

*A mis padres, Loreto y Jesús M.
A mi amada Gemma*

Per fer les coses ben fetes, és necessari: primer, l'amor; segon, la tècnica"

Antoni Gaudí

Agradecimientos

El presente trabajo ha representado por diversas razones, una etapa amplia y clave de mi vida. Una etapa en la cual no sólo ha sido un aprendizaje de nuevas materias y conocimientos, sino globalmente ha representado una lección mucho más amplia. Desde el aprendizaje aplicado de lo que significa el esfuerzo, la constancia y la perseverancia, pasando por el significado del rigor y como no también al aprendizaje de como intentar equilibrar en tu vida la dedicación a la familia, el trabajo y un proyecto como el éste.

Querría empezar dando las gracias a dos personas que seguramente me dieron la oportunidad y creyeron en que algún día podría llegar el momento de escribir este trabajo. Probablemente sin ellas, quizás no se hubiera iniciado. A la Dra Rosa Claramunt Vallespir por animarme y guiarme en la etapa inicial de empezar este proyecto y por haberme dado ese “empujoncito de ánimo”. Al Dr. Jose Ignacio Borrell por haberme dado la oportunidad y haberse preocupado en buscar un sitio donde ubicar dicho proyecto.

A los directores de la tesis la Dra. Violeta Isabel Pérez y del Dr. Jordi Teixidó y que son los actores principales de este guión, quisiera agradecerles primeramente su paciencia, su dedicación y su amistad a lo largo de estos años. Especialmente a ti Violeta porque además de haber sido un honor y un placer aprender de ti, agradecerte tu dedicación a este proyecto, el cual solamente medido en unidades de e-mails, te ha ocupado largos tiempos de dedicación. Y si a esto añadimos, que la mayor parte de ellos eran enviados a altas horas de la noche, tu dedicación ha sido doble.

A mi familia en general y a mis padres Loreto y Jesús en particular. Especialmente a mi padre al cual querría dedicar esta tesis. Padre: “Tú me has animado, me has acompañado y has seguido el proyecto hasta prácticamente el final, pero para mi desgracia no vas a poderme acompañarme en el punto final del proyecto, momento el cual, tanta ilusión me hacía compartir contigo. Tú fuiste quien me enseñó mis primeros conocimientos de química, por eso sé la ilusión que te hacía este proyecto y a mí me hacía feliz saber de tú ilusión, espero que allá donde estés te llegue ésta dedicatoria y mi cariño”.

A mi amada Gemma por infinitas razones. Tú fuiste la que me empujaste a empezar y la que menos dudas ha tenido en que dicho proyecto lo finalizaría, eres la que “lo has sufrido” durante

el largo tiempo dedicado y eres la que estará a mi lado hasta el último trámite final. “Tu has estat el veritable motoret d’aquest treball i ho ets de la meva vida”, nadie más que tú se merece un “titulín”, por tus enseñanzas del Mendeley, por los artículos, pero también por escuchar un día tras otro mis ideas, dudas y reflexiones, mi stress, mis cambios de humor y siempre sin reproche alguno, y sobre todo por hacerlo sin faltarme tu incansable sonrisa. “A tu i al meu pare us dedico molt especialment aquest treball”.

A los amigos, compañeros que siempre se han interesado por la evolución de mi trabajo y que me han animado siempre que nos veíamos, a todos vosotros también os doy las gracias.

Finalmente quisiera agradecer a las bibliotecarias en general y a Montse en particular, por el trabajo que hacen, por su efectividad y dedicación en el rastreo de artículos por las bases de datos más recónditas existentes. Así mismo, por el soporte continuo en el uso de las herramientas de citación que hacen más llevadero el camino bibliográfico.

A todos vosotros muchas gracias.

Abreviaciones

A_{2A}AR: Receptor Adenosina A_{2A}

AUC: Área bajo la curva ROC (Area Under the Curve)

aAUC: Media de las Áreas bajo la curva ROC (average Area Under the Curve)

ADME: Absorción, Distribución, Metabolismo y Excreción

aRMSD: Promedio de la raíz Cuadrada de la Media de Diferencias al Cuadrado (average Root Mean Square Deviation)

β₂AR: Receptor Adrenérgico β₂

CXCL12: Ligando natural de la subfamilia de las quimiocinas que se une al receptor CXCR4.

DM: Dinámica Molecular

ECL2 : Bucle extracelular n°2. Une por la cara extracelular las hélices 4 y 5 (Extracellular Loop 2)

FFT: Fast Fourier Transform

GEM: Grup d'Enginyeria Molecular

GPCRs: G Protein Coupled Receptors

HIF-1: Factor Inducible por Hipoxia (Hypoxia Inducible Factor 1)

HQSAR: Holographic Quantitative Structure Activity Relationship

HTS: High Throughput Screening

IC50: Concentración inhibitoria cincuenta (Inhibitory Concentration fifty)

ICL1: Bucle intracelular n°1 (Intracellular Loop 1)

ICL2: Bucle intracelular n°2 (Intracellular Loop 2)

ICL3: Bucle intracelular n°3 (Intracellular Loop 3)

pAUC_{x%}: Área parcial bajo la curva ROC al %X de falsos positivos (partial Area Under the Curve)

NSCLC: Cáncer de pulmón de células no pequeñas (Non-Small Cell Lung Cancer)

SCLC: Cáncer de pulmón de células pequeñas (Small Cell Lung Cancer)

SDF-1 (CXCL12): Factor Derivado del estroma-1 (Stromal Cell Derived Factor-1)

SHED: SHannon Entropy Descriptors.

RMSD: Raíz Cuadrada de la Media de Diferencias al Cuadrado (Root Mean Square Deviation)

ROC: Curva ROC (Receiver-operating characteristic curve)

T4L: Lisozima de fusión T4

TAMRA: Carboxitetrametilrodamina

TM: Proteína de Transmembrana.

TOX: Toxicidad

VS: Cribado virtual (Virtual Screening)

WHIM: Síndrome de verrugas, hipogammaglobulinemia, infecciones y mielocatexis (Warts, Hypogammaglobulinemia, Infections, and Myelokathexis syndrome).

Sumario

Los métodos de descubrimiento de nuevos fármacos han evolucionado recientemente gracias a la resolución de las estructuras proteicas las cuales actúan como dianas terapéuticas responsables de enfermedades o desregulaciones biológicas. Estas estructuras proteicas tridimensionales, conjuntamente con el desarrollo de nuevas técnicas computacionales están permitiendo el desarrollo acelerado de nuevos compuestos candidatos a convertirse en fármacos.

El presente trabajo se inicia proponiendo un nuevo método que permita mejorar la elección de compuestos candidatos a ser inhibidores de una “diana difícil” aunque bien conocida, como es el receptor VEGFR-2, partiendo de su estructura tridimensional cristalizada y de compuestos inhibidores conocidos de dicha diana.

La resolución tridimensional de la estructura del receptor CXCR4 mediante cristalografía de rayos X, en el año 2010, ha supuesto un avance importante de cara a mejorar el diseño de compuestos inhibidores del VIH, así como de compuestos antitumorales, enfermedades en las que interviene de forma determinante el receptor CXCR4. Así pues, los modelos de cribado virtual desarrollados anteriormente al 2010 en el laboratorio de diseño molecular del IQS (GEM) han sido generados a partir de modelos creados por homología a otras proteínas GPCRs y/o basados únicamente en la forma de ligandos conocidos. De este modo, partiendo de las diferentes estructuras proteicas publicadas de CXCR4, se ha evaluado cuál de dichas estructuras presenta la conformación que distingue mejor los compuestos antagonistas activos de compuestos inactivos. Además, se han evaluado múltiples métodos de cribado virtual de CXCR4 basados en la estructura, en la forma del ligando y mediante modelos farmacofóricos. Una vez obtenida la mejor estructura de CXCR4 y los mejores métodos de cribado virtual retrospectivo, se realizan cribados virtuales prospectivos de una nueva quimioteca generada combinatoriamente, basada en análogos de estructuras previamente desarrolladas en el laboratorio de diseño molecular del IQS.

Adicionalmente se ha estudiado el comportamiento alostérico del receptor CXCR4 frente a moduladores antagonistas de pequeño tamaño y moduladores alostéricos agonistas de naturaleza peptídica. CXCR4 se califica como “diana difícil” debido al gran tamaño del sitio activo ortostérico, juntamente con el amplio número de funciones reguladoras en las que interviene el receptor CXCR4. Por ello la modulación alostérica en CXCR4 se ha estudiado utilizando diferentes aproximaciones, como son: *docking* ciego, *docking* proteína-proteína, *docking* por subsitios y dinámica molecular.

Sumari

Els mètodes de descobriment de nous fàrmacs han evolucionat recentment gràcies a la resolució de les estructures proteiques que actuen com a dianes terapèutiques responsables de malalties o desregulacions biològiques. Aquestes estructures proteiques tridimensionals, juntament amb el desenvolupament de noves tècniques computacionals permeten el desenvolupament accelerat de nous compostos candidats a esdevenir fàrmacs.

El present treball s'inicia proposant un nou mètode que millora l'elecció de compostos candidats a ser inhibidors d'una "diana difícil", però ben coneguda com és el receptor VEGFR-2, a partir de la seva estructura tridimensional cristal·litzada, així com de compostos inhibidors coneguts de l'esmentada diana.

La resolució tridimensional de l'estructura CXCR4 mitjançant cristal·lografia de raigs X a l'any 2010, ha esdevingut un avenç important a l'hora de millorar el disseny de compostos inhibidors del VIH, així com compostos antitumorals, malalties en les que intervé de forma determinant el receptor CXCR4. Així doncs, els models de cribratge virtual desenvolupats abans del 2010 dins el laboratori de disseny molecular de l'IQS (GEM) han estat generats a partir de models creats per homologia vers a altres proteïnes GPCRs i/o s'han basat solament en la forma de lligands coneguts. D'aquesta forma, a partir de les diferents estructures proteiques publicades de CXCR4, s'ha avaluat quina d'aquestes estructures presenta la conformació que distingeix millor els compostos antagonistes actius dels compostos inactius. A més, s'han avaluat múltiples mètodes de cribratge virtual de CXCR4 basats en l'estructura, en la forma del lligand i mitjançant models farmacofòrics. Una vegada obtinguda la millor estructura de CXCR4 i els millors mètodes de cribratge virtual retrospectiu, es realitzen cribratges virtuals prospectius d'una nova quimioteca generada de forma combinatòria, basada en estructures anàlogues prèviament desenvolupades al laboratori de disseny molecular de l'IQS.

Addicionalment, s'ha estudiat el comportament al·lostèric del receptor CXCR4 davant de moduladors antagonistes petits i moduladors al·lostèrics agonistes de naturalesa pèptica. CXCR4 és qualificada com a una "diana difícil" per la gran mida del seu lloc actiu ortostèric, així com per l'ampli número de funcions reguladores en les que intervé el receptor. Per això la modulació al·lostèrica en CXCR4 s'ha estudiat utilitzant diferents aproximacions com són el *docking* cec, *docking* proteïna-proteïna, *docking* per sublocs d'unió i dinàmica molecular.

Summary

Drug discovery methods have recently emerged thanks to the resolution of protein structures which act as therapeutic targets responsible for diseases or biological deregulations. These three dimensional structures in combination with the development of new computational techniques are accelerating the development of new candidates to become drug compounds.

This work starts with the proposal of a new method that improves the selection of candidates to become inhibitors of a well-known “difficult target” such as VEGFR-2 receptor. This method is based on the crystal structure of the receptor and also by a number of inhibitors known for this target.

CXCR4 crystal structure was solved in 2010 by X-ray crystallography and this has been an important event in order to improve the molecular design of HIV inhibitors, as well as anticancer compounds, diseases where CXCR4 receptor is involved. Therefore, virtual screening models developed in the laboratory of molecular design of IQS (GEM) were generated using homology models from other GPCRs and/or based on ligand shape techniques. In this sense, taking into consideration all published CXCR4 crystal structures, it has been evaluated which of them shows the most suitable conformation to distinguish antagonists actives from inactives. Moreover, different virtual screening methods have also been evaluated such as structure based methods, ligand based methods and pharmacophoric models. Once obtained the most suitable structure and the best retrospective virtual screening methods, a prospective virtual screening has been carried out using a new combinatorial library of chemical structures. This new library is based on analogous structures previously generated in the laboratory of molecular design of IQS (GEM).

In addition, the allosteric behaviour of CXCR4 receptor has been studied *versus* small antagonist modulators and *versus* peptidomimetic agonist modulators. CXCR4 is classified as a “difficult target” due to the large size of its extracellular pocket that the orthosteric binding site is placed as well as the diverse number of biochemical regulations where the receptor mediates. Thus, the allosteric modulation of CXCR4 has been studied using different approaches such as blind docking, protein-protein docking, docking by subsites and molecular dynamics.

Indice

INTRODUCCION.....	1
A. CRIBADO VIRTUAL DE COMPUESTOS FARMACOLOGICOS	1
A.1 Propiedades estructurales que determinan el potencial farmacológico.	4
A.2 Identificación y caracterización de la diana farmacológica.	7
A.3 Descubrimiento de fármacos para una proteína diana conocida.	8
A.4 Descubrimiento de fármacos para una proteína diana desconocida.	10
B. RECEPTOR CELULAR CXCR4.....	14
B.1 Receptores GPCRs.	14
B.2 El receptor quimiocina CXCR4.....	20
OBJETIVOS	27
1. FUNDAMENTOS TEORICOS	28
1.1 Modelización molecular.	28
1.2 Modelos basados en la estructura.	34
1.3 Modelos farmacofóricos.....	42
1.4 Descriptores moleculares.....	44
1.5 Búsqueda por similitud (<i>Similarity search</i>).....	46
1.6 Diseño de librerías de compuestos químicos (quimiotecas)	48
1.7 Dinámica molecular	51
2. DISCUSION DE RESULTADOS	55
2.1 Estudio y propuesta de un método de cribado virtual que mejore la recuperación de activos frente a inactivos para “dianas difíciles”.....	55
2.2 Evaluación funcional de la estructura proteica de CXCR4 mediante cribado virtual retrospectivo.	61
2.3 Cribado virtual retrospectivo de antagonistas de CXCR4. Comparación del método de cribado virtual basado en la estructura (<i>docking</i>) con métodos basados en la forma tridimensional del ligando (<i>Shape matching</i>) y modelos farmacofóricos generados a partir de las interacciones proteína-ligando.	75
2.4 Cribado virtual prospectivo de CXCR4. Generación de una quimioteca combinatoria de compuestos para su cribado virtual prospectivo. Propuesta de nuevas moléculas potenciales antagonistas del receptor CXCR4.....	81
2.5 Estudio de la modulación alostérica de agonistas y antagonistas de CXCR4.	89
CONCLUSIONES	100
REFERENCIAS	102
ARTICULOS	A1
APENDICES	B1

INTRODUCCION

A. CRIBADO VIRTUAL DE COMPUESTOS FARMACOLOGICOS

En los últimos 10-15 años se ha incrementado de manera exponencial el uso de aproximaciones computacionales en el campo del diseño de fármacos. El fundamento básico es el propuesto en 1909 por Ehrlich¹, el cual cuando expuso la idea de que una enfermedad podría ser curada inhibiendo a una proteína diana de forma específica¹.

No fue hasta los años 70 y principios de los 80, cuando empezaron a emerger métodos modernos basados en la determinación cristalina de estructuras proteicas. Al mismo tiempo, empezaron a desarrollarse cálculos computacionales *ab initio* aplicados a la química² (Fig. A.1.1), a desarrollarse la mecánica molecular mediante campos de fuerza (*Force fields*) aplicados a moléculas complejas como proteínas y ácidos nucleicos³ y posteriormente sofisticados gráficos por ordenador que permitían un análisis visual de las interacciones de estas proteínas y las posibles moléculas ligando.

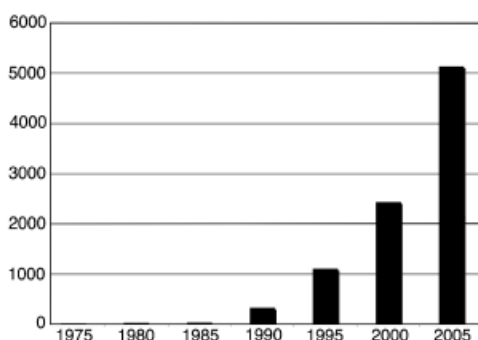


Fig. A.1.1. Crecimiento exponencial del uso de métodos computacionales en diseño de fármacos (*Drug Design*).

Este desarrollo, ha comportado múltiples avances, como son la publicación por internet de bancos de datos de proteínas, de secuencias genómicas, determinadas experimentalmente por cristalografía por rayos X y más recientemente por resonancia magnética nuclear. Estas estructuras tridimensionales han sido informatizadas (p.ej. ficheros PDB) y publicadas en bases de datos de libre acceso, como son el RCSB *Protein Databank*⁴ o el *European Bioinformatic Institute*⁵. Igualmente, se han desarrollado bases de datos de diferentes compuestos químicos, que recogen las actividades inhibitoras *in vivo* o *in vitro* de dichos compuestos. Estas bases de datos contienen los pertinentes valores de inhibición (IC50) clasificados para la proteína diana contra la cual han sido ensayados, como por ejemplo, *Binding Database*⁶ o la base de datos *DUD*⁷ (*Directory of Useful Decoys*) de la Universidad de California. Esta última, permite obtener conjuntos de compuestos activos clasificados por proteína diana, así como conjuntos de moléculas con características y propiedades físico-químicas similares a las activas, pero que supuestamente son inactivas

(*decoys*). En este caso, los conjuntos activos y *decoys* pueden utilizarse para validar modelos de cribados virtuales de forma eficiente, mostrando la selectividad del modelo a la hora de escoger ligandos activos frente a los *decoys*, al mismo tiempo que evitan tendencias en los resultados.

Paralelamente se han desarrollado múltiples programas computacionales comerciales y no comerciales, aplicados al Diseño de fármacos (*Drug Design*). Estos programas son capaces de utilizar las bases de datos de ligandos y proteínas y construir modelos de predicción de actividades biológicas de los ligandos, así como modelar tridimensionalmente proteínas partiendo de su secuencia proteica, o predecir los modos de unión e interacciones de las proteínas con los ligandos, de realizar cribados virtuales de quimiotecas y/o de generar dinámicas moleculares de complejos proteína-proteína o proteína-ligando.

Este desarrollo tecnológico ha sido impulsado básicamente por dos motivos, por un lado observamos que los sistemas biológicos son uno de los sistemas más complejos que encontramos en el amplio ámbito de la investigación y por otro lado observamos los elevados costes que representa el descubrimiento y lanzamiento de un nuevo fármaco al mercado.

En referencia a la complejidad de los sistemas biológicos, la investigación se enfrenta a difíciles problemas como predicción de eficacia, la actividad, la toxicidad o la biodisponibilidad, campos donde la química computacional puede aportar soluciones que den resultados satisfactorios y en tiempos mucho más reducidos que la investigación tradicional de laboratorio.

Por otra parte la investigación médico-farmacéutica se enfrenta a una presión muy elevada en reducir sus costes de investigación y por tanto sus costes de lanzamiento de nuevos fármacos. Se estima que el lanzamiento de un nuevo fármaco tarda de 7 a 12 años hasta salir al mercado y que sus correspondientes costes oscilan entre los 500-2000 millones de dólares⁸, aunque otras fuentes apuntan a entre 1300 y 1700 millones de dólares⁹. Además, se estima que típicamente de cada 8000 compuestos químicos probados en animales, sólo uno es probado en seres humanos y que de éstos, sólo uno de cada cinco compuestos aprueban la fase de ensayos clínicos con éxito⁸. Este escenario ha generado una presión en la industria farmacéutica que ha propiciado fusiones y creación de corporaciones múltiples con el fin de compartir estos costes y diluir los riesgos de inversión asociados. Así pues, estas cifras nos muestran los retos¹⁰ presentes que afronta la innovación en el descubrimiento de fármacos tanto a nivel químico, como biológico, farmacológico, farmacocinético y toxicológico, y como la química computacional puede aportar soluciones a estos problemas, reduciendo dichos costes, así como reduciendo los tiempos de investigación.

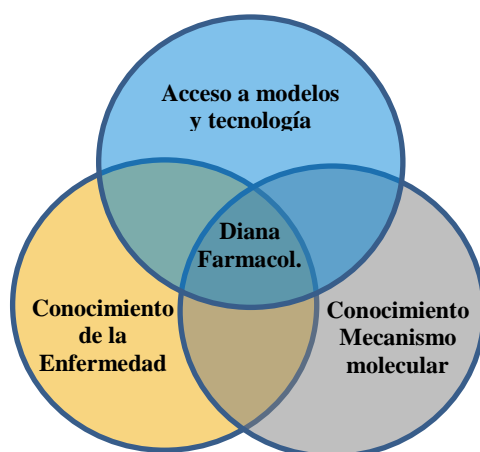


Fig. A.1.2. Factores clave en la estrategia de identificación de dianas farmacológicas

Una de las primeras cuestiones a responder en el diseño de fármacos, es saber qué tipo de compuestos son de interés para llevar a cabo el desarrollo de farmacológico. Típicamente los criterios más relevantes para que series de compuestos químicos pasen a la fase de desarrollo son¹¹:

- Ser activo a nivel bioquímico y en ensayos celulares.
- La actividad dependiente de la concentración.
- Valores umbrales de IC_{50} (por debajo de micromolar o nanomolar).
- Comprensión de la relación estructura-actividad de la serie de compuestos.
- Conocimiento de cinéticas de unión.
- Selectividad
- Estructura y pureza establecida
- Estabilidad.
- Sintéticamente viable.
- Serie patentable.
- Posibilidad de ser optimizado, creando derivados, etc.
- Medida de la solubilidad.
- Medida del Log D.
- Predicción de metabolitos.

- Medida o predicción de variables farmacocinéticas.
- Predicción de su toxicidad.
- Conocimiento de posibles efectos secundarios.

Estos criterios experimentales, indican en una primera etapa, la viabilidad de poder convertir una serie de compuestos químicos en fármacos. La discusión y las evidencias del porque estos compuestos pueden tener éxito o fracasar, nos muestran indicios sobre las características de la estructura molecular, los cuales son importantes en el diseño de fármacos.

A.1 Propiedades estructurales que determinan el potencial farmacológico.

En el diseño de fármacos se han desarrollado una serie de métricas cuyo fin es identificar si los compuestos estudiados son farmacológicos (*drug-like*). Estas métricas se basan principalmente en el tamaño molecular, propiedades moleculares y los grupos funcionales existentes de los compuestos candidatos. Es importante destacar que las métricas farmacológicas no indican si un compuesto será un buen fármaco contra una enfermedad, sin embargo, compuestos candidatos que no cumplan los criterios farmacológicos generalmente no serán buenos fármacos, bien sea por excesiva toxicidad, pobre biodisponibilidad, etc.

Uno de los criterios más frecuentemente utilizados es la Regla de cinco de Lipinsky¹², una regla empírica que establece que para que un compuesto tenga biodisponibilidad al ser administrado oralmente, debe cumplir al menos 3 de las 5 reglas.

1. El compuesto no debe contener más de cinco donadores de enlaces por puentes de hidrógeno.
2. El compuesto no debe tener más de diez aceptores de enlaces por puentes de hidrógeno.
3. El compuesto debe poseer un peso molecular igual o inferior a 500.
4. El compuesto debe poseer un coeficiente de reparto octanol-agua ($\log P$) inferior a 5.

Partiendo de esta regla, diversas variantes y modificaciones han aparecido. Entre ellas cabe citar:

Schneider y Baringhaus los cuales amplían la regla de Lipinsky con 3 reglas adicionales:

-El compuesto no debe tener más de 5 enlaces rotables (esta regla es conocida como la 5ª regla de Lipinsky).

- El compuesto no debe tener una área polar superficial superior a 120 \AA^2 puesto que no sería suficientemente lipofílico como para entrar en la sangre vía absorción intestinal.

-El compuesto no debe tener una solubilidad acuosa ($\log S$) inferior a -4.

Oprea *et al.* Proponen unas reglas también muy parecidas a las de Lipinsky, tras analizar un número elevado de fármacos aprobados:

- El peso molecular ≤ 460
- La hidrofobicidad calculada $4 \leq C \log P \leq 4.2$
- La solubilidad $\log S \leq -5$
- El número de enlaces rotables ≤ 10
- El número de donadores de enlaces por puentes de hidrógeno ≤ 5
- El número de aceptores de enlaces por puentes de hidrógeno ≤ 9

Estos parámetros descritos son fácilmente calculables a nivel computacional, por lo que el filtrado de quimiotecas de compuestos candidatos a fármacos, se convierte en una tarea muy útil a la hora de pre-cribar dichas quimiotecas.

Otro aspecto importante en la relación entre estructura química y su viabilidad farmacológica, reside en los grupos funcionales existentes en dichos compuestos, principalmente, en lo que se refiere a que grupos funcionales deben ser incluidos en quimiotecas combinatorias. Así Bemis *et al.* presentaron estudios^{13,14} con las características comunes en la base de datos CMC¹⁵ formada por más de 5000 compuestos farmacológicos. El estudio se centró tanto en los núcleos de dichos compuestos¹³, como en sus sustituyentes¹⁴, comparándose características comunes que se presentaban entre ellos. Por ejemplo, se observó que el 94% de los compuestos tenían grupos funcionales cíclicos y sólo un 6% era acíclicos, también se determinó que el promedio de sustituyentes por molécula es de 4 y que de estos sustituyentes, el más común hallado es el carbonilo, seguido por el radical metilo y el hidroxilo (Fig. A.1.3.). De hecho, algunas compañías tienen listas de grupos funcionales restringidos los cuales evitan incorporar en fármacos. Algunos de estos grupos restringidos¹¹ son: tioureas, esterés, amidas, disulfuros, tioles, β -lactamas, o-nitro, alcoxi-piridinio, benzofenona, oxadiazina, fluorenona o acilhidroquinonas, aunque en algunos casos como por ejemplo el de las β -lactamas, existen discrepancias debido al amplio rango de bacterias que exhiben resistencia a los antibióticos β -lactámicos. Los grupos funcionales ésteres y las amidas son hidrolizados en el cuerpo humano por esterásas y proteásas, lo que explica su pobre elección al ser incorporados en fármacos.

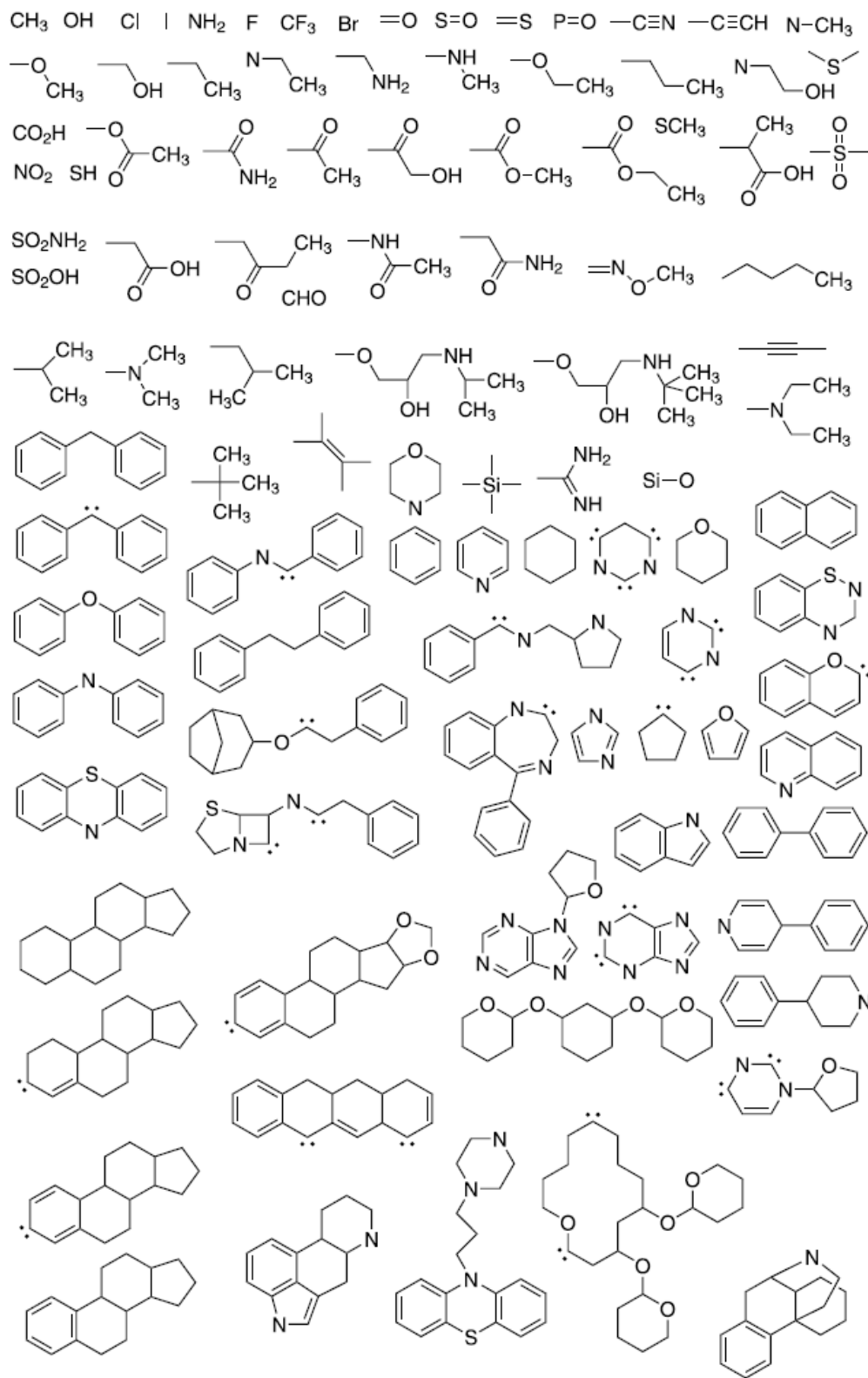


Fig. A.1.3. Núcleos y grupos funcionales frecuentes en compuestos farmacológicos (Representación de D.C.Young.¹¹)

Otra propiedad importante al diseñar compuestos farmacológicos es su especificidad frente a la diana terapéutica. La posibilidad de que un fármaco se una a varias proteínas, o a una proteína que no es la diana terapéutica en cuestión, aumenta la posibilidad de que se den efectos secundarios no deseados en pacientes y probablemente de carácter severo. De este modo, llamamos ligandos

promiscuos (*promiscuous binders*)¹⁶ a aquellos compuestos que pueden unirse o inhibir un número elevado de proteínas¹⁷. Se cifra que cada medicamento puede interactuar en promedio, con 6 dianas farmacológicas¹⁸.

A.2 Identificación y caracterización de la diana farmacológica.

Con la finalización del proyecto genoma humano¹⁹, se dispone de la secuencia primaria de aminoácidos de todas las proteínas potenciales del cuerpo humano. Sin embargo, el conocimiento de la secuencia primaria no indica donde una proteína es expresada, o como activan una vía metabólica. El diseño molecular, es una herramienta que ayuda a conocer estas funciones. Las técnicas de diseño molecular basadas en la estructura de la proteína (*structure-based*) requieren de su estructura tridimensional. Existen diferentes métodos de determinación de estructuras proteicas, las más habituales son mediante cristalografía de rayos X²⁰, mediante RMN²¹, o bien mediante la generación de modelos por homología²², creados a partir de la secuencia de dicha proteína.

Una vez conocida o modelada la estructura tridimensional de la proteína, el siguiente reto que se plantea es conocer el sitio, o sitios activos de la proteína y los mecanismos que se producen cuando a ésta se le une un fármaco.

La localización del sitio activo de la proteína generalmente se determina bien porque esta proteína ha sido co-cristalizada conjuntamente con un ligando y por tanto, la estructura tridimensional lo incorpora *de facto* en el sitio activo, bien mediante el uso de métodos computacionales que escanean la superficie de la proteína, en busca de cavidades cuyo volumen denota la posibilidad de alojar ligandos de tamaño similar, o bien mediante técnicas de mutagénesis dirigida que permitan determinar los aminoácidos clave que establecen interacciones.

Una proteína cuya estructura tridimensional ha sido determinada, puede presentar una conformación activa o inactiva, dependiendo de si ésta se ha co-cristalizado junto a un ligando agonista o antagonista. Además, en el caso de receptores celulares, cuya flexibilidad en la estructura es muy elevada, su estructura determinada experimentalmente puede mostrar una conformación de las múltiples conformaciones posibles que la proteína-receptor puede adoptar. Estas conformaciones proteicas dependen de aspectos como la cinética de reacción de la unión proteína-ligando, o bien de si el ligando puede generar efectos inducidos sobre la proteína, por ejemplo debido a su gran tamaño, o también pueden depender de la existencia de un ligando de naturaleza alostérica, entre otros. Estos mecanismos que afectan a la conformación de un receptor pueden ser simulados mediante dinámicas moleculares que predicen las trayectorias de cada uno de los átomos que conforman el sistema a predecir.

A.3 Descubrimiento de fármacos para una proteína diana conocida.

Si la proteína diana contra la que actuar es conocida, se dispone de su estructura tridimensional y además el fármaco a diseñar actúa de manera competitiva frente al sustrato o bien frente ligando natural o endógeno, se suele utilizar como estrategia *in silico*, el llamado diseño de fármacos basado en la estructura (*structure-based*)²³. La estrategia basada en la estructura tiene como principal característica el identificar compuestos principales (*lead compounds*) mediante cribado virtual realizado a través de la técnica de *docking*. Como resultado, el *docking* devuelve las mejores conformaciones de los compuestos cribados en el sitio activo de la proteína y al mismo tiempo otorga unos *scorings* que permiten realizar un ranking de dichos compuestos en función de las energías halladas²⁴ (Fig. A.1.4.).

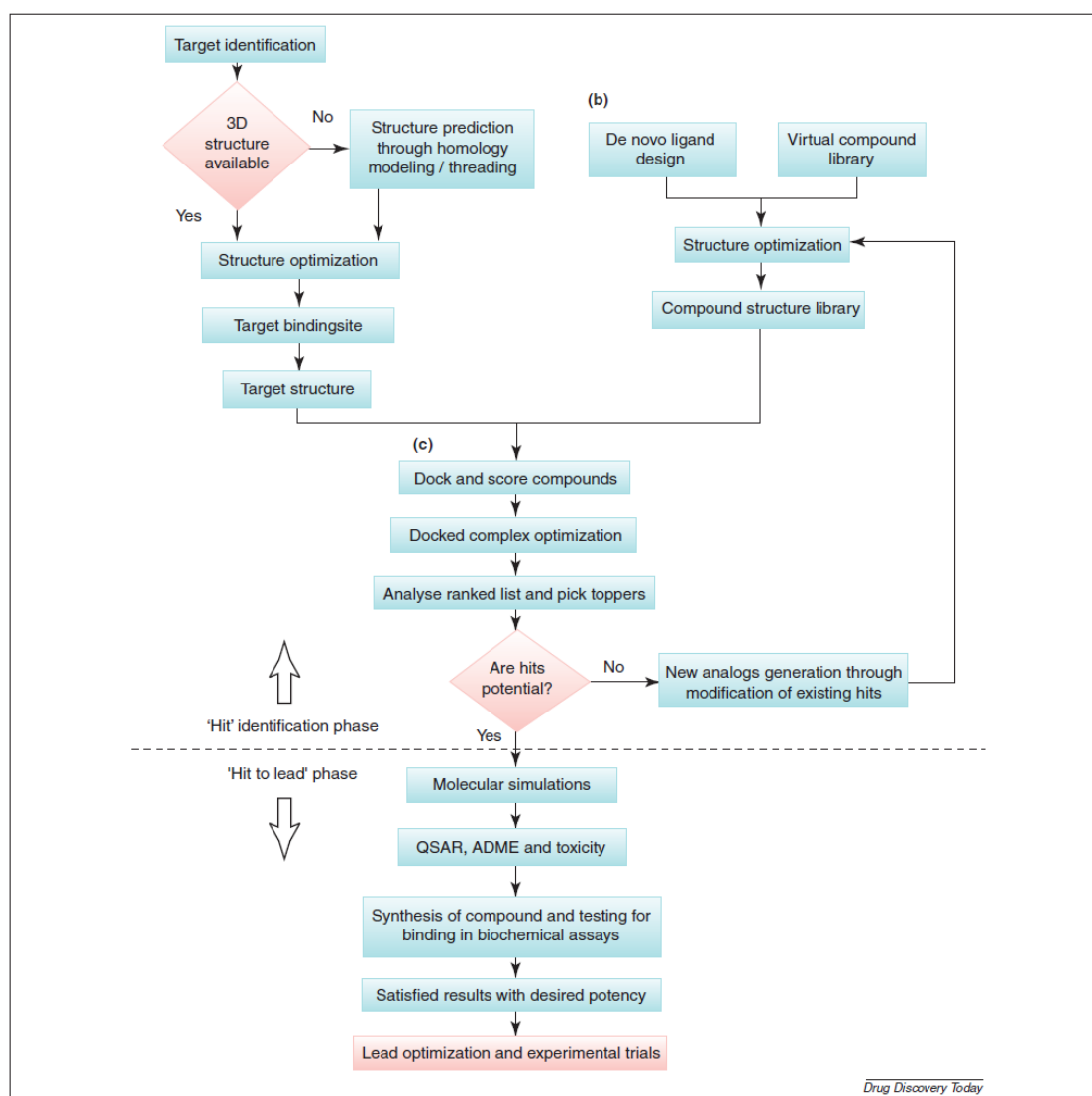


Fig. A.1.4. Etapas del proceso de diseño de fármacos basado en la estructura. (a) Etapa de generación y/o preparación de la estructura e identificación del sitio activo. (b) Generación y preparación de la quimioteca de compuestos candidatos a cribar.

(e) Etapa de cribado virtual de la quimioteca en el sitio activo de la diana terapéutica, mediante *docking* para identificar los compuestos “hits” (Representación de *Kalyaanamoorthy et al.*²³)

La gran ventaja del método de *docking* es su habilidad de representar el modo de unión del ligando en la estructura diana. Dicho proceso explora y predice tanto la orientación, como la conformación del ligando (predicción de poses)²⁵ dentro del sitio activo de la estructura diana. Habitualmente se usa el símil de “la mano y el guante” para mostrar cómo se produce la interacción del ligando y la estructura proteica. La búsqueda conformacional del ligando se realiza simultáneamente a la exploración de las diferentes orientaciones del ligando en el sitio activo de la proteína, con el fin de identificar el mejor modo de unión. Esta búsqueda se realiza utilizando diferentes algoritmos numéricos como son: algoritmos genéticos, de Monte Carlo, algoritmos evolutivos, *Simulated Annealing* algoritmos, aproximaciones empíricas, algoritmos *knowledge-based*, algoritmo de SIS, el algoritmo de Hammerhead, o bien por la aproximación de la transformada rápida de Fourier (FTT). El algoritmo utilizado define la función de *scoring* que otorga a cada pose explorada un valor o *score* en función de la energía proteína-ligando calculada. Esta energía proteína-ligando puede ser calculada como energía de enlace, como energía libre de Gibbs o como la medida expresada de un valor numérico. El resultado que el *docking* devuelve es una pose definida por una conformación, una posición en el sitio activo de la proteína y un correspondiente *score* que indica su energía de interacción con la proteína. Estos resultados se traducen en que para las poses calculadas se nos permite descubrir las interacciones proteína ligando que se establecen, que a su vez, nos indican los residuos proteicos clave, el tipo de interacción química establecida, así como los grupos funcionales del ligando que las establecen. En cambio, el valor del *score* básicamente sólo nos permite clasificar las poses en función de la energía, pues su aplicación para correlacionar energías de enlace con actividad bioquímica aún es un reto no solventado al completo.

A pesar de que el concepto de *docking* puede parecer simple, éste lleva implícita una elevada complejidad. Por ejemplo, cuando en lugar de realizar *docking* rígido de la proteína, se realizan *dockings* flexibles en donde una parte de la proteína adquiere cierta flexibilidad, o en el caso de los *dockings* proteína-proteína, donde la naturaleza del ligando es de origen peptídico y por tanto el número de átomos y el número de conformaciones posibles aumentan exponencialmente. En estos casos el cálculo del *scoring* cobra mayor complejidad y requiere mayor coste computacional.

Finalmente, en la última etapa del *docking* se ordenan las poses devueltas en base a sus *scorings* obtenidos. De este modo se obtiene una clasificación jerarquizada de los compuestos, llamados *hits*, con mayores garantías de pasar a la siguiente fase el proceso (Fig. A.1.4.).

Existe la posibilidad de que en el proceso inicial, se conozca la estructura de la diana terapéutica y su sitio activo, pero en cambio no se conozca ningún compuesto que compita con el ligando ortostérico

a partir del cual desarrollar una posible quimioteca. En estos casos se puede realizar previamente un análisis de diseño de fármacos basado en fragmentos (*Fragment-based Drug Design*)²⁶. Esta técnica se basa en utilizar fragmentos pequeños de fármacos, con el fin de obtener un mapa de elementos farmacóforos clave, que muestren afinidades de unión con la proteína más elevadas. El objetivo de esta técnica es simple y se basa en poder recoger en una sola entidad ligando, los diferentes farmacóforos hallados clave mediante los fragmentos, de manera que al sumar sus interacciones con la proteína, deberían sumarse las contribuciones individuales halladas.

A.4 Descubrimiento de fármacos para una proteína diana desconocida.

Actualmente existen casos donde la proteína diana no es conocida al iniciarse el proyecto de investigación y su determinación es contemplada como una parte más del proceso de descubrimiento de fármacos. El disponer de la estructura proteica y poder utilizar técnicas basadas en la estructura, aumenta las posibilidades de éxito en el proceso de investigación. Sin embargo, esto no siempre es posible. Los receptores celulares ubicados en la superficie de la membrana plasmática, como son los GPCRs, presentan una elevada dificultad a la hora de ser cristalizados y aunque cada vez son más las estructuras determinadas estructuralmente, aún quedan muchos receptores por determinar y no siempre se conoce cuál es el receptor que interviene en la disfunción biológica. En estos casos, la técnica de *docking* no puede ser utilizada y se recurren a las técnicas basadas en el ligando (*ligand-based drug design*). Las técnicas basadas en el ligando se basan en el principio de similitud, por el que se constata que estructuras similares tienen propiedades similares.

Los métodos basados en el ligando ofrecen una serie de ventajas cuyas consideraciones deben tenerse en cuenta²⁷:

- Rapidez: La construcción de modelos por homología, por ejemplo en GPCRs, es un proceso lento, que idealmente requiere de ensayos de mutagénesis dirigida con el fin de incorporar información sobre la estructura. Adicionalmente, las estructuras obtenidas acostumbran a tener sucesivos procesos iterativos de inspección gráfica de su estructura y a ser refinadas mediante dinámicas moleculares.

Por el contrario, las aproximaciones basadas en el ligando pueden aplicarse de forma prácticamente inmediata, al identificar un conjunto de ligandos activos contra una diana farmacológica.

- Aplicabilidad: En la determinación estructural de proteínas se obtiene una única conformación de la proteína. Dependiendo de si la cristalización es realizada con un ligando agonista o antagonista, la conformación de la proteína será activa o inactiva respectivamente. Si tenemos en cuenta que la mayoría de estructuras que se han determinado sólo se

encuentran en una conformación, activa o inactiva, podemos encontrarnos que para diseñar agonistas dispongamos de una estructura en forma inactiva y que para diseñar antagonistas nos encontremos que la estructura cristalizada esté en conformación activa. Este es un problema típico al usar estrategias basadas en la estructura, el cual no se presenta al utilizar métodos basados en el ligando.

- Complementariedad: Las técnicas basadas en el ligando pueden ser un buen complemento a las técnicas basadas en la estructura, ya que la aplicación de diferentes técnicas pueden dar como resultado candidatos pertenecientes a diferentes familias químicas y por tanto permitir ampliar el espectro de resultados.

Las técnicas basadas en el ligando, requieren por definición, el conocimiento de uno o varios compuestos que actúen como ligandos de la diana farmacológica en cuestión. Preferiblemente estos ligandos deberían ser potentes y no deberían ser péptidos de elevado tamaño.

Para realizar cribados virtuales utilizando técnicas basadas en el ligando, es necesario utilizar bases de datos de estructuras químicas, bien sean en 2D o 3D, aunque su conversión final a 3D es indispensable. Estas quimiotecas deben prepararse convenientemente contemplando las posibles formas tautoméricas, protoméricas y enantioméricas de las estructuras que la componen.

Existen 3 tipos de métodos basados en el ligando:

- Métodos de similitud 2D: Son métodos basados en descriptores de similitud 2D (*fingerprints*), que permiten la ordenación de compuestos en base a la similitud con la molécula definida como modelo²⁸. Típicamente, las comparaciones que se realizan otorgan el valor 0 cuando no existe similitud y el valor 1 cuando la similitud es elevada.
- Métodos de similitud 3D: Los métodos de similitud 3D extienden la búsqueda de subestructuras 2D, ampliándola a la disposición relativa en el espacio 3D. De este modo la búsqueda de similitud 3D, además de buscar las subestructuras apropiadas, toma en cuenta las distancias entre ellas. Estos métodos, también se basan en descriptores, en este caso, descriptores 3D. Uno de los inconvenientes que presentan estos descriptores, es que solamente representan una conformación fija, mientras que las moléculas son flexibles a la hora de unirse a una proteína. Para evitar este contratiempo es necesario generar múltiples conformaciones. Estos métodos pueden incluir algoritmos *Shape Matching*, *fingerprints* basados en forma 3D (*shape-based fingerprints*), descriptores moleculares, *fingerprints* farmacofóricos, etc.
- Farmacóforos: La definición oficial de farmacóforo, según la IUPAC²⁹ se detalla como un “conjunto de características estéricas y electrónicas, que describen propiedades químicas las

cuales son necesarias para asegurar las interacciones óptimas con dianas biológicas específicas, con el fin de bloquear o favorecer respuestas biológicas”. Un farmacóforo recoge geoméricamente las interacciones clave de las conformaciones que muchos de los compuestos activos adoptan, cuando estos se unen a una proteína. Conceptualmente se asume que los compuestos activos se unen a la proteína diana del mismo modo y que por tanto se superponen. Igualmente se asume que la protonación y tautomería de los activos es conocida y que existe suficiente diversidad de familias de activos, de modo que las características moleculares comunes que el farmacóforo recoge, permiten correlacionarlas con la actividad³⁰.

Las características moleculares en tres dimensiones (Fig. A.1.5.), que pueden describir a un farmacóforo pueden ser, dadores o aceptores de hidrógenos en puentes de hidrógeno, grupos ácidos o básicos, anillos aromáticos, o centros de cargas positivas o negativas, características hidrofóbicas, etc.

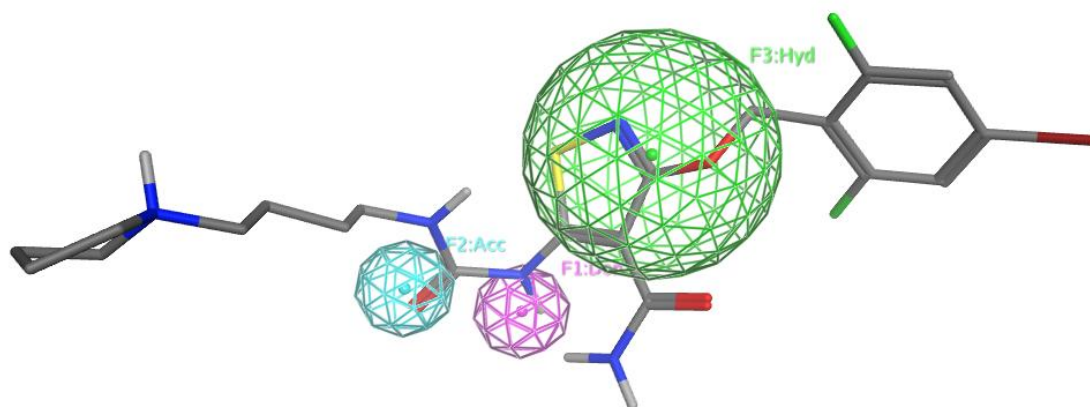


Fig. A.1.5. Farmacóforo creado sobre un ligando co-cristalizado que contiene 3 características moleculares. F1 es una característica dadora de hidrógenos, F2 es una característica aceptoras de hidrógenos y F3 es una característica hidrofoba que presenta un mayor radio de interacción.

Uno de los principales retos del modelado de farmacóforos es la modelización de la flexibilidad del ligando y la necesidad de una eficiente generación de conformaciones, al crear las poses de enlace que estas moléculas pequeñas adoptan cuando se unen a macromoléculas. A su vez, deben mantener reducido el número de conformaciones a creadas y disminuir los tiempos de cálculo conformacional.

En el presente trabajo se han utilizado tanto técnicas basadas en estructura de la proteína, como el *docking* rígido de la proteína, para ligandos orgánicos de pequeño tamaño, así como el *docking*

flexible proteína-proteína. También se han utilizado técnicas basadas en el ligando como son la generación de farmacóforos, de descriptores 3D (*shape matching*) y de descriptores de similitud 2D como el índice de Tanimoto. Así mismo, se han utilizado combinaciones de ambas técnicas. Se ha creado y validado un protocolo de cribado virtual para “dianas difíciles” consistente en tres pasos 1) *docking*, 2) post procesado farmacofórico de las poses halladas en el *docking* 3) búsqueda de similitud: ordenación de las poses que superan el farmacóforo por similitud 2D Tanimoto. Se ha testado este protocolo para "dianas difíciles" con la diana VEGFR2 y se ha aplicado posteriormente a la diana CXCR4.

B. RECEPTOR CELULAR CXCR4

El receptor celular CXCR4 forma parte de la familia de receptores quimiocinas, que a su vez forma parte del gran grupo de los receptores asociados a proteínas G o GPCRs. Este receptor cobra importancia a mediados de los años 90s, a partir del descubrimiento de su mediación como co-receptor celular en el proceso de entrada del virus VIH al interior de la célula^{31,32}. Este hecho marca su investigación y estudio que posteriormente muestra su relevancia en múltiples procesos biológicos.

La familia de los receptores quimiocinas, a la cual pertenece, se caracteriza por hallarse ubicados en la superficie de los leucocitos y sobre todo por disponer de una estructura compuesta de un dominio de 7 hélices alfa enrolladas. Este rasgo fundamental caracteriza a las quimiocinas y al resto de miembros de la gran familia de GPCRs, como se detalla brevemente en los siguientes apartados.

B.1 Receptores GPCRs.

La superfamilia de proteínas GPCR es una de las mayores dentro del genoma humano. En el año 2003 se calculaba³³ que un 30-40% de los fármacos del mercado tenían como objetivo estos receptores³⁴. Por otra parte, se calcula que en el proteoma humano existen 3000 posibles dianas farmacológicas y que de éstas, el 97% corresponden a GPCRs, canales iónicos, transportadores receptores nucleares y enzimas. Actualmente, ya se considera que los receptores GPCRs tienen un espectro farmacológico más amplio que los fármacos dirigidos a enzimas, canales iónicos o bien receptores nucleares³⁵. Adicionalmente, de los más de 800 GPCRs existentes, si se excluyen los receptores olfativos existen unos 400 miembros de los cuales, en el mercado sólo se encuentran fármacos dirigidos a 30 GPCRs³³. Estos datos nos muestran la importancia de los receptores GPCRs así como las excelentes oportunidades de esta familia de receptores a la hora de hallar nuevos fármacos.

Tal y como indica su nombre, estos receptores interactúan con proteínas G, también llamadas proteínas de unión guanina nucleótidos y su principal característica es que todos ellos poseen 7 tramos, llamados dominios de transmembrana (TM), que están formados por un número de aminoácidos que oscila de 25 a 35, de relativa hidrofobicidad y que se expanden en la membrana plasmática en forma de α -hélices enrolladas en sentido antihorario.

En las transmembranas de los GPCRs se producen procesos de transducción de señal celular de 2 tipos: procesos de reconocimiento de ligandos extracelulares y procesos de propagación de la señal generada en el sitio de unión ligando-receptor, hacia los aminoácidos de la transmembrana citoplasmáticos. Cabe resaltar que mientras los procesos de reconocimiento de ligandos son muy específicos del subtipo de receptor, los procesos de propagación de las señales hacia la parte interior de la transmembrana tienen mecanismos comunes. Esto se debe al alto grado de conservación de

secuencia en la parte central y citoplasmática de las hélices de las transmembranas, en comparación con el dominio extracelular de las mismas³⁶.

Estos receptores median en muchas de las funciones intracelulares e intercelulares. La transducción de señal de los receptores GPCR con las correspondientes proteínas G, puede ser activada por una amplia variedad de inductores como son: hormonas, neurotransmisores, quimiocinas como el CXCR4, iones como el Ca²⁺, los olores o la luz. Para la mayoría de GPCRs los correspondientes ligandos endógenos son conocidos y dichos ligandos al unirse al receptor suelen provocar un cambio conformacional, permitiendo la unión de proteínas efectoras tales como las proteínas G o las β -arrestinas. A su vez, este proceso suele desembocar en la activación de una cascada de proteínas de señalización intracelular, que como resultado conlleva una respuesta fisiológica. De este modo, un desequilibrio en el cuerpo de agentes químicos, puede alterar la regulación normal de una respuesta fisiológica, lo que hace las GPCRs unas atractivas dianas de numerosas enfermedades.

2.1.1 Clasificación de GPCRs por familias: Debido al gran número de receptores GPCR existente, es necesaria su clasificación y agrupación atendiendo a sus características fisiológicas o estructurales. Attwood y Findley³⁷ realizaron la primera clasificación de más de 240 GPCRs del tipo rodopsina. Casi al mismo tiempo, Kolakowski³⁸ realizó una clasificación por familias o clanes, de la A a la F y que ha permitido el posterior desarrollo de bases de datos específicas como GPCRdb³⁹, que en la actualidad cuenta con más de 7000 entradas de proteínas-diana de diversas especies animales, o como la base de datos SEVENS⁴⁰ con más de 64000 genes correspondientes a 56 secuencias de eucariotas, entre otras, repasadas exhaustivamente por Adams *et al.*⁴¹. En el año 2003 Fredriksson⁴² *et al.* clasificaron los receptores GPCRs humanos en 5 familias atendiendo a su análisis filogenético, es decir en base a la relación de los genes y a su proximidad evolutiva. Estas 5 familias *Glutamate; Rhodopsin; Adhesion; Frizzled taste* y *Secretin* es lo que llamamos el sistema de clasificación GRAFS⁴³ (Fig. B.1.1). Además de estas 5 familias cabe añadir una sexta, llamada *Otros receptores 7TM*, donde se engloban diferentes receptores (originalmente 23 secuencias proteicas) que no son asignados a ninguna de las anteriores familias. Cada una de estas familias se subdivide en sucesivas subfamilias, así por ejemplo, cabe destacar la familia de la rodopsina como la familia de mayor tamaño, la cual contiene a su vez 19 subfamilias filogenéticas⁴⁴, entre las cuales se clasifican 278 receptores no olfativos, 396 receptores olfativos⁴⁵ y 11 receptores huérfanos por clasificar.

