Universidad Nacional del Nordeste
Facultad de Ciencias Exactas y Naturales y Agrimensura

Diseño de inhibidores de la cruzipaína del *Trypanosoma cruzi*

TESIS DOCTORAL

Adriano Martín Luchi

Corrientes, Argentina

2024

Universidad Nacional del Nordeste

Facultad de Cs. Exactas y Naturales y Agrimensura

Doctorado en Biología

“Diseño de inhibidores de la cruzipaína del *Trypanosoma cruzi*”

Tesis Doctoral

Para optar por el título de Doctor de la UNNE en Biología

Doctorando: Lic. Adriano Martín Luchi

Director: Dra. Peruchena, Nélida María

Sub-Director: Dr. Angelina, Emilio Luis

Corrientes, Argentina

2024

Prefacio

"Esta tesis se presenta como parte de los requisitos para optar al grado académico de Doctor en Biología de la Universidad Nacional del Nordeste y no ha sido presentada previamente para la obtención de otro título en esta Universidad u otras. La misma contiene los resultados obtenidos en investigaciones llevadas a cabo en el Laboratorio de Estructura Molecular y Propiedades (FACENA-UNNE-CONICET), durante el período comprendido entre Septiembre 2016 y Abril 2024, bajo la dirección de la Dra. Peruchena, Nélida María y codirección del Dr. Emilio Luis Angelina con financiamiento del CONICET".

Agradecimientos

A mi Directora Dra. Nélida Peruchena por abrirme las puertas del Laboratorio de Estructura Molecular y Propiedades (LEMyP), por su confianza, consejos académicos y de la vida.

A mi Co-director Dr. Emilio Angelina por su paciencia y dedicación; sobre todo con alguién que no venía del área de química; y principalmente por haberme dado la posibilidad de llevar a cabo este desafío.

A mis compañeros del Laboratorio de Estructura Molecular y Propiedades (LEMyP) y del IQUIBA-NEA, por su buena predisposición y apoyo.

A Leo y Germán por haberme ayudado a destrabar las últimas campañas de cribado virtual.

A Clara, por ser de gran apoyo durante tantos años de doctorado. Por los mates, las charlas y los consejos.

A mis padres por no dejarme nunca bajar los brazos.

A Vicky, mi compañera de vida.

Trabajos publicados bajo el desarrollo de esta tesis

Artículos directamente relacionados con el trabajo de tesis

- Luchi, A. M., Villafañe, R. N., Gómez Chávez, J. L., Bogado, M. L., Angelina, E. L., & Peruchena, N.M. (2019). Combining Charge Density Analysis with Machine Learning Tools to Investigate the Cruzain Inhibition Mechanism. ACS Omega, 4(22), 19582–19594. https://doi.org/10.1021/acsomega.9b01934

- Prepubliacion: Luchi AM, Gomez Chávez JL, Villafañe RN, Conti GA, Pérez ER, Angelina E, Peruchena, N. (2022). Graph neural networks and molecular docking as two complementary approaches for virtual screening: a case study on Cruzain. ChemRxiv. 2022; doi:10.26434/chemrxiv-2022-btz77

Artículos indirectamente relacionados con el trabajo de tesis

- Luchi, A., Angelina, E., Bogado, L., Andujar, E. L., Enriz, R. D. & Peruchena, N.M. (2016). Halogen bonding in biological context: a computational study of D2 dopamine receptor: Halogen Bonding in Biological Context: A Computational Study of D2 Dopamine Receptor.

- Luchi, A., Angelina, E., Bogado, L., Forli, S., Olson, A., & Peruchena, N. (2018). Flap-site Fragment Restores Back Wild-type Behaviour in Resistant Form of HIV Protease. Molecular informatics, 37(12), e1800053. https://doi.org/10.1002/minf.201800053

Resumen

La enfermedad de Chagas, causada por el protozoo parásito *T. cruzi*, es la enfermedad tropical transmisible de mayor prevalencia en América Latina. Fue catalogada por la Organización Mundial de la Salud (OMS), como una de las tantas "enfermedades olvidadas" o desatendidas. A pesar de ser una enfermedad de alta prevalencia, identificada hace más de 100 años y que afecta a millones de personas en todo el mundo, al día de hoy no existen tratamientos eficaces para la enfermedad de Chagas. Esto se debe principalmente a que las drogas actuales presentan severas limitaciones y la cura de esta enfermedad depende de la fase en la que se encuentre el paciente, así como de las condiciones fisiológicas y susceptibilidad de la cepa de *T. cruzi.*

A su vez, el diseño de fármacos demanda de un gran esfuerzo, tanto económico como de tiempo para las grandes industrias farmacéuticas encargadas del desarrollo de drogas. En este sentido, la enfermedad de Chagas afecta mundialmente a 7 millones de personas, de las cuales, la mayor parte cuenta con bajos recursos o vive en países subdesarrollados, lo que significa una contraposición contra la inversión hacia el desarrollo de nuevas alternativas farmacológicas. Es por ello que los esfuerzos para la búsqueda de nuevos candidatos a fármacos, al menos en las primeras etapas, recaen principalmente en entidades financiadas por el estado, tales como Universidades y Centros de Investigación.

Uno de los pasos esenciales en la búsqueda o diseño de fármacos contra *T. cruzi* es la identificación de moléculas diana que estén involucradas en rutas metabólicas importantes en el parásito y que al inhibirlas provoquen una disminución en los niveles de parasitemia. En línea con lo anterior, la cruzípaina (Cz), también conocida como cruzaína, es la cisteíno proteasa principal de *T. cruzi*, encargada de la invasión celular, y se expresa en todas las formas de desarrollo de diferentes aislados de *T. cruzi*. A su vez, se ha documentado que los inhibidores de la proteasa bloquean la proliferación y la metaciclogénesis de amastigotes y promastigotes *in-vitro*.

Por otro lado, el desarrollo de fármacos inicia con la identificación de compuestos que se unen a un blanco terapéutico o que muestran actividad biológica en ensayos de tamizaje.

Considerando que el número de moléculas orgánicas que son sintéticamente factibles está por encima de 10^{60} moléculas (espacio químico), resulta imposible su análisis sin el uso de técnicas computacionales.

La química computacional es una rama de la química que permite simular numéricamente estructuras, reacciones químicas, interacciones moleculares entre otros, facilitando así el estudio de los fenómenos físico-químicos, y la resolución de problemas que serían más difíciles y costosos de abordar desde el punto de vista experimental. En el diseño de fármacos se utiliza el modelado computacional para predecir cómo nuevos candidatos a fármacos aumentarán o disminuirán su afinidad al blanco molecular y de ese modo filtrar librerías de compuestos.

Existen diversos filtros que se utilizan para llevar a cabo el cribado virtual los cuales pueden variar según la complejidad de la base de datos y la información experimental de la que se disponga. Si se conoce la estructura tridimensional (3D) del receptor, se sugiere un cribado basado en la estructura. Si solo se conocen los compuestos activos, pero no el receptor, entonces la búsqueda se hace basada en el ligando. En este trabajo de tesis, al contar con la estructura 3D del blanco molecular, al igual que una serie de compuestos activos, se combinaron ambos tipos de enfoques no sólo para facilitar la búsqueda de nuevos candidatos capaces de inhibir a Cz, sino también para desarrollar metodologías alternativas para el estudio de interacciones en el sitio catalítico enzimas y puesta a punto de las herramientas quimioinformáticas utilizadas.

La tesis se organizó en 7 capítulos. El primero contiene una introducción, el segundo contiene una descripción de la metodología al cual le siguen luego cuatro capítulos de presentación y discusión de los resultados. Finalmente el último capítulo contiene las principales conclusiones generales.

En la primera parte de resultados, mediante la compilación de una biblioteca de estructuras tridimensionales de Cz unida a ligandos covalentes, se realiza un acercamiento desde el punto de vista estructural. Se analiza de forma dinámica las interacciones intermoleculares que tienen lugar en el sitio catalítico de la enzima, mediante la utilización de simulaciones de Dinámicas

Moleculares (DM). En esta primera etapa, se trata de entender los patrones de interacción frecuentes para la mayoría de los inhibidores compilados. Se utilizan herramientas de análisis computacional menos costosas como las “Huellas dactilares de interacción” (*SIFt*) y herramientas de análisis mecanocuánticas como la Teoría Cuántica de Átomos en Moléculas (*QTAIM*) que permiten detectar otras interacciones, que se escaparían en un simple análisis geométrico y de distancia.

En la segunda parte, utilizando de QTAIM los elementos topológicos de la densidad de carga que describen las interacciones en los complejos Cz-ligando, se entrena un modelo de clasificación de aprendizaje supervisado SVM-RFE capaz de discriminar entre las interacciones presentes en los complejos de los inhibidores más activos (interacciones de tipo activo) y las que ocurren en los menos activos (interacciones de tipo inactivo). Se obtienen interacciones relevantes que estabilizan una conformación particular de Cz.

Las interacciones antes mencionadas se conectan a algoritmos de *docking* para mejorar la función de puntuación y guiar las predicciones de dicho estudio.

En línea con lo anterior, luego se realiza un *docking* retrospectivo utilizando una biblioteca de ligandos conocidos. Se utilizan algoritmos de *docking* molecular, normalmente utilizados en campañas de búsqueda de fármacos, así como también Redes Neuronales Gráficas. Se propone una estrategia combinada para explotar los beneficios de ambas herramientas, es decir, la capacidad de los modelos de redes neuronales, en particular GCN, para capturar relaciones complejas de los datos y la interpretabilidad del *docking* molecular basado en la estructura, para evaluar virtualmente la biblioteca AID 1478 contra la cruzipaina.

En última instancia se realiza un cribado prospectivo de una biblioteca compilada *in-house* de 7 millones de compuestos, donde se aplican de filtros pre y pos-*docking* para reducir progresivamente el número de compuestos que se seleccionan en cada etapa del proceso, obteniéndose 18 candidatos a ser adquiridos comercialmente para futuros ensayos experimentales.

Abstract

Chagas disease, caused by the parasitic protozoan *T. cruzi*, is the most prevalent transmissible tropical disease in Latin America. It was classified by the World Health Organization (WHO) as one of the many "neglected" or overlooked diseases. Despite being a highly prevalent disease, identified over 100 years ago and affecting millions of people worldwide, there are currently no effective treatments for Chagas disease. This is mainly due to the severe limitations of current drugs, and the cure for this disease depends on the stage of the patient, as well as the physiological conditions and susceptibility of the *T. cruzi* strain.

Furthermore, drug development requires significant effort, both economically and in terms of time, for the major pharmaceutical industries responsible for drug development. In this regard, Chagas disease affects 7 million people worldwide, the majority of whom have low resources or live in underdeveloped countries, which poses a counterbalance against investment in the development of new pharmacological alternatives. Therefore, efforts to search for new drug candidates, at least in the early stages, mainly rely on entities primarily funded by the state, such as universities and research centers.

One essential step in the search or design of drugs against *T. cruzi* is the identification of target molecules involved in important metabolic pathways in the parasite, which, when inhibited, cause a decrease in parasitemia levels. In line with this, cruzipain (Cz), also known as cruzain, is the main cysteine protease of T. cruzi, responsible for cell invasion, and is expressed in all developmental forms of different *T. cruzi* isolates. It has been documented that protease inhibitors block the proliferation and metacyclogenesis of amastigotes and promastigotes in vitro.

On the other hand, drug development begins with the identification of compounds that bind to a therapeutic target or show biological activity in screening assays. Considering that the number of organically feasible molecules is above 10^{60} molecules (chemical space), their analysis is impossible without the use of computational techniques.

Computational chemistry is a branch of chemistry that allows the numerical simulation of structures, chemical reactions, molecular interactions, among others, facilitating the study of

physicochemical phenomena and the resolution of problems that would be more difficult and costly to address experimentally. Computational modeling is used in drug design to predict how new drug candidates will increase or decrease their affinity to the molecular target, thereby filtering compound libraries.

There are various filters used to conduct virtual screening, which may vary depending on the complexity of the database and the experimental information available. If the three-dimensional (3D) structure of the receptor is known, structure-based screening is suggested. If only the active compounds are known but not the receptor, then the search is ligand-based. In this thesis work, having the 3D structure of the molecular target, as well as a series of active compounds, both types of approaches were combined not only to facilitate the search for new candidates capable of inhibiting Cz but also to develop alternative methodologies for studying interactions in the enzyme catalytic site and refining the chemoinformatics tools used.

In the first part, by compiling a library of three-dimensional structures of Cz bound to covalent ligands, an approach is made from a structural point of view. The intermolecular interactions taking place in the enzyme's catalytic site are dynamically analyzed using Molecular Dynamics (MD) simulations. In this initial stage, the goal is to understand the frequent interaction patterns for most of the compiled inhibitors. Less expensive computational analysis tools such as Interaction Fingerprints (SIFt) and mechanistic analysis tools like Quantum Theory of Atoms in Molecules (QTAIM) are used to detect other interactions that would escape simple geometric and distance analysis.

In the second part, using QTAIM's topological elements of charge density describing interactions in the Cz-ligand complexes, a supervised learning classification model SVM-RFE capable of discriminating between interactions present in the complexes of the most active inhibitors (active-type interactions) and those occurring in the less active ones (inactive-type interactions) is trained. Relevant interactions stabilizing a particular conformation of Cz are obtained.

The aforementioned interactions are then connected to docking algorithms to improve the scoring function and guide the predictions of this study.

In line with the above, a retrospective docking is then performed using a library of known ligands. Molecular docking algorithms, commonly used in drug discovery campaigns, as well as Graph Neural Networks, are used. A combined strategy is proposed to exploit the benefits of both tools, i.e., the ability of neural network models, particularly GCN, to capture complex data relationships and the interpretability of structure-based molecular docking, to virtually evaluate the AID 1478 library against cruzipain.

Finally, a prospective screening of an in-house compiled library of 7 million compounds is carried out, where pre- and post-docking filters are applied to progressively reduce the number of compounds selected at each stage of the process, resulting in 18 candidates to be commercially acquired for future experimental assays.

ÍNDICE

Lista de abreviaturas

ACE *angiotensin-converting enzyme* (Enzima convertidora de angiotensina)

EC Enfermedad de Chagas

B_1R *Braquinin-1 Receptor* (Receptor de Bradiquinina-1)

B_2R *Braquinin-2 Receptor* (Receptor de Bradiquinina-2)

BCPs Bond Critical Points (Puntos críticos de enlace)

BPs Bond Paths (Caminos de enlace)

BZ Benznidazol

CP Cisteína proteasa

CV Cribado Virtual

Cz Cruzipaina

Cz-Inh Complejo Cruzipaina-inhibidor

CVBE Cribado Virtual Basado en la Estructura

CVBL Cribado Virtual Basado en el Ligando

CVP Cribado Virtual Prospectivo

CVR Cribado Virtual Retrospectivo

DFAC Descubrimiento de Fármacos Asistido por Computadoras

DFT Density Functional Theory (Teoría del funcional de la densidad)

DNDi *Drug for Neglected Disease initiative* (Iniciativa Medicamentos para Enfermedades Olvidadas o desatendidas)

DM Dinámica Molecular

FDA *Food and Drug Administration* (Administración de Alimentos y Medicamentos de los Estados Unidos)

IAS *Interatomic Surface* (Superficie Interatómica)

ICP Inhibidor de cisteína proteasas

IC50 Concentración de inhibición 50

Ki Constante de inhibición enzimática

NFX Nifurtimox

ns Nanosegundos

ML *Machine Learning* (Aprendizaje automático)

MMPBSA Molecular Mechanics Poisson-Boltzmann Surface Area

OMS Organización Mundial de la Salud

PDB *Protein data bank* (Banco de datos de proteínas)

QSAR *Quantitative Structure-Activity Relationship* (Relación cuantitativa estructura-actividad)

QTAIM *Quantum Theory of Atoms in Molecules* (Teoría Cuántica de Átomos en Moléculas)

RVE *Recursive feature elimination* (Algoritmo, Eliminación Recursiva de Características)

RMSD *Root-mean-square deviation* (Desviación cuadrática media)

RMSF *Root-mean-square fluctuation* (Fluctuación cuadrática media)

SMD Steered Molecular Dynamics (Dinámica Molecular Dirigida)

SVM *Support Vector Machines* (Máquinas Vectoriales de Soporte)

SH Sulfato de heparano

T. cruzi *Trypanosoma cruzi*

CAPÍTULO I

“Introducción”

1.1 Enfermedad de Chagas

La enfermedad de Chagas (EC) o Trypanosomiasis americana, causada por el protozoo flagelado *Trypanosoma cruzi* (TC) y transmitida principalmente por la picadura de un insecto hematófago de la subfamilia Triatominae conocido como vinchuca, chinche o barbeiro, es una de las mayores problemáticas sanitarias de América Latina. De acuerdo a estimaciones de la Organización Mundial de la Salud (OMS), 7 millones de personas padecen la enfermedad de manera crónica, de las cuales alrededor de 7000 mueren cada año (World Health Organization 2020). Por ser una enfermedad de las denominadas olvidadas o desatendidas, ya que afectan a las regiones más pobres del mundo, actualmente se encuentran disponibles solo dos medicamentos aprobados y reconocidos por la FDA (*Food and Drug Administration*), los cuales acarrean una serie de problemas en cuanto a efectos secundarios y efectividad en pacientes en estadíos crónicos de la enfermedad (de Souza, de Oliveira, y Andricopulo 2017). Por tales motivos, la búsqueda de nuevos candidatos a fármacos así como también el desarrollo de nuevas terapias antichagásicas es de suma necesidad para nuestra población.

1.1.1 Descubrimiento e historia

El descubrimiento de Carlos Chagas de la trypanosomiasis americana (1909) fue uno de los hallazgos más exitosos de la historia de la medicina tropical (Coura and Borges-Pereira 2010).

La historia natural de la enfermedad de Chagas comienza millones de años atrás como una enfermedad enzoótica entre animales salvajes y así lo sigue siendo en determinadas regiones como por ejemplo, el Amazonas.

Cuando la humanidad se aventuró en áreas naturales donde la infección era relevante, ésta se empezó a transmitir de manera accidental hacia los humanos, constituyendo así la antropozoonosis actual. A causa de la deforestación excesiva, el establecimiento de más áreas de cultivo y ganadería en exceso desde hace más de 300 años en América latina, aquellos triatominos que fueron dejados sin sus fuentes de alimento, debido a la remoción de los animales salvajes, empezaron a colonizar áreas lindantes a asentamientos humanos. De este modo, los

insectos se adaptaron a su nuevo nicho, incluyendo la alimentación de sangre humana y de animales domésticos.

Los triatominos, sus reservorios y vectores existieron en la naturaleza por millones de años. Los trypanosomas primitivos eran parásitos monogenéticos de insectos que no se alimentaban de sangre. Cuando los insectos adquieren el comportamiento hematófago, los trypanosomatidos evolucionaron bajo cambios morfológicos y funcionales, desarrollando un flagelo y una membrana ondulante para la circulación en el torrente sanguíneo de vertebrados (Levine 1973).

Los triatominos se conocen desde el siglo XVI (Lent & Wygodzinsky 1979) pero sus adaptaciones a entornos peridomésticos es más reciente. A pesar de que la infección hacia humanos es conocida gracias a estudios realizados en momias de entre 4000-9000 años de edad (Guhl et al. 1999; Aufderheide et al. 2004), la enfermedad de Chagas comienza a hacerse endémica a partir de la incidencia antrópica de los espacios naturales desde hace más de 3 siglos (Coura 2007; Coura y Dias 2009).

1.1.2 Transmisión

La enfermedad de Chagas, una infección que inicialmente afectaba a animales salvajes (enzootica), se transformó en una antropozoonosis cuando a través de la acción predatoria de los seres humanos ocupando nuevos espacios físicos, se invadió ecotopos salvajes removiendo los flora y fauna salvaje que allí vivían (Coura and Borges-Pereira 2010).

Los triatominos (vectores de la enfermedad de Chagas) pudieron adaptarse fácilmente a los nuevos entornos allí construidos, constituyendo un nuevo hábitat, donde le fue fácil proveerse de alimento a través la sangre humana, estableciendo ciclos intercomunicados que alternan los entornos salvajes con los domiciliarios.

En particular, la forma más común de transmisión de la enfermedad a los seres humanos en zonas endémicas se da a través de las heces de los vectores triatominos redúvidos infectados conocidos como chinche (en El Salvador), vinchuca (Ecuador, Bolivia, Chile y Argentina), chipo (Venezuela), pito (Colombia), chirimacha (Perú) y barbeiro (Brasil). Son insectos hematófagos que pertenecen a la subfamilia Triatominae (Familia Reduviidae) donde en Argentina el principal vector de esta enfermedad es el *Triatoma infestans.* Debido a que los insectos triatominos tienen

el tracto digestivo recto, inmediatamente después de alimentarse con la sangre de seres humanos y animales defecan y al rascarse la persona o el animal permite que el parásito se introduzca en la sangre, replicándose principalmente en tejidos musculares y nerviosos. Esta transmisión es conocida como *transmisión vectorial* y comprende básicamente 3 ciclos: doméstico, peridoméstico y selvático (WHO 2012).

En relación con el ciclo doméstico, la estructura de las casas rurales o periurbanas en zonas endémicas las hace especialmente vulnerables a la infección por vía vectorial, debido a que los insectos de hábitos nocturnos pueden habitar en rendijas o agujeros de viviendas, techos de paja, paredes de adobe, bodegas y viviendas precarias en zonas rurales o suburbanas. Por otra parte, la estrecha asociación entre los habitantes y los animales domésticos establece una fuente de sangre abundante y de fácil acceso, por lo que el interior de las viviendas conforma un hábitat muy favorable para los insectos (WHO 2012; Bellera 2014; Cerny 2016).

El ciclo peridoméstico conecta a los ciclos selváticos y domésticos e involucra especies como marsupiales, roedores y otro mamíferos que ocasionalmente entran y salen de las viviendas y triatominos selváticos diferentes de *Triatoma infestans* (Coura and Dias 2009; WHO 2012; Bellera 2014).

Por último, el ciclo selvático involucra a diferentes triatominos selváticos que infectan a numerosas especies de mamíferos salvajes terrestres o arbóreos. (Rodrigues Coura and De Castro 2002; Bellera 2014).

Por otro lado, la enfermedad de Chagas puede ser transmitida de forma congénita, es decir de madre a hijo, por tal motivo, las mujeres embarazadas deben realizarse el análisis para la detección temprana de Chagas. La transmisión de madres chagásicas a su hijo durante el embarazo o el parto ocurre con una incidencia del 0,5% en áreas no endémicas y del 9% en áreas endémicas. Sin embargo, los recién nacidos con diagnosis positiva para la enfermedad, que fueran tratados durante el primer año de vida alcanzan casi un 90% de posibilidades de ser desparasitados mediante tratamiento con benznidazol; además de no presentar efectos secundarios adversos en la mayoría de los casos (Altcheh et al. 2011).

Por último, en zonas donde hay una gran cantidad de infectados, *T. cruzi* también puede ser transmitido a humanos mediante transfusiones sanguíneas o trasplantes de órganos de donantes

infectados, clínicamente sanos (asintomáticos) o ingesta de alimentos contaminados con *T. cruzi* (Haberland et al. 2013; Angheben et al. 2015; Cerny 2016).

1.1.3 Ciclo de vida del parásito

Trypanosoma cruzi, el agente causal de la enfermedad de Chagas, es un protozoo de la clase Kinetoplastea perteneciente a la familia Trypanosomatidae, el cuál posee un cuerpo alargado provisto de un flagelo y una membrana ondulante, estructuras que le permiten su movilización dentro del torrente sanguíneo. Además, cuenta con una sola y gran mitocondria, cuyo material genético constituye la estructura denominada cinetoplasto (kDNA) y representa el 20% del total del ADN del parásito. (Cerny 2016)

T. cruzi es un parásito con un ciclo de vida cerrado que involucra vertebrados e invertebrados y que asume tres distintas morfologías y características bioquímicas durante su ciclo de vida (Bern 2015; Oliveira et al. 2020).

Los dos estadios morfológicos replicativos, epimastigote y amastigote. El primero está presente en el intestino medio del insecto y el segundo, en el citoplasma de una célula de mamífero.

Por otra parte, los tripomastigotes, una forma no replicativa que puede invadir activamente las células de mamíferos, está presente en el torrente sanguíneo o en el intestino posterior del insecto (tripomastigotes metacíclicos) (Brener 1973).

Las características de cada una de las fases morfológicas se describen a continuación:

- Amastigote: esférico u ovalado, es la forma reproductiva en el interior de las células mamíferas.
- Epimastigote: alargado y con el cinetoplasto localizado con anterioridad al núcleo, es la forma reproductiva en el tracto digestivo de los invertebrados y en medios de cultivo.
- Tripomastigote: también alargado, pero con el cinetoplasto localizado posteriormente al núcleo. Se encuentra en la sangre de los mamíferos (en donde es el estadio infectante), así´ como también en las heces de los insectos. Esta forma no se divide.

T. cruzi comprende distintos linajes y morfologías, infectividad, virulencia y patogenicidad (Seco-Hidalgo, De Pablos, and Osuna 2015).

El parásito realiza su ciclo entre el vector y el huésped (Figura 1.1):

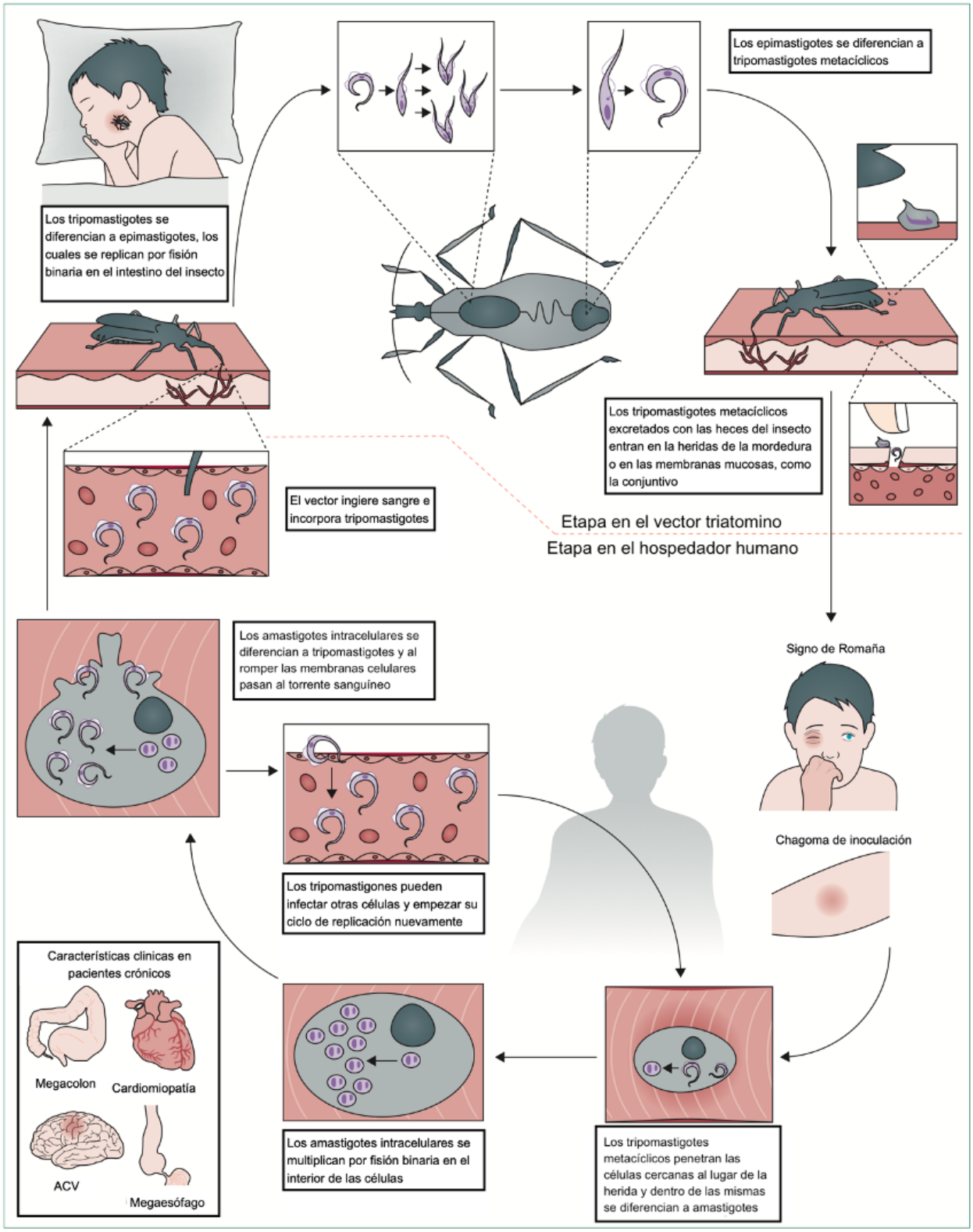


Figura 1.1 Ciclo de vida del *Tripanosoma cruzi,* donde se diferencian los diferentes estadios morfológicos del parásito dentro de sus posibles hospedadores (adaptado de Pérez-Molina y Molina 2018 y Oliveira et al. 2020).

En el humano el ciclo de vida de *T. cruzi* se inicia cuando un insecto hematófago infectado pica a un individuo y defeca cerca de la herida (etapa inferior en Figura.1.1), los tripomastigotes metacíclicos que se encuentran en las heces, entran en el huésped a través de la herida o por membranas mucosas. Una vez dentro del hospedador, los tripomastigotes invaden células cercanas al lugar de inoculación.

En el citoplasma de las células nucleadas, los tripomastigotes se diferencian a amastigotes, los cuales entran en una segunda etapa reproductiva que duplica su número en aproximadamente 12 h por un periodo de 4 o 5 días. Después de la reproducción, una gran cantidad de amastigotes se encuentran dentro de la célula infectada, formándose pseudoquistes. Luego de un número determinado de duplicaciones, se convierte de nuevo en tripomastigote, las células se lisan y se liberan al torrente sanguíneo. Los tripomastigote vuelven a infectar otras células repitiéndose el ciclo replicativo y quedan disponibles para infectar a un vector que se alimente del huésped (Bern, 2015).

En el insecto, el ciclo se inicia cuando un triatómino se alimenta de un huésped infectado y algunos tripomastigotes pasan a él junto con la sangre. Cuando las formas de tripomastigotes que circulaban en la sangre alcanzan el intestino medio del insecto, se diferencian en epimastigotes replicativos. Los epimastigotes migran al extremo posterior del tracto digestivo del insecto, donde se diferencian en tripomastigotes metacíclicos infecciosos que no se dividen y que se expulsan junto con las heces del vector, pudiendo infectar a un nuevo huésped durante la picadura (Tyler and Engman 2001; Bern 2015; Centers for Disease Control and Prevention 2015).

1.1.4 Aspectos clínicos de la enfermedad

Las manifestaciones clínicas de la enfermedad pueden clasificarse en dos etapas. La etapa aguda de corta duración y la etapa crónica, más prolongada en el tiempo y de desarrollo asintomático de la enfermedad (Cerny 2016).

La etapa aguda, puede ocurrir en cualquier momento de la vida del ser humano y si bien es asintomática en la mayoría de los casos (Pérez-Molina and Molina 2018), se caracteriza por presentar una parasitemia circulante alta, debido a que los parásitos invaden y se multiplican en

diferentes células del hospedador, como macrófagos, músculo liso y estriado y neuronas. Algunos pacientes presentan fiebre, inflamación del sitio de inoculación, edema palpebral unilateral (signo de Romaña, cuando la vía conjuntiva es la vía de entrada del parásito, ver Figura 1.1), adenopatía y hepatoesplenomegália. La fase aguda dura aproximadamente de 4 a 8 semanas y la parasitemia decrece sustancialmente desde los 90 días en adelante. Casos severos de fases agudas fueron reportados en el 1-5% de los casos e incluyen manifestaciones tales como miocarditis, derrame pericardico y meningoencefalitis (riesgo de mortalidad aproximado del 0.2 al 0.5%) (Bellera 2014; Pérez-Molina and Molina 2018).

La fase aguda normalmente suele pasar de forma espontánea después de un tiempo y los pacientes pasan a estar infectados de manera crónica si no son tratados. En lo que respecta a la segunda fase de la enfermedad, aproximadamente el 30-40% de los pacientes infectados crónicamente pueden desarrollar patologías relacionadas con afecciones orgánicas (mayormente miocardiopatía o mega-vísceras [megaesófago, megacolon] ver Figura 1.1). Sin un tratamiento adecuado, por lo general la enfermedad de Chagas es mortal en los estadios crónicos debido a la miocardiopatía que ocasiona (Cerny 2016).

1.1.5 Epidemiología

La enfermedad de Chagas es la enfermedad tropical transmisible de mayor prevalencia en América Latina. Fue catalogada por la Organización Mundial de la Salud (OMS), como una de las tantas “enfermedades olvidadas” o desatendidas, al igual que otras enfermedades tropicales infecciosas como la leishmaniasis o la enfermedad del sueño. Esto se debe a que afectan principalmente a las poblaciones mas pobres del planeta, representando de este modo un esfuerzo no lucrativo para las grandes compañías farmacéuticas en cuanto a la investigación y desarrollo de nuevas terapias o medicamentos para su tratamiento (WHO 2012; Bellera 2014).

La enfermedad se encuentra asociada a poblaciones de bajos recursos, donde las condiciones de vivienda son desfavorables. Esto ocurre principalmente en zonas rurales de toda Latinoamérica y en los cinturones de pobreza alrededor de las grandes ciudades, donde las personas infectadas que migran buscando mejores oportunidades de trabajo no tienen posibilidad de acceder a una atención médica adecuada. A su vez, la enfermedad de Chagas

coexiste con otras enfermedades que son más sintomáticas, razón por la cual frecuentemente pasa desapercibida (Cerny 2016).

Por otro lado, a finales del siglo XX, la enfermedad de Chagas tuvo una expansión a nivel mundial, producto de las grandes oleadas migratorias. En particular, se produjo la migración de individuos infectados hacia áreas no endémicas, como es el caso de Estados Unidos, Europa, Australia y Japón, provocando allí aumentos significativos en el número de casos. La globalización de la enfermedad ha obligado a países no endémicos a implementar medidas de prevención como así también discutir nuevas estrategias para su control (Bellera 2014).

Cabe destacar que a causa del gran número de animales silvestres que sirven de reservorio del parásito, la enfermedad no puede erradicarse. Es por esto que los objetivos de control consisten en bajar la tasa de transmisión comunitaria y lograr que la población portadora del parásito tenga acceso temprano a los centros de salud.

En relación con el número de personas infectadas, la OMS aduce que la incidencia anual de la enfermedad es de 28.000 casos en todo el continente americano, y se calcula que en el mundo hay entre 6 y 7 millones de personas infectadas por *T. cruzi*, provocando unas 12.000 muertes anuales. Donde, su costo económico ha sido estimado en unos 7 mil millones de dólares anuales (WHO 2012; 2020).

Por último, en Argentina, las poblaciones más afectadas por la enfermedad corresponden a aquellas que habitan la región del Gran Chaco (Gaspe et al. 2018). Esto se debe principalmente a los altos niveles de infestación domiciliaria de *Triatoma infestans*, a pesar de las múltiples campañas de control basadas en insecticidas realizadas durante casi 70 años (Gürtler et al. 2007).

Si bien en los últimos 25 años ha habido una disminución en el número de infectados (Cerny 2016), las últimas cifras sobre la enfermedad de Chagas en Argentina sugieren que unas 2.300.000 personas están expuestas al parásito, 1.500.000 están infectadas (3,7% de la población) y más de 370.000 sufren de problemas cardíacos relacionados con la enfermedad (Klein et al. 2017). Según la Organización Panamericana de la Salud (OPS), en los últimos 15 años, el perfil epidemiológico ha cambiado, con la transmisión congénita superando a la vectorial y transfusional como la principal fuente de nuevos casos.

El Programa Nacional de Chagas (PNCh) en Argentina trabaja en mejorar el diagnóstico temprano y el tratamiento adecuado de las infecciones agudas y crónicas, a través de la creación y distribución de guías de atención y el suministro gratuito de medicamentos tripanocidas desde el Ministerio de Salud hacia los programas provinciales. Estos medicamentos se distribuyen en hospitales y centros de atención primaria según la demanda. El sistema de salud argentino ofrece cobertura gratuita y universal, complementada con seguros sociales y privados. Sin embargo, la prescripción de tratamientos ha sido históricamente baja, especialmente en el primer nivel de atención (Cardozo 2016).

1.1.7 Tratamiento

A pesar de ser una enfermedad de alta prevalencia, identificada hace más de 100 años y que afecta a millones de personas en todo el mundo, los tratamientos disponibles para la enfermedad de Chagas presentan una eficacia relativa. Esto se debe principalmente a que las drogas actuales presentan severas limitaciones y la cura de esta enfermedad depende de la fase en la que se encuentre el paciente, así como de las condiciones fisiológicas y susceptibilidad de la cepa de *T. cruzi* (Pérez-Morales et al. 2012). Desde hace más de cuatro décadas solo existen dos drogas disponibles en el mercado para su tratamiento: Nifurtimox (1965) y beznidazol (1970).

Nifurtimox (NFX) es un nitrofurano N-(3-metil-1,1-dioxo-1,4-tiazinan-4-il)-1-(5-nitro-2-furil)-metanimina y benznidazol (BZ) es un nitroimidazol N-benzil-2-(2-nitroimidazol-1-il)-acetamida (ver Figura 1.2).

Si bien hasta el momento el mecanismo de acción de ambas drogas no es del todo claro (Carneiro et al. 2017; Thakare et al 2021; García-Huertas et al. 2021), se sabe que NFX actúa por la vía de la reducción del grupo nitro a radicales aniónicos inestables, lo cual produce una reacción que genera la producción de metabolitos de oxígeno reducido altamente tóxicos; *T. cruzi* es deficiente en mecanismos de detoxificación para metabolitos de oxígeno (particularmente peróxido de hidrógeno) siendo por tanto muy sensible al estrés oxidativo a comparación de las células de algún hospedador vertebrado (Urbina y Docampo 2003; Maya et al. 2004; 2007). Por otro lado, BZ estaría involucrado en la modificación covalente de macromoléculas, tales como ADN, proteínas y lípidos, mediante intermediarios de nitroreducción.

A su vez, se sabe que BZ afecta el metabolismo de tripanotiona e inhibe la NADH-fumarato reductasa en el parásito (Turrens et al. 1996; Murta et al. 1999; Urbina y Docampo 2003; Maya et al. 2004; 2007; Montalto De Mecca et al. 2008).

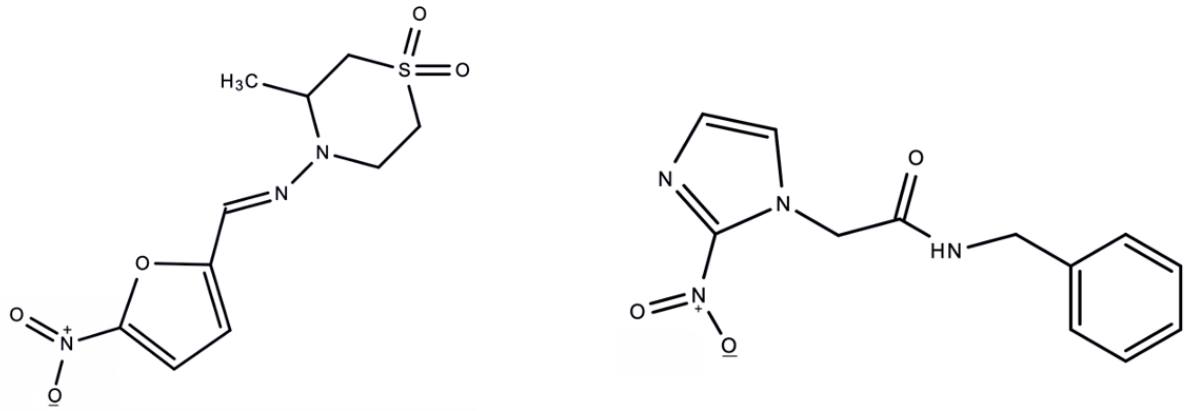


Figura 1.2. Estructura de los fármacos disponibles para el tratamiento de la enfermedad de Chagas. Ambos son nitroheterociclos, que dañan las estructuras celulares del parásito mediante su reducción intracelular y la consecuente producción de especies reactivas del oxígeno (Carneiro et al. 2017).

El tratamiento con estos fármacos requiere protocolos de extensa duración, altos costos y no siempre pueden realizarse en su totalidad debido a los serios efectos secundarios que obligan a la interrupción de su administración (Aldasoro et al. 2018).

En cuanto a los efectos adversos, aquellos pacientes tratados con estos fármacos presentan, con frecuencia, problemas dermatológicos como urticaria o dermatitis, anorexia, náuseas, insomnio, linfoadenopatías y neuropatías (Jackson et al. 2010; Pinazo et al. 2010; Aldasoro et al. 2018).

A pesar de que el desarrollo de nuevos fármacos ha crecido en los últimos años y a más de 100 años del descubrimiento de esta enfermedad, seguimos sin un tratamiento efectivo contra la infección por *T. cruzi.* La resistencia a las drogas actualmente disponibles, los efectos adversos y las limitaciones terapéuticas que BZ y NFX presentan, generan la necesidad de encontrar nuevos fármacos o principios activos capaces de prevenir, paliar o mitigar los efectos de esta enfermedad en la sociedad.

1.1.8 Alternativas quimioterapéuticas y perspectivas futuras en el diseño de fármacos antichagásicos

El diseño de fármacos demanda de un gran esfuerzo, tanto económico como de tiempo para las grandes industrias farmacéuticas encargadas del desarrollo de drogas (figura 1.3) (Roemer y Krysan 2014). A su vez, la enfermedad de Chagas afecta mundialmente a 7 millones de personas, de las cuales, la mayor parte cuenta con bajos recursos o vive en países subdesarrollados, lo que significa una contraposición contra la inversión hacia el desarrollo de nuevas alternativas farmacológicas.

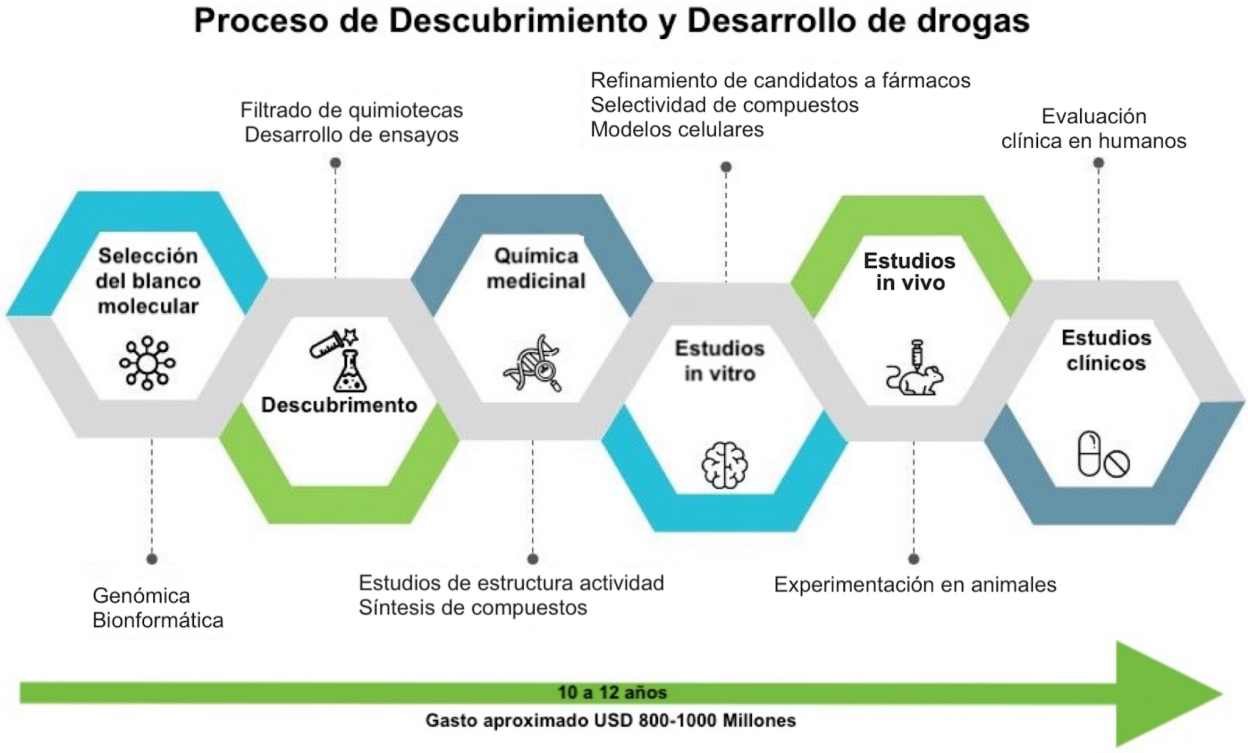


Figura 1.3 Proceso de Descubrimiento y Desarrollo de drogas (Adaptado de Duelen R. et al. 2019).

Sin embargo, hace unos 20 años aproximadamente los esfuerzos por encontrar nuevos candidatos que cumplan con requerimientos para la producción de drogas antichagásicas viene creciendo de manera significativa. Esto se debe principalmente a i) la aparición de nuevos investigadores interesados en la enfermedad, ii) desarrollos tecnológicos que permitieron probar nuevas estrategias para la investigación y el desarrollo de fármacos capaces de combatir la Enfermedad de Chagas (EC) y por último iii) la evolución en la eficacia de nuevas entidades

químicas para el tratamiento de pacientes con la EC en el estadio indeterminado de la enfermedad (Chatelain 2015).

A su vez, el programa de descubrimiento y optimización de compuestos líderes de la Iniciativa "Medicamentos para Enfermedades Desatendidas" (DND*i* siglas en inglés para *Drug for Neglected Disease Initiative*) para la EC, que reúne esfuerzos de un conjunto de instituciones público privadas, y por último, la formación del Consorcio para el descubrimiento de drogas antichagásicas creado en Estados Unidos en 2010 y fundado por el *U.S. National Institutes of Health (NIH)* que incluye un conjunto de laboratorios en Estados Unidos asociados con otros a nivel global (Chatelain 2015; Cristovão-Silva et al. 2019).

De acuerdo con la DND*i*, el tratamiento actual para la EC presenta problemas, como períodos largos de tratamiento (30-60 días), toxicidad dependiente de la dosis, baja tasa de adherencia y falta de una formulación pediátrica (Drugs for Neglected Diseases Initiative 2019). Es por ello que DND*i* delineó los criterios que se consideran aceptables e ideales para los posibles fármacos para la enfermedad de Chagas de la siguiente manera: para ser activo contra todas las cepas de T. cruzi, la eficacia clínica debe ser superior al benznidazol en todas las fases de la enfermedad, no debe tener contraindicaciones o interacciones farmacológicas, y tiene que ser administrado por vía oral (Zingales et al. 2014; Carneiro et al. 2017). La Tabla 1 resume el panorama actual del desarrollo de fármacos contra la EC.

Las actividades de investigación destacan la participación de los socios públicos y privados en la selección de bibliotecas químicas, optimización compuestos líderes de nuevas series para el tratamiento de la enfermedad de Chagas.

Entre las actividades de transición, el enfoque se centra en la realización de estudios preclínicos, estudios de fase I y II con nuevos compuestos, donde se verifican propiedades farmacocinéticas y toxicológicas de la droga.

Por último, en cuanto a las actividades de desarrollo, constan de ensayos clínicos controlados y testeados mediante la utilización de placebos a fin de corroborar posibles efectos beneficiosos del fármaco contra la enfermedad de Chagas (Torrico et al. 2023). En Argentina la compañía Elea Laboratories se encarga de la fabricación de una versión genérica de BZ bajo el nombre

comercial de Abarax. A su vez, se están implementando formulaciones pediátricas de BZ y NFX (Zingales et al. 2014).

Tabla 1.1 Estado actual de fármacos evaluados para el tratamiento de la enfermedad de Chagas

	Investigación		>>	Transición			>>	Desarrollo
		Ensayos		Estudios clínicos				
Fármaco	Optimización	in vitro	in vivo	Estudios preclínicos	Fase I	Fase II	Fase III	Implementación
BZ	√	√	√	√	√	√	√	√
NFX	√	√	√	√	√	√	√	√
POSA	√	√	√	√	√	√	-	-
RAVU	√	√	√	√	√	√	-	-
ITRA	√	√	√	√	√	√	-	-
KETO	√	√	√	√	√	x	-	-
VORI	√	√	√	√	√	-	-	-
ALBA	√	√	√	√	√	-	-	-
DO8701	√	√	√	√	-	-	-	-
TAK-187	√	√	√	√	-	-	-	-
K-777	√	√	√	√	x	-	-	-
FENARI	√	√	√	√	Planeado	-	-	-
FEXINI	√	√	√	√	√	√	-	-
MILTEFO	√	√	√	√	√	-	-	-
EDELFO	√	√	-	-	-	-	-	-
ILMOFO	√	√	-	-	-	-	-	-
NANO BZ	√	√	√	√	-	-	-	-
SELENIUM	√	√	√	√	√	√	en proceso	-
ALOPU	√	√	√	√	√	x	-	-
AMIO	√	√	√	√	√	en proceso	-	-
SCYX-7158	√	√	√	√	en proceso	-	-	-

ALBA = albaconazol; ALOPU = alopurinol; AMIO = amiodarona; BZ = benznidazol; EDELFO = edelfosina; FENARI = fenarimol; FEXINI = fezinidazol; ILMOFO = ilmofosina; ITRA = itraconazol; KETO = ketoconazol; MILTEFO = miltefosina; NANO BZ = benznidazol nanoformulado; NFX = nifurtimox; POSA = posaconazol; RAVU = ravuconazol; SCYX-7158 = oxaborole; VORI = voriconazol; X = interrumpido.

A pesar de los esfuerzos, recientemente se han concluido dos ensayos clínicos de fase II que evalúaron la actividad antiparasitaria de los inhibidores de la biosíntesis de ergosterol posaconazol y ravuconazol (E1224, profármaco) en pacientes crónicos (Molina et al. 2014), pero con resultados desfavorables. A su vez, un fungicida no tóxico fenarimol y un inhibidor de la cisteína proteasa (cruzipaína), el fexinidazol, entre otros, se encuentran en estudios preclínicos (ver figura 1.3).

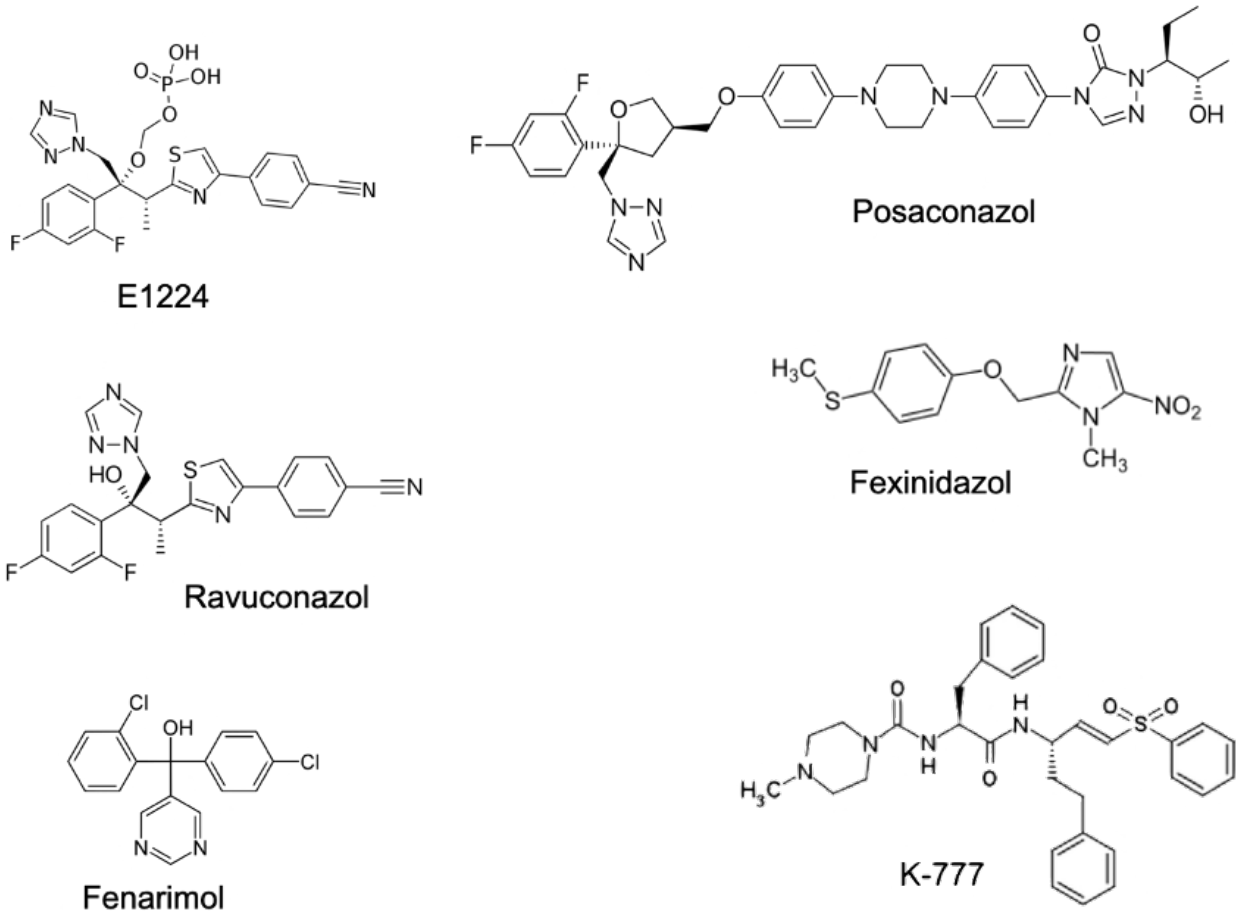


Figura 1.4 Estructura química de algunos candidatos sometidos a estudios clínicos avanzados.

Fexinidazol es un 5-nitroimidazol con potente actividad tripanocida que ha sido redescubierto a través de una extensa minería de compuestos. Puede inducir altos niveles de cura parasitológica en ratones infectados con cepas de T. cruzi susceptibles a BZ, parcialmente resistentes y

resistentes en estadíos experimentales agudos y crónicos de la EC. Esto significa una mejora importante en comparación con el tratamiento estándar actual con BZ (Maria Terezinha Bahia et al. 2012; Maria T. Bahia et al. 2014).

Por otro lado, los derivados de K-777 adquieren gran relevancia, por ser inhibidores de cisteína-proteasas y en particular de la cruzipaína (Cz). La cruzípaina, también conocida como cruzaína, es la cisteína proteasa principal de T. cruzi, encargada de la invasión celular, y se expresa en todas las formas de desarrollo de diferentes aislados de T. cruzi (Urbina y Docampo 2003; Cazzulo, Stoka, y Turk 2005). A su vez, se ha documentado que los inhibidores de la proteasa bloquean la proliferación y la metaciclogénesis de amastigote y promastigote in vitro, reduciendo significativamente la parasitemia y aumentando la supervivencia animal en modelos murinos en ambas etapas de infección con *T. cruzi* (Engel et al. 1998).

Por las razones que se detallan con anterioridad se seleccionó a Cz como el blanco molecular a estudiar en el presente trabajo de tesis y es por ello que en el siguiente apartado se describen las características estructurales y funcionales más relevantes de la cisteína proteasa de T. cruzi, Cz y de los inhibidores descubiertos hasta la fecha.

1.2 Cruzipaína, la principal cisteína proteasa del T. cruzi

1.2.1 Aspectos generales de Cz

Uno de los pasos esenciales en la búsqueda o diseño de fármacos contra *T. cruzi* es la identificación de moléculas diana que estén involucradas en rutas metabólicas importantes en el parásito y que al inhibirlas provoquen una disminución en los niveles de parasitemia.

La cruzipaína (también conocida como cruzaína en su forma recombinante o GP57/51) es la cisteína proteasa (PC) principal y mejor caracterizada del Trypanosoma cruzi, y es un blanco terapéutico para la enfermedad de Chagas. Esta enzima, con un peso molecular de ~41 kDa, corresponde a una glicoproteína perteneciente a la superfamilia de la papaína (Berti y Storer 1995), compartiendo altos índices de similaridad con la rodesaína, una de las cisteina proteasas encontradas en *Trypanosoma brucei* (J. J. Cazzulo et al. 1990; A. E. Eakin et al. 1990). La enzima está codificada por múltiples genes (130 en la cepa Tul 2) (A. E. Eakin et al. 1992), organizados

en *tandem* en diferentes cromosomas, separados por espacios intergénicos de algo más de 400 pares de bases (O. Campetella et al. 1992), y es expresada en todos los estadíos del ciclo de vida del parásito (O. Campetella et al. 1990).

Cz tiene diferentes localizaciones en los cuatro estadios principales de *T. cruzi*, dependiendo del ciclo de vida y de las diversas funciones que realiza (Mckerrow et al. 2009). A través de un estudio de la actividad peptidasa en fracciones obtenidas por ultracentrifugación de epimastigotes, se vio que Cz se localizaba en los lisosomas (E. Bontempi et al. 1989). Este hallazgo fue confirmado más tarde por microscopía de inmunoelectrónica, que también demostró su presencia en la superficie de epimastigotes y amastigotes (P.S. Doyle et. al 2011). Por otro lado, en tripomastigotes, se localiza en los lisosomas y en el bolsillo flagelar, lo que sugiere un mecanismo para la secreción de la enzima en esta etapa del ciclo de vida (Souto-Padron et al. 1990).

Desde el punto de vista estructural, Cz consta de una única cadena polipeptídica que posee un dominio catalítico de 215 aminoácidos y un dominio C-terminal de 130 aminoácidos que se elimina de forma autocatalítica haciendo madurar a la enzima (U. Hellman et al. 1991). Este último dominio es característico de las cisteína proteasas de Tipo I de los trypanosomátidos y otras proteasas de parásitos (A. E. Eakin et al. 1992).

Figura 1.5 Representación esquemática de dominios C-terminal y catalítico en algunas cisteína proteasas (Adaptado de Verma et al. 2016).

Con respecto al dominio catalítico remanente, el mismo es altamente homólogo a las catepsinas S y L de mamífero y, en menor grado, a la papaína. Este dominio se encuentra formado a su vez por dos dominios que dejan entre ellos el surco correspondiente al sitio activo (Figura 1.6). El dominio izquierdo presenta tres regiones de hélices alfa, una de las cuales contiene el residuo Cys25 en su extremo N-terminal. El dominio derecho presenta principalmente una disposición de hojas beta antiparalelas. Por otro lado, la tríada catalítica, altamente conservada en la superfamilia de la papaína (Berti et al. 1995), se forma por los residuos de Cys25, His162 y Asn182. Por último, el sitio activo presenta diferentes subsitios o sub-bolsillos (S1-S4 y S1´-S3´) de unión al sustrato que se encuentran en la hendidura entre los dos dominios (Figura 1.6)(Huang et al. 2003).

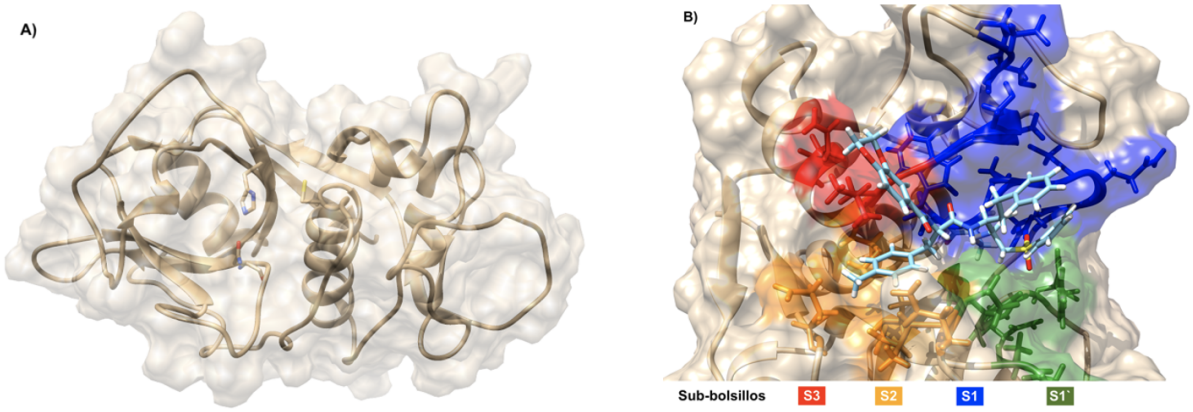


Figura 1.6 Estructura 3D de Cz, a) triada catalítica y b) sub-bolsillos con inhibidor unido covalentemente.

1.2.2 Mecanismo de acción de Cz

Actualmente existen dudas en relación al rol que cumplen los diferentes residuos de la triada catalítica en la formación del complejo enzima-sustrato (Zhai et al. 2018; Luchi et al. 2019). Por otro lado, se establece que los residuos desencadenantes de la catálisis únicamente son cisteína-histidina pudiendo encontrarse como especies neutras Cys-SH: His o en forma ionizada tiolato-imidazolio Cys-S-: H-His+ en la enzima libre, pero el estado de protonación inicial de esta díada catalítica ha sido objeto de considerable debate (Storer et al. 1994; Paasche et al. 2014).

Tomando como estado inicial ambos residuos en su forma neutra, el grupo imidazol de la histidina polariza el tiol de la cisteína permitiendo la desprotonación aún en condiciones neutras o débilmente ácidas. El par iónico S-imidazolio que se produce es altamente nucleofílico. Luego de la unión del sustrato se forma un complejo no covalente de Michaelis. El anión tiolato ataca al doble enlace del grupo carbonilo que se romperá (Figura 1.7a). Luego se forma un complejo tetraédrico el cual es estabilizado por el hueco del oxianión (Figura 1.7b). A continuación el complejo es acilado por la enzima y el primer producto es liberado. La hidrólisis del complejo acil-enzima da como resultado la formación del segundo intermediario tetraédrico (Figura 1.7c). Por último el intermediario colapsa y el ácido es liberado regenerando la enzima (Figura 1.7d) (Leung et al. 2000; Bellera 2014).

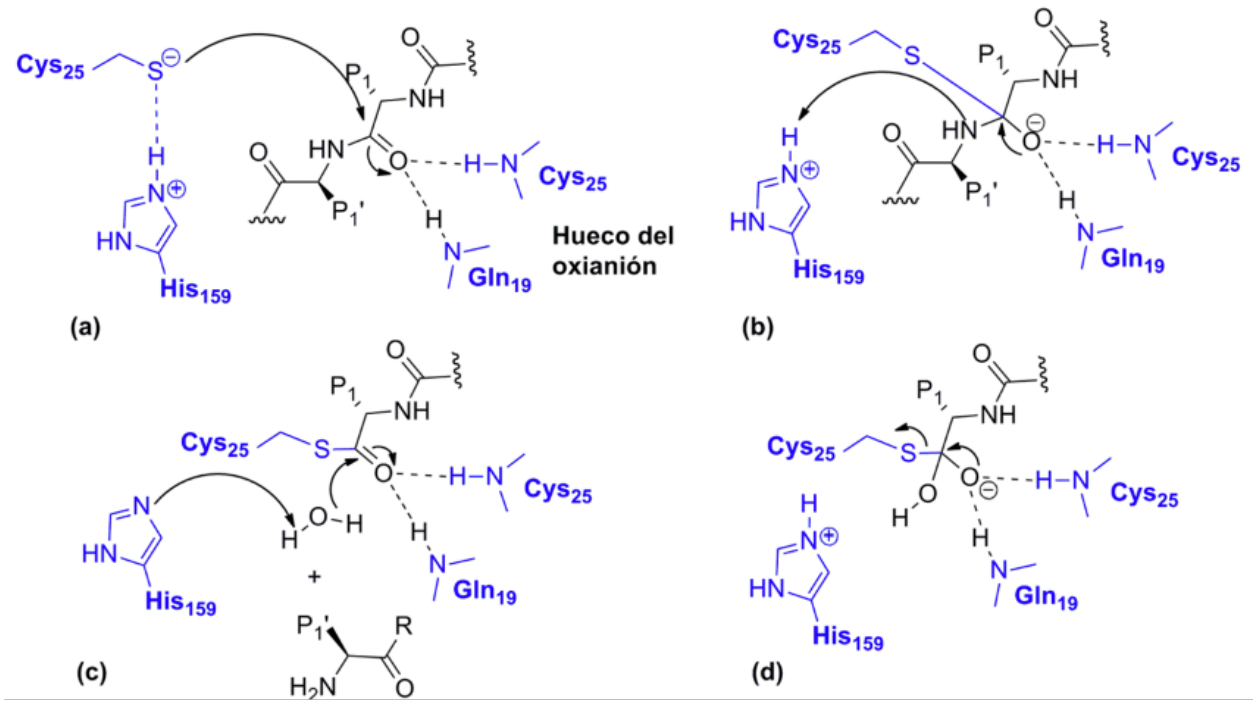


Figura 1.7 Mecanismo de acción propuesto para la Cz (tomado de Bellera 2014).

1.2.3 Invasión celular mediada por Cz

Varios estudios indican la importancia de Cz para la invasión de la célula huésped, basándose en los efectos de: i) inhibidores sintéticos (Bonaldo et al. 1991; Meirelles et al. 1992; Harth et al.

1993), ii) inhibidores endógenos, como la chagasina (Santos 2005) o iii) en los fenotipos expresados por parásitos genéticamente modificados, que expresan diferentes niveles de la enzima (Duschak et al. 2001; Aparicio et al. 2004).

Se han caracterizado dos vías para la invasión celular mediada por Cz, que difieren en su dependencia de las proteínas cininas.

En la primera vía propuesta, Cz genera bradiquinina a través de la escisión del quininógeno. Los ensayos con células que expresan altos niveles de receptores de bradiquinina 2 (B_2R) mostraron que *T. cruzi* podía inducir aumentos transitorios de Ca^{2+}. Las vías mediadas por bradiquinina, a través de B_1R y B_2R, están relacionadas con el desarrollo de una respuesta edematogénica, contribuyendo a la patología de la enfermedad de Chagas (Todorov et al. 2003). La respuesta inflamatoria generada conduce a la fuga de plasma y la acumulación de quininógenos, lo que facilita la invasión celular y la proliferación de T. cruzi (Andrade et al. 2012). Además, la liberación de quinina por cruzaína está modulada por heparan sulfato (HS). En presencia de HS, K_{cat} aumenta y K_m disminuye, mejorando la actividad enzimática (Figura 1.8) (Paula et al. 2001).

En segundo lugar, se ha descrito una vía de invasión independiente de quininas, que no se ve afectada por los antagonistas de B_2R o B_1R. En esta vía solo se propone que el poder infeccioso de *T. cruzi* está directamente relacionado con los niveles de expresión de Cz. Por otro lado, experimentos adicionales indicaron que la enzima procesa un sustrato liberado por tripomastigotes, formando la sustancia realmente responsable de la invasión. De manera similar a los mecanismos descritos anteriormente, esto también implicaba la movilización de Ca^{2+} (Aparicio et al. 2004).

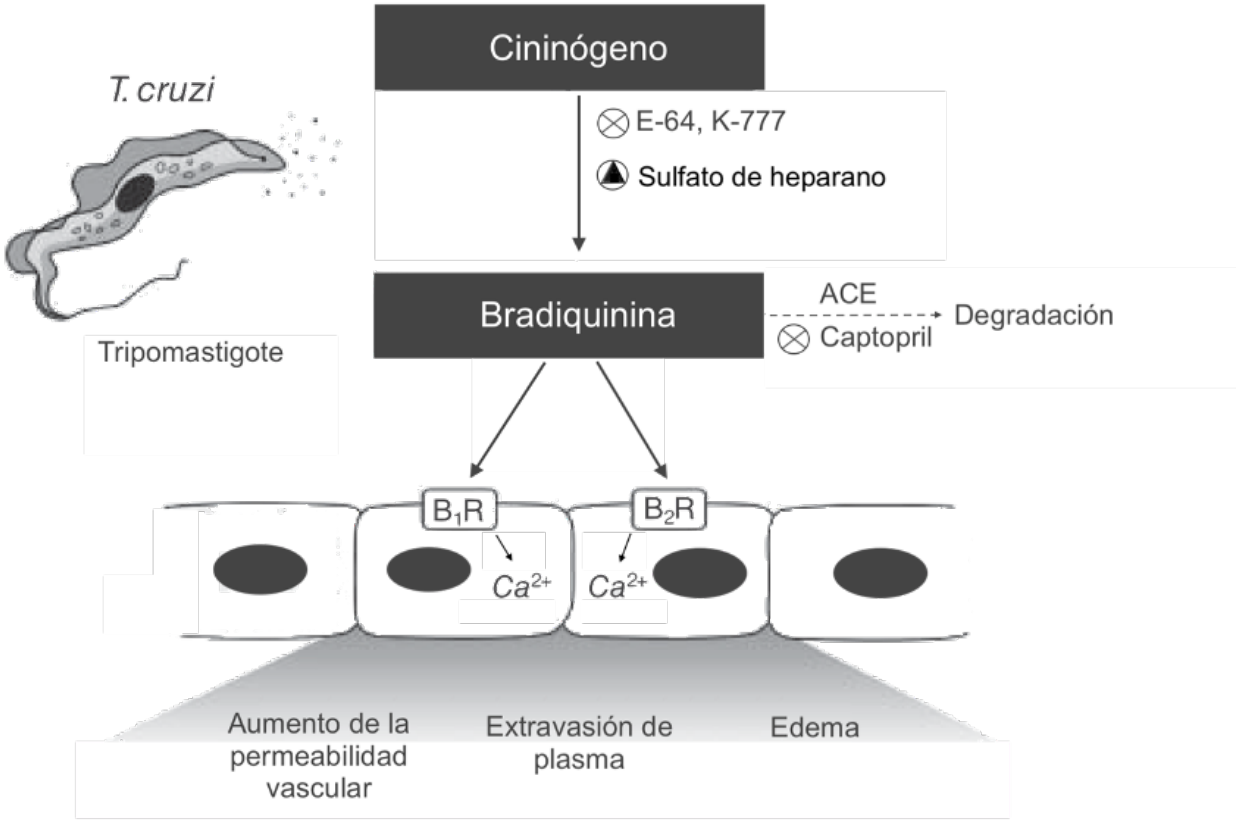


Figura 1.8 Cz promueve la invasión celular a través de una vía mediada por quininas. La enzima genera bradiquinina a partir del quininógeno. Esto puede ser inhibido por E-64 y K777 (⊗) y potenciada por heparán sulfato (▲). La bradiquinina actúa sobre los receptores de bradiquinina tipo 1 y 2 (B_1R y B_2R), lo que conduce a la liberación de Ca^{2+} y aumenta la permeabilidad vascular, el edema y la extravasión de plasma. El captopril, un inhibidor de la enzima convertidora de angiotensina (ACE, siglas en inglés para *angiotensin-converting enzyme*), reduce la degradación de la bradiquinina y potencia la invasión de *T. cruzi* (Figura adaptada de da Silva, et al. 2016).

