusión, la combinación de estos enfoques bioinformáticos permite mejorar significativamente la identificación de sustratos en BacCYPs, reduciendo el espacio de búsqueda y proporcionando una base sólida para estudios experimentales posteriores. Con el crecimiento continuo de las bases de datos estructurales y la mejora de herramientas de predicción, se espera que estos métodos sean cada vez más precisos y aplicables a una gama más amplia de citocromos P450 bacterianos.

Evaluación de la Flexibilidad Conformacional del Sitio Activo Mediante Docking Molecular

5.1 Evaluación de la Flexibilidad Conformacional del Sitio Activo Mediante Docking Molecular

En este capítulo, se aplicarán los conceptos y metodologías desarrollados en los capítulos anteriores para evaluar los cambios conformacionales que experimenta una proteína en respuesta a la unión de un ligando. Partiendo de una estructura proteica sin ligando (también denominada apo), se llevará a cabo un estudio de docking con el fin de predecir y analizar los ajustes estructurales que la proteína adopta al interactuar con el ligando, llegando al complejo final en la forma que es denominada holo.

El objetivo principal es caracterizar estas modificaciones conformacionales y obtener la estructura de la proteína en su estado holo (también llamado **bound)**, comparándola con su conformación inicial en ausencia del ligando. Para ello, se emplearán herramientas de docking molecular. Adicionalmente, partiendo de una estructura sin ligando, se busca predecir, en la medida de lo posible, los cambios conformacionales que facilitan la interacción del sitio activo con el ligando.

Este proceso conlleva varios desafíos, entre ellos el fenómeno conocido como **"ajuste inducido"** (***Figura 33***), en el cual la presencia del ligando puede provocar cambios conformacionales significativos en la proteína. Este efecto puede influir en la precisión del docking, ya que las estructuras iniciales sin ligando podrían no reflejar adecuadamente el estado conformacional más favorable para la interacción con el ligando.

Esperamos que este análisis contribuya a una mejor comprensión de la flexibilidad conformacional de la proteína y su relación con la afinidad y selectividad hacia distintos ligandos, proporcionando información relevante para el diseño de fármacos o la ingeniería de proteínas con funciones específicas.

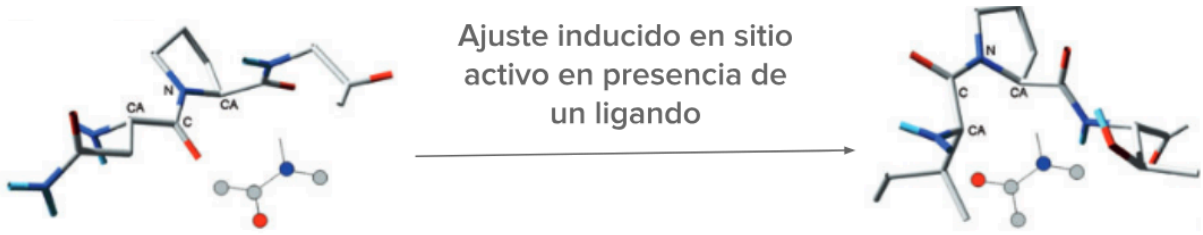


Figura 33 La unión de un ligando provoca cambios conformacionales en la proteína, adaptando su estructura para optimizar la interacción. Esto sugiere que la proteína no es completamente rígida, sino que su flexibilidad juega un papel clave en el reconocimiento y la afinidad por el ligando.

5.2 Métodos

Para llevar a cabo la caracterización del sitio activo se utilizaron varios script bioinformáticos algunos de desarrollo propio, cuyas principales ideas se encuentran descritas en los métodos de esta tesis. Con el objetivo de analizar estos cambios conformacionales entre holo y apoproteínas se busco un set de de datos donde esté confirmado el mismo experimentalmente, Para ello nos basamos en el dataset del paper ***"Identification of Cryptic Binding Sites Using MixMD with Standard and Accelerated Molecular Dynamics".*** [65] Este trabajo explora la eficacia de la técnica MixMD (dinámica molecular con solventes mixtos) para identificar sitios de unión crípticos en proteínas. Los sitios crípticos son regiones de unión que no son evidentes en las estructuras de proteínas sin ligando (apo) y que requieren reordenamientos conformacionales para ser accesibles. Este dataset contiene 11 estructuras que sufren un claro cambio conformacional en presencia de un ligando.

El objetivo de este capítulo es , tomando como punto de partida este conjunto de datos, analizar una serie de parámetros con el fin de recuperar la proteína unida al ligando partiendo de la estructura Apo , para ello se requirió el desarrollo de diferentes scripts de python con el fin de automatizar todos los pasos necesarios. Entre los puntos más importantes se encuentran:

- Análisis exploratorio de Angulos Chi
- Evaluación de Ángulos e interacciones
- Reconstrucción del Sitio Activo

5.3 Análisis Exploratorio Ángulos de torsión de los residuos

Como primer paso se analizaron los cambios conformaciones que pueden suceder en los ángulos torsionales de los residuos de los aminoácidos que conforman el sitio activo. Para ello se evaluaron los ángulos χ1 χ2 χ3 χ4 χ5 para un conjunto de 422 estructuras obteniéndose las distribuciones correspondientes para cada ángulo y tipo de residuo. En el caso del ángulo χ_1, se identificaron tres regiones de mayor población en todos los residuos (no se notaron diferencias significativas entre los valores de los ángulos de distintos residuos, salvo para Gly y Pro, que por su estructura quedan fuera de este análisis) Sin embargo, diferentes residuos presentan diferentes poblaciones relativas (o preferencias) por cada una de las posibles configuraciones, como se muestra en la ***Figura 34***. Estas regiones fueron clasificadas como de **alta, media y baja probabilidad**, según el número de observaciones registradas en cada una de ellas.

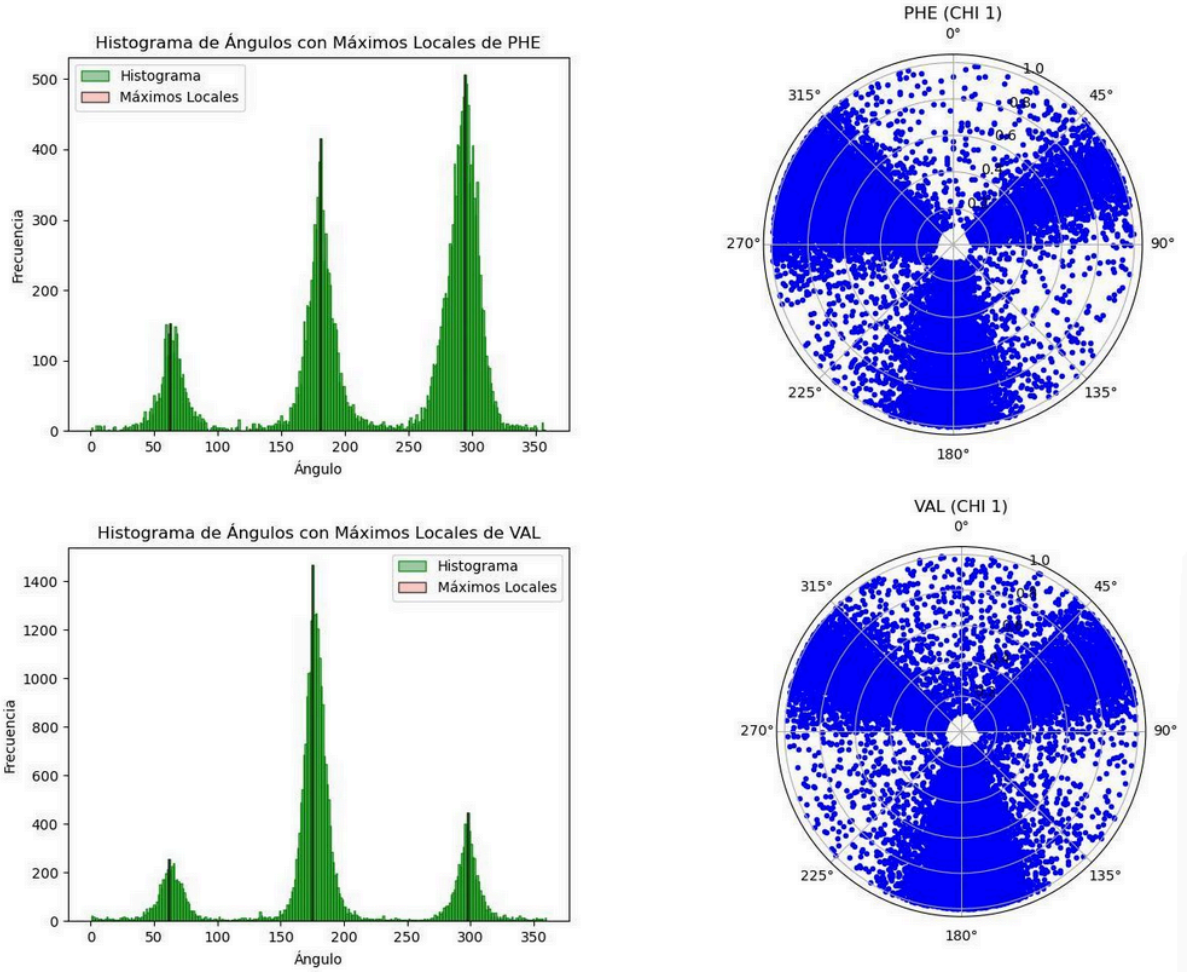


Figura 34 Se observa la distribución de ángulos en los residuos ***PHE*** y ***VAL***, identificándose en ambos casos tres regiones de mayor probabilidad. Cada una de estas regiones fue clasificada como de **alta, media o baja probabilidad**, de acuerdo con su frecuencia.

A partir de estos resultados, se construyó un clasificador utilizando el método de **k-means**, agrupando los valores de los ángulos en tres categorías según su probabilidad de aparición. Esta clasificación permitió asignar cada valor de ángulo a un grupo de alta, media o baja probabilidad. Posteriormente, el clasificador fue utilizado para comparar los ángulos torsionales entre las estructuras Apo y Holo de todos los residuos del sitio activo, evaluando si los cambios observados correspondían a transiciones hacia configuraciones más, o menos probables.

5.4 Set de Evaluación - Re-Docking

A partir de las 11 proteínas que integran el dataset, se realizaron los dockings correspondientes para ambas estructuras (Apo y Holo) usando el ligando de la estructura Holo. Para ello se alinearon ambas estructuras (Apo , sobre Holo) y así poder utilizar las mismas coordenadas de las grillas de Autodock, los alineamientos se evaluaron usando el RMSD entre ambas estructuras. En prácticamente todos los casos dieron valores menores a 4A, a excepción de 3 casos B-Secretates, C-MET y TIE-2 , los cuales no se lograron buenos alineamientos, esta información se resume en la ***Tabla 12***.

Sistema	Apo	Holo	Ligando	RMSD
Adipocyte	1ALB	1LIC	HDS	0.55
Androgen Receptor	2AM9	2PIQ	RB1	2.89
B - Lactasa	1JWP	1PZQ	CBT	1.01
B - Secretasa	1W50	3IXJ	586	17.83
C-MET	1R1W	3F82	353	15.50
Guanylate	1EX6	1GKY	5GP	4.01
HSP 90	2QFO	2WI7	2KL	3.55
PTP1B	2F6V	1T49	892	3.41
Ricin	1RTC	1BR6	PT1	0.57
TIE-2	1FVR	2OO8	RAJ	19.55
UAPI	2YOC	2YQS	UD1	3.9

Tabla 12 PDB ID's de las estructuras de referencia (Apo y Holo), ligando correspondiente y valores de RMSD obtenidos del alineamiento estructural entre ambas conformaciones.

Una vez alineadas las estructuras, se procedió a realizar los docking, tanto en su modalidad estándar como con bias, utilizando tanto las estructuras **holo** como **apo**. El objetivo de este análisis fue evaluar si existían diferencias significativas en los resultados del docking entre ambos casos, bajo la hipótesis de que, si las estructuras presentan variaciones conformacionales relevantes para la unión del ligando, los resultados del docking no deberían ser equivalentes. En otras palabras ***esperamos que los docking solamente den buenos resultados en la estructura holo*** (que es efectivamente un re-docking) y no así en la apo.

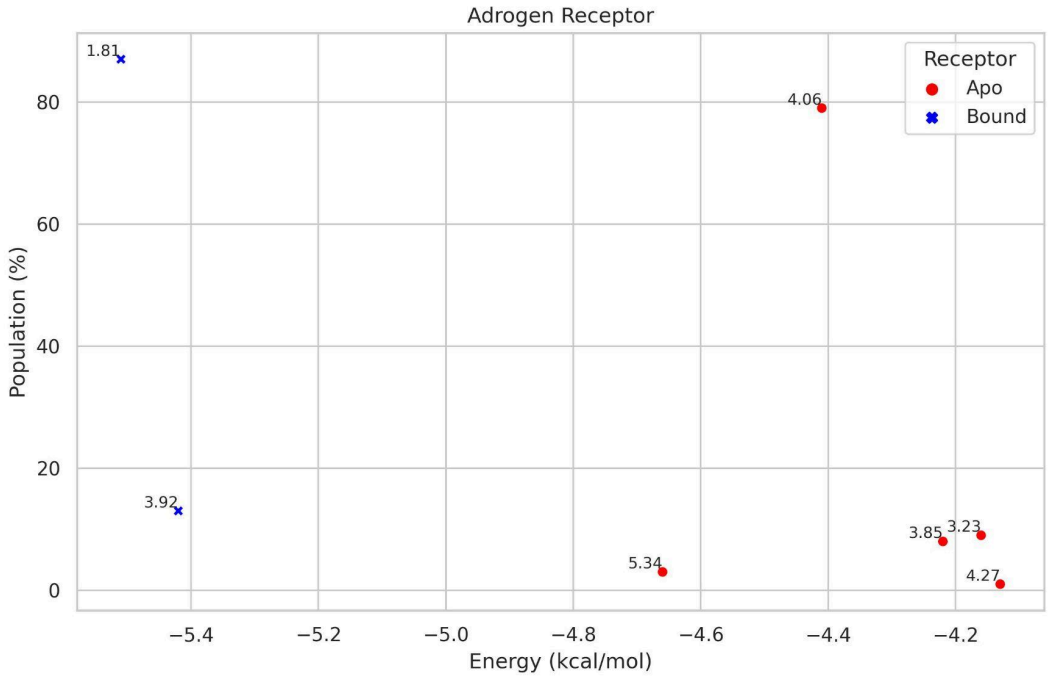


Figura 35 un docking correcto en el caso de bound poca población y energía bajas , mientras en el caso apo una mayor variedad de poses y menor energía libre en el caso apo.

Los resultados del docking se representaron mediante los gráficos de población versus energía, como fue explicado previamente. En este caso, se combinaron los resultados obtenidos para ambas estructuras (Apo y Holo). En la mayoría de los sistemas analizados se observaron dos poblaciones diferenciadas: una con menor energía de docking, asociada principalmente a las estructuras Holo, que además presentaba outliers claramente identificables en el cuadrante superior izquierdo; y una segunda población, generalmente con mayor número de poses, que podía o no contener outliers. Cabe destacar que, en la mayoría de los casos para las estructuras Holo, el outlier correspondió a una pose correcta (es decir, con un RMSD respecto a la referencia inferior a 2 Å), mientras que en las estructuras Apo, muchas veces el outlier correspondió a una pose errónea.

A modo de ejemplo, en el caso del **receptor de andrógeno *(Figura 35)*** se observa un **docking correcto** (RMSD < 2A) en la conformación **holo**, caracterizado por una población grande y una mayor variabilidad de poses, junto con una menor energía libre de acoplamiento. En contraste, en la conformación **apo**, se identifica una población menor que además posiciona al ligando incorrectamente (RMSD de 4A), con una distribución más dispersa, y energías de acoplamiento generalmente menos favorables.