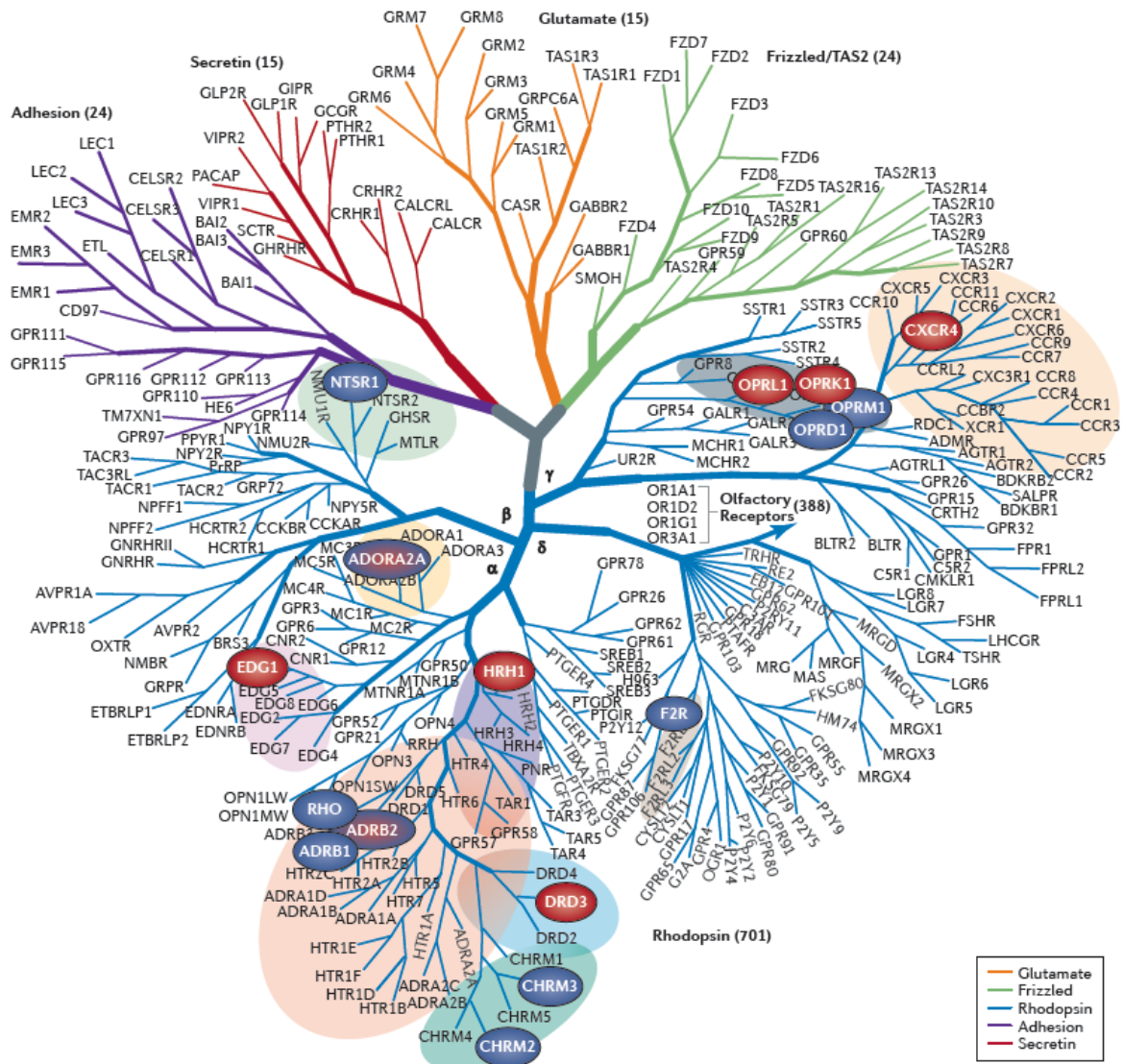


Fig. B.1.1. Representación mostrada por *Stevens et al.*³⁴ del árbol filogenético con las 5 grandes familias GPCRs y el número global de componentes. Los miembros con estructuras determinadas en alta resolución se muestran destacados en color: el receptor CXC4; los receptores β_1 y β_2 adrenérgico (ADRB1 y ADRB2); la A_{2A} adenosina (ADORA2A); la dopamina D3 (DRD3); el receptor acetilcolina muscarínico M_2 y M_3 (CHRM2 y CHARM3); histamina H_1 (HRH1); el opioide tipo κ , tipo δ y tipo μ (OPRD1, OPRK1, OPRM1); nociceptina/orfanina FQ (OPRL1); la rodopsina (RHO); la proteinasa activada-1 (F2R); el receptor esfingosina-1-fosfato (EDG1) y el receptor neurotensina (NTSR1).

Otra clasificación habitual y alternativa a GRAFS, corresponde a la clasificación de los receptores GPCR en mamíferos al agruparlos atendiendo a su similitud en las secuencias proteicas⁴⁶.

- Clase A o clase 1: Abarca un gran número de receptores similares a la rodopsina. Se subclasifican en receptores sensibles a la luz como la rodopsina o la trombina, los receptores adrenérgicos como la adrenalina o el β_2 adrenérgico, A_2 adenosina, angiotensina, las quimiocinas o la gran subfamilia de receptores olfativos. Se distinguen por tener un conjunto de aminoácidos (Asp-Arg-Tyr) altamente conservados en la parte citoplasmática del tercer dominio de transmembrana y que juegan un papel vital para la estabilidad del receptor y para

la mediación en los cambios conformacionales necesarios para la activación del receptor. Se considera que el 25% de los fármacos existentes en el mercado actúan apuntando a dianas de esta subfamilia⁴⁷.

- Clase B o clase 2: Comprende receptores del tipo secretina como el VIP, PACAP, la secretina, el glucagon, α -latrotoxina. Este grupo de receptores presenta como características, una elevada similitud de secuencia en el extremo N-terminal a los receptores *Adhesion* y también en este extremo N-terminal presentan un dominio para la unión de hormonas. Su clasificación filogenética subdivide los 15 receptores en 4 subgrupos. Acostumbran a unirse a proteínas G, como son las proteínas del tipo G_s, las cuales activan la adenilatociclasa.
- Clase C: Son los receptores tipo glutamato metabotrópico (m-Glu), los receptores γ -amino ácido butírico (GABA), receptores activados por arginina, receptores sensibles al Ca⁺² extracelular, algunas feromonas y algunos receptores gustativos. Son llamados también receptores de la familia glutamato. Entre estos se incluyen los receptores de los principales neurotransmisores. Su característica principal es su excepcionalmente larga región extracelular N-terminal, que contiene entre 500 y 600 aminoácidos. Esta región N-terminal ubica el sitio de unión ortostérico y por tanto juega un papel crucial en la activación del receptor y en la unión con el ligando⁴⁸.
- Clase D: Comprende los receptores tipo Frizzled. Este grupo incluye dos subgrupos distintos, los receptores frizzled y los TAS2. Los receptores TAS2 regulan la expresión del gusto en las papilas gustativas y en el epitelio palatal como receptores de gusto amargo⁴⁹, aunque también han sido detectados en los pulmones y en el tracto intestinal⁵⁰. Presentan diferencias con los receptores TAS1 clasificados en la familia de los receptores glutamato. Como característica estructural los receptores TAS2 tienen un extremo N-terminal muy corto que difícilmente puede ubicar un sitio activo para la unión de ligandos⁴². Los receptores Frizzled tienen una longitud de 500 a 700 aminoácidos y actúan como receptores de las proteínas *Wnt* así como de otros ligandos⁵¹. Estructuralmente presentan como característica propia, un extremo N-terminal exofacial, compuesto por un dominio rico en cisteínas seguido por una región hidrofílica de 40-100 aminoácidos⁵². Los genes Frizzled que sintetizan estas proteínas, son esenciales para el desarrollo embrionario, juegan un papel decisivo en la regulación de la polaridad de la célula, en la formación de la sinapsis neural, etc.
- Clase E: Está formada por un subgrupo de receptores feromonas.
- Clase F: formada por los receptores Dictyostelium cAMP⁵³

La familia de la rodopsina contiene, atendiendo a la clasificación por clanes, 5 subgrupos principales de los cuales destacan: el α -grupo rodopsina, el β -grupo rodopsina, el γ -grupo rodopsina y el δ -grupo

rodopsina. A su vez, dentro del γ -grupo rodopsina (Fig B.1.1), encontramos una nueva subdivisión formada por 4 clústeres. De entre de estos 4 clústeres, el de mayor tamaño es el clúster formado por las quimiocinas. Las quimiocinas son unas proteínas pequeñas de 8 a 13 kDa y solubles en agua, las cuales a su vez se dividen en dos subfamilias según la disposición de las 2 cisteínas existentes en el extremo N-terminal. Si estos residuos cisteína se hallan adyacentes se llaman: quimiocinas tipo C-C, como es el receptor CCR5 o bien si las dos cisteínas tienen un aminoácido entre medio se llaman quimiocinas C-X-C tal y como es el receptor CXCR4.

Las quimiocinas también se suelen clasificar según su funcionalidad, en este caso se habla de quimiocinas inflamatorias o quimiocinas homeostáticas, en función de si se inducen durante un proceso de inflamación atrayendo a las células inflamatorias o bien si son segregadas por células estromales.

En los últimos años ha crecido rápidamente el número de estructuras GPCR cristalizadas. Estos receptores han sido cristalizados en forma de complejos con antagonistas, agonistas parciales, agonistas totales, agonistas selectivos (*biased agonists*) y moduladores alostéricos. Los principales receptores GPCRs determinados^{34,54} son los siguientes Receptores de clase A (Fig. B.1.1): el receptor CXCR4⁵⁵, los receptores β 1 y β 2 adrenérgico^{56,57}, la A2A adenosina⁵⁸, la dopamina D3⁵⁹, el receptor acetilcolina muscarinico M2⁶⁰ y M3⁶¹, histamina H1⁶², el opioide tipo κ ⁶³, tipo δ y tipo μ , nociceptina/orfanina FQ⁶⁴, la rodopsina⁶⁵, la proteinasa activada-1⁶⁶, el receptor esfingosina-1-fosfato⁶⁷, la serotonina 5-HT_{1B}⁶⁸ y la 5-HT_{2B}⁶⁹, el receptor CCR5⁷⁰, el receptor neurotensina⁷¹, el receptor purinérgico P2Y₁₂R⁷², el receptor humano GPR40 o FFAR1⁷³. Receptores de clase B: receptor corticotropina tipo 1 (CFR₁R)⁷⁴, el receptor glucagon⁷⁵. Receptores clase C: los receptores glutamato 1 (mGluR1)^{76,77}, 3 (mGluR3)⁷⁸, 5 (mGluR5)⁷⁹ y 7 (mGluR7)⁷⁸. Receptores clase F: cinco receptores Smoothened humano^{80,81,82}

La determinación estructural tridimensional de estos GPCRs ha permitido y permitirá desarrollar métodos de modelado con el fin de profundizar y mejorar el conocimiento de los mecanismos de activación, inhibición, determinación de sitios activos en estas proteínas de membrana⁴¹. Estos avances suponen verdaderamente una nueva era en el conocimiento estructural de los GPCRs y en general, de la biología química⁸³.

2.1.2 Mecanismos de acción o inhibición basados en la funcionalidad de los GPCRs:

Las proteínas GPCRs se caracterizan por tener uno o varios sitios activos dentro de su estructura. Al sitio activo principal, donde se une el ligando natural, se le conoce como sitio activo ortostérico o endógeno. La existencia de múltiples sitios activos en los GPCRs va ligada a la posibilidad de tener múltiples estados activados ligados a múltiples funcionalidades o respuestas por parte de estos. El

descubrimiento de fármacos basado en GPCRs⁸⁴, principalmente se ha centrado en el diseño de compuestos cuya estructura les permite competir con el ligando natural o endógeno sin activar el receptor. Estos compuestos se conocen como antagonistas. Si el diseño de fármacos se basa en compuestos que sintéticamente activan las vías de señal de los GPCRs, imitando a los ligandos endógenos, hablamos de fármacos agonistas. En cambio, si estos compuestos se unen a la proteína en sitios topográficamente distintos al sitio activo, donde se une el ligando endógeno, hablamos de compuestos alostéricos⁸⁵. Los compuestos o ligandos alostéricos pueden modular la respuesta de los ligandos naturales o endógenos, potenciándola (moduladores alostéricos positivos), o bien disminuyéndola (moduladores alostéricos negativos), así como regulando la respuesta de forma neutra (moduladores alostéricos neutros). Existen también un tipo de compuestos alostéricos híbridos, llamados ago-alostéricos⁸⁶ los cuales presentan propiedades como agonistas alostéricos, además de modular alostéricamente al ligando ortostérico tanto en su eficacia y/o su potencia. Los compuestos ago-alostéricos generalmente actúan como co-agonistas, es decir, añaden eficacia y a su vez actúan mejorando la potencia agonista del ligando endógeno, por tanto presentan cooperatividad entre el sitio alostérico y el ortostérico, puesto que el ligando ortostérico puede unirse al receptor al igual que lo hace el modulador ago-alostérico⁸⁷. Sorprendentemente, los sitios activos de todos los moduladores ago-alostéricos caracterizados, presentan superposición con los sitios activos ortostéricos de los ligandos agonistas endógenos⁸⁶. Este punto es de vital importancia dentro del presente trabajo, principalmente y como se muestra en el capítulo dedicado al alosterismo en CXCR4.

El efecto de los moduladores ago-alostéricos puede ser positivo respecto a la eficacia (co-agonista) y positivo en cuanto a su potencia, favoreciendo el agonista endógeno. Sin embargo, también puede ser negativo o inhibidor en términos de potencia, al bloquear la salida del agonista endógeno, mientras que puede ser positivo en términos de eficacia, en lo que se refiere a su papel como co-agonista.

Otras tipologías híbridas de posibles ligandos de GPCR son: los ligandos bitópicos o dualestéricos⁸⁸ los cuales pueden actuar simultáneamente en el sitio ortostérico a la vez que en el sitio alostérico de una misma proteína GPCR. Este tipo de ligandos ofrece teóricamente las ventajas de poder mejorar la afinidad y/o selectividad de la GPCR diana y en principio no requiere de la presencia del ligando agonista endógeno para activar el mecanismo esperado⁸⁹. También existen ligandos bivalentes⁹⁰ los cuales se unen simultáneamente en los sitios activos de receptores adyacentes, principalmente en estructuras diméricas. El diseño computacional de este tipo de ligandos acostumbra a realizarse mediante la construcción de dos farmacóforos químicos unidos por un espaciador o bien mediante péptidos, como es el caso de DV1-K-(DV3), el cual se uniría a dos receptores de CXCR4 de forma simultánea⁹¹.

Por último, una de las nuevas formas terapéuticas de regular GPCRs es mediante el uso de ligandos selectivos (*biased ligands*)⁹², estos ligandos se caracterizan por activar un receptor, seleccionando la respuesta de todas las posibles que el receptor activa de forma selectiva. Es decir, los ligandos selectivos permiten diferenciar para un mismo receptor, una respuesta estimulándola, de otra respuesta inhibiéndola, dadas las múltiples funciones que un mismo receptor puede realizar⁹³.

B.2 El receptor quimiocina CXCR4.

2.2.1 Características del receptor CXCR4: El receptor CXCR4 se clasifica dentro de los receptores GPCR, en la familia de la Rodopsina. Dentro de esta familia, se subclasifica en el grupo Rodopsina- γ , que a su vez se divide en varios clústeres donde encontramos el clúster de las quimiocinas. Las quimiocinas se subdividen en dos subclústeres que son las quimiocinas C-C y las quimiocinas C-X-C. Existen 8 receptores C-X-C, entre los cuales hallamos al receptor CXCR4. El sistema humano contiene más de 40 quimiocinas de las cuales 18 son receptores. Los receptores quimiocinas se definen por su habilidad de inducir de forma direccional la migración de células hacia el gradiente de citoquinas quimiotácticas (quimiotaxis).

El receptor CXCR4 es una proteína de membrana que se activa mediante el proceso de trasducción de señal a través de la quimiocina homeostática de pequeño tamaño: CXCL12, también llamada factor SDF-1 o factor derivado del estroma-1. CXCL12 es el único ligando natural de CXCR4 pero no es el único receptor al que activa, puesto que también actúa como ligando endógeno del receptor CXCR7. A su vez, el receptor CXCR7 puede formar⁹⁴ heterodímeros con CXCR4 lo que nos indica su estrecha relación en los procesos de activación e inhibición de ambos receptores. En humanos existen^{95,96} 6 isoformas del factor derivado estroma-1 que son: SDF-1 α , SDF-1 β , SDF-1 γ , SDF-1 δ , SDF-1 ϵ , SDF-1 ϕ . La isoforma SDF-1 α codifica 89 aminoácidos mientras que la SDF-1 β codifica 93 aminoácidos de los cuales, los primeros 89 son idénticos que en SDF-1 α . SDF-1 γ codifica 119 aminoácidos, SDF-1 δ codifica 120, mientras que SDF-1 ϵ codifica 90 y SDF-1 ϕ 100 aminoácidos, respectivamente⁹⁶. Este factor derivado del estroma juega un papel determinante en el desarrollo de la hematopoyesis, o formación de los componentes celulares de la sangre. Esta funcionalidad tan vital del receptor CXCR4, implica su participación en numerosos procesos de regulación celular. A continuación y de forma abreviada, destacamos algunos de los procesos y enfermedades⁹⁷ en donde media su regulación:

- Una función importante del complejo CXCR4-CXCL12 está relacionada con la reparación de tejidos, concretamente en la reparación del sistema isquémico. Debido a la falta de oxígeno (hipoxia), se ha observado⁹⁸ que el factor HIF-1 induce la expresión de CXCL12 en las zonas

isquémicas de forma directamente proporcional a la carencia de oxígeno. Adicionalmente el factor HIF-1 potencia la expresión y función del receptor CXCR4 en células malignas, este hecho explicaría el mecanismo de cómo el receptor CXCR4 es inducido durante el crecimiento tumoral celular, el cual permitiría a las células neoplásticas emerger de áreas con bajos niveles de oxígeno⁹⁹.

- El ligando natural CXCL12 actúa atrayendo y ayudando⁹⁹ a la supervivencia de las células B de la leucemia linfática crónica, las cuales se acumulan en la sangre, en la médula, así como en el sistema linfóide. De este modo, el complejo CXCR4-CXCL12 se convierte así en una diana terapéutica para tratar las leucemias más frecuentes. Igualmente, otras enfermedades ligadas a la maduración de las células B y en las que interviene el receptor CXCR4, son los mielomas múltiples, así como otros linfomas.
- En el cáncer de pecho también se observa una mayor expresión del ligando CXCL12¹⁰⁰, generada por los fibroblastos estrómicos en estos tejidos.
- Uno de los cánceres más agresivos que forman metástasis del neoplasma, es el cáncer de pulmón de células pequeñas (SCLC), en este tipo de cáncer, al igual que en los cánceres gastrointestinales¹⁰¹ o el cáncer de pulmón de células no pequeñas (NSCLC), también se observan una mayor expresión del receptor CXCR4¹⁰².

Es importante indicar que diferentes tipos de cánceres, expresan diferentes tipos de quimiocinas, sin embargo el receptor CXCR4 es el único que se expresa en la mayoría de cánceres, así como en el proceso de metástasis¹⁰³. Como mínimo existen 23 tipos de cánceres cuyas células tumorales expresan CXCR4¹⁰⁴.

- En procesos inflamatorios, las quimiocinas juegan un papel relevante al ser expresadas tanto en células vasculares endoteliales, como en epiteliales. Concretamente, la expresión del mRNA del receptor CXCR4 se realiza en células epiteliales y endoteliales¹⁰⁵ de múltiples tejidos y por tanto, juega un papel clave en la fisiología y patología inflamatoria.
- Relacionado con los procesos anti-inflamatorios, el receptor CXCR4 en células grasas o lipocitos, parece promover el consumo energético previniendo el excesivo reclutamiento de leucocitos en el tejido adiposo y limitaría por tanto, la obesidad. Este rol defensivo de CXCR4 en el tejido adiposo, se produciría al estimular su señal y/o su expresión, lo que se traduciría en un incremento del consumo energético¹⁰⁶.
- El receptor CXCR4 está involucrado en procesos de insuficiencia renal aguda. La sobreexpresión de CXCR4 en células madre mesenquimatosas derivadas de la médula ósea,

produce una migración de dichas células hacia la zona afectada de insuficiencia renal aguda, reparándola¹⁰⁷.

- En diabetes tipo 2 se han hallado pacientes con reducción de niveles de células que expresan CXCR4 y CD34/CXCR4, de forma correlacionada a la severidad de la enfermedad¹⁰⁸.
- Se han hallado mutaciones en CXCR4 causantes de la enfermedad poco común de inmunodeficiencia síndrome de WHIM^{109,110}, una neutropenia crónica asociada con hiperplasia de médula ósea (myelokathexis). Los pacientes que la padecen exhiben una creciente susceptibilidad a las infecciones bacterianas y virales.
- Además de ser uno de los co-receptores en la entrada del virus HIV, el receptor CXCR4 parece intervenir en la entrada de varios virus del ADN y ARN como son: Herpesvirus, Epstein Barr virus, citomegalovirus, Poxvirus o el virus del papiloma humano (HPVs)¹¹¹.

Todas estas tipologías y patologías en las que participa el receptor CXCR4, hacen de él una diana terapéutica muy importante, por lo que el desarrollo de un mismo inhibidor de CXCR4 puede resultar efectivo en el tratamiento múltiple de estas enfermedades^{112,113}.

Además, el receptor CXCR4 habitualmente forma homo y hetero-oligómeros, esta oligomerización¹¹⁴ parece jugar un papel importante en la regulación alostérica de su comunicación celular¹¹⁵. Concretamente, se ha reportado¹¹⁶ que el co-receptor CXCR4 se dimeriza con CCR2 y CCR5 y que ambos complejos muestran cooperación negativa con sus ligandos, tanto *in vivo* como *in vitro*, con las consecuentes implicaciones que conlleva en la eficacia de sus fármacos.

Los mayores obstáculos a la hora de caracterizar las estructuras proteicas de los GPCR, han sido debidos a la inestabilidad que presentan al estar fuera de la célula, junto al hecho de que pueden adoptar diversos estados conformacionales. Estas características se traducen en que los GPCRs presentan una gran fragilidad y flexibilidad a la hora de intentar obtener cristales de calidad y por tanto se hacen especialmente difíciles de cristalizar.

2.2.2 Estructura proteica del receptor CXCR4: Hasta finales del año 2010 las estructuras utilizadas en diseño molecular del receptor CXCR4 habían sido creadas mediante métodos de homología, tomando como modelo la estructura proteica de la Rodopsina Bovina^{117,118}, una proteína GPCR de clase A que fue resuelta mediante cristalografía de rayos X y cuyo PDB (1F88) fue publicado en el año 2000 por Palczewski *et al.*³⁶. Posteriormente Rasmussen *et al.*⁵⁷ consiguieron determinar el receptor β_2 adrenérgico, otro GPCR de clase A que también ha sido utilizada para modelar¹¹⁹ el co-receptor CXCR4. Finalmente en el 2010, Wu *et al.*⁵⁵ caracterizaron también por cristalografía de rayos X, un total de 5 estructuras del receptor quimiocina CXCR4, con resoluciones de 2.5 a 3.2 Å y en forma de

monómeros y de dímeros. De estas 5 estructuras (3ODU, 3OE0, 3OE6, 3OE8, 3OE9), 1 de ellas (3OE0), presenta como ligando antagonista unido, un péptido cíclico (CVX15) y las 4 estructuras restantes, contienen un compuesto orgánico derivado de la isotiourea (IT1t) (Fig. B.2.1.). Las 4 estructuras que tienen como ligando antagonista la molécula orgánica de pequeño tamaño, se encuentran altamente conservadas, presentando los carbonos- α un *RMSD* igual a 0.6 Å. Así mismo, la estructura 3OE0, que contiene el péptido CVX15 presenta diferencias conformacionales en referencia a las estructuras cuyo ligando unido es la molécula orgánica, siendo el *RMSD* de los carbonos- α igual a 0.9 Å. Estas diferencias conformacionales inducidas por el ligando¹²⁰ cíclico peptido, podrían explicar la unión cooperativa observada en ciertos ligandos naturales de CXCR4, así como los efectos de moduladores alostéricos.

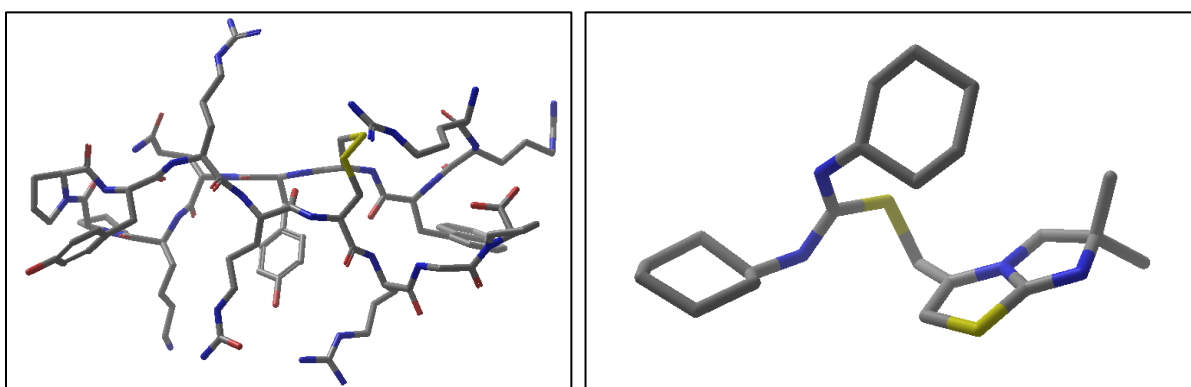


Fig. B.2.1. Ligandos antagonistas co-cristalizados por Wu *et al.* al caracterizar el receptor CXCR4. A la izquierda se observa el péptido cíclico CVX15 y a la derecha el derivado de isotiourea IT1t.

Las principales diferencias que presentan estas estructuras respecto a otros receptores GPCR, cuya estructura también ha sido determinada experimentalmente, son: la hélice I extracelular, si se compara con el receptor β_2 adrenérgico (β_2 AR) se encuentra elevada respecto al eje central del receptor unos 9 Å y si se compara con el receptor adenosina A_{2A} (A_{2A} AR), la diferencia es de unos 3 Å. La hélice II presenta un giro cerrado poco habitual, en el aminoácido Pro92, obteniéndose una rotación de 120° de su extremo extracelular. Este giro puede generar un hueco equivalente a un residuo proteico, el cual al alinear su secuencia con otra secuencia equivalente generada por un modelo de homología y que no contemple dicha rotación, puede generar que mismos residuos proteicos se encaren erróneamente. Así mismo, los extremos intracelulares y extracelulares de la hélice IV del receptor CXCR4 se desvían 5 Å y 3 Å respectivamente de las posiciones consensuadas de otros GPCRs y el extremo final extracelular de la hélice V es un giro más largo en el caso del receptor CXCR4. La hélice VI cuya forma es similar en todas las estructuras, presenta en el extremo final extracelular un desplazamiento de 3 Å respecto a los receptores A_{2A} AR y β_2 AR. Finalmente, la parte extracelular final de la hélice VII es dos vueltas helicoidales más larga que otras estructuras GPCRs.

Además de estas diferencias concretas halladas en cada uno de los dominios de transmembrana,

cabe destacar que globalmente, la estructura del CXCR4 posee el sitio activo que presenta una cavidad más alargada, de mayor apertura y ubicada más cerca de la superficie extracelular, en comparación a otras estructuras GPCRs. Estas características estructurales del receptor CXCR4 y de su sitio activo, denotan una singularidad especial en su actividad y afinidad con otros ligandos y por consiguiente, al aplicar técnicas de diseño molecular, motivo por el cual, se clasifica como una “diana difícil”.

Funcionalmente, se ha constatado la separación física de las funciones de unión y de propagación de la señal dentro del co-receptor, planteándose pues, un modelo de dos sitios activos diferentes. El sitio activo 1 ubicado en el núcleo de la quimiocina, sería el de unión del receptor y el sitio activo 2, ubicado en el extremo N-terminal, sería el responsable de transmitir la comunicación a la proteína G.

Otra constatación del modelo de diferentes sitios activos, es que tanto el ligando péptido como la molécula orgánica de pequeño tamaño, ocupan la misma cavidad donde se ubica el sitio activo, comportándose al mismo tiempo como competidores ortostéricos del ligando natural CXCL12 y bloqueándolo. La molécula orgánica, debido a su menor tamaño, ocupa la zona más profunda de la cavidad extracelular (Fig B.2.2.), lo que generaría diferentes modos de enlace si se compara¹²¹ con los generados por el ligando cíclopéptido CVX15. Esta cavidad extracelular se ubica entre la hélice extracelular 2 (ECL2) y cercana al extremo N-terminal (Fig B.2.2.).

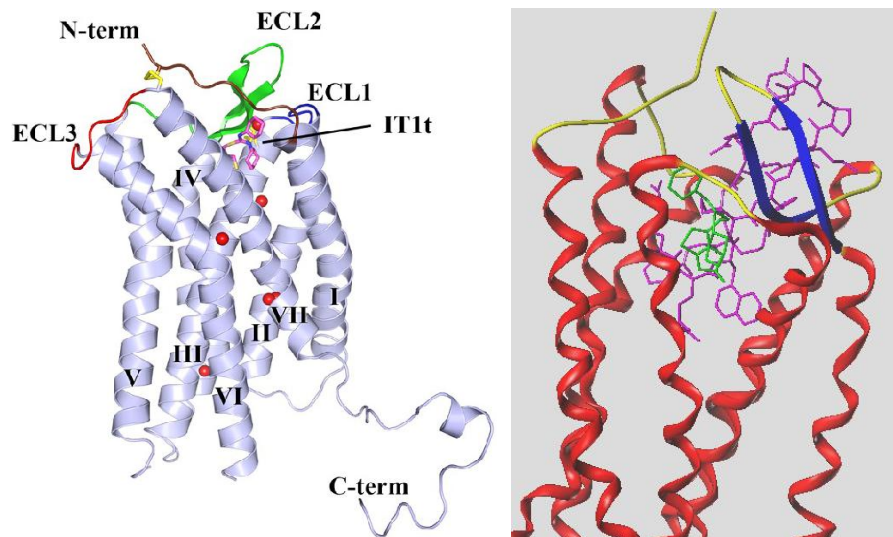


Fig. B.2.2. Estructura proteica del co-receptor celular CXCR4 y ubicación del sitio activo de unión proteína-ligando (a la izquierda y en color en verde). Representación extraída de *Wu et al.*⁵⁵ Se detallan las 7 hélices o dominios helicoidales así como el extremo N-terminal extracelular con sus correspondientes bucles extracelulares (ECL1, ECL2 y ECL3) y el extremo C-terminal intracelular. Ubicado entre el extremo N-terminal y el bucle extracelular 2, se halla el ligando IT1t (en verde a la derecha) y el ligando péptido CVX15 (en lila a la derecha).

Concretamente, el bucle extracelular ECL2, es uno de los bucles extracelulares menos conservado. Ello se constata al comparar¹²² las proteínas GPCRs de clase A cuyas estructuras han sido resueltas

(CXCR4, adrenérgico β_2 , adenosina A_{2A} y rodopsina bovina) ya que la superposición de dichas estructuras, presenta notables diferencias entre sí, tanto a nivel de la longitud de las cadenas, como entre las diferentes secuencias (Fig. B.2.3.).

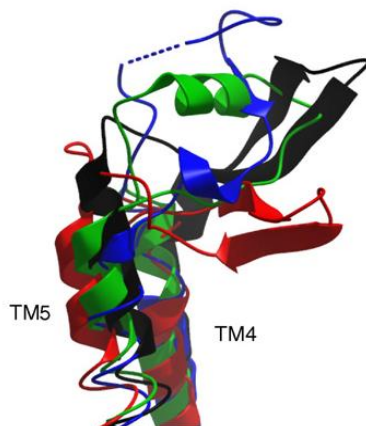


Fig. B.2.3. Superposición de los bucles extracelulares ECL2 de las estructuras GPCRs-clase A. En verde se presenta el esqueleto del receptor adrenérgico β_2 (PDB 2RH1), en color azul la adenosina A_{2A} (PDB 3EML), en rojo la rodopsina bovina (PDB 1U19) y en negro el CXCR4 (3ODU). Representación extraída de *Peeters et al.*¹²²

Contrariamente al concepto de que cadenas proteicas más conservadas indican mayor importancia biológica, el bucle ECL2 se asocia tanto a aspectos de estructura del receptor, como de comunicación celular, así como de selectividad y reconocimiento del ligando. También se considera que en el ECL2 se ubican los sitios activos de unión de los ligandos tanto ortostéricos como alostéricos¹²³.

El bucle extracelular ECL2 puede sufrir cambios conformacionales que parecen ser causados por puentes disulfuros formados adicionalmente y que activarían el receptor. Así, por ejemplo el receptor GPCR de clase A: M_2 muscarínico forma puentes disulfuro extra entre la hélice TM7 y el propio ECL2¹²⁴, esta flexibilidad conformacional actuaría creando conformaciones del bucle ECL2 abiertas, que permiten entrar el ligando y tras la entrada de éste, adoptar una conformación cerrada y así estabilizar el ligando favoreciendo interacciones adicionales nuevas. Así pues el bucle ECL2 actúa como un regulador negativo del receptor, que lo mantiene en estado inactivo, hasta la activación inducida de un agonista.

Igualmente, existen evidencias de que el bucle ECL2 en receptores GPCR clase A, como es el caso del receptor M_2 muscarínico¹²⁴ o de la adenosina A_1 ⁴, puede alojar un sitio activo alostérico. Así, en el caso del receptor M_2 muscarínico se ha demostrado que varios residuos proteicos contribuyen en la interacción alostérica de moduladores de dicho receptor. En el caso del receptor adenosina A_1 , Narlawar *et al.*¹²⁵ proponen que el sitio activo alostérico se halla ubicado muy próximo al segundo bucle extracelular.

Cabe indicar que los sitios activos alostéricos pueden hallarse en regiones del receptor no conservadas y que fueron originadas al azar durante la evolución. Por tanto, es factible que estos sitios activos puedan hallarse en la superficie extracelular del receptor, en bucles poco conservados como el ECL2.

El laboratorio del Grup d'Enginyeria Molecular IQS (GEM) inició la investigación del receptor CXCR4 a partir del proyecto de la Fundación IrsiCaixa en colaboración con el laboratorio de retrovirología del Hospital Germans Trias i Pujol en Badalona, que se centraba en la búsqueda de inhibidores del Virus de Inmunodeficiencia Humana (VIH). Esta colaboración cristalizó en el proyecto de la Fundació La Marató de TV3, así como en el Programa Nacional de Biomedicina (Ministerio de Educación y Ciencia, SAF2007-63622-C02-01) de "Diseño y síntesis de nuevos inhibidores de entrada del VIH". Más recientemente y tras identificar el papel fundamental que desempeña el receptor CXCR4 a nivel oncológico, el laboratorio GEM mantiene relación con el Hospital Clínic de Barcelona en programas de investigación enfocados a terapias anticáncer.

OBJETIVOS

Los objetivos planteados en el presente trabajo de investigación se centran en los siguientes puntos:

Estudio y propuesta de un método de cribado virtual que mejore la recuperación de activos frente a inactivos para “dianas difíciles”. Se definen dianas difíciles como aquellas que poseen un amplio sitio activo, con la posibilidad de acomodar compuestos de diferentes maneras. Aplicación retrospectiva de dicho método de cribado al receptor VEGFR-2.

Estudio funcional de la estructura proteica de CXCR4 a partir de las cinco estructuras cristalográficas publicadas hasta la fecha y de un modelo previamente generado por homología en el laboratorio del Grup d'Enginyeria Molecular IQS (GEM). Análisis mediante cribado virtual retrospectivo basado en la estructura de la proteína (*docking*) del modelo cristalográfico que mejor distingue activos de inactivos. Estudio de los mejores parámetros de *docking* para el cribado virtual de antagonistas de CXCR4. Análisis de la conformación de CXCR4 que mejor propicia la unión de diferentes familias de inhibidores antagonistas al receptor.

Cribado virtual retrospectivo de antagonistas de CXCR4. Comparación del método de cribado virtual basado en la estructura (*docking*) con otros métodos de cribado virtual de antagonistas de CXCR4 basados en la forma tridimensional del ligando (*shape matching*) y modelos farmacofóricos generados a partir de las interacciones proteína-ligando.

Cribado virtual prospectivo de antagonistas de CXCR4. Generación de una quimioteca combinatoria de compuestos para su cribado virtual prospectivo. Aplicación del método de cribado virtual propuesto para “dianas difíciles”, así como los métodos de cribado virtual de antagonistas de CXCR4 más eficientes según el análisis retrospectivo. Propuesta de nuevas moléculas potenciales antagonistas del receptor CXCR4.

Estudio de la modulación alostérica de agonistas y antagonistas de CXCR4. Ubicación de los posibles sitios activos alostéricos de CXCR4. Estudio de las interacciones clave en la unión de los moduladores alostéricos y la proteína CXCR4.

1. FUNDAMENTOS TEORICOS

1.1 Modelización molecular.

Se suele definir la modelización molecular como una aplicación computacional que tiene por objetivo generar, calcular y predecir estructuras y sistemas moleculares, así como sus propiedades asociadas y su comportamiento. Las técnicas que se usan en la modelización molecular se basan en la química teórica y en datos experimentales cuyo análisis permite realizar dicha predicción. El nexo de unión entre la parte experimental y la teórica, permite extraer resultados de un modelo, comparar resultados experimentales de un sistema, comparar predicciones teóricas para un modelo, ayudar a entender e interpretar observaciones experimentales y suministrar información no disponible desde el punto de vista experimental. En este sentido, las técnicas computacionales han revolucionado la modelización molecular, hasta el punto que sin el uso de computadoras, la mayoría de cálculos no podrían ser realizados.

La modelización molecular utiliza métodos de mecánica cuántica, conocidos como métodos de química teórica y al mismo tiempo también utiliza métodos como son los métodos de mecánica molecular, de minimización, las simulaciones, el análisis conformacional y otras aproximaciones computacionales que predicen y ayudan a entender el comportamiento de sistemas moleculares.

1.1.1 Mecánica cuántica: Los métodos de mecánica cuántica representan explícitamente los electrones de un sistema, mediante el cálculo de la distribución electrónica y por tanto se permiten deducir las propiedades físico-químicas, principalmente en reacciones químicas donde los enlaces se forman y se rompen. Esta es una de las principales características que diferencian la mecánica cuántica de los métodos empíricos, como por ejemplo, los basados en campos de fuerza (*Force fields*).

El punto de partida en mecánica cuántica es la ecuación de Schrödinger, que utiliza el Hamiltoniano [1.1] el cual incluye la energía cinética electrónica y nuclear, así como el potencial del sistema (repulsión nuclear, repulsión electrónica y atracción electrón-núcleo). La ecuación de Schrödinger describe el movimiento de una partícula de masa m (electrón) en el espacio y el tiempo bajo la influencia de un campo externo (V), como puede ser el potencial electrostático de los núcleos de los átomos de una molécula y del resto de electrones. Si el potencial (V) es independiente del tiempo, la ecuación de Schrödinger puede ser descrita como:

$$\left\{-\frac{\hbar^2}{2m}\nabla^2 + V\right\}\psi(r) = E\psi(r); \quad \text{si } \mathcal{H} = -\frac{\hbar^2}{2m}\nabla^2 + V; \quad \mathcal{H}\psi = E\psi \quad [1.1]$$

En general, la ecuación de Schrödinger sólo puede resolverse exactamente para unos pocos problemas, como es el caso de una partícula en una caja, el oscilador armónico, el átomo de hidrógeno, etc. Además, la resolución de la ecuación de Schrödinger en átomos con más de un electrón es un proceso matemático muy complicado y no puede ser resuelta con exactitud, por lo que requiere del uso de aproximaciones. La aproximación de Born-Oppenheimer, considera el núcleo fijo y por tanto desprecia su energía cinética y se centra en los movimientos electrónicos, por lo que en cada cambio del núcleo se resuelve la ecuación de Schrödinger para los electrones solos, en el campo del núcleo: $E_{tot}=E_{(electrones)} + E_{(núcleo)}$. Para el cálculo de los orbitales moleculares en sistemas de varios electrones, se utilizan las ecuaciones de Hartree-Fock que considera todos los electrones iguales y contempla las interacciones entre electrones, expresándolas como integrales de intercambio y de Coulomb¹²⁶.

Cuando hablamos de métodos cuánticos *ab initio*, nos referimos a cálculos que usan las ecuaciones de Hartree-Fock y Roothaan-Hall las cuales no ignoran o bien aproximan las integrales de algunos de los términos del Hamiltoniano (\mathcal{H}). Por el contrario, cuando nos referimos a los métodos semi-empíricos, éstos realizan simplificaciones utilizando ciertos parámetros para las integrales, o bien ignoran términos del Hamiltoniano¹²⁷. A efectos prácticos podemos simplificar las diferencias existentes entre estos métodos, indicando que los métodos *ab initio* tienen en cuenta todos los electrones y que por tanto su uso se limita a sistemas con un máximo de decenas de átomos. Estos métodos requieren potentes ordenadores para efectuar sus cálculos. En cambio, los métodos semi-empíricos, son métodos mucho más rápidos de calcular, simplifican el sistema ignorando la existencia de ciertos electrones y por tanto su uso es aplicable a sistemas compuestos por centenares de átomos.

Como ejemplos de métodos *ab initio*, tenemos los diferentes métodos post-Hartree-Fock¹²⁸, Moller-Plesset (MPn)¹²⁹, Configuración interacción (CI)¹³⁰, Clústeres acoplados (CC)¹³¹, método MCSCF¹³², método MRCI¹³³, método NEVPT¹³⁴, entre otros. Mientras que como ejemplos de métodos semi-empíricos tenemos CNDO/2¹³⁵, MNDO¹³⁶, INDO¹³⁷, NDDO¹³⁸.

1.1.2 Mecánica molecular: La mayoría de los problemas y retos que afronta el diseño molecular desafortunadamente son sistemas demasiado grandes y complejos como para utilizar la mecánica cuántica. Aun e ignorando algunos de los electrones de un sistema, como hacen los métodos semi-empíricos cuánticos, un elevado número de átomos y/o moléculas requiere un número excesivamente elevado de cálculos y por tanto de tiempo computacional. Bajo esta premisa, se desarrollan los métodos basados en los campos de fuerza o *Force fields*, los cuales ignoran los movimientos electrónicos y calculan la energía del sistema únicamente en función de las posiciones de los núcleos. Estos métodos aplicables a sistemas de gran tamaño, pueden devolver resultados igual de precisos que los cálculos cuánticos más sofisticados, sin embargo la mecánica molecular no puede predecir propiedades que dependen directamente de la distribución electrónica en una molécula.

La mecánica molecular se basa en un modelo simple de interacciones existentes dentro de un sistema. En dicho sistema se producen contribuciones de procesos como el de los alargamientos de los enlaces, el proceso de apertura y cierre de los ángulos de enlace y las rotaciones de los enlaces simples.

Los supuestos en que se basa la mecánica molecular son¹³⁹:

- 1- Los núcleos y los electrones se agrupan juntos y son tratados como átomo-partículas unidas.
- 2- Los átomos-partículas son tratados como esferas.
- 3- Los enlaces entre partículas son considerados como muelles.
- 4- Las interacciones entre estos átomo-partículas se tratan utilizando funciones derivadas de la mecánica clásica.
- 5- Las funciones de potencial individuales son usadas para describir los diferentes tipos de interacción.
- 6- Las funciones de energía potencial se basan en parámetros empíricos que describen las interacciones entre conjuntos de átomos.
- 7- Las funciones de potencial y los parámetros utilizados para evaluar las interacciones los llamamos *Force fields*.
- 8- La suma de interacciones determina la conformación de los átomos-partículas.

Los *Force Fields* expresan la energía mediante funciones de potencial que incluyen términos de energía internos, de la propia molécula asociados a átomos unidos covalentemente y términos de energía externos, como la interacción con otras moléculas, principalmente interacciones de Van der Waals y electrostáticas (Fig. 1.1.1.). De forma genérica las expresiones comúnmente utilizadas de los *Force fields* en compuestos biomoleculares¹⁴⁰ son:

$$U(R) = U(R)_{interna} + U(R)_{externo} \quad [1.2]$$

$$U(R)_{interna} = \sum_{\text{enlaces}} K_b(b - b_0)^2 + \sum_{\text{angulos}} K_\theta(\theta - \theta_0)^2 + \sum_{\text{diedros}} K_\chi[1 + \cos(n\chi - \sigma)] + \sum_{\substack{\text{diedros} \\ \text{impropios}}} K_{imp}(\varphi - \varphi_0)^2$$

$$U(R)_{externo} = \sum_{\substack{\text{pares átomos} \\ \text{no enlazados}}} \left(\epsilon_{ij} \left[\left(\frac{R_{\min,ij}}{r_{ij}} \right)^{12} - \left(\frac{R_{\min,ij}}{r_{ij}} \right)^6 \right] + \left(\frac{q_i q_j}{\epsilon_D r_{ij}} \right) \right)$$

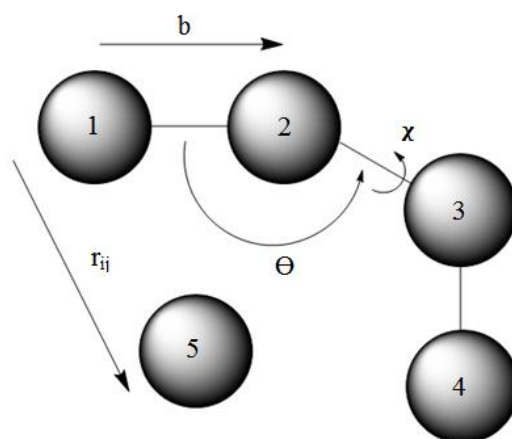


Fig. 1.1.1. Definición básica de parámetros en *Force fields* en una molécula de 4 átomos y una molécula de 1 átomo. Los términos de energía internos en la molécula de 4 átomos vienen dados por: las longitudes de enlace (b) entre 1-2, 2-3, 3-4, los ángulos (Θ) entre 1-2-3 y 2-3-4, las torsiones (χ) entre cada enlace simple y las interacciones externas con el átomo 5 vienen dadas por las diferentes distancias de no-enlace (r_{ij}) entre 5 y 1,2,3,4.

Existe numerosos *Force fields* cuyo uso depende en gran parte de la tipología del compuesto a tratar. Entre ellos destacamos: la familia MM (MM2, MM3 y MM4)¹⁴¹, Merck Molecular Force Field (MMFF)¹⁴², TRIPOS¹⁴³, AMBER¹⁴⁴, CHARMM¹⁴⁵; GROMOS¹⁴⁶, OPLS¹⁴⁷ y CFF¹⁴⁸.

Los *Force fields* utilizados en el presente trabajo son MMFF94, TRIPOS, CHARMM y la versión CHARMM22.

- MMFF94: El *Force field* MMFF94¹⁴² es muy similar al MM3 pero difiere en su aplicación en procesos de fase condensada en dinámicas moleculares. Obtiene precisiones similares a MM en moléculas de pequeño tamaño y es aplicable también a proteínas y otros sistemas biológicos significantes. Está diseñado para ser aplicado a compuestos farmacológicos en los que se tratan las energías conformacionales con precisión. Es adecuado para utilizarse en cálculos fase gas y en fase condensada. Su cálculo comprende términos de energía internas propias del enlace como son: los enlaces, ángulos, alargamientos del enlace, flexiones fuera del plano y ángulos diedros, así como términos no propios del enlace, como son las energías de Van der Waals y las electrostáticas.
- TRIPOS: El *Force field* Tripos es ajustable y calcula¹⁴⁹ el potencial energético en base a las contribuciones de enlace, de la flexión de los ángulos, flexiones de átomos fuera del plano, de la energía de torsión y de la energía de Van der Waals. De forma opcional permite incluir la contribución electrostática, un término energético multiajustable y los términos energéticos de restricción a la distancia, de restricción al ángulo, de restricción a la torsión y de restricción al rango energético.
- CHARMM: El *Force field* fue parametrizado en base a datos experimentales por Karplus et al¹⁵⁰. CHARMM es utilizado tanto en moléculas pequeñas, como en complejos solvatados y

macromoléculas en los cuales ofrece buenos resultados en simulaciones y cálculos. Su fórmula del potencial incorpora ciertas variaciones respecto a la formula general [1.2], puesto que incorpora un término del equilibrio de Urey-Bradley ($U_{\text{Urey-Bradley}}$), así como el termino de corrección numérica (CMAP), para la cadena principal proteica y ciertas modificaciones en el cálculo del potencial externo o intramolecular.

$$U_{\text{Urey-Bradley}} = \sum K_{UB}(S - S_0)^2 \quad ; \quad U_{\text{CMAP}} = \sum_{\text{residuos}} U_{\text{CMAP}}(\phi, \psi) \quad [1.3]$$

$$U(R)_{\text{externo}} = \sum_{\substack{\text{pares átomos} \\ \text{no enlazados}}} \left(\epsilon_{ij} \left[\left(\frac{R_{\text{min},ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{\text{min},ij}}{r_{ij}} \right)^6 \right] + \left(\frac{q_i q_j}{4\pi\epsilon_0\epsilon r_{ij}} \right) \right)$$

Los *Force field* CHARMM22 y CHARMM27 son los *Force field* más extendidos en biomoléculas como proteínas, ácidos nucleicos, lípidos y de forma limitada a carbohidratos¹⁴⁰.

<i>Ab initio</i>	<i>Semi-empíricos</i>	<i>Mecánica molecular</i>
En cuenta todos los electrones	Simplifica ignorando algunos electrones	Ignora todos los átomos. Sólo se consideran núcleos
Limitado a decenas de átomos con potentes ordenadores	Limitado a centenares de átomos	Moléculas que contienen miles de átomos
Puede aplicarse a inorgánicos, orgánicos, organometálicos y fragmentos moleculares	Puede aplicarse a inorgánicos, orgánicos, organometálicos y pequeños oligómeros.	Puede aplicarse a inorgánicos, orgánicos, oligonucleótidos, péptidos, sacáridos y metalo-orgánicos
Extensible a medio en vacío o en solvente implícito	Extensible a medio en vacío o en solvente implícito	Extensible a medio en vacío o en solvente implícito o explícito
Aplicable al estado fundamental, de transición y excitados. Permite estudiar la formación y rotura de enlaces (reacción)	Aplicable a estado fundamental, de transición y excitados	Solo aplicable al estado fundamental. No permite estudiar la formación y rotura de enlaces (reacción)

Fig. 1.1.2. Comparativa entre métodos ab initio, semi-empíricos y técnicas de mecánica molecular¹³⁹.

1.1.2.1 Metodos de minimización: Una de las tareas básicas en mecánica molecular es la minimización de la energía estructural de una molécula, cuando esta se modifica al variar las posiciones de sus átomos. Esto significa minimizar la energía total de deformación, representada por la ecuación del *Force field*, respecto a variables independientes como son las coordenadas cartesianas de los átomos. La minimización energética de estructuras moleculares debe tener en cuenta las siguientes consideraciones:

- La configuración más estable de una molécula puede ser hallada minimizando su energía libre de Gibbs (G).
- Típicamente, la energía se minimiza asumiendo que el efecto de la entropía puede negligirse.
- En el mínimo del potencial energético, la fuerza neta sobre cada átomo desaparece y por tanto la configuración es estable.

- Como la energía cero es arbitral, la energía calculada en la minimización es relativa. Por tanto, sólo es útil comparar energías de diferentes configuraciones, pero de mismos compuestos químicos.
- Es difícil determinar si un mínimo energético particular es el mínimo energético global, el cual es el punto energético más bajo, donde la fuerza es cero y la matriz de la segunda derivada es positiva y definida, para determinarlo es necesario explorar toda la superficie de potencial.

Los métodos más utilizados se dividen en dos categorías: basados en el algoritmo del gradiente descendiente (*Steepest descent*) que utilizan la primera derivada, incluidos otros métodos homólogos, como el gradiente conjugado (*Conjugate gradient*). La segunda categoría incluye los métodos basados en el método de *Newton-Raphson* y que utilizan la segunda derivada. Los métodos basados en *Newton-Raphson* tienen un coste computacional muy elevado y por este motivo se han desarrollado otros métodos de segunda derivada como el Quasi-Newton, los métodos *truncated-Newton* o bien *Broydon-Fletcher-Goldfarb-Shanno (BFGS)* los cuales realizan una estimación en el proceso iterativo.

El principal problema que presenta el método del gradiente descendiente (*Steepest descent*) es la determinación del tamaño apropiado del incremento iterativo, usado en los movimientos de los átomos de cada ciclo derivativo. El tamaño de estos incrementos determina la eficacia de la minimización y la calidad del resultado. En cambio, este método presenta una relativa facilidad de cambiar de *Force field*.

El método del gradiente conjugado utiliza para dirigir la minimización ambos gradientes, el gradiente en curso y la dirección previa buscada. Esta característica hace converger la búsqueda más rápidamente que el método del gradiente descendiente y optimiza el incremento iterativo. Sin embargo, el número de ciclos iterativos que éste método requiere es proporcional al número de átomos y el tiempo por cada ciclo es proporcional al cuadrado del número de átomos, lo que indica el coste computacional que el método requiere.

Los métodos de minimización según *Newton-Raphson* utilizan para localizar el mínimo energético, la curvatura de la superficie que el cálculo de la energía de deformación genera. Los cálculos son considerablemente más complejos que los métodos de primera derivada, pero utilizan la información disponible de forma más completa y convergen más rápidamente.

La estrategia más general al trabajar con proteínas o ácidos nucleicos, es utilizar el método del gradiente descendiente en los primeros 500-1000 ciclos y después utilizar el gradiente conjugado o el método de *Newton-Raphson* para completar la minimización. Habitualmente para converger el resultado se utiliza el gradiente RMS ($RMS < 0.10$) o bien las diferencias energéticas.

El resultado del mínimo calculado es a menudo utilizado como punto de partida en cálculos sucesivos, bien para eliminar posibles deformaciones existentes en la molécula, o bien como preparación previa a una simulación de dinámica molecular.

1.2 Modelos basados en la estructura.

Actualmente, el diseño de fármacos basado en la estructura es una aproximación muy utilizada en el descubrimiento farmacológico. Además, el uso de métodos computacionales se ha extendido ampliamente en varias de las etapas del proceso del diseño de fármacos. Entre estos métodos computacionales y basados en la estructura encontramos el *docking* molecular como uno de los más utilizados.

1.1.1 *Docking*: Podemos definir el *docking* molecular como un algoritmo de cálculo que determina como un compuesto se une en el sitio activo de una proteína diana. Como resultado, este método define tanto la conformación, como la orientación de la molécula ligando respecto a la proteína-receptor y también su valor de *scoring*. Definimos *scoring* como el valor de energía entre la proteína y el ligando, el cual puede ser expresado como la energía de unión, la energía libre o bien mediante un valor numérico cualitativo que representa la medida de la energía. El *docking* molecular proteína-ligando es una potente herramienta que permite estudiar y comprender las interacciones que se establecen entre la proteína diana y el ligando.

El método de *docking* se considera como uno de los métodos más eficaces a la hora de predecir si un compuesto puede ser un buen inhibidor de una diana macromolecular. Es por ello que una de las principales aplicaciones de *docking* es su uso en cribados virtuales¹⁵¹ de quimiotecas que contienen un gran número de compuestos, de manera que los compuestos con mejores resultados serán candidatos a ser sintetizados y analizados en el laboratorio.

El algoritmo del *docking* explora automáticamente el sitio activo de la proteína, intentando unir el ligando a la proteína ensayando diferentes orientaciones y posiciones (poses). Para cada pose ensayada, el método calcula al mismo tiempo su *scoring* energético.

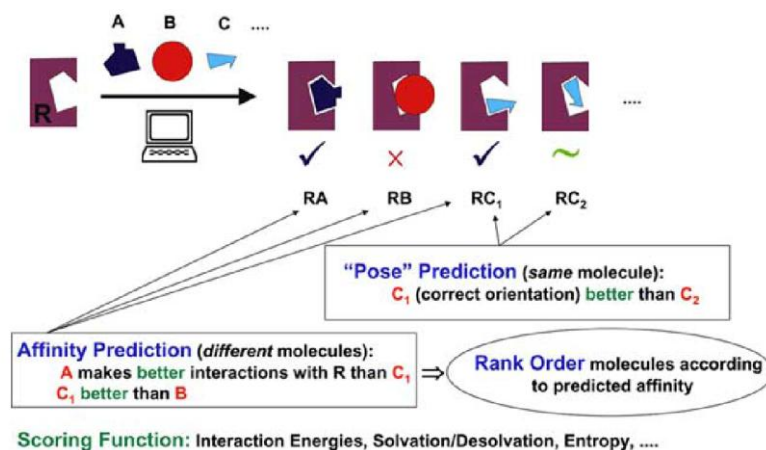


Fig. 1.2.1. Ilustración¹⁵² esquemática del *docking* molecular y del uso del *scoring* para clasificar los ligandos A,B,C y las diferentes poses de C en el receptor R.