1.2.4 Inhibidores de Cz

En las últimas décadas se identificaron diferentes familias de compuestos químicos capaces de inhibir a la Cz. Estas familias de compuestos incluían pequeñas moléculas peptídicas y no peptídicas que interactúan con la enzima de manera reversible o irreversible (Du et al. 2002; Choe et al. 2005) (ver tabla 1.2).

Entre los inhibidores peptídicos, generalmente se encuentran dipéptidos o tripéptidos que poseen un grupo funcional electrofílico “cabeza de guerra” o *warhead* que posibilita la unión covalente a la cisteína del sitio activo mediante un ataque nucleofílico (Roush et al. 2001; Choe

et al. 2005; Bryant et al. 2009). A pesar de la alta afinidad que presentan estos compuestos, se encontraron dificultades principalmente vinculadas a la carencia de selectividad, alta reactividad, toxicidad y baja biodisponibilidad oral de los mismos; por tal motivo, investigaciones anteriores han puesto especial énfasis en la identificación de inhibidores no peptídicos, de tipo droga (drug-like) y preferentemente reversibles (Capaci Rodrigues et al. 2010).

Las vinilsulfonas son una clase ampliamente estudiada para la que se han descrito derivados peptídicos (Götz et al. 2004; Chen et al. 2008) y no peptídicos (Bryant et al. 2009). Entre los ejemplos mas notables, el inhibidor K-777, que se caracterizó originalmente por el Centro Sandler para la Investigación de Enfermedades Parasitarias Tropicales en la Universidad de California, San Francisco, fue el inhibidor de cisteína proteasa más prometedor.

Mediante estudios experimentales realizado sobre ratones infectados con *T. cruzi,* se ha demostrado que el inhibidor K-777 bloquea eficazmente la actividad de la Cz, produciendo mejoras en ratones en fase aguda y a su vez, elimina la parasitemia en ratones con infección crónica (Urbina y Docampo 2003). K-777 es un inhibidor que promueve la acumulación de la enzima no procesada en el aparato de Golgi, lo que conduce a la inviabilidad del parásito (Engel et al. 1998).

En un modelo agudo de infección en perros, K-777 se vio que no promueve la cura parasitológica pero reduce significativamente el daño cardíaco inducido por el parásito (Barr et al. 2005). Desafortunadamente, el desarrollo de este compuesto se interrumpió en estudios preclínicos debido a problemas de tolerabilidad a dosis bajas en primates y perros (DNDi 2014) y hasta el momento no se cuenta con nuevos fármacos en etapas avanzadas capaces de mitigar la enfermedad.

Tabla 1.2 Resumen de los inhibidores de Cz desarrollados a lo largo de los años: se resumen los tipos de compuestos, el rango de actividades y el número de moléculas en cada conjunto (Martinez-Mayorga et al. 2015).

Nº moléculas	Rango de actividad IC50 (uM)	Tipos de moléculas
116	0,019-1000	Tiosemicarbazonas
20	0,024-1	Tetrahidropiranos
5	0,01-10	Dipeptidil epoxy ester
3	0,004-10	Vinil sulfonas
9	32-100	Hidrazonas y benzofuranos
9	85-150	Pirimidinas
22	0,2-1000	Oxidiazoles
23	0,21-77,5	Benzimidazol y/o tiosemicarbazonas

1.3 Objetivos generales

El objetivo general de este plan de trabajo es la búsqueda de nuevos inhibidores reversibles de la Cruzipaina (Cz).

1.4 Objetivos particulares

- Recopilar la información (estructural y de inhibición) disponibles sobre inhibidores de Cz para generar descriptores basados en la densidad electrónica y sus propiedades locales asociadas.

- Utilizar herramientas de aprendizaje no supervisado para extraer relaciones entre elementos topológicos de la densidad de carga y los valores de inhibición de ligandos conocidos.

- Entrenar una Red Neuronal Gráfica capaz de establecer relaciones no lineales entre los datos estructurales y biológicos utilizando una biblioteca de ligandos conocidos.

- Realizar un cribado virtual retrospectivo mediado por puntos farmacofóricos, a modo de docking guiado, utilizando bibliotecas de ligandos con anotaciones de afinidad.

- Cribar una biblioteca de ligandos para priorizar nuevas moléculas capaces de unirse al sitio activo de Cz.

Referencias del capítulo 1

Aldasoro, E., E. Posada, A. Requena-Méndez, A. Calvo-Cano, N. Serret, A. Casellas, S. Sanz, D. Soy, J. Pinazo, and J. Gascon. 2018. “What to Expect and When: Benznidazole Toxicity in Chronic Chagas’ Disease Treatment.” *Journal of Antimicrobial Chemotherapy* 73 (4): 1060–67. https://doi.org/10.1093/jac/dkx516.

Altcheh, Jaime, Guillermo Moscatelli, Samanta Moroni, Facundo Garcia-Bournissen, and Hector Freilij. 2011. “Adverse Events after the Use of Benznidazole in Infants and Children with Chagas Disease.” *Pediatrics* 127 (1): e212–18. https://doi.org/10.1542/peds.2010-1172.

Andrade, Daniele, Rafaela Serra, Erik Svensjö, Ana Paula C. Lima, Erivan S. Ramos Junior, Fabio S. Fortes, Ana Carolina F. Morandini, et al. 2012. “Trypanosoma Cruzi Invades Host Cells through the Activation of Endothelin and Bradykinin Receptors: A Converging Pathway Leading to Chagasic Vasculopathy.” *British Journal of Pharmacology* 165 (5): 1333–47. https://doi.org/10.1111/j.1476-5381.2011.01609.x.

Angheben, Andrea, Lucia Boix, Dora Buonfrate, Federico Gobbi, Zeno Bisoffi, Simonetta Pupella, Giorgio Gandini, and Giuseppe Aprili. 2015. “Chagas Disease and Transfusion Medicine: A Perspective from Non-Endemic Countries.” *Blood Transfusion* 13 (4): 540–50. https://doi.org/10.2450/2015.0040-15.

Aparicio, Isabela M., Julio Scharfstein, and Ana Paula C.A. Lima. 2004. “A New Cruzipain-Mediated Pathway of Human Cell Invasion by Trypanosoma Cruzi Requires Trypomastigote Membranes.” *Infection and Immunity* 72 (10): 5892–5902. https://doi.org/10.1128/IAI.72.10.5892-5902.2004.

Aufderheide, AF, W Salo, M Madden, J Streitz, J Buikstra, F Guhl, B Arriaza, et al. 2004. “A 9,000-Year Record of Chagas’ Disease.” *Proc Natl Acad Sci USA* 101 (7): 2034–39.

Bahia, Maria T., Alvaro F.S. Nascimento, Ana Lia Mazzeti, Luiz F. Marques, Karolina R. Gonçalves, Ludmilla W.R. Mota, Lívia F. De Diniz, et al. 2014. “Antitrypanosomal Activity of Fexinidazole Metabolites, Potential New Drug Candidates for Chagas Disease.” *Antimicrobial Agents and Chemotherapy* 58 (8): 4362–70. https://doi.org/10.1128/AAC.02754-13.

Bahia, Maria Terezinha, Isabel Mayer de Andrade, Tassiane Assíria Fontes Martins, Álvaro

Fernando da Silva do Nascimento, Lívia de Figueiredo Diniz, Ivo Santana Caldas, André Talvani, Bernadette Bourdin Trunz, Els Torreele, and Isabela Ribeiro. 2012. “Fexinidazole: A Potential New Drug Candidate for Chagas Disease.” *PLoS Neglected Tropical Diseases* 6 (11). https://doi.org/10.1371/journal.pntd.0001870.

Barr, S. C., K. L. Warner, B. G. Kornreic, J. Piscitelli, A. Wolfe, L. Benet, and J. H. McKerrow. 2005. “A Cysteine Protease Inhibitor Protects Dogs from Cardiac Damage during Infection by Trypanosoma Cruzi.” *Antimicrobial Agents and Chemotherapy* 49 (12): 5160–61. https://doi.org/10.1128/AAC.49.12.5160-5161.2005.

Belaunzarán, María Laura. 2015. “Chagas Disease: Globalization and New Hope for Its Cure.” *Revista Argentina de Microbiologia* 47 (2): 85–87. https://doi.org/10.1016/j.ram.2015.04.001.

Bellera, Carolina. 2014. “Busqueda Racional de Nuevos Fármacos Antichagásicos Inhibidores de La Cruzipaína.” Universidad Nacional de La Plata.

Bern, Caryn. 2015. “Chagas’ Disease.” *New England Journal of Medicine* 373 (5): 456–66. https://doi.org/10.1056/NEJMra1410150.

Berti, Paul J., and Andrew C. Storer. 1995. “Alignment/Phylogeny of the Papain Superfamily of Cysteine Proteases.” *Journal of Molecular Biology* 246 (2): 273–83. https://doi.org/10.1006/jmbi.1994.0083.

Bonaldo, Myrna C., Luiz Ney d’Escoffier, Jussara M. Salles, and Samuel Goldenberg. 1991. “Characterization and Expression of Proteases during Trypanosoma Cruzi Metacyclogenesis.” *Experimental Parasitology* 73 (1): 44–51. https://doi.org/10.1016/0014-4894(91)90006-I.

Bontempi, Esteban, Javier Martinez, and Juan José Cazzulo. 1989. “Subcellular Localization of a Cysteine Proteinase from Trypanosoma Cruzi.” *Molecular and Biochemical Parasitology* 33 (1): 43–47. https://doi.org/10.1016/0166-6851(89)90040-6.

Brener, Z. 1973. “Biology of Trypanosoma Cruzi.” *Annual Review of Microbiology* 27 (1): 347–82. https://doi.org/10.1146/annurev.mi.27.100173.002023.

Bryant, Clifford, Iain D Kerr, Moumita Debnath, Kenny K H Ang, Joseline Ratnam, Rafaela S Ferreira, Priyadarshini Jaishankar, et al. 2009. “Bioorganic & Medicinal Chemistry Letters Novel Non-Peptidic Vinylsulfones Targeting the S2 and S3 Subsites of Parasite Cysteine

Proteases." *Bioorganic & Medicinal Chemistry Letters* 19 (21): 6218–21. https://doi.org/10.1016/j.bmcl.2009.08.098.

C. Storer, Andrew, and Robert Ménard. 1994. "Catalytic Mechanism in Papain Family of Cysteine Peptidases." *Methods in Enzymology* 244 (C): 486–500. https://doi.org/10.1016/0076-6879(94)44035-2.

Campetella, O., J. Henriksson, U. Åslund, A. C.C. Frasch, U. Pettersson, and J. J. Cazzulo. 1992. "The Major Cysteine Proteinase (Cruzipain) from Trypanosoma Cruzi Is Encoded by Multiple Polymorphic Tandemly Organized Genes Located on Different Chromosomes." *Molecular and Biochemical Parasitology* 50 (2): 225–34. https://doi.org/10.1016/0166-6851(92)90219-A.

Campetella, Oscar, Javier Martínez, and Juan José Cazzulo. 1990. "A Major Cysteine Proteinase Is Developmentally Regulated in Trypanosoma Cruzi." *FEMS Microbiology Letters* 67 (1–2): 145–50. https://doi.org/10.1111/j.1574-6968.1990.tb13852.x.

Capaci Rodrigues, Giseli, Alcino Palermo Aguiar, Joao Lidio da Silva Goncalves Vianez, Andrew Macrae, Ana Cristina Nogueira de Melo, and Alane Beatriz Vermelho. 2010. "Peptidase Inhibitors as a Possible Therapeutic Strategy for Chagas Disease." *Current Enzyme Inhibition* 6 (4): 183–94. https://doi.org/10.2174/157340810794578506.

Cardozo, E. (2016). Análisis de los conocimientos, actitudes y prácticas de médicos del primer nivel de atención de la Argentina respecto al tratamiento etiológico de pacientes con enfermedad de Chagas crónica 2012.

Carneiro, Cláudia Martins, Adrián Sánchez-Montalvá, Rodrigo Correa de Oliveira, Policarpo Ademar Sales Junior, Silvane Maria Fonseca Murta, Fernando Salvador, and Israel Molina. 2017. "Experimental and Clinical Treatment of Chagas Disease: A Review." *The American Journal of Tropical Medicine and Hygiene* 97 (5): 1289–1303. https://doi.org/10.4269/ajtmh.16-0761.

Cazzulo, Juan Jose, Ulf Hellman, Roberto Couso, and Armando J.A. Parodi. 1990. "Amino Acid and Carbohydrate Composition of a Lysosomal Cysteine Proteinase from Trypanosoma Cruzi. Absence of Phosphorylated Mannose Residues." *Molecular and Biochemical Parasitology* 38 (1): 41–48. https://doi.org/10.1016/0166-6851(90)90203-X.

Cazzulo, Juan, Veronika Stoka, and Vito Turk. 2005. “The Major Cysteine Proteinase of Trypanosoma Cruzi: A Valid Target for Chemotherapy of Chagas Disease.” *Current Pharmaceutical Design* 7 (12): 1143–56. https://doi.org/10.2174/1381612013397528.

Centers for Disease Control and Prevention. 2015. “CDC - Chagas Disease - Biology.” Centers for Disease Control and Prevention. 2015. https://www.cdc.gov/parasites/chagas/biology.html.

Cerny, N. 2016. “Antígenos de Trypanosoma Cruzi Para La Inmunoterapia de La Infección.” Facultad de Farmacia y Bioquímica. Universidad de Buenos Aires. http://repositoriouba.sisbi.uba.ar/gsdl/collect/posgrauba/index/assoc/HWA_1179.dir/1179.PDF.

Chatelain, Eric. 2015. “Chagas Disease Drug Discovery: Toward a New Era.” *Journal of Biomolecular Screening* 20 (1): 22–35. https://doi.org/10.1177/1087057114550585.

Chen, Yen Ting, Ricardo Lira, Elizabeth Hansell, James H. McKerrow, and William R. Roush. 2008. “Synthesis of Macrocyclic Trypanosomal Cysteine Protease Inhibitors.” *Bioorganic and Medicinal Chemistry Letters* 18 (22): 5860–63. https://doi.org/10.1016/j.bmcl.2008.06.012.

Choe, Youngchool, Linda S. Brinen, Mark S. Price, Juan C. Engel, Meinolf Lange, Corinna Grisostomi, Scott G. Weston, et al. 2005. “Development of α-Keto-Based Inhibitors of Cruzain, a Cysteine Protease Implicated in Chagas Disease.” *Bioorganic and Medicinal Chemistry* 13 (6): 2141–56. https://doi.org/10.1016/j.bmc.2004.12.053.

Coura, José Rodrigues. 2007. “Chagas Disease: What Is Known and What Is Needed--a Background Article.” *Memorias Do Instituto Oswaldo Cruz* 102: 113–22.

Coura, José Rodrigues, and José Borges-Pereira. 2010. “Chagas Disease: 100 Years after Its Discovery. A Systemic Review.” *Acta Tropica* 115 (1–2): 5–13. https://doi.org/10.1016/j.actatropica.2010.03.008.

Coura, José Rodrigues, and JC Dias. 2009. “Epidemiology, Control and Surveillance of Chagas Disease: 100 Years after Its Discovery.” *Memorias Do Instituto Oswaldo Cruz* 104: 31–40.

Cristovão-Silva, Ana Catarina, Maria Carolina Accioly Brelaz-De-Castro, Ana Cristina Lima Leite, Valéria Rego Alves Pereira, and Marcelo Zaldini Hernandes. 2019. “Chagas Disease

Treatment and Rational Drug Discovery: A Challenge That Remains." *Frontiers in Pharmacology* 10 (JULY): 1–6. https://doi.org/10.3389/fphar.2019.00873.

DNDi. 2014. "DNDi Portfolio K777 (Chagas)." 2014. https://dndi.org/research-development/portfolio/k777/.

Doyle PS, Zhou YM, Hsieh I, Greenbaum DC, McKerrow JH, et al. (2011) *The Trypanosoma cruzi Protease Cruzain Mediates Immune Evasion*. PLOS Pathogens 7(9): e1002139. https://doi.org/10.1371/journal.ppat.1002139

Drugs for Neglected Diseases Initiative. 2019. "The BENDITA Study: Chagas Disease." www.dndi.org.

Du, Xiaohui, Chun Guo, Elizabeth Hansell, Patricia S. Doyle, Conor R. Caffrey, Tod P. Holler, James H. McKerrow, and Fred E. Cohen. 2002. "Synthesis and Structure−Activity Relationship Study of Potent Trypanocidal Thio Semicarbazone Inhibitors of the Trypanosomal Cysteine Protease Cruzain." *Journal of Medicinal Chemistry* 45 (13): 2695–2707. https://doi.org/10.1021/jm010459j.

Duelen, R., Corvelyn, M., Tortorella, I., Leonardi, L., Chai, Y.C., Sampaolesi, M. (2019). Medicinal Biotechnology for Disease Modeling, Clinical Therapy, and Drug Discovery and Development. In: Matei, F., Zirra, D. (eds) Introduction to Biotech Entrepreneurship: From Idea to Business. Springer, Cham. https://doi.org/10.1007/978-3-030-22141-6_5

Duschak, Vilma G., Mirella Ciaccio, Julio R. Nasser, Miguel A. Basombrío, and Miguel A. Basombrio. 2001. "Enzymatic Activity, Protein Expression, and Gene Sequence of Cruzipain in Virulent and Attenuated Trypanosoma Cruzi Strains." *The Journal of Parasitology* 87 (5): 1016. https://doi.org/10.2307/3285225.

Eakin, A. E., A. A. Mills, G. Harth, J. H. McKerrow, and C. S. Craik. 1992. "The Sequence, Organization, and Expression of the Major Cysteine Protease (Cruzain) from Trypanosoma Cruzi." *Journal of Biological Chemistry* 267 (11): 7411–20. https://pubmed.ncbi.nlm.nih.gov/1559982/.

Eakin, Ann E., Jacques Bouvier, Judy A. Sakanari, Charles S. Craik, and James H. McKerrow.

1990. “Amplification and Sequencing of Genomic DNA Fragments Encoding Cysteine Proteases from Protozoan Parasites.” *Molecular and Biochemical Parasitology* 39 (1): 1–8. https://doi.org/10.1016/0166-6851(90)90002-4.

Engel, Juan C., Patricia S. Doyle, Ivy Hsieh, and James H. McKerrow. 1998. “Cysteine Protease Inhibitors Cure an Experimental Trypanosoma Cruzi Infection.” *Journal of Experimental Medicine* 188 (4): 725–34. https://doi.org/10.1084/jem.188.4.725.

FDA. “FDA-US/Chagas.”

García-Huertas, P., & Cardona-Castro, N. (2021). Advances in the treatment of Chagas disease: Promising new drugs, plants and targets. *Biomedicine & Pharmacotherapy*, *142*,

Gaspe, M. S., Provecho, Y. M., Fernandez, M. P., Vassena, C. V., Santo Orihuela, P. L., & Gürtler, R. E. (2018). Beating the odds: Sustained Chagas disease vector control in remote indigenous communities of the Argentine Chaco over a seven-year period. PLoS *neglected tropical diseases*, 12(10), e0006804.

Götz, Marion G., Conor R. Caffrey, Elizabeth Hansell, James H. McKerrow, and James C. Powers. 2004. “Peptidyl Allyl Sulfones: A New Class of Inhibitors for Clan CA Cysteine Proteases.” *Bioorganic & Medicinal Chemistry* 12 (19): 5203–11. https://doi.org/10.1016/j.bmc.2004.07.016.

Guhl, F, C Jaramillo, GA Vallejo, R Yockteng, F Cárdenas-Arroyo, G Fornaciari, B Arriaza, and AC Aufderheide. 1999. “Isolation of Trypanosoma Cruzi DNA in 4,000-Year-Old Mummified Human Tissue from Northern Chile.” *Am J Phys Anthropol* 108 (4): 401–7.

Gürtler, R. E., Kitron, U., Cecere, M. C., Segura, E. L., & Cohen, J. E. (2007). Sustainable vector control and management of Chagas disease in the Gran Chaco, Argentina. *Proceedings of the National Academy of Sciences*, 104(41), 16194-16199.

Haberland, Annekathrin, Silvia Gilka Munoz Saravia, Gerd Wallukat, Reinhard Ziebig, and Ingolf Schimke. 2013. “Chronic Chagas Disease: From Basics to Laboratory Medicine.” *Clinical Chemistry and Laboratory Medicine*. https://doi.org/10.1515/cclm-2012-0316.

Harth, Guenter, Norma Andrews, Alea A. Mills, Juan C. Engel, Robert Smith, and James H. McKerrow. 1993. “Peptide-Fluoromethyl Ketones Arrest Intracellular Replication and Intercellular Transmission of Trypanosoma Cruzi.” *Molecular and Biochemical Parasitology*

58 (1): 17–24. https://doi.org/10.1016/0166-6851(93)90086-D.

Hellman, Ulf, Christer Wernstedt, and Juan José Cazzulo. 1991. “Self-Proteolysis of the Cysteine Proteinase, Cruzipain, from Trypanosoma Cruzi Gives a Major Fragment Corresponding to Its Carboxy-Terminal Domain.” *Molecular and Biochemical Parasitology* 44 (1): 15–21. https://doi.org/10.1016/0166-6851(91)90216-S.

Huang, Lily, Linda S. Brinen, and Jonathan A. Ellman. 2003. “Crystal Structures of Reversible Ketone-Based Inhibitors of the Cysteine Protease Cruzain.” *Bioorganic and Medicinal Chemistry* 11 (1): 21–29. https://doi.org/10.1016/S0968-0896(02)00427-3.

Jackson, Yves, Emilie Alirol, Laurent Getaz, Hans Wolff, Christophe Combescure, and François Chappuis. 2010. “Tolerance and Safety of Nifurtimox in Patients with Chronic Chagas Disease.” *Clinical Infectious Diseases* 51 (10): e69–75. https://doi.org/10.1086/656917.

Lent, Herman, and Pedro W. Wygodzinsky. 1979. “Revision of the Triatominae (Hemiptera, Reduviidae), and Their Significance as Vectors of Chagas’ Disease.” *Bulletin of the AMNH*. Vol. 163, art 3.

Leung, Donmienne, Giovanni Abbenante, and David P. Fairlie. 2000. “Protease Inhibitors: Current Status and Future Prospects.” *Journal of Medicinal Chemistry*. American Chemical Society . https://doi.org/10.1021/jm990412m.

Levine, Norman D. 1973. “The Trypanosomes of Mammals. A Zoological Monograph. Cecil A. Hoare. Blackwell, Oxford, England, 1972 (U.S. Distributor, Davis, Philadelphia). Xviii, 750 Pp.” *Science* 179 (4068): 60 LP – 60. https://doi.org/10.1126/science.179.4068.60.

Luchi, Adriano M., Roxana N. Villafañe, J. Leonardo Gómez Chávez, M. Lucrecia Bogado, Emilio L. Angelina, and Nelida M. Peruchena. 2019. “Combining Charge Density Analysis with Machine Learning Tools to Investigate the Cruzain Inhibition Mechanism.” *ACS Omega* 4 (22): 19582–94. https://doi.org/10.1021/acsomega.9b01934.

Martinez-Mayorga, Karina, Kendall G. Byler, Ariadna I. Ramirez-Hernandez, and Diana E. Terrazas-Alvares. 2015. “Cruzain Inhibitors: Efforts Made, Current Leads and a Structural Outlook of New Hits.” *Drug Discovery Today* 20 (7): 890–98. https://doi.org/10.1016/j.drudis.2015.02.004.

Maya, Juan Diego, Bruce K. Cassels, Patricio Iturriaga-Vásquez, Jorge Ferreira, Mario

Faúndez, Norbel Galanti, Arturo Ferreira, and Antonio Morello. 2007. “Mode of Action of Natural and Synthetic Drugs against Trypanosoma Cruzi and Their Interaction with the Mammalian Host.” *Comparative Biochemistry and Physiology - A Molecular and Integrative Physiology*. Elsevier Inc. https://doi.org/10.1016/j.cbpa.2006.03.004.

Maya, Juan Diego, Andrés Rodríguez, Laura Pino, Adriana Pabón, Jorge Ferreira, Mario Pavani, Yolanda Repetto, and Antonio Morello. 2004. “Effects of Buthionine Sulfoximine Nifurtimox and Benznidazole upon Trypanothione and Metallothionein Proteins in Trypanosoma Cruzi.” *Biological Research* 37 (1): 61–69. https://doi.org/10.4067/S0716-97602004000100007.

Mckerrow, J H, P S Doyle, J C Engel, L M Podust, S A Robertson, R Ferreira, T Saxton, et al. 2009. “Two Approaches to Discovering and Developing New Drugs for Chagas Disease.” *Memorias Do Instituto Oswaldo Cruz*. https://doi.org/10.1590/S0074-02762009000900034.

Meirelles, Maria Nazareth L., Luiz Juliano, Euridice Carmona, Suelen G. Silva, Elizabeth M. Costa, Ana C.M. Murta, and Julio Scharfstein. 1992. “Inhibitors of the Major Cysteinyl Proteinase (GP57/51) Impair Host Cell Invasion and Arrest the Intracellular Development of Trypanosoma Cruzi in Vitro.” *Molecular and Biochemical Parasitology* 52 (2): 175–84. https://doi.org/10.1016/0166-6851(92)90050-T.

Molina, Israel, Jordi Gońez I Prat, Fernando Salvador, Begoña Treviño, Elena Sulleiro, Núria Serre, Diana Pou, et al. 2014. “Randomized Trial of Posaconazole and Benznidazole for Chronic Chagas’ Disease.” *New England Journal of Medicine* 370 (20): 1899–1908. https://doi.org/10.1056/NEJMoa1313122.

Montalto De Mecca, María, Laura C. Bartel, Carmen Rodríguez De Castro, and José A. Castro. 2008. “Benznidazole Biotransformation in Rat Heart Microsomal Fraction without Observable Ultrastructural Alterations: Comparison to Nifurtimox-Induced Cardiac Effects.” *Memorias Do Instituto Oswaldo Cruz* 103 (6): 549–53. https://doi.org/10.1590/s0074-02762008000600007.

Murta, S. M.F., C. Ropert, R. O. Alves, R. T. Gazzinelli, and A. J. Romanha. 1999. “In-Vivo Treatment with Benznidazole Enhances Phagocytosis, Parasite Destruction and Cytokine Release by Macrophages during Infection with a Drug-Susceptible but Not with a Derived

Drug-Resistant Trypanosoma Cruzi Population." *Parasite Immunology* 21 (10): 535–44. https://doi.org/10.1046/j.1365-3024.1999.00251.x.

Oliveira, Antonio Edson R., Viviane Grazielle-Silva, Ludmila R.P. Ferreira, and Santuza M.R. Teixeira. 2020. "Close Encounters between Trypanosoma Cruzi and the Host Mammalian Cell: Lessons from Genome-Wide Expression Studies." *Genomics* 112 (1): 990–97. https://doi.org/10.1016/j.ygeno.2019.06.015.

Paasche, Alexander, Andreas Zipper, Simon Schäfer, John Ziebuhr, Tanja Schirmeister, and Bernd Engels. 2014. "Evidence for Substrate Binding-Induced Zwitterion Formation in the Catalytic Cys-His Dyad of the SARS-CoV Main Protease." *Biochemistry* 53 (37): 5930–46. https://doi.org/10.1021/bi400604t.

Paula, Ana, C A Lima, Paulo C Almeida, Ivarne L S Tersariol, Veronica Schmitz, Alvin H Schmaier ¶, Luiz Juliano, et al. 2001. "Heparan Sulfate Modulates Kinin Release by Trypanosoma Cruzi through the Activity of Cruzipain*." https://doi.org/10.1074/jbc.M108518200.

Pérez-Molina, José A., and Israel Molina. 2018. "Chagas Disease." *The Lancet* 391 (10115): 82–94. https://doi.org/10.1016/S0140-6736(17)31612-4.

Pérez-Morales, Deyanira, Humberto Lanz-Mendoza, Gerardo Hurtado, Rodrigo Martínez-Espinosa, and Bertha Espinoza. 2012. "Proteomic Analysis of Trypanosoma Cruzi Epimastigotes Subjected to Heat Shock." *Journal of Biomedicine and Biotechnology* 2012. https://doi.org/10.1155/2012/902803.

Pinazo, María Jesús, Elías G. Cañas, Jose Ignacio Elizalde, Magdalena García, Joaquim Gascón, Fausto Gimeno, Jordi Gomez, et al. 2010. "Diagnosis, Management and Treatment of Chronic Chagas' Gastrointestinal Disease in Areas Where Trypanosoma Cruzi Infection Is Not Endemic." *Gastroenterologia y Hepatologia* 33 (3): 191–200. https://doi.org/10.1016/j.gastrohep.2009.07.009.

Rodrigues Coura, José, and Solange L. De Castro. 2002. "A Critical Review on Chagas Disease Chemotherapy." *Memorias Do Instituto Oswaldo Cruz* 97 (1): 3–24. https://doi.org/10.1590/S0074-02762002000100001.

Roemer, Terry, and Damian J Krysan. 2014. "Antifungal Drug Development: Challenges, Unmet

Clinical Needs, and New Approaches." *Cold Spring Harbor Perspectives in Medicine*. National Academies Press (US). https://doi.org/10.1101/cshperspect.a019703.

Roush, William R., Jianming Cheng, Beth Knapp-Reed, Alejandro Alvarez-Hernandez, James H. McKerrow, Elizabeth Hansell, and Juan C. Engel. 2001. "Potent Second Generation Vinyl Sulfonamide Inhibitors of the Trypanosomal Cysteine Protease Cruzain." *Bioorganic and Medicinal Chemistry Letters* 11 (20): 2759–62. https://doi.org/10.1016/S0960-894X(01)00566-2.

Santos, C. C. 2005. "Chagasin, the Endogenous Cysteine-Protease Inhibitor of Trypanosoma Cruzi, Modulates Parasite Differentiation and Invasion of Mammalian Cells." *Journal of Cell Science* 118 (5): 901–15. https://doi.org/10.1242/jcs.01677.

Seco-Hidalgo, Víctor, Luis Miguel De Pablos, and Antonio Osuna. 2015. "Transcriptional and Phenotypical Heterogeneity of Trypanosoma Cruzi Cell Populations." *Open Biology* 5 (12). https://doi.org/10.1098/rsob.150190.

Silva, E.B., do Nascimento Pereira, G.A. and Ferreira, R.S. da. 2016. "Trypanosomal Cysteine Peptidases: Target Validation and Drug Design Strategies." In *Comprehensive Analysis of Parasite Biology: From Metabolism to Drug Discovery*, 121–45.

Souto-Padron, T., O. E. Campetella, J. J. Cazzulo, and W. De Souza. 1990. "Cysteine Proteinase in Trypanosoma Cruzi: Immunocytochemical Localization and Involvement in Parasite-Host Cell Interaction." *Journal of Cell Science* 96 (3): 485–90. https://jcs.biologists.org/content/96/3/485.long.

Souza, Anacleto S. de, Marcelo T. de Oliveira, and Adriano D. Andricopulo. 2017. "Development of a Pharmacophore for Cruzain Using Oxadiazoles as Virtual Molecular Probes: Quantitative Structure–Activity Relationship Studies." *Journal of Computer-Aided Molecular Design* 31 (9): 801–16. https://doi.org/10.1007/s10822-017-0039-0.

Thakare, R., Dasgupta, A., & Chopra, S. (2021). "Update on nifurtimox for treatment of Chagas disease". *Drugs Today,* 57(4), 251-263.

Todorov, Alex G., Daniele Andrade, João B. Pesquero, Ronaldo Carvalho Araujo, Michael Bader, John Stewart, Lajos Gera, et al. 2003. "Trypanosoma Cruzi Induces Edematogenic Responses in Mice and Invades Cardiomyocytes and Endothelial Cells in Vitro by Activating

Distinct Kinin Receptor Subtypes (B_1 /B_2).” *The FASEB Journal* 17 (1): 73–75. https://doi.org/10.1096/fj.02-0477fje.

Torrico, F., Gascón, J., Ortiz, L., Pinto, J., Rojas, G., Palacios, A., Barreira, F., Blum, B., Schijman, A. G., Vaillant, M., Strub-Wourgaft, N., Pinazo, M. J., Bilbe, G., & Ribeiro, I. (2023). A Phase 2, Randomized, Multicenter, Placebo-Controlled, Proof-of-Concept Trial of Oral Fexinidazole in Adults With Chronic Indeterminate Chagas Disease. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 76(3), e1186–e1194. https://doi.org/10.1093/cid/ciac579.

Turrens, Julio F., Benjamin P. Watts, Li Zhong, and Roberto Docampo. 1996. “Inhibition of Trypanosoma Cruzi and T. Brucei NADH Fumarate Reductase by Benznidazole and Anthelmintic Imidazole Derivatives.” *Molecular and Biochemical Parasitology* 82 (1): 125–29. https://doi.org/10.1016/0166-6851(96)02722-3.

Tyler, K. M., and D. M. Engman. 2001. “The Life Cycle of Trypanosoma Cruzi Revisited.” In *International Journal for Parasitology*, 31:472–81. Pergamon. https://doi.org/10.1016/S0020-7519(01)00153-9.

Urbina, Julio A., and Roberto Docampo. 2003. “Specific Chemotherapy of Chagas Disease: Controversies and Advances.” *Trends in Parasitology*. Elsevier Ltd. https://doi.org/10.1016/j.pt.2003.09.001.

Verma, S., Dixit, R., & Pandey, K. C. (2016). *Cysteine proteases: modes of activation and future prospects as pharmacological targets*. Frontiers in pharmacology, 7, 107.

WHO. 2012. “Research Priorities for Chagas Disease, Human African Trypanosomiasis and Leishmaniasis.” https://apps.who.int/iris/bitstream/handle/10665/77472/WHO_TRS_975_eng.pdf.

World Health Organization. 2020. “WHO-Chagas.” 2020. https://www.who.int/es/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis).

Zhai, Xiang, and Thomas D. Meek. 2018. “Catalytic Mechanism of Cruzain from Trypanosoma Cruzi As Determined from Solvent Kinetic Isotope Effects of Steady-State and Pre-Steady-State Kinetics.” *Biochemistry* 57 (22): 3176–90. https://doi.org/10.1021/acs.biochem.7b01250.

Zingales, Bianca, Michael A. Miles, Carolina B. Moraes, Alejandro Luquetti, Felipe Guhl, Alejandro G. Schijman, and Isabela Ribeiro. 2014. “Drug Discovery for Chagas Disease Should Consider Trypanosoma Cruzi Strain Diversity.” *Memorias Do Instituto Oswaldo Cruz* 109 (6): 828–33. https://doi.org/10.1590/0074-0276140156.

CAPÍTULO II

“Metodología”

2.1 Descubrimiento de fármacos asistido por computadoras (DFAC)

El proceso que involucra el descubrimiento o diseño de fármacos *de novo* es un proceso costoso y complejo. Este proceso tiene su origen en remedios herbales que datan de milenios; y sólo desde el siglo pasado las drogas tuvieron un origen semisintético (Newman and Cragg 2007; M. Lourenco, M. Ferreira, and S. Branco 2012).

Bajo estas premisas, los compuestos seleccionados a menudo carecen de potencia y seguridad y, por lo tanto, deben optimizarse. Si bien históricamente este fue un proceso de prueba y error (Wikberg et al. 2011; Reardon 2013) pronto se desarrollaron estrategias racionales para mejorar la potencia de los posibles candidatos (Guha et al. 2011; Gao et al. 2013). Al igual que con cualquier procedimiento de manejo de datos, las computadoras se han convertido en una herramienta más destacada y ubicua en el descubrimiento de fármacos desde la década de 1980 (Kaul 1998). La unión entre la investigación computacional y la farmacéutica se suele designar como Diseño de Fármacos Asistido por Computadora (CADD siglas en inglés para *Computer Aided Drug Design*) (Song et al. 2009; Veselovsky et al. 2014).

El objetivo principal de CADD es acelerar y racionalizar el proceso de descubrimiento de fármacos al tiempo que se reducen los costos (Taft et al. 2008). En particular, el objetivo de la fase más temprana en el descubrimiento racional es la identificación de compuestos mediante diversas técnicas de cribado virtual.

2.1.1 Cribado virtual

El desarrollo de fármacos se inicia con la identificación de compuestos que se unen a un blanco terapéutico o que muestran actividad biológica en ensayos de tamizaje. Aquellas moléculas que muestran actividad biológica son llamadas *hits* (Song, Lim, and Tong 2009).

Considerando que el número de moléculas orgánicas que son sintéticamente factibles está por encima de 10^{60} moléculas (espacio químico), es evidente que su análisis sería muy complejo sin el uso de técnicas computacionales (Lucas et al. 2015).

El cribado virtual (CV) es un filtrado computacional (*in silico*) de moléculas para seleccionar candidatos (*hits* computacionales) para su evaluación experimental (Scior et al. 2012; Saldívar-González et al. 2017). De esta manera, el cribado virtual reduce significativamente el número de

ensayos biológicos que se harían si no hubiera una selección de compuestos. Sin embargo, es un proceso predictivo que debe integrarse con ensayos experimentales que validen las predicciones de los ensayos *in-silico* (Lionta et al. 2014).

Existen diversos filtros que se utilizan para llevar a cabo el cribado virtual los cuales pueden variar según la complejidad de la base de datos y la información experimental de la que se disponga (Figura 2.1). Por ejemplo, si se conoce la estructura tridimensional (3D) del receptor se sugiere un cribado basado en la estructura (CVBE, Cribado virtual basado en la estructura). Si solo se conocen los compuestos activos, pero no el receptor, entonces la búsqueda se hace basada en el ligando (CVBL, Cribado Virtual Basado en el Ligando). Si se conoce la estructura 3D del receptor y de los compuestos activos se pueden combinar ambos tipos de filtros para facilitar la búsqueda, por ejemplo, se puede realizar un primer tamizaje empleando técnicas de CVBL que son computacionalmente menos demandantes, seguido de un CVBE.

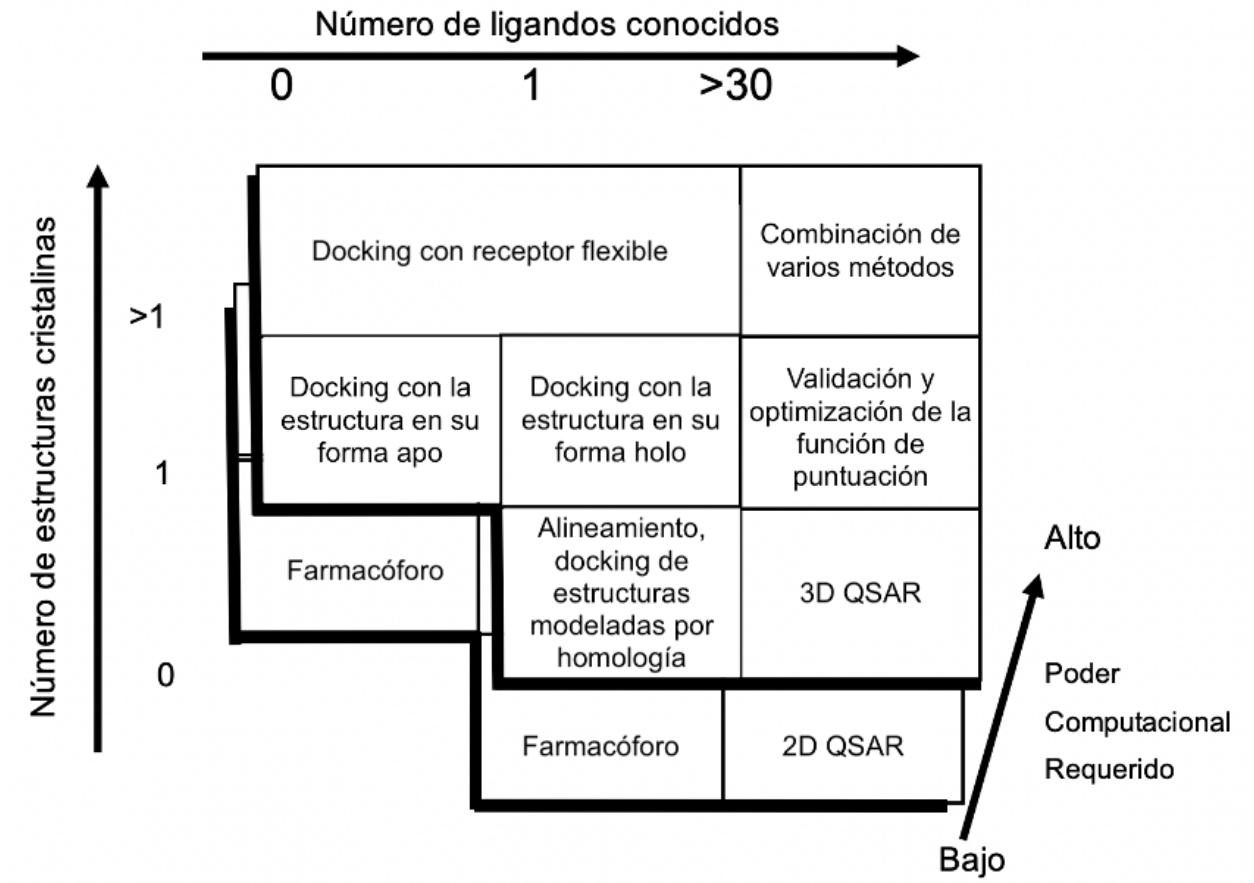


Figura 2.1 La elección de los métodos de cribado virtual está guiada por tres parámetros: número de moléculas pequeñas activas conocidas, número de estructuras cristalinas proteicas disponibles y recursos informáticos (adaptado de Seifert y Lang 2007)

2.1.1.1 Cribado Virtual Basado en el Ligando (CVBL)

El descubrimiento de fármacos basado en el ligando, o descubrimiento de fármacos indirectos, se basa en el conocimiento de moléculas que tienen cierta actividad sobre un blanco de interés, para la búsqueda de nuevos candidatos potencialmente activos (Hassan Baig et al. 2016). Estos métodos resultan de gran utilidad cuando no se dispone de la estructura del receptor, o bien, al ser menos costoso computacionalmente, constituyen un filtro previo para el Cribado Virtual Basado en la Estructura (Gil Redondo 2010).

Existen diversos métodos que se emplean para este tipo de enfoque, los cuales se agrupan en tres clases: 1) búsqueda por similitud, mediante el uso de huellas dactilares moleculares (*fingerprints*) 2D; 2) métodos farmacofóricos, donde se definen una serie de restricciones tridimensionales que deben cumplir los ligandos; y 3) métodos basados en aprendizaje automático (*machine learning*), donde se generan reglas de clasificación a partir de un conjunto conocido de ligandos activos e inactivos (Gil Redondo 2010; Hassan Baig et al. 2016) (Figura 2.2).

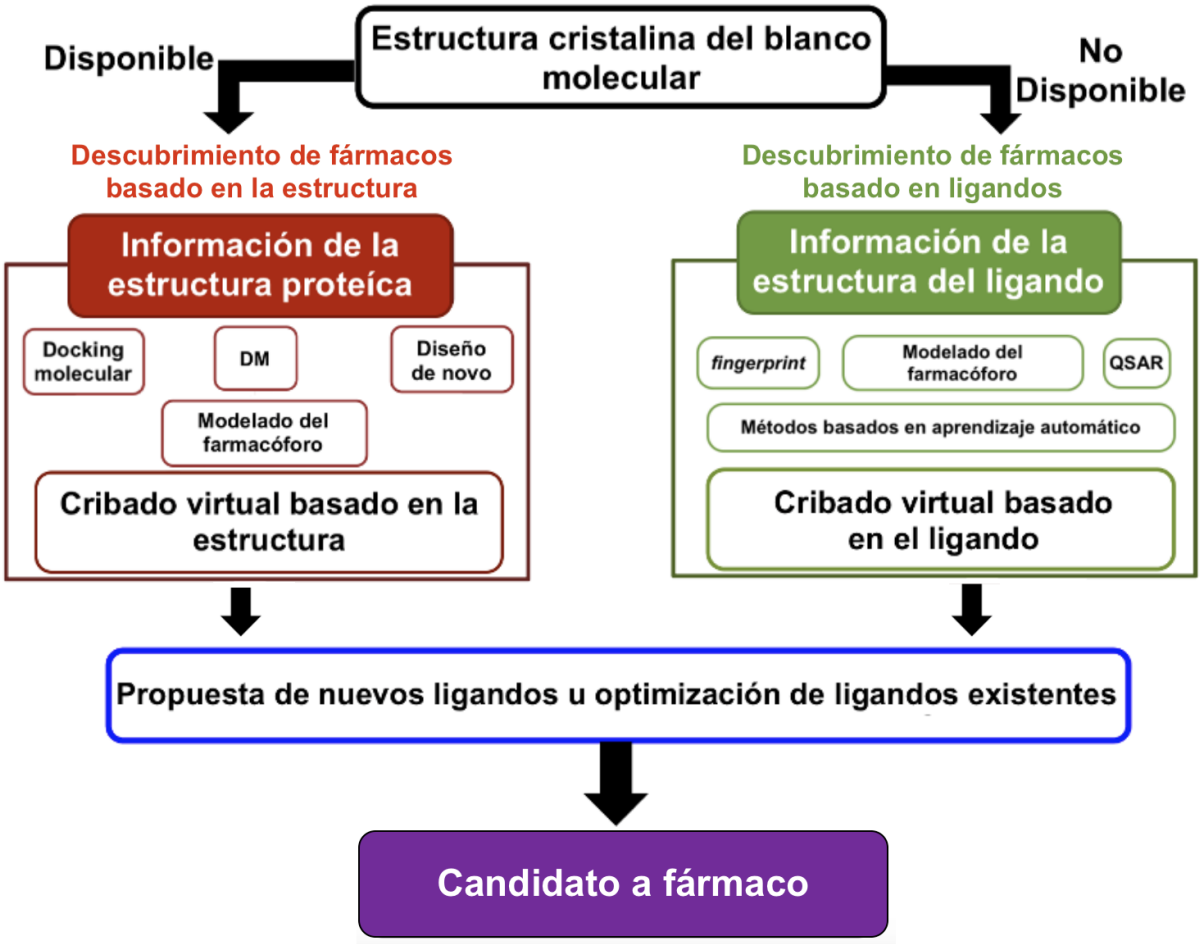


Figura 2.2 Diferentes enfoques utilizados para la búsqueda de nuevos candidatos a fármacos.

2.1.1.2 Cribado Virtual Basado en la Estructura (CVBE)

CVBE se apoya en el conocimiento de la estructura del blanco molecular y en la habilidad de los algoritmos de acoplamiento molecular (*docking*) para predecir modo y energías de unión de compuestos en las bases de datos, de tal manera que solo los compuestos mejor "ranqueados" son corroborados experimentalmente (Gil Redondo 2010).

El acoplamiento molecular es el método más utilizado en CVBE por su balance entre precisión y costo computacional. Sin embargo, puede recurrirse a otros métodos de cribado basado en la estructura dependiendo del balance que se busque entre estas dos variables contrapuestas. Por ejemplo, los modelos farmacofóricos basados en la estructura del blanco molecular permiten un tamizaje rápido de bases de datos de compuestos a expensas de una reducción en la precisión mientras que, en el otro extremo, las simulaciones de Dinámica Molecular (DM) logran una predicción mucho más precisa del modo y energías de unión pero a un coste computacional mucho más alto, lo cual limita su aplicación en campañas de cribado virtual aunque hoy en día,

con el surgimiento de la computación acelerada mediante GPUs ya empieza a hablarse de "Cribado Virtual Basado en Dinámica Molecular" (Ge et al. 2013).

A continuación se describen de manera breve las herramientas de *Docking* molecular y Dinámica molecular.

2.1.1.2.1 *Docking* Molecular

El acoplamiento molecular o *Docking* consiste en la determinación computacional de la afinidad de unión entre una estructura proteica y un ligando. Este método implica un muestreo de todas las poses posibles que puede adoptar el ligando en el bolsillo de unión de la proteína a fin de obtener la geometría de unión óptima, medida por las funciones de puntuación del algoritmo de *docking* (Gilson y Zhou 2007). El acoplamiento de moléculas pequeñas se puede realizar generalmente de tres maneras: (a) acoplamiento rígido, en el que el blanco molecular y el ligando se tratan de forma rígida; (b) acoplamiento flexible del ligando, en el que el blanco molecular se mantiene rígido; o (c) acoplamiento flexible, en el que tanto la proteína como el ligando se consideran flexibles (Mohan et al. 2005). Los protocolos de acoplamiento molecular también se pueden definir como una combinación de un algoritmo de búsqueda y una función de puntuación (Hassan Baig et al. 2016). Esta metodología se basa en que el algoritmo de búsqueda proporcione soporte y libertad a la coordinación proteína-ligando para permitir un muestreo preciso y exhaustivo de los posibles modos de unión.

2.1.1.2.2 Dinámica Molecular

La simulación de dinámica molecular (DM), es una de las principales herramientas para el estudio teórico de las moléculas biológicas (Hansson, Oostenbrink, y Van Gunsteren 2002). Este proceso calcula computacionalmente el comportamiento de un sistema molecular con respecto al tiempo. En MD, la mecánica newtoniana se aplica para calcular la trayectoria de un sistema (van Gunsteren y Berendsen 1990). La dinámica molecular ha proporcionado una gran cantidad de información detallada sobre las variaciones y los cambios conformacionales dentro de las proteínas y los ácidos nucleicos. Estos métodos computacionales ahora se usan comúnmente para investigar el comportamiento dinámico de las moléculas biológicas y sus complejos.

2.1.1.3 Cribado Virtual Retrospectivo (CVR)

En general los métodos empleados en el CBVE como *docking* se parametrizan utilizando un conjunto de datos estructurales de diferentes blancos moleculares que sean lo mas diverso posibles de tal manera que generalicen bien para cualquier sistema en estudio. Sin embargo, en muchos casos estos parámetros no son los adecuados para nuestro sistema en particular y por lo tanto es necesario ajustarlos empleando un conjunto de entrenamiento mas focalizado en el blanco molecular en cuestión. Esto es particularmente cierto para las funciones de puntuación (*scoring*) de los algoritmos de acoplamiento molecular (Pham y Jain 2008).

Ésta puesta a punto puede lograrse poniendo a prueba la habilidad de las técnicas de CVBE para recuperar a partir de una biblioteca de compuestos aquellos que se sabe son activos por el blanco molecular en estudio, lo que se conoce como Cribado Virtual Retrospectivo (CVR) a diferencia del Cribado Virtual Prospectivo (CVP) que consiste en la búsqueda de nuevos inhibidores.

Típicamente, las bibliotecas que se utilizan en el CVR se construyen sembrando unos pocos compuestos activos frente al blanco molecular en estudio en una base de datos de compuestos inactivos (señuelos) que son estructuralmente diferentes a los activos, pero con similares propiedades fisicoquímicas. El cribado virtual de esas bases de datos permite evaluar la performance del modelo en base a su habilidad para seleccionar los compuestos activos de entre los inactivos (Graves, Brenk, y Shoichet 2005). De esta manera también se puede ajustar el modelo eligiendo aquellos parámetros que muestran el mejor desempeño en los experimentos de cribado virtual de dichas bibliotecas (Figura 2.3)

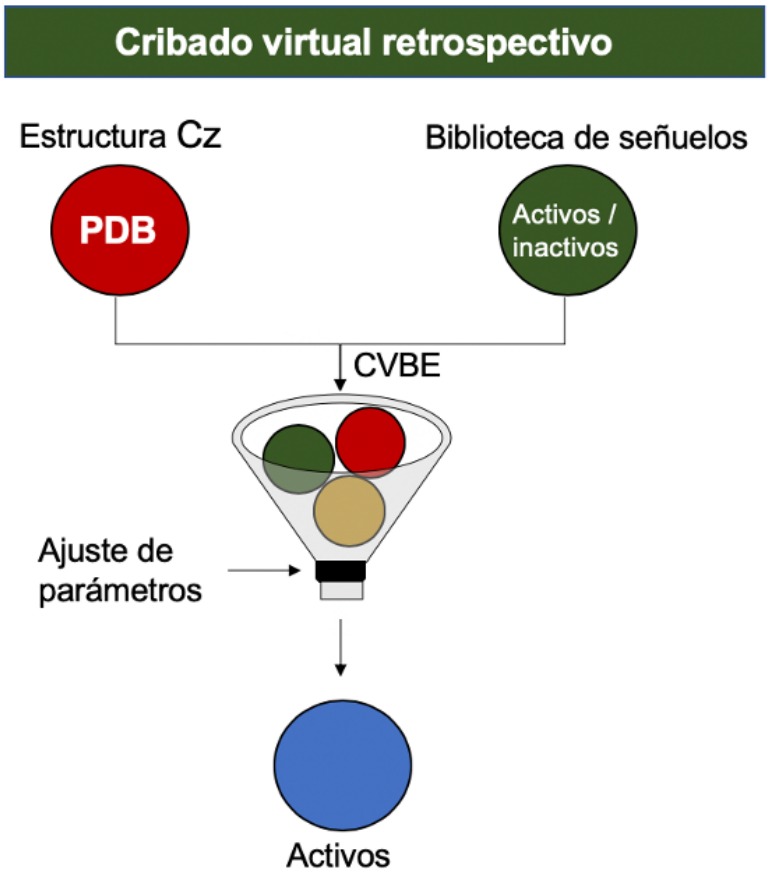


Figura 2.3 Pasos seguidos para la obtención de ligandos activos siguiendo el enfoque CVBE.

2.1.1.3.1 Biblioteca de Señuelos

La biblioteca de señuelos se puede construir de dos formas. En el enfoque más común mencionado anteriormente, algunos compuestos activos se siembran en una base de datos más grande de moléculas seleccionadas aleatoriamente (y supuestamente inactivas) con propiedades fisicoquímicas similares, pero con topologías 2D diferentes. Sin embargo, en proyectos de química medicinal reales, las moléculas a menudo son congenéricas, ya que provienen de síntesis paralela y, en consecuencia, son estructuralmente más similares entre sí (Li et al. 2020). Por lo tanto, sería más razonable seleccionar señuelos de acuerdo con su similitud con las moléculas activas (Triballeau, Bertrand, y Acher 2006). Dado que en este enfoque, los señuelos son más propensos a ser activos que las moléculas seleccionadas al azar (de hecho suelen tener cierta actividad marginal), uno debe estar seguro de que estos son

realmente inactivos (o marginalmente activos), para lo cual es necesario conocer sus actividades (valores de Ki, etc.).

En esta tesis se siguió este último enfoque para evaluar el rendimiento de los algoritmos de *docking*. Como los inhibidores de Cz portan un grupo reactivo que les permite unirse covalentemente a la enzima, la mayoría de las moléculas ensayadas para inhibición de Cz son bastante activas más allá de la afinidad propia del segmento tipo peptídico a través del cual ocurre el reconocimiento molecular. Por lo tanto, resulta difícil encontrar señuelos estructuralmente relacionados que resulten completamente inactivos. Así, nuestra biblioteca de señuelos (*decoys*) consistirá en compuestos muy activos (activos propiamente dicho) sembrados en un conjunto de compuestos marginalmente activos a moderadamente activos (que serán considerados como inactivos).

2.1.1.4 Puesta a punto de las técnicas de CVBE

Para mejorar el rendimiento de las técnicas de CVBE se siguió el protocolo descrito en la figura 2.4.

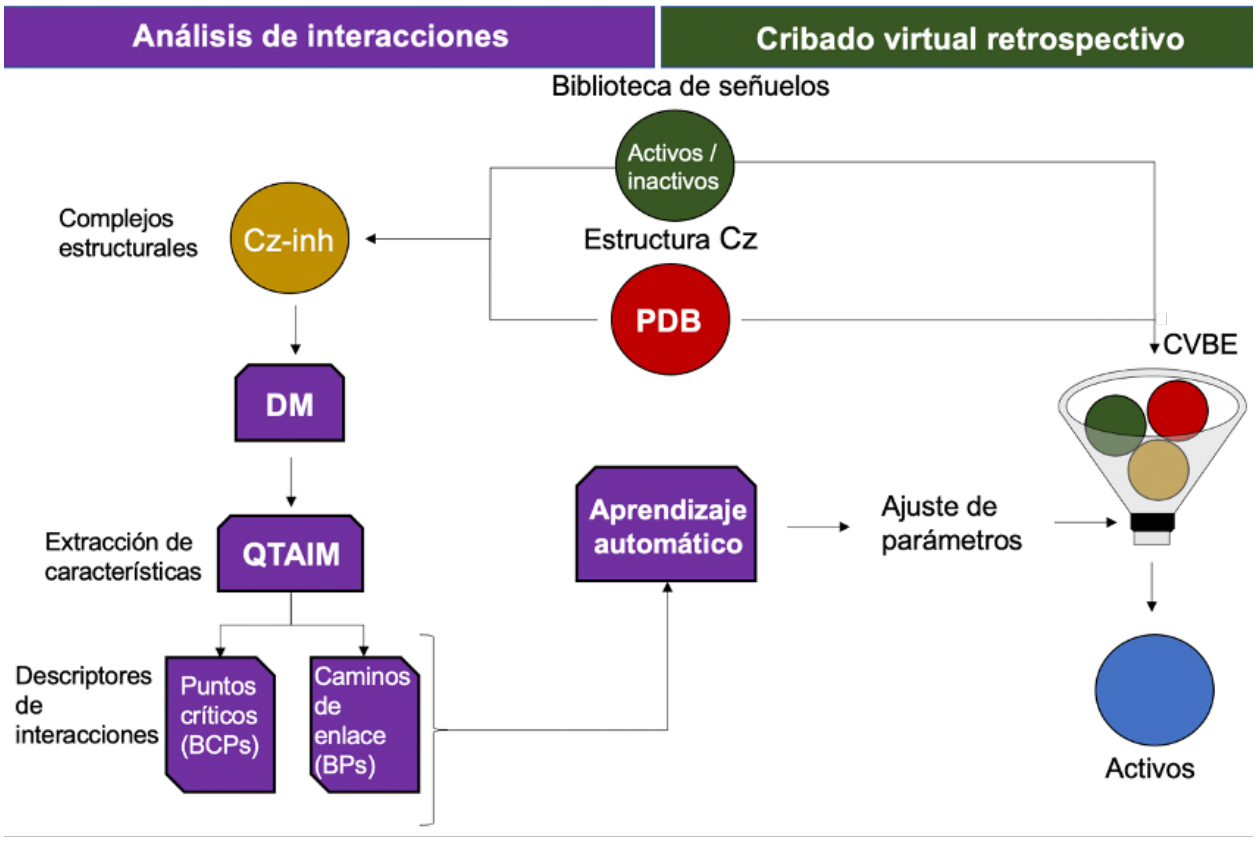


Figura 2.4 Protocolo seguido para la puesta a punto de las técnicas de CVBE.

La mayoría de los inhibidores de cisteína proteasas, consisten en un segmento peptídico para el reconocimiento molecular unido a un electrófilo o cabeza de guerra que se une covalentemente al residuo de cisteína del sitio activo. Como cada residuo en el péptido ocupa un subsitio específico en la enzima que está determinado por su posición en la secuencia peptídica (i.e. los residuos P1`, P1, P2 y P3 se ubican en sus correspondientes subsitios S1´, S1, S2 y S3) es posible predecir con bastante precisión el modo de unión mediante simulaciones de DM. De esta manera es factible aplicar un enfoque basado en la estructura aún para aquellos inhibidores que no hayan sido co-cristalizados con Cz.

Sobre los complejos estructurales Cz-Inhibidor (Cz-inh) obtenidos mediante simulaciones de DM, se realizó un análisis detallado de las interacciones intermoleculares. En particular nos interesaba: a) encontrar cuales son las interacciones mínimas necesarias para que ocurra la inhibición de Cz (farmacóforo) y b) discriminar entre interacciones que estabilizaban y desestabilizaban los complejos Cz-Inh.

Como descriptores de estas interacciones se utilizaron dos elementos que se derivan de la topología de la densidad electrónica en el contexto de la Teoría Cuántica de Átomos en Moléculas (QTAIM): el punto crítico de enlace (Bond Critical Points, BCPs) y los caminos de enlace (Bond Paths, BPs).

Debido a la intrincada red de caminos de enlace que se establecen en los complejos intermoleculares, extraer información de estos “grafos biomoleculares” solamente por inspección visual no es tarea trivial, ergo, resulta mucho más conveniente entonces recurrir a las herramientas de aprendizaje automático para su análisis.

El entrenamiento de modelos predictivos empleando el aprendizaje automático nos permitió identificar las interacciones que estabilizan / desestabilizan los complejos Cz-Inh como así también establecer cuales son las interacciones mínimas necesarias para la inhibición de Cz.

Esta información de las interacciones fue tenida en cuenta para mejorar las predicciones de los experimentos de CVR. A partir del conocimiento de las interacciones mínimas necesarias para la inhibición de Cz se construyeron modelos farmacofóricos que, cuando son acoplados con un algoritmo molecular permiten seleccionar a partir de una base de datos aquellos compuestos que

satisfacen las características estereolectrónicas dictadas por el farmacóforo. Por otro lado, la información sobre las interacciones que estabilizan/desestabilizan los complejos Cz-Inh se incorporan en forma de potenciales e interacción, en los mapas-grilla pre-calculados de los algoritmos de docking.

Los detalles metodológicos sobre cómo se implementan estas mejoras en los métodos de CVBE se describen con más detalle más adelante.

2.1.1.5 Cribado Virtual Prospectivo (CVP)

Una vez evaluados y ajustados los métodos de CVBE en los experimentos de CVR estos son "alimentados" con una base de datos de compuestos desconocidos (Cribado Virtual Prospectivo) con el objeto de seleccionar aquellos más promisorios como potenciales inhibidores de Cz, los cuales son luego corroborados experimentalmente. Dado el gran número de compuestos disponibles en bases de datos públicas, generalmente se realiza un CVBL previo al CVBE con el objeto de reducir el número de compuestos en base a medidas rápidas de similitud con inhibidores conocidos del blanco molecular (Cz) (Figura 2.5).

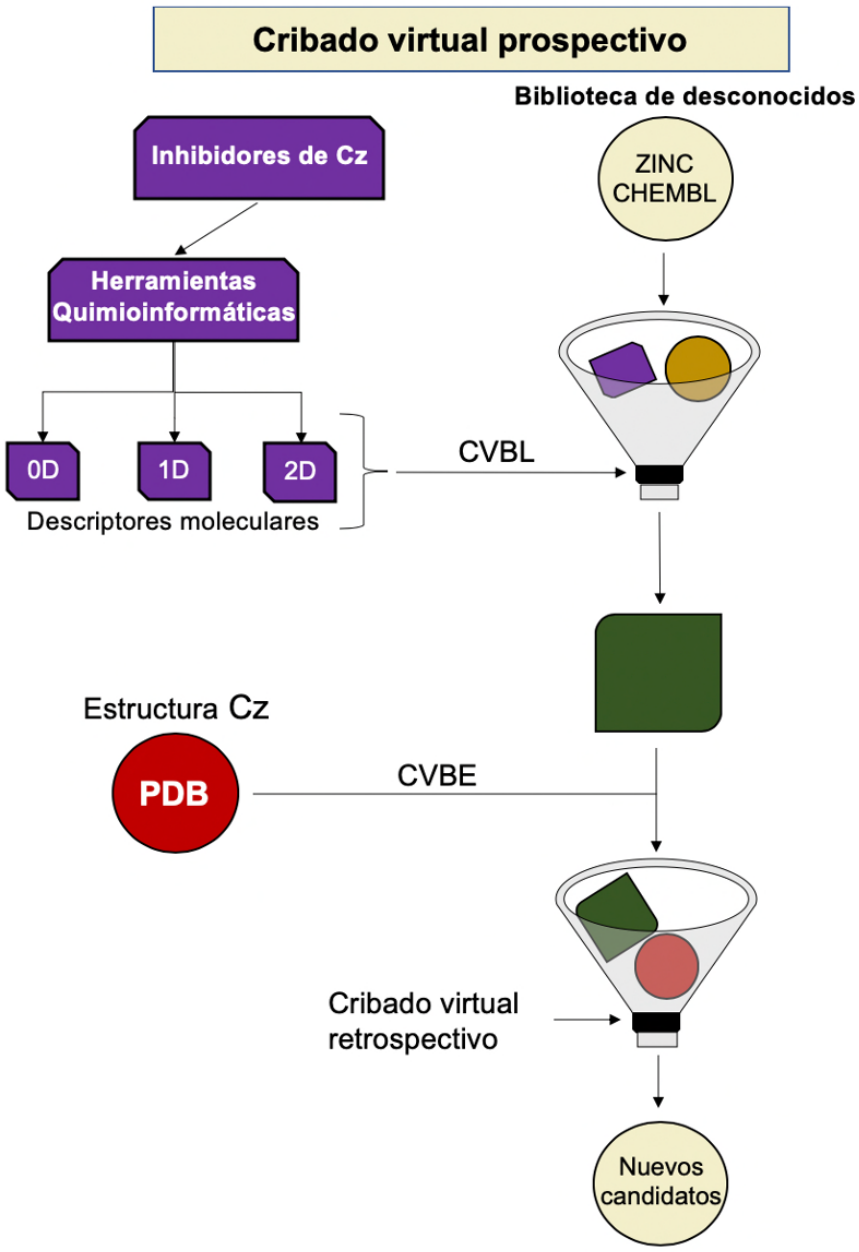


Figura 2.5 Pasos seguidos para realizar el CVP.

A continuación, se revisa brevemente la Teoría de Átomos en Moléculas (QTAIM) y sus aplicaciones para el análisis de interacciones en complejos biomoleculares.

2.2 Teoría Cuántica de Átomos en Moléculas (QTAIM)

La Teoría Cuántica de Átomos en Moléculas (QTAIM, por sus siglas en inglés) es una teoría interpretativa de la cual los principales elementos de la estructura molecular – átomos y enlaces

– son una expresión natural de la distribución de la densidad de carga electrónica $\rho(r)$ de un sistema.

La terminología de QTAIM fue revisada extensivamente en la literatura estándar (Bader 1990). A continuación, revisamos brevemente algunos conceptos básicos que se necesitan para la discusión de los resultados.

2.2.1 Conceptos básicos

La distribución de $\rho(r)$ de una molécula es una distribución de probabilidad que describe la forma promedio en que la carga electrónica se distribuye en el espacio real en el campo atractivo por los núcleos atómicos.

En el contexto de QTAIM, la estructura molecular se pone de manifiesto a través del estudio del campo del vector gradiente asociado a la densidad electrónica, $\nabla\rho(r)$ que define la topología de la densidad electrónica.

Los puntos estacionarios de la distribución donde $\nabla\rho(r) = 0$ (puntos críticos) y los caminos de gradiente que se originan y terminan en estos puntos definen los átomos y enlaces de la estructura molecular.

Dos átomos interactuantes comparten tres elementos topológicos que están relacionados entre sí: un punto, una línea y una superficie. El primer elemento es el punto crítico de enlace (BCP, bond critical point) que se encuentra entre cualquier par de núcleos interactuantes. En cada BCP se originan dos trayectorias únicas de $\nabla\rho(r)$ que terminan en cada uno de los núcleos vecinos. Estas trayectorias definen una línea a lo largo de la cual $\rho(r)$ es un máximo con respecto a cualquier línea vecina. Esta línea, que constituye el segundo elemento, es el camino de enlace (BP, bond path). Finalmente, el conjunto de trayectorias que terminan en un BCP, definen la superficie interatómica (IAS) que separa las cuencas atómicas de los átomos vecinos (Figura 2.6).