Este hallazgo sugiere que, a nivel estructural, los sitios de unión presentan diferencias significativas. Dado que la composición en aminoácidos es la misma en ambos casos, se puede inferir que estos cambios son producto de diferencias en la orientación espacial de los

residuos. Esto sugiere también que la presencia del ligando estabiliza una conformación más adecuada para la interacción, en línea con el fenómeno de **ajuste inducido**.

Es importante notar, que en algunos pocos casos se puede observar que el resultado del docking fue malo en ambas estructuras. Con el fin de mejorar el docking para estos casos y avanzar en el análisis conformacional que es el objetivo del capítulo, se aplicó la técnica de Bias Docking a todos los casos. Para ello se analizaron las interacciones presentes en cada ligando en la versión Holo, y se procedió a seleccionar las 2 más representativas de cada caso para sobre ellas aplicar el bias, los resultados obtenidos se resumen en la ***Tabla 13 ,*** y muestran la ventaja de utilizar el Bias-docking, ya que el mismo logra resultados consistentes en los casos donde el docking normal falla.

Sistema	Población (%) (Bias)	Población (%) (Holo)	Población (%) (Apo)	Energía (Bias)	Energía (Holo)	Energía (Apo)
Adipocyte	79	75	75	-7.84	-6.04	-5.43
Androgen Receptor	98	87	63	-5.51	-5.01	-4.41
B - Lactasa	100	22	48	-7.79	-6.67	-4.57
B - Secretasa	13	10	7	-6.28	-6.28	-6.25
C-MET	98*	98	9	-13.24	-13.14	-4.22
Guanylate	86	70	52	-12.64	-8.93	-5.82
HSP 90	50	47	60**	-11.02	-9.61	-8.14
PTP1B	94	88	7	-16.03	-10.79	X
Ricin	91	36	30	-13.41	-6.05	-4.02
TIE-2	21	50	17	X	-9.45	X
UAPI	66	41	15	-17.57	-11.64	-2.43

Tabla 13 Resumen de los resultados de docking. Se indican los porcentajes de población de clústeres y las energías medias de unión para cada sistema, comparando las versiones Bias, Holo y Apo. Las celdas en **verde** representan casos donde el docking mejoró significativamente tras la aplicación del bias, mientras que las celdas en **rojo** corresponden a sistemas donde no se logró resolver adecuadamente la predicción de unión.

A modo de ejemplo, en la ***Figura 36***, se muestra el complejo **Ricin–PT1**, donde se evidencia el efecto de la aplicación del bias sobre los resultados de docking para la predicción de pose. En el **panel izquierdo**, correspondiente al docking convencional, para la estructura Holo se observa que, si bien se encuentra la pose correcta (RMSD = 0,95 Å), esta aparece con relativamente **baja población (ca. 38%)**. En contraste, sobre la estructura Apo, el docking no logra identificar poses relevantes, obteniendo principalmente poses con un **alto RMSD** y baja calidad. Luego de aplicar el **Bias** (panel derecho), se evidencia un **enriquecimiento significativo** en la predicción correcta: la pose correspondiente al estado Holo se recupera ahora con un **RMSD de 0,92 Å** (ligeramente menor que en el caso anterior), y con una **población superior al 80%.** Como es de esperarse los resultados tambi**én mejoran ligeramente para la estructura Apo.**

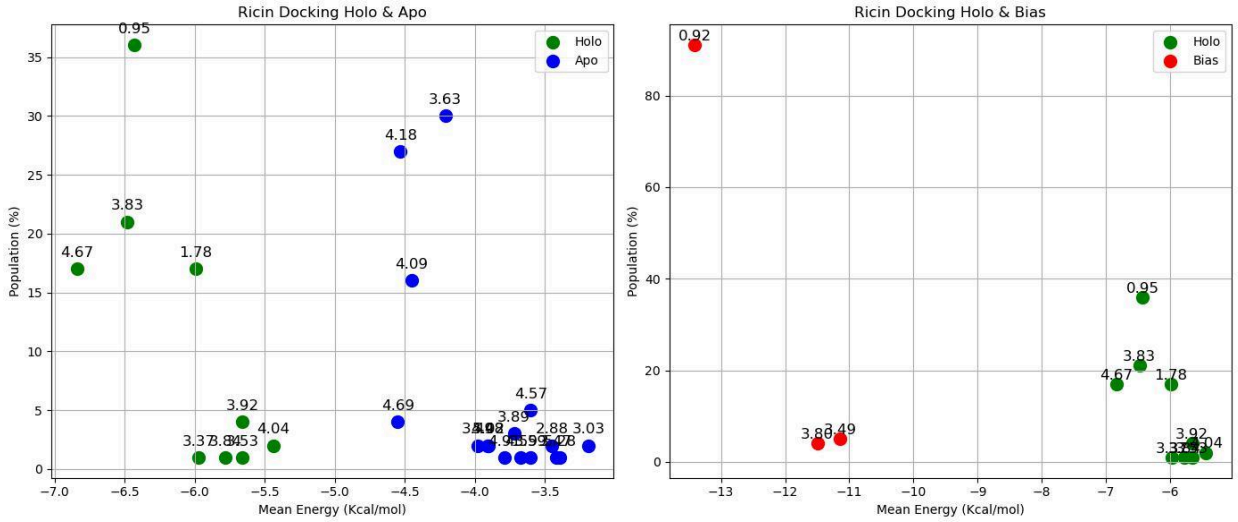


Figura 36 Se observa una **mejora sustancial** en el caso con **bias** (panel derecho de la imagen), lo que refuerza la estabilidad del docking y la precisión en la predicción de poses. Además, esta configuración amplifica aún más las diferencias en los valores de **energía libre** respecto al caso **Apo**, destacando la influencia del estado conformacional inicial en los resultados obtenidos en el docking.

En resumen, los resultados del docking confirman que en la mayoría de los casos nuestra hipótesis de trabajo de que los cambios estructurales entre las conformaciones Apo y Holo pueden impactar significativamente en la capacidad de unión del ligando. En la ***Figura 37*** se resume la variación de la población entre el mejor caso obtenido para la estructura Holo (eligiendo el mejor caso, ya sea normal o bias) y la estructura Apo. Como puede observarse existe un continuo en la magnitud del cambio de población, desde valores cercanos a 30% hasta valores superiores al 80% (barras azules). Buscando entonces aquellos sistemas donde el cambio sea tan significativo que podemos decir que solamente la proteína holo es capaz de revelar mediante el docking la unión del sustrato, y este falla en la Apo, nos quedamos con aquellos sistemas donde la diferencia de población superará al 25%, y donde se confirmara que en la holo, el docking representa la pose correcta (o sea con un RSMD menor a 2). Con este criterio el sistema B-Secreta es entonces el único excluido por no superar el umbral. Otra variable analizada fue la **variación del RMSD** entre las poses obtenidas (en la misma figura

pero en color naranja). Estos resultados muestran que, en la mayoría de los sistemas evaluados, reflejada en una disminución del RMSD respecto a la estructura de referencia, es decir que en los casos Holo, el bias encuentra poses más cercanas al cristal que las Apo.

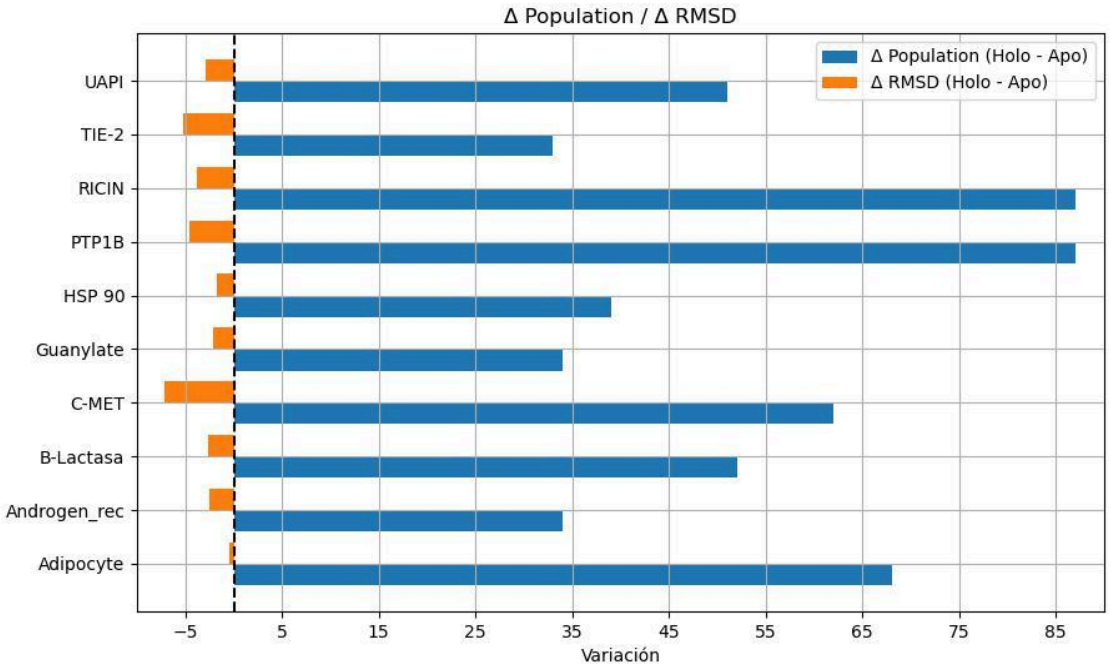


***Figura 37*:** Variación de la población y del RMSD entre estructuras Apo y Holo. Se muestra la diferencia en la población del cluster más estable (Holo - Apo) en azul, y la variación del RMSD respecto a la estructura de referencia (Apo - Holo) en naranja, para cada sistema analizado. Los resultados reflejan que en la mayoría de los casos los cambios estructurales entre las conformaciones Apo y Holo impactan significativamente en la capacidad de unión del ligando, con una disminución general del RMSD en los sistemas Holo, indicando una mayor cercanía a la pose cristalográfica.

5.5 Analisis Sitio Activo

Se procedió al análisis del sitio activo tanto en las estructuras Apo como en las Holo. Se definió como **sitio activo** a todo conjunto de residuos cuyos átomos se encuentran a una distancia de hasta 10 Å del centro de masa del ligando. Todo residuo que presente al menos un átomo dentro de dicho radio fue considerado parte del sitio activo.

Se analizaron tres variables principales asociadas al sitio activo:

1. Interacciones intermoleculares.

2. Volumen del sitio activo.

3. Ángulos conformacionales de los residuos involucrados.

Este análisis se realizó de manera sistemática para todos los sistemas incluidos en la presente tesis. Sin embargo, como ejemplo representativo, y por una cuestión de espacio, se presenta a continuación el análisis detallado realizado para el complejo **Ricin–PT1**, con el objetivo de ilustrar paso a paso la metodología aplicada. En este caso, el sitio activo se definió a partir del centro de masa del ligando en su conformación original (complejo Holo), y las mismas coordenadas de referencia fueron empleadas para delimitar el sitio activo en la estructura Apo, asegurando así la comparabilidad entre ambos estados conformacionales.

En primer lugar, se evaluaron las interacciones entre el sitio activo y el ligando. Para el receptor en su conformación Holo se utilizó directamente la estructura cristalográfica, mientras que en el caso del receptor Apo se partió de la "mejor" conformación obtenida mediante docking. Este análisis se realizó empleando un ***script*** desarrollado en el marco de la presente tesis. En el caso del ligando **PT1**, el script identifica *hot points* (regiones candidatas a participar en interacciones del tipo donador/aceptor de puentes de hidrógeno o interacciones aromáticas) tanto en el ligando como en los residuos del sitio activo del receptor. A continuación, evalúa la geometría de las posibles interacciones mediante el cálculo de ángulos. En la ***Figura 38*** se presenta una salida representativa del análisis, donde pueden observarse varios puntos de interés en el ligando PT1 y la tabla de resultados correspondientes.

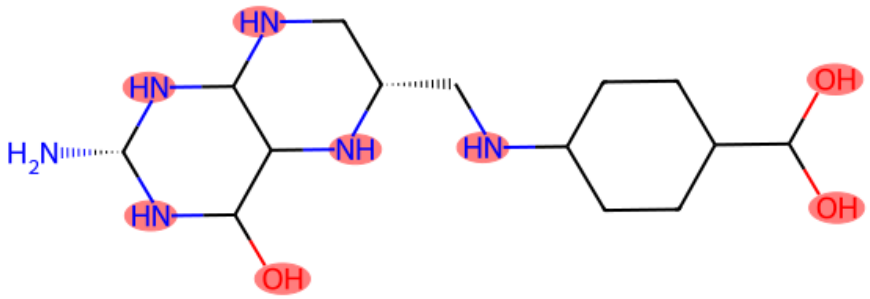


Figura 38 Ligando PT1 con puntos de interacción, tanto dadores como aceptores de H^+

Una vez calculadas las posibles interacciones (puentes de hidrógeno, interacciones aromáticas) a partir de los ángulos y distancias correspondientes, se construye la ***Tabla 14***, que describe las interacciones identificadas, su tipo, los ángulos involucrados y los átomos participantes tanto del receptor como del ligando. De esta manera, se obtiene una **caracterización detallada de los residuos** que participan activamente en el sitio activo. La ***Figura 38*** complementa esta información al mostrar los grupos funcionales del ligando **PT1** que actúan como donadores o aceptores de puentes de hidrógeno, los cuales fueron identificados en las interacciones descritas en la tabla. Así, se establece una correspondencia directa entre los átomos resaltados en la figura y las interacciones específicas con residuos del sitio activo reportadas en la tabla.

Residuo	Átomo	Distancia	Tipo	Angulo
ARG 180	NH2	2.837	Aceptor	121°
ASN 78	ND2	2.77	Aceptor	111°
VAL 81	N	3.00	Aceptor	127°
VAL 81	O	3.04	Dador	141°
Tyr 80	Ring Center	3.12	Aromático	71°

Tabla 14 Caracterización de las interacciones con el ligando PT1. Se detallan los residuos, los átomos participantes, las distancias de interacción, el tipo de interacción (aceptor, dador o aromática) y los ángulos correspondientes. Esta información complementa los puntos de interacción mostrados en la Figura 35, permitiendo una descripción precisa de las interacciones moleculares en el sitio activo.

De la comparativa entre ambos casos (holo vs apo) podemos concluir que, como era de esperar hay una desaparición de la mayoría de las interacciones en el caso Apo respecto al Holo, lo que confirma la diferencia en la conformación del sitio activo, y por donde las interacciones con el ligando entre ambas estructuras, el resumen se encuentra en la ***Tabla 15***.

Interacción	Holo	Apo
Puente Hidrógeno	11	4
Aromática	1	0
Total	12	4

Tabla 15 Resumen de las interacciones encontradas entre el receptor y el ligando PT-1 , se observa la pérdida de las mismas en el caso Apo.

Junto con las interacciones, se caracterizaron los residuos pertenecientes al sitio activo, se identificaron cuales residuos lo integran en ambos contextos estructurales (***Figura 39***). El análisis indicó que los residuos involucrados eran, en su mayoría, coincidentes en ambas estructuras. No obstante, se observaron pequeñas diferencias: el residuo **GLY 212** se aleja del sitio activo en la conformación Apo, mientras que **GLN 173**, ausente inicialmente, se aproxima e ingresa en dicha región. Este comportamiento se invierte en la estructura Holo, reflejando un cambio inducido por la presencia del ligando.