La mayoría de métodos de *dockings* consideran la proteína como una estructura rígida¹⁵³, si bien existen también protocolos que consideran flexibilidad parcial de las cadenas laterales¹⁵⁴, o bien flexibilidad limitada a ciertos dominios¹⁵⁵ de la proteína.

En referencia al ligando, los métodos de *docking* generan y ubican las poses en el sitio activo (predicción de poses) mediante algoritmos de muestreo del ligando. Existen tres tipos de estos algoritmos. Los algoritmos basados en la forma del sitio activo (*shape matching*)¹⁵⁶ los cuales alinean las poses en base a la forma del sitio activo. Son los algoritmos más sencillos y se basan en la complementariedad de la superficie del sitio activo, con la superficie del ligando. Otro tipo de algoritmos es el llamado de búsqueda sistemática (*systematic search*)¹⁵⁷, que sistemáticamente rotan todos los enlaces rotables del ligando en base a unas restricciones que acotan la búsqueda. Por último, los algoritmos estocásticos (*stochastic algorithms*)¹⁵⁸ muestrean las conformaciones y orientaciones del ligando mediante constantes cambios al azar del espacio rotacional y traslacional del ligando. Existen principalmente cuatro tipos de algoritmos estocásticos: de Monte Carlo, métodos *Tabu Search*, algoritmos evolutivos y algoritmos de optimización de enjambre o nube de partículas (*swarm optimization methods*)¹⁵⁹.

La búsqueda conformacional es un proceso que puede realizarse de forma independiente antes del *docking*¹⁶⁰ o integrado como una parte del *docking*¹⁶¹.

Los algoritmos de muestreo del ligando utilizados en el presente estudio son:

- *AutoDock Vina*: Método de muestreo del ligando basado en algoritmos estocásticos del tipo *Iterated LocalSearch Global Optimizer* y que consisten en mutar el ligando y optimizar localmente utilizando el método de Broyden-Fletcher-Goldfarb-Shanno (BFGS)¹⁶².
- *Glide*: Este método se basa en un algoritmo de muestreo del ligando mediante búsqueda sistemática (*systematic search*) de la orientación, conformación y posición en el espacio, del

ligando. La búsqueda exhaustiva utiliza aproximaciones y truncamientos con el fin de obtener tiempos computacionales aceptables¹⁵⁷.

- *Gold*: Se basa en el muestreo del ligando mediante un algoritmo estocástico. El método de cálculo está basado en un algoritmo genético (GA) por el cual se modifican y optimizan los ángulos diedros de los enlaces del ligando que pueden rotar las geometrías de los anillos del ligando, y los ángulos diedros de los OH y NH₃⁺ de la proteína. Se parte al inicio del *docking*, de valores de estos parámetros al azar¹⁶³.
- *LigandFit*: La generación de conformaciones se realiza mediante métodos estocásticos del tipo Monte Carlo, en donde se varían los ángulos de torsión. *LigandFit* es un método de *docking* basado en *Shape matching*, en donde la forma del ligando es comparada con la forma del sitio activo y si se considera aceptable se calcula su energía de *docking*¹⁶⁴.
- *MOE Docking*: El *docking* integrado en MOE¹⁶⁵ realiza una búsqueda estocástica utilizando métodos del tipo *Tabu Search*, que son una versión modificada de las búsquedas Monte Carlo y que se caracterizan por la realización de pequeños cambios al azar, determinando que cambios son tabú¹⁶⁶. La búsqueda permite realizar el *docking* flexible en el sitio activo, o realizar el *scoring* en una base de datos¹¹.
- *Surflex*: es un método que utiliza un algoritmo automático de flexibilidad molecular total, que combina la función de *scoring* empírica Hammerhead, conjuntamente con una búsqueda conformacional de poses en el sitio activo, basada en la similitud de la superficie molecular (similitud morfológica) y que permite la rápida generación de poses¹⁵⁶.

Muchos ensayos bioquímicos determinan la constante de inhibición K_i (constante del tipo Arrhenius), por lo que es de esperar que la energía de enlace calculada por el *docking* debería ser proporcional a dicha constante. Algunas funciones de *scoring* presentan una relativa correlación con el valor del $\ln K_i$, aunque la mayoría de funciones solamente muestran un ranking cualitativo sobre que compuestos son mejores y cuales peores. Esta falta de correlación es precisamente una de las mayores debilidades de los métodos de *docking*, conjuntamente con el elevado coste computacional que requiere, principalmente cuando se le añade flexibilidad a la proteína.

Otro de los aspectos críticos a la hora de realizar el *docking* es la conveniencia de considerar las moléculas de agua como una parte flexible del sitio activo de la proteína y por tanto ser consideradas críticas a la hora de establecer interacciones proteína-ligando, o bien, ser consideradas como “artefactos” durante el proceso de cristalización y por tanto su contribución a la energía de interacción ser nula. Algunos procedimientos de *docking* incluyen el cálculo de posiciones favorables de moléculas de agua, previo a realizar el *docking*, con el fin de analizar la idoneidad de estas moléculas en el sitio activo. En otros casos se utilizan funciones de solvatación con el fin de corregir

la energía para el compuesto ligando expuesto al solvente. Sin embargo se debe tener en cuenta que no es apropiado aplicar términos de solvatación a átomos ubicados en el sitio activo pero no expuestos al solvente.

Asimismo, otra limitación que se puede presentar un *docking* es que moléculas de gran tamaño, como compuestos peptidomiméticos, por el hecho de tener hipotéticamente más interacciones con la proteína receptor, puedan presentar mejores valores de *scoring* que moléculas pequeñas.

1.1.2 Funciones de *Scoring*: Las funciones de *Scoring* calculan el valor de la energía de afinidad entre la proteína y el ligando, por lo que determinan la precisión del algoritmo. La velocidad y la exactitud son aspectos importantes en dichas funciones. Existen numerosas funciones de *scoring* desarrolladas en las últimas décadas, habitualmente estas funciones se agrupan en función de su naturaleza en tres métodos: funciones basadas en *Force fields*, empíricas (*empirical*) y *knowledge-based*.

Las funciones basadas en *Force fields* descomponen la energía de interacción del ligando en términos de interacción, como son las energías de van der Waals, las energías electrostáticas, y las energías propias del ligando, que se componen de las energías de torsión y de flexión del enlace. Generalmente son derivadas de los *Force fields* AMBER y CHARMM. Estas funciones presentan como mayores retos el tratamiento de los efectos de solvatación producidos por las moléculas de agua y los efectos entrópicos.

En las funciones empíricas, la energía de enlace se calcula sumando de forma ponderada y empíricamente la suma de términos energéticos, como es la energía de van der Waals, la energía electrostática, las energías generadas por enlaces de puente de hidrógeno, términos de solvatación, de entropía del ligando y de hidrofobicidad.

Las funciones *knowledge-based* son funciones que se derivan de la información estructural determinada experimentalmente en complejos proteína-ligando. Estas funciones generalmente ofrecen un buen balance entre precisión y velocidad de cálculo, principalmente al compararlas con las funciones *Force field*. Uno de los principales retos de estos métodos es el cálculo preciso de los efectos entrópicos del ligando.

En el presente estudio se han utilizado las siguientes funciones de *scoring*:

- Affinity dG: La función de MOE Affinity dG es una función de *scoring* basada en *Force field* que incluye contribución de energía libre, interacciones de enlace puente de hidrógeno, interacciones iónicas, interacciones de metal-ligando, interacciones hidrofóbicas, interacciones de entre átomos hidrofóbicos y átomos polares e interacciones generales entre 2 átomos¹⁶⁵.

- Alpha HB: La función Alpha HB de MOE es una combinación lineal de dos términos. El primero mide como se ajusta geoméricamente el ligando al sitio activo y el segundo término mide los efectos de los enlaces por puentes de hidrógeno. Ambos términos son sumados para todo el conjunto de átomos.¹⁶⁵
- ASE: Es una función de MOE que evalúa el *scoring* de forma proporcional a la suma de Gaussianas para todos los átomos del ligando y del receptor, así como para todos los pares de átomos del ligando y esferas alfa definidas¹⁶⁵.
- AutoDock: Utiliza el método de *Broyden-Fletcher-Goldfarb-Shanno (BFGS)* para la optimización local¹⁶², un método *quasi Newton* que usa el valor del *scoring* así como su gradiente durante la optimización.
- ChemScore: La función empírica de ChemScore se basa en la suma de las contribuciones lipofílicas, de enlaces por puentes de hidrógeno, de interacciones metal-ligando, de flexibilidad del ligando y de su energía libre.
- DockScore: La función *DockScore* se basa en las aproximaciones *Force field*. Principalmente incorpora la energía de interacción proteína-ligando y la energía interna del ligando calculada como energía de van der Waals. La energía de interacción engloba la suma de energías de van der Waals y energía electrostática, que en ambos casos son suavizadas para evitar tendencias¹⁶⁷.
- GlideScore: La función de *scoring* de Glide se basa en la función empírica de ChemScore. Las principales modificaciones respecto a ChemScore se producen en las contribuciones de puentes de hidrógeno y en las interacciones metal-ligando. Además, GlideScore introduce las contribuciones de Coulomb y de van der Waals entre el receptor y el ligando, así como introduce términos de solvatación con el fin de poder hacer *docking* explícito de las moléculas de agua en el sitio activo¹⁵⁷.
- GoldScore: La función de GoldScore es una función empírica basada en *force fields* de mecánica molecular que incluye términos de contribución de enlaces por puente de hidrógenos, entre la proteína y el ligando, e intramoleculares del propio ligando, así como también fuerzas de van der Waals intramoleculares del ligando y del ligando con el receptor¹⁶³.
- Jain: La función *Jain* es una función de *scoring* empírica que incluye principalmente, términos de efectos de entropía y solvatación del ligando, además de términos de interacciones hidrofóbicas y de polaridad, tanto atractiva como repulsiva¹⁶⁸. Los términos de lipofilia y de interacciones polares son representados por la suma ponderada de una función de Gauss y sigmoidea. La función incluye penalizaciones para contactos muy próximos entre la proteína y el ligando.

- LigScore: Las funciones *LigScore1* y *LigScore2* son funciones empíricas desarrolladas mediante funciones genéticas de aproximación y que incluyen interacciones de van der Waals, interacciones polares entre el ligando y la proteína y penalizaciones en la desolvatación, atribuidas a la interacción de átomos polares del ligando con la proteína y viceversa¹⁶⁹.
- London dG: Es una función de *scoring* implantada en MOE basada en *Force field* que incluye términos de entropía y de solvatación, además de las contribuciones debidas a la flexibilidad del ligando, a los enlaces de hidrógeno ideales, incluyendo una corrección en base a su “imperfección”, y a la contribución metal-ligando.
- PLP (*Piecewise Linear Potential*): Es una función de *scoring* empírica de *LigandFit* que incorpora términos estéricos y de interacciones de puentes de hidrógeno, combinado con un término intramolecular¹⁷⁰.
- PMF (Potentials of Mean Force): La función de *scoring* PMF es una función *Knowledge-based* que incorpora las interacciones proteína ligando mediante potenciales entre átomos del ligando y del receptor, de forma que los términos de solvatación y de entropía son tratados implícitamente¹⁷¹.
- TotalScore Surflex: utiliza la función de *scoring* empírica Hammerhead. La función incluye contribuciones hidrofóbicas, polares, de repulsión entre átomos polares, de entropía del ligando por pérdida rotacional o traslacional al ubicar el ligando en el sitio activo y de solvatación. Además, la función incluye un *Crash Score* que penaliza los efectos estéricos entre la proteína y el ligando así como entre átomos del mismo ligando¹⁴⁹.

Existen programas de *docking* que no siempre obtienen la mejor resolución de la predicción de afinidad con la función de *scoring* y que por tanto, reevalúan dicha precisión mediante un *rescoring*, de manera que las poses resultantes del *docking* son recalculadas mediante una o varias funciones alternativas de *scoring*. De hecho, el *rescoring* estrictamente es un postprocesado que tiene por objeto mejorar la predicción de la afinidad. Dentro del *rescoring*, podemos considerar el *Consensus Scoring* como una forma de mejorar los errores, corregir las tendencias y el ruido estadístico de las funciones de *scoring*. En cribados virtuales, en lugar de utilizar una función simple de *scoring*, se pueden combinar diversas funciones de *scoring* con el fin de clasificar los potenciales candidatos farmacológicos. La premisa del *consensus scoring* es que mientras más funciones de *scoring* acuerden que un compuesto es activo, más fiable debe ser la predicción¹⁵².

Además de las funciones *consensus*, otra técnica híbrida que mejora los resultados de las funciones de *scoring* es el método llamado *clustering and entropy-based*¹⁷² que incorpora efectos entrópicos al dividir las diferentes poses en clústeres. La contribución entrópica de cada clúster es medida mediante el espacio configuracional que las poses cubren, o bien, mediante el número de poses

existentes en cada clúster. Este método es aplicado por ejemplo, al realizar *dockings* ciegos de un ligando en una proteína.

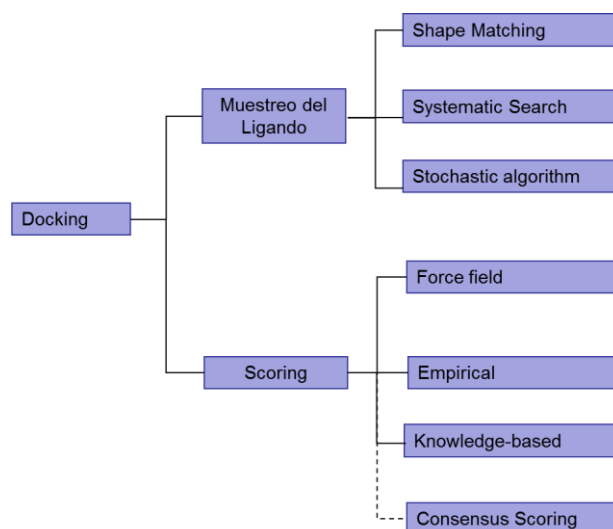


Fig. 1.2.2. Clasificación general de métodos utilizados en *docking* molecular y funciones de *scoring*.

Así pues, existe un número importante de funciones de *scoring* y cada una de ellas realiza diferentes aproximaciones al problema de realizar el ranking de cada pose que devuelve el *docking* según las afinidades estimadas. Entre los problemas más importantes a los que enfrenta la función de *scoring* hallamos, el problema de que las estructuras proteicas rígidas pueden tener orientados incorrectamente los hidroxilos y que residuos como las histidinas y lisinas pueden presentar diferentes estados de protonación en función del pH. Igualmente, los efectos entrópicos son considerados de forma muy aproximada o negligidos, al igual que el potencial molecular electrostático de la proteína y el momento dipolar del ligando. Normalmente, los ligandos de gran tamaño presentan afinidades menores que las esperadas al sumar las interacciones favorables y en muchos casos las funciones de *scoring* toleran ligandos demasiado grandes. La desolvatación de los ligandos no suele ser considerada correctamente sobreestimando en muchos casos la afinidad de compuestos polares^{173,174}.

1.1.3 Docking Proteína-Proteína: En la mayoría de procesos biológicos se producen interacciones proteína-proteína, este hecho indica la importancia del conocimiento estructural de estos complejos a la hora de entender cómo trabajan las unidades biomoleculares y como cumplen sus funciones. De hecho, el número esperado de complejos biomoleculares excede el número de proteínas que conforman el proteoma, en un valor como mínimo de un orden de magnitud¹⁷⁵. Estos complejos presentan en muchos casos una dificultad extrema al predecir su estructura mediante métodos experimentales como cristalografía de rayos X o NMR, de ahí la importancia de utilizar aproximaciones computacionales, como el *docking* proteína-proteína, en el proceso de predecir la

estructura tridimensional de un complejo. En este sentido las herramientas computacionales como el *docking* proteína-proteína, que predice los modos de unión y la energía libre entre estructuras proteicas individuales, son necesarias para complementar los métodos experimentales, pues ayudan a identificar las interacciones proteína-proteína y la determinación estructural del complejo. Los métodos de *docking* proteína-proteína han avanzado notablemente en las últimas dos décadas. Sin embargo, la predicción de estructuras proteína-proteína es un problema difícil, con retos importantes que solventar. Cabe destacar el papel fundamental que juega la comunidad CAPRI (<http://capri.ebi.ac.uk>) al monitorizar y evaluar la precisión de los métodos de *docking*¹⁷⁶.

De forma similar al *docking* proteína-ligando en moléculas pequeñas, el *docking* proteína-proteína consta de dos importantes procesos: el muestreo de conformaciones y el *scoring*. En el proceso de muestreo se realiza la búsqueda de posibles orientaciones y/o conformaciones entre las dos macromoléculas. Esta búsqueda básicamente se divide en dos etapas, el muestreo de las orientaciones manteniendo la proteína rígida (*rigid-body*) y el muestreo conformacional de las moléculas, en donde una vez hallada la orientación manteniendo la proteína rígida, se considera la flexibilidad de las proteínas para generar las conformaciones de unión.

La función de *scoring* mide la afinidad entre las dos moléculas unidas, permitiendo efectuar el ranking de las conformaciones halladas en el muestreo. La naturaleza de las funciones de *scoring* también es similar a las de *docking* proteína-ligando y se agrupan en funciones basadas en: *Force field*, *knowledge-based* y empíricas (*empirical*) o bien por combinación entre las mismas¹⁷⁷.

Las estrategias de búsqueda de orientaciones proteína-proteína se dividen en tres categorías: de búsqueda exhaustiva global (*exhaustive global search*), de búsqueda local de formas complementarias (*local shape feature matching*) y de búsqueda al azar (*randomized search*).

A su vez esta clasificación de búsquedas se subdivide dependiendo del algoritmo de cálculo utilizado¹⁷⁸.

- Algoritmos *Exhaustive global search*. En la búsqueda exhaustiva se explora traslacionalmente y rotacionalmente el espacio 3D de forma independiente. Una proteína se mantiene fija y la proteína móvil rota alrededor de la proteína fija explorando completamente el espacio rotacional 3D. El proceso se repite para cada punto traslacional definido de la proteína fija. Forman parte de este grupo los algoritmos: *FFT correlation*, *Spherical Fourier transform based search* y *Direct search in Cartesian space*.
- Algoritmos *Local shape feature matching*. Las proteínas son representadas en formas moleculares como las superficies Connolly y a su vez, los algoritmos buscan la complementariedad de las formas tridimensionales entre las dos proteínas. Pertenecen a este grupo los algoritmos *Distance geometry algorithm*, *Geometric hashing* y *Genetic algorithm*.

- Algoritmos *Randomized search*: Mantienen una proteína fija mientras que la proteína móvil se ubica al azar alrededor del sitio activo, en búsquedas locales, o bien alrededor de toda la proteína móvil en el caso de búsquedas globales. Posteriormente las orientaciones son optimizadas y refinadas mediante algoritmos estocásticos. Forman parte de este grupo los algoritmos *Monte Carlo search*, *Particle swarm optimization* y *Genetic algorithm*.

Existen otras aproximaciones llamadas de *post-docking* las cuales estrictamente hablando, no son algoritmos de búsqueda, pero que permiten refinar las búsquedas realizadas por otros protocolos de búsqueda proteína-proteína. Generalmente estos protocolos toman las mejores conformaciones proteína-proteína calculadas por el algoritmo de búsqueda y las optimizan, volviendo a realizar su ranking mediante funciones de *scoring* más sofisticadas¹⁷⁹ que pueden incorporar flexibilidad explícita de la proteína e incluso información biológica.

Los métodos de *docking* proteína-proteína utilizados en el presente trabajo son:

- HADDOCK: Es un protocolo de *docking* semi-flexible que utiliza algoritmo *Randomized search*, especialmente diseñado para cuando se dispone de información bioquímica o biofísica de ensayos experimentales de mutagénesis, o de NMR sobre interacciones protéicas existentes¹⁸⁰.
- FlexPepDock: El protocolo de alta resolución FlexPepDock de Rosetta está especialmente concebido para el *docking* de péptidos puesto que realiza *docking* flexible tanto del esqueleto, como de las cadenas laterales. Utiliza el algoritmo de búsqueda *Randomized search* de Monte Carlo con minimización¹⁸¹.
- ZDock: Es una función de *docking* rígido, que realiza búsqueda rotacional y traslacional del tipo *exhaustive global search* y que utiliza el algoritmo de cálculo FFT para acelerar la búsqueda traslacional.

1.3 Modelos farmacofóricos.

Los modelos farmacofóricos son utilizados habitualmente en cribados virtuales, diseño *de novo* y también en otras aplicaciones como la optimización de candidatos (*lead optimization*) y en el diseño de fármacos multidiana, ADME/Tox, quimiogenómica¹⁸², etc. Principalmente su mayor aplicación ha sido desarrollada ante la inexistencia o desconocimiento de la estructura terapéutica involucrada en la disfunción biológica. Los modelos farmacofóricos se clasifican en *ligand-based* o bien *structure-based* en función de si las interacciones de partida utilizadas para su construcción, son extraídas de un ligando utilizado como modelo, o bien establecidas con la estructura diana.

Los farmacóforos basados en el ligando se generan partiendo de las interacciones más esenciales entre uno o varios ligandos representativos y la diana farmacológica. Para cada ligando debe explorarse su flexibilidad, la cual queda recogida en el conjunto de conformaciones que posteriormente son utilizadas para alinear los diferentes ligandos entre si y poder determinar así las características físico-químicas comunes que definirán el farmacóforo. La flexibilidad del ligando puede ser explorada mediante los métodos llamados pre-enumeración (*pre-enumerating*), en donde las múltiples conformaciones generadas para cada molécula son pre-calculadas y guardadas en una base de datos, o bien, por el método *on-the fly*, cuyas conformaciones son generadas durante el proceso de modelaje del farmacóforo¹⁸³. Cabe indicar que los métodos de pre-enumeración devuelven mejores resultados que los métodos *on-the fly*¹⁸⁴. De hecho, un buen generador de conformaciones debe satisfacer tres condiciones: generar eficientemente las conformaciones de unión, que una molécula de pequeño tamaño adopta al unirse a una diana macromolecular. Además debe reducir el número de conformaciones de baja energía con el fin de evitar una generación muy numerosa de conformaciones y por último evitar tiempos computacionales elevados.

En referencia a la segunda etapa del modelado de farmacóforos, como es el alineamiento molecular de los ligandos, cabe distinguir dos categorías de métodos, los métodos basados en puntos (*point-based*) y los métodos basados en propiedades (*property-based*). Los algoritmos *point-based* diferencian los puntos generados en base a si son átomos, fragmentos, o bien características farmacofóricas y estos son superpuestos generalmente mediante un ajuste de mínimos cuadrados. La limitación más importante en los métodos *point-based* es la necesidad de predefinir puntos de anclaje, que en el caso de ligandos no similares puede ser problemático.

Los algoritmos *property-based* generan los alineamientos utilizando descriptores moleculares, generalmente representados mediante funciones Gaussianas y la optimización del alineamiento es llevada a cabo a través de la medida de la similitud del solapamiento intermolecular¹⁸³.

Los farmacóforos *structure-based*, como su nombre indica, tienen en cuenta la estructura proteica del sitio activo de la diana farmacológica. Estos protocolos analizan la complementariedad de las características químicas del sitio activo en el espacio tridimensional. Los farmacóforos *structure-based* se pueden construir basándose en el complejo macromolécula-ligando o bien basándose únicamente en la macromolécula¹⁸³.

Los farmacóforos *structure-based* son la opción preferida, pero presentan el inconveniente de que requieren del conocimiento de la estructura proteica 3D y por tanto no pueden ser aplicados en casos en los cuales no se conoce dicha estructura (Fig. 1.3.1.).

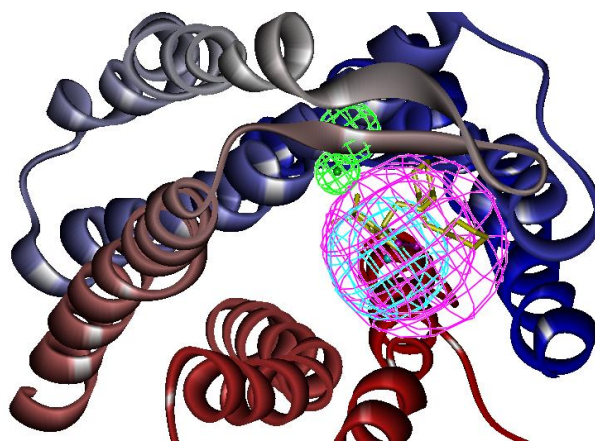


Fig. 1.3.1. Farmacóforo *structure-based* en el cual las características químicas (*features*) son generadas a partir de puntos del sitio activo que pueden presentar interacciones.

Las características químicas (Fig. 1.3.1.) más habituales en compuestos farmacológicos¹⁸⁵ y que se utilizan para modelar un farmacóforo son del tipo: características de átomo, de enlace, de aceptor de hidrógeno, de dador de hidrógeno, de centroide, de hidrofobicidad, de hidrofobicidad aromática, hidrofobicidad alifática, ionizable negativo, ionizable positivo, de carga positiva, de carga negativa, de anillo aromático, volumen de exclusión, restricción de distancia, restricción de ángulo, restricción de ángulo diedro, característica de forma tridimensional, de interacción con metal entre otras.

En el presente estudio los protocolos de modelaje de farmacóforos utilizados son:

- LigandScout: Genera farmacóforos *structure-based* basándose en el complejo proteína-ligando, e incorpora una completa definición de características químicas, las cuales se basan en el algoritmo de alineamiento basado en puntos (*point-based*) del tipo “*pattern matching*”¹⁸⁶.
- MOE: Genera farmacóforos *ligand-based* y *structure-based*. Utiliza varios métodos de pre-enumeración para desarrollar las conformaciones del ligando, como son las dinámicas moleculares, los métodos estocásticos o bien búsqueda sistemática, mientras que como algoritmo de alineamiento de ligandos utiliza métodos *property-based*¹⁸³.

1.4 Descriptores moleculares.

Los descriptores moleculares se definen como el resultado final de un procedimiento lógico y matemático, el cual transforma la información química codificada en una representación simbólica de una molécula, en un número o un resultado útil de un experimento estandarizado¹⁸⁷. Las propiedades moleculares comúnmente utilizadas en las ecuaciones QSAR son conocidas como descriptores moleculares. Las ecuaciones QSAR permiten predecir algunas propiedades moleculares, tales como la actividad de una molécula, mediante correlaciones realizadas a partir de los descriptores moleculares. Igualmente, los descriptores moleculares son utilizados para filtrar u ordenar quimiotecas de

compuestos moleculares en base a una serie de propiedades fisicoquímicas que caracterizan su similitud.

Existen un gran número de descriptores, lo cuales se clasifican¹¹ como:

- Descriptores constitucionales: Estos descriptores indican características simples de la molécula, como puede ser su peso molecular, el número de heteroátomos que la molécula contiene, o bien número de átomos de flúor, o de cloro, número de anillos aromáticos, dobles enlaces, número de enlaces que pueden rotar, número de centros quirales, etc. Estos descriptores son utilizados tanto en ecuaciones QSAR, como en filtrado de bases de datos, cuando los valores a predecir dependen del tamaño de la molécula.
- Descriptores topológicos¹⁸⁸: son índices que dan información sobre el orden de los enlaces y la conectividad intramolecular, dando información sobre cómo es de rígida o flexible una molécula. Ejemplos de estos descriptores son el índice de Weiner, que describe variaciones en las propiedades físico-químicas de los alcanos, el índice de conectividad, que diferencia los enlaces dentro de la misma molécula, el índice de Randic, que describe las ramificaciones en moléculas orgánicas, el índice de Hosoya, que agrupa tipología de enlaces disjuntos, el índice de Harary, los índices de Kier, índice de Balaban, el índice de Zagreb o el índice de Hall entre otros.
- Descriptores electrostáticos¹⁸⁹: los índices electrostáticos caracterizan las distribuciones de carga moleculares tales como índices de polaridad, multipolos y polarizabilidad. Uno de los descriptores electrostáticos más utilizados es el área superficial polar topográfica (TPSA) que indica la fracción de superficie molecular compuesta por grupos polares frente a grupos apolares. Otro descriptor electrostático muy utilizado es el coeficiente de partición octanol-agua (ClogP) que indica la hidrofilia/lipofilia de una estructura. Otros descriptores electrostáticos existentes son los descriptores de aceptores de puentes de hidrógeno y dadores de puentes de hidrógeno, el parámetro de polaridad submolecular, el índice de carga total, cargas parciales positivas y negativas relativas, o cargas totales positivas o negativas.
- Descriptores geométricos¹⁹⁰: Caracterizan el tamaño de la molécula y su superficie topográfica y por tanto requieren las coordenadas 3D de los átomos que conforman la molécula. Ejemplos de estos descriptores son los momentos de inercia, volumen molecular, área superficial molecular, el área superficial total accesible al solvente (TSASA), así como las diferentes áreas superficiales de van der Waals.
- Descriptores cuánticos¹¹: Los descriptores cuánticos dan información sobre la estructura electrónica de la molécula. Ejemplos de estos descriptores son: la energía HOMO y LUMO, la refractividad, la energía total, el potencial de ionización, la afinidad electrónica, la energía de

protonación, la energía de solvatación, el componente de van der Waals de la energía potencial, etc.

Existen también descriptores combinados o híbridos que tienen en cuenta simultáneamente descriptores electrostáticos y geométricos o bien topográficos como el índice topográfico-electrónico T^E que conecta parámetros de polaridad con topografía molecular expresada en distancias¹⁹⁰.

Las áreas de aplicación donde más se han aplicado los descriptores y las técnicas QSAR en el campo de las interacciones químico-biológicas son: en las reglas de Lipinski¹², en el campo ADME, en la predicción de toxicidades¹⁹¹ y en la farmacocinética, donde el rol que juega la lipofilia de las moléculas es un factor determinante.

En el presente estudio se han utilizado diversos descriptores, básicamente para filtrar quimiotecas según las reglas de Lipinski y para validar los *test-set* utilizados en los cribados virtuales retrospectivos, garantizando que las propiedades físico-químicas de las estructuras que lo forman son similares y que por tanto el cribado virtual realizado, no obedece a diferencias en dichas propiedades/descriptores de sus integrantes.

1.5 Búsqueda por similitud (*Similarity search*).

En la gestión de bases de datos de estructuras químicas, el uso de la similitud molecular es una de las herramientas más utilizadas. Típicamente ante la existencia de una molécula con interés biológico, a menudo se intentan identificar otras moléculas similares, por ejemplo extraídas de una base de datos, que al ser ensayadas experimentalmente proporcionen similares o mejores resultados que los obtenidos para la molécula original. Este tipo de filtrado de quimiotecas ha permitido que los métodos de búsqueda por similitud se hayan considerado como métodos de cribado virtual de rápido cálculo.

Las ventajas más destacables de los métodos de búsqueda por similitud son²⁸ básicamente tres:

- Necesitan poca información para realizar la exploración y no requieren de ninguna hipótesis sobre que parte de la molécula es la responsable de la actividad. Son por lo tanto métodos aplicables en etapas iniciales del descubrimiento de fármacos, cuando a menudo no se dispone de mucha información sobre la diana molecular.
- Los costes computacionales que demandan la mayoría de métodos de similitud son muy bajos, lo que permite realizar de forma regular, búsquedas en quimiotecas de gran tamaño.
- La premisa de que las moléculas con estructuras similares presentan actividad biológica similar es generalmente válida¹⁹².

Por contra, la efectividad de los métodos de búsqueda por similitud puede variar de forma importante de una actividad biológica a otra diferente, por lo que debe considerarse como un método preliminar que es apropiado cuando sólo se dispone de una molécula bioactiva conocida.

El fundamento de la búsqueda por similitud se basa en el uso de descriptores. De este modo la molécula target se caracteriza por unos o varios descriptores. Estos descriptores son comparados con el conjunto de descriptores calculado para cada una de las moléculas que conforman la quimioteca. La comparativa permite el cálculo de la medida de similitud entre la estructura target y los elementos de la quimioteca, así como la correspondiente ordenación de estos elementos en función de dicha similitud. En la lista ordenada obtenida, los compuestos iniciales del ranking son los de mayor interés.

La representación de varios descriptores moleculares en forma de cadenas de Bits es uno de los métodos más comunes de búsqueda por similitud y se conocen como *molecular fingerprints*. Los *fingerprints* en combinación con otras métricas, como es el índice de Tanimoto¹⁹³, son utilizados para cuantificar el grado de similitud entre la estructura target y las estructuras de la base de datos (Fig. 1.5.1).

En la práctica existen *fingerprints* 2D y 3D. En los *fingerprints* 2D o también llamados binarios, en cada posición del bit se contabiliza la presencia (mediante un 1) o ausencia (mediante un 0) de la característica que mide el descriptor para cada molécula. Como ejemplos de *fingerprints* 2D tenemos las cadenas Molecular Access System (MACCS) que tiene variantes de 166 bits, de 164, de 320 y de 960 bits, o los fingerprints UNITY o SLN de 988 o 992 bits, los fingerprints Barnard Chemical Information (BCI) que contienen 1052 bits, o los fingerprints PubChem de 881 bits^{11,194}.

Mediante estos *fingerprints* es posible calcular coeficientes como el de Tanimoto, el cual contabiliza y calcula el número de bits igual a 1 entre dos moléculas (Fig. 1.5.1).

$$\text{Indice de Tanimoto } Tc = \frac{n_{i,j}}{n_i + n_j - n_{i,j}}$$

$$\text{Indice de Dice } Dc = \frac{2 n_{i,j}}{n_i + n_j}$$

$$\text{Indice de Cosine } Cc = \frac{n_{i,j}}{(n_i \cdot n_j)^{1/2}}$$

$$\text{Distancia de Hamming } HD_{ij} = n_i + n_j - 2n_{i,j}$$

$$\text{Distancia Euclidiana } ED_{ij} = (n_i + n_j - 2n_{i,j})^{1/2}$$

$$\text{Distancia promedio } D = \frac{\sum_{i=1}^n \sum_{j=1}^n D_{ij}}{n(n-1)}$$

Fig. 1.5.1. Métricas de similitud y funciones de distancia utilizadas para comparar las cadenas de bits (*fingerprints*). n_i y n_j son el número de bits fijados para las moléculas i y j . $n_{i,j}$ es el número de bits en común para las dos moléculas. D_{ij} es la distancia entre las moléculas i y j . Mientras que D es la distancia promedio y n el número total de moléculas.

Existen otros *fingerprints* que no son binarios, por ejemplo HQSAR (*Holographic Quantitative Structure Activity Relationship*) que utiliza *fingerprints* no binarios para correlacionar estructura con actividad, en lugar de para calcular similitudes.

Los *fingerprints* 3D dependen de la forma 3D que la estructura química adopta según su conformación. Estos descriptores se pueden considerar como un método intermedio entre los *fingerprints* 2D y los modelos farmacofóricos. Los *fingerprints* 3D se clasifican como *fingerprints* rígidos o flexibles, según si miden las distancias en una única conformación, o bien si miden todas las distancias posibles, en base a los incrementos de rotación de los enlaces rotables existentes¹⁹⁵.

En el presente estudio se ha utilizado principalmente el índice de Tanimoto y los *fingerprints* 2D *Molecular Access System* (MACCS).

1.6 Diseño de librerías de compuestos químicos (quimiotecas)

El diseño de librerías de compuestos químicos se basa en la creación de extensas listas de estructuras químicas, generadas de forma combinatoria. A menudo, en fases iniciales del diseño de fármacos se exploran diversas quimiotecas virtuales, con el fin de explorar amplios espectros de compuestos químicos, sin embargo, actualmente el uso de quimiotecas virtuales se ha focalizado principalmente como estrategia de búsqueda de *hits* frente una, o preferiblemente, varias dianas biológicas¹⁹⁶. Es por ello, que el diseño óptimo de quimiotecas debe tener en consideración tanto la diversidad química, como la diversidad de posibles dianas biológicas, además de cumplir propiedades farmacológicas. La diversidad química normalmente es investigada con profundidad, mientras que la diversidad de dianas que puede cubrir una quimioteca es un campo que aún requiere desarrollo.

El proceso de diseño de librerías virtuales de compuestos debe iniciarse mediante recopilación de información, referente a los miembros de la familia de dianas a la que investigar. Dicha información puede obtenerse gracias al conocimiento de las estructuras tridimensionales de la familia diana, o bien, mediante información basada en ligandos activos conocidos, como pueden ser sus descriptores moleculares, sus características farmacofóricas, el conocimiento de subestructuras activas (*fragment-based*), o también mediante el conocimiento de propiedades bioquímicas, que presentan diferencias para un mismo ligando, al ser ensayados en diferentes familias de dianas.

Habitualmente los métodos basados en el ligando han tenido mayor repercusión gracias a su amplia aplicación y su eficacia computacional. En casos como la superfamilia de receptores GPCR en donde el

número de estructuras proteicas determinadas experimentalmente es pequeño y muy recientes, los métodos basados en el ligando (*ligand-based*) han sido los más frecuentes. En cambio, los métodos basados en la estructura (*structure-based*) han sido utilizados en casos como el de las quinasas, en donde el número de estructuras determinadas es considerable y por tanto la generación de farmacóforos que recogen las características de sus sitios activos, es factible.

La diversidad de dianas es uno de los aspectos más importantes que una quimioteca debe recoger, tanto por su capacidad de cubrir diferentes espacios químicos, como por abarcar diferentes dianas farmacológicas. La medida de la diversidad puede realizarse mediante el grado de cobertura y la medida de la tendencia (*bias*) que la quimioteca obtiene tanto para las dianas biológicas, como de la propia quimioteca. Esta medida se realiza mediante la distribución del número de dianas que cada molécula de la quimioteca cubre (Fig. 1.6.1.).

La cobertura de dianas (T_{cov}) indica la proporción de dianas (*targets*) con posesión de al menos un ligando y la cobertura química (C_{cov}) indica la proporción de ligandos que cubren al menos una diana. La medida de la tendencia (*bias*) refleja la desviación de la uniformidad en la distribución de ligandos para cada diana (T_{bias}) y de forma equivalente, muestra la desviación de la distribución de dianas para cada ligando (C_{bias})¹⁹⁷ calculada mediante los descriptores de entropía de Shannon (SHED)¹⁹⁸ (Fig. 1.6.1.).

La medida de la promiscuidad de un ligando frente a varias dianas terapéuticas es otro aspecto importante a valorar¹⁹⁹ en quimiotecas. El grado de promiscuidad de los compuestos que conforman una quimioteca puede medirse mediante el parámetro *Prom10*. Este parámetro otorga valores entre 0 y 1 según si todas las moléculas de la quimioteca son selectivas (valor=0), o bien si tienden a ser activas con al menos 10 dianas (valor=1).

En base a estos indicadores, la selectividad de una quimioteca se puede evaluar¹⁹⁷ a partir de la fórmula [1.4]:

$$Score = \frac{\{T_{cov} + C_{cov} + (1 - T_{bias}) + (1 - C_{bias}) + Prom10\}}{5} \quad [1.4]$$

El diseño óptimo de quimiotecas, que pretende cubrir toda una familia de proteínas, debe presentar máximas coberturas de las dianas así como máxima cobertura química de sus compuestos y al mismo tiempo debe presentar desviaciones mínimas de la distribución de dianas y de ligandos.

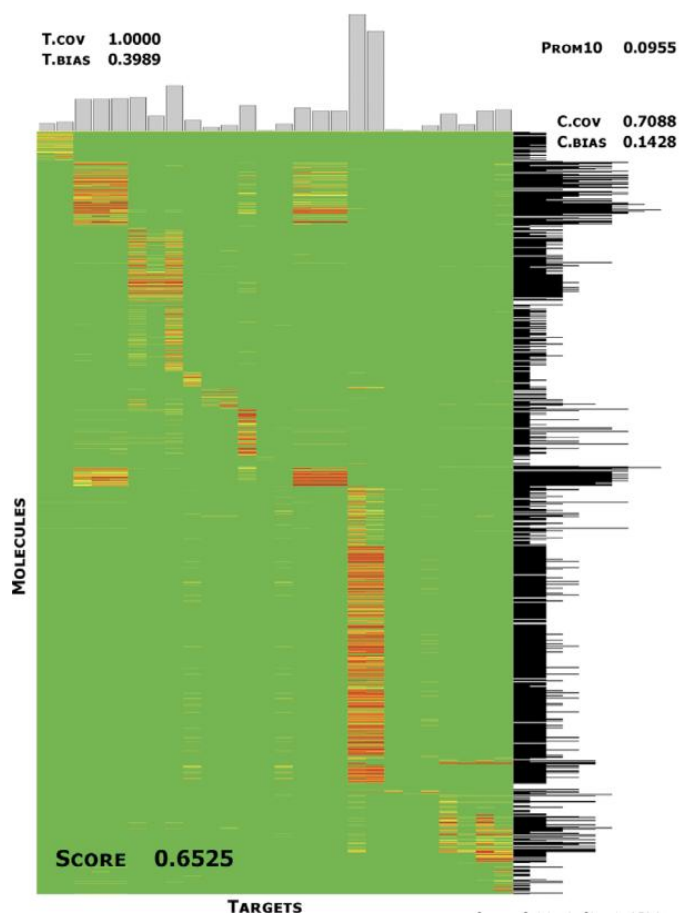


Fig. 1.6.1. Evaluación de la cobertura y medida de la tendencia de una quimioteca. El gráfico representa las dianas terapéuticas (*Targets*) en el eje de abscisas y las moléculas que conforman la quimioteca en el eje de ordenadas. Escala de color de rojo a verde que indica la distancia Euclídea de los descriptores SHED, entre compuestos de la quimioteca (filas) con la molécula bioactiva escogida por diana (columnas). El color rojo indica distancias igual a 0 y el verde indica distancias cercanas al umbral máximo definido. La distribución de cobertura de dianas (T_{cov}) y su tendencia (T_{bias}) se muestra para cada diana en la parte superior del gráfico y para cada molécula de la quimioteca a la derecha del gráfico así como su (C_{cov}) y (C_{bias}). Representación obtenida de Gregori-Puigjané *et al*¹⁹⁷.

Otra aproximación a la diversidad química en el uso de quimiotecas es a través de la diversidad de estructuras químicas *scaffolds*, que caracteriza a un grupo de moléculas con núcleos estructurales comunes. La diversidad de *scaffolds* se basa en el principio que, *scaffolds* que enriquecen listas de compuestos activos (*scaffolds* activos) en cribados virtuales, permiten guiar los procesos de cribados de compuestos y además, facilitan la síntesis combinatoria de activos²⁰⁰.

La clave a la hora de construir librerías de compuestos, se basa en poder utilizar *scaffolds* privilegiados que permitan realizar un diseño inteligente de las mismas. Ello requiere tener en cuenta el conocimiento de la actividad del *scaffolds*, las propiedades farmacológicas y los tests de cribados efectivos²⁰¹.

En el presente estudio se ha creado una quimioteca combinatoria mediante el módulo CombiGlide²⁰² de Schrödinger partiendo de *scaffolds* activos, accesibles de acuerdo con las estrategias sintéticas que

viene utilizando el laboratorio GEM IQS durante años, en su línea de investigación de inhibidores de VIH.

1.7 Dinámica molecular

La dinámica molecular (DM) es un método de simulación del movimiento de partículas que permite obtener una perspectiva de la dinámica del sistema, motivo por el cual, es ampliamente usado para biomoléculas. Una DM ideal, reproduciría en el espacio y el tiempo los movimientos internos y trayectorias moleculares que se suceden, desencadenados por los procesos bioquímicos. La resolución de las DM en el espacio se mide en angstroms (Å) y la resolución en el tiempo en femtosegundos.

Habitualmente en sistemas macromoleculares, los métodos basados en simulaciones son utilizados en tres tipologías de aplicaciones²⁰³:

- Como método de muestreo del espacio conformacional, con el objetivo de refinar estructuras.
- Como método de descripción de un sistema en equilibrio, incluyendo propiedades estructurales y trayectorias.
- Para examinar las dinámicas actuales, muestreando el espacio conformacional y representando el desarrollo del sistema en el tiempo.

En los 2 primeros casos tanto los métodos de Monte Carlo de mecánica estadística, como las simulaciones por DM pueden ser utilizados, en cambio en el tercer caso donde las trayectorias y su evolución en el tiempo son primordiales, sólo las DM pueden aportar la información requerida.

Las DM utilizan aproximaciones basadas en la física Newtoniana, para simular los movimientos atómicos. En el caso de sistemas moleculares complejos, como es el caso de biomoléculas, la descripción mecano cuántica de dichos movimientos y las reacciones que se llevan a cabo, es muy compleja y queda fuera del alcance de cualquier supercomputador. Es gracias a las aproximaciones como las DM como se pueden superar estas limitaciones, estimando las fuerzas que actúan en cada uno de los átomos del sistema. Básicamente, tal y como se ha descrito al inicio de este capítulo, las principales fuerzas que intervienen sobre un átomo, provienen de otros átomos enlazados y átomos no enlazados próximos (Fig. 1.1.1.). El desglose de estas energías proviene de los enlaces químicos que vibran, de las interacciones establecidas en los ángulos de enlace y en los ángulos diedros y para los átomos no enlazados, las interacciones existentes pueden ser de van der Waals y electrostáticas [1.2]. La parametrización de estos términos de energía, la recogen los *Force Fields*, que como su nombre indica, describen las contribuciones de las fuerzas de los átomos, que gobiernan las DM. Los *Force Fields* más utilizados en DM son AMBER¹⁴⁴, CHARMM¹⁴⁵ y GROMOS¹⁴⁶, éstos actúan calculando todas las fuerzas que afectan a cada uno de los átomos del sistema, de manera que una vez calculadas, los

átomos se mueven a posiciones conforme las leyes de Newton. Este proceso de cálculo se repite típicamente millones de veces durante una DM.

Una de las principales ventajas que las DM ofrecen en el diseño de fármacos, es precisamente la simulación *in silico* del movimiento de las proteínas. Los modelos cristalográficos existentes determinados experimentalmente por cristalografía de rayos X o bien por RMN, son modelos estáticos y la mayoría de técnicas de descubrimiento de fármacos no contemplan su movilidad. Principalmente en proteínas de membrana, como en los GPCRs, dada su elevada flexibilidad, las DM pueden simular los efectos inducidos que se producen cuando los ligandos se unen a las proteínas, o bien identificar sitios activos no triviales de detectar, como son los llamados sitios activos crípticos²⁰⁴, o los sitios activos alostéricos y por tanto sus consecuentes conformaciones proteína-ligando generadas²⁰⁵.

Las DM también son usadas para describir el proceso en que un ligando entra en el sitio activo, así como para predecir los modos de unión del ligando en la proteína y su selectividad frente a varias dianas. Precisamente una de las ventajas que las MD ofrecen es la precisión en el cálculo de la energía de unión o energía libre. Métodos computacionales rápidos, como el *docking*, ofrecen en la mayoría de casos, predicciones pobres de energías de unión. De forma general en DM, se define la energía libre de unión, como la diferencia de energía entre el ligando libre en solución y la energía del ligando formando complejo con la proteína.

Las simulaciones en proteínas de membrana son habituales en estudios computacionales de biología molecular. La estabilidad de estas proteínas viene gobernada por membranas plasmáticas de origen heterogéneo, hidrofóbico e hidrofílico cuya modelización es un requisito en el estudio de la estructura y la función de las proteínas de membrana. Las bicapas lipídicas son una de las mejores representaciones de las membranas plasmáticas²⁰⁶ y conjuntamente con las proteínas de membrana han sido utilizadas a modo de complejo explícitamente²⁰⁷ en DM, ofreciendo aproximaciones muy detalladas, pero con costes computacionales muy altos debido a los millones de átomos que intervienen en la simulación. Para reducir estos tiempos computacionales, se han desarrollado de forma alternativa modelos implícitos²⁰⁸ que tratan la influencia del solvente y de la membrana lipídica de forma ponderada, definiendo potenciales electrostáticos continuos para describir la energías de solvatación electrostáticas del soluto y los potenciales electrostáticos²⁰⁹.

La energía de solvatación en membranas implícitas se expresa generalmente como la suma de contribuciones electrostáticas (ΔG_{elec}) y no polares (ΔG_{np}) [1.5]:

$$\Delta G_{solv} = \Delta G_{elec} + \Delta G_{np} \quad [1.5]$$

En donde el término no polar depende del área superficial del soluto (S) y del coeficiente de la tensión superficial (γ) [1.6]:

$$\Delta G_{np} = \gamma S \quad [1.6]$$

El potencial electrostático puede resolverse mediante diversas aproximaciones, entre estas, el modelo generalizado de Born, que derivada de la ecuación de Born aplicada a iones [1.7]:

$$\Delta G_{elec} = -\frac{1}{2} \tau \sum_{\alpha\beta} \frac{q_{\alpha}q_{\beta}}{\sqrt{r^2 + R_{\alpha}^{GB}R_{\beta}^{GB} \exp(-r_{\alpha\beta}^2/4R_{\alpha}^{GB}R_{\beta}^{GB})}} \quad [1.7]$$

Donde R_{α}^{GB} es el radio efectivo de Born de los átomos α y β ; y τ es la constante dieléctrica en el interior del soluto²⁰⁹.

Si una simulación por el contrario se realiza mediante solvatación explícita de la proteína, es necesario definir unas condiciones de contorno periódicas, con el fin de reducir el espacio ocupado por el solvente y también el coste computacional requerido. Así, una molécula de agua que durante la simulación tenga una trayectoria fuera del volumen delimitado, reentrará en el sistema por el lado opuesto.

Cuando las partículas se mueven durante la DM, las fuerzas que actúan sobre ellas deben ser reevaluadas. Con el fin de acelerar las simulaciones, habitualmente se definen distancias de corte (*distance cutoff*) que se aplican al realizar el cálculo de las interacciones, de manera que interacciones con distancias superiores a la distancia de corte, no son calculadas. En simulaciones con solventes implícitos se fijan valores *cutoff* mayores que el tamaño del sistema definido, mientras que en simulaciones con solventes explícitos se utilizan métodos como Particle-Mesh Ewald²¹⁰ (PME) para la energía electrostática.

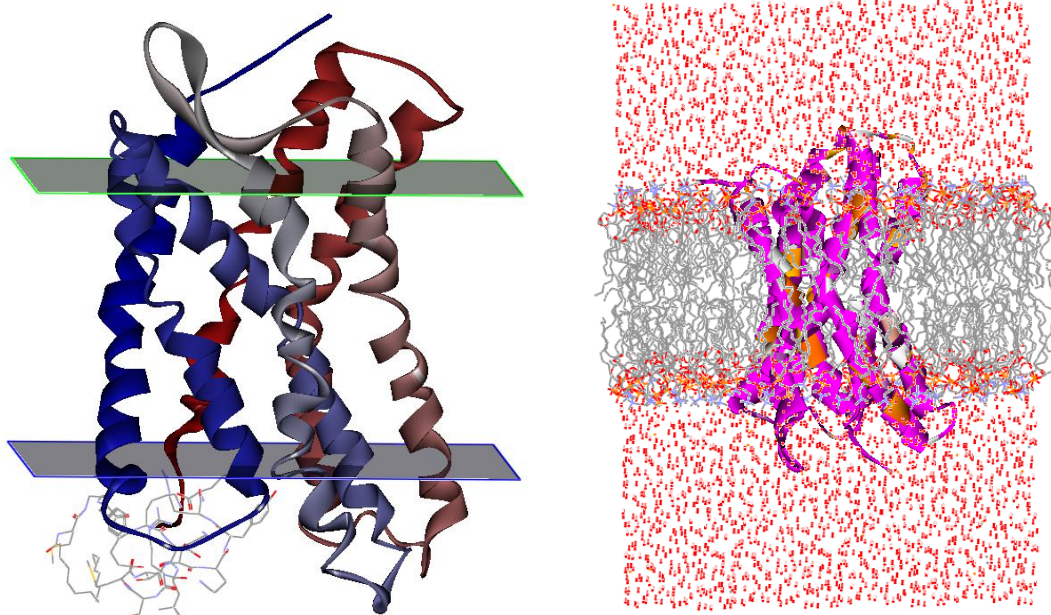


Fig. 1.7.1. Representaciones de la proteína de membrana CXCR4 en modelo de solvente implícito (izquierda) y en modelo explícito con membrana lipídica (derecha).

Al iniciar la DM, generalmente se parte de una estructura estática determinada experimentalmente, lo que requiere una etapa de minimización, con el fin de encontrar un mínimo energético local y se equilibra el sistema, con el fin de optimizar la estructura de partida. Posteriormente se añade energía

al sistema hasta alcanzar la temperatura deseada y se vuelve a equilibrar el sistema. En la etapa de calentamiento se restringe el sistema, para evitar la deformación de la estructura proteica.

Los principales retos de las DM son el desarrollo de *Force Fields* más precisos, así como el elevado coste computacional que requieren estas simulaciones, que en algunos casos precisan de incrementos de tiempo de 1 microsegundo, o incluso a poder ser de milisegundos²¹¹, con el fin de obtener muestreos conformacionales del espacio adecuados. Además, las DM asumen que el sistema está en equilibrio, por lo que tiempos de equilibrio insuficientes, previos a la etapa de producción en la simulación, pueden conllevar limitaciones en la exactitud de los resultados.

Una de las carencias que presentan las simulaciones por DM, son las predicciones de sistemas que involucran átomos metálicos, debido a la existencia de efectos cuánticos importantes causados por la formación y rotura de enlaces químicos²¹². Para solventar estas carencias se han propuesto *Force Fields* que incluyen cálculos mecánico-cuánticos como por ejemplo, en simulaciones que intervienen sitios activos de enzimas o metaloenzimas²¹³. Igualmente, la polarización causada por un flujo de electrones es también un efecto mecano cuántico generalmente ignorado por las DM y que algunos *Force Fields* intentan incorporar²¹⁴.

En el presente estudio se han realizado simulaciones por DM utilizando CHARMM, para refinar la búsqueda de sitios activos alostéricos e identificar las principales interacciones proteína-péptido de carácter alostérico que se establecen.

2. DISCUSION DE RESULTADOS

A continuación se presentan los resultados obtenidos para cada uno de los objetivos planteados en el presente estudio. Los 5 objetivos planteados han dado lugar a 4 artículos que se exponen al final del presente trabajo dentro del apartado “Artículos” y también a un quinto artículo, resultante de una comunicación oral en una conferencia internacional que se adjunta en el apartado “Apéndices”.

2.1 Estudio y propuesta de un método de cribado virtual que mejore la recuperación de activos frente a inactivos para “dianas difíciles”.

Los cribados virtuales retrospectivos, se basan en la compilación y cribado virtual de un *test-set* de moléculas activas conocidas, conjuntamente con moléculas supuestamente inactivas o *decoys*. Este conjunto de moléculas debe tener características estructurales y fisicoquímicas parecidas, con el propósito de que su discriminación virtual no sea trivial.

Existen numerosas técnicas de cribado virtual¹⁵³, que a su vez se dividen en múltiples métodos, como son las múltiples funciones de *docking*, con sus múltiples funciones de *scoring*, etc. Dependiendo de las características de la diana farmacológica y de las características del método seleccionado, el resultado del método de cribado virtual puede dar resultados más o menos selectivos al clasificar los compuestos activos frente a los *decoys*. Uno de los puntos conflictivos a la hora de realizar un cribado virtual retrospectivo se encuentra en el cribado virtual de “dianas difíciles”. Se definen “dianas difíciles” como aquellas que poseen un amplio sitio activo, con la posibilidad de acomodar una determinada familia o varias familias de compuestos (*scaffolds* significativamente diferentes) de diversas maneras (diferentes modos de unión), como los “*multi-conformer binders*” y “*multisite binders*”¹⁷. Dichas dianas pueden unir diversas familias de compuestos con estructura significativamente diferente, por lo que es difícil que el cribado sea selectivo a la hora de reconocer las diversas estructuras como activas frente a los *decoys* y el rendimiento del cribado suele ser notablemente bajo. Contrariamente, en el cribado de “dianas fáciles”, dichas dianas se unen a una familia de compuestos activos de estructura y conformación determinada, por lo que todas las moléculas que contengan ese tipo de estructura (tipo de interacciones) podrían considerarse como supuestamente activas a diferencia de los *decoys* que no mostrarían esta estructura química ni interacciones. En esta tesis se ha desarrollado un protocolo de tres pasos (*docking*, postfiltrado farmacofórico y posterior clasificación por similitud molecular) para mejorar el rendimiento del cribado de “dianas difíciles”. El protocolo se ha aplicado al cribado retrospectivo del receptor VEGFR-2, diana con un amplio sitio activo, que es capaz de acomodar “*multi-conformer binders*”²¹⁵. Previamente

aproximaciones de cribado virtual aplicadas a inhibidores de VEGFR-2 muestran que es difícil alcanzar un alto rendimiento en el cribado de compuestos para esta diana²¹⁶⁻²¹⁸. Esta diana farmacológica está involucrada en la enfermedad del cáncer a través del receptor CXCR4 puesto que éste induce la sobreexpresión de VEGFR que produce angiogénesis tumoral²¹⁹, siendo por tanto de interés para el laboratorio del Grup d'Enginyeria Molecular IQS (GEM).