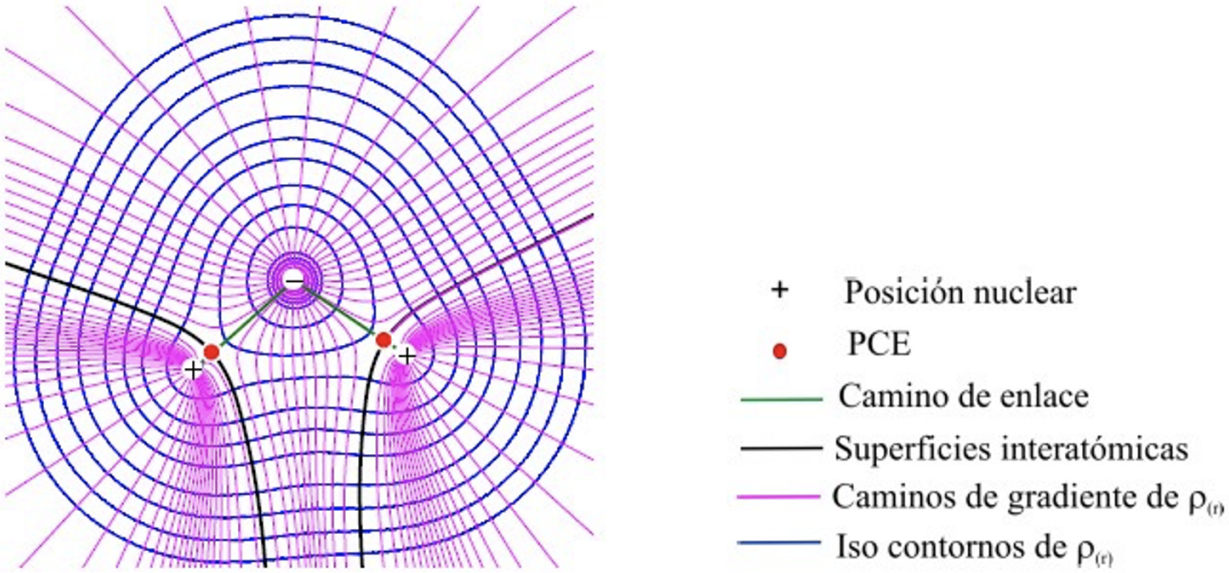


Figura 2.6 Mapa del campo vector del gradiente $\nabla\rho(r)$ para el plano que contiene los núcleos de oxígeno e hidrógeno de la molécula de agua, superpuesto con un mapa de contorno de $\rho(r)$.

2.2.2 QTAIM en complejos biomoleculares

Calcular la densidad electrónica de un sistema biomolecular completo, incluyendo su entorno, resulta prácticamente imposible aún hoy en día a pesar de los avances que se registran continuamente en cuanto a la capacidad de cálculo. El enfoque actual para tratar estos sistemas de gran tamaño empleando métodos mecano-cuánticos consiste en modelarlos “en capas” o empleando un enfoque híbrido QM/MM (figura 2.7) en el cual el ligando y los residuos del bolsillo de unión son tratados mecano-cuánticamente (QM) y los residuos circundantes del blanco molecular son tratados empleando un campo de fuerza de mecánica molecular (MM) o bien son descartados (MM=vacío).

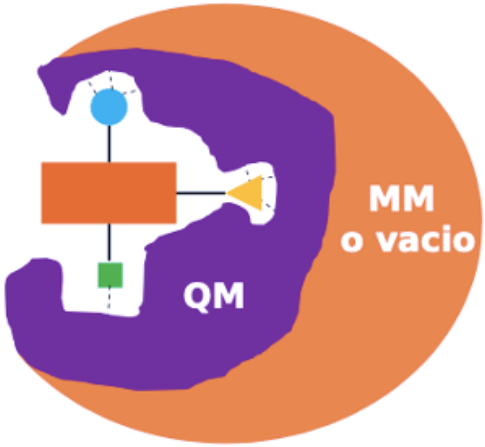


Figura 2.7 Enfoque híbrido QM/MM para el estudio de complejos biomoleculares. Coloración del sitio activo en violeta y resto de proteína anaranjado.

2.2.3 QTAIM como descriptor de las interacciones intermoleculares

La densidad de carga electrónica en el punto crítico (ρ_b) de un enlace de hidrógeno aumenta de manera aproximadamente lineal al aumentar la energía de estabilización (ΔE) del complejo yendo de enlaces de hidrógeno débiles a moderados y fuertes como se muestra en la Figura 2.8. La relación aproximadamente lineal entre ρ_b y ΔE puede hacerse extensible a otros tipos de interacciones de capa cerrada como por ejemplo los enlaces de halógeno.

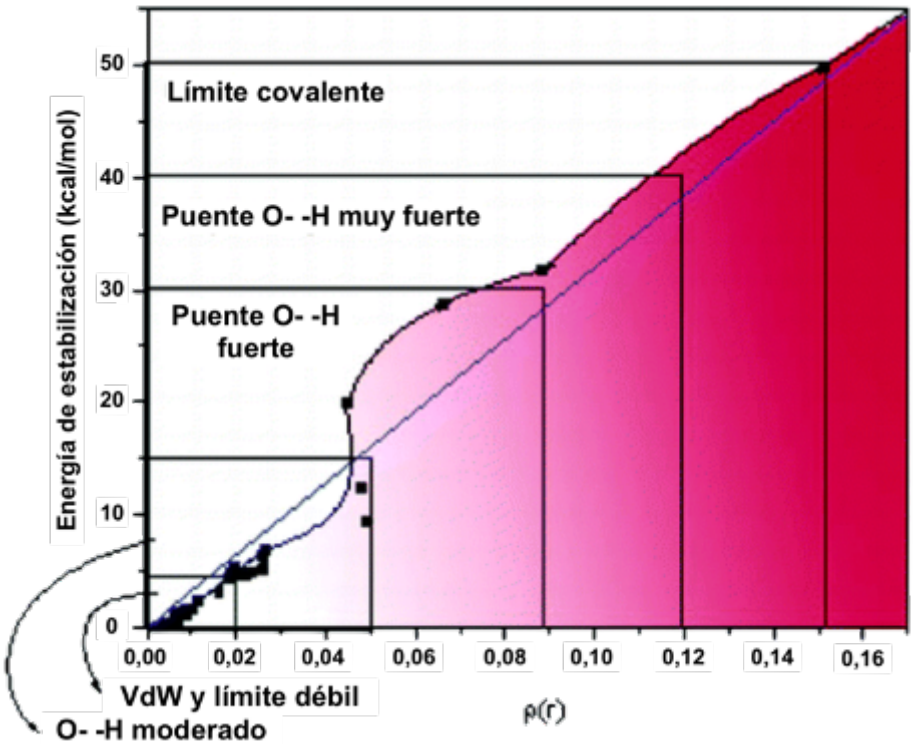


Figura 2.8 Relación aproximadamente lineal entre la densidad electrónica en el punto crítico del enlace de hidrógeno y la energía de estabilización en complejos débil, moderada y fuertemente enlazados (adaptado de Parthasarathi, Subramanian, y Sathyamurthy 2006).

Por otra parte, en el caso de complejos biomoleculares, que involucran más de una interacción intermolecular se ha demostrado que la suma de los valores de densidad electrónica en los puntos críticos de las interacciones intermoleculares $(\sum\rho_b)$ muestra una relación más o menos lineal con la energía de unión del complejo. Esta relación en general mejora cuando también se tienen en cuenta las interacciones intramoleculares tanto del ligando como del blanco molecular.

Pero el principal atractivo de la teoría QTAIM radica en que permite particionar la interacción total en contribuciones por átomo o grupos de átomos lo cual la hace particularmente útil en Química Medicinal para el análisis, diseño y optimización de compuestos líderes. Esto se muestra esquemáticamente en la figura 2.9.

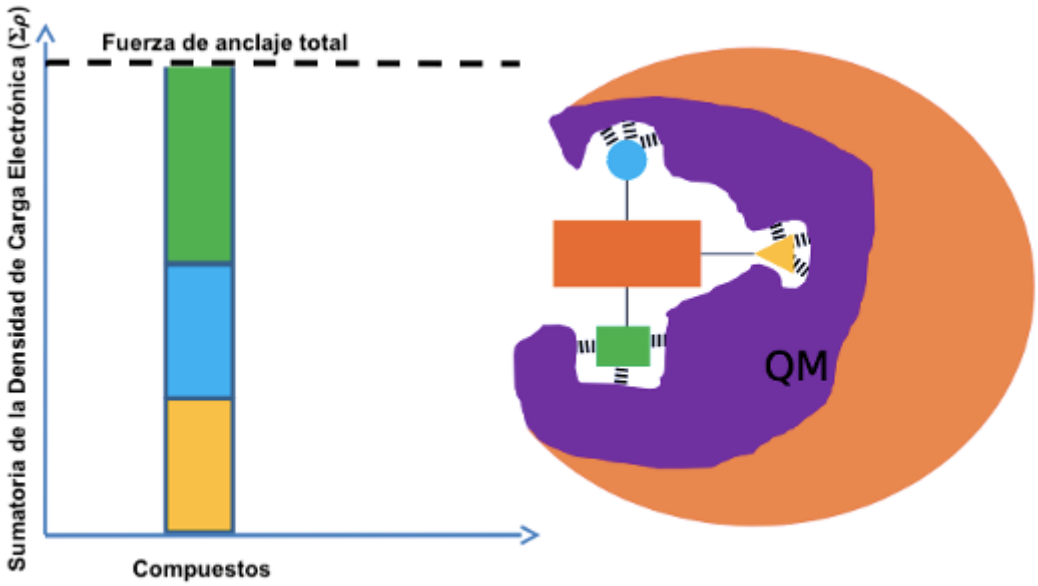


Figura 2.9 Análisis de las contribuciones de diferentes grupos de átomos del ligando al anclaje total de la molécula en el sitio activo.

2.2.4 Aplicabilidad de la teoría QTAIM en el DFAC

En las campañas de cribado virtual para la búsqueda de nuevos *hits* se evalúan el modo y energías de unión de cientos a miles o millones de compuestos con lo cual se requiere de

métodos rápidos como el docking molecular. Evidentemente la aplicación de QTAIM en esta primera instancia del DFAC no es factible. Sin embargo, en etapas posteriores cuando el número de candidatos se ha reducido significativamente, puede emplearse QTAIM para optimizar la estructura de los *hits* iniciales encontrados. Esto se esquematiza en la figura 2.10. La posibilidad de analizar la fuerza de anclaje de cada grupo funcional separadamente permite identificar los sitios de anclaje débiles, lo cual hace de QTAIM una herramienta muy útil en la etapa optimización de *hits*.

QTAIM también puede resultar útil en estudios retrospectivos en los cuales se trata de relacionar la afinidad de los hits con las interacciones específicas que se establecen con el blanco molecular. A su vez la información de las interacciones puede utilizarse para ajustar los parámetros de los algoritmos de docking para mejorar sus rendimientos en nuevas rondas de cribado virtual.

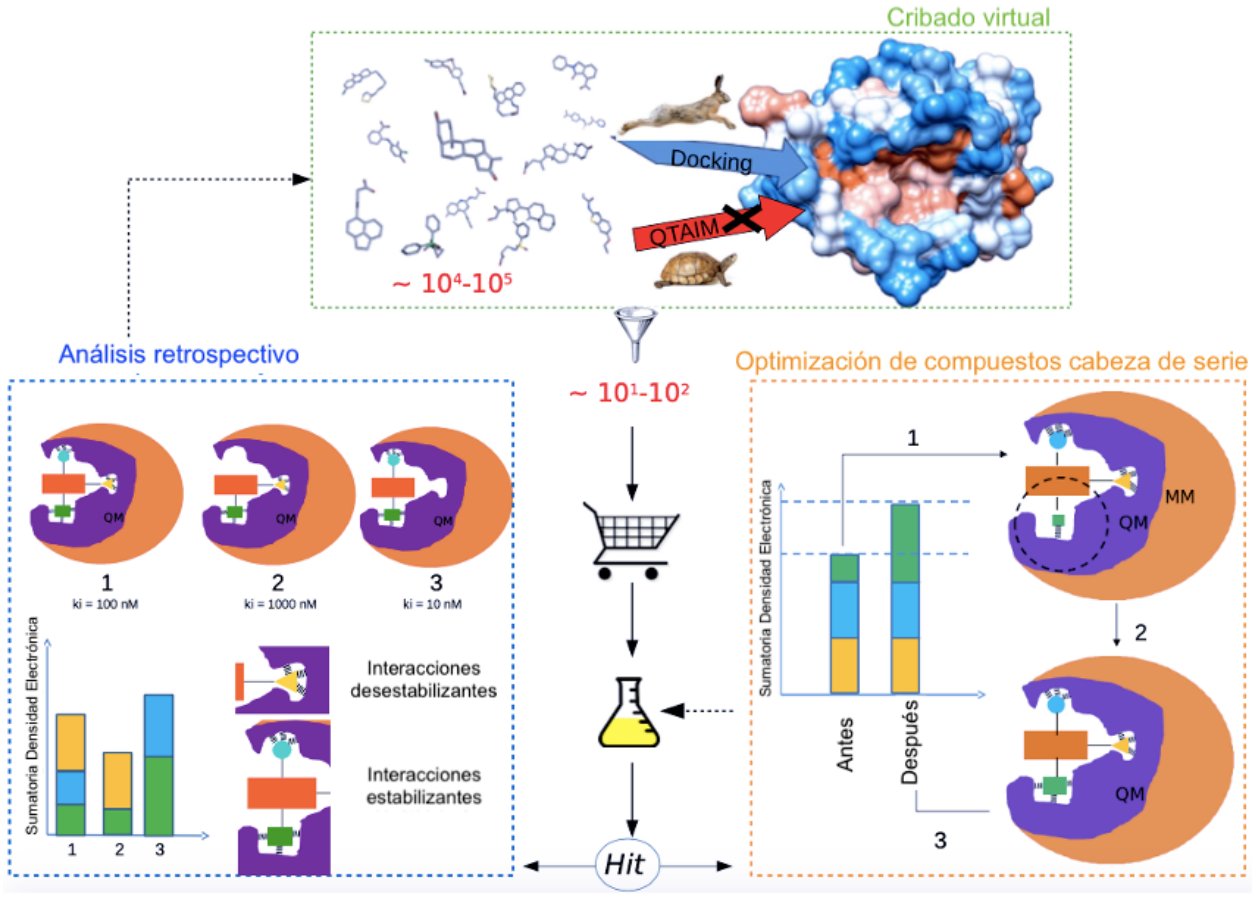


Figura 2.10 Estudios llevados a cabo utilizando QTAIM en el contexto DFAC.

Debido a la intrincada red de interacciones que se establecen en complejos biomoleculares los estudios de relación entre la actividad de los hits y las interacciones que estabilizan o desestabilizan dichos complejos resulta difícil de realizar por simple inspección visual. Por ello en esta tesis se propone aplicar herramientas de aprendizaje automático para extraer las interacciones más relevantes que permitan explicar las actividades in vitro.

2.3 Aprendizaje Automático

El aprendizaje automático (ML, del inglés *Machine Learning*) es una ciencia dedicada a hacer que las computadoras aprendan, sin ser directamente programadas para realizar una tarea (Samuel 1959). El aprendizaje automático se centra en el desarrollo de programas informáticos que pueden cambiar cuando se exponen a nuevos datos.

Los algoritmos del aprendizaje automático se clasifican a menudo como supervisados o no supervisados (Bhavsar et al. 2017). Los algoritmos supervisados pueden extraer inferencias del conjunto de datos.

En el aprendizaje supervisado, se nos proporciona un conjunto de datos y ya sabemos como debería ser nuestro resultado correcto, teniendo la idea de que existe una relación entre la entrada y la salida.

Por el contrario, el aprendizaje no supervisado nos permite abordar problemas con poca o ninguna idea de cómo deberían ser nuestros resultados. Podemos derivar la estructura de los datos agrupándolos en función de las relaciones entre las variables.

A continuación se explican los conceptos básicos utilizados dentro del aprendizaje supervisado como los modelos de clasificación lineal y Maquinas Vectoriales de Soporte tomados de Gereon 2018 y Müller 2017.

2.3.1 Aprendizaje supervisado

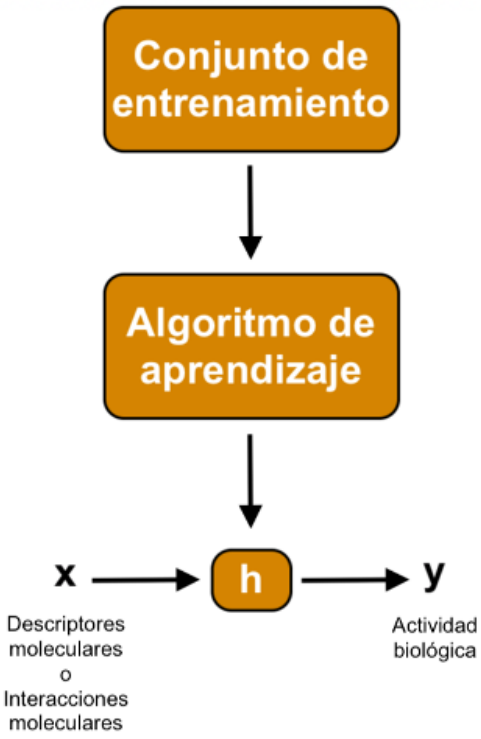
Supongamos que tenemos un conjunto de m moléculas para las cuales se conocen sus actividades biológicas sobre algún blanco terapéutico y cierta característica que de cuenta de su

estructura química (descriptor molecular 0D-3D) o bien de su modo de interacción con el blanco molecular:

Característica	Actividad
$x^{(1)}$	$y^{(1)}$
$x^{(2)}$	$y^{(2)}$
$x^{(3)}$	$y^{(3)}$
$x^{(4)}$	$y^{(4)}$
:	:
$x^{(m)}$	$y^{(m)}$

Las x(i) denotan las variables de “entrada” (características), e y(i) la “salida” o variable objetivo que estamos tratando de predecir (actividad). Un par (x(i), y(i)) se llama un ejemplo de entrenamiento, y el conjunto de *m* ejemplos que usaremos para aprender {(x(i), y(i)); i = 1, …,m} se llama un conjunto de entrenamiento.

Para describir el problema de aprendizaje supervisado un poco más formalmente, nuestro objetivo es, dado un conjunto de entrenamiento, aprender una función $h: X \rightarrow Y$ de manera que $h(x)$ sea un “buen” predictor del valor correspondiente de y (Esquema 1). Por razones históricas, esta función *h* se denomina hipótesis.



Esquema 1 Pasos seguidos para realizar predicciones utilizando algoritmos de aprendizaje supervisado.

Cuando la variable objetivo que estamos tratando de predecir es continua (ej. IC50, *Ki*, % inhibición, etc.) llamamos al problema de aprendizaje un problema de regresión. Cuando y puede asumir solo una pequeña cantidad de valores discretos (ej. Compuestos etiquetados como activos e inactivos), lo llamamos un problema de clasificación.

Nosotros aplicamos estas ideas para entrenar un modelo que nos permita predecir si un compuesto es activo o no como inhibidor de Cz. Este es claramente un problema de clasificación. Veamos entonces en más detalle en qué consiste un modelo de clasificación en el contexto del aprendizaje supervisado.

2.3.1.1 Modelo de clasificación lineal

Por ahora, nos centraremos en el problema de clasificación binaria en el que, *Y* puede tomar solo dos valores, 0 y 1. 0 también se denomina clase negativa (ej. Compuestos inactivos), y 1 es la clase positiva (ej. Compuestos activos).

Podríamos abordar el problema de clasificación ignorando el hecho de que *Y* es una variable que toma valores discretos y usar una regresión lineal como hipótesis *h* para tratar de predecir *Y* dado *X:*

$$y \approx h = \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 \quad (1)$$

Donde los θ_i son los parámetros (o pesos) de la función lineal y los x_i son las características o variables de entrada.

Sin embargo, intuitivamente no tiene sentido que *h* tome valores mayores que 1 o menos que 0 cuando sabemos que $y \in \{0, 1\}$. Para corregir esto, vamos a cambiar la forma de nuestra hipótesis h para satisfacer $0 \leq h \leq 1$. Esto se logra introduciendo la función lineal (1) en una "función logística" o "función sigmoidea" $g(z)$:

$$h = g(z) = \frac{1}{1+e^{-z}} \quad (2)$$

donde z es la función lineal

$$z = \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 \quad (3)$$

La siguiente figura muestra como se ve la función sigmoidea:

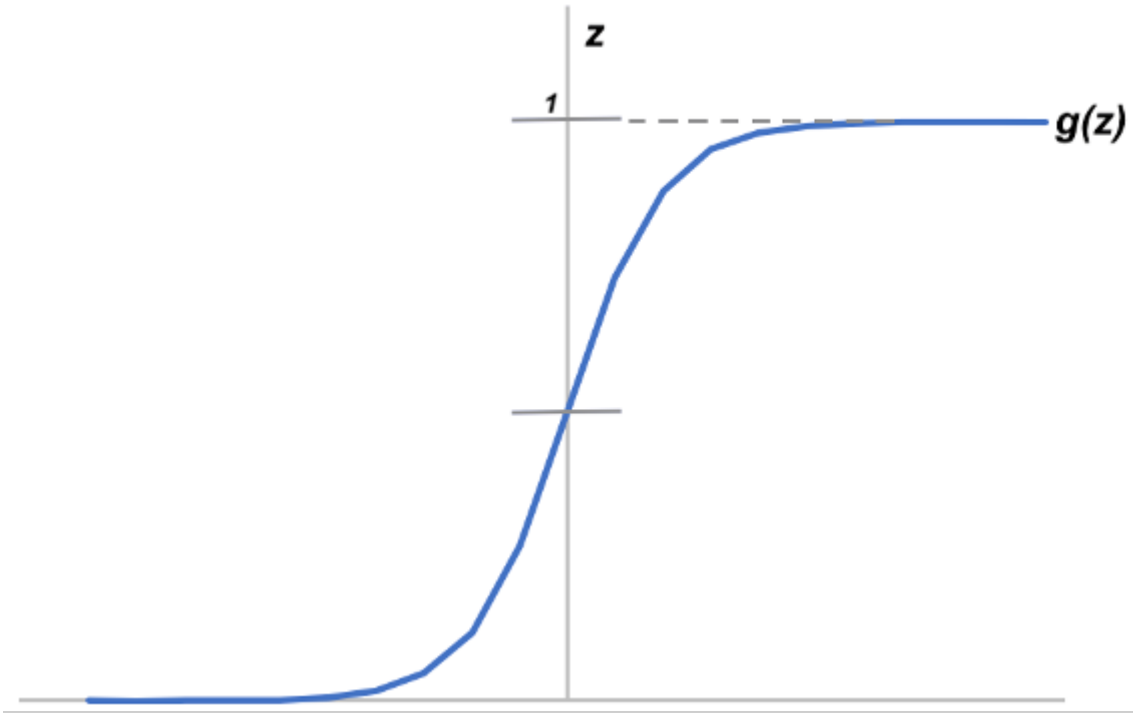


Figura 2.11 Curva sigmoidea.

La función $g(z)$ que se muestra aquí, asigna cualquier número real al intervalo. (0, 1), lo que la hace útil para transformar una función que toma valores arbitrarios en otra más adecuada para la clasificación. Se ve también en la figura que la función logística $g(z)$ toma un valor mayor o igual a 0.5 cuando z es mayor o igual a cero.

El valor de la hipótesis h, representado por $g(z)$ se interpreta como la probabilidad de que la variable de salida y sea 1. Por ejemplo, $h = g(z) = 0.7$ da una probabilidad del 70% de que y sea 1. Así, podemos traducir la salida de la función de hipótesis de la siguiente manera:

$h \geq 0.5 \rightarrow y = 1$ (h predice que el ejemplo pertenece a la clase positiva)
$h < 0.5 \rightarrow y = 0$ (h predice que el ejemplo pertenece a la clase negativa)

Para entender mejor cómo el modelo de clasificación hace estas predicciones, se introduce la noción de "límite de decisión". El límite de decisión es la línea que separa el área donde $y = 0$ de donde $y = 1$. La siguiente figura muestra en magenta el límite de decisión para un conjunto de entrenamiento formado por clases positivas (cruces) y negativas (círculos).

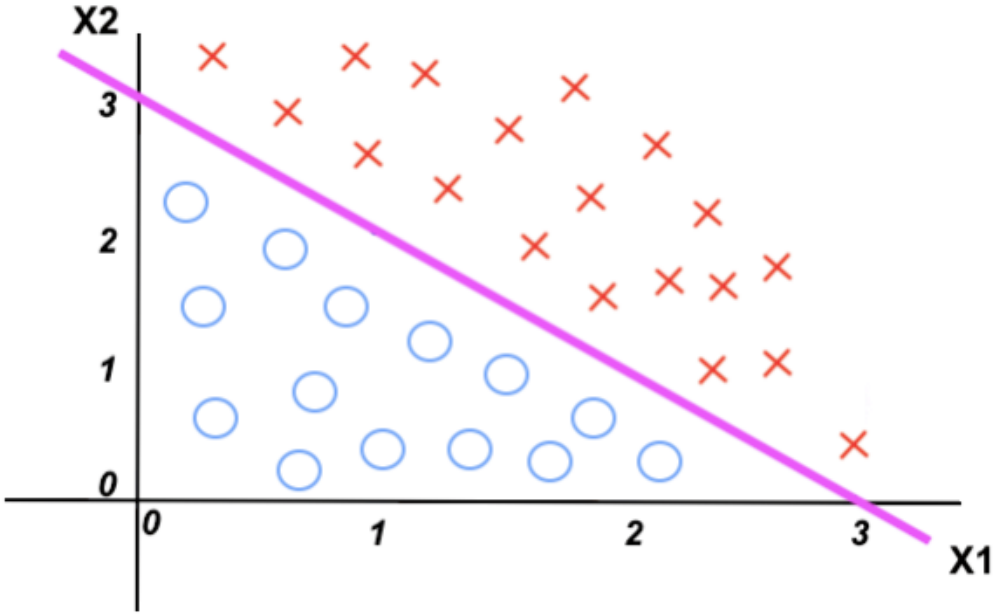


Figura 2.12 Ejemplo de límite de decisión. Hiperplano separador de ambas clases.

El límite de la decisión está definido por la hipótesis h. Supongamos que luego de ajustar el modelo de clasificación con nuestro conjunto de entrenamiento llegamos a la hipótesis que tiene la forma $h = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$, con parámetros $\theta_0 = -3, \theta_1 = 1\ y\ \theta_2 = 1$.

Dado que $h = g(z)$ predice la clase positiva cuando $z \geq 0$ (ver función sigmoidea), el límite de decisión está dado por los valores de x_1 y x_2 que satisfacen la siguiente ecuación:

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0 \tag{4}$$

Reemplazando los valores de los parámetros θ_i y reordenando queda que $x_1 + x_2 = 3$. Es decir que todos aquellos ejemplos del conjunto de entrenamiento con $x_1 + x_2 \geq 3$ caen por encima del límite de decisión y por tanto serán clasificados como clase positiva y aquellos ejemplos con $x_1 + x_2 \quad < 3$ se encuentran por debajo de la línea magenta en la figura y por lo tanto pertenecen a la clase negativa.

2.3.1.2 Máquinas Vectoriales de Soporte (SVM)

Consideremos diferentes límites de decisión para separar un conjunto de datos linealmente separables en clases positivas y negativas:

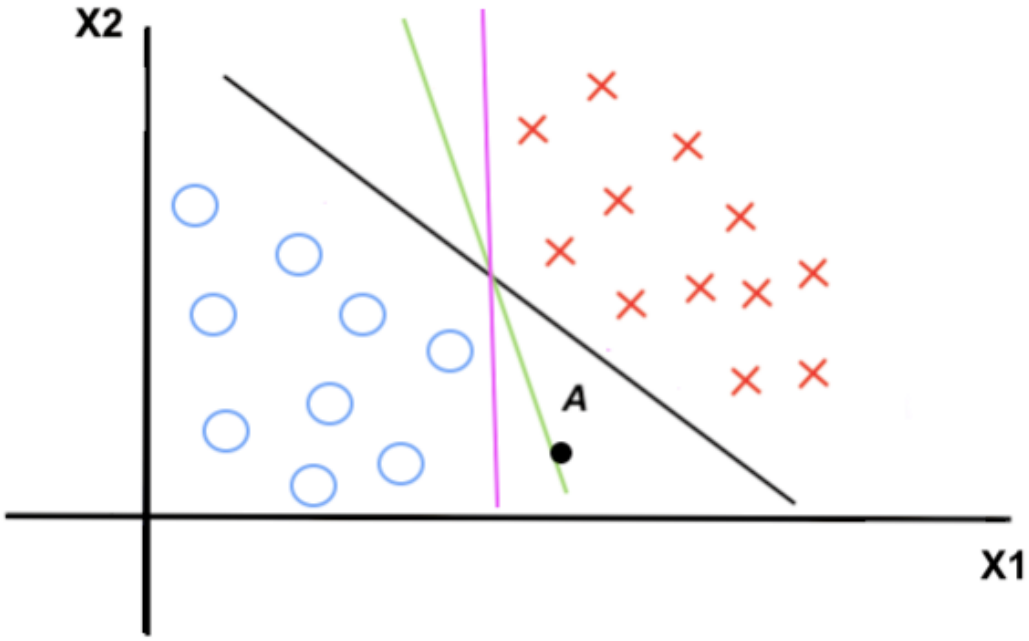


Figura 2.13 Ejemplo de clasificación.

Intuitivamente, las líneas de decisión en verde y magenta no parecen ser la mejor opción para separar las clases positivas (círculos) de las negativas (cruces). Un ejemplo de la clase positiva

con coordenadas en el punto A sería mal clasificado como clase negativa según límites de decisión. Por el contrario, la línea en negro parece ser un clasificador mas robusto porque presenta una mayor distancia (o margen) con respecto a los ejemplos de entrenamiento mas próximos de ambas clases.

Esta es la idea detrás de SVM (Cortes 1995), es decir, elegir el límite de decisión de tal manera de maximizar se distancia (o margen) a cualquiera de los ejemplos más próximos a cada lado de la línea.

Mientras que un modelo de regresión logística podría haber encontrado cualquiera de los límites de decisión dibujados en la figura anterior, SVM elige el límite de decisión de mayor margen.

La siguiente figura muestra los márgenes de un clasificador SVM representados con líneas azules a ambos lados del límite de decisión.

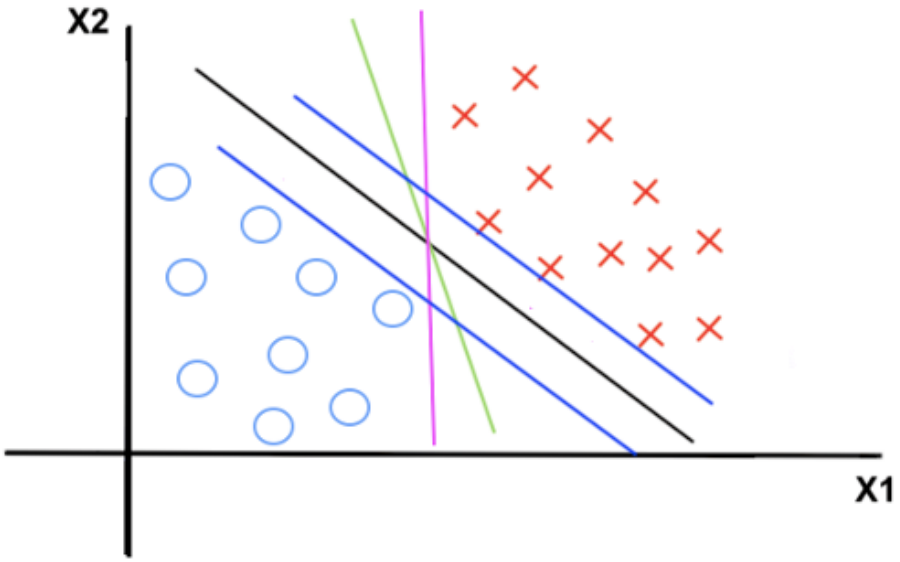


Figura 2.14 Márgenes máximos de clasificación (lineas azules) y límite de decisión (linea negra).

Para entender cómo el algoritmo de optimización SVM encuentra el límite de decisión como máximo margen consideremos la siguiente función de hipótesis de la ecuación (1), por simplicidad supongamos que θ_0 el término de intersección es igual a cero

$$h = \theta_1 \cdot x_1 + \theta_2 \cdot x_2 \tag{5}$$

Vectorialmente esta ecuación se puede representar como un producto escalar de dos vectores

$$\hat{\theta}^T \cdot \hat{x} = [\theta_1 \theta_2] \cdot [x_1\ x_2\ ] = \theta_1 \cdot x_1 + \theta_2 \cdot x_2 \tag{6}$$

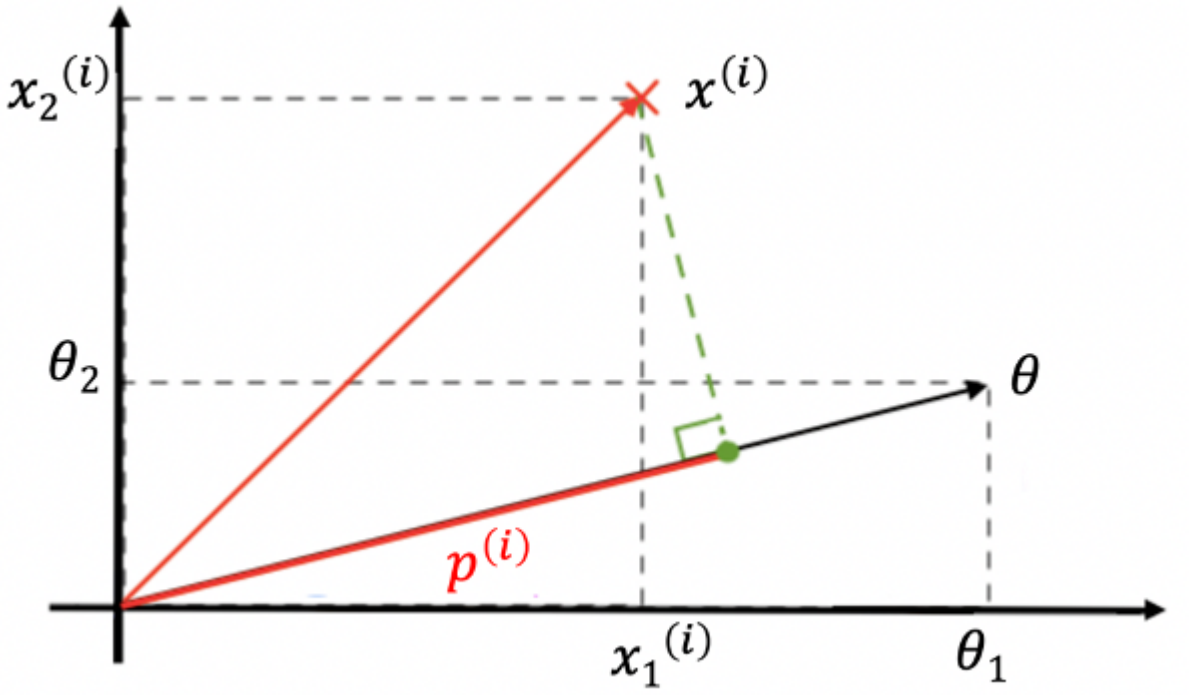


Figura 2.15 Representación de dos vectores.

Geométricamente el producto escalar de dos vectores se puede interpretar como la proyección de un vector sobre el otro, como se muestra en la siguiente figura para el caso concreto de un ejemplo de entretenimiento $x^{(i)}$.

De acuerdo con la interpretación geométrica, el producto escalar se puede escribir como

$$\hat{\theta}^T \cdot \hat{x}^{(i)} =\cdot [x_1\ x_2\ ] = p^{(i)} \cdot \|\hat{\theta}\| \tag{7}$$

Donde $p^{(i)}$ es la proyección del vector $x^{(i)}$ sobre el vector θ multiplicado por el módulo de este último. Es importante notar que $p^{(i)}$ tiene signo, si el ángulo que forman ambos vectores es mayor a 90º $p^{(i)}$ será negativo.

Veamos ahora como la representación geométrica de la hipótesis se relaciona con la noción de máximo margen. Supongamos que luego de entrenar un modelo de clasificación encontramos el límite de decisión indicando con una línea verde

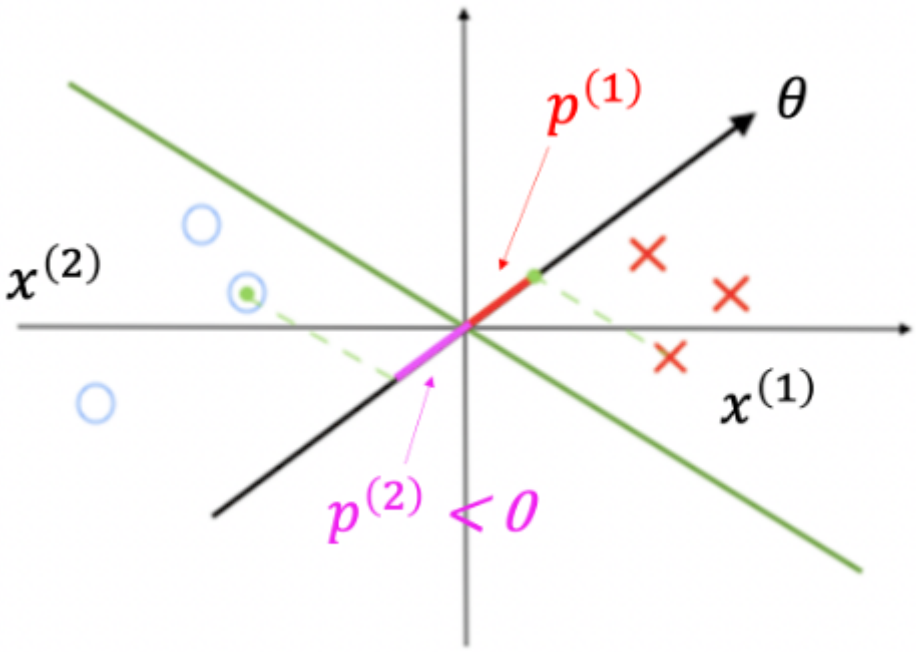


Figura 2.16 Representación del vector θ, junto con vectores soportes suboptimos.

Es posible demostrar que el vector de parámetros θ es siempre perpendicular a la línea de decisión tal como se indica en la figura. La figura también muestra las proyecciones sobre θ de los vectores $x^{(1)}$ y $x^{(2)}$ correspondientes a los ejemplos de entrenamiento más próximos al límite de decisión en ambas clases. La magnitud de $p^{(1)}$ y $p^{(2)}$ indican que tan amplio es el margen entre el límite de decisión y los ejemplos de entrenamiento más próximos mientras que su signo indica si la clase proyectada es positiva o negativa.

El hecho de que el límite de decisión pase por el origen del sistema de coordenadas es consecuencia de que el término de intersección θ_0 es nulo.

Ahora tracemos otro límite de decisión distinto y proyectemos los ejemplos de entrenamiento marginales sobre el vector de parámetros θ

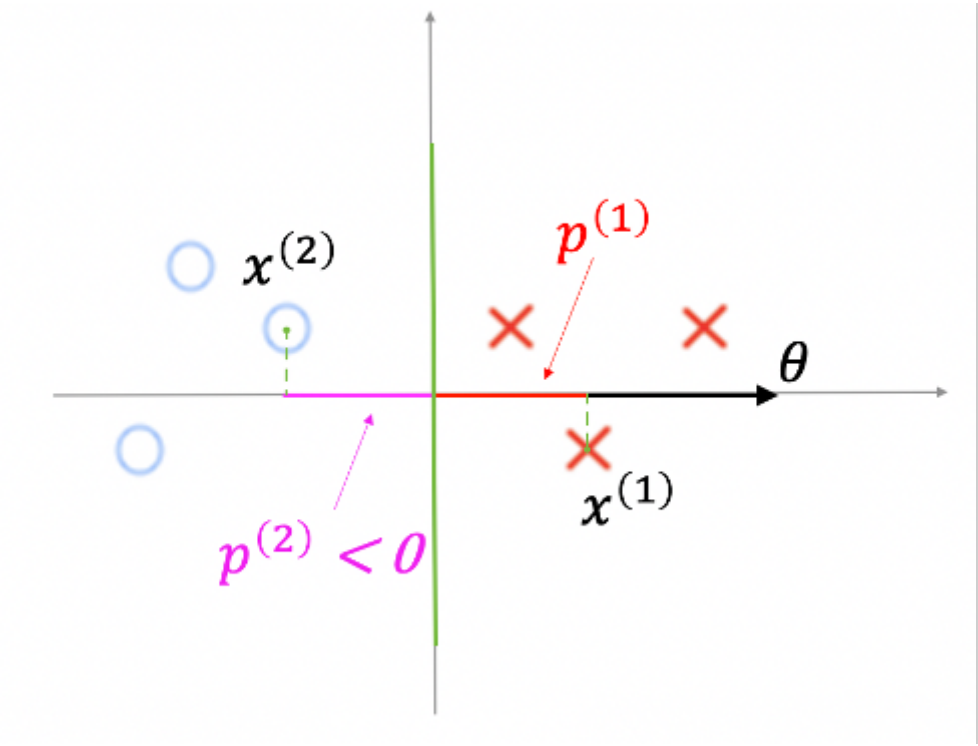


Figura 2.17 Representación del vector θ junto con vectores de soporte.

Claramente, este segundo límite de decisión es un clasificador más robusto que el primero lo cual se evidencia por el mayor margen indicado por la magnitud de los vectores $p^{(1)}$ y $p^{(2)}$.

Los fundamentos expuestos resultan útiles para luego poder interpretar los modelos de clasificación entrenados empleando librerías de aprendizaje automático como sklearn (Pedregosa et al. 2011) en Python (Van Rossum, G., y Drake Jr 1995) que funcionan como verdaderas “cajas negras”.

Cuando se entrena un modelo de clasificación SVM empleando estas librerías, el resultado es una serie de coeficientes que definen el modelo. ¿Pero qué significan estos coeficientes? ¿Cuál es la conexión con los fundamentos vistos anteriormente? En efecto, dichos coeficientes representan el vector de parámetros θ que vimos anteriormente. Repasemos cuales son las propiedades de este vector y cómo éstas se pueden aprovechar para extraer información del modelo entrenado:

1) El vector θ siempre es perpendicular al límite de decisión que separa las diferentes clases.

2) La proyección de cualquiera de los ejemplos del conjunto de entrenamiento sobre θ indica a qué clase pertenece el ejemplo. Si el producto escalar de ambos resulta en un valor positivo el ejemplo pertenece a la clase positiva, de lo contrario pertenece a la clase negativa.

3) Finalmente, la dirección de θ da información acerca de la importancia de cada descriptor o característica en la construcción del modelo de clasificación. Por ejemplo, en la figura previa θ apunta en la dirección de x_1 (es decir el límite de decisión es ortogonal a x_1) lo cual significa que solo el coeficiente θ_1 es no nulo y por tanto solo x_1 es útil para discriminar entre las clases (dado que θ_2 es nulo). En otras palabras, la magnitud de cada coeficiente o peso es indicativo de la importancia de la correspondiente característica en la definición del modelo de clasificación.

2.3.2 Aprendizaje profundo

A diferencia de lo que se describió anteriormente, donde la metodología de aprendizaje automático se basaba principalmente en la utilización de algoritmos de regresión o árboles de decisión, en esta sección se describirá lo que se considera una evolución sofisticada y matemáticamente compleja de los algoritmos de aprendizaje automático.

En particular, el aprendizaje profundo (DL, del inglés *Deep Learning*) (LeCun et al. 2015) es un subconjunto de ML basado en redes neuronales artificiales que utilizan múltiples capas para extraer progresivamente características más complejas de la entrada sin procesar. Debido a su capacidad para aprender de los datos y el entorno, DL y las redes neuronales (RN), también conocidas como redes neuronales artificiales (RNA), llamadas así por su representación artificial del funcionamiento de un sistema nervioso humano, se han convertido en una de las técnicas más utilizadas para el descubrimiento de fármacos (Nag et al. 2022).

La figura 2.18 representa el funcionamiento de una sola neurona y de una RN típica con varias neuronas conectadas. Como puede observarse, una neurona, también llamada perceptrón recibe un vector de entrada de tamaño *n*, donde a cada entrada se le asigna un peso *W* relacionado con la importancia de algunas respecto a las demás. La función de agregación permite calcular

un valor único a partir de las entradas y de los pesos correspondientes. Por otro lado, la función de activación asocia a cada valor agregado un único valor de salida h_w, dependiendo del umbral.

La función de la salida más básica es la siguiente:

$$h_w(x) = activación\ (x^t w) \tag{8}$$

Donde x es la matriz de entrada, x^t es la traspuesta de la matriz de entrada y w es la matriz de pesos de las entradas.

De este modo, cada neurona generará una función de salida que será pasado a la capa siguiente o se utilizará como salida de la red.

Por otro lado, una red neuronal artificial es una estructura que tiene varios vectores de entrada I = [i1, i2, i3, ..., i_n] y un vector de salida apropiado O = [o1, o2, o3,..., o_m] junto con varios unidades elementales conectadas, conocidas como neuronas. La red inicialmente recibe información en forma de vectores de entrada y tiene como objetivo procesar o aprender de ella. Desde aquí, los datos pasan por una o más capas ocultas que comprenden los patrones ocultos en los datos mediante cálculos y realizan las transformaciones correspondientes. La función de activación o la función de transferencia también actúa sobre los datos procesados para capturar la relación no lineal entre las entradas y también convertirla en una salida más utilizable (Gurney 2018). El desempeño de una RNA depende de la cantidad de capas, la cantidad de neuronas, la función de transferencia, la presencia de un umbral para la activación de cada neurona y la forma en que las neuronas están interconectadas (Puri et al. 2016).

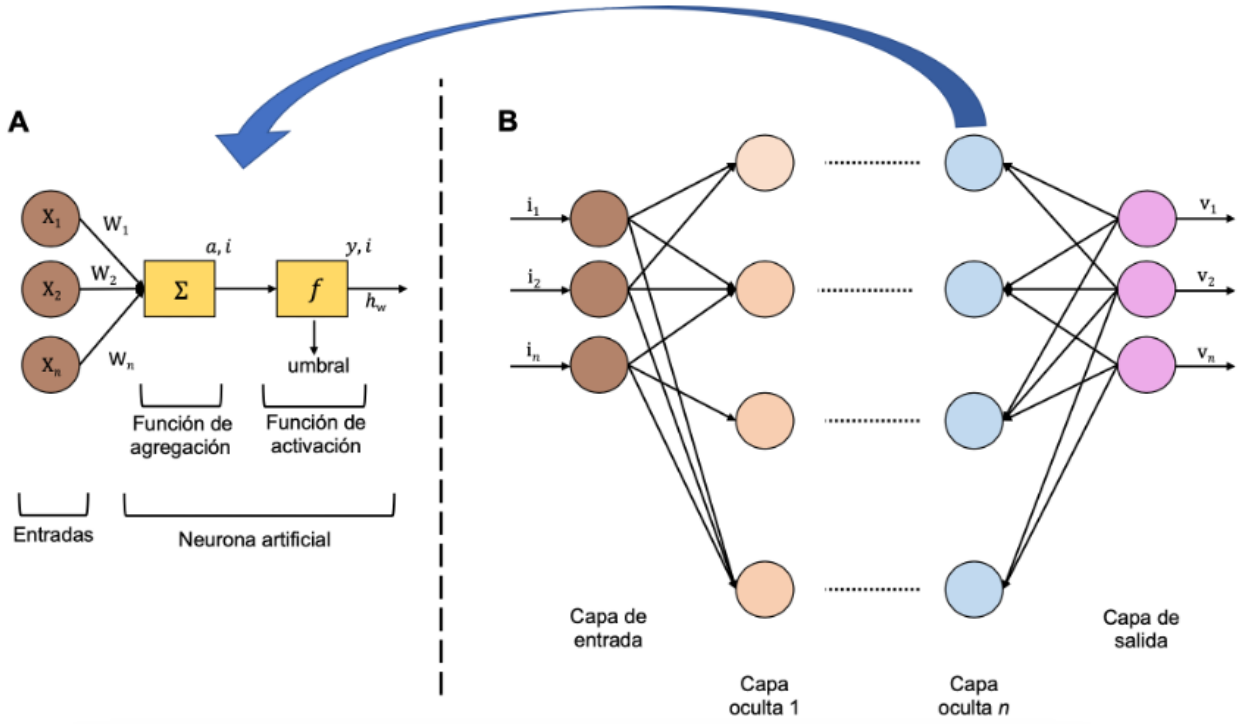


Figura 2.18 A) Diagrama esquemático del funcionamiento de una sola neurona en redes neuronales artificiales (RNA). **B)** RNA, donde se observan diversos vectores de entrada (i1, i2..) y salida (v1, v2..) y sus interconexiones se representan como neuronas. Estas neuronas interconectadas ayudan a mantener la arquitectura de la RNA. El rendimiento de trabajo de una RNA depende de la cantidad de capas, la cantidad de neuronas, la función de transferencia, la presencia de un sesgo y la forma en que las neuronas están interconectadas. Figura adaptada de Nag, S et al. 2022.

Tal como se mencionó anteriormente, las RNA debido a su complejidad tienen la capacidad para aprender de los datos y el entorno. Este proceso de aprendizaje se conoce como entrenamiento, el cual se lleva a cabo mediante la utilización de un subconjunto de datos, donde se dispone de una colección de información de entradas y salidas conocidas. Mediante un proceso iterativo se introducen ejemplos a la red y se compara el resultado obtenido con el esperado y se calcula la diferencia entre ambos. El resultado de esta comparación actualiza el peso de cada conexión siguiendo una fórmula denominada función de actualización la cuál se describe a continuación:

$$W_{i,j}^{(t)} = W_{i,j}^{(t-1)} + \eta(y_j - y`_j)x \qquad (9)$$

Donde $W_{i,j}$ es el peso de la entrada i en la neurona j; x_i es el valor de la posición i de la entrada; y'_i es la salida de la neurona en la iteración de aprendizaje actual; y_j es la salida esperada en la neurona y η es el ritmo de aprendizaje (Moure Ortega, 2021).

Referencias del capítulo 2

Bader, R. 1990. Atoms in molecules: a quantum theory (AIM). Oxford University Press.

Bhavsar, Parth, Ilya Safro, Nidhal Bouaynaya, Robi Polikar, and Dimah Dera. 2017. *Machine Learning in Transportation Data Analytics*. *Data Analytics for Intelligent Transportation Systems*. Elsevier Inc. https://doi.org/10.1016/B978-0-12-809715-1.00012-2.

Cortes, C., Vapnik, V. Support-vector networks. Mach Learn 20, 273–297 (1995). https://doi.org/10.1007/BF00994018

Gao, Mingshan, Lei Duan, Jinfeng Luo, Lianwen Zhang, Xiaoyun Lu, Yan Zhang, Zhang Zhang, et al. 2013. "Discovery and Optimization of 3-(2-(Pyrazolo[1,5- a]Pyrimidin-6-Yl) Ethynyl)Benzamides as Novel Selective and Orally Bioavailable Discoidin Domain Receptor 1 (DDR1) Inhibitors." *Journal of Medicinal Chemistry* 56 (8): 3281–95. https://doi.org/10.1021/jm301824k.

Ge, Hu, Yu Wang, Chanjuan Li, Nanhao Chen, Yufang Xie, Mengyan Xu, Yingyan He, et al. 2013. "Molecular Dynamics-Based Virtual Screening: Accelerating the Drug Discovery Process by High-Performance Computing." *Journal of Chemical Information and Modeling* 53 (10): 2757–64. https://doi.org/10.1021/ci400391s.

Gereon, A. (2018). Hands-on Machine Learning with Scikit-Learn and Tensor Flow. *O'Reily Media Inc*., USA.

Gil Redondo, Rubén. 2010. "Desarrollo y Utilización de Métodos Computacionales En La Mejora Del Proceso de Obtención de Nuevos Fármacos." Universidad Autónoma de Madrid.

Gilson, Michael K., and Huan-Xiang Zhou. 2007. "Calculation of Protein-Ligand Binding Affinities." *Annual Review of Biophysics and Biomolecular Structure* 36 (1): 21–42. https://doi.org/10.1146/annurev.biophys.36.040306.132550.

Graves, Alan P., Ruth Brenk, and Brian K. Shoichet. 2005. "Decoys for Docking." *Journal of Medicinal Chemistry* 48 (11): 3714–28. https://doi.org/10.1021/jm0491187.

Guha, Rajarshi, and Andreas Bender. 2011. *Computational Approaches in Cheminformatics and Bioinformatics.* Edited by Rajarshi Guha and Andreas Bender. *Computational Approaches in Cheminformatics and Bioinformatics*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118131411.

Gunsteren, Wilfred F. van, and Herman J.C. Berendsen. 1990. “Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry.” *Angewandte Chemie International Edition in English*. John Wiley & Sons, Ltd. https://doi.org/10.1002/anie.199009921.

Hansson, Tomas, Chris Oostenbrink, and Wilfred F. Van Gunsteren. 2002. “Molecular Dynamics Simulations.” *Current Opinion in Structural Biology*. Elsevier Ltd. https://doi.org/10.1016/S0959-440X(02)00308-1.

Hassan Baig, Mohammad, Khurshid Ahmad, Sudeep Roy, Jalaluddin Mohammad Ashraf, Mohd Adil, Mohammad Haris Siddiqui, Saif Khan, Mohammad Amjad Kamal, Ivo Provazník, and Inho Choi. 2016. “Computer Aided Drug Design: Success and Limitations.” *Current Pharmaceutical Design* 22 (5): 572–81. https://doi.org/10.2174/1381612822666151125000550.

Kaul, Pushkar N. 1998. “Drug Discovery: Past, Present and Future.” *Progress in Drug Research*. Birkhauser Verlag AG. https://doi.org/10.1007/978-3-0348-8833-2_1.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

Li, H., Sze, K. H., Lu, G., & Ballester, P. J. (2020). Machine-learning scoring functions for structure-based drug lead optimization. Wiley Interdisciplinary Reviews: *Computational Molecular Science*, 10(5), e1465.

Lionta, Evanthia, George Spyrou, Demetrios Vassilatis, and Zoe Cournia. 2014. “Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances.” *Current Topics in Medicinal Chemistry* 14 (16): 1923–38. https://doi.org/10.2174/1568026614666140929124445.

Lucas, Xavier, Björn A. Grüning, Stefan Bleher, and Stefan Günther. 2015. “The Purchasable Chemical Space: A Detailed Picture.” *Journal of Chemical Information and Modeling* 55 (5): 915–24. https://doi.org/10.1021/acs.jcim.5b00116.

M. Lourenco, A., L. M. Ferreira, and P. S. Branco. 2012. “Molecules of Natural Origin, Semi-Synthesis and Synthesis with Anti-Inflammatory and Anticancer Utilities.” *Current Pharmaceutical Design* 18 (26): 3979–4046. https://doi.org/10.2174/138161212802083644.

Mohan, Venkatraman, Alan Gibbs, Maxwell Cummings, Edward Jaeger, and Renee DesJarlais. 2005. “Docking: Successes and Challenges.” *Current Pharmaceutical Design* 11 (3): 323–33. https://doi.org/10.2174/1381612053382106.

Müller, A.C. 2017. Introduction to machine learning with python.

Nag, Sagorika, Anurag T.K. Baidya, Abhimanyu Mandal, Alen T Mathew, Bhanuranjan Das, Bharti Devi, and Rajnish Kumar. 2022. “Deep Learning Tools for Advancing Drug Discovery and Development.” *3 Biotech*. https://doi.org/10.1007/s13205-022-03165-8.

Newman, David J., and Gordon M. Cragg. 2007. “Natural Products as Sources of New Drugs over the Last 25 Years.” *Journal of Natural Products*. American Chemical Society . https://doi.org/10.1021/np068054v.

Ortega Moure, Alfonso. 2021. “Definición , Tipologías y Casos de Uso de Graph Neural Networks Para El Aprendizaje Basado En Relaciones.” Universitat Oberta de Catalunya.

Parthasarathi, R., V. Subramanian, and N. Sathyamurthy. 2006. “Hydrogen Bonding without Borders: An Atoms-in-Molecules Perspective.” *Journal of Physical Chemistry A* 110 (10): 3349–51. https://doi.org/10.1021/jp060571z.

Pedregosa, Fabian, Vincent Michel, Olivier Grisel OLIVIERGRISEL, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Jake Vanderplas, et al. 2011. “Scikit-Learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot.” *Journal of Machine Learning Research*. Vol. 12. http://scikit-learn.sourceforge.net.

Pham, Tuan A., and Ajay N. Jain. 2008. “Customizing Scoring Functions for Docking.” *Journal of Computer-Aided Molecular Design* 22 (5): 269–86. https://doi.org/10.1007/s10822-008-9174-y.

Puri, Munish, Aum Solanki, Timothy Padawer, Srinivas M. Tipparaju, Wilfrido Alejandro Moreno, and Yashwant Pathak. 2016. “Introduction to Artificial Neural Network (ANN) as a Predictive Tool for Drug Design, Discovery, Delivery, and Disposition.” *Artificial Neural Network for Drug Design, Delivery and Disposition*, 3–13. https://doi.org/10.1016/B978-0-12-801559-9.00001-6.

Reardon, Sara. 2013. “Project Ranks Billions of Drug Interactions.” *Nature* 503 (7477): 449–50. https://doi.org/10.1038/503449a.

Rossum, G., & Drake Jr, F. L. Van. 1995. “Python Reference Manual.” *Centrum Voor Wiskunde En Informatica Amsterdam*.

Saldívar-González, Fernanda, Fernando D. Prieto-Martínez, and José L. Medina-Franco. 2017. “Descubrimiento y Desarrollo de Fármacos: Un Enfoque Computacional.” *Educacion Quimica* 28 (1): 51–58. https://doi.org/10.1016/j.eq.2016.06.002.

Samuel, A L. 1959. “Some Studies in Machine Learning Using the Game of Checkers.” *IBM Journal of Research and Development* 44 (1–2): 207–19. https://doi.org/10.1147/rd.441.0206.

Scior, Thomas, Andreas Bender, Gary Tresadern, Jose L. Medina-Franco, Karina Martínez-Mayorga, Thierry Langer, Karina Cuanalo-Contreras, and Dimitris K. Agrafiotis. 2012. “Recognizing Pitfalls in Virtual Screening: A Critical Review.” *Journal of Chemical Information and Modeling*. American Chemical Society. https://doi.org/10.1021/ci200528d.

Seifert, Markus H J, and Martin Lang. 2007. “Essential Factors for Successful Virtual Screening,” 63–72.

Song, Chun Meng, Shen Jean Lim, and Joo Chuan Tong. 2009. “Recent Advances in Computer-Aided Drug Design.” *Briefings in Bioinformatics* 10 (5): 579–91. https://doi.org/10.1093/bib/bbp023.

Taft, Carlton A., Vinicius Barreto Da Silva, and Carlos Henrique Tomich De Paula Da Silva. 2008. “Current Topics in Computer-Aided Drug Design.” *Journal of Pharmaceutical Sciences*. John Wiley and Sons Inc. https://doi.org/10.1002/jps.21293.

Triballeau, Nicolas, Hugues Olivier Bertrand, and Francine Acher. 2006. “Are You Sure You Have a Good Model?” In *Pharmacophores and Pharmacophore Searches*, 325–64. https://doi.org/10.1002/3527609164.ch15.

Veselovsky, A. V., M. S. Zharkova, V. V. Poroikov, and M. C. Nicklaus. 2014. “Computer-Aided Design and Discovery of Protein-Protein Interaction Inhibitors as Agents for Anti-HIV Therapy.” *SAR and QSAR in Environmental Research* 25 (6): 457–71.

https://doi.org/10.1080/1062936X.2014.898689.

Wikberg, Jarl E.S., Ola Spjuth, Martin Eklund, and Maris Lapins. 2011. “Chemoinformatics Taking Biology into Account: Proteochemometrics.” In *Computational Approaches in Cheminformatics and Bioinformatics*, 57–92. John Wiley and Sons. https://doi.org/10.1002/9781118131411.ch3.

CAPÍTULO III

“Análisis estructural de Cz e inhibidores conocidos”

3.1 Introducción

La determinación tridimensional de estructuras proteicas ha demostrado ser información invaluable, ya que complementa la información biológica y bioquímica de otros tipos de experimentos. A su vez, la información estructural es una herramienta crucial para el diseño racional de medicamentos, ya que mediante estas técnicas se pueden ahorrar aproximadamente el 50% del costo del descubrimiento de fármacos (Stevens 2004).

Bajo este paradigma sienta sus bases el Descubrimiento de Fármacos Basado en la Estructura (DFAC) ya que se basa en el conocimiento de la estructura tridimensional de la diana biológica obtenida a través de diversos métodos como la cristalografía de rayos X o la espectroscopia de RMN (Jhoti y Leach 2007).

Usando la estructura del blanco molecular, los candidatos a fármacos pueden ser modelados para que se unan con alta afinidad y selectividad (Mauser y Guba 2008). Para ello, la identificación del sitio de enlace es el primer paso en el diseño basado en la estructura (Yuan, Pei, y Lai 2013). Si la estructura de la diana se determina en presencia de un ligando unido, entonces el ligando debe ser observable en la estructura, en cuyo caso la ubicación del sitio de unión es trivial. Sin embargo, puede haber sitios de unión alostéricos desocupados que pueden ser de interés. Además, puede ser que solo las estructuras de apoproteína (proteína sin ligando) estén disponibles y la identificación confiable de sitios desocupados que tengan el potencial de unir ligandos con alta afinidad no sea trivial. En resumen, la identificación del sitio de unión generalmente se basa en la identificación de superficies cóncavas en la proteína que pueden acomodar moléculas del tamaño del fármaco que también poseen "puntos calientes" apropiados (superficies hidrofóbicas, sitios de enlace de hidrógeno, etc.) que impulsan la unión del ligando (Leis, Schneider, y Zacharias 2010; Yuan, Pei, y Lai 2013).

En el caso de Cz, a partir de la primera estructura tridimensional cristalizada (Gillmor, Craik, y Fletterick 1997), numerosos han sido los esfuerzos llegando a obtener actualmente 27 entradas asociadas a este blanco molecular en el Protein Data Bank (rcsb.org) donde Cz se ha cristalizado con inhibidores reversibles e irreversibles. Por lo tanto, Cz se presenta como un objetivo atractivo para el desarrollo de posibles terapias para el tratamiento de la enfermedad mediante el uso de un enfoque basado en la estructura (Barbosa da Silva et al. 2019).

Sin embargo, para cristalizar una proteína se la somete a condiciones muy particulares, por ejemplo, baja temperatura, distinta fuerza iónica respecto a su ambiente “natural”, entre otras. Incluso la proteína en el cristal puede adoptar una conformación en donde algunos aminoácidos están demasiado cerca (McPherson y Gavira 2014).

Todo lo anterior hace que la estructura cristalográfica muchas veces no sea representativa de la estructura a 298°K, pH fisiológico, ni a fuerza iónica de un entorno biorrelevante, con lo cual muchas veces sucede que la estructura cristalográfica resulta ser tan solo una aproximación de una proteína en su estado natural (Wang 2013). Si bien esta estructura puede aportar alguna idea de su comportamiento en condiciones biológicas relevantes, no siempre es concluyente. Es por eso que se necesita llevar la estructura del cristal de la proteína a las condiciones más próximas al entorno (Mirjalili y Feig 2013; Feig 2016). Tomando como punto de partida lo dicho anteriormente, los cristales deben ser pretratados para luego interpretar las interacciones que allí ocurren. Es por ello que los complejos ligando-proteína se someten a simulaciones de dinámicas moleculares a fin de optimizar dichas estructuras.

Por otra parte, el estudio de las interacciones intermoleculares a partir de diferentes métodos proporciona información crucial para la caracterización de puntos calientes de interacción en el sitio activo. En este contexto, la aplicación de la metodología QTAIM en entornos biomoleculares permite detectar interacciones no direccionales, por ejemplo, aquellas que involucran electrones π en anillos aromáticos, entre otros contactos débiles e inusuales que de otro modo se perderían en un análisis meramente geométrico de las interacciones (Angelina et al. 2014).

3.2 Metodología

3.2.1 Construcción de biblioteca de ligandos conocidos

Se realizó la compilación de una biblioteca de complejos Cz-inh de estructura conocida. Las estructuras de los complejos Cz-inh que fueron determinadas experimentalmente se descargaron del *Protein Data Bank* (PDB, rcsb.org).

3.2.2 Dinámicas Moleculares

Todas las simulaciones de complejos Cz-inh se llevaron a cabo con el paquete de software Amber14 (David A. Case et al. 2005; D. A. Case et al. 2014) a una temperatura de 300 K y se extendieron hasta 50 ns de tiempo total de simulación en una caja periódica octaédrica truncada de moléculas de agua TIP3P. Se usó el campo de fuerza amber ff14SB para los residuos de proteínas (Maier et al. 2015). Se utilizó el software Antechamber del paquete Amber-Tools para generar los parámetros del inhibidor con el campo de fuerza GAFF y se calcularon cargas RESP (Wang et al. 2004).

3.2.3 Análisis de las interacciones intermoleculares

Se seleccionó una estructura representativa de cada complejo Cz-inh. Para ello, se buscó la estructura correspondiente al mínimo de energía potencial de las trayectorias DM. Sobre dichas estructuras, se realizaron los siguientes análisis:

- El primero consistió en el estudio de interacciones basándose únicamente en distancias interatómicas, siguiendo el enfoque utilizado por Deng et al. 2004. Esta metodología se basa en la construcción de una huella digital (*fingerprint*) de interacción que traduce la información de unión estructural 3D de un complejo proteína-ligando en una cadena binaria unidimensional (comprendida por 1 denotando presencia y 0 ausencia de una determinada característica). Para ello, se extrajeron y clasificaron siete tipos diferentes de interacciones que ocurren en cada residuo correspondiente al bolsillo de unión, las cuales se enumeran a continuación: (i) Interacciones débiles (del tipo dispersión de London); (ii) Nube π (cara a cara); (iii) Nube π (borde a cara); (iv) Puente de hidrógeno (proteína dador/ ligando aceptor); (v) Puente de hidrógeno (proteína aceptor/ ligando dador); (vi) Puente salino (proteína +/ ligando -); y (vii) Puente salino (proteína -/ ligando +).

Por otro lado se utilizó el coeficiente de Tanimoto (Tc) como medida cuantitativa de la similitud de la cadena de bits (Willett, Barnard, y Downs 1998). El Tc entre dos cadenas de bits A y B se define como:

$$Tc(A,B) = |A \cap B| / |A \cup B|$$

donde $|A \cap B|$ representa el número de interacciones o *bits* encendidos comunes entre los complejos A y B; y $|A \cup B|$ representa el número de interacciones o *bits* encendidos en uno de los complejos A o B. El valor del coeficiente de Tanimoto oscila entre 0 y 1, donde 0 representa máxima disimilitud (es decir ningún bit o tipo interacción en común) y 1 representa máxima similitud (todos los bits de la huella dactilar en común).

Este análisis fue llevado a cabo mediante la utilización de las herramientas quimioinformáticas incluidas en el software OpenBabel (O'Boyle et al. 2011).

- El segundo consistió en el cálculo de la densidad de carga. Debido a que los cálculos mecano-cuánticos aún son irrealizables desde el punto de vista computacional para los complejos biomoleculares completos, se construyeron modelos reducidos a partir de las estructuras de mínimo de energía potencial. Se incluyeron un total de 28 residuos (~570 átomos) en los modelos reducidos: el inhibidor y los residuos circundantes en un volumen esférico de aproximadamente 5 Å centrado en los átomos del inhibidor. La densidad de carga se calculó con ayuda del paquete Gaussian 09 (Frisch et al. 2016) mediante la metodología DFT con un funcional híbrido corregido por dispersión M06-2x y el conjunto base 6-31G (d). El análisis topológico de la densidad de carga se realizó luego con el software Multiwfn (Lu y Chen 2012).

3.3 Resultados y Discusión

3.3.1 Análisis de la secuencia de Cz

Primeramente se realizó un análisis comparativo de la identidad de la cadena principal de aminoácidos de Cz con enzimas pertenecientes a la familia de proteasas de diferentes organismos utilizando el servidor Clustal Omega (Sievers and Higgins 2014). Las cinco enzimas con mayor identidad se muestran en la Figura 3.1.

Según los datos obtenidos a partir de la matriz de identidad, se puede establecer que Cz tiene las siguientes identidades con las diferentes proteasas analizadas: rodesaína perteneciente al parásito Trypanosoma brucei 71%, falcipaína perteneciente al parásito Plasmodium falciparum 37%, papaína perteneciente a Carica papaya 38% y las enzimas de catepsinas humanas L1 y B 46% y 28% respectivamente.

El alto grado de identidad entre estas estructuras, en particular con rodesaína, sugiere que los inhibidores de la Cz pueden potencialmente tener actividad en estas enzimas (H. Wiggers 2011). Esto se evidencia aún más visualmente, haciendo una superposición de las estructuras tridimencionales de ambos blancos moleculares (Figura 3.2).

