UNIVERSIDAD NACIONAL DEL NORDESTE

Facultad de Ciencias Exactas y Naturales y Agrimensura

IMPLEMENTACIÓN DE MÉTODOS DE MECÁNICA CUÁNTICA PARA MODELAR LA INTERACCIÓN LIGANDO-PROTEINA Y SU APLICACIÓN EN EL DESCUBRIMIENTO DE FÁRMACOS LÍDERES

TESIS QUE PRESENTA

Lic. María Gabriela Aucar

PARA OBTENER EL TITULO DE

Doctor de la Universidad Nacional del Nordeste en Física

Bajo la dirección del DR. CAVASOTTO, CLAUDIO (IIMT)

y la codirección del DR. ROMERO, RODOLFO HORACIO (UNNE)

Corrientes, Argentina 2021

A mis padres

Agradecimientos

Quisiera agradecer a todos los que hicieron posible que hoy cierre con profunda alegría esta etapa importante de mi vida. No hubiera llegado hasta acá sin el aporte de todos los que se fueron haciendo presentes en este tiempo.

Agradezco en primer lugar a Claudio, mi director de Tesis. Gracias por la formación de calidad que me supiste brindar con dedicación en estos años, por ayudarme a descubrir el mundo de la Química Computacional. Por confiar en mis capacidades, guiándome en cada etapa del doctorado, y por el apoyo que me diste tanto a nivel profesional como personal.

A mis padres, a quienes dedico esta Tesis. Les estaré siempre agradecida. Porque no solo me dieron la vida sino que supieron guiarme en mis primeros pasos, y darme aliento en los más difíciles. Les agradezco el esfuerzo que hicieron por acompañar mi camino de crecimiento personal y profesional. Por ser ejemplos de constancia y perseverancia. Por ayudarme a aspirar alto.

Quiero agradecer a mis hermanos: Fran, Agus, Emi, Cielo, Juan y Espe. Les agradezco profundamente a cada uno, por la compañía y el apoyo que me dieron en estos años, cada uno con sus matices. Esta etapa, estuvo marcada por profundos dolores que, juntos, se hicieron más livianos, y profundas alegrías que, juntos, se multiplicaron. Gracias por los momentos de vida compartidos. Agus, te agradezco tu gratuidad, tu ayuda concreta en toda esta etapa del doctorado, y el apoyo que me das. Emi, mi hermana mayor. Gracias, porque más allá de las diferencias, realidades y distancia, se que puedo contar con vos en todas.

A Nico, mi esposo y compañero de vida. Por sacar de mi la mejor versión. Por tu paciencia sin límites. Por ayudarme a transitar con alegría esta etapa de mi carrera, haciendo más suave el recorrido. Por este último tiempo, por la paciencia desmedida, por alentarme tan de cerca confiando siempre en mi, aún cuando más costaba, a seguir para adelante y llegar a la meta. Por todo lo que cada día me das, gracias.

Quiero extender mi agradecimiento también a mis primeros compañeros de oficina: Fernando, Leo, Marcos, Agus, Carlos, Pato, Tere, Diego, Ale. Gracias por hacer que la

oficina sea más que un espacio laboral, gracias por los momentos de alegrías compartidas, que voy a recordar siempre.

A Euge, Mabel, y Caro. Por la generosidad con la que supieron atender mis necesidades concretas durante el doctorado, pero más aún por la calidez humana con la que me acompañaron siempre. Gracias por el afecto de madres que me dieron.

A Rodolfo, mi codirector, por hacer posible con tu disponibilidad y buena voluntad que hoy pueda presentar esta Tesis.

A Nati. Gracias por las discusiones, consejos y correcciones que supiste transmitirme. Y sobre todo, por la disponibilidad y predisposición de siempre.

A Marie. Gracias por tus aportes y transmisión de conocimiento, de manera cálida y cercana, en mi primer tiempo de doctorado viviendo en Capital.

A toda la familia Contreras. Gracias por abrirme las puertas de su casa, y de su familia, que hizo que costara un poco menos mi aterrizaje en Buenos Aires. Les estaré siempre agradecida por su generosidad.

A Cristi. Gracias por estar más que presente en estos años de doctorado a pesar de la distancia. Por compartir la vida de becaria doctoral, con todas sus hazañas y darle siempre una cuota de humor a esta etapa, alentándome a no bajar los brazos.

A mis compañeros de oficina del IBioBa: Ivan, Dani, Sol, Gabi, Diego y Fiore. Gracias por recibirme con la mejor onda en el grupo y estar siempre dispuestos a dar una mano.

A mis compañeros de oficina del IIMT: Meli, Pablo, Agos. Y a los vecinos de oficina: Connie, Cande, Marce, Romi, Mai, Maxi y Pau. Gracias a cada uno por la recepción que me dieron desde el primer momento, y más que nada por el gran aliento de este último tiempo.

Agradezco al Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) por la posibilidad, gracias al soporte económico brindado, de realizar mi doctorado.

A las Instituciones que me brindaron no sólo el lugar físico para que desarrollara mi doctorado, sino también la calidad humana de todos sus miembros. En primer lugar al Instituto de Modelado e Innovación Tecnológica (IIMIT) y a la Universidad Nacional del Nordeste (UNNE), donde me formé y pasé mis primeras etapas del doctorado, no sólo como profesional sino como persona. Gracias por dejarme su impronta. Al Instituto de Biomedicina de Buenos Aires (IBioBa), y al Instituto de Investigaciones en Medicina Traslacional (IIMT-Universidad Austral). El recorrido por cada uno de ellos fue dejando en mi momentos por los que voy a estar siempre agradecida.

Por último quiero agradecer al Centro de Cómputo del Instituto IMIT, y a su administrador José Ríos, al Centro de Computación de Alto Rendimiento (CeCAR), el cluster TUPAC del Centro de Simulación Computacional para Aplicaciones Tecnológicas (CSC), el cluster Mendieta del Centro de Computación de Alto Desempeño de la Universidad Nacional de Córdoba (CCAD) y a cada uno de sus administradores por el enorme soporte técnico que me supieron brindar para que pueda sacar adelante mis cálculos.

Resumen

Debido a los grandes avances computacionales ocurridos en los últimos diez años se ha ido introduciendo, como paso crucial en el descubrimiento de fármacos líderes, la implementación de métodos computacionales como herramienta complementaria de los estudios experimentales. El crecimiento exponencial de desarrollos metodológicos y el poder de cálculo alcanzado en este período hicieron posible la implementación de métodos basados en Mecánica Cuántica para realizar cálculos en sistemas biomoleculares, permitiendo así alcanzar una mayor precisión en la descripción de interacciones proteína-ligando. Se presenta en esta Tesis un estudio teórico y computacional orientado a mejorar el *scoring* y el cálculo de otras propiedades mediante el modelado mecano-cuántico de sistemas biomoleculares en solución.

En la primera parte de esta Tesis se describen brevemente los fundamentos teóricos del modelado molecular y se detallan las características principales de las simulaciones computacionales utilizadas para realizar los cálculos presentados. Del mismo modo, dentro de esta primera parte se aborda el desarrollo teórico y las metodologías de cálculo de la energía libre de unión, propiedad esencial para la descripción de interacciones proteína-ligando. Seguidamente, se presenta el Cribado Virtual, una herramienta ampliamente utilizada en la comunidad y de actual desarrollo para la identificación de potenciales ligandos dentro de una base de datos de miles de moléculas.

Partiendo de un estudio comprensivo de los métodos usados para calcular energías libres de unión y describir sistemas biomoleculares en fase acuosa, se implementaron métodos de Mecánica Cuántica en complejos proteína-ligando. El nivel de teoría elegido para la descripción de dichos sistemas fue el de métodos semi-empíricos de Mecánica Cuántica, ya que pueden ser aplicados de manera eficiente en cálculos que involucren sistemas moleculares de miles de átomos, como las proteínas.

Uno de los aspectos que debe ser abordado en primer lugar para la correcta descripción de sistemas biomoleculares es la inclusión de un modelo de solvatación preciso. El modelo de solvente limita la precisión de los cálculos arrojados por métodos semi-empíricos, por lo que se requiere de parámetros optimizados del solvente para mejorar la precisión. Para alcanzar este objetivo se re-parametrizó el modelo de solvente continuo

Conductor like Screening Model (COSMO) para tres Hamiltonianos semiempíricos. Los nuevos parámetros son de gran importancia, dado que se pueden incluir posteriormente en el estudio de complejos proteína-ligando en solución.

Seguidamente, se aplicaron metodologías de cálculo de energía libre en estudios de investigación relacionados con distintas etapas del diseño de un fármaco. La metodología *Molecular Mechanics/Quantum Mechanics-COSMO* (MM/QM-COSMO) fue utilizada para guiar la etapa de optimización de un candidato líder. Por otra parte, se desarrolló una función de *scoring* cuántica para ser aplicada en un contexto de Cribado Virtual (CV).

En la segunda parte de esta Tesis se presentan, en primer lugar, los resultados de la re-parametrización del modelo de solvente continuo COSMO. Para la obtención de dichos parámetros optimizados se incorporó la componente no polar de la energía de solvatación, cuyo valor no está incluido en el modelo. Se utilizaron tres metodologías distintas para el cálculo de dicha componente, analizando comparativamente los resultados obtenidos con cada una. Se presenta también el análisis del impacto del tamaño del conjunto de entrenamiento usado en la re-parametrización.

Es de destacar que una mejor descripción de las interacciones proteína-ligando, provista por métodos cuánticos, podría permitir la identificación del modo correcto de unión del ligando en el sitio activo de un receptor con mayor precisión. Se utilizó entonces el método MM/QM-COSMO, para discriminar la pose de una molécula candidato entre dos conformaciones isoenergéticas resultantes de un proceso de *docking*. Para efectuar los cálculos cuánticos de energía se usó el programa MOPAC, cuyo desarrollo está orientado específicamente a la implementación de métodos semi-empíricos.

Como eje central de ésta Tesis, se presenta el desarrollo y aplicación de una nueva función de *scoring* con métodos semiempíricos, para ser empleada en un contexto de Cribado Virtual automatizado. Esta metodología permite identificar ligandos dentro de una librería química de gran tamaño, incrementando el número de potenciales ligandos dentro del grupo de *hits* seleccionados. Para validar los protocolos desarrollados para la aplicación de dicha metodología se empleó como sistema de estudio un conjunto de 15 complejos proteína-ligando con valores experimentales de afinidad. La determinación de la correlación entre los cálculos de energía libre de unión y los resultados experimentales permitió orientar los esfuerzos para mejorar la precisión de la función de *scoring* cuántica en esta dirección. Luego, fue evaluada la calidad de dicha función en cuanto a su capacidad de separar correctamente ligandos de no ligandos. Para ello se llevó a cabo un estudio retrospectivo en cinco receptores de interés farmacológico, con un conjunto de ligandos conocidos extraídos de una base de datos de libre acceso.

Nomenclatura

ADMET	Absorción, Distribución, Metabolismo, Eliminación y Toxicidad
CDK2	*Cyclin Dependent Kinase 2*. Quinasa dependiente de Ciclina 2
COX1	*Ciclo-oxigenase 1*. Ciclo-oxigenasa 1
CV	Cribado Virtual
DM	*Docking* Molecular
ER	*Estrogen Receptor*. Receptor de Estrógeno
FN	*False Negatives*. Falsos Negativos
FP	*False Positives*. Falsos Positivos
FS	Función de *Scoring*
FEP	*Free Energy Perturbation*
HTS	*High-Throughput Screening*
MD	*Molecular Dynamics*. Dinámica Molecular
SSES	*Scaled Surface Excluding Solvent*. Superficie excluyente del solvente escalada
SASA	*Solvent Accesible Surface Area*. Área superficial accesible al solvente
SBVS	*Structure Based Virtual Screening*. Cribado virtual basado en la estructura del receptor
TN	*True Negatives*. Verdaderos Negativos
TP	*True Positives*. Verdaderos Positivos
QM	*Quantum Mechanics*. Mecánica Cuántica

Índice general

Parte I

Introducción

Introducción

El *reconocimiento molecular* es un mecanismo ligado a muchos procesos biológicos importantes, como la señalización celular, la catálisis y el transporte celular. Este mecanismo se refiere a la unión de dos moléculas, como por ejemplo la unión de ligandos específicos con un receptor determinado, usualmente una proteína. El ligando es una molécula complementaria que se une al receptor. Una mayor comprensión de estos procesos de asociación molecular resulta crucial en la industria farmacéutica y en el ámbito académico, para favorecer el desarrollo de métodos que permitan el descubrimiento de nuevos fármacos de manera rápida y eficaz.

La introducción de un nuevo medicamento en el mercado está asociada a un proceso lento y costoso, que puede durar hasta diez años y se ha estimado puede alcanzar los 2600 millones de dólares desde sus inicios hasta su finalización.[1] Hasta hace unos años, dicho proceso se iniciaba principalmente mediante el cribado experimental automatizado (*High-Throughput Screening*, HTS) para identificar compuestos activos.[2] Sin embargo, la baja tasa de éxito, el costo elevado y el tiempo demandado por las técnicas de HTS han hecho que las mismas pierdan parte del protagonismo de la década pasada.[2,3] Al mismo tiempo, las herramientas de diseño de fármacos asistido por computadoras comenzaron a cobrar mayor importancia gracias a distintos factores, como el crecimiento del número de estructuras tridimensionales resueltas experimentalmente en los últimos años, y el gran avance en materia computacional (desarrollos de hardware y de algoritmos matemáticos). La principal contribución de los métodos computacionales durante los años '80, consistió en la optimización de compuestos a través de la información extraída de estructuras experimentales de complejos cristalográficos proteína-ligando.

En el ámbito académico, en los últimos veinte años, se produjo un crecimiento considerable en el uso de una herramienta computacional, conocida como Cribado Virtual (CV), usando *Docking* Molecular (DM) de receptor-molécula pequeña.[4,5] Las técnicas de CV surgen a finales de los años '90 con el propósito principal de identificar *in silico*, a partir de una librería de millones de moléculas, aquellas que tienen mayor probabilidad de unirse a un receptor determinado. El CV consta de dos etapas: En la etapa de *docking* se coloca una molécula en el sitio de unión del receptor y se caracteriza la pose que adopta,

es decir, su posición, orientación y conformación dentro de dicha región. Una vez seleccionada la pose para cada molécula de la base de datos, una función de *scoring* evalúa la misma, a fin de ordenar las moléculas en función de la afinidad calculada. El resultado de un proceso de CV es entonces una lista de moléculas, ordenada de acuerdo a la puntuación asignada. Usualmente, un pequeño porcentaje de moléculas de la lista es seleccionado para continuar el proceso. Éstas, denominadas *hits*, deben ser evaluadas experimentalmente para confirmar que se unen al receptor, y evaluar su actividad biológica. El CV es usado normalmente en las primeras etapas del descubrimiento de un fármaco, con el objetivo de generar una sub-librería enriquecida con potenciales ligandos.[6] El conjunto de aproximaciones computacionales aplicadas en un contexto de CV representan un complemento a los experimentos, por lo que el diseño de fármacos asistido por computadoras se ha constituido en una etapa determinante en el proceso del desarrollo de un fármaco.[7,8]

Un aspecto crucial en el proceso de asociación de un complejo proteína-ligando, en el contexto del diseño de un fármaco, es la determinación de la estructura y de las propiedades de dicho sistema.[9] En particular, la energía libre de unión[1] es una propiedad de gran relevancia para la industria farmacéutica.[10] Los desarrollos metodológicos orientados al cálculo de esta magnitud física presentan una serie de limitaciones y desafíos actuales, por lo que la correcta estimación de la misma, representa un área de creciente investigación en el campo de la biofísica computacional. Uno de los factores condicionantes que limitan la precisión de los cálculos de energía libre es la necesidad de incorporar la flexibilidad de la proteína en el diseño de fármacos basados en la estructura del receptor. Esto se debe principalmente a la dificultad para explorar la hipersuperficie de energía de altas dimensiones mediante los algoritmos computacionales disponibles. Otro de los desafíos actuales está relacionado con el modelo de solvente empleado para describir las interacciones soluto-solvente y los efectos originados por la presencia de moléculas de agua en el sitio de unión del receptor.

El principal objetivo de la presente Tesis es la optimización y evaluación de métodos de Mecánica Cuántica (*Quantum Mechanics*, QM) para el modelado preciso de interacciones proteína-ligando. Se propone para ello la descripción completa del sistema en forma cuántica. En el cribado virtual automatizado, en particular, el aporte más importante de esta Tesis es el uso de métodos semi-empíricos de QM (SQM) para desarrollar una función de *scoring* precisa, capaz de identificar ligandos dentro de una librería química de grandes dimensiones y generar una sub-librería enriquecida con potenciales ligandos. En este contexto, la nueva metodología representa una mejora frente a las funciones de *scoring* existentes en cuanto al modelo físico empleado para describir el sistema (métodos

[1] La formación de un complejo proteína-ligando tiene asociada una variación en la energía libre de Gibbs, o energía libre de unión, que incluye tanto la variación entálpica como los cambios entrópicos del sistema.

SQM) y el tratamiento de los efectos del solvente con un modelo de solvente continuo optimizado.

Un breve recorrido por los desarrollos teóricos y metodológicos para calcular afinidades de unión, permite poner en contexto la investigación realizada en este trabajo de Tesis. En los últimos años se han presentado numerosos estudios teóricos, así como también desarrollos de algoritmos de cálculo eficientes para estimar energías libres de afinidad. Entre los métodos que alcanzan mayor precisión se encuentran el método de Perturbación de Energía Libre (*Free Energy Perturbation*, FEP) y el método de Integración Termodinámica (*Thermodynamic Integration*, TI).[11] Ambos utilizan, para el solvente, una representación completa e incorporan la flexibilidad conformacional mediante un tratamiento riguroso de los grados de libertad del sistema proteína-ligando. A pesar de la exactitud de este tipo de metodologías, el costo computacional involucrado es alto, por lo que son usados para la optimización de candidatos líderes. Otras metodologías con un enfoque más eficiente desde el punto de vista computacional fueron desarrolladas para ser usadas en un contexto de cribado virtual automatizado (*High Throughput Docking*, HTD) con el objetivo de priorizar moléculas para su evaluación experimental.[12] Entre las metodologías más precisas, como FEP y TI, y los cálculos aproximados de HTD, se encuentran los métodos basados en puntos extremos, en donde el cálculo de energía libre de unión se lleva a cabo usando solamente los estados inicial y final del sistema. El solvente es representado de manera aproximada mediante un modelo de solvente continuo, y se incorpora la flexibilidad conformacional. Desde el punto de vista computacional, son menos costosos que los métodos FEP y TI. Se pueden mencionar dentro de este grupo los métodos *Molecular Mechanics Poisson-Boltzmann Surface Area* (MM-PBSA) y *Molecular Mechanics Generalized Born Surface Area* (MM-GBSA).[13,14] A pesar de haber demostrado resultados promisorios, también se han detectado una serie de fallas en la determinación de energías libres con dichas metodologías. En cuanto a los métodos de HTD numerosos estudios de investigación realizados en los últimos años han demostrado la necesidad de nuevas funciones de puntuación que describan con mayor precisión la interacción de proteína-ligando, la influencia del solvente en dicha interacción, y el cambio entrópico al pasar de la conformación libre en solución a la conformación unida.

La mayor parte de los desarrollos metodológicos antes mencionados emplean aproximaciones basadas en la mecánica clásica. Estas metodologías, aún si han presentado resultados exitosos,[15,16] no tienen en cuenta efectos importantes para la correcta descripción de las interacciones involucradas, como por ejemplo la polarización electrónica y transferencia de carga. Por otro lado, presentan una transferabilidad reducida de sus parámetros por lo que dependen de la disponibilidad de un conjunto de entrenamiento suficientemente grande.

En los últimos 10 años, el desarrollo de métodos basados en QM y su aplicación al estudio de sistemas biomoleculares ha registrado un notable crecimiento, permitiendo una mejora en la descripción de las interacciones proteína-ligando y en la determinación de afinidades de unión. La principal ventaja de éste tipo de métodos es que son de validez general, sistemáticamente mejorables, y pueden ser usados en moléculas no estándar sin necesidad de una parametrización. Es importante remarcar que la formulación QM incluye todas las contribuciones de energía de manera directa, contrariamente a los métodos de mecánica clásica, lo que permite describir efectos de polarización electrónica, transferencia de carga, entre otros. El trabajo de Merz *et al.*[15,17] fue pionero en incorporar cálculos basados en la formulación QM en la descripción de interacciones proteína-ligando. Una revisión detallada de los desarrollos de métodos QM para el cálculo de energías de unión ligando-proteína, se encuentra en la Ref. 18; una revisión realizada recientemente por Yilmazer y Korth, que presenta los desarrollos de métodos SQM con inclusión de correcciones de dispersión y enlaces de hidrógeno se puede encontrar en la Ref. 19. A pesar de los últimos avances encontrados en este campo, el uso de aproximaciones basadas en QM en el diseño racional de fármaco sigue siendo un desafío actual.[19–21]

Dado que la mayoría de los procesos bioquímicos ocurre en un entorno acuoso, para mejorar el modelado de sistemas biomoleculares en solución resulta imprescindible contar, en primer lugar, con un método robusto y computacionalmente eficiente para describir los efectos de la solvatación sobre las propiedades moleculares. Los modelos de solvente continuo son de gran utilidad para ser usados en combinación con métodos semi-empíricos, ya que los cálculos cuánticos con modelos de solvente explícito resultarían extremadamente costosos. Sin embargo, es necesario contar con parámetros atómicos optimizados para alcanzar una mayor precisión en los mismos. En la presente Tesis se escogió el modelo de solvente continuo COSMO para mejorar su exactitud, optimizando sus parámetros atómicos para reproducir valores experimentales de energías libres de hidratación de moléculas pequeñas neutras. Los cálculos se realizaron en el programa MOPAC2012,[22] con el algoritmo de escalamiento lineal MOZYME[23], que permite trabajar de manera eficiente con métodos SQM en sistemas de miles de átomos.

En una segunda parte, se implementó el método MM/QM-COSMO para calcular energías libres de unión, en un estudio de investigación realizado en colaboración con grupos experimentales orientado al descubrimiento de nuevos inhibidores de la entrada del virus del Dengue a la célula (ver Ref. 24). Se identificaron compuestos con actividad biológica confirmada mediante un procedimiento de cribado virtual basado en la estructura del receptor. Dos conformaciones resultantes de *docking* fueron encontradas con energías similares. El método MM/QM-COSMO fue usado entonces para calcular diferencias de energía libre de unión entre ambas poses del ligando, para determinar la más favorable y guiar así el proceso de optimización del candidato líder.

Una contribución importante de esta Tesis es el desarrollo de una nueva función de *scoring* cuántica. Se incluye en esta Tesis tanto el estudio teórico subyacente como la validación apropiada de dicha metodología. El *score* formulado aproxima la energía libre de unión de proteína-ligando mediante una suma de términos con significado físico. La correlación encontrada entre valores experimentales de constantes de inhibición y cálculos de energías libre para un sistema de estudio considerado permitieron sugerir formas funcionales para el *score*. En dicho estudio de investigación, se tomó el receptor *Cyclin Dependent Kinase 2* (CDK2) en complejo con 15 inhibidores, con estructuras cristalográficas disponibles, para calcular valores de energía libre de unión con un método cuántico. Los métodos semi-empíricos usados en la metodología propuesta en esta Tesis para calcular energía libre de unión, describen con suficiente precisión las interacciones de dispersión y polarización. Por otro lado fueron incluidos los efectos de energía de deformación, energía de solvatación y contribución entrópica. Luego, se midió la correlación de los cálculos cuánticos con los valores experimentales de constantes de inhibición conocidos para los complejos. Los resultados de correlación obtenidos permitieron definir la forma funcional del *score* cuántico.

Posteriormente, se definieron distintos protocolos para la aplicación de la nueva función de *scoring* cuántica, desarrollada en esta Tesis. Los mismos, fueron implementados en un estudio retrospectivo realizado sobre cinco receptores con estructura cristalográfica y una librería química de ligandos conocidos y *decoys*.[2] La calidad de dicha función fue evaluada en cuanto a la capacidad de identificar ligandos en dicha librería de moléculas. El factor de enriquecimiento fue la métrica usada para determinar la precisión de la metodología. Los resultados encontrados indican que la función presentada en este estudio de investigación representa una alternativa precisa a los métodos clásicos, para generar una sub-librería enriquecida en potenciales ligandos en un contexto de HTD.

[2]Un *decoy* es una molécula que, se asume, no se une al receptor, por lo que se la considera un no-ligando de la base de datos. Posee propiedades fisicoquímicas similares a la de un ligando pero es estructuralmente diferente.

Parte II

Fundamentos Teóricos

Capítulo 1

Modelado Molecular

La descripción y el estudio de fenómenos relacionados con los sistemas moleculares se puede realizar por medio de formulaciones matemáticas y herramientas computacionales orientadas a facilitar los cálculos y efectuar predicciones. Para ello, es necesario definir en un modelo que permita describir de manera idealizada, las características y propiedades de un sistema o proceso. El modelado molecular, consiste precisamente en la descripción del comportamiento de sistemas moleculares siguiendo diferentes aproximaciones.

La aplicación de metodologías computacionales permite establecer una conexión entre los estudios experimentales y los modelos teóricos, generando así un complemento importante para el estudio y descubrimiento de nuevos fenómenos de interés.

En el presente Capítulo, se describen brevemente los niveles de teoría comúnmente empleados en los modelos para la descripción de sistemas biomoleculares. En primer lugar, se presentan los fundamentos más generales de la formulación de Mecánica Cuántica (*Quantum Mechanics*, QM), y la Mecánica Molecular fundada en las leyes de la Mecánica Clásica (*Molecular Mechanics*, MM). Seguidamente, se mencionan las maneras en que se puede representar el sistema en estudio, de acuerdo al nivel de teoría empleado. En una segunda parte del Capítulo, se desarrolla el concepto de optimización de geometría, dada la importancia de la obtención de propiedades moleculares a partir de estructuras de equilibrio, presentando también los métodos de minimización más empleados en el área de Química Computacional.

1.1. Mecánica Cuántica

La Mecánica Cuántica permite obtener valores precisos de las propiedades de un sistema molecular a partir del conocimiento de la distribución electrónica. Ésta se puede

hallar resolviendo la ecuación de Schröedinger independiente del tiempo,

$$H\Psi(\mathbf{r}) = E\Psi(\mathbf{r}) \tag{1.1}$$

donde H es el operador Hamiltoniano, Ψ es la función de onda, $\mathbf{r}$ representa el vector coordenada de los átomos, y E es la energía total del sistema.

Expresando el Hamiltoniano en función de las contribuciones de energía cinética y potencial, se tiene:

$$H = -\frac{\hbar^2}{2m}\nabla^2 + V \tag{1.2}$$

donde $\hbar$ es la constante de Planck, ∇ es el operador Laplaciano en coordenadas cartesianas, y V es un campo externo independiente del tiempo.

La Ec. 1.1 pertenece al tipo de ecuaciones conocidas como ecuaciones de autovalores. Para resolver la misma, se debe hallar el valor de E y las autofunciones Ψ correspondientes. Esto se puede realizar de manera exacta únicamente para sistemas de un electrón. Para sistemas de dos o más electrones, se debe recurrir a soluciones aproximadas. Como consecuencia fundamental, se tiene que la función de onda puede adoptar más de una forma funcional siendo la forma más general, una serie infinita de funciones. La función de onda total debe cumplir además con la condición de ortonormalidad, de modo que

$$\int \Psi_i \Psi_j d\mathbf{r} = \delta_{ij} \tag{1.3}$$

donde δ_{ij} es la delta de Kronecker.

Resolviendo la Ec. 1.1 para la función de onda Ψ_j, multiplicando a la izquierda por Ψ_i e integrando para todo el espacio, se tiene

$$\int \Psi_i H \Psi_j d\mathbf{r} = \int \Psi_i E \Psi_j d\mathbf{r} = E_j \delta_{ij} \tag{1.4}$$

donde se ha empleado la condición de ortonormalidad (Ec. 1.3).

Para hallar los valores de energía E_j, es necesario determinar la forma de la función de onda Ψ_j.

Principio Variacional Dado que no es posible hallar una solución exacta a la ecuación de Schröedinger para un sistema de muchos cuerpos, se debe establecer un criterio para determinar cuándo una función de onda propuesta es mejor que otra. El *principio variacional* expresa que la energía calculada a partir de una función de onda aproximada, será siempre mayor a la verdadera energía del sistema. Por lo tanto, la mejor función de

onda será aquella para la cual la energía alcance el valor mínimo,

$$\frac{\langle\Psi|H|\Psi\rangle}{\langle\Psi|\Psi\rangle} \geq E_0 \tag{1.5}$$

Aproximación de Born-Oppenheimer

La aproximación de Born-Oppenheimer parte del fundamento de que los núcleos de los sistemas moleculares poseen una masa mucho mayor que la masa de los electrones. Se propone entonces una separación del movimiento nuclear y la energía potencial de interacción entre los núcleos de los términos electrónicos en el Hamiltoniano molecular. De esta forma, la función de onda molecular se puede escribir como el producto de una parte nuclear y una parte electrónica.

La energía y función de onda moleculares se obtienen resolviendo la ecuación de Schröedinger,

$$H\psi(q_e, q_N) = E\psi(q_e, q_N) \tag{1.6}$$

donde $\psi(q_e, q_N)$ es la función de onda molecular, que depende de las coordenadas electrónicas y nucleares (q_e y q_N, respectivamente).

Mediante la aproximación de Born-Oppenheimer podemos concentrarnos en el movimiento electrónico, considerando los núcleos fijos (dado que su velocidad es mucho menor a la de los electrones). Por lo tanto, la ecuación de Schröedinger se puede resolver para los electrones en el campo de los núcleos,[25]

$$(H_{el} + V_{NN})\psi_{el} = U\psi_{el} \tag{1.7}$$

donde ψ_{el} representa la función de onda electrónica, que depende de las variables electrónicas (q_e) y de forma paramétrica de las coordenadas nucleares (q_N). El término H_{el} de la Ec. 1.7 representa la energıa electrónica (que depende de las posiciones nucleares), y V_{NN} es la energía de repulsión entre núcleos. La energía U es la energía electrónica que incluye la repulsión inter-nuclear ($U = E_{el} + V_{NN}$).

La aproximación de Born-Oppenheimer permite definir una superficie de energía potencial, a partir de la variación de energía electrónica para diferentes coordenadas nucleares fijas. Por lo general, las propiedades electrónicas se calculan a partir de geometrías de equilibrio sobre dicha superficie.

Ecuaciones de Hartree-Fock

El método de Hartree-Fock (HF) se basa en un modelo de partícula independiente, en el que la función de onda se aproxima como el producto antisimetrizado de espín-orbitales mono electrónicos. En este modelo se consideran las interacciones de Coulomb entre electrones, mediante un término efectivo mono electrónico, en el que cada electrón experimenta un potencial de campo medio, originado por los demás núcleos y electrones.

Una vez que se define la forma de la función de onda, y de la energía en función de operadores de un electrón, como es el caso de la energía de HF, se debe determinar de qué manera encontrar los orbitales. Suponemos que se toma una combinación lineal arbitraria de orbitales (espaciales) de un electrón para cada orbital molecular:

$$\psi_i = \sum c_{\nu i}\phi_\nu \tag{1.8}$$

donde los orbitales de un electrón ϕ_ν son comúnmente llamados funciones de base, que por lo general corresponden a orbitales atómicos.

De acuerdo al principio variacional, el mejor conjunto de coeficientes será aquel para el cual la energía alcance el valor mínimo para todos los coeficientes,

$$\langle\psi_1\psi_2\psi_3...|H|\psi_1\psi_2\psi_3...\rangle \equiv E_{HF}[\psi_1\psi_2\psi_3...] \geq E_0 \tag{1.9}$$

obteniendo de esta forma los ψ_i que minimizan el funcional $E_{HF}[\psi_i]$.

Para hallar la energía del estado fundamental de acuerdo al método de HF, se deben resolver las ecuaciones de Roothaan-Hall, mediante un procedimiento autoconsistente (*self-consistent field*, SCF)[26]. Expresando dichas ecuaciones en forma matricial se tiene que,

$$\mathbf{FC} = \epsilon\mathbf{SC} \tag{1.10}$$

donde $\mathbf{F}$ es la matriz de Fock, $\mathbf{C}$ representa la matriz de coeficientes, $\mathbf{S}$ la matriz de solapamiento y $\boldsymbol{\epsilon}$ el vector de energıa de los espín-orbitales moleculares.

La matriz $\mathbf{F}$ contiene contribuciones monoelectrónicas ($\mathbf{h}$), y bi-electrónicas dadas por las matrices de Coulomb ($\mathbf{J}$) y de intercambio ($\mathbf{K}$),

$$\mathbf{F} = \mathbf{h} + \mathbf{J} - \mathbf{K} \tag{1.11}$$

A diferencia de los orbitales moleculares, las funciones de base no deben cumplir con la condición de ortornomalidad. Esto implica que dos funciones de base pueden estar localizadas en diferentes átomos (ϕ_μ y ϕ_ν) sin que se anule el solapamiento, $S_{\mu\nu}$=$\phi_\mu\phi_\nu$.

Las contribuciones del Hamiltoniano monoelectrónico (**h**), requieren el cálculo de integrales que involucran funciones de base de hasta dos centros (dependiendo de si las mismas estén centradas en el mismo átomo o no). Contrariamente, las integrales bi-electrónicas incluídas en las matrices de Coulomb y de intercambio (**J** y **K**) pueden incluir hasta cuatro funciones de base localizadas en cuatro centros diferentes.

La energía de HF del estado fundamental estará dada por

$$E_{HF} = \langle \psi | \mathbf{F} | \psi \rangle \tag{1.12}$$

donde el estado fundamental molecular ψ contiene los orbitales mono electrónicos moleculares optimizados, asociados a las energías individuales, ϵ_i, que se obtienen como solución de la Ec. 1.10.

La teoría de Hartree-Fock puede ser aplicada para efectuar cálculos de mecánica cuántica en sistemas moleculares. Los cálculos de orbitales moleculares de mecánica cuántica se pueden dividir en dos grandes categorías: los métodos *ab initio* y los métodos *semi-empíricos* (*Semi-empirical Quantum Mechanical Methods*, SQM). Los métodos *ab initio* se refieren a cálculos que resuelven la ecuación de Hartree-Fock sin despreciar ni aproximar las integrales, y consideran todos los términos del Hamiltoniano. Contrariamente, los métodos *semi-empíricos* se caracterizan por ignorar o incluir parámetros para calcular algunas de dichas integrales, simplificando considerablemente los cálculos. A continuación se presentan brevemente las principales características de estos últimos.

1.2. Métodos Semi-empíricos

Un cálculo Hartree-Fock escala a la cuarta potencia con el número de átomos del sistema, es decir con el número de funciones de base. La parte más demandante del cálculo SCF HF con métodos *ab initio* está relacionada precisamente con el manejo de las integrales. La forma más directa de reducir el esfuerzo computacional consiste en despreciar o aproximar de alguna manera dichas integrales. Los métodos SQM fueron desarrollados con este objetivo, reduciendo considerablemente el tiempo de cómputo en primer lugar al incorporar de manera explícita únicamente los electrones de valencia del sistema. Estos métodos emplean el menor conjunto de funciones de base necesario para acomodar los electrones de valencia en un átomo neutro. Por lo general, se usan orbitales de Slater del tipo s, p y en ocasiones d.

Una característica común a la mayoría de los métodos SQM es que la matriz de solapamiento **S** (en la Ec. 1.10) es igual a la matriz identidad **I**. Los elementos que corresponden al solapamiento entre orbitales atómicos sobre diferentes átomos son cero. Esto

implica que **FC=SCE** se transforma en **FC=CE**, tomando automáticamente la forma matricial estándar.

Muchos de los métodos semi-empíricos se basan en la aproximación conocida como Solapamiento Diferencial Nulo (*Zero Differential Overlap*, ZDO), en la que los productos de funciones de base átomicas diferentes se anulan,

$$\phi_\mu \phi_\nu = 0 \tag{1.13}$$

donde ϕ_μ corresponde al orbital atómico sobre el centro μ y ϕ_ν al orbital atómico sobre el centro ν. De esta manera la matriz **S** se convierte en la matriz identidad. Las integrales bi-electrónicas de tres y cuatro centros son nulas bajo esta aproximación. Las demás integrales se ajustan a valores experimentales por medio de parámetros.

El primer método que implementó la aproximación ZDO fue desarrollado en el año 1965 por Pople *et al.*[27]. Esta nueva aproximación, denominada *Complete Neglect of Differential Overlap* (CNDO), asigna un parámetro γ a las integrales bi-electrónicas centradas en distintos átomos, y depende de la naturaleza de dichos átomos y de su distancia internuclear y no del tipo de orbital. De la misma manera, existen distintos métodos que parten de la base de las aproximaciones ZDO, que se caracterizan por el número de integrales no consideradas y el tipo de parametrización realizada. Entre ellos se pueden mencionar el método *Neglect of Diatomic Differential Overlap* (NDDO), en el que se toma en cuenta una reducción de la carga nuclear debido a los electrones internos. Si además se desprecian las integrales bi-electrónicas que no son de tipo Coulomb, se llega al método *Intermediate Neglect of Differential Overlap* (INDO), mientras que el método *Complete Neglect of Differential Overlap* (CNDO) desprecia todas las integrales bielectrónicas.

En el presente trabajo de Tesis, se utilizaron los métodos semiempíricos RM1[28], PM6-D3H4[29–31] y PM7[32], que parten de las mismas aproximaciones efectuadas por el método *Modified Neglect of Diatomic Overlap* (MNDO), desarrollado por Dewar y Thiel[33], basado en NDDO. Todas las integrales son calculadas de manera aproximada utilizando parámetros ajustados con valores experimentales. Se toma en cuenta la repulsión electrostática y la estabilización de intercambio.

De esta manera, resulta manejable el estudio de sistemas moleculares de mayores dimensiones a los tratados por métodos *ab initio*, que resuelven de manera rigurosa la ecuación de Schröedinger. Aún si la precisión alcanzada por los métodos semi-empíricos es menor, permiten investigar sistemas biomoleculares de manera eficiente.

1.3. Mecánica Molecular

Muchos de los problemas relacionados al modelado molecular y más específicamente cuando uno trabaja con biomoléculas comprenden un sistema de un número demasiado elevado de partículas para ser estudiado mediante métodos de QM. Ya que dichos métodos requieren la consideración de los electrones del sistema, aún si no todos se incluyen explícitamente (como en el caso de los métodos semi-empíricos), el número de partículas a considerar sigue siendo muy grande y el tiempo requerido para los cálculos es alto.

Los métodos de Mecánica Molecular modelan el sistema molecular ignorando el movimiento electrónico. Solamente los núcleos son considerados de manera explícita, y se los representa por medio de esferas conectadas por resortes (que caracterizan los enlaces). El sistema se describe por medio de campos de fuerza (*Force Field*, FF) que incluyen los distintos tipos de interacción por medio de parámetros, y los movimientos de los átomos se rigen por las leyes de la Mecánica Clásica. Cabe mencionar que los métodos de MM no pueden proporcionar información acerca de propiedades que dependan de la distribución electrónica.

1.3.1. *Campo de Fuerzas*

La forma más general de la energía potencial definida por un FF tiene en cuenta las fuerzas intra- e inter moleculares de un sistema a través de un modelo matemático sencillo, presentado esquemáticamente en la Fig. 1.1. Se puede observar que la energía potencial $U(\mathbf{r})$ está dada por: un término que modela la interacción entre pares de átomos enlazados mediante un potencial armónico (enlaces), un segundo término definido como una suma sobre ángulos de la molécula, modelado también por un potencial armónico (ángulos de enlace), un potencial torsional que da cuenta del cambio de energía debido a la rotación de enlaces (torsiones angulares), y finalmente dos contribuciones correspondientes a los términos no-enlazantes, que son calculadas para pares de átomos separados por al menos tres enlaces. Éstos últimos términos son usualmente determinados usando un potencial electrostático de Coulomb para las interacciones electrostáticas y un potencial de Lennard-Jones para las interacciones de van der Waals. Existen otros FF más robustos que poseen términos adicionales, pero de todas maneras contienen los términos que aparecen en la Fig. 1.1

Los distintos FF son desarrollados generalmente para su aplicación en simulaciones de dinámica molecular, minimización de energía, o modelado de péptidos y proteínas. Cada FF se distingue por su forma funcional y conjunto de parámetros correspondientes. Una característica importante que representa una ventaja de este tipo de metodologías es que

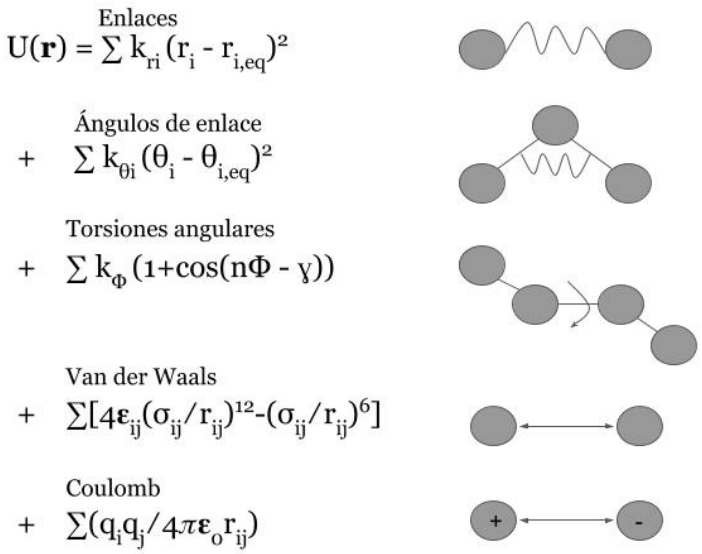


Figura 1.1: Energía potencial y representación gráfica de los términos comúnmente empleados en un *force field*.

el mismo grupo de parámetros puede ser empleado para describir moléculas relacionadas, sin necesidad de tener que definir nuevos parámetros para cada molécula individual. Por último, es importante mencionar el concepto de "tipo de átomo" (del inglés *atom type*), común a la mayoría de los FF. Antes de realizar un cálculo, se asigna un *atom type* a cada átomo del sistema, el cual contiene información acerca de la carga, estado de hibridación y entorno local del mismo. Los parámetros de un FF se expresan en función de estos *atom type*, que pueden ser obtenidos por medio de ajustes con valores experimentales o ser derivados por medio de cálculos cuánticos.

1.4. Representación del sistema

En el modelado molecular se pueden emplear distintos niveles de teoría para representar el sistema y determinar sus propiedades.

Mecánica Cuántica Para describir correctamente las reacciones químicas y otras propiedades electrónicas, como por ejemplo la transferencia de carga, la polarización electrónica, la ruptura y formación de nuevos enlaces, es necesario recurrir a métodos de QM. Sin embargo, éstos se restringen al estudio de sistemas de unos pocos cientos de átomos debido a la dificultad y elevado costo computacional requeridos por los mismos.

Las aproximaciones específicas empleadas para resolver la ecuación de Schröedinger

(Ec. 1.1) definen a su vez distintos tipos de métodos: los *ab initio*, *semi-empíricos* (SQM), y de *funcionales de la densidad* (DFT). En el primer caso, el modelo teórico se define por la combinación del método empleado para aproximar el Hamiltoniano (HF, MMP2, etc) y el conjunto de funciones de base empleado para aproximar la función de onda (STO, 6-31G(d), etc). Los métodos SQM recurren a un conjunto de funciones de base por defecto (conjunto minimal) y el Hamiltoniano define el modelo (CNDO, AM1, RM1, PM7). Los métodos de DFT se especifican por los diferentes funcionales de densidad (BLYP, PBE, entre otros).

Mecánica Molecular Aún si la descripción dada por la formulación de QM es precisa y de validez general, la dimensión y complejidad conformacional de los sistemas biomoleculares exige un tratamiento capaz de considerar simultáneamente miles de átomos y permitir simulaciones de escalas temporales de centenas a millares de nanosegundos. Esto se puede alcanzar empleando métodos de Mecánica Molecular (MM) basados en campos de fuerza (FF). La principal característica de dichos modelos es que no tienen en cuenta de manera explícita los electrones del sistema, cuyos efectos se incluyen de manera implícita en el FF. Debido a esta consideración y a la parametrización de los FF, poseen una mayor eficiencia computacional. Sin embargo, una desventaja de los mismos es que usualmente los FF están limitados a una clase de moléculas empleadas en la parametrización. Por otro lado, no es posible calcular propiedades que dependan de la estructura electrónica, como por ejemplo formación y ruptura de enlaces químicos.

QM/MM Los métodos combinados de Mecánica Cuántica/Mecánica Molecular (QM/MM por sus siglas en inglés) surgen en el año 1976, en el que Warshel and Levitt introducen este concepto[34]. Para modelar macromoléculas biológicas de gran tamaño, se deduce que una aproximación adecuada viene dada por la combinación de métodos de QM y MM de manera tal que, para la región donde se producen las modificaciones que tienen que ver con propiedades electrónicas (por ejemplo un ligando y el sitio de unión del receptor) se recurre a métodos QM y para los alrededores (por ejemplo, proteína y solvente) un método MM. Los métodos resultantes se denominan comúnmente métodos híbridos o combinados QM/MM. Se han realizado numerosos estudios con esta metodología en la última década. Las referencias [18, 35], hacen una revisión completa de los avances desarrollados y sus aplicaciones.

En la Fig. 1.2 se puede apreciar un esquema representativo de la división del sistema en sus partes QM y MM. El sistema completo (S) se divide en una región interna (I) que es tratada mecano cuánticamente, y una parte externa (E) que se describe mediante campos de fuerza de mecánica molecular. Estas regiones se denominan regiones QM y MM, respectivamente. Debido a la fuerte interacción entre las regiones QM-MM, la energía total

no es la suma de las energías de los subsistemas. Es necesario tener en cuenta los términos de acoplamiento, teniendo especial cuidado en el contorno entre ambos subsistemas sobre todo si éste corta enlaces covalentes. Se define como región de contorno a la región en la que los métodos QM y MM son modificados o aumentados de alguna manera.

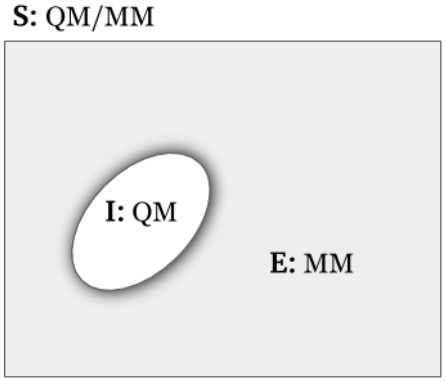


Figura 1.2: División del sistema (S) en una parte interna (I) y una parte externa (E). El anillo negro que rodea al sistema interno representa la región de contorno.

1.5. Optimización de geometría

Las moléculas poseen vibraciones térmicas alrededor de estructuras de equilibrio, por lo tanto para poder realizar comparaciones con valores experimentales, es importante poder determinar las estructuras de mínima energía. Para ello, resulta de gran interés contar con métodos computacionales precisos de optimización de geometría. A continuación se describe brevemente el proceso para llevar a cabo una optimización de geometría, estableciendo la relación con la forma funcional de energía del sistema. El objetivo de esta Sección no es presentar una descripción completa de los algoritmos, por lo que se detallan únicamente los fundamentos básicos de las metodologías comúnmente empleadas.

Superficie de Energía Potencial

La energía potencial es una función multidimensional de las coordenadas de un sistema. La manera en que varía dicha energía con las coordenadas genera lo que se conoce como superficie de energía potencial, o hipersuperficie. Para un sistema de N átomos la energía es una función de $3N$ coordenadas cartesianas, o $3N$-6 coordenadas internas. Por esta razón, es imposible visualizar la superficie de energía completa excepto para aquellos casos más simples, en los que el sistema dependa de una o dos coordenadas.

Lo que resulta de especial interés en el modelado molecular es el conocimiento de los puntos de mínima energía sobre dicha hipersuperficie, dado que corresponden a conformaciones estables del sistema; cualquier movimiento con respecto a un mínimo dará una configuración con una energía más alta. Se pueden encontrar numerosos mínimos en la superficie de energía (*mínimos locales*). El mínimo de energía más baja de toda la hipersuperficie se conoce como *mínimo de energía global*. Los algoritmos de minimización permiten identificar las geometrías del sistema (estructuras de equilibrio) que corresponden a un punto estacionario sobre la hipersuperficie de energía potencial. El punto más alto localizado entre dos mínimos se conoce como *punto de ensilladura*, en los que el ordenamiento de los átomos corresponde a una *estructura de transición*. Tanto los mínimos como los puntos de ensilladura son puntos estacionarios sobre la superficie de energía.