En el artículo anexoado: “Improving VEGFR-2 Docking-Based Screening by Pharmacophore Postfiltering and Similarity Search Postprocessing” (Artículo 1) se realiza el cribado virtual del *test-set* publicado en DUD database⁷ de la proteína quinasa VEGFR-2, que actúa como receptor involucrado en enfermedades como el cáncer²²⁰, según el protocolo de tres pasos mencionado. Inicialmente, para verificar que el *test-set* utilizado cumple el requisito de poseer propiedades fisicoquímicas similares entre activos y *decoys*, se evalúa la similitud de los compuestos en: peso molecular, número de enlaces rotables, número de aceptores y dadores de hidrógenos, así como el coeficiente de partición y el área superficial polar de cada molécula (Fig. 2.1.1.).

	Activos	Decoys
Peso Mol.	238.3/692.6 (86.5)	325.4/474.2 (21.6)
Num. Enlaces rot.	0.000/0.294 (0.064)	0.057/0.286 (0.037)
H. aceptores	1/7 (1.3)	0/7 (0.9)
H. dadores	0/5 (1.2)	0/4 (0.7)
Coef. Partición	0.25/6.79 (1.17)	1.33/5.74 (0.72)
Area polar superficial	34.2/374.2 (54.2)	19.1/288.2 (37.6)

Fig. 2.1.1. Propiedades moleculares del *test-set* de VEGFR-2 publicado en la base de datos DUD. Los valores indicados corresponden al valor máximo y mínimo y entre paréntesis al valor promedio.

Como herramientas de cribado virtual, se inicia el estudio utilizando las funciones de *scoring* de MOE¹⁶⁵ (*ASE*, *Affinity dG*, *Alpha HB* y *London dG*) con las correspondientes funciones de *placement* (*Alpha PMI*, *Alpha Triangle*, *Proxi Triangle* y *Triangle Matcher*) .

Para evaluar dichas funciones de *scoring* y escoger la más adecuada para la diana VEGFR-2 se realiza una validación de dichas funciones mediante *Self-Docking* y *Cross-Docking*. El *Self-Docking* consiste en realizar el *docking* de los ligandos co-cristalizados, en la propia estructura en la cual se hallan formando complejo, midiendo los valores del RMSD para las 10 mejores poses clasificadas según la función de *scoring* evaluada, también se tiene en cuenta el número de poses con valores de RMSD²²¹ inferiores a los 2 Å. El *Cross-Docking* en cambio realiza el *docking* de los ligandos co-cristalizados no en la estructura donde se halla formando complejo, sino que su *docking* se realiza en el resto de estructuras en donde existe otro ligando co-cristalizado, pero ocupando el mismo sitio activo. Para realizar el *Cross-Docking* las estructuras proteicas (Fig. 2.1.2.) deben ser superpuestas con el fin de poder medir el RMSD entre las 10 mejores poses clasificadas por la función de *scoring* evaluada, respecto a la conformación determinada experimentalmente.

PDB	Resolución (Å)
1Y6A	2.10
1YWN	1.71
2OH4	2.05
2P2H	1.95
2QU5	2.95
2RL5	2.65
3BE2	1.75
3C7Q	2.10
3CJF	2.15
3EWH	1.60
3CJG	2.25

Fig. 2.1.2. Resolución y características de las estructuras VEGFR-2 utilizadas para hacer el análisis *Self-Docking* y *Cross-Docking*.

Del análisis de *Self-Docking* y *Cross-Docking* se obtiene que la función de *placement Alpha Triangle* conjuntamente con la función de *scoring Alpha HB* devuelven los mejores resultados tanto de *Self-Docking* (aRMSD=1.3 Å y 10 poses con RMSD<2 Å) como de *Cross-Docking* (aRMSD=3.0 Å).

Debido a la existencia de varias estructuras cristalizadas de VEGFR-2 en RCBS Protein Data Bank²²², que a su vez contienen diferentes ligandos co-cristalizados, para realizar la selección de la estructura de trabajo, se analizan las resoluciones cristalográficas conjuntamente con los resultados de *Self-Docking* y *Cross-Docking*. Del *Self-Docking* se evalúa la pose con mejor RMSD, la media, el promedio y la desviación estándar de las 10 poses con mejores valores de RMSD. Del *Cross-Docking* se analiza para cada ligando y cada PDB el mejor valor de RMSD y el promedio de las 10 mejores poses que el *Cross-Docking* devuelve. El análisis se realiza utilizando la mejores funciones halladas de *placement, Alpha Triangle* y de *scoring, Alpha HB*. Como resultado se obtiene que la estructura 1YWN es una de las estructuras con mejor resolución cristalográfica (1.71 Å), con mejores valores de *Self-Docking* (RMSD de la mejor pose: 0.8 Å y aRMSD=1.4 Å y una desviación estándar en el RMSD de las 10 mejores poses de 0.3 Å) y buenos valores de *Cross-Docking* (la media de aRMSD para los 10 *dockings* correspondientes a los 10 ligandos co-cristalizados es de 5.3 Å), aunque los valores de *Cross-Dockings* son inferiores a los de las estructuras 3BE2 (media aRMSD=5.1 Å) y 3EWH (media aRMSD=4.3 Å). Estas estructuras 3BE2 y 3EWH tienen similares resoluciones cristalográficas: 1.75 Å y 1.6 Å respectivamente (Fig. 2.1.2.), pero peores valores de *Self-Docking* (RMSD de la mejor pose: 1.2 Å y 1.8 Å y aRMSD=2.6 Å y 2.9 Å con unas desviaciones estándar en el RMSD de las 10 mejores poses de 0.9 Å y 2.9 Å, respectivamente).

Una vez seleccionada las funciones de *docking* y la estructura tridimensional de trabajo, se procede a realizar el cribado virtual retrospectivo de la quimioteca DUD de activos y *decoys* de VEGFR-2. El *docking* de la quimioteca preparada, al efectuar su ranking en base a la función de *scoring Alpha HB*, devuelve una curva ROC^{223,224} bastante pobre, por debajo de la clasificación aleatoria (AUC=0.33), con

unos resultados de pronta recuperación de activos²²⁵ al 1%, 5% y 10% de *decoys* escrutados muy bajos (0.00%; 2.56% y 6.41% respectivamente) (Fig. 2.1.4.).

Si el mismo *docking* en lugar de ser clasificado por la función de scoring, es clasificado utilizando el índice de Tanimoto calculado mediante *fingerprints* MACCS 2D, tomando como modelo el ligando co-cristalizado en 1YWN (derivado 2-anilino-5-2-aril-oxazol), la curva ROC mejora y discrimina los activos de los *decoys* de forma más selectiva que una clasificación *random* (AUC=0.70), con unos resultados de pronta recuperación de activos al 1%, 5% y 10% de *decoys* escrutados, de 14.10%; 24.36% y 32.05% respectivamente (Fig. 2.1.4.).

Como procedimiento de mejora del cribado virtual, se propone realizar posterior al *docking*, un cribado farmacofórico de las poses devueltas por el *docking*. Para construir el farmacóforo, se toman todas las estructuras cristalizadas de VEGFR-2 con resolución inferior a 3 Å (Fig. 2.1.2.) que contienen un ligando co-cristalizado y se alinean y superponen todas las estructuras proteicas, de manera que los ligandos quedan superpuestos en base su posición original respecto al sitio activo de la proteína. Mediante el módulo *Pharmacophore Consensus* de MOE¹⁶⁵ se realiza la búsqueda de características farmacofóricas que definen las interacciones comunes entre todos los ligandos superpuestos y los aminoácidos del sitio activo de VEGFR-2. La búsqueda de características farmacofóricas proteína-ligando da como resultado la detección de 5 interacciones clave, cuyos volúmenes de interacción se optimizan de forma manual incrementando consecutivamente cada uno de dichos volúmenes, con el fin de maximizar el número de activos cribados frente a *decoys*. De las 5 características farmacofóricas, 2 de ellas (interacciones aromáticas o de anillo π) se muestran críticas a la hora de optimizar su volumen de interacción (Fig 2.1.3.)

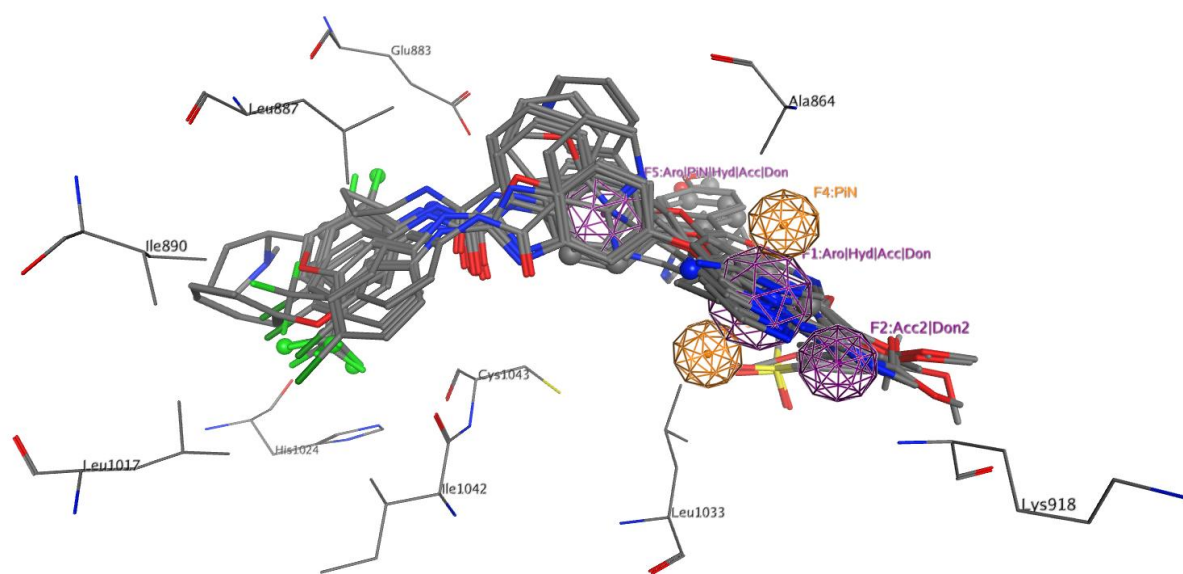


Fig. 2.1.3. Detalle y definición de las 5 características farmacofóricas que definen el farmacóforo generado a partir de interacciones proteína-ligando halladas entre la proteína VEGFR-2 y los ligandos co-cristalizados de VEGFR-2 superpuestos. La superposición de ligandos se realiza alineando y superponiendo los PDB de VEGFR-2 quedando los respectivos ligandos co-cristalizados superpuestos en base a su posición en el sitio activo.

El farmacóforo más selectivo retiene el 59% de los activos y el 20.8% de los decoys existentes presentando unos valores de pronta recuperación de activos al 1%, 5% y 10% de *decoys* escrutados de 10.26%; 24.36% y 34.62% respectivamente (Fig. 2.1.4.).

El uso del filtro farmacofórico, constata una mejora en la recuperación de activos frente a *decoys*, clasificando igualmente las poses según la energía que la función de *scoring* calcula. Así pues, el siguiente paso a realizar es clasificar las poses filtradas por el farmacóforo utilizando *fingerprints* de similitud MACCS 2D, como es el índice de Tanimoto. Para calcularlo se toma como molécula de referencia el ligando co-cristalizado en 1YWN. La clasificación por índice de Tanimoto mejora la pronta recuperación de activos frente a *decoys* presentando unos valores al 1%, 5% y 10% de *decoys* escrutados de: 16.67 %; 30.77% y 41.03% respectivamente (Fig. 2.1.4.).

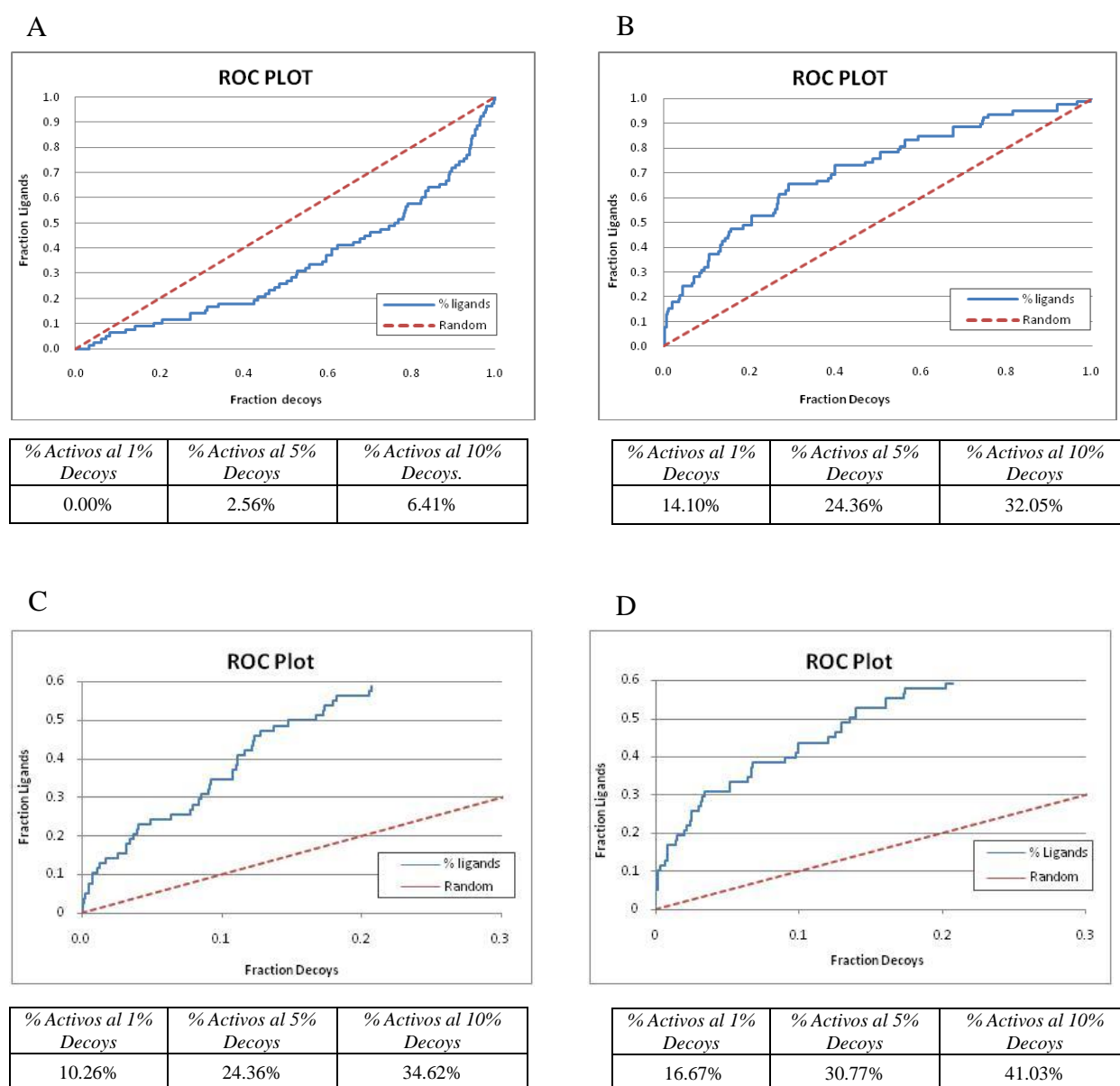


Fig. 2.1.4. Evolución del cribado virtual retrospectivo de la quimioteca DUD de VEGFR-2. Se muestran las curvas ROC y los valores de pronta recuperación de activos, al 1%, al 5% y al 10% de decoys cribados, al realizar: **A.** Sólo *docking* con la mejor función determinada por *seldocking* y *crossdocking*. **B.** *Docking* con ordenación posterior de las poses por similitud 2D mediante índice de Tanimoto **C.** *Docking* con posterior cribado farmacofórico **D.** *Docking* con postcribado farmacofórico y

posterior ordenación por similitud 2D mediante índice de Tanimoto.

Los resultados muestran que aplicando este protocolo de tres pasos, la discriminación de activos frente a *decoys* mejora significativamente, principalmente en la pronta recuperación de activos (% Activos al 1% *Decoys*=16.7%), en comparación a los resultados iniciales obtenidos por *docking* con clasificación según la función de *scoring* (% Activos al 1% *Decoys*=0%). Igualmente, es importante señalar, que el método propuesto es más adecuado a la hora de encontrar correctos modos de unión, que utilizando tan sólo la función de *scoring* proveniente del *docking*. El método proporciona poses que superan el farmacóforo construido en base a las interacciones proteína-ligando de diferentes conformaciones de ligandos activos para VEGFR-2. Por tanto, este método garantiza correctos modos de unión de los ligandos ensayados, en el sitio activo de VEGFR-2 y puede pasar a ser aplicado al cribado virtual de otras 'dianas difíciles', como es el caso de CXCR4, estudiada en esta tesis.

2.2 Evaluación funcional de la estructura proteica de CXCR4 mediante cribado virtual retrospectivo.

La determinación experimental de la estructura del receptor CXCR4 llevada a cabo por primera vez, a finales de 2010 por parte de Wu *et al.*⁵⁵ ha supuesto un avance de vital importancia tanto para el conocimiento real de la estructura tridimensional del receptor, así como de cara a las posibilidades que se abren de hallar nuevos inhibidores de dicha proteína.

Existen un total de cinco estructuras cristalográficas de CXCR4 publicadas hasta la fecha en el *Protein Data Bank*⁴, determinadas todas ellas mediante difracción de rayos X. Las cinco estructuras tienen diferentes resoluciones cristalográficas y a su vez contienen dos ligandos diferentes co-cristalizados. El ligando IT1t (Fig. B.2.1), que aparece unido al receptor en cuatro de las cinco estructuras, IT1t es una molécula de pequeño tamaño, de peso molecular 406.7 g/mol y es un derivado de la isotiourea²²⁶. En cambio el ligando co-cristalizado CVX15²²⁷ (Fig. B.2.1) es un péptido cíclico formado por 16 residuos proteicos y cuyo peso molecular es de 2138.6 g/mol.

Igualmente las diferentes estructuras publicadas se presentan en forma monomérica, homodimérica y homotrimérica (Fig. 2.2.1.)

<i>PDB</i>	<i>Resolución (Å)</i>	<i>Ligando</i>	<i>Estructura Receptor</i>
3ODU	2.5	IT1t	Dimérica
3OE0	2.9	CVX15	Monomérica
3OE6	3.2	IT1t	Monomérica
3OE8	3.1	IT1t	Trimérica
3OE9	3.1	IT1t	Dimérica

Fig. 2.2.1. Resolución y características de las estructuras publicadas por Wu *et al.*

Hasta la fecha en que se publicaron estas estructuras experimentales, el receptor CXCR4 había sido modelado por homología²²⁸. El modelado de proteínas por homología se realiza partiendo de la secuencia de aminoácidos de la proteína cuya estructura tridimensional se pretende modelar (*target*), así como de una o varias proteína cuya estructura tridimensional es conocida y que además posee una secuencia proteica similar a la proteína *target* (*template*). Partiendo de la alineación de secuencias proteicas, así como de la similitud existente entre ellas, se puede construir un modelo tridimensional, llamado modelo por homología.

Así pues el receptor CXCR4 ha sido modelado partiendo de otras proteínas GPCR homólogas, principalmente de la rodopsina bovina^{118,229,121} o el β_2 adrenérgico^{230,119,231,232}.

De este modo con este estudio se pretende:

-Hallar las diferencias y semejanzas estructurales existentes entre las estructuras publicadas en el *Protein Data Bank* y el modelo de CXCR4 previamente generado por homología en el laboratorio del Grup d'Enginyeria Molecular IQS (GEM)¹¹⁸.

-Estudio funcional de la estructura proteica de CXCR4 a partir de las cinco estructuras cristalográficas de CXCR4 publicadas hasta la fecha y del modelo previamente generado por homología en el GEM. Análisis mediante cribado virtual retrospectivo basado en la estructura de la proteína (*docking*) del modelo cristalográfico que mejor distingue activos de inactivos. Análisis de la conformación de CXCR4 que mejor propicia la unión de diferentes familias de inhibidores antagonistas al receptor. Uno de los efectos a tener en cuenta es que por ejemplo el PDB 3OEO, contiene como ligando unido al receptor, un péptido susceptible de generar cambios conformacionales inducidos sobre la estructura del receptor y por tanto modificar el resultado del cribado virtual.

-Evaluación y determinación de los mejores parámetros de *docking* para el cribado virtual de antagonistas de CXCR4, con el fin de poder realizar nuevos cribados prospectivos de moléculas candidatas a inhibir el receptor CXCR4.

2.2.1 Análisis de las diferencias y semejanzas estructurales existentes entre los PDB publicados y el modelo de CXCR4 previamente generado por homología en el GEM: La evaluación genérica de las diferencias existentes entre las distintas estructuras del receptor CXCR4, incluyendo la estructura creada por homología, se realizó mediante la medición del *RMSD*²³³ de los átomos voluminosos o pesados de dichas estructuras, una vez alineadas sus secuencias proteicas.

El *RMSD* permite expresar la similitud estructural y se calcula a partir de las distancias (*d*) existentes entre pares de átomos equivalentes (*n*), de las 2 estructuras a comparar:

$$RMSD = \sqrt{\frac{\sum_i d_i^2}{n}}$$

Siendo el valor de *RMSD*=0 para dos estructuras idénticas, el valor de *RMSD* depende²²¹ del tamaño molecular. Así en la comparación de 2 moléculas pequeñas, pueden obtenerse fácilmente valores bajos de *RMSD*, mientras que en un caso como el que nos ocupa, donde se comparan 2 estructuras proteicas de gran tamaño, los valores de *RMSD* pueden ser elevados. Independientemente de la magnitud del valor, en este caso vamos a utilizar el valor del *RMSD* para comparar estructuras iguales, por lo que los valores obtenidos van a servir para fines comparativos, puesto que todas las estructuras

diméricas y triméricas se prepararon para comparar uno sólo de sus monómeros. Por este motivo se eliminaron las cadenas proteicas duplicadas de un mismo PDB, moléculas de agua y ligandos co-cristalizados, dejando únicamente 1 cadena proteica por estructura. Para realizar una comparación estricta de la estructura proteica de CXCR4 se eliminó previamente la proteína de fusión TL4 que se halla cristalizada conjuntamente a CXCR4 en todos los PDB publicados. La Lisozima T4 es una proteína de fusión utilizada debido a su elevada facilidad de cristalización^{234,235}. Por este motivo es ampliamente usada en la cristalización de proteínas de membrana GPCR en donde se ha utilizado con éxito en receptores como el $\beta 2$ adrenérgico ($\beta 2AR$)^{235,57} la adenosina A2A⁵⁸ además de CXCR4⁵⁵.

Para el cálculo del *RMSD*, se tuvieron en cuenta todos los α -carbonos. Se alinearon todas las secuencias de las seis estructuras y se superpusieron las estructuras proteicas (Fig. 2.2.2.) mediante MOE 2009.10¹⁶⁵ obteniéndose la matriz de valores *RMSD* (Fig. 2.2.3.).

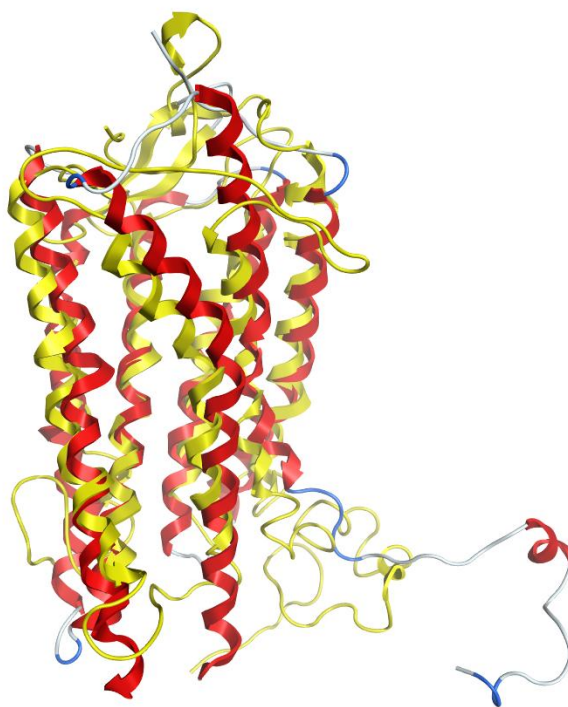


Fig. 2.2.2. Superposición de las estructuras PDB 3ODU (rojo) y del modelo de homología (amarillo).

La comparativa de valores *RMSD* entre estructuras, muestra como la estructura generada por homología (CXCR4congen) presenta valores de *RMSD* más elevados, de 6 Å de *RMSD* respecto al resto de PDBs. Por el contrario, el PDB 3OE6 junto con la estructura 3OE8, son las estructuras que presentan valores más bajos, con unos valores de *RMSD* que oscilan entre 1.26 y 1.47 Å para 3OE6 y entre 1.10 y 1.52 Å para 3OE8, lo cual indicaría que estas estructuras presentan mayor similitud con el resto de estructuras. Es decir, la estructura 3OE6, representaría una conformación *consensus* de las diferentes

conformaciones de CXCR4, independientemente de las variaciones en *RMSD* debida a su resolución cristalográfica.

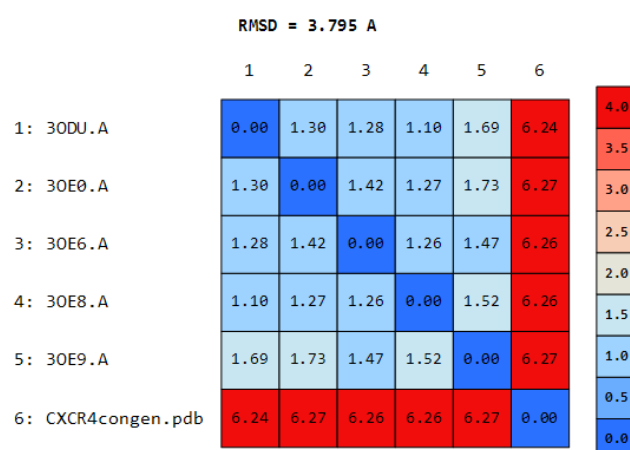


Fig. 2.2.3. Matriz *RMSD*. Valores de *RMSD* en Å, obtenidos al comparar las seis estructuras entre sí, una vez alineadas las secuencias y superpuestas todas las estructuras. Los valores de *RMSD* están codificados por colores, según la escala indicada a la derecha de la matriz.

El cálculo del *RMSD* representa una medida de similitud estructural global, entre 2 o más estructuras. Sin embargo, a nivel farmacológico no aporta información sobre si la conformación de la proteína es activa o inactiva, ni sobre cuál es el dominio o dominios que presentan actividad dentro de la proteína. Para comparar como afectan las diferencias estructurales al sitio de unión (*binding site*) y por tanto como se recoge la actividad de la proteína, debemos basarnos en comparativas que se centren en la actividad de ésta y en el sitio de unión de la misma.

El principal sitio de unión del co-receptor CXCR4, tal y como se ha comentado, está ubicado en el extremo N-terminal, junto al bucle extracelular 2 (ECL2). La superposición del PDB 3OE0, que contiene el ciclopéptido CVX15, juntamente con cualquiera de los PDBs que contiene el antagonista IT1t, permite identificar visualmente una cavidad extracelular de gran tamaño. En dicha cavidad, se alojan ambos ligandos, ocupando diferentes volúmenes aunque parcialmente superpuestos (Fig. 2.2.4). Diversos estudios mutagénicos confirman interacciones entre los residuos proteicos Asp171, el Asp262 y Glu288 con ligandos tales como el biciclamo AMD3100^{236,237,238} los monociclamos AMD3465 y AMD3529²³⁹ o el ligando no ciclamo AMD11070²⁴⁰. Estos tres residuos también interactúan con ligandos péptidicos como el T-140^{241,242} compuesto por 14 residuos proteicos o el ciclopentapéptido antagonista FC131²⁴³.

Al analizar las interacciones que presentan los ligandos co-cristalizados: IT1t (derivado de isotiourea de pequeño tamaño) y CVX15 (péptido de 16 residuos proteicos) en sus respectivas estructuras, se observa que el residuo Glu288 presenta interacción en ambos casos y en todas las estructuras. Sin

embargo los residuos Asp171 y el Asp262 sólo interactúan con el ligando de mayor tamaño CVX15 (Fig. 2.2.4) en la estructura 3OE0.

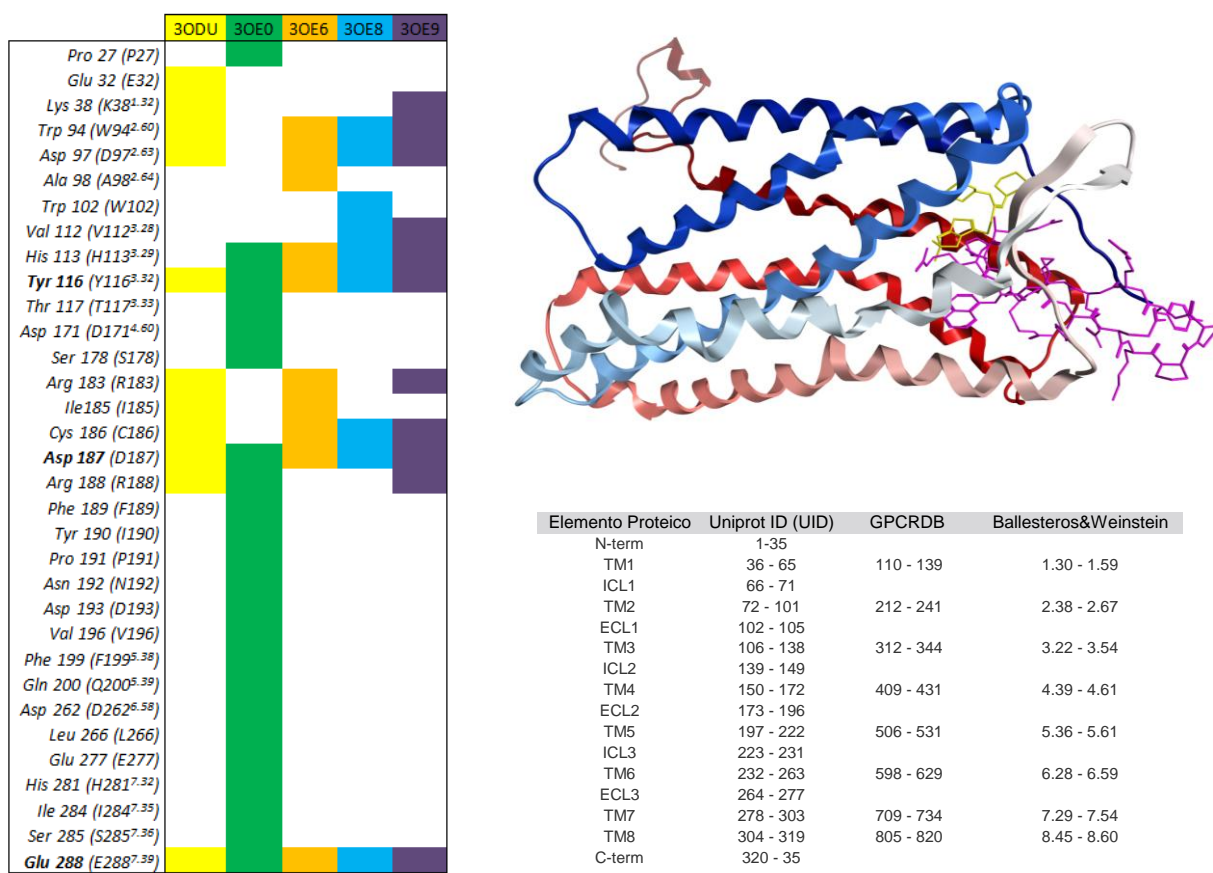


Fig. 2.2.4 A la izquierda se muestra el mapa de interacciones proteína ligando de las estructuras cristalográficas. La estructura cristalina 3OE0 cuyo ligando peptídico (CVX15) es de gran volumen, presenta múltiples y diferentes interacciones en comparación con el resto de estructuras co-cristalizadas con el ligando IT1t de menor tamaño. Los residuos proteicos Tyr116, Asp187 y Glu288 (negrita) presentan interacción de tipo polar y ácida, con los ligandos co-cristalizados en todas las estructuras. Los PDBs 3ODU, 3OE6, 3OE8 y 3OE9 presentan pequeñas variaciones en las interacciones proteína-ligando. En la parte superior derecha se muestra la superposición de los PDBs 3ODU y 3OE6 y sus respectivos ligandos IT1t (estructura en amarillo) y CVX15 (estructura en rosa) al alinear las estructuras. Las notaciones de los residuos proteicos entre paréntesis corresponden a la notación Ballesteros-Weinstein²⁴⁴. En la tabla inferior derecha se muestran las notaciones Ballesteros-Weinstein para los residuos de CXCR4 según su ubicación proteica y su correlación con la numeración Uniprot ID (UID)²⁴⁵

Los residuos proteicos Asp171, Asp262 y Glu288 se consideran aminoácidos clave según los estudios de mutagénesis dirigida^{229,236,237,239,243} y también según el mapa de interacciones proteína ligando. Si se procede a alinear y superponer de nuevo las cinco estructuras cristalizadas, mediante los residuos Trp94, Asp97, Asp171, Asp262 y Glu288 que definen la cavidad extracelular donde se unen los ligandos, se pueden analizar las diferencias estructurales existentes concretamente en dicha cavidad de la proteína que aloja el sitio activo (Fig. 2.2.5).

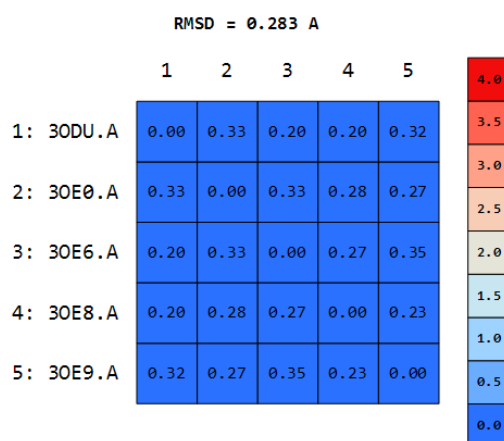


Fig. 2.2.5. Matriz de valores *RMSD* para los C en Å codificada por colores. Los valores son obtenidos al alinear los cinco PDB y superponer los residuos proteicos Trp94, Asp97, Asp171, Asp262 y Glu288 de todas las estructuras proteicas.

La superposición de todos los PDBs a través de los residuos de la cavidad extracelular, indica pequeñas diferencias entre estructuras, a pesar de que ninguno de los casos se sobrepasan los 2 Å de diferencia en las distancias. En el caso la estructura 3OE0, ésta presenta valores ligeramente más elevados de *RMSD* en comparación al resto de estructuras. La posible explicación a estos valores más elevados podría ser debida a efectos inducidos generados por el ligando CVX15, que al ser de mayor tamaño que IT1t, puede inducir mayores cambios conformacionales en la proteína.

2.2.2 Cribado virtual retrospectivo de las estructuras de CXCR4: En el apartado anterior se han analizado las diferencias estructurales entre los PDBs existentes del receptor celular CXCR4 así como el modelo generado por homología¹¹⁸. A continuación, se analizan las diferencias existentes entre las seis estructuras proteicas desde un punto de vista funcional de la proteína. Dicho análisis se encuentra detallado en el artículo anexoado *“Impact of the CXCR4 structure on docking-based virtual screening of HIV entry inhibitors”* (Artículo 2) donde se evaluó para cada estructura su eficacia a la hora de cribar moléculas activas de moléculas inactivas, mediante la técnica de cribado virtual por *docking* y utilizando un total de 14 funciones de *scoring* diferentes. Es decir, se analizaron los diferentes modelos proteicos de CXCR4 mediante *docking* retrospectivo. Utilizando una quimioteca virtual compilada por Pérez-Nueno *et al.* que consta de 248 moléculas antagonistas de CXCR4 conocidas y 4696 moléculas supuestamente inactivas o *decoys*¹⁷³, y que fueron extraídos¹¹⁸ de la Maybridge Screening Collection²⁴⁶. Estos *decoys* presentan propiedades moleculares similares a las de los activos (Fig. 2.2.6). El análisis detallado de la distribución de activos y *decoys*, mediante curvas ROC, para las propiedades moleculares consideradas más importantes, como son las cargas parciales debido a la importancia de las interacciones carga-carga en las proteínas GPCR con los ligandos, se halla detallado en la

Supplementary Information del artículo “Impact of the CXCR4 structure on docking-based virtual screening of HIV entry inhibitors”.

	Ligands	Decoys
Peso Molecular:	727.1/306.5 (74.3)	613.9/452.2 (34.9)
Nº de enlaces simples rotables	18/0 (4.0)	24/0 (2.2)
Nº de Aceptores de puentes de hidrogeno	6/1 (1.2)	13/0 (1.6)
Nº de Donadores de puentes de hidrogeno	5/0 (1.3)	9/0 (1.0)
Atomos Hidrofóbicos	46 /15 (4.6)	43/9 (3.7)
Coef. De particion Octanol-agua	7.6/-2.8 (2.3)	13.9/-2.8 (1.8)
Area Polar Superficial:	485.5/25.8 (60.0)	561.5/0.0 (91.7)

Fig. 2.2.6. Tabla de propiedades moleculares de las 248 moléculas activas y de los 4696 *decoys* que conforman el *test-set*. Los valores tabulados para cada propiedad corresponden al valor máximo y valor mínimo encontrado tanto para los activos como para los *decoys*, así como sus correspondientes desviaciones estándar (valor entre paréntesis).

Los 248 activos que forman el *test-set* o quimioteca de ensayo, cumplen la diversidad química necesaria para garantizar que la selectividad del *docking* no está determinada por una composición homogénea de activos cuyas características similares permiten separar fácilmente los activos de los inactivos. De este modo el *test-set* se compuso mayoritariamente por cuatro familias de antagonistas del receptor CXCR4 y que son estructuralmente diferentes. Las familias químicas que componen los 248 compuestos químicos activos son: Compuestos derivados de AMD3100^{236,247,248} (81 compuestos, 32% del total de activos), tetrahydroquinolinaminas²⁴⁹ (123 compuestos, 49.6% del total de activos), macrociclos²⁵⁰ (19 compuestos, 7.7% del total de activos), KRH^{251,252} (23 compuestos, 9.3% del total de activos) además de 2 compuestos péptidos (0.8% del total de activos) que completan la lista de 248 activos.

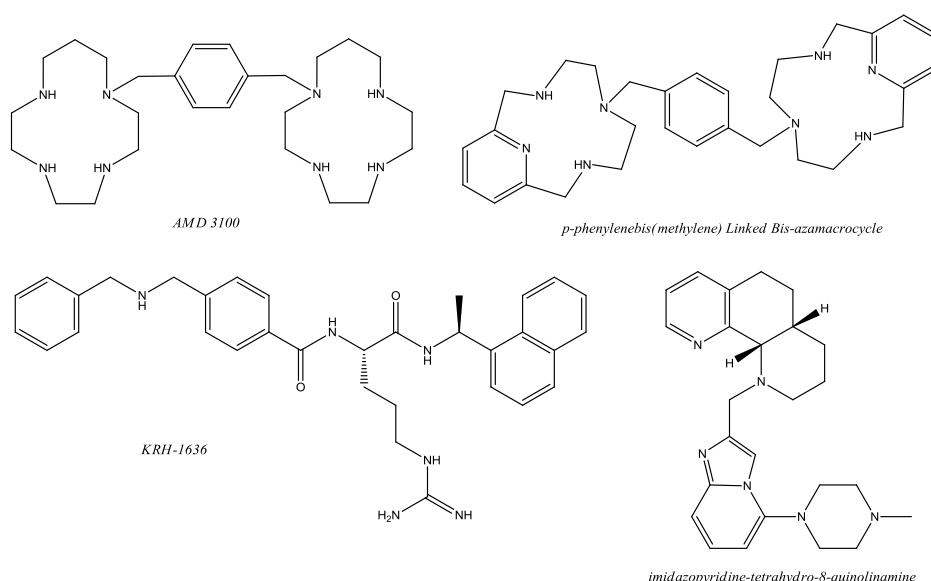


Fig. 2.2.7. Principales compuestos representativos de las familias estructurales químicas que componen el *set* de activos de la quimioteca utilizada para realizar el *docking* retrospectivo de las estructuras proteicas de CXCR4. Arriba a la izquierda el

compuesto AMD 3100 como representante de la familia de derivados de AMD, en la parte superior derecha un derivado macrocíclico, en la parte inferior izquierda un derivado de Kureha (KRH) y a su derecha un derivado de la familia de las tetrahydroquinolinaminas.

De manera similar y con el fin de garantizar la mayor diversidad en el número y en la tipología de las funciones de *scoring* utilizadas en el cálculo del *docking* retrospectivo, se escogieron un total de 12 funciones de *scoring* representativas de los tres tipos de funciones de *scoring*²⁵³:

-Funciones de *scoring* basadas en *Force-field* (force-field based), como la función *GoldScore*¹⁶³ y *DockScore*^{164,254} de LigandFit.

-Funciones de *scoring* empíricas, como las funciones *ChemScore*¹⁶³, *Surflex-TotalScore* de Sybyl X¹⁴⁹, *Glide*¹⁵⁷ de Schrödinger, así como las funciones implementadas en Discovery Studio *LigScore1-Dreiding*¹⁶⁹, *LigScore2-Dreiding*¹⁶⁹, *PLP1*, *PLP2*^{170,255} y *Jain*²⁵⁵.

-Funciones de *scoring knowledge-based* como la función *PMF* o la función híbrida *AutoDock Vina*¹⁶², que es una aproximación basada en la combinación de potenciales *knowledge-based* y términos empíricos.

Previo al *docking* retrospectivo, el *test-set* de activos e inactivos fue convenientemente preparado mediante el software MOE. Para ello, se protonaron todas las moléculas a pH fisiológico, se añadieron las cargas atómicas parciales Gasteiger²⁵⁶ a los enlaces σ y a los sistemas π no conjugados y la estructura geométrica fue minimizada y optimizada mediante el uso del Merck Molecular *Force field* MMFF94¹⁴².

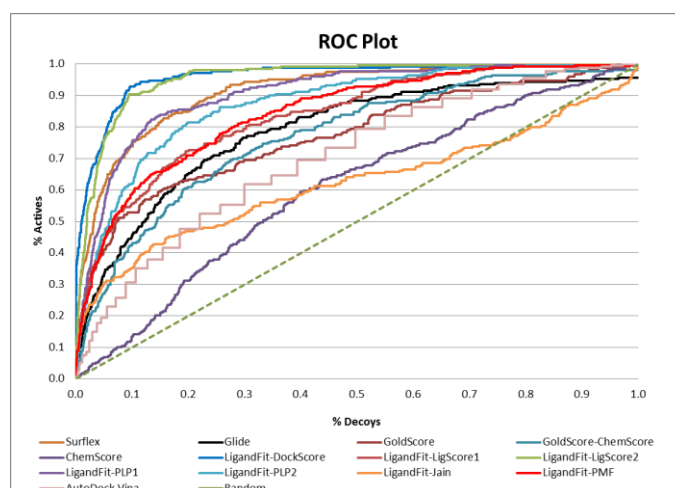
Las estructuras proteicas fueron preparadas mediante el propio módulo existente para tal función en cada software utilizado y que mayoritariamente consiste en: añadir hidrógenos puesto que estos no son determinados en la técnica de cristalización, eliminación de las cadenas homodiméricas y homotriméricas idénticas, escogiendo aquellas cadenas que al alinearse entre ellas presentaban pérdida de residuos proteicos.

Igualmente se eliminaron las moléculas químicas utilizadas como coadyuvantes en el proceso de cristalización, así como las moléculas de agua existentes en los PDBs 3ODU y 3OE0 (contiene 6 molec. de agua), puesto que los PDB 3OE6, 3OE8 y 3OE9 no contienen moléculas de agua co-cristalizadas. La eliminación de las moléculas de agua se verificó (*Supplementary Information* Artículo 2) al comparar los cribados virtuales obtenidos al mantener y eliminar las moléculas de agua co-cristalizadas presentes en el PDB de mejor resolución, 3ODU, al mismo tiempo que se compararon dichos resultados con los resultados obtenidos para el PDB 3OE6, cuyo cribado virtual fue el más efectivo en términos globales. Dicha comparación reveló que para todas las funciones de *scoring* los valores de pronta recuperación de activos (*Early Recovery*) eran peores cuando se mantenían las moléculas de

agua en la estructura proteica a cribar, que cuando las moléculas de agua eran eliminadas. Así mismo, se constató que los resultados para 3ODU eran peores que los obtenidos para la estructura 3OE6.

Como definición de sitio de unión de la proteína donde realizar el *docking* se definió la misma región en donde se encontraban los ligandos co-cristalizados de las estructuras CXCR4. Concretamente y al igual como se indicó en el apartado anterior, basándose en los estudios de mutagénesis dirigida, se definieron los residuos proteicos Asp171, Asp262 y Glu288 como los residuos que delimitan la región, en donde ensayar los *dockings* retrospectivos para las 5 estructuras proteicas y el modelo de homología.

Los parámetros utilizados en el análisis de las diferentes estructuras se hallan detallados en el Artículo 2, así como todos los resultados obtenidos. Estos resultados muestran la estructura monomérica 3OE6 como la mejor para discriminar compuestos activos de inactivos en el cribado virtual retrospectivo y por tanto, como mejor PDB a utilizar en *dockings* prospectivos para la búsqueda de nuevos inhibidores. Al comparar la pronta recuperación de activos (*Early Recovery*) obtenidos, tanto al analizar los factores de enriquecimiento, como el porcentaje de activos al 1%, 5% y 10% de inactivos recuperados, el PDB 3OE6 obtiene los resultados más altos y por tanto más selectivos. Igualmente, el análisis de las curvas ROC obtenidas y de los diferentes parámetros asociados a dicha curva (AUC²²³, aAUC, pAUC²⁵⁷) vuelven a mostrar que la estructura 3OE6 es la que produce resultados más selectivos (Fig.2.2.8.). Desde el punto de vista de predicción de poses la estructura 3OE6 obtiene, después de la estructura 3ODU, los mejores resultados.



3OE6	% Falsos positivos Cribados	% Verdaderos Positivos	pAUC
	1%	49.6%	0.70
LigandFit - DockScore	5%	77.4%	0.80
	10%	93.5%	0.86

Fig. 2.2.8. En la parte superior se muestran las curvas ROC superpuestas para el modelo cristalizado 3OE6. Este modelo, obtiene los mejores resultados en el *docking* retrospectivo. En la tabla inferior se muestran los parámetros asociados a la curva

ROC para la función de *scoring* que obtiene mejores resultados (LigandFit-DockScore) al 1%, 5% y 10% de decoys cribados, obteniendo un AUC global de 0.96.

De este análisis podemos concluir que la estructura proteica 3OE6, a pesar de presentar la peor resolución cristalográfica, parece presentar la mejor conformación proteica a la hora de ubicar compuestos activos en lugar de inactivos. Esta observación se apoya en el análisis realizado en la Fig. 2.2.3. y Fig. 2.2.5. en donde se observa que la estructura 3OE6 es de las estructuras publicadas de CXCR4, la que presenta menores diferencias respecto a la globalidad de estructuras (estructura *consensus*) tanto a nivel del sitio activo, como a nivel como de su estructura global (Fig. 2.2.5.). Este hecho indicaría que la conformación del sitio activo en el PDB 3OE6 se hallaría en una conformación más activada²⁵⁸ para alojar antagonistas de CXCR4.

La estructura 3OE6, por ser la estructura cuyos resultados obtenidos se consideran los más selectivos, fue utilizada también para realizar el análisis de recuperación de familias de activos (*Scaffold retrieval analysis*). Es decir, se ha valorado el porcentaje de compuestos activos, según a la familia química a la que pertenecen, al cribar el 1%; 5% y 10% de la base de datos. Este cálculo se realizó para las 13 funciones de *scoring* y se muestra en la Fig. 2.2.9.

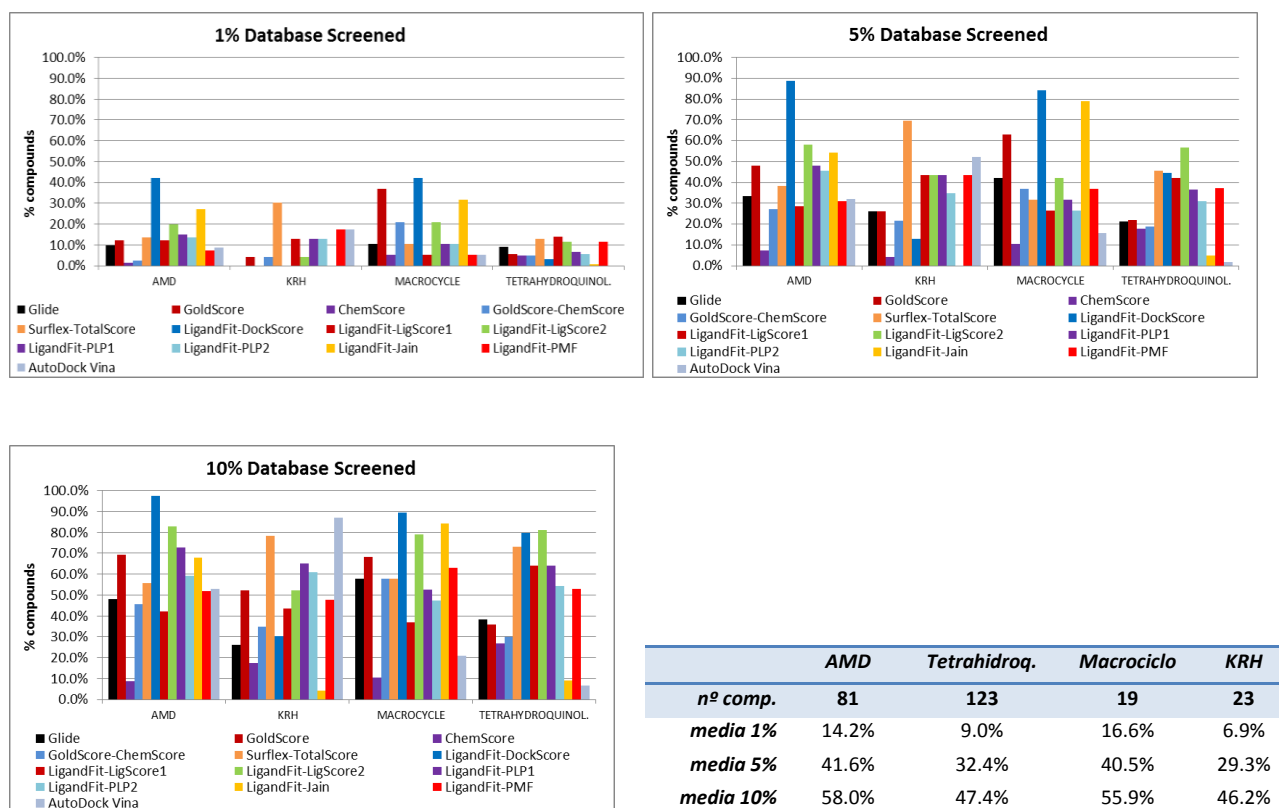


Fig. 2.2.9. Recuperación de familias de activos (*Scaffold Retrieval analysis*) al clasificarlos al 1%, 5% y 10% de la base de datos cribada utilizando la estructura 3OE6 según las 13 funciones de *scoring* al utilizar la estructura 3OE6. En la tabla inferior derecha se muestran los valores promedio, para todas las funciones de *scoring*, del porcentaje de compuestos recuperados de cada familia al cribar al 1%; 5% y 10% de la base de datos.

Los resultados obtenidos de la recuperación de activos por familias, nos permiten afirmar que el *docking* retrospectivo no selecciona sólo compuestos de una determinada familia, sino que recupera moléculas activas indistintamente de la familia a la que pertenecen. Incluso a los valores más bajos de la pronta recuperación (*Early Recovery*) las familias de activos son recuperadas en porcentajes similares entre ellas (Fig. 2.2.9.).

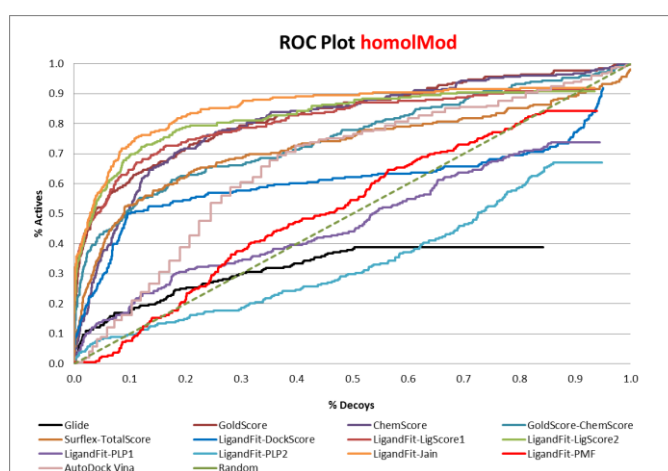
Desde el punto de vista de análisis de las funciones de *scoring* utilizadas, atendiendo a las que mejor clasifican la globalidad de activos frente a inactivos, la función de *scoring Dock_Score* de LigandFit, es la función de *scoring* que devuelve mejores resultados para todas las estructuras CXCR4, tanto a nivel de factores de enriquecimiento (*Early Recovery*), como de valores de curvas ROC: AUC, aAUC y pAUC. Los valores AUC son superiores a 0.9 utilizando *Dock_Score* para todos los modelos, excepto en el caso del modelo por homología.

Al comparar estos resultados con los obtenidos utilizando el modelo de homología se observan ciertas diferencias. Mientras que para las funciones de *scoring* GoldScore, ChemScore, y GoldScore con *rescoring* ChemScore se obtienen altos factores de enriquecimiento, al igual que buenos valores de AUC (0.83; 0.81 y 0.75) y pAUC (0.73; 0.64 y 0.68 al 10% de falsos positivos extraídos), para el resto de funciones de *scoring* los resultados obtenidos tanto en el AUC y pAUC son más discretos (Fig.2.2.10.). Sin embargo, los factores de enriquecimiento (*Early Recovery*) para el modelo por homología, devuelven valores similares a los obtenidos para las estructuras experimentales. Dichos resultados nos conducen a la hipótesis de que ambos, el modelo por homología y el PDB 3OE6 discriminan bien activos de inactivos debido a su capacidad de capturar de una manera más general una conformación característica compatible con el conocido comportamiento alostérico de CXCR4, la cual es capaz de unir mediante *docking* un amplio rango de ligandos activos, estructuralmente diferentes, discriminándolos bien de los inactivos. Estos resultados, a su vez soportan la noción de que si no se tiene información suficiente, como era el caso para la diana CXCR4 años atrás, los modelos por homología pueden ser útiles.