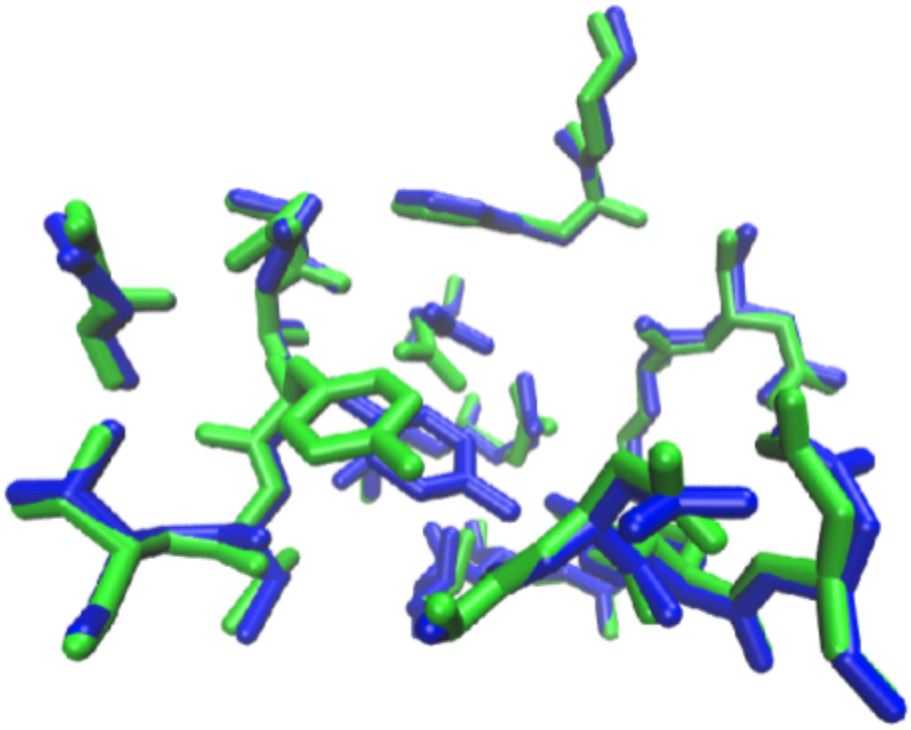


Figura 39 *Residuos detectados como parte del sitio activo En verde se representa la conformación en la estructura Holo y en azul la correspondiente a la estructura Apo. Se observa, en particular, una diferencia marcada en la orientación de los residuos* ***TYR 80*** *y* ***GLU 177****, que presentan las mayores variaciones angulares, indicando un movimiento significativo asociado a la transición entre los estados Apo y Holo.*

Una vez identificados los residuos que conforman el sitio activo, se procedió a evaluar las variaciones conformacionales mediante el análisis de los ángulos **χ** (chi) de cada residuo. Para ello, se calcularon las diferencias en los ángulos χ1 a χ5 —según correspondiera a cada aminoácido— entre las estructuras Apo y Holo. Este análisis permitió cuantificar los cambios en la orientación de las cadenas laterales de los residuos involucrados.

Lo primero que se observa es que no todos los residuos presentan modificaciones significativas en sus ángulos de torsión y orientación "χi". Muchos mantienen una conformación estable entre ambas condiciones, lo que sugiere que su posición en el sitio activo no se ve afectada por la presencia del ligando. Sin embargo, en ciertos casos se detectan variaciones considerables, indicando una posible reorganización estructural inducida por la unión del ligando. En la ***Figura 40*** se representa la comparación de los ángulos χ_1 de cada residuo, donde el valor correspondiente a la estructura apo se ubica en el eje Y, y el de la estructura holo en el eje X. Por lo tanto, aquellos puntos que se sitúan sobre la diagonal corresponden a residuos que no modifican significativamente su orientación, mientras que los puntos alejados de la diagonal reflejan un cambio conformacional, evidenciado por una diferencia angular mayor. En este caso particular, se observa que solo **dos residuos** presentan un cambio

notable en su orientación caracterizada por χ_1: **GLU 177**, que se aleja ligeramente de la diagonal, y otro residuo identificado como **ASP 75**, que exhibe una desviación mucho más marcada. Este tipo de análisis se realiza de manera análoga para cada uno de los ángulos χ considerados para todos los residuos del sitio activo.

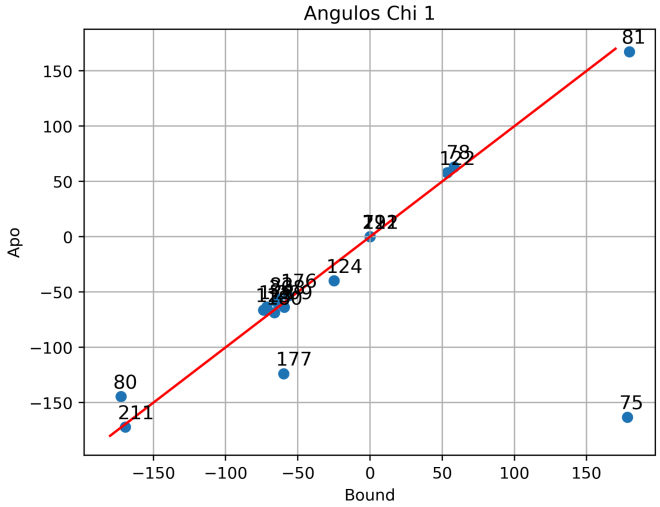


Figura 40 Comparación de los ángulos χ_1 de los residuos del sitio activo entre las estructuras Apo (eje Y) y Holo (eje X) para el sistema Ricin-PT1. Los puntos ubicados sobre la diagonal indican residuos que mantienen su orientación entre ambas condiciones, mientras que los puntos alejados de la diagonal reflejan cambios conformacionales, evidenciando variaciones en los ángulos χ_1 asociadas a la unión del ligando.

Es interesante destacar, que estos cambios conformacionales pueden estar asociados a la formación o pérdida de interacciones específicas, como puentes de hidrógeno, interacciones hidrofóbicas o contactos aromáticos, y constituyen un aspecto clave en la caracterización de sitios activos dinámicos o crípticos, en el **heatmap** (***Figura 41***) se resume la variación angular observada en los ángulos **χ1 a χ5** para todos los residuos del sitio activo, comparando las estructuras Apo y Bound del sistema Ricin–PT1. Las celdas coloreadas representan diferencias angulares expresadas en grados, donde los tonos rojos indican cambios positivos y los tonos azules, cambios negativos. Las celdas en gris indican ausencia de variación.

Los residuos que presentan las mayores diferencias conformacionales son **GLU 177**, que muestra una variación significativa en tres de sus ángulos (χ1 = 64.0°, χ2 = –118.6°, χ3 = 104.3°), y **TYR 80**, con un cambio marcado en χ2 (–168.2°). También se observan diferencias notables en **TYR 123** (χ2 = 65.1°), **ASN 122** (χ2 = –60.5°), y **ARG 180** (variaciones moderadas en χ2 = 15.2°, χ3 = 18.3° y χ4 = –51.9°).

Por otro lado, varios residuos como **ALA 79**, **GLY 121**, **ILE 172**, **ASP 124** y **VAL 82** no muestran variaciones angulares, lo que sugiere que sus cadenas laterales se mantienen en una conformación estable independientemente de la presencia del ligando.

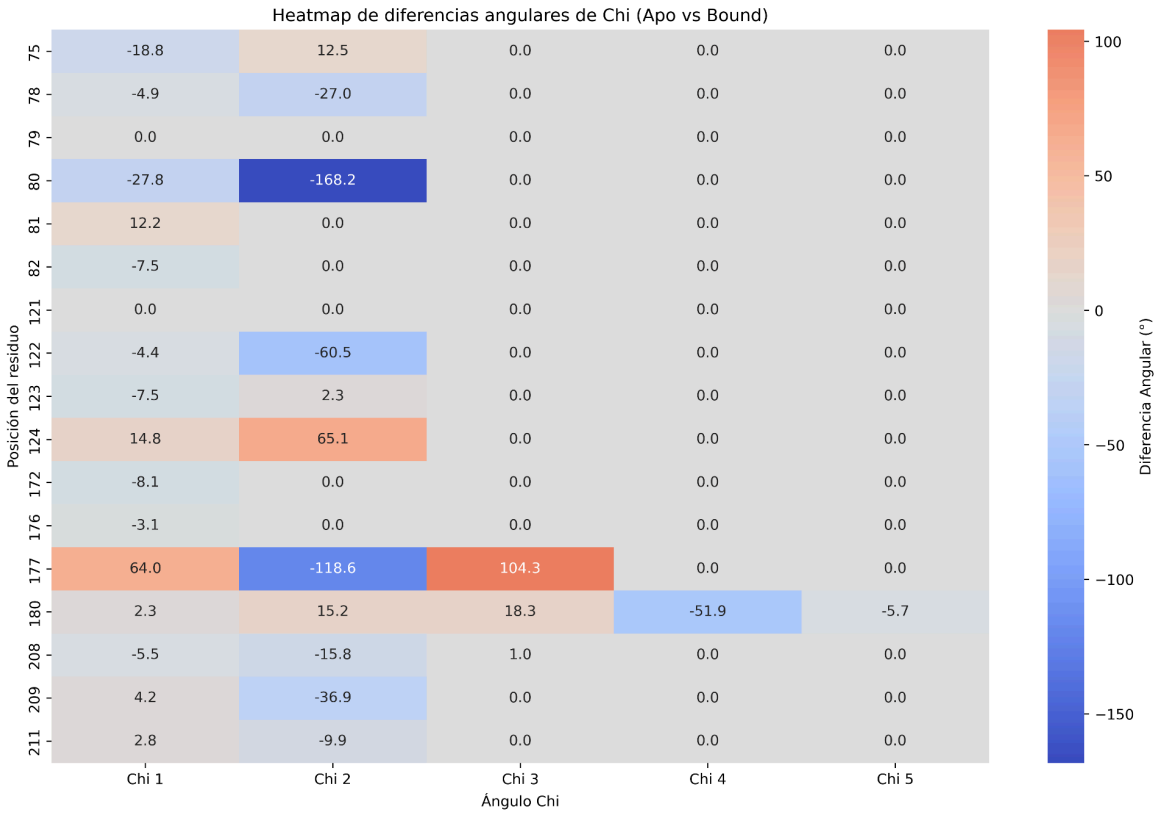


Figura 41 Diferencias angulares de los ángulos Chi (χ1–χ5) entre las conformaciones Apo y Bound de una proteína. Cada celda representa la diferencia angular (en grados) entre las conformaciones correspondientes para un residuo específico (eje Y) y un ángulo Chi específico (eje X).

También se analizó el movimiento de los ángulos que participan del backbone de la proteína (φ, ψ, ω). En el caso particular de Ricin-PT1 (***Figura 42***), los mayores cambios conformacionales se concentran en los ángulos φ y ψ, responsables de la flexión del esqueleto polipeptídico. Esta observación se mantuvo de manera consistente en los diez modelos analizados, donde, si bien se detectaron pequeñas variaciones en los ángulos del backbone, los desplazamientos fueron generalmente acotados. Tal como se muestra en la Figura 42, estas fluctuaciones sugieren que el ajuste estructural del sitio activo ocurre principalmente a través de **rotaciones del backbone** (φ y ψ), mientras que los ángulos ω (plano peptídico) y τ (torsiones complementarias) permanecen mayormente conservados.

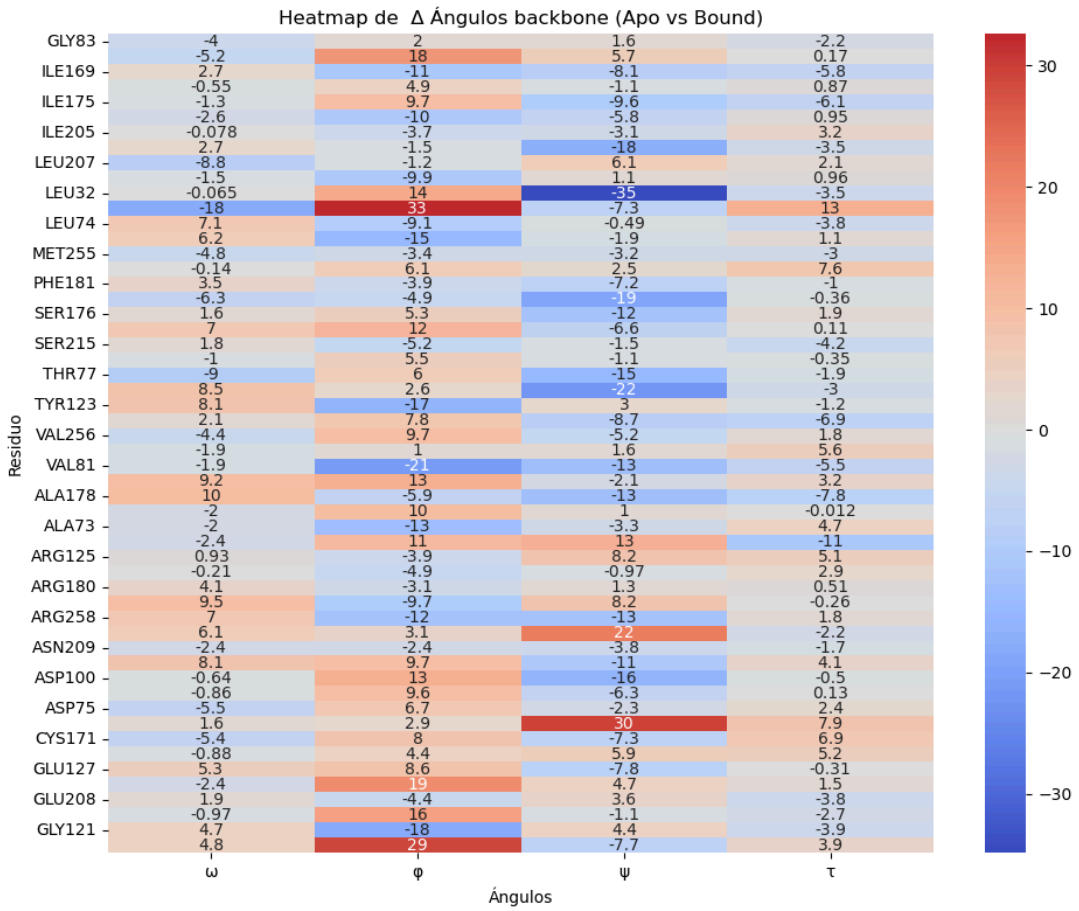


Figura 42 Heatmap de las diferencias de ángulos del backbone (Δφ, Δψ, Δω, Δτ) entre las conformaciones Apo y Holo para Ricin-PT1. Se observa que los mayores cambios se concentran en los ángulos φ y ψ, asociados a la flexión del esqueleto polipeptídico, mientras que ω (plano peptídico) y τ (torsión complementaria) muestran variaciones menores. Esta tendencia se repite de forma consistente en los modelos analizados.

Del análisis de esta clasificación de ángulos, se evaluó si existía alguna tendencia sistemática en los cambios conformacionales observados, como por ejemplo una mayor frecuencia de ángulos de baja probabilidad en la conformación Holo (con el ligando presente), o una dirección preferencial en las transiciones entre categorías de probabilidad (alta, media o baja). Si bien se observaron casos de transiciones entre estados de distinta probabilidad —por ejemplo, de alta a baja, o de baja a media—, no se identificó una tendencia generalizable que explique dichos desplazamientos. En conjunto, y como conclusión podemos decir que los cambios observados no siguen una dirección clara hacia configuraciones más o menos probables, lo que sugiere un comportamiento altamente dependiente del contexto estructural de cada sistema.

5.6 Análisis del volumen de sitio activo Ricin-PT1.

A partir del análisis de las diferencias estructurales entre los estados Apo y Bound, se identificaron los ángulos con mayor variación dentro del sitio activo. Como siguiente paso, se propuso **modificar el sitio activo en la estructura Apo**, ajustando aquellos ángulos que presentaron mayores cambios, con el objetivo de **hacerlo lo más similar posible al estado Holo**. Esta aproximación permite evaluar si dichas modificaciones estructurales son suficientes para favorecer o estabilizar la conformación observada en presencia del ligando.