1.5.1. Minimización de energía

El término minimización se refiere a la identificación de los puntos estacionarios de la función de energía potencial. Esto es, encontrar los valores de las variables de una función f, para los cuales dicha función alcanza el valor mínimo. La derivada primera de la energía con respecto a cada una de las variables es cero en los puntos estacionarios. En adelante nos concentraremos únicamente en la localización de *mínimos* (que corresponden a derivadas segundas positivas de energía) para la función de energía potencial. Dicha función puede ser del tipo FF, o bien de la resolución de la ecuación de Schröedinger (QM, SQM), siendo las variables las coordenadas cartesianas o internas de los átomos. Los mínimos se localizan con métodos numéricos que cambian gradualmente las coordenadas para producir configuraciones con energías cada vez más bajas, hasta encontrar la de menor energía.

El punto de partida para un programa de minimización consiste en un conjunto de coordenadas iniciales del sistema, que pueden ser obtenidas por técnicas experimentales (como rayos X, cristalografía, NMR) o por técnicas computacionales. En otros casos, se pueden utilizar también métodos teóricos como algoritmos de búsqueda conformacional, para generar un conjunto de estructuras iniciales, a partir de las cuales se realiza una optimización de energía para encontrar el mínimo sobre la hipersuperficie de energía potencial.

Los algoritmos de minimización se pueden clasificar en dos grupos: los que usan derivada de energía con respecto a las coordenadas y los que no. Los primeros, denominados *métodos derivativos*, proporcionan información acerca de la forma de la superficie de energía y pueden mejorar la eficiencia con la que se localiza el mínimo. No es posible identificar un único método de minimización como el mejor método para todos los proble-

mas encontrados en el modelado molecular. El algoritmo ideal es aquel que proporciona el punto de mínima energía lo más rápido posible, con el menor requerimiento de memoria.

1.5.2. Métodos derivativos

Los métodos de minimización pueden ser derivativos de primer orden o de segundo orden, según se calcule la primera derivada de la función de energía o la primera y segunda derivadas. Una optimización de geometría o minimización derivativa de primer orden toma como punto de partida una estructura molecular determinada y realiza un paso sobre la superficie. Luego calcula la energía y la derivada en dicho punto (gradiente). La dirección del gradiente indica la máxima reducción de energía sobre la hipersuperficie, mientras que su magnitud indica la inclinación de la pendiente. De esta manera determina en qué dirección y de qué magnitud será el paso siguiente. Las derivadas segundas indican la curvatura de la función, indicando de esta manera dónde cambiará de dirección la función.

Los algoritmos de minimización de primer orden usados con mayor frecuencia en modelado molecular son los métodos *steepest descent* (SD) y *conjugate gradient* (CG). A través de la modificación de los átomos, acercan el sistema al punto de mínimo. La configuración de partida para cada iteración es la obtenida en el paso previo. Para la primera iteración se usa como punto de partida la configuración inicial del sistema proporcionada por el usuario.

El método SD realiza los movimientos en dirección paralela a la fuerza neta, que corresponde a dar un paso hacia abajo en energía. Para $3N$ coordenadas cartesianas, esta dirección se puede representar de manera mas conveniente como un vector unitario de $3N$ dimensiones, $\mathbf{d_k}$, por lo tanto[26]:

$$d_k = -\frac{\mathbf{g_k}}{|g_k|} \tag{1.14}$$

donde $\mathbf{d_k}$ es el gradiente. De esta manera se define la dirección a lo largo de la cual moverse, y luego es necesario decidir la magnitud del paso a lo largo del gradiente. Esto se puede realizar mediante una búsqueda lineal o definiendo un paso de dimensión arbitraria para efectuar la búsqueda.

En el método CG, la dirección de movimiento en un punto se calcula a partir del gradiente en dicho punto y del vector de dirección del paso previo,

$$d_k = -\mathbf{g_k} + \gamma_k \mathbf{v}_{k-1} \tag{1.15}$$

donde γ es una constante. El primer paso se calcula al igual que en el método SD, es decir en la dirección del gradiente.

Los métodos derivativos de segundo orden usan además del gradiente, la derivada segunda de la función. Dentro de éstos, el método de *Newton-Raphson* es el más simple. En este caso se requiere el cálculo de la matriz de las segundas derivadas de la función con respecto a las coordenadas, o matriz Hessiana y su inversa. Por lo tanto un requisito de dicho método es contar con una matriz Hessiana definida positiva. Existen numerosas variantes de este tipo de métodos que evitan el cálculo de la matriz completa de derivadas segundas. Otras variaciones utilizan la misma matriz para pasos sucesivos del algoritmo en los que únicamente el gradiente es recalculado en cada iteración. Dentro de estos últimos se encuentra el método de *Low memory-Broyden-Fletcher-Goldfarb-Shanno* (L-BFGS), algoritmo usado por defecto en el programa de química computacional MOPAC[22] para sistemas de más de 100 variables. El algoritmo utilizado para optimización de geometrías con menos de 100 variables es el denominado *Eigenvector Following* (EF).[36]

1.5.3. Criterios de convergencia

Una optimización finaliza con la localización del mínimo de energía. En problemas reales de modelado molecular, es difícil determinar con exactitud la localización de dicho mínimo o punto de ensilladura. Lo que se espera hallar es una aproximación al verdadero mínimo. Por lo tanto, resulta importante tomar algunos criterios para decidir cuándo el cálculo de minimización se acerca suficientemente al mínimo y se puede dar por terminado. Estos se denominan criterios de convergencia. En primer lugar, se puede calcular la energía de una iteración a otra y finalizar la minimización cuando la diferencia de energía entre pasos sucesivos cae por debajo de un límite especificado. Otra alternativa es monitorizar el cambio de coordenadas y terminar el cálculo cuando la diferencia entre configuraciones sucesivas es suficientemente pequeña. También resulta útil monitorizar el valor máximo del gradiente para asegurar que la minimización ha relajado adecuadamente todos los grados de libertad.

El uso de más de un criterio de convergencia previene una identificación errónea del mínimo. Por ejemplo, en un valle casi plano sobre la superficie de energía potencial, las fuerzas pueden ser cercanas a cero mientras que los pasos calculados son bastante grandes cuando la optimización se mueve hacia la parte más baja del valle. O, en regiones extremadamente empinadas, el tamaño del paso puede hacerse muy pequeño mientras que las fuerzas toma valores muy grandes.

1.6. Simulaciones computacionales

A partir de una minimización de energía efectuada sobre un sistema se obtienen geometrías de equilibrio. En algunos casos, la información extraída de una minimización es suficiente para calcular con precisión las propiedades del sistema. Si se pudieran encontrar todas las configuraciones del sistema sobre la hipersuperficie de energía potencial sería posible determinar todas las propiedades termodinámicas del sistema a partir de la función de partición de la termodinámica estadística. Esto es factible únicamente para moléculas con un número reducido de átomos, o para un grupo pequeño de moléculas en fase gaseosa. Sin embargo, en el modelado molecular muchas veces se desea comprender y predecir propiedades de sistemas moleculares de cientos a miles de átomos en fase solución, como por ejemplo sistemas biomoleculares. En estos casos, las mediciones experimentales se realizan sobre muestras macromoleculares que poseen un número elevado de átomos o moléculas, y una gran cantidad de mínimos de energía cercanos entre sí. De esta manera, resulta inviable una exploración completa de la superficie de energía.

Los métodos de simulación computacional son herramientas que permiten estudiar y predecir propiedades de sistemas macroscópicos, generando pequeñas réplicas de los mismos con un número manejable de átomos o moléculas. Una simulación genera configuraciones representativas de estas réplicas permitiendo el cálculo de propiedades estructurales y termodinámicas del sistema de manera computacional. Por otra parte, permiten también estudiar la evolución temporal del sistema proporcionando información detallada acerca de los cambios conformacionales del sistema.

En este Capítulo se presentan brevemente los fundamentos de dos de los métodos de simulaciones computacionales más usados en el modelado molecular, empleados en este trabajo de Tesis: el método de Dinámica Molecular (*Molecular Dynamics*, MD) y el método de Monte Carlo (MC). La elección del método depende del objetivo que se tenga y la capacidad de recursos computacionales de los que se dispone. Las simulaciones de MD se describen en primer lugar, continuando más adelante con MC. Se mencionan brevemente las características principales de ambas metodologías.

1.6.1. Simulación Temporal: Dinámica Molecular

Las simulaciones de dinámica molecular (MD) permiten estudiar la evolución temporal del sistema. El conjunto de posiciones atómicas y velocidades se deriva de manera secuencial aplicando las ecuaciones de movimiento de Newton. Una simulación, primero determina la fuerza sobre cada átomo ($\mathbf{F}_i$), igual al gradiente negativo de la energía

potencial

$$\mathbf{F} = -\frac{\partial V}{\partial \mathbf{r}_i} \tag{1.16}$$

siendo V la energía potencial y $\mathbf{r}_i$ la posición del átomo i.

Seguidamente, las posiciones y velocidades atómicas se pueden calcular integrando las ecuaciones diferenciales de la segunda ley de Newton,

$$\frac{d^2 x_i}{dt^2} = \frac{F_{x_i}}{m_i} \tag{1.17}$$

donde m_i es la masa de la partícula, x_i es la coordenada y F_{xi} es la fuerza aplicada sobre la partícula en esa dirección.

Las fuerzas que actúan sobre cada átomo son calculadas en cada paso de la simulación, y combinadas con las posiciones y velocidades actuales para generar las nuevas posiciones y velocidades en un tiempo posterior. Se asume que la fuerza que actúa sobre los átomos, es constante durante el intervalo de tiempo. Luego, los átomos se mueven a la nueva posición y se vuelve a calcular la fuerza, y así sucesivamente.

Una simulación de dinámica molecular genera de este modo una *trayectoria*, que describe cómo cambian las posiciones, velocidades y aceleraciones de una partícula con el tiempo, a partir de la cual se pueden determinar valores promedios de distintas propiedades del sistema. Las simulaciones de dinámica molecular usualmente tienen una duración en la escala de microsegundos.

Métodos de integración

Uno de los métodos más empleados en simulaciones de MD para integrar las ecuaciones de movimiento es el *algoritmo de Verlet.*[37] Éste algoritmo, emplea las posiciones y aceleraciones de las partículas en un tiempo t, y calcula las nuevas posiciones en un tiempo posterior $\mathbf{r}(t + \delta t)$, a partir de un paso anterior $\mathbf{r}(t - \delta t)$,

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)\delta t^2 + \cdots \tag{1.18}$$

$$\mathbf{r}(t - \delta t) = \mathbf{r}(t) - \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)\delta t^2 - \cdots \tag{1.19}$$

Sumando estas dos ecuaciones se obtiene,

$$\mathbf{r}(t + \delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \delta t) + \mathbf{a}(t)\delta t^2 \tag{1.20}$$

Como se puede notar en la Ec. 1.20, las velocidades no aparecen explícitamente en el algoritmo de Verlet. Una de las formas de calcularlas consiste en dividir por 2δt, la diferencia

entre las posiciones en los tiempos $t + \delta t$ (Ec. 1.18) y $t - \delta t$ (Ec. 1.19) para obtener,

$$\mathbf{v}(t) = \frac{[\mathbf{r}(t + \delta t) - \mathbf{r}(t - \delta t)]}{2\delta t} \tag{1.21}$$

Los pasos que se deben seguir para implementar este algoritmo son, en primer lugar el cálculo de la fuerza en el tiempo t, $\mathbf{F}(t)$, para la posición $\mathbf{r}$(t). Luego se emplea dicha posición, junto con la posición del paso previo $\mathbf{r}(t - \delta t)$, y la fuerza $\mathbf{F}(t)$, para calcular la nueva posición $\mathbf{r}(t - \delta t)$ de acuerdo a la Ec. 1.20. El valor de $\mathbf{r}(t - \delta t)$ se puede obtener a partir de un desarrollo en serie de Taylor,

$$\mathbf{r}(\mathbf{t} - \delta\mathbf{t}) = \mathbf{r}(t) - \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)\delta t^2 + \cdots \tag{1.22}$$

La Ec. 1.22 se puede truncar luego del primer término, de modo que $\mathbf{r}(-\delta t)$=$\mathbf{r}(0) - \mathbf{v}(0)\delta t$.

Intervalos temporales

Un punto importante en una simulación de MD es la elección del tamaño del paso temporal δt (*time step*) elegido para la evaluación de la energía y las fuerzas que actúan sobre el sistema. Cuanto menor sea el *time step* más parecida será la trayectoria al resultado de la integración analítica; sin embargo esto genera un incremento en el esfuerzo computacional y la trayectoria puede llegar a cubrir solamente una pequeña región del espacio de fases. Por otro lado, un *time step* corto requerirá un mayor tiempo de simulación computacional para una longitud de cálculo dada. Contrariamente, una elección de pasos temporales demasiado grandes puede conducir a inestabilidades en el algoritmo de integración, originadas por el colapso de alta energía entre átomos que se superponen al ocupar las mismas coordenadas atómicas. Se debe buscar entonces un balance adecuado entre cubrir adecuadamente el espacio de fases, y simular la trayectoria correcta.

Usualmente, se establece como límite superior para el *time step* aquel que permite incorporar adecuadamente los movimientos de más alta frecuencia, es decir los más rápidos del sistema, por ejemplo la vibración de un enlace C-H. Esta posee un período del orden de 10 fs. Para moléculas flexibles, se considera generalmente un paso temporal de aproximadamente un décimo de tiempo del período de movimiento más corto. Comúnmente se emplea un *time step* de 1 fs en las simulaciones de dinámica molecular. Dado que este tipo de movimientos de alta frecuencia no influye en el comportamiento general del sistema, se puede aplicar una restricción a fin de eliminar estos grados de libertad, dejando fijos los enlaces involucrados. Esto permite considerar un *time step* mayor, de 2 fs.

Una limitación del método de MD es que los tiempos de las trayectorias obtenidas a partir de simulaciones computacionales, son por lo general mucho menores a los tiempos

en los que ocurren los procesos químicos y físicos reales de interés (la mayoría de dichos fenómenos se dan a partir de tiempos superiores a los nanosegundos).

Preparación y producción de una simulación

Para efectuar una simulación de MD, es necesario determinar en primer lugar la configuración inicial del sistema. Esta puede ser obtenida a partir de datos experimentales o de modelos teóricos, por medio de una minimización de energía sobre una geometría determinada. Por lo general, las velocidades iniciales son seleccionadas de manera aleatoria a partir de una distribución de Maxwell-Boltzmann a la temperatura elegida para la simulación. También pueden ser elegidas a partir de otro tipo de distribuciones, como una Gaussiana o una distribución uniforme.

Una vez que se cuenta con la configuración inicial del sistema y fueron asignadas las velocidades a los átomos, se inicia la simulación. Se calcula la fuerza que actúa sobre cada átomo en cada paso, a partir de la derivada de la función de energía potencial. La fuerza ejercida sobre los átomos incluye distintos términos de acuerdo al campo de fuerzas usado.

Equilibración La primera etapa de una MD se conoce como fase de *equilibración del sistema*. El objetivo de la misma es alcanzar el equilibrio a partir de la configuración inicial, para asegurar que las propiedades son calculadas para un sistema en el equilibrio. El período requerido para esta etapa depende del sistema en estudio y de la propiedad que se desea calcular. Para determinar cuándo el sistema alcanza el equilibrio, se monitorean distintos parámetros, como por ejemplo la energía cinética, energía total, las velocidades, la temperatura y presión. Así, la energía potencial promedio para el tiempo simulado debe ser constante, para un sistema cercano al equilibrio.

Producción Una vez que el sistema ha alcanzado el equilibrio, comienza la etapa de *producción*, en la que se efectúa una exploración del espacio de fase. Durante esta etapa, se puede extraer información para su análisis posterior. Por lo general, las posiciones, energías y velocidades de las distintas conformaciones se guardan a intervalos regulares de tiempo, para ser usadas al finalizar la simulación para el cálculo de otras propiedades. Los tiempos de simulación dependen del sistema, y de la escala temporal de la propiedad en estudio.

1.6.2. Simulación Estocástica: Monte Carlo

En una simulación de MD existe una correlación temporal entre las distintas configuraciones alcanzadas por el sistema durante la trayectoria realizada. Contrariamente, en una simulación de Monte Carlo, cada configuración depende únicamente de su predecesora, independientemente de cualquier otra configuración que haya sido visitada por el sistema. El método de MC genera configuraciones aleatorias y usa un conjunto de criterios para decidir cuándo aceptar y cuando descartar una nueva configuración generada. Dichos criterios aseguran que la probabilidad de obtener una configuración determinada es igual al factor de Boltzmann $e^{-V(\mathbf{r}^N)/k_BT}$. $V(\mathbf{r}^N)$ se calcula por medio de la función de energía potencial. De esta forma, las configuraciones de baja energía tendrán mayor probabilidad de ocurrencia que las que posean energías mayores. Para cada configuración aceptada, se calculan las propiedades deseadas y al finalizar la simulación, el valor de las mismas se obtiene por un promedio sobre el número de valores calculados, M:

$$\langle A \rangle = \frac{1}{M} \sum_{i=1}^{N} A(\mathbf{r}^N) \tag{1.23}$$

donde A es una propiedad determinada, y $\langle\rangle$ indica un promedio sobre microestados.

En una simulación de MC, se genera una nueva configuración, alterando aleatoriamente una o más variables del sistema. La energía de la nueva configuración se calcula por medio de la energía potencial. Si la energía de la nueva configuración es menor que la energía de la configuración precedente, entonces la nueva configuración es aceptada. Si en cambio la energía es mayor, se acepta la nueva con $p = e^{-(V_n(\mathbf{r}^N)-V_a(\mathbf{r}^N))/k_BT}$. El subíndice n se utiliza para indicar la *nueva* configuración y a para identificar la *anterior*. Luego, se genera un número aleatorio entre 0 y 1, y se lo compara con el factor de Boltzman. Si el número aleatorio es mayor, se retiene la configuración inicial para la siguiente iteración, rechazando la última configuración; si el número aleatorio es menor, entonces el movimiento es aceptado y la nueva configuración se convierte en el próximo estado. Este procedimiento tiene el efecto de permitir movimientos a estados de mayor energía. Cuanto más bajo sea el valor de la diferencia entre energías potenciales, mayor será la probabilidad de que el movimiento sea aceptado.

Capítulo 2

Modelos de Solvente

La mayor parte de los procesos químicos ocurren en un entorno acuoso, motivo por el cual resulta de gran importancia considerar los efectos del solvente en el estudio de sistemas biológicos a nivel molecular. En particular, la energía libre de solvatación puede tener contribuciones significativas a la energía libre de unión de un complejo proteína-ligando. Estos conceptos serán desarrollados más adelante en esta Tesis.

Para considerar los efectos del solvente, éste puede ser representado usando diferentes modelos, entre los que se pueden mencionar el *modelo de solvente explícito* y *modelo de solvente continuo*. En el primer caso, se consideran moléculas de solvente individuales, por lo que este modelo es uno de los más exactos. Sin embargo, el costo computacional de tratar el solvente de forma explícita es demasiado elevado, debido al gran número de moléculas implicadas. Este tipo de modelos es utilizado generalmente en simulaciones de MD y MC, que emplean FFs para describir el sistema. En este caso, no se incluyen efectos de polarización electrónica mutua entre el soluto y el solvente, debido a que las cargas se consideran fijas.

En el modelo de solvente continuo se representa al solvente como un medio homogéneo polarizable, caracterizado por el valor de una constante dieléctrica ϵ, en el que se genera una cavidad donde se sitúa el soluto. Este tipo de modelo es incapaz de modelar efectos de solvatación de corto alcance, como por ejemplo enlaces de hidrógeno o una orientación preferencial de las moléculas de solvente cercanas al soluto. Los distintos métodos que emplean este tipo de modelo se diferencian entre sí por el tratamiento de la densidad de carga del soluto, el cálculo de la interacción del soluto con el dieléctrico que lo rodea y la construcción de la cavidad del soluto[38]. La representación del solvente de manera continua reduce significativamente el costo computacional asociado a los métodos de solvatación explícita.

En el presente Capítulo se desarrollan las características principales de los modelos de

solvente continuo, y en particular las pertenecientes al grupo de aproximaciones conocidas como *Carga Superficial Aparente* (ASC)[39]. Dentro de este grupo se encuentra el modelo de solvente continuo desarrollado por Klamt y Schüürman, denominado *COnductor-Like Screening MOdel* (COSMO)[40] que trata el solvente como un conductor, lo que simplifica considerablemente los cálculos. Para solventes de constante dieléctrica alta, el error que se introduce por esta aproximación es despreciable. El modelo COSMO puede ser usado en combinación con los métodos de QM. Su implementación eficiente con métodos semi-empíricos resulta especialmente útil para tratar sistemas de cientos a miles de átomos, con un adecuado balance entre eficiencia y precisión. Por este motivo, se empleó el modelo COSMO en el presente trabajo de Tesis, para el cálculo de energías de solvatación.

2.1. Energía libre de solvatación

El proceso de solvatación consiste, siguiendo la definición dada por Ben-Naim[41,42], en transferir una partícula de soluto M en una posición fija en fase gaseosa a una posición fija en solución S, a presión, temperatura y composición química constantes. La energía libre de Gibbs de solvatación (ΔG_{solv}) es el cambio de energía requerido para transferir la molécula de soluto M entre ambas fases (considerando una concentración estándar de 1 M para ambos entornos), teniendo en cuenta las interacciones soluto-solvente y el cambio interno producido en ellos durante el proceso de solvatación. En particular, si el solvente es agua, se la denomina *energía libre de hidratación* (ΔG_{hidr}). A efectos del cálculo, se puede expresar como la suma de las siguientes contribuciones[39]:

$$\Delta G_{hidr} = \Delta G_{elec} + \Delta G_{disp} + \Delta G_{cav} + \Delta G_{conc} \tag{2.1}$$

donde ΔG_{elec} es la energía electrostática total, que incluye la interacción del soluto M con el solvente. El último término de la ecuación 2.1 tiene en cuenta el cambio de concentración al pasar de fase gaseosa a fase solución. Debido a que en los valores experimentales se usa 1 mol/L tanto para la fase gaseosa como para la fase acuosa, se considera a lo largo de este trabajo $\Delta G_{conc} = 0$. La componente de dispersión, ΔG_{disp}, representa la contribución proveniente de los términos de interacción dispersión-repulsión entre el soluto y las moléculas del solvente. La energía libre de cavitación ΔG_{cav} es un término positivo que corresponde a la energía requerida para formar una cavidad adecuada para el soluto, dentro del solvente. Usualmente, estas contribuciones se incluyen mediante un término denominado energía libre de solvatación no polar, G_{np},

$$\Delta G_{np} = \Delta G_{disp} + \Delta G_{cav} \tag{2.2}$$

Los términos de la Ec. 2.1 no pueden ser determinados experimentalmente de manera separada, siendo ΔG_{hidr} la única cantidad mesurable.

Se mencionan a continuación las características más importantes de los métodos que emplean un modelo de solvente continuo. Seguidamente se detallan las diferentes formas de calcular las componentes de ΔG_{hidr}, empezando por la energía electrostática y siguiendo con los términos de dispersión y cavitación, a los que se denomina contribuciones no electrostáticas.

2.1.1. Modelo de Solvente Continuo

La metodología utilizada para modelar el solvente como un continuo se puede dividir en dos grandes grupos, según el nivel de teoría empleado para describir el sistema. Un primer grupo se caracteriza por la descripción cuántica del soluto, incluyendo la interacción con el medio a través de una descripción continua de este último (métodos continuos de mecánica cuántica). Otras aproximaciones describen el soluto como una distribución de cargas polarizable clásica, combinada con una descripción continua del medio que la rodea (métodos continuos de mecánica clásica). El factor común para ambos grupos es el denominado "campo de reacción", que es el campo eléctrico generado por el solvente polarizado, que a su vez influye en la interacción con el soluto. En adelante se referirá a éste como potencial de interacción.

Los modelos de solvente continuo se basan en la definición de un Hamiltoniano efectivo para el soluto M, $\hat{H}_M$, que de acuerdo a la aproximación de Born-Oppenheimer, depende de las coordenadas de los N_e electrones $\mathbf{q}\equiv\mathbf{q}_1,...,\mathbf{q}_{N_e}$ y paramétricamente de las coordenadas de los N_n núcleos, $\mathbf{Q}\equiv\mathbf{Q}_1,...,\mathbf{Q}_{N_n}$ y puede ser expresado como:[39]

$$\hat{H}_M(\mathbf{q};\mathbf{Q}) = \hat{H}^0_M(\mathbf{q};\mathbf{Q}) + \hat{V}^{int} \tag{2.3}$$

donde $\hat{H}^0_M$ es el Hamiltoniano electrónico en vacío y $\hat{V}^{int}$ es el campo de reacción o potencial de interacción, que será definido más adelante. Éste depende de las coordenadas electrónicas, de la geometría del soluto (a través de las coordenadas nucleares), de la distribución de cargas total ρ_M (modificada por los efectos del solvente) y del valor de la constante dieléctrica del solvente ϵ,

$$\hat{V}^{int} = \hat{V}^{int}(\mathbf{q};\mathbf{Q};\rho_{\mathbf{M}};\epsilon) \tag{2.4}$$

La formulación cuántica del modelo de solvente continuo requiere el tratamiento simultáneo de dos tipos de problemas: En primer lugar, el cálculo de la distribución electrónica ρ_M con núcleos fijos en presencia del campo electrostático generado por el

dieléctrico polarizado. Y en segunda instancia, el problema electrostático de determinar el potencial de interacción ($\hat{V}^{int}$) que a su vez depende también de la distribución de cargas ρ_M. Esto se reduce a un problema no lineal dado que el campo eléctrico y el potencial de interacción dependen a su vez de la distribución de cargas del soluto. La solución se puede determinar mediante un proceso iterativo autoconsistente, determinando ρ_M por métodos cuánticos y resolviendo el problema electrostático clásico para hallar el potencial. Por otra parte, los métodos clásicos utilizan dos tipos de aproximación para resolver el problema electrostático: consideración de cargas fijas (modelos rígidos), por lo tanto el efecto de polarización del soluto se ignora, o consideración de polarización del soluto a primer orden, mediante una función de distribución para ρ_M (modelos polarizables).

Cavidad La cavidad es un concepto involucrado en todos los modelos de solvente continuo. La creación de la misma en el solvente tiene un costo de energía asociado (ΔG_{cav} en Ec. 2.1). Teniendo en cuenta que las moléculas por lo general son de forma irregular, se pueden definir dos tipos de cavidad. Ambos se basan en la consideración de que existen porciones sobre la periferia de las moléculas que no pueden ser alcanzadas por el solvente. Se definen así la denominada *Superficie de Exclusión del Solvente* (SES) y la *Superficie Accesible al Solvente* (SAS). La SES, llamada también *Superficie de van der Waals*, tiene pequeños huecos en los que las moléculas de solvente no pueden ingresar. Una descripción más adecuada corresponde a la superficie marcada por una partícula esférica de solvente, de un radio determinado, al rodar sobre la superficie de van der Waals definida por las moléculas del soluto. Esta superficie es denominada *Superficie Accesible al Solvente* (SAS). En la práctica se utiliza en muchas ocasiones la SES en lugar de la SAS dado que la generación de esta última resulta más costosa computacionalmente. La forma de la cavidad es importante para obtener valores de energías de solvatación comparables a los valores experimentales. La diferencia conceptual entre ambas cavidades (SES y SAS) se puede apreciar con mayor claridad en la Fig. 2.1.

2.1.2. Contribución electrostática

Una vez que la forma y la dimensión de la cavidad han sido definidas se procede a resolver el problema electrostático clásico. El potencial total está dado por la suma del potencial generado por la distribución de cargas del soluto V^M, y el potencial de interacción $V^{int}(\mathbf{r})$,

$$V(\mathbf{r}) = V^M(\mathbf{r}) + V^{int}(\mathbf{r}) \tag{2.5}$$

Existen diferentes aproximaciones, utilizadas por los métodos clásicos y cuánticos, para describir el campo de reacción del solvente y su relación con ρ_M y resolver así el problema electrostático. Éstas se pueden clasificar como[39] (1) Carga Superficial Aparente (ASC), (2)

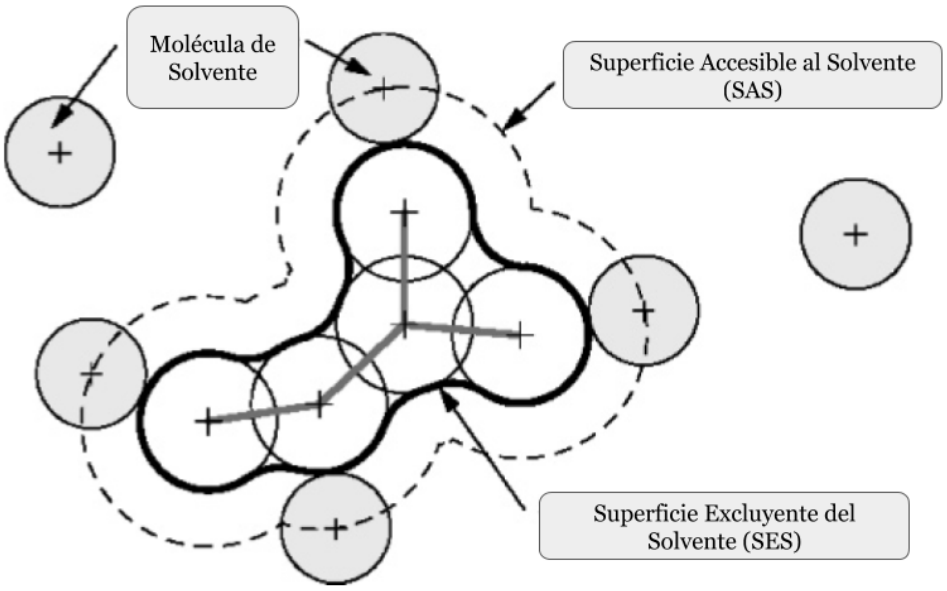


Figura 2.1: Superficie Accesible al Solvente (SAS) trazada por el centro de una partícula esférica que representa la molécula de solvente. La Superficie Excluyente del Solvente (SES) es el contorno marcado por la molécula de solvente de prueba al rodar por la superficie de van der Waals generada por las moléculas del soluto.

Expansión Multipolar (MPE), (3) Aproximación Generalizada de Born (GBA), (4) Carga Imagen (IMC), (5) Métodos de Elementos Finitos (FEM), y (6) Métodos de Diferencia Finita (FDM).

En el presente Capítulo nos concentraremos en las metodologías basadas en la aproximación ASC. En este caso el potencial de interacción $V^{int}(\mathbf{r})$ se puede describir en términos de una distribución de carga aparente $\sigma(\mathbf{r}_s)$, localizada sobre la superficie de la cavidad

$$V^{int}(\mathbf{r}) = \int \frac{\sigma(\mathbf{r}_s)}{|\mathbf{r} - \mathbf{r}_s|} d^2(\mathbf{r}_s) \tag{2.6}$$

donde usamos el símbolo $\mathbf{r}_s$ para el vector posición a fin de enfatizar que la distribución de carga se limita a la superficie de la cavidad.

El potencial de la Ec. 2.6 es precisamente el potencial de reacción. La densidad de carga σ está relacionada con la constante dieléctrica y el campo eléctrico perpendicular a la superficie, generado por la distribución de carga dentro de la cavidad (Φ_{in}),

$$\sigma = \frac{(\epsilon - 1)}{4\pi\epsilon} \nabla\Phi_{in} \cdot \mathbf{n} \tag{2.7}$$

Para cavidades esféricas o elipsoidales, la Ec. 2.7 puede ser resuelta de manera analítica. Para otro tipo de cavidades la misma se resuelve de manera numérica por medio de una discretización de la superficie en un número apropiado de elementos finitos. La superficie

se divide en una suma de elementos finitos de área ΔA_k que contienen una carga q_k en la posición $\mathbf{r}_{s_k}$:

$$q_k = \Delta A_k \sigma(\mathbf{r}_{s_k}) \tag{2.8}$$

de manera que la Ec. 2.6 se puede reescribir como

$$V^{int}(\mathbf{r}) = \sum_k \frac{q_k}{|\mathbf{r} - \mathbf{r}_{s_k}|} \tag{2.9}$$

El potencial de interacción $V^{int}(\mathbf{r})$ es calculado para la distribución de carga molecular (Ec. 2.7). Este potencial se suma al Hamiltoniano del soluto y se procede a resolver la ecuación de Schröedinger,[43]

$$(\hat{H}_M^0 + \frac{1}{2}\hat{V}^{int}) \left|\Psi\right\rangle = E \left|\Psi\right\rangle \tag{2.10}$$

donde el operador de interacción $\hat{V}^{int}$, sumado al Hamiltoniano de la molécula o soluto (H_M^0) acopla la respuesta electrostática del solvente con la distribución de cargas del soluto, y E es la energía electrónica del soluto más la contribución electrostática a la energía de solvatación. El factor $\frac{1}{2}$ aparece al tener en cuenta la teoría de dieléctricos que establece que el trabajo de carga para un dieléctrico sujeto a un campo externo es igual a la mitad de la energía de interacción. En este caso particular, el trabajo gastado en polarizar el solvente es equivalente a la mitad de la energía de interacción soluto-solvente. La contribución electrostática se calcula de la siguiente manera

$$\Delta G_{ele} = G_{ele} - E^0, \tag{2.11}$$

donde E^0 es la energía libre electrostática del soluto aislado y G_{ele} es la energía libre electrostática en solución

$$\begin{aligned} E^0 &= \langle \Psi^0 | \hat{H}_M^0 | \Psi^0 \rangle \\ G_{ele} &= \langle \Psi | \hat{H} | \Psi \rangle \end{aligned} \tag{2.12}$$

La función de onda Ψ puede ser determinada mediante una minimización variacional del valor esperado de $\hat{H}$,

$$\langle \Psi | \, \hat{H}_M^0 + \frac{1}{2}\hat{V}^{int} \left|\Psi\right\rangle . \tag{2.13}$$

La función de onda del soluto y el campo de reacción son mutuamente dependientes y por lo tanto deben ser optimizados simultáneamente. En la aproximación de Hartree-Fock esto se realiza mediante una modificación de la matriz de Fock, cuyos elementos incluyen

una contribución $\hat{V}^{int}$ para el término monoelectrónico.

$$F_{\mu\nu} = F^0_{\mu\nu} + \langle\mu|\, \hat{V}^{int}\, |\nu\rangle\,. \tag{2.14}$$

2.1.3. Contribución no electrostática

Para solventes polares como el agua, el término dominante en la interacción soluto-solvente es el electrostático. Sin embargo, existe una importante contribución a la energía libre de hidratación que no puede ser despreciada, que proviene de la energía libre de solvatación *no polar*, G_{np}.

Las distintas metodologías desarrolladas para calcular G_{np}, recurren a parámetros geométricos del soluto y propiedades generales del solvente. Los términos de dispersión y cavidad pueden ser calculados de manera separada. En este caso, la componente de cavitación se puede calcular siguiendo la teoría de partícula escalada de Pierotti para cavidades esféricas. De acuerdo a este modelo, la energía de formación de la cavidad está dada por la suma de energías de cavitación de un átomo aislado, $\Delta G_{cav}(R_i)$, multiplicadas por el cociente entre la superficie atómica expuesta al solvente, A_i, y la superficie total de la molécula, accesible al solvente $4\pi R_i^2$,

$$\Delta G_{cav} = \sum_i^N \frac{A_i}{A_T} \Delta G_{cav}(R_i), \tag{2.15}$$

donde R_i es el radio del i-ésimo átomo de la molécula de soluto. En esta expresión, $\Delta G_{cav}(R_i)$ es la energía libre de cavitación del átomo i calculada según el formalismo de Pierotti, A_i es la superficie atómica expuesta al solvente del átomo i, A_T es la superficie molecular total expuesta al solvente y N es el número de átomos total.

De manera análoga, la contribución de dispersión se puede calcular también como una suma de contribuciones atómicas individuales:

$$\Delta G_{disp} = \sum_{i=1}^{N} G_{disp,i} = \sum_{i=1}^{N} \gamma_i A_i \tag{2.16}$$

donde $G_{disp,i}$ es la contribución de la energía de dispersión del átomo i, a la energía libre de solvatación y γ_i es el coeficiente de tensión superficial atómico.

Las Ecs. 2.15 y 2.16 expresan la componente no polar de la energía libre de solvatación como una suma de términos correspondientes a átomos individuales. Estos términos dependen directamente de la exposición al solvente que tenga cada uno de ellos en la molécula de soluto.

Un modelo simplificado incluye las contribuciones de cavitación y dispersión en un único término, proporcional al área superficial accesible al solvente (A),[38]

$$G_{np} = G_{cav} + G_{disp} = \sum_{i=1}^{N} \gamma_i A_i, \tag{2.17}$$

donde A_i es el área superficial accesible al solvente para todos los átomos del tipo i y γ_i es un parámetro atómico de solvatación determinado empíricamente ajustado con respecto a valores experimentales de energía libre de solvatación para cada tipo atómico.

El término no polar puede expresarse también en su forma más simple empleando un modelo empírico, como el producto del área superficial total de la cavidad A (superficie accesible al solvente), y un coeficiente de tensión superficial efectivo γ_{ef} más una constante empírica b,

$$G_{np} = G_{cav} + G_{disp} = \sum_{i=1}^{N} \gamma A_i + b \equiv \gamma_{ef} SASA \tag{2.18}$$

Una formulación más precisa para calcular el trabajo requerido para crear la cavidad de volumen y forma apropiados, fue desarrollada por Pierotti.[44] En la presente Tesis se utiliza una forma simplificada de la formulación original:[45]

$$\begin{aligned} G_{cav} &= RT \left\{ -ln(1-y) + \frac{3y}{1-y} \left(\frac{R(M)}{R_{solv}} \right) + \right. \\ &\left. \left[\frac{3y}{1-y} + \frac{9}{2} \left(\frac{y}{1-y} \right)^2 \right] \left(\frac{R(M)}{R_{solv}} \right)^2 \right\} \end{aligned} \tag{2.19}$$

donde se introduce la función auxiliar y, dada por

$$y = \frac{\pi}{6} (2R_{solv})^3 \rho \tag{2.20}$$

Los parámetros usados son el cociente de los radios de las esferas rígidas de soluto ($R(M)$) y solvente (R_{solv}) y la densidad numérica reducida del solvente $\rho = N/V_{solv}$, donde N es el número de Avogadro y V_{solv} el volumen molar del líquido.

La antes mencionada contribución de la energía de cavitación fue calculada basada en la teoría de partícula escalada desarrollada por Reiss[46] y adaptada por Pierotti. De acuerdo a este modelo, la energía de formación de la cavidad está dada por la suma de energías de cavitación atómicas, $\Delta G_{cav}(R_i)$, multiplicadas por el cociente entre el área

atómica expuesta al solvente, A_i, y el área molecular total de la superficie accesible al solvente $4\pi R_i^2$,

$$\Delta G_{cav} = \sum_i \frac{A_i}{4\pi R_i^2} \Delta G_{cav}(R_i), \tag{2.21}$$

donde R_i es el radio del i-ésimo átomo de la molécula de soluto. Los términos de cavitación atómicos, $\Delta G_{cav}(R_i)$, son calculados por medio de la Ec. 2.19.

Es de esperar que el formalismo de Pierotti de buenos resultados para solutos y solventes de forma esférica y pequeño tamaño. Sin embargo, para solventes de forma no esférica, es necesario definir un radio efectivo para las moléculas del solvente. Para trabajar con solutos de forma compleja se puede recurrir a la extensión del formalismo de Pierotti realizado por Claverie.[47]

2.2. Energía de hidratación en el modelo COSMO

En esta Tesis se empleó el método de solvente continuo COSMO. Su principal característica es la representación del solvente como un medio conductor ($\epsilon = \infty$), una aproximación válida para solventes con $\epsilon \gg 1$, como el agua[1]. Esto modifica las condiciones de contorno del problema electrostático, ya que el campo eléctrico se anula en la superficie de la cavidad. A partir de esta condición se sigue que la carga superficial aparente, σ, está determinada por el valor local del potencial electrostático en lugar de por la componente normal de su gradiente. Para recuperar los efectos del valor finito de la constante dieléctrica del medio, la densidad de carga ideal, σ^*, correspondiente a $\epsilon = \infty$, es multiplicada por una función de ϵ adecuada, es decir

$$\sigma(\mathbf{r}_s) = f(\epsilon)\sigma^*(\mathbf{r}_s) \tag{2.22}$$

La función de escalamiento $f(\epsilon)$, se expresa como

$$f(\epsilon) = \frac{\epsilon - 1}{\epsilon + k} \tag{2.23}$$

con k pequeño.[48] En el trabajo original sobre COSMO,[40] Klamt sugiere $k = 0.5$ para moléculas neutras, remarcando que k depende de la forma de la cavidad y de la distribución de las cargas en el soluto. A partir de las cargas determinadas de este modo para el solvente y del conocimiento de la distribución de carga de la molécula se puede calcular la energía de interacción entre el solvente y la molécula de soluto.

[1] La constante dieléctrica del agua tiene un valor de ϵ=78.5.

Este método emplea los radios atómicos de van der Waals para construir la cavidad alrededor del soluto. Su precisión se puede mejorar significativamente optimizando los radios atómicos y coeficientes de tensión superficial para reproducir la energía libre de hidratación de moléculas pequeñas.

Utilizando ahora el modelo COSMO para determinar la contribución electrostática, se pueden definir los siguientes pasos: la generación de una cavidad dentro del solvente, la inserción del soluto no cargado y la generación de la distribución de cargas del soluto. De esta manera la energía libre de hidratación puede ser expresada en función de sus componentes como[48]

$$\Delta G_{hidr}(R_{solv}, R_i, \gamma, \epsilon) = \Delta G_{elec}(R_{solv}, R_i, \epsilon) + \Delta G_{disp}(\gamma, R_{solv}, R_i) + \Delta G_{cav}(R_{solv}, R_i) \tag{2.24}$$

La Ec. 2.24 es la Ec. 2.1 con dependencias explícitas para los distintos parámetros.[49]

Para determinar la componente electrostática por métodos cuánticos, la función de onda electrónica para el sistema se determina por medio de la ecuación de Schrödinger

$$(\hat{H}_0 + \hat{V}^{int}) \left|\Psi\right\rangle = E \left|\Psi\right\rangle , \tag{2.25}$$

donde $\hat{H}_0$ es el Hamiltoniano del soluto en fase gaseosa y $\hat{V}^{int}$ es el operador del potencial de interacción. Asumiendo que la superficie de la cavidad puede ser discretizada en un gran número N de elementos de superficie, de manera tal que la carga inducida, q_k, sobre el elemento de superficie k puede ser considerada una constante a ser determinada iterativamente, el operador del campo perturbativo se calcula por medio de la Ec. 2.9. Por tanto, la componente electrostática de la energía libre de hidratación toma la forma[43]

$$\Delta G_{ele} = \langle \hat{H}_0 + \frac{1}{2} \sum_{k,e} \frac{q_k}{|\vec{r}_e - \vec{r}_k|} \rangle + \frac{1}{2} \sum_{k,a} \frac{q_k Z_a}{|\vec{R}_a - \vec{r}_k|} - \langle \hat{H}_0 \rangle_0 \tag{2.26}$$

donde Z_a son las cargas nucleares atómicas del soluto; el último término corresponde a la energía de la fase gaseosa; las sumas sobre e, a y k, se refieren a los electrones, núcleos y cargas superficiales, respectivamente; y el valor promedio del tercer término se calcula usando la función de onda de la fase gaseosa. En la Ec. 2.26 se tiene en cuenta el trabajo realizado para polarizar el solvente.

Al calcular las cargas superficiales con el modelo COSMO, la constante dieléctrica del medio adquiere un valor $\epsilon = \infty$. En este caso el potencial electrostático debido al soluto y al solvente se anula sobre el la superficie del soluto. Asumiendo que un conjunto de cargas del soluto **Q** induce cargas **q'** sobre la superficie de la cavidad, el potencial sobre cada elemento superficial k se puede expresar en forma vectorial como **Φ=BQ+Aq'**, donde **A**

y **B** son las funciones de green de superficie-superficie, y superficie-soluto, que dependen únicamente de la geometría del sistema y de la distribución de la constante dieléctrica.[50] La condición de que **Φ** se anule en cada elemento superficial implica que

$$\mathbf{q}' = -\mathbf{A}^{-1}\mathbf{B}\mathbf{Q}. \tag{2.27}$$

Los efectos de tener una constante dieléctrica con un valor finito se pueden recuperar multiplicando **q'** por el factor de escalamiento $f(\epsilon)$ (Ec. 2.23),

$$\mathbf{q} = f\mathbf{q}' \tag{2.28}$$

La Ec. 2.26 se resuelve de manera iterativa actualizando en cada ciclo el valor de la carga mediante Ec. 2.27, siendo la función de onda Ψ y las cargas superficiales q_k mutuamente dependientes.

El valor de los parámetros atómicos, radios atómicos, radio del solvente y coeficientes de tensión superficial, modifican tanto la contribución electrostática como la contribución no electrostática a la energía libre de hidratación. En la Parte II correspondiente a los desarrollos de la presente Tesis se describe el proceso de optimización llevado a cabo para dichos parámetros, con el objetivo de mejorar la exactitud del cálculo de esta propiedad.

Capítulo 3

Energía libre de unión proteína-ligando

Las macromoléculas involucradas en procesos biológicos, como las proteínas, no actúan como entidades estáticas y aisladas.[51] Por el contrario, en solución acuosa se encuentran en constante movimiento y pueden presentar numerosas interacciones con otras proteínas, ácidos nucleicos, membranas, moléculas pequeñas y moléculas de solvente. Dichas interacciones usualmente muestran un alto grado de especificidad y una gran afinidad, fenómeno que se conoce como *reconocimiento molecular*. Éste consiste precisamente en la habilidad que tienen las moléculas para distinguir una de otra provocando su asociación o repulsión. Fundamentalmente, la gran mayoría de los procesos biológicos esta relacionada con la asociación molecular.

Partiendo de los fundamentos de la termodinámica general, la unión no covalente entre dos moléculas interactuantes se puede describir mediante el cambio de entalpía y entropía del sistema, que a su vez contribuyen a la variación de energía libre de Gibbs mediante

$$\Delta G = \Delta H - T\Delta S \tag{3.1}$$

donde el término entálpico (ΔH) está asociado a la contribución correspondiente a las interacciones presentes en el sistema y el entrópico (-TΔS) está asociado con las restricciones de movimiento de las moléculas al pasar del estado libre en solución a sus conformaciones en el estado unido. Los cambios relativos en ΔG pueden ser interpretados como una medida de cuán favorable es un proceso y cómo se distribuye el sistema en el equilibrio entre los estados unido y no unido. La energía libre ΔG es la diferencia de dos términos muy grandes (Ec. 3.1) cuya determinación revela por lo general un problema de compensación, de modo tal que resulta de gran importancia la información obtenida a partir del cambio de energía libre total y por otro lado el cambio de sus componentes

entálpicas y entrópicas. Sin embargo, la determinación precisa de las mismas representa un desafío para la biofísica computacional, debido a que un cálculo exacto de dichas cantidades involucra una correcta descripción del modelo físico subyacente y una exploración exhaustiva de la hipersuperficie de energía del sistema.

Se desprende entonces que el conocimiento cuantitativo del proceso de unión es esencial para comprender mejor el reconocimiento molecular. Esto requiere una comprensión detallada de las interacciones involucradas, sus contribuciones a la formación del complejo y la determinación de los cambios conformacionales que pueden ocurrir en el mismo.