Cabe destacar, que en el *docking* retrospectivo realizado con LigandFit de Discovery Studio para el modelo de homología, al definir como sitio de unión o *binding site* la cavidad principal donde aparecen los ligandos co-cristalizados, se obtuvo como resultado del *docking* que sólo unas pocas moléculas (319 de las 4796 que forman el *test-set*), son capaces de interactuar con la proteína. Esto se debe a que Discovery Studio es muy sensible a la hora de establecer interacciones proteína-ligando en función de la geometría y tamaño del sitio de unión a la hora de interactuar con los ligandos. Por este motivo y con el objetivo de mejorar el resultado del *docking*, se procedió a dividir el *binding site* inicialmente definido, en varios niveles de partición y por tanto en subsitios activos o *binding subsites*. Esta

herramienta permite realizar una búsqueda exhaustiva en sitios activos de gran tamaño como es el caso de CXCR4 y que presentan una gran apertura, como en el caso del modelo creado por homología²⁵⁴. De este modo, al dividir el *binding site* en 2 niveles de partición se mejoró el número de moléculas que recupera el *docking* (4574 del total de 4797) y al dividirlo en 3 niveles de partición se mejoró aún más (4646 del total de 4797) tal y como se muestra en las curvas ROC para las funciones de Ligandfit en la Fig. 2.2.10. detalladas en el artículo referenciado.

La observación de que en el mismo sitio de unión pueden existir varios subsitios de unión o bien, varios modos de unión en el receptor CXCR4, es coherente con las observaciones de Kawatkat *et al.*¹²¹, Neves *et al.*¹¹⁹ Lam *et al.*²³² y Choi *et al.*²⁵⁹.



homolMod	% Falsos positivos Cribados	% Verdaderos Positivos	pAUC
	1%	34.7% / 24.2%	0.61 / 0.59
GoldScore/ GoldScore&ChemScore	5%	52.0% / 43.1%	0.70 / 0.66
	10%	61.3% / 51.2%	0.73 / 0.68

Fig. 2.2.10. En la parte superior se muestran las curvas ROC superpuestas para el modelo creado por homología. Este modelo, obtiene buenos resultados para las funciones de *scoring* GoldScore, ChemScore, GoldScore con *rescoring* ChemScore y Jain y en general buenos resultados de *Early Recovery* (significativamente buenos al 1% y 5%). En la tabla inferior se muestran los parámetros asociados a la curva ROC para las funciones de *scoring* GoldScore y GoldScore con *rescoring* ChemScore al 1%, 5% y 10% de *decoys* cribados, obteniendo un AUC global de 0.83 y 0.75 respectivamente.

2.2.3 Análisis del cribado virtual retrospectivo por subsites.

La ejecución del *docking* retrospectivo utilizando varios niveles de partición del sitio activo permite realizar el análisis de recuperación de familias de activos (*Scaffold retrieval analysis*) por subsites. Este análisis permite identificar 2 aspectos:

-Identificar posibles subsites en los cuales se ubican las diferentes moléculas activas al realizar el *docking* y por tanto determinar en donde se establecen las interacciones proteína-ligando.

-El análisis de la diversidad de los activos que se distribuyen por cada *subsites*, clasificados en función de la familia estructural química a la que pertenecen.

Desde el punto de vista teórico, la definición de varios niveles de partición de un sitio activo en Discovery Studio al realizar un *docking*, implica la exploración de los diferentes *subsites* definidos individualmente en cada partición. Es decir, cuando definimos 3 *niveles de partición* del sitio activo, en realidad ello implica explorar tre posibilidades de partición, realizando 1 sola partición (*subsite 1_1*), explorar 2 particiones (*subsite 1_2* y *2_2*) y 3 particiones (*1_3*; *2_3* y *3_3*) y lo que conlleva la realización de 6 *dockings*.

En el apartado anterior se realizó el *docking* retrospectivo del modelo CXCR4 creado por homología definiendo diferentes *subsites*, con el objetivo de mejorar el cribado. A continuación se llevó a cabo el mismo análisis de cribado virtual por *subsites* para el PDB experimental que mejor discrimina activos de inactivos (3OE6), y se compararon ambos resultados. Para ello, se alineó previamente la estructura 3OE6, a la estructura creada por homología para obtener la misma definición de particiones. Los resultados obtenidos para las 2 estructuras proteicas se muestran en la Fig. 2.2.11.

<i>Subsite</i>	1_1	1_2	1_3	2_2	2_3	3_3
<i>nº molec. HomolMod^a</i>	2	100	9	1	116	0
<i>nº molec. 3OE6^b</i>	96	11	12	115	12	1
<i>AMD^c</i>	1.2%-33.3%	50.6%-1.2%	7.4%-4.9%	1.2%-56.8%	32.1%-7.4%	0.0%-0.0%
<i>Tetrahydroquinolina</i>	0.0%-39.8%	37.4%-6.5%	0.8%-4.1%	0.0%-42.3%	57.7%-6.5%	0.0%-0.8%
<i>Macrociclo</i>	0.0%-26.3%	21.1%-5.3%	0.0%-10.5%	0.0%-52.6%	57.9%-5.3%	0.0%-0.0%
<i>KRH</i>	0.0%-56.5%	34.8%-4.3%	8.7%-4.3%	0.0%-30.4%	34.8%-0.0%	0.0%-0.0%

Fig. 2.2.11. Clasificación de activos recuperados por el *docking* retrospectivo al definir 3 niveles de partición idénticos en el modelo proteico 3OE6 y en el modelo creado por homología.

^a número total de moléculas recuperadas en cada *subsite* al ejecutar el *docking* sobre el modelo proteico generado por homología.

^b número total de moléculas recuperadas en cada *subsite* al ejecutar el *docking* sobre el modelo proteico 3OE6.

^c porcentaje de activos recuperados de la familia AMD, ubicados en cada *subsite* tanto para el modelo creado por homología (primer valor) como para la estructura 3OE6 (segundo valor separado por guión).

Según este análisis, parecen existir diferencias notables entre ambos modelos a la hora de ubicar las diferentes familias de compuestos activos por *subsites*. Así, mientras para el modelo creado por homología los *subsites* donde más interaccionan los activos son el *subsite 1_2* y *2_3*, para la estructura 3OE6, los *subsites* con mayor número de ligando ubicados son: *1_1* y *2_2*. Los *subsites 2_2* y *2_3* son equivalentes, puesto que corresponden a *subsitios* equivalentes resultantes de dividir el sitio activo en 2 y 3 particiones respectivamente (Fig. 2.5.4.). En cambio el *subsite 1_1* corresponde a la parte central

del sitio activo completo, es decir con 1 partición. Estas pequeñas diferencias muestran como variaciones estructurales existentes entre dos modelos, pueden generar interacciones proteína-ligando diferentes, tanto al comparar las interacciones establecidas (nº de molec. Totales por *subsite*), como al comparar la distribución de las familias por *subsite* según la estructura.

A la vista de todos estos resultados podemos concluir que CXCR4 es una proteína con un amplio sitio activo (Fig. 2.2.12.) en el que se unen antagonistas de diversas familias (estructuralmente diferentes), lo cual queda corroborado por la mejora de resultados de *docking* al dividir el sitio activo en *subsites* (como si cada subsitio se tratase de un sitio activo diferente). Ello nos lleva a definir CXCR4 como una “diana difícil” a la que se unen “*multisite binders*”, a la cual podríamos intentar aplicar el protocolo de mejora de cribado de 3 pasos presentado en el capítulo anterior para cribar prospectivamente moléculas candidatas a antagonistas de CXCR4 (véase capítulo 2.4)

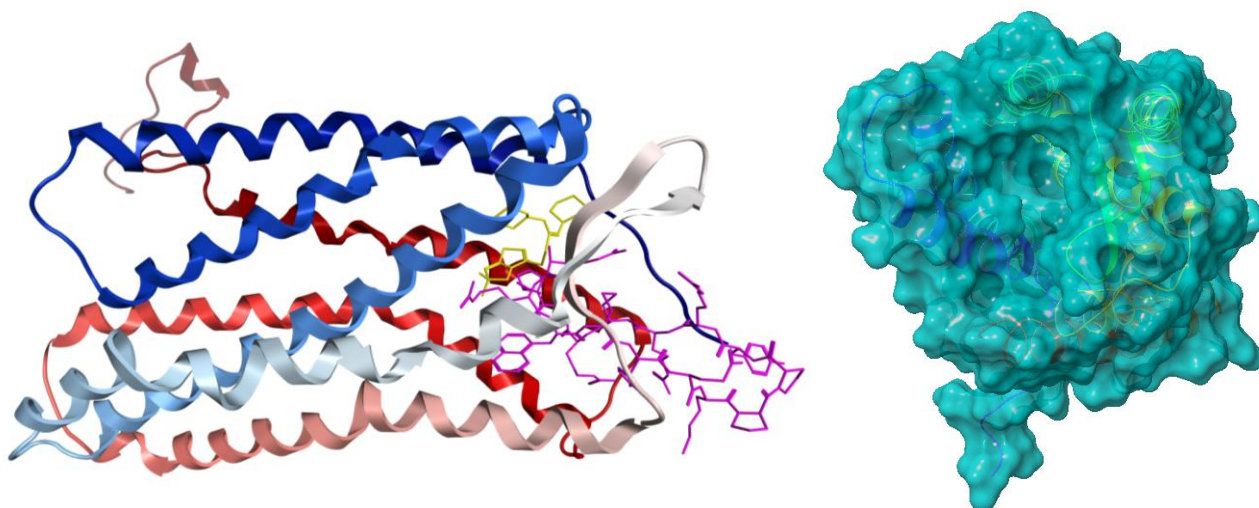


Fig. 2.2.12. A la izquierda se muestra la estructura proteica de CXCR4 representada en hélices, que contiene los ligandos IT1t (amarillo) y el péptido de gran tamaño CVX15 (rosa). Ambos ligandos se encuentran alojados en la cavidad extracelular de CXCR4 que contiene el sitio activo ortostérico. A la derecha se muestra el dominio extracelular del receptor CXCR4 y el detalle de la cavidad delimitada entre los bucles extracelulares.

2.3 Cribado virtual retrospectivo de antagonistas de CXCR4. Comparación del método de cribado virtual basado en la estructura (*docking*) con métodos basados en la forma tridimensional del ligando (*Shape matching*) y modelos farmacofóricos generados a partir de las interacciones proteína-ligando.

En este capítulo se comparan los resultados obtenidos en el apartado anterior con otros métodos de cribado virtual retrospectivo de antagonistas de CXCR4. Se plantea pues ahora el estudio comparativo entre métodos basados en la estructura (*docking*), métodos basados en la forma tridimensional del ligando (*shape matching*) y métodos basados en cribados farmacofóricos, aplicados al cribado virtual retrospectivo en CXCR4. En total, se han considerado 18 métodos diferentes de clasificación y filtrado de compuestos activos e inactivos.

En el artículo anexo: “Highly Specific and Sensitive Pharmacophore Model for Identifying CXCR4 Antagonists. Comparison with *Docking* and Shape-Matching Virtual Screening Performance” (Artículo 3) se detalla el estudio de la comparativa de cribado retrospectivo aplicando dichos métodos, así como su validación.

En este trabajo, primeramente se ha compilado una nueva quimioteca de 228 compuestos antagonistas de CXCR4, correspondientes a 20 diferentes familias de inhibidores y cuyos valores de actividad son inferiores a 59 μM . De forma complementaria, como compuestos presuntos inactivos, se ha utilizado la misma quimioteca de 4696 extraída de la Maybridge Screening Collection database, también utilizada en el capítulo anterior.

El *test-set* formado por los activos y *decoys* fue comparado frente a otros inhibidores conocidos de diferentes dianas del VIH que no son CXCR4. El objetivo es garantizar que las propiedades moleculares sean semejantes a las de otros inhibidores de HIV y que por tanto, su discriminación mediante los métodos de cribado virtual no sea realizada atendiendo a diferencias en estas propiedades. Para ello, se extrajeron de la base de datos Binding DB⁶, 928 inhibidores integrasa, 2374 inhibidores proteasa, 2484 inhibidores de la transcriptasa reversa y 354 inhibidores de CCR5. En referencia a las propiedades se compararon (Fig.2.3.1.) el peso molecular, el número de enlaces rotables, el número de posibles aceptores y dadores de hidrógenos, el número de átomos hidrofóbicos y el área superficial polar total.

Molécula	Peso Molecular	Número de enlaces rotables, <i>b_1rotN</i>	Número de átomos aceptores de hidrógeno, <i>a_acc</i>	Número de átomos dadores de hidrógeno, <i>a_don</i>	Número de átomos hidrofóbicos, <i>ahyd</i>	Area Superficial polar total, TPSA
228 CXCR4 antagonistas	455.2 (115.5)	7.0 (3.8)	2.4 (1.4)	0.9 (1.1)	24.9 (6.2)	73.5 (32.2)
4696 decoys	497.4 (45.5)	5.8 (2.2)	3.2 (1.6)	0.9 (0.9)	23.7 (4.1)	83.5 (33.4)
354 CCR5 antagonistas	560.9 (86.6)	9.3 (3.6)	3.3 (1.6)	0.8 (0.8)	30.3 (4.7)	74.0 (31.5)
928 HIV inh. integrasa	465.7 (299.0)	8.1 (10.8)	4.7 (3.4)	2.5 (3.9)	20.4 (3.9)	138.3 (131.1)
2374 HIV inh. proteasa	597.3 (115.1)	11.69 (4.1)	4.8 (2.1)	3.0 (1.7)	31.0 (5.9)	123.6 (39.8)
2484 HIV inh. inh. transcriptasa reversa	388.1 (217.0)	5.9 (6.9)	3.5 (2.2)	1.3 (1.8)	17.7 (10.2)	81.7 (65.0)

Fig. 2.3.1. Comparativa de propiedades moleculares de activos de CXCR4 frente a supuestos inactivos utilizados en el *test-set*, así como otros activos de HIV. Los valores entre paréntesis corresponden a la desviación estándar.

La estructura PDB 3OE6, fue utilizada como modelo de CXCR4, por ser la estructura más eficiente a la hora de seleccionar activos frente a supuesto inactivos, de acuerdo con el cribado virtual realizado en el apartado 2.1. Esta estructura contiene como ligando el compuesto isotiourea IT1t²⁶⁰ (Fig. 2.2.12.) y en base a las interacciones proteína ligando 3OE6-IT1t se desarrolló un farmacóforo con ayuda del módulo generación automática de farmacóforos *structure-based* de LigandScout 3.1¹⁸⁶, el cual tiene en cuenta a la hora de crear las características químicas tanto al ligando como a la proteína. Este protocolo identificó 7 características farmacofóricas (*features*) clave de dichas interacciones: 2 características de carga en los nitrógenos del grupo tiourea, 2 enlaces de hidrógeno dadores con los residuos Cys286 y Glu288 y 3 características hidrofóbicas que representan la alta hidrofobia del sitio activo de CXCR4. El ajuste final de las características farmacofóricas fue realizado posteriormente de forma manual, suprimiendo las características no consideradas clave para el conjunto de posibles inhibidores de CXCR4, obteniéndose el farmacóforo final mostrado en la figura Fig.2.3.2., el cual no contiene ningún volumen de exclusión.

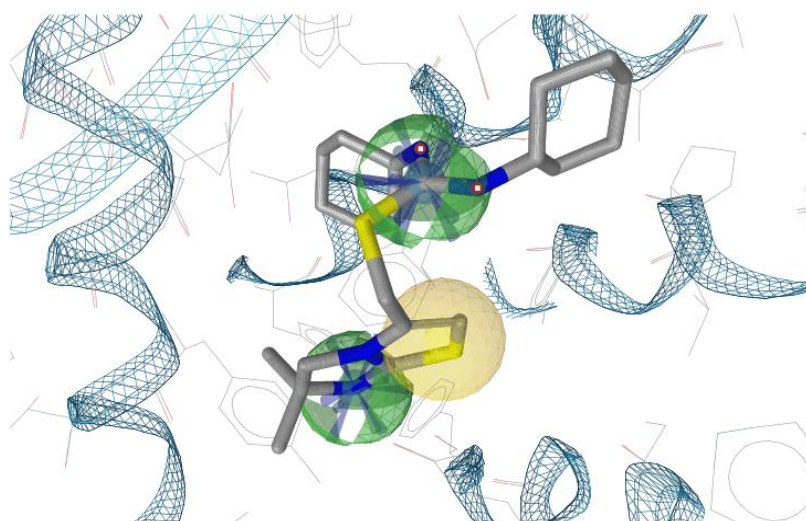


Fig. 2.3.2. Detalle de las características farmacofóricas generadas a partir de las interacciones proteína-ligando entre 3OE6 y la isotiourea IT1t.

El filtrado farmacofórico de los 228 activos se realizó partiendo de las conformaciones que el *docking* DockScore predecía y generando posteriormente, 25 conformaciones para cada uno de los activos.

Como funciones de *docking*, para realizar el cribado virtual basado en la estructura, se utilizaron las funciones de *scoring* utilizadas en el apartado 2.1: GoldScore, ChemScore, y GoldScore con *rescoring* ChemScore, Glide, TotalScore de Surflex, AutoDockVina y las funciones de *scoring* de Ligandfit: LigScore1-Dreiding, LigScore2-Dreiding, PLP1; PLP2, Jain, PMF y DockScore.

Por último, para completar la comparativa se evaluaron las funciones basadas en la forma tridimensional de los ligandos PARASURF/PARAFIT^{261,262}, ROCS²⁶³ y HPCC²⁶⁴. La utilidad de estos métodos se comprobó según su mejor o peor habilidad de superponer la forma tridimensional de cada compuesto de la base de datos, sobre la del compuesto de referencia isotiourea extraído de la estructura cristalográfica 3OE6. Para cada compuesto de la base de datos se superpusieron las mismas 25 conformaciones utilizadas para el filtro farmacofórico. Para cada compuesto, la conformación que mostró la mejor superposición con la molécula de referencia fue seleccionada y calculado su valor de superposición mediante una función de *scoring*.

Las curvas ROC resultantes de los cribados virtuales realizados para los diferentes métodos se muestran en la figura Fig.2.3.3.

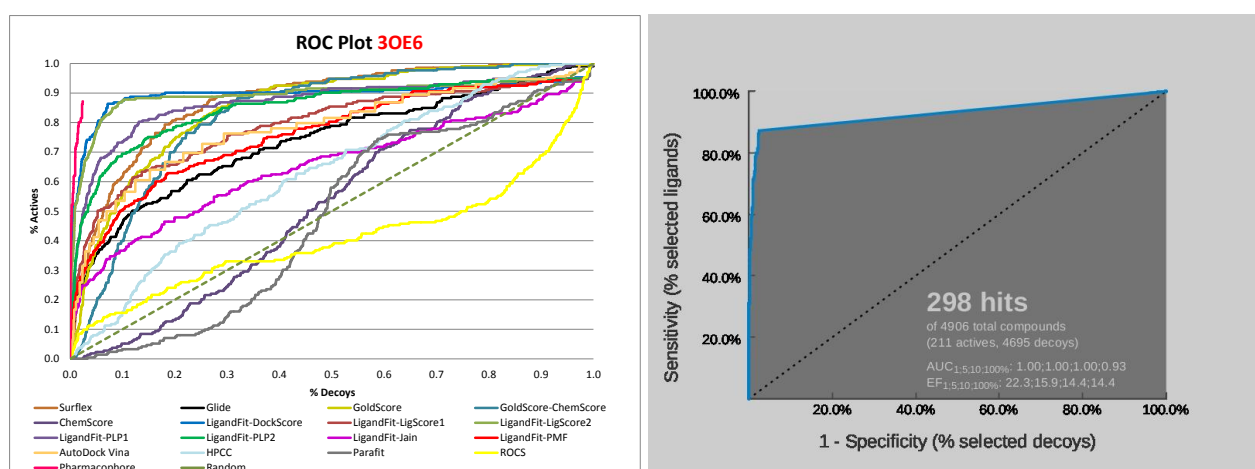


Fig. 2.3.3. A la izquierda se muestra la comparativa de las curvas ROC generadas por los métodos basados en la estructura (*docking*), por métodos basados en la forma tridimensional del ligando y mediante generación de farmacóforo creado según las interacciones 3OE6-IT1t. A la derecha se muestra el detalle de la curva ROC para el farmacóforo, realizando la hipótesis de proyección proporcional en la recuperación de los activos no devueltos por el farmacóforo.

Como se puede apreciar en la Fig.2.3.3, los métodos que obtienen mejores curvas ROC y mayor recuperación de activos a bajos porcentajes de falsos activos recuperados, son la función de *docking* DockScore y principalmente el farmacóforo construido basado en las interacciones CXCR4-IT1t. Los valores de pronta recuperación de activos se muestran en Fig.2.3.4.

	% Falsos positivos cribados	% Activos	EF (Factor Enriquecimiento)	pAUC	AUC
DockScore	1%	52.6%	22.6	0.72	0.88
	2%	62.0%	20.2	0.75	
	5%	77.9%	13.4	0.81	
	100%				
Farmacóforo	1%	71.1%	22.3	0.75	0.93 ^a
	2%	80.6%	21.1	0.82	
	5%	- ^a	15.9	0.88	
	100%				

Fig. 2.3.4. Valores de pronta recuperación de activos para los métodos más eficientes (DockScore y farmacóforo) en el cribado virtual retrospectivo de activos de CXCR4.

^a El valor AUC es estimado en base a la clasificación random de activos e inactivos no devueltos por el farmacóforo. La curva ROC no llega al 100% de la base de datos cribada puesto que sólo un 2.4% de falsos activos son devueltos por el farmacóforo (6.1% del total de la base de datos a cribar).

Debido a que el farmacóforo devuelve los mejores valores de pronta recuperación de activos, se evaluó su sensibilidad y especificidad frente a inhibidores pertenecientes a las diferentes dianas del VIH: integrasa, proteasa, transcriptasa reversa y CCR5. El cálculo de la sensibilidad se realizó considerando como activos a los inhibidores de las diferentes dianas del VIH y como inactivos los 4696 decoys de CXCR4. La especificidad fue determinada considerando como activos los 228 inhibidores de CXCR4 y como inactivos a los inhibidores de las diferentes dianas del VIH. Los resultados obtenidos muestran una elevada sensibilidad y especificidad del farmacóforo generado, tal y como recogen los datos indicados en la Fig.2.3.5.

		CXCR4 antagonistas +Maybridge Decoys	CCR5 inh. +Maybridge Decoys	HIV integrasa inh. + Maybridge Decoys	HIV proteasa inh. + Maybridge Decoys	HIV transcrip. reversa inh. + Maybridge Decoys
EFrel	1%	95.9	60.0	58.9	87.1	35.2
	2%	90.8	56.0	50.0	83.7	30.8
	5%	68.6	- ^a	- ^a	72.2	- ^a
pAUC	1%	0.75	0.55	0.51	0.52	0.50
	2%	0.82	0.58	0.52	0.53	0.50
	5%	0.88	- ^a	- ^a	0.54	- ^a

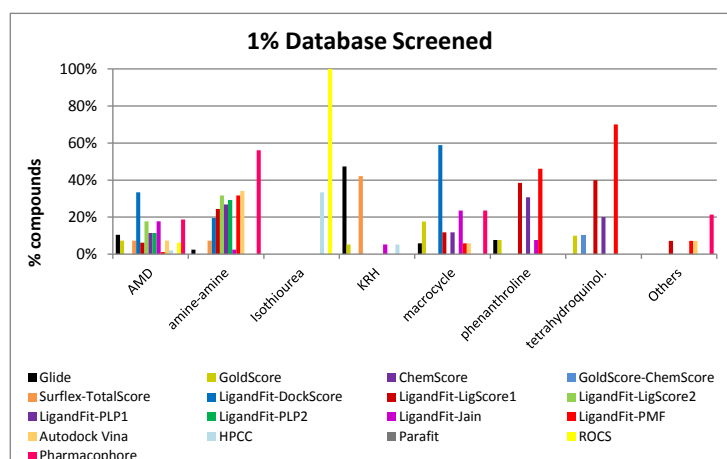
		CXCR4 antagonistas + CCR5 inh.	CXCR4 antagonistas +HIV integrasa inh.	CXCR4 antagonistas +HIV proteasa inh.	CXCR4 antagonistas +HIV transcrip. reversa inh.
EFrel	1%	100.0	100.0	96.0	100.0
	2%	100.0	100.0	92.2	100.0
	5%	- ^a	- ^a	68.0	- ^a
pAUC	1%	0.55	0.69	0.66	0.76
	2%	0.58	0.69	0.67	0.83
	5%	- ^a	- ^a	0.70	- ^a

^a Ausencia de moléculas devueltas por el farmacóforo al porcentaje indicado.

Fig. 2.3.5. Tablas sensibilidad (arriba) y de especificidad (abajo) del filtro farmacóforico aplicado a inhibidores de CXCR4 así como a otros inhibidores del VIH de dianas diferentes.

Cabe señalar como los métodos basados en la forma tridimensional del ligando, devuelven los peores valores asociados a las curvas ROC. Esto puede ser debido al gran tamaño de la cavidad de CXCR4 que permite alojar múltiples subsitos activos en dicha cavidad. De este modo, diferentes familias de activos, que poseen diferentes formas tridimensionales, podrían unirse en diferentes subsitos activos dentro de la misma cavidad de la proteína. De este modo, la elección de una única forma tridimensional activa como referencia y sobre la que comparar los compuestos de una base de datos, contemplaría únicamente un tipo de familias de compuestos (tipo isotiourea) ligado a un sitio activo y por tanto no contemplaría el resto de familias activas, que al tener diferentes formas tridimensionales, pueden unirse al receptor en diferentes subsitos activos.

Por este motivo, se estudió la recuperación de familias de activos (*Scaffold Retrieval analysis*), del mismo modo que se realizó en el apartado anterior 2.1 para su correspondiente quimioteca de activos. De esta forma es posible verificar para esta nueva quimioteca, como se recuperan las diferentes familias de activos, por parte de los diferentes métodos analizados. En la Fig. 2.3.6. se observa que sólo los métodos DockScore y el filtro farmacofórico recuperan todas las familias de activos al 5% de la base de datos cribada mientras que los métodos basados en la forma tridimensional del ligando (PARASURF/PARAFIT, ROCS y HPCC) además de devolver los peores resultados, estos métodos únicamente recuperan activos de las familias isotiourea y AMD, recuperando el 0% del resto de familias al cribar el 5% y el 10% de la base de datos. En cambio, en los casos de DockScore y del filtro farmacofórico, ninguna familia queda excluida por ambos métodos, quedando la mayoría de familias recuperadas con valores superiores al 90%, al cribar el 10% de la base de datos.



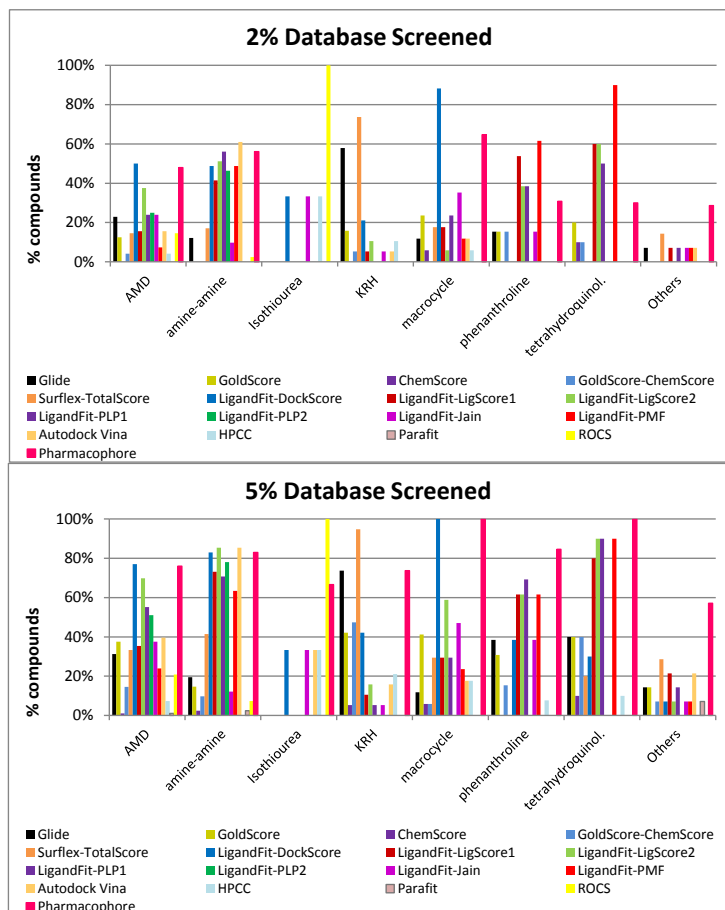


Fig. 2.3.6. Recuperación de familias de activos (*Scaffold Retrieval analysis*) al clasificarlos al 1%, 2% y 5% de la base de datos cribada utilizando la estructura 3OE6 según las 13 funciones de *scoring*, las 3 funciones *Shape matching* y el farmacóforo proteína-ligando.

2.4 Cribado virtual prospectivo de CXCR4. Generación de una quimioteca combinatoria de compuestos para su cribado virtual prospectivo. Propuesta de nuevas moléculas potenciales antagonistas del receptor CXCR4.

El uso de compuestos polinitrogenados es una de las estrategias más habituales en la búsqueda de compuestos activos para diferentes dianas terapéuticas. En el caso de CXCR4, se han desarrollado numerosos compuestos polinitrogenados como son: la familia de compuestos derivados de monociclamos y biciclamos, entre ellos el AMD3100²⁴⁸, AMD3165²⁵⁰, AMD3329²⁵⁰, AMD3389²³⁹ u otros compuestos AMD no ciclamos, como los derivados de AMD11070²⁶⁵. Además existe una variedad de compuestos antagonistas de CXCR4 también polinitrogenados como son los derivados dipicolilamina-Zn(II)²⁶⁶, fenantrolinas²⁶⁷, los compuestos desarrollados por Kureha²⁶⁸, por Taigen²²⁶, o derivados de tetraquinolina^{269,270} (Fig. 2.4.1.) los cuales también han desarrollado compuestos análogos mediante desarrollo de quimiotecas²⁷¹.

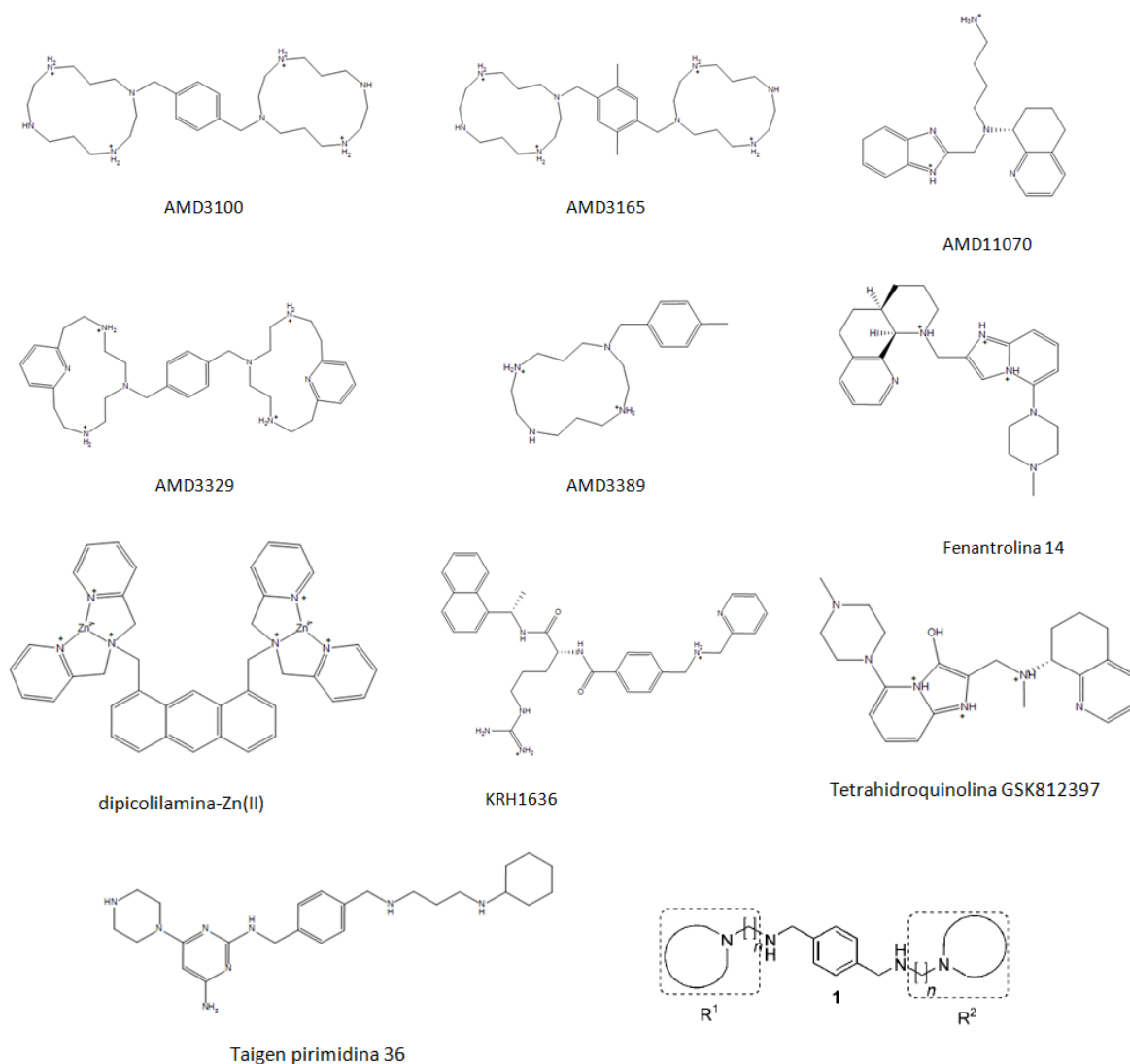


Fig. 2.4.1. Compuestos polinitrogenados representativos de las familias químicas de antagonistas de CXCR4 y estructura genérica (abajo a la derecha) de los compuestos diamino descrito por Perez-Nueno *et al*²⁷².

Trabajos anteriores en el laboratorio GEM IQS (Pettersson *et al.*^{273,274}) llevaron a proponer igualmente el desarrollo de quimiotecas basadas en compuestos diamino derivados, con características estructurales similares a las existentes en la molécula AMD3100 (Fig. 2.4.1.). Estas semejanzas se basan en la existencia de 2 átomos de nitrógeno a cada lado del fenileno central y en que las distancias entre dichos átomos de nitrógeno fueran similares a las existentes en el AMD3100.

Basándonos en las estructuras diamino generadas anteriormente en el laboratorio GEM por Pettersson *et al.*, se propone desarrollar una nueva quimioteca de compuestos diamino-derivados y estudiar la recuperación (*Early Recovery*) de dichos compuestos al realizar el *docking* de los mismos y comparar su *scoring* con otros activos y supuestos inactivos, utilizados en los modelos de los puntos 2.1 y 2.2 del presente trabajo.

La quimioteca a construir se basa en la combinación de 78 ciclohexilaminas 2-sustituidas, ubicadas a ambos extremos de diferentes núcleos derivados de los compuestos diamino desarrollados por Pettersson *et al.* (Fig. 2.4.2.).

El procedimiento realizado para construir la quimioteca combinatoria consta de los siguientes pasos:

1. Preparación de las 78 ciclohexilaminas mediante LigPrep²⁷⁵, generando los posibles estereoisómeros de estas estructuras. La preparación de las 78 ciclohexilaminas genera 88 estructuras diferentes.
2. Generación de los 3 subnúcleos de la quimioteca (Fig. 2.4.2.) mediante el módulo CombiGlide²⁰² de Schrödinger. Las 88 estructuras preparadas son combinadas con cadenas lineales de 2, 3 y 4 átomos de carbono. La combinación se realiza a través de la unión del nitrógeno de la piperidina completando el subnúcleo. En total se generan 264 estructuras (quimioteca de subnúcleos).
3. Creación de las estructuras de los 6 núcleos base a combinar, definiendo las 2 aminas terminales como puntos de unión (Fig. 2.4.2.).
4. Generación de la quimioteca combinatoria de 6 núcleos mediante combinación de la quimioteca de subnúcleos (264 estructuras) con cada uno de los 6 núcleos base, teniendo en cuenta que cada subnúcleo puede unirse por las 2 aminas terminales de cada núcleo. En total se generan 69696 (264^2) moléculas por cada núcleo, utilizando el módulo CombiGlide²⁰² de Schrödinger.
5. Las 6 quimiotecas se unen en una única quimioteca única, la cual es filtrada según las reglas de Lipinski²⁷⁶ y los compuestos resultantes son preparados mediante Discovery Studio¹⁶⁷ utilizando el protocolo *Prepare Ligands* con el fin de realizar el *docking* de la quimioteca resultante.

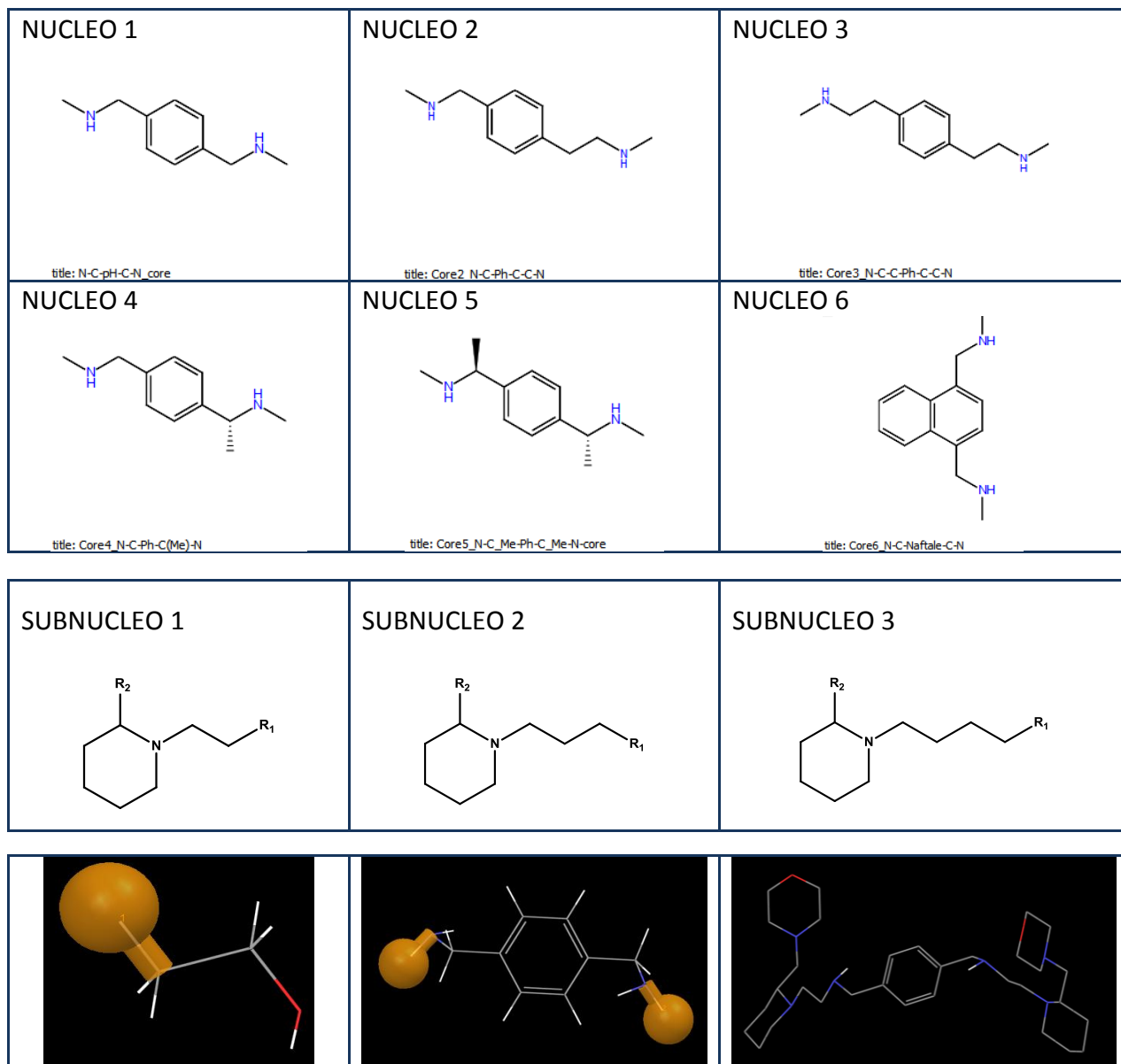


Fig. 2.4.2. Estructuras de los 6 núcleos de diamino compuestos (tabla superior) y de los 3 subnúcleos que contienen las 78 ciclohexilaminas (tabla central) así como un ejemplo de generación de diamina combinatoria (tabla inferior). La quimioteca combinatoria es generada por unión de los 6 núcleos, a través de sus 2 aminas terminales, con los 3 subnúcleos a través de R_1 . Los subnúcleos son cadenas de $n=2, 3$ y 4 átomos de carbono lineales unidos al nitrógeno de la ciclohexilamina. R_2 indica el sustituyente variable de la ciclohexilamina. R_1 indica el punto de unión con los núcleos a combinar, pero no forma parte de la molécula final.

En el cuadro inferior se muestra de izquierda a derecha la secuencia de construcción de la quimioteca combinatoria. A la izquierda se muestra el subnúcleo 1, de 2 átomos de carbono y el anclaje (esfera naranja) por donde se unen las ciclohexilaminas. Posteriormente dicho subnucleo se unirá al núcleo por los átomos marcados en rojo y otro en blanco, eliminándose dichos átomos. En el cuadro inferior central aparece el núcleo 1 con sus 2 anclajes (bolas naranjas) ubicados en cada amina terminal, por donde se unirán simultáneamente la quimioteca de subnúcleos. A la derecha se muestra una molécula generada combinatoriamente del núcleo 1 y del subnúcleo 1.

Una vez generadas, filtradas y preparadas las estructuras de la nueva quimioteca, se puede proceder a clasificarlas y/o filtrarlas mediante métodos de cribado virtual. Las moléculas que tengan una prona

recuperación (*Early Recovery*) mayor son retenidas y consideradas con probabilidad de ser activas. Dichas moléculas serán buenas candidatas a ser sintetizadas.

Con el objetivo de tener una amplia variedad de métodos de clasificación, que elimine posibles tendencias en la clasificación según la tipología del método utilizado, se plantea utilizar el método de cribado virtual propuesto para “dianas difíciles” (dado que según el análisis detallado en el capítulo 2.2, CXCR4 cumple con las características de “diana difícil”) consistente en *docking*, postfiltrado farmacofórico y ordenación de poses por similitud 2D, y variaciones de éste, utilizando métodos de clasificación y filtrado *structure-based*, *ligand-based* (mezclando los anteriores) de acuerdo con los mejores parámetros hallados en el análisis retrospectivo del cribado virtual de antagonistas de CXCR4.

De este modo, la quimioteca combinatoria de diamino compuestos se clasificó (Fig. 2.4.3.) mediante los métodos:

-Función de *docking* LigandFit¹⁶⁴ y ordenando las conformaciones resultantes del *docking* mediante la función de *scoring* DockScore tal y como se ha determinado en el punto 2.2. y 2.3 del presente trabajo.

-Cribado de estructuras según el farmacóforo descrito por Perez-Nueno *et al.*²⁷², llamado posteriormente Farmacoforo 1, y posterior ranking de las estructuras resultantes mediante la función de *scoring* DockScore.

-Cribado de estructuras según el Farmacóforo 1 con posterior ranking de las estructuras resultantes, mediante el *RMSD* que calcula el farmacóforo, el cual cuantifica cómo se ajusta cada molécula a dicho farmacóforo.

-Ordenación *consensus* entre la clasificación que devuelve para cada molécula la función DockScore y la ordenación que devuelve el Farmacóforo 1 al ser ordenada por el *RMSD* del farmacóforo. La ordenación *consensus* se calcula realizando la media entre las posiciones que ocupa una molécula en cada clasificación:

$$Val. Consensus (mol_i) = \frac{ranking DockScore (mol_i) + ranking farmacof rmsd (mol_i)}{2}$$

[2.1]

-Cribado de estructuras según el farmacóforo desarrollado en el capítulo 2.3 basado en las interacciones proteína ligando 3OE6-IT1t y generado utilizando LigandScout (llamado posteriormente Farmacóforo 2). La clasificación de las poses cribadas se clasifican posteriormente mediante la función de *scoring* DockScore.

-Cribado de estructuras según el Farmacóforo 2 y posterior ranking de las estructuras resultantes, mediante el *RMSD* que calcula el farmacóforo.

-Ordenación *consensus* entre la clasificación que devuelve la función DockScore para cada molécula y la ordenación que devuelve el Farmacóforo 2 al ordenarlas por el *RMSD* que calcula el farmacóforo.

- Ordenación *consensus* entre los rankings que devuelven: el cribado del Farmacóforo 1 al clasificar por su *RMSD*, la ordenación del *RMSD* del Farmacoforo 2 y la clasificación que devuelve el *docking* clasificado según DockScore.

$$Val.Consensus (mol_i) = \frac{ranking\ rmsd\ Pet.(mol_i) + ranking\ rmsd\ Harm.P (mol_i) + ranking\ DockScore (mol_i)}{3}$$

[2.2]

En la figura 2.4.3. se recogen los rankings de las 20 primeras estructuras de la quimioteca combinatoria, al ser clasificadas según los diferentes métodos acabados de describir.

Estructuras amino-amino combinatorias con RANKING en las 20 primeras posiciones							
Scoring	Farmacóforo 1			Farmacóforo 2			
DockScore	DockScore + Farmacóforo 1	Facmacofofo 1 (rmsd)	Consensus DockScore + rmsd	DockScore + Farmacóforo 2	Facmacofofo 2 (rmsd)	Consensus DockScore + rmsd	DockScore + Farmacóforo 1 + Farmacóforo 2
1429	2199	122	4285	1429	4023	2623	3941
66	775	4451	4226	66	3891	3904	3996
3847	4663	2372	1102	3847	3850	2101	4393
2199	4801	656	2993	2199	3996	2793	3191
5986	2939	1188	4456	5986	4063	2451	2485
947	38	1722	3941	947	4033	2044	4141
2084	6310	1217	181	2084	3928	3191	5571
775	3014	5808	4393	775	3113	2149	3857
4293	2733	4928	273	4293	4006	3885	3952
2483	1102	181	675	2483	4017	4034	3040
4663	4715	1901	488	4663	2090	3952	3402
4801	2999	943	542	4801	2769	3014	4210
2939	6309	3549	224	2939	2309	2318	4801
4306	4226	1100	6717	4306	3896	2622	43
38	4285	3356	4693	38	2799	3186	2451
6310	1055	2761	2011	6310	3901	3596	2209
3014	2963	309	5944	3014	3031	3941	3574
99	4456	5169	4071	99	4021	2996	2926
4881	1118	3070	5571	4881	3862	2752	6522
2733	2623	5064	1271	2733	2617	2110	730

*Coinciden todos con DockScore

Fig. 2.4.3. Clasificación de los diamino compuestos creados combinatoriamente, según los 8 diferentes métodos de ranking utilizados. Los compuestos se enumeran a través de su número identificativo en la quimioteca (números en rojo). Las repeticiones de mismos compuestos hallados por diferentes métodos, se muestran resaltando su casilla en un mismo color (p.e. el compuesto 4801 se halla en los rankings de DockScore en la posición nº12, en DockScore + filtro Farmacoforo 1 (rmsd) en la posición nº4 y en DockScore + Farmacóforo 1 (rmsd) + Farmacóforo 2 (rmsd) en posición nº13.). Los 8 métodos usados se clasifican en 4 grupos diferentes: El grupo que clasifica únicamente por *scoring* con DockScore, el segundo grupo aplica el Farmacóforo 1 (rmsd) en todos los métodos, el tercer grupo aplica el Farmacóforo 2 (rmsd) en todos los métodos y el cuarto grupo es un método *consensus* que contempla los 3 grupos anteriores.

De la aplicación de estos métodos, cabe tener en cuenta las siguientes consideraciones:

-El Farmacóforo 1 al ser aplicado al *test-set* de activos/inactivos utilizado en capítulo 2.1, devuelve 100 activos del total de 248 activos que constan en el *test-set* (40.3%) y sólo 58 inactivos de los 4696 posibles inactivos (1.3%). Igualmente, al aplicar este farmacóforo al *test-set* utilizado en el capítulo 2.3, devuelve 93 activos de los 228 activos totales (40.8%). Cabe indicar que el set de supuestos inactivos usados en ambos *test-sets* es idéntico.

Al aplicar el Farmacóforo 1 a la quimioteca combinatoria generada únicamente 3850 estructuras fueron devueltas por el farmacóforo.

-El Farmacóforo 2 cribó 211 activos del total de 228 activos (92.5%) y 87 inactivos de los 4696 inactivos totales (1.8%) para el *test-set* utilizado y analizado en el capítulo 2.3.

Al aplicar el Farmacóforo 2 a la quimioteca combinatoria creada, un total de 6754 estructuras fueron devueltas por el farmacóforo.

		<i>DockScore</i>	<i>Farmacoforo 1</i>	<i>Farmacoforo 2</i>
EF	1%	22.6	17.6	22.3
	2%	20.2	15.6	21.1
pAUC	1%	0.72	0.63	0.75
	2%	0.75	-1	0.82

¹No existen valores al 2% de falsos positivos ya que el farmacóforo recupera 1.1% de falsos activos.

Fig. 2.4.4. Tabla comparativa de valores de pronta recuperación (*Early Recovery*) al aplicar los 3 modelos de clasificación y cribado de poses al *test-set* de 228 activos y 4696 inactivos (*test-set* capítulo 2.3).

-Al aplicar rankings basados en similitud molecular 2D Tanimoto como criterio de postordenación en cribados farmacofóricos (tercer paso del protocolo para “dianas difíciles”), encontramos que todas las estructuras que componen la quimioteca combinatoria dan valores de Tanimoto bajo random. Este hecho se produce tanto al escoger como compuesto de referencia la isotiourea IT1t, como el AMD3100. Así pues, este criterio de ordenación no es útil en este caso, debido a la existencia de múltiples familias de compuestos activos con diferentes similitudes moleculares, entre ellas, la familia de compuestos diamino derivados generada. La existencia de diferentes familias y diferentes similitudes moleculares, muy probablemente está asociada a la existencia de múltiples sitios y subsitios activos existentes en la cavidad extracelular de CXCR4. En el apartado 2.2. y dentro del artículo: *Impact of the CXCR4 structure on docking-based virtual screening of HIV entry inhibitors* se muestra la diferente distribución de las familias de activos (*chemotype families*) por subsitios de unión, al realizar el *docking* de estos activos dividiendo la cavidad extracelular de CXCR4 en diferentes subsitios. Igualmente, en el apartado 2.3 y dentro de los resultados obtenidos en el artículo: *Highly Specific and Sensitive Pharmacophore and sensitive pharmacophore model for identifying CXCR4 antagonists. Comparison with docking and shape-matching*

virtual screening performance se observa como los métodos de cribado virtual basados en *Shape Matching* (ROCS, Parafit y HPCC) devuelven pobres curvas ROC y pobres *Early Recoveries*, éste hecho reforzaría la idea de distintas similitudes moleculares ligada a múltiples sitios de unión. Cada subsitio de unión tendría su molécula de referencia sobre la que comparar similitud 2D.

Del análisis de los métodos utilizados para clasificar la quimioteca combinatoria, podemos concluir:

- Al ordenar la quimioteca combinatoria por la función de *scoring* DockScore y mezclar dicha quimioteca con el set de activos y supuestos inactivos utilizados anteriormente, observamos que la mayoría de compuestos de la quimioteca quedan ordenados entre los activos del *test-set*. Este hecho indica su potencial a ser compuestos activos en la inhibición de CXCR4.
- El filtro farmacofórico 1 puede ayudar a reducir el gran número de posibles candidatos a sintetizar que existen. Sin embargo, cabe destacar que dicho farmacóforo elimina el 60% de activos de los dos *test-sets* utilizados en el presente trabajo, con el riesgo de poder eliminar determinadas familias químicas activas.
- Las moléculas ordenadas en las primeras posiciones que el Farmacóforo 1 devuelve, aparecen repetidas en otros rankings (Fig. 2.4.3).
- El ranking con menor repetición de estructuras corresponde al Farmacóforo 2, ordenado por el RMSD que en el cribado se calcula.
- Las ordenaciones más efectivas en términos de EF y pAUC son las del Farmacóforo 2 y DockScore.
- Las moléculas que más se repiten en los diferentes rankings (Fig.2.4.5) son: ID 3014 y la 4801 que aparecen en 4 de los 8 rankings realizados. La estructura I.D. 3941 aparece en 3 rankings. Esta estructura aparece en primer lugar en la clasificación *Consensus* generada entre el Farmacóforo 1 al clasificar por su RMSD, la ordenación del RMSD del Farmacóforo 2 y la clasificación que devuelve el *scoring* de DockScore. Igualmente varias de las estructuras que DockScore ubica entre las 20 primeras, se encuentran en la clasificación del cribado del Farmacóforo 1 ordenado por su RMSD, así como en el cribado del Farmacóforo 2, ordenado por DockScore.

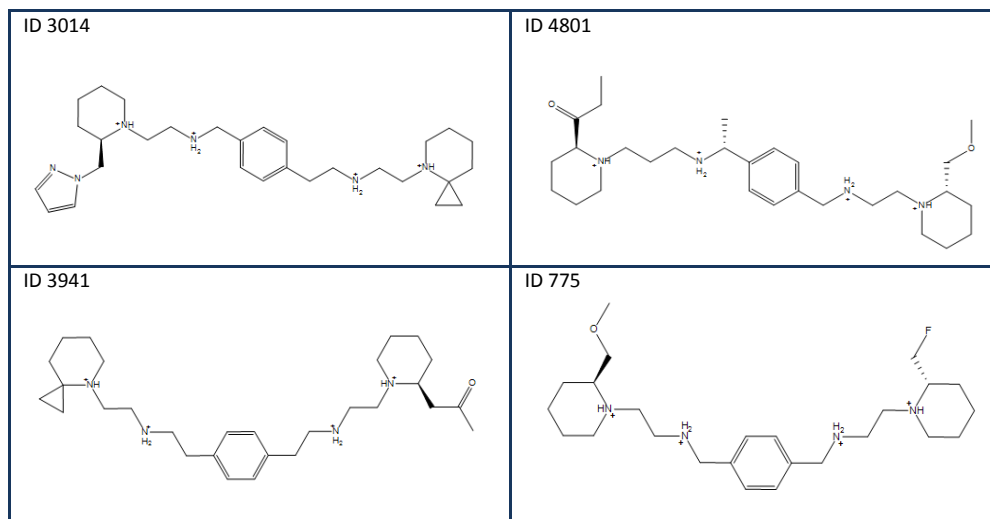


Fig. 2.4.5. Estructuras de la quimioteca combinatoria que aparecen más repetidas en las primeras posiciones del ranking utilizando múltiples métodos de cribado virtual.

Finalmente indicar, que se espera que elementos de esta quimioteca tengan actividad como antagonistas de CXCR4, tal y como recientemente estructuras análogas, como ICT5024^{277,278}, han demostrado contra el tratamiento del cáncer por metástasis. Es por ello que en laboratorio de diseño molecular del IQS (GEM) se está desarrollando la síntesis de las moléculas más asequibles sintéticamente.

2.5 Estudio de la modulación alostérica de agonistas y antagonistas de CXCR4.

La regulación alostérica en las células es un fenómeno habitual, asociado a transiciones conformacionales y funcionales de las proteínas, que habilita la actividad de un sitio activo de la proteína, para modular la función en otro sitio activo distinto.

El alosterismo se produce mediante interacciones no-covalentes a través de iones, lípidos, proteínas, ADN, ARN así como también mediante fármacos^{279,280}. Este proceso dinámico se da en proteínas con una única cadena proteica, en complejos macromoleculares de múltiples cadenas y también en ADN y ARN. Estos receptores cuando se ven sometidos a cambios conformacionales de naturaleza alostérica, pueden propagar dichos cambios a subsiguientes procesos bioquímicos actuando en otras proteínas. Estos cambios en cascada, a su vez, pueden generar perturbaciones o mutaciones alostéricas, existiendo la posibilidad de que se eliminen o generen nuevos sitios activos alostéricos, los cuales pueden desarrollar modificaciones post-traslacionales que deriven así mismo, en enfermedades²⁸¹. Sin embargo, este mismo proceso también presenta la oportunidad de permitir regular malos funcionamientos celulares, mediante la acción de ligandos alostéricos que en este caso actuarían como fármacos.