Mediante la técnica de **Convex Hull** se calcularon el **volumen** y el **área** del sitio activo. Para ello, se utilizó como punto de referencia el **centro de masa de cada residuo** perteneciente al sitio activo, tanto en la estructura Apo como en la Holo. A partir de estos puntos, se generó la envolvente convexa que engloba el espacio tridimensional ocupado por el conjunto de residuos en cada caso.

Se realizaron los cálculos para el sistema Ricin-PT1, una vez definida la geometría del poliedro (***Figura 43***), se calcularon sus propiedades métricas: **el área superficial** y el **volumen encerrado**, los cuales sirven como indicadores del grado de apertura o compactación del sitio activo. Este enfoque permite comparar cuantitativamente cómo varía la conformación general del sitio activo en ausencia y presencia del ligando.

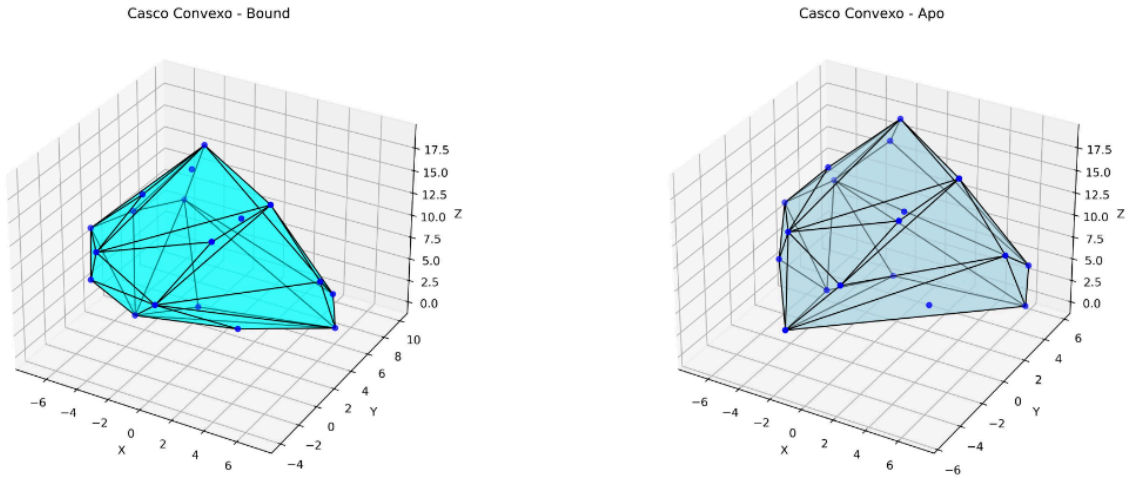


Figura 43 Representación tridimensional del volumen del sitio activo definido mediante el método de Convex Hull, utilizando como referencia el centro de masa de cada residuo del sitio activo, tanto para la estructura Bound (izquierda) como para la estructura Apo (derecha). Las superficies generadas permiten comparar cuantitativamente la geometría general del sitio activo en ausencia y presencia del ligando, evaluando diferencias en área superficial y volumen encerrado como indicadores del grado de apertura o compactación estructural.

A partir de los valores obtenidos para el área y volumen, se calcularon dos parámetros geométricos adicionales que permiten caracterizar su forma y grado de compactación: la **relación Área/Volumen (A/V)** y el **índice de compacidad (*Tabla 16*)**. El índice de compacidad

es adimensional y proporciona una idea de **qué tan compacto o esférico** es un objeto, Un valor **más alto** indica una forma **más esférica o compacta** (siendo 1 un esfera perfecta); un valor **más bajo** sugiere una forma más irregular o extendida.Mientras que la **relación Área/Volumen (A/V)** te da una idea de **cuánta superficie hay expuesta por unidad de volumen** del sitio activo, valores altos son sitios más abiertos mientras que valores más bajos son espacios más cerrados.
Estas diferencias sugieren que el sitio activo en el **Holo** presenta una geometría **ligeramente más compacta y esférica**, con **menor exposición superficial relativa**, en comparación con el Apo, que exhibe un bolsillo más **abierto o expandido**. Esta tendencia es coherente con una reorganización y ligera ampliación del sitio activo inducida por la unión del ligando.

Ricin-PT1	Holo	Apo	Δ (Holo - Apo)
Área	621.319	568.066	53.253
Volumen	1039.72	920.232	119.488
Compacidad	0.510	0.523	**-0.013**
A/V	0.5716	0.6176	-0.0460

Tabla 16

5.7 Análisis Global del volumen de los sitios activos

Habiendo mostrado el análisis de la comparación del sitio activo entre ambas estructuras (holo y apo), para el sistema Ricin-PT1, pasamos ahora a analizar los resultados obtenidos para todo el set. **Con el fin de evaluar si existía algún patrón general en el comportamiento conformacional del sitio activo frente a la unión del ligando**, se analizaron en todo el dataset los mismos parámetros estructurales calculados previamente: **área**, **volumen** y **compacidad** del sitio activo. Este análisis comparativo permitió observar tendencias comunes o particulares entre los diferentes sistemas estudiados. Los resultados se resumen en las ***Figura 45*** ,donde se observa las variaciones de volumen y área entre ambos sitios y la ***Figura 44*** , donde se observa la compacidad y la relación A/V

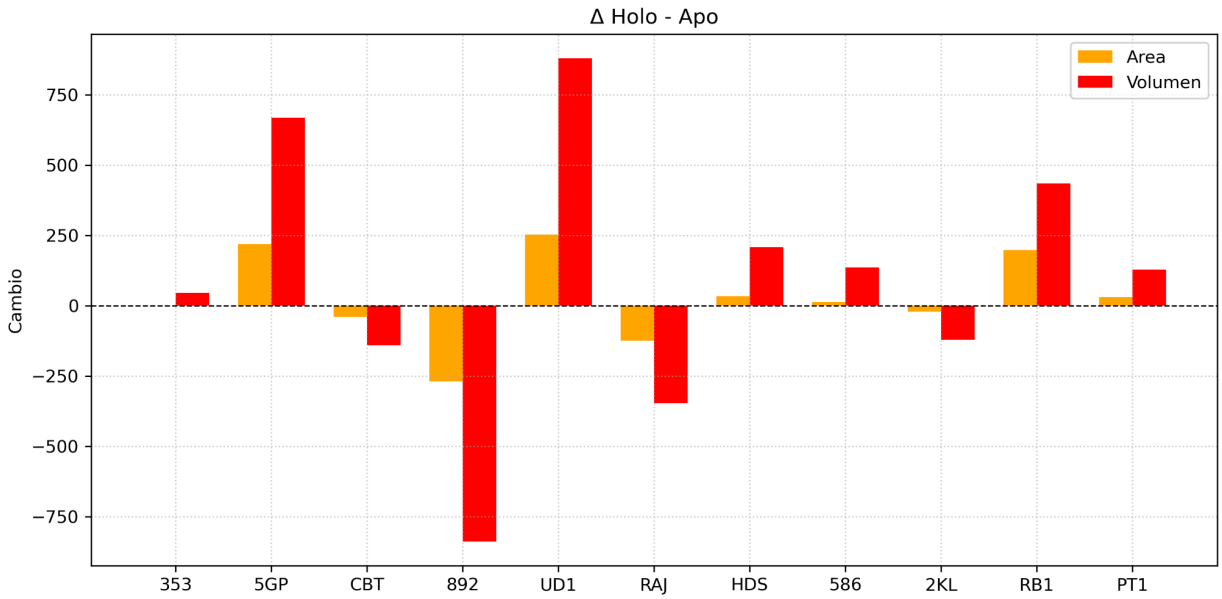


Figura 45 *Variaciones en el área superficial y el volumen del sitio activo entre las estructuras Bound (Holo) y Apo para los diferentes sistemas analizados. Un valor positivo indica una expansión del sitio activo tras la unión del ligando, mientras que un valor negativo sugiere una contracción o cierre del mismo.*

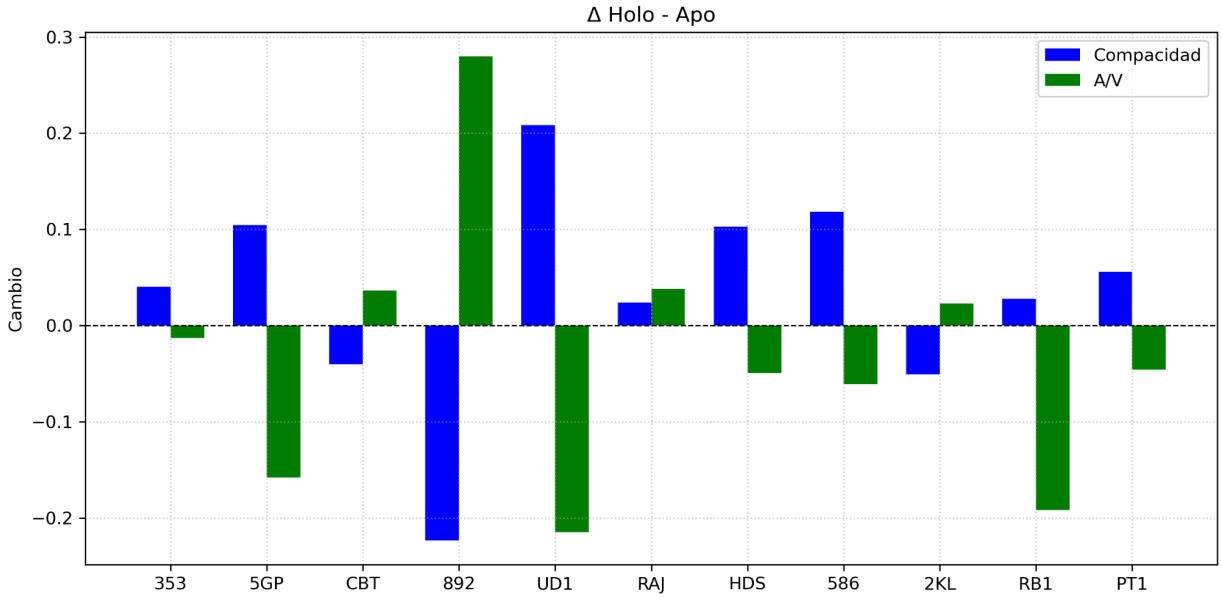


Figura 44 *Variaciones en la compacidad y en la relación área/volumen (A/V) del sitio activo entre las estructuras Bound (Holo) y Apo para los diferentes sistemas analizados. Los valores representan la diferencia (Bound – Apo). Un aumento en la compacidad sugiere un sitio más cerrado y organizado tras la unión del ligando, mientras que una disminución en A/V indica una reducción en la exposición superficial relativa del pocket.*

Es interesante destacar que no se identificó un patrón uniforme entre los casos analizados. Por el contrario, **cada par proteína-ligando mostró un comportamiento distintivo**, reflejando distintos tipos de reorganización conformacional frente a la unión del ligando. Esto sugiere que la respuesta estructural del sitio activo es altamente dependiente del contexto molecular de cada sistema, al menos para el conjunto de sistemas estudiados.

El valor **Δ Holo – Apo**, permite cuantificar el **efecto estructural neto de la unión del ligando** sobre el sitio activo. Un **Δ negativo** indica que el valor correspondiente (ya sea volumen, área o compacidad) **disminuye al pasar de Apo a Holo**, lo cual sugiere un **colapso o cierre del sitio activo** como resultado de la interacción con el ligando. Esto puede reflejar un mecanismo de tipo *ajuste inducido*, donde el sitio se adapta estructuralmente para acomodar al ligando, disminuyendo su accesibilidad. En cambio, un **Δ positivo** implica una **expansión o apertura del sitio activo**, posiblemente asociada a una reorganización estructural que permite una mejor acomodación del ligando o la exposición de nuevas regiones de interacción. Por lo tanto, el análisis de estos valores Δ proporciona información valiosa sobre los **cambios conformacionales inducidos por la unión del ligando** y la **dinámica funcional del pocket**.

5.8 Reconstrucción del Sitio Activo - Ricin-PT1

Habiendo analizado que el docking es incapaz de obtener la pose correcta sobre la estructura apo, cuando el cambio conformacional entre esta y la holo es significativo y habiendo caracterizado de manera sistematica las diferencias entre ambas estructuras, el último objetivo de la prueba de concepto de la hipótesis del presente capítulo, involucra tomar como punto de partida la estructura Apo y realizar una serie de cambios que permitieran generar conformaciones que acercan la misma a la estructura holo, siendo esta en principio desconocida, y utilizar el docking como herramienta reveladora de que la conformaciòn que se ha generado efectivamente se asemeja a la holo. Dicho de otro modo, lo que quiero mostrar en este capítulo de tesis es que partiendo de la estructura apo y conociendo los movimientos potenciales del sitio activo y el ligando, el docking es capaz de revelarnos cuando hemos encontrado la estructura holo.

En este contexto, el siguiente paso consistió entonces en reconstruir el sitio activo a partir de la información recolectada en las etapas previas de comparación entre las estructuras Holo y Apo. Para ello, se desarrolló un script en Python que modifica los ángulos chi de los residuos del sitio Apo, ajustándose a valores cercanos a los correspondientes en la estructura Holo. Se decidió modificar únicamente aquellos residuos cuya diferencia en el ángulo chi superará los 10 grados. A modo de ejemplo, en la ***Figura 46*** se muestra el “nuevo” sitio activo Apo del receptor de andrógenos (en color rojo), superpuesto sobre la estructura Holo correspondiente. Puede observarse que el sitio adopta una disposición casi idéntica en los residuos modificados.

Otra medida que se re calculo para confirmar la modificación del sitio activo fue el área y Volumen para el sitio modificado, se observa una disminución en los deltas de ambas áreas, en este caso no llegan a ser exactamente iguales probablemente por que en esta medición se

incluyen residuos que no fueron modificados y también los modificados no fueron variados sus backbones (***Tabla 17***)

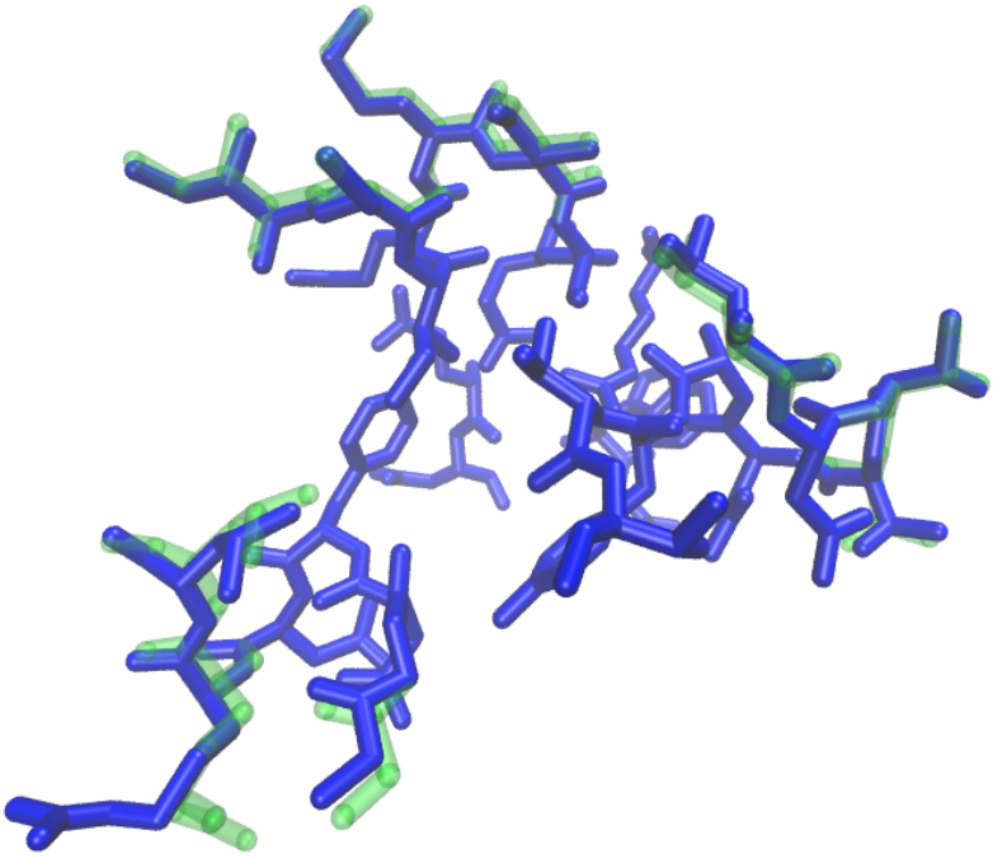


Figura 46 Superposición de los sitios activos de la proteína Ricin. En azul se muestra el sitio reconstruido a partir de la estructura Apo (1RTC), mientras que en verde se representa el sitio original correspondiente a la estructura Holo (1BR6). Se observa que los ángulos modificados en la estructura reconstruida quedan totalmente superpuestos en el sitio activo, adoptando la misma posición que en la estructura Holo.