El objetivo principal de este Capítulo es presentar la comprensión actual acerca del reconocimiento molecular, en el contexto de la asociación proteína-ligando. Se introducen en primer lugar los fundamentos termodinámicos que gobiernan la asociación macromolecular en sistemas biológicos. Esto comprende la definición de energía libre de unión, de sus componentes entálpicas y entrópicas, y de las diferentes contribuciones de energía, con su correspondiente formulación teórica. Se menciona también el papel que juega la flexibilidad de proteína y ligando en la unión de ambas partes, así como los efectos del solvente (desplazamiento y reorganización, e interacciones soluto-solvente). En una segunda parte del Capítulo se describen los métodos computacionales aplicados en la actualidad para calcular la energía libre de unión. Se describen brevemente aquellos que incluyen el solvente de manera explícita, y se basan en un cálculo exhaustivo de la hipersuperficie de energía, como los métodos *Free Energy Perturbation* (FEP) y *Thermodynamic Integration* (TI), y con más detalle los métodos utilizados en el presente trabajo de Tesis, que se caracterizan por integrar el solvente de manera continua, recurriendo únicamente al cálculo de energía de puntos extremos. Dentro de estos métodos aproximados, se encuentran el método *Molecular Mechanics-Poisson Boltzmann Superficial Area* (MM-PBSA) y el método *Molecular Mechanics-Generalized Born Superficial Area* (MM-GBSA).

3.1. Energía Libre de Unión

3.1.1. Entalpía y entropía

La asociación no-covalente de dos macromoléculas está gobernada por los principios de la termodinámica general. La reacción química que nos interesa es la asociación entre un receptor, en particular una proteína (P) y un ligando (L) en solución acuosa, para formar un complejo proteína-ligando (PL):

$$P + L \rightleftarrows PL \tag{3.2}$$

El proceso descripto por la reacción 3.2 tiene asociado un cambio en la energía libre de Gibbs. Esta función de estado se puede escribir como la suma de un término entálpico y un término entrópico (Ec. 3.1), donde ΔG es el cambio de energía libre de la reacción, ΔH y ΔS corresponden a los cambios en entalpía y entropía respectivamente y T es la temperatura absoluta del sistema. La unión proteína-ligando es favorable cuando el cambio total de energía libre es negativo.[52] Resulta de gran importancia en el estudio de la asociación molecular, el conocimiento de la información combinada acerca de los cambios de energía libre, entalpía y entropía del sistema, y no solamente la información acerca del cambio total de energía libre. Esto se debe a que por lo general las componentes entálpicas y entrópicas son de gran magnitud y de signo contrario conduciendo en algunos casos a un problema de compensación. Calculando las contribuciones de manera separada, se puede obtener una comprensión más detallada del fenómeno de asociación representado por la Ec. 3.2.

La contribución entálpica a la energía libre de unión proviene de los distintos tipos de interacciones entre las partes, como por ejemplo electrostáticas, de van der Waals, enlaces de hidrógeno, entre otras. El cambio de entropía, por otra parte, está asociado con las restricciones de movimiento que sufren ambas partes al pasar del estado libre en solución, al estado unido, y por lo tanto involucra el cambio en los grados de libertad del sistema. También incluye la entropía del solvente originada por la reorganización del mismo luego de la inserción del ligando en la cavidad formada en el receptor.

La estimación de las contribuciones entrópicas en el proceso de unión proteína-ligando es uno de los mayores desafíos de la biofísica computacional relacionada al cálculo de energía libre de unión. Esto se debe en primer lugar a que la entropía requiere en principio un muestreo del espacio de fases completo y por lo tanto es computacionalmente muy costoso. En segundo lugar, cuando se considera el cambio de entropía, es claro que una incerteza pequeña en la variación de este valor (al ser multiplicada por T) se traduce en un error grande en la energía libre total, por lo que resulta necesario determinar su valor con la mayor precisión posible. Por último, los términos de entropía (rotacionales, traslacionales, torsionales, vibracionales) por lo general están correlacionados, lo que dificulta el cálculo preciso de sus contribuciones de manera separada. En cuanto a los efectos de solvatación, tales como la reorganización del solvente, el efecto hidrofóbico y la pérdidas de moléculas de agua fuertemente unidas durante el proceso de asociación, se sabe que éstos pueden contribuir significativamente a la energía libre de unión, sin embargo es difícil determinar su valor de manera adecuada.

3.2. Afinidad de unión y equilibrio

Es importante mencionar que la variación de energía libre de una reacción química está directamente relacionada con la concentración de reactantes y productos. Se define la constante de asociación de la reacción como el cociente de las concentraciones de cada componente. En el equilibrio queda determinada por:

$$K_{eq} = \frac{[PL]_{eq}}{[P]_{eq}[L]_{eq}} C^{\circ} \tag{3.3}$$

donde $[P]_{eq}$, $[L]_{eq}$ y $[PL]_{eq}$ son las concentraciones de equilibrio de proteína libre, ligando libre y complejo proteína-ligando, respectivamente y C° es la concentración estándar. K_{eq} es la constante de asociación en el equilibrio, utilizada para expresar la afinidad de unión de la reacción. Su valor puede ser determinado experimentalmente. La constante de asociación está relacionada con la energía libre de unión ΔG por medio de la siguiente expresión[53]:

$$\Delta G = \mu^{\circ}_{PL} - \mu^{\circ}_{P} - \mu^{\circ}_{L} + RT\ln\frac{[PL]}{[P][L]} C^{\circ} \tag{3.4}$$

donde ΔG es el cambio de energía libre de la reacción, μ° es el potencial químico estándar. Para las reacciones bioquímicas, las condiciones estándar se definen generalmente como 25°C (298 K), reactantes y productos presentes a una concentración de 1 M y presión de 1 atm. La Ec. 3.4 puede ser re-expresada como[54],

$$\Delta G = \Delta G^{\circ} + RT\ln\frac{[PL]}{[P][L]} C^{\circ} \tag{3.5}$$

donde ΔG° es el cambio de energía libre asociado con la reacción en condiciones estándar, R es la constante de los gases, T es la temperatura absoluta. En el equilibrio, asumiendo que la reacción ocurre a temperatura y presión constantes, la energía libre de Gibbs alcanza un valor mínimo, de modo que $\Delta G = 0$. De esta manera, la constante de equilibrio queda directamente relacionada con el cambio de energía libre de Gibbs estándar,

$$\Delta G^{\circ} = -RT\ln K_{eq} \tag{3.6}$$

Despejando K_{eq} se obtiene que,

$$K_{eq} = \frac{[PL]_{eq}}{[P]_{eq}[L]_{eq}} = e^{-\Delta G^{\circ}/RT} \tag{3.7}$$

La constante de equilibrio de una reacción se puede calcular de esta manera a partir de la energía libre estándar, y viceversa. Se puede notar a partir de la Ec. 3.7 que una

variación de K_{eq} de 10 veces a temperatura ambiente corresponde a una variación de ΔG° de 1.4 kcal·mol^{-1} que es menor al valor de energía libre de un enlace de hidrógeno débil. Por otra parte, el error en la medición de la constante de equilibrio es usualmente del 10 al 20 % lo que implica un error en ΔG° de 0.1-0.25 kcal·mol^{-1}.[55] Esto indica la necesidad de una determinación precisa del valor de energía libre.

En adelante se eliminará el superíndice "°" de las cantidades termodinámicas, asumiendo siempre que nos referimos a estados estándar.

Al ser un potencial termodinámico, la función de energía libre depende únicamente de los estados final e inicial del sistema. Como fue mencionado al inicio de este Capítulo, los estados configuracionales más importantes de la unión ligando-proteína son el unido y no unido, que representan precisamente los estados final e inicial del sistema respectivamente. En este caso, el cambio de energía libre de la reacción en solución ($\Delta G_{u,sol}$) puede ser determinado como:

$$\Delta G_{u,sol} = G(PL)_{sol} - [G(P)_{sol} + G(L)_{sol}] \tag{3.8}$$

Una manera de obtener la energía libre del complejo proteína-ligando $G(PL)_{sol}$, relativa a la de proteína y ligando libres en solución [$G(P)_{sol}$ y $G(L)_{sol}$, respectivamente], consiste en simular la transferencia de ambas partes separadas en fase gaseosa a la fase en solución con la correspondiente reorganización del solvente. La diferencia de energía libre ($\Delta G_{u,sol}$) es difícil de calcular de manera directa, pero se puede determinar aplicando un ciclo termodinámico como el de la Fig. 3.1, de manera que:

$$\begin{aligned} \Delta G_{u,sol} &= \Delta G_{u,gas} + \Delta\Delta G_{solv} \\ &= \Delta G_{u,gas} + \Big[\Delta G(PL)_{solv} - \Delta G(P)_{solv} - \Delta G(L)_{solv}\Big] \end{aligned} \tag{3.9}$$

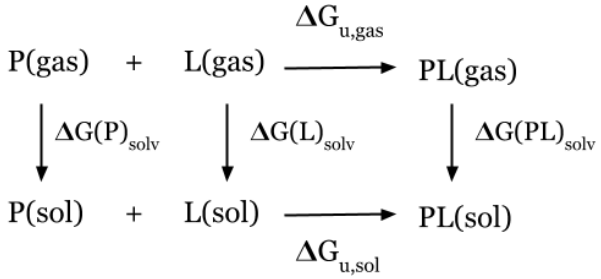


Figura 3.1: Ciclo termodinámico para calcular la energía libre de unión en solución para un ligando (L) que se une a una proteína (P) para formar un complejo (PL).

donde $\Delta G_{u,gas}$ corresponde a la energía libre de unión entre proteína (P) y ligando (L) en fase gaseosa y $\Delta G(P)_{solv}$, $\Delta G(L)_{solv}$ y $\Delta G(PL)_{solv}$ son los términos de energía libre de solvatación de la proteína libre, ligando libre y complejo proteína-ligando, respectivamente. La Ec. 3.9 muestra cómo la energía libre de unión se puede descomponer

en un término de interacción en fase gaseosa y contribuciones de solvatación del complejo proteína-ligando y las partes interactuantes no unidas.

3.2.1. Formulación desde la Termodinámica Estadística

Partiendo de la termodinámica estadística se puede obtener una formulación teórica más rigurosa de la energía libre de unión. El cambio de energía libre estándar, de la asociación en solución acuosa de una proteína (P) y un ligando (L), $P + L \rightleftarrows PL$, puede ser expresado como el cociente de las integrales de configuración[54],

$$\Delta G = -RT\ln\Big(\frac{C^\circ}{8\pi^2}\frac{Z_{PL}}{Z_P Z_L}\Big) + P^\circ \Delta V_{PL} \tag{3.10}$$

donde ΔG es la energía libre de unión del complejo proteína-ligando en condiciones estándar, R es la constante de los gases, T es la temperatura absoluta, C° es la concentración estándar y $P^\circ \Delta V_{PL}$ es el trabajo presión-volumen asociado al cambio de dimensión del sistema al pasar de dos moléculas libres al complejo unido. Este último término es generalmente despreciable en agua a 1 atm.

La integral de configuración Z_x se puede expresar como[56]

$$Z_x = \int J(\mathbf{q}) \mathrm{e}^{-\beta E_x(\mathbf{q})} d\mathbf{q} \tag{3.11}$$

donde $\mathbf{q}$ representa el conjunto de $3N$-6 coordenadas internas, $J(\mathbf{q})$ es el Jacobiano de la transformación de coordenadas cartesianas a internas, $\beta = 1/kT$ y $E_x(\mathbf{q}) = U_x(\mathbf{q}) + W_x(\mathbf{q})$, donde $U_x(\mathbf{q})$ es la energía potencial de las especies X en fase gas, y $W_x(\mathbf{q})$ es la energía de solvatación efectiva de X, que incorpora los efectos de grados de libertad del solvente. Aparte del término presión-volumen, $W_x(\mathbf{q})$ representa el trabajo de transferir el soluto X en la conformación ($\mathbf{q}$) de la fase gas al solvente. Cuando consideramos el complejo PL, el conjunto de coordenadas internas incluye las de L, P, y las seis coordenadas externas relativas ζ de L con respecto a P, por lo tanto $\mathbf{q} \equiv (\zeta, \mathbf{q}_P, \mathbf{q}_L)$. En ese caso, la integral de configuración en la Ec. 3.11 incluye las configuraciones para P y L cuando están unidos.

Asumiendo que en el complejo los grados internos de libertad accesibles de las partes (P y L) están débilmente acopladas con los movimientos relativos de traslación y cuerpo rígido rotacional, la energía se puede separar como[56],

$$E_{PL}(\zeta, \mathbf{q}_P, \mathbf{q}_L) = E_{PL}(\zeta) + E_{PL}(\mathbf{q}_P, \mathbf{q}_L) \tag{3.12}$$

Un observable macroscópico se puede calcular como un promedio sobre la cantidad mi-

croscópica correspondiente pesada por la distribución de probabilidad de Boltzmann, que en este caso estará dada por $p(\mathbf{q}) = \exp(\text{-}\beta E(\mathbf{q})/Z)$ para una configuración $\mathbf{q}$ determinada. Generando una colección representativa de configuraciones, la suma sobre todos los estados se puede aproximar por un promedio sobre un conjunto finito de configuraciones. Este conjunto de configuraciones es denominado "conjunto estadísitico" (del inglés, *ensemble*). Teniendo en cuenta la hipótesis ergódica, y reemplazando la energía dada por la Ec. 3.12 en la Ec. 3.11, la energía libre de unión de un complejo proteína-ligando (Ec. 3.10) se puede re formular como una suma de contribuciones aditivas independientes[56],

$$\Delta G^{PL} = \Delta\langle U\rangle + \Delta G^{solv} - T\Delta S^{\mathrm{RB}} - T\Delta S^{\mathrm{int}} \tag{3.13}$$

donde $\langle ...\rangle$ representa un promedio sobre el ensamble, ΔG^{solv} es el cambio en energía libre de solvatación, y el cambio de entropía se separa en un término que representa la asociación de cuerpo rígido ΔS^{RB} y otro término asociado a los grados de libertad internos S^{int}, dados por

$$\Delta S^{\mathrm{RB}} = \frac{1}{T}\langle E_{PL}\rangle + R\ln\Big(\frac{C^{\circ}}{8\pi^2} z^{\mathrm{RB}}\Big) \tag{3.14}$$

$$z^{\mathrm{RB}} = \int e^{-\beta E_{PL}(\zeta)} d\zeta \tag{3.15}$$

$$S^{\mathrm{int}} = -R\int J(\mathbf{q}) p(\mathbf{q}) \ln p(\mathbf{q}) d\mathbf{q} \tag{3.16}$$

La integral de entropía interna (Ec. 3.16) está relacionada con la entropía conformacional asociada con el número de rotámeros accesibles y la entropía vibracional del sistema. Se debe mencionar que el cambio en la entropía del solvente está incluida en el término ΔG^{solv}.

Existen otras formas de particionar la energía libre de unión en componentes aditivas partiendo de la deducción de la Ec. 3.13. Este tipo de aproximaciones es utilizado por métodos que recurren a campos de fuerza de mecánica molecular para determinar las componentes, por lo que se las conoce como tipo *Force Field*. Se pueden considerar por ejemplo, las contribuciones electrostáticas y las de solvatación, calcular cada una de ellas y combinar los resultados para obtener el cambio de energía libre. Según la naturaleza de las interacciones y la importancia que tiene cada una de ellas para un sistema particular de estudio, se adoptan diferentes componentes para la energía libre de unión como por ejemplo,

$$\Delta G_{u,sol} = \Delta G_{int} + \Delta G_{conf} + \Delta G_{solv} - T\Delta S \tag{3.17}$$

La Ec. 3.17 tiene en cuenta de manera explícita las interacciones proteína-ligando

que surgen de la cercanía entre ambas partes (ΔG_{int}), los cambios conformacionales en la proteína y el ligando durante el proceso de unión (ΔG_{conf}). Se puede notar que la suma de los dos primeros términos es equivalente al primer término de la Ec. 3.13. Se consideran también la contribución debida a la influencia del solvente (ΔG_{solv}) y la entropía asociada a las restricciones de movimiento generadas por la unión de ambas partes (ΔS).

Se mencionan a continuación distintos tipos de interacciones cuyas contribuciones a la energía libre de unión son significativas.

Energía electrostática Las interacciones electrostáticas involucradas en el proceso de unión de un complejo proteína-ligando se pueden clasificar como: carga-carga, carga-dipolo y dipolo-dipolo. En los métodos que emplean una descripción basada en Mecánica Molecular usualmente se utilizan cargas puntuales para representar el sistema. En este caso las interacciones electrostáticas típicas carga-carga son aquellas que se generan entre átomos cargados, grupos funcionales del ligando o cadenas laterales de la proteína, como las cargadas positivamente (grupos amino, lisina, arginina, histidina) y cargadas negativamente (grupos carboxilos, grupos fosfatos, grupos ácidos). Las mismas tienen una influencia importante en el cambio de entalpía asociado con el proceso de unión.

Interacciones de van der Waals Son de gran importancia para la estructura e interacción de moléculas biológicas, y en particular para la unión de un complejo proteína-ligando. Las fuerzas de van der Waals pueden ser tanto atractivas como repulsivas. Las primeras involucran dipolos inducidos que se generan por las fluctuaciones de densidades de carga que ocurren entre átomos cercanos neutros que no están unidos por enlaces covalentes. Las interacciones repulsivas ocurren cuando la distancia entre dos átomos resulta muy pequeña, debido al principio de exclusión de Pauli. Entre todos los tipos de interacción intermolecular, las de van der Waals son las más débiles. Varían en un rango de 0.5 a 1.0 kcal·mol^{-1} (de 2 a 4 kJ·mol^{-1}) para interacciones átomo-átomo.[57] Sin embargo, debido a la gran cantidad de interacciones de este tipo que ocurren en la formación de un complejo proteína-ligando, su contribución total a la energía libre resulta significativa[58].

Enlaces de hidrógeno Los enlaces de hidrógeno son interacciones atractivas no covalentes entre un hidrógeno unido covalentemente a un grupo electronegativo ("donor"), y otro átomo electronegativo como un oxígeno o nitrógeno ("aceptor"). Una característica particular de este tipo de interacciones es que la distancia interatómica entre los átomos involucrados es menor a la suma de sus radios de van der Waals y por lo general involucran un número limitado de partes interactuantes. Esto se puede interpretar como un tipo de

valencia. Este tipo de enlace direccional es crucial para las moléculas biológicas, ya que tiene una gran importancia en la determinación de sus estructuras tridimensionales y de sus asociaciones intermoleculares. La fuerza de un enlace de hidrógeno en sistemas biológicos varía entre 1 y 3 kcal·mol^{-1} (4-13 kJ·mol^{-1}), siendo menor a la de los enlaces iónicos o covalentes (aproximadamente 100 kcal/mol para un enlace covalente hidrógeno-carbono).[57] Sin embargo, por el mismo hecho de ser una interacción débil, se pueden romper y volver a formar rápidamente en un proceso de unión, lo que facilita la asociación macromolecular y la actividad biológica. Otro factor que aporta gran importancia a este tipo de interacciones es su especificidad. Ésta se relaciona con las reglas geométricas que debe cumplir un enlace de hidrógeno, como ser sus orientaciones, longitudes y preferencias angulares.

3.3. Cálculos de energía libre de unión

En la última década se han realizado numerosos esfuerzos con el propósito de alcanzar mayor precisión en el cálculo de la energía libre de unión, de manera computacionalmente eficiente. Esto ha sido potenciado por el incremento exponencial de recursos computacionales, que se tradujo favorablemente en innovaciones metodológicas. Al mismo tiempo, es posible estudiar una mayor cantidad de sistemas debido dichos avances. Sin embargo, los cálculos de energía libre y concretamente su aplicación a sistemas macromoleculares siguen siendo problemáticos por distintos factores, entre los cuales se pueden mencionar: el modelo de energía subyacente, el tamaño del sistema de interés (usualmente miles de átomos) y la flexibilidad, ya que se basan en la exactitud de la función de energía potencial del sistema y en una exploración exhaustiva de la hipersuperficie de energía. Otro gran desafío viene dado por la dificultad en el tratamiento del solvente, que involucra la correcta descripción de la influencia de las moléculas de agua del sitio de unión y un modelo de solvente continuo suficientemente preciso.

3.3.1. Métodos computacionales

En los últimos 20 años se ha visto un avance considerable en el desarrollo teórico y de algoritmos para el cálculo de afinidades de unión, que van desde estimaciones rápidas para ser usadas en un contexto de *High Throughput Docking* (HTD) hasta los cálculos mucho más lentos pero de mayor precisión, como los de simulaciones de dinámica molecular basados en campos de fuerza clásicos, como FEP. Para energías libres de unión relativas, los mismos pueden alcanzar una precisión de alrededor de 1.2 kcal mol^{-1}.[59] Los métodos TI también se encuentran dentro de este grupo. A pesar de la precisión que alcanzan, son

computacionalmente costosos. La mayoría de estas aplicaciones se basan en campos de fuerzas de mecánica molecular, pero en los últimos años se han desarrollado y aplicado de manera exitosa, métodos de mecánica cuántica para sistemas biomacromoleculares en el contexto del diseño y descubrimiento de nuevos fármacos[18,56,60,61].

La verdadera potencialidad de los métodos de energía libre es que las diferencias en energías libres pueden ser obtenidas con una precisión estadística de pocas kcal/mol, a un costo computacional razonable. La correlación con los valores calculados experimentalmente para la afinidad de unión dependerá de la correcta descripción de la superficie de energía.

3.3.1.1. Perturbación de Energía Libre (FEP)

Este método se utiliza para calcular diferencias de energía libre entre dos estados de una simulación de dinámica molecular, considerando el solvente de manera explícita. Estos dos estados pueden ser por ejemplo una proteína y un ligando en su estado libre (A) y un complejo proteína-ligando (B), o dos complejos proteína-ligando con ligandos diferentes.

La diferencia de energía libre puede ser directamente calculada por la ecuación de Zwanzig,[62]

$$\Delta G = -\frac{1}{\beta}\ln\langle e^{-\beta((E_B(\mathbf{p},\mathbf{R})-E_A(\mathbf{p},\mathbf{R}))}\rangle \tag{3.18}$$

donde A y B representan los dos estados, ΔG es la diferencia entre energías libres de ambos estados, $\beta = 1/k_BT$ donde k_B es la constante de Boltzmann y T es la temperatura absoluta, E es la energía del sistema y $\langle ... \rangle$ denota el promedio sobre el conjunto estadístico. Para efectuar el cálculo se introducen estados intermedios entre A y B, que pueden ser descriptos en términos de un parámetro de acoplamiento λ ($0 \leq \lambda \leq 1$). La aproximación más simple consiste en efectuar una interpolación lineal, pero también se pueden utilizar otras relaciones más complicadas,[63]

$$E_\lambda = \lambda E_B + (1-\lambda)E_A \tag{3.19}$$

El número de estados intermedios se elige de manera tal que el promedio sobre el ensamble se realice sobre cambios de energía comparables a k_BT. El cambio de energía libre entre dos puntos vecinos estará dado por la Ec. 3.19, y el cambio total de energía será la suma de dichos términos:

$$\Delta G = -\frac{1}{\beta}\ln\langle e^{-\beta(\Delta E_\lambda))}\rangle \tag{3.20}$$

El método FEP ha sido ampliamente utilizado para calcular diferencias de energía libre.[64–66] Sin embargo, la convergencia de este tipo de cálculos es crítica. El principal

problema está relacionado con la dificultad de muestrear el espacio conformacional, sobre todo para sistemas grandes como proteínas, debido a un solapamiento insuficiente de las densidades de espacios de fase involucradas. FEP converge adecuadamente sólo cuando la diferencia entre los dos estados es suficientemente pequeña; esto requiere usualmente que la perturbación se divida en un conjunto de pequeños pasos que son calculados de manera independiente. Esto implica un costo computacional muy alto.

3.3.1.2. Integración Termodinámica (TI)

Dada una función de energía como la de la Ec. 3.19, tanto la función de partición como la energía libre serán funciones del parámetro λ,

$$G(\lambda) = -k_B T \ln Z(\lambda) \tag{3.21}$$

Diferenciando la Ec. 3.21 con respecto a λ y realizando la integral entre los estados final e inicial del sistema,

$$G(1) - G(0) = \int_0^1 \left\langle \frac{\delta E(\lambda)}{\delta \lambda} \right\rangle \mathrm{d}\lambda \tag{3.22}$$

donde se ha empleado el promedio sobre el ensamble. El lado izquierdo de la Ec. 3.22 es la diferencia de energía libre, mientras que el lado derecho se puede aproximar por una suma discreta, de modo que:

$$G(B) - G(A) = \sum_i \left\langle \frac{\delta E(\lambda)}{\delta \lambda} \right\rangle \Delta\lambda_i \tag{3.23}$$

El uso de la ecuación 3.23 para calcular diferencias de energía libre ΔG se conoce como Integración Termodinámica (TI).[67] A diferencia de FEP, TI se basa en un promedio sobre la derivada de la función de energía con respecto al parámetro de acoplamiento λ, y no en un promedio sobre diferencias finitas de las funciones de energía. El costo computacional de realizar el promedio es despreciable comparado con el costo de generar el ensamble, y por lo tanto el mismo ensamble puede ser usado para calcular la diferencia de energía libre ya sea por la Ec. 3.20 o la Ec. 3.23.

3.3.1.3. MM-PBSA/MM-GBSA

Un problema fundamental de los métodos FEP y TI es su gran demanda computacional, sumado a la dificultad que tienen para alcanzar la convergencia. El poder de cómputo que requieren está relacionado con dos factores principales. El primero es el tratamiento explícito del solvente y el segundo es la exploración exhaustiva de la hipersuperficie de energía potencial que se realiza para determinar la diferencia de energía libre entre dos

estados del sistema. Una alternativa atractiva a estos métodos computacionalmente intensivos son los denominados métodos de puntos extremos, que consideran únicamente los estados final e inicial en los cálculos de energía libre. Al igual que FEP y TI, se derivan de primeros principios. Una diferencia importante es que la contribución del solvente se tiene en cuenta utilizando métodos de solvente continuo. Estas aproximaciones representan entonces un compromiso entre eficiencia y precisión. Los métodos *Molecular Mechanics Poisson-Boltzmann Surface Area* (MM-PBSA) y *Molecular Mechanics Generalized Born Surface Area* (MM-GBSA) son los más comúnmente utilizados.

En MM-PBSA/GBSA la energía libre de un complejo proteína-ligando se calcula empleando una ecuación de tipo *Force Field* (Ec. 3.13). Por lo general los cálculos se basan en un conjunto conformacional generado por simulaciones de Dinámica Molecular (MD) con solvente explícito. De esta manera se incluye en el cálculo de energía libre el cambio conformacional del sistema. Una vez efectuada la simulación computacional, se realiza una re-evaluación de energía usando un modelo de solvente continuo, sobre las estructuras del conjunto conformacional extraídas de la trayectoria de dinámica molecular.

En la práctica, se calcula la energía libre de un complejo y sus componentes, siguiendo la *aproximación de trayectoria única* (Fig. 3.2). Esta consiste en realizar una única simulación de dinámica molecular y extraer de la misma las conformaciones del complejo proteína-ligando, por lo que las conformaciones de las partes individuales en el estado libre se obtienen eliminando la proteína o ligando de cada una de las configuraciones, correspondientemente. A pesar de que en este caso no se tiene en cuenta la flexibilidad conformacional que ocurre durante la unión, esta aproximación conduce a resultados precisos siendo más adecuada para sistemas en los que no se esperan encontrar cambios conformacionales significativos entre las estructuras unida y no unida.[14]

Una vez realizada la simulación de MD, la energía libre de unión se calcula para cada conformación como la diferencia dada por la Ec. 3.13. La energía libre para cada componente individual $G(PL)_{sol}$, $G(P)_{sol}$ y $G(L)_{sol}$ se aproxima de la siguiente manera:

$$G(X)_{sol} = E_{MM} + G_{solv} - TS \tag{3.24}$$

donde X=PL, P o L. La energía libre en vacío E_{MM} se determina a partir de un campo de fuerzas de mecánica molecular para cada componente. La energía libre de solvatación G_{solv}, se puede separar en una componente polar (G^{pol}_{solv}) y una no polar (G^{np}_{solv}), donde la primera se estima a través de la solución de diferencias finitas de la ecuación de Poisson-Boltzmann (PB) usando un modelo de solvente continuo. El método MM-GBSA utiliza el modelo de solvente basado en la teoría de Born Generalizada (ver Cap. 2). El término G^{solv}_{np} se calcula por medio de la aproximación *Gamma SASA* definida en el Capítulo 2 (Ec. 2.17). Por último, la contribución de la entropía puede ser determinada mediante un

análisis de modos normales o análisis quasi-armónico para la componente vibracional (Ec. 3.16) mientras que para la componente roto-traslacional se puede utilizar la Ec. 3.14.

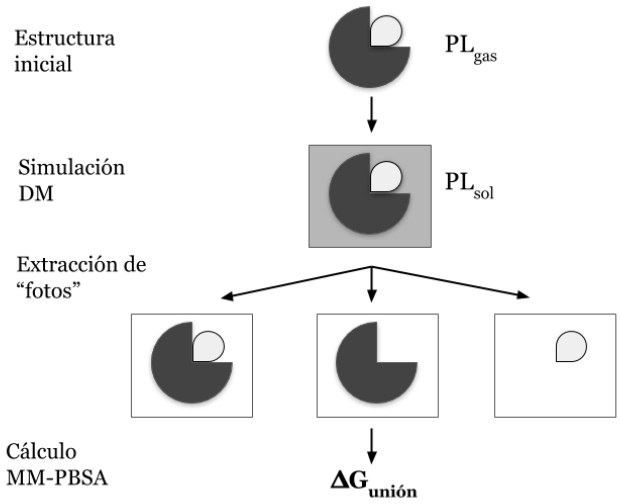


Figura 3.2: Cálculo MM-PBSA para determinación de energía libre de unión de acuerdo a la aproximación de trayectoria única.

La Ec. 3.13 puede ser aplicada entonces para calcular las energías libres MM-PBSA y MM-GBSA en una trayectoria de dinámica molecular de un complejo proteína-ligando. Para ello se extraen de dicha simulación un grupo de estructuras o configuraciones y se eliminan las moléculas de agua e iones. La selección se realiza con un intervalo de tiempo, definido previamente de manera de garantizar que no estén correlacionadas.[56] Se emplea para el cálculo de energía la aproximación de trayectoria única mencionada en la Sección anterior, por lo que las conformaciones de ligando y proteína libres se obtienen a partir del complejo de la trayectoria de dinámica molecular eliminando la proteína o ligando de cada una de las configuraciones. En esta aproximación, la energía de deformación de proteína y ligando se desprecian. Bajo estas condiciones se ha demostrado que la aproximación de trayectoria única reduce el ruido en cálculos de puntos extremos.

Capítulo 4

Cribado Virtual

En el largo proceso del descubrimiento de un fármaco, resulta de gran importancia contar con técnicas que permitan identificar en forma temprana, compuestos que se unan a un determinado blanco terapéutico, partiendo de grandes librerías químicas. A principio de los años 90', las técnicas experimentales de Cribado de Alto Rendimiento (*High Throughput Screening*, HTS) fueron introducidas en la industria farmacéutica como una técnica innovadora que permitiría realizar ensayos de actividad biológica de un gran número de compuestos contra un receptor determinado, de manera rápida y eficaz. Sin embargo, fueron apareciendo distintas debilidades con el tiempo, como un bajo rendimiento y un alto costo debido a un gran número de falsos positivos (FP) detectados (moléculas que no se unen al receptor, identificadas erróneamente como ligandos), y a que muchos compuestos seleccionados en este proceso fallaban en la etapa de optimización, debido a propiedades farmacocinéticas inadecuadas[2]. Esta baja tasa de éxito, junto con el costo económico de estas técnicas, hicieron que pierdan parte del protagonismo de la década pasada. Al mismo tiempo, el diseño de fármacos asistido por computadoras comenzó a cobrar mayor importancia, favorecido por el desarrollo teórico de técnicas de modelado molecular, avances en los algoritmos y mayor capacidad de cómputo.[68]

Aparecen así a finales de los 90' las técnicas de Cribado Virtual (CV), conformadas por un conjunto de métodos computacionales empleados con la finalidad de seleccionar/identificar a partir de un grupo de cientos de miles de moléculas, aquellas que tienen mayor probabilidad de unirse a un receptor determinado, priorizando las mismas para su posterior síntesis y evaluación experimental.[2] De esta manera, el CV representa un complemento al HTS y un criterio para la priorización de la síntesis y/o la adquisición de quimiotecas (bases de datos moleculares). Estas técnicas son económicas, rápidas y, hoy día, permiten considerar un número de compuestos *in silico* del orden de millones (cifra prohibitiva experimentalmente). Típicamente en un proceso de CV, una quimioteca virtual que contiene unas 10^4-10^8 estructuras es sucesivamente filtrada y reducida a una

colección de unas 100-1000 moléculas.

Los métodos computacionales de CV se pueden dividir en dos grandes grupos:[69] los basados en la estructura de ligandos conocidos (*Ligand-Based Virtual Screening*, LBVS), y basados en la estructura del receptor (*Structure-Based Virtual Screening*, SBVS). Resumidamente, podemos decir que LBVS toma como punto de partida la estructura de un ligando y realiza una búsqueda de candidatos dentro de una librería química virtual, en función de la semejanza que posea con el mismo, asumiendo el principio de propiedad-semejanza. Dicho principio establece que moléculas similares tendrán propiedades similares,[70] lo que significa que se debe contar de antemano con al menos una molécula o grupo de moléculas que se unan al blanco biomolecular. Por otro lado, SBVS toma en cuenta el conocimiento acerca de la estructura tridimensional (3D) del blanco -obtenida ya sea experimentalmente o de manera computacional- para identificar compuestos activos contra el mismo.

En este Capítulo se describen ambos métodos LBVS y SBVS, poniendo un mayor énfasis en este último, que fue utilizado en la presente Tesis. Se describen más adelante los pasos comprendidos en una implementación de un protocolo de SBVS para la búsqueda de un fármaco líder, haciendo hincapié en los tipos de función de *scoring* utilizadas por los métodos actuales. Luego se define una métrica comúnmente usada para evaluar la calidad de un método de CV.

4.1. LBVS

En este tipo de métodos se utilizan las propiedades topológicas, farmacofóricas, y fisicoquímicas de ligandos conocidos para identificar moléculas en una base de datos, que puedan unirse a un determinado receptor. Lo que se determina es la similitud de los compuestos, con respecto a compuestos activos de referencia contra un receptor determinado, o que muestra propiedades de interés. Este tipo de métodos resulta menos costoso que los basados en la estructura del receptor. Por este motivo, se emplean principalmente cuando el número de moléculas de la base de datos inicial es muy grande.

4.1.1. Métodos basados en similitud

Una forma de realizar la búsqueda de similitudes entre compuestos es utilizar patrones en la estructura de los mismos para compararlos. Luego se realiza la asignación de un puntaje a cada molécula de la base de datos, de acuerdo a la similitud con el ligando, que permite generar una lista ordenada o ranking, dejando en los primeros puestos aquellas con mayor probabilidad de ser activas. Estas son priorizadas para su posterior síntesis o adquisición, seguida por una evaluación experimental.

Otro tipo de métodos que se basa en el principio de similitud efectúa una comparación entre potenciales electrostáticos. Utilizando el potencial electrostático de un ligando de referencia, se pueden encontrar moléculas de una base de datos que posean una distribución electrostática similar, y que por lo tanto puedan complementarse con el entorno electrostático del receptor. Esto implica que pueden tener probabilidades de generar una acción contra el mismo.

4.1.2. Farmacóforos basados en ligandos

Se denomina $farmacóforo$ al conjunto de características necesarias para asegurar la interacción óptima con un determinado receptor conduciendo (o bloqueando) una respuesta biológica. A los farmacóforos obtenidos a partir de un grupo de ligandos se los denomina, *farmacóforos basados en ligandos*. Éstos incorporan características comunes a un conjunto de ligandos que presentan actividad biológica contra un blanco terapéutico común. Se asume entonces, que dichas características son responsables de la actividad del ligando, y el farmacóforo es empleado para buscar moléculas que tengan una distribución similar de características en la librería química que será utilizada en el CV.

En este trabajo no nos concentraremos en este tipo de metodologías, por lo que se refiere al lector a las recientes revisiones que se han realizado sobre las mismas [71, 72], para una profundización del tema.

4.2. SBVS

A pesar de que usualmente compuestos similares poseen actividades similares, puede ocurrir que algunas modificaciones en un compuesto pueden perjudicar las interacciones proteína-ligando, conduciendo a una pérdida de actividad. Si el principio de similaridad fuera aplicado en casos de este tipo, las predicciones probablemente serían equivocadas. Para evitar este tipo de problemas que pueden aparecer en los métodos basados en ligandos, es importante considerar la información que se puede obtener del receptor.

Los métodos SBVS se utilizan en los casos en los que la estructura 3D de un blanco se encuentra disponible, ya sea de manera experimental o por medio de un modelado por homología. Una de las ventajas de este tipo de métodos con respecto a LBVS, es la posibilidad de encontrar candidatos a fármacos novedosos, ya que la búsqueda no se realiza en función de ligandos pre-existentes. Por otro lado, también permite determinar, al menos en principio, la estructura del complejo receptor-ligando, lo cual resulta de gran utilidad para la etapa de optimización del líder.

Dentro de los métodos basados en la estructura del receptor, se pueden mencionar el *docking molecular* (DM), al cual nos referiremos con más detalle en este Capítulo, el *diseño de novo* y el método de *farmacóforos basado en la estructura*.

El método de *docking* de proteína-ligando es una de las técnicas más empleadas en el cribado virtual y en el diseño racional de fármacos. Esta metodología utiliza la estructura tridimensional de la proteína para predecir de qué manera se unen los compuestos de una librería química en el sitio de unión. Después de efectuar la búsqueda conformacional, se asigna un puntaje o *score* a la pose con el propósito de generar un ranking para las moléculas de la base de datos. De esta manera, se impone un filtro estructural en una librería química, con el objetivo de priorizar moléculas para su síntesis y evaluación experimental.

El *diseño de novo* consiste en tomar un determinado grupo químico (al que se denomina fragmento) y colocarlo en el sitio de unión de un receptor. Luego, este fragmento se somete a un proceso de *docking* y *scoring*, es decir una búsqueda de la mejor conformación a la que se le asigna un determinado puntaje de acuerdo a su afinidad. El proceso continúa uniendo el fragmento a otros y repitiendo el paso anterior, o bien haciéndolo crecer en el espacio disponible que se tenga. La principal ventaja de este método es que se pueden diseñar compuestos nuevos, pero que pueden presentar dificultades a la hora de sintetizar los mismos.

Por último, la *búsqueda de farmacóforos basados en la estructura*, es similar al proceso de farmacóforos basado en ligandos. La principal diferencia radica en la manera de obtener la distribución de características.

SBVS basado en DM

Para realizar un proceso de SBVS basado en *docking*, se deben considerar los siguientes pasos:

- Construcción de la librería de compuestos
- Preparación de la proteína
- Definición del sitio de unión
- Búsqueda conformacional (etapa de *docking*)
- Asignación del puntaje a la pose de cada molécula (etapa de *scoring*)
- Re-*scoring* y optimización
- Selección de *hits*

4.2.1. Construcción de la librería de compuestos

Para poder realizar un cribado virtual es importante contar con una librería química virtual de compuestos lo más diversa posible. Actualmente existen más de 15 bases de datos químicas públicas disponibles, conformadas por varios miles a millones de moléculas. Entre las más conocidas, se pueden mencionar ChEMBL,[73] con 1,4 millones de compuestos y ZINC[74] que contiene 230 millones de moléculas disponibles comercialmente con estructuras 3D y 750 millones en formato 2D. En un proceso de cribado virtual resulta demasiado costoso computacionalmente evaluar tanta cantidad de compuestos. Puede ocurrir que dentro de esas librerías se encuentren moléculas con propiedades no deseadas y que por lo tanto no pasen la etapa de optimización del líder. Por lo tanto, es común efectuar un filtro en dichas librerías antes de realizar los cálculos de *docking*. Dentro de estos filtros se pueden mencionar, por ejemplo, aquellos que descartan compuestos reactivos, tóxicos, o que pueden tener propiedades físicas desaconsejables como baja solubilidad.

4.2.2. Preparación de la proteína

La estructura tridimensional de la proteína se extrae usualmente de la base de datos *Protein Data Bank* (PDB).[75] Antes de realizar un SBVS se debe preparar la proteína de manera cuidadosa, teniendo en cuenta distintas consideraciones. En primer lugar es necesario agregar átomos de hidrógeno a la proteína que, por lo general, no están resueltos en las estructuras de rayos-X. Las posiciones de dichos átomos deben asegurar que se alcance el patrón de enlaces de hidrógeno más favorable. Esta etapa requiere fijar el estado de protonación correcto de residuos ionizables que pueden encontrarse en la región donde se une el ligando al receptor, y por lo tanto pueden modificar el resultado del *docking*. Por lo general, una vez agregados los hidrógenos, se debe minimizar la estructura para resolver eventuales choques estéricos entre átomos cercanos. Por otro lado es importante determinar de manera correcta la orientación y estados de protonación de los residuos His, que pueden encontrarse en una conformación neutra o cargada. La orientación de átomos pesados de los residuos Gln y Asn debe ser examinada con cuidado, debido a que los átomos de oxígeno y nitrógeno en dichos residuos pueden presentar posiciones ambigüas. También es importante corroborar la correcta asignación de cargas efectuada para los residuos Asp, Glu, Lys, Arg e His.[76]

4.2.3. Determinación del Sitio de Unión

Una vez que la estructura de la proteína ha sido preparada se debe identificar el sitio de unión, es decir la región donde se unen los ligandos a la proteína, para delimitar la

zona en la que se realizará el *docking* de compuestos. En los complejos con estructura cristalográfica conocida, el sitio de unión se puede identificar fácilmente como la región que comprende los residuos situados a una determinada distancia del ligando. En caso de no contar con dicha estructura, o si se tiene en cambio la estructura de la proteína cristalizada sin ligandos, se puede recurrir a datos experimentales disponibles, o a programas específicos para la determinación del mismo.[77,78]

Los límites de la región de la proteína en la cual se debe realizar el *docking* de cada molécula, restringen el espacio ocupado por las poses. Esto puede tener consecuencias en la eficiencia del método en general. Si la región seleccionada para efectuar la búsqueda conformacional es muy pequeña, puede ocurrir que algunos potenciales ligandos de la base de datos, al ser de mayor tamaño no se puedan ubicar correctamente y por lo tanto sean descartados de manera errónea. Por el contrario, si el sitio es demasiado grande el método puede tomar mucho tiempo evaluando regiones del espacio conformacional innecesariamente.

4.2.4. *Docking* Molecular en CV

La parte computacionalmente más demandante de un SBVS corresponde al *docking* de moléculas de la librería química en el sitio de unión del receptor. Ésta metodología, intenta simular de manera computacional lo que ocurre en un proceso de reconocimiento molecular. El *Docking* Molecular ha sido ampliamente utilizado para determinar el modo de unión (o *pose*) de moléculas pequeñas en complejo con un receptor. Sin embargo, uno de sus mayores potenciales está en su aplicación en un proceso de alto rendimiento (*High Throughput Docking*, HTD). En este tipo de metodologías, se pueden identificar dos etapas: primero se realiza una exploración de las diferentes conformaciones que puede adoptar una molécula pequeña en el sitio de unión del receptor mediante la modificación de sus posiciones y orientaciones, con el objetivo de predecir la(s) pose(s) de preferencia, es decir la(s) más estable(s) para la formación del complejo (etapa de *docking*). Idealmente, la *pose* correcta corresponde al mínimo de energıa en la hipersuperficie de energıa potencial. Una vez seleccionada la *pose* de preferencia para cada molécula de la librería, una *función de scoring* (FS) asigna un puntaje (*score*) que permite evaluar la calidad de dicha pose con respecto a las demás de la base de datos, generando de este modo una lista ordenada o ranking (etapa de *scoring*). Las moléculas ubicadas en los primeros puestos, serán las que tengan mayor probabilidad de unirse al receptor. Se puede decir entonces que, la etapa de *docking* discrimina entre diferentes conformaciones de la misma molécula para predecir la pose de preferencia, mientras que la etapa de *scoring* ordena las moléculas *dockeadas* de la librería química generando un *ranking* de acuerdo al *score*. Estas etapas se encuentran esquematizadas en la Figura 4.1.

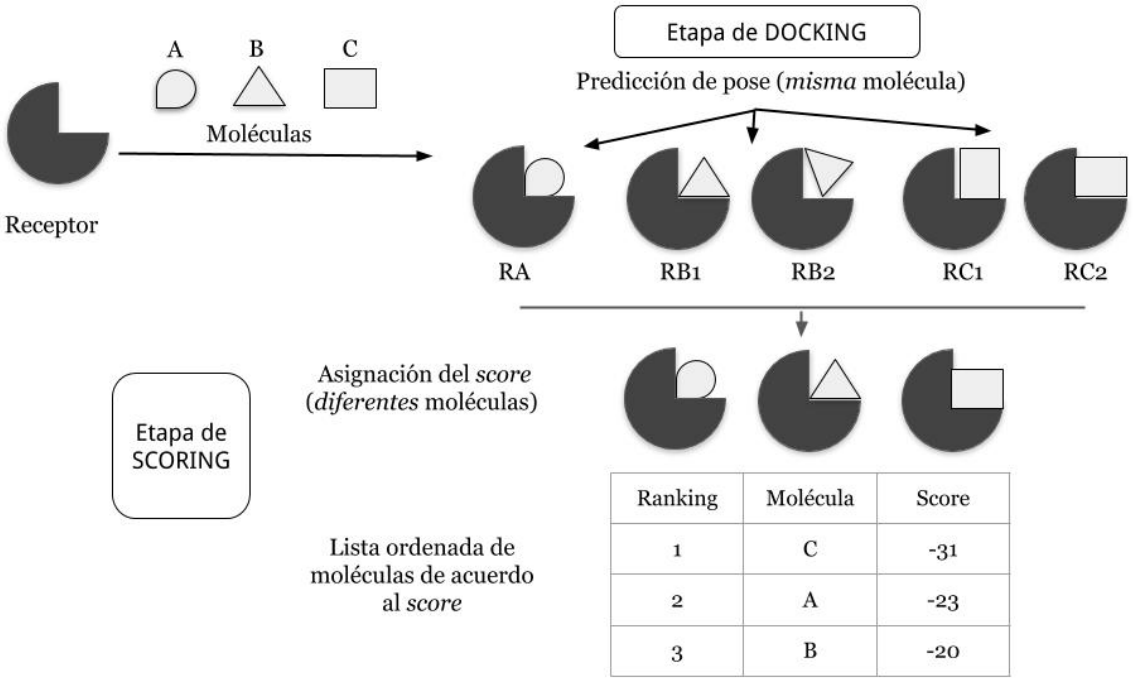


Figura 4.1: Esquema ilustrativo de las etapas de *docking* y *scoring* comprendidas en *Docking* Molecular.

4.2.5. Función de *Scoring*

El papel más importante de la *función de scoring* (FS) es la identificación de potenciales *hits* o candidatos a fármacos líderes para una proteína, a través de un CV de una base de datos. Una buena FS debería ser capaz de discriminar correctamente ligandos de no-ligandos, siendo el objetivo principal de un HTD la obtención de un sub-conjunto enriquecido de potenciales ligandos, a partir de una librería química de miles de moléculas.

A pesar de que algunos programas utilizan la misma función para realizar la búsqueda conformacional (o determinación de la *pose*) y la asignación del *score*, la función de *scoring* debe realizar la evaluación de millones de compuestos de manera rápida y precisa. En la actualidad, éste es uno de los mayores desafíos de esta metodología ya que ambas etapas (*docking* y *scoring*) están relacionadas. Si la pose resultante del *docking* no es correcta, tampoco será confiable el valor del *score*. Por otro lado, debido a la gran cantidad de aproximaciones involucradas, las funciones de *scoring* presentan dificultades para describir correctamente las interacciones entre proteína y moléculas pequeñas. Por este motivo, se puede encontrar una gran cantidad de FP y perder una gran cantidad de verdaderos positivos como resultado de un proceso de selección de *hit* de un CV. A pesar de los continuos desarrollos en ésta área, el desarrollo de una función de *scoring* adecuada para este tipo de aplicaciones sigue siendo un desafío actual.

Clasificación de las Funciones de *Scoring*

Las funciones de *scoring* se pueden clasificar como: basadas en campos de fuerza, empíricas, y basadas en un potencial estadístico[68,79]. Recientemente se ha publicado un nuevo esquema de clasificación agregando funciones de *scoring* basadas en *machine learning* a la lista antes mencionada. A continuación se mencionarán las características más generales de cada grupo.