Una de las principales características de los ligandos alostéricos, es que estos interactúan en sitios activos de la molécula receptora (sitio alostérico), que son topográficamente distintos del sitio activo ortostérico común. Parece ser que la arquitectura de las proteínas estaría construida y distribuida espacialmente, mediante una serie de redes tridimensionales interconectadas llamadas sectores, en donde los aminoácidos desarrollan el papel básico de estructura y funcionalidad. Los sectores se conectarían en milisegundos a través de los residuos proteicos activando el sitio activo de una proteína y zonas más superficiales de ésta (Fig. 2.5.1.), las cuales se convertirían en zonas de regulación alostérica²⁸². Este punto es clave en el desarrollo del estudio realizado tal y como se muestra a continuación.

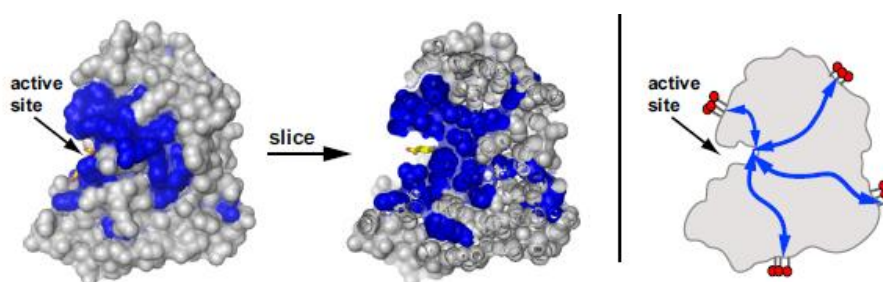


Fig. 2.5.1. Representación²⁸² de los residuos que forman el sector proteico visto desde la perspectiva de la superficie proteica (izquierda), visto desde la disección realizada por el núcleo proteico (centro) y la representación (a la derecha) de la red de conexión establecida (líneas azules) que conectan físicamente el sitio activo de la proteína con zonas superficiales que pueden convertirse en puntos de regulación alostérica, las cuales serían inducidas por cambios conformacionales.

Una manifestación clave que denota una interacción alostérica es la alteración en la afinidad de un ligando ortostérico al unirse a su sitio activo correspondiente. Esta alteración puede generar o bien un

aumento en la velocidad de asociación del ligando a la proteína o por el contrario una disminución de dicha velocidad⁸⁵.

El fenómeno del alosterismo, también se constata en las proteínas GPCR, a pesar de que en la actualidad la mayoría de fármacos existentes son ligandos que se unen al sitio activo ortostérico. A día de hoy, ya existen más de 100 moduladores alostéricos reportados para esta familia de proteínas²⁸³. En las proteínas GPCR los sitios activos alostéricos se encuentran estructuralmente y evolutivamente menos conservados que los sitios activos ortostéricos, presentando estos últimos una alta conservación. *Wang et al.*²⁸⁴ muestran que al superponer los sitios activos ortostéricos de 60 GPCRs diferentes, se observa una alta superposición estructural y de secuencia. Estas observaciones nos conducen a pensar que la regulación alostérica, abre un emergente abanico de oportunidades para modular proteínas como los receptores GPCR²⁸⁵ tal y como se ha comentado anteriormente.

Según su modo directo o indirecto de acción sobre la proteína GPCR objetivo, los moduladores alostéricos de GPCRs se clasifican²⁸⁶ como: *in-Target*, cuando actúan uniéndose directamente a la proteína GPCR objetivo, o bien como *Off-Target* si el modulador alostérico, se une a una proteína GPCR que a su vez se une al GPCR objetivo para formar un heterodímero. También hablamos de moduladores Off-Target cuando éste se une a una proteína, la cual se une al GPCR objetivo formando un complejo receptor.

El presente estudio de alosterismo de CXCR4, se basa en los trabajos de *Mosi et al.*²⁶⁵ que proponen el compuesto antagonista AMD11070²⁸⁷, también conocido como AMD070 ó 070, (Fig. 2.5.2.) como posible inhibidor alostérico de la interacción de CXCR4 con el ligando natural o endógeno, SDF-1 (CXCL12). Este antagonista orgánico actuaría de forma selectiva sin inhibir otras quimiocinas como son: CXCR3, CCR1, CCR2b, CCR4, CCR5 o CCR7²⁸⁸. Concretamente, AMD11070 inhibe la respuesta de las isoformas SDF-1 α y SDF-1 β (CXCL12a y CXCL12b respectivamente) de diferente manera y con diferentes valores de IC₅₀ lo que nos indicaría la existencia de un mecanismo mixto de inhibición.

De similar forma, el compuesto orgánico GSK812397²⁸⁹ (Fig. 2.5.2.) fue desarrollado a partir de AMD11070, mediante modificación iterativa de su estructura, con el fin de mejorar su actividad anti-viral. GSK812397 es un antagonista no competitivo inhibidor del HIV y actúa de forma selectiva sobre CXCR4. Este compuesto podría unirse al receptor en una ubicación separada y diferente al sitio activo por donde se une el HIV, de forma que una vez unido, indujera un cambio conformacional de CXCR4 de manera que no fuera posible termodinámicamente la entrada del virus²⁶⁹.

Los péptidos y sus derivados, también llamados compuestos peptidomiméticos, son unos de los compuestos más utilizados como reguladores alostéricos y también sucede en el caso del receptor CXCR4²⁹⁰. Entre estos compuestos peptidomiméticos reguladores de GPCRs encontramos las llamadas pepducinas²⁹¹. Las pepducinas son péptidos cuya longitud de cadena habitual es de 10 a 20 aminoácidos.

Son derivados de secuencias proteicas de bucles intracelulares de los GPCRs diana y que además contienen unido a uno de sus extremos, uno o varios radicales de carácter lipídico, como puede ser el radical palmitato²⁹². Estos radicales precisamente dotan a las pepducinas de capacidad de penetrar a través de la bicapa lipídica como es la membrana plasmática celular. Las pepducinas presentan la característica de ser moduladores alostéricos de las proteínas GPCRs puesto que al unirse al receptor afectan a la habilidad de éste en sufrir cambios conformacionales posteriores, bloqueando su conformación²⁹³. Además las pepducinas acostumbran a presentar un comportamiento como compuestos alostéricos selectivos o *biased* tal y como *Quoyer et al.*²⁹⁴ han determinado para la pepducina ATI-2341 en CXCR4. Los autores muestran como ATI-2341 favorece selectivamente la unión de CXCR4 con las proteínas $G\alpha_i$, pero no favorece la unión con las proteínas $G\alpha_{13}$ ni tampoco con las β -arrestinas, a diferencia del ligando natural SDF-1 que promueve la unión y activación de las 3 familias de proteínas G_i , G_{13} , β -arrestina1 y β -arrestina2. *Janz et al.*²⁹⁵ reportaron 3 pepducinas agonistas (ATI-2341, ATI-2755, ATI-2766) análogas y competitivas entre sí, al unirse a CXCR4, pero que contrariamente, no compiten con el péptido T-140, ni con el ligando natural CXCL12. Para verificar estos comportamientos de competitividad, se utilizó la pepducina ATI-2766 (Fig. 2.5.2.), que es un derivado de la ATI-2341²⁹⁶.

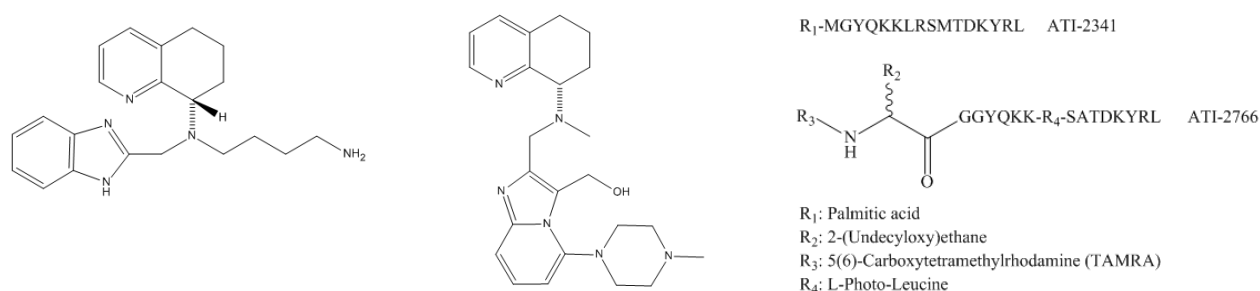


Fig. 2.5.2. Compuestos orgánicos inhibidores alostéricos de CXCR4. De izquierda a derecha los compuestos orgánicos antagonistas AMD11070, GSK812397 y de los compuestos peptidomiméticos análogos: Pepducina ATI-2341 y ATI-2766 agonistas de CXCR4.

La pepducina ATI-2766 a diferencia de la ATI-2341, posee un aminoácido menos, tiene diferentes radicales lipídicos y presenta fotoafinidad, lo que le permite llevar a cabo experimentos fotoquímicos de entrecruzamiento (*crosslinking*).

Los resultados de este estudio se han publicado en el artículo anexoado "*Studying the Binding Interactions of Allosteric Agonists and Antagonists of the CXCR4 Receptor.*" (Artículo 4) el cual estudia el alosterismo en CXCR4 mediante las posibles interacciones que se producen entre el receptor CXCR4 y los agonistas ATI-2341 así como con los antagonistas AMD11070, GSK812397 acabados de describir. Se ubican los posibles sitios activos alostéricos de CXCR4 y se estudian las interacciones clave en la unión de los moduladores alostéricos a dicha proteína.

2.4.1 Interacciones alostéricas CXCR4-pepducina ATI-2341.

Las interacciones alostéricas se han estudiado realizando en primer lugar, un *docking* ciego proteína-peptido mediante el programa ZDock²⁹⁷ el cual realiza una búsqueda rotacional y traslacional de la proteína de menor tamaño (ligando) alrededor de la proteína de mayor tamaño (receptor). Como proteína receptor se ha utilizado la estructura de CXCR4, 3OE6 y como ligando la pepducina ATI-2341, cuya conformación de partida ha sido estabilizada mediante dinámica molecular con solvatación implícita GBSW utilizando *Force Field* CHARMM2¹⁴⁵. La estructura de 3OE6 fue preparada completando en primer lugar, el bucle ICL3 incompleto, el cual se presenta originalmente unido a la proteína de fusión TL4.

La elección de la pepducina ATI-2341, como molécula en la que estudiar con detalle sus interacciones con CXCR4, se debe a que esta pepducina únicamente contiene un radical de origen no-peptídico el cual, es además de menor tamaño que los 2 existentes en el caso de ATI-2766. Este hecho favorece la utilización de protocolos de *docking* flexible péptido-proteína, los cuales no contemplan interacciones entre radicales de tipo no peptídico y por tanto, el número de posibles interacciones no tenidas en cuenta, así como los posibles impedimentos, debe ser menor. Por este motivo también se han comparado los dos protocolos de *docking* flexible FlexPepDock y HADDOCK (semi-flexible), con dinámica molecular de la pepducina ATI-2341 en CXCR4. La DM, utilizada como procedimiento de estudio de interacciones proteína-ligando^{298,299}, sí mantiene el radical no peptídico ácido palmítico. De este modo, se pretende comparar los resultados obtenidos por los tres métodos³⁰⁰, y también corroborar el papel de este radical lipófilo al interactuar con CXCR4.

Los resultados obtenidos se pueden considerar comparables a la hora de establecer las principales interacciones en los bucles intracelulares ICL1 y ICL3 principalmente y en menor modo con ICL2 (Fig. 2.5.3.). Los residuos Leu69, Arg70, Ser71, Met72 del bucle ICL1 de CXCR4 establecen en todos los métodos ensayados, interacciones con la pepducina ATI-2341. Igualmente Arg134 del ICL2 así como Ala237, Leu238 y Thr240 del bucle ICL3 también establecen interacciones en los tres métodos usados. Entre estos, para la Arg134 y la Ala237 los tres métodos coinciden al proponer enlaces por puente de hidrógenos.

La DM encuentra débiles interacciones del radical no peptídico de naturaleza lipofílica de ATI-2341, con los residuos apolares del ICL3: Gly212 y Ile213 así como con la Gly212 (Fig. 2.5.3.). Este hecho confirmaría que el radical lipídico de las pepducinas, básicamente desempeña el papel de facilitar el atravesar la bicapa lipídica de la transmembrana celular, no estableciendo interacciones importantes con CXCR4. Esta observación coincide con la realizada de forma coetánea al artículo aquí presentado, por Zhang *et al.*³⁰¹ que publican la conformación determinada experimentalmente de una pepducina formada por 19 residuos extraídos del ICL3 y unidos a un extremo palmitato, conjuntamente con el GPCR PAR1. Dicha

conformación se une al receptor PAR1 en el ICL3 y presenta el extremo palmitato paralelo a la hélice TM6 en la zona de transmembrana de forma muy similar a la conformación hallada por DM (Fig. 2.5.3.).

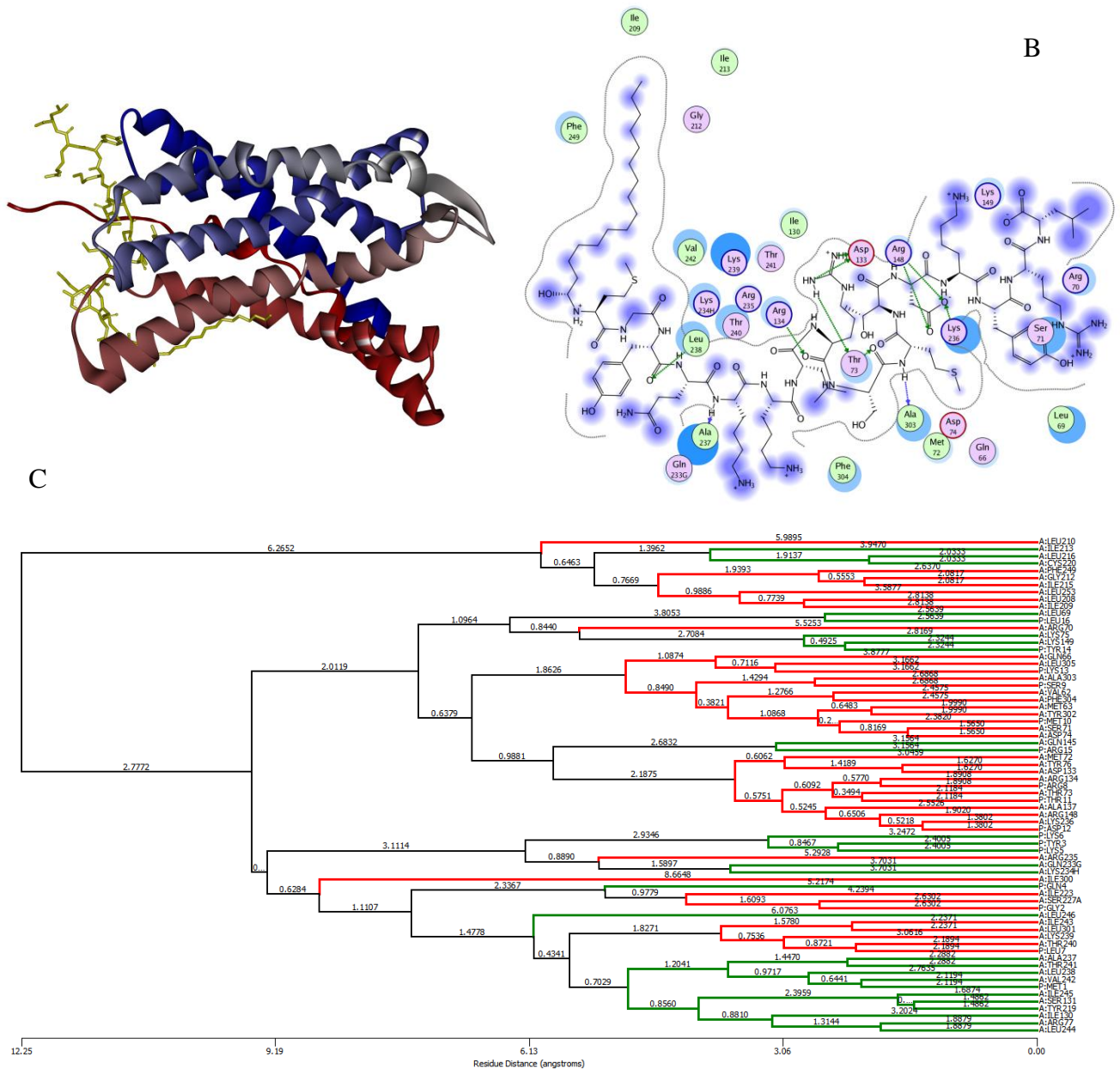


Fig. 2.5.3. **A.** Conformación más estable de la pepducina ATI-2341 (amarillo) obtenida en la estructura 3OE6 de CXCR4, mediante DM con membrana implícita GBSW de 28.5 Å y solvatación. La estructura de CXCR4 se representa en escala de color N-terminal (azul) a C-terminal (rojo) y el bucle ICL1 corresponde al giro β que se muestra en color azul. **B.** Interacciones proteína-ligando establecidas entre la conformación de ATI-2341 y CXCR4. **C.** Gráfico filograma que muestra las distancias entre los residuos de CXCR4 (residuos cadena A) y la conformación más estable de la pepducina ATI-2341 (residuos cadena B). Las distancias coloreadas en negro indican clústeres de residuos ubicados a distancias superiores a 5 Å.

Los residuos de CXCR4 que interactúan con las conformaciones halladas mediante los 3 métodos ensayados, coinciden en parte con la secuencia de 16 aminoácidos de CXCR4 del bucle ICL1, que van de los residuos Met63 hasta Leu78, a partir de la cual se han diseñado las pepducinas ATI-2341 y ATI-2766, principalmente con los residuos Leu69, Arg70, Ser71 y Met72, aunque probablemente también con los

residuos Gln66, Thr73 y Asp74. Este bucle intracelular (ICL1) se considera el dominio principal de CXCR4 donde se une la peptidina ATI-2341, de hecho, todos los métodos utilizados, además de establecer interacciones directas con estos residuos, ubican la peptidina ATI-2341 a menos de 5 Å (Fig. 2.5.3.) de estos residuos de CXCR4.

2.4.2 Interacciones de CXCR4 con los antagonistas AMD11070 y GSK812397.

Los inhibidores del VIH, AMD11070 y GSK812397 son compuestos antagonistas de CXCR4 cuyo mecanismo de acción indica que podrían modular el receptor CXCR4 de forma alostérica^{265,288,269}. Bajo esta premisa, se ha estudiado mediante *docking* ciego con Autodock Vina¹⁶², la región en la cual dichos compuestos presentan mayor afinidad de interacción. Posteriormente, se ha refinado la búsqueda de interacciones en dicha región mediante *docking* exhaustivo por subsitios (*subsites*) (Fig 2.5.4.). La región de CXCR4 que presenta mayores interacciones corresponde a la cavidad extracelular donde el ligando endógeno CXCL12 se une a CXCR4 y la función de *docking* utilizada para refinar dichas interacciones fue LigandFit¹⁶⁴. Esta función de *docking*, es la función de *docking* utilizada en los apartados 2.2, 2.3 y 2.4 del presente trabajo y que presenta mejor recuperación de activos de CXCR4 frente posibles falsos activos.

El análisis por *subsites* se realiza debido al gran volumen de la cavidad definida a explorar, 1520 Å³, en comparación a la estructura con mayor volumen de Van der Waals, así como por los indicios que apuntan a la existencia de sitios alostéricos en la cavidad extracelular de GPCRs²⁸⁵. De las 32 estructuras generadas (isómeros, protómeros y estereoisómeros) de AMD11070 y GSK812397, la que posee mayor volumen de Van der Waals tiene 438 Å³, siendo el volumen medio de las 32 estructuras de 402.7 Å³. La definición de *subsites* coincide con la realizada en el apartado 2.1, con la diferencia de que el volumen total definido de sitio activo se ha maximizado y por tanto también la de los *subsites* definidos (Fig 2.5.4.). El motivo de maximizar el sitio activo ha sido con el objetivo de abarcar todos los posibles *subsites* existentes adyacentes al sitio activo ortostérico.

Los resultados obtenidos mediante *docking* por subsitios indican que los subsitios parcialmente superpuestos 1-2 y 1-3, son los subsitios con mayor densidad de conformaciones estables. Especialmente, el subsitio 1-2 presenta la mejor conformación. Al realizar el *docking* de AMD11070 y GSK812397, en el PDB 3ODU (datos no publicados), éste presenta el mayor número de conformaciones ocupando dicho subsitio, por lo que globalmente se le considera el subsitio candidato a establecer interacciones de tipo alostérico. Cabe destacar sin embargo, que el subsitio 1-2 (2 particiones), incluye los subsitios 1-3 y 3-3 generados al establecer la tercera partición. Las diferencias entre conformaciones ubicadas en los subsitios 1-2 y 1-3 y entre 1-2 y 3-3 pueden indicar subsitios distintos parcialmente superpuestos. En este sentido, el ligando cristalizado isotiourea (IT1t), antagonista no alostérico, se halla ubicado dentro del sitio de unión conocido como *minor site* y que ocupa el subsitio definido como 3-3, lo que nos evidenciaría la existencia de mínimo 2 subsitios activos adyacentes o parcialmente superpuestos al sitio activo *minor*.

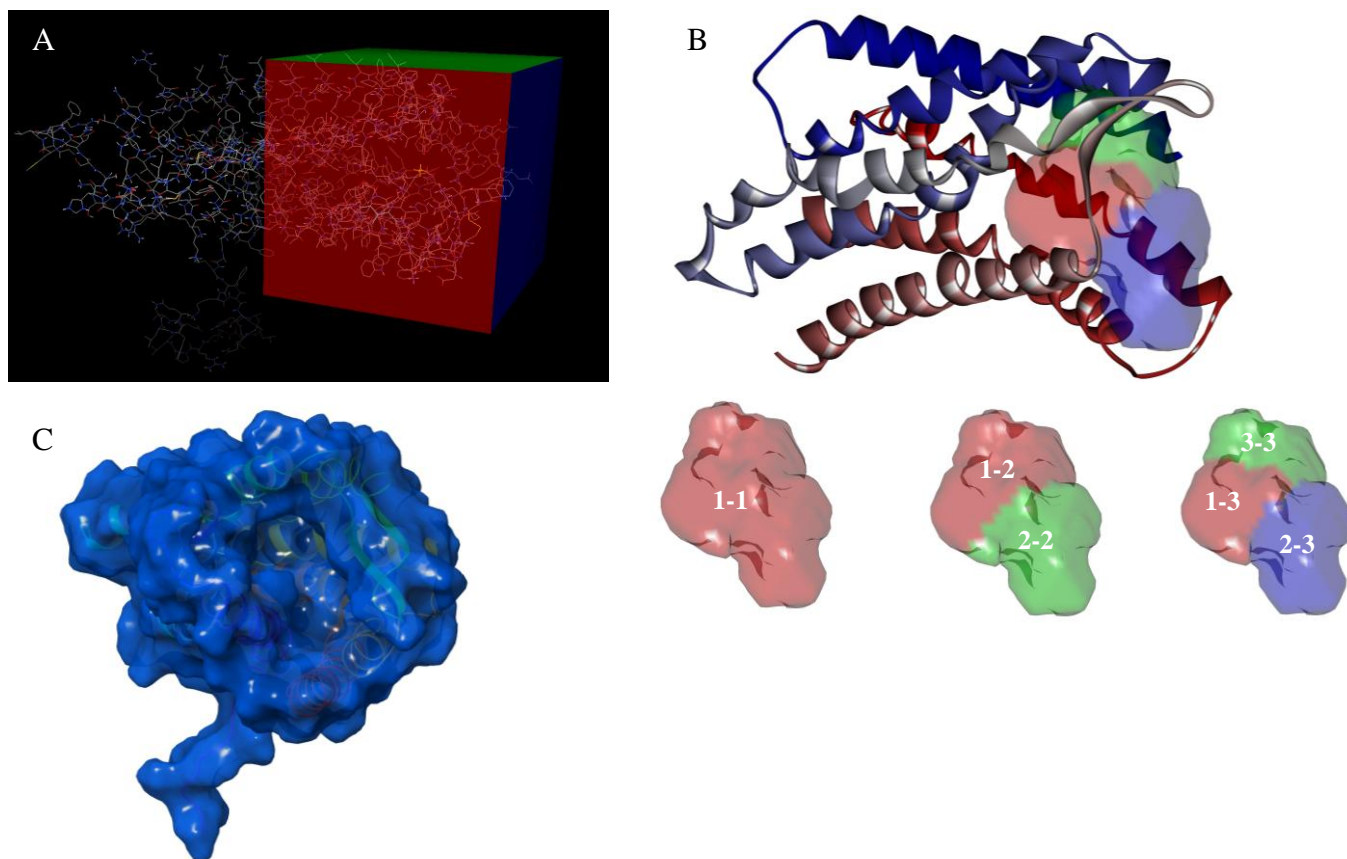


Fig. 2.5.4. **A.** Volumen (caja) que engloba las hélices de CXCR4 donde realizar el *docking* ciego mediante AutoDock Vina. El dominio proteico no incluido dentro de la caja del *docking* corresponde a la lisozima de fusión T4L^{234,235}. **B.** Representación del sitio de unión donde refinar las interacciones proteína-ligando mediante *docking* por subsitios. Se establecen 3 niveles de partición del sitio de unión (parte inferior de B) que generan un total de 6 subsitios: 1-1 (1ª partición); 1-2 y 2-2 (2ª partición) y 1-3; 2-3 y 3-3 (3ª partición). **C.** Visión frontal del dominio extracelular de CXCR4, donde se aprecia la cavidad principal de la proteína delimitada por los bucles extracelulares.

El análisis de las interacciones obtenidas tanto para AMD11070 como para GSK812397 en el subsitio 1-2, muestra a los residuos proteicos Asp187 y Glu288 de CXCR4, como los aminoácidos más determinantes a la hora de establecer interacciones (Fig 2.5.5.). Principalmente, Glu288 y las interacciones de carga, las cuales son clave para la proteína CXCR4 tal y como se ha determinado en los apartados anteriores 2.2 y 2.3.

Los residuos hidrofóbicos His203, Tyr255, His281 son también considerados importantes al delimitar y establecer interacciones con ambos compuestos, en el subsitio hidrofóbico propuesto (Fig 2.5.5. y Fig 2.5.6.).

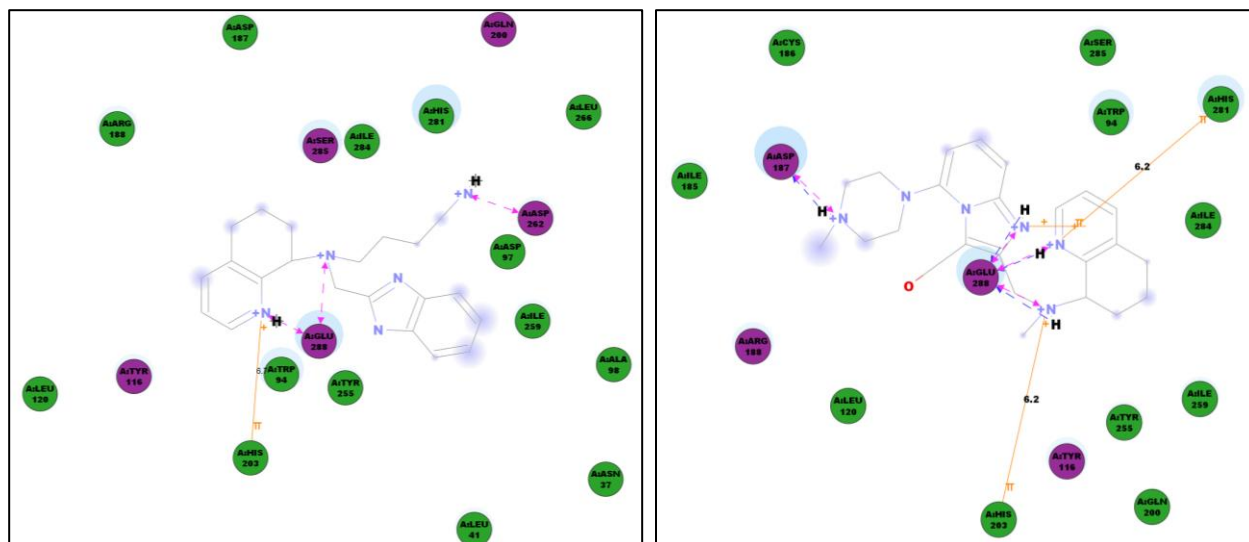


Fig. 2.5.5. Interacciones proteína-ligando establecidas en las conformaciones más estables de los compuestos orgánicos AMD11070 (izquierda) y GSK812397 (derecha) en el subsitio 1-2, al refinar interacciones mediante LigandFit.

Las interacciones π se muestran en color naranja (residuos His203 en AMD11070 así como His203 y His281 en GSK812397). Las interacciones de carga se muestran en líneas discontinuas con flecha a ambos lados y de color rosa (Asp262 y Glu288 en AMD11070 y Asp187 y Glu288 para GSK812397). Las interacciones de puentes de hidrógeno se muestran en líneas discontinuas de color azul con la flecha dirigida hacia el dador de electrones (Asp187 y Glu288 en GSK812397). Los residuos mostrados en círculos de color lila indican contribución polar, de carga o de puentes de hidrógeno, mientras que los residuos en color verde muestran contribución de tipo van der Waals. Los halos alrededor de los residuos proteicos (Asp94, Tyr116; His281, Ser285 y Glu288 en AMD11070 y Asp94, Tyr116, Asp187, Arg188, His281 y Glu288 en GSK812397) indican superficies accesibles a solventes.

Así pues, tal y como se argumenta en el artículo cuarto correspondiente a este capítulo, se propone la existencia de un sitio alostérico extracelular muy cercano al sitio activo ortostérico, en donde ambos sitios activos pueden compartir ciertos residuos proteicos a la hora de establecer interacciones con los respectivos ligandos, lo que indicaría cierta superposición de ambos sitios activos (Fig 2.5.6.). Esta proximidad y superposición de los sitios activos ortostérico y alostérico, caracterizaría la modulación antagonista del receptor. De hecho, el concepto de proximidad de sitios activos ortostérico y alostérico es propuesto por *Christopolous et al.*³⁰² Y por *Birdsall et al.*³⁰³ para los receptores muscarínicos y es planteado de forma genérica por *Schwartz et al.*⁸⁶ para moduladores ago-alostéricos⁸⁷.

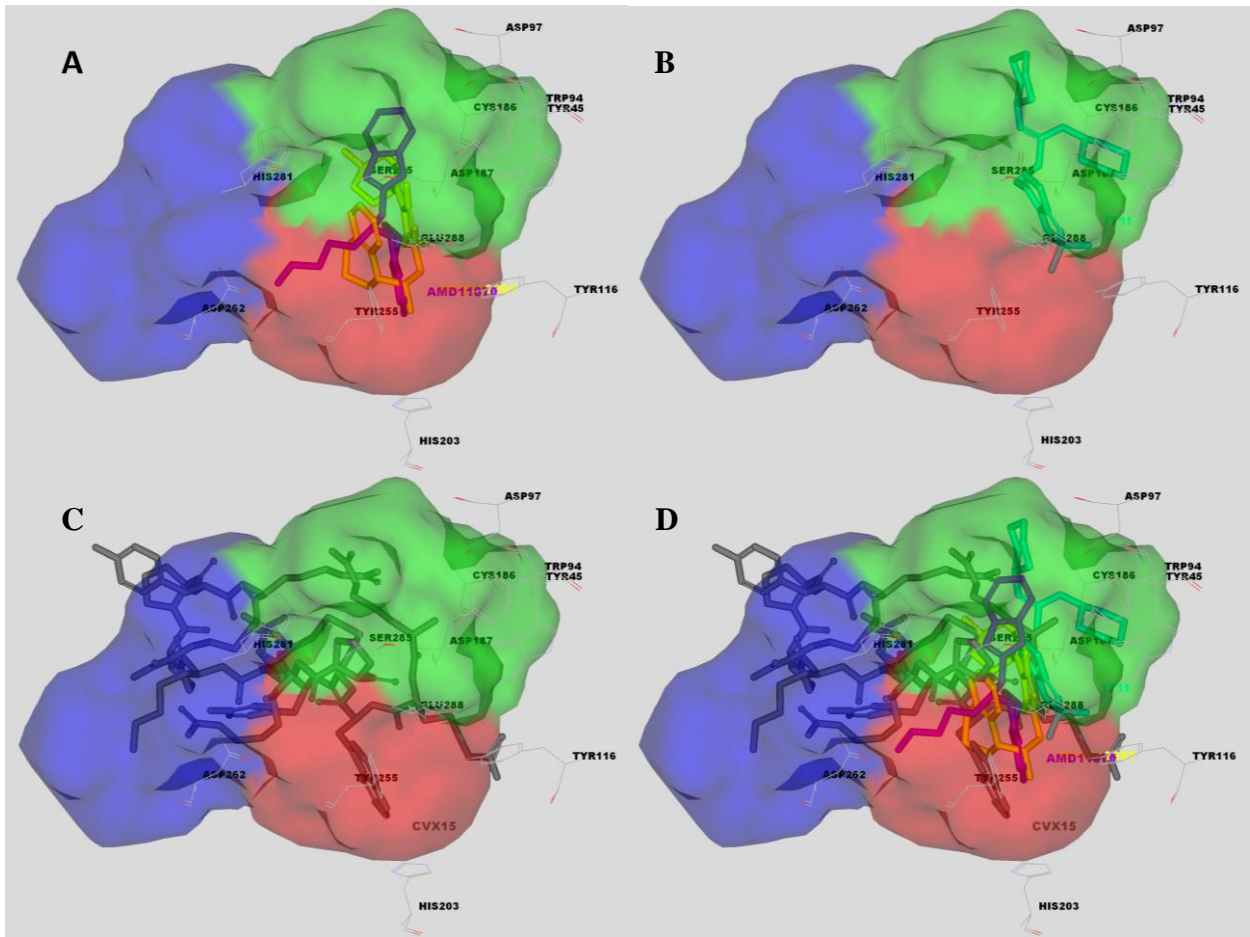


Fig. 2.5.6. A. Distribución de poses de AMD11070 y GSK812397 resultantes del *docking* en el subsitio 1-2. (verde y rojo) B. Ligando co-cristalizado IT1t que ocupa el subsitio 3-3 (verde) y que correspondería al sitio ortostérico conocido como *minor*. C. Ligando co-cristalizado CVX15 de naturaleza peptídica que ocupa el sitio activo llamado *major* y que ocupa la parte central de la cavidad que contiene los sitios activos, ocupando los subsitios 2-3 (azul) y parte del subsitio 1-2 (verde y rojo) superponiéndose al sitio activo alostérico. D. Superposición de los diferentes ligandos: AMD11070, GSK812397, IT1t y CVX15. E. Detalle de los 4 ligandos que se unen en los diferentes sitios activos de CXCR4. En verde la peptidina ATI-2341 en el sitio activo intracelular alostérico, en lila el péptido CVX15 ubicado en el *major site*, en amarillo el ligando IT1t ubicado en el *minor site* y en color azul y en verde oscuro los compuestos AMD11070 y GSK812397 ocupando el sitio alostérico superpuesto al sitio activo ortostérico.

Funcionalmente, podemos indicar que las pepducinas son agonistas alostéricos de CXCR4 y que al actuar en la parte intracelular del receptor, favorecerían la unión del receptor con las proteínas G activando la conformación de la proteína. Su potencial uso como fármacos se focaliza en tratamientos que requieran una mayor expresión de CXCR4, como puede ser la formación de vasos sanguíneos que se requiere en la regeneración de tejidos, mediante aumento de la angiogénesis. En cambio, AMD11070 y GSK812397 son inhibidores no competitivos de CXCR4-CXCL12, que inhiben negativamente dicha unión y por tanto actúan en la parte extracelular del receptor. El uso terapéutico de estos compuestos se focaliza en enfermedades que sobreexpresan CXCR4-CXCL12, como son la leucemia, la metástasis, los cánceres en general³⁰⁴ o los procesos inflamatorios³⁰⁵.

CONCLUSIONES

1. Se ha establecido un método de mejora de recuperación de activos frente inactivos en cribados virtuales para “dianas difíciles”, definidas como proteínas con amplios sitios activos, utilizando un protocolo de tres pasos: *docking*, cribado de las poses resultantes mediante un filtro farmacológico, y posterior ordenación de las poses mediante índice de similitud 2D de Tanimoto. El método mejora sensiblemente la pronta recuperación de activos frente a *decoys* y además permite la correcta predicción de modos de unión proteína-ligando tal y como se muestra para el caso de la “diana difícil” VEGFR-2, la cual une “*multiconformer binders*”. Cabe decir que si el sitio activo tiene un volumen significativamente grande como para albergar diferentes *subsitios* (como es el caso de CXCR4, la cual une “*multisite binders*”), es mejor aplicar el método de *docking* por subsitios (como si cada subsitio se tratase de un sitio activo diferente) que el protocolo de mejora establecido, ya que en el último de los pasos asume la existencia de una única molécula de referencia con la que comparar/medir similitud 2D.

2. Se identifica la estructura 3OE6, como la estructura más selectiva al discriminar moléculas activas de presuntas inactivas de CXCR4, al utilizar la función de *scoring* DockScore de LigandFit. Se ha determinado la estructura cristalizada de CXCR4 y la función de *docking* y *scoring* más selectivas para el cribado virtual *structure-based* de antagonistas de CXCR4. Mediante cribado virtual retrospectivo, ensayando un total de 14 funciones de *scoring* diferentes en 2 *test-set* de activos y presuntos inactivos, se han analizado las 5 estructuras publicadas de CXCR4 y un modelo creado por homología. La discriminación obtenida, devuelve valores elevados de pronta recuperación de activos y no diferencia los compuestos activos en función de las familias químicas a las que pertenecen.

3. Los métodos basados en la forma del ligando no recuperan tan bien los activos, de los inactivos. Esto es debido a la probable existencia de múltiples sitios y/o subsitios de unión en la cavidad principal de CXCR4. Se han comparado métodos de cribado virtual basados en la estructura (*docking*), con otros métodos de cribado virtual de antagonistas de CXCR4, basados en la forma tridimensional del ligando (*shape matching*) y también con modelos farmacofóricos generados a partir de las interacciones proteína-ligando, con el objetivo de proponer el mejor método de cribado virtual. Los cribados virtuales retrospectivos se han realizado utilizando un segundo *test-set* de inhibidores compilado para CXCR4. Los métodos basados exclusivamente en la forma del ligando presentan los peores resultados, dado que estos métodos se basan en la elección de una única forma tridimensional de referencia sobre la que superponer las moléculas de la base de datos. La existencia de múltiples sitios/subsitios de unión, permiten múltiples formas tridimensionales funcionalmente activas, diferentes a las de la forma de referencia establecida en el cálculo de los métodos *shape matching*.

4. Se mejora de la recuperación de activos frente inactivos al utilizar *docking* con post-filtrado farmacológico en el cribado virtual de antagonistas de CXCR4. Utilizando las conformaciones de los

activos devueltos por la función de *docking*, con mejor pronta recuperación de activos (*Early Recovery*) y mejor AUC, se ha construido un filtro farmacóforo para CXCR4. El farmacóforo mejora la pronta recuperación de activos y la AUC del *docking*, presentando al mismo tiempo una elevada sensibilidad y especificidad, frente a otros inhibidores de diferentes dianas anti-HIV.

5. Se ha generado una quimioteca de candidatos a inhibidores de CXCR4 y se han propuesto una serie de moléculas candidatas a ser potenciales inhibidores de CXCR4. Mediante la técnica de generación de quimiotecas virtuales se ha creado una quimioteca combinatoria de compuestos diamino derivados con el fin de realizar su posterior cribado prospectivo. Los compuestos han sido generados por combinación de diferentes núcleos, con diferentes subnúcleos y diferentes aminas terminales. La quimioteca resultante se ha filtrado y clasificado utilizando los mejores parámetros según el análisis retrospectivo de cribado virtual para antagonistas de CXCR4. La clasificación de compuestos se ha realizado utilizando diversos criterios de *scoring*, así como la combinación de ellos (*consensus scoring*).

6. Se ha determinado la región intracelular adyacente al ICL1 como un dominio de interacción entre el receptor CXCR4 y el péptido-derivado, pepducina ATI-2341. Se establecen las principales interacciones en el bucle ICL1 interaccionando también con el bucle ICL3 y en menor manera con el ICL2. El estudio de la modulación alostérica de agonistas peptomiméticos de CXCR4 y sus principales interacciones se ha desarrollado mediante *docking* ciego rotacional y traslacional (*ZDock*). Posteriormente se han refinado las interacciones alostéricas en dicho dominio intracelular utilizando diferentes métodos. Los métodos utilizados, con el objetivo de comparar y predecir las interacciones alostéricas establecidas más relevantes, son *flexible docking* (FlexPepDock y HADDOCK) así como dinámica molecular.

7. Se han establecido las interacciones de los compuestos orgánicos AMD11070 y GSK812397, cuyos modos de acción indican ser potenciales moduladores alostéricos de CXCR4. En primer lugar se establecen las regiones de mayor afinidad de interacción, mediante *docking* ciego con AutoDock Vina. Posteriormente, para la región de mayor probabilidad, que corresponde a la cavidad extracelular donde se une CXCL12, se han refinado y determinado las interacciones mediante *docking* exhaustivo por subsitios (LigandFit).

8. Del análisis de las interacciones de CXCR4 con la pepducina ATI-2341, AMD11070 y GSK812397 se propone la existencia de 4 posibles sitios activos en CXCR4: un sitio alostérico en el dominio intracelular del receptor, un sitio activo extracelular conocido como *major binding site* donde aparece unido el péptido co-cristalizado CVX15, un segundo sitio activo extracelular conocido como *minor binding site* donde se une el ligando co-cristalizado IT1t y un cuarto subsitio activo entre el *major binding site* y el *minor binding site* y solapado a este último, que podría actuar como sitio activo alostérico y donde se unirían los antagonistas AMD11070 y GSK812397 cuyo comportamiento se podría semejar al de compuestos ago-alostéricos.

REFERENCIAS

- (1) Parascandola, J. The Theoretical Basis of Paul Ehrlich's Chemotherapy. *J. Hist. Med. Allied Sci.* **1981**, *36*, 19–43.
- (2) Clementi, E. Ab Initio Computational Chemistry. *J. Phys. Chem.* **1985**, *89*, 4426–4436.
- (3) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. An All Atom Force Field for Simulations of Proteins and Nucleic Acids. *J. Comput. Chem.* **1986**, *7*, 230–252.
- (4) Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S. The Protein Data Bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2002**, *58*, 899–907.
- (5) Emmert, D. B.; Stoehr, P. J.; Stoesser, G.; Cameron, G. N. The European Bioinformatics Institute (EBI) Databases. *Nucleic Acids Res.* **1994**, *22*, 3445.
- (6) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein–ligand Binding Affinities. *Nucleic Acids Res.* **2006**, *35*, D198.
- (7) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
- (8) Merino, A.; Bronowska, A. K.; Jackson, D. B.; Cahill, D. J. Drug Profiling: Knowing Where It Hits. *Drug Discov. Today* **2010**, *15*, 749–756.
- (9) Collier, R. Drug Development Cost Estimates Hard to Swallow. *CMAJ* **2009**, *180*, 279–280.
- (10) Bennani, Y. L. Drug Discovery in the next Decade: Innovation Needed ASAP. *Drug Discov. Today* **2012**, *17 Suppl*, S31–S44.
- (11) Young, D. C. *Computational Drug Design: A Guide for Computational and Medicinal Chemists*; 2009.
- (12) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **2012**, *64*, 4–17.
- (13) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
- (14) Bemis, G. W.; Murcko, M. A. Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095–5099.
- (15) Hansch, C.; Sannes, P. C.; Taylor, J. B. *Comprehensive Medicinal Chemistry Vol. 6*, 1990.
- (16) Seidler, J.; McGovern, S. L.; Doman, T. N.; Shoichet, B. K. Identification and Prediction of Promiscuous Aggregating Inhibitors among Known Drugs. *J. Med. Chem.* **2003**, *46*, 4477–4486.
- (17) Pérez-Nueno, V. I.; Ritchie, D. W. Identifying and Characterizing Promiscuous Targets: Implications for Virtual Screening. *Expert Opin. Drug Discov.* **2012**, *7*, 1–17.
- (18) Mestres, J.; Gregori-Puigjané, E. The Topology of Drug–target Interaction Networks: Implicit Dependence on Drug Properties and Target Families. *Mol. Biosyst.* **2009**, *5*, 1051–1057.
- (19) Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A.; Gocayne, J. D.; Amanatides, P.; Ballew, R. M.; Huson, D. H.; Wortman, J. R.; Zhang, Q.; Kodira, C. D.; Zheng, X. H.; Chen, L.; Skupski, M.; Subramanian, G.; Thomas, P. D.; Zhang, J.; Gabor Miklos, G. L.; Nelson, C.; Broder, S.; Clark, A. G.; Nadeau, J.; McKusick, V. A.; Zinder, N.; Levine, A. J.; Roberts, R. J.; Simon, M.; Slayman, C.; Hunkapiller, M.; Bolanos, R.; Delcher, A.; Dew, I.; Fasulo, D.; Flanigan, M.; Florea, L.; Halpern, A.; Hannenhalli, S.; Kravitz, S.; Levy, S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Threideh, J.; Beasley, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramouliswaran, I.; Charlab, R.; Chaturvedi, K.; Deng, Z.; Di Francesco, V.; Dunn, P.; Eilbeck, K.; Evangelista, C.; Gabrielian, A. E.; Gan, W.; Ge, W.; Gong, F.; Gu, Z.; Guan, P.; Heiman, T. J.; Higgins, M. E.; Ji, R. R.; Ke, Z.; Ketchum, K. A.; Lai, Z.; Lei, Y.; Li, Z.; Li, J.; Liang, Y.; Lin, X.; Lu, F.; Merkulov, G. V.; Milshina, N.; Moore, H. M.; Naik, A. K.; Narayan, V. A.; Neelam, B.; Nusskern, D.; Rusch, D. B.; Salzberg, S.; Shao, W.; Shue, B.; Sun, J.; Wang, Z.; Wang, A.; Wang, X.; Wang, J.; Wei, M.; Wides, R.; Xiao, C.; Yan, C.; Yao, A.; Ye, J.; Zhan, M.; Zhang, W.; Zhang, H.; Zhao, Q.; Zheng, L.; Zhong, F.; Zhong, W.; Zhu, S.; Zhao, S.; Gilbert, D.; Baumhueter, S.; Spier,

- G.; Carter, C.; Cravchik, A.; Woodage, T.; Ali, F.; An, H.; Awe, A.; Baldwin, D.; Baden, H.; Barnstead, M.; Barrow, I.; Beeson, K.; Busam, D.; Carver, A.; Center, A.; Cheng, M. L.; Curry, L.; Danaher, S.; Davenport, L.; Desilets, R.; Dietz, S.; Dodson, K.; Doup, L.; Ferreira, S.; Garg, N.; Gluecksmann, A.; Hart, B.; Haynes, J.; Haynes, C.; Heiner, C.; Hladun, S.; Hostin, D.; Houck, J.; Howland, T.; Ibegwam, C.; Johnson, J.; Kalush, F.; Kline, L.; Koduru, S.; Love, A.; Mann, F.; May, D.; McCawley, S.; McIntosh, T.; McMullen, I.; Moy, M.; Moy, L.; Murphy, B.; Nelson, K.; Pfannkoch, C.; Pratts, E.; Puri, V.; Qureshi, H.; Reardon, M.; Rodriguez, R.; Rogers, Y. H.; Romblad, D.; Ruhfel, B.; Scott, R.; Sitter, C.; Smallwood, M.; Stewart, E.; Strong, R.; Suh, E.; Thomas, R.; Tint, N. N.; Tse, S.; Vech, C.; Wang, G.; Wetter, J.; Williams, S.; Williams, M.; Windsor, S.; Winn-Deen, E.; Wolfe, K.; Zaveri, J.; Zaveri, K.; Abril, J. F.; Guigó, R.; Campbell, M. J.; Sjolander, K. V.; Karlak, B.; Kejariwal, A.; Mi, H.; Lazareva, B.; Hatton, T.; Narechania, A.; Diemer, K.; Muruganujan, A.; Guo, N.; Sato, S.; Bafna, V.; Istrail, S.; Lippert, R.; Schwartz, R.; Walenz, B.; Yooseph, S.; Allen, D.; Basu, A.; Baxendale, J.; Blick, L.; Caminha, M.; Carnes-Stine, J.; Caulk, P.; Chiang, Y. H.; Coyne, M.; Dahlke, C.; Mays, A.; Dombroski, M.; Donnelly, M.; Ely, D.; Esparham, S.; Fosler, C.; Gire, H.; Glanowski, S.; Glasser, K.; Glodek, A.; Gorokhov, M.; Graham, K.; Gropman, B.; Harris, M.; Heil, J.; Henderson, S.; Hoover, J.; Jennings, D.; Jordan, C.; Jordan, J.; Kasha, J.; Kagan, L.; Kraft, C.; Levitsky, A.; Lewis, M.; Liu, X.; Lopez, J.; Ma, D.; Majoros, W.; McDaniel, J.; Murphy, S.; Newman, M.; Nguyen, T.; Nguyen, N.; Nodell, M.; Pan, S.; Peck, J.; Peterson, M.; Rowe, W.; Sanders, R.; Scott, J.; Simpson, M.; Smith, T.; Sprague, A.; Stockwell, T.; Turner, R.; Venter, E.; Wang, M.; Wen, M.; Wu, D.; Wu, M.; Xia, A.; Zandieh, A.; Zhu, X. The Sequence of the Human Genome. *Science* **2001**, *291*, 1304–1351.
- (20) Usón, I.; Sheldrick, G. M. Advances in Direct Methods for Protein Crystallography. *Curr. Opin. Struct. Biol.* **1999**, *9*, 643–648.
- (21) Clore, G. M.; Schwieters, C. D. Theoretical and Computational Advances in Biomolecular NMR Spectroscopy. *Curr. Opin. Struct. Biol.* **2002**, *12*, 146–153.
- (22) Hillisch, A.; Pineda, L. F.; Hilgenfeld, R. Utility of Homology Models in the Drug Discovery Process. *Drug Discov. Today* **2004**, *9*, 659–669.
- (23) Kalyaanamoorthy, S.; Chen, Y.-P. P. Structure-Based Drug Design to Augment Hit Discovery. *Drug Discov. Today* **2011**, *16*, 831–839.
- (24) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
- (25) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455–1474.
- (26) Hajduk, P. J.; Greer, J. A Decade of Fragment-Based Drug Design: Strategic Advances and Lessons Learned. *Nat. Rev. Drug Discov.* **2007**, *6*, 211–219.
- (27) Mannhold, R., Kubinyi, H., & Folkers, G. *Ligand Design for G Protein-Coupled Receptors*; John Wiley & Sons, 2006.
- (28) Sheridan, R. P.; Kearsley, S. K. Why Do We Need so Many Chemical Similarity Search Methods? *Drug Discov. Today* **2002**, *7*, 903–911.
- (29) Wermuth, C.-G.; Robin Ganellin, C.; Lindberg, P.; Mitscher, L. A. Chapter 36 – Glossary of Terms Used in Medicinal Chemistry (IUPAC Recommendations 1997). In *Annual Reports in Medicinal Chemistry*; 1998; Vol. 33, pp. 385–395.
- (30) Cramer, R. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
- (31) Oberlin, E.; Amara, A.; Bachelier, F.; Bessia, C.; Virelizier, J. L.; Arenzana-Seisdedos, F.; Schwartz, O.; Heard, J. M.; Clark-Lewis, I.; Legler, D. F.; Loetscher, M.; Baggiolini, M.; Moser, B. The CXC Chemokine SDF-1 Is the Ligand for LESTR/fusin and Prevents Infection by T-Cell-Line-Adapted HIV-1. *Nature* **1996**, *382*, 833–835.
- (32) Endres, M. J.; Clapham, P. R.; Marsh, M.; Ahuja, M.; Turner, J. D.; McKnight, A.; Thomas, J. F.; Stoebenau-Haggarty, B.; Choe, S.; Vance, P. J.; Wells, T. N. .; Power, C. A.; Sutterwala, S. S.; Doms, R. W.; Landau, N. R.; Hoxie, J. A. CD4-Independent Infection by HIV-2 Is Mediated by Fusin/CXCR4. *Cell* **1996**, *87*, 745–756.
- (33) Schlyer, S.; Horuk, R. I Want a New Drug: G-Protein-Coupled Receptors in Drug Development. *Drug Discov. Today* **2006**, *11*, 481–493.
- (34) Stevens, R. C.; Cherezov, V.; Katritch, V.; Abagyan, R.; Kuhn, P.; Rosen, H.; Wüthrich, K. The GPCR Network: A Large-Scale Collaboration to Determine Human GPCR Structure and Function. *Nat. Rev. Drug Discov.* **2013**, *12*, 25–34.

- (35) Vogt, I.; Mestres, J. Drug-Target Networks. *Mol. Inform.* **2010**, *29*, 10–14.
- (36) Mirzadegan, T.; Benkö, G.; Filipek, S.; Palczewski, K. Sequence Analyses of G-Protein-Coupled Receptors: Similarities to Rhodopsin. *Biochemistry* **2003**, *42*, 2759–2767.
- (37) Attwood, T. K.; Findlay, J. B. C. Fingerprinting G-Protein-Coupled Receptors. *Protein Eng. Des. Sel.* **1994**, *7*, 195.
- (38) Kolakowski Jr, L. F. GCRDb: A G-Protein-Coupled Receptor Database. *Receptors Channels* **1994**, *2*, 1–7.
- (39) Horn, F.; Weare, J.; Beukers, M. W.; Horsch, S.; Bairoch, A.; Chen, W.; Edvardsen, O.; Campagne, F.; Vriend, G. GPCRDB: An Information System for G Protein-Coupled Receptors. *Nucleic Acids Res.* **1998**, *26*, 275.
- (40) Suwa, M.; Sugihara, M.; Ono, Y. Functional and Structural Overview of G-Protein-Coupled Receptors Comprehensively Obtained from Genome Sequences. *Pharmaceuticals* **2011**, *4*, 652–664.
- (41) Adams, R.; Worth, C. L.; Guenther, S.; Dunkel, M.; Lehmann, R.; Preissner, R. Binding Sites in Membrane Proteins-Diversity, Druggability and Prospects. *Eur. J. Cell Biol.* **2012**, *91*, 326–339.
- (42) Fredriksson, R.; Lagerstrom, M. C.; Lundin, L. G.; Schioth, H. B. The G-Protein-Coupled Receptors in the Human Genome Form Five Main Families. Phylogenetic Analysis, Paralogue Groups, and Fingerprints. *Mol. Pharmacol.* **2003**, *63*, 1256.
- (43) Schiöth, H. B.; Fredriksson, R. The GRAFS Classification System of G-Protein Coupled Receptors in Comparative Perspective. *Gen. Comp. Endocrinol.* **2005**, *142*, 94–101.
- (44) Joost, P.; Methner, A. Phylogenetic Analysis of 277 Human G-Protein-Coupled Receptors as a Tool for the Prediction of Orphan Receptor Ligands. *Genome Biol* **2002**, *3*, 1–16.
- (45) De March, C. A.; Golebiowski, J. A Computational Microscope Focused on the Sense of Smell. *Biochimie* **2014**, *107 Pt A*, 3–10.
- (46) Bockaert, J.; Pin, J. P. Molecular Tinkering of G Protein-Coupled Receptors: An Evolutionary Success. *EMBO J.* **1999**, *18*, 1723–1729.
- (47) Congreve, M.; Langmead, C. J.; Mason, J. S.; Marshall, F. H. Progress in Structure Based Drug Design for G Protein-Coupled Receptors. *J. Med. Chem.* **2011**, *54*, 4283–4311.
- (48) Bräuner-Osborne, H.; Wellendorph, P.; Jensen, A. a. Structure, Pharmacology and Therapeutic Prospects of Family C G-Protein Coupled Receptors. *Curr. Drug Targets* **2007**, *8*, 169–184.
- (49) Behrens, M.; Foerster, S.; Staehler, F.; Raguse, J.-D.; Meyerhof, W. Gustatory Expression Pattern of the Human TAS2R Bitter Receptor Gene Family Reveals a Heterogenous Population of Bitter Responsive Taste Receptor Cells. *J. Neurosci.* **2007**, *27*, 12630–12640.
- (50) Tan, J.; Abrol, R.; Trzaskowski, B.; Goddard, W. a. 3D Structure Prediction of TAS2R38 Bitter Receptors Bound to Agonists Phenylthiocarbamide (PTC) and 6-N-Propylthiouracil (PROP). *J. Chem. Inf. Model.* **2012**, *52*, 1875–1885.
- (51) Wang, H.; Liu, T.; Malbon, C. C. Structure-Function Analysis of Frizzleds. *Cell. Signal.* **2006**, *18*, 934–941.
- (52) Huang, H.-C.; Klein, P. S. The Frizzled Family: Receptors for Multiple Signal Transduction Pathways. *Genome Biol.* **2004**, *5*, 234.
- (53) Raisley, B.; Zhang, M.; Hereld, D.; Hadwiger, J. A. A cAMP Receptor-like G Protein-Coupled Receptor with Roles in Growth Regulation and Development. *Dev. Biol.* **2004**, *265*, 433–445.
- (54) Shonberg, J.; Kling, R. C.; Gmeiner, P.; Löber, S. GPCR Crystal Structures: Medicinal Chemistry in the Pocket. *Bioorg. Med. Chem.* **2014**, *23*, 3880–3906.
- (55) Wu, B.; Chien, E. Y. T.; Mol, C. D.; Fenalti, G.; Liu, W.; Katritch, V.; Abagyan, R.; Brooun, A.; Wells, P.; Bi, F. C. Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists. *Science (80-.)*. **2010**, *330*, 1066–1071.