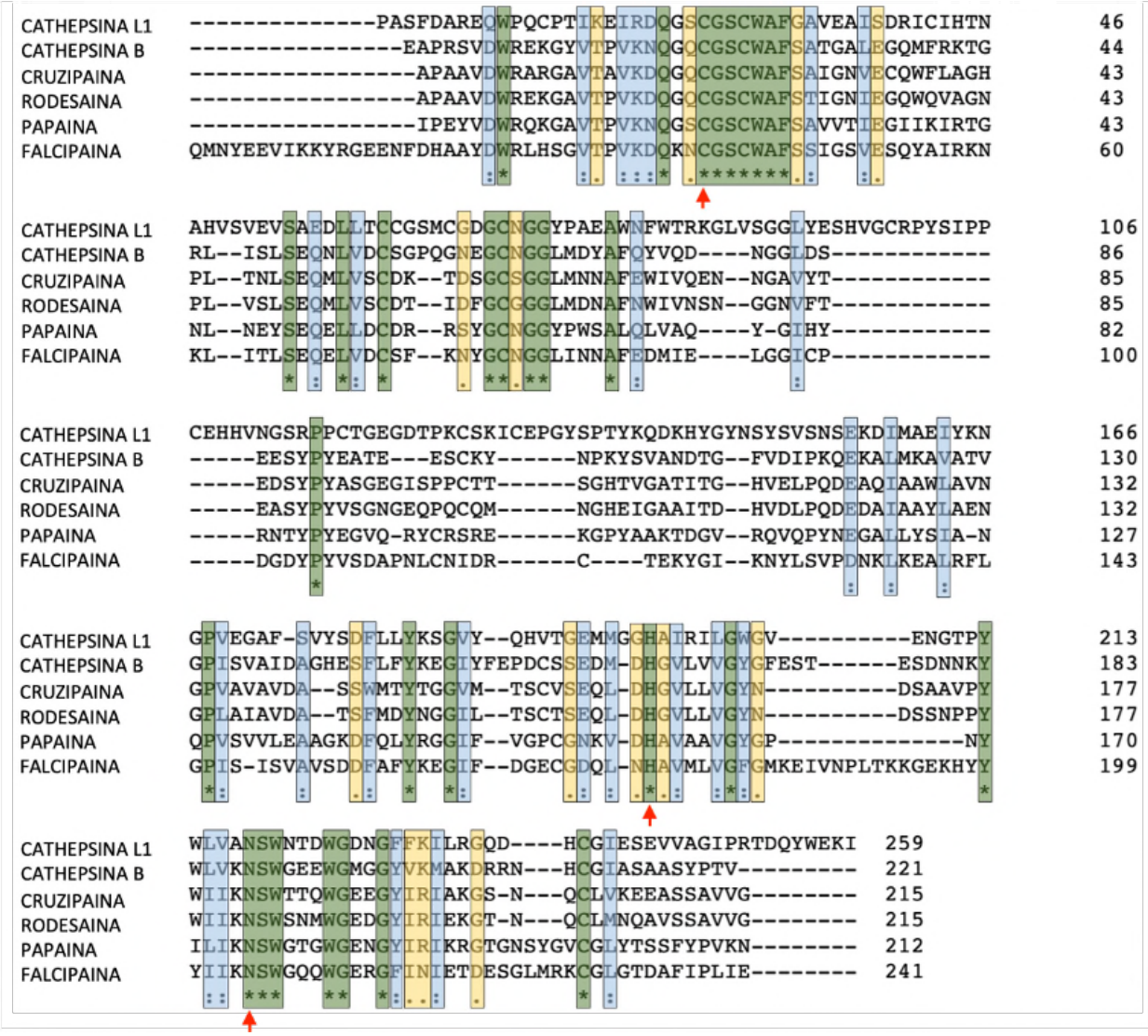


Figura 3.1 Alineamiento de estructuras primarias de enzimas proteicas de diferentes organismos con cruzaína. Los rectángulos verdes denotan conservación de aminoácidos en todas las estructuras. Los amarillos representan conservación en por lo menos cuatro cadenas y los celestes en por lo menos tres. Las flechas rojas indican los residuos de la triada catalítica, los cuales se encuentran conservados en todas las cadenas.

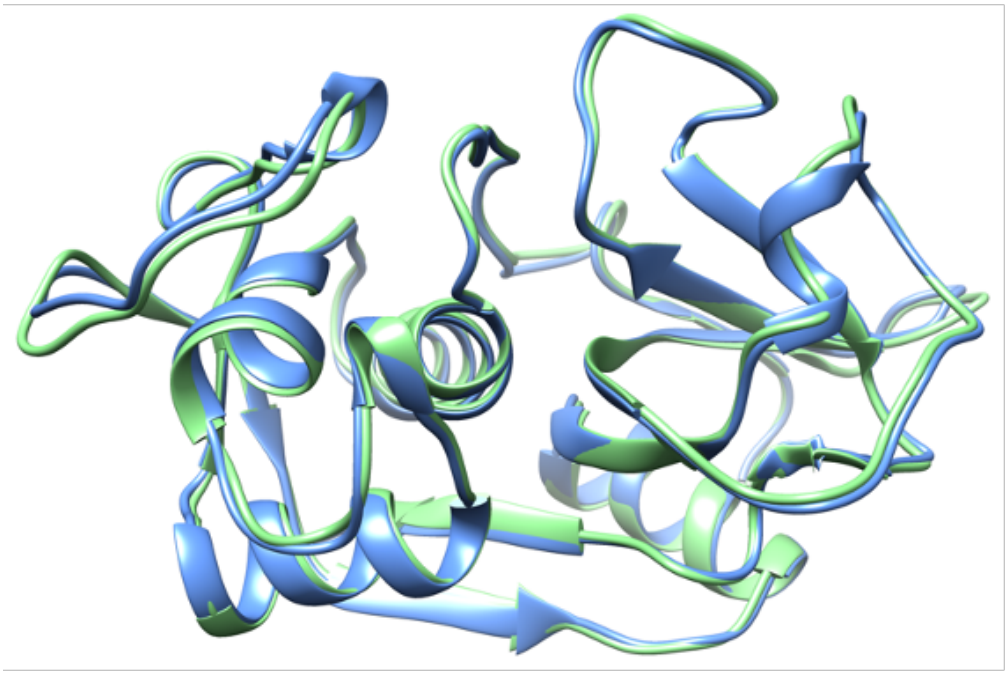
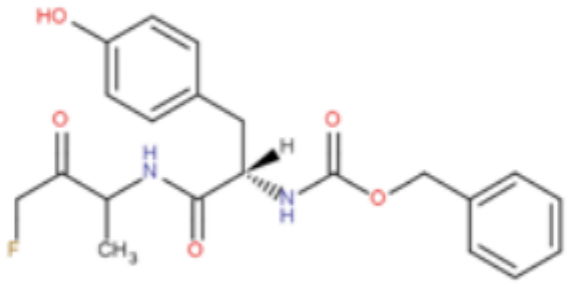
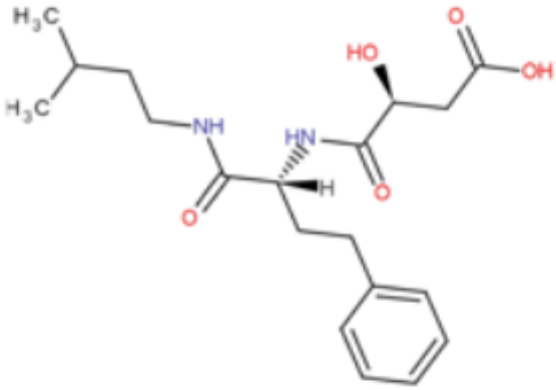
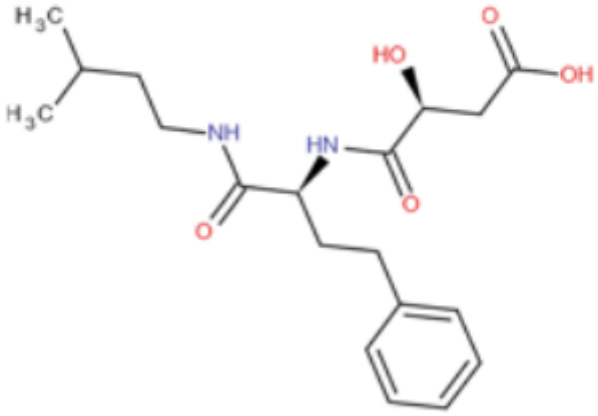


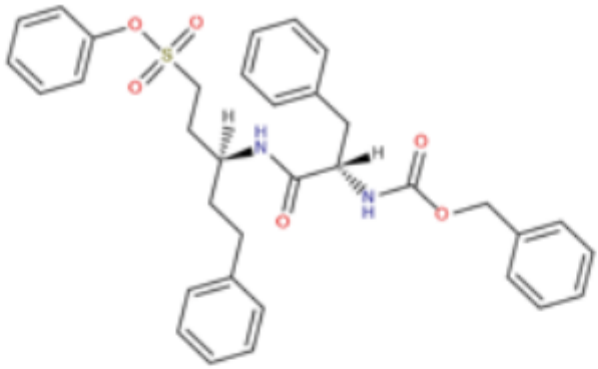
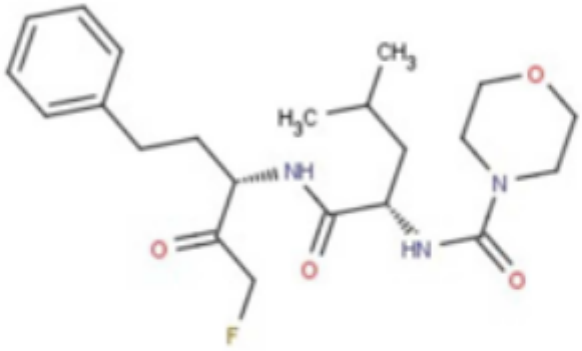
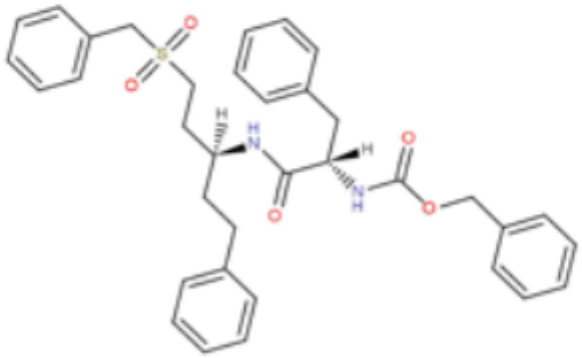
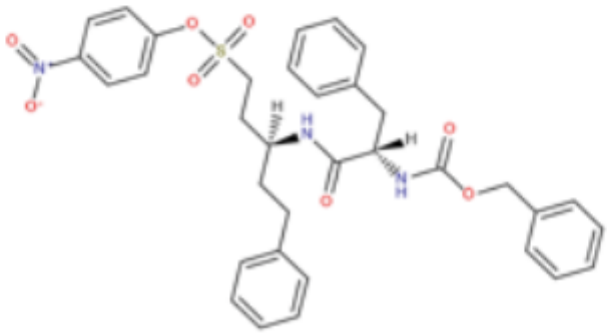
Figura 3.2 Superposición de estructuras terciarias de Cz (verde) y rodesaína (azul). Códigos PDB: 2OZ2 y 2P86 respectivamente. Figura construida con el software UCSF Chimera (Pettersen et al. 2004).

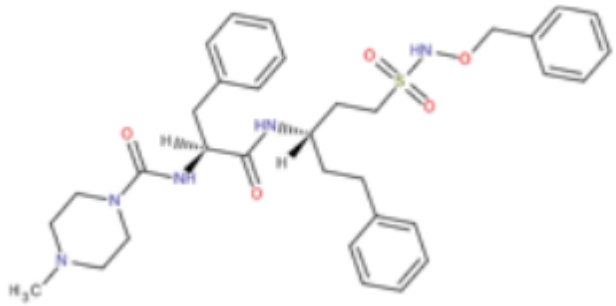
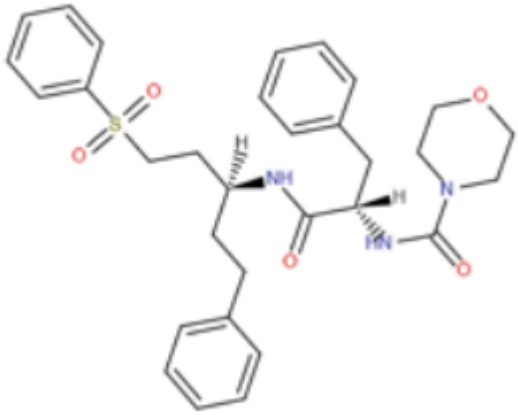
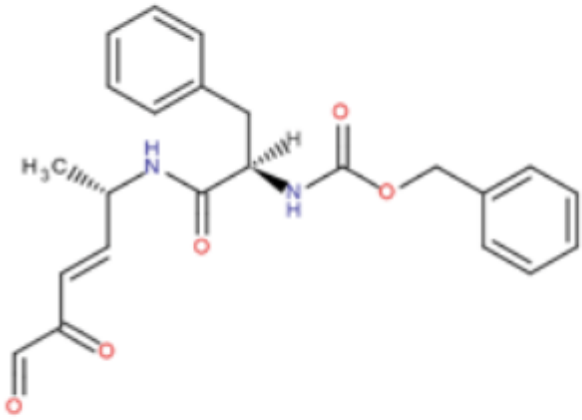
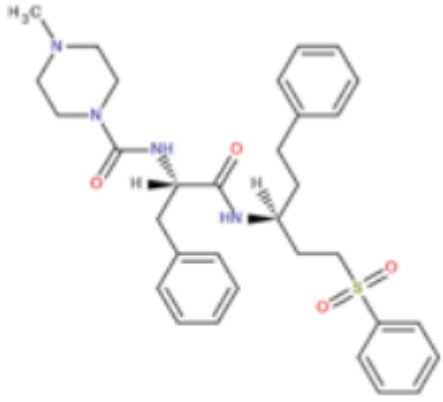
3.3.2 Biblioteca de complejos

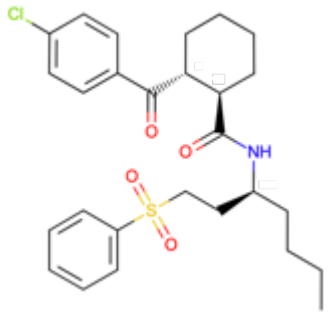
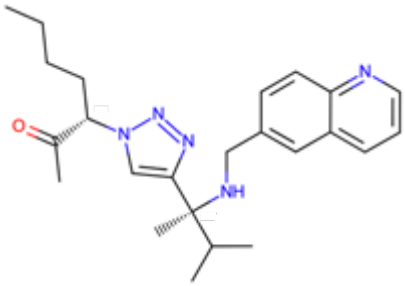
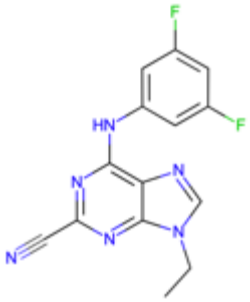
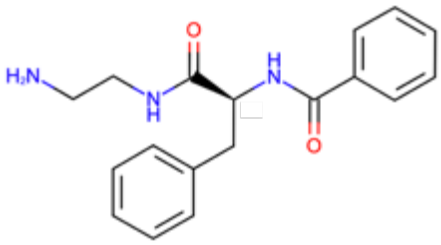
Se buscó en la base de datos de proteínas (PDB) estructuras cristalinas de Cz formando complejos Cz-inh. Se encontraron un total de 27 estructuras depositadas en la base de datos, de las cuales solo se seleccionaron aquellas estructuras con inhibidores unidos covalentemente. Los datos cristalográficos correspondientes a las estructuras depositadas en el PDB se muestran en la tabla 3.1.

Tabla 3.1 Información referida a los cristales seleccionados para la conformación de la biblioteca de complejos Cz-inh.

Código PDB	Resolución (Å)	Ligando co-cristalizado	Familia de compuesto (según el grupo reactivo)
1AIM	2,00		Fluorometilcetona
1EWL	2,00		Hidroximetilcetona
1EWM	2,00		Hidroximetilcetona

1EWO	2,10		Vinilsulfona
1EWP	1,75		Fluorometilcetona
1F2A	1,60		Vinilsulfona
1F2B	1,80		Vinilsulfona

1F2C	2,00		Vinilsulfona
1F29	2,15		Vinilsulfona
1U9Q	2,30		Alfacetoester
2OZ2	1,95		Vinilsulfona (K777)

3HD3	1,75		Vinilsulfona
3IUT	1,20		Metilcetona
3I06	1,10		Derivado de nitrilo
4QH6	3,13		Derivado de nitrilo

Un parámetro importante en los estudios cristalográficos es la resolución de los cristales, expresados en Å, donde los valores más pequeños indican resoluciones más altas. A altas resoluciones (<1,5 Å), el modelo es probablemente más del 95% una consecuencia de los datos observados. Sin embargo, a bajas resoluciones (> 2,5 Å), el modelado de detalles en estructuras

de proteínas es mucho más subjetivo que basado en información experimental (H. J. Wiggers et al. 2013).

Cabe destacar que de todas las estructuras depositadas en la base de datos, no todas contaban con el respectivo dato experimental de porcentaje de inhibición IC50 o K_i (constante de inhibición) para los inhibidores co-cristalizados con la proteína, siendo en su mayoría derivados de vinilsulfonas.

3.3.3 Análisis de drogabilidad y sub-bolsillos

La drogabilidad (del inglés *druggability*) es un término utilizado en el descubrimiento de fármacos para describir un blanco molecular (como una proteína) que se sabe o se prevé que se una con alta afinidad a un fármaco. Además, por definición, la unión del fármaco a un objetivo susceptible de fármaco debe alterar la función del mismo con un beneficio terapéutico para el paciente (Cheng et al. 2007).

La drogabilidad se predice utilizando diferentes métodos que se basan en relaciones evolutivas, propiedades estructurales 3D u otros descriptores (Al-Lazikani et al. 2007).

Desde el punto de vista del DFAC, se define el término drogabilidad como una estimación cuantitativa de los sitios de unión y las afinidades por un fármaco potencial que actúa sobre una proteína diana específica (Bakan et al. 2012). Estas aproximaciones se basan en la disponibilidad de estructuras 3D determinadas experimentalmente o modelos de homología de alta calidad. Existen varios métodos para esta evaluación de la drogabilidad, pero todos constan de tres componentes principales (Nayal y Honig 2006; Seco et al. 2009; Halgren 2009; Bakan et al. 2012):

- Identificación de cavidades o bolsillos en la estructura.
- Calcular propiedades fisicoquímicas y geométricas del bolsillo.
- Evaluar cómo estas propiedades se ajustan a un conjunto de entrenamiento de objetivos conocidos.

En el caso de Cz, diversos estudios se han enfocado en tratar de definir el bolsillo de unión y los diferentes sub-bolsillos que conforman el sitio catalítico (Turk et al. 1998; Bryant et al. 2009; Durrant et al. 2010), pudiéndose caracterizar 4 sub-bolsillos (S1, S1`, S2, S4) bien definidos, a

los que se agregan otros (S1`, S2`, S3`, S3) dependiendo del ajuste conformacional adoptado por Cz inducido por diferentes ligandos (Rodrigues Sartori et al. 2019; Luchi et al. 2019). A su vez, las porciones del ligando que se unen a esos determinados sub-bolsillos se les asignan nombres (P1, P2, P3, etc). Ver Figura 3.3 y Tabla 3.2.

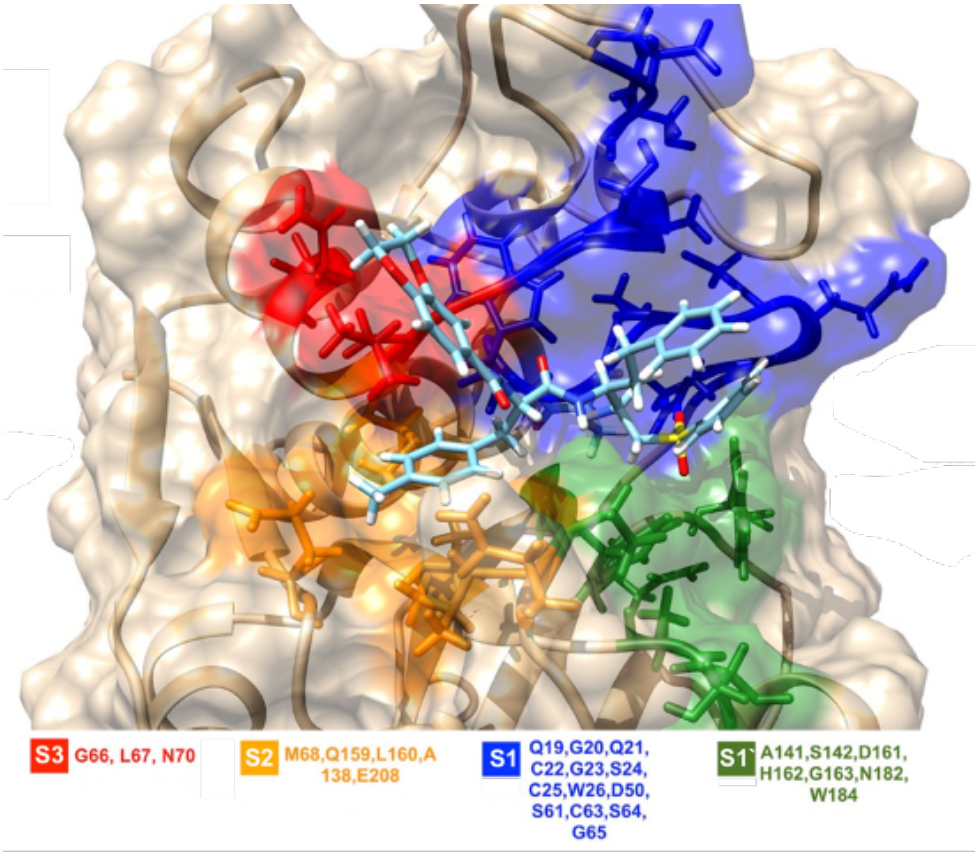


Figura 3.3 Estructura 3D del sitio catalítico de Cz con sus sub-bolsillos. Los residuos correspondientes a los diferentes sub-bolsillos se seleccionaron tomando una distancia de corte de 5 Å desde cada átomo del ligando.

Tabla 3.2 Características de los sub-bolsillos de Cz

Sub-bolsillos	Volumen ($Å^3$)	Área ($Å^2$)	Nº de residuos
S1	385,65	341,00	13
S1`	121,86	266,43	7
S2	143,04	223,02	5
S3	119,43	142,74	3

3.3.4 Análisis dinámico de Cz e interacciones con ligandos de unión covalente

Como se ha visto, los inhibidores seleccionados, son de tipo “suicidas” ya que se unen de forma irreversible al sitio activo de Cz. Esto los convierte en inactivadores del blanco molecular, pero que en ocasiones pueden causar efectos secundarios indeseables, pudiéndose unir a otras proteínas similares (Young 2009; Müller, S., Cerdan, R., & Radulescu 2016).

Para indagar más acerca del comportamiento de los diferentes inhibidores en el sitio catalítico de Cz se realizaron corridas de DM con el fin de observar la dinámica de la enzima y los ligandos y las interacciones intermoleculares que tenían lugar en estos complejos luego de que los ligandos ya se han unidos al sitio catalítico.

Los cálculos de dinámica molecular incorporan las ecuaciones de movimiento de Newton para simular la vibración y el movimiento molecular dentro de un solvente. A su vez, en una simulación de DM clásica que utiliza campos de fuerza estándar, no se puede simular la formación o ruptura de enlaces químicos. Es por ello, que se construyeron modelos de Cz-inh que constaban de la proteína de interés con un ligando unido covalentemente, para luego someterlo a simulaciones de DM. Estos modelos tenían la particularidad de que primeramente se debía parametrizar el ligando para que sea reconocido por el campo de fuerza utilizado.

Como puede observarse en la Figura 3.4 el cálculo de RMSF (*random mean square fluctuation*) nos proporciona una idea de cómo se mueven los Cα de los residuos de la proteína de interés en función del tiempo, pudiéndose ver en este caso una estabilidad generalizada en Cz. Solo algunos segmentos correspondientes a *loops* poseen alta movilidad pero los mismos se

encuentran lejos del sitio catalítico. Además podemos observar que los residuos correspondientes a la triada catalítica son los que menos movilidad presentan (flechas rojas en la Figura 3.4) indicando la presencia de interacciones importantes tanto inter como intramoleculares, las cuales estabilizan estos residuos y las discutiremos más adelante.

Por otro lado puede observarse cierta estabilidad de los residuos correspondientes al sitio catalítico indicando que los mismos disminuyen su movilidad al estar implicados en interacciones con los diferentes inhibidores.

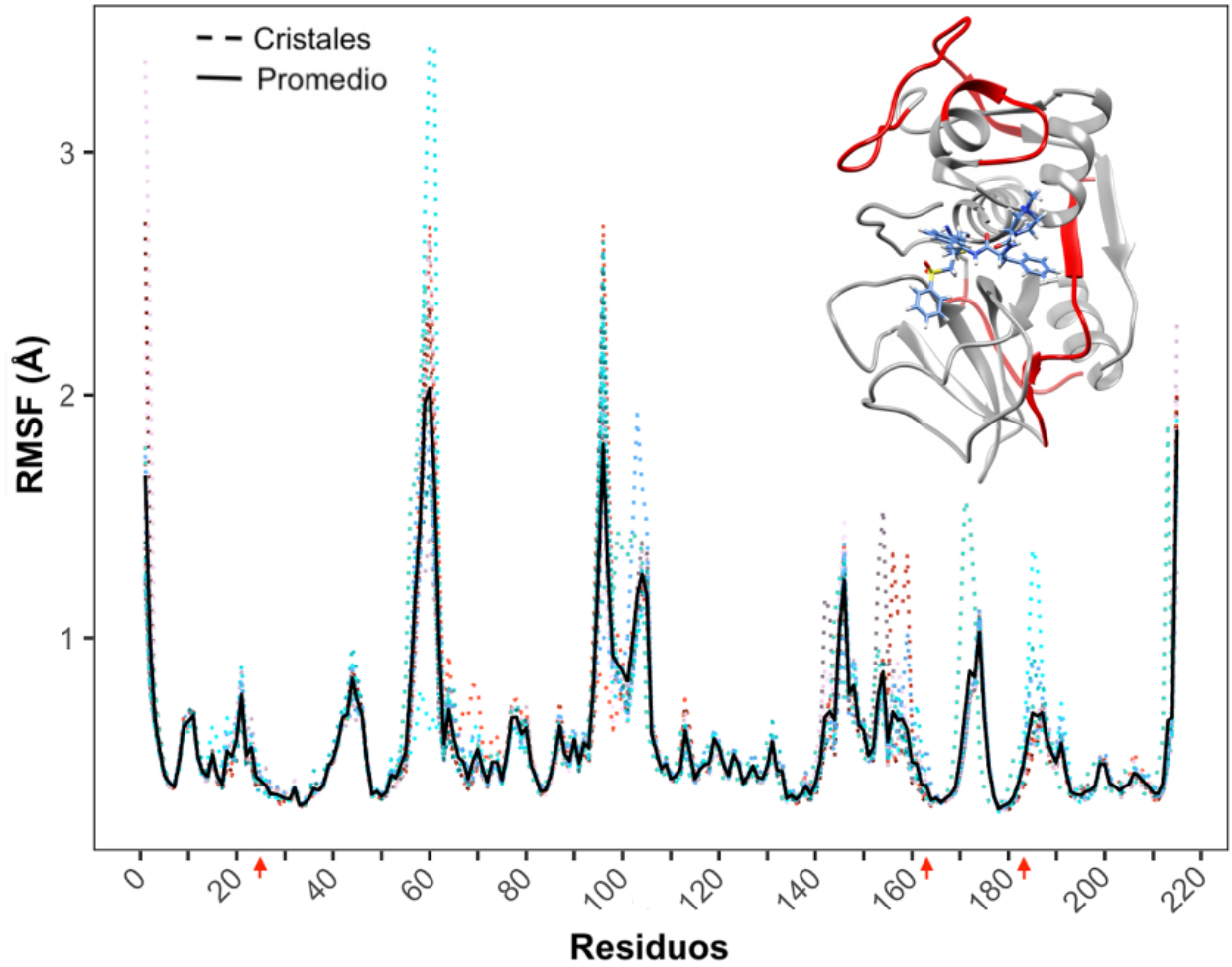


Figura 3.4 RMSF correspondiente a las DMs de Cz unida de manera covalente con los diferentes inhibidores seleccionados. Las líneas de puntos representan cada una de las corridas de DM de los complejos Cz-inh. La línea negra representa el promedio de todas las DMs. Las flechas rojas indican los residuos pertenecientes a la triada catalítica. En el vértice superior derecho del gráfico se muestra la estructura tridimencional de Cz con los segmentos de mayor movilidad coloreados de rojo.

A su vez, se seleccionaron las estructuras del mínimo de energía potencial de las trayectorias DM de los complejos Cz-Inh como una estructura representativa única sobre la que se realizó el análisis de interacciones, siguiendo dos enfoques diferentes que se detallan a continuación.

3.3.4.1 Análisis de huellas dactilares (*fingerprints)* de interacción

A partir de las estructuras de mínimo de energía potencial, extraídas de las trayectorias de dinámica molecular se realizó el primer análisis de interacciones siguiendo la metodología SIFt del inglés *Structural Interaction Fingerprint* (Deng, Chuaqui, y Singh 2004). Esta primera aproximación, de relativo bajo costo computacional, se basa en el cómputo de distancias interatómicas en el bolsillo de unión del receptor a fin de encontrar interacciones relevantes.

La clave de este enfoque es la generación de una huella digital de interacción que traduce la información de unión estructural 3D de un complejo proteína-ligando en una cadena binaria unidimensional. Cada huella dactilar representa el "perfil de interacción estructural" del complejo que se puede utilizar para organizar, analizar y visualizar la gran cantidad de información codificada en los complejos ligando-receptor (Deng, Chuaqui, y Singh 2004).

- En este caso se tomaron 28 residuos correspondientes al bolsillo de unión del receptor y se calcularon siete tipos de interacciones para cada uno de ellos. (i) Interacciones débiles (del tipo dispersión de London); (ii) Nube π (cara a cara); (iii) Nube π (borde a cara); (iv) Puente de hidrógeno (proteína dador/ ligando aceptor); (v) Puente de hidrógeno (proteína aceptor/ ligando dador); (vi) Puente salino (proteína +/ ligando -); y (vii) Puente salino (proteína -/ ligando +).

Además de realizar la caracterización de interacciones por residuo, también se realizaron agrupamientos para ver qué inhibidores explotaban las mismas interacciones en el sitio activo.

Como puede verse, la Figura 3.5 resume la información correspondiente a interacciones interatómicas de los diferentes complejos Cz-inh para cada sub-bolsillo.

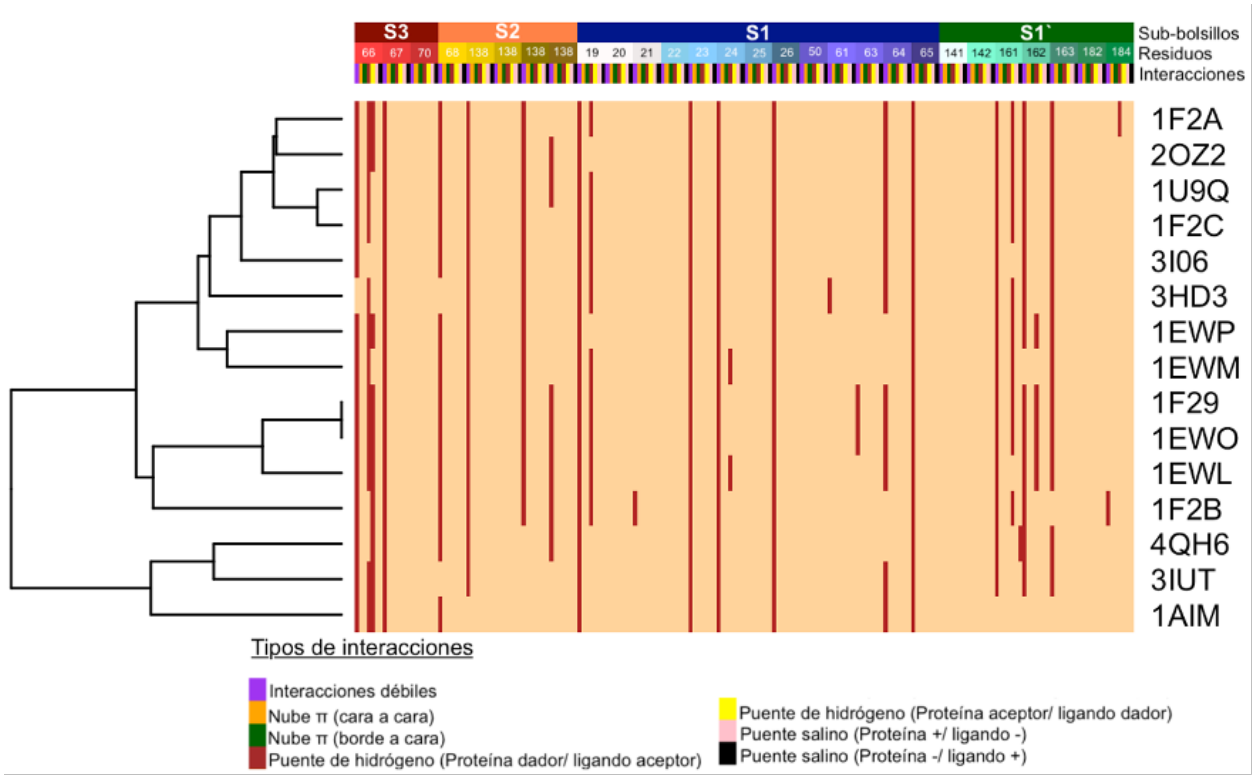


Figura 3.5 Interacciones interatómicas de los diferentes inhibidores en el sitio activo de cada complejo Cz-inh. Cada huella dactilar (*fingerprint*) se representa como una línea en el mapa de calor en el medio de la figura, y solo los *bits* prendidos se muestran como marcas marrones (los mismos forman líneas verticales al estar presentes en la mayoría de los complejos). Del lado izquierdo del mapa de calor se muestra el resultado del agrupamiento jerárquico para cada complejo. La línea superior sobre el mapa de calor indica la ubicación de los residuos en cada sub-bolsillo del sitio catalítico. En la línea media se representa cada residuo del sitio catalítico. La línea inferior contempla dentro de cada residuo los siete *bits* con los diferentes tipos de interacciones analizadas. Esto último se representa mediante siete bloques más pequeños con diferentes colores los cuales se aclaran en la parte inferior de la figura.

A partir del análisis de la figura anterior, no se pueden establecer agrupamientos por familia de compuestos correspondiente a los patrones de interacción obtenidos. Esto puede deberse a que los inhibidores, en su mayoría al ser péptido-miméticos, presentan cadenas laterales similares y solo varían en el *warhead* o cabeza de guerra para el ataque nucleofílico (Barbosa da Silva, do Nascimento Pereira, y Ferreira 2016)

A su vez, se puede observar que los residuos 66 correspondiente al sub-bolsillo S3 y el residuo 161 correspondiente al sub-bolsillo S1´ poseen interacciones tipo “Puente de Hidrógeno” en casi todos los complejos analizados en este trabajo de tesis.

Por otro lado, se puede ver que la mayoría de los inhibidores de la familia vinilsulfona, explotan una interacción puente de hidrógeno más con el residuo GLN-19.

Al estudiar estas interacciones desde el punto de vista estructural se puede ver que la cadena principal de GLY-66 funciona como dador y aceptor de enlaces puentes de hidrógeno, mientras que el oxígeno carboxílico de la cadena principal de ASP-161 funciona como aceptor del enlace de hidrógeno en el bolsillo S1´como lo muestra la Figura 3.6. Estos resultados coinciden con trabajos previos de diferentes autores (Ferreira et al. 2010; Santos et al. 2019).

En el caso de la interacción con GLN-19, el hidrógeno de su cadena lateral actúa como dador del enlace de hidrógeno con un oxígeno de los inhibidores en la posición P1`. En este caso, diferentes autores han visto que esta interacción es explotada por otras familias de inhibidores como tiosemicarbazonas e hidrazonas, siendo parte de una de las más importantes a nivel del sub-bolsillo S1´ (dos Santos Filho et al. 2009; Caputto et al. 2011).

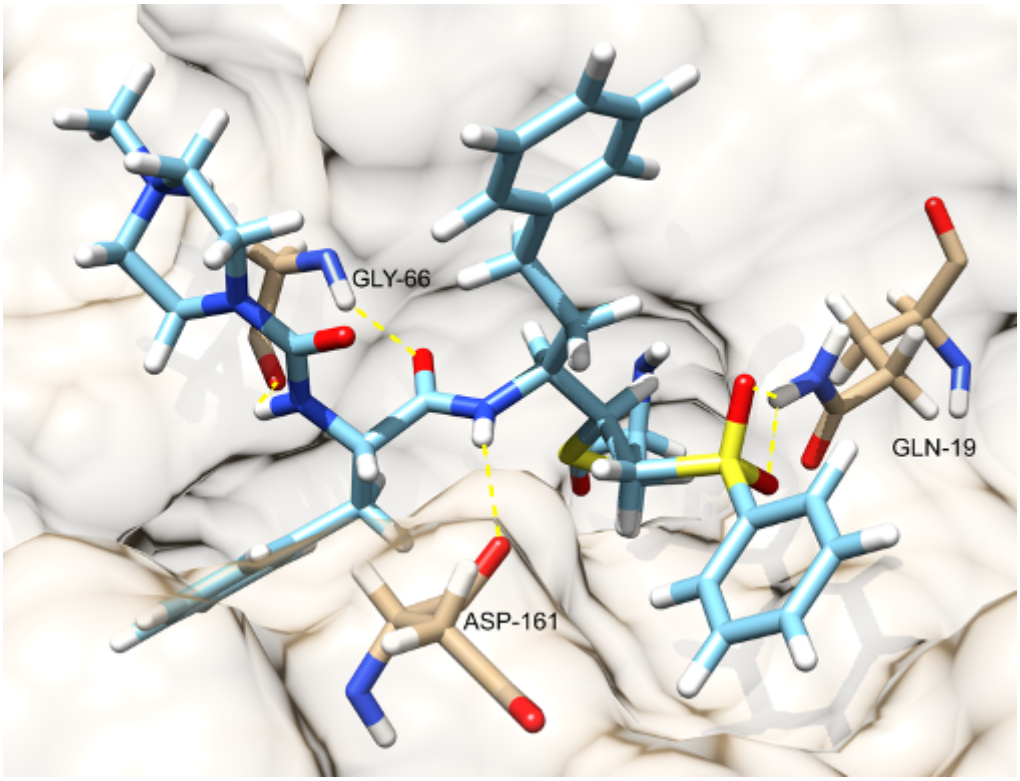


Figura 3.6 Interacciones importantes identificadas en el análisis por huellas dactilares de interacción en el complejo Cz-inh con el ligando (en color cían) correspondiente al cristal 2OZ2. Las interacciones entre pares de átomos se marcan con líneas amarillas punteadas.

Por otro lado, al analizar las distancias interatómicas para estas 4 interacciones, se puede ver que las mismas se mantienen a lo largo del tiempo de simulación en los diferentes complejos. Esto nos habla de la importancia de estas para acomodar a los diferentes inhibidores en el sitio activo, constituyendo verdaderos puntos calientes de interacciones en el sitio catalítico. La Figura 3.7 muestra las distancias interatómicas promedio para cada interacción a lo largo del tiempo de simulación.

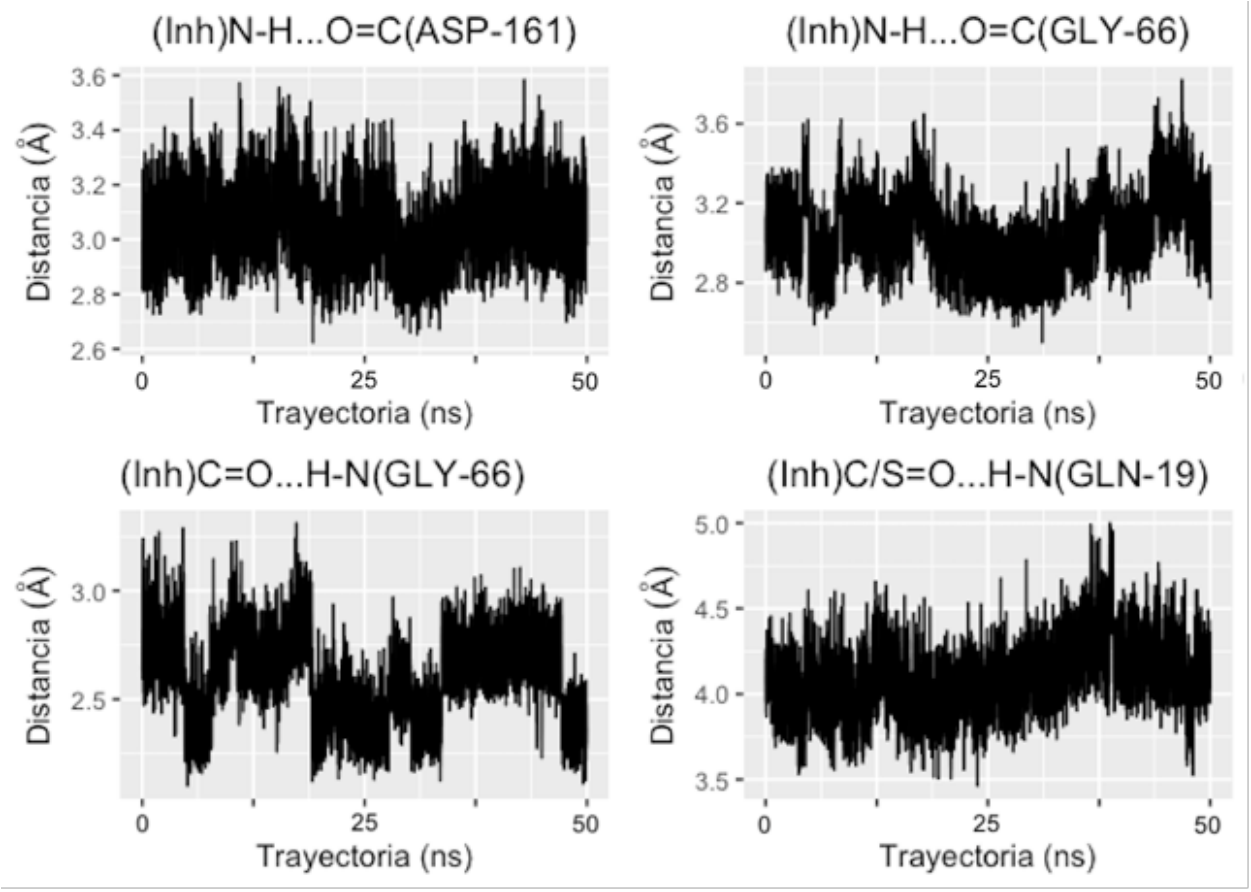


Figura 3.7 Distancia interatómica media a lo largo de las trayectorias de dinámica molecular para las interacciones relevantes obtenidas mediante el análisis de huellas dactilares de interacción. C/S en el gráfico inferior derecho hace referencia a que en esa posición puede existir un átomo de carbono o uno de azufre unido al oxígeno interactuante.

De este modo, estas 4 interacciones en su conjunto forman un primer acercamiento entre las interacciones que tienen mayor relevancia en el sitio catalítico de la cruzipaína (puntos calientes de interacción). Para indagar aún más se procede al segundo análisis de interacciones.

3.3.4.2 Analisís QTAIM

Si bien el análisis basado en el estudio de patrones de huellas dactilares de interacción (SIFt) nos da un primer acercamiento a lo que estaría ocurriendo en el sitio activo de nuestros complejos proteína-ligando de interés, la aplicación de la metodología QTAIM en entornos biomoleculares permite detectar interacciones no direccionales, por ejemplo, aquellas que involucran electrones π en anillos aromáticos, entre otros contactos débiles e inusuales que de otro modo se perderían en un análisis meramente geométrico de las interacciones (Angelina et al. 2014).

A su vez, el atractivo de la teoría QTAIM radica en que permite separar la interacción total en contribuciones por átomo o grupos de átomos lo cual la hace particularmente útil en Química Medicinal para el análisis, diseño y optimización de compuestos líderes.

Por lo antes mencionado y al igual que en estudio de *fingerprints* se partió de las estructuras de mínimo de energía potencial para el cálculo de la densidad de carga y el posterior análisis de interacciones bajo la teoría QTAIM. En particular nos interesaba: a) corroborar la presencia de las interacciones encontradas anteriormente; b) encontrar cuales son las interacciones mínimas necesarias para que ocurra la inhibición de Cz (farmacóforo) y c) medir la fortaleza de las interacciones y establecer la fortaleza de anclaje de los inhibidores en el sitio activo.

Como descriptores de estas interacciones se utilizaron dos elementos que se derivan de la topología de la densidad electrónica en el contexto de la Teoría Cuántica de Átomos en Moléculas (QTAIM): el punto crítico de enlace (Bond Critical Points, BCPs) y los caminos de enlace (Bond Paths, BPs).

La Figura 3.8 muestra un grafo molecular de la densidad electrónica, donde se puede ver el conjunto de interacciones obtenidas mediante la teoría QTAIM.

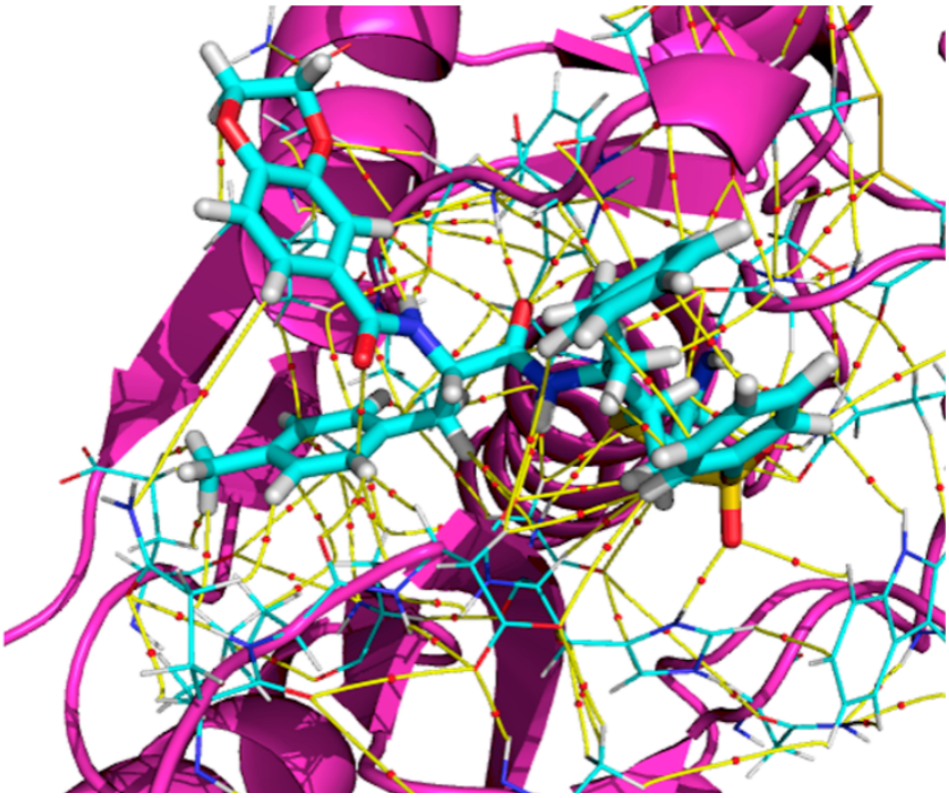


Figura 3.8 Red de interacciones en la estructura del complejo Cz-inhibidor. Los elementos topológicos de densidad de carga que describen las interacciones no covalentes se representan con pequeños círculos rojos (BCP) y líneas amarillas que conectan cada BCP con ambos átomos que interactúan (BP). Se consideran interacciones tanto intermoleculares (Cz-Inh) como intramoleculares (Cz-Cz e Inh-Inh). La estructura de la proteína se representa tridimensionalmente como un conjunto de hélices y láminas.

Debido a la intrincada red de caminos de enlace que se establecen en complejos biomoleculares, extraer información de estos “grafos biomoleculares” solamente por inspección visual no es tarea trivial; resulta mucho más conveniente entonces recurrir a otras herramientas para su análisis.

3.3.4.2.1 Análisis de mapas de calor (*heatmaps)* de basados en valores de densidad electrónica

Con la finalidad de encontrar patrones de interacción por sub-bolsillos en el sitio de unión se realizó el análisis de los *heatmaps.* Este análisis se basa en valores de densidad electrónica y el mismo se realizó segmentando a los ligandos según los sub-bolsillos que ocupaban. Es decir, a aquellas porciones del ligando que interactúan mayoritariamente con S1 se lo llamó P1. Esto no necesariamente indica que el segmento P1 solo interactúe con residuos del sub-bolsillo S1 ya

que por proximidad también podría interactuar con cadenas laterales de residuos correspondientes a otros sub-bolsillos como se verá más adelante.

Segmento P1 / Sub-bolsillo S1

La Figura 3.9 muestra el mapa de calor basado en valores de densidad electrónica para las interacciones interatómicas que tiene lugar en el entre el segmento del inhibidor que ocupa la cavidad correspondiente al sub-bolsillo S1.

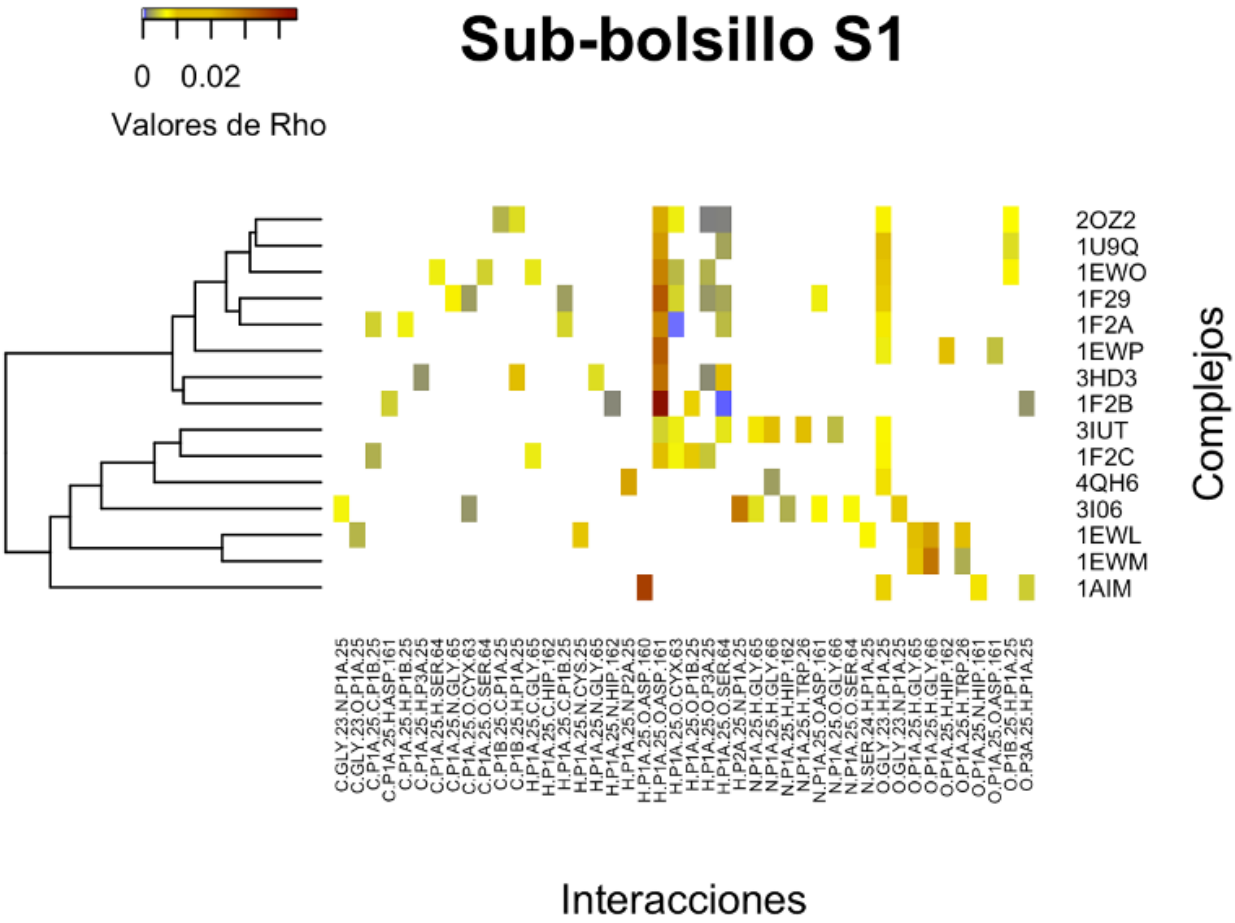


Figura 3.9 Mapa de calor en función de los valores de densidad de carga para las diferentes interacciones protagonizadas por segmentos P1 de los inhibidores y residuos del sitio activo, mayoritariamente aquellos del sub-bolsillo S1.

Se puede observar como las vinilsulfonas se agrupan en un conjunto en la parte superior del agrupamiento, mientras que el segundo grupo de *warhead* más numeroso (las metilcetonas) se agrupan en la parte inferior. Esto podría deberse a la similitud en estructura en el P1 de las estructuras cristalinas derivadas de estos dos grupos numerosos.

Las interacciones que adquieren mayor relevancia en este sub-bolsillo catalítico son: (P1)H…O(ASP-161), (P1)H…O(CYX-63), (P1)H…O(SER-64), (P1)H…O(GLY-23). Donde se ve que los oxígenos correspondientes a la cadena principal de la SER 64 y de la CYX 63 forman puentes de hidrógeno con los hidrógenos no polares del anillo fenilo. Por otro lado, la interacción más importante (en términos de valores de densidad de carga) se da entre el hidrógeno del grupo amida y el oxígeno carboxílico del ASP 161. Esta última ya fue caracterizada anteriormente mediante el análisis de huellas dactilares de interacción en la sección 3.3.4.1.

En relación con lo anterior, se puede establecer que las interacciones con mayor fortaleza (valores más altos de ρ) se dan en aquellos inhibidores que poseen un grupo fenilo en el P1. Este hallazgo coincide con resultados previos donde se sostiene que algunos inhibidores con una cadena lateral de homofenilalanina en P1, tienden a ser más resistentes al metabolismo, mostrando una alta potencia en ensayos experimentales (Huang, Lee, y Ellman 2002).

Segmento P1` / Sub-bolsillo S1`

La Figura 3.10 muestra el mapa de calor basado en valores de densidad electrónica para las interacciones interatómicas que tiene lugar en el entre el segmento del inhibidor que ocupa la cavidad correspondiente al sub-bolsillo S1`.

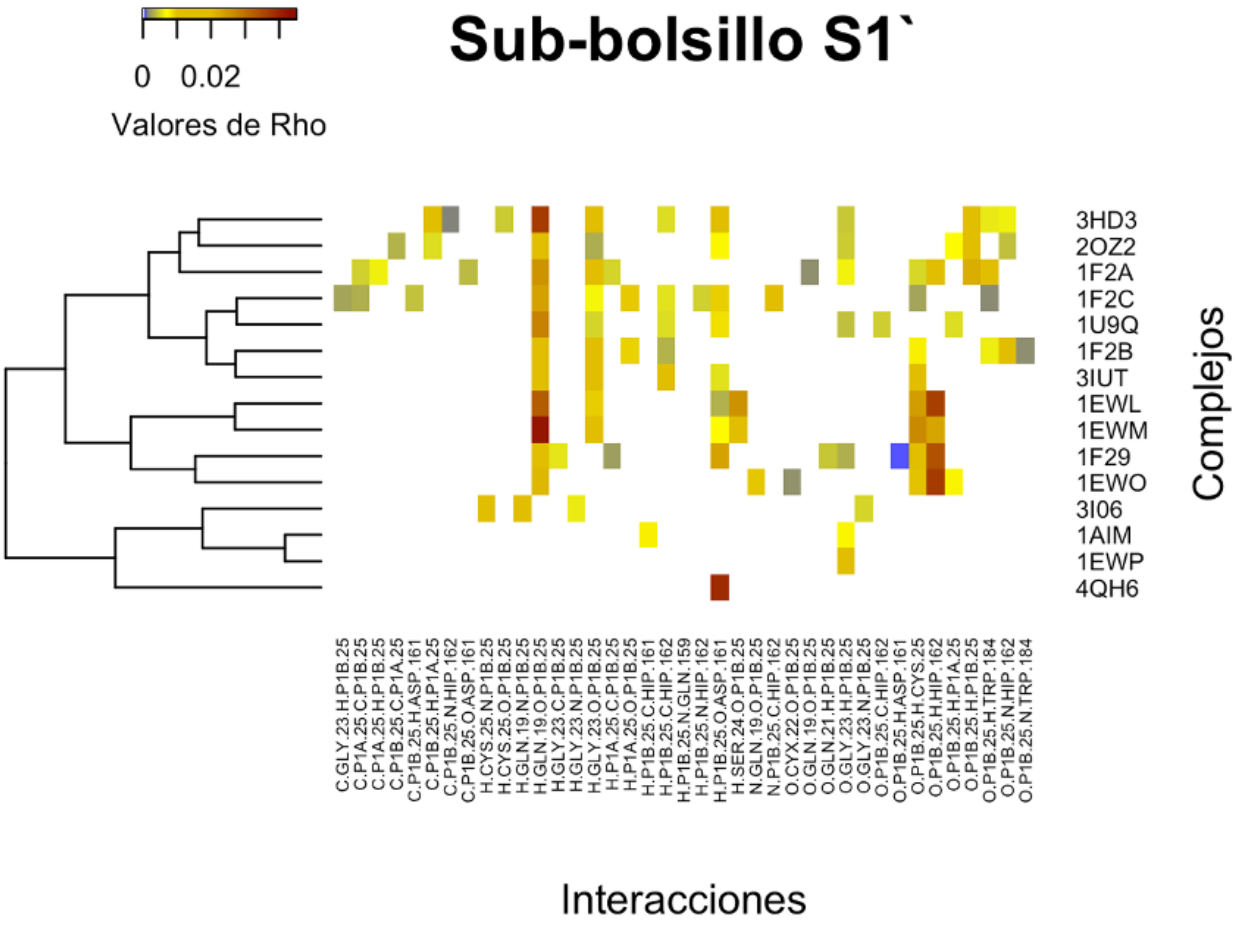


Figura 3.10 Mapa de calor en función de los valores de densidad de carga para las diferentes interacciones protagonizadas por segmentos P1` de los inhibidores y residuos del sitio activo, mayoritariamente aquellos del sub-bolsillo S1`.

Las interacciones más importantes en este sub-bolsillo son (P1`)O…H(GLN-19), (P1`)O…H(GLY-23), (P1`)H…C(HIP-162).

El grupo sulfonilo de las vinilsulfonas así como también el carbonilo de los demás inhibidores juegan un rol fundamental en la estabilidad de los ligandos en este sub-bolsillo. Los mismos forman interacciones con el/los hidrógenos del grupo amida de GLN-19, con el hidrógeno de la cadena lateral de la GLY-23.

Por otro lado, el hidrógeno del grupo fenilo de P1` interactúa con el N del anillo imidazol de la histidina.

Cabe destacar que el puente salino formado entre los derivados nitrados y la cadena lateral del ASP-161 es de gran importancia.

Segmento P2 / Sub-bolsillo S2

La Figura 3.11 muestra el mapa de calor basado en valores de densidad electrónica para las interacciones interatómicas que tiene lugar en el entre el segmento del inhibidor que ocupa la cavidad correspondiente al sub-bolsillo S2.

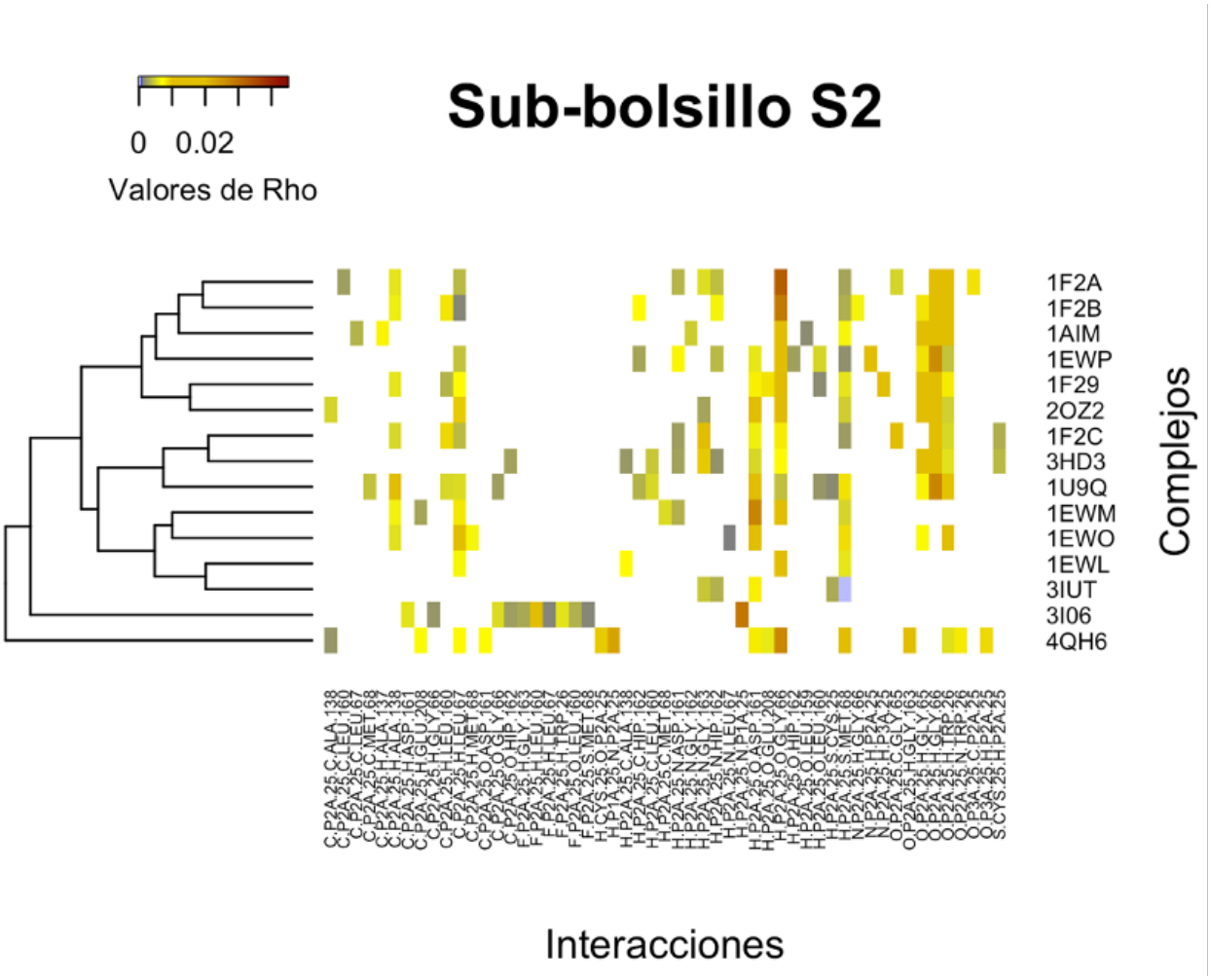


Figura 3.11 Mapa de calor en función de los valores de densidad de carga para las diferentes interacciones protagonizadas por segmentos P2 de los inhibidores y residuos del sitio activo, mayoritariamente aquellos del sub-bolsillo S2.

El bolsillo S2, es uno de los más grandes y presenta características de naturaleza predominantemente hidrofóbica. A su vez, los segmentos de inhibidores en P2 generalmente cuentan con la presencia de L-Phe en dicha la subunidad y, en menor grado, L-Leu, L-Val y L-Met, favoreciendo la actividad (Huang, Lee, y Ellman 2002; Choe et al. 2005). Sin embargo, es

importante señalar que la Cz también acepta grupos cargados positivamente en P2 (Gillmor, Craik, y Fletterick 1997).

En concordancia con lo anterior, las interacciones más importantes en el bolsillo S2 son las siguientes: (P2)H…S(Met-68), (P2)H…O(Gly-66), (P2)C…H(Leu-67), (P2)C…H(Ala-138).

Por otro lado, algo llamativo es la ausencia de interacciones con la cadena lateral de Glu 208, ya que la misma se encuentra desplazada, acercando el carbono beta al sub-bolsillo, lo cual da a pensar que el efecto entrópico de la mayoría de los ligandos con anillos aromáticos es mucho mayor que algún enlace de H con la cadena lateral involucrada. En los únicos cristales donde se ve esa interacción son el 1F29 y el 4QH6.

Segmento P3 / Sub-bolsillo S3

La Figura 3.12 muestra el mapa de calor basado en valores de densidad electrónica para las interacciones interatómicas que tiene lugar en el entre el segmento del inhibidor que ocupa la cavidad correspondiente al sub-bolsillo S3

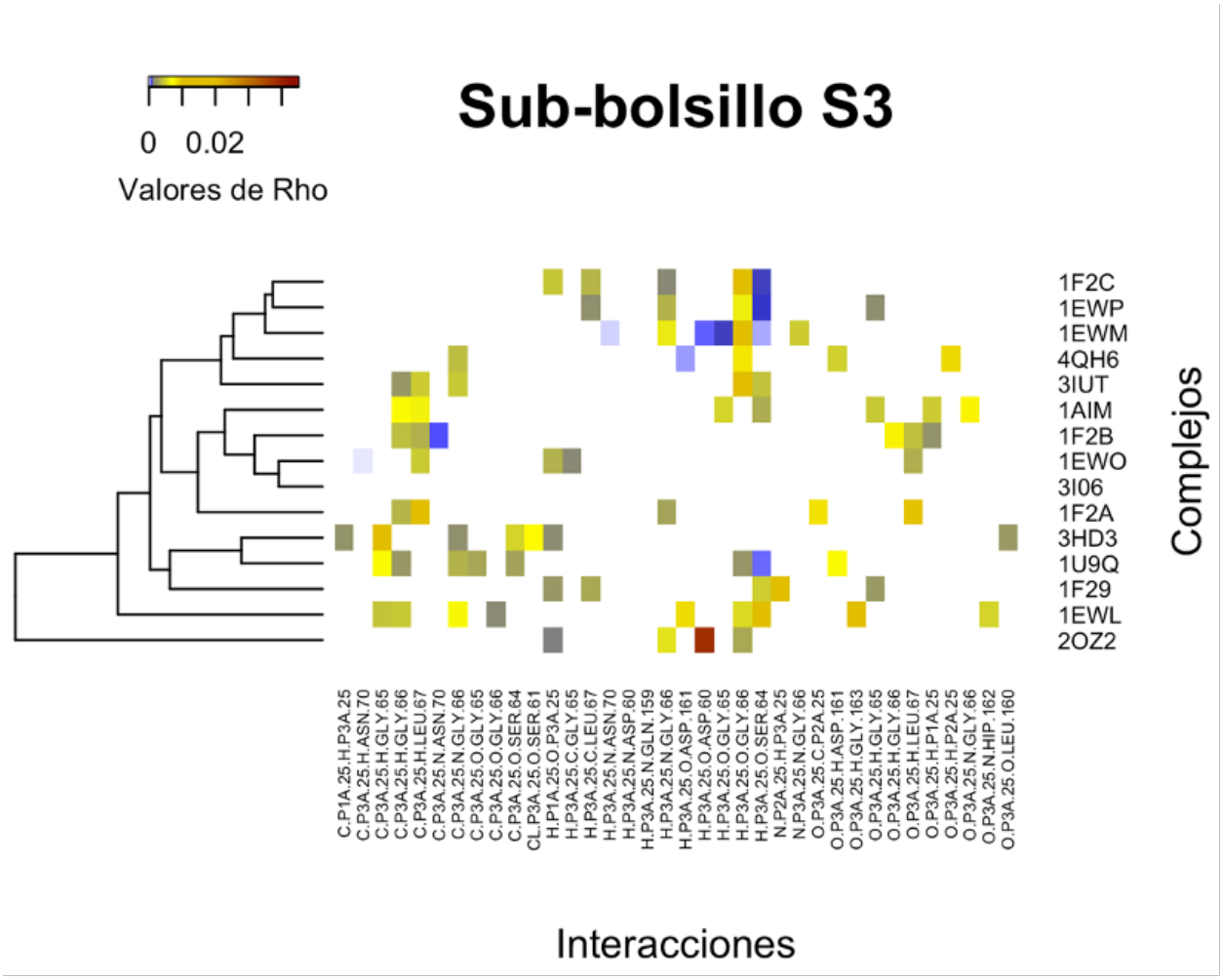


Figura 3.12 Mapa de calor en función de los valores de densidad de carga para las diferentes interacciones protagonizadas por segmentos P3 de los inhibidores y residuos del sitio activo, mayoritariamente aquellos del sub-bolsillo S3.

Si bien en este bolsillo no se observan interacciones generales de manera tan notable como en los demás, se puede establecer que los residuos SER-64, GLY-65 y GLY-66 juegan un rol preponderante en la estabilización del segmento P3. Las interacciones más importantes en este bolsillo son: (P3)H…O(GLY-66), (P3)H…O(SER-64).

Se puede apreciar que las principales interacciones se dan con el backbone de las glicinas 66 y 65 y con la cadena lateral del residuo 64. Cabe destacar el puente salino que se forma en un solo caso entre ASP-60 y N metilpiperazina del inhibidor 2OZ2, constituyendo la interacción de mayor fortaleza en el bolsillo analizado.