Basadas en campos de fuerza (*force-field based*) Este tipo de funciones de *scoring* contiene términos de energía de interacción atómica con significado físico, incluyendo interacciones electrostática, de van der Waals, enlaces de hidrógeno. Dichos términos permiten monitorizar la calidad geométrica del modo de unión, con lo cual penalizarán las moléculas cuya pose, previamente determinada, no sea correcta de acuerdo a los patrones de interacción detectados. En el contexto de *docking*, los campos de fuerza más utilizados son AMBER (*Assisted Model Building with Energy Refinement*),[80,81] CHARMM(*Chemistry at Harvard Macromolecular Mechanics*),[82] GAFF (*General Amber Force Field*)[83] compatible con el campo de fuerzas AMBER, con parámetros definidos para moléculas orgánicas, desarrollado específicamente para ser utilizado en diseño racional de fármacos.

La precisión de este tipo de funciones de *scoring* está sujeta a la forma de la función de energía potencial y a los parámetros empleados por el campo de fuerzas. Recientemente, se han desarrollado también nuevas funciones basadas en mecánica cuántica, que permiten la correcta descripción de fenómenos como la polarización electrónica, transferencia de carga e interacciones covalentes. Este tipo de funciones emplea una combinación de métodos de mecánica molecular y mecánica cuántica (QM/MM) y mejoran la precisión de los basados en FF, pero también se incrementa el costo computacional requerido.

Una función de *scoring* típica basada en FF es la implementada en el programa DOCK,[84] que emplea parámetros de energía tomados del campo de fuerzas AMBER. La forma funcional está dada por la suma de componentes de energía de van der Waals y un término electrostático,

$$E = \sum_i \sum_j \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} \right) \tag{4.1}$$

donde r_{ij} es la distancia entre el átomo i de la proteína y el átomo j del ligando, A_{ij} y B_{ij} son los parámetros de van der Waals, y q_i y q_j son las cargas atómicas. La constante dieléctrica introducida en el término de Coulomb da cuenta implícitamente del efecto del solvente.

Otros ejemplos de FS basadas en campos de fuerza son AutoDock[85,86], GOLD[87] entre otras.

Empíricas Las FS empíricas estiman la energía libre de unión de un complejo como la suma de términos de energía con distintos pesos asociados,

$$\Delta G_u = \sum_i c_i \Delta G_i \tag{4.2}$$

donde ΔG_i son los diferentes términos de energía y c_i son los coeficientes que se determinan por medio de una regresión, con respecto a datos experimentales de afinidad de unión de un conjunto de complejos proteína-ligando con estructuras experimentales conocidas. Algunos programas de *docking* conocidos, como por ejemplo FlexX[88] y Glide[89,90] utilizan una función de *scoring* empírica del tipo

$$\Delta G_u = c_{vdw}\Delta G_{vdw} + c_{hb}\Delta G_{hb} + c_{elec}\Delta G_{elec} + c_{solv}\Delta G_{solv} + c_{tor}\Delta G_{tor} \tag{4.3}$$

Los distintos términos de energía corresponden a la energía de van der Waals (ΔG_{vdw}), enlace de hidrógeno (ΔG_{hb}), energía de interacción electrostática (ΔG_{elec}), solvatación (ΔG_{solv}) y energía de ángulos torsionales (ΔG_{tor}).

Estas FS se caracterizan por su bajo costo computacional, asociado a la suma de términos relativamente simples de energía. Sin embargo, no permiten establecer una buena relación entre afinidades de unión y las estructuras cristalográficas. Algunos ejemplos de FS empíricas conocidas y ampliamente utilizadas son,[91] ChemScore,[92] ICM,[93,94] Surflex,[95] X-Score,[96] entre otras.

Basadas en un potencial estadístico Estas FS se basan en la información disponible sobre la estructura tridimensional del sistema, y emplean potenciales de energía que se derivan precisamente de la información obtenida a partir de dichas estructuras determinadas experimentalmente. El potencial se construye a partir de la frecuencia con la que se encuentra un par de átomos a una distancia r, en una base de datos relacionando de esta manera la frecuencia con la interacción de dichos átomos[97]. Estos potenciales son utilizados para asignar un puntaje o *score* a la calidad de la pose de *docking* en un complejo proteína-molécula. La precisión de las funciones de *scoring* basadas en un potencial estadístico depende de la base de datos estructural, es decir de la cantidad de observaciones estadísticamente significativas de contactos entre un par de tipo de átomos en particular.

Representan la afinidad como una suma de interacciones de pares de átomos proteína-ligando. Estos potenciales se derivan a partir de complejos de estructuras conocidas de

la base de datos PDB, donde las distribuciones de probabilidad de distancias interatómicas entre diferentes pares de tipos de átomo proteína-ligando se convierten, asumiendo distribuciones energéticas tipo Boltzmann, en funciones de potencial. La energía libre de interacción se calcula sumando las contribuciones de los pares de átomos dentro de una cierta distancia.

La ventaja de este tipo de FS frente a las dos descriptas anteriormente radica en el buen compromiso entre precisión y costo computacional que poseen. Las FS basadas en un potencial estadístico son más adecuadas para la predicción de poses que para la predicción de afinidad de unión, debido a que se derivan exclusivamente a partir de datos estructurales y no utilizan datos experimentales de afinidad. Como ejemplos de este tipo de funciones se pueden mencionar DrugScore[16,98] y PMF.[99,100]

Basadas en Machine-Learning A diferencia de las funciones de *scoring* clásicas que asumen una forma funcional matemática, las FS basadas en Aprendizaje Automático utilizan una variedad de algoritmos de aprendizaje automático, como redes neuronales, *random forest*, etc. Este grupo de funciones de *scoring* ha demostrado significativas mejoras frente a los métodos anteriores.[101–103] Por lo general, son mayormente utilizadas para una re-evaluación, efectuada sobre los *hits* encontrados por el proceso de *docking*. Esto se debe a que dependen del conjunto de entrenamiento empleado.

Las funciones de *scoring* juegan un papel importante en un proceso de SBVS, por lo que se han ido desarrollando diversos tipos de funciones a lo largo de los últimos años. Sin embargo es difícil encontrar una función de *scoring* que pueda arrojar buenos resultados para cualquier tipo de sistema proteína-ligando, por lo que éste sigue siendo un desafío en estudios de *docking*. Todas las funciones de *scoring* tiene sus ventajas y desventajas.[104]

4.2.6. Re-scoring y Optimización

Como resultado de un proceso de SBVS, se obtienen la *pose* y el *score* predichos para un gran número de moléculas en complejo con un receptor. Esto se refleja en un ranking de moléculas, en el que se sitúan en primer lugar las de mayor *score*. Solamente un pequeño número de dichos compuestos será sometido a una evaluación experimental, por lo que la selección cuidadosa de los *hits* es un paso fundamental del SBVS, a fin de reducir el número de FP. Esto ocurre por las deficiencias de las funciones de *scoring* para predecir afinidad de manera precisa, debido al número de aproximaciones involucradas, como por ejemplo una estimación poco precisa de la energía de solvatación e interacciones receptor-molécula pequeña, y la consideración de receptor rígido.

Por este motivo, resulta importante contar con estrategias posteriores al DM para

aumentar las posibilidades de que las moléculas que continúen a la etapa de optimización, sean verdaderos ligandos. Entre ellas se puede mencionar la aplicación de filtros estructurales, e inspección visual.[105] Una vez aplicados estos filtros, se puede continuar el proceso con una etapa de re-*scoring* de las moléculas seleccionadas con métodos que incorporan una descripción más precisa de las distintas contribuciones a la afinidad de unión, como energías de solvatación, interacciones electrostáticas y entropía.[106] La energía de solvatación puede ser calculada en este caso con modelos de solvente continuo como el de Poisson-Boltzmann (PB) o el Generalizado de Born (GB). Un ejemplo de este tipo de métodos es MM-PBSA o el método MM-GBSA.[13,14,107,108] Estas aproximaciones, más precisas, son inadecuadas para la primera etapa del CV debido al costo computacional que implican, siendo aplicadas en mayor medida para efectuar estimaciones de afinidades de unión en etapas posteriores.[109] Seguidamente, se puede realizar un análisis más exhaustivo mediante la determinación de energías libres de unión de manera aún más precisa. En este caso se emplean métodos más robustos de energía libre como los métodos alquímicos FEP y TI, que a su vez demandan mayor tiempo y poder de cómputo.

4.2.7. Selección de hits

Una vez finalizada la etapa de selección e inspección de las moléculas (*hits*) se procede a su compra o síntesis química en el laboratorio, para realizar su evaluación experimental. El número de compuestos seleccionados para esta etapa depende de los recursos disponibles. La etapa final de un proceso de CV es la confirmación de la actividad predicha de los *hits* seleccionados, con ensayos bioquímicos y biológicos específicos, que determinan si la molécula se une al receptor y con qué afinidad.

4.3. Evaluación de un Protocolo de Cribado Virtual

La calidad de un algoritmo de *docking* en CV puede ser evaluada teniendo en cuenta dos aspectos: la fidelidad de la pose de *docking* y el enriquecimiento de la base de datos.[5,6] Antes de efectuar un SBVS basado en DM, es importante realizar un estudio retrospectivo, para conocer el poder predictivo del método, evaluando su habilidad para separar compuestos activos de inactivos. En este caso, la base de datos sobre la que se efectúa el CV está conformada por inhibidores conocidos que se unen al receptor, y por un alto porcentaje de no ligandos. Para que este tipo de estudios resulte más confiable, sería necesario tener la comprobación experimental de las moléculas inactivas de dicho conjunto, es decir de las que no se unen al receptor. Como esta confirmación es difícil de encontrar, se recurre a los denominados *decoys*. Estos son moléculas que, se asume,

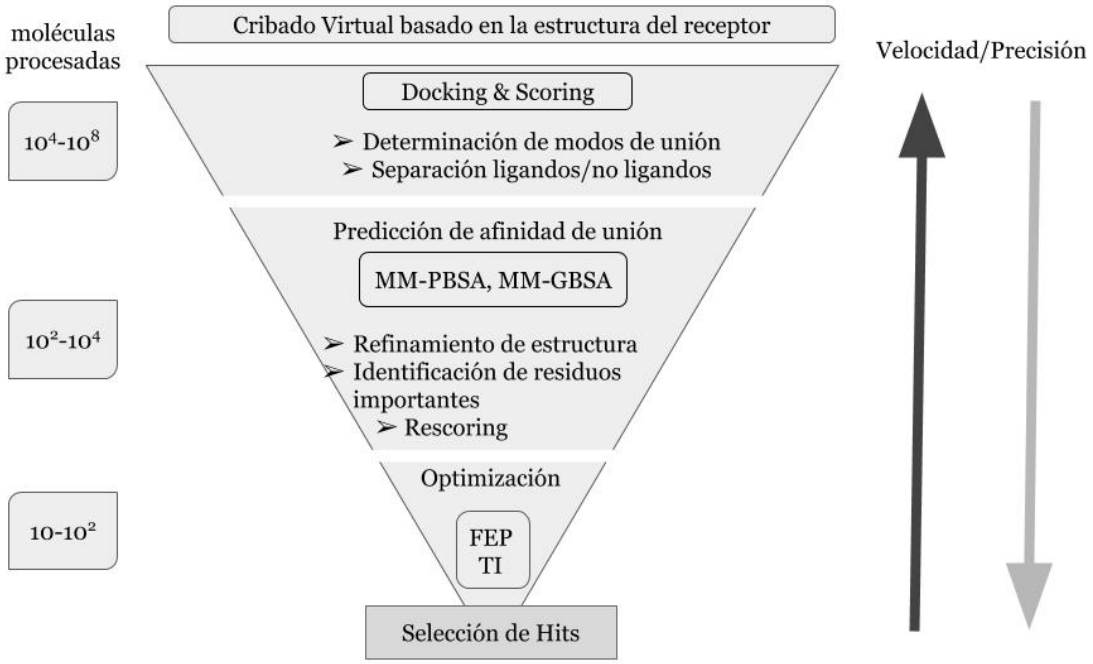


Figura 4.2: Esquema ilustrativo del proceso de re-*scoring* posterior a un CV basado en la estructura del receptor.

no se unen al receptor por lo que son consideradas no-ligandos de la base de datos. Una forma de generar estas moléculas consiste en seleccionar aquellas que posean propiedades fisicoquímicas similares a la de los ligandos activos, pero que sean estructuralmente diferentes. La base de datos DUD-E (*Directory of Useful Decoys Enhanced*) se encuentra públicamente disponible y fue desarrollada siguiendo los principios anteriormente expuestos. Posee actualmente un conjunto de 50 *decoys* para cada una de las moléculas activas de aproximadamente 130 receptores diferentes.

La calidad de una función de *scoring* en un proceso de CV se determina entonces por su capacidad de asignar un mejor *score* a las moléculas activas frente a las inactivas, enriqueciendo el número de potenciales ligandos en los primeros puestos del *ranking* de la base de datos.[110–112]

Métricas de Evaluación

Para medir la calidad de un proceso retrospectivo de SBVS se han propuesto distintos tipos de métricas.[113] Las más utilizadas son: *Enrichment Factor* (EF), *Receiver Operating Characteristic* (ROC) y *Area Under the Receiver Operating Characteristic Curve* (AUC). Solamente nos referiremos en esta Sección al EF, ya que ha sido el criterio de evaluación usado en esta Tesis.

Antes de concentrarnos en las medidas de evaluación propiamente dichas, es necesario

establecer algunas definiciones. En un proceso de CV, partiendo de una librería química virtual de N moléculas, las primeras n del ranking generado por el *score* pasarán a conformar una sub-librería donde se espera encontrar mayormente compuestos que se unan al receptor. Las primeras X moléculas del ranking generado por el *docking* son denominadas *positivos* y las moléculas descartadas *negativos*. Si la actividad de dichas moléculas es conocida, se agrega la clasificación de *verdaderos* y $falsos$. De esta manera se tienen dentro de las n moléculas o *hits*, compuestos activos o "verdaderos positivos" (*true positives*, TP) y compuestos inactivos o *decoys*, "falsos positivos" (*false positives*, FP). Por otro lado, entre las moléculas que no han sido seleccionadas encontraremos también compuestos activos, que se denominan "falsos negativos" (*false negatives*, FN), mientras que los *decoys* no seleccionados, "verdaderos negativos" (*true negatives*, TN).

4.3.1. Factor de enriquecimiento (EF)

La segunda parte del proceso de *docking* en un CV consiste asignar un *score* a las poses de las moléculas de la base de datos. La función de *scoring* empleada para tal fin debe ser lo suficientemente rápida y precisa como para asignar una mayor puntuación a las moléculas activas. El objetivo central de esta etapa en un estudio retrospectivo es identificar correctamente los compuestos activos frente a los inactivos o *decoys*. El parámetro que describe cuánto mejora la tasa de *hits* comparada con una selección aleatoria de activos dentro de la base de datos se conoce como factor de enriquecimiento. Este se define como el cociente entre el número de compuestos activos en un subconjunto seleccionado de moléculas y el número total de compuestos activos elegidos de manera aleatoria en el conjunto total,

$$\mathrm{EF}_{\%} = \frac{\mathrm{TP/n}}{\mathrm{A/N}} = \left(\frac{\mathrm{TP}}{\mathrm{TP+FN}}\right) \times \frac{\mathrm{N}}{\mathrm{n}} \tag{4.4}$$

donde TP es el número de inhibidores conocidos encontrados dentro del sub-grupo de n moléculas seleccionadas al X% del *ranking* la base de datos, generado de acuerdo al *score*, A es el número total de moléculas activas en la librería química de N moléculas, y FN es el número de activos que no fueron seleccionados dentro del subgrupo.

Es importante mencionar que la selección apropiada de *decoy* es un paso previo importante en el proceso de evaluación de una función de *scoring*. Como se describió anteriormente, éstos deben poseer propiedades fisicoquímicas similares a las de los compuestos activos de la base de datos, y al mismo tiempo deben ser químicamente diferentes. Imponiendo la primera condición, se evita que el EF sea erróneamente alto debido a una separación de propiedades físicas simples, mientras que la segunda condición reduce la probabilidad de que el conjunto de *decoys* contenga moléculas que se pueden unir al receptor, generando una reducción del EF.

4.3.2. Curvas ROC

La curva ROC representa la tasa de TP en función de la tasa de FP. La curva se obtiene graficando la "señal de actividad" (taza de TP, Se) en función del "ruido detectado" (tasa de FP, 1-Sp) a varios umbrales de actividad. En la Fig. 4.3, se puede observar un gráfico teórico de las curvas ROC. Ésta se puede interpretar como una curva de probabilidad. Un protocolo de CV que posee una performance de distribución aleatoria, genera una curva ROC que tiende a la línea diagonal Se=1-Sp, mientras que un protocolo de CV capaz de detectar la señal correcta tendrá una gráfica ROC que se curva por encima de la diagonal. Para distribuciones ideales, la cuva crece verticalmente desde el extremo inferior izquierdo y luego corre horizontalmente hacia el extremo superior derecho. Por lo tanto cuanto mas tienda la curva ROC hacia la esquina superior izquierda, mejor será la calidad del protocolo de CV para separar compuestos activos de inactivos.

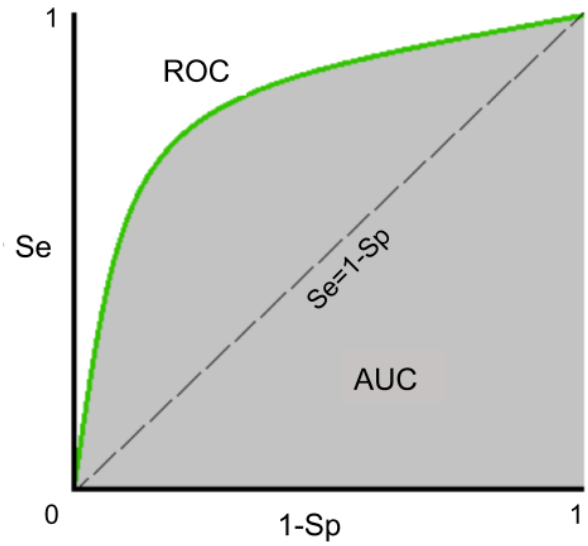


Figura 4.3: Ejemplo de curvas ROC.

4.3.3. Gráficos AUC

Otra forma de interpretar los resultados de los gráficos ROC es por medio del cálculo del área bajo la curva ROC (AUC). Esta métrica es útil para obtener la probabilidad de que el método de cribado asigne un puesto más alto en el ranking a compuestos activos elegidos de manera aleatoria que a compuestos inactivos. El AUC a un porcentaje específico de la librería rankeada se calcula como la suma de todos los rectángulos formados

por los valores de Se y 1-Sp para diferentes umbrales:

$$\mathrm{AUC} = \sum_{i}^{n} \left[(\mathrm{Se_{i+1}})(\mathrm{Sp_{i+1}} - \mathrm{Sp_i}) \right] \tag{4.5}$$

El valor de AUC es igual a 1 para una distribución ideal, mientras que una distribución aleatoria tendrá un valor de AUC de 0.5. Los protocolos de CV cuyo rendimiento sea mejor que aleatorio, dará valores de AUC entre 0,5 y 1, mientras que los protocolos de CV que asignen mejor *score* a compuestos inactivos tendrá valores de AUC menores que 0,5. Un modelo excelente tiene un valor de AUC cercano a 1, lo que significa que posee una buena medida de separabilidad. Un modelo pobre tendrá un valor de AUC cercano a 0, lo que significa que tiene una mala medida de separabilidad. Es decir, predice compuestos activos como inactivos y viceversa. Cuando el valor de AUC es 0,5, esto implica que el modelo no tiene capacidad de efectuar una separación. Si AUC es 0,7 esto significa que hay un 70 % de probabilidad de que el modelo sea capaz de distinguir entre compuestos activos e inactivos.

4.3.4. Factores condicionantes del docking-scoring

Si bien el método de DM representa un filtro importante en un proceso de CV, se pueden mencionar cuatro limitaciones importantes de esta metodología: El número de moléculas de la librería química sobre la que se efectúa el CV es muy grande, por lo tanto una evaluación de todas las conformaciones posibles requiere un elevado costo computacional que por lo general, excede los recursos disponibles. En segundo lugar, la flexibilidad del receptor se desprecia, por el mismo motivo. Y por último, el cálculo de afinidad de unión estimado por la función de *scoring* no es exacto, principalmente por la dificultad en la determinación precisa de energía libre de solvatación y la incorrecta o nula contribución entrópica considerada. Los factores antes mencionados pueden conducir a un número alto de FP.

Flexibilidad de receptor y ligando A pesar de que se conoce la influencia que tiene la consideración de la flexibilidad de ligando y proteína en el proceso de unión proteína-molécula pequeña, éste es un aspecto que presenta desafíos de implementación en los métodos actuales.[114–116] Una forma de incorporar la flexibilidad consiste en usar un conjunto de estructuras del receptor, ya sea tomando varias estructuras experimentales de la misma proteína. El problema de esto se encuentra en la cantidad de tiempo que involucra el cálculo, ya que se debe realizar un proceso de *docking* para cada una de dichas estructuras del receptor, para la librería de moléculas.

La flexibilidad puede ser introducida totalmente para el ligando, mientras que el receptor se considera rígido en la mayoría de los casos. El tratamiento flexible del receptor tiene un costo computacional muy alto, por lo que se usan aproximaciones como el introducir movilidad en las cadenas laterales de algunos aminoácidos a partir de librerías de rotámeros (GOLD), o la construcción de una geometría difusa que engloba distintas conformaciones (módulo FlexE de FlexX 103 , AUTODOCK 104).

Moléculas de agua La incerteza para incorporar las moléculas de agua de manera explícita en el proceso del *docking* puede conducir a resultados erróneos en cuanto a la estructura y afinidad de la pose resultante. La principal dificultad de este aspecto está en la correcta determinación del número de moléculas de agua involucradas en las interacciones proteína-ligando en el sitio de unión de manera rápida y la predicción de la posición que toman en el sitio de unión. Han sido numerosos los intentos orientados al desarrollo de nuevas metodologías que incorporen moléculas de agua en un proceso de *docking*. Algunas se basan en Dinámica Molecular, pero resultan muy costosas computacionalmente para ser aplicadas en un contexto de CV.[115,117,118] Otros programas han incorporado la consideración de la orientación de moléculas de agua durante el *docking*, pero al dejarlas como partes fijas de la estructura de la proteína, no tienen en cuenta el efecto sobre las mismas causado por la diversidad de tamaño, forma y propiedades fisicoquímicas de cada ligando. Estas metodologías, aún si han logrado un gran avance con respecto al tratamiento de moléculas de agua en un proceso de *docking*, hacen que hoy en día no sean aplicables en un contexto de CV.

Parte III

Desarrollo

Capítulo 5

Reparametrización del modelo COSMO

En el modelado computacional de sistemas biológicos, es sumamente importante describir correctamente de qué manera influye el solvente sobre las propiedades moleculares. En particular, en el área del diseño de fármacos asistido por computadoras, es fundamental contar con metodologías que permitan calcular de manera precisa los efectos de la solvatación en sistemas de interés. Esto permite relacionar los cálculos teóricos con valores experimentales, y efectuar predicciones para propiedades en un entorno acuoso.

El presente trabajo de Tesis comprende el estudio de interacciones proteína-ligando en fase solución. Por este motivo se buscó en primer lugar, un modelo de solvente adecuado para ser implementado con métodos cuánticos semi-empíricos, para la determinación de energías libres de solvatación en complejos proteína-ligando.

La precisión de los métodos QM que incluyen el solvente con un modelo continuo, depende en gran medida de la disponibilidad de radios atómicos y coeficientes de tensión superficial optimizados. Una manera de obtener dichos parámetros consiste en realizar una optimización de los mismos sobre un conjunto pequeño de moléculas con valores experimentales conocidos de energías libres de hidratación.

Se realizó una re-parametrización del modelo de solvente continuo COSMO para los métodos semi-empíricos RM1, PM6 y PM7, incluyendo la componente de energía de solvatación no polar. Los parámetros geométricos (radios y coeficientes de tensión superficial) fueron re-optimizados para reproducir energías libres de hidratación de un conjunto de moléculas pequeñas con valores experimentales conocidos (*conjunto de entrenamiento*) empleando dos procedimientos de optimización distintos. El conjunto de entrenamiento antes mencionado fue constituído por moléculas neutras, debido a que las moléculas ionizadas poseen mayor error experimental. Se obtuvieron nuevos parámetros atómicos, que

fueron validados posteriormente en un grupo de moléculas distinto al usado en la parametrización (*conjunto de validación*), determinando el error absoluto promedio con respecto a los valores experimentales conocidos de energías libres de hidratación.

Este Capítulo se organiza de la siguiente manera: se describe la metodología empleada para calcular la energía libre de hidratación de moléculas pequeñas, en la optimización de los parámetros. Luego se detallan las bases de datos usadas y los procedimientos seguidos para la optimización. Al finalizar el Capítulo, se encuentran los nuevos parámetros obtenidos, junto con los valores de las medidas estadísticas que se usaron para determinar la calidad de los mismos.

5.1. Metodología

En el Capítulo **??**, han sido presentadas las dos formas de modelar el solvente. Los cálculos cuánticos, usados en combinación con un modelo de solvente explícito, son muy demandantes para ser aplicados en sistemas biológicos, lo que hace que el uso de modelos de solvatación implícitos resulte más atrayente. En este trabajo de Tesis, se emplearon métodos de mecánica cuántica semi-empíricos (SQM) en combinación con el modelo de solvente continuo COSMO. El nivel de teoría usado fue elegido por el número de átomos de los sistemas biomoleculares que serán estudiados. Por otra parte, el hecho de emplear métodos cuánticos permite incluir los efectos de polarización mutua de soluto y solvente. Entre los modelos de solvente continuo que pueden ser empleados en combinación con métodos SQM, el modelo COSMO es útil por por su eficiencia computacional en el estudio de sistemas con cientos a miles de átomos.

El programa elegido para realizar la optimización de los parámetros fue MOPAC2012[22]. Este es un paquete gratuito para uso académico, especializado en métodos SQM, que permite calcular propiedades de sistemas moleculares de grandes dimensiones, en combinación con el modelo de solvente continuo COSMO, gracias a la implementación del algoritmo de escalamiento lineal MOZYME[23].

La energía libre de hidratación para moléculas pequeñas fue calculada como la diferencia de energía libre en fase solución y fase gas, por medio de la siguiente ecuación:

$$\Delta G_{hidr} = G^{COSMO} - E^{gas} + \Delta G_{np} \tag{5.1}$$

Con los valores de energía de interacción electrónica e interacción nuclear se calcula la componente electrostática de la energía de hidratación G^{COSMO}, a la que se debe agregar la contribución no polar ΔG_{np}. Se usaron los Hamiltonianos RM1, PM6 y PM7 para

calcular G^{COSMO}. El término no polar se obtuvo empleando las metodologías γSASA, γ_{ef}, y Claverie-Pierotti.

En la formulación γSASA, G_{np} se calcula como,

$$G_{np} = \sum_i \gamma_i Ai \equiv \gamma SASA \tag{5.2}$$

donde γ_i y A_i corresponden al coeficiente de tensión superficial, y a la superficie accesible al solvente para el átomo i, respectivamente.

Alternativamente, se define una aproximación similar a la de la Ec. 5.2 tomando un único coeficiente de tensión superficial para todos los átomos de la molécula,

$$G_{np} = \gamma_{ef} A + b \equiv \gamma_{ef} SASA \tag{5.3}$$

donde A es el área superficial accesible al solvente para toda la molécula y b es una constante. La denominación γ_{ef} está asociada al único coeficiente de tensión superficial usado para todos los átomos de la molécula

También fue empleada la metodología más precisa de Claverie-Pierotti (Ec. 2.21) para calcular el término no polar,

$$G_{np} = \sum_i \frac{A_i}{4\pi R_i^2} \Delta G_{cav}(R_i) \equiv CP \tag{5.4}$$

donde R_i es el radio del i-ésimo átomo de la molécula de soluto. Los valores de $\Delta G_{cav}(R_i)$ se calculan empleando la Ec. 2.19.

Definición de la cavidad: SES y SAS Para el cálculo de la contribución no polar, es importante definir correctamente la superficie que será usada para el cálculo de G_{np}. En MOPAC, la cavidad COSMO se construye a partir de los radios de van der Waals para los átomos que conforman la molécula, por lo que la superficie accesible al solvente es por definición, la SES (ver Fig. 2.1). La SAS se puede obtener de manera aproximada a partir de los valores extraídos del cálculo COSMO en MOPAC, multiplicando el valor de la SES por un factor de escalamiento. De esta forma queda determinada la superficie escalada excluyente del solvente (*Solvent Excluding Scaled Surface*, SSES),

$$SASA \simeq \sum_i A_{i,COSMO} \frac{(R_{solv} + r_i)^2}{(r_i)^2} \equiv SSES \tag{5.5}$$

donde $A_{i,COSMO}$ es el valor de la SES calculada por MOPAC, para el átomo i, R_{solv} es el radio del solvente y r_i el radio del átomo i.

Por otra parte, se utilizó la librería FreeSASA[119] para calcular el valor de SASA. En esta Tesis, las superficies SSES y SAS fueron usadas para calcular el término no polar de energía de solvatación.

5.1.1. Bases de datos

Como todo método que involucra la determinación de parámetros particulares, la calidad del modelo empleado depende fuertemente de la cantidad y diversidad del conjunto de moléculas utilizado en el proceso de optimización, denominado *conjunto de entrenamiento*. Esto requiere que el conjunto de moléculas que se utiliza para efectuar la parametrización, posea diversidad en cuanto a sus propiedades físico-químicas.

Hasta la actualidad, se han realizado entrenamientos de gran escala y muchas validaciones de modelos de solvente implícito con campos de fuerza clásicos, pero no se han llevado a cabo las validaciones necesarias para modelos que emplean métodos SQM. Por ejemplo, en el caso del modelo COSMO implementado en MOPAC, se sabe que el conjunto de parámetros empleados por defecto no es el más adecuado, habiéndose demostrado que se pueden encontrar mejores resultados al aumentar dichos radios o realizar una optimización de los mismos.[120]

A continuación, se mencionan los conjuntos de entrenamiento y validación elegidos para la re-parametrización del modelo COSMO.

Conjunto de entrenamiento

En esta Tesis se usaron dos conjuntos entrenamiento. Estos fueron compilados por el grupo de trabajo de Cavasotto et al.,[50] en el cual se desarrolla la presente Tesis, y por el grupo de trabajo del Profesor F. J. Luque.[121] El primer conjunto contiene energías de hidratación experimentales para 507 moléculas pequeñas neutras que poseen átomos de C, O, H, N, I, P, Br, F, S y Cl mientras que el de Luque et al. cuenta con 81 moléculas neutras con átomos de C, O, H, N, P, Br, F, S y Cl.

Conjunto de validación

Debido a que el interés del presente trabajo radica en la aplicación del método para la correcta descripción de interacciones proteína-ligando, el conjunto de validación contiene específicamente moléculas con propiedades similares a fármacos. El mismo fue publicado por la competencia SAMPL4,[122] perteneciente a la serie SAMPL, y está conformado

por 47 moléculas pequeñas neutras con valores experimentales conocidos de energías de hidratación.

La competencia SAMPL4 está caracterizada por ser una evaluación 'a ciegas', es decir que los datos experimentales de una propiedad en particular se dan a conocer a los participantes de la competencia luego de que hayan efectuado sus propias predicciones, a fin de evaluar la calidad de sus metodologías.[123–125] En este caso en particular, se realizaron predicciones de energías libres de hidratación determinadas por J. P. Guthrie.[126] Debido a que no se dispone de geometrías moleculares para este grupo de moléculas, fueron construidas y optimizadas con el Hamiltoniano RM1 en fase gaseosa, en el grupo de investigación en el que se desarrolla el presente trabajo de Tesis.

5.1.2. Evaluación estadística

Se efectuaron diferentes medidas estadísticas para comparar las predicciones teóricas con los valores experimentales. Dichas medidas son, la raíz cuadrada del coeficiente de correlación lineal (R) por medio del cual se desea observar cuál es la relación entre los valores teóricos y experimentales de energías de hidratación, el error absoluto promedio (*mean absolut error*, MAE) y el error raíz cuadrático medio (*root-mean-square error*, $RMSE$).

El MAE es el promedio sobre el conjunto de moléculas, de los valores absolutos de las diferencias entre la predicción y el valor experimental de energía de hidratación. Al ser una función de puntuación lineal, todas las diferencias individuales se ponderan por igual en el promedio. El $RMSE$ es una regla de puntuación cuadrática que mide la magnitud promedio del error. La diferencia entre la predicción y los correspondientes valores observados es elevada al cuadrado, y luego promediada sobre la muestra. Finalmente se toma la raíz cuadrada del promedio. Dado que los errores se elevan al cuadrado antes de ser promediados, el $RMSE$ da un peso relativamente alto a los errores grandes. Por esta razón, el $RMSE$ es una medida estadística útil cuando los errores grandes son particularmente indeseables. El $RMSE$ siempre será mayor o igual que el MAE, cuanto mayor sea la diferencia entre ellos, mayor será la variación en los errores individuales en la muestra. Si $RMSE = MAE$ todos los errores son de la misma magnitud. ΔG_i es la variación de energía libre entre las fases en solvente y en gas de cada molécula. El subíndice i recorre las diferentes moléculas del conjunto estudiado.

$$R = \frac{n\sum \Delta G_i(calc)\Delta G_i(exp) - \sum \Delta G_i(calc)\sum \Delta G_i(exp)}{\sqrt{n\sum \Delta G_i(calc)^2 - (\sum \Delta G_i(calc))^2}\sqrt{n\sum \Delta G_i(exp)^2 - (\sum \Delta G_i(exp))^2}}, \quad (5.6)$$

$$MAE = \frac{1}{N}\sum_{i}^{N}\Big[|\Delta G_i(exp) - \Delta G_i(calc)|\Big], \tag{5.7}$$

$$RMSE = \left(\frac{1}{N}\sum_{i}^{N}\Big[\Delta G_i(exp) - \Delta G_i(calc)\Big]^2\right)^{0,5}, \tag{5.8}$$

donde N indica el número de miembros del conjunto de moléculas.

5.1.3. Procedimientos de optimización

Para la optimización de parámetros se realizaron dos procedimientos. En una primera parte de la investigación, se recurrió a un método particular de los llamados algoritmos evolutivos, conocido como Algoritmo Genético (AG), propuesto por John Holland.[127] Los métodos SQM usados en la re-parametrización fueron RM1 y PM6. Para el término no polar se emplearon las metodologías γSASA y CP. La re-parametrización se llevó a cabo en el grupo de moléculas de Forti, y en el grupo de moléculas de Cavasotto. El conjunto de validación elegido fue, en ambos casos, SAMPL4,

Los resultados obtenidos con γSASA, a pesar de ser una metodología más simple que la propuesta por CP, condujo a resultados prometedores. Por tal motivo, se realizó una nueva optimización de parámetros atómicos para los Hamiltonianos RM1 y PM7, empleando en este caso el algoritmo de minimización local de Powell[128] y calculando la contribución no polar con los métodos γSASA y γ_{ef}. Los valores del área SES fuero extraídos de las salidas de los cálculos efectuados en MOPAC, y el área de la SAS fue calculada con la librería FreeSASA. Se usaron los conjuntos de entrenamiento y validación de Forti y SAMPL4, respectivamente.

Optimización con Algoritmo Genético

Los algoritmos evolutivos son métodos de búsqueda conformacional, que permiten efectuar una optimización de parámetros minimizando o maximizando una *función de adaptación*. En cada iteración, el algoritmo trabaja con una población de individuos, donde cada uno representa un punto en el espacio de soluciones para un problema en particular. Luego, el desempeño de cada individuo es evaluado por la función de adaptación, que ordena de mejor a peor, los miembros de la población. La población inicial evoluciona sucesivamente hacia mejores regiones del espacio de búsqueda mediante procesos proba-

bilísticos. Para encontrar los óptimos globales, los algoritmos de optimización hacen uso de dos técnicas: a) explorar áreas desconocidas en el espacio de búsqueda, y b) explotar el conocimiento obtenido de puntos previamente evaluados. En este trabajo de Tesis, se maximizó la función de adaptación, definida como la inversa del MAE (Ec. 5.7).

Un cromosoma, en analogía a la naturaleza, es una secuencia particular de genes. En nuestro caso, los genes son los parámetros a optimizar, es decir los radios y coeficientes de tensión superficial. Inicialmente, se generó una población de 100 individuos. Sus cromosomas, es decir los valores de los parámetros, fueron generados aleatoriamente entre ciertos límites. Se evaluó el desempeño de cada individuo, siendo los de menor MAE los que tienen mayor probabilidad de subsistir. Se seleccionó entonces la mejor mitad. Luego surge el entrecruzamiento entre individuos, donde se eligen dos de manera aleatoria y se entrecruzan sus genes también aleatoriamente, lo cual genera nuevos individuos. Este paso se repite hasta obtener nuevamente un número total de 100 individuos.

Para los radios atómicos, se tomaron como punto de partida los valores por defecto de MOPAC y los optimizados por Forti et al[121] separadamente. El rango de variación establecido para los mismos fue ±0,15 de su valor inicial. Para los coeficientes de tensión superficial, el valor inicial fue de -0,1 con un rango de variación entre 0 y -0,1, En la metodología γSASA se consideró un valor de 0,1 para el parámetro γ, con una posible variación entre 0 y 0,1, El rango de variación para el radio del solvente fue de ±0,2 de un valor inicial de 1,2, Fueron impuestas además las siguientes restricciones,

$$\begin{aligned} R(C) > R(N) > R(O) > R(F) \\ R(F) < R(Cl) < R(Br) \\ R(P) > R(N) \\ R(S) > R(O) \end{aligned}$$

Este proceso de selección fue repetido 200 veces. Debido a que la exploración que realiza el método de algoritmo genético sobre el espacio muestral de soluciones posibles es de forma aleatoria, a los individuos de la última generación se les aplicó luego una minimización local. Se utilizó para tal fin el método de Powell implementado en una librería de Python llamada $lmfit$. Se aplicaron 200 pasos de dicho método sobre cada uno de los individuos.

Finalmente, de este conjunto total de individuos se seleccionaron aquellos que presentaron mejor respuesta a la evaluación de las medidas estadísticas mencionadas.

Cabe mencionar que la optimización se efectuó sobre dos conjuntos de moléculas de

entrenamiento. Luego se evaluó la energía de hidratación sobre el conjunto de validación, con tres grupos de parámetros diferentes. Se reportarán en este trabajo, únicamente los resultados de aquel que proporcionó el menor MAE en la evaluación sobre dicho conjunto de prueba.

Tabla 5.1: R iniciales [§]

	MOPAC	Forti
H	1,30	1,00
C	2,00	1,88
N	1,83	1,88
O	1,72	1,75
P	2,11	2,25
S	2,16	2,19
F	1,72	1,69
Cl	2,05	2,25
Br	2,16	2,44
Rsolv	1,10	1,10

[§] Valores de R en Å.

Optimización con Algoritmo de Powell

En una segunda instancia del trabajo de investigación, se optimizaron los parámetros atómicos de modo que el MAE alcance el valor mínimo en el conjunto de entrenamiento, empleando el algoritmo de minimización local de Powell,

$$F(\mathbf{x}) = \sum_i [\Delta G^i_{exp} - \Delta G^i_{hidr}(\mathbf{x})]^2 + \alpha ||\mathbf{x} - \mathbf{x_0}||^2 \tag{5.9}$$

donde F es la función error y $\mathbf{x}$ es el vector de los parámetros a optimizar.

La optimización se realizó en tres ciclos sucesivos, usando el conjunto de parámetros optimizados al finalizar cada ciclo, como punto de partida para el siguiente. En cada iteración, fue modificado el conjunto de parámetros y evaluada la diferencia de energía entre los valores experimentales y los calculados. Se incluyó una restricción parabólica (α=0,1) para evitar que los parámetros se alejen del valor obtenido en la optimización del primer ciclo.[50]

El proceso de optimización fue repetido para los Hamiltonianos RM1 y PM7, conduciendo a dos conjuntos de parámetros específicos para cada método. Los radios atómicos y el parámetro γ fueron parametrizados simultáneamente en el conjunto total de moléculas de entrenamiento. Para los radios atómicos, se tomaron como punto de partida los valores por defecto de MOPAC y los optimizados por Forti et al.[121] separadamente (Tabla 5.1), mientras que el valor inicial elegido para el parámetro γ fue de 0,005 kcal/mol/Å^2, con b

inicial igual a 1 (se probó también un valor de $b = 0$, pero la re-parametrización fue mejor en el primer caso).

5.2. Análisis y Discusión de Resultados

Las energías libres de hidratación fueron calculadas para las 47 moléculas pertenecientes a la competencia SAMPL4, cuyas conformaciones iniciales fueron extraídas en su mayoría del conjunto de Mobley.[129] Sólo 5 estructuras iniciales se extrajeron del banco de datos Pubchem.[130] Una vez obtenidas dichas estructuras, se efectuó una optimización en fase gaseosa con los Hamiltonianos RM1 y PM6 que fueron empleadas posteriormente para efectuar el cálculo de energía libre de hidratación.

5.2.1. Re-parametrización con AG

Los análisis de resultados obtenidos mediante el primer tipo de procedimiento de optimización, se pueden dividir en tres etapas. En primer lugar, se realizó la comparación entre dos metodologías empleadas para el cálculo de la componente no polar de la energía de hidratación, usando en ambos casos el Hamiltoniano RM1, Seguidamente se efectuaron los cálculos con el Hamiltoniano PM6 en combinación con el método más preciso para el término no polar, según los resultados obtenidos con RM1, Por último, se analizó el impacto que tiene el tamaño del conjunto de entrenamiento empleado para la parametrización, en la precisión alcanzada por el método. Una vez efectuados dichos análisis, se compararon los resultados con las predicciones publicadas por los distintos grupos participantes de la competencia SAMPL4,[131]

Componente no polar

Se eligieron dos metodologías para calcular el término no polar, a fin de evaluar cuál arroja mayor precisión en el cálculo de energías libres de hidratación para el grupo de moléculas con valores experimentales conocidos. En primer lugar, se consideraron por separado los términos de energía de cavitación y de dispersión, usando la formulación de Claverie-Pierotti (Ec. 2.21). Luego, se efectuaron los cálculos siguiendo la formulación γ SASA (Ec. 2.17).

Tabla 5.2: Medidas estadísticas de diferentes estimaciones de energías libres de hidratación para el conjunto de 47 moléculas de SAMPL4[§]

Metodologías	**RMSE**	**MAE**	**R**
Claverie-Pierotti	1,75	1,40	0,91
γSASA	1,89	1,46	0,89

[§] Valores de MAE y RMSE calculados en kcal/mol.

Como se puede observar en la Tabla 5.14, el valor más bajo del MAE (1,40 kcal/mol) se obtuvo con la formulación CP. Este resultado, fue muy cercano al obtenido con γSASA (1,46 kcal/mol). Por otro lado, el valor del MAE es muy cercano al del RMSE en ambos métodos, lo que estaría indicando que gran parte de los errores individuales son de la misma magnitud. La diferencia es menor para la metodología Claverie-Pierotti.

Comparación entre Hamiltonianos RM1 y PM6

Habiendo comparado dos métodos para calcular la componente no polar de la energía libre de hidratación, se observó que el desarrollado por Claverie-Pierotti permite una mejor predicción de la energía libre de hidratación para el grupo de moléculas de SAMPL4, Se eligió entonces dicha metodología para el término no polar, usando ahora el Hamiltoniano PM6 para calcular la componente electrostática a la energía de hidratación del mismo conjunto de moléculas. Los resultados obtenidos se presentan en la Tabla 5.3.

Tabla 5.3: Medidas estadísticas de predicciones de energía libre de hidratación para el conjunto de moléculas de SAMPL4 utilizando los Hamiltonianos RM1 y PM6.[§]

Hamiltoniano	**RMSE**	**MAE**	**R**
RM1	1,75	1,40	0,91
PM6	2,11	1,72	0,86

[§] Valores de MAE y RMSE calculados en kcal/mol.

Las predicciones de ambas metodologías son comparables entre sí, lo cual era esperable dadas las características comunes de los mismos. Sin embargo, se puede observar que RM1 es levemente superior a PM6 en cuanto a la precisión alcanzada (ver Tabla 5.3). El MAE de RM1 es de 1,40 kcal/mol frente a un valor de 1,72 kcal/mol para PM6. Del mismo modo, el error RMSE otorga ventaja al primer Hamiltoniano. En lo que respecta al coeficiente de correlación, se puede afirmar que la regresión lineal tiene un mejor ajuste para RM1 en comparación con PM6.

Dado que ambos Hamiltonianos fueron empleados en este trabajo para reproducir una propiedad con la cual no han sido parametrizados, se puede observar que PM6 no

presenta una mejora con respecto a RM1 para este caso particular, contrariamente a lo que ocurre para otras propiedades como ser momentos dipolares y longitudes de enlace.[29]

Influencia del conjunto de entrenamiento sobre la precisión de la metodología

En la presente Sección se presentan los resultados comparativos del impacto que tiene el número de moléculas del conjunto de entrenamiento, y su diversidad, en la precisión de una metodología de parametrización de un modelo. Se compararon las predicciones realizadas para las energías de hidratación de las moléculas del conjunto de validación, SAMPL4, usando dos conjuntos de entrenamiento diferentes.

Tabla 5.4: Medidas estadísticas de diferentes estimaciones de energías libres de hidratación evaluadas sobre dos conjuntos de entrenamiendo diferentes, siguiendo la formulación de Claverie-Pierotti para la contribución no electrostática[§]

Conjunto	**RMSE**	**MAE**	**R**
Forti	1,75	1,40	0,91
Cavasotto	1,70	1,30	0,91

[§] Valores de MAE y RMSE calculados en kcal/mol.

En la Tabla 5.4, se puede observar que para el conjunto de parámetros optimizados sobre el conjunto de entrenamiento con un número mayor de moléculas (Cavasotto), se alcanza un valor de MAE más bajo que al emplear un conjunto más reducido de moléculas. Aún si la diferencia entre las medidas estadísticas obtenidas con ambos conjuntos de entrenamiento no es significativa, estos resultados dan un indicio de que el primer conjunto de entrenamiento es más adecuado para realizar las optimizaciones correspondientes para el modelo y alcanzar una mayor precisión.

Parámetros optimizados: análisis

En la Tabla 5.5 se presentan los resultados de los parámetros optimizados en el conjunto de entrenamiento de moléculas de Forti, con los Hamiltonianos RM1 empleando las formulaciones CP y γSASA para la componente no polar, y el Hamiltoniano PM6 con el cálculo de CP. Las diferencias en valores de radios atómicos para los dos Hamiltonianos, son pequeñas pero suficientes como para realizar un impacto de ciertas kcal/mol en las energías de hidratación. Se puede observar, que el radio atómico del Br es el más dependiente del Hamiltoniano, seguido por los radios del O y F. Cabe destacar que únicamente el radio del C presenta el mismo valor para ambos Hamiltonianos.

Tabla 5.5: Parámetros atómicos optimizados.[§]

Parámetros	RM1-CP	RM1-γSASA	PM6-CP
γ(Br)	-0,026	0,008	-0,035
γ(C)	-0,028	0,005	-0,025
γ(Cl)	-0,024	0,012	-0,028
γ(F)	-0,003	0,071	-0,008
γ(H)	-0,137	-0,099	-0,121
γ(N)	-0,027	0,056	-0,017
γ(O)	-0,001	0,104	-0,006
γ(P)	-0,006	0,053	0,005
γ(S)	-0,027	0,009	-0,012
R(Br)	2,582	2,644	2,661
R(C)	2,030	2,030	2,030
R(Cl)	2,406	2,549	2,457
R(F)	1,706	1,623	1,651
R(H)	0,881	0,775	0,882
R(N)	1,890	1,749	1,882
R(O)	1,706	1,623	1,782
R(P)	2,246	2,181	2,286
R(S)	2,483	2,489	2,453

[§] Valores de R en Å y γ en kcal/mol Å^2.