- (56) Warne, T.; Serrano-Vega, M. J.; Baker, J. G.; Moukhametzianov, R.; Edwards, P. C.; Henderson, R.; Leslie, A. G. W.; Tate, C. G.; Schertler, G. F. X. Structure of a beta1-Adrenergic G-Protein-Coupled Receptor. *Nature* **2008**, *454*, 486–491.
- (57) Rasmussen, S. G. F.; Choi, H.-J.; Rosenbaum, D. M.; Kobilka, T. S.; Thian, F. S.; Edwards, P. C.; Burghammer, M.; Ratnala, V. R. P.; Sanishvili, R.; Fischetti, R. F.; Schertler, G. F. X.; Weis, W. I.; Kobilka, B. K. Crystal Structure of the Human beta2 Adrenergic G-Protein-Coupled Receptor. *Nature* **2007**, *450*, 383–387.
- (58) Jaakola, V.-P.; Griffith, M. T.; Hanson, M. A.; Cherezov, V.; Chien, E. Y. T.; Lane, J. R.; IJzerman, A. P.; Stevens, R. C. The 2.6 Angstrom Crystal Structure of a Human A2A Adenosine Receptor Bound to an Antagonist. *Science (80-.)*. **2008**, *322*, 1211–1217.
- (59) Chien, E. Y. T.; Liu, W.; Zhao, Q.; Katritch, V.; Han, G. W.; Hanson, M. A.; Shi, L.; Newman, A. H.; Javitch, J. A.; Cherezov, V.; Stevens, R. C. Structure of the Human Dopamine D3 Receptor in Complex with a D2/D3 Selective Antagonist. *Science* **2010**, *330*, 1091–1095.
- (60) Haga, K.; Kruse, A. C.; Asada, H.; Yurugi-Kobayashi, T.; Shiroishi, M.; Zhang, C.; Weis, W. I.; Okada, T.; Kobilka, B. K.; Haga, T.; Kobayashi, T. Structure of the Human M2 Muscarinic Acetylcholine Receptor Bound to an Antagonist. *Nature* **2012**, *482*, 547–551.
- (61) Kruse, A. C.; Hu, J.; Pan, A. C.; Arlow, D. H.; Rosenbaum, D. M.; Rosemond, E.; Green, H. F.; Liu, T.; Chae, P. S.; Dror, R. O.; Shaw, D. E.; Weis, W. I.; Wess, J.; Kobilka, B. K. Structure and Dynamics of the M3 Muscarinic Acetylcholine Receptor. *Nature* **2012**, *482*, 552–556.
- (62) De Graaf, C.; Kooistra, A. J.; Vischer, H. F.; Katritch, V.; Kuijter, M.; Shiroishi, M.; Iwata, S.; Shimamura, T.; Stevens, R. C.; de Esch, I. J. P.; Leurs, R. Crystal Structure-Based Virtual Screening for Fragment-like Ligands of the Human Histamine H(1) Receptor. *J. Med. Chem.* **2011**, *54*, 8195–8206.
- (63) Beuming, T.; Sherman, W. Current Assessment of Docking into GPCR Crystal Structures and Homology Models: Successes, Challenges, and Guidelines. *J. Chem. Inf. Model.* **2012**, *52*, 3263–3277.
- (64) Thompson, A. A.; Liu, W.; Chun, E.; Katritch, V.; Wu, H.; Vardy, E.; Huang, X.-P.; Trapella, C.; Guerrini, R.; Calo, G.; Roth, B. L.; Cherezov, V.; Stevens, R. C. Structure of the Nociceptin/orphanin FQ Receptor in Complex with a Peptide Mimetic. *Nature* **2012**, *485*, 395–399.
- (65) Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A.; Trong, I. L.; Teller, D. C.; Okada, T.; Stenkamp, R. E. Crystal Structure of Rhodopsin: AG Protein-Coupled Receptor. *Science (80-.)*. **2000**, *289*, 739.
- (66) Zhang, C.; Srinivasan, Y.; Arlow, D. H.; Fung, J. J.; Palmer, D.; Zheng, Y.; Green, H. F.; Pandey, A.; Dror, R. O.; Shaw, D. E.; Weis, W. I.; Coughlin, S. R.; Kobilka, B. K. High-Resolution Crystal Structure of Human Protease-Activated Receptor 1. *Nature* **2012**, *492*, 387–392.
- (67) Hanson, M. A.; Roth, C. B.; Jo, E.; Griffith, M. T.; Scott, F. L.; Reinhart, G.; Desale, H.; Clemons, B.; Cahalan, S. M.; Schuerer, S. C.; Sanna, M. G.; Han, G. W.; Kuhn, P.; Rosen, H.; Stevens, R. C. Crystal Structure of a Lipid G Protein-Coupled Receptor. *Science* **2012**, *335*, 851–855.
- (68) Wang, C.; Jiang, Y.; Ma, J.; Wu, H.; Wacker, D.; Katritch, V.; Han, G. W.; Liu, W.; Huang, X.-P.; Vardy, E.; McCorvy, J. D.; Gao, X.; Zhou, X. E.; Melcher, K.; Zhang, C.; Bai, F.; Yang, H.; Yang, L.; Jiang, H.; Roth, B. L.; Cherezov, V.; Stevens, R. C.; Xu, H. E. Structural Basis for Molecular Recognition at Serotonin Receptors. *Science* **2013**, *340*, 610–614.
- (69) Wacker, D.; Wang, C.; Katritch, V.; Han, G. W.; Huang, X.-P.; Vardy, E.; McCorvy, J. D.; Jiang, Y.; Chu, M.; Siu, F. Y.; Liu, W.; Xu, H. E.; Cherezov, V.; Roth, B. L.; Stevens, R. C. Structural Features for Functional Selectivity at Serotonin Receptors. *Science* **2013**, *340*, 615–619.
- (70) Tan, Q.; Zhu, Y.; Li, J.; Chen, Z.; Han, G. W.; Kufareva, I.; Li, T.; Ma, L.; Fenalti, G.; Li, J.; Zhang, W.; Xie, X.; Yang, H.; Jiang, H.; Cherezov, V.; Liu, H.; Stevens, R. C.; Zhao, Q.; Wu, B. Structure of the CCR5 Chemokine Receptor-HIV Entry Inhibitor Maraviroc Complex. *Science* **2013**, *341*, 1387–1390.
- (71) White, J. F.; Noinaj, N.; Shibata, Y.; Love, J.; Kloss, B.; Xu, F.; Gvozdenovic-Jeremic, J.; Shah, P.; Shiloach, J.; Tate, C. G.; Grisshammer, R. Structure of the Agonist-Bound Neurotensin Receptor. *Nature* **2012**, *490*, 508–513.
- (72) Zhang, K.; Zhang, J.; Gao, Z.-G.; Zhang, D.; Zhu, L.; Han, G. W.; Moss, S. M.; Paoletta, S.; Kiselev, E.; Lu, W.; Fenalti, G.; Zhang, W.; Müller, C. E.; Yang, H.; Jiang, H.; Cherezov, V.; Katritch, V.; Jacobson, K. A.; Stevens, R. C.; Wu, B.; Zhao, Q. Structure of the Human P2Y12 Receptor in Complex with an Antithrombotic Drug. *Nature* **2014**, *509*, 115–118.

- (73) Srivastava, A.; Yano, J.; Hirozane, Y.; Kefala, G.; Gruswitz, F.; Snell, G.; Lane, W.; Ivetac, A.; Aertgeerts, K.; Nguyen, J.; Jennings, A.; Okada, K. High-Resolution Structure of the Human GPR40 Receptor Bound to Allosteric Agonist TAK-875. *Nature* **2014**, *513*, 124–127.
- (74) Hollenstein, K.; Kean, J.; Bortolato, A.; Cheng, R. K. Y.; Doré, A. S.; Jazayeri, A.; Cooke, R. M.; Weir, M.; Marshall, F. H. Structure of Class B GPCR Corticotropin-Releasing Factor Receptor 1. *Nature* **2013**, *499*, 438–443.
- (75) Siu, F. Y.; He, M.; de Graaf, C.; Han, G. W.; Yang, D.; Zhang, Z.; Zhou, C.; Xu, Q.; Wacker, D.; Joseph, J. S.; Liu, W.; Lau, J.; Cherezov, V.; Katritch, V.; Wang, M.-W.; Stevens, R. C. Structure of the Human Glucagon Class B G-Protein-Coupled Receptor. *Nature* **2013**, *499*, 444–449.
- (76) Kunishima, N.; Shimada, Y.; Tsuji, Y.; Sato, T.; Yamamoto, M.; Kumasaka, T.; Nakanishi, S.; Jingami, H.; Morikawa, K. Structural Basis of Glutamate Recognition by a Dimeric Metabotropic Glutamate Receptor. *Nature* **2000**, *407*, 971–977.
- (77) Wu, H.; Wang, C.; Gregory, K. J.; Han, G. W.; Cho, H. P.; Xia, Y.; Niswender, C. M.; Katritch, V.; Meiler, J.; Cherezov, V.; Conn, P. J.; Stevens, R. C. Structure of a Class C GPCR Metabotropic Glutamate Receptor 1 Bound to an Allosteric Modulator. *Science* **2014**, *344*, 58–64.
- (78) Muto, T.; Tsuchiya, D.; Morikawa, K.; Jingami, H. Structures of the Extracellular Regions of the Group II/III Metabotropic Glutamate Receptors. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 3759–3764.
- (79) Doré, A. S.; Okrasa, K.; Patel, J. C.; Serrano-Vega, M.; Bennett, K.; Cooke, R. M.; Errey, J. C.; Jazayeri, A.; Khan, S.; Tehan, B.; Weir, M.; Wiggin, G. R.; Marshall, F. H. Structure of Class C GPCR Metabotropic Glutamate Receptor 5 Transmembrane Domain. *Nature* **2014**, *511*, 557–562.
- (80) Wang, C.; Wu, H.; Katritch, V.; Han, G. W.; Huang, X.-P.; Liu, W.; Siu, F. Y.; Roth, B. L.; Cherezov, V.; Stevens, R. C. Structure of the Human Smoothed Receptor Bound to an Antitumour Agent. *Nature* **2013**, *497*, 338–343.
- (81) Wang, C.; Wu, H.; Evron, T.; Vardy, E.; Han, G. W.; Huang, X.-P.; Hufeisen, S. J.; Mangano, T. J.; Urban, D. J.; Katritch, V.; Cherezov, V.; Caron, M. G.; Roth, B. L.; Stevens, R. C. Structural Basis for Smoothed Receptor Modulation and Chemoresistance to Anticancer Drugs. *Nat. Commun.* **2014**, *5*, 4355.
- (82) Weierstall, U.; James, D.; Wang, C.; White, T. A.; Wang, D.; Liu, W.; Spence, J. C. H.; Bruce Doak, R.; Nelson, G.; Fromme, P.; Fromme, R.; Grotjohann, I.; Kupitz, C.; Zatsepin, N. A.; Liu, H.; Basu, S.; Wacker, D.; Han, G. W.; Katritch, V.; Boutet, S.; Messerschmidt, M.; Williams, G. J.; Koglin, J. E.; Marvin Seibert, M.; Klinker, M.; Gati, C.; Shoeman, R. L.; Barty, A.; Chapman, H. N.; Kirian, R. A.; Beyerlein, K. R.; Stevens, R. C.; Li, D.; Shah, S. T. A.; Howe, N.; Caffrey, M.; Cherezov, V. Lipidic Cubic Phase Injector Facilitates Membrane Protein Serial Femtosecond Crystallography. *Nat. Commun.* **2014**, *5*, 3309.
- (83) Granier, S.; Kobilka, B. A New Era of GPCR Structural and Chemical Biology. *Nat. Chem. Biol.* **2012**, *8*, 670–673.
- (84) Congreve, M.; Marshall, F. The Impact of GPCR Structures on Pharmacology and Structure-Based Drug Design. *Br. J. Pharmacol.* **2010**, *159*, 986–996.
- (85) May, L. T.; Leach, K.; Sexton, P. M.; Christopoulos, A. Allosteric Modulation of G Protein-coupled Receptors. *Annu. Rev. Pharmacol. Toxicol.* **2007**, *47*, 1–51.
- (86) Schwartz, T. W.; Holst, B. Allosteric Enhancers, Allosteric Agonists and Ago-Allosteric Modulators: Where Do They Bind and How Do They Act? *Trends Pharmacol. Sci.* **2007**, *28*, 366–373.
- (87) Smith, N. J.; Bennett, K. A.; Milligan, G. When Simple Agonism Is Not Enough: Emerging Modalities of GPCR Ligands. *Mol. Cell. Endocrinol.* **2011**, *331*, 241–247.
- (88) Kamal, M.; Jockers, R. Bitopic Ligands: All-in-One Orthosteric and Allosteric. *FI000 Biol. Rep.* **2009**, *1*, 77.
- (89) Lane, J. R.; Sexton, P. M.; Christopoulos, A. Bridging the Gap: Bitopic Ligands of G-Protein-Coupled Receptors. *Trends Pharmacol. Sci.* **2013**, *34*, 59–66.
- (90) Hiller, C.; Kühhorn, J.; Gmeiner, P. Class A G-Protein-Coupled Receptor (GPCR) Dimers and Bivalent Ligands. *J. Med. Chem.* **2013**, *56*, 6542–6559.
- (91) Xu, Y.; Duggineni, S.; Espitia, S.; Richman, D. D.; An, J.; Huang, Z. A Synthetic Bivalent Ligand of CXCR4 Inhibits HIV Infection. *Biochem. Biophys. Res. Commun.* **2013**, *435*, 646–650.

- (92) Rominger, D. H.; Cowan, C. L.; Gowen-MacDonald, W.; Violin, J. D. Biased Ligands: Pathway Validation for Novel GPCR Therapeutics. *Curr. Opin. Pharmacol.* **2014**, *16*, 108–115.
- (93) Shonberg, J.; Lopez, L.; Scammells, P. J.; Christopoulos, A.; Capuano, B.; Lane, J. R. Biased Agonism at G Protein-Coupled Receptors: The Promise and the Challenges—a Medicinal Chemistry Perspective. *Med. Res. Rev.* **2014**, *34*, 1286–1330.
- (94) Levoye, A.; Balabanian, K.; Baleux, F.; Bachelier, F.; Lagane, B. CXCR7 Heterodimerizes with CXCR4 and Regulates CXCL12-Mediated G Protein Signaling. *Blood* **2009**, *113*, 6085–6093.
- (95) Nagasawa, T.; Nakajima, T.; Tachibana, K.; Iizasa, H.; Bleul, C. C.; Yoshie, O.; Matsushima, K.; Yoshida, N.; Springer, T. A.; Kishimoto, T. Molecular Cloning and Characterization of a Murine Pre-B-Cell Growth-Stimulating Factor/stromal Cell-Derived Factor 1 Receptor, a Murine Homolog of the Human Immunodeficiency Virus 1 Entry Coreceptor Fusin. *Proc. Natl. Acad. Sci.* **1996**, *93*, 14726–14729.
- (96) Yu, L.; Cecil, J.; Peng, S.-B.; Schrementi, J.; Kovacevic, S.; Paul, D.; Su, E. W.; Wang, J. Identification and Expression of Novel Isoforms of Human Stromal Cell-Derived Factor 1. *Gene* **2006**, *374*, 174–179.
- (97) Tamamura, H.; Tsutsumi, H.; Fujii, N. The Chemokine Receptor CXCR4 as a Therapeutic Target for Several Diseases. *Mini Rev. Med. Chem.* **2006**, *6*, 989–995.
- (98) Ceradini, D. J.; Kulkarni, A. R.; Callaghan, M. J.; Tepper, O. M.; Bastidas, N.; Kleinman, M. E.; Capla, J. M.; Galiano, R. D.; Levine, J. P.; Gurtner, G. C. Progenitor Cell Trafficking Is Regulated by Hypoxic Gradients through HIF-1 Induction of SDF-1. *Nat. Med.* **2004**, *10*, 858.
- (99) Burger, J. A.; Kipps, T. J. CXCR4: A Key Receptor in the Crosstalk between Tumor Cells and Their Microenvironment. *Blood* **2006**, *107*, 1761.
- (100) Fernandis, A. Z.; Prasad, A.; Band, H.; Klösel, R.; Ganju, R. K. Regulation of CXCR4-Mediated Chemotaxis and Chemoinvasion of Breast Cancer Cells. *Oncogene* **2004**, *23*, 157–167.
- (101) Lucia, L.; Francesca, T.; Franco, M.; Pia, L. T.; Evaristo, M. Chemokine Receptor CXCR4: Role in Gastrointestinal Cancer. *Crit. Rev. Oncol. Hematol.* **2013**, *88*, 696–705.
- (102) Burger, J. A.; Stewart, D. J.; Wald, O.; Peled, A. Potential of CXCR4 Antagonists for the Treatment of Metastatic Lung Cancer. *Expert Rev. Anticancer Ther.* **2011**, *11*, 621–630.
- (103) Ramsey, D. M.; McAlpine, S. R. Halting Metastasis through CXCR4 Inhibition. *Bioorg. Med. Chem. Lett.* **2012**, *23*, 20–25.
- (104) Balkwill, F. Cancer and the Chemokine Network. *Nat. Rev. Cancer* **2004**, *4*, 540–550.
- (105) Murdoch, C.; Monk, P. N.; Division, A. F. Functional Expression of Chemokine Receptor CXCR4 on Human Epithelial Cells. *Immunology* **1999**, *98*, 36–41.
- (106) Barlic-Dicen, J. Chemokine Receptors on the Defensive – the Surprising Role of CXCR4 in Brown Adipose Tissue. *Receptors & Clinical Investigation*, 2014, *2*, 10–14800/rci.397.
- (107) Liu, N.; Patzak, A.; Zhang, J. CXCR4-Overexpressing Bone Marrow-Derived Mesenchymal Stem Cells Improve Repair of Acute Kidney Injury. *Am. J. Physiol. Renal Physiol.* **2013**, *305*, F1064–F1073.
- (108) Egan, C. G.; Lavery, R.; Caporali, F.; Fondelli, C.; Laghi-Pasini, F.; Dotta, F.; Sorrentino, V. Generalised Reduction of Putative Endothelial Progenitors and CXCR4-Positive Peripheral Blood Cells in Type 2 Diabetes. *Diabetologia* **2008**, *51*, 1296–1305.
- (109) Hernandez, P. A.; Gorlin, R. J.; Lukens, J. N.; Taniuchi, S.; Bohinjec, J.; Francois, F.; Klotman, M. E.; Diaz, G. A. Mutations in the Chemokine Receptor Gene CXCR4 Are Associated with WHIM Syndrome, a Combined Immunodeficiency Disease. *Nat. Genet.* **2003**, *34*, 70–74.
- (110) Diaz, G. A.; Gulino, A. V. WHIM Syndrome: A Defect in CXCR4 Signaling. *Curr. Allergy Asthma Rep.* **2005**, *5*, 350–355.
- (111) Arnolds, K. L.; Spencer, J. V. CXCR4: A Virus’s Best Friend? *Infect. Genet. Evol.* **2014**, *25*, 146–156.

- (112) Liang, Z.; Zhan, W.; Zhu, A.; Yoon, Y.; Lin, S.; Sasaki, M.; Klapproth, J. M. A.; Yang, H.; Grossniklaus, H. E.; Xu, J. Development of a Unique Small Molecule Modulator of CXCR4. *PLoS One* **2012**, *7*, e34038–e34038.
- (113) Portella, L.; Vitale, R.; De Luca, S.; D'Alterio, C.; Ieranò, C.; Napolitano, M.; Riccio, A.; Polimeno, M. N.; Monfregola, L.; Barbieri, A.; Luciano, A.; Ciarmiello, A.; Arra, C.; Castello, G.; Amodeo, P.; Scala, S. Preclinical Development of a Novel Class of CXCR4 Antagonist Impairing Solid Tumors Growth and Metastases. *PLoS One* **2013**, *8*, e74548.
- (114) Maurice, P.; Kamal, M.; Jockers, R. Asymmetry of GPCR Oligomers Supports Their Functional Relevance. *Trends Pharmacol. Sci.* **2011**, *32*, 514–520.
- (115) Wang, J.; Norcross, M. Dimerization of Chemokine Receptors in Living Cells: Key to Receptor Function and Novel Targets for Therapy. *Drug Discov. Today* **2008**, *13*, 625–632.
- (116) Sohy, D.; Parmentier, M.; Springael, J. Y. Allosteric Transinhibition by Specific Antagonists in CCR2/CXCR4 Heterodimers. *J. Biol. Chem.* **2007**, *282*, 30062.
- (117) Huang, X.; Shen, J.; Cui, M.; Shen, L.; Luo, X.; Ling, K.; Pei, G.; Jiang, H.; Chen, K. Molecular Dynamics Simulations on SDF-1 [alpha]: Binding with CXCR4 Receptor. *Biophys. J.* **2003**, *84*, 171–184.
- (118) Perez-Nueno, V. I.; Ritchie, D. W.; Rabal, O.; Pascual, R.; Borrell, J. I.; Teixido, J. Comparison of Ligand-Based and Receptor-Based Virtual Screening of HIV Entry Inhibitors for the CXCR4 and CCR5 Receptors Using 3D Ligand Shape Matching and Ligand–receptor Docking. *J. Chem. Inf. Model.* **2008**, *48*, 509–533.
- (119) Neves, M. A. C.; Simões, S.; Sá e Melo, M. L. Ligand-Guided Optimization of CXCR4 Homology Models for Virtual Screening Using a Multiple Chemotype Approach. *J. Comput. Aided. Mol. Des.* **2010**, *24*, 1023–1033.
- (120) Deupi, X.; Standfuss, J. Structural Insights into Agonist-Induced Activation of G-Protein-Coupled Receptors. *Curr. Opin. Struct. Biol.* **2011**, *21*, 541–551.
- (121) Kawatkar, S. P.; Yan, M.; Gevariya, H.; Lim, M. Y.; Eisold, S.; Zhu, X.; Huang, Z.; An, J. Computational Analysis of the Structural Mechanism of Inhibition of Chemokine Receptor CXCR4 by Small Molecule Antagonists. *Exp. Biol. Med. Biol. Med.* **2011**, *236*, 844–850.
- (122) Peeters, M. C.; van Westen, G. J. P.; Li, Q.; IJzerman, A. P. Importance of the Extracellular Loops in G Protein-Coupled Receptors for Ligand Recognition and Receptor Activation. *Trends Pharmacol. Sci.* **2011**, *32*, 35–42.
- (123) Unal, H.; Karnik, S. S. Domain Coupling in GPCRs: The Engine for Induced Conformational Changes. *Trends Pharmacol. Sci.* **2011**, *33*, 79–88.
- (124) Avlani, V. A.; Gregory, K. J.; Morton, C. J.; Parker, M. W.; Sexton, P. M.; Christopoulos, A. Critical Role for the Second Extracellular Loop in the Binding of Both Orthosteric and Allosteric G Protein-Coupled Receptor Ligands. *J. Biol. Chem.* **2007**, *282*, 25677–25686.
- (125) Narlawar, R.; Lane, J. R.; Doddareddy, M.; Lin, J.; Brussee, J.; IJzerman, A. P. Hybrid Ortho/Allosteric Ligands for the Adenosine A1 Receptor. *J. Med. Chem.* **2010**, *53*, 3028–3037.
- (126) Izmaylov, A. F.; Scuseria, G. E.; Frisch, M. J. Efficient Evaluation of Short-Range Hartree-Fock Exchange in Large Molecules and Periodic Systems. *J. Chem. Phys.* **2006**, *125*, 104103.
- (127) Leach, A. R. *Molecular Modelling: Principles and Applications*; 2nd ed.; Pearson Education, 2001.
- (128) Pople, J. A.; Nesbet, R. K. Self-Consistent Orbitals for Radicals. *J. Chem. Phys.* **1954**, *22*, 571.
- (129) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46*, 618–622.
- (130) David Sherrill, C.; Schaefer, H. F. The Configuration Interaction Method: Advances in Highly Correlated Approaches. *Adv. Quantum Chem.* **1999**, *34*, 143–269.
- (131) Kümmel, H. G. A Biography of the Coupled Cluster Method. *Int. J. Mod. Phys. B* **2003**, *17*, 5311–5325.
- (132) Nakano, H. Quasidegenerate Perturbation Theory with Multiconfigurational Self-consistent-field Reference Functions. *J. Chem. Phys.* **1993**, *99*, 7983–7992.

- (133) Knowles, P.; Werner, H. An Efficient Method for the Evaluation of Coupling Coefficients in Configuration Interaction Calculations. *Chem. Phys. Lett.* **1988**, *145*, 514–522.
- (134) Angeli, C.; Bories, B. Third-Order Multireference Perturbation Theory: The N-Electron Valence State Perturbation-Theory Approach. *J. Chem. Phys.* **2006**, *124*, 054108.
- (135) Pople, J.; Segal, G. Approximate Self Consistent Molecular Orbital Theory. III. CNDO Results for AB₂ and AB₃ Systems. *J. Chem. Phys.* **1966**, *44*, 3289–3296.
- (136) Dewar, M.; Thiel, W. Ground States of Molecules. 38. The MNDO Method. Approximations and Parameters. *J. Am. Chem. Soc.* **1977**, *99*, 4899–4907.
- (137) Gordon, M.; Pople, J. Approximate Self Consistent Molecular Orbital Theory. VI. INDO Calculated Equilibrium Geometries. *J. Chem. Phys.* **1968**, *49*, 4643–4650.
- (138) Hofmann, H.; Birner, P. Success of the NDDO Method to Predict Conformations in Conjugated Systems. *Chem. Phys. Lett.* **1976**, *37*, 608–610.
- (139) Ramachandran, K.; Deepa, G. ; Namboori, K. . *Computational Chemistry and Molecular Modeling*; Springer Science & Business Media, 2008.
- (140) Mackerell, A. Empirical Force Fields for Biological Macromolecules: Overview and Issues. *J. Comput. Chem.* **2004**, *25*, 1584–1604.
- (141) Lii, J.; Allinger, N. Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 2. Vibrational Frequencies and Thermodynamics. *J. Am. Chem. Soc.* **1989**, *111*, 8566–8575.
- (142) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.
- (143) Clark, M.; Cramer III, R. D.; Van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *10*, 982–1012.
- (144) Cornell, W.; Cieplak, P.; Bayly, C. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- (145) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
- (146) Gunsteren, WF van; Daura, Xavier; Marck, A. E. GROMOS Force Field. *Encycl. Comput. Chem.* **1998**.
- (147) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.
- (148) Maple, J. R.; Hwang, M.-J.; Stockfisch, T. P.; Dinur, U.; Waldman, M.; Ewig, C. S.; Hagler, A. T. Derivation of Class II Force Fields. I. Methodology and Quantum Force Field for the Alkyl Functional Group and Alkane Molecules. *J. Comput. Chem.* **1994**, *15*, 162–182.
- (149) SYBYL. X 2.0, Tripos International,. 1699 South Hanley Rd., St. Louis, MO 63144, USA, 2012.
- (150) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (151) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual Screening Using Protein-Ligand Docking: Avoiding Artificial Enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793–806.
- (152) Kroemer, R. T. Structure-Based Drug Design: Docking and Scoring. *Curr. Protein Pept. Sci.* **2007**, *8*, 312–328.

- (153) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.
- (154) Jones, G.; Willett, P.; Glen, R. C.; Leach, a R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727–748.
- (155) Cavasotto, C. N.; Abagyan, R. a. Protein Flexibility in Ligand Docking and Virtual Screening to Protein Kinases. *J. Mol. Biol.* **2004**, *337*, 209–225.
- (156) Jain, A. N. Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *J. Med. Chem.* **2003**, *46*, 499–511.
- (157) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Daniel, T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.
- (158) Thomsen, R.; Christensen, M. MolDock: A New Technique for High-Accuracy Molecular Docking. *J. Med. Chem.* **2006**, *49*, 3315–3321.
- (159) Huang, S. Y.; Zou, X. Advances and Challenges in Protein-Ligand Docking. *Int. J. Mol. Sci.* **2010**, *11*, 3016–3034.
- (160) McGann, M. FRED Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2011**, *51*, 578–596.
- (161) Neves, M.; Totrov, M.; Abagyan, R. Docking and Scoring with ICM: The Benchmarking Results and Strategies for Improvement. *J. Comput. Aided. Mol. Des.* **2012**, *26*, 675–686.
- (162) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461.
- (163) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein–ligand Docking Using GOLD. *Proteins Struct. Funct. Bioinforma.* **2003**, *52*, 609–623.
- (164) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: A Novel Method for the Shape-Directed Rapid Docking of Ligands to Protein Active Sites. *J. Mol. Graph. Model.* **2003**, *21*, 289–307.
- (165) MOE. (*Molecular Operating Environment*); Chemical Computing Group Inc.: Montreal, Canada, 2009.
- (166) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949.
- (167) *Discovery Studio*; Accelrys, Inc: San Diego, 2009.
- (168) Jain, A. Scoring Noncovalent Protein-Ligand Interactions: A Continuous Differentiable Function Tuned to Compute Binding Affinities. *J. Comput. Aided. Mol. Des.* **1996**, *10*, 427–440.
- (169) Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M. LigScore: A Novel Scoring Function for Predicting Binding Affinities. *J. Mol. Graph. Model.* **2005**, *23*, 395–407.
- (170) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. R.; Fogel, L. J.; Freer, S. T. Molecular Recognition of the Inhibitor AG-1343 by HIV-1 Protease: Conformationally Flexible Docking by Evolutionary Programming. *Chem. Biol.* **1995**, *2*, 317–324.
- (171) Muegge, I.; Martin, Y. C. A General and Fast Scoring Function for Protein–Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, *42*, 791–804.
- (172) Chang, M.; Richard, K. Empirical Entropic Contributions in Computational Docking: Evaluation in APS Reductase Complexes. *J. Comput Chem.* **2011**, *29*, 1753–1761.
- (173) Graves, A. P.; Brenk, R.; Shoichet, B. K. Decoys for Docking. *J. Med. Chem.* **2005**, *48*, 3714–3728.
- (174) Stroud, R. M.; Finer-Moore, J. *Computational and Structural Approaches to Drug Discovery: Ligand-Protein Interactions*; Royal Society of Chemistry, 2008.

- (175) Bonvin, A. M. J. J. Flexible Protein–protein Docking. *Curr. Opin. Struct. Biol.* **2006**, *16*, 194–200.
- (176) Gray, J. J.; Moughon, S. E.; Kortemme, T.; Schueler-Furman, O.; Misura, K. M. S.; Morozov, A. V.; Baker, D. Protein–protein Docking Predictions for the CAPRI Experiment. *Proteins Struct. Funct. Bioinforma.* **2003**, *52*, 118–122.
- (177) Moal, I. H.; Moretti, R.; Baker, D.; Fernández-Recio, J. Scoring Functions for Protein-Protein Interactions. *Curr. Opin. Struct. Biol.* **2013**, *23*, 862–867.
- (178) Huang, S. Search Strategies and Evaluation in Protein–protein Docking: Principles, Advances and Challenges. *Drug Discov. Today* **2014**, *19*, 1081–1096.
- (179) Pierce, B.; Weng, Z. ZRANK: Reranking Protein Docking Predictions with an Optimized Energy Function. *Proteins Struct. Funct. Bioinforma.* **2007**, *67*, 1078–1086.
- (180) Dominguez, C.; Boelens, R.; Bonvin, A. M. J. J. HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information. *J. Am. Chem. Soc.* **2003**, *125*, 1731–1737.
- (181) London, N.; Raveh, B.; Cohen, E.; Fathi, G.; Schueler-Furman, O. Rosetta FlexPepDock Web Server—high Resolution Modeling of Peptide–protein Interactions. *Nucleic Acids Res.* **2011**, *39*, W249–W253.
- (182) Klabunde, T. Chemogenomic Approaches to Drug Discovery: Similar Receptors Bind Similar Ligands. *Br. J. Pharmacol.* **2007**, *152*, 5–7.
- (183) Yang, S.-Y. Pharmacophore Modeling and Applications in Drug Discovery: Challenges and Recent Advances. *Drug Discov. Today* **2010**, *15*, 444–450.
- (184) Wolber, G.; Seidel, T.; Bendix, F.; Langer, T. Molecule–Pharmacophore Superpositioning and Pattern Matching in Computational Drug Design. *Drug Discov. Today* **2008**, *13*, 23–29.
- (185) Zuccotto, F. Pharmacophore Features Distributions in Different Classes of Compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1542–1552.
- (186) Wolber, G.; Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **2005**, *45*, 160–169.
- (187) Todeschini, Roberto, and V. C. *Molecular Descriptors for Chemoinformatics. Volumes I & II*; John Wiley & Sons, Inc., 2009.
- (188) Randić, M.; Zupan, J. On Interpretation of Well-Known Topological Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 550–560.
- (189) Kroemer, R. T.; Hecht, P.; Liedl, K. R. Different Electrostatic Descriptors in Comparative Molecular Field Analysis: A Comparison of Molecular Electrostatic and Coulomb Potentials. *J. Comput. Chem.* **1996**, *17*, 1296–1308.
- (190) Katritzky, a R.; Gordeeva, E. V. Traditional Topological Induces Vs Electronic, Geometrical, and Combined Molecular Descriptors in Qsar/Qspr Research. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835–857.
- (191) Tetko, I. V.; Bruneau, P.; Mewes, H.-W.; Rohrer, D. C.; Poda, G. I. Can We Estimate the Accuracy of ADME-Tox Predictions? *Drug Discov. Today* **2006**, *11*, 700–707.
- (192) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley New York, 1990.
- (193) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (194) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71*, 58–63.
- (195) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
- (196) Rose, S.; Stevens, A. Computational Design Strategies for Combinatorial Libraries. *Curr. Opin. Chem. Biol.* **2003**, *7*, 331–339.

- (197) Gregori-Puigjané, E.; Mestres, J. Coverage and Bias in Chemical Library Design. *Curr. Opin. Chem. Biol.* **2008**, *12*, 359–365.
- (198) Gregori-Puigjané, E.; Mestres, J. SHED: Shannon Entropy Descriptors from Topological Feature Distributions. *J. Chem. Inf. Model.* **2006**, *46*, 1615–1622.
- (199) Violeta I. Pérez-Nuño, Vishwesh Venkatraman, L. M. and D. W. R. Predicting Drug Promiscuity Using Spherical Harmonic Surface Shape-Based Similarity Comparisons. *Open Conf. Proc. J.* **2011**, *2*, 113–129.
- (200) Gillet, V. J. New Directions in Library Design and Analysis. *Curr. Opin. Chem. Biol.* **2008**, *12*, 372–378.
- (201) Welsch, M. E.; Snyder, S. a.; Stockwell, B. R. Privileged Scaffolds for Library Design and Drug Discovery. *Curr. Opin. Chem. Biol.* **2010**, *14*, 347–361.
- (202) CombiGlide. Version 2.4, Schrödinger, LLC, New York, NY, 2010.
- (203) Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646–652.
- (204) Schames, J.; Henchman, R. Discovery of a Novel Binding Trench in HIV Integrase. *J. Med. Chem.* **2004**, *47*, 1879–1881.
- (205) Grant, B. J.; Lukman, S.; Hocker, H. J.; Sayyah, J.; Brown, J. H.; McCammon, J. A.; Gorfe, A. a. Novel Allosteric Sites on Ras for Lead Generation. *PLoS One* **2011**, *6*, e25711.
- (206) Smith, S. O.; Song, D.; Shekar, S.; Groesbeek, M.; Ziliox, M.; Aimoto, S. Structure of the Transmembrane Dimer Interface of Glycophorin A in Membrane Bilayers†. *Biochemistry* **2001**, *40*, 6553–6558.
- (207) Petrache, H. I.; Grossfield, A.; MacKenzie, K. R.; Engelman, D. M.; Woolf, T. B. Modulation of Glycophorin A Transmembrane Helix Interactions by Lipid Bilayers: Molecular Dynamics Calculations. *J. Mol. Biol.* **2000**, *302*, 727–746.
- (208) Yuzlenko, O.; Lazaridis, T. Membrane Protein Native State Discrimination by Implicit Membrane Models. *J. Comput. Chem.* **2013**, *34*, 731–738.
- (209) Im, W.; Feig, M.; Brooks, C. L. An Implicit Membrane Generalized Born Theory for the Study of Structure, Stability, and Interactions of Membrane Proteins. *Biophys. J.* **2003**, *85*, 2900–2918.
- (210) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J Chem Phys* **1995**, *103*, 8577–8593.
- (211) Borhani, D. W.; Shaw, D. E. The Future of Molecular Dynamics Simulations in Drug Discovery. *J. Comput. Aided. Mol. Des.* **2012**, *26*, 15–26.
- (212) Durrant, J. D.; McCammon, J. A. Molecular Dynamics Simulations and Drug Discovery. *BMC Biol.* **2011**, *9*, 71.
- (213) Hong, G.; Cornish, a. J.; Hegg, E. L.; Pachter, R. On Understanding Proton Transfer to the Biocatalytic [Fe - Fe]H Sub-Cluster in [Fe-Fe]H₂ases: QM/MM MD Simulations. *Biochim. Biophys. Acta - Bioenerg.* **2011**, *1807*, 510–517.
- (214) Cieplak, P.; Dupradeau, F. Polarization Effects in Molecular Mechanical Force Fields. *J. Phys. Condens. Matter* **2009**, *21*, 333102.
- (215) Rahimi, N. VEGFR-1 and VEGFR-2: Two Non-Identical Twins with a Unique Physiognomy. *Front. Biosci.* **2006**, *11*, 818–829.
- (216) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How To Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information. *J. Chem. Inf. Model.* **2009**, *49*, 678–692.
- (217) Tawa, G. J.; Baber, J. C.; Humblet, C. Computation of 3D Queries for ROCS Based Virtual Screens. *J. Comput. Aided. Mol. Des.* **2009**, *23*, 853–868.
- (218) Zhou, Z.; Felts, A. K.; Friesner, R. A.; Levy, R. M. Comparative Performance of Several Flexible Docking Programs and Scoring Functions: Enrichment Studies for a Diverse Set of Pharmaceutically Relevant Targets. *J. Chem. Inf. Model.* **2007**, *47*, 1599–1608.

- (219) Liang, Z.; Brooks, J.; Willard, M.; Liang, K. CXCR4/CXCL12 Axis Promotes VEGF-Mediated Tumor Angiogenesis through Akt Signaling Pathway. *Biochem. Biophys. Res. Commun.* **2007**, *359*, 716–722.
- (220) Polverino, A.; Coxon, A.; Starnes, C.; Diaz, Z.; DeMelfi, T.; Wang, L.; Bready, J.; Estrada, J.; Cattley, R.; Kaufman, S. AMG 706, an Oral, Multikinase Inhibitor That Selectively Targets Vascular Endothelial Growth Factor, Platelet-Derived Growth Factor, and Kit Receptors, Potently Inhibits Angiogenesis and Induces Regression in Tumor Xenografts. *Cancer Res.* **2006**, *66*, 8715.
- (221) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. 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.* **2008**, *22*, 213–228.
- (222) Bernstein, F.; Koetzle, T.; Williams, G. Protein Data Bank, Atomic Coordinate and Bibliographic Entry Format Description. *J. Mol. Biol.* **1977**, *112*, 535–542.
- (223) Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874.
- (224) Clark, R. D.; Webster-Clark, D. J. Managing Bias in ROC Curves. *J. Comput. Aided. Mol. Des.* **2008**, *22*, 141–146.
- (225) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual Screening Using Protein-Ligand Docking: Avoiding Artificial Enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793–806.
- (226) Mosley, C. A.; Wilson, L. J.; Wiseman, J. M.; Skudlarek, J. W.; Liotta, D. C. Recent Patents Regarding the Discovery of Small Molecule CXCR4 Antagonists. *Expert Opin. Ther. Patients* **2009**, *19*, 23–38.
- (227) Moncunill, G.; Armand-Ugon, M.; Clotet-Codina, I.; Pauls, E.; Ballana, E.; Llano, A.; Romagnoli, B.; Vrijbloed, J. W.; Gombert, F. O.; Clotet, B. Anti-HIV Activity and Resistance Profile of the CXC Chemokine Receptor 4 Antagonist POL3026. *Mol. Pharmacol.* **2008**, *73*, 1264.
- (228) Šali, A.; Blundell, T. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993**, *234*, 779–815.
- (229) Wong, R. S. Y.; Bodart, V.; Metz, M.; Labrecque, J.; Bridger, G.; Fricker, S. P. Comparison of the Potential Multiple Binding Modes of Bicyclam, Monocyclam, and Noncyclam Small-Molecule CXC Chemokine Receptor 4 Inhibitors. *Mol. Pharmacol.* **2008**, *74*, 1485–1495.
- (230) Vilar, S.; Ferino, G.; Phatak, S. S.; Berk, B.; Cavasotto, C. N.; Costanzi, S. Docking-Based Virtual Screening for Ligands of G Protein-Coupled Receptors: Not Only Crystal Structures but Also in Silico Models. *J. Mol. Graph. Model.* **2011**, *29*, 614–623.
- (231) Roumen, L.; Sanders, M.; Vroling, B.; de Esch, I. J. P.; de Vlieg, J.; Leurs, R.; Klomp, J. P. G.; Nabuurs, S. B.; de Graaf, C. In Silico Veritas: The Pitfalls and Challenges of Predicting GPCR-Ligand Interactions. *Pharmaceuticals* **2011**, *4*, 1196–1215.
- (232) Lam, A. R.; Bhattacharya, S.; Patel, K.; Hall, S. E.; Mao, A.; Vaidehi, N. Importance of Receptor Flexibility in Binding of Cyclam Compounds to the Chemokine Receptor CXCR4. *J. Chem. Inf. Model.* **2010**, *51*, 137–147.
- (233) Carugo, O.; Pongor, S. A Normalized Root Mean Square Distance for Comparing Protein Three Dimensional Structures. *Protein Sci.* **2001**, *10*, 1470–1473.
- (234) Bill, R. M.; Henderson, P. J. F.; Iwata, S.; Kunji, E. R. S.; Michel, H.; Neutze, R.; Newstead, S.; Poolman, B.; Tate, C. G.; Vogel, H. Overcoming Barriers to Membrane Protein Structure Determination. *Nat. Biotechnol.* **2011**, *29*, 335–340.
- (235) Zou, Y.; Weis, W. I.; Kobilka, B. K. N-Terminal T4 Lysozyme Fusion Facilitates Crystallization of a G Protein Coupled Receptor. *PLoS One* **2012**, *7*, e46039.
- (236) Hatse, S.; Princen, K.; Vermeire, K.; Gerlach, L. O.; Rosenkilde, M. M.; Schwartz, T. W.; Bridger, G.; De Clercq, E.; Schols, D. Mutations at the CXCR4 Interaction Sites for AMD3100 Influence Anti-CXCR4 Antibody Binding and HIV-1 Entry. *FEBS Lett.* **2003**, *546*, 300–306.
- (237) Gerlach, L. O.; Skerlj, R. T.; Bridger, G. J.; Schwartz, T. W. Molecular Interactions of Cyclam and Bicyclam Non-Peptide Antagonists with the CXCR4 Chemokine Receptor. *J. Biol. Chem.* **2001**, *276*, 14153–14160.

- (238) Rosenkilde, M. M.; Gerlach, L. O.; Jakobsen, J. S.; Skerlj, R. T.; Bridger, G. J.; Schwartz, T. W. Molecular Mechanism of AMD3100 Antagonism in the CXCR4 Receptor. *J. Biol. Chem.* **2004**, *279*, 3033.
- (239) Rosenkilde, M. M.; Gerlach, L. O.; Hatse, S.; Skerlj, R. T.; Schols, D.; Bridger, G. J.; Schwartz, T. W. Molecular Mechanism of Action of Monocyclam versus Bicyclam Non-Peptide Antagonists in the CXCR4 Chemokine Receptor. *J. Biol. Chem.* **2007**, *282*, 27354–27365.
- (240) Pettersson, S.; Pérez-Nueno, V. I.; Mena, M. P.; Clotet, B.; Esté, J. A.; Borrell, J. I.; Teixidó, J. Novel Monocyclam Derivatives as HIV Entry Inhibitors: Design, Synthesis, Anti-HIV Evaluation, and Their Interaction with the CXCR4 Co-Receptor. *ChemMedChem* **2010**, *5*, 1272–1281.
- (241) Tamamura, H.; Xu, Y.; Hattori, T.; Zhang, X.; Arakaki, R.; Kanbara, K.; Omagari, A.; Otaka, A.; Ibuka, T.; Yamamoto, N. A Low-Molecular-Weight Inhibitor against the Chemokine Receptor CXCR4: A Strong Anti-HIV Peptide T140. *Biochem. Biophys. Res. Commun.* **1998**, *253*, 877–882.
- (242) Trent, J. O.; Wang, Z.; Murray, J. L.; Shao, W.; Tamamura, H.; Fujii, N.; Peiper, S. C. Lipid Bilayer Simulations of CXCR4 with Inverse Agonists and Weak Partial Agonists. *J. Biol. Chem.* **2003**, *278*, 47136–47144.
- (243) Våbenø, J.; Nikiforovich, G. V.; Marshall, G. R. Insight into the Binding Mode for Cyclopentapeptide Antagonists of the CXCR4 Receptor. *Chem. Biol. Drug Des.* **2006**, *67*, 346–354.
- (244) Ballesteros, J. A.; Weinstein, H. Integrated Methods for the Construction of Three-Dimensional Models and Computational Probing of Structure-Function Relations in G Protein-Coupled Receptors. *Methods Neurosci.* **1995**, *25*, 366–428.
- (245) Nagel, K.; Jimeno-Yepes, A.; Rebholz-Schuhmann, D. Annotation of Protein Residues Based on a Literature Analysis: Cross-Validation against UniProtKb. *BMC Bioinformatics* **2009**, *10*, S4.
- (246) Maybridge Bringing life to drug discovery, M. D.; England, F. S. I. No Title, 2005.
- (247) De Clercq, E. Recent Advances on the Use of the CXCR4 Antagonist Plerixafor (AMD3100, Mozobil™) and Potential of Other CXCR4 Antagonists as Stem Cell Mobilizers. *Pharmacol. Ther.* **2010**, *128*, 509–518.
- (248) Bridger, G. J.; Skerlj, R. T.; Thornton, D.; Padmanabhan, S.; Martellucci, S. A.; Henson, G. W.; Abrams, M. J.; Yamamoto, N.; Vreese, K. D. Synthesis and Structure-Activity Relationships of Phenylenebis (methylene)-Linked Bis-Tetraazamacrocycles That Inhibit HIV Replication. Effects of Macrocyclic Ring Size and Substituents on the Aromatic Linker. *J. Med. Chem.* **1995**, *38*, 366–378.
- (249) Bridger, G.; Skerlj, R.; Kaller, A.; Harwig, C.; Bogucki, D.; Wilson, T. R.; Crawford, J.; Mceachern, E. J.; Atsma, B.; Nan, S. Chemokine Receptor Binding Heterocyclic Compounds. U.S. Patent No. 6,750,348., 2004.
- (250) Bridger, G. J.; Skerlj, R. T.; Padmanabhan, S.; Martellucci, S. A.; Henson, G. W.; Struyf, S.; Witvrouw, M.; Schols, D.; De Clercq, E. Synthesis and Structure-Activity Relationships of Phenylenebis (methylene)-Linked Bis-Azamacrocycles That Inhibit HIV-1 and HIV-2 Replication by Antagonism of the Chemokine Receptor CXCR4. *J. Med. Chem.* **1999**, *42*, 3971–3981.
- (251) Ichiyama, K.; Yokoyama-Kumakura, S.; Tanaka, Y.; Tanaka, R.; Hirose, K.; Bannai, K.; Edamatsu, T.; Yanaka, M.; Niitani, Y.; Miyano-Kurosaki, N. A Duodenally Absorbable CXC Chemokine Receptor 4 Antagonist, KRH-1636, Exhibits a Potent and Selective Anti-HIV-1 Activity. *Proc. Natl. Acad. Sci.* **2003**, *100*, 4185–4190.
- (252) Yamazaki, T.; Saitou, A.; Ono, M.; Yokoyama, S.; Bannai, K.; Hirose, K.; Yanaka, M. Novel Nitrogenous Compound and Use Thereof." World Patent WO/2003/029218, 2003.
- (253) Wang, R.; Lu, Y.; Fang, X.; Wang, S. An Extensive Test of 14 Scoring Functions Using the PDBbind Refined Set of 800 Protein-Ligand Complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114–2125.
- (254) Montes, M.; Miteva, M. A.; Villoutreix, B. O. Structure-Based Virtual Ligand Screening with LigandFit: Pose Prediction and Enrichment of Compound Collections. *Proteins Struct. Funct. Bioinforma.* **2007**, *68*, 712–725.
- (255) Bustanji, Y.; Taha, M. O.; Yousef, A. M.; G. Al-Bakri, A. Berberine Potently Inhibits Protein Tyrosine Phosphatase 1B: Investigation by Docking Simulation and Experimental Validation. *J. Enzyme Inhib. Med. Chem.* **2006**, *21*, 163–171.
- (256) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity--a Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219–3228.

- (257) Lasko, T. A.; Bhagwat, J. G.; Zou, K. H.; Ohno-Machado, L. The Use of Receiver Operating Characteristic Curves in Biomedical Informatics. *J. Biomed. Inform.* **2005**, *38*, 404–415.
- (258) Mason, J. S.; Bortolato, A.; Congreve, M.; Marshall, F. H. New Insights from Structural Biology into the Druggability of G Protein-Coupled Receptors. *Trends Pharmacol. Sci.* **2012**, *33*, 249–260.
- (259) Choi, W. T.; An, J. Biology and Clinical Relevance of Chemokines and Chemokine Receptors CXCR4 and CCR5 in Human Diseases. *Exp. Biol. Med.* **2011**, *236*, 637–647.
- (260) Thoma, G.; Streiff, M. B.; Kovarik, J.; Glickman, F.; Wagner, T.; Beerli, C.; Zerwes, H. G. Orally Bioavailable Isothioureas Block Function of the Chemokine Receptor CXCR4 in Vitro and in Vivo. *J. Med. Chem.* **2008**, *51*, 7915–7920.
- (261) Lin, J. H.; Clark, T. An Analytical, Variable Resolution, Complete Description of Static Molecules and Their Intermolecular Binding Properties. *J. Chem. Inf. Model.* **2005**, *45*, 1010–1016.
- (262) Ritchie, D. W.; Kemp, G. J. L. Protein Docking Using Spherical Polar Fourier Correlations. *Proteins Struct. Funct. Bioinforma.* **2000**, *39*, 178–194.
- (263) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape. *J. Comput. Chem.* **1998**, *17*, 1653–1666.
- (264) Cai, W.; Xu, J.; Shao, X.; Leroux, V.; Beautrait, A.; Maigret, B. SHEF: A vHTS Geometrical Filter Using Coefficients of Spherical Harmonic Molecular Surfaces. *J. Mol. Model.* **2008**, *14*, 393–401.
- (265) Mosi, R. M.; Anastassova, V.; Cox, J.; Darkes, M. C.; Idzan, S. R.; Labrecque, J.; Lau, G.; Nelson, K. L.; Patel, K.; Santucci, Z. The Molecular Pharmacology of AMD11070: An Orally Bioavailable CXCR4 HIV Entry Inhibitor. *Biochem. Pharmacol.* **2011**, *83*, 472–479.
- (266) Tamamura, H.; Ojida, A.; Ogawa, T.; Tsutsumi, H.; Masuno, H.; Nakashima, H.; Yamamoto, N.; Hamachi, I.; Fujii, N. Identification of a New Class of Low Molecular Weight Antagonists against the Chemokine Receptor CXCR4 Having the Dipicolylamine–Zinc(II) Complex Structure. *J. Med. Chem.* **2006**, *49*, 3412–3415.
- (267) Catalano, J. G.; Gudmundsson, K. S.; Svolto, A.; Boggs, S. D.; Miller, J. F.; Spaltenstein, A.; Thomson, M.; Wheelan, P.; Minick, D. J.; Phelps, D. P.; Jenkinson, S. Synthesis of a Novel Tricyclic 1,2,3,4,4a,5,6,10b-Octahydro-1,10-Phenanthroline Ring System and CXCR4 Antagonists with Potent Activity against HIV-1. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 2186–2190.
- (268) Iwasaki, Y.; Akari, H.; Murakami, T.; Kumakura, S.; Dewan, M. Z.; Yanaka, M.; Yamamoto, N. Efficient Inhibition of SDF-1 α -mediated Chemotaxis and HIV-1 Infection by Novel CXCR4 Antagonists. *Cancer Sci.* **2009**, *100*, 778–781.
- (269) Jenkinson, S.; Thomson, M.; McCoy, D.; Edelstein, M.; Danehower, S.; Lawrence, W.; Wheelan, P.; Spaltenstein, A.; Gudmundsson, K. Blockade of X4-Tropic HIV-1 Cellular Entry by GSK812397, a Potent Noncompetitive cxcr4 Receptor Antagonist. *Antimicrob. Agents Chemother.* **2010**, *54*, 817–824.
- (270) Truax, V. M.; Zhao, H.; Katzman, B. M.; Prosser, A. R.; Alcaraz, A. A.; Saindane, M. T.; Howard, R. B.; Culver, D.; Arrendale, R. F.; Gruddanti, P. R.; Evers, T. J.; Natchus, M. G.; Snyder, J. P.; Liotta, D. C.; Wilson, L. J. Discovery of Tetrahydroisoquinoline-Based CXCR4 Antagonists. *ACS Med. Chem. Lett.* **2013**, *4*, 1025–1030.
- (271) Narumi, T.; Tanaka, T.; Hashimoto, C.; Nomura, W.; Aikawa, H.; Sohma, A.; Itotani, K.; Kawamata, M.; Murakami, T.; Yamamoto, N.; Tamamura, H. Pharmacophore-Based Small Molecule CXCR4 Ligands. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 4169–4172.
- (272) Pérez-Nueno, V. I.; Pettersson, S.; Ritchie, D. W.; Borrell, J. I.; Teixido, J. Discovery of Novel HIV Entry Inhibitors for the CXCR4 Receptor by Prospective Virtual Screening. *J. Chem. Inf. Model.* **2009**, *49*, 810–823.
- (273) Pettersson, S.; Pérez-Nueno, V. I.; Ros-Blanco, L.; Puig de La Bellacasa, R.; Rabal, M. O.; Batllori, X.; Clotet, B.; Clotet-Codina, I.; Armand-Ugón, M.; Esté, J. Discovery of Novel Non-cyclam Polynitrogenated CXCR4 Coreceptor Inhibitors. *ChemMedChem* **2008**, *3*, 1549–1557.
- (274) Pettersson, S.; Pérez-Nueno, V. I.; Mena, M. P.; Clotet, B.; Esté, J. A.; Borrell, J. I.; Teixidó, J. Novel Monocyclam Derivatives as HIV Entry Inhibitors: Design, Synthesis, anti-HIV Evaluation, and Their Interaction with the CXCR4 Co-receptor. *ChemMedChem* **2010**, *5*, 1272–1281.