Ricin - PT1	Área	Volumen
Δ Holo - Apo	30.765	127.851
Δ Holo - Apo Mod	51.179	100.03

Tabla 17. Se observa el Δ de variación de las Áreas y volúmenes entre los Caso Holo - Apo y Holo - Mod Apo , se observa una disminución en el caso modificado confirmando la modificación y similitud del sitio.

Finalmente, sobre la estructura modificada se realizó el docking del ligando con el objetivo de evaluar si era posible reproducir el resultado observado en la estructura Holo , en el ***Figura 48*** se presenta el resultado, donde se observa una mejora tanto en la energía de unión como en la población del clúster principal, en comparación con la versión Apo. Sin embargo, estos valores aún no alcanzan los niveles obtenidos con la estructura Holo. Asimismo, se evidenció una mejora en el valor de RMSD, que pasó de 4.31 a 3.48 Å. Un análisis más detallado de la interacción (***Figura 47***) revela que el ligando se posiciona correctamente dentro del sitio activo, aunque adopta una pose invertida respecto a la orientación observada en la estructura Holo.

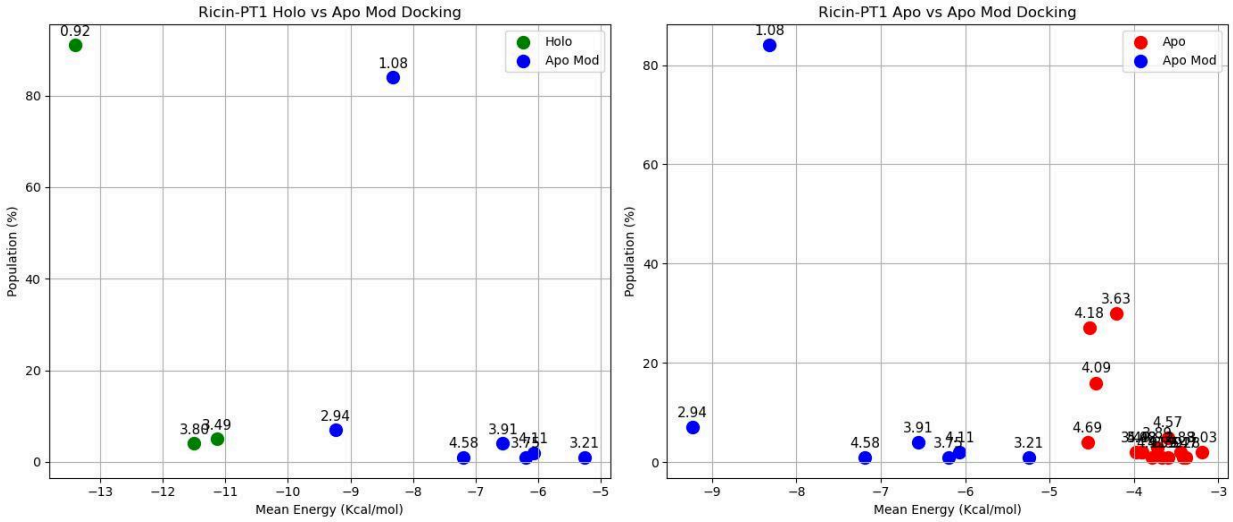


Figura 48 Comparación de los resultados de docking entre la estructura Holo con bias (azul) y la estructura modificada (verde). Se representa la energía media de los clústeres frente a su población. Los valores numéricos corresponden al RMSD respecto a la estructura Holo.

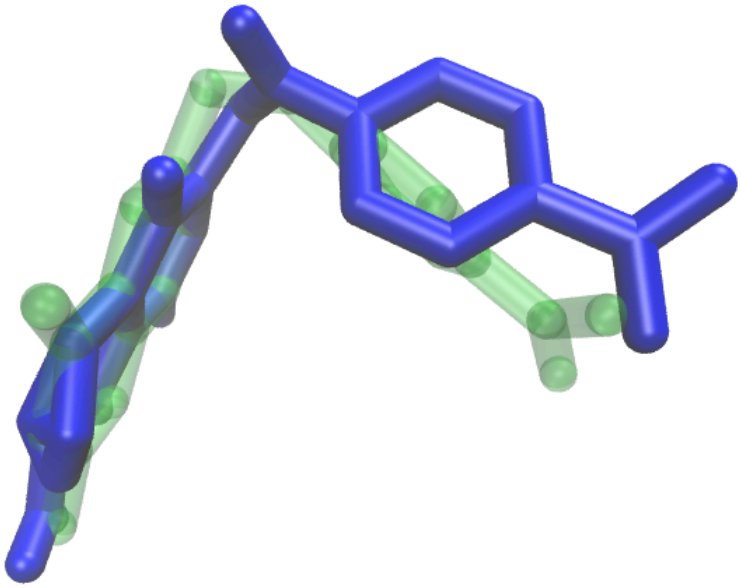


Figura 47 Comparación de la orientación del ligando tras el redocking sobre el sitio modificado, se observa la estructura Holo original (verde) y en Azul el resultado sobre el sitio modificado , con un rmsd de 1.08. Se observa que la modificación del sitio activo se asemeja al original

Por último se evaluaron las interacciones entre el sitio activo y el ligando, encontrando la misma interacción que se encontraba originalmente pero interactuando con otro átomo del ligando (***Figura 49***) , en la ***Tabla 18*** se resumen las interacciones encontradas en los 3 casos (Holo , Apo y Apo mod)

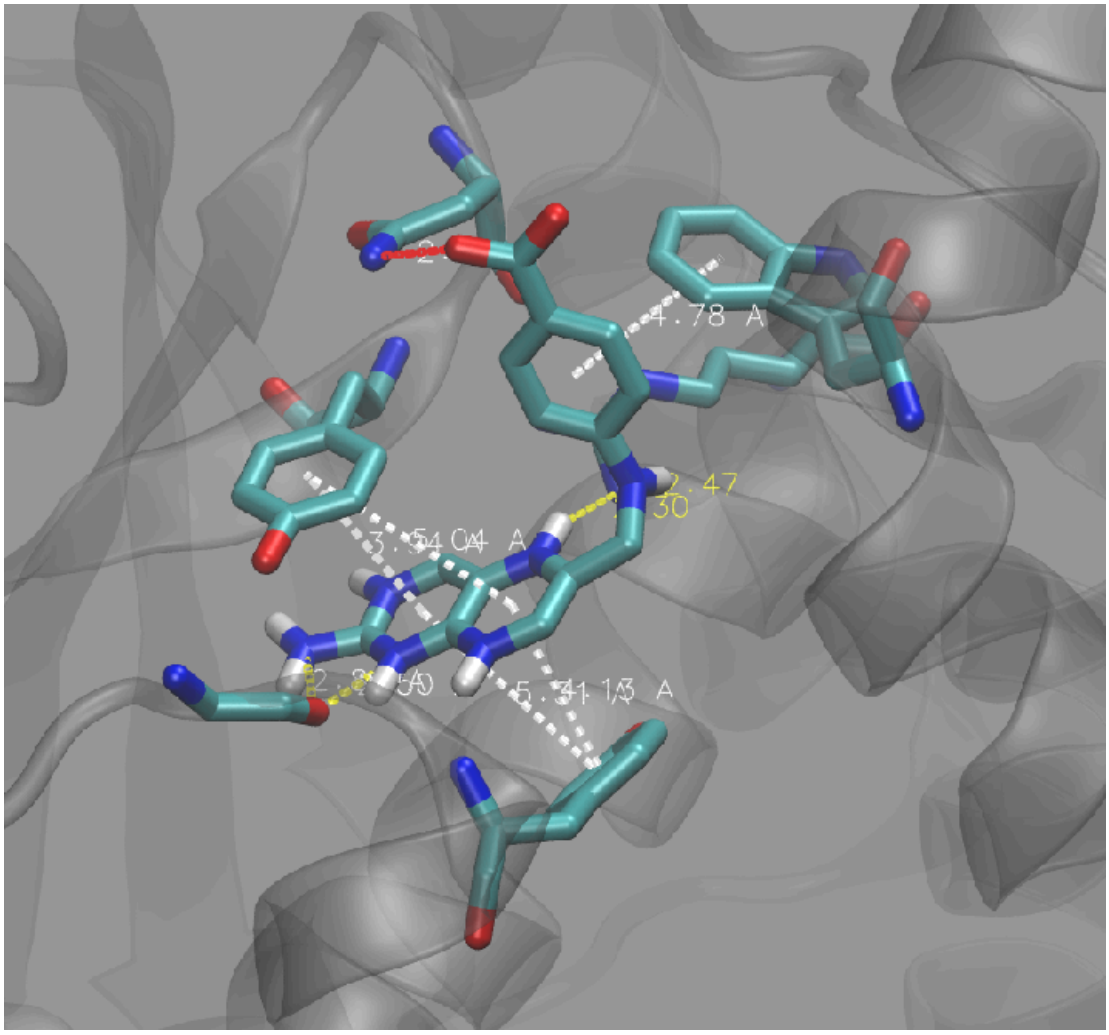


Figura 49 Interacción entre el ligando y el sitio Apo modificado

Interacción	Holo	Apo Mod	Apo
Puente Hidrógeno	7	6	4
Aromática	3	5	0
Total	10	11	4

Tabla 18 Resumen comparativo de las interacciones identificadas entre el receptor y el ligando en las tres conformaciones analizadas: Holo (estructura cristalográfica), Apo Mod y Apo (mejor pose de docking). Se indican la cantidad de interacciones por tipo (puentes de hidrógeno y aromáticas).

De esta etapa puede concluirse que es posible aproximarse a la conformación observada en la estructura Holo partiendo de la estructura Apo. Si bien, en este ejemplo, no se logra replicar exactamente la estructura original, se obtiene una configuración similar de manera rápida y sin necesidad de recurrir a simulaciones de dinámica molecular o procesos de minimización de energía. No obstante, el método aún presenta margen de mejora, especialmente en lo referido a la incorporación de variaciones en el backbone y al perfeccionamiento en la modificación de los ángulos, los cuales todavía generan ciertos artefactos estructurales.

5.9 Reconstrucción del Sitio Activo - Global

Esta modificación del sitio activo se aplicó de manera sistemática a todo el conjunto de sistemas analizados. En la ***Tabla 19*** se resumen los resultados de los docking realizados sobre los sitios activos modificados para cada sistema. En líneas generales, se observa una mejora significativa respecto a los resultados obtenidos sobre las estructuras Apo originales, con valores de energía de unión, población de clústeres, y lo más importante, valores de RMSD contra la estructura de referencia más similares a los obtenidos en las estructuras Holo. Esto muestra que es posible a partir del análisis de la estructura apo y los movimientos usuales observados entre conformaciones holo y apo, transformar una en la otra.

Sistema	Población (%) (Bias)	RMSD Apo M.	Interacciones Apo M.	RMSD Holo	Interacciones Holo
Adipocyte	49	1.18	2	1.01	2
Androgen Receptor	43	1.52	1	1.48	1
B - Lactasa	84	2.34	1	2.33	2
C-MET	55	3.53	6	0.81	6
Guanylate	99	2.10	7	2.15	6
HSP 90	87	3.90	3	4.10	3
PTP1B	48	4.82	6	1.12	5
Ricin	84	1.08	11	0.95	10
TIE-2	98	0.93	5	0.76	5

UAPI	67	4.33	3	0.98	7

Tabla 19 Resumen comparativo de los resultados de docking obtenidos para cada sistema utilizando la estructura Apo modificada. Con % Población ,RMSD e interacción es del caso modificado, y su comparativa con el caso Holo

En la ***Figura 50*** se compara el RMSD de las poses obtenidas respecto a la estructura cristalográfica para cada sistema, excluyendo tres casos específicos para los que no se logró una recuperación adecuada. En el resto de los sistemas analizados, la modificación del sitio activo permitió recuperar poses muy similares a la cristalográfica, Los casos en los cuales no se logró una mejora sustancial se detallan en el apartado correspondiente, evidenciando las limitaciones del enfoque en ciertos contextos estructurales más complejos.
En la figura se compara la variación entre los RMSD de la estructura original (Verde) y Apo modificado (Verde). En la misma se observa en casi todos los casos valores similares lo que hace intuir una poses similares, salvo en 3 casos (C-MET , PTP1B y UAPI) donde no se obtienen estructuras similares. Es interesante destacar que en estos tres casos los ligandos son grandes, que poseen mayores puntos de interacción y mayor número de ángulos posibles, siendo probablemente más sensibles a las variaciones del sitio activo.

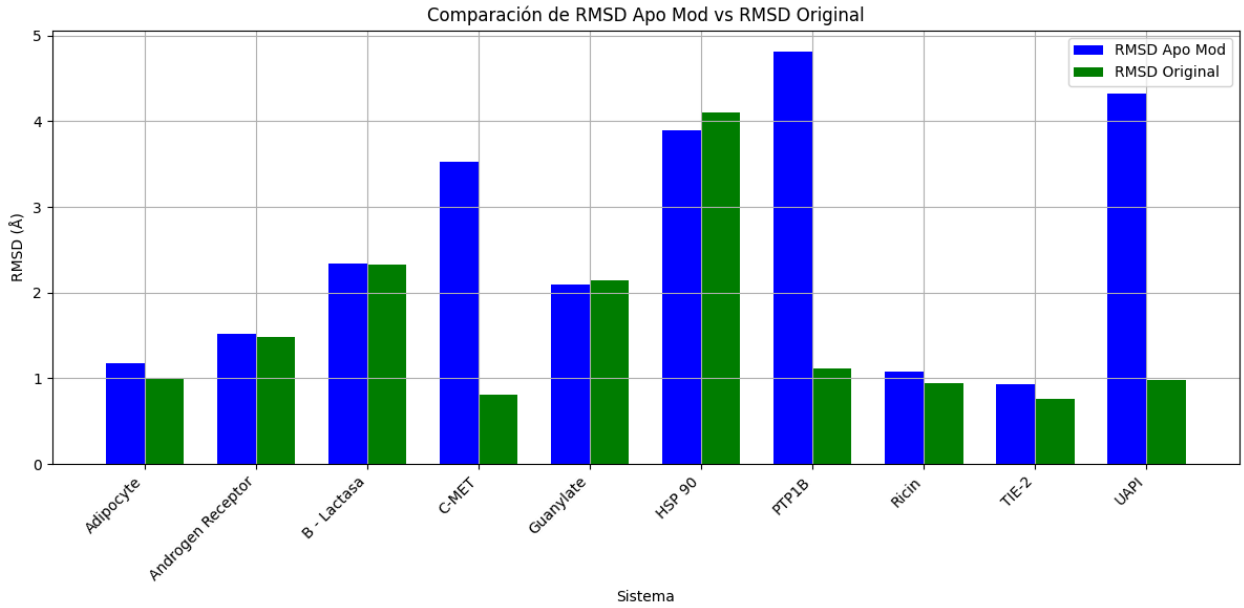


Figura 50 Comparación de los valores de RMSD obtenidos para las poses generadas mediante docking sobre la estructura original (en azul) y la estructura Apo modificada (en verde) respecto a la conformación cristalográfica de referencia. En la mayoría de los casos, la modificación racional del sitio activo permitió obtener poses más cercanas a la estructura experimental.