Tabla 5.6: Parámetros atómicos optimizados con el Hamiltoniano RM1 sobre dos conjuntos de entrenamiento diferentes.[§]

Parámetros	RM1-Forti	RM1-Cavasotto
γ(Br)	-0,026	-0,015
γ(C)	-0,028	-0,019
γ(Cl)	-0,024	-0,016
γ(F)	-0,004	0,004
γ(H)	-0,137	-0,256
γ(N)	-0,027	-0,012
γ(O)	-0,001	-0,007
γ(P)	-0,006	0,046
γ(S)	-0,027	-0,013
R(Br)	2,582	2,520
R(C)	2,030	2,030
R(Cl)	2,406	2,463

Tabla 5.6: Parámetros atómicos optimizados con el Hamiltoniano RM1 sobre dos conjuntos de entrenamiento diferentes.[§]

Parámetros	RM1-Forti	RM1-Cavasotto
R(F)	1,706	1,750
R(H)	0,882	0,857
R(N)	1,890	1,835
R(O)	1,706	1,750
R(P)	2,246	2,100
R(S)	2,483	2,240

[§] Valores de R en Å y γ en kcal/mol Å^2.

Tabla 5.7: Medidas estadísticas de las mejores estimaciones de energías libres de hidratación para el conjunto de moléculas de SAMPL4

Metodologías	**RMSE**	**MAE**	**R**	**ref**
QM+solv imp	1,23	0,87	0,98	[132]
PB-single conf	1,23	0,94	0,95	[133],[134]
AM1-BCC+GAFF	1,22	1,00	0,96	[135],[136]
AM1-BCC+GAFF	1,26	1,00	0,95	[137]
COSMO-CP-CCRM1	1,70	1,30	0,91	
COSMO-CP-FPM6	2,11	1,72	0,86	

[§] Valores de error en kcal/mol.

En la Tabla 5.7 se presentan los cuatro métodos con mejores medidas estadísticas de todos los participantes en la competencia SAMPL4, junto con dos de los presentados en este trabajo. El primero de ellos realiza un tratamiento mecano-cuántico para el soluto (QM) y recurre a la representación del solvente mediante un modelo de solvente continuo. La energía de interacción soluto-solvente también es calculada como la suma de una componente electrostática más los términos correspondientes a dispersión y repulsión, sin embargo el cálculo se efectúa recurriendo a la Teoría de los Funcionales de la Densidad (DFT) y al potencial de Lennard-Jones. La energía de cavitación se tiene en cuenta por medio de la expresión de Pierotti[44]. El solvente se describe por medio de un modelo conocido como Campo de Fuerzas de Simulación del Líquido (LSFF, por sus siglas iniciales en inglés) con un conjunto de parámetros independientes del entorno atómico.

5.2.2. Re-parametrización con Powell

Se presentan a continuación los resultados obtenidos para la re-parametrización con el algoritmo de minimización local, con los Hamiltonianos RM1 y PM7, empleando dos

metodologías diferentes para la contribución no polar.

En primer lugar se aplicó la formulación γ_{ef}SASA (Ec. 5.2). Se usaron separadamente los valores de SSES y SASA para calular el área superficial accesible al solvente.

RM1/γ_{ef}SSES: El modelo COSMO fue re-parametrizado incluyendo el término correspondiente a la contribución no polar. El valor de esta componente fue calculado en primer lugar con la Ec. 5.3. Se tomó el valor calculado de SSES, con el valor inicial $b = 0$ para la constante. Esta metodología se identifica en adelante con la denominación γ_{ef}SSES.

Fueron tomados como radios iniciales, por un lado los usados por defecto en MOPAC para COSMO, y por otro los radios optimizados de Forti, comparando los resultados de ambas parametrizaciones. En la Tabla 5.8 se encuentran los resultados de los paráme-

Tabla 5.8: Conjunto de valores optimizados con el método γSSES ($b = 0$) para el Hamiltoniano RM1, tomando como punto de partida los radios de MOPAC (γSSES$_M$) y por otro lado los radios optimizados de Forti (γSSES$_F$).[§]

Parámetros	γ_{ef}SSES$_M$	γ_{ef}SSES$_F$
γ	0,0025	0,0046
b	0,381	0,186
H	1,002	0,900
C	1,983	1,875
N	1,705	1,733
O	1,740	1,762
P	2,379	2,407
S	2,317	2,246
F	2,020	1,899
Cl	2,350	2,395
Br	2,283	2,431

[§] Valores de R en Å.

tros optimizados, tomando como valores iniciales los radios de MOPAC (γ_{ef}SSES$_M$) y los radios de Forti (γ_{ef}SSES$_F$), separadamente. Se puede observar que el conjunto de parámetros iniciales, influye notablemente sobre los resultados finales de los parámetros optimizados.

Tabla 5.9: Medidas estadísticas de las estimaciones de energías libres de hidratación para el conjunto de entrenamiento de Forti, con la metodología γ_{ef}SSES ($b = 0$), usando dos conjuntos diferentes de radios iniciales.

Radios iniciales	**RMSE**	**MAE**	**R^2**
Mopac	1,28	0,93	0,74
Forti	1,16	0,87	0,79

[§] Valores de MAE y RMSE calculados en kcal/mol.

Los valores obtenidos para las distintas medidas estadísticas sobre el conjunto de entrenamiento con los parámetros optimizados se presentan en la Tabla 5.9. Como se puede observar, el MAE alcanzado partiendo de los radios optimizados de Forti es menor al encontrado si se usan los parámetros por defecto de MOPAC. La diferencia entre MAE y RMSE es menor para el primer caso por lo que los errores de las moléculas individuales son similares en magnitud en dicho caso. Dado que los parámetros optimizados partiendo de los radios de Forti arrojan un MAE menor, se tomaron dichos valores como punto de partida para los estudios realizados posteriormente.

Se empleó entonces nuevamente la metodología γ_{ef}SSES y el conjunto de entrenamiento de Forti, variando el parámetro inicial b, tomándolo igual a 0 o 1 como punto de partida en cada caso. La parametrización fue realizada nuevamente con el Hamiltoniano RM1,

Tabla 5.10: Conjunto de radios tomados de Forti, y valores optimizados con el método γSSES (con $b = 1$) para el Hamiltoniano RM1,[§]

Parámetros	Forti	RM1/γSSES
γ	0,0050	0,0013
b	1,00	1,1637
H	1,00	0,899
C	1,88	1,956
N	1,88	1,710
O	1,75	1,755
P	2,25	2,440
S	2,19	2,238
F	1,69	1,987
Cl	2,25	2,503
Br	2,44	2,342

[§] Valores de R en Å.

Tabla 5.11: Medidas estadísticas de diferentes estimaciones de energías libres de hidratación para el conjunto de entrenamiento de Forti, y tomando como radios iniciales los optimizados por Forti. Para el término no polar se empleó el valor de SSES.

b inicial	**RMSE**	**MAE**	**R^2**
0	1,16	0,87	0,79
1	1,07	0,78	0,82

[§] Valores de MAE y RMSE calculados en kcal/mol.

Los resultados se encuentran presentados en la Tabla 5.10, junto a los valores iniciales de parámetros de Forti. Como se puede observar en la Tabla 5.11, el valor de MAE se reduce considerablemente de 0,87 a 0,78 al emplear un valor inicial de 1 para el parámetro b.

RM1/γSASA: Comparación de SSES y SASA Al comparar los resultados obtenidos empleando diferentes parámetros iniciales, se observó que los parámetros optimizados usando los radios iniciales tomados de Forti conducen a una mayor precisión en la determinación de energías libres de hidratación.

Se efectuaron entonces los cálculos con el valor de SASA en lugar de SSES, para la contribución no polar, continuando con el Hamiltoniano RM1, Se compararon nuevamente los resultados de tomar $b = 0$ o $b = 1$ inicialmente.

Tabla 5.12: Medidas estadísticas de diferentes estimaciones de energías libres de hidratación para el conjunto de entrenamiento de Forti, tomando como radios iniciales los optimizados por Forti y la SASA.

Metodología	**b inicial**	**RMSE**	**MAE**	**R^2**
γ_{ef}SASA	0	1,12	0,85	0,80
	1	1,06	0,77	0,82

§ Valores de MAE y RMSE calculados en kcal/mol.

Observando los resultados presentados en la Tabla 5.12 se puede afirmar que, en términos generales, para el Hamiltoniano RM1 el mejor conjunto de parámetros corresponde a los optimizados con un valor inicial de b=1 y el valor del área superficial accesible al solvente. El valor del error absoluto promedio (MAE=0,77) indica sobre este conjunto de moléculas que los parámetros alcanzan una buena precisión en la determinación de energías libres de hidratación.

Comparación entre Hamiltonianos RM1 y PM7

Contribución no polar: γ_{ef} y γSASA

Habiendo comparado la metodología γ_{ef}SASA con γ_{ef}SSES para calcular la contribución no electrostática sobre un mismo conjunto de moléculas, se observó que la primera permite una mejor predicción de la energía libre de hidratación. Se efectuaron entonces los cálculos de esta propiedad con dicho método recurriendo a los Hamiltonianos RM1 y PM7. Los resultados de los optimizados con RM1 y PM7 usando γ_{ef}SASA, tomando como punto de partida los valores optimizados de Forti, se presentan más adelante en la Tabla 5.16.

Finalmente, se efectuó la re-parametrización empleando la metodología γSASA, esto es, con distintos coeficientes de tensión superficial para cada átomo. Se emplearon los Hamiltonianos RM1 y PM7 en el conjunto de entrenamiento de Forti.

Tabla 5.13: Parámetros usados en COSMO para su implementación con los parámetros optimizados por Forti como conjunto inicial, junto con los parámetros optimizados para los Hamiltonianos RM1 y PM7.[§]

Parámetros	Forti	RM1/γ_{ef}SASA	PM7/γ_{ef}SASA
γ	0,005	0,002	0,006
b	1,00	1,097	1,389
H	1,00	0,836	1,099
C	1,88	1,967	1,906
N	1,88	1,720	1,854
O	1,75	1,778	1,818
P	2,25	2,422	2,542
S	2,19	2,242	2,489
F	1,69	1,968	1,953
Cl	2,25	2,396	2,201
Br	2,44	2,476	2,227

[§] Valores de R en Å.

A partir de un análisis de los resultados presentados en la Tabla 5.16 se puede ver que la metodología que más se ajusta a los resultados experimentales del conjunto de validación en este caso, es el del Hamiltoniano PM7 con la metodología γ_{ef}SASA para el término no polar. Comparando la calidad de los resultados obtenidos con ambos Hamiltonianos se puede concluir que una mejor predicción es alcanzada por PM7 frente a RM1,

Tabla 5.14: Medidas estadísticas de diferentes estimaciones de energías libres de hidratación para el conjunto de entrenamiento de Forti, con la metodología γSASA, y los Hamiltonianos RM1 y PM7.

Hamiltoniano	**RMSE**	**MAE**	$\mathbf{R^2}$
RM1	1,04	0,76	0,83
PM7	0,98	0,75	0,92

[§] Valores de MAE y RMSE calculados en kcal/mol.

Tabla 5.15: Parámetros atómicos optimizados con el método γSASA para los Hamiltonianos RM1 y PM7, tomando como punto de partida los radios optimizados por Forti.[§]

Parámetros	RM1/γSASA	PM7/γSASA
R(H)	0,9362	0,9999
R(C)	1,8665	1,8119
R(N)	1,7590	1,8944
R(O)	1,7725	1,7778
R(F)	1,6544	1,7024
R(P)	2,2295	2,2499
R(S)	2,1933	2,2900
R(Cl)	2,3258	2,3042
R(Br)	2,4950	2,5050
γ(H)	-0,0088	0,0130
γ(C)	0,0101	0,0166
γ(N)	0,0085	0,0151
γ(O)	0,0102	0,0255
γ(F)	0,0290	0,0282
γ(P)	0,0212	0,0763
γ(S)	0,0138	0,0445
γ(Cl)	0,0079	0,0074
γ(Br)	0,0027	0,0001

[§] Valores de R en Å y γ en kcal/mol Å^2.

Tabla 5.16: Error absoluto promedio para energías libres de solvatación, (MAE con respecto a valores experimentales) en el conjunto de entrenamiento de Forti et al., y en el conjunto de validación de SAMPL4,[§]

Metodología	**Forti**	**SAMPL4**
RM1/γ_{ef}SASA	0,78	1,42
PM7/γ_{ef}SASA	0,83	1,30
RM1/γSASA	0,76	1,50
PM7/γSASA	0,75	1,36

[§] Valores de MAE y RMSE calculados en kcal/mol.

Capítulo 6

Energía Libre y Función de Scoring con Métodos Cuánticos

Tanto el desarrollo como la aplicación de modelos para predecir la energía libre de unión de manera precisa y eficaz representan una tarea ardua en el área del diseño racional de fármacos asistido por computadoras.[138]

La necesidad de contar con nuevas metodologías que permitan una mejor comprensión de la asociación molecular sigue siendo un desafío actual. En el año 2011, en línea con los estudios realizados hasta el momento, se desarrolló una nueva metodología para predecir la energía libre de unión en sistemas biomoleculares. El trabajo, elaborado en el contexto de la investigación que dio origen a esta Tesis, se encuentra publicado en la Ref. 56. La metodología desarrollada, MM/QM-COSMO, surge como una aproximación alternativa a los métodos MM-PB(GB)SA existentes, introduciendo una re-evaluación cuántica de energía sobre geometrías extraídas de una dinámica molecular, minimizadas con métodos SQM. Su aplicación en el estudio de complejos proteína-péptido con valores experimentales de energía libre de unión demuestra el potencial del método para predecir energías libres de unión relativas, en contraposición a los métodos basados en campos de fuerza.

En esta Tesis, se tomó como punto de partida la metodología MM/QM-COSMO teniendo en cuenta el balance adecuado entre precisión y eficiencia computacional alcanzados por la misma. Se adaptó la metodología para ser aplicada en un trabajo de investigación, realizado en colaboración con grupos experimentales, para la identificación de nuevos inhibidores contra el virus del Dengue. Los resultados de dicho trabajo se presentan más adelante, en el Capítulo **??**. La implementación fue llevada a cabo con el objetivo de determinar la conformación de menor energía entre dos poses de *docking*, con cálculos cuánticos de energía libre, teniendo en cuenta que son de validez general y mayor precisión que los realizados con métodos de MM. En este Capítulo se presentan los

aspectos más relevantes de la metodología MM/QM-COSMO, de utilidad para esta Tesis.

6.1. Energía libre de unión en MM/QM-COSMO

Los métodos de puntos extremos representan una alternativa más atractiva a los métodos computacionalmente intensivos (FEP, TI) y a las aproximaciones rápidas pero poco precisas para estimar energías libres de unión como *Docking* Molecular, ya que se han obtenido resultados prometedores a un costo computacional mucho menor. Además, dichos métodos están fundados en la mecánica estadística y tienen en cuenta la flexibilidad conformacional, lo que ha inspirado a ciertos autores a usar funciones basadas en mecánica clásica en combinación con el modelo de solvente PBSA, para la determinación de energías libres de unión. Tanto MM-PBSA como MM-GBSA combinan simulaciones de MD en solvente explícito, con una re-evaluación de las trayectorias utilizando un modelo de solvente continuo, para calcular energías de unión promediadas para las configuraciones extraídas de las dinámicas de los estados unido y no unido.[139]

Una limitación importante de este tipo de métodos en el cálculo de energías libres de unión, es que se ignora el efecto de la transferencia de carga, debido al campo de fuerzas de mecánica molecular empleado. Esto conduce a un balance inadecuado entre la interacción electrostática intermolecular y la contribución electrostática a la energía de solvatación. Una mejora de la descripción de la función de energía potencial podría proveer un mejor balance de las interacciones electrostáticas. Los métodos de mecánica cuántica son sistemáticamente mejorables y por ser de validez general no requieren una parametrización dependiente del sistema, por lo que pueden ser implementados con el propósito de alcanzar dicho objetivo.

El método de puntos extremos basado en una formulación cuántica MM/QM-COSMO toma en consideración las premisas antes mencionadas para calcular energías libres de unión en sistemas biomoleculares, agregando además una mejor descripción para los cambios entrópicos. La deducción del método se encuentra detallada en la referencia [56], por lo que en la presente Sección solamente se abordarán los aspectos fundamentales, necesarios para la obtención de las componentes de energía libre de unión.

De acuerdo a la Ec. 3.13, la energía libre de unión de un complejo proteína-ligando se puede calcular como,

$$\Delta G^{PL} = \Delta\langle U\rangle + \Delta G^{solv} - T\Delta S^{\mathrm{RB}} - T\Delta S^{\mathrm{int}} \tag{6.1}$$

donde $\langle ...\rangle$ representa un promedio sobre el conjunto estadístico, ΔG^{solv} es el cambio en energía libre de solvatación, y el cambio de entropía se separa en un término que

representa la asociación de cuerpo rígido ΔS^{RB} y otro término asociado a los grados de libertad internos S^{int}, dados por

$$\Delta S^{\mathrm{RB}} = \frac{1}{T}\langle E_{PL}\rangle + R\ln\Big(\frac{C^{\circ}}{8\pi^2} z^{\mathrm{RB}}\Big) \tag{6.2}$$

$$z^{\mathrm{RB}} = \int e^{-\beta E_{PL}(\zeta)} d\zeta \tag{6.3}$$

$$S^{\mathrm{int}} = -R \int J(\mathbf{q}) p(\mathbf{q}) \ln p(\mathbf{q}) d\mathbf{q} \tag{6.4}$$

6.1.1. Metodología

Estructuras para el cálculo de energía MM/QM-COSMO

El método MM/QM-COSMO toma como punto de partida para la evaluación de la energía libre de cada componente del sistema, distintas configuraciones alcanzadas por el mismo durante una simulación de MD. Se recurre a la aproximación de trayectoria única descripta en el Capítulo 3 y se extraen las estructuras (a las que en adelante denominaremos *configuraciones*) a intervalos regulares de tiempo a fin de asegurar que sean no correlacionados. A continuación, es necesario efectuar una minimización de energía cuántica sobre dichas geometrías, que fueron generadas por campos de fuerza de mecánica clásica. Esto se debe a la diferencia existente entre las hipersuperficies de energía clásica y cuántica.[18] El número de ciclos de optimización con métodos semiempíricos realizados sobre cada estructura dependerá del poder de cómputo del que se dispone y la consideración de haber alcanzado una convergencia adecuada. Finalmente se calcula el calor de formación del complejo, del ligando libre y de la proteína libre a partir de la estructura del complejo proteína-ligando minimizada extrayendo por separado las estructuras de las componentes (P y L) y realizando un cálculo de energía sobre las mismas. En analogía con los cálculos clásico de MM/PBSA la energía libre de unión en solución se calcula como (Ec. 3.13):

$$\Delta G_{u,sol} \;=\; \Delta\langle G^{COSMO}\rangle - T\Delta S^{RB} - T\Delta S^{int} \tag{6.5}$$

donde $\langle ...\rangle$ representa el promedio sobre el conjunto de configuraciones extraídas de la simulación de MD.

Finalmente la energía de unión del complejo se obtiene promediando los valores de ΔG en cada una de las estructuras minimizadas, para cada componente (PL, P, L).

Determinación de G^{COSMO}

Como la dimensión de los sistemas biomoleculares es de varios miles de átomos, los métodos cuánticos *ab initio* son computacionalmente muy costosos para ser aplicados al sistema en su totalidad. Por ello MM/QM-COSMO emplea Hamiltonianos SQM para el cálculo de energía combinados con el modelo de solvente continuo COSMO, presentado en el Capítulo 2. La energía libre en solución para cada una de las componentes de un complejo proteína-ligando, se puede calcular mediante un ciclo termodinámico (Fig. 3.1) como:[140]

$$G(X)_{sol} = G(X)_{gas} + \Delta G(X)_{solv} \tag{6.6}$$

donde X = PL, P o L, y $G(X)_{gas}$ incluye la energía electrónica del sistema en fase gas a 0 K y las correcciones termo estadísticas (contribuciones de vibración, rotación y traslación a 298 K), que pueden ser obtenidas mediante un modelo de oscilador armónico de rotor rígido (RRHO), además de contener también la energía de punto cero.

Para cálculos semiempíricos, se define el calor de formación $\Delta H(X)_{f,298}$ como la energía requerida para formar un mol del sistema en fase gaseosa a 298 K, a partir de sus elementos en el estado estándar. Por otra parte, están parametrizados usando datos de referencia de sistemas químicos en su estado estándar a 298 K, por lo que tanto la energía de punto cero como los términos de corrección de energía internos están incluidos en el calor de formación, a través de los parámetros. Sustituyendo entonces $G(X)_{sol} \rightarrow \Delta H(X)_{f,298}$ en la Ec. 6.6, la energía libre de Gibbs de la componente X será:[140]

$$G(X)_{sol} = G(X)_{gas} + \Delta G(X)_{solv} = \Delta H(X)_{f,298} \tag{6.7}$$

donde $G(X)_{sol}$ es la energía libre en solución de la componente X y $\Delta H_{f,298}(X)$ corresponde al calor de formación a 298 K en solvente.

En este caso, los términos de energía correspondientes a la solvatación están incluidos en los términos de energía electrónica y nuclear para el sistema solvatado. Si se usa el modelo COSMO para la descripción del solvente, se debe agregar a la Ec. 6.7 el término de energía de solvatación no polar $E(X)^{\mathrm{np}}_{solv}$, debido a que éste modelo tiene en cuenta únicamente la componente electrostática de la energía de solvatación. De esta manera resulta,

$$G(X)^{COSMO}_{sol} = E(X)^{COSMO} + E(X)^{\mathrm{np}}_{solv} \tag{6.8}$$

donde $\mathrm{G(X)}^{\mathrm{COSMO}}_{sol}$ representa la energía libre de la componente X (donde X = P, L o PL) calculada con métodos SQM, que incluye tanto la energía en vacío ($G(X)_{gas}$) como la energía libre de solvatación ($\Delta G(X)_{solv}$) y la componente entrópica del solvente. El valor de $E(X)^{COSMO}$ es determinado por medio del cálculo del calor de formación (Ec. 6.7).

Reemplazando G_{sol}^{COSMO} (Ec. 6.8) para cada componente (PL, P y L) en la Ec. 6.5 y considerando la aproximación γSASA para el término de energía de solvatación no polar $E(X)_{solv}^{\mathrm{np}}$ (Ec. 2.17), se obtiene,

$$\begin{aligned} \Delta G_{u,sol} &= E(PL)^{COSMO} - E(P)^{COSMO} - E(L)^{COSMO} \\ &+ \gamma\mathrm{SASA}(PL) - \gamma\mathrm{SASA}(P) - \gamma\mathrm{SASA}(L) - T\Delta S^{RB} - T\Delta S^{int} \end{aligned} \tag{6.9}$$

donde ΔS^{RB} y ΔS^{int} corresponden a la variación de entropía del sistema. Para simplificar notación se eliminó el subíndice *sol* que indica el cálculo sobre una estructura en solución, asumiendo en adelante que siempre se cumple esta condición.

La Ec. 6.9 ha sido aplicada originalmente para el cálculo de energía libre de unión de un sistema proteína-ligando. La correlación obtenida con valores experimentales fue mejor a la encontrada con el método MM-PBSA. La metodología MM/QM-COSMO puede ser implementada también en la etapa de re-*scoring*, luego de un proceso de CV, con el objetivo de reducir el número de falsos positivos. Dado que permite determinar la afinidad de unión con mayor precisión que los métodos clásicos, también puede ser empleada para identificar la pose de menor energía de una molécula que ha sido sometida a una evaluación de molecular *docking*.

Por lo general, los sistemas biomoleculares comprenden miles de átomos, por lo que se deben aplicar distintas aproximaciones para obtener el calor de formación de cada componente con un adecuado requerimiento computacional y un menor tiempo. Según sea el objetivo de la implementación de la metodología, se buscará la alternativa más adecuada considerando el compromiso entre tiempo y precisión.

6.2. Función de Scoring cuántica

Las interacciones no covalentes originadas en la formación de un complejo proteína-ligando tienen un papel preponderante en la determinación de su estructura y sus propiedades. Si éstas no se describen correctamente el valor de energía libre de unión empleado para aproximar el *score* en la metodología de *Molecular Docking* éste no será adecuado para distinguir los ligandos de los no ligandos dentro de un conjunto de moléculas pequeñas, con suficiente precisión.

En el Capítulo 4 se describieron brevemente los distintos tipos de funciones de *scoring* desarrolladas hasta el momento. La principal limitación de las mismas tiene que ver con una descripción aproximada de las interacciones no covalentes involucradas en la formación

del complejo proteína-ligando. Además de las numerosas maneras de aproximar la energía libre de unión mediante funciones de *scoring* clásicas, el trabajo de Merz fue pionero al incorporar una aproximación basada en la mecánica cuántica para calcular afinidad de unión de ligandos y metaloenzymas con una precisión razonable.[15] La principal ventaja de los métodos cuánticos es que incorporan de manera directa efectos que no pueden ser correctamente modelados por los métodos clásicos y que son de vital importancia en la determinación de las estructuras y propiedades de los complejos proteína-ligando, como por ejemplo los efectos de transferencia de carga y polarización. En este trabajo de Tesis se desarrolló una nueva función de *scoring* cuántica, para ser aplicada en un contexto de Cribado Virtual basado en la estructura del receptor.

Empleando el método MM/QM-COSMO, la energía libre de unión (ΔG_u) de un complejo proteína-ligando (PL) se puede expresar como

$$\Delta G_u = \Delta\langle G^{COSMO}\rangle - T\Delta S \tag{6.10}$$

donde la diferencia en el primer término se calcula entre los estados unido (PL) y no unido (P, L), $\langle ...\rangle$ representa el promedio sobre trayectorias de dinámica molecular clásicas (DM) minimizadas con métodos cuánticos y G^{COSMO} es la energía calculada con métodos cuánticos, que incluye el término de solvente continuo. El segundo término representa el cambio de entropía de la proteína y ligando durante la formación del complejo.

Debido a que (ΔG_u) en la Ec. 6.10 es muy costosa para ser usada en la asignación de un puntaje o *score*, y conformación del *ranking* de grandes librerías químicas de moléculas pequeñas en HTD, se definió una función de *scoring* de mecánica cuántica (SSC) como una aproximación a la Ec. 6.10, de modo que

$$SSC = \Delta G^{COSMO} - T\Delta S \tag{6.11}$$

donde el promedio sobre las trayectorias de DM fueron reemplazadas por un cálculo QM autoconsistente sobre la estructura PL resultante del *docking*, y las estructuras libres no unidas de P y L.

La función SSC se puede reescribir desglosando el primer término, de manera tal que las contribuciones de la deformación de proteína y ligando queden expresadas de manera explícita, al igual que la energía de interacción entre ambas partes una vez formado el complejo,

$$SSC = \Delta E_{int} + \Delta G_{def}(P) + \Delta G_{def}(L) + \Delta G_{np} - T\Delta S \tag{6.12}$$

donde ΔE_{int} es la energía de interacción en fase solución de proteína-molécula pequeña, $\Delta G_{def}(P)$ y $\Delta G_{def}(L)$ son las contribuciones provenientes de la deformación de la proteína y la molécula pequeña, respectivamente, al pasar de la conformación libre en solución a la conformación unida en el complejo.

En nuestra función de puntuación o *score*, la energía de interacción (ΔE_{int}) viene dada por la diferencia de energía entre los estados unido (PL) y no unido (P, L) calculadas con un hamiltoniano semiempírico sobre las geometrías resultantes de un proceso de *docking*. El solvente se representa con el modelo de solvente continuo COSMO, de manera que la energía de solvatación está directamente incluida en el cálculo del calor de formación. Este término representa usualmente la mayor contribución al *score*. Se utilizaron en el presente trabajo los métodos semiempíricos PM6-D3H4 y PM7 para calcular la componente electrostática.[30,31] La contribución no polar de energía de solvatación, fue calculada con la aproximación γSASA con un único parámetro efectivo de tensión superficial (Ec. 2.18).

Los términos $\Delta G_{def}(P)$ y $\Delta G_{def}(L)$ dan cuenta de los cambios conformacionales que sufren la proteína y el ligando durante el proceso de unión. $\Delta G_{def}(X)$ puede estimarse mediante una minimización local o por medio de un método de búsqueda conformacional más robusto. La minimización es rápida, sin embargo conduce al mínimo local más cercano. Por otro lado los métodos de búsqueda conformacional, como simulaciones de Dinámica Molecular o Monte Carlo permiten explorar la hipersuperficie de energía permitiendo de esta manera alcanzar una mayor probabilidad de encontrar la conformación de menor energía y considerar con mayor precisión el término de deformación. Sin embargo, dado que el protocolo desarrollado tiene por objetivo su implementación en un contexto de Cribado Virtual, debe ser suficientemente rápido ya que requiere la evaluación de energía de miles a millones de compuestos.

La contribución no polar de energía de solvatación, fue calculada con la aproximación γSASA con un único parámetro efectivo de tensión superficial (Ec. 2.18). Por último, el término entrópico ($T\Delta S$) da cuenta de la entropía configuracional del sistema, que incluye tanto la entropía conformacional debida a los grados de libertad rotacionales y translacionales y la entropía vibracional, relacionada con los modos normales de vibración de la molécula. En nuestra función de *score* empleamos para estimar ΔS la aproximación de enlaces rotables. En este caso, los cambios de entropía se originan por los grupos químicos que tienen libertad de movimiento en el estado inicial, es decir cuando proteína y ligando no están unidos, y están rígidos en el complejo. Por otra parte, el análisis vibracional consiste en el cálculo de las frecuencias vibracionales de las moléculas no unidas y el complejo. Luego cada una las vibraciones es trasladada a contribuciones a la energía libre de unión por medio de la termodinámica estadística. Debido a que el cálculo

de este término requiere una demanda computacional muy elevada con respecto a los demás, utilizar esta aproximación resulta poco práctico.

Los términos de interacción y deformación se calculan como,

$$\Delta E_{int} = E(PL)^{COSMO}_{PL_{min}} - E(P)^{COSMO}_{PL_{min}} - E(L)^{COSMO}_{PL_{min}} \tag{6.13}$$

y

$$\Delta G_{def}(X) = E(X)^{COSMO}_{PL_{min}} - E(X)^{COSMO}_{X_{min}}, \tag{6.14}$$

donde X = P o L. El subíndice indica sobre qué estructura se calcula la energía.

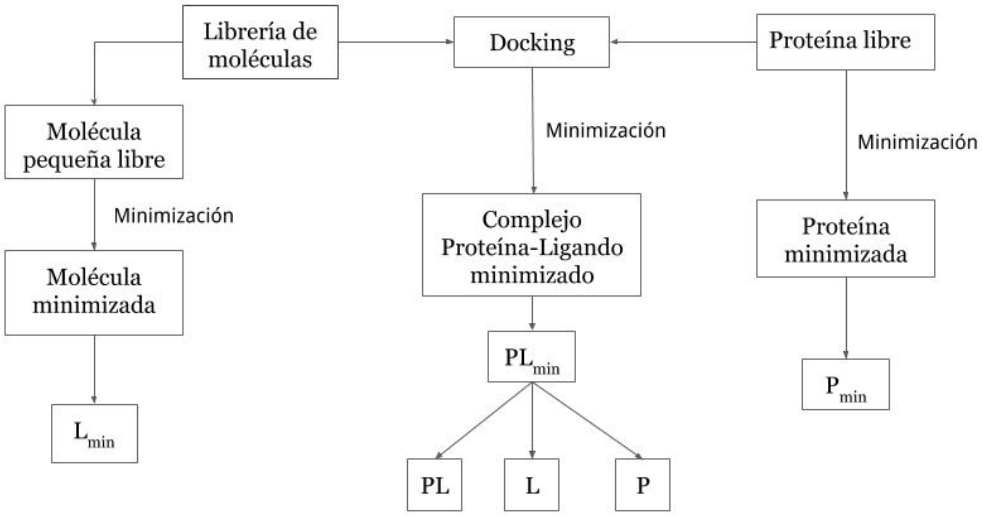


Figura 6.1: Esquema ilustrativo de los pasos efectuados para calcular la energía de unión aproximada por el *score* cuántico.

En el esquema de la Fig. 6.1 se identifican los cálculos de minimización efectuados sobre las diferentes estructuras y cómo se obtienen los distintos términos de las Ecs. 6.13 y 6.14. $E(P)^{COSMO}_{P_{min}}$ representa la energía electrónica calculada con COSMO que equivale al calor de formación de la estructura de la proteína que se obtiene luego de efectuar una minimización de la misma en el estado no ligado en solución (estructura P_{min} en Fig. 6.1), mientras que $E(P)^{COSMO}_{PL_{min}}$ es el calor de formación de la estructura de la proteína extraída del complejo proteína-ligando previamente minimizado en solución (estructura PL_{min} en Fig. 6.1).

Para calcular G^{COSMO} en el contexto de un CV es necesario recurrir a distintas aproximaciones para reducir el costo computacional y el tiempo invertido en dicho cálculo, siendo que las etapas de *docking* y *scoring* deben realizarse en un corto período de tiempo.

Es por este motivo que se propone efectuar una reducción del sistema considerando en el cálculo únicamente los átomos de la proteína de un entorno cercano al ligando, y todos los átomos del ligando.

6.2.1. Criterio de validación

La nueva función de *scoring* desarrollada debe ser validada en cuanto a su desempeño en un proceso de SBVS. El criterio comúnmente empleado es la evaluación de la capacidad de dicha función para seleccionar potenciales ligandos (o *hits*) dentro de una librería química de moléculas que contiene un número elevado de *decoys*, para un determinado receptor. El factor de enriquecimiento (EF) fue la medida empleada para cuantificar dicha evaluación (ver Ec. 4.4).

El EF arroja el número de ligandos activos en un determinado porcentaje de la base de datos rankeada, dividido por el número de ligandos que se esperaría encontrar si estos fueran elegidos de manera aleatoria en la base de datos completa. Es decir, si el 10 % de los ligandos conocidos fueran encontrados dentro del 1 % superior del ranking de moléculas de la base de datos, el factor de enriquecimiento en ese punto (EF1) sería igual a 10. Cuanto mayor sea el valor del EF a un porcentaje fijo de la base de datos ordenada, mejor es la calidad de la función de *scoring*.

Para efectuar la validación antes mencionada se realiza un estudio retrospectivo. En este trabajo se tomaron para tal fin cinco receptores, con sus respectivas librerías químicas de ligandos y *decoys* extraídas de diferentes bases de datos. En los casos en que una molécula de la librería química presenta diferentes estados de protonación o quiralidad, se le asignó un score a cada estado y se eligió el de menor score para conformar la lista de moléculas sobre la que se calculó el EF. Tanto la implementación como los resultados son presentados en el Capítulo 9.

Capítulo 7

Aplicación de MM/QM-COSMO para discriminación de poses

Tomando como base los desarrollos presentados en el Capítulo **??** se extendió la aplicación de la metodología MM/QM-COSMO a un sistema biológico de interés. El estudio de investigación fue realizado en colaboración con grupos experimentales, con el objetivo de encontrar inhibidores novedosos y potentes contra la entrada del virus del Dengue a la célula. Mediante CV y *diseño de novo* se encontraron moléculas con probada actividad. Luego de la identificación de dichos compuestos se caracterizaron sus conformaciones en el sitio activo. Esto brinda información crucial para las etapas de *hit-to-lead* o *lead optimization*, en las que se realizan modificaciones para mejorar la potencia de dichos compuestos. Para efectuar correctamente la caracterización es muy importante conocer el modo de unión de los ligandos en el sitio activo del receptor. Al no contar con estructuras cristalográficas, se implementó un *Docking* Molecular. Como resultado de este proceso se encontraron para una de las moléculas dos poses con energías similares, y conformaciones totalmente opuestas. Por este motivo, se realizaron cálculos de energía libre de unión con el método MM/QM-COSMO, a fin de distinguir la mejor pose. La componente electrostática de la energía de unión fue determinada con dos Hamiltonianos SQM, mientras que el solvente fue modelado utilizando el solvente continuo COSMO. Se estudió además la influencia de moléculas de agua que interactúan de manera directa con el ligando y la proteína. Fueron establecidos distintos protocolos con el objetivo de alcanzar mayor precisión en el cálculo de energía libre, teniendo en cuenta el compromiso entre eficiencia y costo computacional.

7.1. Identificación de antivirales para el Dengue

El Dengue es un virus transmitido por mosquitos que se ha convertido en los últimos años en una preocupación para la salud a nivel mundial. Actualmente, la enfermedad producida por el mismo es endémica en países de América Latina. Sin embargo, no se han aprobado hasta el momento drogas antivirales específicas que permitan controlar la infección del virus. Resulta entonces de gran importancia la identificación y el desarrollo de drogas antivirales efectivas que puedan combatir la creciente expansión de la enfermedad. La proteína de envoltura *E* del virus del Dengue (*Dengue Envelope Protein-E*, DENV-E) está directamente involucrada con el ingreso del mismo a la célula. La estructura cristalográfica de dicha proteína revela un hueco hidrofóbico ocupado por una molécula pequeña (n-octil-β-D-glucósido, β-OG), por lo que se propuso el bolsillo de unión de la misma como un blanco adecuado para el desarrollo de nuevos inhibidores que impidan la entrada del virus.

En este estudio se realizó una búsqueda de moléculas que pudieran unirse al sitio de unión de β-OG siguiendo dos metodologías, SBVS y diseño *de novo*. 23 compuestos estructuralmente diferentes fueron seleccionados luego del proceso de CV, y se sintetizaron 10 compuestos resultantes del diseño *de novo*. Luego, se efectuaron ensayos biológicos de dichos compuestos y se encontraron dos con marcada actividad antiviral, que fueron estudiados posteriormente mediante simulaciones de dinámica molecular. Este trabajo sirvió de base para el desarrollo de nuevos y potentes inhibidores contra la entrada del virus del Dengue a la célula.

Cuando las estructuras experimentales 3D de un complejo proteína-inhibidor no se encuentran disponibles se puede recurrir al método de *Docking* Molecular para predecir las mismas. En las primeras etapas del diseño racional de un fármaco, resulta crucial el conocimiento de la pose que adopta un determinado ligando en el sitio de unión del receptor. A partir de la misma, se pueden realizar distintos estudios computacionales para identificar las principales interacciones y sugerir posibles modificaciones químicas para incrementar la potencia de dicho ligando mediante la síntesis de nuevos derivados, en la etapa de *lead-optimization*.

Descubrimiento de ligandos basado en *Docking*

El proceso de SBVS se llevó a cabo con el programa ICM debido a su efectividad en la búsqueda de candidatos líderes en otros blancos.[141] Se tomó como punto de partida la estructura cristalográfica de la proteína de envoltura *E* del virus del Dengue (DENV-E) extraída de la base de datos PDB, junto con una librería química virtual de 110,000

moléculas pequeñas tomadas de la base de datos Maybridge. En primer lugar, se aplicó un filtro ADME-Tox[1] a dicho conjunto con el propósito de retener únicamente moléculas no tóxicas y con características farmacológicas apropiadas.[142] Seguidamente, los compuestos fueron sujetos a tres procesos independientes de cribado virtual de alto rendimiento (HTD) con el objetivo de generar una mayor diversidad de poses. Por cada compuesto se retuvo únicamente la pose de mejor puntaje de *docking*. Luego de una inspección visual sobre las primeras 500 moléculas del ranking, se compraron o sintetizaron 23 compuestos estructuralmente diferentes sobre los que se efectuaron los estudios *in vitro*. El proceso se encuentra representado en el esquema de la Fig. 7.1.

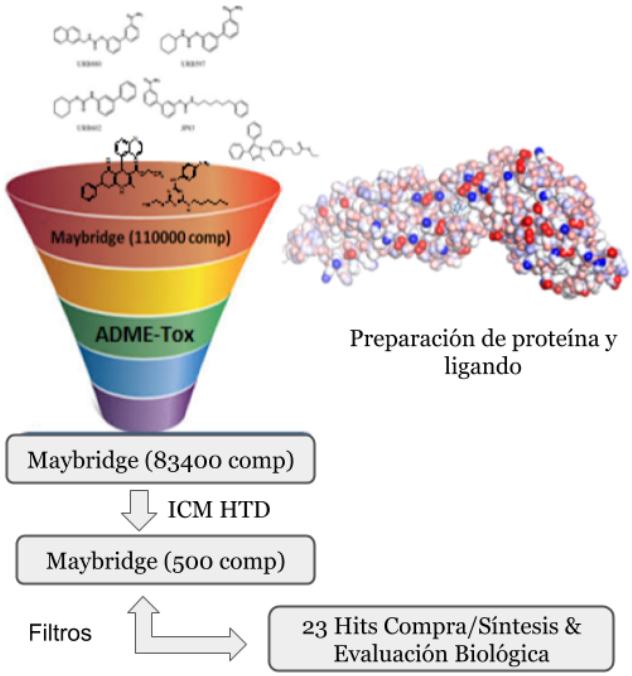


Figura 7.1: Esquema ilustrativo del proceso de SBVS para la identificación de inhibidores de DENV.

Los ensayos biológicos permitieron identificar 5 *hits* con actividad antiviral contra DENV-E, con un valor de EC_{50} en un rango micromolar bajo. En particular se obtuvo un compuesto con marcada actividad antiviral, $EC_{50} = 3,1\mu M$ (compuesto **2**, Fig. 7.2-(A)). La inspección de la pose predicha por *docking* para este compuesto revela que el mismo se une de manera similar a la molécula β-OG, en un área rodeada por los amino ácidos Thr48, Glu49, Ala50, Phe193, Thr268, Ile270, Phe279 y Thr280,

[1]En farmacología, el acrónimo ADME significa Absorción, Distribución, Metabolismo y Excreción y se utiliza para describir la disposición de un compuesto farmacéutico en el organismo. Esos cuatro criterios tienen una influencia directa sobre el nivel del fármaco y su farmacocinética al ser expuesto a los tejidos y por tal razón, influencian el rendimiento y actividad farmacológica del compuesto.

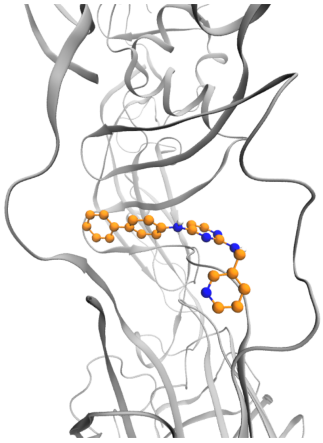
Identificación de ligandos empleando Diseño *de novo*

Partiendo de la estructura cristalográfica reportada, se diseñaron también nuevas moléculas orgánicas pequeñas que pudieran unirse a la proteína E de manera similar al ligando β-OG. Se generó una librería química virtual partiendo del fragmento NH_3 elegido como *core* y posicionado en el sitio de unión de manera tal que formara un puente de hidrógeno con el grupo carbonilo de la Thr48, Se tomaron en consideración para realizar la síntesis los diez mejores compuestos evaluados. Luego se determinó experimentalmente la actividad antiviral de los mismos. El más activo ($EC_{50} = 7,8\mu M$) fue seleccionado para continuar con el proceso de optimización (compuesto **dv7**, Fig. 7.2-(B)).

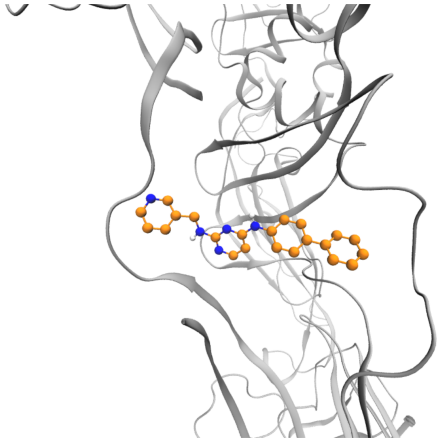
(A) (B)

Figura 7.2: Estructuras representativas tomadas de las simulaciones de dinámica molecular de los compuestos **2** (A) y **dv7** (B) dentro del sitio de unión de la proteína E.

La pose del ligando **dv7** en el sitio de unión de la proteína fue determinada empleando el programa ICM para efectuar el *docking*. Teniendo en cuenta que distintas funciones de *docking* generan poses distintas se resolvió emplear también el programa AutoDock Vina para la predicción de la misma. Como resultado se encontraron dos conformaciones con la menor energía de *docking* en cada caso, rotadas 180° entre ellas (Fig. 7.3). Debido a la gran diferencia entre la pose de menor energía de *docking* para cada programa, resulta de gran interés evaluar la afinidad de unión usando métodos cuánticos para así identificar la conformación correspondiente al mínimo de energía con un mayor nivel de precisión.



(a) Pose de **dv7** resultante del *docking* en ICM.



(b) Pose de **dv7** resultante del *docking* en AutoDock Vina.

Figura 7.3: Poses del compuesto **dv7** en complejo con la proteína *E* en el sitio de unión, con energías de *docking* similares.

Seguidamente, se realizaron simulaciones de dinámica molecular para confirmar la estabilidad de los compuestos más prometedores hallados por SBVS y Diseño *de novo* en el sitio de unión de la proteína, determinar las principales interacciones proteína-ligando y calcular la energía de unión de las dos poses obtenidas para el compuesto **dv7**.

7.1.1. Detalles computacionales

7.1.1.1. *Docking* de alto rendimiento (HTD)

Todas las simulaciones de *docking* fueron realizadas partiendo de la estructura cristalográfica de la proteína de envoltura del virus del dengue tipo 2, unida al detergente n-octil-β-D-glucósido (β-OG), (PDB: 1OKE). El sistema molecular fue descripto en términos de coordenadas torsionales usando el campo de fuerzas ECCEPP/3 implementado en el programa ICM (versión 3,7-2c, MolSoft LLC, La Jolla, CA). Las cargas de los ligandos fueron calculadas con el campo de fuerzas MMFF. Se agregaron átomos de hidrógeno a la estructura del receptor seguido de una minimización local de energía. El *docking* fue realizado en el sitio de unión de β-OG luego de eliminar las moléculas de agua y co-factores, usando una aproximación ligando flexible: receptor rígido implementada en ICM, en la que el receptor se representa por seis mapas de energía potencial: electrostático, enlace de hidrógeno, hidrofóbico y de van der Waals. En el algoritmo de *docking* se considera la flexibilidad del ligando dentro del campo del receptor y se lo somete a una minimización global de energía, con lo cual son minimizadas tanto la energía intramolecular como la intermolecular del ligando. Luego del *docking* de cada molécula se asignó un *score* de *docking* empírico de acuerdo al modo en que se ajusta su posición con el sitio de unión

de la proteína. Para mejorar la convergencia del paso de minimización de energía global, el *docking* fue realizado tres veces y la pose de mejor *score* por molécula se utilizó para continuar los análisis.

7.1.1.2. Diseño *de novo*

Las estructuras iniciales fueron generadas por el programa *BOMB* partiendo del archivo PDB 1OKE; se removió el ligando y se lo reemplazó por amonio, que fue usado por el programa para crear los análogos en el sitio de unión. Se fueron agregando sucesivamente distintos grupos químicos, y evaluando en cada paso el ajuste de los mismos en el sitio de unión hasta obtener una molécula adecuada.

Para cada molécula generada se realizó una búsqueda conformacional. Para ello fueron optimizados los ángulos dihedros del confórmero junto con sus posiciones y orientaciones en el sitio de unión, mediante el campo de fuerzas OPLS-AA para la proteína y OPLS/CM1A para el análogo. El confórmero de menor energía resultante se evaluó con una función de *scoring* similar a las usadas en *docking*, para predecir la afinidad. Los diez mejores análogos encontrados fueron sintetizados y evaluados experimentalmente.

7.1.1.3. Simulaciones de Dinámica Molecular

Se estudió el comportamiento dinámico de los ligandos en el sitio de unión de la proteína mediante simulaciones de dinámica molecular. La estructura de partida en cada caso fue la correspondiente a la pose resultante del *docking*. Las simulaciones fueron realizadas para cada compuesto a 300 K con moléculas de agua explícitas, con el programa GROMACS v4,6[143,144] con el campo de fuerzas Amber99SB para proteínas.[145] Los parámetros para las moléculas pequeñas se obtuvieron mediante la interfaz AnteChamber PYthon Parser InterfacE (ACPYPE) usando el campo de fuerzas GAFF.[83]

Preparación y evolución temporal del sistema Luego de solvatar y neutralizar el sistema, se realizó una minimización de energía por medio del algoritmo *steepest−descent*, hasta que la fuerza máxima decayó a 1000 [kJ mol^{-1} nm^{-1}]. Luego se equilibró durante 100 ps el sistema completo con un ensamble NVT, seguido de 500 ps de equilibración NPT. La temperatura se mantuvo constante a un valor de 300 K usando el termostato modificado de Berendsen con una constante de acoplamiento de 0,1 ps. La presión se mantuvo constante a 1 bar en todas las direcciones con constante de acoplamiento de 0,5 ps. Finalmente el sistema equilibrado fue sujeto a 100 ns de simulaciones de dinámica molecular a 300 K. A partir de las trayectorias obtenidas, se analizaron las interacciones más importantes originadas en cada sistema.