Por último, algo notable a considerar es que el residuo carboxibencilo (Cbz) en P3 está presente en varios potentes inhibidores irreversibles conocidos de la cruzipaína (varios analizados en este

trabajo de tesis), y su eliminación provoca una pérdida drástica de actividad, lo que hace pensar que este tipo de estructuras son beneficiosas para las interacciones en S3. De manera similar, la sustitución de Cbz por un acetilo (Choe et al. 2005) o por una morfolina no presentaron resultados beneficiosos en estudios de inhibición, mientras que la sustitución por 3-piridinilo aumentó la actividad (Huang, Lee, y Ellman 2002).

3.3.4.3 Puntos calientes e interacciones mínimas

Los análisis realizados en las secciones 3.3.4.1 y 3.3.4.2 proporcionaron información relevante para la identificación de "puntos calientes", es decir, regiones en la superficie de la proteína que proporcionan la mayor parte de la afinidad de unión (Defelipe et al. 2018).

A su vez, se pudieron establecer cuales son las interacciones mínimas que aparecen en la mayoría de los ligandos estudiados en este trabajo de tesis. En concordancia, el conjunto de estas interacciones constituyen un modelo farmacofórico. Según lo establece la IUPAC, un farmacóforo es “un conjunto de características estéricas y electrónicas que son necesarias para asegurar la óptima interacción supramolecular con un blanco biológico específico y para activar (o bloquear) su respuesta biológica” (Wermuth et al. 1998). Por otro lado, en el diseño de fármacos asistido por computadoras, los modelos farmacofóricos se utilizan para definir rasgos esenciales de una o más moléculas con la misma actividad biológica. A partir de los anterior, una base de datos de diversas familias de compuestos puede ser explorada a fin de identificar rasgos similares en relación a los tipos de interacciones observadas (Mukherjee et al. 2008).

Es así como se pudieron extraer las siguientes interacciones como marcadores de puntos farmacofóricos:

Tabla 3.2 Tabla con el detalle de las interacciones más importantes halladas a través de los análisis anteriormente realizados. C/S en la tabla hace referencia a que en esa posición puede existir un átomo de carbono o uno de azufre unido al oxígeno interactuante. Las filas con sombreado gris indican las interacciones con mayores valores de ρ.

Interacción	Sub-bolsillos	Detalle	Valores de densidad para los complejos ($\underline{\rho}$)
1	S1	(P1)N-H…O=C(ASP-161)	0.0176
2	S1	(P1)C-H…O=C(GLY-23)	0.0059
3	S1`	(P1`)C/S=O…H-N(GLN-19)	0.0180
4	S1`	(P1`)C/S=O…H-N(HIS-162)	0.0095
5	S2	(P2)C…H-C(LEU-67)	0.0047
6	S2	(P2)N-H…O=C(GLY-66)	0.0140
7	S2	(P2)C-H…S(MET-68)	0.0046
8	S2	(P2)C=O…H-C(GLY-65)	0.0052
9	S2	(P2)C=O…H-N(GLY-66)	0.0109
10	S2	(P2)C=O…H-C(TRP-26)	0.0065
11	S3	(P3)C-H…O=C(GLY-66)	0.0043

Como puede verse en la tabla anterior las interacciones 1, 3, 6 y 9 fueron reconocidas por ambos tipos de análisis de interacciones aplicados anteriormente, lo que indica la relevancia que tendrían para la estabilización de los complejos Cz-inh. A su vez, dichas interacciones poseen elevados valores de ρ en los puntos críticos de enlace.

Por otro lado, la figura 3.13 muestra las distancias interatómicas promedio para cada interacción a lo largo del tiempo de simulación.

A partir del análisis de dichas distancias a lo largo de la dinámica molecular se puede evidenciar una “competencia” entre las interacciones 3 y 4. Esto se debe a que las cadenas laterales de los residuos GLN-19 e HIS-162 interactúan con el mismo grupo de átomos del segmento P1` del inhibidor. Este hallazgo es importante ya que dichas interacciones implicarían un rearreglo estructural de la triada catalítica en el bolsillo de unión.

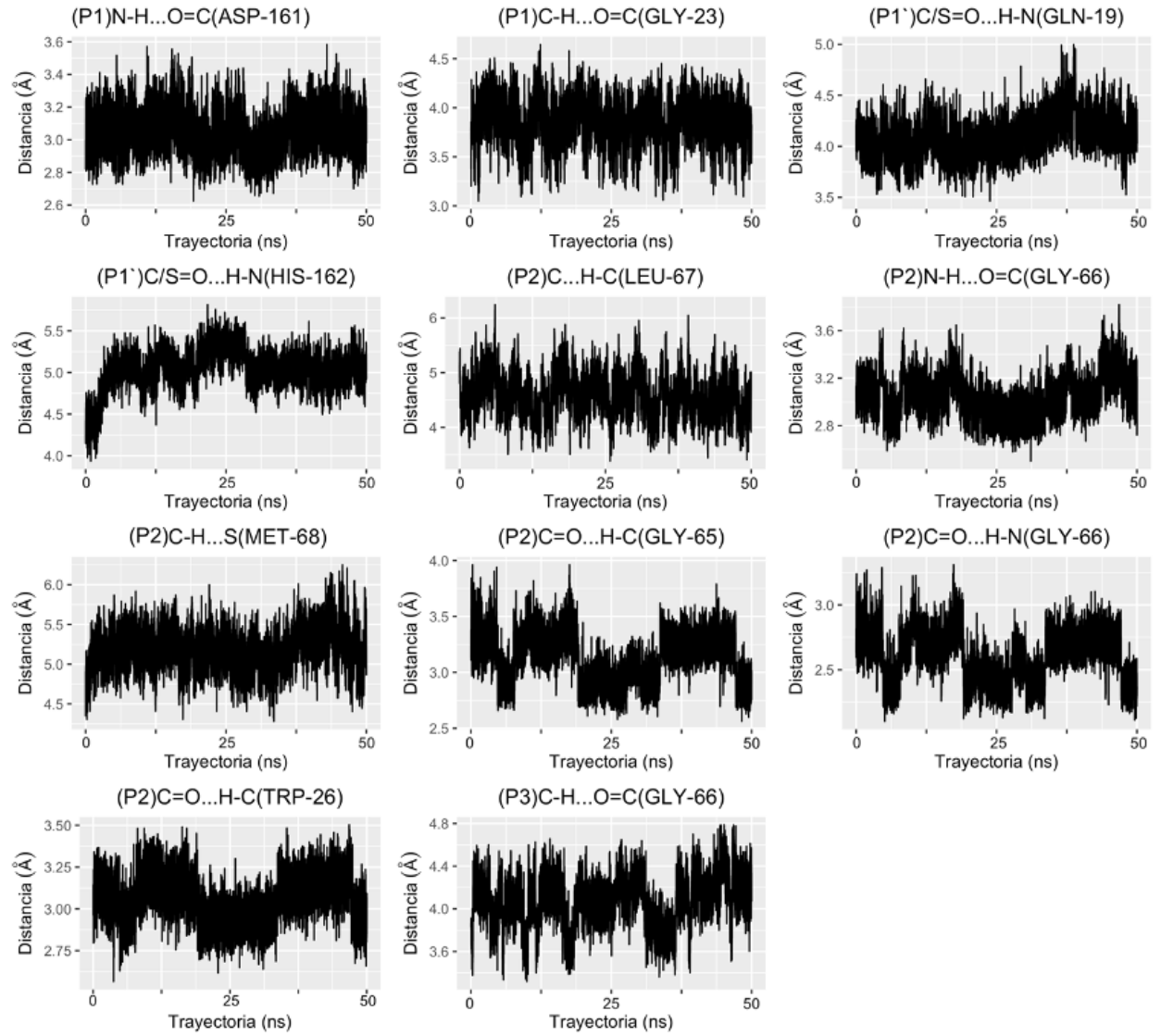


Figura 3.13 Distancia interatómica media a lo largo de las trayectorias de dinámica molecular para las interacciones relevantes obtenidas mediante los análisis de huellas dactilares de interacción y QTAIM. C/S en algunos gráficos hace referencia a que en esa posición puede existir un átomo de carbono o uno de azufre unido al oxígeno interactuante.

3.4 Conclusiones

En este capítulo se realizó un primer acercamiento al estudio de la cruzipaína del *T. cruzi* utilizando un enfoque estructural. A su vez se compiló una biblioteca de inhibidores conocidos y se analizaron de forma dinámica las interacciones intermoleculares que tienen lugar en el sitio catalítico.

A partir del análisis de la secuencia de aminoácidos y emparejamiento con diversas cisteíno-proteasas de la superfamilia de la papaína se pudo obtener el porcentaje de identidad con de Cz con otras proteasas. Se pudo ver que Cz posee un elevado grado de identidad (71%) con la enzima rodesaína del *T. brucei*, lo cuál se pudo evidenciar estructuralmente, obteniendo una superposición casi total, a partir sus estructuras terciarias.

Por otro lado, se compiló una biblioteca de ligandos conocidos que se unen covalentemente a la Cz en el sitio catalítico. De estos inhibidores, la gran mayoría fueron vinilsulfonas, seguidos en menor número por derivados de cetonas, ésteres y nitrilos. Desafortunadamente, no todos los inhibidores seleccionados contaban con los valores correspondientes de inhibición (K_i o IC50).

A su vez, los esfuerzos de este capítulo fueron centrados en entender los patrones de interacción frecuentes en la mayoría de los inhibidores compilados. Primeramente se aplicó la teoría de "Huellas dactilares de interacción" (*SIFt*), un método de relativo bajo costo computacional que permite analizar a partir de distancias interatómicas átomos interactuantes. En segundo lugar se aplicó la Teoría Cuántica de Átomos en Moléculas (*QTAIM*) que nos permitió detectar otras interacciones, que se escaparían de un simple análisis geométrico y de distancia, además de otorgarnos los valores de densidad electrónica para las interacciones en los diferentes complejos.

A partir de sendos análisis pudimos caracterizar diversas interacciones obteniendo "puntos calientes de interacción" en la cavidad catalítica de la enzima. A su vez, se pudieron identificar cuatro interacciones relevantes que serán tenidas en cuenta a la hora de parametrizar el campo de fuerza de los algoritmos de docking en un cribado virtual retrospectivo.

Este capítulo proporciona una ventana hacia la comprensión del comportamiento dinámico de los complejos Cz-inh estudiados. Además nos informa sobre las interacciones que tienen lugar en el sitio catalítico de cruzipaína, proporcionando información relevante para futuras campañas de cribado virtual retrospectivo.

Referencias del capítulo 3

Al-Lazikani, B, A Gaulton, G Paolini, J Lanfear, J Overington, and A Hopkins. 2007. “The Molecular Basis of Predicting Druggability.” In *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*, edited by G Wess, SL Schreiber, and TM Kapoor, 804–23. Weinheim: Wiley-VCH.

Angelina, Emilio L., Sebastiàn A. Andujar, Rodrigo D. Tosso, Ricardo D. Enriz, and Ǹlida M. Peruchena. 2014. “Non-Covalent Interactions in Receptor-Ligand Complexes. A Study Based on the Electron Charge Density.” *Journal of Physical Organic Chemistry* 27 (2): 128–34. https://doi.org/10.1002/poc.3250.

Bakan, Ahmet, Neysa Nevins, Ami S. Lakdawala, and Ivet Bahar. 2012. “Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules.” *Journal of Chemical Theory and Computation* 8 (7): 2435–47. https://doi.org/10.1021/ct300117j.

Barbosa da Silva, Elany, Elfriede Dall, Peter Briza, Hans Brandstetter, and Rafaela Salgado Ferreira. 2019. “Cruzain Structures: Apocruzain and Cruzain Bound to S-Methyl Thiomethanesulfonate and Implications for Drug Design.” *Acta Crystallographica Section F Structural Biology Communications* 75 (6): 419–27. https://doi.org/10.1107/S2053230X19006320.

Bryant, Clifford, Iain D Kerr, Moumita Debnath, Kenny K H Ang, Joseline Ratnam, Rafaela S Ferreira, Priyadarshini Jaishankar, et al. 2009. “Bioorganic & Medicinal Chemistry Letters Novel Non-Peptidic Vinylsulfones Targeting the S2 and S3 Subsites of Parasite Cysteine Proteases.” *Bioorganic & Medicinal Chemistry Letters* 19 (21): 6218–21. https://doi.org/10.1016/j.bmcl.2009.08.098.

Burley, Stephen K., Helen M. Berman, Charmi Bhikadiya, Chunxiao Bi, Li Chen, Luigi Di Costanzo, Cole Christie, et al. 2019. “RCSB Protein Data Bank: Biological Macromolecular Structures Enabling Research and Education in Fundamental Biology, Biomedicine, Biotechnology and Energy.” Nucleic Acids Research. 2019. https://doi.org/10.1093/nar/gky1004.

Caputto, María E., Lucas E. Fabian, Diego Benítez, Alicia Merlino, Natalia Ríos, Hugo

Cerecetto, Graciela Y. Moltrasio, Albertina G. Moglioni, Mercedes González, and Liliana M. Finkielsztein. 2011. “Thiosemicarbazones Derived from 1-Indanones as New Anti-Trypanosoma Cruzi Agents.” *Bioorganic & Medicinal Chemistry* 19 (22): 6818–26. https://doi.org/10.1016/j.bmc.2011.09.037.

Case, D. A., V. Babin, Josh Berryman, R. M. Betz, Q. Cai, D. S. Cerutti, T. E. Cheatham III, et al. 2014. “Amber 14.” https://orbilu.uni.lu/handle/10993/16614.

Case, David A., Thomas E. Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M. Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J. Woods. 2005. “The Amber Biomolecular Simulation Programs.” *Journal of Computational Chemistry*. NIH Public Access. https://doi.org/10.1002/jcc.20290.

Cheng, Alan C., Ryan G. Coleman, Kathleen T. Smyth, Qing Cao, Patricia Soulard, Daniel R. Caffrey, Anna C. Salzberg, and Enoch S. Huang. 2007. “Structure-Based Maximal Affinity Model Predicts Small-Molecule Druggability.” *Nature Biotechnology* 25 (1): 71–75. https://doi.org/10.1038/nbt1273.

Choe, Youngchool, Linda S. Brinen, Mark S. Price, Juan C. Engel, Meinolf Lange, Corinna Grisostomi, Scott G. Weston, et al. 2005. “Development of α-Keto-Based Inhibitors of Cruzain, a Cysteine Protease Implicated in Chagas Disease.” *Bioorganic & Medicinal Chemistry* 13 (6): 2141–56. https://doi.org/10.1016/j.bmc.2004.12.053.

Defelipe, Lucas A., Juan Pablo Arcon, Carlos P. Modenutti, Marcelo A. Marti, Adrián G. Turjanski, and Xavier Barril. 2018. “Solvents to Fragments to Drugs: MD Applications in Drug Design.” *Molecules* 23 (12). https://doi.org/10.3390/molecules23123269.

Deng, Zhan, Claudio Chuaqui, and Juswinder Singh. 2004. “Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions.” *Journal of Medicinal Chemistry* 47 (2): 337–44. https://doi.org/10.1021/jm030331x.

Durrant, Jacob D., Henrik Keränen, Benjamin A. Wilson, and J. Andrew McCammon. 2010. “Computational Identification of Uncharacterized Cruzain Binding Sites.” *PLoS Neglected Tropical Diseases* 4 (5). https://doi.org/10.1371/journal.pntd.0000676.

Feig, Michael. 2016. “Local Protein Structure Refinement via Molecular Dynamics Simulations

with LocPREFMD." *Journal of Chemical Information and Modeling* 56 (7): 1304–12. https://doi.org/10.1021/acs.jcim.6b00222.

Ferreira, Rafaela S., Anton Simeonov, Ajit Jadhav, Oliv Eidam, Bryan T. Mott, Michael J. Keiser, James H. McKerrow, David J. Maloney, John J. Irwin, and Brian K. Shoichet. 2010. "Complementarity between a Docking and a High-Throughput Screen in Discovering New Cruzain Inhibitors." *Journal of Medicinal Chemistry* 53 (13): 4891–4905. https://doi.org/10.1021/jm100488w.

Frisch, M. J., G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, et al. 2016. "Gaussian 09, Revision A.02." Gaussian, Inc.

Gillmor,', Sarah A, Charles S Craik, and Robert J Fletterick3. 1997. "Structural Determinants of Specificity in the Cysteine Protease Cruzain." *Protein Science*. Vol. 6. Cambridge University Press.

Gillmor, Sarah A, Charles S Craik, and Robert J Fletterick. 1997. "Structural Determinants of Specificity in the Cysteine Protease Cruzain." *Protein Science* 6 (8): 1603–11. https://doi.org/10.1002/pro.5560060801.

Halgren, Thomas A. 2009. "Identifying and Characterizing Binding Sites and Assessing Druggability." *Journal of Chemical Information and Modeling* 49 (2): 377–89. https://doi.org/10.1021/ci800324m.

Huang, Lily, Alice Lee, and Jonathan A. Ellman. 2002. "Identification of Potent and Selective Mechanism-Based Inhibitors of the Cysteine Protease Cruzain Using Solid-Phase Parallel Synthesis." *Journal of Medicinal Chemistry* 45 (3): 676–84. https://doi.org/10.1021/jm010333m.

Jhoti, Harren, and Andrew R. Leach. 2007. *Structure-Based Drug Discovery*. *Structure-Based Drug Discovery*. Springer Netherlands. https://doi.org/10.1007/1-4020-4407-0.

Leis, Simon, Sebastian Schneider, and Martin Zacharias. 2010. "In Silico Prediction of Binding Sites on Proteins." *Current Medicinal Chemistry* 17 (15): 1550–62. https://doi.org/10.2174/092986710790979944.

Lu, Tian, and Feiwu Chen. 2012. "Multiwfn: A Multifunctional Wavefunction Analyzer." *Journal of Computational Chemistry* 33 (5): 580–92. https://doi.org/10.1002/jcc.22885.

Luchi, Adriano M., Roxana N. Villafañe, J. Leonardo Gómez Chávez, M. Lucrecia Bogado, Emilio L. Angelina, and Nelida M. Peruchena. 2019. “Combining Charge Density Analysis with Machine Learning Tools to Investigate the Cruzain Inhibition Mechanism.” *ACS Omega* 4 (22): 19582–94. https://doi.org/10.1021/acsomega.9b01934.

Maier, James A., Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. 2015. “Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB.” *Journal of Chemical Theory and Computation* 11 (8): 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

Mauser, H., and W. Guba. 2008. “Recent Developments in de Novo Design and Scaffold Hopping.” *Current Opinion in Drug Discovery and Development*. https://europepmc.org/article/med/18428090.

McPherson, Alexander, and Jose A. Gavira. 2014. “Introduction to Protein Crystallization.” *Acta Crystallographica Section F:Structural Biology Communications*. International Union of Crystallography. https://doi.org/10.1107/S2053230X13033141.

Mirjalili, Vahid, and Michael Feig. 2013. “Protein Structure Refinement through Structure Selection and Averaging from Molecular Dynamics Ensembles.” *Journal of Chemical Theory and Computation* 9 (2): 1294–1303. https://doi.org/10.1021/ct300962x.

Mukherjee, Subhendu, Shuchi Nagar, Sanchita Mullick, Arup Mukherjee, and Achintya Saha. 2008. “Pharmacophore Mapping of Arylbenzothiophene Derivatives for MCF Cell Inhibition Using Classical and 3D Space Modeling Approaches.” *Journal of Molecular Graphics and Modelling* 26 (5): 884–92. https://doi.org/10.1016/j.jmgm.2007.06.003.

Müller, S., Cerdan, R., & Radulescu, O. 2016. *A Comprehensive Analysis of Parasite Biology: From Metabolism to Drug Discovery. A Comprehensive Analysis of Parasite Biology: From Metabolism to Drug Discovery*. Wiley-VCH Verlag. https://doi.org/10.1002/9783527694082.

Nayal, Murad, and Barry Honig. 2006. “On the Nature of Cavities on Protein Surfaces: Application to the Identification of Drug-Binding Sites.” *Proteins: Structure, Function, and Bioinformatics* 63 (4): 892–906. https://doi.org/10.1002/prot.20897.

O’Boyle, Noel M., Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. 2011. “Open Babel: An Open Chemical Toolbox.” *Journal of*

Cheminformatics 3 (10): 33. https://doi.org/10.1186/1758-2946-3-33.

Pettersen, Eric F., Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. 2004. “UCSF Chimera - A Visualization System for Exploratory Research and Analysis.” *Journal of Computational Chemistry* 25 (13): 1605–12. https://doi.org/10.1002/jcc.20084.

Rodrigues Sartori, Geraldo, Andrei Leitão, Carlos A Montanari, and Charles A Laughton. 2019. “Ligand-Induced Conformational Selection Predicts the 5 Selectivity of Cysteine Protease Inhibitors.” *BioRxiv*, August, 744953. https://doi.org/10.1101/744953.

Santos Filho, José Mauricio dos, Ana Cristina Lima Leite, Boaz Galdino de Oliveira, Diogo Rodrigo Magalhães Moreira, Milena S. Lima, Milena Botelho Pereira Soares, and Lucia Fernanda C.C. Leite. 2009. “Design, Synthesis and Cruzain Docking of 3-(4-Substituted-Aryl)-1,2,4-Oxadiazole-N-Acylhydrazones as Anti-Trypanosoma Cruzi Agents.” *Bioorganic and Medicinal Chemistry* 17 (18): 6682–91. https://doi.org/10.1016/j.bmc.2009.07.068.

Santos, Lucianna H., Birgit J. Waldner, Julian E. Fuchs, Glaécia A.N. Pereira, Klaus R. Liedl, Ernesto R. Caffarena, and Rafaela S. Ferreira. 2019. “Understanding Structure-Activity Relationships for Trypanosomal Cysteine Protease Inhibitors by Simulations and Free Energy Calculations.” *Journal of Chemical Information and Modeling* 59 (1): 137–48. https://doi.org/10.1021/acs.jcim.8b00557.

Seco, Jesus, F. Javier Luque, and Xavier Barril. 2009. “Binding Site Detection and Druggability Index from First Principles.” *Journal of Medicinal Chemistry* 52 (8): 2363–71. https://doi.org/10.1021/jm801385d.

Sievers, Fabian, and Desmond G. Higgins. 2014. “Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences.” *Methods in Molecular Biology* 1079: 105–16. https://doi.org/10.1007/978-1-62703-646-7_6.

Silva, Elany Barbosa da, Gla??cia Aparecida do Nascimento Pereira, and Rafaela Salgado Ferreira. 2016. “Trypanosomal Cysteine Peptidases: Target Validation and Drug Design Strategies.” *A Comprehensive Analysis of Parasite Biology: From Metabolism to Drug Discovery*, 121–45. https://doi.org/10.1002/9783527694082.ch5.

Stevens, Raymond C. 2004. “Long Live Structural Biology.” *Nature Structural and Molecular*

Biology. Nature Publishing Group. https://doi.org/10.1038/nsmb0404-293.

Turk, Dušan, Gregor Gunčar, Marjetka Podobnik, and Boris Turk. 1998. “Revised Definition of Substrate Binding Sites of Papain-Like Cysteine Proteases.” *Biological Chemistry* 379 (2): 137–47. https://doi.org/10.1515/bchm.1998.379.2.137.

Wang, Junmei. 2013. “Molecular Dynamics Simulations of a Protein Crystal.” *Citation: Wang J* 2: 117. https://doi.org/10.4172/2167-7662.1000e117.

Wang, Junmei, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case. 2004. “Development and Testing of a General Amber Force Field.” *Journal of Computational Chemistry* 25 (9): 1157–74. https://doi.org/10.1002/jcc.20035.

Wermuth, C. G., C. R. Ganellin, P. Lindberg, and L. A. Mitscher. 1998. “Glossary of Terms Used in Medicinal Chemistry (IUPAC Recommendations 1998).” *Pure and Applied Chemistry* 70 (5): 1129–43. https://doi.org/10.1351/pac199870051129.

Wiggers, Helton. 2011. “Integração de Métodos in Silico e in Vitro Para o Planejamento de Inibidores Da Enzima Cruzaína,” 177.

Wiggers, Helton J., Josmar R. Rocha, William B. Fernandes, Renata Sesti-Costa, Zumira A. Carneiro, Juliana Cheleski, Albérico B.F. da Silva, et al. 2013. “Non-Peptidic Cruzain Inhibitors with Trypanocidal Activity Discovered by Virtual Screening and In Vitro Assay.” *PLoS Neglected Tropical Diseases* 7 (8). https://doi.org/10.1371/journal.pntd.0002370.

Willett, Peter, John M. Barnard, and Geoffrey M. Downs. 1998. “Chemical Similarity Searching.” *Journal of Chemical Information and Computer Sciences* 38 (6): 983–96. https://doi.org/10.1021/ci9800211.

Young, David C. 2009. *Computational Drug Design*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9780470451854.

Yuan, Yaxia, Jianfeng Pei, and Luhua Lai. 2013. “Binding Site Detection and Druggability Prediction of Protein Targets for Structure- Based Drug Design.” *Current Pharmaceutical Design* 19 (12): 2326–33. https://doi.org/10.2174/1381612811319120019.

CAPÍTULO IV

“Combinación del análisis de densidad de carga con herramientas de aprendizaje automático para investigar el mecanismo de inhibición de Cz”

4.1 Introducción

En el capítulo anterior se caracterizó estructuralmente a la Cz del *T. cruzi.* A su vez, a partir de diversas metodologías se han podido analizar las interacciones que tienen lugar en el sitio catalítico de la enzima al unirse una serie de inhibidores co-cristalizados.

Entre los inhibidores de Cz los más abundantes en nuestro estudio fueron aquellos derivados de vinilsulfona. Este hallazgo no es casual ya que este grupo de compuestos puede exhibir una buena selectividad y un desarrollo prospectivo favorable a pesar de la naturaleza irreversible de la inhibición. En relación con lo anterior, Jaishankar (2008) sintetizó y determinó las constantes de inhibición contra Cz de una serie de análogos de vinilsulfona estrechamente relacionados con K-777, un inhibidor de Cz (PDB: 2OZ2). En particular, este grupo investigó cómo las sustituciones en los fragmentos P2 y P3 de K-777 modifican las actividades contra Cz.

A raíz de estos resultados, en este capítulo de tesis aprovechamos la relación estructura-actividad entre los análogos de vinilsulfona descritos por Jaishankar (2008) pero desde una perspectiva basada en la estructura, es decir, mediante el estudio de las interacciones moleculares en el sitio de unión de la enzima, con el fin de obtener algunas pistas sobre el mecanismo de inhibición enzimática. Cabe destacar en este punto que esto no fue posible realizar en el capítulo anterior debido a que la mayoría de los ligandos carecían de sus respectivos valores experimentales de inhibición.

Como descriptor de interacciones moleculares en complejos de vinilsulfonas con Cz, se empleó el valor de densidad de carga en el punto crítico de enlace, (Bader 1990).

A su vez, debe considerarse que el procesamiento de una cantidad tan masiva de datos no debe hacerse "a mano", es decir, mediante la inspección visual de los grafos moleculares. Si esto ocurriera, se pasará por alto mucha información "oculta" bajo los datos de densidad de carga.

En consecuencia, para este capítulo de tesis se emplearon herramientas de aprendizaje automático para automatizar el proceso de extracción de información de grafos moleculares de densidad de carga y de este modo, explotar exhaustivamente los datos obtenidos.

4.2 Metodología

4.2.1 Protocolo de simulación

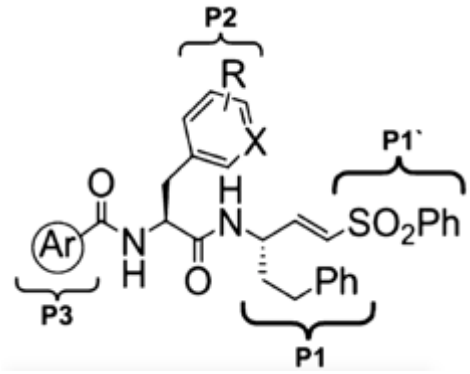
Se simularon una serie de 17 análogos de vinilsulfona, estrechamente relacionados con K-777, de los cuales se contaba con las constantes de inhibición contra Cz (Jaishankar et al. 2008). La Tabla 4.1 resume toda esta información.

Aunque la estructura experimental de estos análogos de vinilsulfona unidos covalentemente a Cz no se ha determinado todavía (excepto para K-777, código de PDB: 2OZ2), para los inhibidores de Cz del tipo peptídicos, se puede realizar una estimación inicial razonablemente precisa del modo de unión del inhibidor. De este modo, se construyeron "a mano" los complejos, colocando cada residuo de la secuencia del inhibidor P1', P1, P2 y P3 en el correspondiente sub-bolsillo enzimático S1', S1, S2 y S3 respectivamente, como se muestra en el esquema 4.1.

Las coordenadas iniciales del complejo se tomaron de la estructura de Cz unido a K-777 (código de PDB: 2OZ2) (Kerr et al. 2009). Se realizaron sustituciones en los residuos P2 y P3 de K-777 para obtener los análogos informados por Jaishankar (2008), formando un total de 17 complejos. Los mismos posteriormente fueron refinados por simulaciones de DM.

Todas las simulaciones de complejos Cz-inh se llevaron a cabo con el paquete de software Amber14 (David A. Case et al. 2005; D. A. Case et al. 2014) a una temperatura de 300 K y se extendieron hasta 50 ns de tiempo total de simulación en una caja periódica octaédrica truncada de moléculas de agua TIP3P. Se usó el campo de fuerza amber ff14SB para los residuos de proteínas (Maier et al. 2015). Se utilizó el software Antechamber del paquete Amber-Tools para generar los parámetros del inhibidor con el campo de fuerza GAFF y se calcularon cargas RESP (Wang et al. 2004).

Esquema 4.1 Estructura de vinilsulfona con los sitios de sustitución denominados Ar, R y X [a]



[a] Los segmentos del inhibidor que se unen a los sub-bolsillos enzimáticos S1, S1 ′, S2 y S3 se denominan P1, P1 ′, P2 y P3, respectivamente

Tabla 4.1 Inhibidores de Cz análogos de vinilsulfona reportados por Jaishankar (2008)[a]

Compuestos[b]	R	X	Ar	K_i (nM)
9d	4-Me	CH	DHBD	19
7d	4-Me	CH	4-CF_3Ph	45
6b	3-Me	CH	3,5-DiFPh	50
9b	3-Me	CH	DHBD	71
9a	H	CH	DHBD	80
7b	3-Me	CH	4-CF_3Ph	92
7a	H	CH	4-CF_3Ph	97
8c	3-CF_3	CH	2-piridil	150
4c	3-CF_3	CH	*N*-MePip	170
4a (K-777)	H	CH	*N*-MePip	220
8a	H	CH	2-piridil	250
8b	3-Me	CH	2-piridil	280
6d	4-Me	CH	3,5-DiFPh	350
6a	H	CH	3,5-DiFPh	980
8d	4-Me	CH	2-piridil	1700
4b	3-Me	CH	*N*-MePip	3300
4e	H	N	*N*-MePip	3600

[a] *N*-MePip: N-metil piperazina; DHBD: 2,3-dihidro-1,4-benzodiox-in-6-ilo; y 3,5-DiFPh: 3,5-diFluorofenilo. [b] La nomenclatura de los compuestos se extrajo de Jaishankar (2008)

4.2.2 Teoría cuántica de átomos en moléculas

En el contexto de la Teoría Cuántica de Átomos en Moléculas (QTAIM) el mapeo del campo del vector gradiente sobre la distribución compleja de la densidad de carga electrónica da lugar a tres elementos topológicos (un punto, una línea, una superficie). Entre los elementos topológicos, un punto crítico de enlace (BCP) y los caminos de enlace (BP), que conectan los átomos interactuantes, son indicadores inequívocos de la existencia de una interacción o de enlace (Bader 1990).

A partir de las trayectorias de DM de los complejos Cz-inh, se seleccionaron las estructuras del mínimo de energía potencial como una estructura representativa única sobre la que se realizó el análisis de densidad de carga. Debido a que los cálculos mecánicos cuánticos precisos todavía no son posibles para los complejos biomoleculares completos (por ser costosos computacionalmente), se construyeron modelos reducidos a partir de las estructuras de mínimo de energía potencial. Se incluyeron un total de 28 residuos (~570 átomos) en los modelos reducidos: el inhibidor derivado de vinilsulfona y los residuos circundantes en un volumen esférico de aproximadamente 5 Å centrado en los átomos del inhibidor. La densidad de carga se calculó con ayuda del paquete Gaussian 09 (Frisch et al. 2016) mediante la metodología del funcional de la densidad, DFT, con un funcional híbrido corregido por dispersión M06-2x y el conjunto base 6-31G(d). El análisis topológico de la densidad de carga se realizó luego con el software Multiwfn (Lu y Chen 2012).

4.2.3 Máquinas de vectores de soporte – eliminación recursiva de características (SVM-RFE)

Los valores de densidad de carga asociados a 319 interacciones no covalentes por complejo, obtenidos del análisis QTAIM se utilizaron como características (variables) para entrenar un clasificador de SVM lineal. Los SVM son modelos de aprendizaje supervisado con algoritmos de aprendizaje asociados que analizan los datos utilizados para el análisis de clasificación y regresión (Li, Liang, and Xu 2009).

Si los datos no son separables por un hiperplano, pueden mapearse en espacios de características de mayor dimensionalidad donde la separación lineal de ejemplos positivos y negativos podría ser posible (es decir, el llamado truco del *kernel*). Sin embargo, a diferencia de los modelos lineales, los modelos SVM entrenados en espacios de *kernel* de alta dimensión tienen características de “cajas negras” y, en general, es difícil racionalizar el rendimiento del modelo (Rodriguez-Pérez, Vogt, and Bajorath 2017). Por lo tanto, en este capítulo de tesis, nos limitamos a un algoritmo de SVM lineal porque nuestro principal interés era descubrir las relaciones entre las características (es decir, las interacciones moleculares) y las actividades biológicas para comprender, en última instancia, el mecanismo de inhibición enzimática. No obstante, es importante tener en cuenta que el análisis de características que contribuyen a las predicciones sólo tiene sentido si el modelo alcanza un nivel de rendimiento razonablemente alto.

Es bien sabido que cuando el número de características (interacciones) es grande y el número de ejemplos de entrenamiento es comparativamente pequeño, surge el riesgo de sobreajuste. Por lo tanto, para saltear el problema de la alta dimensionalidad y las escasas muestras de nuestro conjunto de datos, el algoritmo de SVM se combinó con uno RFE durante el entrenamiento del modelo.

SVM-RFE es un algoritmo de selección de características basado en la eliminación hacia atrás de características con pesos más bajos. En cada iteración, el modelo SVM se entrena con el subconjunto actual de características (interacciones), el peso (|w|) de cada característica (interacción) se calcula de acuerdo con el clasificador SVM, las características (interacciones) se clasifican de acuerdo con |w|, y luego, se eliminan las características con menos relevancia para la clasificación (Lin et al. 2018). El modelo fue implementado para discriminar entre las interacciones de tipo activo y las interacciones de tipo inactivo, presentes en los complejos formados por los inhibidores más activos y las que ocurren en los menos activos.

SVM-RFE y la validación cruzada estratificada doble se implementaron con la ayuda del módulo scikit-learn de Python (Pedregosa et al. 2011).

4.2.4 Análisis dinámico de correlación cruzada

Se calculó la matriz de correlación que describe cómo se relacionan las interacciones entre sí entre los complejos Cz-Inh. Esto se realizó a partir de los datos de densidad de carga obtenidos de los cálculos de QTAIM. Solo las interacciones con “importancias” superiores a 2.0 en el clasificador SVM-RFE se consideraron para el análisis de correlación.

4.3 Resultados y Discusión

4.3.1 Densidad de carga electrónica local como descriptor de interacciones moleculares en los complejos Cz-Inh

Para describir las interacciones moleculares en los complejos modelados de Cz-inh, se realizó el análisis topológico de densidad de carga en el marco del QTAIM sobre los complejos refinados.

Brevemente, este análisis consiste básicamente en el mapeo del campo del vector de gradiente sobre la densidad de carga precalculada del complejo, $\nabla\rho$. A partir de este mapeo, surgen los elementos topológicos de densidad de carga. Entre los elementos topológicos, los puntos críticos de enlace y los caminos de enlace que conectan a los átomos interactuantes, son indicadores inequívocos de la existencia de una interacción de enlace.

Como ya se mostró previamente, la Figura 4.1 muestra los caminos de enlace (BP) y los puntos críticos de enlace (BCP) asociados a las interacciones no covalentes (interacciones Cz-Inh así como Cz-Cz e Inh-Inh) en uno de los complejos estudiados aquí.

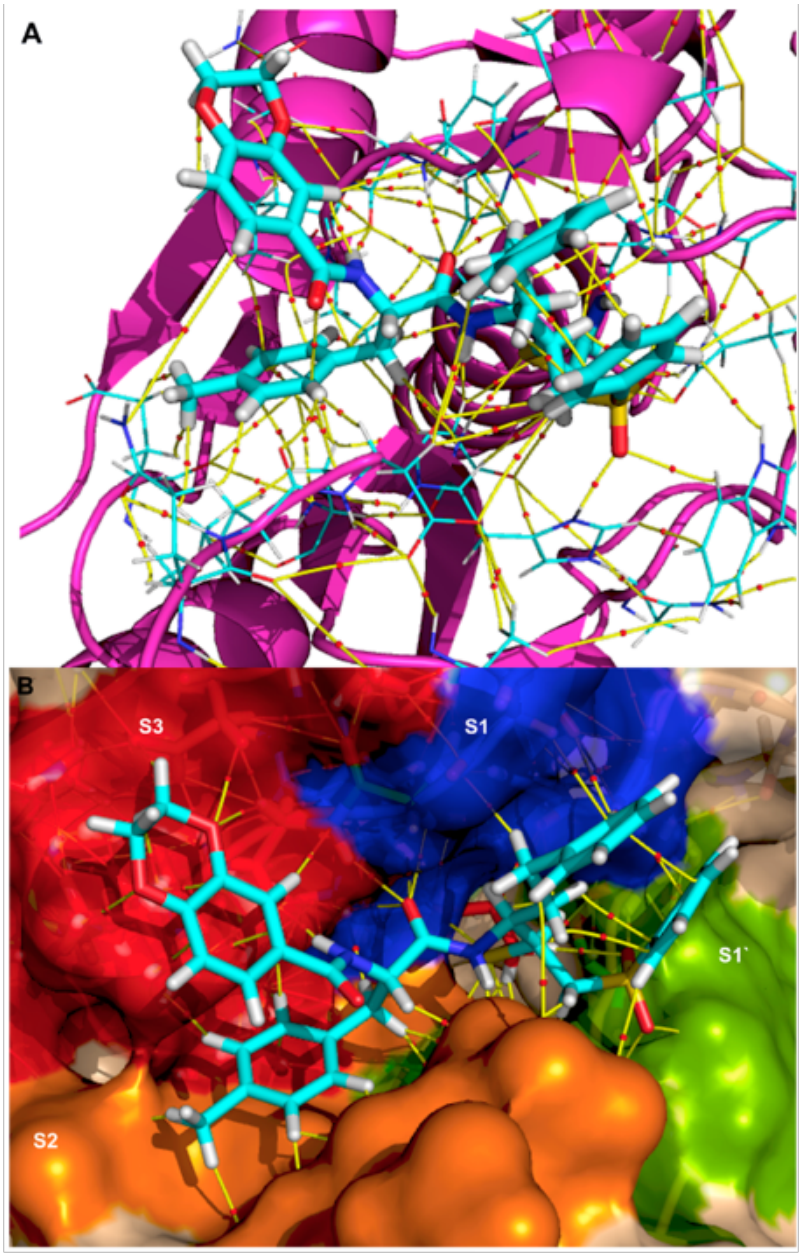


Figura 4.1 Vista de la intrincada red de interacciones en la estructura del complejo Cz-9d. Los elementos topológicos de densidad de carga que describen las interacciones no covalentes se representan con pequeños círculos rojos (BCP) y líneas amarillas que conectan cada BCP con ambos átomos que interactúan (BP). Se consideran interacciones tanto intermoleculares (Cz-Inh) como intramoleculares (Cz-Cz e Inh-Inh). La estructura de la proteína se representa tridimensionalmente como un conjunto de hélices y láminas en (A) y de manera superficial en (B), donde cada color de la superficie representa un sub-bolsillo diferente dentro de la hendidura de unión de Cz.

4.3.2 Entrenamiento de un clasificador de interacciones basado en los datos de densidad de carga

En este punto, se contaba con los elementos topológicos de la densidad de carga que describen las interacciones en los complejos Cz-Inh y los datos de actividad asociados a los correspondientes inhibidores.

Aprovechando estos datos, se buscaron las interacciones favorables (para estabilizar el complejo), lo que podría explicar la mayor afinidad de unión de los inhibidores más activos y las interacciones desfavorables (o menos favorables) que dominan la unión de los complejos menos activos. A su vez, se buscaron las interacciones implicadas en el mecanismo de acción de la enzima utilizando un modelo lineal supervisado basado en las interacciones complejas y sus correspondientes actividades inhibitorias.

El análisis QTAIM de complejos biomoleculares a menudo da lugar a redes de interacciones muy densas y complejas. Al inspeccionar la Figura 4.1, se hace evidente que un análisis comparativo de una red tan intrincada de interacciones para un conjunto de complejos Cz-Inh no se puede realizar mediante la inspección visual de los grafos moleculares. Si es así, se pasará por alto mucha de la información "oculta" bajo los datos de densidad de carga.

En cambio, para este trabajo, se aplicaron herramientas de aprendizaje automático a fin de automatizar el proceso de extracción de información de los grafos moleculares de densidad de carga y explotar exhaustivamente los datos obtenidos.

Como explica Fujita (2016), dependiendo del tipo particular de pregunta científica que deba responderse, el modelo predictivo puede ser más o menos complejo. Los modelos lineales relativamente simples son más fáciles de interpretar en términos de interacciones moleculares, aunque su poder predictivo es limitado. Los modelos no lineales más complejos tienen un mayor poder de predicción, pero a su vez menos interpretables.

Además, los conjuntos de datos en los que hay menos individuos/entidades observadas que variables son cada vez más frecuentes, gracias a la creciente facilidad de observación de las variables, junto con el alto costo de repetir las observaciones en algunos contextos (por ejemplo, *microarrays* de ADN) (Jolliffe y Cadima 2016). Por ejemplo, Guyon y col. (2002) construyeron un clasificador de SVM para seleccionar un subconjunto de genes biológicamente relevantes para el cáncer a partir de patrones amplios de datos de expresión génica, registrados en *microarrays* de ADN. Para ello, utilizaron una cantidad relativamente pequeña de ejemplos de entrenamiento de pacientes con cáncer y normales.

Es bien sabido que cuando el número de variables es grande y el número de individuos de entrenamiento es comparativamente pequeño (como en el caso de la referencia anterior y en nuestro caso), surge el riesgo de sobreajuste.

El problema del sobreajuste se puede reducir midiendo la importancia de las variables y seleccionando un subconjunto de estas más discriminativo. La eliminación de variables redundantes o irrelevantes puede mejorar la precisión del modelo, la capacidad de generalización e incluso el costo computacional en algunos casos (Bolón-Canedo y Alonso-Betanzos 2019).

SVM es una técnica de clasificación que utiliza vectores de soporte para maximizar la distancia entre las dos clases. Los coeficientes del modelo representan las coordenadas vectoriales que son ortogonales al hiperplano y su dirección indica la clase predicha. El tamaño absoluto de los coeficientes (ponderaciones) se puede usar para determinar la importancia de la variable para la tarea de separación de datos (Guyon et al. 2002).

Por otro lado, SVM-RFE es un algoritmo de selección de características hacia atrás basado en SVM. SVM-RFE se ha aplicado ampliamente en muchos campos, incluidos genómica, proteómica, metabolómica y otras situaciones, donde los datos presentan una gran cantidad de variables y las muestras son escasas (Lin et al. 2018).

RFE comienza con el conjunto completo de variables, crea el modelo SVM y evalúa la precisión. Los predictores menos importantes se borran y el modelo se vuelve a calcular (Butcher y Smith 2020).

En este capítulo de tesis, se entrenó un modelo SVM-RFE con la información de densidad de carga derivada de QTAIM sobre las interacciones moleculares de los 17 complejos Cz-Inh para seleccionar características relevantes para la tarea de clasificación, lo que podría ayudar a comprender el mecanismo de acción enzimática. Los inhibidores se etiquetaron como activos o inactivos de acuerdo con un valor umbral de decisión de 170 nM de actividad inhibidora, lo que aseguró clases equilibradas.

SVM-RFE se construyó con un conjunto de datos que contenía inicialmente 319 interacciones, donde luego, las variables (interacciones) menos relevantes se eliminaron iterativamente mediante un procedimiento de selección retrospectivo.

El análisis de características que contribuyen a las predicciones sólo tiene sentido si el modelo alcanza un nivel de rendimiento razonablemente alto. Por lo tanto, para monitorear la precisión del modelo durante la eliminación retrospectiva de características, se realizó una validación cruzada estratificada doble. En la validación cruzada estratificada, la distribución de clases de cada iteración se conserva para todo el conjunto de datos (Zhou y Tuck 2007).

La Figura 4.2 muestra la precisión media de la validación cruzada del modelo en función del número de características seleccionadas por el procedimiento SVM-RFE. Además, se muestra la variación de la precisión entre cada iteración.

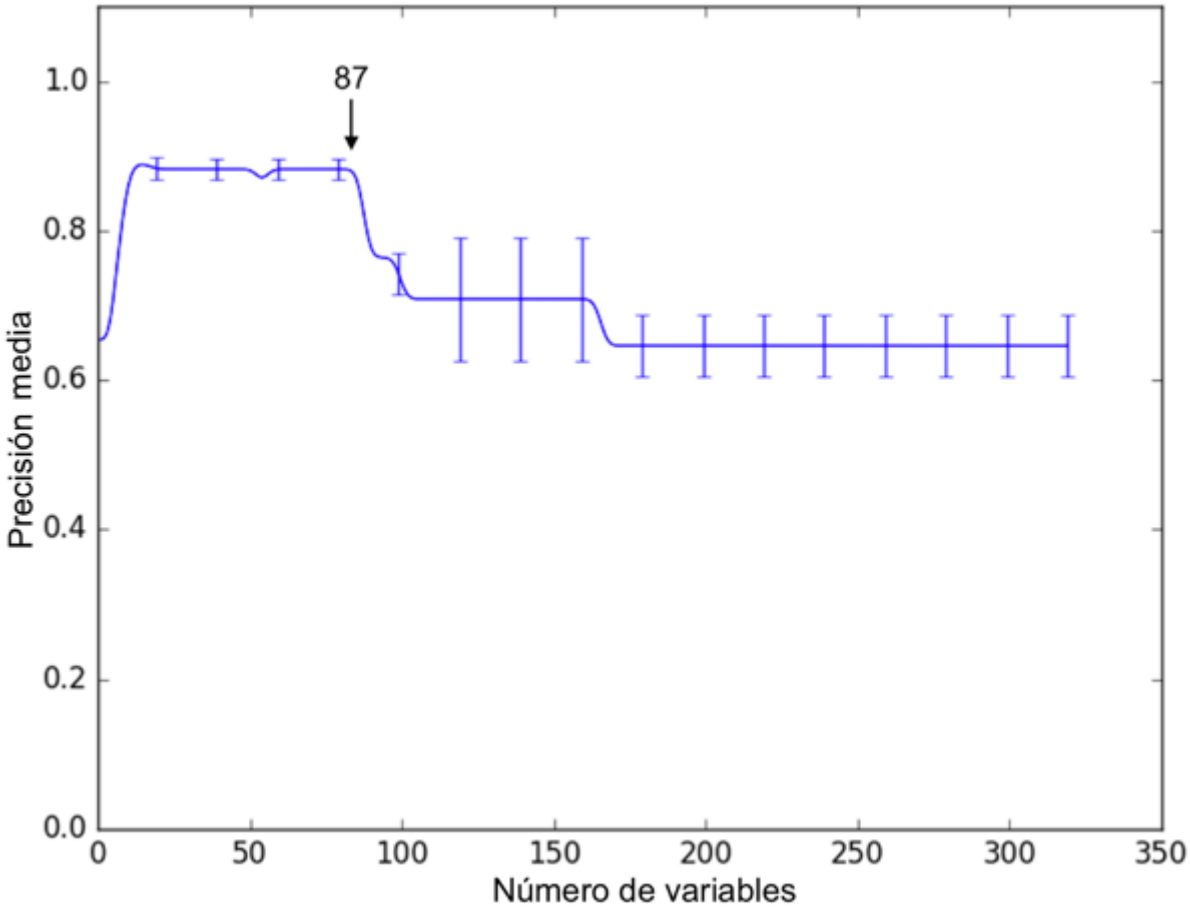


Figura 4.2 Proceso iterativo de eliminación retrospectiva de variables y entrenamiento del modelo SVM con las variables restantes. La precisión media del modelo SVM se representa en función del número de variables. Las barras de error representan las variaciones de los valores de precisión entre las iteraciones.

Como se puede ver en la figura, la precisión media del modelo aumenta a medida que el número de variables desciende hasta 87, cuando se alcanza el rendimiento máximo (87,75%). Además,

se debe tener en cuenta que la varianza de la precisión entre las dos iteraciones disminuye a un valor mínimo en la región de meseta entre ~20 y 87 características.

Por debajo de ~20 características, la precisión media comienza a disminuir nuevamente, lo que indica que el modelo de clasificación se vuelve demasiado simple como para discriminar entre compuestos de clases activas e inactivas.

Por lo tanto, para el análisis posterior de las interacciones relevantes, seleccionamos el modelo SVM entrenado con un subconjunto de las 87 mejores variables porque una mayor reducción del número de variables no implica un aumento en el rendimiento del modelo.

El diagrama de barras en la Figura 4.3 muestra las principales interacciones (variables) que fueron utilizadas por el modelo final para hacer las clasificaciones. En la figura solo se muestran los coeficientes de las variables con valores absolutos superiores a 2,0.

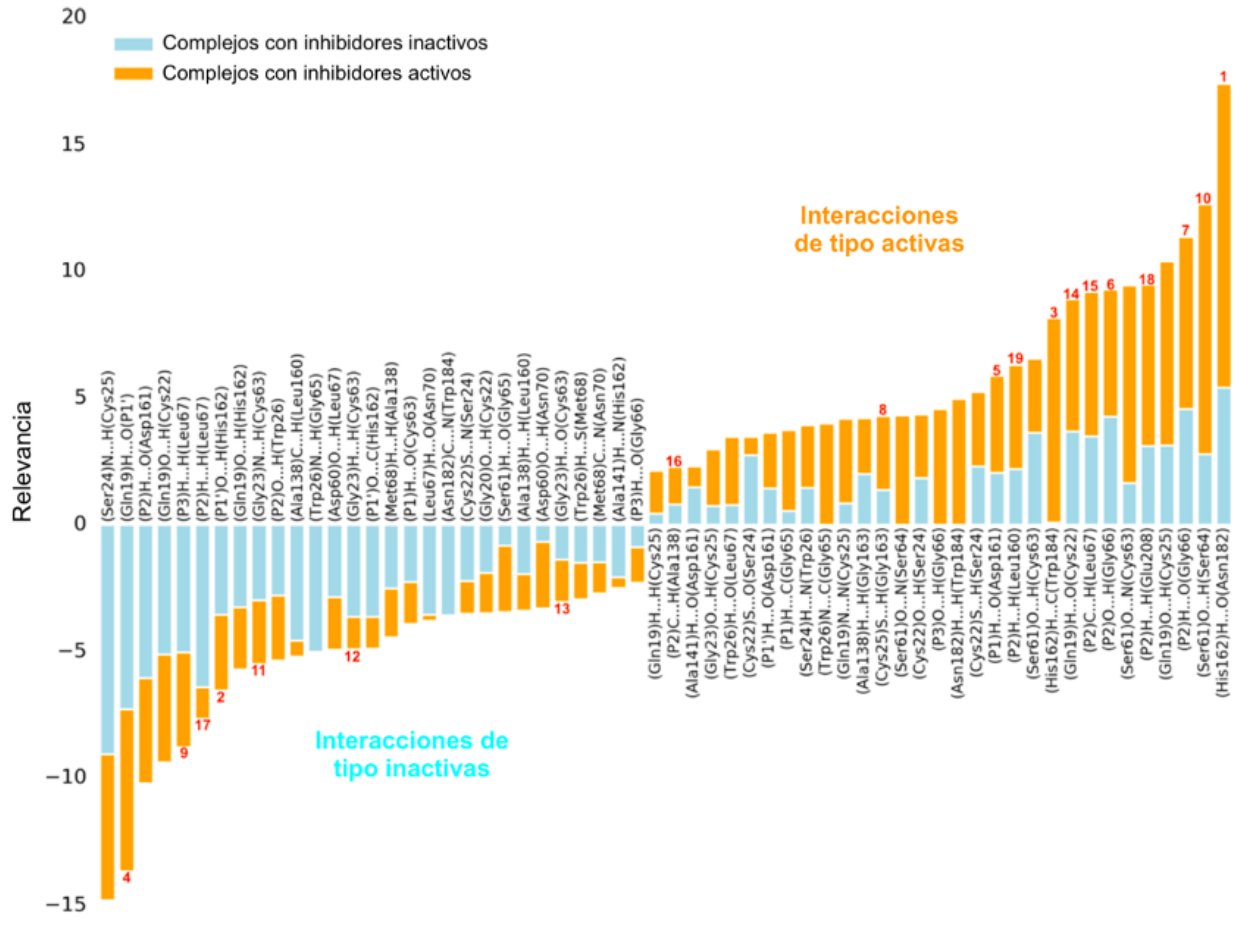


Figura 4.3 Principales interacciones (variables) seleccionadas por el modelo SVM para realizar las clasificaciones de clases. Los números en rojo indican las interacciones discutidas en el texto (secciones 4.3.5, 4.3.6 y 4.3.7).

La altura total de las barras apiladas en la figura anterior representa la importancia de la interacción para la tarea de clasificación, mientras que cada categoría dentro de la barra representa la contribución de la densidad de carga de las dos clases (activa e inactiva en naranja y celeste, respectivamente) a la importancia general de la interacción.

Como se puede ver en la figura, las interacciones con coeficientes positivos tienen en general mayores contribuciones de compuestos etiquetados como activos, mientras que lo contrario ocurre para las interacciones con coeficientes negativos, es decir, sus contribuciones más importantes provienen de compuestos etiquetados como inactivos.

Por lo tanto, utilizando un modelo de clasificación de SVM lineal simple e interpretable junto con un procedimiento RFE, es posible extraer información útil sobre cuáles son las interacciones más importantes para discriminar entre compuestos activos e inactivos (o menos activos) frente a Cz.

4.3.3 Matriz de correlación basada en interacciones a partir de datos de densidad de carga

Aunque el modelo de clasificación entrenado ayuda a recuperar las interacciones relevantes de los gráficos moleculares de densidad de carga de los complejos Cz-Inh, no necesariamente proporciona información sobre cómo estas interacciones entran en juego juntas, es decir, cómo se correlacionan entre sí para llevar a la enzima a un estado conformacional particular.

Diferentes inhibidores podrían formar diferentes interacciones que a su vez podrían estabilizar diferentes estados conformacionales de la enzima. En particular, nos interesaba saber si podría haber una relación entre la actividad del compuesto contra Cz y la conformación de la enzima estabilizada. Esta información podría ser muy útil para elegir la estructura de partida adecuada en futuras campañas de cribado virtual basadas en estructuras.

En consecuencia, la matriz de correlación que describe cómo se relacionan las interacciones entre sí entre los complejos Cz-Inh se calculó a partir de los datos de densidad de carga (Figura 4.4). Solo se consideraron para el análisis de correlación las interacciones con importancia superior a 2,0 en el clasificador SVM.

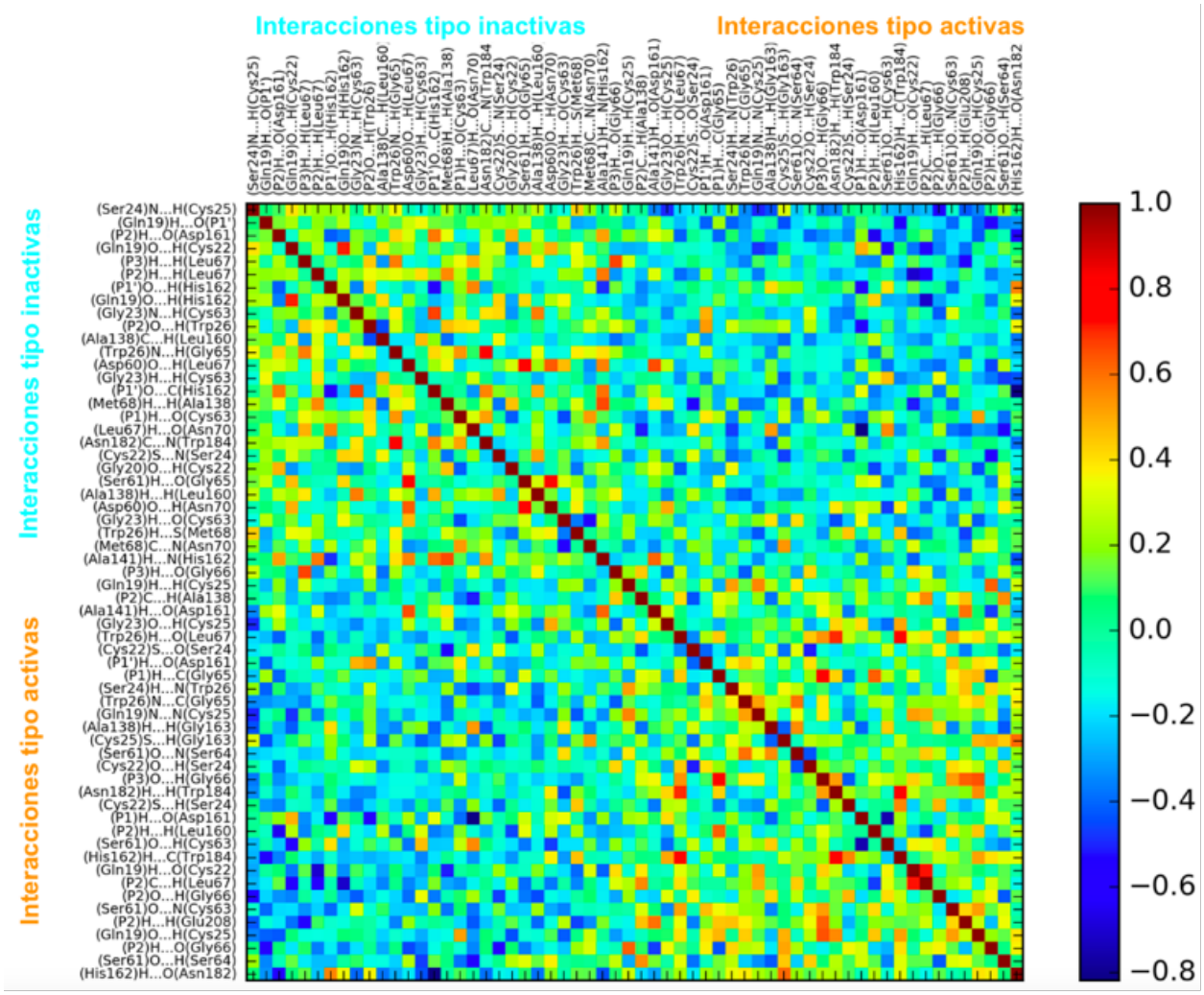


Figura 4.4 Matriz de correlación basada en datos de densidad de carga de interacciones en los complejos Cz-Inh estudiados.

La Figura 4.4 muestra que existe una clara anticorrelación (es decir, valor negativo) entre interacciones de tipo activo e inactivo, es decir, entre interacciones que prevalecen en complejos de compuestos etiquetados como activos e inactivos, respectivamente. Esto significa que a medida que las primeras interacciones se vuelven más fuertes, las últimas se debilitan. Este hallazgo sugiere que los compuestos activos e inactivos (menos activos) estabilizan diferentes conformaciones de Cz.

4.3.4 Grafos moleculares de densidad de carga

La Figura 4.5 muestra la superposición estructural de los complejos Cz-6b y Cz-8d correspondientes a compuestos de clases activas e inactivas, respectivamente. Las interacciones que se forman/rompen (o simplemente se fortalecen/debilitan) en la comparación entre ambos complejos se representan a través de sus correspondientes elementos topológicos de densidad de carga (es decir, los BCP y los BP). Los valores de densidad de carga para las interacciones discutidas se muestran en la Tabla 2, en la sección anexo, al final del capítulo.

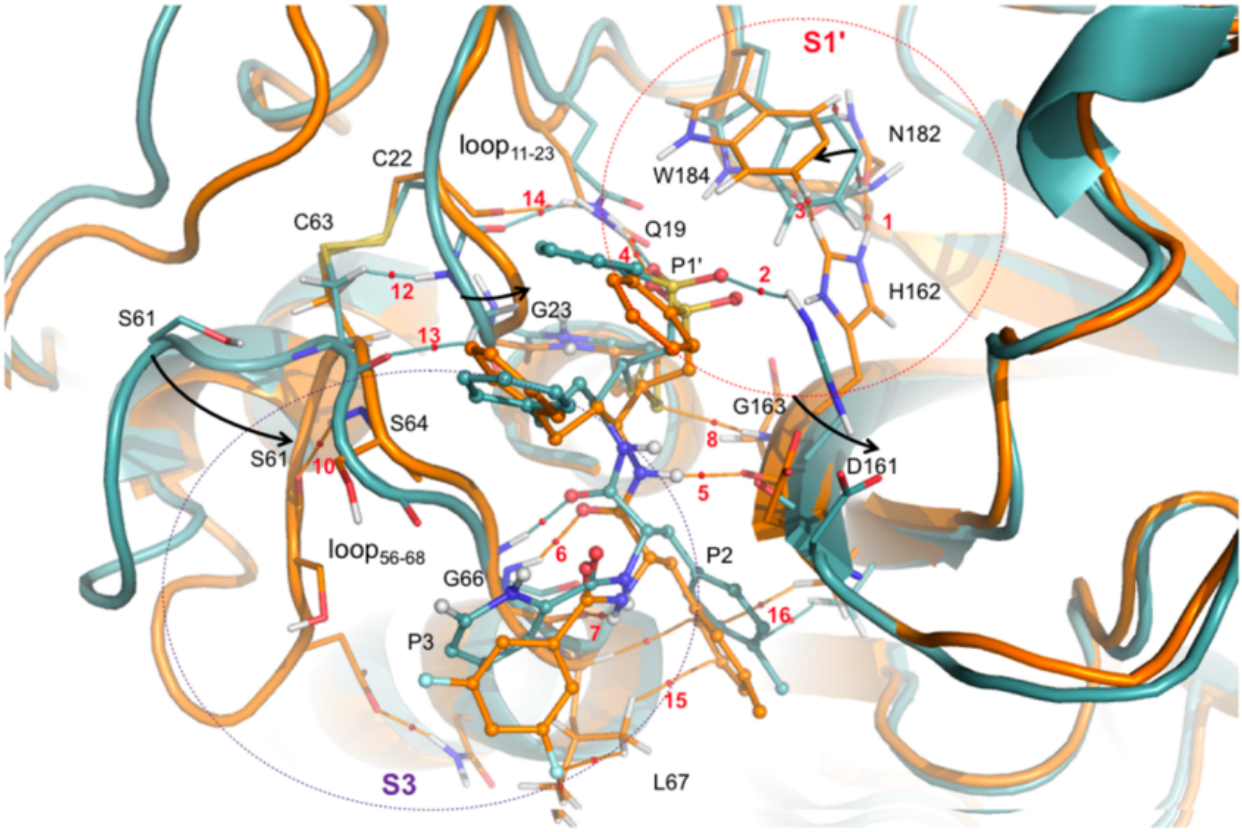


Figura 4.5 Superposición estructural de complejos Cz-6b (anaranjado) y Cz-8d (cian). También se representan elementos topológicos de densidad de carga para interacciones atómicas: los caminos de enlace (BP) que conectan los núcleos se representan en naranja y azul para Cz-6b y Cz-8d, respectivamente. Los puntos críticos de enlace (BCP) se muestran como pequeñas esferas rojas. Los números en rojo indican las interacciones más significativas (lo mismo que en la Figura 4.3). Las flechas indican el desplazamiento de la cadena principal de la proteína entre los complejos Cz-8d y Cz-6b

Entre las interacciones que prevalecen en el grupo de inhibidores más activos de Cz, el enlace N-H…O=C entre cadenas laterales de His162 (protonada) y Asn182 en el subsitio enzimático S1′

es el más relevante para la tarea de clasificación, según el diagrama de barras de la Figura 4.3. Esta interacción se puede identificar como interacción 1 en el grafo molecular de la Figura 4.5 y en el gráfico de barras de la Figura 4.3.

Como es bien sabido, la interacción 1 facilita la formación del par iónico tiolato-imidazolio (Cys25)S−…+H−N(His162) necesario para la catálisis. Por tanto, es notable que esta interacción esté formada por compuestos de la clase activa, como el compuesto 6b, pero no por compuestos de la clase inactiva, como el compuesto 8d. Esto significa que los compuestos etiquetados como activos imitan mejor al complejo enzima-sustrato porque son capaces de acomodar la maquinaria catalítica como si estuviera a punto de escindirse un enlace del sustrato.

En complejos con inhibidores de la clase inactiva, la cadena lateral de la His162 se desplaza lejos de Asn182, acercándose hacia el grupo sulfonilo de los inhibidores en posición P1'. Este desplazamiento forma una fuerte interacción (P1')S=O…$^{+}$H-N(His162) que es una de las principales características del modelo de aprendizaje automático (ML) entre compuestos etiquetados como inactivos (interacción 2 en las Figuras 4.3 y 4.5). En estos complejos, un anillo indol cercano a Trp184 ocupa el espacio donde los residuos His162 y Asn182 interactuarían en complejos activos.

Por el contrario, en los complejos de compuestos de la clase activa, se observan cambios opuestos: la interacción 2 se debilita y la His162 se mueve hacia Asn182 para formar la interacción 1.

Sin embargo, antes de que se pueda establecer la interacción 1, el anillo Trp184 debe primero desocupar la región entre los residuos His162 y Asn182. Al hacerlo, Trp se aleja de Asn182 y termina justo encima del anillo His162 donde la nube de electrones Trp forma una interacción de apilamiento C−H…π con un átomo de hidrógeno no polar de His. Esta interacción, etiquetada

como 3 en las Figuras 4.3 y 4.5, también es considerada por el modelo SVM-RFE como una de las principales interacciones entre la clase activa de inhibidores.

Creemos que estos hallazgos recuperados con la ayuda de un modelo de ML son significativos porque Trp184 es un residuo altamente conservado entre las cisteína proteasas lisosómicas pertenecientes a la superfamilia de papaína, y en trabajos anteriores se lo catalogaba como el "orquestador" de la tríada catalítica Cys25-His162-Asn182 porque se cree que juega un papel crítico en la escisión del sustrato al orientar la maquinaria catalítica enzimática (Arafet, Świderek, y Moliner 2018). En consecuencia, con base en nuestros resultados y hallazgos previos, proponemos que Trp184 podría actuar como un "interruptor" para la formación de la interacción.

Continuando con el análisis de interacciones relevantes en el subsitio S1′, se puede ver en la Figura 4.5 que el grupo sulfonilo del inhibidor se mantiene en su lugar en la entrada del subsitio S1′. Esto ocurre gracias a la formación de dos interacciones O…H fuertes entre ambos átomos de O sulfonílicos. La primera con un átomo de H del anillo imidazol His162 (interacción 2) y la segunda con un átomo de H del grupo amida de la cadena lateral Gln19 (interacción 4). Ambas interacciones son características relevantes entre la clase inactiva de inhibidores como se evidencia en la Figura 4.3. Esto significa que estas interacciones son más fuertes en complejos con inhibidores de la clase inactiva y se rompen o debilitan en complejos con inhibidores de la clase activa.

Aparentemente, cuando las vinilsulfonas están fuertemente unidas a través de las interacciones 2 y 4, como en el caso de los compuestos etiquetados como inactivos, las regiones restantes del inhibidor no encajan bien dentro de la hendidura de unión y, por lo tanto, no pueden establecer otras interacciones importantes que ayuden a unir la cadena principal del inhibidor. Más concretamente, la cadena principal de los residuos P1 y P2 del inhibidor no encaja adecuadamente en la parte estrecha de la hendidura de unión formada por los átomos de la cadena principal en subsitio enzimático S1 (ver más abajo, Figura 4.6).

Para que el inhibidor encaje bien en la hendidura de unión de la enzima, debe poder alterar la disposición de los residuos dentro del subsitio S1'. En otras palabras, debe ser capaz de romper

o debilitar las interacciones 2 y 4 que mantienen firmemente el resto de vinilsulfona P1' en la entrada del sub-bolsillo S1'.

Estos reordenamientos implican el desplazamiento de His162 hacia Ans182 y la formación de la interacción 1 posterior como se explicó anteriormente (Figura 4.5). También implica la retracción de la cadena lateral Gln19 como se describe a continuación. Los reordenamientos de la cadena lateral Gln19 se pueden ver más claramente en la Figura 4.6 que muestra la superposición estructural de complejos con los inhibidores 9d y 4b de la clase activa e inactiva, respectivamente.