Por último, se comparó el número de interacciones formadas en las poses obtenidas a partir de la estructura original y la estructura modificada (***Figura 51***). En general, se observó un número **prácticamente equivalente de interacciones** en ambos casos, con algunas excepciones puntuales en las que de todos modos distintos átomos del ligando interactuaron con los mismos

residuos del sitio activo. Estas diferencias podrían atribuirse a la **flexibilidad intrínseca de ciertos residuos voluminosos y flexibles (i.e Leu o Ile)**, que permiten ajustes locales sin afectar significativamente el patrón global de interacción.

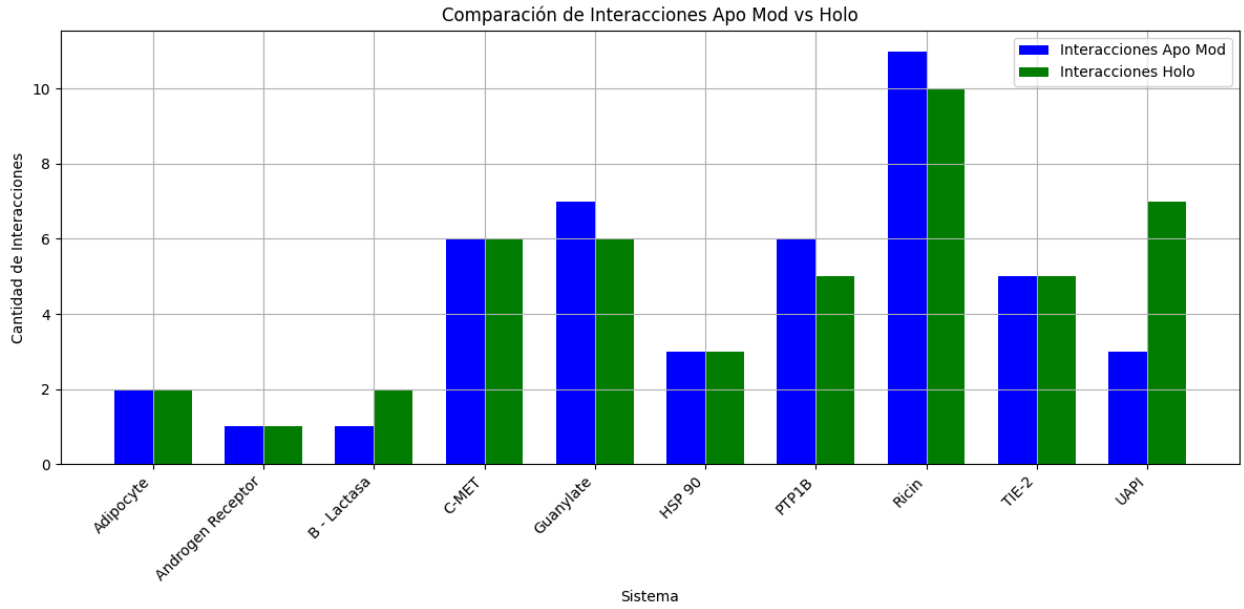


Figura 51 Comparación del número total de interacciones proteína-ligando observadas para cada sistema, obtenidas a partir del docking sobre la estructura Holo (en verde) y la estructura Apo modificada (en azul). En general, se observan cantidades similares de interacciones en ambos casos, con algunas variaciones atribuidas a ligeros desplazamientos de los ligandos o a la flexibilidad local de residuos voluminosos del sitio activo.

5.10 Conclusión

En este capítulo se realizó un **análisis pormenorizado del sitio activo**, centrado en los cambios conformacionales que ocurren ante la presencia de un ligando, con el objetivo de evaluar si es posible identificar, al menos parcialmente, las modificaciones estructurales que **facilitan la interacción entre el sitio activo y el ligando**, utilizando como herramienta principal el **docking molecular, y su versión sesgada (bias docking)**.

A partir de los resultados obtenidos, se observa que las **modificaciones del sitio activo** ocurren principalmente a nivel de la **orientación de los residuos que lo conforman**. Estos cambios no siguen un patrón uniforme ni predecible: pueden involucrar tanto regiones altamente conservadas como zonas con mayor flexibilidad estructural, sin que se identifique una tendencia clara. Algo similar se observa al analizar los cambios en el **área superficial y/o volumen** del pocket; si bien se podría suponer que la unión del ligando induce una apertura del sitio activo, los datos no muestran un comportamiento sistemático que lo confirme.

No obstante, al aplicar modificaciones racionales sobre los ángulos de torsión (particularmente los ángulos χ de residuos clave), fue posible **reconstruir eficazmente el sitio**

activo en su conformación ligada, recuperando **poses e interacciones altamente similares a las presentes en la estructura cristalográfica de referencia**.

Estos resultados demuestran el potencial de esta estrategia para mejorar la predicción de unión en ausencia de estructuras Holo, aunque también subrayan la necesidad de desarrollar herramientas que permitan **automatizar este procedimiento**, evaluando de manera sistemática las posibles conformaciones que pueden adoptar los residuos del sitio activo. Este constituye un desafío relevante para futuras investigaciones, especialmente en el contexto de distintas **familias de proteínas con alta flexibilidad local**.

Discusión y Perspectivas Futuras

6.1 Discusión

En esta tesis doctoral se desarrollaron y aplicaron estrategias avanzadas en docking molecular con el objetivo de mejorar las limitaciones clásicas asociadas a la rigidez del receptor y la precisión de la predicción proteína-ligando. El trabajo se estructuró en tres grandes ejes: primero, se optimizó el virtual screening sobre el complejo bacteriano LolCDE mediante la identificación de sitios de solvente y la consecuente implementación de Bias Docking, mejorando así la selección de inhibidores potenciales para esta proteína. En segundo lugar, se integraron el modelado estructural mediante AlphaFold, análisis filogenéticos, y docking molecular para predecir sustratos de citocromos P450 bacterianos (BacCYPs) directamente a partir de su secuencia, logrando inferencias funcionales robustas. Finalmente, se abordó de manera explícita el impacto de la flexibilidad del sitio activo en el éxito del docking, demostrando que la modificación controlada de estructuras Apo puede aproximar los resultados a los obtenidos con conformaciones Holo.

En el capítulo dedicado al ***Virtual Screening en LolCDE*** se abordó uno de los principales desafíos en el diseño de antibióticos: la identificación de inhibidores efectivos frente a blancos moleculares complejos, ya identificados y seleccionados. Recordemos que LolCDE es una proteína presente en bacterias Gram-negativas, responsable del transporte de lipoproteínas desde la membrana interna hacia el periplasma. Este complejo forma parte de un sistema esencial para la viabilidad bacteriana, resultando en un blanco atractivo para el desarrollo de nuevos antimicrobianos debido a su rol fundamental en la biogénesis de la membrana externa y, especialmente, por estar ausente en células humanas, lo cual minimiza el riesgo de efectos adversos derivados de la inhibición cruzada.El docking molecular convencional aplicado a estos sistemas presenta limitaciones significativas, principalmente debido a la baja definición de los sitios de unión y a la falta de consideración de la dinámica conformacional de las proteínas. Estas restricciones afectan negativamente la precisión del virtual screening, generando una elevada proporción tanto de falsos positivos, cómo de falsos negativos.

Para superar estas limitaciones, se implementó una estrategia combinada que integró el mapeo de sitios de solvente mediante la realización de simulaciones de dinámica molecular en presencia de cosolventes, seguido de la incorporación de esta información en los cálculos de docking mediante la estrategia de Bias Docking. Se aplicaron mapas de energía modificados para favorecer interacciones con sitios tanto hidrofílicos cómo hidrofóbicos previamente detectados en las simulaciones, buscando así mejorar la especificidad y relevancia biológica de las poses predichas. Los resultados obtenidos, y el análisis de controles positivos, evidenciaron que esta aproximación permitió reducir significativamente el número de falsos positivos, uno de los mayores problemas en estrategias de búsqueda virtual. El filtrado de los datos a través del análisis de Z-scores, combinado con la evaluación de las interacciones proteína-ligando,

permitió optimizar la identificación de compuestos de interés. Actualmente, la confirmación experimental de estos resultados se encuentra en curso, llevada a cabo por un grupo colaborador, lo que permitirá validar la eficacia del enfoque propuesto.

En este capítulo, la investigación se centró en el análisis de los BacCYPs, proteínas clave en las rutas biosintéticas de productos naturales microbianos. Determinar su rol preciso en la producción de estos metabolitos es esencial para comprender, no solo su biosíntesis, sino también su enorme potencial biotecnológico. En los últimos años, la disponibilidad de genomas microbianos ha aumentado exponencialmente, dando lugar a la anotación actual de más de 100.000 secuencias de BacCYPs. Sin embargo, en la mayoría de los casos, dicha anotación se limita a la asignación a una familia de proteínas, en este caso BacCYP, sin información sobre su función específica o tipo de sustrato.

Ante este escenario, predecir *in silico* el rango de posibles sustratos y reacciones catalizadas por cada BacCYP se presenta como un desafío complejo y estratégico. Este capítulo muestra cómo una estrategia integradora que combina técnicas de *docking* molecular con herramientas bioinformáticas (como el análisis filogenético y del contexto genómico), y quimio informáticas (como la comparación de estructuras químicas usando el índice de Tanimoto), puede aportar soluciones efectivas a este interesante problema. Esta aproximación multidimensional permite acotar de manera significativa el universo de sustratos posibles, generando hipótesis funcionales fundamentadas, para luego avanzar en su validación experimental.

Los resultados obtenidos muestran que la identificación de los sustratos y la capacidad predictiva del método desarrollado representan un avance sustancial. Sin embargo, es importante remarcar que, aunque la estrategia demuestra gran potencial, los hallazgos deben interpretarse como una prueba de concepto. En este sentido, se destaca que los continuos avances en genómica, transcriptómica y metabolómica permitirán enriquecer las bases de datos de referencia, lo cual fortalecerá y enriquecerá las predicciones de sustrato, y mejorará sustancialmente el enfoque propuesto en esta tesis.

Desde un punto de vista más general, la estrategia desarrollada pone de manifiesto el valor de los métodos bioinformáticos que integran diferentes capas de información: por un lado, el caudal creciente de secuencias y anotaciones funcionales; por el otro, el detalle estructural derivado de modelos tridimensionales de proteínas. Esta convergencia permite ir más allá de la asignación genérica de una proteína a una familia, y comenzar a abordar, con mayor precisión, su función bioquímica específica y su rol en un contexto biológico específico.

Además del interés académico, la predicción de sustratos en BacCYPs tiene implicancias biotecnológicas concretas. Muchos de estos enzimas participan en la biosíntesis de compuestos bioactivos con aplicaciones potenciales como antibióticos, antitumorales o biosurfactantes. Comprender qué sustratos pueden modificar o transformar estos compuestos, no solo permite descifrar rutas metabólicas desconocidas, sino también identificar blancos para ingeniería metabólica o diseño racional de nuevas moléculas. En este contexto, herramientas como AlphaFold —capaz de generar modelos estructurales de alta precisión incluso en ausencia de estructuras experimentales—, y el uso de contextos génicos conservados, representan una oportunidad concreta para extender este tipo de análisis a BacCYPs aún no caracterizados.

En resumen, este capítulo plantea una metodología robusta y escalable para abordar la predicción de sustratos en BacCYPs, con potencial de aplicación en estudios funcionales, descubrimiento de nuevos productos naturales y biotecnología racional.

Finalmente, el tercer problema tratado se encuentra vinculado a la **flexibilidad del sitio activo, donde aún** persisten numerosos desafíos abiertos. Abordar la representación más realista de la flexibilidad estructural en los protocolos de docking constituye una línea de investigación crítica en el área de desarrollo de inhibidores.

Considerando que los métodos tradicionales de docking molecular asumen una conformación rígida del receptor, lo cual constituye una limitación importante cuando el éxito de la unión depende de cambios conformacionales sutiles o significativos, especialmente en el sitio activo. En primer lugar, en la presente tesis se analizó comparativamente la variabilidad conformacional de sitios activos entre estructuras Apo y Holo. Se caracterizaron las diferencias en los ángulos chi de residuos del sitio activo, y se realizaron modificaciones racionales de las estructuras Apo, orientadas a reproducir parcialmente las características observadas en las estructuras Holo. Se evaluó el impacto de estas modificaciones sobre los resultados de docking, incluyendo la energía de unión, el RMSD contra la estructura de referencia y población de los clústeres de cada pose obtenida. La estrategia adoptada consistió en identificar y modificar racionalmente los principales cambios en el sitio activo observados entre estructuras Apo y Holo, centrándose especialmente en los ángulos χ de residuos críticos para la unión. Posteriormente, se analizaron los efectos de estas modificaciones sobre la calidad y estabilidad de las poses obtenidas.

Los resultados obtenidos demostraron que, aunque no siempre es posible reproducir completamente el comportamiento de una estructura Holo partiendo de su forma Apo, la introducción de ajustes conformacionales dirigidos permite mejorar sustancialmente la predicción de unión, o sea los resultados del docking. En particular, se observó una disminución significativa del RMSD de las poses respecto a la estructura de referencia y un aumento en la ocupación de clústeres representativos del estado Holo. Estos hallazgos refuerzan la noción de que la flexibilidad local en el sitio activo, incluso aplicada de manera parcial y dirigida, puede ser suficiente para potenciar la performance del docking en escenarios donde el receptor experimenta cambios conformacionales al interactuar con el ligando.

Es importante destacar, que si bien en la presente tesis el número de casos estudiados fue limitado, esto permite analizarlos en detalle, y buscar, como muestran los resultados, generar una primer prueba de concepto, que demuestra la validez y el potencial de la estrategia seleccionada y la hipótesis de trabajo subyacente.

En conjunto, este estudio destaca la importancia de considerar explícitamente la flexibilidad del sitio activo en protocolos de docking, proponiendo un enfoque práctico para mejorar predicciones de unión partiendo de estructuras Apo, sin necesidad de recurrir inicialmente a simulaciones de dinámica molecular extensivas. Esta línea de trabajo constituye además una base conceptual sólida para futuros desarrollos orientados a la integración de flexibilidad en esquemas de docking de alta eficiencia.