7.1.1.4. Cálculos Semiempíricos

Se utilizó el software MOPAC 2012 para realizar los cálculos de estructura electrónica, junto con la aproximación de escalamiento lineal MOZYME debido a las dimensiones de los sistemas estudiados. Los Hamiltonianos semiempíricos PM6-D3H4 y PM7 fueron empleados para el cálculo del calor de formación, con el modelo de solvente continuo COSMO también implementado en MOPAC. Fueron realizados distintos tipos de cálculo según los protocolos mencionados en la sub-Sección 6.1.1 del presente Capítulo. El sistema fue minimizado 25 y 50 ciclos.

Estructuras de partida: Configuraciones de Dinámica Molecular

Para realizar los cálculos de energía con métodos semiempíricos para las dos poses del compuesto **dv7**, se tomaron en total 100 estructuras de la simulación de dinámica molecular, con un intervalo de 10 ps entre cada una de ellas, a partir de los primeros 50 ns de simulación. Luego se evaluó la energía de cada conformación mediante SQM y se tomaron en consideración los valores promediados para el ensamble.

7.2. Discusión de resultados

Mediante la determinación del valor de energía libre de unión con el método MM/QM-COSMO se pudo identificar como la pose de menor energía para el compuesto **dv7** la correspondiente al *docking* efectuado con el programa AutoDock Vina. Por este motivo fue seleccionada únicamente esta pose para realizar la caracterización *in silico*, presentada a continuación. La implementación y los resultados de los cálculos cuánticos se encuentran más adelante en el presente Capítulo.

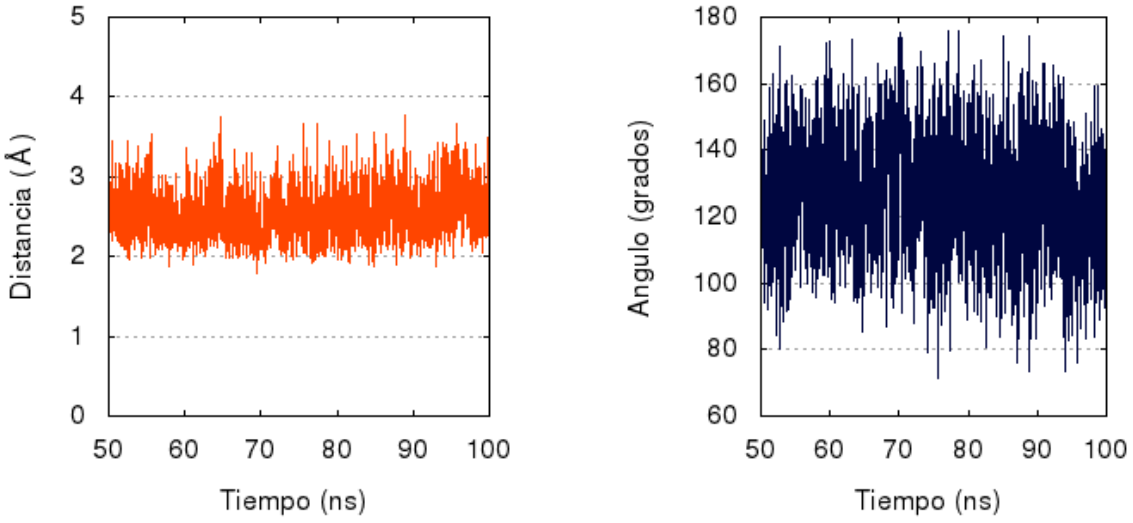
7.2.1. Caracterización *in silico* de la proteína E en complejo con los compuestos 2 y dv7

Una vez seleccionados los mejores candidatos obtenidos con las dos metodologías empleadas, se realizó una simulación de dinámica molecular de 100 ns para confirmar la estabilidad de dichos compuestos (**2** y **dv7**) en el sitio de unión de la proteína E de envoltura del dengue. Esto permitió además caracterizar las interacciones entre ligando y proteína, e investigar el papel que juegan en este caso las moléculas de agua que pueden estar involucradas mediando dicha interacción (ya que las mismas fueron omitidas durante la etapa de *docking*).

En el caso del compuesto **2**, la pose resultante del *docking* fue usada como conformación inicial para las simulaciones, mientras que para el compuesto **dv7** se tomaron las dos poses de energía de *docking* similar como punto de partida realizando 100 ns de dinámica molecular en ambos sistemas iniciales.

Figura 7.4: Interacción de **2** con los amino ácidos del sitio de unión y moléculas de agua. Por simplicidad, los nitrógenos del esqueleto no se muestran, excepto aquellos que interactúan directamente con el ligando. Se muestran sólo los hidrógenos polares. Los enlaces de hidrógeno se muestran como una línea de esferas coloreadas. Código de color: **2**, carbonos amarillos; DENV *E* carbonos blancos; oxígenos, rojos; nitrógenos, azul; azúfre, verde; hidrógenos polares, gris.

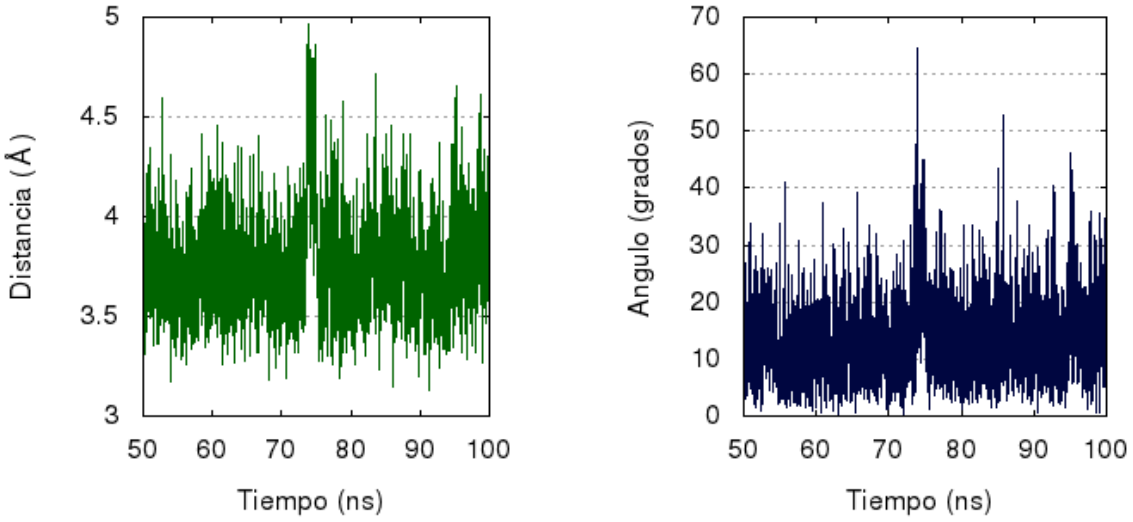
En la Fig. 7.4 se puede ver una configuración representativa tomada de la dinámica para el compuesto **2** con la proteína *E* donde se observan las interacciones predominantes entre los átomos del ligando y los residuos del sitio de unión de la proteína, así como también las moléculas de agua que interactúan estabilizando dicho compuesto.



(A) (B)

Figura 7.5: Distancia (A) y ángulo (B) del enlace de hidrógeno como función del tiempo entre el átomo de H de la Ala50 y el átomo aceptor N del ligando **2**.

La formación de interacciones de enlaces de hidrógeno entre las cadenas laterales de la proteína y el ligando, contribuyen significativamente a la estabilidad del complejo. Estas interacciones fueron determinadas y evaluadas en la simulación por medio de análisis geométrico de ángulos y distancias de enlace. En el caso de **2** el sistema es estabilizado por un enlace de hidrógeno dinámico moderado entre el átomo de nitrógeno de la piridina del ligando y el HN del residuo Ala50; como se observa en la Fig. 7.5 este enlace permanece estable luego de los primeros 50 ns de la simulación, con una distancia interatómica promedio de 2,5 Å y un ángulo donor de aproximadamente 130°. El oxígeno en la amida está expuesto de manera intermitente al solvente a través de un canal angosto.



(A) (B)

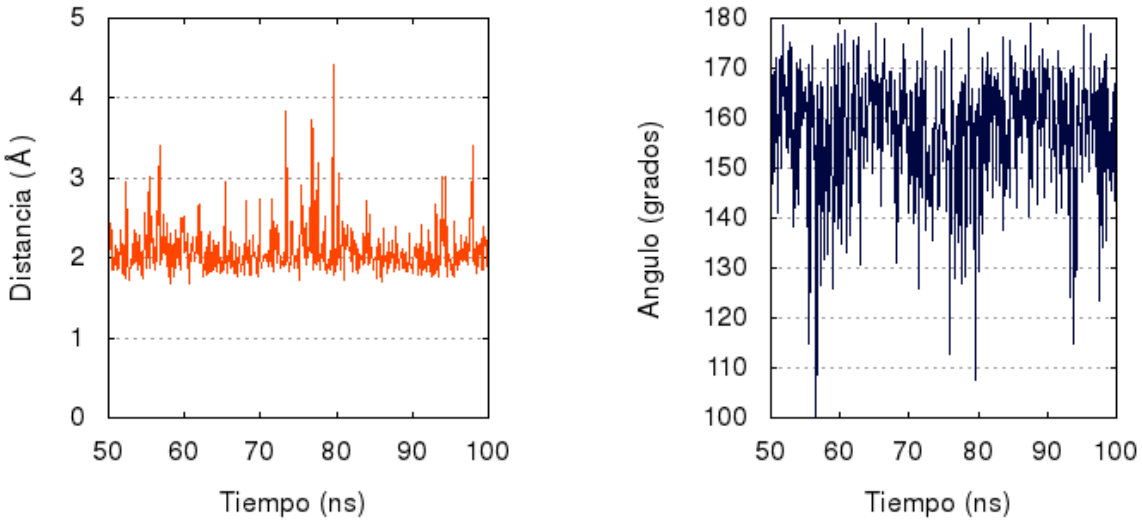
Figura 7.6: Distancia (A) y ángulo (B) del enlace π-π de **2**.

Durante la simulación se encontraron ciertas moléculas de agua dentro del sitio de unión, mediando la interacción entre el ligando y la proteína. Una molécula de agua con casi 100 % de ocupancia en los últimos 50 ns establece puentes de hidrógeno con el N de la piridina terminal, el grupo hidroxilo de la Tyr137, el grupo carboxílico O de la Thr280 y otra molécula de agua que pertenece a un pequeño cluster conservado de tres moléculas de agua escondidas que también interactúan con el HN de la Leu191 e His282, el O carboxílico de Thr189 y la cadena lateral de His282 (Fig. 7.4). También se pudo identificar durante la simulación una interacción estable π-π entre el anillo de la piridina terminal y la Phe193, con una distancia promedio entre centroides de los anillos de 3,7 Å y un ángulo promedio entre planos de 14° (Fig. 7.6).

Figura 7.7: Interacción de **dv7** con los amino ácidos del sitio de unión y moléculas de agua. Por simplicidad, los nitrógenos del esqueleto, oxígenos carbonilos no se muestran, excepto aquellos que interactúan directamente con el ligando. Se muestran sólo los hidrógenos polares. Los enlaces de hidrógeno se muestran como una línea de esferas coloreadas. Código de color: **dv7**, carbonos amarillos; DENV *E* carbonos blancos; oxígenos, rojos; nitrógenos, azul; azúfre, verde; hidrógenos polares, gris.

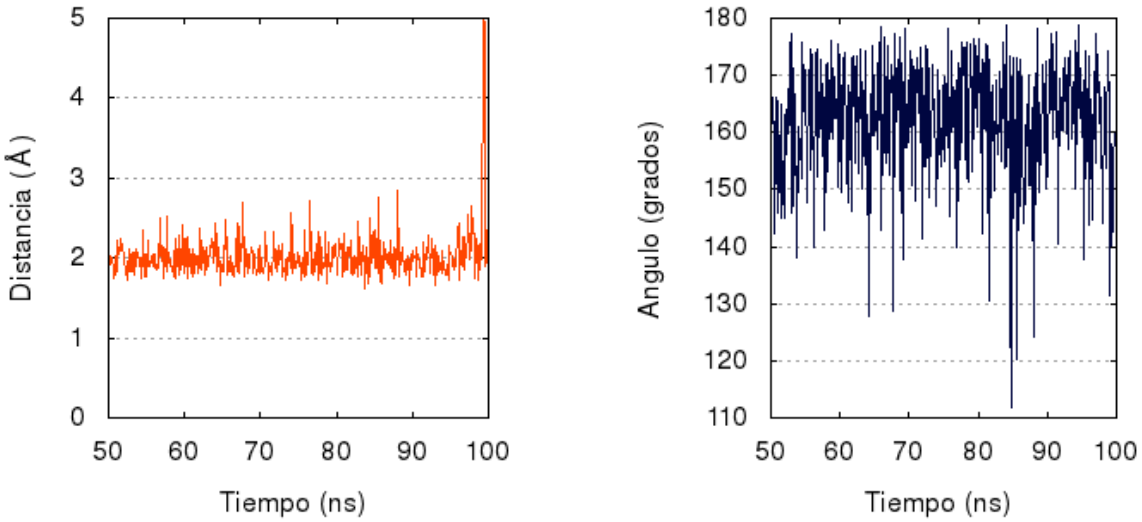
En el caso del compuesto **dv7**, la inspección visual de la dinámica del compuesto muestra que el mismo permanece completamente dentro de la cavidad sin partes enteramente expuestas al solvente, para ambas poses iniciales. El grupo amino está en contacto con el residuo Ala50 y establece un enlace de hidrógeno estable con el átomo de O. Otro puente de hidrógeno se pudo identificar entre el O de la Lys51 y el NH del ligando (Fig. 7.7). Un extremo del ligando posee mayor movilidad y se encuentra semi expuesto al solvente; en el último tercio de la simulación establece contactos hidrofóbicos con los amino

ácidos Gln200 y Met201,



(A) (B)

Figura 7.8: Distancia (A) y ángulo (B) del enlace de hidrógeno como función del tiempo entre el átomo de H del ligando **dv7** y el átomo de O de la Ala50,



(A) (B)

Figura 7.9: Distancia (A) y ángulo (B) del enlace de hidrógeno como función del tiempo entre el átomo de H del ligando **dv7** y el átomo de O de la Lys51,

7.2.2. Discriminación de poses de *Docking* con MM/QM-COSMO

A fin de identificar la pose de menor energía para el compuesto **dv7** de manera precisa, se aplicó la metodología desarrollada en el Capítulo **??**. Se efectuó una re-evaluación de las componentes de energía a nivel SQM usando el modelo COSMO, para distintas

conformaciones extraídas de una simulación de MD realizada a partir de las poses del *docking*. La energía libre total en solución (Ec. 6.7) fue calculada como el valor promedio de energía libre en cada conformación evaluada, para el complejo, la proteína y el ligando.

Se establecieron distintos protocolos de cálculo, con el objetivo de encontrar un balance adecuado entre precisión y eficiencia computacional. En primer lugar, se consideró el sistema completo proteína-ligando, conformado por 12258 átomos para el dímero de la proteína E y 46 átomos para el ligando. Fueron seleccionadas 10 estructuras a partir de los 50,5 ns de simulación, utilizando la herramienta *trjconv* del programa GROMACS, habiendo eliminado previamente las moléculas de agua. Seguidamente, el sistema fue minimizado 50 ciclos con el Hamiltoniano PM7 y el modelo COSMO. Luego, se calculó la energía libre en la conformación minimizada del complejo, la proteína y el ligando, tomando la diferencia correspondiente y promediando para todas las estructuras ($\langle \Delta G^{COSMO} \rangle$). Por otra parte, se calculó la energía total a partir de la trayectoria de MD con métodos de mecánica clásica, con GROMACS, y finalmente se realizó la evaluación de energía con el método MM-PBSA.

En la Tabla 7.1 se presenta el *ranking* de las dos poses del compuesto **dv7** (dv7-A y dv7-B), según la energía libre estimada por cada metodología. Se observa como resultado, una inversión en el ordenamiento de las poses entre los métodos de *Docking* y MM, con respecto a MM-PBSA y MM/QM-COSMO.

Tabla 7.1: Comparación del *ranking* asignado para las poses del compuesto **dv7** según el valor de afinidad para 10 estructuras evaluadas por métodos de *Docking*, Mecánica Molecular con solvente explícito (MM), Mecánica Molecular con solvente implícito (MM-PBSA) y semiempíricos con solvente implícito (MM/QM-COSMO). El número 1 se asigna a la pose de menor energía.

Pose	*Docking*	MM	MM-PBSA	MM/QM-COSMO
dv7-A	1	1	2	2
dv7-B	2	2	1	1

Luego, se incorporaron moléculas de agua cercanas a la proteína y el ligando, dada la importancia de dichas moléculas en la interacción con ambas partes. El criterio seguido para la selección de las mismas, fue el de retener todas aquellas moléculas que tuvieran al menos un átomo a una distancia de 4 Å del ligando y de la proteína simultáneamente. Se hicieron cálculos con una única estructura de **dv7** comparando los resultados obtenidos con distintas aproximaciones:

(1) Cálculo de energía QM-COSMO sin optimización previa de la estructura.

(2) Cálculo de energía QM-COSMO sobre la estructura inicial, considerando las moléculas de agua como parte del receptor.

(3) Optimización del complejo, proteína con aguas, y ligando, partiendo de la estructura inicial y posterior cálculo de energía QM-COSMO.

(4) Optimización del complejo, proteína con aguas, y ligando partiendo de la estructura inicial, restringiendo la minimización únicamente a los átomos localizados a 5 Å del ligando seguida del cálculo de energía QM-COSMO.

Fue seleccionada la tercera metodología, de acuerdo a los valores de energía libre obtenidos para una configuración, para extender su aplicación a un conjunto de 10 configuraciones de la dinámica. La optimización realizada fue de 50 ciclos sobre todo el sistema. Para simplificar la notación, llamaremos en adelante pose **1** y pose **2** a las poses **dv7-A** y **dv7-B**, respectivamente. Se compararon los resultados obtenidos con las metodologías MM-PBSA y MM/QM-COSMO (Tabla 7.2). La incorporación de moléculas de agua se ve reflejada en los valores más negativos de energía de interacción para ambas poses. El orden de las poses sigue siendo el mismo al encontrado sin tener en cuenta dichas moléculas en el cálculo. Se puede notar, que los valores de energía para ambas poses sin consideración de aguas explícitas, es similar al encontrado con el método MM-PBSA. Se puede concluir entonces que el método MM/QM-COSMO, es adecuado para describir las principales interacciones entre receptor y ligando.

Tabla 7.2: Comparación de valores de energía para las poses **1** y **2** del ligando dv7, con los métodos MM-PBSA y MM/QM-COSMO.

Pose	MM-PBSA (kcal/mol)	MM/QM-COSMO (kcal/mol)	
		sin aguas	con aguas
1	-36,7 +/- 2,0	-38,2 +/-1,3	-58,9 +/-9,6
2	-28,2 +/- 4,6	-27,4 +/-1,8	-50,4 +/-8,3

Para reducir el tiempo de cómputo y poder ampliar el estudio a un número mayor de estructuras, se realizaron algunas modificaciones en la implementación de la metodología. En primer lugar, se consideró únicamente un monómero de la proteína, lo que resulta en una reducción de aproximadamente 6000 átomos en el sistema total. Se efectuó también una comparación entre el número de ciclos de minimización realizados (tomando 50 y 25 ciclos), y por último se compararon los resultados de una minimización completa del sistema, frente a la minimización del sitio de unión de la proteína y ligando, dejando fijos los demás átomos. En cuanto a las moléculas de agua explícitas, se tomaron nuevos criterios de selección para incorporar únicamente aquellas que establecen una mediación de la interacción entre proteína y ligando. Se aplicaron los siguientes protocolos:

- **I)** Cálculo de energía sin consideración de moléculas de agua explícitas.
- **II)** Se seleccionaron las moléculas de agua que tienen cualquiera de sus átomos a una distancia menor a 3,5 Å de algún átomo de la proteína y 3,5 Å de algún átomo del

ligando, cualquiera sea dicho átomo. De todas las que cumplen la condición anterior fueron seleccionadas las 7 moléculas más cercanas al ligando.

- **III)** Se seleccionaron las moléculas de agua con distancia de 3,5 Å entre el oxígeno del agua y átomos pesados (que no fueran átomos de carbono) de proteína y ligando simultáneamente.
- **IV)** Se seleccionaron las moléculas de agua con distancia de 3,5 Å entre el oxígeno del agua y átomos pesados de proteína y ligando simultáneamente.

El tiempo de cómputo se redujo aproximadamente a la mitad, al pasar de considerar del dímero al monómero. El error encontrado también fue menor en este último caso. El número total de átomos del sistema es 12635 para el dímero y 6150 para el monómero. Los valores de energía siguen la misma tendencia que los cálculos con todos los átomos del complejo proteína-ligando. Esto permitió reducir el sistema y continuar realizando los análisis en un tiempo más accesible.

Los resultados de las metodologías anteriormente expuestas, para el sistema reducido, se presentan en la Tabla 7.3. Para las dos poses, el error se incrementa levemente al reducir de 50 a 25 ciclos la minimización del sistema. El ordenamiento de las poses es el mismo independientemente del número de ciclos de minimización, para los cuatro protocolos seguidos. Se puede considerar por lo tanto que la minimización alcanza un valor estable con un menor número de ciclos de minimización, lo cual asegura que es posible determinar la energía de las poses en un menor tiempo.

Los distintos criterios adoptados para la selección de moléculas de agua que son introducidas luego en el cálculo de energía de unión influyen directamente en el valor de la misma. Cuando no se incorporan moléculas de agua de manera explícita (protocolo I), la diferencia entre las dos poses del ligando es mayor a los casos en que las moléculas de agua son tenidas en cuenta.

Tabla 7.3: Comparación de valores de energía para dos poses del ligando dv7, método MM/QM-COSMO. Minimización del sitio de unión del monómero, 50 y 25 ciclos en 10 configuraciones.

Pose	Ciclos	MM/QM-COSMO (kcal/mol)			
		I	II	III	IV
1	50	-36,4 +/- 1,5	-52,4 +/- 1,7	-45,2 +/- 1,3	-50,3 +/- 1,7
1	25	-34,5 +/- 1,5	-50,0 +/- 1,9	-43,6 +/- 1,4	-48,9 +/- 2,1
2	50	-24,6 +/- 1,5	-45,2 +/- 2,7	-40,1 +/- 2,2	-44,3 +/- 2,3
2	25	-22,9 +/- 1,4	-42,9 +/- 2,8	-38,3 +/- 2,0	-41,3 +/- 1,9

Se eligió el protocolo IV para ampliar el número de configuraciones evaluadas a 50, y se volvieron a comparar los resultado de una minimización de 50 y 25 ciclos (Tabla 7.4). El error obtenido fue similar en ambos casos. Nuevamente, la pose **1** fue la de menor energía.

Tabla 7.4: Comparación de valores de energía para dos poses del ligando dv7 con el método MM/QM-COSMO. Minimización del sitio de unión del monómero, 50 y 25 ciclos en 50 configuraciones.

Pose	MM/QM-COSMO (kcal/mol)	
	Ciclos	
	50	25
1	-52,3 +/- 1,0	-51,8 +/- 1,1
2	-49,3 +/- 1,2	-49,2 +/- 1,2

Se procedió entonces a la aplicación de los protocolos sin aguas explícitas, y con aguas seleccionadas a partir del protocolo IV, para extender la metodología a 100 configuraciones extraídas de la simulación de MD. Se realizó además un cálculo de energía con moléculas de agua explícita, pero sin minimización previa en el sistema (protocolo V), es decir calculando la energía directamente de las estructuras tomadas de la dinámica. Los resultados se presentan en la Tabla 7.5. Se puede observar una reducción notable de los errores obtenidos, al incrementar el número de conformaciones tomadas para la evaluación de energía. Para el protocolo V, la diferencia de energía encontrada entre las poses es menor a las halladas por los protocolos I y IV. La incorporación de moléculas de agua en la evaluación de energía, aún si la consideración de las mismas es importante para la descripción correcta de las interacciones en el sistema, dificulta en cierta medida la distinción de la mejor pose, al reducir la diferencia de energía entre ambas conformaciones.

Tabla 7.5: Comparación de valores de energía para dos poses del ligando dv7 con el método MM/QM-COSMO. Minimización del sitio de unión del monómero, 50 y 25 ciclos en 100 configuraciones.

Pose	MM/QM-COSMO (kcal/mol)			MM-PBSA (kcal/mol)
	I	IV	V	
1	-37,1 +/- 0,5	-51,9 +/- 0,8	-45,6 +/- 0,8	-42,0 +/- 0,4
2	-26,3 +/- 0,7	-49,0 +/- 0,9	-44,3 +/- 0,8	-28,0 +/- 0,3

Se pudo predecir, con la serie de estudios realizados con métodos SQM, que la pose **1** es la más estable para el compuesto activo **dv7**. Por otro lado, la estructura resultante del *docking* se encuentra cerca del mínimo de energía local, por lo que el cálculo de energía con métodos cuánticos puede ser efectuado directamente sobre la pose resultante del *docking* sin minimización previa (protocolo IV) encontrando de esta manera una metodología menos costosa computacionalmente y de precisión equivalente a las que incluyen minimización cuántica.

Capítulo 8

Cálculo de energía libre en complejos proteína-ligando

La función de *scoring* cuántica presentada en el Capítulo **??** fue definida de tal manera que el *score* es una aproximación de la energía libre de unión. Es necesario entonces, como primera validación de la metodología, corroborar que las estimaciones realizadas se acerquen a valores experimentales de afinidad para complejos proteína-ligando.

En este Capítulo se presentan los resultados de un estudio retrospectivo realizado, empleando los Hamiltonianos SQM PM7 y PM6-D3H4, junto con el modelo COSMO. El *score* cuántico fue asignado a un grupo de 15 inhibidores de la proteína CDK2 con estructura cristalográfica y valor de afinidad conocidos. El objetivo más amplio de esta aplicación se enmarca en la posterior implementación de la nueva función de *scoring* cuántica, en un protocolo de cribado virtual basado en la estructura del receptor. La principal motivación de este tipo de estudios es la necesidad de contar con métodos más precisos para la separación de ligandos/no ligandos en un proceso de CV, por lo que las aproximaciones empleadas deben contemplar el compromiso existente entre precisión, y requerimientos computacionales.

La calidad de los métodos cuánticos para la estimación de energías libres de afinidad relativas condujo a resultados alentadores para aplicaciones posibles en el marco del diseño racional de un fármaco. Esto permite aplicar la función de *scoring* a complejos proteína-ligando cuyas estructuras cristalográficas se desconocen.

8.1. Proteína CDK2 en complejo con 15 inhibidores

Cuando se tiene un conjunto de ligandos con estructuras cristalográficas en unión con el receptor y valores de afinidad experimentales conocidos es posible generar un *ranking* de acuerdo a la estimación del valor de afinidad relativa. El hecho de contar con información experimental permite evaluar la capacidad de la metodología para predecir diferencias de energía libre entre compuestos bio-activos, para así establecer posibles modificaciones y alcanzar una mayor precisión. Es de esperar que la correlación entre los valores calculados por el *score* y la energía de unión medida experimentalmente no sea buena, debido a que usualmente se desprecian contribuciones muy importantes en los cálculos para reducir el enorme costo computacional que implicarían, debido al gran tamaño de los sistemas biológicos estudiados. Esto permite realizar optimizaciones para poder aplicar la misma a un proceso de HTD, en el que se cuenta con miles a millones de compuestos a los que se debe asignar un *score* de manera rápida y eficiente.

Preparación de los complejos

Las estructuras cristalográficas de la proteína CDK2 en complejo con 15 inhibidores diferentes fueron extraídas de la base de datos PDB. En el caso particular de los complejos 1pkd, 1h1p, 1h1s, 1ogu y 2x1n, los inhibidores están unidos a la forma activa de la proteína, es decir que CDK2 se encuentra unida a ciclina A3, Para dichos complejos se tomó en consideración únicamente el monómero CDK2, eliminando la ciclina A3, Posteriormente, se agregaron átomos de hidrógeno a las estructuras efectuando una minimización sobre dichos átomos, usando el FF AMEBR-ff03 y GAFF, implementados en el programa Chimera. Las constantes de inhibición de todos los complejos proteína-ligandos analizados se encuentran disponibles en la literatura.

8.2. Protocolos de aplicación

Para realizar los cálculos de *score* sobre las estructuras de los complejos fueron aplicados distintos protocolos, yendo desde una menor a mayor complejidad en los mismos. Consecuentemente, los protocolos más precisos conllevan un tiempo mayor de cálculo.

Partiendo de la Ec. 6.12, se consideraron distintas aproximaciones para evaluar la precisión de la metodología empleando para ello algunos de los términos del *score*, o

incluyendo en el cálculo todas las componentes,

$$Score = \Delta E_{int} + \Delta G_{def}(P) + \Delta G_{def}(L) + \Delta G_{np} - T\Delta S \tag{8.1}$$

donde en este caso, ΔE_{int} es la energía de interacción en agua del complejo proteína-inhibidor, $\Delta G_{def}(P)$ y $\Delta G_{def}(L)$ son las contribuciones provenientes de la deformación de la proteína y el ligando, respectivamente, al pasar de la conformación libre a la conformación que adopta en el complejo, ΔG_{np} es la contribución de la energía no polar y el último término $-T\Delta S$ es el cambio de entropía.

La energía fue calculada con los Hamiltonianos PM7 y PM6-D3H4, con el modelo de solvente continuo COSMO en el programa MOPAC.

Antes de realizar la evaluación del *score* se efectuó una minimización de 100 ciclos para los átomos de la proteína localizados a una distancia menor o igual a 7 Å del inhibidor, con cada uno de los Hamiltonianos de manera separada.

La deformación de la proteína y ligando ($\Delta G_{def}(X)$) fue calculada sobre las estructuras minimizadas como

$$\Delta G_{def}(X) = E(X)^{COSMO}_{PL_{min}} - E(X)^{COSMO}_{X_{min}} \tag{8.2}$$

donde X = P o L, y el subíndice (PL_{min}, X_{min}) indica sobre cual estructura se realizó el cálculo de energía. Es decir, se tomó la diferencia de energía entre la geometría del ligando o la proteína en el complejo proteína-ligando ($E(X)^{COSMO}_{PL_{min}}$), y la energía del ligando o la proteína libre, optimizados en agua ($E(X)^{COSMO}_{X_{min}}$). En este caso, se utilizó el modelo COSMO para la optimización en fase solvente.

La componente no polar de la energía de solvatación fue calculada con la metodología γSASA.

El término entrópico (-TΔS) fue determinado usando una aproximación de oscilador armónico de rotor rígido, con un potencial empírico. Según esta aproximación, las biomoléculas ocupan un único pozo armónico de potencial. Se sabe que dicha aproximación no es muy precisa y puede introducir errores para sistemas con muchos pozos de potencial accesibles, sin embargo es comúnmente empleada para evaluar la contribución de entropía en sistemas biomoleculares.

A continuación, se presentan las distintas estimaciones de energía libre de afinidad realizada por el *score* cuántico, el cual es denominado en adelante ΔG_{calcx}, donde x indica

el número de protocolo,

$$\Delta G_{calc1} = \Delta E_{int} + \Delta G_{def}(L) + \Delta G_{np} \quad (8.3)$$

$$\Delta G_{calc2} = \Delta E_{int} + \Delta G_{def}(L) + \Delta G_{np} - T\Delta S_{ff09} \quad (8.4)$$

$$\Delta G_{calc3} = \Delta E_{int} + \Delta G_{def}(P) + \Delta G_{def}(L) + \Delta G_{np} - T\Delta S_{ff09} \quad (8.5)$$

$$\Delta G_{calc4} = \Delta E_{int} + \Delta G_{def}(L) + \Delta G_{np} - T\Delta S_{ff03} \quad (8.6)$$

$$\Delta G_{calc5} = \Delta E_{int} + \Delta G_{def}(P) + \Delta G_{def}(L) + \Delta G_{np} - T\Delta S_{ff03} \quad (8.7)$$

8.3. Análisis y Discusión de resultados

En esta sección se presentan los resultados de la aplicación de la función de *scoring*, para la proteína CDK2 en complejo con 15 inhibidores conocidos. Las estructuras cristalográficas de los complejos se encuentran disponibles, al igual que los valores experimentales de las contantes de inhibición K_i. En primer lugar, se realizaron los análisis tomando como punto de partida para la evaluación del *score* las estructuras optimizadas de los complejos, para los átomos del sitio de unión, definido como la región comprendida a una distancia de 7 Å de cada inhibidor. El cálculo fue realizado con los Hamiltonianos PM7 y PM6-D3H4, en combinación con el modelo de solvente continuo COSMO, que incluye únicamente la componente electrostática de la energía de solvatación. Por esta razón, se agregó en el *score* la contribución de la componente no polar con el método γSASA.

PM7

La energía de interacción (ΔE_{int}) de cada uno de los inhibidores en complejo con la proteína CDK2 fue calculada con el Hamiltoniano PM7, en primer lugar considerando todos los átomos del complejo proteína-ligando. Los resultados para los distintos términos del *score* se presentan en la Tabla 8.1.

Tabla 8.1: Valores experimentales y términos calculados para 15 complejos proteína-ligando de CDK2, con el Hamiltoniano PM7.

Complejo	ΔE_{int}	ΔG_{np}	ΔG_{calc1}	$T\Delta S_{ff09}$	ΔG_{calc2}	ΔG_{calc3}	$\ln K_i$
1aq1	-88,5	-1,7	-86,0	-16,5	-69,5	-62,4	-19,66
1e1x	-54,7	-1,3	-55,1	-12,4	-42,7	-39,6	-13,55
1h1p	-50,3	-1,3	-50,5	-18,0	-32,5	-26,6	-11,33
1h1s	-82,2	-1,7	-81,6	-25,4	-56,2	-49,0	-18,93
1ogu	-79,4	-1,7	-79,1	-21,4	-57,7	-52,9	-17,55
1pkd	-78,9	-1,8	-80,0	-17,7	-62,3	-62,6	-17,32
1pxj	-44,2	-1,1	-44,2	-20,3	-23,9	-20,0	-11,94
1pxl	-57,2	-1,6	-57,8	-27,0	-30,8	-27,0	-15,05
1pxm	-68,5	-1,5	-67,1	-38,6	-28,6	-26,4	-16,63
1pxn	-74,3	-1,5	-72,6	-25,8	-46,8	-42,1	-16,47
1pxp	-58,0	-1,6	-58,3	-19,7	-38,5	-36,5	-15,33
2a4l	-58,8	-1,7	-58,7	-0,5	-58,2	-56,2	-13,63
2exm	-46,1	-1,1	-46,9	-18,8	-28,1	-25,2	-9,46
2fvd	-77,6	-1,7	-75,7	-24,6	-51,1	-43,9	-19,62
2x1n	-53,1	-1,6	-52,7	-13,7	-39,0	-33,2	-17,59

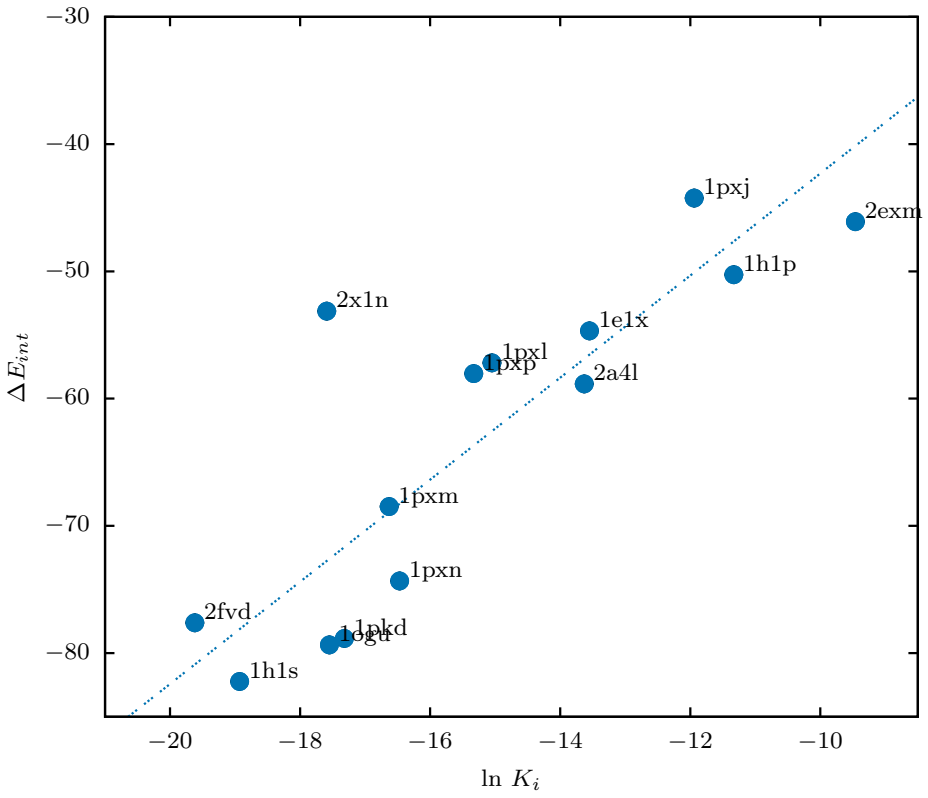


Figura 8.1: Correlación entre energía de interacción ΔE_{int} y constantes de inhibición ln (K_i) en kcal/mol, con r^2=0,74.

En la Fig. 8.1, se pueden observar los valores de ΔE_{int}, en función de la constante de inhibición K_i (ln K_i). La correlación hallada es buena (r^2=0,74). Si se eliminan del grupo de inhibidores los compuestos con mayor desviación con respecto a los valores

experimentales (denominados *outliers*) que en este caso corresponden a 2x1n y 1pxj, la correlación se incrementa notablemente (r^2=0,89).

La correlación entre los valores experimentales de la constante de inhibición y la estimación de energías libres de unión aproximada por el *score*, debería ser mejor al incluir las demás contribuciones, es decir la energía de solvatación no polar, las diferencias de energía de deformación y la contribución entrópica. La correlación encontrada para ΔE_{int} puede ser originada por una compensación entre los términos que no fueron incluidos. Cuando los términos de solvatación no polar, y de deformación del ligando son incorporados en el *score* (ΔG_{calc1}), la correlación con los valores experimentales empeora, con r^2=0,72 (Fig. 8.2). Nuevamente, eliminando los *outliers* de la lista, se encuentra una mejor correlación (r^2=0,87). Estos resultados indican, que en el caso de la proteína CDK2, los términos de deformación del ligando pueden ser ignorados.

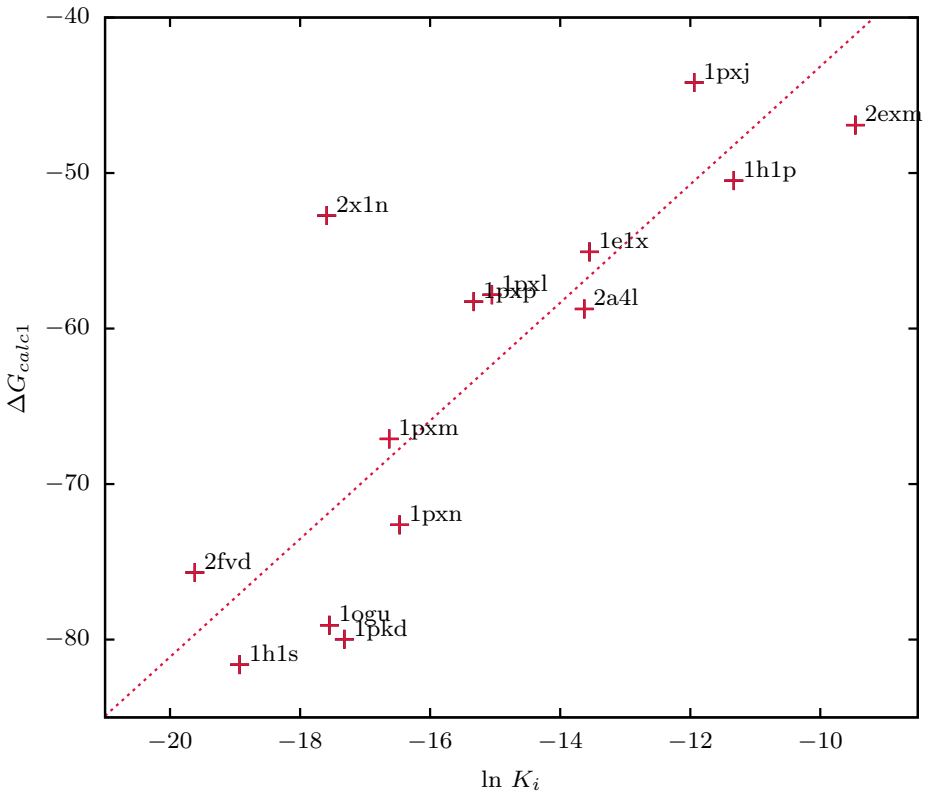


Figura 8.2: Correlación entre el *score* calculado como ΔG_{calc1} y constantes de inhibición ln (K_i) en kcal/mol, con r^2 =0,72.

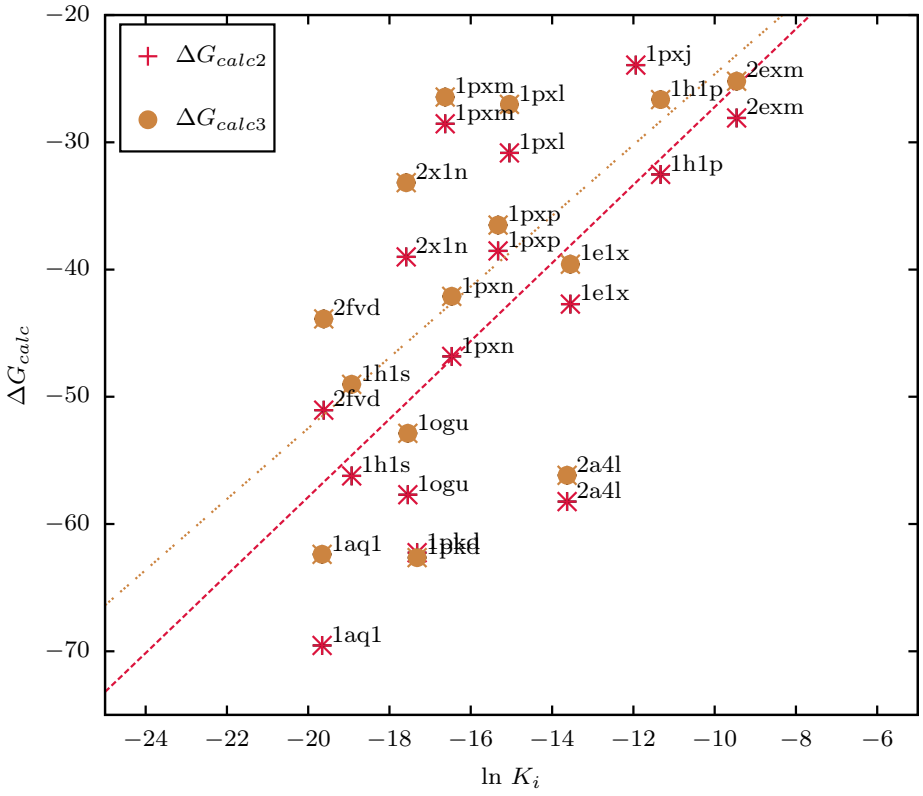


Figura 8.3: Correlación entre el *score* calculado como ΔG_{calc2} y ΔG_{calc3}, y constantes de inhibición ln (K_i) en kcal/mol, con $r^2 = 0{,}44$ y $r^2 = 0{,}38$ respectivamente.

Por último, agregando la contribución de entropía al *score*, calculada con el FF AMBER ff09 (ΔG_{calc2} y ΔG_{calc3}), la correlación resulta considerablemente más baja que en los casos anteriores, $r^2 = 0{,}44$ si se tiene en cuenta únicamente la deformación del ligando, y $r^2 = 0{,}38$ si se incluye además la deformación de la proteína (Fig. 8.3). Usando en cambio el FF AMBER ff03 (Tabla 8.2), la correlación mejora en ambos casos, $r^2 = 0{,}51$ para ΔG_{calc4} y $r^2 = 0{,}49$ para ΔG_{calc5} (Fig. 8.4).

Los últimos resultados de correlación encontrados con todos los términos del *score* incluidos indican que la estimación del término entrópico con la aproximación de un único pozo de potencial armónico, no es completamente adecuada. Un sistema biológico puede tener numerosos mínimos accesibles sobre la hipersuperficie de energía potencial, con lo cual la aproximación empleada no describe correctamente el comportamiento del sistema.

En la Tabla 8.3 se presentan los valores de r^2 para cada uno de los protocolos implementados con el Hamiltoniano PM7, Se observa, como se ha mencionado, que considerando únicamente el término de energía de interacción, la correlación es mejor que si se incluyen los términos restantes. Una gran disminución en los valores de correlación se produce al incorporar la componente entrópica ($r^2 = 0{,}44$ para ΔG_{calc2} y $r^2 = 0{,}38$ para ΔG_{calc3}). Eliminando nuevamente los complejos con mayor desviación mencionados anteriormente, para ΔE_{int} se obtiene una correlación $r^2 = 0{,}89$.

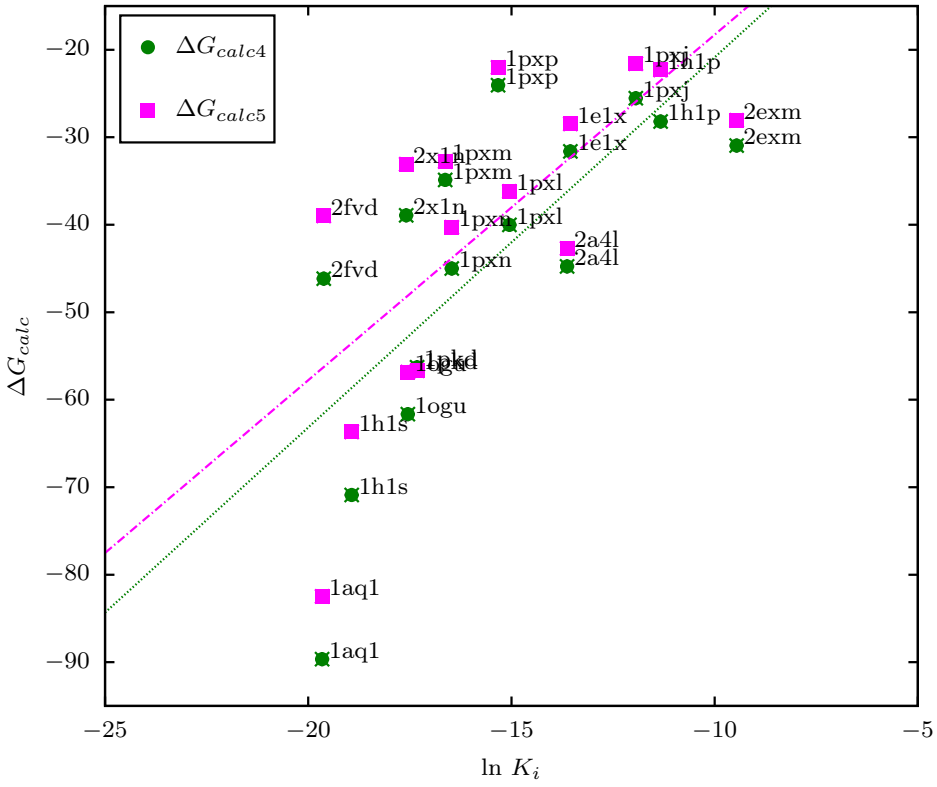


Figura 8.4: Correlación entre el *score* calculado como ΔG_{calc4} y ΔG_{calc5}, y constantes de inhibición (ln K_i) en kcal/mol, con $r^2 = 0{,}51$ y $r^2 = 0{,}49$, respectivamente.

Tabla 8.2: Valores experimentales y términos calculados para 15 complejos proteína-ligando de CDK2, con el Hamiltoniano PM7, empleando el FF AMBER ff03 para el término entrópico (ΔG_{calc4} y ΔG_{calc5}).