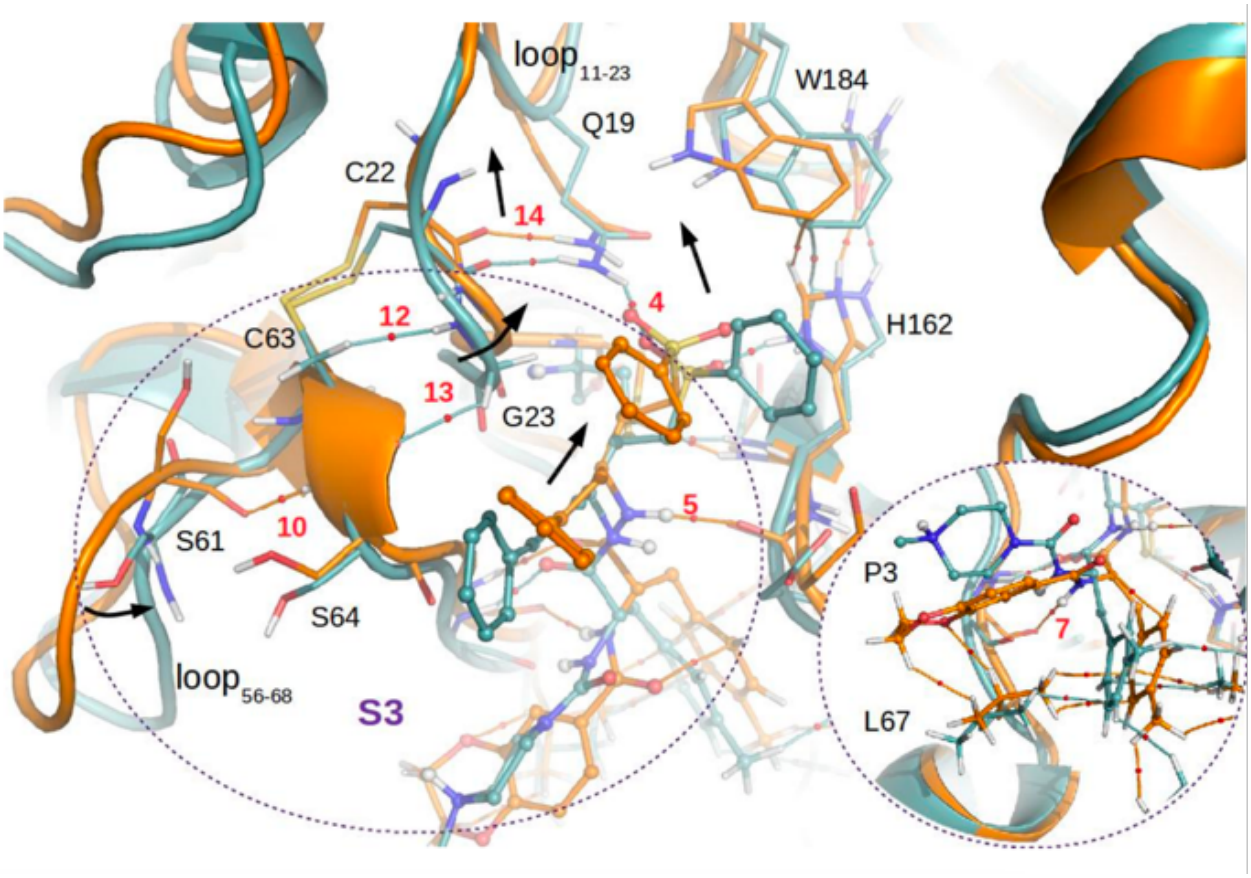


Figura 4.6 Superposición estructural de complejos Cz-9d (anaranjado) y Cz-4b (cian). Los elementos topológicos de densidad de carga para interacciones atómicas también se representan: los BP que conectan los núcleos se representan en naranja y azul claro para Cz-9d y Cz-4b, respectivamente. Los BCP se muestran con pequeñas esferas rojas. Las flechas indican el desplazamiento de la cadena principal de la proteína entre los complejos Cz-4b y Cz-9d. Las interacciones del residuo P3 con Leu67 se destacan en el recorte en la parte inferior derecha.

Dependiendo de los complejos analizados, algunos inhibidores de la clase activa parecen impulsar a los residuos Gln19 e His162 de modo que el segmento vinilsulfona P1' pueda penetrar un poco más profundamente en el subsitio S1' (Figura 4.6). Por lo tanto, Gln19 y His162 podrían actuar como “moduladores” al permitir selectivamente la entrada al subsitio P1' sólo a los inhibidores más activos.

Después de este reordenamiento en el subsitio S1', la cadena principal de los inhibidores en las regiones P1 y P2 encajan bien en la región estrecha de la hendidura de unión. Esto se evidencia en las interacciones entre las cadenas principales (P1)N-H…O=C(Asp161), (P2)C=O…H-N(Gly66) y (P2)N-H…O=C(Gly66) que se forman o mejoran en complejos con inhibidores etiquetados como activos y son algunas de las características más relevantes dentro de la clase activa, según el modelo SVM-RFE (interacciones 5, 6 y 7, respectivamente en las Figuras 4.3, 4.5 y 4.6).

Estas interacciones también se consideran un sello distintivo del evento de reconocimiento del sustrato en las cisteína proteasas (Turk et al. 1998). Además, la formación de la interacción (Cys25)S…H-N(Gly163) (interacción 8) ayuda a tirar de la cadena principal del inhibidor (que está unida covalentemente a Cys25) hacia la parte inferior de la región estrecha de la hendidura de unión a enzima, contribuyendo así al mejor ajuste general del inhibidor que se observa en los complejos de los compuestos más activos.

Debido a que los análogos de vinilsulfona reportados por Jaishankar (2008) difieren sólo en los residuos P2 y P3, la explicación de por qué algunos inhibidores pueden inducir los reordenamientos de residuos requeridos dentro del subsitio P1′ y algunos otros no deben estar relacionados de alguna manera con las interacciones que se establecen en los subsitios S2/S3.

4.3.4.1. Interacciones en el subsitio S3

El anillo P3 de los compuestos etiquetados como activos e inactivos interactúa de diferentes formas con Leu67, un residuo clave en el subsitio S3 (ver Figuras 4.5 y 4.6). Los inhibidores de la clase activa tienen grupos ricos en electrones en P3 y, por tanto, tienden a actuar como aceptores de enlaces H con la cadena lateral de Leu67. Esto se evidencia, por ejemplo, por interacciones como Leu(67)C-H…π(P3) y Leu(67)C-H…F(P3) en las que la nube de electrones

o los pares libres del flúor del anillo 3,5-difluorofenilo (compuesto 6b) actúan como aceptores (Figura 4.5). En el otro caso, los pares libres del oxígeno del anillo 2,3-dihidro-1,4-benzodioxina (compuesto 9d) actúan como aceptores del enlace H en la interacción Leu(67)C-H…O(P3) (Figura 4.6). Desafortunadamente, estas interacciones no son recuperadas por el modelo SVM-RFE (si esto ocurriera, el modelo estaría sobreajustando los datos de densidad de carga) ya que no hay un patrón de enlace H único para Leu67 (es decir, hay diferentes aceptores de enlace H).

Por otro lado, los compuestos de la clase inactiva tienen anillos en la región P3 deficientes en electrones (2-piridinio y N-metil piperazina en las series 8 y 4, respectivamente, ver Tabla 4.1 al comienzo del capítulo), por lo que solo pueden formar contactos de dihidrógeno con la cadena lateral de Leu67. Este tipo de interacciones se recuperan mediante el modelo ML y constituyen una de las características más importantes entre los complejos con inhibidores etiquetados como inactivos (interacción 9 en la Figura 4.3).

Desde el punto de vista mecanicista, un fuerte anclaje del anillo P3 a la cadena lateral de Leu67 en complejos de compuestos activos podría tirar del inhibidor hacia la parte inferior de la hendidura de unión, lo que permitiría la formación de la interacción 7 entre las cadenas principales del residuo P2 del inhibidor y Gly66, (P2)N-H…O=C(Gly66) (Figuras 4.3, 4.5 y 4.6). Como se argumentó anteriormente, la formación de la interacción 7, junto con las interacciones 5, 6 y 8, son indicadores de un buen ajuste de la estructura del inhibidor dentro de la hendidura de unión a la enzima.

Además de este efecto directo de la interacción P3 con Leu67 sobre el anclaje de la estructura del inhibidor, parece existir también un mecanismo indirecto por el cual las interacciones P3 en el subsitio S3 influyen en el modo de unión del inhibidor. Estos hallazgos explicarían por qué los ligandos mejor rankeados en campañas de docking cuentan anillos ricos en electrones propiciando interacciones sobre Leu67 (Tripathi et al. 2023, de Almeida et al. 2023).

En complejos de compuestos etiquetados como activos, los residuos Ser61 y Ser64 del mismo bucle que Leu67 (es decir, $bucle_{56-68}$) forman un enlace de hidrógeno C=O…H-N que estabiliza un giro cerrado entre ambos residuos. Esta interacción (Ser61)C=O…H-N(Ser64) (etiquetada como interacción 10 en las Figuras 4.3, 4.5 y 4.6) es recuperada por el modelo SVM-RFE como la segunda variable más importante entre las interacciones de tipo activo para la clasificación de

compuestos etiquetados como activos/inactivos en función de los valores de K_i. Es probable que la estabilidad de la interacción 10 esté relacionada al menos en parte con el tipo de interacciones que forma el anillo de los inhibidores en la región P3 con la cadena lateral de Leu67. Un patrón de interacción de dihidrógeno inestable entre P3 y Leu67 (es decir, a través de la interacción 9), como en los complejos de compuestos etiquetados como inactivos, podría perturbar la conformación del bucle$_{56\text{-}68}$, conduciendo así a la ruptura observada de la interacción 10.

Por el contrario, los enlaces H estables entre P3 y Leu67, como en los complejos de compuestos de la clase activa, podrían ayudar a mantener más firme el bucle, contribuyendo así a preservar el giro Ser61→Ser64 en su forma cerrada.

Además, la conformación del giro Ser61→Ser64 parece definir cómo va a interactuar el bucle$_{56-68}$ con los elementos estructurales de la proteína circundante, como el bucle$_{11-23}$ cercano.

En complejos de compuestos de la clase inactiva, hay varias interacciones recuperadas por el modelo SVM-RFE como interacciones de tipo inactivo que podrían ayudar a mantener los bucles 56-68 y 11-23 juntos (es decir, (Cys63)H…N(Gly23), (Cys63)H…H(Gly23) y (Cys63)O…H(Gly23), etiquetadas como interacciones 11, 12 y 13, respectivamente). Por otro lado, tras la formación de la interacción 10 en complejos de compuestos de la clase activa, se produce un reordenamiento conformacional en el bucle$_{56\text{-}68}$ que de alguna manera provoca la ruptura de las interacciones 11, 12 y 13 que mantenían unidos ambos bucles. A medida que el bucle$_{11\text{-}23}$ se aleja del bucle$_{56\text{-}68}$, el primer bucle arrastra la cadena lateral de Gln19 a través de una interacción con la cadena principal de ese bucle (interacción 14, Figuras 4.3, 4.5 y 4.6). Mientras Gln19 se arrastra hacia atrás, su cadena lateral adquiere una conformación retorcida en la que Gln19 se aleja aún más del inhibidor. Como consecuencia, la interacción 4 entre la cadena lateral de Gln19 y el átomo de oxígeno del grupo sulfonilo del inhibidor se debilita. Como se discutió anteriormente, el reordenamiento de la cadena lateral Gln19 parece ser crítico para el posicionamiento adecuado del sustrato dentro de la hendidura de unión de Cz y la formación de las interacciones 5, 6, 7 y 8.

4.3.4.2. Interacciones en el subsitio S2

Entre todos los subsitios que abarcan la hendidura de unión Cz, el único que es lo suficientemente profundo como para merecer el nombre del sub-bolsillo es S2. En el sub-bolsillo S2, el anclaje de compuestos de la clase activa es impulsado principalmente por interacciones π…H entre la nube de electrones del anillo P2 y los hidrógenos no polares donados por los residuos Leu67 y Ala138 en ambos lados del sub-bolsillo. Estas interacciones, denominadas 15 y 16, respectivamente, han sido seleccionadas por el modelo SVM-RFE como características importantes entre los compuestos etiquetados como activos (ver Figuras 4.3, 4.5 y 4.6). Por otro lado, los compuestos de la clase inactiva no forman las interacciones 15 y 16 o son mucho más débiles. En cambio, el anillo P2 de estos compuestos forma contactos de dihidrógeno con Leu67 (interacción 17), lo que resalta la ubicación incorrecta del anillo P2 dentro del sub-bolsillo S2.

Al reunir las interacciones analizadas para los residuos P2 y P3, es evidente que Leu67 juega un papel clave en el anclaje adecuado de ambos residuos a la hendidura de unión de Cz.

La Figura 4.7 muestra la superposición estructural de complejos con los compuestos 6b y 6a de clases activas e inactivas, respectivamente. Estos compuestos sólo difieren en el sustituyente en el residuo P2, por lo que son adecuados para estudiar diferencias en los patrones de interacción que pueden atribuirse directamente a la estructura P2.

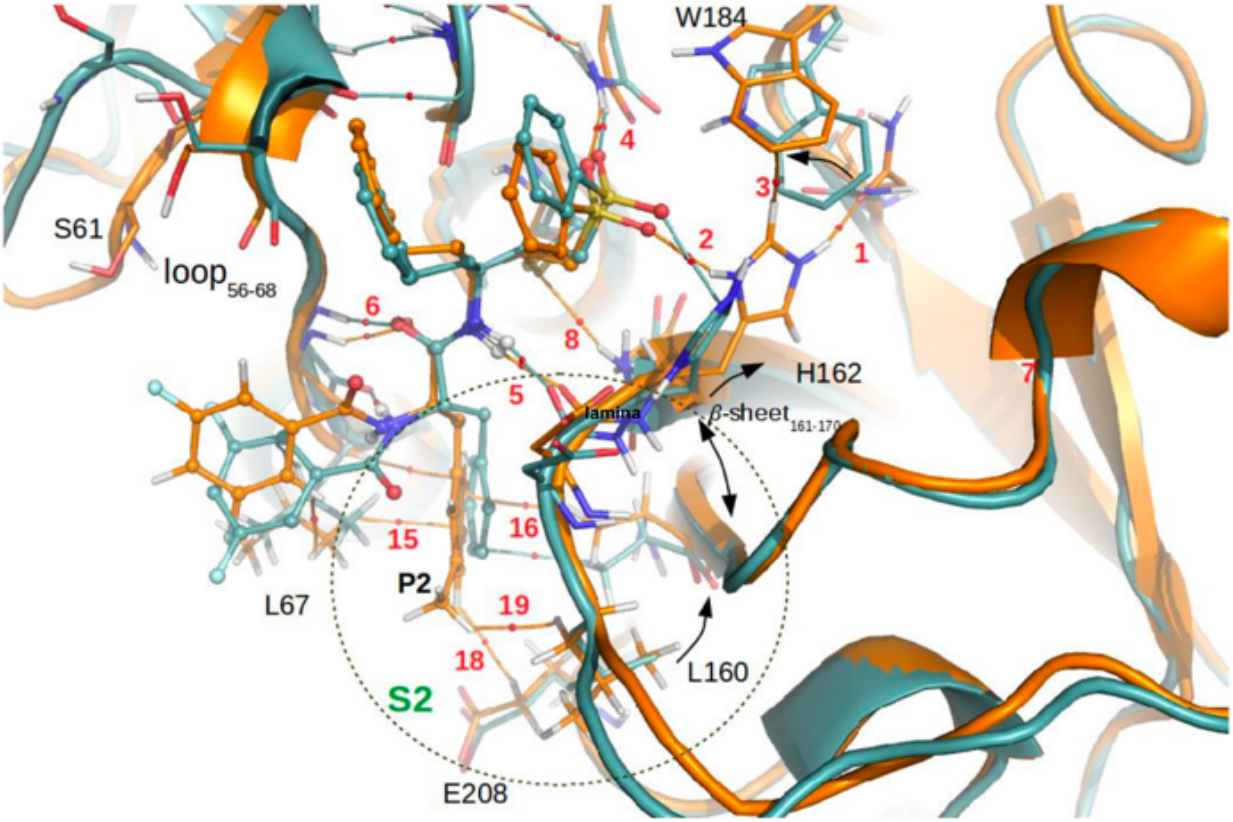


Figura 4.7 Superposición estructural de complejos Cz-6b (anaranjado) y Cz-6a (cian). También se representan los elementos topológicos de densidad de carga para las interacciones atómicas: los BP que conectan los núcleos se representan en anaranjado y azul para Cz-6b y Cz-6a, respectivamente. Los BCP se muestran con pequeñas esferas rojas. Las flechas indican el desplazamiento de la estructura de la proteína entre los complejos Cz-6b y Cz-6a.

Además de impulsar interacciones con Leu67 y Ala138, la mayoría de los compuestos activos también forman otras interacciones que vale la pena señalar. Así, por ejemplo, la interacción (P2)H…H(Glu208) entre dos átomos de H no polares de las cadenas laterales P2 y Glu208, respectivamente, fue seleccionada por el modelo SVM-RFE como una característica relevante entre los inhibidores de la clase activa (interacción 18 en las Figuras 3 y 7). Glu208 se encuentra en la parte inferior del sub-bolsillo S2; por tanto, la interacción de los inhibidores de la clase activa con ese residuo indica que pueden alcanzar dicha región distal del subsitio S2, mientras que los inhibidores de la clase inactiva, en general, no pueden.

Cerca de Glu208, hay otro residuo, Leu160, que también es el objetivo de la mayoría de los inhibidores activos a través de interacciones de dihidrógeno (P2)H…H(Leu160) que también son

recuperados por el modelo ML como una característica relevante entre los complejos de compuestos marcados como activos (interacción 19 en las Figuras 4.3 y 4.7).

Es poco probable que las fuerzas de atracción estén detrás de la formación de estas interacciones de dihidrógeno, ya que son más sugerentes de choques estéricos entre átomos hidrófobos del ligando y la enzima. Estos choques sutiles de dihidrógeno generalmente son las huellas dejadas por fuerzas repulsivas más fuertes que han sido aliviadas por el desplazamiento de los residuos involucrados. Por lo tanto, al inspeccionar estas interacciones de dihidrógeno, se pueden rastrear las translocaciones de residuos o los cambios de conformación que podrían haber ocurrido como consecuencia de un choque estérico anterior más fuerte. En particular, las interacciones de dihidrógeno 18 y 19 son las huellas de los desplazamientos de las cadenas laterales de Glu208 y Leu160, respectivamente, inducidos por sustituyentes en la posición 3 o 4 del anillo inhibidor en posición P2 (ver Tabla 4.1). Por el contrario, los compuestos que no llevan un sustituyente en el anillo P2, la mayoría de ellos pertenecientes a la clase inactiva, no alcanzan la pared distal/fondo del subsitio S2, por lo que no forman interacciones 18 y 19. Estas interacciones, y en particular la interacción 19 parecen estar relacionadas con reordenamientos de residuos en los subsitios S1 y S1'. A medida que el sustituyente en el anillo P2 empuja la cadena lateral de Leu160, también perturba las interacciones de la cadena principal entre la lámina-$\beta_{161-170}$ cercana y la lámina-$\beta_{135-139}$ que interactúan formando una horquilla. Como consecuencia, la cadena principal de la lámina-$\beta_{161-170}$ se "libera" parcialmente y los residuos al final de esa lámina, es decir, Asp161, His162 y Gly163 experimentan un movimiento hacia atrás concertado que los coloca en una posición adecuada para la formación de las interacciones 5 y 8 en el subsitio S1 y desencadenan reordenamientos en el subsitio S1′. Estos reordenamientos, que involucran a His162, finalmente conducen a la formación de la interacción 1, como se discutió anteriormente.

Tomando el conjunto de interacciones del inhibidor en el sub-bolsillo S2 y el subsitio S3, las primeras parecen gobernar los cambios conformacionales que ocurren en la pared derecha de la hendidura de unión (es decir, aquellos que involucran la lámina-$\beta_{161-170}$), mientras que las interacciones del residuo P3 en el subsitio S3 impulsa principalmente los cambios conformacionales en la pared izquierda (es decir, aquellos relacionados con el bucle$_{56-68}$). Ambos

cambios conformacionales finalmente conducen a reordenamientos de los residuos His162 y Gln19 en el sitio S1'. Dichos cambios permiten el posicionamiento adecuado de la “cabeza de guerra” vinilsulfona, y a su vez promueven la formación de interacciones críticas para la inhibición entre las cadenas principales del inhibidor y la pared del sitio catalítico.

Sin embargo, debe tenerse en cuenta que la división del estudio del mecanismo de inhibición por subsitios de la proteína podría ser una simplificación excesiva. Esto se debe principalmente a que las interacciones en diferentes subsitios podrían estar relacionadas entre sí, es decir, los cambios conformacionales observados podrían depender no solo de los sustituyentes en P2 y P3 sino también en la combinación de ambos.

4.3.5 Modelo conformacional de dos estados finales para Cz compatible con las simulaciones de DM.

En la Sección 4.3.3, separamos las interacciones que son más frecuentes en los complejos de la mayoría de los inhibidores activos (es decir, interacciones de tipo activo) de las que son más comunes en complejos de compuestos de la clase inactiva (es decir, interacciones de tipo inactivo). Luego, en la Sección 4.3.4, a través del análisis de correlación, dimos un paso más para concluir que las interacciones de tipo activo e inactivo estabilizan dos conformaciones opuestas de la enzima.

Para apoyar aún más esta hipótesis, analizamos algunas interacciones de tipo activo e inactivo a lo largo de las trayectorias de DM de los complejos Cz-Inh.

La Figura 4.8 muestra histogramas de distancia obtenidos a partir de las simulaciones DM de complejos Cz-6b y Cz-8d. Estos gráficos corresponden a varias interacciones consideradas por el modelo SVM-RFE como características importantes para la estabilización de las conformaciones enzimáticas de estado final activas o inactivas. Además, se muestra un histograma para el ángulo de torsión de la cadena lateral Gln19.

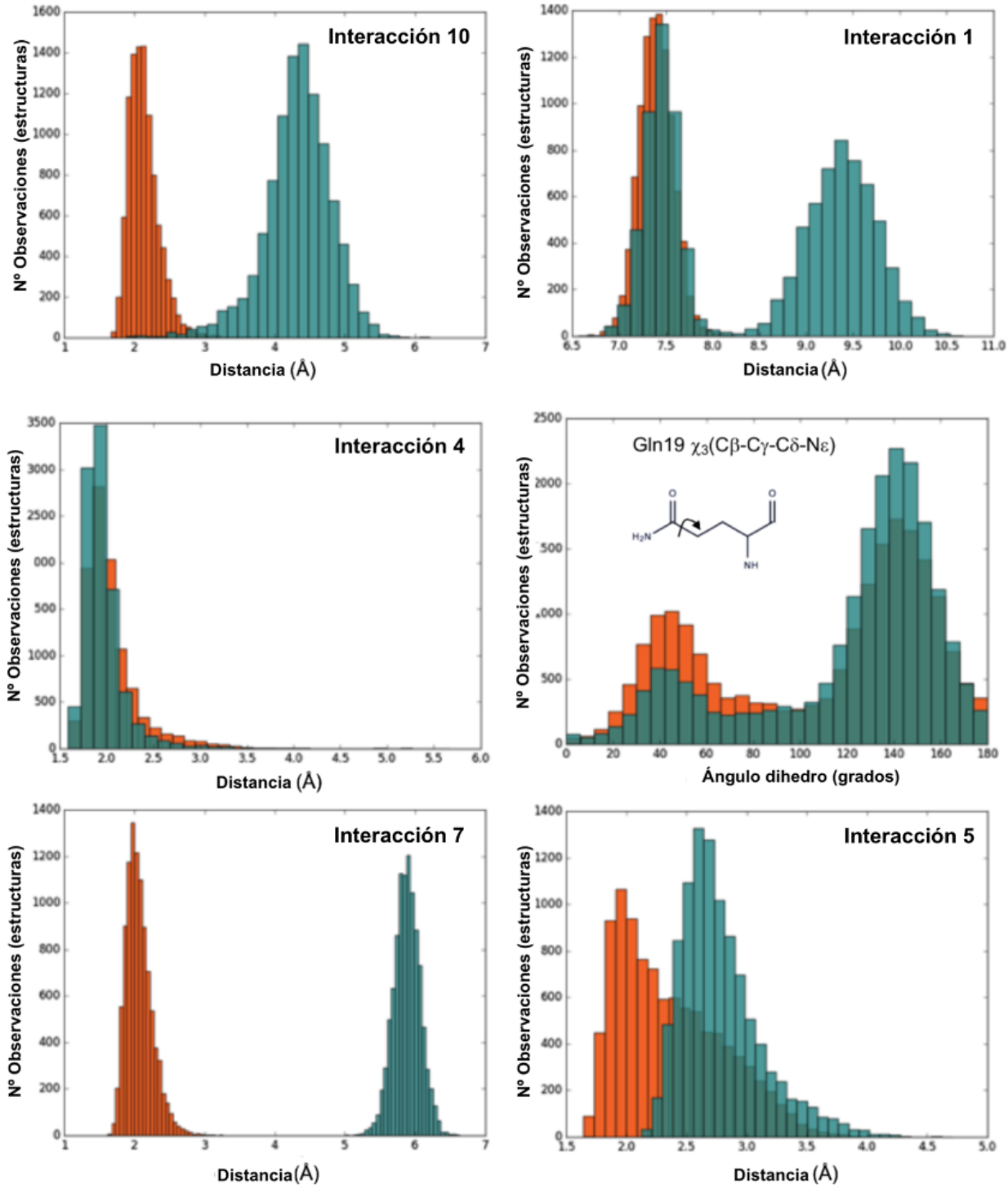


Figura 4.8 Histogramas de distancia para interacciones seleccionadas de los complejos Cz-6b (anaranjado) y Cz-8d (azul). Además, se muestra el histograma para el ángulo de torsión de la cadena lateral Gln19. Las distancias correspondientes a la interacción 1 se midieron entre los centros de masa de los residuos involucrados.

Como se evidencia en la Figura 4.8, varias interacciones muestran una distribución bimodal de frecuencias en las que se forman o se rompen, lo que concuerda con el modelo conformacional

de dos estados finales propuesto basado en el análisis de densidad de carga de estructuras seleccionadas de diferentes simulaciones de DM. La distribución de distancias de la interacción 1 en el complejo Cz-8d hace evidentes los dos estados conformacionales del residuo His162. En ese complejo, His162 está aproximadamente la mitad del tiempo lejos de Asn182 como en la conformación de Cz estabilizada por inhibidores menos activos. Por otro lado, en el complejo Cz-6b, His162 está cerca de Asn182 durante todo el tiempo de simulación, favoreciendo así la formación de la interacción 1 como en la conformación estabilizada por la mayoría de los inhibidores de Cz activos.

Además, la interacción 10 que está involucrada en la conformación del bucle$_{56-68}$ se forma la mayor parte del tiempo de la simulación en el complejo Cz-6b, estabilizando así la forma cerrada del bucle. Esto no ocurre en los complejos inactivos ya que las mismas interacciones se rompen durante la simulación del complejo Cz-8d.

Como se discutió anteriormente, como consecuencia de la reorganización del bucle$_{56-68}$ al pasar del complejo con inhibidores de Cz menos activos a la mayoría de los activos, el bucle$_{11-23}$ también se desplaza hacia arriba y arrastra con él la cadena lateral Gln19 a través de la interacción 14 (no se muestra). Concretamente, el movimiento de arrastre implica la torsión de la cadena lateral Gln19 que se evidencia por la población bimodal del ángulo de torsión X^3 donde la conformación torcida está representada por la distribución alrededor de 40°.

Puede verse en la Figura 4.8 que la cadena lateral de Gln19 permanece más tiempo retorcida en el complejo Cz-6b que en Cz-8d. En cuanto a la interacción 4 entre la amida de la cadena lateral Gln19 y el átomo de oxígeno del grupo sulfonilo de los inhibidores, la misma permanece formada todo el tiempo de simulación. Sin embargo, la distribución de la distancia se desplaza ligeramente hacia distancias de interacción más grandes en el complejo Cz-6b, lo que probablemente sea una consecuencia de la torsión duradera de la cadena lateral Gln19 que la coloca más lejos de los oxígenos sulfonílicos. Como se discutió anteriormente, el debilitamiento de la interacción 4 podría contribuir a un mejor ajuste general del inhibidor dentro del sitio catalítico de Cz.

Finalmente, las distribuciones de distancia correspondientes a las interacciones 7 y 5 muestran que 6b está más firmemente unido que 8d a la cadena principal de Gly66 y Asp161,

respectivamente, lo que también está en línea con el análisis de densidad de carga anterior en estructuras seleccionadas de diferentes simulaciones DM.

4.3.6. Descomposición de la afinidad de unión por sub-bolsillos

Dado que la densidad de carga, medida en el punto crítico de interacción, es una propiedad topológica local, podemos calcular la contribución de un subconjunto de dichos valores de densidad de carga a la fuerza de anclaje total del inhibidor. De esa manera, podríamos saber en cuál de los sub-bolsillos enzimáticos, es necesario mejorar las interacciones con el inhibidor.

La figura 4.9 muestra la sumatoria de la densidad de carga por sub-bolsillo. Es decir, los valores de BCP por sub-bolsillo en los complejos Cz-Inh estudiados.

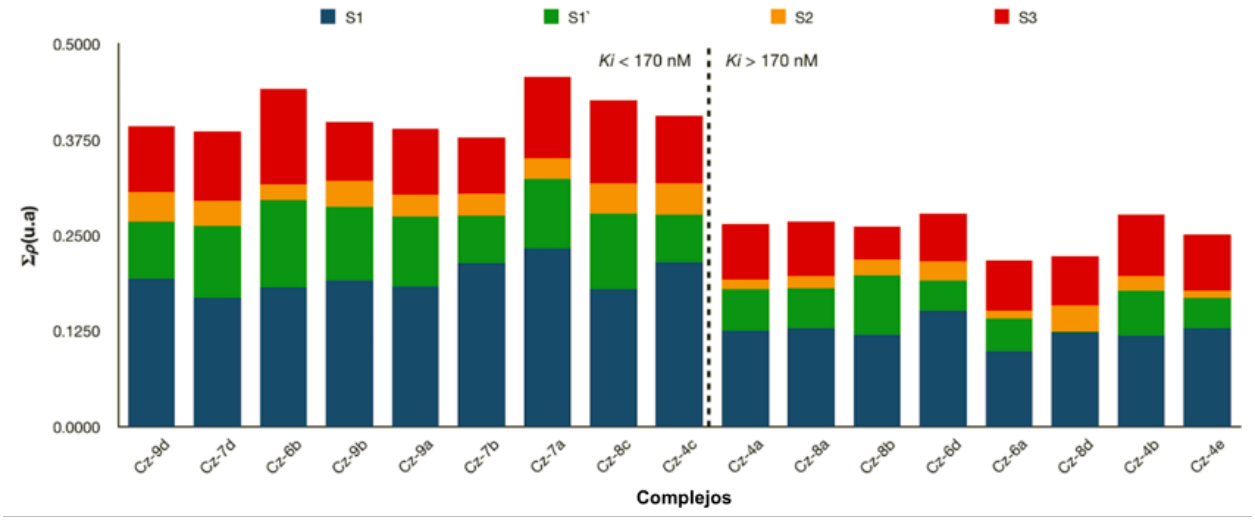


Figura 4.9 Suma de los valores de densidad de carga en los BCP debido a interacciones intra-intermoleculares en complejos Cz-inh. Los valores se dividen en cuatro contribuciones correspondientes a los sub-bolsillos S1 (azul), S1′ (verde), S2 (anaranjado) y S3 (rojo). De izquierda a derecha, los complejos se ordenan en valores crecientes de K_i. Los complejos se dividen, con una línea de puntos, en dos grupos según el valor del umbral de decisión (K_i 170 nM) utilizado en la sección SVM. La nomenclatura utilizada se extrajo de Jaishankar (2008).

Como se puede ver en la Figura 4.9, al pasar de los inhibidores menos activos a los más activos, la fuerza de anclaje del inhibidor mejora no en un sub-bolsillo en particular sino en todos los sub-bolsillos enzimáticos. Este hallazgo está en línea con nuestros resultados anteriores. Hemos visto que la sustitución en las posiciones P2 y P3 del inhibidor no sólo induce cambios en los

sub-bolsillos de enzima S3 y S2, sino que todo el sitio catalítico se ve “afectado” por tales sustituciones. Esta fuerte comunicación entre los diferentes sub-bolsillos enzimáticos anticipa que la optimización de interacciones por separado en cada uno de los sub-bolsillos de la enzima podría ser difícil de lograr. De manera similar, un enfoque basado en fragmentos para el diseño de fármacos también sería un desafío por la misma razón. Al seguir este enfoque, para el descubrimiento de nuevos inhibidores de Cz, uno presumiblemente comenzaría con un pequeño fragmento capaz de unirse al sub-bolsillo S2 (es decir, el sub-bolsillo más fácil de apuntar) y desde allí tendría que agrandarse hacia los sub-bolsillos vecinos, ya sea por el enfoque de crecimiento de fragmentos o de enlace de fragmentos. Debido a la fuerte interrelación entre sub-bolsillos que hemos mostrado a lo largo de este trabajo, no hay garantía de que al agrandar el fragmento en S2 hacia S3, por ejemplo, se mantengan las interacciones anteriores en S2.

4.4 Conclusiones

En esta parte de la tesis se ha realizado un estudio mecano-cuántico riguroso de la red de interacciones moleculares en 17 complejos de Cz con inhibidores conocidos de la enzima, con el objetivo de interpretar las diferencias de actividad en términos de densidades electrónicas, y así indagar/investigar sobre el mecanismo de inhibición de la enzima. Este tipo de estudio no había sido realizado previamente. Para extraer la información se utilizaron herramientas de aprendizaje automático.

QTAIM proporcionó los elementos topológicos de la densidad de carga que describen las interacciones en los complejos Cz-Inh. En este punto, con más de trescientas interacciones por complejo, se entrenó un modelo de clasificación de aprendizaje supervisado con RFE capaz de discriminar entre las interacciones presentes en los complejos de los inhibidores más activos (interacciones de tipo activo) y las que ocurren en los menos activos (interacciones de tipo inactivo). Además, el modelo también proporcionó información sobre la importancia de cada interacción, es decir, cuáles son las interacciones más importantes para discriminar entre complejos con inhibidores activos e inactivos (o menos activos) frente a Cz.

Nuestro modelo nos permitió señalar 19 interacciones principales, tanto intermoleculares como intramoleculares, que podrían explicar los principales cambios en los complejos analizados.

Entre las interacciones intermoleculares, las interacciones entre cadenas principales 5, 6, 7 y 8, así como las interacciones de los residuos de inhibidores en posiciones P2 y P3 con la cadena lateral Leu67, juegan un papel clave en el anclaje adecuado de la mayoría de los inhibidores activos en el sitio catalítico de la enzima. Desafortunadamente, no se encontró una relación cuantitativa entre la estructura y los datos de actividad al considerar sólo las interacciones intermoleculares.

Teniendo en cuenta también las interacciones intramoleculares y con la ayuda del modelo SVM-RFE para separar las interacciones activas de las inactivas, surge un mecanismo más indirecto de inhibición enzimática que implica amplios cambios conformacionales dentro de la estructura de la proteína.

Estos cambios conformacionales de la proteína ocurren en ambas "paredes" del sitio catalítico, promovidos por interacciones intermoleculares en los sitios S2 y S3. A su vez, las interacciones en el sub-bolsillo S2 desencadenan cambios conformacionales en la lámina-$\beta_{161-170}$ (pared derecha), mientras que las interacciones en el subsitio S3 impulsan principalmente cambios conformacionales en el bucle$_{56-68}$ (pared izquierda). Ambos cambios conformacionales conducen finalmente a un reordenamiento de los residuos His162 y Gln19 en el subsitio S1' que permite el posicionamiento adecuado del segmento vinilsulfona y la formación de interacciones clave entre las cadenas principales de los inhibidores tipo péptido y los residuos que delimitan el sitio catalítico.

Por otro lado, este estudio también nos permitió comprender cuán importante es el papel del Trp184 altamente conservado, permitiendo la formación de la interacción 1 que conduce a la activación de la histidina catalítica. La "actividad de activación" de Trp184 es crucial para la correcta disposición espacial de la tríada catalítica. Las diferentes interacciones "orquestadas" por este residuo determinan la activación/inactivación de la maquinaria proteica.

En este sentido, se encontró que la mayoría de los inhibidores de Cz activos inducen una conformación en la que están presentes interacciones consideradas como distintivas del evento de reconocimiento del sustrato. Es importante destacar que esta estructura de Cz "activada" puede usarse en experimentos de docking molecular, en el contexto de campañas de detección

virtual prospectivas para hallar inhibidores de Cz altamente activos en bases de datos de compuestos.

Además, entre las interacciones relevantes que estabilizan la conformación Cz "activada", las interacciones intermoleculares como 5, 6 y 7 podrían conectarse a los algoritmos de *docking* para mejorar la función de puntuación y guiar las predicciones de dicho estudio.

Finalmente, a lo largo de este estudio, se obtuvo una idea de la fuerte comunicación que existe entre los sub-bolsillos enzimáticos. Esta propiedad podría ayudar en la elección del mejor enfoque a seguir en las campañas de detección prospectivas. Probablemente un enfoque basado en fragmentos no sea la mejor opción en este caso debido a la fuerte comunicación entre los sub-bolsillos de la enzima.

Toda la información recopilada se tendrá en cuenta en los siguientes estudios prospectivos destinados a buscar nuevos inhibidores de Cz.

Anexo capítulo *IV*

Tabla 2. Valores de densidad de carga local (ρ, en unidades atómicas) correspondientes a las interacciones discutidas en el capítulo. Los números de interacción coinciden con los de las Figuras 4.3, 4.5, 4.6 y 4.7.

	Valores de interación (u.a.)																	
Complejos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16	17	18	19
9d-Cz	0.0248	0.0218	0.0069	0.0313	0.0063	0.0198	0.01783	0.0128	0.0048	0.0296	0.0000	0.0028	0.0060	0.0398	0.0101	0.0000	0.0160	0.0080
7d-Cz	0.0325	0.0000	0.0089	0.0082	0.0283	0.0233	0.04068	0.0186	0.0065	0.0350	0.0038	0.0000	0.0087	0.0068	0.0059	0.0041	0.0069	0.0096
6b-Cz	0.0345	0.0217	0.0155	0.0171	0.0141	0.0273	0.03734	0.0193	0.0031	0.0307	0.0000	0.0000	0.0000	0.0272	0.0040	0.0000	0.0029	0.0057
9b-Cz	0.0373	0.0233	0.0000	0.0185	0.0342	0.0229	0.03104	0.0223	0.0000	0.0117	0.0068	0.0044	0.0058	0.0308	0.0051	0.0030	0.0130	0.0120
9a-Cz	0.0317	0.0328	0.0050	0.0195	0.0122	0.0268	0.03491	0.0095	0.0000	0.0033	0.0000	0.0042	0.0262	0.0318	0.0053	0.0000	0.0101	0.0025
7b-Cz	0.0000	0.0000	0.0029	0.0229	0.0221	0.0161	0.02233	0.0000	0.0000	0.0375	0.0137	0.0000	0.0205	0.0387	0.0094	0.0000	0.0116	0.0026
7a-Cz	0.0303	0.0202	0.0000	0.0336	0.0043	0.0149	0.01761	0.0117	0.0070	0.0372	0.0037	0.0047	0.0206	0.0264	0.0056	0.0000	0.0029	0.0127
8c-Cz	0.0242	0.0000	0.0000	0.0330	0.0115	0.0274	0.04080	0.0087	0.0085	0.0000	0.0000	0.0072	0.0201	0.0257	0.0077	0.0000	0.0031	0.0000
4c-Cz	0.0282	0.0338	0.0000	0.0188	0.0018	0.0197	0.03322	0.0281	0.0069	0.0572	0.0000	0.0000	0.0124	0.0337	0.0112	0.0000	0.0094	0.0000
4a-Cz	0.0262	0.0424	0.0000	0.0235	0.0198	0.0142	0.02663	0.0102	0.0177	0.0025	0.0045	0.0047	0.0065	0.0336	0.0053	0.0047	0.0017	0.0000
8a-Cz	0.0000	0.0000	0.0000	0.0311	0.0195	0.0214	0.02906	0.0172	0.0109	0.0115	0.0097	0.0064	0.0119	0.0243	0.0068	0.0059	0.0020	0.0000
8b-Cz	0.0308	0.0339	0.0000	0.0317	0.0105	0.0168	0.01944	0.0114	0.0023	0.0123	0.0046	0.0077	0.0238	0.0090	0.0038	0.0088	0.0062	0.0079
6d-Cz	0.0026	0.0000	0.0004	0.0298	0.0116	0.0171	0.01648	0.0029	0.0000	0.0254	0.0046	0.0053	0.0257	0.0163	0.0040	0.0083	0.0141	0.0000
6a-Cz	0.0000	0.0000	0.0000	0.0265	0.0047	0.0178	0.03103	0.0045	0.0000	0.0171	0.0037	0.0065	0.0000	0.0296	0.0051	0.0000	0.0007	0.0052
8d-Cz	0.0000	0.0128	0.0000	0.0278	0.0000	0.0310	0.01675	0.0011	0.0045	0.0000	0.0000	0.0299	0.0075	0.0193	0.0071	0.0056	0.0071	0.0111
4b-Cz	0.0238	0.0559	0.0000	0.0378	0.0166	0.0225	0.02303	0.0084	0.0000	0.0000	0.0000	0.0066	0.0097	0.0210	0.0026	0.0033	0.0052	0.0042
4e-Cz	0.0264	0.0368	0.0000	0.0236	0.0057	0.0281	0.02357	0.0064	0.0141	0.0000	0.0061	0.0000	0.0124	0.0310	0.0000	0.0000	0.0000	0.0000

Referencias del capítulo 4

Arafet, Kemel, Katarzyna Świderek, and Vicent Moliner. 2018. “Computational Study of the Michaelis Complex Formation and the Effect on the Reaction Mechanism of Cruzain Cysteine Protease.” ACS Omega 3 (12): 18613–22. https://doi.org/10.1021/acsomega.8b03010.

Bader, R. F. W. Atoms in Molecules: A Quantum Theory; OxfordUniversity Press, 1990

Bolón-Canedo, Verónica, and Amparo Alonso-Betanzos. 2019. “Ensembles for Feature Selection: A Review and Future Trends.” *Information Fusion* 52 (December): 1–12. https://doi.org/10.1016/j.inffus.2018.11.008.

Butcher, Brandon, and Brian J. Smith. 2020. “Feature Engineering and Selection: A Practical Approach for Predictive Models.” *The American Statistician* 74 (3): 308–9. https://doi.org/10.1080/00031305.2020.1790217.

Case, D. A., V. Babin, Josh Berryman, R. M. Betz, Q. Cai, D. S. Cerutti, T. E. Cheatham III, et al. 2014. “Amber 14.” https://orbilu.uni.lu/handle/10993/16614.

Case, David A., Thomas E. Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M. Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J. Woods. 2005. “The Amber Biomolecular Simulation Programs.” *Journal of Computational Chemistry*. NIH Public Access. https://doi.org/10.1002/jcc.20290.

de Almeida, G. C., de Oliveira, G. B., da Silva Monte, Z., Costa, É. C. S., da Silva Falcão, E. P., Scotti, L., & de Melo, S. J. (2023). Structure-based design, optimization of lead, synthesis, and biological evaluation of compounds active against Trypanosoma cruzi. *Chemical Biology & Drug Design*, 102(4), 843-856.

Frisch, M. J., G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, et al. 2016. “Gaussian 09, Revision A.02.” Gaussian, Inc.

Fujita, Toshio, and David A. Winkler. 2016. “Understanding the Roles of the ‘Two QSARs.’” *Journal of Chemical Information and Modeling*. American Chemical Society. https://doi.org/10.1021/acs.jcim.5b00229.

Guyon, Isabelle, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. “Gene Selection for Cancer Classification Using Support Vector Machines.” *Machine Learning* 46 (1–3): 389–

422. https://doi.org/10.1023/A:1012487302797.

Jaishankar, Priyadarshini, Elizabeth Hansell, Dong Mei Zhao, Patricia S. Doyle, James H. McKerrow, and Adam R. Renslo. 2008. “Potency and Selectivity of P2/P3-Modified Inhibitors of Cysteine Proteases from Trypanosomes.” *Bioorganic and Medicinal Chemistry Letters* 18 (2): 624–28. https://doi.org/10.1016/j.bmcl.2007.11.070.

Jolliffe, Ian T., and Jorge Cadima. 2016. “Principal Component Analysis: A Review and Recent Developments.” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374 (2065): 20150202. https://doi.org/10.1098/rsta.2015.0202.

Kerr, Iain D., Ji H. Lee, Christopher J. Farady, Rachel Marion, Mathias Rickert, Mohammed Sajid, Kailash C. Pandey, et al. 2009. “Vinyl Sulfones as Antiparasitic Agents and a Structural Basis for Drug Design.” *Journal of Biological Chemistry* 284 (38): 25697–703. https://doi.org/10.1074/jbc.M109.014340.

Li, Hongdong, Yizeng Liang, and Qingsong Xu. 2009. “Support Vector Machines and Its Applications in Chemistry.” *Chemometrics and Intelligent Laboratory Systems* 95 (2): 188–98. https://doi.org/10.1016/j.chemolab.2008.10.007.

Lin, Xiaohui, Chao Li, Yanhui Zhang, Benzhe Su, Meng Fan, and Hai Wei. 2018. “Selecting Feature Subsets Based on SVM-RFE and the Overlapping Ratio with Applications in Bioinformatics.” *Molecules* 23 (1). https://doi.org/10.3390/molecules23010052.

Lu, Tian, and Feiwu Chen. 2012. “Multiwfn: A Multifunctional Wavefunction Analyzer.” *Journal of Computational Chemistry* 33 (5): 580–92. https://doi.org/10.1002/jcc.22885.

Maier, James A., Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. 2015. “Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB.” *Journal of Chemical Theory and Computation* 11 (8): 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

Pedregosa, Fabian, Vincent Michel, Olivier Grisel OLIVIERGRISEL, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Jake Vanderplas, et al. 2011. “Scikit-Learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot.” *Journal of Machine Learning Research*. Vol. 12. http://scikit-learn.sourceforge.net.

Rodriguez-Pérez, Raquel, Martin Vogt, and Jürgen Bajorath. 2017. “Support Vector Machine Classification and Regression Prioritize Different Structural Features for Binary Compound Activity and Potency Value Prediction.” *ACS Omega* 2 (10): 6371–79. https://doi.org/10.1021/acsomega.7b01079.

Tripathi, R. K. P., Dey, R., & Das, N. (2023). Identification of natural lead molecules as potential Trypanosoma cruzi cruzipain inhibitors and decoding the interaction mechanism for the treatment of Chagas disease: a computational biology analysis. *Natural Product Research, 1–5*. https://doi.org/10.1080/14786419.2023.2256018

Turk, Dušan, Gregor Gunčar, Marjetka Podobnik, and Boris Turk. 1998. “Revised Definition of Substrate Binding Sites of Papain-Like Cysteine Proteases.” *Biological Chemistry* 379 (2): 137–47. https://doi.org/10.1515/bchm.1998.379.2.137.

Wang, Junmei, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case. 2004. “Development and Testing of a General Amber Force Field.” *Journal of Computational Chemistry* 25 (9): 1157–74. https://doi.org/10.1002/jcc.20035.

Zhou, Xin, and David P. Tuck. 2007. “MSVM-RFE: Extensions of SVM-RFE for Multiclass Gene Selection on DNA Microarray Data.” *Bioinformatics* 23 (9): 1106–14. https://doi.org/10.1093/bioinformatics/btm036.

CAPÍTULO V

“Cribado virtual retrospectivo de una biblioteca de ligandos”

5.1 Introducción

El objetivo principal del proceso de descubrimiento de fármacos es encontrar un compuesto químico que pueda ajustarse/encajar tanto geométrica como químicamente en una cavidad específica del blanco molecular. Los métodos de diseño de fármacos convencionales incluyen el cribado automático de alto rendimiento (HTS, del inglés *High Throughput Screening*) de bibliotecas de moléculas pequeñas para identificar, mediante pruebas biológicas y ensayos de afinidad, aquellas capaces de unirse a un determinado objetivo molecular, generalmente un receptor proteico o una enzima. Aunque las campañas de HTS filtran desde decenas de miles hasta millones de compuestos, dado que el espacio químico es tan amplio (~10^{60} moléculas), cualquier colección de compuestos cubre una porción insignificante de este espacio. Esto explica por qué la mitad de los experimentos de HTS fallan. Además, HTS es una técnica costosa y requiere una inversión sustancial en infraestructura y desarrollo de ensayos. Las desventajas antes mencionadas, así como el atractivo de un enfoque más determinista (es decir, más racional) para combatir las enfermedades, dieron origen al Diseño de Fármacos Asistido por Computadora (DFAC) (Baldi 2010).

La idea detrás del cribado virtual es probar primero los compuestos computacionalmente para reducir la cantidad de candidatos que necesitan ser evaluados experimentalmente, reduciendo así el tiempo y el costo de los experimentos físicos (Liao et al. 2013).

Actualmente, el *docking* molecular es la técnica de cribado virtual (CV) más popular para priorizar compuestos candidatos a partir de grandes conjuntos de datos. Sin embargo, a pesar de ser ampliamente utilizado, la precisión de esta metodología es sustancialmente baja en comparación con otras técnicas computacionales más robustas (por ejemplo Dinámica molecular o métodos híbridos QM/MM) como resultado de una compensación entre precisión y la velocidad requerida en campañas de cribado de alto rendimiento. Como consecuencia, los resultados del CV a menudo están plagados de altas tasas de falsos positivos, es decir, muchos compuestos que se clasifican como activos frente a una determinada proteína, en realidad no muestran actividad (N. Deng et al. 2015).

A su vez, una de las técnicas para tratar de mejorar el nivel de acierto del *docking* consiste en tratar de recuperar, a partir de una biblioteca de compuestos, aquellos que se sabe son activos por el blanco molecular en estudio, lo que se conoce como Cribado Virtual Retrospectivo (CVR).

Típicamente, las bibliotecas que se utilizan en el CVR se construyen sembrando unos pocos compuestos activos frente al blanco molecular en estudio en una base de datos de compuestos inactivos (señuelos) que son estructuralmente diferentes a los activos, pero con similares propiedades fisicoquímicas.

Por otro lado, recientemente han surgido técnicas de aprendizaje profundo (DL, del inglés *Deep Learning*) como una alternativa prometedora a los enfoques de *docking* molecular. Los métodos de aprendizaje automático (ML, del inglés *Machine Learning*) se vienen empleando hace ya un tiempo para desarrollar nuevas funciones de scoring (FS) que superen las FS de los algoritmos de docking molecular (Li et al. 2020). Sin embargo, existen dos diferencias clave entre los modelos DL y ML: solo los primeros pueden aprender automáticamente las características de los datos y también logran un poder expresivo mucho mayor debido a sus arquitecturas de red inherentemente profundas (Lim et al. 2019). Ambas propiedades pueden ser explotadas para mejorar las predicciones de actividad en campañas de cribado virtual de moléculas.

El propósito de entrenar modelos de DL en problemas relacionados con el descubrimiento de fármacos es dilucidar las relaciones estructura-actividad correctas a partir de los datos existentes, de manera similar al CVR. En términos prácticos, esto significa encontrar una función f capaz de realizar el mapeo $Y = f(X)$ entre la estructura X y las actividades biológicas Y de los compuestos químicos.

Los enfoques de DL basados en redes neuronales con capas completamente conectadas (Fully Connected Layers) se han utilizado ampliamente para codificar las características moleculares en forma de vectores, pero en la práctica, estos enfoques inevitablemente descartan parte de la información estructural (Na, Chang, y Kim 2020). En cambio, las Redes Convolucionales basadas en Grafo (GCN, del inglés *Graph Convolutional Network*) , diseñadas para trabajar directamente en grafos, son la elección natural para procesar la información contenida en las estructuras químicas, ya que los átomos y los enlaces en las moléculas se pueden representar como nodos y bordes de grafos, respectivamente.

Las GCNs han ganado mucha atención en varias aplicaciones químicas, como la predicción de propiedades moleculares (Wieder et al. 2020; Korolev et al. 2020), diseño molecular (Mercado et al. 2021), reacciones químicas (Coley et al. 2019), entre otras. Por ejemplo, Ryu et al. (2018) desarrollaron varios tipos de GCN para la predicción de propiedades fisicoquímicas de moléculas , desde una GCN estándar con una arquitectura muy similar a la propuesta por Kipf y Welling (2016) hasta versiones aumentadas que incorporan mecanismos de atención y de compuerta (gate).

A primera vista, la predicción de las propiedades fisicoquímicas de las moléculas (es decir, el coeficiente de partición octanol-agua (logP), el área de superficie topológica (TPSA), etc., que están codificadas en gran medida en la propia topología de los ligandos, parece ser una tarea menos compleja en comparación con la predicción de sus actividades farmacológicas, que dependen en gran parte de las interacciones proteína-ligando.

Sin embargo, Sakai et al. (2021) demostraron recientemente que las GCN que se basan únicamente en la información estructural 2D de los compuestos pueden predecir no solo las propiedades fisicoquímicas, sino incluso la actividad de los compuestos contra un blanco molecular particular de interés. Han demostrado que los modelos de GCN construidos únicamente a partir de la información estructural bidimensional de los compuestos presentaron un alto grado de predictibilidad de la actividad contra 127 blancos moleculares diversos de la base de datos ChEMBL. Por lo tanto, si hay suficientes datos experimentales disponibles y hay suficientes capas ocultas de nodos, una simple representación 2D podría predecir cuantitativamente la actividad con una precisión satisfactoria.

Teniendo en cuenta el estado del arte en modelos de DL para la predicción de actividad, decidimos realizar un estudio de evaluación comparativa de una red convolucional basada en grafos (GCN) frente a un cribado de *docking* molecular, para la predicción de actividad de compuestos sobre cruzipaina.

Es importante tener en cuenta que el *docking* molecular es un enfoque basado en la estructura, que tiene en cuenta explícitamente las interacciones moleculares del ligando con la proteína, mientras que la GCN que hemos entrenado solo considera la topología del ligando (es decir, el enfoque basado en el ligando).

Hay algunas implementaciones de GCN donde se entrenan por separado las redes para los pares interactuantes , ya sea proteína/ligando (Torng y Altman 2019) o proteína / proteína (Fout et al. 2017), que luego se pasan a una red neuronal del tipo Capas Totalmente Conectadas (FCL, del inglés Fully Connected Layer) para hacer la predicción de afinidad de unión. Este enfoque se asemeja a las soluciones de *docking* molecular en las que un ligando se acopla sobre la estructura del blanco molecular en estudio y luego una función de scoring clasifica las poses (en esta analogía, la FCL sería la función de scoring). Sin embargo, ninguna de esas implementaciones tiene en cuenta realmente las interacciones intermoleculares de manera explícita.

Por otro lado, Lim et al. (2019) propusieron un modelo de GCN que puede extraer interacciones intermoleculares como características de grafo (features) directamente de la información estructural 3D de la pose de unión proteína-ligando. Sin embargo, esta implementación podría requerir la disponibilidad de cientos de estructuras 3D de complejos proteína-ligando conocidas para entrenar la red. En el caso de nuestro blanco de estudio, Cz, solo había 37 estructuras depositadas en el Protein Data Bank (a mayo de 2022), lo que es una cantidad insuficiente para arquitecturas de aprendizaje profundo “hambrientas de datos”.

Por el contrario, las implementaciones de GCN basadas en ligandos se basan únicamente en la topología de las moléculas de ligandos y sus correspondientes anotaciones de actividad sobre el blanco molecular en cuestión, las cuales están ampliamente disponibles para cruzipaína en bases de datos públicas como PubChem, CHEMBL, entre otras.

En consecuencia, en este capítulo llevamos a cabo un CVR de un gran conjunto de compuestos de un ensayo de cribado cuantitativo de alto rendimiento para inhibidores de cruzipaina (PubChem AID 1478), mediante ambos enfoques, *docking* molecular basado en estructura y GCN basado en ligando.

El objetivo del estudio fue evaluar comparativamente la capacidad de ambos enfoques para realizar la tarea de clasificación (es decir, discriminar entre compuestos activos e inactivos) y cómo podrían complementarse entre sí para reducir los falsos positivos y los falsos negativos en el contexto de las campañas de cribado virtual.

5.2 Metodología

La figura 5.1 muestra el procedimiento general para el cribado virtual retrospectivo (CVR) de ligandos con actividad sobre Cz a partir de la base de datos AID-1478 mediante la utilización de los dos enfoques, GCN (CVR-GCN) y *docking* molecular (CVR-*docking*). Para CVR-GCN, el conjunto de datos debe dividirse en (al menos) conjuntos de entrenamiento y prueba, y solo el último debe usarse para la evaluación del modelo. En cuanto CVR-*docking*, en principio se puede utilizar todo el conjunto de datos para la evaluación del rendimiento, debido a que los compuestos son "ranqueados" con una función de scoring integrada al algoritmo de *docking* que fue previamente calibrada.

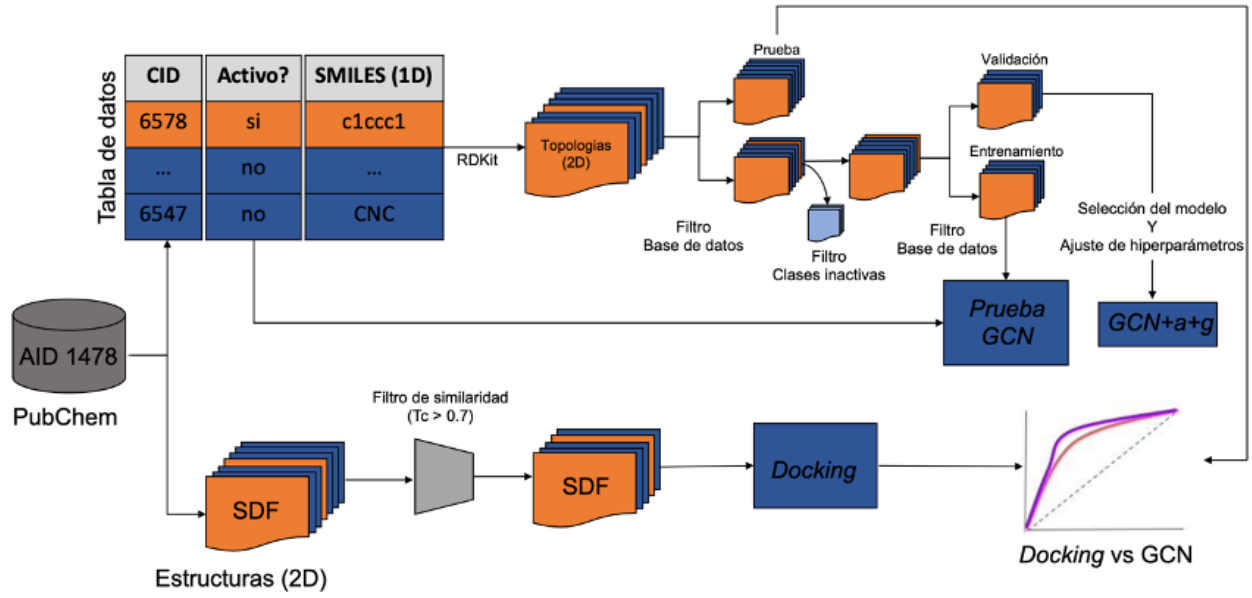


Figura 5.1 Flujo de trabajo para la detección virtual de ligandos activos, mediada por GCN y *docking* molecular, a partir de la base de datos AID-1478.

5.2.1 Conjunto de datos AID-1478

La base de datos de PubChem AID-1478 está compuesta por 197.846 moléculas obtenidas a partir en un experimento cuantitativo de cribado de alto rendimiento (qHTS, del inglés *Quantitative High-Throughput Screening*) contra cruzipaina (https://pubchem.ncbi.nlm.nih.gov/bioassay/1478). Se compiló de PubChem una tabla de datos que contiene información referente a cada molécula, como ser, un identificador del compuesto (por ejemplo: CID 1388525), la representación unidimensional de las estructuras (SMILES

isoméricas) y la actividad para cada compuesto en el conjunto de datos. También se recuperaron de la base de datos las estructuras 3D en formato SDF (ver Fig. 5.1).

Después de eliminar las moléculas duplicadas y las moléculas con actividad no concluyente, quedaron un total de 193.973 compuestos, de los cuales solo 848 eran inhibidores reales de cruzipaina.

5.2.2 *Docking* molecular

El *docking* molecular de los compuestos en el sitio activo de una estructura de la cruzipaina (Cz) recuperada del *Protein Data Bank* (código PDB 2OZ2) se realizó con el software de *docking* rDock (Ruiz-Carmona et al. 2014). El ligando, de tipo peptídico co-cristalizado con Cz en esa estructura, se usó para definir la región activa para las corridas de *docking*.

Además del conjunto de datos AID-1478 completo y preprocesado, se construyó un subconjunto de datos filtrado por similitud mediante la eliminación de compuestos inactivos muy similares a los activos que el algoritmo de *docking* podría no distinguir correctamente. Para ello, mediante la utilización del software openbabel (O'Boyle et al. 2011) se calcularon las fingerprints 1D de cada compuesto, descartando aquellos inactivos con un coeficiente de similitud de Tanimoto mayor a 0,70 con respecto a los activos. De este proceso, solo 77.719 ligandos inactivos pasaron el filtrado.

Las poses de *docking* se clasificaron de acuerdo a los valores obtenidos a partir de la función de *scoring* por defecto en rDock. La capacidad del algoritmo para priorizar los ligandos activos entre los compuestos mejor clasificados se evaluó para diferentes valores de corte de scores de docking, mediante el cálculo del área bajo la curva ROC (AUC).

5.2.3. Preparación de conjuntos de datos para la creación de modelos de GCN

Las representaciones moleculares 2D introducidas en el modelo de GCN se construyeron a partir de sus SMILES isoméricos 1D (ver figura 5.1), con la ayuda de RDKit (https://www.rdkit.org) en Python.

Un problema con el conjunto de datos AID-1478 es el grave desbalance de clases entre los ligandos activos (~800) e inactivos (~200K), lo que podría llevar al modelo de GCN a ignorar por completo la clase minoritaria, en la que las predicciones son más importantes para priorizar los compuestos activos.

Para superar este inconveniente, realizamos un submuestreo aleatorio de la clase mayoritaria para reducir el número de ejemplos inactivos hasta lograr una distribución de clases en una proporción 1:2 (es decir, 1 ligando activo cada 2 ligandos inactivos). Es importante tener en cuenta que el cambio en la distribución de clases sólo se aplicó al conjunto de datos de entrenamiento. El submuestreo no se aplicó al conjunto de prueba utilizado para evaluar el rendimiento del modelo (Figura 5.1).

El conjunto de datos se dividió en dos sets, para entrenamiento y prueba del modelo (split 80/20), respectivamente. Se aplicó un *split* estratificado para preservar las mismas proporciones de ejemplos en cada clase, como en el conjunto de datos original.

Luego de realizar el submuestreo en el conjunto reservado para entrenamiento, este se dividió nuevamente en conjuntos de entrenamiento y validación (división 80/20), y el rendimiento del modelo se evaluó en el set de validación durante el entrenamiento (figura 5.1).

Todos estos pasos de preparación de datos se realizaron con la biblioteca de aprendizaje automático para Python, scikit-learn (Pedregosa et al. 2011).

5.2.4 Red convolucional gráfica (GCN)

5.2.4.1 Representación gráfica de moléculas

Los átomos y enlaces en una molécula se pueden representar mediante nodos y aristas en un grafo del tipo $G = (A, X)$, donde: (i) X es una matriz de características de entrada de dimensiones NxF (N es el número de nodos y F es el número de características de entrada para cada nodo) y (ii) A es una matriz de NxN , definida como la matriz de adyacencia A de G (Kipf y Welling 2016), que contiene la conectividad de los átomos en la molécula.

La figura 5.2 muestra la representación gráfica de la *n*-propilamina. La matriz de adyacencia A, así como las características de entrada de cada átomo, incluido el tipo de átomo, la cantidad de

hidrógenos unidos, la cantidad de enlaces y la aromaticidad se calcularon con el módulo de quimioinformática de Python RDKit.

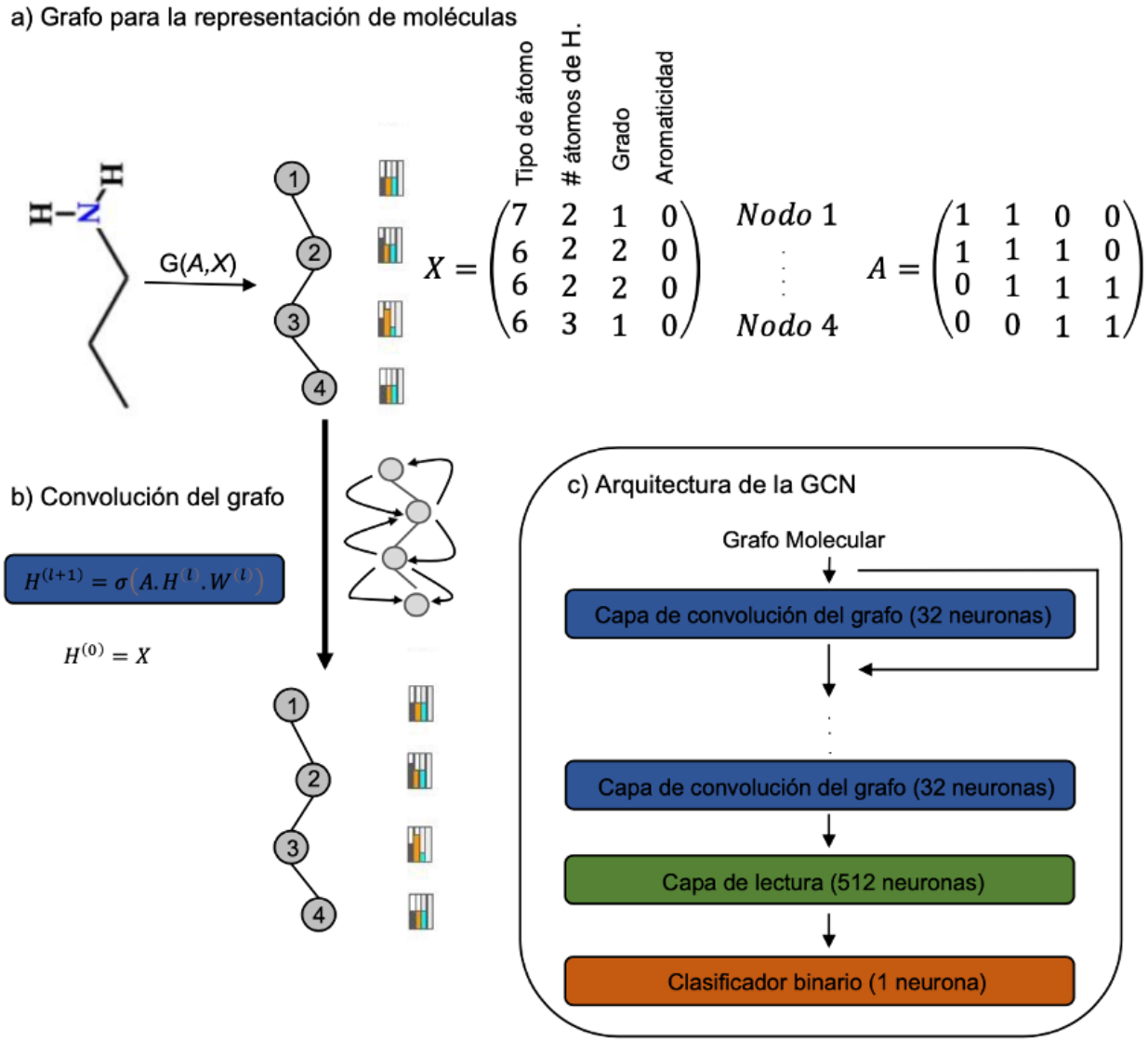


Figura 5.2 (a) Representación gráfica de la molécula de *n*-propilamina. (b) Convolución posterior para propagar las características del átomo a través de los nodos vecinos. (c) Arquitectura general de la GCN standard o vanilla, la cual consta de varias capas convolucionales con saltos de conexión (*skip-connection*), seguidas de una capa de lectura (*readout*) que resume todos los nodos del grafo para su clasificación posterior. Los detalles sobre las versiones de GCN aumentadas se pueden encontrar en (Ryu et al. 2018).

5.2.4.2 Convolución del grafo

La convolución del grafo se realiza aplicando la regla de propagación

$$H^{(l+1)} = \sigma\left(A \cdot H^{(l)} \cdot W^{(l)}\right) \quad (5.1)$$

donde σ, $H^{(l)}$ y $W^{(l)}$ denotan la función de activación, la matriz de características y la matriz de peso en la capa l, respectivamente (con $H^{(0)} = X$, es decir, la matriz de características de entrada). En cada capa, las características $H^{(l)}$ se agregan para formar las características de la siguiente capa $H^{(l+1)}$ utilizando la regla de propagación σ (Fig. 5.2b).