6.2 Perspectivas Futuras

Existen muchos estilos y tipos de tesis doctorales, algunas resuelven un problema específico, otras proponen un desarrollo tecnológico, otras utilizan herramientas existen para responder interrogantes sobre la biología. El presente trabajo de tesis, me permitió durante su desarrollo navegar los diferentes aspectos y abordajes de las mismas, siempre centrado en un objetivo específico, mejorar los métodos de docking. Sin embargo, habiendo llegando al final del recorrido, la misma se presenta como una tesis que, habiendo trabajado tres problemas específicos:

i) búsqueda de inhibidores de LolCDE,

ii) identificación de los sustratos de los BacCYPs

iii) tratamiento de la flexibilidad del receptor en los esquemas de docking

Representa un sòlido punto de partida, y prueba de concepto bioinformática (o in-silico) para el desarrollo y avance futuro del abordaje de estos tres ejes, tanto por nuevas técnicas bioinformáticas, como por su inesquivable verificación experimental. En este contexto, la principal perspectiva futura que surge del trabajo es la **validación experimental** de los resultados obtenidos, especialmente en lo referido a los sustratos predichos en el capítulo 3 y a los compuestos/sustratos identificados en el capítulo 4. Asimismo, el enfoque de reconstrucción racional del sitio activo explorado en el capítulo 5, cuya eficacia fue evaluada mediante simulaciones de *docking*, representa una línea prometedora para ser desarrollada y validada en profundidad en trabajos posteriores.

En relación con los potenciales inhibidores del complejo LolCDE identificados en esta tesis, los resultados obtenidos representan un primer paso hacia la exploración de este blanco terapéutico poco explotado, pero de gran relevancia en bacterias Gram-negativas. Los compuestos seleccionados constituyen *hits* iniciales que, si bien fueron priorizados mediante criterios estructurales y energéticos, requieren validación experimental mediante ensayos de afinidad y actividad antibacteriana. En este sentido, la implementación de pruebas de inhibición de crecimiento, ensayos de MIC y validaciones funcionales en modelos celulares representaría

un avance clave. Además, la estrategia utilizada en este capítulo —que incluye filtrado basado en similitud, modelado estructural y *docking* racional— podría aplicarse en el futuro a otros sistemas de transporte lipoproteico o complejos homólogos en otras especies patógenas, extendiendo así el impacto metodológico de este enfoque. Esperamos, que en el contexto de la colaboración con GARDP algunos de los mismos sean evaluados en el futuro cercano.

Un desafío a futuro es mejorar la integración de enfoques computacionales en el pipeline farmacéutico, habiendo ya algunos casos como el de la abaucina [66] demuestra que la IA y el **docking** pueden integrarse exitosamente en la búsqueda de antibióticos. La combinación de *machine learning* para explorar el espacio químico, sumada a herramientas como la predicción de estructura proteica (AlphaFold) y simulaciones de acoplamiento molecular, permite identificar y optimizar *hits* de manera más rápida y eficiente que los métodos tradicionales

Este enfoque computacional, siempre validado con experimentos rigurosos, probablemente generará más candidatos focalizados en dianas específicas de patógenos difíciles. A medida que mejoren los modelos predictivos, podríamos ver una enorme aceleración en la **tasa de descubrimiento de lead compounds** contra objetivos antes considerados "intratables" en bacterias Gram-negativas.

Por su parte, la identificación de los sustratos de BacCYPs puede enmarcarse en proyectos de **descubrimiento de productos naturales**, **anotación funcional de genomas microbianos** y **bioingeniería metabólica** orientada a la producción de compuestos bioactivos. Dada la enorme cantidad de BacCYPs sin caracterizar presentes en genomas bacterianos, la predicción *in silico* de sus sustratos constituye una herramienta estratégica para priorizar blancos experimentales, acelerar la anotación funcional y abrir nuevas rutas de exploración metabólica.

Desde una perspectiva **biotecnológica**, los BacCYPs representan enzimas de interés por su capacidad para realizar reacciones químicas difíciles de reproducir sintéticamente, como hidroxilaciones, epoxidaciones y desmetilaciones. Por ello, la caracterización de nuevos BacCYPs con funciones bien definidas puede contribuir al desarrollo de plataformas de **biocatálisis selectiva**, así como al **rediseño de rutas metabólicas** para producir antibióticos, antifúngicos, hormonas vegetales o metabolitos secundarios con valor industrial o farmacéutico.

Asimismo, este tipo de estudios se alinea con iniciativas globales como la **bioprospección de microbiomas** (por ejemplo, en suelos, ambientes extremos o microbiota de insectos) y el desarrollo de **biosíntesis racional de compuestos naturales**, por lo que futuras colaboraciones interdisciplinarias con laboratorios de microbiología, química de productos naturales, o biocatálisis podrían acelerar la transición de estas predicciones *in silico* hacia aplicaciones concretas.

Los resultados obtenidos en esta tesis constituyen una **prueba de concepto bioinformática** para la predicción de sustratos de BacCYPs, utilizando un enfoque integrador que combina análisis filogenético, contexto genómico y modelado estructural con *docking*

racional. Esta estrategia permitió reducir significativamente el universo de posibles sustratos para enzimas sin caracterización previa, facilitando hipótesis funcionales fundamentadas que pueden guiar futuras validaciones experimentales. En este sentido, el diseño de experimentos que incluyan la **expresión heteróloga de los BacCYPs priorizados**, la realización de **ensayos de biotransformación con sustratos candidatos** y el análisis de productos por **técnicas analíticas** (como LC-MS o RMN), representa el siguiente paso lógico en la transición desde el modelo *in silico* hacia su validación *in vitro*.

Más allá del valor académico de la anotación funcional, los BacCYPs tienen un **alto potencial biotecnológico**. Por ejemplo, los P450 de Clase VII —como CYP102A1 (BM3)— han sido utilizados como plataformas autosuficientes para biocatálisis industrial, y mediante ingeniería dirigida han demostrado la capacidad de realizar **oxidaciones complejas de forma regio- y estereoselectiva**, lo que los convierte en herramientas atractivas para la síntesis de derivados farmacéuticos [67]. La estrategia de predicción estructural presentada en esta tesis podría aplicarse directamente al diseño de nuevas variantes de BacCYPs, optimizadas para modificar sustratos específicos de interés industrial o farmacológico, especialmente en contextos donde se busca funcionalización selectiva difícil de lograr por medios químicos tradicionales.

Además, estudios recientes han documentado el papel de los BacCYPs en **procesos de biodegradación ambiental**, mostrando que ciertas enzimas bacterianas pueden catalizar reacciones de deshalogenación o desnitración de contaminantes orgánicos persistentes, incluso bajo condiciones de bajo oxígeno. Esta capacidad de los BacCYPs para oxidar, reducir o detoxificar compuestos tóxicos posiciona a estos enzimas como **candidatos clave para aplicaciones en biorremediación**, en particular si se combinan con enfoques de ingeniería de enzimas o biología sintética. En este contexto, la metodología desarrollada en este trabajo podría emplearse para identificar BacCYPs con potencial en biodegradación ambiental, guiando la selección racional de enzimas para ensayos funcionales orientados a la eliminación de pesticidas, solventes clorados u otros xenobióticos [68].

Finalmente, y en relación con el último capítulo, como emerge de los resultados de la presente tesis, uno de los principales desafíos identificados a lo largo de este trabajo es la representación adecuada de la **flexibilidad del receptor** en protocolos de docking molecular. La incapacidad de muchos métodos tradicionales para capturar los cambios conformacionales que acompañan la unión de ligandos limita la precisión en la predicción de poses y afinidades de unión, especialmente en sistemas donde la transición entre estados Apo y Holo implica ajustes significativos en el sitio activo. La importancia de considerar explícitamente la flexibilidad del receptor se ha vuelto cada vez más evidente, y las estrategias basadas en modificaciones racionales de la estructura Apo, como las aplicadas en esta tesis, representan un primer paso en esa dirección. Sin embargo, superar esta limitación de manera sistemática y generalizable requerirá integrar nuevas herramientas que permitan modelar de forma más realista la dinámica estructural de las proteínas.

En este contexto, el desarrollo reciente (y su enorme potencial a futuro) de modelos de **inteligencia artificial** para la predicción de estructuras proteicas, y en particular el surgimiento

de **AlphaFold**[70], ha transformado el panorama de la biología estructural. AlphaFold no solo permite obtener modelos de alta precisión para proteínas individuales, sino que también ha abierto la posibilidad de explorar diferentes estados conformacionales, incluyendo formas alternativas relacionadas con la función biológica o la unión de ligandos. La reciente generación de **AlphaFold-Multistate Models**[69] y desarrollos afines sugiere que será posible en un futuro cercano predecir no solo estructuras estáticas, sino también **paisajes conformacionales** accesibles para cada proteína.

Integrar estas capacidades de predicción estructural avanzadas con estrategias de docking molecular, como el utilizado en la presente tesis, abre un nuevo horizonte para abordar el problema de la flexibilidad del receptor. La combinación de **modelos de IA** que capturen la diversidad conformacional natural de las proteínas, junto con técnicas como **Bias Docking** y **ensemble docking**[71] permitirá construir pipelines de búsqueda virtual que incorporen de manera explícita la flexibilidad del receptor como variable clave. De este modo, el futuro del docking molecular no radica únicamente en mejorar las funciones de puntaje, sino en redefinir el concepto mismo de receptor, pasando de entidades rígidas a **poblaciones conformacionales dinámicas**. Este cambio de paradigma será esencial para aumentar la aplicabilidad y el impacto del docking en el descubrimiento racional de nuevos fármacos y en la caracterización funcional de proteínas de interés biomédico e industrial.

Desde una perspectiva general, quiero además destacar, que la expansión del uso de **modelos de IA** en otros ámbitos del estudio proteína-ligando también está creciendo rápidamente. En particular, existen desarrollos recientes que aplican **redes neuronales profundas** para mejorar directamente el proceso de **docking per se**[72] reemplazando algoritmos tradicionales de búsqueda conformacional. Asimismo, la IA se está utilizando para **rescoring** de resultados de docking, aplicando modelos entrenados en grandes bases de datos experimentales para, de este modo, mejorar la discriminación entre poses correctas e incorrectas, abordando así uno de los cuellos de botella históricos de los métodos de docking.

Finalmente, una línea de expansión particularmente prometedora es la aplicación de **modelos generativos de IA**[73] —como modelos de difusión, redes generativas adversariales (GANs) o modelos auto-regresivos— para **diseñar nuevos ligandos de manera dirigida a pockets proteicos específicos**. Estos modelos permiten crear estructuras químicas novedosas optimizadas para encajar en sitios activos predeterminados, transformando el proceso de descubrimiento de ligandos en un enfoque de diseño racional asistido por IA. De esta manera, la IA no solo se limita a analizar y optimizar interacciones existentes, sino que también comienza a **generar nuevos espacios químicos** adaptados a objetivos biológicos específicos, abriendo un nuevo paradigma en la investigación y el desarrollo de fármacos.

En conjunto, la convergencia de avances en modelado estructural, predicción de interacciones y diseño de moléculas mediante inteligencia artificial promete redefinir el futuro del docking molecular, incrementando su precisión, alcance y aplicabilidad en el descubrimiento racional de nuevos agentes terapéuticos.

Publicaciones

Durante el tiempo de realización de esta tesis doctoral se originaron las siguientes publicaciones:

- **Juan Manuel Prieto,** Jorge Lannot, Camila Clemente, Carlos Modenutti, Adrian Turjanski and Marcelo A. Martí. How and **Why does Knowledge-based Biased Docking improve Molecular Docking Performance?".** Manuscrito en preparación para ser enviado a **J Chem Inf Mode en 2025**.

- **Juan Manuel Prieto**; Gustavo Schottlender; Camila M. Clemente; Rafael Betanzos; Dario Fernández Do Porto; Marcelo A. Marti. Computer-Aided Drug Discovery and Design | Book chapter: ***Docking and Bias Docking.*** 2024.Structure-Based Drug Design. Computer-Aided Drug Discovery and Design, vol 2. Springer DOI: 10.1007/978-3-031-69162-1_5

- Santiago Sosa; Alan M. Szalai; Lucía F. Lopez; **Juan Manuel Prieto**; Cecilia Zaza; Aleksandra K. Adamczyk; Hernán R. Bonomi; Marcelo A. Marti; Guillermo P. Acuna; Fernando A. Goldbaum et al. ***Monitoring Dynamic Conformations of a Single Fluorescent Molecule Inside a Protein Cavity.*** Small Methods (2025). DOI: 10.1002/smtd.202402114

- Gustavo Schottlender; **Juan Manuel Prieto**; Miranda Clara Palumbo; Florencia A. Castello; Federico Serral; Ezequiel J. Sosa; Adrián G. Turjanski; Marcelo A. Martì; Darío Fernández Do Porto. ***From drugs to targets: Reverse engineering the virtual screening process on a proteomic scale.*** Frontiers in Drug Discovery 2022-10-20 | Journal article DOI: 10.3389/fddsv.2022.969983 *Part of* ISSN: 2674-0338

- Gustavo Schottlender; **Juan Manuel Prieto**; Camila Clemente; Claudio David Schuster; Victoria Dumas; Darío Fernández Do Porto; Marcelo Adrian Martí. ***Bacterial cytochrome P450s: a bioinformatics odyssey of substrate discovery. Frontiers in Microbiology 2024-02-07 DOI: 10.3389/fmicb.2024.1343029 Part of ISSN: 1664-302X***

- Camila M. Clemente, **Juan M. Prieto**, and Marcelo Martí.***Unlocking Precision Docking for Metalloproteins.*** Journal of Chemical Information and Modeling 2024 64 (5), 1581-1592 DOI: 10.1021/acs.jcim.3c01853

Referencias

[1] Kubinyi, H. "***Chance Favors the Prepared Mind - From Serendipity to Rational Drug"***

[2] Talele, T.; Khedkar, S.; Rigby, A. "***Successful Applications of Computer Aided Drug Discovery: Moving Drugs from Concept to the Clinic"***. Curr. Top. Med. Chem. 2010, 10 (1), 127–141.

[3] Amzel, L. M. (1998). ***"Structure-based drug design".*** *Current opinion in biotechnology*, *9*(4), 366-369.

[4] Jorgensen, W. L. (2009). ***"Efficient drug lead discovery and optimization".*** *Accounts of chemical research*, *42*(6), 724-733.

[5]Arcón, Juan Pablo. (2018). ***Determinantes moleculares de la interacción droga-proteína : uso de cosolventes como moléculas de prueba. (Tesis Doctoral. Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales.).*** Recuperado de https://hdl.handle.net/20.500.12110/tesis_n6457_Arcon

[6] Mattos, C.; Bellamacina, C. R.; Peisach, E.; Pereira, A.; Vitkup, D.; Petsko, G. A.; Ringe, D. ***Multiple Solvent Crystal Structures: Probing Binding Sites, Plasticity and Hydration.*** J. Mol. Biol. 2006, 357 (5), 1471–1482.

[7] R. D. Taylor et al., ***A review of protein-small molecule docking methods***, J. Comput. Aided Mol. Des. 16,151–166 (2002).

[8]Oshiro, C. M.; Kuntz, I. D.; Dixon, J. S. ***Flexible Ligand Docking Using a Genetic Algorithm***. J. Comput. Aided. Mol. Des. 1995, 9 (2), 113–130.

[9]Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. ***Development and Validation of a Genetic Algorithm for Flexible Docking.*** J. Mol. Biol. 1997, 267 (3), 727–748.

[10] Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. ***Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function***. J. Comput. Chem. 1998, 19 (14), 1639–1662.

[11] Hart, T. N.; Read, R. J. A ***Multiple-Start Monte Carlo Docking Method***. Proteins: Struct. Funct. Bioinf. 1992, 13 (3), 206–222

[12] Goodsell, D. S.; Olson, A. J. ***Automated Docking of Substrates to Proteins by Simulated Annealing***. Proteins: Struct. Funct. Bioinf. 1990, 8 (3), 195–202.

[13] Truchon JF, Bayly CI. ***Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem.*** J Chem Inf Model. 2007 Mar-Apr;47(2):488-508. doi: 10.1021/ci600426e. Epub 2007 Feb 9. PMID: 17288412.