Complejo	ΔG_{calc1}	ΔG_{calc4}	ΔG_{calc5}	lnK_i
1aq1	-86,0	-89,7	-82,5	-19,66
1e1x	-55,1	-31,6	-28,5	-13,55
1h1p	-50,5	-28,2	-22,3	-11,33
1h1s	-81,6	-70,9	-63,7	-18,93
1ogu	-79,1	-61,7	-56,8	-17,55
1pkd	-80,0	-56,3	-56,7	-17,32
1pxj	-44,2	-25,5	-21,6	-11,94
1pxl	-57,8	-40,0	-36,2	-15,05
1pxm	-67,1	-34,9	-32,8	-16,63
1pxn	-72,6	-45,0	-40,3	-16,47
1pxp	-58,3	-24,1	-22,1	-15,33
2a4l	-58,7	-44,8	-42,7	-13,63
2exm	-46,9	-31,0	-28,1	-9,46
2fvd	-75,7	-46,2	-39,0	-19,62
2x1n	-52,7	-38,9	-33,1	-17,59

Una vez que fue evaluada la correlación con los distintos protocolos implementados para el cálculo del *score* cuántico procedimos a recortar el sistema en estudio, a fin de

Tabla 8.3: Valores de correlación entre *score* y constantes de inhibición para 15 inhibidores de la proteína CDK2, Cálculos efectuados con PM7.

Score	$\mathbf{r}^2$	
	total	sin *outliers*
ΔE_{int}	0,74	0,89
ΔG_{calc1}	0,72	0,87
ΔG_{calc2}	0,44	0,74
ΔG_{calc3}	0,38	0,67
$\Delta E_{int} + \Delta G_{def} + \Delta G_{np}$	0,66	0,87

reducir el costo computacional implicado en la optimización de geometrías y evaluación de energía con métodos SQM en los complejos. Para disminuir el número total de átomos se consideraron como parte del sistema únicamente aquellos residuos de la proteína que tuvieran al menos un átomo a una distancia igual o menor a 8 Å del inhibidor. El conjunto de todos los residuos encontrados de esta forma se usó para definir un sitio de unión común a todos los complejos. Luego, se recortó cada uno de los complejos según la región previamente delimitada, eliminando del cálculo el resto de la proteína. Seguidamente, se realizó una optimización de 100 ciclos para el nuevo sistema reducido, es decir los átomos del sitio de unión de la proteína junto con los de cada inhibidor. Los resultados hallados con este procedimiento, se encuentran en las Tablas 8.4-8.5.

Tabla 8.4: Valores de los términos calculados para 15 complejos proteína-ligando de CDK2, sobre el sistema reducido, con el Hamiltoniano PM7.

Complejo	ΔE_{int}	ΔG_{np}	ΔG_{calc1}	$T\Delta S_{ff09}$
1aq1	-82,3	-1,7	-79,7	-16,5
1e1x	-59,7	-1,2	-59,4	-12,4
1h1p	-54,1	-1,2	-53,8	-18,0
1h1s	-69,9	-1,6	-69,0	-25,4
1ogu	-75,9	-1,0	-75,1	-21,4
1pkd	-79,4	-1,7	-76,5	-17,7
1pxj	-34,6	-1,0	-34,6	-20,3
1pxl	-64,5	-1,5	-64,8	-27,9
1pxm	-64,1	-1,4	-63,7	-38,6
1pxn	-65,4	-1,5	-65,2	-25,8
1pxp	-72,3	-1,5	-72,4	-19,7
2a4l	-63,1	-1,7	-62,8	-0,5
2exm	-46,9	-1,0	-46,4	-18,8
2fvd	-74,4	-1,7	-72,7	-24,6
2x1n	-58,5	-1,5	-57,9	-13,7

Tabla 8.5: Valores de correlación entre *score* y constantes de inhibición para 15 inhibidores de la proteína CDK2. Cálculos efectuados con PM7, empleando el FF AMBER ff09 para el término entrópico (ΔG_{calc2} y ΔG_{calc3}) y AMBER ff03 (ΔG_{calc4} y ΔG_{calc5}).

Complejo	ΔG_{calc2}	ΔG_{calc3}	ΔG_{calc4}	ΔG_{calc5}	$\ln K_i$
1aq1	-63,2	-60,7	-83,3	-80,8	-19,66
1e1x	-47,0	-40,7	-35,9	-29,6	-13,55
1h1p	-35,8	-31,6	-31,5	-27,2	-11,33
1h1s	-43,6	-36,7	-58,3	-51,3	-18,93
1ogu	-53,7	-50,5	-57,7	-54,4	-17,55
1pkd	-58,8	-56,7	-52,8	-50,7	-17,32
1pxj	-14,3	-4,2	-15,9	-5,8	-11,94
1pxl	-37,8	-35,1	-46,9	-44,3	-15,05
1pxm	-25,2	-21,7	-31,5	-28,0	-16,63
1pxn	-39,4	-37,1	-37,6	-35,3	-16,47
1pxp	-52,7	-45,8	-38,2	-31,3	-15,33
2a4l	-62,2	-55,7	-48,8	-42,2	-13,63
2exm	-27,5	-14,3	-30,4	-17,2	-9,46
2fvd	-48,1	-44,4	-43,2	-39,6	-19,62
2x1n	-44,2	-27,1	-44,1	-27,0	-17,59

Tabla 8.6: Valores de correlación entre *score* y constantes de inhibición para 15 inhibidores de la proteína CDK2. Cálculos efectuados con el Hamiltoniano PM7. Se usaron los radios por defecto de MOPAC.

Score	$\mathbf{r}^2$	
	total	sin *outliers*
ΔE_{int}	0,74	0,89
ΔG_{calc1}	0,72	0,87
ΔG_{calc2}	0,44	0,74
ΔG_{calc3}	0,38	0,67
$\Delta E_{int} + \Delta G_{def} + \Delta G_{np}$	0,66	0,87

Tabla 8.7: Valores de correlación entre *score* y constantes de inhibición para 15 inhibidores de la proteína CDK2, Cálculos efectuados con el Hamiltoniano PM7, y radios atómicos optimizados para COSMO para el sistema completo y reducido. La entropía fue calculada con el FF AMBER ff03 y ff09.

Sistema	***Score***	$\mathbf{r}^2$	
		ff03	ff09
Completo	$\Delta E_{int} + \Delta G_{def}(L) + \Delta G_{np} - T\Delta S$	0,51	0,44
	$\Delta E_{int} + \Delta G_{def} + \Delta G_{np} - T\Delta S$	0,49	0,38
Reducido	$\Delta E_{int} + \Delta G_{def}(L) + \Delta G_{np} - T\Delta S$	0,48	0,28
	$\Delta E_{int} + \Delta G_{def} + \Delta G_{np} - T\Delta S$	0,50	0,32

Al realizar el recorte del sistema, los tiempos de cálculo disminuyen considerablemente sin que se observe una disminución notable en la correlación con los valores experimentales. Estos resultados indican un buen compromiso entre calidad y eficiencia, lo que permite considerar la posibilidad de ampliar la metodología del *score* cuántico, a sistemas biomoleculares de interés en el contexto de un cribado virtual automatizado basado en *docking*.

PM6-D3H4

Al igual que en el caso anterior, las estructuras de los 15 inhibidores en complejo con CDK2 fueron evaluadas para asignar un *score* cuántico, usando en este caso el Hamiltoniano PM6-D3H4. Los valores de correlación entre las estimaciones de energías libres de unión y los valores experimentales de constantes de inhibición fueron similares a los encontrados para el Hamiltoniano PM7.

Los resultados de los distintos términos calculados sobre cada estructura se pueden ver en las Tablas 8.8, 8.9 y 8.10. En la Tabla 8.9, la componente de entropía fue calculada con el FF AMBER ff09 mientras que en la Tabla 8.10 se presentan los valores hallados usando el FF AMBER ff03.

Tabla 8.8: Valores de los términos calculados para 15 complejos proteína-ligando de CDK2, con el Hamiltoniano PM6-D3H4.

Complejo	ΔE_{int}	ΔG_{np}	ΔG_{calc1}	$T\Delta S_{ff09}$
1aq1	-65,1	-1,3	-63,4	-16,5
1e1x	-54,4	-0,9	-54,3	-12,4
1h1p	-32,3	-0,9	-32,2	-18,0
1h1s	-67,3	-1,2	-65,0	-25,4
1ogu	-67,0	-1,2	-64,8	-21,4
1pkd	-59,7	-1,3	-59,8	-17,7
1pxj	-42,2	-0,7	-41,4	-20,3
1pxl	-46,5	-1,1	-45,7	-27,0
1pxm	-50,9	-1,1	-49,8	-38,6
1pxn	-67,6	-1,1	-64,1	-25,8
1pxp	-47,2	-1,1	-46,7	-19,7
2a4l	-36,5	-1,2	-34,8	-0,5
2exm	-42,0	-0,8	-42,3	-18,8
2fvd	-72,7	-1,2	-70,2	-24,6
2x1n	-55,2	-1,1	-54,1	-13,7

Tabla 8.9: Valores experimentales y términos calculados para 15 complejos proteína-ligando de CDK2, con el Hamiltoniano PM6-D3H4 incluyendo término entrópico con el FF AMBER ff09.

Complejo	ΔG_{calc2}	ΔG_{calc3}	$\ln K_i$
1aq1	-46,90	-41,46	-19,66
1e1x	-41,93	-38,09	-13,55
1h1p	-14,21	-11,58	-11,33
1h1s	-39,62	-28,74	-18,93
1ogu	-43,46	-38,51	-17,55
1pkd	-42,11	-39,94	-17,32
1pxj	-21,10	-23,21	-11,94
1pxl	-18,70	-13,37	-15,05
1pxm	-11,25	-2,22	-16,63
1pxn	-38,34	-28,39	-16,47
1pxp	-26,97	-22,09	-15,33
2a4l	-34,27	-31,68	-13,63
2exm	-23,50	-21,82	-9,46
2fvd	-45,57	-33,95	-19,62
2x1n	-40,35	-34,59	-17,59

Tabla 8.10: Valores experimentales y términos calculados para 15 complejos proteína-ligando de CDK2, con el Hamiltoniano PM6-D3H4 incluyendo término entrópico con el FF AMBER ff03.

Complejo	ΔG_{calc4}	ΔG_{calc5}	$\ln K_i$
1aq1	-67,0	-61,6	-19,66
1e1x	-30,8	-27,0	-13,55
1h1p	-9,9	-7,3	-11,33
1h1s	-54,3	-43,4	-18,93
1ogu	-47,4	-42,5	-17,55
1pkd	-36,2	-34,0	-17,32
1pxj	-22,7	-24,8	-11,94
1pxl	-27,9	-22,5	-15,05
1pxm	-17,6	-8,6	-16,63
1pxn	-36,5	-26,6	-16,47
1pxp	-12,5	-7,6	-15,33
2a4l	-20,8	-18,2	-13,63
2exm	-26,4	-24,7	-9,46
2fvd	-40,7	-29,1	-19,62
2x1n	-40,3	-34,5	-17,59

Con el Hamiltoniano PM6-D3H4, la correlación encontrada para el término de energía de interacción (ΔE_{int}) con respecto a los valores experimentales de constantes de inhibición es de 0,71, siendo menor al valor de correlación para dicho término con PM7. Esto ocurre de la misma manera para los demás términos (ver Tabla 8.11). Si se incluyen la deformación de la proteína y el término de entropía con FF AMBER ff09, la correlación en este caso deja de ser significativa ($r^2 = 0{,}19$).

Tabla 8.11: Valores de correlación entre *score* y constantes de inhibición para 15 inhibidores de la proteína CDK2. Cálculos efectuados con PM6-D3H4.

Score	$\mathbf{r}^2$	
	total	sin *outliers*
ΔE_{int}	0,71	0,76
ΔG_{calc1}	0,70	0,76
ΔG_{calc2}	0,40	0,61
ΔG_{calc3}	0,19	0,38
$\Delta E_{int} + \Delta G_{def} + \Delta G_{np}$	0,55	0,61

En la Tabla 8.12, se observa que los valores del coeficiente de correlación para los cálculos de *score* que incluyen la variación de entropía con el FF AMBER ff03 son más altos que si se emplea en cambio el FF AMBER ff09, para ambos Hamiltonianos. Por otra parte, la comparación de r^2 entre Hamiltonianos indica que en PM7 la influencia de la deformación de la proteína es despreciable, contrariamente a lo que ocurre con PM6-D3H4.

Tabla 8.12: Valores de correlación entre *score* y constantes de inhibición para 15 inhibidores de la proteína CDK2, empleando los Hamiltonianos PM6-D3H4 y PM7. El término entrópico se calculó con el FF AMBER ff03 y ff09.

Hamiltoniano	***Score***	$\mathbf{r}^2$	
		ff03	ff09
PM6-D3H4	$\Delta E_{int} + \Delta G_{def}(L) + \Delta G_{np} - T\Delta S$	0,51	0,40
	$\Delta E_{int} + \Delta G_{def} + \Delta G_{np} - T\Delta S$	0,34	0,19
PM7	$\Delta E_{int} + \Delta G_{def}(L) + \Delta G_{np} - T\Delta S$	0,51	0,44
	$\Delta E_{int} + \Delta G_{def} + \Delta G_{np} - T\Delta S$	0,49	0,38

Capítulo 9

Función de *Scoring* cuántica: Aplicación

El buen rendimiento de un proceso de *High Throughput Docking* (HTD), está ligado a la determinación precisa de la pose del ligando en el sitio de unión del receptor (calidad de *pose*) y a la correcta discriminación entre ligandos y no-ligandos de una librería química de moléculas (calidad de *ranking*). De este modo, en un HTD de buen rendimiento, se reduce la probabilidad de encontrar falsos positivos en el grupo de *hits* seleccionados para realizar la evaluación experimental.

Se sabe actualmente que los programas de *docking* utilizados en un SBVS pueden predecir la conformación de menor energía de un complejo con un alto grado de confiabilidad. Sin embargo, el desarrollo de funciones de *scoring* rápidas, a la vez que suficientemente precisas, constituye actualmente un área de activa investigación.[10,91]

En este Capítulo se presenta una función de *scoring* basada en MC y su evaluación mediante la determinación del Factor de Enriquecimiento (EF), para 5 sistemas pertenecientes a distintas familias de proteínas, que poseen diferentes características del sitio de unión, presencia de co-factores y moléculas de agua, y factores de enriquecimiento calculados con un método de HTD tradicional. Las poses de las moléculas en el sitio activo del receptor fueron generadas con funciones de *docking* clásicas. Luego, se asignó un *score* calculado con métodos cuánticos a cada una. Con la aplicación de la función de *scoring* cuántica, se construyó el *ranking* para las moléculas de la base de datos y se calculó el factor de enriquecimiento a distintos porcentajes de la lista ordenada de moléculas. Los resultados obtenidos por la nueva función de *scoring* basada en MC, superan en gran medida los valores de enriquecimiento obtenidos con métodos de *docking* tradicionales.

9.1. Metodología

Sistemas y bases de datos

En un estudio retrospectivo que se realiza con el fin de validar un método de *docking*, es importante que los sistemas elegidos sean diversos en cuanto a funcionalidad y características del sitio de unión, si luego se pretende aplicar la metodología a un grupo de sistemas distinto.[97] Por otro lado, se debe contar con una estructura cristalográfica de buena resolución o con una estructura obtenida mediante un modelado por homología.[146,147] También es importante tener en cuenta que los ligandos de la librería química empleada deben, en lo posible, tener diversidad en cuanto a características como flexibilidad, número de posibles enlaces de hidrógeno, polaridad, entre otras.

Siguiendo los criterios antes mencionados, se extrajeron de la base de datos PDB las estructuras de las siguientes proteínas: Quinasa dependiente de Ciclina 2 (Cyclin-Dependent Kinase 2, CDK2), Receptor de Estrógeno α (Estrogen Receptor, ESR1), Ciclooxigenasa-1 (Cyclooxygenase, COX1), Neuranimidasa (Neuranimidase, NRAM) y Proteína de choque térmico 90 α (Heat Shock Protein 90 α, HSP90a). Todas las moléculas de agua y co-factores fueron eliminados, excepto para NRAM (se incluyó un átomo de Ca^{2+} localizado a 8 Å del ligando unido), y HSP90a (fueron incluídas cuatro moléculas de agua cristalográficas).

Las librerías de *docking* para cada proteína están conformadas por un conjunto de ligandos y *decoys*. Estos últimos tienen propiedades físico-químicas similares a las de los ligandos, pero difieren en su estructura 2D. Los ligandos y *decoys* correspondientes a cada receptor fueron extraídos de las bases de datos *Directory of Useful Decoys)* (DUD), NRLiSt o *Directory of Useful Decoys-Enhanced* (DUD-E) de la siguiente manera: CDK2, DUD; ESR1, NRLiSt; COX1, NRAM y HSP90a, DUD-E. En la Tabla 9.1 se presenta el número de ligandos y *decoys* para cada receptor.

Tabla 9.1: Receptores usados en la evaluación de la función de *scoring* presentada en este trabajo de Tesis.

Receptor	Código PDB	Dimensión[1]	Ligandos	Decoys	Ligandos/Decoys
CDK2	1FVV	4900	50	1779	1/36
ESR1	3ERT	4300	133	6555	1/49
COX1	2OYU	8930	156	6935	1/44
NRAM	1B9V	5979	98	6199	1/63
HSP90a	1UYG	3304	88	4848	1/55

[1] Número de átomos del receptor

Generación de poses de docking

Para asignar un *score* a las moléculas de la base de datos, se requiere el conocimiento de la *pose* que adoptan las mismas en el sitio activo del receptor. En la presente Tesis, dicha pose fue generada mediante un *docking* con los programas *Internal Coordinate Mechanics* (ICM) y AutoDock Vina (AD Vina). El primero de ellos utiliza una representación en coordenadas internas, que definen la geometría de la molécula (enlaces covalentes, longitudes de enlace y ángulos planos, ángulos de torsión y seis coordenadas de posición). ICM fue diseñado para predecir conformaciones de baja energía de moléculas a través de una búsqueda conformacional en el espacio torsional (ángulos diedros). Por otra parte, AutoDock Vina es un software de acceso libre y gratuito, que posee una calidad razonable para la predicción de poses y fue desarrollado para realizar *docking* molecular. Ambos programas, emplean un campo de fuerzas clásico y se diferencian en la función de energía de *docking* utilizada para evaluar las distintas conformaciones y en la función de energía de *scoring* empleada.

Para realizar el *docking* se utilizaron los parámetros predeterminados tanto en ICM como en AutoDock Vina, considerando proteína rígida en ambos casos. Para cada molécula se seleccionó la conformación de menor energía de *docking* para la re-evaluación con la nueva función de *scoring* cuántica, excepto para el análisis de poses realizado con el programa AutoDock Vina, para el cual se utilizaron las primeras tres o cinco poses de más baja energía de cada molécula.

Minimización del complejo (PL) y estado libre de proteína (P) y ligando (L)

Luego de realizar el *docking* con ICM, se efectuó una minimización de energía en dicho programa para los complejos proteína-molécula pequeña. Para cada receptor, todos los ángulos diedros de los aminoácidos que se encuentran a 4 Å de cualquiera de los ligandos de la librería química correspondiente, fueron considerados libres. Luego, para cada complejo PL se realizó una minimización de energía, imponiendo una restricción armónica sobre los átomos pesados con respecto a la conformación inicial. Posteriormente se realizó una minimización de energía de la proteína y molécula pequeña aisladas, para generar los estados libres o no unidos.

Recorte del sistema

Con el objetivo de reducir el costo computacional que involucran los cálculos cuánticos, se definió un sistema reducido para cada receptor. Para ello, se incluyeron en una lista todos los aminoácidos comprendidos en una región de hasta 8 Å de cada molécula sobre la que se efectuó el *docking*. Seguidamente, habiendo efectuado una inspección visual, se agregaron otros aminoácidos a dicha lista a fin de evitar rupturas internas de hélices α o láminas β. Finalmente, se generó el sistema reducido eliminando de la estructura original del receptor todos los aminoácidos que no estuvieran incluidos en la lista, agregando un átomo de hidrógeno a los átomos de C y N terminales.

Cálculos de Mecánica Cuántica

Los cálculos de MC fueron realizados con los Hamiltonianos semiempíricos PM6-D3H4 y PM7, usando el programa MOPAC2012 junto con su módulo de escalamiento lineal MOZYME. La contribución a la energía de solvatación en entorno acuoso fue calculada usando el modelo de solvente continuo COSMO, con radios atómicos predeterminados.

Entropía

El cambio de entropía conformacional, originado por la unión de la molécula pequeña al receptor, fue estimado como

$$\Delta S = -R\ln\Omega \tag{9.1}$$

donde se asume que luego de dicha unión, la molécula adopta una única conformación (por lo tanto $S_{unido} = 0$) y Ω es el número de conformaciones en el estado libre.

En este trabajo de Tesis se emplearon dos tipos de aproximaciones para estimar el número de conformaciones en el estado libre. La primera es comúnmente usada en la comunidad, y establece que el cambio de entropía del ligando al pasar de la conformación libre en solución a la que adquiere en el sitio de unión de la proteína, es proporcional al número de enlaces rotables que posee dicha molécula.[148–150] Por lo tanto $\Omega = 3^N$ de modo que

$$\Delta S_{rotb} = -R\ln(3)N_{rotb} \tag{9.2}$$

La segunda aproximación utilizada consiste en realizar una búsqueda conformacional para el ligando libre, de modo tal que la variación de entropía entre las conformaciones de ligando libre (N_{conf}) y la conformación asume el mismo en el sitio de unión de la proteína,

está dada por

$$\Delta S_{nconf} = -R\ln(N_{conf}) \tag{9.3}$$

Para obtener el número de conformaciones se realizó una búsqueda de Monte-Carlo con minimización local de energía en el espacio torsional usando ICM, recolectando todas las conformaciones distintas dentro de un rango de energía de 3 kcal/mol. Esta última aproximación fue empleada debido a que la primera sobre estima el número de conformaciones de baja energía, y por lo tanto la entropía.[56]

9.1.1. Función de *Scoring*

La función de *scoring* cuántica *Semiempirical Scoring-COSMO* (SSC) presentada en este trabajo de Tesis se define como,

$$SSC = \Delta E_{int} + \Delta G_{def}(P) + \Delta G_{def}(L) + \Delta G_{np} - T\Delta S \tag{9.4}$$

La Ec. 9.4 es idéntica a la Ec. 6.12, cuyos términos han sido previamente definidos (ver Ec 6.13, Ec. 6.14, Ec. 2.18, Ec. 9.2 y Ec. 9.3).

Se definieron dos tipos de funciones de *scoring*, según la estructura tomada como referencia para realizar los cálculos de energía i) SSC1, los cálculos cuánticos se realizaron directamente sobre los complejos PL del *docking*, es decir sobre la estructura sin minimizar y ii) SSC2, se realizó una minimización clásica sobre la estructura del *docking*, y se calculó la energía sobre dicha estructura.

Cuando las contribuciones de deformación de proteína y ligando se encuentran incluídas en el cálculo de SSC (segundo y tercer término de la Ec. 9.4), se agrega el subíndice "d" ($SSC1_d$ y $SSC2_d$).

En la Fig 9.1 se esquematiza de manera simplificada la metodología de *docking* con la incorporación de las funciones de *scoring* cuánticas utilizadas.

Figura 9.1: Esquema de la metodología de *docking-scoring* utilizada en la presente Tesis.

9.2. Resultados

Se eligieron cinco sistemas pertenecientes a distintas familias de proteínas, que poseen diferencias en las propiedades del sitio de unión, en la presencia de co-factores y moléculas de agua cercanas al sitio de unión, y en el factor de enriquecimiento al 1 % calculado luego de un *docking* con AutoDock Vina (Tabla 9.2).

Tabla 9.2: Proteínas usadas para la evaluación de las funciones de *scoring*.

Código PDB	**Co-factor**[1]	**Moléculas de agua**[1]	$\mathbf{EF}_1(\mathbf{EF}_2)^2$
CDK2	-	-	8,0 (5,0)
ESR1	-	-	16,5 (11,0)
COX1	-	-	1,3 (0,7)
NRAM	Ca^{2+}	-	0,0 (0,0)
HSP90a	-	4	0,0 (0,0)

[1] Dentro de una distancia de 4 Å del ligando cristalográfico
[2] Factor de enriquecimiento al 1 % y 2 % correspondientes al *docking* con AutoDock Vina.

La calidad del *scoring* depende fuertemente de que la evaluación del *score* se realice sobre una pose de *docking* correcta. Teniendo en cuenta esta consideración, en este trabajo de Tesis se usaron las dos funciones de *scoring* cuánticas, para calcular el *score* de los complejos PL obtenidos por un *docking* clásico realizado con el programa ICM, debido a que éste genera poses de buena calidad para proteína-molécula.[152] Por otro lado, a fin de comprobar si la baja calidad de la función de *docking* de AutoDock Vina se debe a que la conformación de mínima energía de *docking* (primera pose) no corresponde a la pose correcta, se tomaron para los receptores CDK2, ESR1 y COX1 las primeras 3 y 5 poses del *docking* de cada molécula, para evaluar el *score* en las mismas. Los resultados se encuentran más adelante, en el presente Capítulo.

El *score* de la Ec. 9.4 fue aplicado en primer lugar, sobre la estructura del complejo proteína-ligando resultante del *docking* clásico de ICM, para cada compuesto de la librería química, usando los dos tipos de funciones de *scoring* cuánticas SSC1 y SSC2.

Debido a que usualmente las dimensiones del receptor son muy grandes para realizar cálculos cuánticos de energía, se realizó un recorte sobre el sistema a fin de evaluar SSC sobre los mismos. Este sistema reducido, está conformado por todos los átomos del sitio de unión, definido como la región que comprende a la molécula pequeña y todos los amino ácidos de la proteína que posean al menos un átomo pesado a una distancia de 8 Å de dicha molécula. Esto equivale a reducir aproximadamente 4 veces la dimensión del sistema.

El proceso general seguido para el cálculo del *score* sobre el sistema reducido, se encuentra esquematizado en la Fig. 9.2.

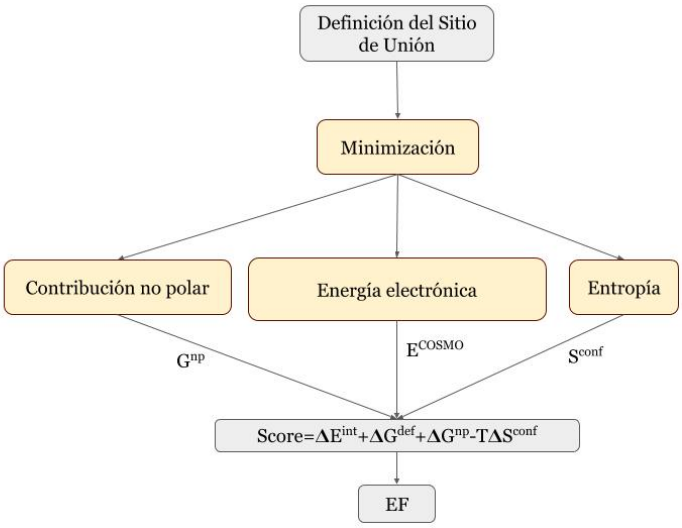
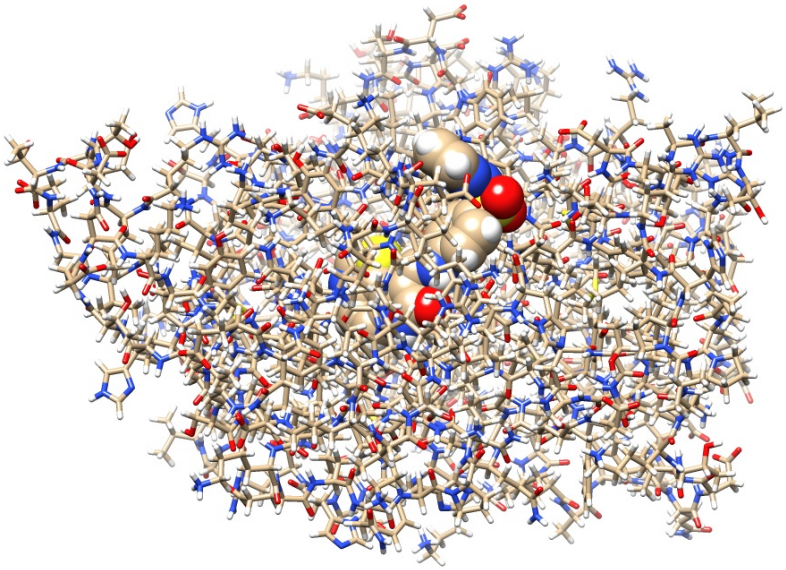
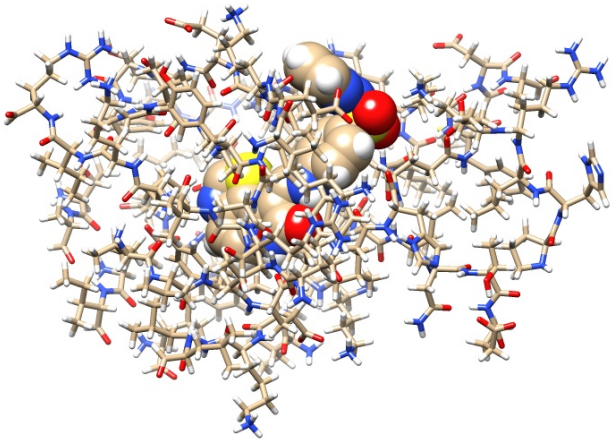


Figura 9.2: Etapa de *scoring* de la metodología desarrollada en la presente Tesis para la evaluación de la función de *scoring* SSC2.

En la Fig. 9.3 se presentan, tomando como ejemplo el receptor CDK2, las representaciones del sistema completo y reducido mencionados anteriormente.



(A) Sistema completo (≈ 5000 átomos) para el complejo proteína-molécula pequeña (Receptor: CKD2).



(B) Sistema reducido (≈ 1800 átomos) para el complejo proteína-ligando CDK2.

Figura 9.3: Representaciones de los sistemas completo (A) y reducido con un radio de corte de 8 Å (B) del complejo proteína-molécula pequeña (Receptor: CDK2).

Para validar la aproximación del recorte del sistema, se evaluó la función de *scoring* cuántica SSC1 para el sistema completo y reducido ($SSC1_c$ y $SSC1_r$, correspondientemente), para los receptores CDK2 y ESR1. Se usaron para el cálculo de energía los Hamiltonianos semiempíricos PM7 y PM6-D3H4 para CDK2 y PM7 para ESR1. En las Tablas 9.3-9.5 se presentan los valores obtenidos para el factor de enriquecimiento, a distintos porcentajes de la base de datos ($\mathbf{EF}_{\%}$).

Tabla 9.3: Valores de factor de enriquecimiento al 1 %, 2 %, 5 %, 10 % y 20 % sobre el conjunto de 1829 moléculas del sistema **CDK2**. Cálculos realizados con el Hamiltoniano **PM7** sobre las poses resultantes del *docking* de ICM.

Función de Scoring	**EF**$_1$	**EF**$_2$	**EF**$_5$	**EF**$_{10}$	**EF**$_{20}$
SSC1$_c$	20,3	13,6	8,8	5,0	2,9
SSC1$_r$	24,1	15,0	9,6	6,0	3,5

Tabla 9.4: Valores de factor de enriquecimiento al 1 %, 2 %, 5 %, 10 % y 20 % sobre el conjunto de 1829 moléculas del sistema **CDK2**. Cálculos realizados con el Hamiltoniano **PM6-D3H4** sobre las poses resultantes del *docking* de ICM.

Función de Scoring	**EF**$_1$	**EF**$_2$	**EF**$_5$	**EF**$_{10}$	**EF**$_{20}$
SSC1$_c$	10,2	10,5	5,9	3,7	2,7
SSC1$_r$	8,1	8,2	6,5	3,7	2,5

Tabla 9.5: Valores de factor de enriquecimiento al 1 %, 2 %, 5 %, 10 % y 20 % sobre el conjunto de 6688 moléculas del sistema **ESR1**. Cálculos realizados con el Hamiltoniano **PM7** sobre las poses resultantes del *docking* de ICM.

Función de Scoring	**EF**$_1$ **(L)**	**EF**$_2$ **(L)**	**EF**$_5$ **(L)**	**EF**$_{10}$ **(L)**	**EF**$_{20}$ **(L)**
SSC1$_c$	31,5	22,5	11	6,1	3,3
SSC1$_r$	35,0	25,4	12,3	6,6	3,5

Contrariamente a lo que ocurre para la evaluación con el Hamiltoniano PM7, en la Tabla 9.4 se puede ver que la función de *scoring* empleando el Hamiltoniano PM6-D3H4 no conduce a resultados prometedores para el sistema CDK2. Al 2 % de la base de datos la función de *scoring* SSC1$_r$ con dicho PM6-D3H4 recupera solamente la mitad de los ligandos recuperados por ICM. Esto indica una baja calidad del método para identificar correctamente ligandos de la librería utilizada. El método PM7 fue seleccionado entonces para la evaluación de las funciones de *scoring* cuánticas en los demás sistemas de estudio.

Como se puede observar en los resultados presentados en las Tablas 9.3-9.5, usar un criterio de corte para reducir las dimensiones del sistema no tiene un impacto importante en los cálculos. Por lo tanto, en este trabajo, se empleó para los demás complejos PL una representación reducida de la proteína para efectuar los cálculos cuánticos correspondientes. En adelante, los resultados de la evaluación de las funciones de *scoring* SSC1 y SSC2 corresponden, en todos los casos, a las efectuadas en el sistema recortado.

La función de *scoring* SSC2, introduce una minimización clásica antes del cálculo cuántico de energía. Esto permite una relajación del complejo PL de *docking* mediante la cual la pose de la molécula el sitio de unión pueda ser corregida. El campo de fuerzas utilizado para la minimización clásica fue MMFF, implementado en ICM.

El primer término de SSC2 puede ser expresado como

$$\Delta E_{int} = E(PL_{8\text{Å}})^{COSMO}_{PL_{min_{4\text{Å}}}} - E(P_{8\text{Å}})^{COSMO}_{PL_{min_{4\text{Å}}}} - E(L)^{COSMO}_{PL_{min_{4\text{Å}}}} \tag{9.5}$$

Usando la misma notación, las contribuciones de deformación de ligando y proteína vienen dadas por

$$\begin{aligned} \Delta G_{def}(P) &= E(P)^{COSMO}_{PL_{min_{4\text{Å}}}} - E(P)^{COSMO}_{P_{min}} \\ \Delta G_{def}(L) &= E(L)^{COSMO}_{PL_{min_{4\text{Å}}}} - E(L)^{COSMO}_{L_{min}} \end{aligned} \tag{9.6}$$

Calculando el enriquecimiento de la base de datos mediante la aplicación de este último protocolo de minimización clásica, se puede observar una notable mejora en los resultados con respecto a la función de *scoring* anterior. Esto se observa especialmente en el caso de los receptores ESR1 y NRAM, donde el $\mathbf{EF}_1$ pasa de 35,0 a 45,0 y 9,2 a 21,4 para las funciones SSC1 y SSC2, respectivamente.

Tabla 9.6: Comparación de valores de enriquecimiento al 1 % [EF_1] de las bases de datos ordenadas, para las funciones de *scoring* cuánticas SSC1 y SSC2, en los sistemas estudiados.

Scoring	$\mathbf{EF}_1$ **($\mathbf{EF}_2$)**				
	CDK2	ESR1	COX1	NRAM	HSP90a
SSC1	24,1 (15,0)	35,0 (25,4)	4,2 (3,2)	9,2 (8,7)	20.8 (14,6)
SSC2	26,3 (16,2)	45,0 (28,8)	2,1 (1,1)	21,4 (15,3)	28,3 (16,1)

A partir de los resultados de la Tabla 9.10 se puede establecer que la relajación de las estructura del complejo proteína-molécula pequeña mediante una minimización, conduce a un mayor enriquecimiento. Es decir que minimizando el sistema se logra recuperar una mayor cantidad de ligandos verdaderos en los primeros lugares de la lista ordenada según el *score*, de la base de datos.

Seguidamente se analizó de manera particular, el efecto que tiene la inclusión de la deformación de la proteína y el ligando en el *scoring*. En las Tablas 9.7-9.9 se presentan los resultados para los receptores CDK2, NEUR y HSP90a.

Tabla 9.7: Valores de factor de enriquecimiento y número de ligandos recuperados al 1 %, 2 %, 5 %, 10 % y 20 % sobre el conjunto de 1829 moléculas del sistema **CDK2**. Cálculos realizados con el Hamiltoniano **PM7** sobre las poses resultantes del *docking* de ICM, con la función de *scoring* SSC2.

Función de Scoring	**EF_1 (L)**	**EF_2 (L)**	**EF_5 (L)**	**EF_{10} (L)**	**EF_{20} (L)**
SSC2	26,3 (13)	16,2 (16)	8,8 (22)	6,0 (30)	3,5 (35)
$SSC2_d$	24,3 (12)	17,2 (17)	9,2 (23)	5,8 (29)	3,6 (36)

Tabla 9.8: Valores de enriquecimiento sobre el conjunto de 6297 moléculas del sistema **NRAM** calculados al 1 %, 2 %, 5 %, 10 % y 20 %. Cálculos realizados con el Hamiltoniano **PM7** sobre las poses resultantes del *docking* de ICM, con la función de *scoring* SSC2.

Función de Scoring	**EF_1 (L)**	**EF_2 (L)**	**EF_5 (L)**	**EF_{10} (L)**	**EF_{20} (L)**
SSC2	21,4 (21)	15,3 (30)	9,0 (44)	6,6 (65)	4,3 (84)
$SSC2_d$	19,4 (19)	16,3 (32)	12,3 (60)	8,0 (78)	4,6 (91)

Tabla 9.9: Valores de enriquecimiento sobre el conjunto de 4936 moléculas del sistema **HSP90a** calculados al 1 %, 2 %, 5 %, 10 % y 20 %. Cálculos realizados con el Hamiltoniano **PM7** sobre las poses resultantes del *docking* de ICM, con la función de *scoring* SSC2.

Función de Scoring	**EF_1 (L)**	**EF_2 (L)**	**EF_5 (L)**	**EF_{10} (L)**	**EF_{20} (L)**
SSC2	28,3 (20)	16,1 (23)	7,2 (26)	4,9 (35)	3,3 (48)
$SSC2_d$	31,1 (22)	16,1 (23)	7,5 (27)	4,4 (32)	3,5 (50)

Analizando el efecto de la inclusión del término de deformación para los receptores, se deduce que la incorporación de dicha contribución no afecta en gran medida la calidad de la función de *scoring* para identificar ligandos. La incorporación del término de deformación del ligando, no aporta una mejora considerable al EF para el receptor CDK2. Dado que el grupo de inhibidores que conforma la librería química de este receptor es más rígido, la consideración de la flexibilidad no influye en gran medida a los resultados de EF. La mejora en los tres receptores se observa para altos porcentajes de la base de datos, (mayores al 2 %) y en HSP90a esto ocurre también para al 1 % y 2 % del ranking, disminuyendo al 5 %.

9.2.1. Valores de EF obtenidos con las distintas funciones de *scoring*

Los resultados del enriquecimiento de las bases de datos, obtenidas para los cinco sistemas correspondientes a las proteínas estudiadas, se presentan en la Tabla 9.10. Dichos valores corresponden a la evaluación efectuada con AutoDock Vina, y con las funciones de *scoring* cuánticas presentadas en esta Tesis. Se omite en la Tabla 9.10 la contribución de entropía, y se presenta más adelante el análisis correspondiente a la inclusión de dicho

término en SSC.

Tabla 9.10: Valores de enriquecimiento al 1 % [EF_1] y 2 % [EF_2, entre paréntesis] de las bases de datos ordenadas, para AutoDock Vina y cuatro esquemas de las funciones de *scoring* cuánticas, para los sistemas estudiados.

Scoring	**EF_1 (EF_2)**				
	CDK2	ESR1	COX1	NRAM	HSP90a
AD Vina	8,0 (5,0)	16,5 (11,0)	1,3 (0,7)	0,0 (0,0)	0,0 (0,0)
SSC1	24,1 (15,0)	35,0 (25,4)	4,2 (3,2)	9,2 (8,7)	20,8 (14,6)
$SSC1_d$	24,1 (18,0)	32,5 (22,9)	5,6 (3,9)	8,2 (9,2)	16,7 (13,2)
SSC2	26,3 (16,2)	45,0 (28,8)	2,1 (1,1)	21,4 (15,3)	27,8 (16,0)
$SSC2_d$	24,3 (17,2)	42,5 (25,8)	2,8 (3,5)	19,4 (15,8)	30,6 (16,0)

Las estructuras de *docking* de los complejos PL empleadas para efectuar el cálculo de SSC2 y $SSC2_d$ fueron previamente minimizadas con un campo de fuerzas de mecánica molecular. Se podría suponer que una minimización cuántica sobre el sistema podría conducir a mejores resultados, sin embargo, el tiempo computacional requerido por dicha metodología haría que la función de *scoring* perdiera su verdadera utilidad.

Para seleccionar los receptores en estudio para este trabajo de Tesis, se tuvieron en cuenta los valores del factor de enriquecimiento calculado para AutoDock Vina en cada uno de ellos. Como se observa en la Tabla 9.10, las funciones de *scoring* cuánticas mejoran notablemente el resultado del EF obtenido con AutoDock Vina. Incluso los cálculos efectuados directamente sobre la estructura del *docking* sin minimización previa, superan ampliamente a los valores arrojados por la función de *docking* de mecánica molecular.

Los resultados de la evaluación de SSC2 permiten inferir que una minimización sobre el complejo proteína-molécula pequeña (incluso si la misma es realizada con un campo de fuerzas de mecánica molecular) permite recuperar un mayor número de ligandos en los primeros puestos de la lista ordenada de la base de datos, generando así un efecto positivo en el cálculo del *scoring* con métodos cuánticos. Continuando con el análisis de SSC2, si se incluye la contribución de la deformación ($SSC2_d$) esto conduce a un empobrecimiento de los resultados para las proteínas ESR1 y NRAM (observando los valores de EF_1). Sin embargo en todos los receptores, excepto en ESR1, el valor del factor de enriquecimiento al 2 % se incrementa, o se mantiene constante, al incorporar la deformación. Lo mismo sucede con EF_1 para CDK2, COX1 y HSP90a. Estos resultados permiten establecer que la inclusión de los términos de deformación en la función de *scoring*, conducen a mejores resultados para el factor de enriquecimiento de la base de datos.

El cambio de entropía conformacional de la molécula pequeña al pasar del estado libre en solución al estado unido luego de la formación del complejo, fue calculado de dos maneras: i) considerando dicho término proporcional al número de enlaces rotables libres

de la molécula (Ω_{conf}=3^N), designado como ΔS_{rot} y ii) calculando Ω_{conf} como el número de conformaciones distintas de baja energía, generadas con un método de búsqueda de Monte Carlo con minimización local de energía. Los resultados del EF incorporando dicho término en las función de *scoring* cuántica SSC2, se presentan en la Tabla 9.11.

Tabla 9.11: Valores de enriquecimiento al 1 % [EF_1] y 2 % [EF_2, entre paréntesis] de las bases de datos ordenadas, para cuatro esquemas de las funciones de *scoring* cuánticas, incluyendo la entropía conformacional del ligando, para los sistemas estudiados.

Scoring	**EF_1 (EF_2)**				
	CDK2	ESR1	COX1	NRAM	HSP90a
$SSC2_d$	26,3 (16,2)	42,5 (25,8)	2,8 (3,5)	19,4 (16,3)	31,1 (16,1)
$SSC2_d$ - $T\Delta S_{conf}$	26,3 (17,2)	41,7 (26,3)	3,6 (3,2)	20,4 (16,8)	31,1 (16,1)
$SSC2_d$ - $T\Delta S_{rot}$	28,4 (19,2)	41,7 (26,7)	2,1 (2,8)	19,4 (16,8)	31,1 (16,1)

A pesar de que la inclusión del término entrópico requiere un mayor costo computacional, y no genera en todos los sistemas un gran impacto en el valor del enriquecimiento, desde un punto de vista teórico es correcto incluir dicha contribución. Se puede observar en la Tabla 9.11 que S_{rot} empeora los valores de EF_1 para ESR1, COX1 y no tiene ningún efecto en CDK2, NRAM y HSP90a. Por el contrario, la entropía con el método de búsqueda conformacional de Monte Carlo S_{conf} tiene un efecto marginal, mejorando levemente los resultados frente a la omisión de dicho término para COX1 y NRAM.

9.2.2. *Scoring* de las poses generadas por *docking* en AutoDock Vina

Se presenta en esta Sección, el análisis de las poses obtenidas por el *docking* realizado en AutoDock Vina. Debido a la baja calidad de la función de *docking* implementada en dicho programa para identificar correctamente los ligandos de la base de datos (Tabla **??**), se analizaron distintas poses para determinar si la pose correcta no es identificada como tal por el programa (primera conformación de la lista determinada según el valor calculado por la función de *docking*) sino que se encuentra más abajo en la lista de conformaciones.

Los resultados se presentan a continuación para los sistemas CDK2, ESR1, y COX1. Para los últimos dos receptores, se tomó un subconjunto de moléculas para realizar la evaluación, manteniendo la misma relación del número de *decoys* por ligandos, que en la base de datos originales. El protocolo de *scoring* cuántico elegido para evaluar la metodología fue SSC1. Se incluyeron en la evaluación del *score* la primera, y luego las tres primeras poses de Vina, para los tres sistemas. En el caso de CDK2, se analizó también el valor alcanzado de EF tomando las 5 primeras poses, cuyos resultados están en la Tabla 9.15.

Tabla 9.12: Valores de enriquecimiento sobre el conjunto de 1829 moléculas del sistema **CDK2** provenientes del *docking* en Vina, calculados al 2 %, 5 %, 10 % y 20 %, con **PM7** para las tres primeras poses, con la función de *scoring* **SSC1**.

Poses	**EF_2 (L)**	**EF_5 (L)**	**EF_{10} (L)**	**EF_{20} (L)**
1	5,9 (7)	4,0 (10)	3,6 (18)	2,5 (24)
3	9,1 (9)	5,7 (14)	3,9 (19)	2,5 (24)

Tabla 9.13: Valores de enriquecimiento sobre el conjunto de 2209 moléculas del sistema **ESR1** provenientes del *docking* en Vina, calculados al 2 %, 5 %, 10 % y 20 %, con **PM7** para las tres primeras poses con **SSC1**.

Poses	**EF_2 (L)**	**EF_5 (L)**	**EF_{10} (L)**	**EF_{20} (L)**
1	5,0 (4)	3,5 (7)	2,8 (11)	1,8 (14)
3	5,0 (4)	2,5 (5)	2,5 (10)	1,6 (13)

Tabla 9.14: Valores de enriquecimiento sobre el conjunto de 2449 moléculas del sistema **COX1** provenientes del *docking* en Vina, calculados al 2 %, 5 %, 10 % y 20 %, con **PM7** para las tres primeras poses con **SSC1**.

Poses	**EF_2 (L)**	**EF_5 (L)**	**EF_{10} (L)**	**EF_{20} (L)**
1	2,0 (3)	1,1 (4)	0,7 (5)	0,5 (8)
3	2,7 (4)	1,9 (7)	1,2 (9)	0,9 (13)

La metodología SSC1 fue elegida por su eficiencia y bajo costo computacional, para los tres sistemas. Se puede ver que en términos generales, la metodología es capaz de identificar un número mayor de ligandos, en el sistema CDK2 y COX1 cuando se considera más de una pose resultante del *docking* en Vina. Para ESR1 en cambio, los resultados tomando 1 o 3 poses no presentan una variación considerable (ver Tabla 9.13).

Siguiendo con este tipo de análisis y tomando CDK2 como sistema de estudio, fueron incluídas las 5 primeras poses en el *scoring* cuántico. Las funciones de *scoring* cuánticas aplicadas en este caso fueron SSC1 y SSC2. Se puede observar en la Tabla 9.15, que el valor de EF disminuye al aumentar a 5 el número de poses. El número de ligandos identificados al 2 % de la base de datos se mantiene constante para SSC1 tomando 3 o 5 poses. Contrariamente, con SSC2 se recuperan 3 ligandos menos al tener en cuenta cinco poses. El número de ligandos al 5 % considerando las tres primeras poses es de 14, y si se consideran las 5 primeras este número se reduce a 12. Se asume entonces que la verdadera pose puede ser identificada entre las tres primeras de la lista. Si aumentamos este valor, esto puede introducir ruido en la metodología.