De esta forma, la capa convolucional del grafo representa cada nodo como un agregado de su vecindad. Para reflejar las características del átomo a largas distancias de un átomo central específico, se debe aplicar un número múltiple de convoluciones de grafos.

5.2.4.3 Arquitectura de la GCN

Las GCNs se pueden diseñar para realizar tareas a nivel de nodo o de grafo completo. La clasificación a nivel de grafo tiene como objetivo predecir la clase para un grafo completo. El aprendizaje de extremo a extremo para esta tarea se puede realizar con una combinación de capas convolucionales de grafos, opcionalmente capas de *dropout* y capas de lectura. Las capas convolucionales son responsables de extraer representaciones de nodos de alto nivel y la capa de lectura agrupa las representaciones de nodos en una única representación para cada grafo.

En este trabajo, implementamos una versión modificada de la GCN standard (vanilla) y las correspondientes versiones aumentadas desarrolladas originalmente por Ryu et al. (2018). La Figura 2c muestra la arquitectura general de la GCN vanilla implementada en este trabajo.

En cuanto a las GCN aumentadas, las mismas incorporan variantes del tipo *gate* (GCN+g), *attention* (GCN+a) o ambos mecanismos simultáneamente (GCN+a+g). Aunque el GCN del tipo vanilla incorpora un salto de conexión (*skip-connection*) para evitar el problema de desvanecimiento/fuga de gradiente, todavía tiene problemas para regular la mejor tasa de actualización. Por lo tanto, también se consideró un mecanismo de conexión de salto controlado (es decir, GCN+g) que se encuentra en la implementación de Ryu et al, (2018). Además, un mecanismo de atención (es decir, GCN+a) permite que el modelo se centre en partes relevantes de las entradas y logre una predicción mejor y más precisa (Xiong et al. 2020).

Dado que los valores de actividad solo están disponibles para los ~800 compuestos activos en el conjunto de datos AID-1478, no es posible entrenar una GCN para realizar predicciones cuantitativas, como estaba previsto en la GCN original. Por lo tanto, en lugar de un problema de regresión, la GCN se adaptó para realizar una clasificación binaria basada en el fenotipo de actividad del compuesto (es decir, activo o inactivo).

Además, los hiperparámetros (número de capas convolucionales de gráficos, tamaño de lote, número de épocas, etc.) de la red se ajustaron manualmente para lograr el mejor rendimiento posible (Tabla 5.1). La función de pérdida monitoreada durante las épocas de entrenamiento fue la entropía cruzada que es una métrica de monitoreo comúnmente utilizada durante los esquemas de clasificación.

La adaptación de la red y el ajuste de parámetros se realizaron a partir del repositorio de GitHub de Ryu 2018 (https://github.com/SeongokRyu/augmented-GCN).

Excepto por la función de activación de la última capa y la función de pérdida, que se modificaron para una tarea de clasificación, y el ajuste de los hiperparámetros, la arquitectura general de las GCN estándar y aumentadas que hemos entrenado sigue siendo la misma que la implementada originalmente por Ryu et al. (2018).

Tabla 5.1 Configuración de hiperparámetros del modelo de aprendizaje profundo. Los mejores valores están resaltados en negrita. GCN: red convolucional de grafos. GCN+a: red convolucional gráfica + *attention*, GCN+g: red convolucional gráfica + *gate*, GCN+a+g: red convolucional gráfica + *attention* + *gate*. En todos los escenarios, el optimizador fue Adam (Kingma y Ba 2015).

Hiperparámetros	Valores
Arquitecturas de red	GCN, GCN+a, GCN+g, **GCN+a+g**
Capas convolucionales	3,4,**5**,6
Tasa de aprendizaje	0.01, **0.001**, 0.0001
Tamaño del lote	50, **100**, 150, 200

El conjunto de datos preprocesados y un Jupyter Notebook de demostración con los mejores hiperparámetros están disponibles en https://github.com/lemyp-cadd/gcn-docking.

5.3 Resultados y Discusión

Ferreira et al. (2010) llevaron a cabo un cribado cuantitativo de alto rendimiento (qHTS) de ~200K compuestos sobre Cz para buscar inhibidores reversibles y competitivos de la enzima. Los resultados de la campaña de qHTS se depositaron en PubChem (AID 1478). En este capítulo, explotamos esa información para evaluar retrospectivamente la capacidad de un algoritmo de *docking* y una red convolucional de grafos (GCN) para priorizar los compuestos activos del conjunto de datos.

5.3.1 CVR mediante la utilización de *docking* molecular (CVR-*docking*)

La figura 5.3 muestra el sitio catalítico de Cz unido a un inhibidor conocido, derivado de vinilsulfona (PDB: 2OZ2).

Para los cálculos de *docking*, se eliminó el inhibidor de la estructura y en su lugar se acoplaron los compuestos de la base de datos AID-1478.

La figura 5.4 muestra el rendimiento del algoritmo de *docking* para discriminar entre compuestos activos e inactivos.

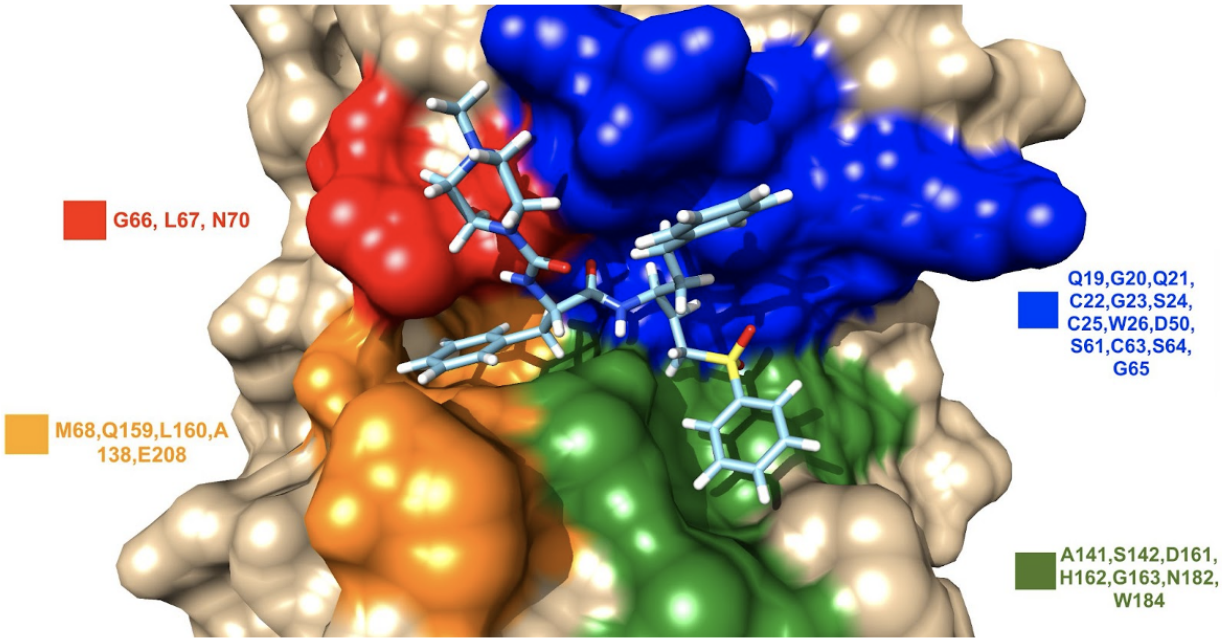


Figura 5.3 Superficie de Cz con un inhibidor del tipo péptidomimetico derivado de vinilsulfona unido a la hendidura de unión de la enzima (PDB 2OZ2). Los residuos que dan forma a los sub-bolsillos S1, S1', S2 y S3 se representan en azul, verde, naranja y rojo, respectivamente.

Como lo demuestra el AUC en la figura 5.4 (curva naranja), el algoritmo de *docking* tiene dificultades para clasificar los compuestos correctamente cuando se analiza todo el conjunto de datos. Es probable que esto se deba a que el conjunto de datos AID-1478 es una colección de compuestos analizados para la inhibición de Cz que no fue diseñado específicamente para evaluar el rendimiento de programas de docking. La inspección visual de los compuestos inactivos clasificados erróneamente como activos (falsos positivos), revela que la alta tasa de error de clasificación podría deberse en parte a la gran similitud estructural de los subconjuntos de compuestos inactivos y activos.

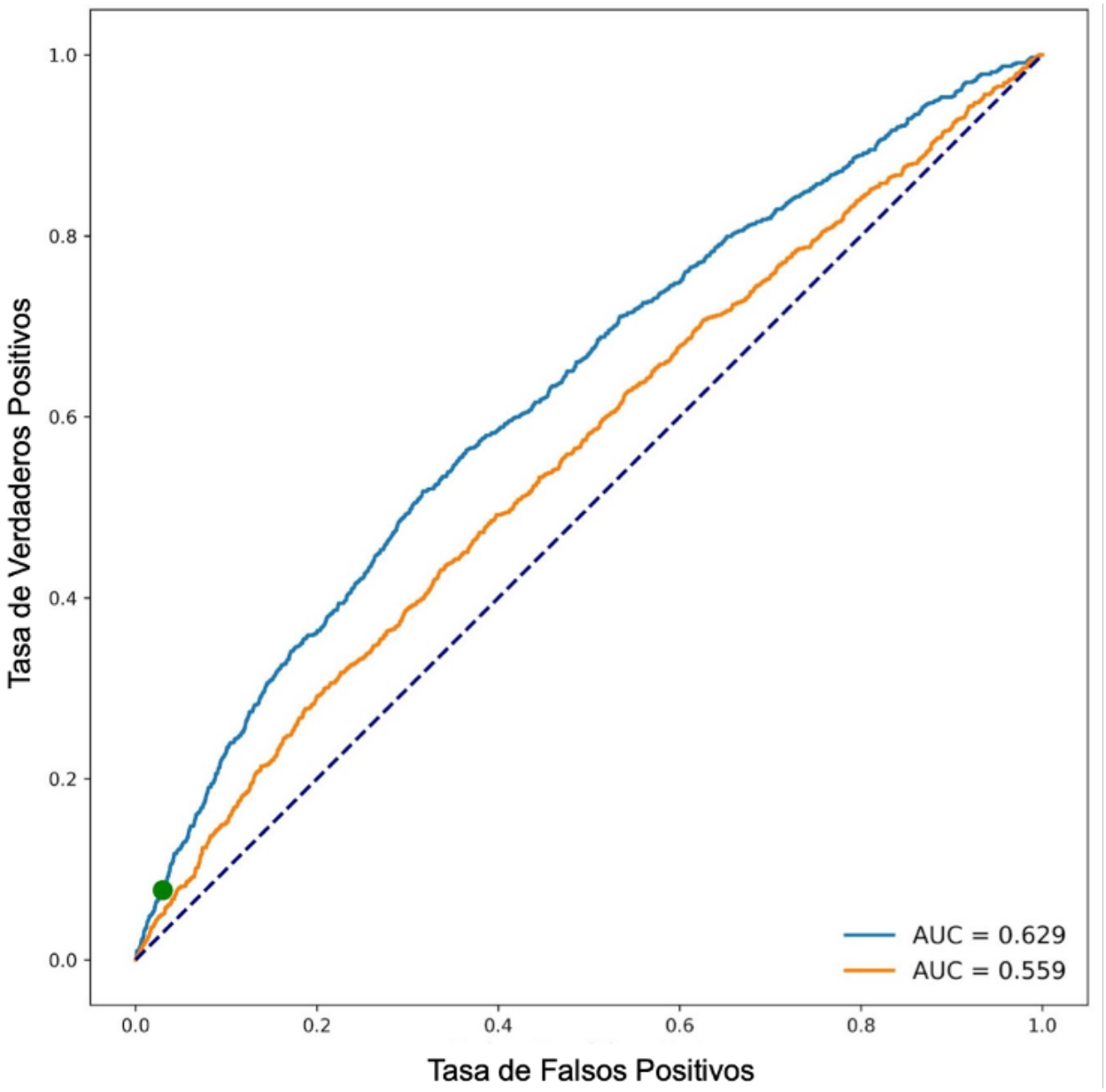


Figura 5.4 Curva ROC y área bajo la curva (AUC) logradas mediante el *docking* molecular del conjunto de datos AID-1478 completo (curva naranja) y filtrado por similitud (curva azul). La línea diagonal discontinua representa el rendimiento igual a la elección aleatoria. El punto verde sobre la curva azul representa el equilibrio entre la Tasa de Falsos Positivos (TFP) y la Tasa de Verdaderos Positivos (TVP) lograda por el 0,3 % mejor clasificado de la base de datos.

En una biblioteca típica para la evaluación comparativa de algoritmos de *docking*, los compuestos inactivos o "señuelos" se seleccionan aleatoriamente de grandes bibliotecas de compuestos de tal manera que los mismos cuenten con propiedades fisicoquímicas similares pero diferentes topologías 2D que los inhibidores reales.

Una de las principales debilidades de las bibliotecas señuelos es que la diferencia entre los dos espacios químicos definidos por los compuestos activos por un lado y los compuestos señuelo por el otro, puede conducir a una sobreestimación artificial del enriquecimiento. En otras palabras, pueden proporcionar una evaluación demasiado optimista del rendimiento del *docking*.

De todas formas, la predicción de actividad en conjuntos de datos qHTS "reales" como en la base de datos AID-1478 parece ser una tarea extremadamente desafiante para los algoritmos de *docking*, a juzgar por el bajo valor de AUC logrado.

En vista de estos hallazgos, decidimos ajustar ligeramente el conjunto de datos AID-1478 filtrando los compuestos inactivos cuyas fingerprints estructurales tienen un coeficiente de similitud de Tanimoto (Tc) mayor o igual que un umbral dado con respecto a los compuestos activos (ver sección 5.2.2). Este procedimiento se asemeja parcialmente a las “recetas” utilizadas comúnmente para la construcción de bibliotecas de señuelos para la evaluación comparativa en campañas de Cribado Virtual Retrospectivo (CVR) (Mysinger et al. 2012).

Al disminuir gradualmente el umbral de similitud, descubrimos que en un Tc = 0,70, el algoritmo de *docking* alcanzó un AUC ~0,63, lo que representa una mejora con respecto al rendimiento del cribado en el conjunto de datos original (figura 5.4, curva azul). Sin embargo, es claro que estos resultados están lejos de ser óptimos y nuestro protocolo requiere algún tipo de reajuste para alcanzar mejores valores de clasificación.

5.3.2 CVR mediante la utilización de GCN

Siguiendo los resultados de Sakai et al. (2021) que demostraron el potencial de las GCNs para la predicción de actividad, en esta parte del trabajo de tesis se implementó una versión modificada de las arquitecturas GCN (vanilla GCN, GCN+a, GCN+g y GCN+a+g) desarrolladas por Ryu et al. (2018) para realizar la clasificación de actividad de los compuestos en el conjunto de datos AID-1478.

La curva de aprendizaje de la figura 5.5 muestra el progreso del entrenamiento de la GCN aumentada tanto por mecanismos de *attention* como de *gate* (GCN+a+g), que fue la que mostró el mejor rendimiento (ver Tabla 5.1).

Como se puede ver en la figura 5.5, después de 40 épocas, la curva de validación alcanza una meseta, por lo cual, aumentar el número de épocas más allá de ese número daría como resultado un sobreajuste del modelo. Por lo tanto, empleamos la estrategia de "detención temprana" (*early stopping*) para reducir el sobreajuste, que es una técnica efectiva y simple para la regularización en el aprendizaje profundo. Esta técnica se basa en que, al entrenar una red neuronal profunda, el error de entrenamiento disminuirá progresivamente, pero no necesariamente ocurrirá lo mismo con el error de validación. El entrenamiento del modelo se detiene cuando la curva de validación comienza a ascender nuevamente (Courville, Goodfellow y Bengio 2016) o no resulta en una mayor mayor reducción del error. .

Si el entrenamiento continúa más allá de ese paso, los parámetros aprendidos cambiarán y el modelo finalmente sobreajustará los datos.

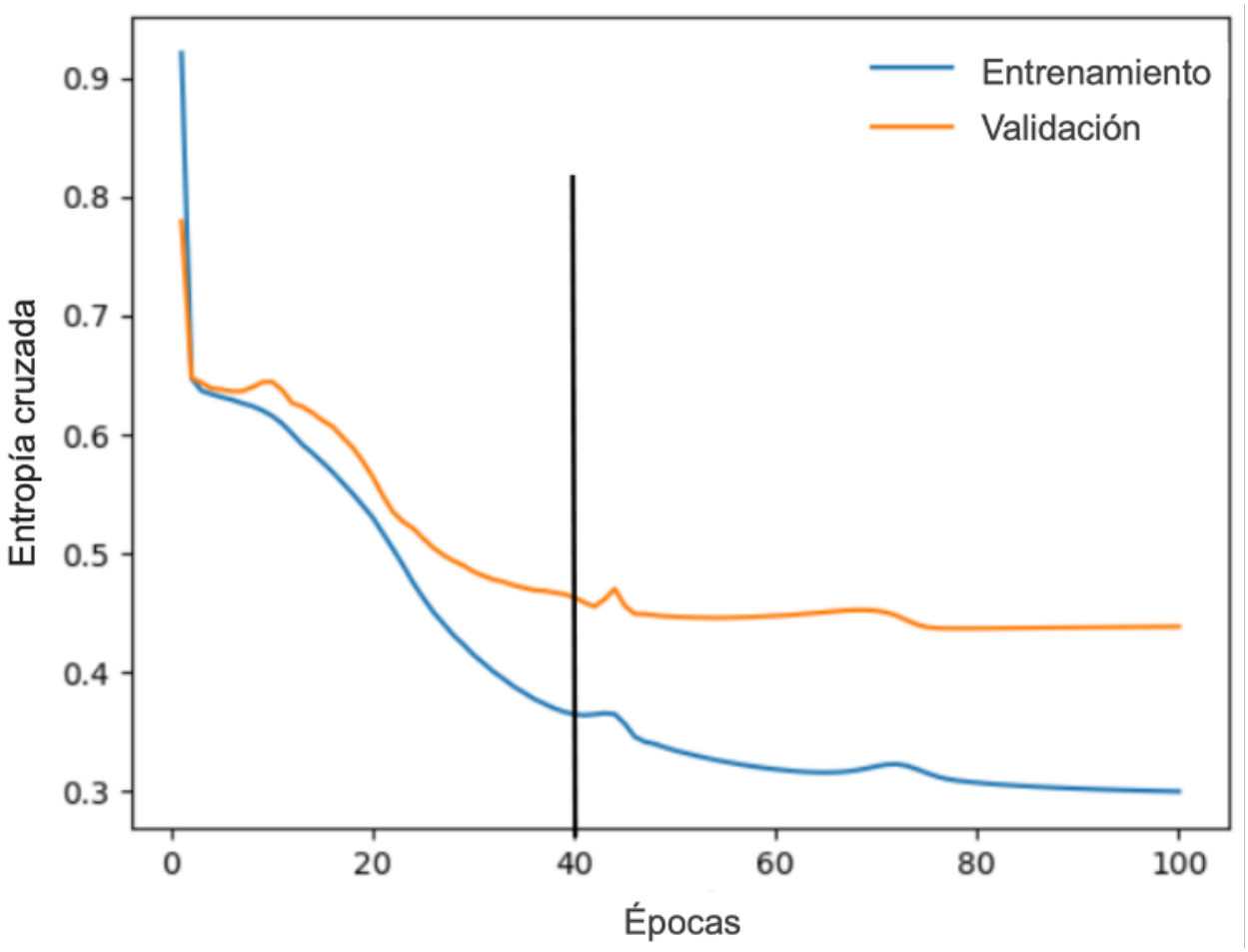


Figura 5.5 Curva de aprendizaje que muestra el progreso del proceso de entrenamiento en conjuntos de entrenamiento y validación. La línea negra muestra cuándo se detiene el entrenamiento.

El rendimiento del modelo de GCN final para clasificar las moléculas en el conjunto de prueba se evaluó mediante la métrica ROC-AUC. Como se muestra en la figura 6.6, la red de grafos GCN+a+g superó al algoritmo de *docking* al priorizar los ligandos activos del conjunto de datos AID-1478.

Es importante señalar que, a diferencia del *docking* molecular, en el caso del cribado mediado por GCN no hubo necesidad de filtrar los compuestos inactivos más similares topológicamente a los activos para lograr un rendimiento aceptable (ver figura 5.1). En otras palabras, el AUC logrado por GCN+a+g corresponde a la evaluación sobre una muestra aleatoria estratificada obtenida a partir del conjunto de datos completo.

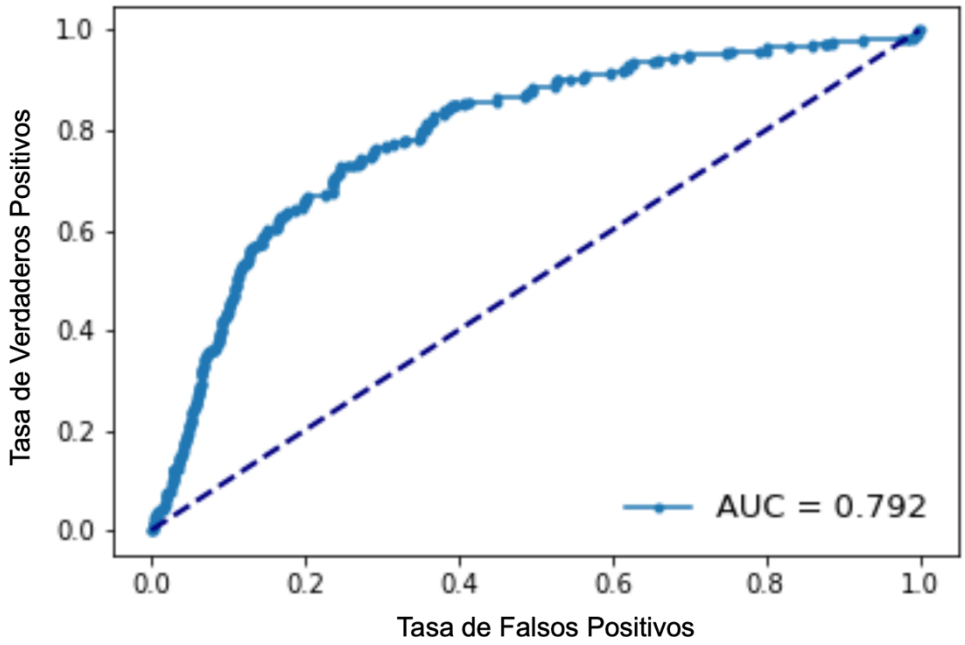


Figura 5.6 Curva ROC y área bajo la curva (AUC) lograda por el modelo GCN en el conjunto de prueba. La línea diagonal discontinua representa la AUC de una selección aleatoria.

Por otro lado, la figura 5.7 muestra la distribución de moléculas en el conjunto de validación. Cada punto representa una molécula coloreada por su clase de actividad (azul para activo, rojo para inactivo). El gráfico se construyó realizando una reducción de dimensionalidad t-SNE en los vectores de salida de la capa de lectura (*readout layer*) de la GCN+a+g. t-SNE tiene en cuenta tanto la estructura local como la estructura global y nos permite observar la presencia de agrupamientos de ligandos (Van Der Maaten y Hinton 2008). En este caso, se usó el conjunto

de validación en lugar del conjunto de prueba debido al severo desequilibrio de clases en este último que dificulta visualizar la distribución de puntos activos en el gráfico t-SNE.

Las moléculas activas e inactivas están bastante bien separadas (con algunos ejemplos mezclados), como lo demuestran los dos grupos de puntos azules y rojos, respectivamente, en la siguiente figura.

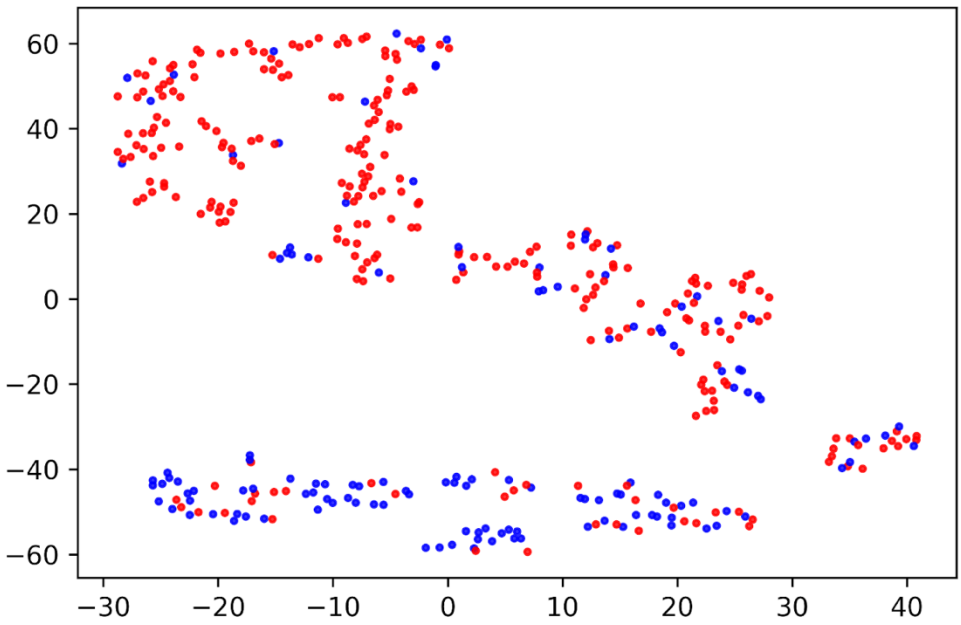


Figura 5.7 Distribución de moléculas activas (azules) e inactivas (rojas) en el conjunto de datos AID-1478 (conjunto de validación) obtenido por reducción de dimensionalidad mediante t-SNE.

Debemos reconocer que estos son resultados son algo sorprendentes. Las técnicas de *docking* molecular son enfoques basados en la estructura que dan cuenta explícitamente de las interacciones moleculares del ligando con la estructura de la proteína, mientras que el modelo de GCN+a+g implementado no lo hace. Por lo tanto, uno esperaría un mejor desempeño de la primera técnica.

Por otro lado, las moléculas son capaces de adoptar diversas conformaciones dependiendo del número de sus grados de libertad, y normalmente sólo están involucradas conformaciones específicas en su modo de acción farmacológico. El modelo GCN entrenado no tiene en cuenta esta información conformacional y, sin embargo, funciona mejor que el *docking* molecular. Presumiblemente, como argumentan Sakai et al. (2021) esto se debe en parte a que la

conformación preferida es inherente a la estructura química en muchos casos, es decir, las topologías 2D ya contienen los determinantes clave de las acciones farmacológicas.

5.3.3 Interpretación de las características atómicas consideradas por la GCN para las relaciones estructura-actividad

Para lograr cierta intuición sobre las características estructurales identificadas por la GCN como determinantes clave para la inhibición de Cz, analizamos las incrustaciones de nodos (*embeddings*) de la última capa de convolución del grafo (figura 5.8).

Debido al correcto mapeo entre entradas y salidas, la GCN produce un espacio de alta dimensión en el que las entradas y salidas de la misma clase están ubicadas en proximidad (Ryu et al., 2018). Mientras que en las incrustaciones a nivel de grafo se observan una separación bastante clara entre los vectores de características de moléculas activas e inactivas (figura 5.7), en las incrustaciones a nivel de nodo no deberíamos esperar una imagen tan clara, porque solo unos pocos átomos son responsables de la actividad. Los nodos restantes conforman grupos químicos comunes presentes tanto en moléculas activas como inactivas. En consecuencia, en la figura 5.8 se observa una distribución mixta de nodos activos e inactivos. Aun así, todavía hay una región enriquecida en nodos activos (área dentro del círculo), que presumiblemente podría contener información útil sobre los requisitos estructurales para la inhibición de Cz.

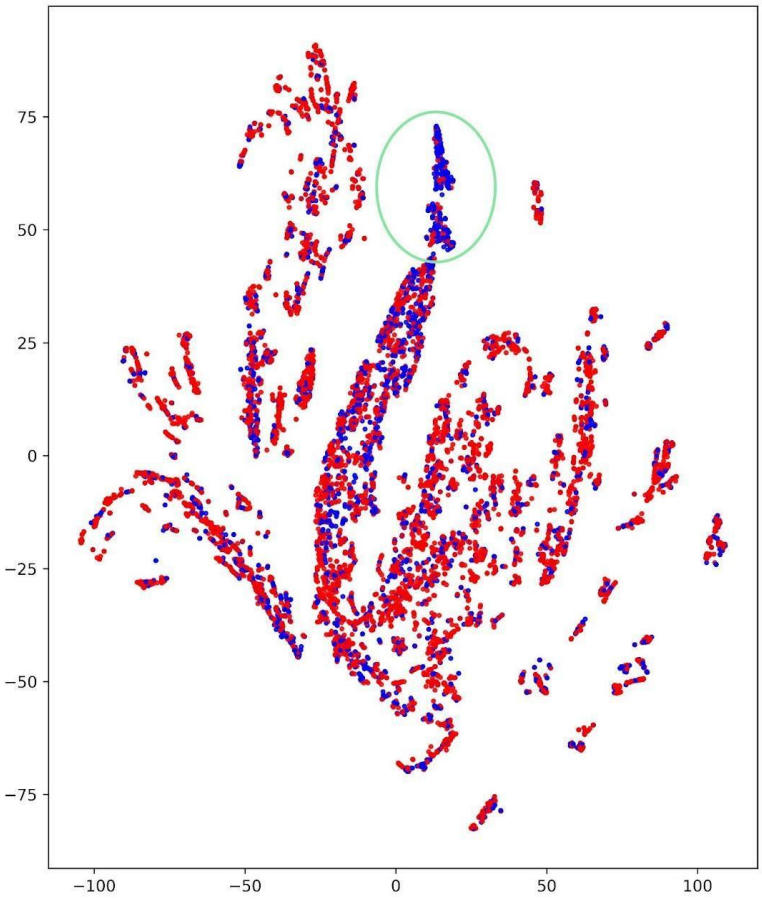


Figura 5.8 Distribución de nodos de moléculas activas (azules) e inactivas (rojas) en el conjunto de datos AID-1478 (conjunto de validación) obtenido por reducción de dimensionalidad t-SNE. El área rodeada por un círculo está enriquecida con nodos activos.

La figura 5.9 muestra diferentes grupos químicos incluidos en los compuestos activos del conjunto de validación. En este conjunto se incluyen inhibidores no covalentes y grupos electrofílicos (cabezas de guerra) que reaccionan covalentemente con el átomo de azufre nucleofílico del residuo de cisteína del sitio activo, Cys 25. A pesar de la diversidad de grupos, la GCN+a+g intenta encontrar un conjunto común de características de los nodos que ayuden a distinguir las moléculas activas de las inactivas. El conjunto de nodos seleccionados por la red como determinantes clave para la actividad se resaltan en rojo sobre las topologías moleculares en la figura 5.9. Estos nodos resaltados pertenecen a la región enriquecida en nodos activos en la figura 5.8 (área dentro de un círculo).

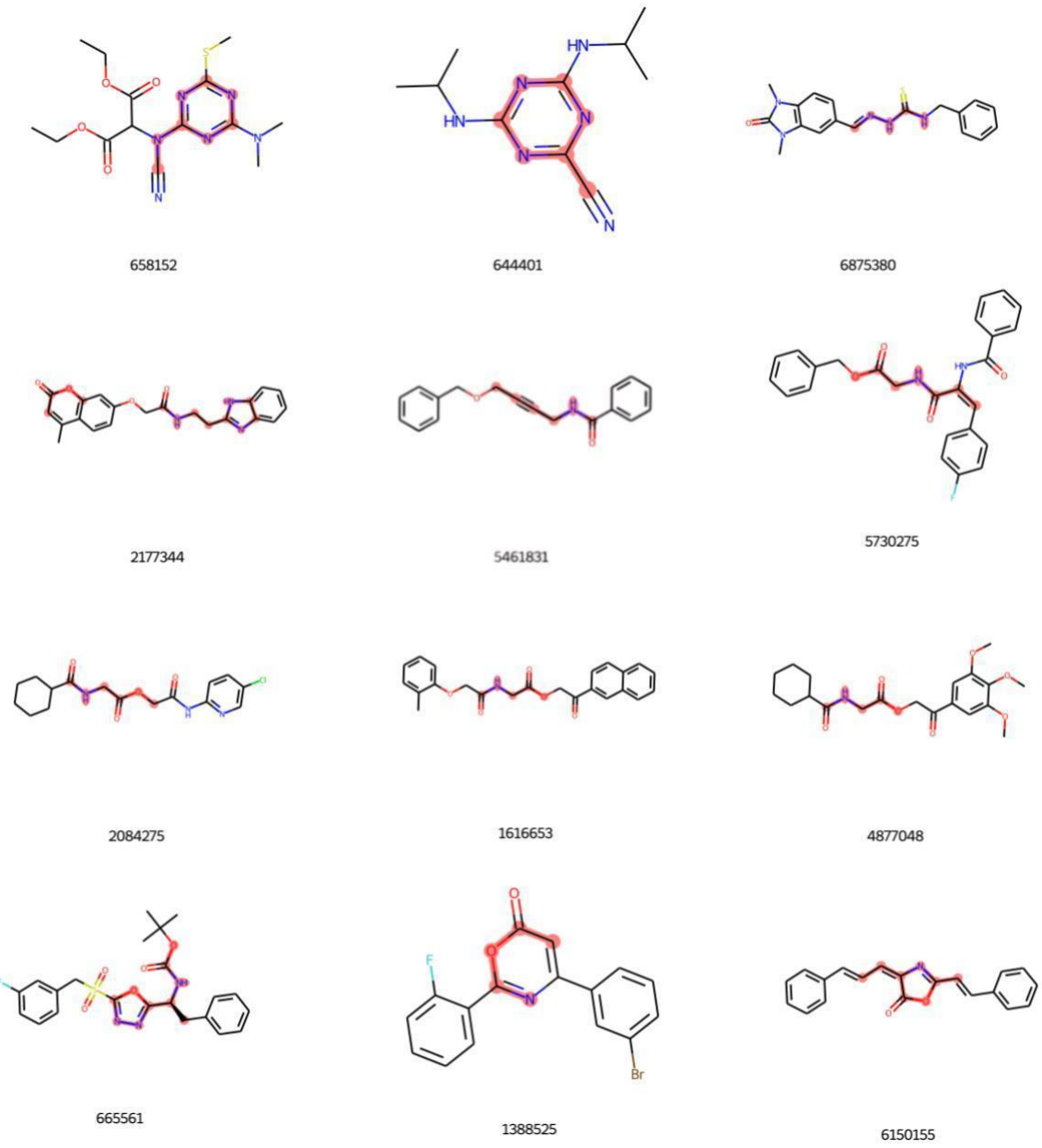


Figura 5.9 Topologías de moléculas activas representativas en el conjunto de validación con sus correspondientes números CID de PubChem. Los átomos resaltados en rojo sobre las topologías de ligandos 2D pertenecen a la región enriquecida en nodos activos (área encerrada en un círculo en la figura 5.8).

La red tiende a seleccionar grupos funcionales que contienen el enlace amida –N(H)–C=O, o algún bioisóstero del mismo, como por ejemplo: carboxilato –O–C=O en ésteres (es decir, CID 4877048), tioamida –N(H)–C=S e hidrazinilideno –N(H)–N=C en tiosemicarbazonas (es decir, CID 6875380 en la figura 5.9), amidina –N(H)–C=N en benzimidazoles (ej. CID 2177344) y en triazinas (ej. CID 658162) y –O–C=N en oxadiazoles (ej. CID 665561).

La red también selecciona grupos con propiedades electrofílicas que pueden reaccionar covalentemente con el átomo de azufre nucleofílico del residuo de cisteína del sitio activo, Cys 25, como grupos carbonilo y tiocarbonilo, así como “cabezas de guerra” de tipo alquino (ej. CID 5461831) y nitrilo (ej. CID 644401).

Los enlaces carbonilo y tiocarbonilo suelen formar parte de grupos electrofílicos que son bioisotéricos con el enlace peptídico, es decir, como en las tiosemicarbazonas y los ésteres. Los enlaces alquino y nitrilo se vuelven bioisostéricos con el enlace peptídico tras la adición del grupo tiol nucleofílico Cys 25 (ver figura 5.10 a continuación). El hecho de que esos grupos reactivos tengan que ser bioisotéricas con el enlace peptídico sugiere que, más allá de sus propiedades reactivas inherentes, también son importantes para el reconocimiento molecular por parte de la enzima y su posterior anclaje al sitio catalítico.

Además, las lactonas como CID 1388525 y CID 6150155 (figura 5.9) se pueden hidrolizar a la forma de cadena abierta que tiene una estructura similar a un péptido que encaja mejor en el sitio catalítico de la enzima (figura 5.10). Estos tipos de moléculas de tipo profármacos, al igual que las moléculas que contienen grupos reactivos, a menudo no son manejadas adecuadamente por los algoritmos de *docking* molecular porque la estructura 3D de la molécula que se une a la enzima difiere de aquella en la base datos de compuestos.

Por otro lado, el GCN+a+g es capaz de detectar los determinantes estructurales clave para la actividad tanto en forma de profármaco o en la forma final unida covalentemente de los compuestos. Esto podría explicar en parte el mejor desempeño del GCN+a+g con respecto al *docking*.

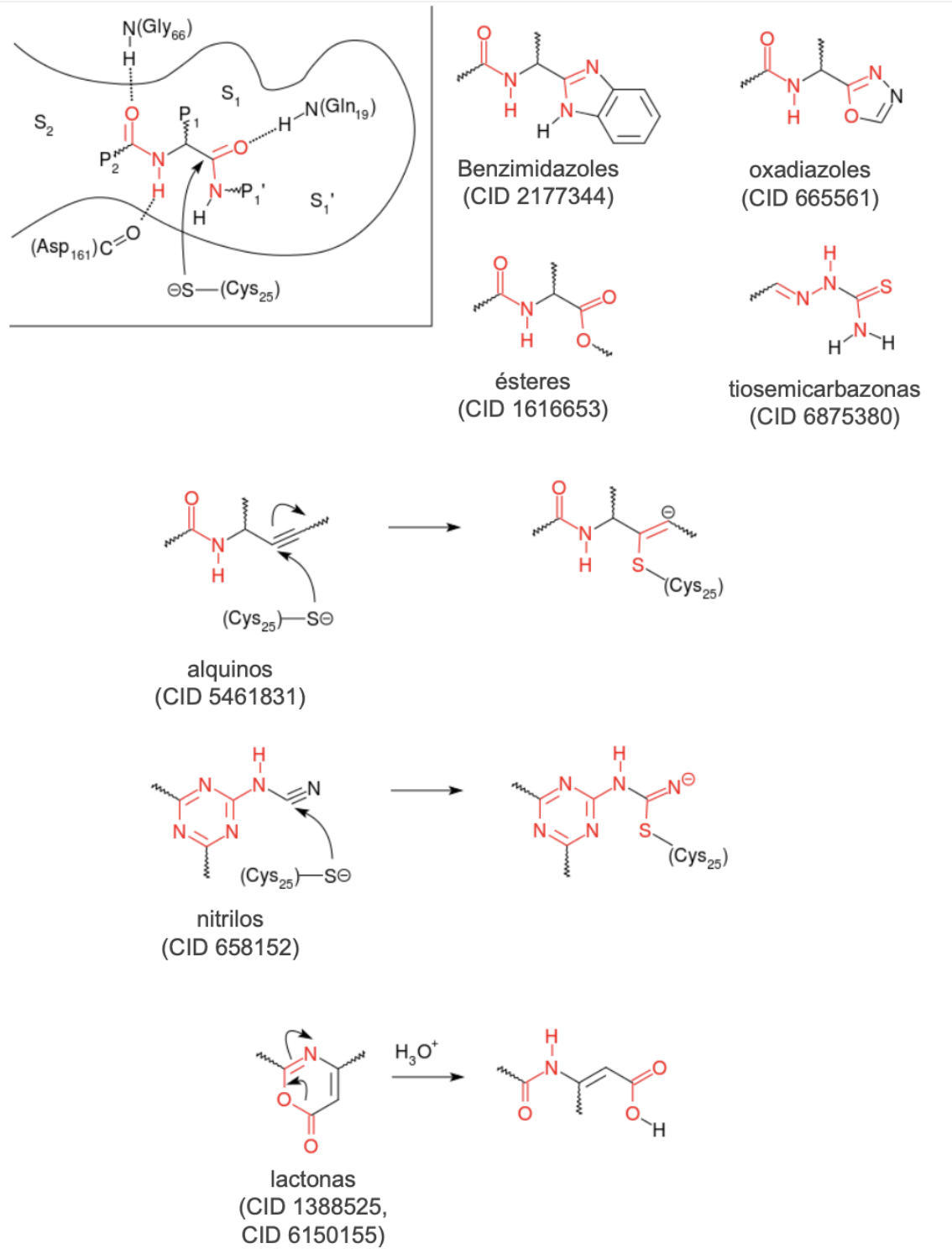


Figura 5.10 Los enlaces peptídicos y los bioisósteros del enlace peptídico se destacan como subestructuras en grupos químicos representativos de los inhibidores de Cruzipaina. Para cada grupo químico, se indica una molécula de ejemplo de la figura 5.9 con el número CID. El recuadro superior izquierdo muestra los residuos de proteínas dentro de la hendidura de unión a enzimas que interactúan con los enlaces peptídicos del sustrato.

La figura 5.10 representa la estructura de los grupos químicos más comunes en el conjunto de validación. La mayoría de estos grupos son inhibidores bien conocidos de Cz (da Silva, E.B., do Nascimento Pereira, G.A. and Ferreira 2016). Los enlaces peptídicos y los bioisósteros del enlace peptídico están coloreados en rojo. Por lo general, dos o más de esos grupos similares a péptidos están resaltados por la red dentro de cada estructura (ver figura 5.9). Estos grupos juntos pretenden imitar los enlaces amida del sustrato peptídico de la enzima, como se muestra en el recuadro de la figura 10 (arriba a la izquierda). Por lo tanto, la red enfatiza la importancia de preservar la estructura similar a un sustrato de los compuestos para mostrar actividad contra frente a Cz.

El recuadro superior izquierdo en la figura 5.10 también muestra los residuos de proteína dentro de la enzima que interactúan con los enlaces peptídicos del sustrato. El enlace peptídico ubicado más a la izquierda encaja en la parte más estrecha de la hendidura de unión de la enzima, entre los sub-bolsillos S1 y S2, y proporciona un fuerte anclaje a través de enlaces de hidrógeno con la cadena principal de Asp161 y Gly66.

Por otro lado, el enlace peptídico que se ubica más a la derecha, es decir, el que es escindido por la enzima, forma enlaces de hidrógeno con el residue Gln19 del agujero oxianiónico. Esa interacción tiene por objeto estabilizar la carga negativa en el oxígeno del carbonilo del enlace peptídico del sustrato, tras la adición del átomo de azufre nucleofílico (Turk et al. 1998).

Por lo tanto, para los grupos químicos similares al sustrato en la figura 5.10, uno debería esperar modos de unión e interacciones equivalentes a las observadas para el sustrato natural de la enzima.

5.3.4 *Docking* molecular guiado por GCN

La información obtenida hasta el momento es de gran utilidad para decirle al algoritmo de docking a qué interacciones debe prestar más atención para mejorar el rendimiento, es por ello que se establecieron una serie de puntos farmacofóricos a fin de guiar los cálculos de cribado virtual.

Los puntos farmacóforicos corresponderían a un arreglo tridimensional de características (estéricas y electrónicas) mínimas necesarias para orientar la formación de interacciones óptimas entre el sitio activo de la enzima y los ligandos bajo estudio.

A su vez, la combinación de puntos farmacofóricos suele enfocarse en la utilización de dos a cuatro, siendo los tripletes aquellos ampliamente utilizados ya que tradicionalmente se han considerado más eficaces en términos de contenido de información frente a complejidad (Silakari y Singh 2021).

De esta manera, los ligandos que tienen las características estructurales resaltadas por GCN+a+g, que les permiten formar las interacciones clave, similares a sustratos (representadas en el recuadro de la figura 5.10), obtendrán una puntuación más alta, mientras que los ligandos que carecen de esas características obtendrán una puntuación más baja, mejorando así el rendimiento general del cribado virtual.

En consecuencia, se incluyeron tres restricciones farmacofóricas obligatorias en el algoritmo de *docking*: a) dos aceptores de enlaces de hidrógeno (EH) colocados en las coordenadas correspondientes para guiar la formación EH con la cadena lateral de Gln19 y la cadena principal de Gly66 y b) un dador de EH. Este último, con la cadena principal de Asp161 (ver figura 5.11).

La figura 5.11 muestra las poses de *docking* de un compuesto activo del conjunto de datos AID 1478 (CID 751269) antes y después de aplicar las restricciones farmacofóricas. Al aplicar las restricciones farmacofóricas, los átomos del ligando se reorganizan para coincidir con los puntos farmacofóricos, representados como esferas en la figura.

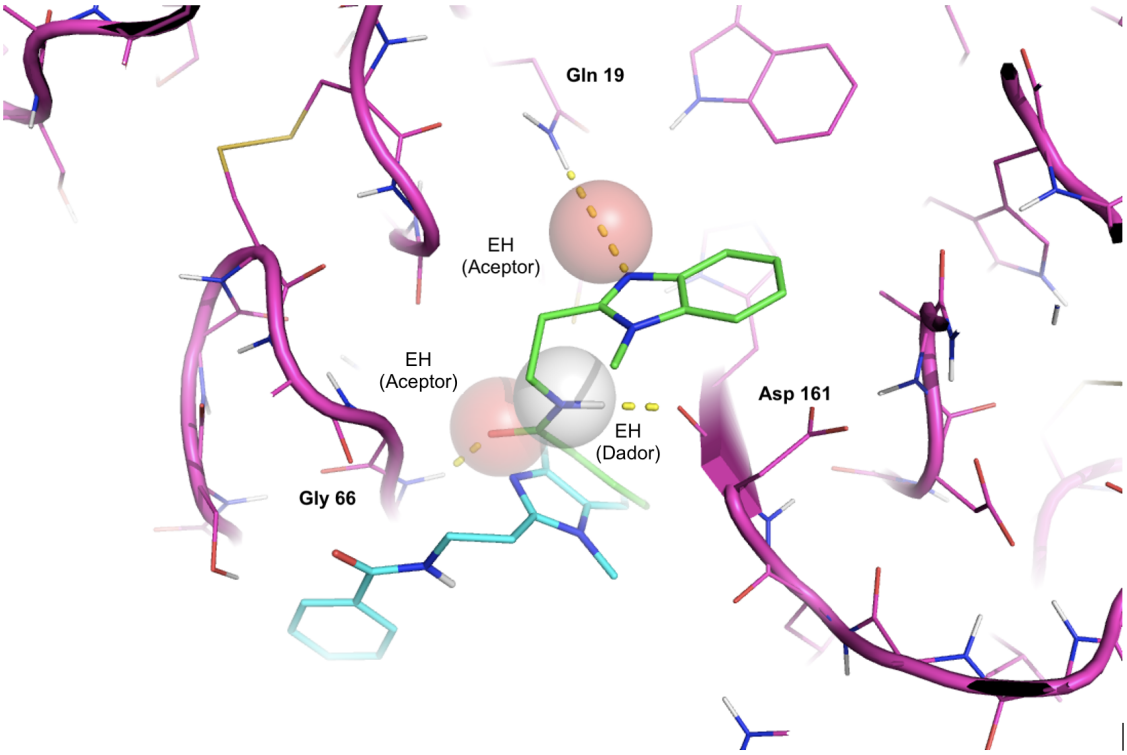


Figura 5.11 Pose de unión de un compuesto activo (CID 751269) del conjunto de datos AID 1478 filtrado por similitud obtenido mediante acoplamiento libre (cian) y restringido (verde). Las restricciones farmacofóricas, representadas como esferas rojas translúcidas (aceptor de enlaces H) y blancas (donante de enlaces H), guían al compuesto para que adopte una conformación similar a un sustrato dentro de la hendidura de unión a la enzima.

Las restricciones farmacofóricas se incorporan al algoritmo de *docking* en forma de sanciones de distancia que agregan un término positivo (es decir, desestabilizador) a la función de puntuación. La penalización por cada restricción se basa en la distancia desde el átomo de ligando coincidente más cercano al centro de restricción del farmacóforo. Por lo tanto, los compuestos que tienen las características estructurales resaltadas por GCN+a+g, es decir, que pueden formar las interacciones clave de tipo sustrato (representadas en el recuadro de la figura 5.10), obtendrán una puntuación más alta, mientras que los ligandos que carecen de esas características obtendrán una clasificación más baja.

Como puede verse en la figura 5.12, el rendimiento del *docking* guiado ha mejorado (AUC= 0,69) en comparación con el acoplamiento imparcial (AUC ~0,63, figura. 5.3).

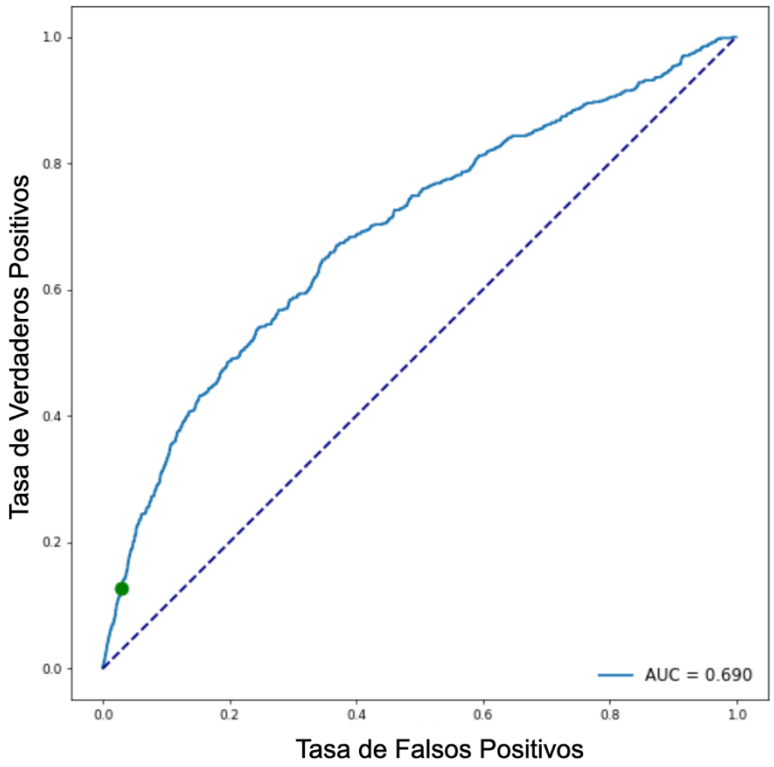


Figura 5.12 *Docking* del conjunto de datos AID 1478 filtrado por similitud guiado por restricciones farmacofóricas que explican las características relevantes aprendidas por GCN+a+g.

5.3.5 GCN como filtro previo al *docking*

Además de guiar el *docking* con las características moleculares aprendidas por GCN, podríamos explotar directamente el poder predictivo de GCN para aumentar el rendimiento del *docking*.

En las campañas de cribado virtual prospectivas, los compuestos de una base de datos desconocida se clasifican de acuerdo a los valores de *docking*, y se prioriza un cierto porcentaje de los compuestos mejor clasificados para las pruebas experimentales.

La GCN entrenada, por otro lado, no podía usarse directamente para clasificar compuestos de la misma manera que el *docking*, porque no fue entrenada con datos de actividad de valor real sino con datos binarios. Sin embargo, debido a su alto rendimiento para la clasificación de compuestos, esta herramienta podría usarse como un filtro previo al *docking*, es decir, para filtrar moléculas inactivas del conjunto de datos. De esta manera, el algoritmo de *docking* se alimentaría con un conjunto de datos enriquecido en compuestos activos, lo que probablemente aumentaría la tasa de aciertos en las posibles campañas de cribado virtual.

La curva ROC naranja en la figura. 5.13 muestra que la estrategia de combinación de GCN seguida del *docking* molecular supera al *docking* estándar del conjunto de datos AID 1478 filtrado por similitud. Los compuestos utilizados para entrenar la GCN se descartaron por adelantado del conjunto de datos para evitar sesgar la evaluación del rendimiento.

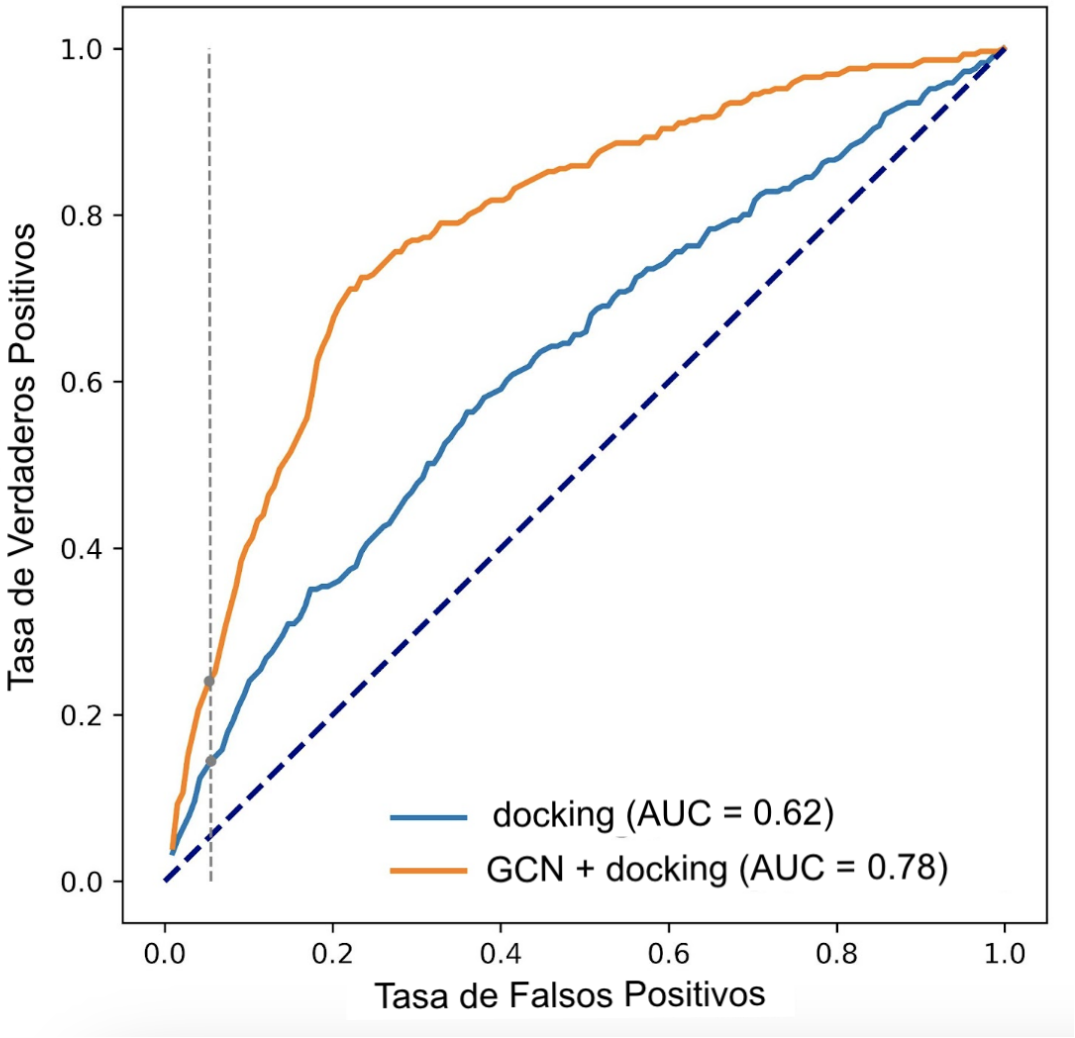


Figura 5.13 Comparación de curvas ROC y área bajo la curva (AUC) logradas mediante la aplicación del protocolo de *docking* estándar y de una estrategia combinada de *docking* + GCN, en el conjunto de datos AID 1478 filtrado por similitud. La intersección de la línea discontinua vertical con las curvas ROC representa la tasa de verdaderos positivos (TVP) lograda por el 5% mejor clasificado de la base de datos.

Como lo demuestra la línea discontinua vertical en la figura 5.13, si se priorizara el mismo porcentaje de compuestos mejor clasificados para las pruebas de ambos procedimientos, se logrará una tasa de aciertos bastante más alta mediante el enfoque combinado de *docking* + GCN, en comparación con el procedimiento de *docking* estándar.

5.3.6 Manejo de inhibidores covalentes

A pesar de la mejora en los resultados obtenidos al aplicar GCN+a+g como prefiltro de *docking*, una gran parte de las moléculas activas aún se encuentran mal clasificadas por el algoritmo. En el experimento anterior, los cálculos de *docking* se ejecutaron directamente en los compuestos que pasaron el filtro GCN, sin ninguna intervención en el algoritmo de *docking*. El rendimiento del *docking* podría mejorarse aún más mediante la aplicación de restricciones farmacofóricas a los compuestos filtrados, pero esto aún puede ser insuficiente para priorizar moléculas similares a profármacos y unidas covalentemente.

Esos grupos quimicos no son tenidos en cuenta por los algoritmos de *docking* que clasifican los compuestos basándose solo en la forma y la complementariedad de carga. Por otro lado, la priorización de compuestos basada únicamente en el resultado de la GCN es inviable como ya se discutió anteriormente. Una posible solución podría ser la de realizar un agrupamiento basado en la similitud de los compuestos predichos por el GCN como activos y luego seleccionar uno o más grupos químicos representativos de cada grupo para los pasos prospectivos, prestando especial atención a aquellos grupos de ligandos con grupos reactivos.

5.4 Conclusiones

Este capítulo de la tesis se inspiró principalmente en el artículo de Ferreira et al. titulado *"Complementarity Between a Docking and a High-Throughput Screen in Discovering New Cruzain Inhibitors*", que se remonta al año 2010. En el mismo, los investigadores realizaron un *docking* paralelo y un cribado de alto rendimiento (HTS) de 197.861 compuestos contra cruzipaina y encontraron que ambas técnicas son complementarias entre sí y que las debilidades del *docking* son ortogonales a las de HTS.

Desde la perspectiva del diseño de fármacos asistido por computadora, no se supone que los programas de *docking* se usen junto con HTS, como en el artículo de referencia, sino en lugar del HTS, debido a los altos costos y tasa de acierto de este último.

Sin embargo, si se usara el *docking* solo para examinar la biblioteca AID 1478, se habrían perdido algunos de los ligandos activos, recuperados solo por el HTS. Por lo tanto, dicha "complementariedad" entre el *docking* y HTS en realidad estaba evidenciando en que los algoritmos de *docking* no eran lo suficientemente precisos en ese momento. Tampoco lo son hoy en día, incluso con el continuo avance en *software* y *hardware* de computadoras, la complejidad inherente de los sistemas biológicos desafía cualquiera de los métodos de modelado molecular actualmente disponibles.

Por otro lado, los modelos de aprendizaje profundo (DL) de última generación, como las GCN, pueden capturar las complejas relaciones no lineales entre los datos estructurales y biológicos, pero carecen de la intuición de los enfoques basados en estructuras.

En este trabajo, propusimos estrategias combinadas para explotar los beneficios de ambos, es decir, la capacidad de los modelos GCN para capturar relaciones complejas de los datos y la interpretabilidad del *docking* molecular basado en la estructura, para evaluar virtualmente la biblioteca AID 1478 contra la cruzipaina.

Al conectar las características atómicas aprendidas por la GCN en el algoritmo de *docking* por medio de restricciones farmacofóricas, se mejoró la capacidad de *docking* para recuperar los ligandos activos.

Además, al aplicar la GCN como un filtro previo, la biblioteca de compuestos se enriquece en moléculas activas y el *docking* posterior de la biblioteca filtrada logra tasas de acierto significativamente más altas.

Las estrategias de combinación que involucran el aprendizaje profundo y las técnicas clásicas de *docking* molecular ofrecen una forma pragmática de eludir las limitaciones técnicas actuales para modelar eventos complejos de unión de proteínas y ligandos mediante enfoques basados en estructuras.

Referencias del capítulo 5

Baldi, A. 2010. “Computational Approaches for Drug Design and Discovery: An Overview.” *Systematic Reviews in Pharmacy* 1 (1): 99–105. https://doi.org/10.4103/0975-8453.59519.

Coley, Connor W., Wengong Jin, Luke Rogers, Timothy F. Jamison, Tommi S. Jaakkola, William H. Green, Regina Barzilay, and Klavs F. Jensen. 2019. “A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity.” *Chemical Science* 10 (2): 370–77. https://doi.org/10.1039/C8SC04228D.

Courville, Aaron, Lan Goodfellow, and Yoshua Bengio. 2016. “Deep Learning.” *MIT Press*, 800.

Deng, Nanjie, Stefano Forli, Peng He, Alex Perryman, Lauren Wickstrom, R. S.K. Vijayan, Theresa Tiefenbrunn, et al. 2015. “Distinguishing Binders from False Positives by Free Energy Calculations: Fragment Screening against the Flap Site of HIV Protease.” *Journal of Physical Chemistry B* 119 (3): 976–88. https://doi.org/10.1021/JP506376Z/ASSET/IMAGES/LARGE/JP-2014-06376Z_0009.JPEG.

Deng, Zhan, Claudio Chuaqui, and Juswinder Singh. 2004. “Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions.” *Journal of Medicinal Chemistry* 47 (2): 337–44. https://doi.org/10.1021/jm030331x.

Ferreira, Rafaela S., Anton Simeonov, Ajit Jadhav, Oliv Eidam, Bryan T. Mott, Michael J. Keiser, James H. McKerrow, David J. Maloney, John J. Irwin, and Brian K. Shoichet. 2010. “Complementarity between a Docking and a High-Throughput Screen in Discovering New Cruzain Inhibitors.” *Journal of Medicinal Chemistry* 53 (13): 4891–4905. https://doi.org/10.1021/jm100488w.

Fout, Alex, Jonathon Byrd, Basir Shariat, and Asa Ben-Hur. 2017. “Protein Interface Prediction Using Graph Convolutional Networks.” *Advances in Neural Information Processing Systems* 30.

Kingma, Diederik P., and Jimmy Lei Ba. 2015. “Adam: A Method for Stochastic Optimization.” *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Kipf, Thomas N., and Max Welling. 2016. “Semi-Supervised Classification with Graph Convolutional Networks.” *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, September. https://doi.org/10.48550/arxiv.1609.02907.

Korolev, Vadim, Artem Mitrofanov, Alexandru Korotcov, and Valery Tkachenko. 2020. “Graph Convolutional Neural Networks as ‘General-Purpose’ Property Predictors: The Universality and Limits of Applicability.” *Journal of Chemical Information and Modeling* 60 (1): 22–28. https://doi.org/10.1021/ACS.JCIM.9B00587/SUPPL_FILE/CI9B00587_SI_001.PDF.

Li, Hongjian, Kam Heung Sze, Gang Lu, and Pedro J. Ballester. 2020. “Machine-Learning Scoring Functions for Structure-Based Drug Lead Optimization.” *Wiley Interdisciplinary Reviews: Computational Molecular Science* 10 (5): e1465. https://doi.org/10.1002/WCMS.1465.

Liao, Chenzhong, Megan L. Peach, Risheng Yao, and Marc C. Nicklaus. 2013. “Molecular Docking and Structure-Based Virtual Screening.” In *In Silico Drug Discovery and Design*, 6–20. Future Medicine Ltd. https://doi.org/10.4155/EBO.13.181.

Lim, Jaechang, Seongok Ryu, Kyubyong Park, Yo Joong Choe, Jiyeon Ham, and Woo Youn Kim. 2019. “Predicting Drug-Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation.” *Journal of Chemical Information and Modeling* 59 (9): 3981–88. https://doi.org/10.1021/ACS.JCIM.9B00387/ASSET/IMAGES/MEDIUM/CI9B00387_0004.GIF.

Maaten, Laurens Van Der, and Geoffrey Hinton. 2008. “Visualizing Data Using T-SNE.” *Journal of Machine Learning Research* 9: 2579–2625.

Mercado, Rocio, Tobias Rastemo, Edvard Lindelof, Gunter Klambauer, Ola Engkvist, Hongming Chen, and Esben Jannik Bjerrum. 2021. “Graph Networks for Molecular Design.” *Machine Learning: Science and Technology* 2 (2): 025023. https://doi.org/10.1088/2632-2153/ABCF91.

Mysinger, Michael M., Michael Carchia, John J. Irwin, and Brian K. Shoichet. 2012. “Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking.”

Journal of Medicinal Chemistry 55 (14): 6582–94. https://doi.org/10.1021/JM300687E/SUPPL_FILE/JM300687E_SI_004.TXT.

Na, Gyoung S., Hyunju Chang, and Hyun Woo Kim. 2020. “Machine-Guided Representation for Accurate Graph-Based Molecular Machine Learning.” *Physical Chemistry Chemical Physics* 22 (33): 18526–35. https://doi.org/10.1039/D0CP02709J.

O’Boyle, Noel M., Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. 2011. “Open Babel: An Open Chemical Toolbox.” *Journal of Cheminformatics* 3 (10): 33. https://doi.org/10.1186/1758-2946-3-33.

Pedregosa, Fabian, Vincent Michel, Olivier Grisel OLIVIERGRISEL, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Jake Vanderplas, et al. 2011. “Scikit-Learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot.” *Journal of Machine Learning Research*. Vol. 12. http://scikit-learn.sourceforge.net.

Ruiz-Carmona, Sergio, Daniel Alvarez-Garcia, Nicolas Foloppe, A. Beatriz Garmendia-Doval, Szilveszter Juhos, Peter Schmidtke, Xavier Barril, Roderick E. Hubbard, and S. David Morley. 2014. “RDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids.” *PLOS Computational Biology* 10 (4): e1003571. https://doi.org/10.1371/JOURNAL.PCBI.1003571.

Ryu, Seongok, Jaechang Lim, Seung Hwan Hong, and Woo Youn Kim. 2018. “Deeply Learning Molecular Structure-Property Relationships Using Attention- and Gate-Augmented Graph Convolutional Network,” May. https://doi.org/10.48550/arxiv.1805.10988.

Sakai, Miyuki, Kazuki Nagayasu, Norihiro Shibui, Chihiro Andoh, Kaito Takayama, Hisashi Shirakawa, and Shuji Kaneko. 2021. “Prediction of Pharmacological Activities from Chemical Structures with Graph Convolutional Neural Networks.” *Scientific Reports 2021 11:1* 11 (1): 1–14. https://doi.org/10.1038/s41598-020-80113-7.

Silakari, Om, and Pankaj Kumar Singh. 2021. “Ligand-Based Pharmacophore Modeling: A Technique Utilized for Virtual Screening of Commercial Databases.” *Concepts and Experimental Protocols of Modelling and Informatics in Drug Design*, January, 203–34. https://doi.org/10.1016/B978-0-12-820546-4.00009-X.

Silva, E.B., do Nascimento Pereira, G.A. and Ferreira, R.S. da. 2016. “Trypanosomal Cysteine Peptidases: Target Validation and Drug Design Strategies.” In *Comprehensive Analysis of Parasite Biology: From Metabolism to Drug Discovery*, 121–45.

Torng, Wen, and Russ B. Altman. 2019. “Graph Convolutional Neural Networks for Predicting Drug-Target Interactions.” *Journal of Chemical Information and Modeling* 59 (10). https://doi.org/10.1021/ACS.JCIM.9B00628.

Turk, Dušan, Gregor Gunčar, Marjetka Podobnik, and Boris Turk. 1998. “Revised Definition of Substrate Binding Sites of Papain-Like Cysteine Proteases.” *Biological Chemistry* 379 (2): 137–47. https://doi.org/10.1515/bchm.1998.379.2.137.

Wieder, Oliver, Stefan Kohlbacher, Mélaine Kuenemann, Arthur Garon, Pierre Ducrot, Thomas Seidel, and Thierry Langer. 2020. “A Compact Review of Molecular Property Prediction with Graph Neural Networks.” *Drug Discovery Today: Technologies* 37 (December): 1–12. https://doi.org/10.1016/J.DDTEC.2020.11.009.

Wu, Zonghan, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2019. “A Comprehensive Survey on Graph Neural Networks.” *IEEE Transactions on Neural Networks and Learning Systems* 32 (1): 4–24. https://doi.org/10.1109/TNNLS.2020.2978386.

Xiong, Zhaoping, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, et al. 2020. “Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism.” *Journal of Medicinal Chemistry* 63 (16): 8749–60. https://doi.org/10.1021/ACS.JMEDCHEM.9B00959/SUPPL_FILE/JM9B00959_SI_001.PDF.

CAPÍTULO VI

“Cribado virtual prospectivo de una biblioteca de ligandos”

6.1 Introducción

En el capítulo anterior, se trabajó sobre una biblioteca de compuestos con actividad conocida sobre Cruzipaina para desarrollar un protocolo de Cribado Virtual (CV) destinado a descubrir nuevos inhibidores de esta enzima.

Se observó que al aplicar un filtrado previo al docking de la biblioteca de compuestos, utilizando una Red Convolucional basada en Grafos (GCN), pre-entrenada únicamente con la estructura 2D de los ligandos, se lograba una mejora significativa en la tasa de recuperación de compuestos activos. El rendimiento mejoraba aún más si adicionalmente se incluían las características aprendidas por la GCN, conocidas como embeddings, a modo de restricciones farmacofóricas para guiar las soluciones del docking.