[14] Juan Pablo Arcon, Carlos P Modenutti, Demian Avendaño, Elias D Lopez, Lucas A Defelipe, Francesca Alessandra Ambrosio, Adrian G Turjanski, Stefano Forli, Marcelo A Marti, ***AutoDock Bias: improving binding mode prediction and virtual screening using known protein–ligand interactions***, *Bioinformatics*, Volume 35, Issue 19, October 2019, Pages 3836–3838

[15] Lionta E, Spyrou G, Vassilatis DK, Cournia Z. ***Structure-based virtual screening for drug discovery: principles, applications and recent advances.*** Curr Top Med Chem. 2014;14(16):1923-38. doi: 10.2174/1568026614666140929124445. PMID: 25262799; PMCID: PMC4443793.

[16]Shoichet, B. ***Virtual screening of chemical libraries***. *Nature* **432**, 862–865 (2004). https://doi.org/10.1038/nature03197

[17]Lavecchia A, Di Giovanni C. ***Virtual screening strategies in drug discovery: a critical review***. Curr Med Chem. 2013;20(23):2839-60. doi: 10.2174/09298673113209990001. PMID: 23651302.

[18]T. Scior et al., J. Chem.***Recognizing Pitfalls in Virtual Screening: A Critical Review.*** Inf. Model. 52, 867–881 (2012)

[19] X. Biarnés et al., J. Comput. Aided Mol. Des. 25,395–402 (2011)

[20] K. Wang et al., J. Comput. Aided Mol. Des. 27,989–1007 (2013)

[21] B. R. Miller III et al., J. Chem. Theory Comput. 8,3314–3321 (2012)

[22] D. B. Kitchen et al., Nat. Rev. Drug Discov. 3,1.935–949 (2004)

[23] D. E. Koshland, Angew.,***The Key–Lock Theory and the Induced Fit Theory***,Chem. Int. Ed Engl. 33,2375–2378 (1995). https://doi.org/10.1002/anie.199423751

[24] Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. A. ***Target Flexibility: An Emerging Consideration in Drug Discovery and Design.*** J. Med. Chem. 2008, 51 (20), 6237–6255.

[25] Radusky et al., 2017; Shoichet, 2004; Ruiz-Carmona et al., 2014

[26] Zhao Y, Li J, Gu H, Wei D, Xu YC, Fu W, Yu Z. ***"Conformational Preferences of π-π Stacking Between Ligand and Protein, Analysis Derived from Crystal Structure Data Geometric Preference of π-π Interaction."*** Interdiscip Sci. 2015 Sep;7(3):211-20. doi: 10.1007/s12539-015-0263-z. Epub 2015 Sep 14. PMID: 26370211.

[27] Kouranov A, Xie L, de la Cruz J, Chen L, Westbrook J, Bourne PE, Berman HM. ***The RCSB PDB information portal for structural genomics.*** Nucleic Acids Res. 2006 Jan 1;34(Database issue):D302-5. doi: 10.1093/nar/gkj120. PMID: 16381872; PMCID: PMC1347482.

[28] Jumper, J., Evans, R., Pritzel, A., et al. (2021). ***Highly accurate protein structure prediction with AlphaFold***. *Nature*, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

[29] Tsuneo Omura and Ryo Sato. ***The Carbon Monoxide-binding Pigment of Liver Microsomes*** I. EVIDENCE FOR ITS HEMOPROTEIN NATURE. THE JOURNAL OF BIOLOGICAL CHEMISTRY, 239(7), 1964.

[30] ***Design.*** J. Recept. Signal Transduct. Res. 1999, 19 (1–4), 15–39.
R. W. Estabrook. ***A PASSION FOR P450s (REMEMBRANCES OF THE EARLY HISTORY OF RESEARCH ON CY-TOCHROME P450).*** Drug Metabolism and Disposition, 31(12):1461–1473, dec 2003. ISSN 0090-9556. doi:10.1124/dmd.31.12.1461.URL http://dmd.aspetjournals.org/cgi/doi/10.1124/dmd.31.12.1461.

[31] I C Gunsalus and G C Wagner. ***Bacterial P-450cam methylene monooxygenase components: cytochromem, putidaredoxin, and putidaredoxin reductase***. Methods in enzymology, 52:166–88, 1978. ISSN 0076-6879. URL http://www.ncbi.nlm.nih.gov/pubmed/672627

[32]McLean, K., Leys, D., Munro, A. (2015). ***Microbial Cytochromes P450. In: Ortiz de Montellano,*** P. (eds) Cytochrome P450. Springer, Cham. https://doi.org/10.1007/978-3-319-12108-6_6

[33] Mary A. Schuler and Stephen G. Sligar. ***Diversities and similarities in P450 systems***. Metal Ions in Life Sciences, 3:1–26, 2007. ISSN 1559-0836.

[34] Girvan HM, Munro AW. ***Applications of microbial cytochrome P450 enzymes in biotechnology and synthetic biology.*** Curr Opin Chem Biol. 2016 Apr;31:136-45. doi: 10.1016/j.cbpa.2016.02.018. Epub 2016 Mar 22. PMID: 27015292.

[35] Kaplan E, Greene NP, Crow A, Koronakis V. ***Insights into bacterial lipoprotein trafficking from a structure of LolA bound to the LolC periplasmic domain.*** Proc Natl Acad Sci U S A. 2018;115: E7389–E7397.

[36] ***https://gardp.org/***

[37] Diogo Santos-Martins; Leonardo Solis-Vasquez;Andreas F Tillack;Michel F Sanner;Andreas Koch*;Stefano Forli*. **Accelerating AutoDock4 with GPUs and Gradient-Based Local Search., J. Chem. Theory Comput. 2021, 10.1021/acs.jctc.0c01006.**

[38] O'Boyle, N.M., Banck, M., James, C.A. *et al.* ***Open Babel: An open chemical toolbox***. *J Cheminform* **3**, 33 (2011). https://doi.org/10.1186/1758-2946-3-33

[39] Landrum, G. 2010. ***"RDKit."*** Q2. https://www.rdkit.org/.

[40] D.A. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, J.T. Berryman, S.R. Brozell, F.S. Carvahol, D.S. Cerutti, T.E. Cheatham, III, G.A. Cisneros, V.W.D. Cruzeiro, T.A. Darden, N. Forouzesh, M. Ghazimirsaeed, G. Giambaşu, T. Giese, M.K. Gilson, H. Gohlke, A.W. Goetz, J. Harris, Z. Huang, S. Izadi, S.A. Izmailov, K. Kasavajhala, M.C. Kaymak, I. Kolossv\'a ry, A. Kovalenko, T. Kurtzman, T.S. Lee, P. Li, Z. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, A. Shajan, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, J. Wang, X. Wu, Y. Wu, Y. Xiong, Y. Xue, D.M. York, C. Zhao, Q. Zhu, and P.A. Kollman (2025), ***Amber 2022***, University of California, San Francisco.

[41] MacKerell, Bashford, Bellott, Dunbrack, Evanseck, Field, Fischer, Gao, Guo, Ha, et al, J Phys Chem, 102, 3586 (1998)

[42] Kapli P, Yang Z, Telford MJ. ***Phylogenetic tree building in the genomic age***. Nat Rev Genet. 2020 Jul;21(7):428-444. doi: 10.1038/s41576-020-0233-0. Epub 2020 May 18. PMID: 32424311.

[43] Bogusz, Marcin and Simon Whelan. ***"Phylogenetic Tree Estimation With and Without Alignment: New Distance Methods and Benchmarking."*** *Systematic Biology* 66 (2016): 218–231.

[44] Criscuolo, A., Gribaldo, S. ***"BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments".*** *BMC Evol Biol* **10**, 210 (2010). https://doi.org/10.1186/1471-2148-10-210

[45] Lefort V, Longueville JE, Gascuel O. "***SMS: Smart Model Selection in PhyML"***. Mol Biol Evol. 2017 Sep 1;34(9):2422-2424. doi: 10.1093/molbev/msx149. PMID: 28472384; PMCID: PMC5850602.

[46] Stéphane Guindon, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, Olivier Gascuel, ***"New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0"***, *Systematic Biology*, Volume 59, Issue 3, May 2010, Pages 307–321, https://doi.org/10.1093/sysbio/syq010

[47] Daniel S. Wigh, Jonathan M. Goodman, Alexei A. Lapkin. ***"A review of molecular representation in the age of machine learning"*** **Advanced Review** **https://doi.org/10.1002/wcms.1603**

[48] ***Natural Product Scores and Fingerprints Extracted from Artificial Neural Networks*** - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Example-of-a-molecule-being-translated-into-a-fingerprint-The-presence-of-specific_fig1_353574903

[49] Bajusz, D., Rácz, A. & Héberger, K. "***Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?"***. *J Cheminform* **7**, 20 (2015). https://doi.org/10.1186/s13321-015-0069-3

[50] O'Boyle, N.M., Banck, M., James, C.A. *et al.* ***Open Babel: An open chemical toolbox***. *J Cheminform* 3, 33 (2011). https://doi.org/10.1186/1758-2946-3-33

[51] Riniker, Sereina, and Gregory A. Landrum. 2015. ***"Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation."*** *Journal of Chemical Information and Modeling* 55 (12): 2562–74.

[52] Tang, X., Chang, S., Zhang, K. *et al.* ***Structural basis for bacterial lipoprotein relocation by the transporter LolCDE***. *Nat Struct Mol Biol* 28, 347–355 (2021). https://doi.org/10.1038/s41594-021-00573-x

[53] Teague Sterling and John J. Irwin *"**ZINC 15 – Ligand Discovery for Everyone**"* Journal of Chemical Information and Modeling 2015 *55* (11), 2324-2337 DOI: 10.1021/acs.jcim.5b00559

[54] Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, Maria Paula Magarinos, Nicolas Bosc, Ricardo Arcila, Tevfik Kizilören, Anna Gaulton, A Patrícia Bento, Melissa F Adasme, Peter Monecke, Gregory A Landrum, Andrew R Leach, ***"The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods",*** *Nucleic Acids Research*, Volume 52, Issue D1, 5 January 2024, Pages D1180–D1192, https://doi.org/10.1093/nar/gkad1004

[55] Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. ***"Reoptimization of MDL Keys for Use in Drug Discovery"*** Nourse Journal of Chemical Information and Computer Sciences 2002 *42* (6), 1273-1280 DOI: 10.1021/ci010132r

[56] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. ***"Open Babel: An open chemical toolbox"***. J Cheminform. 2011 Oct 7;3:33. doi: 10.1186/1758-2946-3-33. PMID: 21982300; PMCID: PMC3198950.

[57] Rozewicki J, Li S, Amada KM, Standley DM, Katoh K. ***MAFFT-DASH: integrated protein sequence and structural alignment***. Nucleic Acids Res. 2019 Jul 2;47(W1):W5-W10. doi: 10.1093/nar/gkz342. PMID: 31062021; PMCID: PMC6602451.

[58] Katoh, Kazutaka, Kazuharu Misawa, Kei‐ichi Kuma, and Takashi Miyata. 2002. ***"MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform."*** *Nucleic Acids Research* 30 (14): 3059–66.

[59] Schuster-Böckler B, Schultz J, Rahmann S. ***"HMM Logos for visualization of protein families."*** BMC Bioinformatics. 2004 Jan 21;5:7. doi: 10.1186/1471-2105-5-7. PMID: 14736340; PMCID: PMC341448.

[60] Werck-Reichhart D, Feyereisen R. ***"Cytochromes P450: a success story".*** Genome Biol. 2000;1(6):REVIEWS3003. doi: 10.1186/gb-2000-1-6-reviews3003. Epub 2000 Dec 8. PMID: 11178272; PMCID: PMC138896.

[61] Eswar N, Eramian D, Webb B, Shen MY, Sali A. ***"Protein structure modeling with MODELLER"***.

[62] Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Krüger FA, Light Y, Mak L, McGlinchey S, Nowotka M, Papadatos G, Santos R, Overington JP. The ChEMBL bioactivity database: an update. Nucleic Acids Res. 2014 Jan;42(Database issue):D1083-90. doi: 10.1093/nar/gkt1031. Epub 2013 Nov 7. PMID: 24214965; PMCID: PMC3965067.

[63] Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Marañón M, Hunter F, Junco L, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux CJ, Segura-Cabrera A, Hersey A, Leach AR. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res. 2019 Jan 8;47(D1):D930-D940. doi: 10.1093/nar/gky1075. PMID: 30398643; PMCID: PMC6323927.

[64]Methods Mol Biol. 2008;426:145-59. doi: 10.1007/978-1-60327-058-8_8. PMID: 18542861.
Bansal P, Morgat A, Axelsen KB, Muthukrishnan V, Coudert E, Aimo L, Hyka-Nouspikel N, Gasteiger E, Kerhornou A, Neto TB, Pozzato M, Blatter MC, Ignatchenko A, Redaschi N, Bridge A. Rhea, the reaction knowledgebase in 2022. Nucleic Acids Res. 2022 Jan 7;50(D1):D693-D700. doi: 10.1093/nar/gkab1016. PMID: 34755880; PMCID: PMC8728268.

[65] Smith Richard D.,Carlson Heather A.,***Identification of Cryptic Binding Sites Using MixMD with Standard and Accelerated Molecular Dynamics.***Journal of Chemical Information and Modeling. Volume 61, Issue 3,Febrero 2021.(1287-1299)

[66] Awan RE, Zainab S, Yousuf FJ, Mughal S. ***AI-driven drug discovery: Exploring Abaucin as a promising treatment against multidrug-resistant Acinetobacter baumannii.*** Health Sci Rep. 2024 Jun 4;7(6):e2150. doi: 10.1002/hsr2.2150. PMID: 38841115; PMCID: PMC11150274.

[67] Adhikari A, Shakya S, Shrestha S, Aryal D, Timalsina KP, Dhakal D, Khatri Y, Parajuli N. ***Biocatalytic role of cytochrome P450s to produce antibiotics: A review.*** Biotechnol Bioeng. 2023 Dec;120(12):3465-3492. doi: 10.1002/bit.28548. Epub 2023 Sep 10. PMID: 37691185.

[68] James B. Y. H. Behrendorff ***"Reductive Cytochrome P450 Reactions and Their Potential Role in Bioremediation"*** Front. Microbiol., 14 April 2021 Volume 12 - 2021 | https://doi.org/10.3389/fmicb.2021.649273

[69] Heo L, Feig M. ***Multi-state modeling of G-protein coupled receptors at experimental accuracy***. Proteins. 2022 Nov;90(11):1873-1885. doi: 10.1002/prot.26382. Epub 2022 May 16. PMID: 35510704; PMCID: PMC9561049.

[70] Jumper J. et al. (2021). ***"Highly accurate protein structure prediction with AlphaFold."*** *Nature.* 596, 583–589.

[71] **Information Decay in Molecular Docking Screens against Holo, Apo, and Modeled Conformations of Enzymes** Susan L. McGovern and Brian K. Shoichet Journal of Medicinal Chemistry **2003** *46* (14), 2895-2907 DOI: 10.1021/jm0300330

[72] McNutt, A.T., Francoeur, P., Aggarwal, R. *et al.* ***"GNINA 1.0: molecular docking with deep learning"***. *J Cheminform* **13**, 43 (2021). https://doi.org/10.1186/s13321-021-00522-2

[73] Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, Ahern W, Borst AJ, Ragotte RJ, Milles LF, Wicky BIM, Hanikel N, Pellock SJ, Courbet A, Sheffler W, Wang J, Venkatesh P, Sappington I, Torres SV, Lauko A, De Bortoli V, Mathieu E, Ovchinnikov S, Barzilay R, Jaakkola TS, DiMaio F, Baek M, Baker D. ***"De novo design of protein structure and function with RFdiffusion"***. Nature. 2023 Aug;620(7976):1089-1100. doi: 10.1038/s41586-023-06415-8. Epub 2023 Jul 11. PMID: 37433327; PMCID: PMC10468394.