El porcentaje de verdaderos ligandos identificados luego del *docking-scoring*, a bajos porcentajes de una base de datos con poses de *docking*, debería ser siempre mayor comparada con el porcentaje identificado por una selección aleatoria para un buen método. Si esto se verifica para porcentajes bajos, el impacto se verá reflejado también en los

Tabla 9.15: Valores de enriquecimiento sobre el conjunto de 1829 moléculas del sistema **CDK2** provenientes del *docking* en Vina, calculados al 2 %, 5 %, 10 % y 20 %, con **PM7** para las 5 primeras poses con **SSC1** y **SSC2**.

Poses	**Función de Scoring**	**EF_2 (L)**	**EF_5 (L)**	**EF_{10} (L)**	**EF_{20} (L)**
3	SSC1	9,1 (9)	5,7 (14)	3,9 (19)	2,5 (24)
	SSC2	16,4 (12)	7,2 (17)	4,7 (22)	2,8 (26)
5	SSC1	9,1 (9)	4,9 (12)	3,7 (18)	2,6 (25)
	SSC2	9,1 (9)	5,7 (14)	3,7 (18)	2,6 (25)

porcentajes altos. Cuanto mayor sea el porcentaje de ligandos conocidos encontrados a un cierto porcentaje de la base de datos previamente ordenada según el *score* (es decir a un porcentaje dado del *ranking* construido), mejor será el proceso de CV en cuanto a la habilidad para enriquecer la lista con potenciales ligandos. Los resultados de enriquecimiento para las bases de datos con moléculas *dockeadas* se calcularon para 4 porcentajes, siendo EF_2, EF_5, EF_{10} y EF_{20} los valores del factor de enriquecimiento al 2, 5, 10 y 20 % de la base de datos.

Scoring de poses de ICM

En los sistemas estudiados (CDK2, ESR1, COX1, NRAM y HSP90a), la calidad del proceso de *docking-scoring*, en cuanto a la obtención de un sub-conjunto enriquecido de ligandos, fue superior para el docking realizado con ICM, en comparación con los resultados usando *docking* con AutoDock Vina.

En relación a futuros desarrollos metodológicos, evidentemente el esfuerzo debe ir en la dirección de una búsqueda de menor costo computacional, manteniendo la eficiencia del método. Una mayor calidad del método de *scoring* se puede alcanzar también, realizando una búsqueda conformacional con otros métodos para incluir la deformación del ligando, como por ejemplo con Monte Carlo.

Scoring de poses de Vina

En una segunda etapa, se analizaron los resultados de la aplicación de los protocolos de *scoring* cuántico sobre las poses de AutoDock Vina. Se tomaron las 3 primeras poses para los receptores CDK2, ESR1 y COX1. Luego, siguiendo con el mismo tipo de análisis se amplió a 5 el número de poses, para el sistema CDK2. En base a los resultados obtenidos se pudo observar que la pose correcta puede ser encontrada entre las primeras tres poses de la lista.

En CDK2, el EF mejora tomando las tres primeras poses y realizando un cálculo de

energía con la función de *scoring* cuántica SSC1 (EF_2=9,1), con respecto al EF alcanzado por Vina (EF_2=5,0). Más aún, el *scoring* cuántico con SSC2 mejora los resultados considerablemente a todos los porcentajes evaluados de la librería. En el sistema ESR1 los cálculos cuánticos mejoran considerablemente la calidad del *ranking* de poses realizado por Vina. Contrariamente, para COX1 el impacto de la aplicación de la función de *scoring* en el factor de enriquecimiento, sobre las poses de Vina, es mucho menor al encontrado en CDK2 y ESR1.

Parte IV

Conclusiones

Conclusiones generales y perspectivas

En esta Tesis se presenta el desarrollo de métodos fundados en la formulación de QM, para su aplicación en la descripción de interacciones proteína-ligando en fase solución, en el contexto del descubrimiento de fármacos líderes.

La aplicación de métodos de Mecánica Molecular es una práctica usual en el modelado de sistemas biológicos, y en particular en el área del diseño de drogas. Sin embargo, las aproximaciones involucradas en dichas aplicaciones poseen serias limitaciones para describir correctamente distintos fenómenos que involucran a los electrones. En el modelado de interacciones proteína-ligando, es muy importante considerar los efectos de dispersión, correlación y polarización electrónica. Dichos efectos están incorporados en la descripción de los sistemas dada por métodos fundados en QM por lo que permitirían un incremento sustancial en la precisión que puede ser alcanzada al describir sistemas biológicos y sus interacciones. Una desventaja de los mismos es el alto costo computacional que traen aparejados. En este sentido los métodos semi-empíricos representan un camino atractivo para su aplicación en sistemas biológicos dado que alcanzan una buena precisión con un costo computacional razonable. A pesar de las limitaciones que pueden encontrarse, la aplicación de métodos cuánticos para modelar las interacciones en complejos proteína-ligando representan un campo de investigación actualmente activo que puede conducir a una mayor precisión en el cálculo de energía libre.

El cálculo preciso de la energía libre de unión es esencial para comprender distintos procesos biológicos importantes, como la asociación molecular de proteína-ligando. En este sentido las aproximaciones computacionales representan un gran aporte para los estudios experimentales ya que permiten efectuar cálculos teóricos en tiempos más reducidos y a un costo mucho menor, a partir de los cuales se pueden realizar inferencias acerca de las observaciones. Para que el cálculo teórico sea confiable es necesario determinar la exactitud de la metodología de cálculo empleada. En esta Tesis se realizó una optimización y validación de métodos basados en mecánica cuántica usados para describir sistemas biológicos en solución acuosa, para conducir a una mayor exactitud y eficiencia

computacional en los mismos. Un aporte importante de este trabajo fue, en particular, el desarrollo de una metodología de *scoring* para ser aplicada en un contexto de cribado virtual automatizado.

Para estudiar sistemas biológicos en solución y calcular sus propiedades, es necesario adoptar una metodología que pueda describir con precisión las interacciones que se producen en fase acuosa. Los cálculos de energía libre de unión dependen en gran medida de una correcta descripción de la influencia del solvente sobre las propiedades del soluto. En este trabajo de Tesis se buscó un modelo de solvente que pudiera ser fácilmente incorporado en una metodología basada en métodos semi-empíricos. El modelo de solvente continuo COSMO se eligió entonces por su eficiencia computacional, ya que está implementado en MOPAC con el algoritmo de escalamiento lineal MOZYME, para el estudio de sistemas de miles de átomos.

Por tanto, se realizó en esta Tesis una re-parametrización del modelo COSMO para los Hamiltonianos semi-empíricos RM1, PM6 y PM7, calculando la componente no polar con distintas metodologías. Comparando los resultados obtenidos para los métodos semi-empíricos reparametrizados en la primera parte del trabajo, se puede afirmar que RM1 alcanza una mayor precisión frente a PM6. Continuando con las optimizaciones efectuadas para RM1 y PM7 con el método de minimización local de Powell, se obtuvieron mejores predicciones de energía de hidratación en las moléculas de SAMPL4 con PM7. La ventaja de dicho Hamiltoniano es que fue desarrollado con el objetivo de mejorar las interacciones no-covalentes. PM7 fue optimizado para reproducir geometrías de referencia y calores de formación. Por esta razón se puede afirmar que se obtienen con dicho Hamiltoniano los parámetros atómicos más adecuados para ser implementados en estudios posteriores para la descripción de sistemas biológicos usando el modelo COSMO.

Para moléculas pequeñas neutras, la precisión alcanzada con los nuevos parámetros obtenidos en este trabajo es comparable a la determinada por otros grupos de investigación cuyos resultados fueron publicados en la competencia SAMPL4.[122] Los resultados obtenidos con los métodos presentados en esta Tesis se encuentran dentro del primer tercio de metodologías con mejores medidas estadísticas. Los valores más bajos de MAE obtenidos para la re-parametrización efectuada con el método de Powell corresponden a PM7 con la componente no polar calculada con las metodologías γ_{ef}SASA y γSASA (1,30 y 1,36 kcal/mol, respectivamente). Estas metodologías aproximan la contribución de energía de solvatación no polar como un término proporcional al área superficial accesible al solvente usando un único coeficiente de tensión superficial (γ_{ef}SASA) o diferentes coeficientes de tensión superficial para cada átomo (γSASA). Los valores de MAE encontrados no se alejan de manera significativa de la precisión química, definida usualmente como predicciones termoquímicas comprendidas en el rango de 1 kcal/mol.

Los análisis efectuados para la re-parametrización del modelo de solvente continuo COSMO permiten afirmar que el modelo propuesto es adecuado para efectuar cálculos de energía de solvatación. Los nuevos parámetros optimizados de PM7 pueden ser posteriormente aplicados en el cálculo de energías libres de unión y en la descripción de interacciones proteína-ligando, en el marco del diseño de fármacos asistido por computadoras.

Entre las distintas metodologías desarrolladas para el cálculo computacional de energías de unión, los métodos de puntos extremos permiten hallar valores que ofrecen el mejor balance entre eficiencia y costo computacional, al considerar en los cálculos únicamente los estados final e inicial del sistema. En esta Tesis se extendió la aplicación de la metodología MM/QM-COSMO para distinguir el modo de unión correcto de un compuesto con probada actividad contra el virus del Dengue, en un proceso de SBVS, para el cual fueron encontradas dos poses de *docking* diferentes con energías similares. La identificación de la pose más favorable es un paso importante en el proceso de optimización del candidato líder. La descripción correcta de las interacciones involucradas permitió determinar la conformación de menor energía con mayor precisión que la arrojada por métodos de mecánica clásica.

Una herramienta que ha ido tomando particular interés en el área del diseño de fármacos asistido por computadoras es el cribado virtual automatizado basado en la estructura del receptor (SBVS). Este tipo de metodologías permite pre-seleccionar moléculas de una librería química, en base a la probabilidad que posean las mismas de unirse al receptor, de acuerdo a una función de *scoring*. Dichas moléculas continúan el proceso del desarrollo de un nuevo fármaco líder, finalizando con la evaluación experimental para confirmar su actividad. De esta forma, se impone un filtro que permitiría ahorrar costos y tiempo en el proceso de descubrimiento de fármacos. El objetivo de la función de *scoring* es entonces incrementar el número de potenciales ligandos en la lista de *hits*, para priorizar dichas moléculas para su evaluación experimental. El notable incremento en los estudios de investigación orientados a definir nuevas funciones de *scoring*, más eficientes que las actualmente empleadas por los programas de *docking*, sugirieron la necesidad de desarrollar una nueva metodología que permita enriquecer la lista de *hits* en un proceso de SBVS. Las numerosas investigaciones referidas al estudio de sistemas biomoleculares con métodos fundados en QM, y los estudios realizados por el grupo de trabajo en el cual se desarrolló este trabajo de Tesis, dieron origen a uno de los ejes centrales de la misma, esto es, a la formulación de una función de *scoring* cuántica basada en métodos SQM en combinación con el modelo de solvente continuo COSMO.

La forma funcional del *score* fue sugerida por un estudio realizado sobre un sistema proteína-ligando con valores experimentales de constantes de inhibición, para calcular

energía libre de unión con métodos cuánticos. Para esta etapa de la investigación se eligió la proteína CDK2 en complejo con 15 inhibidores conocidos con datos experimentales de estructuras cristalográficas y constantes de inhibición. El receptor elegido es un blanco terapéutico atractivo que pertenece a la familia de proteínas Kinasas,[153] las cuales están involucradas en muchos procesos celulares por lo que una desregulación o mutación en las mismas puede asociarse a muchas enfermedades. Una serie de estudios realizados con anterioridad a esta Tesis sobre dicho receptor, usando métodos de mecánica clásica, demostró una baja correlación entre las estimaciones de energía libre de unión realizadas y los valores experimentales de actividad.[154–156] Más aún, en el primero de dichos estudios no se encontró una relación estadística significativa entre las energías calculadas y la actividad siendo muy bajo el valor del coeficiente de correlación alcanzado (r^2 =0,15).[154] Otras metodologías presentaron una correlación de hasta r^2 =0,69.[156] En esta Tesis se usó una metodología basada en métodos semi-empíricos, la cual incluye distintos términos con significado físico, para calcular energías libres de unión. La descripción de interacciones de dispersión, que son importantes en este sistema de estudio,[157,158] deben ser incluidas a través de modificaciones en los Hamiltonianos semi-empíricos.[56] Por otra parte, se agregaron las contribuciones de energía de deformación, desolvatación y entropía. No fueron utilizados en este trabajo parámetros empíricos para ajustar el cálculo de energía, lo que hace posible la aplicación de la metodología a diferentes tipos de complejos proteína-ligando. Una continuación de los estudios realizados sobre este sistema puede ser llevada a cabo para incluir los parámetros atómicos optimizados obtenidos en esta Tesis, a fin de validar la aplicación de los mismos a sistemas biológicos e incrementar la exactitud del cálculo de energía libre.

Se realizó una minimización con métodos SQM sobre las geometrías de cada complejo, dejando fijos los átomos localizados a una distancia mayor a 8 Å del ligando. Posteriormente, se aplicaron distintos protocolos para el cálculo de energía libre de unión de cada complejo. Una vez determinados dichos valores se midió la correlación con los valores experimentales de constantes de inhibición. Para el Hamiltoniano PM7, se encontraron buenos valores de correlación considerando en el cálculo únicamente la energía de interacción electrostática ($r^2 = 0{,}74$), despreciando contribuciones como las energías de deformación y entropía. Agregando la deformación del ligando, este valor no sufrió grandes modificaciones ($r^2 = 0{,}72$). Esto puede estar relacionado con inhibidores poco flexibles. Es notable que al eliminar los dos *outliers* en dichos cálculos la correlación alcanzada es de $r^2 = 0{,}87$. Este mismo valor de correlación arroja el cálculo de energía libre incluyendo deformación de ligando y proteína, eliminando *outliers*. Con el Hamiltoniano PM6-D3H4 los resultados fueron muy similares a los de PM7, alcanzando un valor de correlación de $r^2 = 0{,}71$ para el término de interacción electrostática y de $r^2 = 0{,}70$ si se incluye deformación del ligando. La incorporación del término de deformación de proteína produjo una disminución en los

valores de correlación ($r^2 = 0{,}55$). En ambos Hamiltonianos se observó una disminución considerable de la correlación entre el cálculo de energía libre y las constantes de inhibición, al incluir el término entrópico determinado con el campo de fuerzas clásico ff03 ($r^2 = 0{,}51$ para PM7 y PM6-D3H4). Se puede concluir que la aproximación empleada para calcular dicho término es poco precisa para describir correctamente el comportamiento de los sistemas estudiados.

Un estudio relevante que surgió a partir de esta aplicación, es la determinación del impacto de la contribución del cambio de entropía entre las conformaciones libres y unida, de proteína y ligando en un complejo, en el *score* cuántico. En el cálculo de energía libre para el receptor CDK2 en complejo con 15 inhibidores, se utilizó una aproximación armónica, calculando la variación de entropía con campos de fuerza de mecánica clásica. Los resultados de correlación hallados, indicaron que la inclusión del término entrópico siguiendo dicha aproximación no es adecuada para sistemas biomoleculares. Por este motivo, se modificó la manera de calcular la contribución entrópica. En este caso se tuvo en cuenta la entropía, como un término proporcional a la variación en el número de enlaces rotables. Se observó que la inclusión de la componente entrópica no genera un impacto notable en los resultados. Ésto indica que en este sistema, las variaciones de entropía configuracional tienen menor importancia frente a la entropía vibracional, en acuerdo con los estudios previamente realizados para otros sistemas.[12]

Posteriormente se desarrolló una nueva función de *scoring* cuántica, teniendo en cuenta los resultados hallados para el receptor CDK2, y se aplicó la misma en un estudio retrospectivo de SBVS para cinco sistemas de interés en la industria farmacéutica. Seguidamente se calculó el factor de enriquecimiento (EF) el cual es una medida de la calidad del método para identificar correctamente los ligandos de la base de datos. Un aporte importante de esta metodología es que la reducción realizada en el sistema para efectuar una minimización contribuye en gran medida a la disminución de tiempos de cómputo requerido, sin perder la precisión alcanzada considerando el sistema total.

Para evaluar la calidad de la función de *scoring*, se aplicó la misma sobre las poses resultantes del *docking* generadas con dos programas diferentes, ICM y AutoDock Vina. Se puede afirmar que la determinación correcta de la pose tiene una influencia significativa en la precisión alcanzada por la función de *scoring*. Es decir, una pose mal identificada por la función de *docking*, puede originar que un ligando sea descartado por la consecuente asignación de un bajo *score*.

Se presenta como desarrollo central de este trabajo de Tesis, una nueva función de *scoring* basada en mecánica cuántica, para su aplicación en el cribado virtual automatizado de alto rendimiento (HTD). Cuatro esquemas de dicho *scoring* cuántico ($SSC1$, $SSC1_d$, $SSC2$ y $SSC2_d$) fueron evaluados en cinco sistemas pertenecientes a distintas

familias de proteínas, con características diferentes en cuanto al sitio de unión, presencia de co-factores, moléculas de agua, y valores de enriquecimiento calculados con una función de *docking* estándar. Las estructuras de referencia sobre las que se evaluó el *score* fueron las resultantes de un *docking* de ICM. La evaluación de las funciones de *scoring* cuánticas condujo a una mejora sustancial en cuanto al valor del enriquecimiento encontrado, en comparación con los obtenidos por la función de *docking* estándar de AutoDock Vina. Los cálculos realizados incluso con la función de *scoring* más simple ($SSC1$) muestran excelentes resultados en términos del número de ligandos de la base de datos recuperados al 1 % y 2 %. Los resultados son incluso mejores si se incorpora en la metodología una etapa de minimización clásica sobre las estructuras de *docking* (función de *scoring* SSC2). La contribución de los términos de deformación dan cuenta del cambio sufrido por el receptor y la molécula pequeña al pasar del estado libre en solución al estado unido en el complejo, por lo que deben ser incorporados en la función de *scoring*. En dos de los cinco sistemas estudiados se observa un pequeño empobrecimiento del enriquecimiento cuando se incluyen dichos términos.

Para mejorar la metodología utilizada para modelar sistemas biológicos con métodos cuánticos en esta Tesis se deben tener en cuenta distintas consideraciones: i) la extensión de la evaluación de las funciones de *scoring* cuántica a un mayor número de sistemas; ii) una comparación con otras funciones de *scoring* basadas en Mecánica Molecular; iii) la implementación de distintas estrategias de relajación del sistema; iv) la comparación de los resultados empleando un modelo de solvente continuo distinto; v) un tratamiento más adecuado para la aplicación de métodos cuánticos en sistemas biológicos debe ser desarrollado para el cálculo de entropía.

En términos de tiempo computacional requerido, la función de *scoring* cuántica presentada en este trabajo de Tesis es aproximadamente 10 veces más lenta que la puntuación con un *score* basado en métodos estándar de mecánica molecular. A pesar de ello, los resultados obtenidos para los cinco receptores estudiados pertenecientes a distintas familias de proteínas remarcan el gran potencial del *score* basado en métodos cuánticos. El desarrollo de funciones de *scoring* basadas en métodos cuánticos para su aplicación en procesos de HTD, permitiría proveer de métodos sumamente precisos a la identificación y optimización de moduladores de moléculas pequeñas para receptores relevantes en la industria farmacéutica. Estos desarrollos pueden ser justificados, considerando que los futuros desarrollos teóricos en la mecánica cuántica, así como en algoritmos y hardware computacionales, pueden conducir en un futuro próximo a un reemplazo de los campos de fuerza clásicos (FF) por métodos semiempíricos o métodos DFT.[159]

Bibliografía

[1] R.W. Hansen J.A. DiMasi, H.G. Grabowski. Innovation in the pharmaceutical industry: new estimates of rd costs. *J. Health Econ.*, 47:20–33, 2016.

[2] S. S. Phatak, S. Stephan, and C. N. Cavasotto. High-throughput and in silico screenings in drug discovery. *Exp. Opin. Drug. Discov.*, 4:947–959, 2009.

[3] J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott. Principles of early drug discovery. *Br J. Pharmacol.*, 162:1239–1249, 2011.

[4] C. N. Cavasotto and A. J. Orry. Ligand Docking and Structure-based Virtual Screening in Drug Discovery. *Curr. Top. Med. Chem.*, 7:1006–1014, 2007.

[5] S. K. Burger, D. C. Thompson, and P. W. Ayers. Quantum Mechanics/Molecular Mechanics Strategies for Docking Pose Refinement: Distinguishing between Binders and Decoys in Cytochrome c Peroxidase. *J. Chem. Inf. Model*, 51:93–101, 2011.

[6] A. Gimeno, M. J. Ojeda-Montes, S. Tomás-Hernández, A. Ceretto-Massagué, R. Beltrán-Debón, M. Mulero, G. Pujadas, and S. Garcia-Vallvé. The Light and Dark Sides of Virtual Screening: What is There to Know? *Int. J. Mol. Sciences*, 20:1–24, 2019.

[7] W. L. Jorgensen. The many roles of computation in drug discovery. *Science*, 303:1813–1818, 2004.

[8] B. K. Shoichet. Virtual screening of chemical libraries. *Nature*, 432:862–865, 2004.

[9] Z. Wang, H. Sun, X. Yao, D. Li, L. Xu, Y. Li, S. Tian, and T. Hou. Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys. Chem. Chem. Phys.*, 18:12964–12975, 2016.

[10] C. N. Cavasotto. Binding free energy calculations and scoring in small-molecule docking. In Physico-Chemical and Computational Approaches to Drug Discovery. In F. J. Luque and X. Barril, editors, *Physico-Chemical and Computational Approaches to Drug Discovery*, chapter 8, pages 195–222. Royal Society of Chemistry: London, 1st edition, 2012.

[11] N. Hansen and W. F. van Gunsteren. Practical aspects of free-energy calculations: A review. *J. Chem. Theory Comput.*, 10:2632–2647, 2014.

[12] C. A. Chang, W. Chen, and M. K. Gilson. Ligand configurational entropy and protein binding. *Proc. Natl. Acad. Sci.*, 104:1534–1539, 2007.

[13] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc. Chem. Res.*, 33:889–897, 2000.

[14] T. Hou, J. Wang, Y. Li, and W. Wang. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model*, 51:69–82, 2011.

[15] K. Raha and J. K. M. Merz. Structural basis of dielecric permittivity of proteins: insights from quantum mechanics. *Proceedings of the International School of Physics 'Enrico Fermi'*, 165:193–205, 2004.

[16] H. Gohlke, M. Hendlich, and G. Klebe. Predicting binding modes, binding affinities and 'hot spots' for protein-ligand complexes using a knowledge-based scoring function. *Perspectives in Drug Discovery and Design.*, 20:115–144, 2000.

[17] K. Raha and J. K. M. Merz. Large-scale validation of a quantum mechanics based scoring function: predicting the binding affinity and the binding mode of a diverse set of protein-ligand complexes. *J Med Chem.*, 48:4558–4575, 2005.

[18] U. Ryde and P. Söderhjelm. Ligand-Binding Affinity Estimates Supported by Quantum-Mechanical Methods. *Chem. Rev.*, 116:5520–5566, 2016.

[19] N. D. Yilmazer and M. Korth. Recent Progress in Treating Protein-Ligand Interactions with Quantum-Mechanical Methods. *Int. J. Mol. Sci.*, 17:1–12, 2016.

[20] M. G. Aucar and C. N Cavasotto. Molecular Docking Using Quantum Mechanical Methods. *Methods in Molecular Biology*, 2019.

[21] D. Mucs and R. A. Bryce. The application of quantum mechanics in structure-based drug design. *Expert Opin. Drug Discov.*, 8:263–276, 2013.

[22] J. J . P. Stewart. *MOPAC2009*. Computational Chemistry, Colorado Springs, CO, USA, 2009.

[23] J. S. P. Stewart. Application of Localized Molecular Orbitals to the Solution of Semiempirical Self-Consistent Field Equations. *Int. J. of Quant. Chem.*, 58:133–146, 1996.

[24] E. S. Leal, M. G. Aucar, L.G. Gebhard, N. G. Iglesias, M. J. Pascual, J. J. Casal, A. V. Gamarnik, C. N. Cavasotto, and M. Bollini. Discovery of novel dengue virus entry inhibitors via a structure-based approach. *Bioorganic Medicinal Chemistry Letters*, 27:3851–3855, 2017.

[25] I. N. Levine. *Química Cuántica*. Pearson Education, 2001.

[26] A. R. Leach. *Molecular Modelling: Principles and Applications*. Prentice Hall, 2001.

[27] J. A. Pople, D. P. Santry, and G. A. Segal. Approximate Self-Consistent Molecular Orbital Theory. I. Invariant Procedures. *J. Chem. Phys.*, 43:129–135, 1965.

[28] G. B. Rocha, R. O. Freire, A. M. Simas, and J. J. P. Stewart. RM1: a reparametrization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *J. Comp. Chem.*, 27:1101–1111, 2006.

[29] J. J. P. Stewart. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model*, 13:1173–1213, 2007.

[30] J. Řezáč, J. Fanfrlík, D. Salahub, and P. Hobza. Semiempirical Quantum Chemical PM6 Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes Various Types of Noncovalent Complexes. *J. Chem. Theory Comput.*, 5:1749–1760, 2009.

[31] J. Řezáč and P. Hobza. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. *J. Chem. Theory Comput.*, 8:141–151, 2012.

[32] J. J. P. Stewart. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model*, 19:1–32, 2013.

[33] M. J. S. Dewar and W. Thiel. Ground states of molecules, 38. The MNDO method. Approximations and parameters. *J. Am. Chem. Soc.*, 99:4899–4907, 1977.

[34] A. Warshel and M. Levitt. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, 103:227–249, 1976.

[35] H. M. Senn and W. Thiel. QM/MM Methods for Biomolecular Systems. *Angew. Chem. Int.*, 48:1198–1229, 2009.

[36] J. Baker. An algorithm for the location of transition states. *J. Comp. Chem.*, 7:385–, 1986.

[37] L. Verlet. Computer Experiments on Classical Fluids. I. Thermodynamical Propertiesof Lennard-Jones Molecules. *Phys. Rev.*, 159:98–103, 1967.

[38] J. Tomasi, B. Menucci, and R. Cammi. Quantum mechanical continuum solvation models. *Chem. Rev.*, 105:2999–3093, 2005.

[39] J. Tomasi and M. Persico. Molecular Interactions in Solution: An Overview of Methods Based on Continuous Distributions of the Solvent. *Chem. Rev.*, 94:2027–2094, 1994.

[40] A. Klamt and G. Schüürmann. COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and its Gradients. *J. Chem. Soc. Perkin Trans.*, 2:799–805, 1993.

[41] A. Ben-Naim. Standard thermodynamics of transfer. uses and misuses. *J. Phys. Chem.*, 82:792–803, 1978.

[42] A. Ben-Naim and Y. Marcus. Solvation thermodynamics of nonionic solutes. *J. Chem. Phys.*, 81:2016–2027, 1984.

[43] F. J. Luque, C. Curutchet, J. Muñoz-Muriedas, A. Bidon-Chanal, I. Soteras, A. Morrale, J. L. Gelpi, and M. Orozoco. Continuum solvation models: Dissectin the free energy of solvation. *Phys. Chem. Chem. Phys*, 5:3827–3836, 2003.

[44] R. A. Pierotti. A scaled particle theory of aqueous and nonaqueous solutions. *Chem. Rev.*, 76:717–726, 1976.

[45] M. Prevost, I. T. Oliveira, J. P. Kocher, and S. J. Wodak. Free-energy of cavity formation in liquid water and hexane. *J. Phys. Chem.*, 100:2738–2743, 1996.

[46] H. Reiss, H. L. Frisch, E. Helfand, and J. L. Lebowitz. Aspects of the statistical thermodynamics of real fluids. *J. Chem. Phys.*, 32:119–124, 1960.

[47] P. Claverie. *Intermolecular Interactions: From Diatomics to Biomolecules.* ed. B. Pullman, Wiley, 1978.

[48] F. J. Luque, X. Barril, and M. Orozoco. Fractional description of free energies of solvation. *J. Comput. Aided Mol. Des.*, 13:139–152, 1999.

[49] A. Ben-Naim. On the evolution of the concept of solvation thermodynamics. *A. Journal of Solution Chemistry*, 30:475–487, 2001.

[50] V. M. Animisov and C. N. Cavasotto. Hydration Free Energies Using Semiempirical Quantum Mechanical Hamiltonians and a Continuum Solvent Model with Multiple Atomic-Type Parameters. *J. Phys. Chem.*, 115:7896–7905, 2011.

[51] A. K. Bronowska. Thermodynamics of Ligand-Protein Interactions: Implications for

Molecular Design. In J. C. Moreno Piraján, editor, *Thermodynamics. Interaction Studies - Solids, Liquids and Gases*, chapter 1, pages 1–29. InTech, 1st edition, 2011.

[52] R. Baron and J. A. McCammon. Molecular recognition and ligand association. *Annu. Rev. Phys. Chem.*, 64:151–175, 2013.

[53] J. de Heer. *Phenomenological Thermodynamics with Applications to Chemistry.* Prentice Hall, 1986.

[54] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys. J.*, 72:1047–1069, 1997.

[55] Ajay and M. A. Murcko. Computational methods to predict binding free energy in ligand-receptor complexes. *J. of Med. Chem.*, 38:4953–4967, 1995.

[56] V. M. Animisov and C. N. Cavasotto. Quantum Mechanical Binding Free-energy Calculation for Phosphopeptide Inhibitors of the Lck SH2 Domain. *J. Comput. Chem.*, 32:2254–2263, 2011.

[57] Stryer L. Berg JM, Tymoczko JL. Chemical Bonds in Biochemistry. In *Biochemistry*, chapter 1, pages 1–29. New York: W H Freeman, 5th edition, 2002.

[58] L. Yang, C. Adam, G. S. Nichol, and S. L. Cockroft. How much do van der Waals dispersion forces contribute to molecular recognition in solution? *Nature Chemistry*, 5:1006–1010, 2013.

[59] S. Ehrlich, A. H. Goller, and S. Grimme. Towards full quantum-mechanics-based protein-ligand binding affinities. *Chem. Phys. Chem.*, 18:898–905, 2017.

[60] C. N. Cavasotto, N. S. Adler, and M. G. Aucar. Quantum Chemical Approaches in Structure-Based Virtual Screening and Lead Optimization. *Frontiers in Chemistry*, 6:1–7, 2018.

[61] A. V. Ilatovskiy, R. Abagyan, and I. Kufareva. Quantum mechanics approaches to drug research in the era of structural chemogenomics. *Int. J. Quantum Chem.*, 113:1669–1675, 2013.

[62] R. Zwanzig. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.*, 22:1420–1426, 1954.

[63] D. Frenkel and B. Smith. *Understanding Molecular Simulations.* Academica Press, 1996.

[64] A. Warshel and F. Sussman. Toward Computer-Aided-Site-Directed Mutagenesis of Enzymes. *Proc. Natl. Acad. Sci.*, 83:3806–3810, 1986.

[65] R. Zwanzig. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.*, 22:1420–1426, 1954.

[66] R. Zwanzig. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.*, 22:1420–1426, 1954.

[67] J. G. Kirkwood. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.*, 3:300–313, 1935.

[68] D. B. Kitchen, H. Decornez, and J. R. Furr et al. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.*, 3:935–949, 2004.

[69] D. Wilton, P. Willet, K. Lawson, and G. Mullier. Comparison of Ranking Methods for Virtual Screening in Lead-Discovery Programs. *J. Chem. Inf. Comput. Sci.*, 43(2):469–474, 2003.

[70] M. A. Johnson and G. M. Maggiora. *Concepts and Applications of Molecular Similarity.* Wiley, New York, 1990.

[71] T. I. Oprea and H. Matter. Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol.*, 8:349–358, 2004.

[72] H. Geppert, M. Vogt, and J. Bajorath. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.*, 50:205–216, 2010.

[73] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazaikani, and J. P. Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acid Res.*, 40:1100–1107, 2012.

[74] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman. ZINC: a free tool for discovery chemistry biology. *J. Chem. Inf. Model*, 52(7):1757–1768, 2012.

[75] E. Meyer and W. Hamilton. https://www.rcsb.org/.

[76] A. J. W. Orry and R. Abagyan. Preparation and Refinement of Model Protein–Ligand Complexes. . In A. J. W. Orry and R. Abagyan, editors, *Homology Modeling. Methods in Molecular Biology. (Methods and Protocols)*, chapter 16, pages 351–373. Humana Press, 1st edition, 2011.

[77] J. A. Capra, R. A. Laskowski, J. M. Thornton, M. Singh, and T. A. Funkhouser. Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure. *PLoS Comput. Biol.*, 5:1–18, 2016.

[78] A. Roy and Y. Zhang. Recognizing Protein-Ligand Binding Sites by Global Structural Alignment and Local Geometry Refinement. *Structure Cell Press*, 20:987–997, 2012.

[79] C. N. Cavasotto. Binding free energy calculations and scoring in small-molecule docking. In F. J. Luque and X. Barril, editors, *Physico-Chemical and Computational Approaches to Drug Discovery*, chapter 8, pages 195–222. Royal Society of Chemistry, London, 1st edition, 2012.

[80] W. D. Cornell, P. Cieplak, C. Bayly, I. Gould, I. R. Merz, D. M. Ferguson, and P. A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. of A. Chem. Soc.*, 117:5179–5197, 1995.

[81] C. R. Corbeil and N. Moitessier. Modeling Reality for Optimal Docking of Small Molecules to Biological Targets. *Curr. Computer-Aided Drug Design*, 5:241–263, 2009.

[82] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization and dynamics calculations. *J. Comput. Chem.*, 4:187–217, 1983.

[83] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25:1157–1174, 2004.

[84] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161:269–288, 1982.

[85] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a lamarckian genetic algorithm and empirical binding free energy function. *J. Comput. Chem*, 19:1639–1662, 1998.

[86] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson. Autodock4 and autodocktools4: automated docking with selective receptor flexiblity. *J. Comput. Chem*, 16:2785–2791, 2009.

[87] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Talor. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267:727–748, 1997.

[88] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol*, 261:470–489, 1996.

[89] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, D. E. Shaw, M. Shelley, J. K. Perry, P. Francis, and

P. S. Shenkin. Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Med. Chem.*, 47:1739–1749, 2004.

[90] R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin, and D. T. Mainz. Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.*, 49:6177–6196, 2006.

[91] I. A. Guedes, F. S. S. Pereira, and L. E. Dardenne. Empirical scoring functions for structure-based virtual screening: Applications, critical aspects, and challenges. *Front. Pharmacol.*, 9:1–18, 2018.

[92] M. E. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini, and R. P. Mee. Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.*, 11:425–445, 1997.

[93] R. Abagyan. https://www.molsoft.com, 1985.

[94] R. Abagyan, M. Totrov, and D. Kuznetsov. Icm: a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.*, 15:488–506, 1994.

[95] A. N. Jain. Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. *J. Comput. Aided Mol. Des.*, 10:427–440, 1996.

[96] R. Wang, L. Lai, and S. Wang. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided Mol. Des.*, 16:1126–6196, 2002.

[97] S. Huang, S. Z. Grinter, and X Zou. Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys.*, 12:12899–12908, 2010.

[98] H. F. Velec, H. Gohlke, and G. Klebe. DrugScoreCSD-Knowledge-Based Scoring Function Derived from Small Molecule Crystal Data with Superior Recognition Rate of Near-Native Ligand Poses and Better Affinity Prediction. *J. Med. Chem.*, 48:6296–6303, 2005.

[99] I. Muegge. A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. *Perspectives in Drug Discovery and Design*, 20:99–114, 2000.

[100] I. Muegge. PMF scoring revisited. *J. Med. Chem.*, 49:5895–5902, 2006.

[101] P. J. Ballester and J. B. O. Mitchel. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26:1169–1175, 2010.

[102] M. A. Khamis and W. Gomaa. Comparative assessment of machine-learning scoring functions on PDBbind 2013. *Engineering Applications of Artificial Intelligence*, 45:136–151, 2015.

[103] M. Wójcikowski, P. Siedlecki, and P. J. Ballester. Building machine-learning scoring functions for structure-based prediction of intermolecular binding affinity. In Jr. W. de Azevedo, editor, *Docking Screens for Drug Discovery. Methods in Molecular Biology*, chapter 1, pages 1–12. Humana, New York, NY, 1st edition, 2019.

[104] A. Oda, K. Tsuchida, T. Takakura, N. Yamaotsu, and S. Hirono. Comparison of Consensus Scoring Strategies for Evaluating Computational Models of Protein-Ligand Complexes. *J. Chem. Inf. Model*, 46:380–391, 2006.

[105] A. Ciancetta and S. Moro. Protein-Ligand Docking: Virtual Screening and Applications to Drug Discovery. In C. N. Cavasotto, editor, *In Silico Drug Discovery and Design: Theory, Methods, Challenges, and Applications*, chapter 7, pages 189–208. CRC Press, Boca Raton, FL, USA, 1st edition, 2017.

[106] C. R. W. Guimarães. Rescoring Dokcing Predictions. In R. Baron, editor, *Comptuational Drug Discovery and Design*, chapter 17, pages 255–268. Human Press, Salt Lake City, UT, USA, 1st edition, 2012.

[107] T. Hou, J. Wang, Y. Li, and W. Wang. Assessing the performance of the mm/pbsa and mm/gbsa methods. 1. the accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model*, 51:69–82, 2011.

[108] S. Genheden and U. Ryde. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.*, 10:449–461, 2015.

[109] B. Kuhn, P. Gerber, T. Schulz-Gasch, and M. Stahl. Validation and Use of the MM-PBSA approach for Drug Discovery. *J. Med. Chem.*, 48:4040–4048, 2005.

[110] R. Wang, Y. Lu, X. Fang, and S. Wang. An extensive test of 14 scoring functions using the pdbbind refined set of 800 protein-ligand complexes. *J. Chem. Inf. Comput. Sci.*, 44:2114–2125, 2004.

[111] E. Kellenberg, J. Rodrigo, P. Muller, and D. Rognan. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins*, 57:225–242, 2004.

[112] M. D. Cummings, R. L. DesJarlais, A. C. Gibbs, V. Mohan, and E. P. Jaeger. Co-

marison of automated docking programs as virtual screening tools. *J. Med. Chem.*, 48:962–976, 2005.

[113] J. Kirchmair, P. Markt, S. Distinto, G. Wolber, and T. Langer. Evaluation of the performance of 3d virtual screening protocols: Rmsd comparisons, enrichment assessments, and decoy selection-what can we learn from earlier mistakes? *J. Comput. Aided Mol. Des.*, 22:213–228, 2008.

[114] C. N. Cavasotto and R. A. Abagyan. Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.*, 337:209–225, 2004.

[115] F. Spyrakis and C. N. Cavasotto. Open challenges in structure-based virtual screening: Receptor modeling, target flexibility consideration and active site water molecules description. *Arch. of Biochem. and Biophys.*, 583:105–119, 2015.

[116] C. F. Wong. Flexible receptor docking for drug discovery. *Expert Opin. Drug Discov.*, 10:1189–1200, 2015.

[117] J. Michel, J. Tirado-Rives, and W. L. Jorgensen. Energetics of Displacing Water Molecules from Protein Binding Sites: Consequences for Ligand Optimization. *J. Am. Chem. Soc.*, 131:15403–15411, 2009.

[118] C. R. Corbeil and N. Moitessier. Docking Ligands into Flexible and Solvated Macromolecules. 3. Impact of Input Ligand Conformation, Protein Flexibility, and Water Molecules on the Accuracy of Docking Programs. *J. Chem. Inf. Model*, 49:997–1009, 2009.

[119] S. Mitternacht. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research*, 5:1–10, 2016.

[120] A. V. Marenich, C. J. Cramer, and D. G. Truhlar. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B*, 113:6378–6396, 2009.

[121] F. Forti, X. Barril, F. J. Luque, and M. Orozoco. Extension of the MST Continuum Solvation Model to the RM1 Semiempirical Hamiltonian. *J. Comput. Chem.*, 29:578–587, 2007.

[122] M. T. Geballe and J. P. Guthrie. The SAMPL3 blind prediction challenge: transfer energy overview. *J. Comput. Aided Mol. Des.*, 26:489–496, 2012.

[123] M. T. Geballe, A. G. Skillman, A. Nicholls, J. P. Guthrie, and P. J. Taylor. The SAMPL2 blind prediction challenge: introduction and overview. *J. Comput. Aided Mol. Des.*, 24:259–279, 2010.

[124] A. Nicholls, D. L. Mobley, J. P. Guthrie, J. D. Chodera, and C. I. Bayly. Predicting small-molecule solvation free energies: a blind challenge test for computational chemistry. *J. Med. Chem.*, 51:769–779, 2008.

[125] D. L. Mobley, K. L. Wymer, N. M. Lim, and J. P. Guthrie. Blind prediction of solvatation free energies from the SAMPL4 challenge. *J. Comput. Aided Mol. Des.*, 28:135–150, 2014.

[126] J. P. Guthrie. SAMPL4, a blind challenge for computational solvation free energies: the compounds considered. *J. Comput. Aided Mol. Des.*, 28:151–168, 2014.

[127] J.H. Holland. *Adaptation in Natural and Artificial Systems - 2nd ed.* MIT Press, 1992.

[128] M. J. D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer J.*, 7:155–162, 1964.

[129] D. L. Mobley, K. A. Dill, and J. D. Chodera. Entropy and conformational change in implicit solvent simulations of small molecules. *J. Phys. Chem. B*, 112:938–946, 2008.

[130] J.J.P.Stewart. https://pubchem.ncbi.nlm.nih.gov/, 2009.

[131] F. Forti, C. N. Cavasotto, M. Orozco, X. Barril, and F. J. Luque. A multilevel strategy for the exploration of the conformational flexibility of small molecules. *J. Chem. Theory Comput.*, 8:1808–1819, 2012.

[132] L. Sandberg. Predicting hydration free energies with chemical accuracy: The SAMPL4 challenge. *J. Comput. Aided Mol. Des.*, 28:211–219, 2014.

[133] B. A. Elingson, M. T. Geballe, S. Wlodek, C. I. Bayly, A. G. Skillman, and A. Nicholls. Efficient calculation of SAMPL4 hydration free energies using OMEGA, SZYBKI, QUACPACK, and ZAP TK. *J. Comput. Aided Mol. Des.*, 28:289–298, 2014.

[134] A. Nicholls, S. Wlodek, and J. A. Grant. SAMPL2 and continuum modeling. *J. Comput. Aided Mol. Des.*, 24:135–150, 2010.

[135] A. Jakalian, D. Jack, and C. I. Bayly. Fast, efficient generation of high-qualit atomic charges. AM1(BCC model): II. Parametrization and validation. *J. Comput. Chem.*, 23:1623–1641, 2002.

[136] J. Wang, R. Wolf, J. Cladwell, P. Kollman, and D. Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25:1157–1174, 2011.

[137] H. S. Muddana, N. V. Sapra, A. T. Fenley, and M. K. Gilson. The SAMPL4

hydration challenge: evaluation of partial charge sets with explicit-water molecular dynamics simulations. *J. Comput. Aided Mol. Des.*, 28:277–287, 2014.

[138] W. L. Jorgensen. The many roles of computation in drug discovery. *Science*, 303:1813–1818, 2004.

[139] B. Khun and P. A. Kollman. Binding of a diverse set of ligands to avidin and streptavidin: an accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models. *J. Med. Chem.*, 43:3786–3791, 2000.

[140] J. H. Jensen. Predicting accurate absolute binding energies in aqueous solution: thermodynamic considerations for electronic structure methods. *Phys. Chem. Chem. Phys.*, 17:12441–12451, 2015.

[141] M. A. C. Neces, M. Totrov, and R. Abagyan. Docking and scoring with icm: the benchmarking results and strategies for improvement. *J. Comput. Aided Mol Des.*, 26:675–686, 2012.

[142] S. S. Singh. Preclinical pharmacokinetics: an approach towards safer and efficacious drugs. *Curr. Drug Metab.*, 7:165–182, 2006.

[143] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.*, 91:43–56–461, 1995.

[144] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.*, 4:435–447, 2008.

[145] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, 65:712–725, 2006.

[146] C. N. Cavasotto. Homology Models in Docking and High-Throughput Docking. *Curr. Top. Med. Chem.*, 11:1528–1534, 2011.

[147] E. Gatica and C. N. Cavasotto. Ligand and Decoy Sets for Docking to G Protein-Coupled Receptors. *J. Chem. Inf. Model.*, 52:1–6, 2012.

[148] S. D. Pickett and M. J. Sternberg. Empirical scale of side-chain conformational entropy in protein folding. *J. Mol. Biol.*, 231:825–839, 1993.

[149] M. K. Gilson and H. X. Zhou. Calculating of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.*, 36:21–42, 2007.

[150] C. Gao, M-S. Park, and H. A. Stern. Accounting for Ligand Conformational Restriction in Calculations of Protein-Ligand Binding Affinities. *Biophys J.*, 98:901–910, 2010.

[151] B. Q. Wei, W. A. Baase, L. H. Weaver, B. W. Matthews, and B. K. Shoichet. A model binding site for testing scoring functions in molecular docking. *J. Mol. Biol.*, 322:339–355, 2002.

[152] B. D. Bursulada, M. Totrov, R. Abagyan, and C. L. Brooks. Comprative study of several algorithms for flexible ligand docking. *J. Comput. Aided Mol. Des.*, 17:755–763, 2003.

[153] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298:1912–1934, 2002.

[154] M. Otyepka, I. Bartova, Z. Kriz, and J. Koca. Different mechanisms of cdk5 and cdk2 activation as revealed by cdk5/p25 and cdk2/cyclin a dynamics. *J. Biol. Chem.*, 281:7271–7281, 2006.

[155] J. S. Duca, V. S. Madison, and J. H. Voigt. Cross-docking of inhibitors into cdk2 structures. 1. *J. Chem. Inf. Model.*, 48:659–668, 2008.

[156] C. R. W. Guimarães and M. Cardozo. MM-GB/SA rescoring of docking poses in structure-based lead optimization. *J. Chem. Inf. Model.*, 48:958–970, 2008.

[157] M. Otyepka, V. Krystof, L. Havlicek, V. Siglerova, M. Strnad, and J. Koca. Docking-based development of purine-like inhibitors of cyclin-dependent kinase-2. *J. Med. Chem.*, 43:2506–2513, 2000.

[158] M. Otyepka, Z. Kriz, and J. Koca. Dynamics and binding modes of free cdk2 and its two complexes with inhibitors studied by computer simulations. *J. Biomol. Struct. Dyn.*, 20:141–154, 2002.

[159] S. Grimmme and P. R. Schreiner. Computational Chemistry: The Fate of Current Methods and Future Challenges. *Angew. Chem. Int. Ed. Engl.*, 57:4170–4176, 2018.

Trabajos Publicados

- E. S. Leal*, **M. G. Aucar***[1], L. G. Gebhard, N. G. Iglesias, M. J. Pascual, J. J. Casal, A. V. Gamarnik, C. N. Cavasotto and M. Bollini. Discovery of novel dengue virus entry inhibitors via a structure-based approach. *Bioorganic & Medicinal Chemistry Letters*, 27:3851-3855, 2017.
- M. Bollini, E. S. Leal, N. S. Adler, **M. G. Aucar**, G. A. Fernández, M. J. Pascual, F. Merwaiss, D. E. Alvarez and C. N. Cavasotto. Discovery of Novel Bovine Viral Diarrhea Inhibitors Using Structure-Based Virtual Screening on the Envelope Protein E2, *Frontiers in Chemistry*, 6:1-10, 2018.
- C. N. Cavasotto, N. S. Adler and **M. G. Aucar**. Quantum Chemical Approaches in Structure-Based Virtual Screening and Lead Optimization. *Frontiers in Chemistry*, 6:1-7, 2018.
- C. N. Cavasotto, **M. G. Aucar** and N. S. Adler. Computational chemistry in drug lead discovery and design. *International Journal of Quantum Chemistry*, 119:1-19, 2019.
- **M. G. Aucar** and C. N Cavasotto. Molecular Docking Using Quantum Mechanical Methods. *Methods in Molecular Biology*, 2019. Aceptado.

[1]* Estos autores contribuyeron de igual forma a este trabajo. (Ver: `http://www.sciencedirect.com/science/article/pii/S0960894X17306534?via%3Dihub`)