La incorporación de la GCN dentro del protocolo de Cribado Virtual (CV) ciertamente ha permitido mejorar de manera pragmática el rendimiento del docking, mediante el pre-filtrado de falsos positivos y el guiado mediante restricciones farmacofóricas aprendidas por la GCN.

Sin embargo, como la GCN no fue entrenada con información estructural de la enzima, no arroja pistas sobre cuáles serían las características estructurales que el algoritmo de docking falla en considerar, y que afectan directamente su capacidad de cribado.

En consecuencia, en la primera parte de este capítulo buscamos entender cuáles son estas deficiencias del algoritmo de docking, de manera de poder intervenir más directamente para mejorar su rendimiento.

En la segunda parte, una vez identificadas al menos algunas de las fallas del docking y aplicadas las acciones correctivas correspondientes, nos enfocamos en la búsqueda de nuevos inhibidores de la Cruzipaina.

6.2 Metodología

6.2.1 Calibración de *docking*

6.2.1.1 Biblioteca de compuestos AID-2158

La base de datos de PubChem AID-2158 está compuesta por 599 moléculas obtenidas a partir en un experimento cuantitativo de cribado de alto rendimiento (qHTS, del inglés Quantitative

High-Throughput Screening) contra cruzipaina (https://pubchem.ncbi.nlm.nih.gov/bioassay/2158). Se compiló de PubChem una tabla de datos que contiene información referente a cada molécula, como ser, un identificador del compuesto (por ejemplo: CID: 44142159), la representación unidimensional de las estructuras (SMILES isoméricas) y la actividad para cada compuesto en el conjunto de datos. De las 599 moléculas obtenidas bajo este estudio, 303 son activas frente a Cz.

6.2.1.2 Estructuras cristalográficas

19 estructuras fueron recuperadas del Protein Data Bank (códigos PDB 1AIM, 1EWL, 1EWM, 1EWO, 1EWP, 1F29, 1F2A, 1F2B, 1F2C, 1ME4, 2AIM, 2OZ2, 3HD3, 3KKU, 3I06, 3IUT, 3LXS, 4PI3, 4QH6).

6.2.2 Cribado virtual prospectivo

El cribado virtual prospectivo se llevó a cabo sobre una biblioteca in-house de 7 millones de compuestos siguiendo el esquema descrito en la figura 6.1. En el camino A, a partir de los resultados crudos, se realizó un primer "filtrado" para reducir el número de compuestos, empleando la función de scoring del algoritmo de docking. Las poses o modos de unión de los compuestos son ranqueadas de acuerdo con la función de puntuación de rDock, seleccionando aquellos con mejor energía de interacción y grado de satisfacción de las restricciones farmacofóricas.

En este punto del filtrado, la selección de compuestos se ha reducido a un número razonablemente bajo como para permitir evaluar la estabilidad de la unión empleando herramientas computacionales más precisas.

Los compuestos seleccionados en el *docking*, luego fueron sometidos a simulaciones de Dinámica Molecular Dirigida (Steering Molecular Dynamics, SMD) empleando Dynamic Undocking o DUck (www.ub.edu/bl/undocking). DUck calcula el trabajo necesario para alcanzar un estado "cuasi enlazado" (W_{qb}). Esta propiedad de no equilibrio es sorprendentemente efectiva

en el cribado virtual porque los verdaderos ligandos forman interacciones más resilientes que los señuelos (Ruiz-Carmona et al., 2017).

Aquellos compuestos con un valor de W_{qb} favorable fueron seleccionados para realizar simulaciones de Dinámica Molecular (DM) por triplicado empleando Amber16.

Finalmente, los compuestos que superaron el filtro previo son ranqueados de acuerdo a su energía libre de unión ($\Delta G_{unión}$), estimada a partir de las simulaciones de DM, empleando el protocolo MMPBSA.

Por otro lado, en el camino B se realizó el filtrado de la base de datos a partir de la GCN previamente entrenada (ver capitulo 5).

Para finalizar, aquellos compuestos seleccionados a partir del pos-procesamiento del *docking* fueron comparados con aquellos seleccionados por la GCN llegando así al número final de candidatos.

El cribado virtual de este volumen de compuestos fue posible gracias a que se contó con acceso a los recursos computacionales del Barcelona Supercomputer Center (https://www.bsc.es), en particular los cálculos de docking se corrieron en el cluster MareNostrum IV que cuenta con ~4000 nodos equipados cada uno con dos procesadores de la línea Intel Xeon Platinum. Por último, la aplicación de este protocolo y el cálculo del valor de W_{qb} para los compuestos seleccionados fue posible gracias a que se contó con acceso al cluster CTE-Power del Barcelona Supercomputer Center, que cuenta con 52 nodos cada uno equipado con 4 GPUs NVIDIA V100

(Volta) para acelerar las simulaciones.

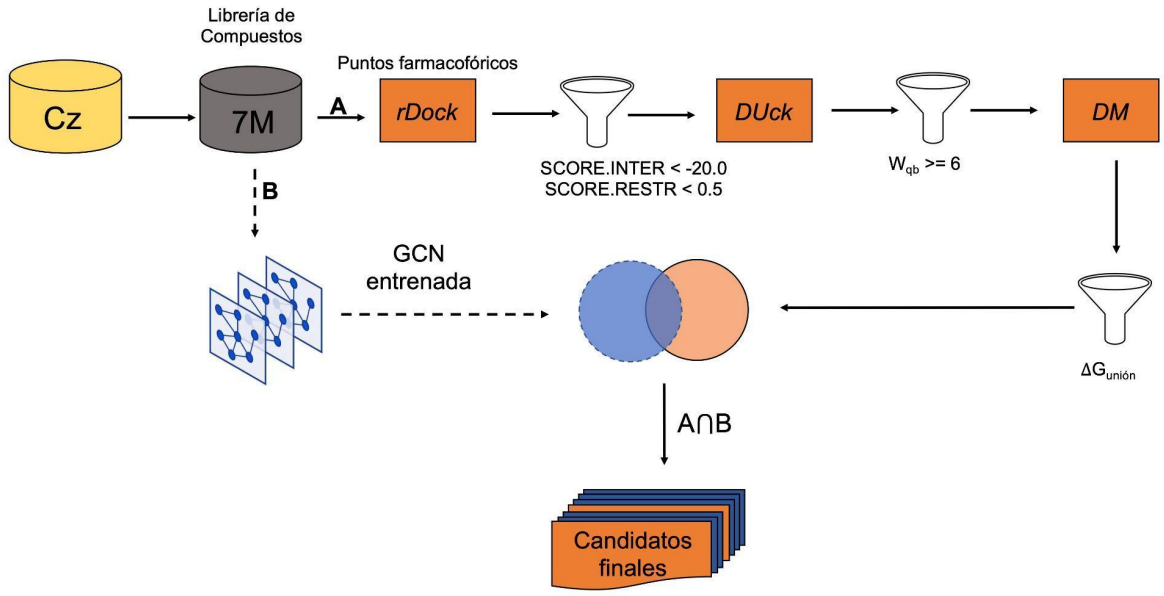


Figura 6.1 Esquema de general del proceso de cribado virtual.

6.3 Resultados y Discusión

6.3.1 Primera parte. Análisis de los factores que afectan el desempeño del *docking* molecular en las campañas de cribado virtual sobre Cruzipaina.

En el capítulo previo, se demostró que el algoritmo de docking por sí solo, sin la ayuda de métodos de aprendizaje profundo, no lograba un buen desempeño al intentar encontrar los compuestos activos en la base de datos AID 1478 (AUC = 0.559). En la primera parte de este capítulo, nuestro objetivo fue intentar comprender cuáles son los factores que afectan su rendimiento, para así poder aplicar las acciones correctivas correspondientes.

Para agilizar el estudio, los siguientes experimentos se realizaron sobre el dataset AID 2158, que abarca un número más manejable de compuestos que el AID 1478 y que por lo tanto permite analizar una mayor cantidad de parámetros de acoplamiento en un menor tiempo.

El AID 2158 consistió en un experimento de cuantificación de alto rendimiento (qHTS) cuyo propósito fue el de validar y confirmar los hits identificados en el AID 1478. Se evaluaron alrededor de 600 compuestos y se confirmaron 300 de ellos como activos.

Aunque el conjunto de datos AID 2158 representa un subconjunto muy pequeño del AID 1478, incluye muchos de los andamiajes estructurales presentes en el conjunto original. Para nuestros propósitos podemos considerar al conjunto de datos AID 2158 como una muestra suficientemente representativa de AID 1478.

En consecuencia, se llevó a cabo un experimento de Cribado Virtual con el conjunto de datos AID 2158 utilizando las mismas condiciones que para el AID 1478. Se empleó el mismo algoritmo de docking (rDock) y la misma estructura del blanco molecular (código PDB: 2OZ2). Al igual que con la base de datos original, el desempeño del algoritmo de docking para priorizar los ligandos activos resultó decepcionante, con un valor muy bajo del Área bajo la Curva (AUC = 0.482).

A continuación, exploramos algunas modificaciones al protocolo de Cribado Virtual tendientes a mitigar el problema del bajo desempeño del docking.

6.3.1.1 Estructura del blanco molecular

Como es bien sabido, los algoritmos de *docking* generalmente tratan a la proteína como una entidad rígida durante el proceso de acoplamiento, basándose en un modelo rudimentario de interacción de tipo “llave-cerradura”. Sin embargo, esta aproximación puede no reflejar con precisión la realidad biológica, ya que las proteínas son moléculas dinámicas que pueden experimentar cambios conformacionales y flexibilidad en respuesta a su entorno y a la unión con ligandos.

Teorías más modernas de interacción ligando-proteína como los modelos de ajuste inducido y de selección conformacional consideran esta variabilidad conformacional de la enzima.

El modelo de ajuste inducido sugiere que tanto el ligando como la proteína pueden cambiar su conformación al interactuar entre sí, lo que resulta en una adaptación mutua para lograr una unión más óptima. El modelo de ajuste inducido podría explicar porque el acoplamiento de un par ligando-proteína, ambos provenientes del mismo cristal (docking nativo) predice generalmente con mayor precisión la pose cristalográfica del ligando que el docking sobre una estructura diferente de la misma proteína (*docking* cruzado) (Sutherland et al., 2017).

Por otro lado, el modelo de selección conformacional postula que la proteína puede existir en diferentes estados conformacionales, y solo ciertas conformaciones son capaces de unirse al ligando.

Una manera simplificada de poner en práctica el modelo de selección conformacional consiste en realizar el acoplamiento sobre varias estructuras diferentes que representan los posibles estados conformacionales que puede adoptar la proteína durante la interacción con el ligando, lo que se conoce como "*ensemble docking*".

El siguiente mapa de calor binario resume los resultados de un experimento de ensemble docking de inhibidores covalentes de Cruzipaina, acoplados sobre la misma estructura del cristal (*docking* nativo) y sobre otras conformaciones de la enzima (*docking* cruzado).

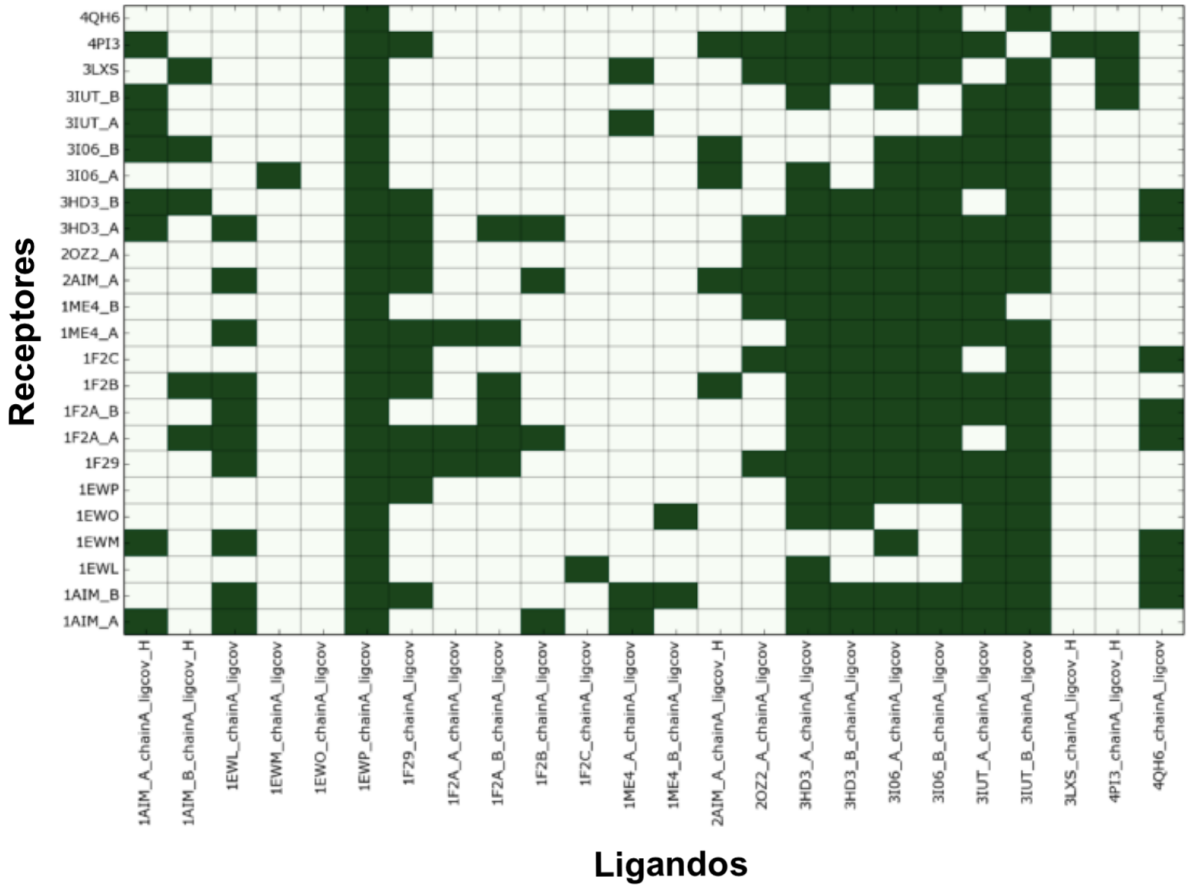


Figura 6.2 Mapa binario con los valores del rendimiento de docking cruzado. Las poses del docking mejor puntuadas con una RMSD <2Å con respecto al modo de unión cristalográfico se representan en verde. En abscisa y ordenadas se representan las diferentes conformaciones cristalográficas de los

ligandos y la enzima, respectivamente La nomenclatura A y B corresponden a diferentes estados de protonación de la His 162.

Como puede observarse en la figura anterior, se destaca la baja proporción de bits coloreados en verde (38%), que representan los casos en los cuales el algoritmo de docking logra predecir el modo de unión con una desviación cuadrática media (RMSD, del inglés Root-Mean-Square Deviation) menor o igual a 2Å con respecto a la pose cristalográfica. Esta observación resalta la fuerte interdependencia que existe en la mayoría de los casos entre la conformación de la enzima, la naturaleza del ligando y la pose que éste adopta.

No obstante, en un experimento de screening virtual típico, donde se criban desde cientos de miles a millones de compuestos, realizar el docking sobre múltiples conformaciones del blanco molecular puede ser computacionalmente costoso y requerir mucho tiempo. Por lo tanto, el número de blancos suele acotarse a unas pocas estructuras que sean lo más representativas posibles de las diferentes conformaciones que puede adoptar la enzima, según el problema en estudio.

En nuestro caso particular, además de la estructura de Cz unida al inhibidor irreversible K-777 (2OZ2), se realizó el acoplamiento sobre otras pocas estructuras que consideramos relevantes. Esto incluye una estructura de Cz unida a un inhibidor no covalente (PDB: 3KKU) ya que deseamos encontrar principalmente este tipo de inhibidores reversibles en nuestras campañas prospectivas. También se consideró la estructura de Cz unida a un inhibidor covalente de tipo nitrilo (PDB: 3I06) debido a que este tipo de “warhead” representa una proporción relativamente importante de los inhibidores covalentes incluidos en el conjunto AID 2158. Se han considerado ambas conformaciones alternativas 3I06_A y 3I06_B, para tener en cuenta las dos posibles conformaciones en que se resolvió el residuo del sitio activo Histidina 162.

La figura 6.3 muestra el desempeño del *docking* sobre las conformaciones seleccionadas de Cz para priorizar los compuestos activos del conjunto de datos AID 2158.

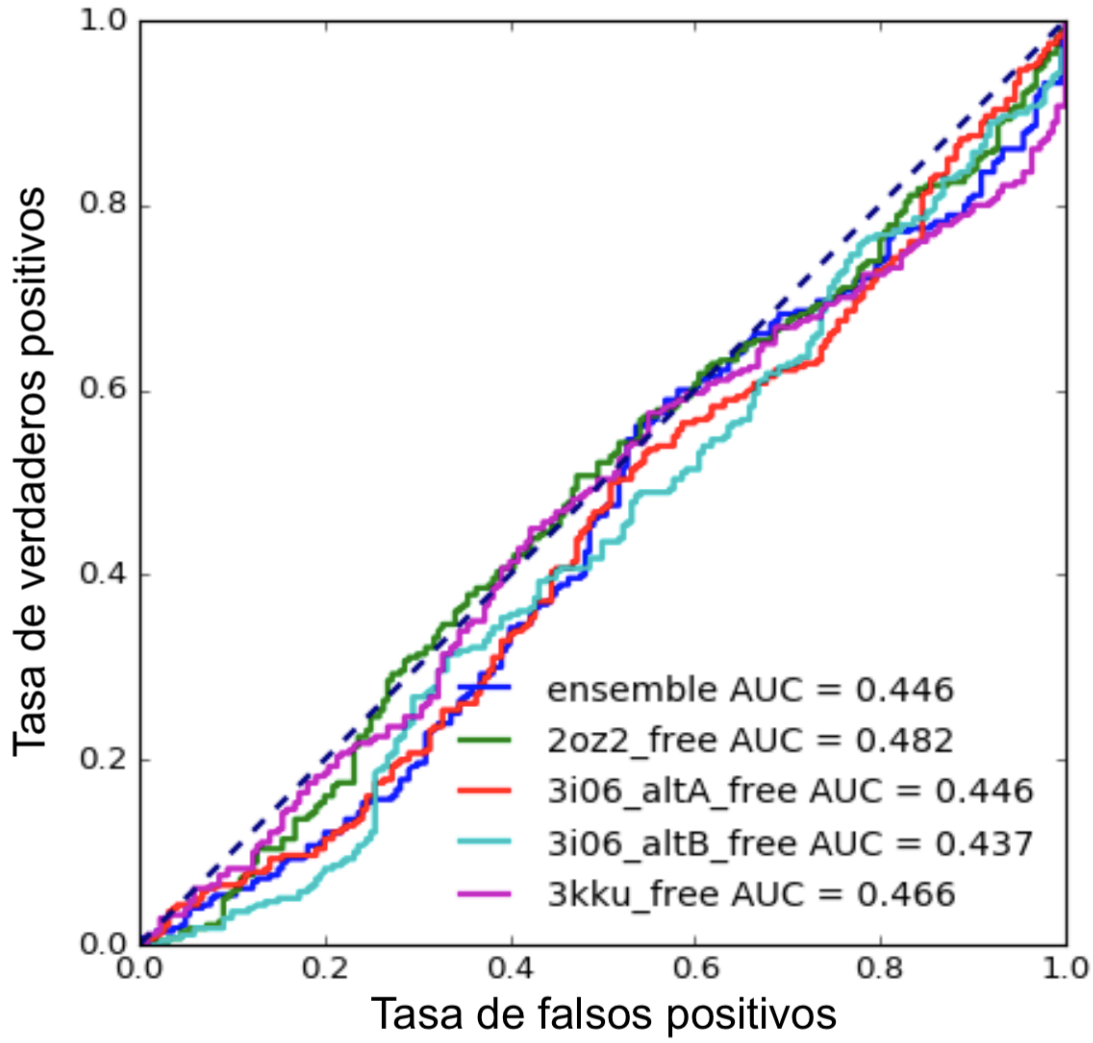


Figura 6.3 Desempeño de las campañas de docking mediante la utilización de diferentes conformaciones de cruzipaina

Resulta evidente, a partir de los bajos valores de AUC, que el algoritmo de docking falla en recuperar los ligandos activos del subconjunto AID 2158. Incluso tras el cálculo del consenso entre los diferentes receptores (*ensemble* AUC) no se logra mejorar su desempeño.

6.3.1.1 Heterogeneidad de la base de datos

Otro aspecto a tener en cuenta entre los factores que pueden influir sobre la capacidad predictiva del algoritmo de *docking* es la heterogeneidad del conjunto de datos que se está cribando. La base de datos AID 1478, del cual el AID 2158 constituye una muestra más o menos representativa, incluye tanto ligandos con grupos electrofílicos o "*warheads*" (ej. nitrilos,

tiosemicarbazonas, ésteres, etc) que se unen covalentemente a la enzima como así también ligandos no covalentes.

A su vez, dentro de estos dos grandes grupos de compuestos, existe una gran diversidad de andamiajes estructurales. Al realizar un análisis de agrupamiento basado en la similaridad de la base de datos AID 2158, utilizando MACCS (Molecular ACCess System) *fingerprints* con un valor de corte del coeficiente de Tanimoto de Tc = 0.70, se identificaron 30 grupos con varios compuestos cada uno, ademas de mas de 100 grupos con un solo elemento.

Teniendo en cuenta la heterogeneidad del *dataset* y para facilitar la búsqueda del origen de la pobre capacidad predictiva del *docking,* decidimos enfocarnos en un grupo más homogéneo de compuestos.

Dado que nuestro interés se centra en inhibidores no covalentes, como primer paso para reducir la heterogeneidad del *dataset*, hemos descartado los compuestos que contienen grupos electrofílicos. Esto ha reducido drásticamente el número de compuestos, ya que la mayoría de los *hits* identificados en el dataset original AID 1478 corresponden a moléculas con grupos reactivos, lo cual tiene sentido considerando la presencia del grupo tiol altamente nucleofílico de las cisteína proteasas.

Tras eliminar los compuestos con grupos electrofílicos, destaca un grupo con 44 derivados de benzimidazol (36 confirmados como activos), entre los cuales se encuentra el inhibidor no covalente B95 co-cristalizado con la cruzipaina (código PDB: 3KKU).

En la figura 6.4, se muestra el resultado del *docking* nativo de B95 superpuesto al modo de unión experimental. Tanto el *docking* libre como el guiado por puntos farmacofóricos (PFs) fueron considerados.

Siguiendo un modelo de ajuste inducido, los residuos del surco de unión deberían complementar perfectamente la estructura del ligando en el cristal, lo que implica que el *docking* no debería tener problemas para identificar la pose cristalográfica de B95. Sin embargo, en ambos casos (con y sin PFs), el docking falla en predecir el modo de unión correcto. Esta misma dificultad también se observó entre los restantes 43 derivados de benzimidazol, donde solo unos pocos adoptan una pose similar al modo de unión cristalográfico (8 y 6 ligandos, con y sin PFs, respectivamente).

Las poses de estos derivados se pueden agrupar en dos modos de unión diferentes: uno en el cual el enlace amida simil-peptídico clave para el reconocimiento molecular se ancla en el sitio S1, formando el típico patrón de enlaces de hidrógeno (Gly66)N-H...O=C-N-H...O=C(Asp161), como en la estructura cristalina 3KKU; y otro modo de unión donde el inhibidor está invertido y el mismo enlace amida se acopla en el sitio S1', formando un patrón de enlace de hidrógeno (Gln19)N-H...O=C-N-H...O=C(Asp161). Nótese que el docking predice principalmente este modo de unión no cristalográfico para el ligando B95 (figura 6.4).

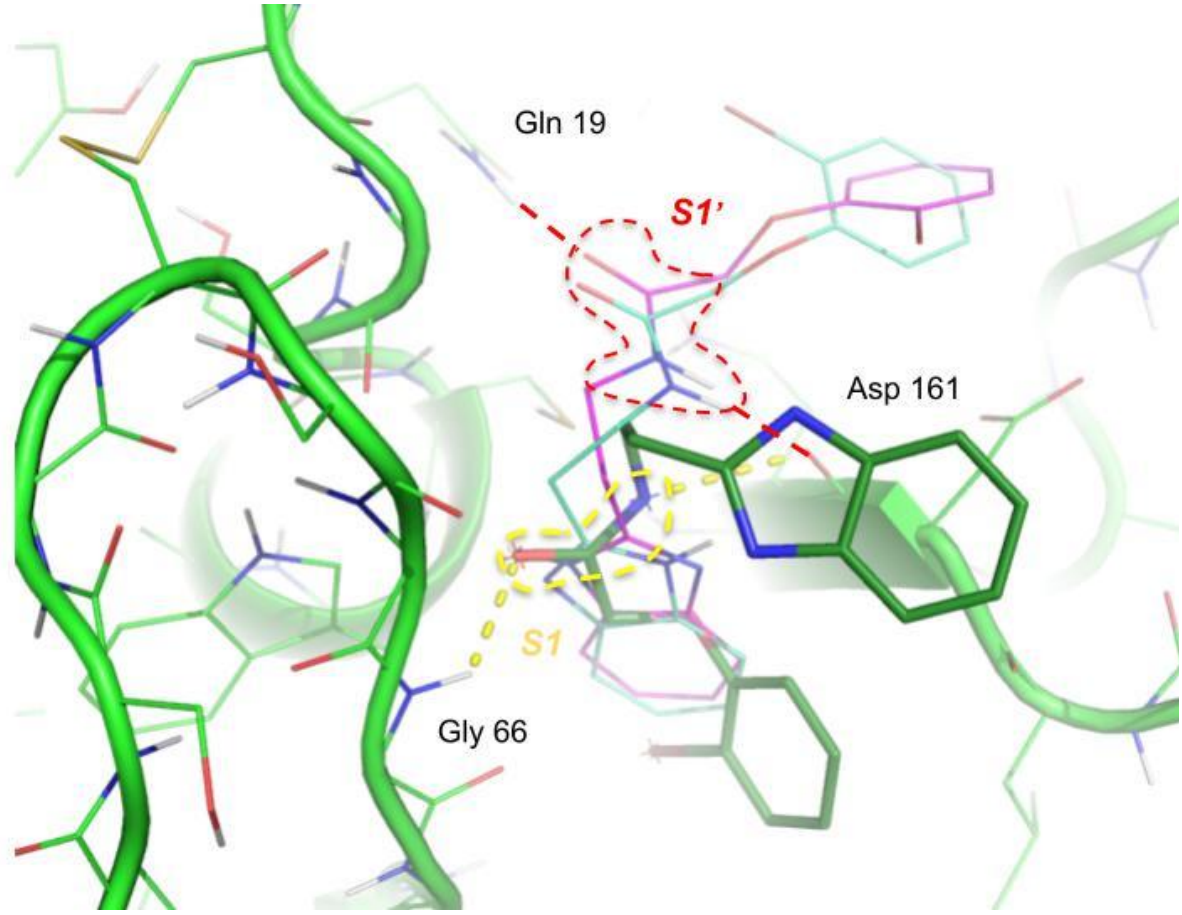


Figura 6.4 Poses del *docking* del ligando B95 con (cian) y sin (magenta) ayuda de puntos farmacofóricos, superpuestas al modo de unión cristalográfico del mismo ligando en el cristal 3KKU (verde).

La simulación por dinámica molecular de esta conformación invertida del inhibidor B95 muestra que ésta no es estable, ya que el inhibidor se desprende rápidamente del surco de unión. Sorpresivamente, después de unos pocos nanosegundos, B95 se ancla nuevamente, pero esta

vez adoptando la pose cristalográfica, con el enlace simil-peptídico unido al sub bolsillo S1 y formando los enlaces de hidrógeno característicos con la cadena principal de Asp 161 y Gly 66 (véase figura 6.5). Notablemente, el inhibidor permanece anclado en esta pose durante el resto del tiempo de simulación, lo que confirma la estabilidad del modo de unión cristalográfico.

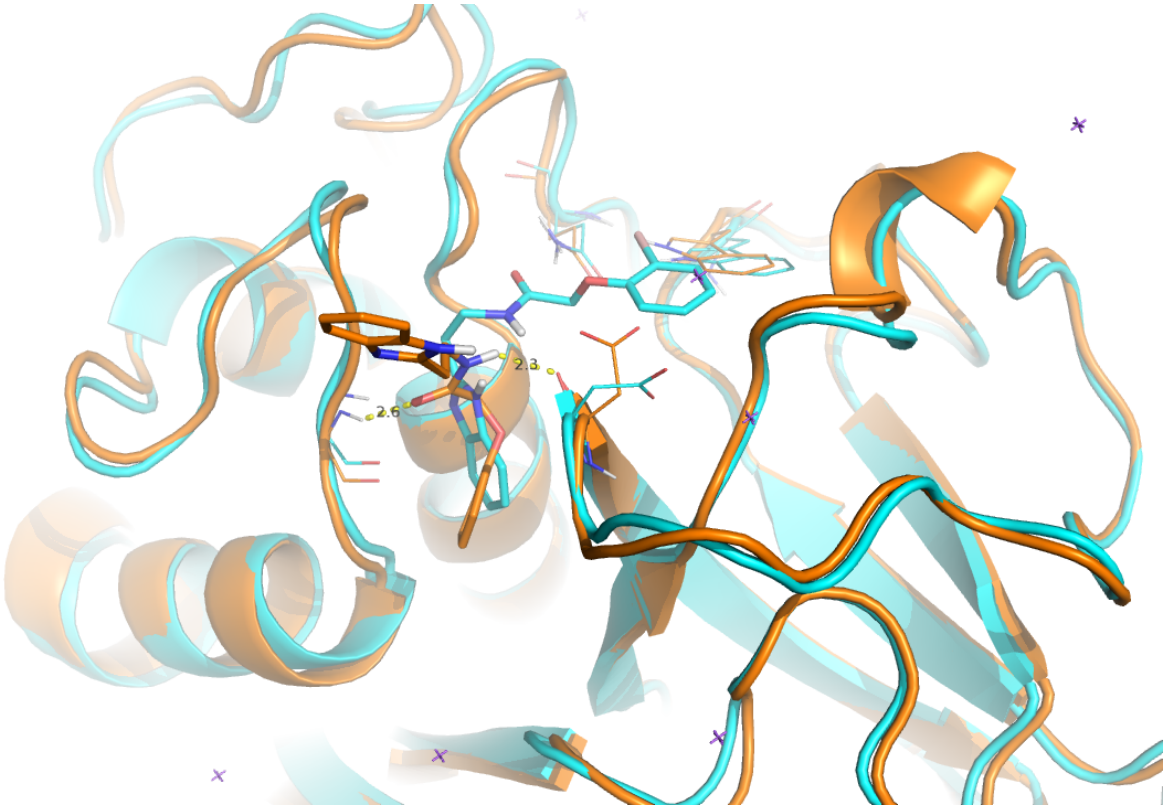


Figura 6.5 Modo de unión de B95 antes (cian) y luego (naranja) de 20 ns de simulación por dinámica molecular de la pose no cristalográfica del inhibidor.

6.3.1.2 Explorando modelos más complejos de interacción ligando-receptor

Los resultados previos resultan desalentadores, ya que nuestro protocolo de *docking* en su estado actual no ha logrado predecir correctamente el modo de unión experimental del ligando sobre su proteína nativa, aún luego de aplicar restricciones farmacofóricas.

Aparentemente, la unión del enlace simil-peptídico del ligando en el sub bolsillo incorrecto S1' estaría mucho más favorecida que el anclaje en el sub bolsillo S1.

Una posible explicación para este comportamiento es que la conformación de la enzima capturada en el cristal 3KKU podría ser muy diferente a la conformación inicialmente reconocida

por el ligando. Bajo esta suposición, podríamos plantear la hipótesis de que el ligando "selecciona" inicialmente una conformación específica de la enzima entre las diversas conformaciones que ésta puede adoptar en su forma libre. Una vez que el ligando se ha unido a esta conformación específica, éste “induce" los cambios en la proteína que finalmente llevan a la estructura observada en el cristal. Estos cambios serían lo suficientemente profundos, de manera tal que cuando se realiza el re-acoplamiento del ligando sobre esta misma estructura final, se obtiene una pose diferente a la esperada.

Este tipo de modelos híbridos, que combinan la selección conformacional y el ajuste inducido, son bastante comunes en el estudio de sistemas biológicos, por ejemplo son utilizados para describir la activación de los receptores acoplados a proteína G (GPCRs) (Zou Y et al. 2019). En estos casos, el ligando se une inicialmente al GPCR en una conformación específica, y a través de cambios conformacionales posteriores, se desencadena la señalización celular.

En consecuencia, para investigar esta hipótesis, a continuación llevamos a cabo extensas simulaciones de dinámica molecular de la forma libre de la enzima. El objetivo fue muestrear las diversas conformaciones que la enzima puede adoptar, con la intención de evaluar posteriormente si alguna de ellas es capaz de unir correctamente a los derivados de benzimidazol, incluyendo el inhibidor B95.

6.3.1.3 Selección estructuras representativas de apo Cruzipaina

La forma apo o libre de Cruzipaina, tal como se encuentra en el cristal 3KKU fue sometida a 0,5 microsegundos de simulación por Dinámica Molecular (DM).

Para tener una perspectiva de la variabilidad conformacional total de la Cruzipaina, y así poder realizar un muestreo conformacional más informado, se realizó un análisis de componentes principales (PCA) de la trayectoria de DM.

Los resultados del PCA muestran que 29% de la variabilidad conformacional total está contenida en las primeras 3 componentes principales (PC1, PC2 y PC3), de las cuales PC1 da cuenta de más de la mitad (15%) mientras que el resto se distribuye entre PC2 (7%) y PC3 (6%). La contribución de las restantes componentes a la varianza es aún menor y se distribuye de

manera casi equitativa entre ellas, con lo cual su aporte individual a la variabilidad conformacional resulta despreciable.

El histograma de la figura 6.6 muestra la proyección de la trayectoria sobre las componentes PC1, PC2 y PC3.

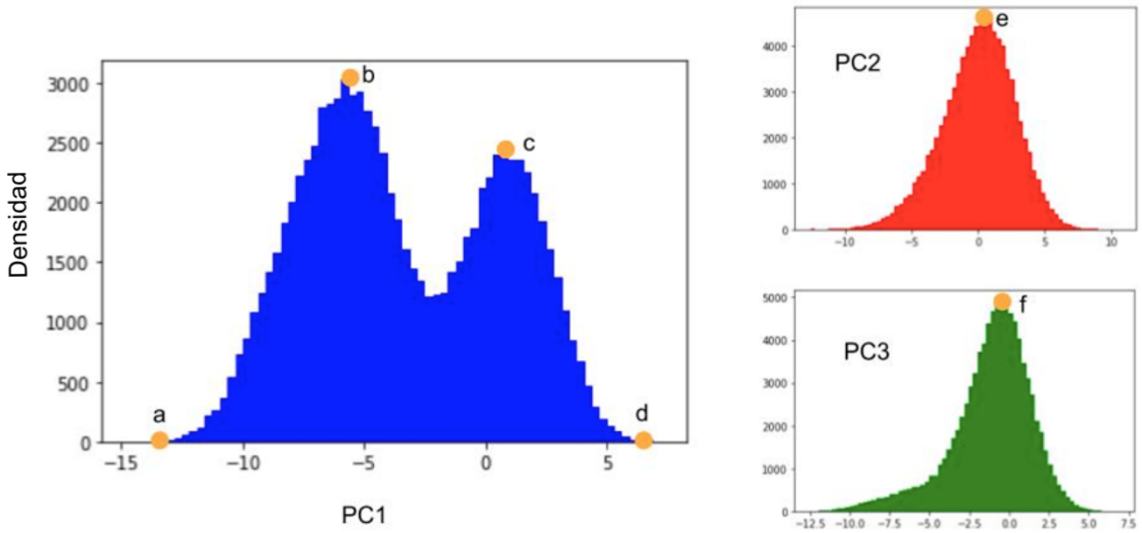


Figura 6.6 Proyección de la trayectoria de la forma apo (libre) de Cruzipaina sobre las componentes principales 1, 2 y 3.

PC1 muestra una distribución bimodal, con un máximo global b y un máximo local c, mientras que PC2 y PC3 presentan un único máximo.

Las estructuras correspondientes a los valores extremos en dichas distribuciones pueden dar una idea de la variabilidad conformacional total de la enzima. La Figura 6.7A muestra ambas estructuras extremas en la distribución de PC1 superpuestas (estructuras en celeste y gris, correspondientes a los puntos extremos a y d en la Figura 6.6, respectivamente).

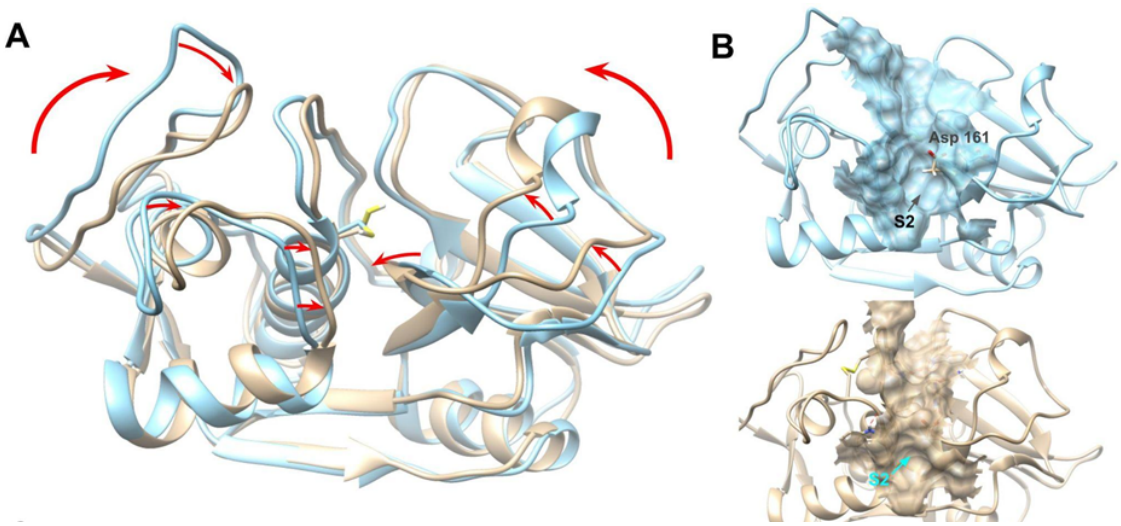


Figura 6.7 Conformaciones extremas de apo Cz en la distribución de PC1. En celeste y gris se muestran las estructuras correspondientes a los puntos extremos a y d en la figura respectivamente.

Las estructuras superpuestas muestran que ambos dominios de la proteína, el dominio de hélices α (izquierdo) y el de láminas β (derecho) experimentan movimientos concertados en dirección opuesta que tienden a cerrar/estrechar la hendidura de unión del sustrato que se encuentra en el medio, entre ambos dominios. Así, podemos asociar a las estructuras en celeste y gris con las conformaciones abierta y cerrada de la enzima, respectivamente.

En la forma cerrada el surco de unión se encuentra totalmente ocluido, especialmente a nivel de los sub bolsillos S1 y S2 (ver figura 6.B), lo que imposibilita el anclaje del inhibidor. Por lo tanto de aquí en más nos enfocamos en la forma “drogable”, es decir la forma de cadena abierta de apo Cz para realizar el *docking* del grupo de derivados de benzimidazol.

Como se observa en la tabla 6.1, el *docking* sin restricciones no arrojó buenos resultados, dado que solo 3 ligandos adoptan una pose similar a la cristalina, con el enlace simil-peptidico anclado en el sub bolsillo S1. Los restantes 41 derivados mostraron poses muy variadas dentro de la hendidura de unión, lo que indica que ya no existe una única pose competitiva como en el caso de 3KKU, sino varias. Esta variabilidad posiblemente se deba a la expansión global del surco de unión provocada por los desplazamientos de las cadenas principales en la forma abierta de la proteína.

Sin embargo, tras la inclusión de las restricciones farmacofóricas para guiar las soluciones de *docking*, los resultados mejoraron notablemente. En este caso los 44 derivados de benzimidazol logran acomodar el enlace simil-peptídico en el bolsillo S1, formando el característico patrón de interacciones (Gly66)N–H···O=C–N–H···O=C(Asp161). Estos resultados se resumen en la tabla 6.1

Tabla 6.1 Cantidad de ligandos unidos correctamente (según modo cristalográfico) frente a las diferentes formas de Cz.

Modelo	**N° poses simil-cristal**
3KKU (sin PFs)	6/44
3KKU (con PFs)	8/44
forma abierta apo Cz (sin PFs)	3/44
forma abierta apo Cz (con PFs)	44/44

Estos resultados sugieren que la conformación abierta de la enzima sería la forma inicialmente “reconocida” por los ligandos de Cz.

Para explicar el origen de esta diferencia entre las poses de *docking* sobre la forma de cadena abierta y la estructura cristalina 3KKU, la figura 6.8 muestra la superposición de las cadenas principales de ambas estructuras proteicas.

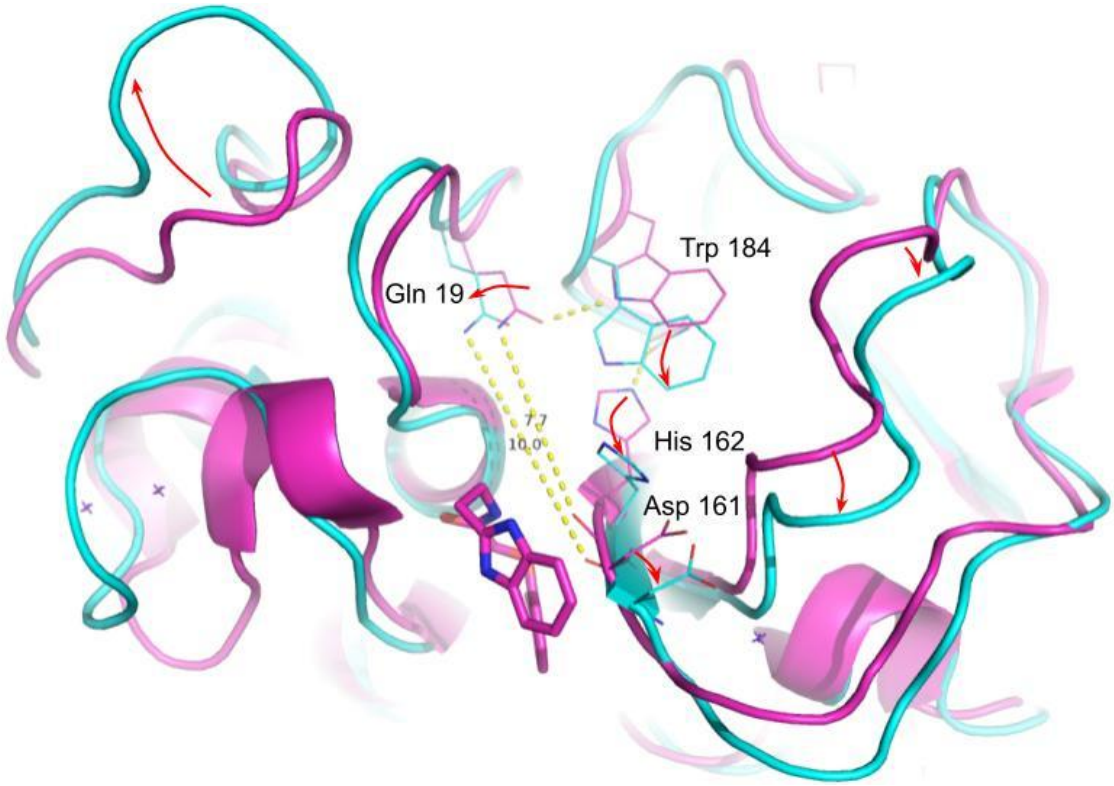


Figura 6.8 Superposición de la estructura 3KKU (magenta) sobre la conformación abierta de apo Cz (celeste).

Como lo indican las flechas en la figura 6.8, los desplazamientos de la cadena principal de la proteína al pasar de la estructura cristalina a la forma abierta en apo Cz provocan, además de una expansión global del surco de unión del sustrato, un aumento de la distancia entre la cadena principal de Asp 161 y la cadena lateral de Gln 19 (de 7,7 a 10,0 Å). En consecuencia, estos residuos ya no se encontrarían a una distancia óptima para actuar como aceptor y dador de hidrógeno frente al enlace simil-peptídico del inhibidor (ver figura 6.5).

Esta diferencia conformacional entre la estructura cristalina y la forma abierta sugiere que en esta última el sub-bolsillo S1' ya no funcionaría como un sitio de unión competitivo con el sub-bolsillo S1, favoreciendo así el acoplamiento en el mismo modo de unión observado en la estructura cristalina.

En consecuencia, incorporamos esta estructura extrema a lo largo de PC1 en nuestras campañas de cribado virtual prospectivo.

Por otro lado, aunque las estructuras extremas a lo largo de la PC1 dan una idea de la variabilidad conformacional global de la proteína, éstas presentan una densidad poblacional muy

baja en el histograma de la figura 6.6 (puntos de mínimo a y d en PC1). En cambio los máximos en la distribución de PC1, PC2 y PC3 (puntos b, c, e y f en la Figura 6.6) corresponden a valores de dichas componentes que son más frecuentemente visitados por Cz a lo largo de la simulación. Por lo tanto, el muestreo de la trayectoria en la proximidad de estos máximos posiblemente recupere conformaciones de la enzima que sean más representativas de las simulaciones.

El análisis de las estructuras de apo Cz muestreadas en la proximidad de los máximos global y local en la distribución de PC1 (y simultáneamente en la proximidad de los máximos en PC2 y PC3), reveló que estos comparten características comunes con las estructuras abierta y cerrada de la enzima, respectivamente, aunque las diferencias no son tan marcadas como entre las estructuras correspondientes a los extremos (no mostrado).

Por lo tanto, la conformación del máximo global a lo largo de PC1 también fue seleccionada para el cribado virtual prospectivo por poseer características comunes con la forma de cadena abierta y ser al mismo tiempo una conformación más representativa de apo Cz.

De aquí en más nos referimos a ambas conformaciones, la forma de cadena abierta y el máximo global a lo largo de PC1, como PC1_OPEN y PC1_MAX, respectivamente.

6.3.2 Segunda parte: Cribado Virtual Prospectivo

La mayoría de los inhibidores de la Cruzipaina reportados hasta la fecha consisten en moléculas que contienen una cabeza electrofílica y actúan mediante una modificación irreversible y covalente de la enzima. Sin embargo, es ampliamente conocido que los inhibidores covalentes suelen presentar problemas de toxicidad debido a su reactividad cruzada con proteínas del hospedador. Por esta razón, en esta tesis hemos decidido enfocarnos en perseguir inhibidores no covalentes de la enzima.

Si bien la Cruzipaina es un blanco molecular prometedor para la búsqueda de nuevas alternativas terapéuticas para la enfermedad de Chagas, ha demostrado ser especialmente esquivo, en particular si se pretende buscar inhibidores no covalentes. Esto se debe a que el surco de unión de la Cruzipaina presenta poca profundidad y alta exposición al solvente, además de mostrar una gran flexibilidad y dinámica. Estas características dificultan el diseño de moléculas que puedan anclarse con alta afinidad sin formar enlaces covalentes.

Conscientes de estas dificultades, hemos optado por emplear diversas estrategias para la búsqueda de compuestos, derivados de las campañas de cribado virtual. De esta manera, buscamos incrementar las posibilidades de descubrir un nuevo inhibidor no covalente de la enzima.

Un primer enfoque consistió en un Cribado Virtual Basado en la Estructura (CVBE) empleando una estrategia de ensemble *docking* y guiado mediante puntos farmacofóricos.

Para tener en cuenta la variabilidad conformacional de la enzima, realizamos el cribado de la biblioteca de compuestos sobre cuatro estructuras diferentes de Cruzipaína. Estas incluyeron las dos estructuras derivadas del análisis de PCA de las simulaciones de la forma apo de Cz, PC1_OPEN y PC1_MAX, así como las estructuras cristalinas de Cz unidas a un inhibidor no covalente (3KKU) y a uno covalente (2OZ2).

Basándonos en los conocimientos acerca de las interacciones principales que guían el correcto anclaje de los ligandos al sitio activo, se introdujeron puntos farmacofóricos al algoritmo de rDock a modo de restricciones para guiar las soluciones del docking.

Empleando este protocolo de "*docking* guiado" se cribó una biblioteca in-house de 7 millones de compuestos adquiribles comercialmente. Se trata de una biblioteca compilada a partir de catálogos comerciales que incluyen 6 proveedores fiables y relativamente económicos.

La figura 6.2 muestra con mayor nivel de detalle el protocolo de cribado virtual y la aplicación de filtros pre y pos-docking para reducir progresivamente el número de compuestos que se seleccionan en cada etapa del proceso.

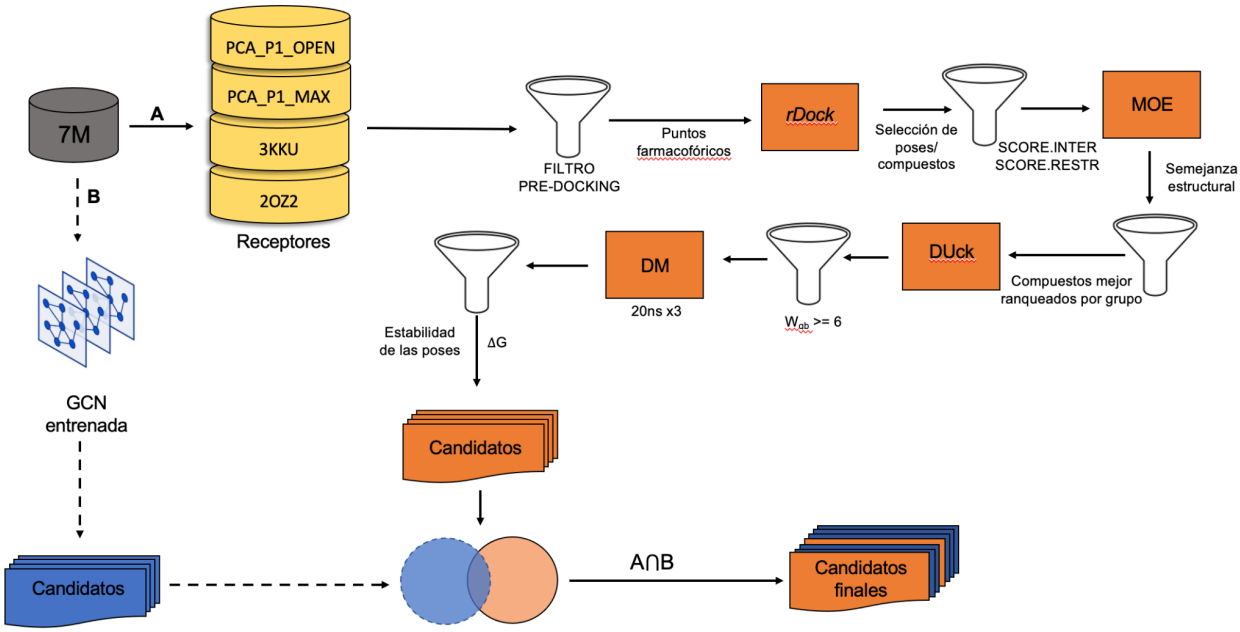


Figura 6.9 Esquema general del procedimiento utilizado para el cribado virtual de una quimioteca utilizando herramientas supervisadas (A) y no supervisadas (B).

6.3.2.1 Filtros pos-*docking* y selección de candidatos

Los candidatos se seleccionaron a partir del análisis de diferentes parámetros obtenidos luego de realizar las corridas de *docking.* En particular se seleccionaron aquellas moléculas que presentaban la mayor puntuación total (SCORE.TOTAL) de acuerdo a la función de puntuación del algoritmo. Seguidamente, las poses de aquellos ligandos mejor "clasificadas" fueron ordenadas de menor a mayor valor en términos de SCORE.INTER y SCORE.RESTR que describen la energía de interacción ligando-proteína y el grado de satisfacción de las restricciones farmacofóricas, respectivamente. Aquellos compuestos con SCORE.INTER > -20 y SCORE.RESTR > 0.5 fueron descartados.

Posteriormente, empleando *fingerprints* (MACCS) para la representación de las moléculas, la lista de compuestos fue reorganizada en base a similitud estructural. Aquellos compuestos con una similaridad >85% de acuerdo con el coeficiente de Tanimoto fueron agrupados dentro del mismo grupo. De cada conjunto de moléculas se seleccionó el representante mejor clasificado de acuerdo con el valor de SCORE.INTER, reduciendo así el número de moléculas. Este agrupamiento se realizó con el *software* MOE (*Molecular Operating Enviroment,* https://www.chemcomp.com/).

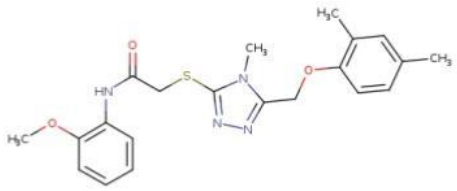
En este punto del filtrado, la selección de compuestos se ha reducido a un número razonablemente bajo como para permitir evaluar la estabilidad de la unión empleando herramientas computacionales más precisas. Así, los complejos de Cz-ligando con los compuestos más promisorios fueron sometidos a simulaciones de Dinámica Molecular Dirigida (SMD del inglés, *Steered Molecular Dynamics*) empleando un protocolo denominado *Dynamic Undocking* o DUck (Majewski, Ruiz-Carmona, y Barril 2018).

DUck calcula el trabajo necesario para alcanzar un estado "cuasi enlazado" (W_{qb}) en el que el ligando acaba de romper el contacto nativo más importante con el receptor. Esta propiedad de no equilibrio es sorprendentemente efectiva en el cribado virtual porque los ligandos verdaderos forman interacciones más resistentes que los señuelos (Majewski, Ruiz-Carmona, y Barril 2018).

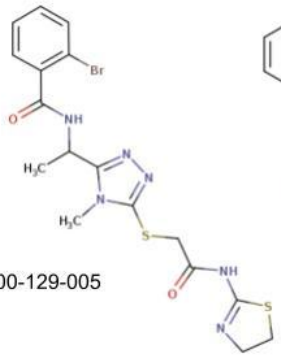
Aquellos compuestos con un valor de W_{qb} >=6 kcal/mol fueron seleccionados para pasar a la siguiente etapa del protocolo. De esta manera se filtraron aún más las moléculas a seleccionar, las cuales fueron posteriormente sometidos a simulaciones de DM libre, seleccionando los compuestos que preservaron el modo de unión del *docking* luego de 20 ns de simulación por triplicado (en 3 réplicas). Estas simulaciones evalúan la estabilidad global del anclaje N (donde N es el número de réplicas, de 1 a 3, en las cuales el ligando permanece anclado en el surco de unión) y la energía libre de unión (ΔG) (ver figura 6.1).

Por último, aquellos compuestos seleccionados a partir del pos-procesamiento del *docking* fueron comparados con aquellos seleccionados por la GCN previamente entrenada (ver capítulo 5), llegando así al número final de candidatos a ensayar en futuros estudios experimentales.

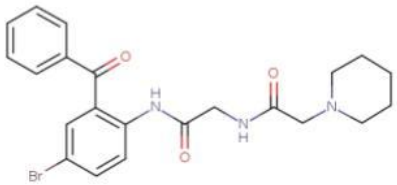
Los ligandos obtenidos a partir del filtrado de la biblioteca *in-house* se presentan a continuación:



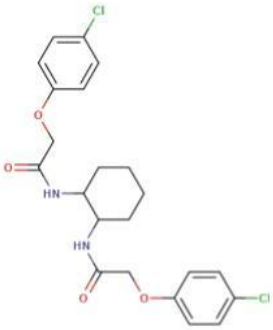
Molport-000-121-481



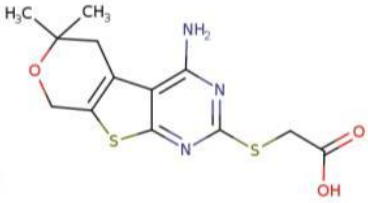
Molport-000-129-005



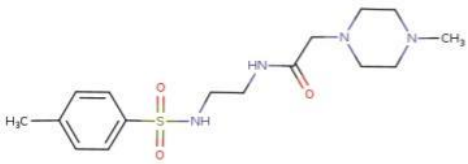
Molport-000-721-497



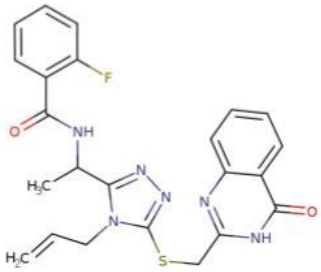
Molport-001-523-757



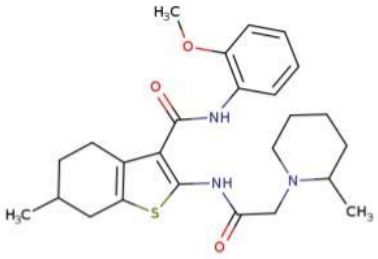
Molport-001-929-225



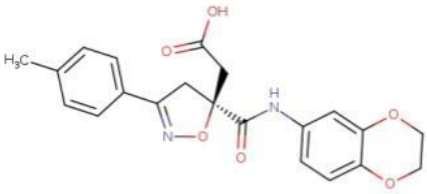
Molport-001-992-930



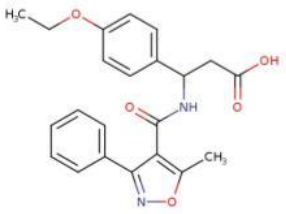
Molport-002-026-751



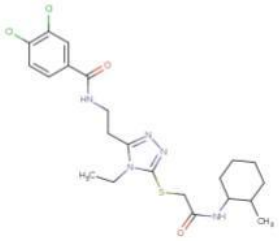
Molport-002-340-324



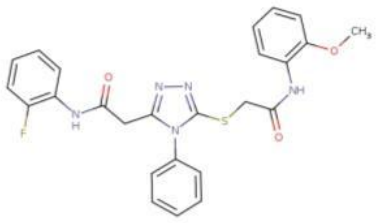
Molport-002-661-407



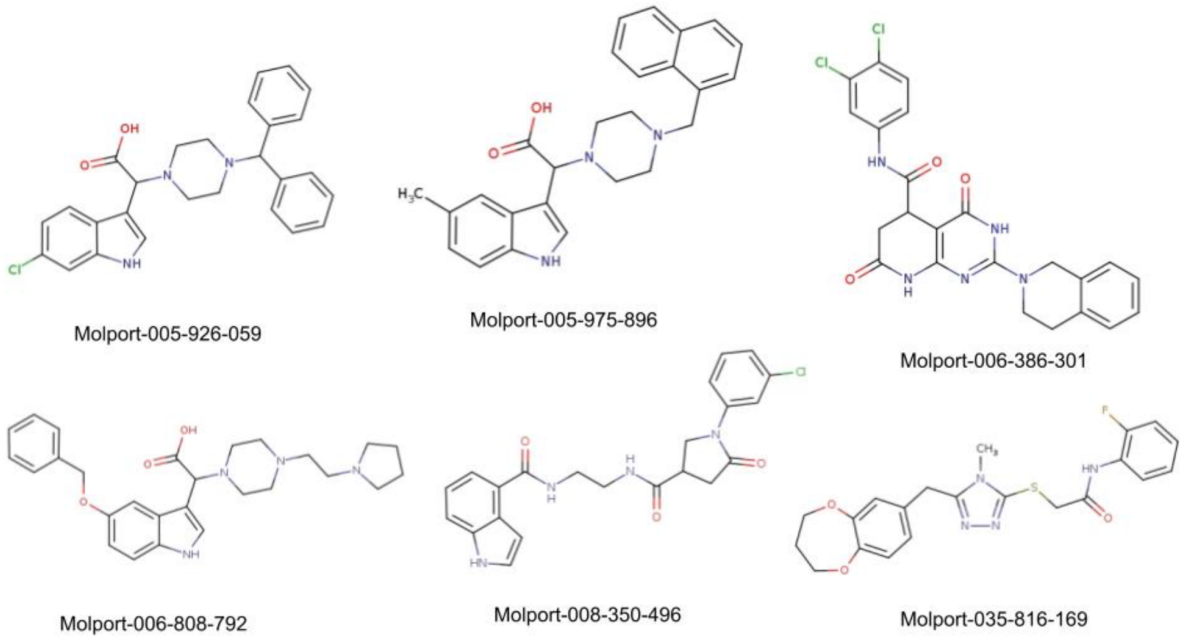
Molport-002-724-447



Molport-002-983-910



Molport-002-989-223



Como se observa, la mayoría de las moléculas seleccionadas mediante este protocolo contienen dos o tres anillos (hetero)aromáticos conectados por “linkers” que incluyen al menos una unión amida. Esta unión simil-peptídica parece ser crucial para el anclaje de los ligandos en la hendidura de unión estrecha situada entre los sub-bolsillos S1 y S2. Es importante destacar que los compuestos seleccionados carecen de grupos electrofílicos capaces de unirse covalentemente a la Cys25. Aunque se han reportado inhibidores no covalentes de la enzima, como los benzimidazoles (Bezerra Morais et al. 2023), la mayoría de los inhibidores de Cruzipaina descubiertos hasta la fecha, al igual que los de otras cisteín proteasas, están basados en cabezas de guerra electrofílicas como las vinilsulfonas, los derivados de nitrilos y las tiosemicarbazonas.

6.4 Conclusiones

Este capítulo de la tesis se enfocó principalmente en el cribado virtual de una librería de 7 millones de compuestos accesibles comercialmente.

En una primera instancia los esfuerzos se centraron en corroborar la exactitud de los algoritmos de docking para recuperar inhibidores conocidos sobre la base de datos AID 2158. Como se pudo observar, en gran medida los resultados de docking estaban sujetos a los que estructura de Cz que se utilizaba, demostrando que el reacomodamiento de los sub-bolsillos, así como el

surco de unión en general juegan un papel crucial en el correcto anclaje de ligandos y una vez allí el surco se cierra priorizando interacciónes que estabilizan los ligandos. A raiz de estos hallazgos se decidió utilizar 4 diferentes tipos de estructuras de Cz que diferian levemente en su estructura 3ría y que junto con los puntos farmacofóricos hacián de guia para el correcto anclaje de moléculas.

Cabe destacar que la incorporación de la GCN dentro del protocolo de Cribado Virtual (CV) nos ha permitido mejorar el rendimiento del docking, mediante el pre-filtrado de falsos positivos y el guiado mediante restricciones farmacofóricas aprendidas por la GCN. A su vez, se observó que al aplicar un filtrado previo al docking de la biblioteca de compuestos, utilizando una Red Convolucional basada en Grafos (GCN), pre-entrenada únicamente con la estructura 2D de los ligandos, se lograba una mejora significativa en la tasa de recuperación de compuestos activos.

Por último, las herramientas para el pos-procesado de las campañas de docking demostraron ser sumamente útiles, ya que permitieron continuar el filtrado de gran cantidad de compuestos, utilizando enfoques de relativo bajo costo computacional.

Este último capítulo demuestra que un enfoque híbrido entre cribado virtual basado en estructura y ligandos, mediado por algoritmos de aprendizaje automático, es posible con el correcto calibrado de las herramientas computacionales utilizadas.

Referencias del capítulo 6

Baldi, A. (2010) Computational Approaches for Drug Design and Discovery: An Overview. Sys. Rev. Pharm. 1(1), 99-105

Bezerra Morais, P. A., & Goulart Trossini, G. H. (2023). Cruzain Inhibitors: State-of-Art of Novel Synthetic Strategies. *Current Organic Chemistry,* 27(4), 243-247.

Majewski, Maciej, Sergio Ruiz-Carmona, and Xavier Barril. 2018. “Dynamic Undocking: A Novel Method for Structure-Based Drug Discovery.” Methods in Molecular Biology (Clifton, N.J.) 1824: 195–216. https://doi.org/10.1007/978-1-4939-8630-9_11.

Ruiz-Carmona, Sergio, Daniel Alvarez-Garcia, Nicolas Foloppe, A. Beatriz Garmendia-Doval, Szilveszter Juhos, Peter Schmidtke, Xavier Barril, Roderick E. Hubbard, and S. David Morley. 2014. “RDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids.” PLOS Co*mputational Biology* 10 (4): e1003571. https://doi.org/10.1371/JOURNAL.PCBI.1003571.

Sutherland JJ, Nandigam RK, Erickson JA, Vieth M. Lessons in molecular recognition. 2. Assessing and improving cross-docking accuracy. J Chem Inf Model. 2007;47:2293–2302. http://dx.doi.org/10.1021/ci700253h

Wiggers, H. J., Rocha, J. R., Cheleski, J., & Montanari, C. A. (2011). Integration of Ligand- and Target-Based Virtual Screening for the Discovery of Cruzain Inhibitors. Molecular Informatics, 30(6-7), 565–578. doi:10.1002/minf.201000146

Zhu, J., Chen, T., Liu, J., Ma, R., Lu, W., Huang, J., ... & Jiang, H. (2009). 2-(3, 4-Dihydro-4-oxothieno [2, 3-d] pyrimidin-2-ylthio) acetamides as a new class of falcipain-2 inhibitors. 3. Design, synthesis and biological evaluation. *Molecules,* 14(2), 785-797.

Zou Y, Ewalt J, Ng HL. Recent Insights from Molecular Dynamics Simulations for G Protein-Coupled Receptor Drug Discovery. Int J Mol Sci. 2019 Aug 29;20(17):4237. doi: 10.3390/ijms20174237. PMID: 31470676; PMCID: PMC6747122.

CAPÍTULO VII

“Conclusiones generales”

La cruzipaina es una de las principales cisteino-proteasas del *T. cruzi* y ha sido señalada por diversos autores como una de las principales moléculas blanco de inhibidores, ya que participa en importantes vías metabólicas del parásito, y su inhibición es crucial para evitar la proliferación del parásito.

Durante el desarrollo de esta tesis se han utilizado distintas metodologías propias de la química computacional, desde la mecánica clásica en las simulaciones de dinámica molecular y *docking* molecular a la mecánica cuántica, así como otras un poco más novedosas como la inteligencia artificial y el aprendizaje no supervisado en las distintas etapas de la búsqueda de nuevos inhibidores.

El trabajo realizado se puede dividir en cuatro líneas generales: análisis estructural de la cruzipaína y estudio de interacciones de ligandos conocidos en el sitio catalítico, combinación del análisis de densidad de carga con herramientas de aprendizaje automático para investigar el mecanismo de inhibición de Cz, cribado virtual retrospectivo de una biblioteca de ligandos y parametrización del algoritmo de docking y por último, cribado prospectivo de una biblioteca de ligandos. Los resultados obtenidos se pueden condensar en las siguientes conclusiones generales:

En este trabajo de tesis se recopiló la información estructural disponible sobre inhibidores de Cz para generar descriptores basados en la densidad electrónica y sus propiedades locales asociadas. La aplicación de la metodología *QTAIM* permitió detectar interacciones no direccionales, por ejemplo, aquellas que involucran electrones π en anillos aromáticos, entre otros contactos débiles e inusuales que de otro modo se perderían en un análisis meramente geométrico de las interacciones. A su vez, bajo esta teoría se pudo descomponer la sumatoria de valores de densidad de carga de interacciones (interacciones totales ligando-receptor) en contribuciones por átomo o grupos de átomos lo cual la hace particularmente útil para entender la importancia de ciertos sub-bolsillos en el anclaje de ligandos.

Más allá de la robustez de QTAIM, como herramienta para el estudio e identificación de interacciones claves en el sitio catalítico, es necesario el uso de otras herramientas computacionales capaces de clasificar y ponderar dichas interacciones siguiendo anotaciones

de afinidad. A su vez, la utilización de métodos tradicionales aplicados en la química computacional no son suficientes debido a la gran cantidad de datos generados lo que requiere la utilización de herramientas basadas en el aprendizaje automático.

Utilizando los elementos topológicos de la densidad de carga que describen las interacciones en los complejos Cz-ligando, se entrenó un modelo de clasificación de aprendizaje supervisado SVM-RFE capaz de discriminar entre las interacciones presentes en los complejos de los inhibidores más activos (interacciones de tipo activo) y las que ocurren en los menos activos (interacciones de tipo inactivo). De este modo se obtuvieron interacciones relevantes que estabilizan una conformación particular de Cz (llamada en esta tesis, conformación activa).

Por último, la combinación de estrategias convencionales utilizadas en campañas de cribado virtual retrospectivo/prospectivo pudieron ser mejoradas mediante la implementación de un enfoque híbrido que contemple el uso de Redes Neuronales Gráficas (como GCN) para capturar relaciones complejas de los datos y realizar mejoras en las diferentes etapas del cribado virtual mediado por *docking* molecular. Mediante este protocolo optimizado de cribado virtual se logró priorizar una colección de 18 candidatos a inhibidores de Cz, los cuales deberán ser ensayados experimentalmente.

En síntesis, la tesis presenta una estrategia integrada que aprovecha las ventajas de dos herramientas clave: los modelos de redes neuronales (GCN) para capturar las relaciones complejas de los datos y la interpretabilidad del docking molecular basado en la estructura. Finalmente, la combinación de enfoques computacionales complementarios, en campañas de cribado virtual destinadas a la búsqueda de nuevos inhibidores de la cruzipaína del Trypanosoma cruzi, constituye la contribución más relevante de este trabajo de tesis doctoral.