- (275) LigPrep. Version 2.4, Schrödinger, LLC, New York, NY, 2010.
- (276) Lipinski, C. A. Lead-and Drug-like Compounds: The Rule-of-Five Revolution. *Drug Discov. Today Technol.* **2004**, *1*, 337–341.
- (277) Ahmed, M.; Ahmet, D.; Basheer, H. Potential of a Novel CXCR4 Chemokine Receptor Antagonist for the Treatment of Metastatic Cancer. *researchgate.net* **2013**, *8*, 78744.
- (278) Vinader, V.; Ahmet, D. S.; Ahmed, M. S.; Patterson, L. H.; Afarinkia, K. Discovery and Computer Aided Potency Optimization of a Novel Class of Small Molecule CXCR4 Antagonists. *PLoS One* **2013**, *8*, e78744.
- (279) Cui, Q.; Karplus, M. Allostery and Cooperativity Revisited. *Protein Sci.* **2008**, *17*, 1295–1307.
- (280) Csermely, P.; Palotai, R.; Nussinov, R. Induced Fit, Conformational Selection and Independent Dynamic Segments: An Extended View of Binding Events. *Trends Biochem. Sci.* **2010**, *35*, 539–546.
- (281) Nussinov, R.; Tsai, C.-J. Allostery in Disease and in Drug Discovery. *Cell* **2013**, *153*, 293–305.
- (282) Reynolds, K. A.; McLaughlin, R. N.; Ranganathan, R. Hot Spots for Allosteric Regulation on Protein Surfaces. *Cell* **2011**, *147*, 1564–1575.
- (283) Conn, P.; Christopoulos, A.; Lindsley, C. Allosteric Modulators of GPCRs: A Novel Approach for the Treatment of CNS Disorders. *Nat. Rev. Drug Discov.* **2009**, *8*, 41–54.
- (284) Wang, C.-I. A.; Lewis, R. J. Emerging Opportunities for Allosteric Modulation of G-Protein Coupled Receptors. *Biochem. Pharmacol.* **2012**, *85*, 153–162.
- (285) Wooten, D.; Christopoulos, A.; Sexton, P. M. Emerging Paradigms in GPCR Allostery: Implications for Drug Discovery. *Nat. Rev. Drug Discov.* **2013**, *12*, 630–644.
- (286) Ballesteros, J.; Ransom, J. In-Target versus off-Target Allosteric Modulators of GPCRs. *Drug Discov. Today Ther. Strateg.* **2007**, *3*, 445–450.
- (287) Moyle, G.; DeJesus, E.; Boffito, M.; Wong, R. S.; Gibney, C.; Badel, K.; MacFarland, R.; Calandra, G.; Bridger, G.; Becker, S. Proof of Activity with AMD11070, an Orally Bioavailable Inhibitor of CXCR4-Tropic HIV Type 1. *Clin. Infect. Dis.* **2009**, *48*, 798–805.
- (288) Fricker, S.; Mosi, R.; Anastassova, V.; Labrecque, J.; Wong, R.; Skerlj, R.; Bridger, G.; Huskens, D.; Schols, D. The Orally Bioavailable Allosteric CXCR4 HIV-1 Entry Inhibitor AMD11070. *Retrovirology* **2012**, *9*, P7.
- (289) Boggs, S.; Elitzin, V. I.; Gudmundsson, K.; Martin, M. T.; Sharp, M. J. Kilogram-Scale Synthesis of the CXCR4 Antagonist GSK812397. *Org. Process Res. Dev.* **2009**, *13*, 781–785.
- (290) Oishi, S.; Fujii, N. Peptide and Peptidomimetic Ligands for CXC Chemokine Receptor 4 (CXCR4). *Org. Biomol. Chem.* **2012**, *10*, 5720–5731.
- (291) Dimond, P.; Carlson, K.; Bouvier, M.; Gerard, C.; Xu, L.; Covic, L.; Agarwal, A.; Ernst, O. P.; Janz, J. M.; Schwartz, T. W. G Protein-coupled Receptor Modulation with Pepducins: Moving Closer to the Clinic. *Ann. N. Y. Acad. Sci.* **2011**, *1226*, 34–49.
- (292) Covic, L.; Gresser, A. L.; Talavera, J.; Swift, S.; Kuliopulos, A. Activation and Inhibition of G Protein-Coupled Receptors by Cell-Penetrating Membrane-Tethered Peptides. *Proc. Natl. Acad. Sci.* **2002**, *99*, 643–648.
- (293) Carlson, K. E.; McMurry, T. J.; Hunt III, S. W. Pepducins: Lipopeptide Allosteric Modulators of GPCR Signaling. *Drug Discov. Today Technol.* **2012**, *9*, e33–e39.
- (294) Quoyer, J.; Janz, J. M.; Luo, J.; Ren, Y.; Armando, S.; Lukashova, V.; Benovic, J. L.; Carlson, K. E.; Hunt, S. W.; Bouvier, M. Pepducin Targeting the C-X-C Chemokine Receptor Type 4 Acts as a Biased Agonist Favoring Activation of the Inhibitory G Protein. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, E5088–E5097.
- (295) Janz, J. M.; Ren, Y.; Looby, R. J.; Kazmi, M. A.; Sachdev, P.; Grunbeck, A.; Haggis, L.; Chinnapen, D.; Lin, A. Y.; Seibert, C.; McMurry, T.; Carlson, K. E.; Muir, T. W.; Hunt III, S. W.; Sakmar, T. P. Direct Interaction between an Allosteric Agonist Pepducin and the Chemokine Receptor CXCR4. *J. Am. Chem. Soc.* **2011**, *40*, 15878–15881.

- (296) Tchernychev, B.; Ren, Y.; Sachdev, P.; Janz, J. M.; Haggis, L.; O'Shea, A.; McBride, E.; Looby, R.; Deng, Q.; McMurry, T. Discovery of a CXCR4 Agonist Pepducin That Mobilizes Bone Marrow Hematopoietic Cells. *Proc. Natl. Acad. Sci.* **2010**, *107*, 22255–22259.
- (297) Chen, R.; Li, L.; Weng, Z. ZDOCK: An Initial-Stage Protein-Docking Algorithm. *Proteins Struct. Funct. Bioinforma.* **2003**, *52*, 80–87.
- (298) Bruno, A.; Costantino, G. Molecular Dynamics Simulations of G Protein-Coupled Receptors. *Mol. Inform.* **2012**, *31*, 222–230.
- (299) Zhang, C.; Du, C.; Feng, Z.; Zhu, J.; Li, Y. Hologram Quantitative Structure Activity Relationship, Docking, and Molecular Dynamics Studies of Inhibitors for CXCR4. *Chem. Biol. Drug Des.* **2014**.
- (300) Rubinstein, M.; Niv, M. Y. Peptidic Modulators of Protein-Protein Interactions: Progress and Challenges in Computational Design. *Biopolymers* **2009**, *91*, 505–513.
- (301) Zhang, P.; Leger, A. J.; Baleja, J. D.; Rana, R.; Corlin, T.; Nguyen, N.; Koukos, G.; Bohm, A.; Covic, L.; Kuliopulos, A. Allosteric Activation of a G Protein Coupled Receptor with Cell-Penetrating Receptor Mimetics. *J. Biol. Chem.* **2015**, *290*, 15785–15798.
- (302) Christopoulos, A.; Kenakin, T. G Protein-Coupled Receptor Allosterism and Complexing. *Pharmacol. Rev.* **2002**, *54*, 323–374.
- (303) Birdsall, N.; Lazareno, S. Allosterism at Muscarinic Receptors: Ligands and Mechanisms. *Mini-Reviews Med. Chem.* **2005**, *5*, 523–543.
- (304) Schrader, A. J.; Lechner, O.; Templin, M.; Dittmar, K. E. J.; Machtens, S.; Mengel, M.; Probst-Kepper, M.; Franzke, A.; Wollensak, T.; Gatzlaff, P.; Atzpodien, J.; Buer, J.; Lauber, J. CXCR4/CXCL12 Expression and Signalling in Kidney Cancer. *Br J Cancer* **2002**, *86*, 1250–1256.
- (305) Nagasawa, T. CXC Chemokine Ligand 12 (CXCL12) and Its Receptor CXCR4. *J. Mol. Med. (Berl)*. **2014**, *92*, 433–439.

ARTICULOS

La presente tesis doctoral está basada, tal y como se ha indicado en el texto, en los artículos siguientes:

Artículo 1 Planesas, J. M.; Claramunt, R. M.; Teixidó, J.; Borrell, J. I.; Pérez-Nueno, V. I. Improving VEGFR-2 Docking-Based Screening by Pharmacophore Postfiltering and Similarity Search Postprocessing. *J. Chem. Inf. Model.* **2011**, 51, 777–787.

Artículo 2 Planesas, J. M.; Pérez-Nueno, V. I.; Borrell, J. I.; Teixidó, J. Impact of the CXCR4 Structure on Docking-Based Virtual Screening of HIV Entry Inhibitors. *J. Mol. Graph. Model.* **2012**, 123–136.

Artículo 3 Karaboga, A. S.; Planesas, J. M.; Petronin, F.; Teixidó, J.; Souchet, M.; Pérez-Nueno, V. I. Highly Specific and Sensitive Pharmacophore Model for Identifying CXCR4 Antagonists. Comparison with Docking and Shape-Matching Virtual Screening Performance. *J. Chem. Inf. Model.* **2013**, 53, 1043–1056.

Artículo 4 Planesas, J. M.; Pérez-Nueno, V. I.; Borrell, J. I.; Teixidó, J. Studying the Binding Interactions of Allosteric Agonists and Antagonists of the CXCR4 Receptor. *J. Mol. Graph. Model.* **2015**, 60, 1–14.

Artículo 1

Improving VEGFR-2 Docking-Based Screening by Pharmacophore Postfiltering and Similarity Search Postprocessing

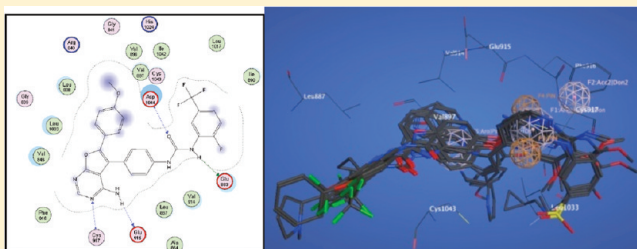
Jesús M. Planesas,[‡] Rosa M. Claramunt,[‡] Jordi Teixidó,[†] José I. Borrell,[†] and Violeta I. Pérez-Nueno^{†,*}

[†]Grup d'Enginyeria Molecular, Institut Químic de Sarrià (IQS), Universitat Ramon Llull, Barcelona, Spain

[‡]Departamento de Química Orgánica y Bio-Orgánica, Facultad de Ciencias, UNED, Senda del Rey 9, E-28040 Madrid, Spain

S Supporting Information

ABSTRACT: Conventional docking-based virtual screening (VS) of chemical databases is based on the ranking of compounds according to the values retrieved by a scoring function (typically, the binding affinity estimation). However, using the most suitable scoring function for each kind of receptor pocket is not always an effective process to rank compounds, and sometimes neither to distinguish between correct binding modes from incorrect ones. To improve actives from decoys selection, here we propose a three-step VS protocol, which includes the conventional docking step, a pharmacophore postfilter step, and a similarity search postprocess. This VS protocol is retrospectively applied to VEGFR-2 (Kdr-kinase) inhibitors. The resulting docking poses calculated using the Alpha HB scoring function implemented in MOE are postfiltered according to defined pharmacophore interactions (structure based). The selected poses are again ranked according to their molecular similarity (MACCS fingerprint) to the cognate ligand. Results show that both the overall and early VS performance improve the application of this protocol.



INTRODUCTION

The vascular endothelial growth factor (VEGF) is responsible for regulation of blood vessel growth processes starting from pre-existing blood vessels (angiogenesis).¹ There is evidence and clinical observations which confirm that processes of abnormal angiogenesis can cause various diseases such as diabetic retinopathy, psoriasis, rheumatoid arthritis, macular degeneration, and cancer. In particular, it is considered that metastasis and growth of solid tumors largely depend on angiogenesis. Among the different kind of factors, angiopoietins² and VEGF are particularly interesting because they are specific factors of endothelial cells and they play important and coordinated roles in vascular development.³

Inside the family of VEGF factors we find their receptors, VEGFR. Those are protein kinases that modulate angiogenesis.⁴ There are different kinds of kinase receptors: VEGFR-1, VEGFR-2 (or KDR/Flk1), VEGFR-3, and VEGF-A to VEGF-E. While VEGFR-1 is critical for the development of angiogenesis and VEGFR3 is restricted to lymphatic endothelial cells, VEGFR-2 seems to mediate in the majority of angiogenesis and pathological effects at high levels (processes like cell proliferation, differentiation of endothelial cells, invasion, or migration).^{5,6} Therefore, the development of VEGFR-2 inhibitors^{7,8} has been manifested as a potent treatment for metastatic tumor growth and cancer.⁹

Conventional structure-based docking¹⁰ drug design aims to predict the preferred conformation (pose) of a molecule (ligand) in relation to a target receptor (protein) when both bind to form

a stable complex. The function which measures and evaluates the goodness-of-fit for protein docking is the scoring function. Recently, several papers have been published^{11,12} to evaluate, in a comparative manner, docking performance for different targets using different scoring functions. Authors evaluate these scoring functions using in silico databases of known ligands and decoys, such as DUD database¹³ or MUV.¹⁴ The purpose of these papers is to study how the different scoring functions can distinguish ligands from decoys (retrospective screening), in order to find the scoring function that should be used to perform prospective docking-based virtual screening (VS) with the highest performance.

This work focuses on the retrospective docking of VEGFR-2 protein with known ligands and decoys¹⁵ obtained from the DUD database. The computational software MOE¹⁶ has been used, and the most suitable scoring function for VEGFR-2 inhibitors has been selected. The objective is to improve VS when scoring functions themselves do not reach high specificity to distinguish actives from decoys, as it happens in many retrospective dockings when dealing with difficult targets, mainly the ones displaying large binding sites, such as VEGFR-2, where it is difficult to accommodate all ligand conformations, and in which multiple ligands can bind in different subsites. Some authors have reported docking-based VS experiments on VEGFR-2 showing poor performance; for example, Kirchmair et al. have reported

Received: July 17, 2010

Published: March 18, 2011

AUC values from 0.43 to 0.48¹⁷ and Tawa et al. have performed¹⁸ VEGFR-2 VS based in Rapid Overlay of Chemical Structures (ROCS) obtaining an AUC of 0.61. These low values are due to the difficulty in finding a common query matching all different

Table 1. Ligands and Decoys Molecular Properties: Summary of the 1D Physicochemical Properties of Active and Decoy Molecules in the VEGFR-2 DUD Screening Database^a

	ligands: 78	decoys: 2479
molecular weight:	238.3/692.6 (86.5)	325.4/474.2 (21.6)
rotatable single bonds	0.000/0.294 (0.064)	0.057/0.286 (0.037)
hydrogen bond acceptors	1/7 (1.3)	0/7 (0.9)
hydrogen bond donors	0/5 (1.2)	0/4 (0.7)
hydrophobic atoms	12/32 (3.7)	11/25 (1.7)
octanol–water partition coeff	0.25/6.79 (1.17)	1.33/5.74 (0.72)
polar surface area	34.2/374.2 (54.2)	19.1/288.2 (37.6)

^aSimilar molecular values between ligands and decoys avoid bias in docking process. Standard deviation values in parentheses.

VEGFR-2 ligands, given that they bind in multiple pocket subsites. Hence, here we propose a three-step VS protocol applied to VEGFR-2 inhibitors, which includes the conventional docking step, a pharmacophore postfilter step, and a similarity search postprocess to improve actives from decoys selection. Our protocol might also work for other difficult targets given that it is taking into account not only docking information but also receptor-based pharmacophore filter and ligand similarity searching, which can help to determine important features for pocket subsite binding.

METHODS

Crystal Structure and Docking Scoring Function Selection. To date, all known protein kinase structures have been determined by protein crystallography.¹⁹ In the RCSB Protein Databank²⁰ we can find more than 20 crystal structures (PDB) for protein kinase VEGFR-2, all determined by X-ray crystallography, with different resolutions and different ligands, and in some cases with some mutations in the protein sequences.

Table 2. Self-Docking Validation^a

PDB ^b	scoring function: ASE				scoring function: Affinity dG				
	Alpha Triangle	Alpha PMI	Proxy Triangle	Triangle Matcher	PDB	Alpha Triangle	Alpha PMI	Proxy Triangle	Triangle Matcher
1Y6A	2.4	5.1	4.6	4.6	1Y6A	4.7	5.1	4.5	3.6
1YWN	1.4	0.9	0.8	0.9	1YWN	2.5	0.9	3.3	3.3
2OH4	0.9	1.2	0.8	0.8	2OH4	0.9	1.2	0.8	0.8
2P2H	3.0	4.5	1.4	1.5	2P2H	3.2	1.4	1.4	1.4
2QU5	0.9	0.7	0.8	0.8	2QU5	1.3	0.7	0.8	0.8
2RL5	0.8	0.3	0.9	0.9	2RL5	1.1	0.2	1.0	0.9
3BE2	1.6	0.9	1.8	1.8	3BE2	3.8	0.9	1.6	1.6
3C7Q	2.4	5.5	2.2	5.5	3C7Q	1.7	5.5	2.7	2.7
3CJF	3.2	5.8	3.3	3.3	3CJF	4.1	6.0	1.2	2.4
3EWH	1.5	3.0	1.7	1.2	3EWH	2.3	3.0	2.8	2.8
3CJG	3.5	7.3	2.2	2.2	3CJG	2.5	7.5	2.9	2.9
aRMSD ^c	2.0	3.2	1.9	2.1	aRMSD	2.6	3.0	2.1	2.1
Inf. 2 Å ^d	6	5	7	7	Inf. 2 Å	4	6	6	5

PDB	Scoring function: Alpha HB				Scoring function: London dG				
	Alpha Triangle	Alpha PMI	Proxy Triangle	Triangle Matcher	PDB	Alpha Triangle	Alpha PMI	Proxy Triangle	Triangle Matcher
1Y6A	1.3	5.1	3.6	3.6	1Y6A	4.5	5.1	3.8	3.4
1YWN	0.8	0.9	0.8	0.8	1YWN	1.8	0.9	0.9	1.6
2OH4	1.5	1.2	0.8	0.8	2OH4	0.9	1.2	0.8	0.8
2P2H	0.8	5.4	1.4	1.4	2P2H	1.2	5.4	2.8	1.4
2QU5	0.8	0.7	1.1	0.8	2QU5	1.3	0.7	0.8	0.8
2RL5	1.1	0.2	0.9	0.9	2RL5	1.1	0.2	1.0	1.0
3BE2	1.2	0.9	1.6	1.6	3BE2	2.1	0.9	1.8	1.6
3C7Q	1.3	5.5	2.4	2.1	3C7Q	4.2	5.5	2.4	2.2
3CJF	3.2	6.0	3.4	3.4	3CJF	3.1	6.0	3.2	3.2
3EWH	1.8	3.0	1.2	1.2	3EWH	2.5	2.8	2.6	2.6
3CJG	0.9	7.5	1.7	1.7	3CJG	1.1	7.1	1.7	1.7
aRMSD	1.3	3.3	1.7	1.7	aRMSD	2.2	3.3	2.0	1.9
Inf. 2 Å	10	5	8	8	Inf. 2 Å	6	5	6	7

^aBest RMSD values obtained when the 11 cognate ligands were docked in their corresponding PDBs protein structures using the different placement functions for each scoring function. Scoring function Alpha HB and placement function Alpha Triangle obtain the lower aRMSD (1.3 Å) and the highest number of best poses whose RMSD value is below or equal 2.0 Å. All values are expressed in angstroms. ^bPDBs used do not contain allosteric ligands.

^cAverage rmsd. ^dNumber of best poses with a rmsd value equal or below 2 Å.

The first VEGFR-2 PDB complex published was 1Y6A (07-06-05), with a crystal resolution of 2.2 Å. This protein structure contains as cognate ligand a 2-anilino-5-(2-aryl)oxazole. This ligand has been assayed and tested²¹ to inhibit VEGFR-2 protein ($IC_{50} = 0.022 \mu M$). We used this structure, together with 10 more recent VEGFR-2 structures published to date, to select the most appropriate MOE scoring function for docking compounds into VEGFR-2 pocket. The 11 VEGFR-2 PDB complexes were downloaded, crystallized water molecules were eliminated in order to avoid possible hindrances during docking runs,²² and the proteins were protonated. MOE allows working with several scoring and placement functions to perform protein–ligand docking such as ASE, Affinity dG, Alpha HB, and London dG. The reliability of these docking scoring functions was assessed by a self-docking and a cross-docking analysis.

From the 11 existent VEGFR-2 PDB complexes, we selected one for our VS experiments by also applying a self-docking analysis as a first step, and a cross-docking study as a second step. Both analyses used the best scoring and placement functions found in the previous scoring function validation. If the average rmsd (aRMSD) of the top 10 ranked poses was equal or below 2.0 Å, the PDB complex was considered suitable for VEGFR-2 docking-based VS.

Self-Docking Analysis. The ligands for each of the 11 VEGFR2 complexes used were docked back into their corresponding protein structures using ASE, Affinity dG, Alpha HB, and London dG as docking-scoring functions and Alpha PMI, Alpha Triangle, Proxy Triangle, and Triangle Matcher as docking-placement functions implemented in MOE. The docking results were evaluated through comparison of the best docked ligands binding modes with the experimental ones. As a measure of docking reliability, the root-mean-square deviation (rmsd)¹¹ was used to compare differences between the atomic distances of the docked poses and the real cocrystallized pose. An evaluation of the most suitable scoring-placement function was assessed attending to the aRMSD and the number of best poses with rmsd values below or equal to 2.0 Å, obtained in both cases for the 11 PDBs.

Cross-Docking Analysis²³. To perform cross-docking evaluation, all VEGFR-2 complexes were aligned with each other, and then all ligands were docked into all complex structures. The docking reliability was evaluated by calculating for each ligand and each complex structure the rmsd between the reference position of the ligand in the experimental protein–ligand complex and that predicted by the docking function in the various VEGFR-2 complex structures. The aRMSD comparison and the number of best poses with rmsd values below or equal to 2.0 Å were carried out equally as described above for the self-docking studies.

Finally, in order to select the working structure we analyzed the statistical values obtained when all ligands were docked into their corresponding structure using the selected docking function and also taking into account their crystallographic resolution.

Dockable Database Preparation. We used VEGFR-2 test set in the DUD database^{11,12,22,24,25} for docking-based VS. This test set contains 78 ligands and 2479 decoys with similar physical properties to the actives, such as molecular weight, octanol–water partition coefficient (logP), polar surface area, number of molecular rotatable single bonds, or minimum number of hydrogen bond donors and acceptors,²⁶ in order to avoid bias in docking results²⁷ (see Table 1). All ligands and

Table 3. Summary of the Cross-Docking Validation^{a–d}

	aRMSD	rmsd std dev	no. of poses $\leq 2 \text{ \AA}$
ASE–Alpha PMI	4.9	2.0	10
ASE–Alpha Triangle	3.7	2.0	29
ASE–Proxi Triangle	3.3	1.7	33
ASE–Triangle Matcher	3.3	1.8	40
Affinity dG–Alpha PMI	4.9	1.9	11
Affinity dG–Alpha Triangle	4.0	2.2	22
Affinity dG–Proxi Triangle	3.6	2.0	29
Affinity dG–Triangle Matcher	3.6	1.9	26
Alpha HB–Alpha PMI	5.1	1.9	5
Alpha HB–Alpha Triangle	3.0	1.6	45
Alpha HB–Proxi Triangle	3.1	1.7	40
Alpha HB–Triangle Matcher	3.1	1.8	44
London dG–Alpha PMI	4.9	1.9	6
London dG–Alpha Triangle	3.1	1.8	41
London dG–Proxi Triangle	3.0	1.6	43
London dG–Triangle Matcher	3.1	1.7	40

^a PDB used do not contain allosteric ligands. ^b All values are expressed in angstroms. ^c Protein–ligand complexes were aligned prior to perform cross-dockings, and the resulting docked poses were compared with their respective cognate pose. ^d Averaged RMSD (aRMSD) values and number of best poses with a RMSD equal or below 2 Å (no. of poses $\leq 2 \text{ \AA}$) obtained when cognate ligands were docked in the different PDBs protein structures using the different scoring functions (bold) and placement functions. See Table S-1 in Supporting Information for details.

decoys were energy minimized using the MMFF94x forcefield in MOE.

Similarity Search and Receptor-Based Pharmacophore.

The ligand similarity search was performed comparing 2D MACCS fingerprints, and the Tanimoto index was used as scoring function.

The receptor-based pharmacophore was built using the “Pharmacophore consensus” tool implemented in MOE, which finds the most common interactions resulting from the alignment of the protein residues in the published VEGFR-2 PDB complexes. Pharmacophore feature radii were manually optimized by executing several dockings followed by pharmacophore postfiltering in order to maximize the ratio ligands/decoys.

Assessment of Docking-Based VEGFR-2 Virtual Screening.

Following the docking calculations, all compounds were sorted into ranked lists based upon their docking score, pharmacophore postfilter score, and similarity postprocessing score. These lists were then used to plot the percentage of known actives found versus the percentage of the ranked database screened (enrichment plot), and to calculate the enrichment factor (EF) at 1%, 5%, and 10% of false positives selected.²⁸ The EF measures the number of known ligands in the top-ranked list, relative to a random selection.²⁹ Moreover, Receiving Operating Characteristic (ROC) plots were calculated. ROC plots represent sensitivity, or the fraction of true positives recovered, versus 1-specificity, or the percentage of false positives recovered. The area under the ROC curve (AUC) gives the probability of ranking a randomly selected active higher than a randomly chosen decoy. It varies between 0 and 1, where 1 represents a perfect ranking (all actives ranked above the decoys), while 0.5 corresponds to a completely random ranking.³⁰

Table 4. PDB Selection: Statistical Values of Pose Prediction for VEGFR-2 PDBs with One Cognate Ligand Using Alpha HB as Scoring Function and Alpha Triangle as Placement Function. (a) RMSD Values Extracted from the Self-Docking Analysis. (b) RMSD Values Extracted from the Cross-Docking Analysis

PDB	resolution ^a (Å)	best rmsd pose (Å)	(a)				no. of poses ^b
			rmsd median (Å)	aRMSD (Å)	rmsd std dev (Å)		
1Y6A	2.10	1.3	4.0	3.8	0.9	10	
1YWN	1.71	0.8	1.4	1.4	0.3	10	
2OH4	2.05	1.5	2.6	2.5	0.6	10	
2P2H	1.95	0.8	2.2	2.3	0.9	10	
2QU5	2.95	0.8	1.4	1.3	0.4	10	
2RL5	2.65	1.1	1.5	1.4	0.2	10	
3BE2	1.75	1.2	2.9	2.6	0.9	10	
3C7Q	2.10	1.3	3.4	3.4	0.9	10	
3CJF	2.15	3.2	4.4	4.2	0.5	10	
3EWH	1.60	1.8	3.0	2.9	0.9	10	
3CJG	2.25	0.9	4.7	4.1	2.0	10	

(b)												
Alpha HB–Alpha Triangle												
Ligands												
PDB	1Y6A	1YWN	2OH4	2P2H	2QU5	2RL5	3BE2	3C7Q	3CJF	3EWH	3CJG	mean aRMSD
1Y6A	^c	(5.3/7.0)	(6.4/7.7)	(1.2/4.3)	(7.0/7.4)	(5.5/6.4)	(5.6/8.6)	(3.0/5.1)	(4.9/5.3)	(5.8/3.1)	(1.1/3.8)	5.9
1YWN	(3.2/5.3)		(1.2/2.6)	(1.6/4.1)	(2.0/7.4)	(1.4/2.3)	(1.4/5.3)	(1.9/6.7)	(2.9/5.6)	(1.3/9.3)	(1.5/3.7)	5.3
2OH4	(3.3/4.4)	(4.9/8.0)		(1.6/3.0)	(1.9/7.4)	(0.9/7.2)	(1.6/9.5)	(2.2/3.7)	(4.3/4.6)	(2.8/7.0)	(1.1/1.7)	5.6
2P2H	(4.7/5.2)	(5.1/8.0)	(4.4/8.5)		(4.8/7.8)	(5.5/6.5)	(6.5/9.3)	(2.8/3.1)	(1.7/2.3)	(5.9/8.2)	(1.7/3.0)	6.2
2QU5	(4.4/4.7)	(5.8/8.8)	(1.7/8.9)	(3.3/8.8)		(1.7/6.1)	(3.0/9.7)	(3.4/8.6)	(4.6/4.8)	(3.0/10.4)	(1.9/2.2)	7.3
2RL5	(4.7/4.3)	(0.8/7.0)	(0.9/7.5)	(1.4/4.6)	(1.8/7.3)		(1.7/8.7)	(2.5/2.9)	(4.5/4.8)	(1.2/7.8)	(0.9/2.1)	5.7
3BE2	(1.8/7.3)	(1.8/2.2)	(2.4/3.4)	(1.0/4.9)	(1.3/3.3)	(1.3/3.7)		(4.0/9.5)	(3.1/4.1)	(2.7/9.9)	(1.3/2.9)	5.1
3C7Q	(4.4/5.9)	(2.1/9.7)	(2.9/7.8)	(4.8/8.3)	(3.0/7.8)	(1.6/2.8)	(1.7/7.3)		(1.9/4.7)	(2.7/4.4)	(4.5/4.4)	6.3
3CJF	(3.1/5.7)	(5.8/7.1)	(4.2/7.3)	(2.5/3.8)	(4.2/7.6)	(4.5/4.3)	(3.6/4.3)	(2.3/6.3)		(3.5/6.3)	(1.6/3.0)	5.6
3EWH	(3.2/3.9)	(1.2/7.7)	(1.0/6.8)	(1.3/2.9)	(2.2/2.6)	(1.4/2.5)	(1.3/2.7)	(3.4/5.7)	(4.1/5.9)		(1.6/2.7)	4.3
3CJG	(4.2/5.7)	(5.5/6.6)	(1.6/7.0)	(1.5/5.1)	(5.6/6.9)	(3.5/5.8)	(5.4/8.5)	(3.2/4.2)	(1.6/3.9)	(5.1/8.9)		6.3

aRMSD:	3.0
std dev:	1.6
no. of poses ≤ 2 Å:	45

^aPDB crystallographic resolution. ^bNumber of poses used to calculate statistical rmsd values. ^c(best rmsd pose/aRMSD) for the best 10 poses.

RESULTS

Docking Function and Crystal Structure Selection. Regarding the docking function selection, among the different MOE scoring and placement functions, we tested 1936 dockings by self-docking and cross-docking validation. According to the self-docking validation (Table 2) the docking function Alpha HB–Alpha Triangle obtained the best results. The function returned 10 rmsd values below to 2.0 Å and also the lowest aRMSD value (1.3 Å) in comparison to the rest of docking functions. For the cross-docking validation, each cocrystallized ligand was docked into the rest of protein structures. Hence, for each cross-docking combination, a rmsd value was obtained, and therefore for each docking function, we computed an aRMSD value and also the number of poses ranked with a rmsd value below or equal to 2.0 Å. All cross-docking detailed results are available in the Supporting Information (Table S-1). Table 3 summarizes the aRMSD, standard deviation, and number of poses with a rmsd

value equal or below 2 Å for all the 16 combinations of scoring and placement functions. All 16 combinations returned aRMSD values equal or above 3.0 Å. Regarding the number of best poses with a rmsd value equal or below 2.0 Å, 4 of the 16 combinations (Alpha HB–Alpha Triangle, Alpha HB–Triangle Matcher, London dG–Alpha Triangle, London dG–Proxy Triangle) were selected as the best functions, with more than 40 poses retrieved. All the aforementioned combinations of scoring and placement functions returned also the lowest aRMSD values (3.0 Å and 3.1 Å). Combining these results with the self-docking analysis, Alpha HB–Alpha Triangle was chosen as the most suitable docking function to work with our set of VEGFR-2 PDBs.

Hence, we used Alpha HB–Alpha Triangle docking function for the PDB selection. Table 4a and 4b shows the rmsd values of pose prediction for all the available VEGFR-2 PDBs extracted from the self-docking (Table 4a) and the cross-docking (Table 4b) analyses. Table 4a shows that the best aRMSD values

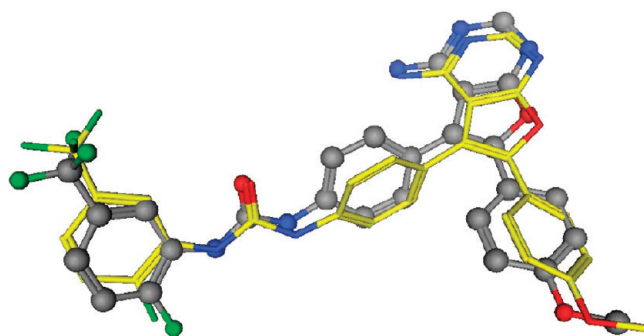


Figure 1. Superposition of the best rmsd docked pose onto the cocrystallized ligand in the VEGFR-2 complex (PDB code: 1YWN). Cocrystallized ligand from PDB 1YWN (carbon atoms in balls and gray color), and best rmsd docked pose (0.8 Å) (carbon atoms in sticks and yellow color) using Alpha HB scoring function and Alpha Triangle placement function implemented in MOE.

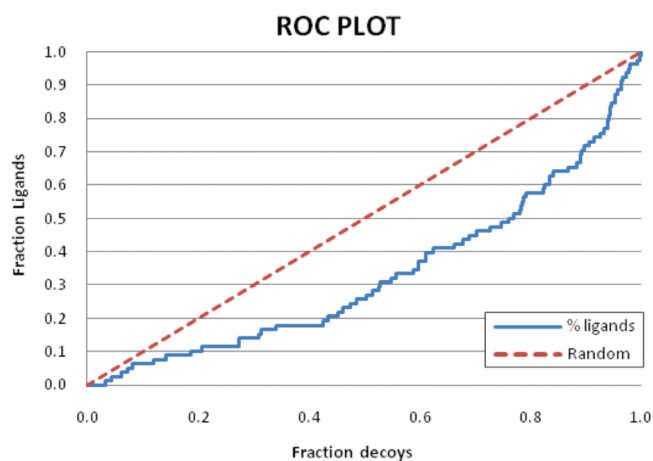
were obtained for the PDBs 2RL5 (1.4 Å), 2QU5 (1.3 Å), and 1YWN (1.4 Å), the three having similar average values for the top 10 ranked poses. PDB 1YWN and 2QU5 retrieved also the best rmsd poses: 0.8 Å (1YWN) (Figure 1) and 0.8 Å (2RL5). Nevertheless, crystallographic resolution was poorer for PDB 2RL5 (2.95 Å) and 2RL5 (2.65 Å) than for PDB 1YWN (1.71 Å). Table 4b shows that in the cross-docking analysis PDBs 3EWH, 3BE2, and 1YWN, respectively, retrieved the best mean aRMSD for the top 10 ranking poses. PDB 1YWN also retrieved the best rmsd pose together with PDBs 2RL5, 3BE2, and 3EWH.

Therefore, taking into account the crystallographic resolution and the analysis of the docking results obtained for each PDB, we defined structure 1YWN as our working structure³¹ to use for VEGFR-2 docking-based VS. This structure has also been used in docking experiments applied to an in-house database by Enyedy et al.³² This VEGFR-2 complex (1YWN) has a substituted 4-aminofuro[2,3-*d*]pyrimidine as cocrystallized ligand ($IC_{50} = 0.003 \mu M$ ³³). Hence, PDB 1YWN protein structure with Alpha HB function and Alpha Triangle placement function were used for the subsequent docking-based VS.

Retrospective Docking of DUD VEGFR-2 Data Set. We docked the VEGFR-2 test set in the DUD database using Alpha HB and Alpha Triangle MOE placement function. All compounds were ranked from higher to lower Alpha HB value. Figure 2 shows the resulting ROC plot. The overall AUC was worse than the random selection (AUC = 0.33), with a very low early recovery performance (% true positives at 1% of false positives: 0%, and % true positives at 5% of false positives: 2.56%).

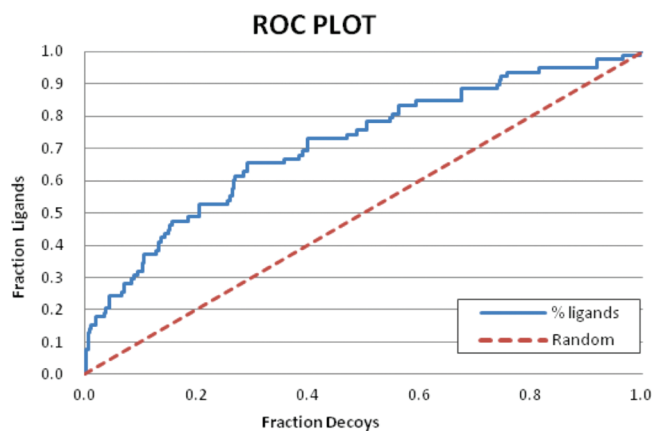
An alternative commonly used to postprocess docking poses is through molecular similarity analysis, using coefficients such as the Tanimoto index.³⁴ Similarity between two molecules is measured, comparing fingerprints or physicochemical properties. In this case, we used the 2D MACCS fingerprints although it is also possible to use 3D properties such as three-dimensional shape of molecules, due to their important role in bioactivity.³⁵ The underlying assumption is that similar structures have similar biological activity,³⁶ taking into account that there is a wide range of different similarity methods.³⁷

Thus, it is expected that similar molecules (molecules with Tanimoto coefficient values close to 1) should have similar activities when they bind to a biological target and they inhibit it.³⁸



% True Pos at 1% of False Pos	% True Pos at 5% of False Pos	% True Pos at 10% of False Pos.
0.00%	2.56%	6.41%

Figure 2. ROC plot obtained when docked poses are ranked by Alpha HB scoring function. AUC = 0.33. The table below the ROC plot shows early recognition values when poses are ranked by Alpha HB energy scoring function.



% True Pos at 1% of False Pos	% True Pos at 5% of False Pos	% True Pos at 10% of False Pos.
14.10%	24.36%	32.05%

Figure 3. ROC plot obtained when docked poses are ranked by MACCS similarity to the cognate ligand (Tanimoto index). AUC = 0.70. The table below the ROC plot shows early recognition values when poses are ranked by MACCS similarity to the cognate ligand.

Jahn et al.²⁴ successfully achieved discrimination between VEGFR-2 ligands and decoys using ligand-based VS methods. The results processed by Tanimoto MACCS similarity fingerprints improved early recognition in comparison with results obtained using only standard scoring functions for ranking compounds (e.g., awROC enrichments at 1% false positive fraction of 9.4 ± 2.2 versus 3.2 ± 1.5 applying similarity postprocess and a docking standard scoring function, respectively). Figure 3 shows VEGFR-2 docking-based VS results ranked by MACCS (Tanimoto coefficient) using the cocrystallized ligand from PDB 1YWN as the query. In our case, it can be

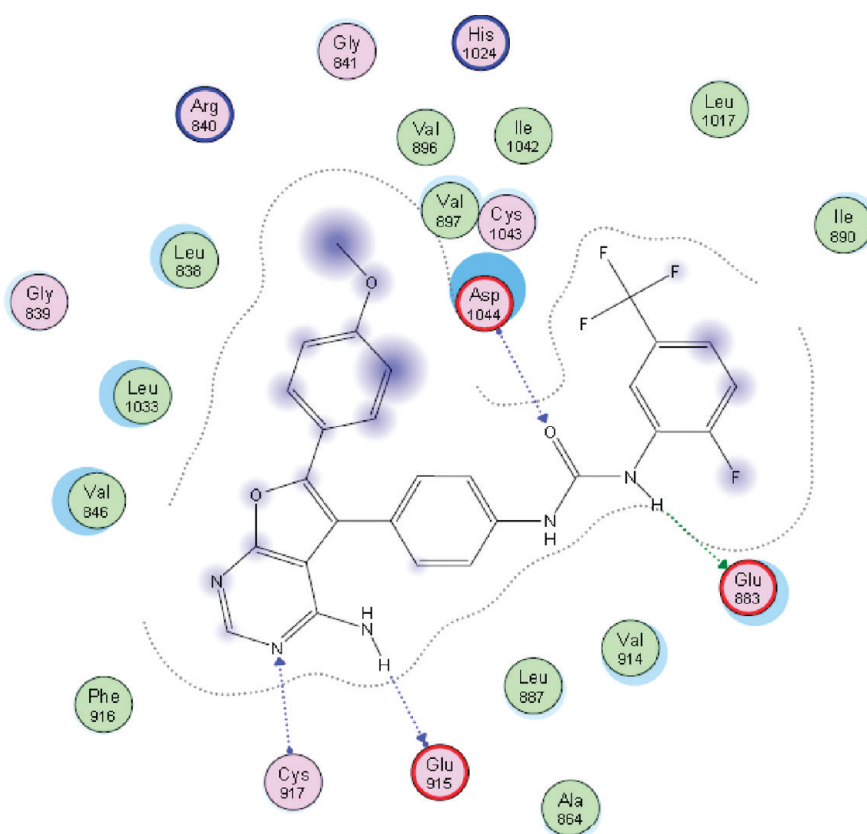


Figure 4. Key interactions between VEGFR-2 binding site residues and the cocrystallized ligand (PDB code: 1YWN). The cognate ligand makes two hydrogen bond donor interactions with residues Glu915 and Glu883 and two hydrogen bond acceptor interactions with residues Cys917 and Asp1044. The dashed line shows the ligand proximity contour. Residues in lilac color indicate polar interactions, and residues in green indicate nonpolar interactions. Ligand atoms with blue contour and residues with turquoise discs surrounding them (Leu838, Val846, Val914, Glu883, Leu1033, and Asp1044) indicate solvent accessibility.

observed that the overall area under the curve (AUC = 0.70) improves significantly the AUC obtained using only Alpha HB scoring function (AUC = 0.33). The early performance also improves substantially (% true positives at 1% of false positives: 0.0% using only Alpha HB scoring function, and % true positives at 1% of false positives: 14.1% with similarity postprocessing).

Pharmacophore Postfiltering Using Protein–Ligand Interaction Information. To improve VEGFR-2 docking-based VS, we used pharmacophore³⁹ postfiltering, that is to say, we filtered all docked poses, taking into account relevant protein–ligand interactions³⁵ (Figure 4). Often ligand-based pharmacophore filtering techniques are used as predocking screening methods. However, we pose-filtered docking results using a receptor-based pharmacophore which captures the essential protein–ligand interactions information.

Docking poses were filtered according to a series of protein–ligand interactions defined as features. All molecules whose docking poses did not satisfy all required interactions were eliminated.

Usually a pharmacophore query is defined following two strategies. One strategy is to perform dockings of several cocrystallized ligands and align these poses. Common interactions found for the different cognate ligands are analyzed and grouped defining the pharmacophore²⁵ query. The second strategy is to use tools for analyzing protein–ligand interactions for several PDB complexes and build a map-histogram for all of them. Common and repeated interactions are used as

pharmacophore features (e.g., PLIF tools implemented in MOE). In this work, we used protein alignment to superpose all cognate ligands. Ligand's superposition generated from the alignment of PDB complexes is more accurate than superposition of docked ligands, e.g., by flexible alignment or other ligand-based superposition methods, where the position (and interacting information) of ligand in reference to the receptor binding pocket is not taken into account. Nevertheless, these strategies are useful approximations⁴⁰ when their PDB complexes are not available for a protein target. Therefore, we aligned protein chains for all VEGFR-2 PDB complexes with crystallographic resolution lower than 3 Å (see Figure 5). Common ligand functional groups and common protein–ligand interactions located inside a threshold volume were used as pharmacophore features. Pharmacophore Consensus implemented in MOE was used for searching the common interactions between the overlapped ligands. The resulting pharmacophore query (Figure 6) was therefore defined on the basis of a mixed strategy, structure based and ligand based. Pharmacophore Consensus search returned five pharmacophore features, some of them showing multiple interactions (features F1, F2, and F5 in Figure 6). Features F1 and F2 indicate hydrogen bond acceptor^{41–43} and hydrogen bond donor⁴⁴ interaction sites with residue Cys 917, since this residue admits both interactions.^{45–47} Features F3, F4, and F5 show both hydrophobic and nonaromatic π -system rings interaction sites, mainly with residues Leu 838, Val 897, Leu 887, Leu 1033, and Cys1043.⁴⁷

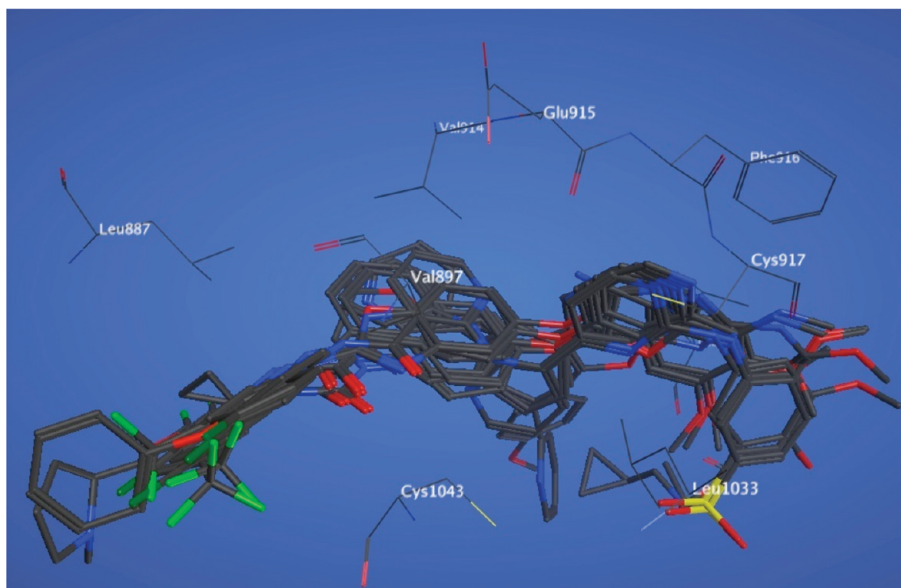


Figure 5. Superposition of cocrystallized ligands by aligning VEGFR-2 PDB complexes with crystallographic resolution lower than 3 Å.

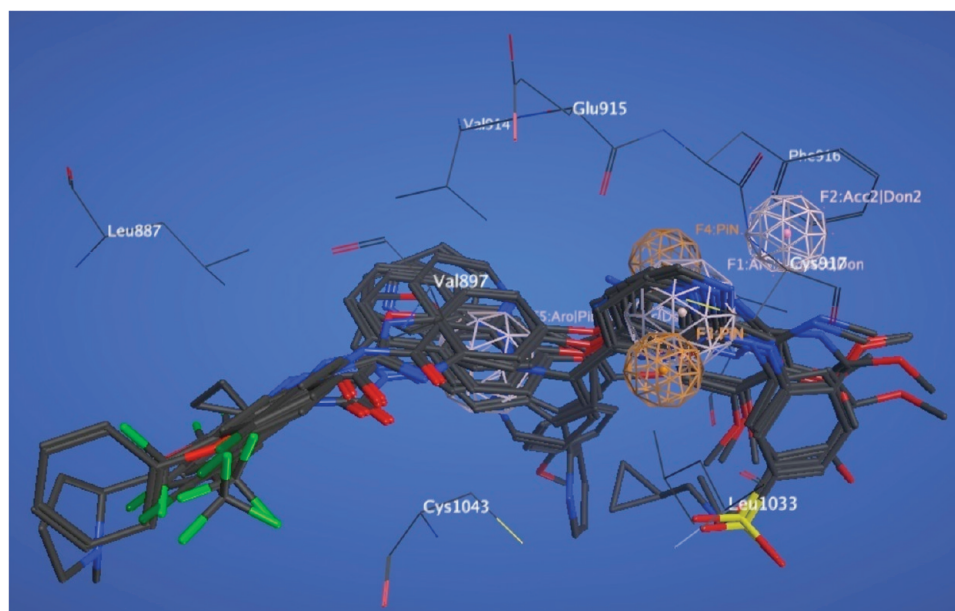


Figure 6. Pharmacophore model resulting from VEGFR-2 PDB complexes alignment and Pharmacophore Consensus analysis. Key interactions (features) defining the pharmacophore are colored in orange (F3 and F4) and pink (F1, F2, and F5). Each feature indicates one or more interactions at the same time. Feature F2 is projected out of feature F1 to facilitate visualization.

Radii of the features were optimized, by running several docking calculations and postfiltering docked poses, changing the pharmacophore radii values, until the maximum ratio of ligands/decoys was reached. The best results were achieved with two main variants of the same pharmacophore query (Figure 6). First, when the radii of the features F4 and F5 were 1.5 Å and 2 Å, respectively, 69.3% of the ligands satisfied the pharmacophore query, but also 35.2% of the decoys satisfied the pharmacophore filter, with an overall enrichment factor of 1.83. Second, if the radii of the features F4 and F5 were 1.0 Å and 1.5 Å, respectively, there were 59.0% of docked and postfiltered ligands, and 20.8% of false positives with an enrichment factor of 2.69 (see Figure 7).

Figure 8 and Figure 9 show the resulting enrichment plot and ROC plot, respectively, when the pharmacophoric filter is applied to docked poses and hits are ranked by their Alpha HB scoring value. Early recovery values are also shown in both cases. It can be observed that the early performance considerably increases using the pharmacophore postfilter (% true positives at 1% of false positives: 0.0% using only Alpha HB scoring function, and % true positives at 1% of false positives: 10.26% with pharmacophore postfilter).

Similarity search postprocessing was also applied to the pharmacophore postfiltered results. Figure 10 shows the resulting ROC plot when pharmacophore postfiltered docked poses

were ranked by MACCS similarity to the congnate ligand using the Tanimoto index. When the three steps (docking, pharmacophore postfilter and similarity search postprocessing) are used, the greatest improvement observed is in the early recovery performance (% true positives at 1% of false positives: 0.0% using only Alpha HB scoring function versus % true positives at 1% of false positives: 16.67% with pharmacophore postfilter and similarity postprocess, % true positives at 5% of false positives: 2.56% using only Alpha HB scoring function versus % true positives at 5% of false positives: 30.77% with pharmacophore postfilter and similarity postprocess, and % true positives at 10% of false positives: 6.41% using only Alpha HB scoring function

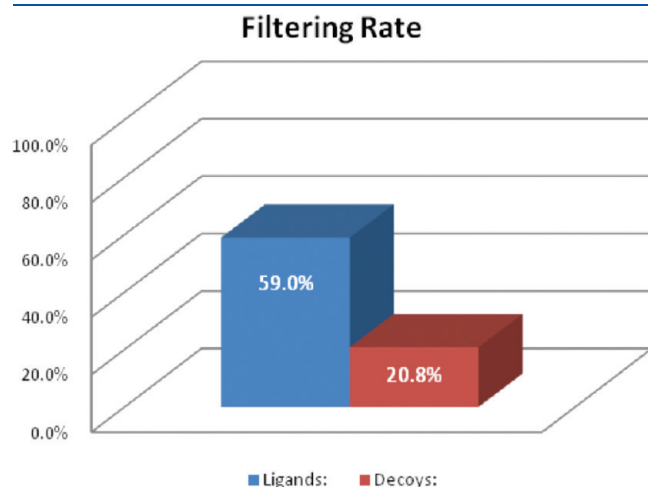


Figure 7. Filtering rate bar chart. Percentage of ligands and decoys recovered when docked poses are postfiltered with a pharmacophore.

versus % true positives at 10% of false positives: 41.03% with pharmacophore postfilter and similarity postprocess).

DISCUSSION

Docking-based VS aims to find the lowest energy binding modes for a set of ligands inside the receptor pocket site. In addition, scoring functions are expected to distinguish active

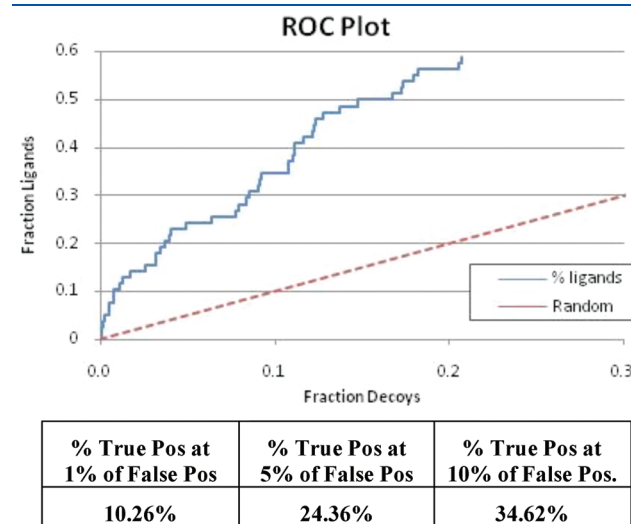


Figure 9. ROC plot obtained when the pharmacophore postfilter is applied to the docked poses ranked by Alpha HB scoring function. The curve only shows ligands and decoys whose docked poses match pharmacophoric requirements. No value is returned for docked poses that do not match the pharmacophore constrains; therefore, they cannot be plotted. In the table, the ROC plot shows early recognition values.

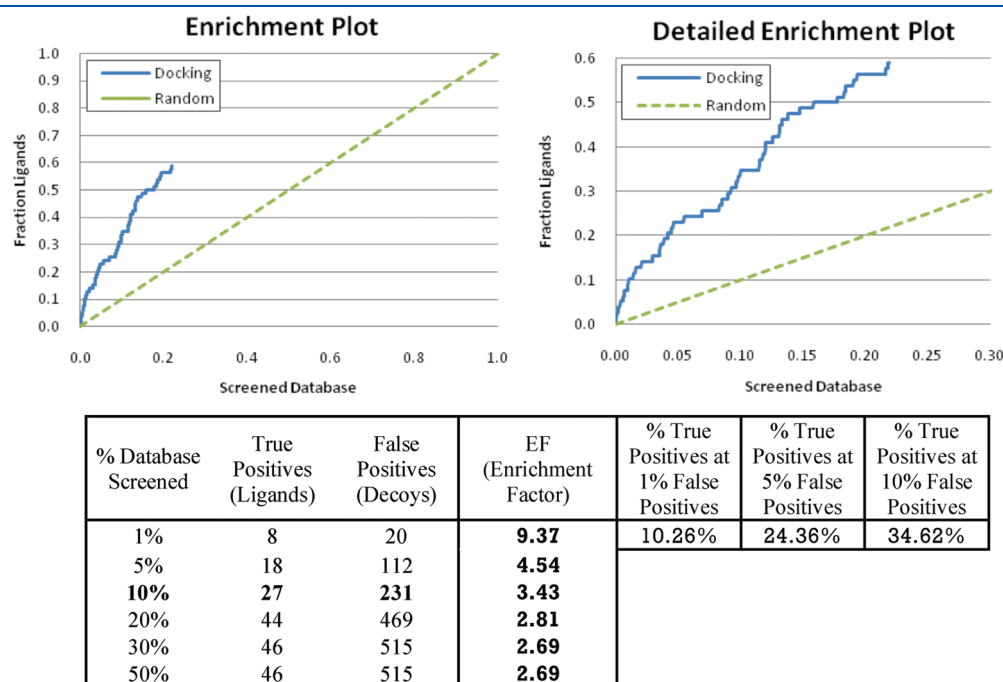
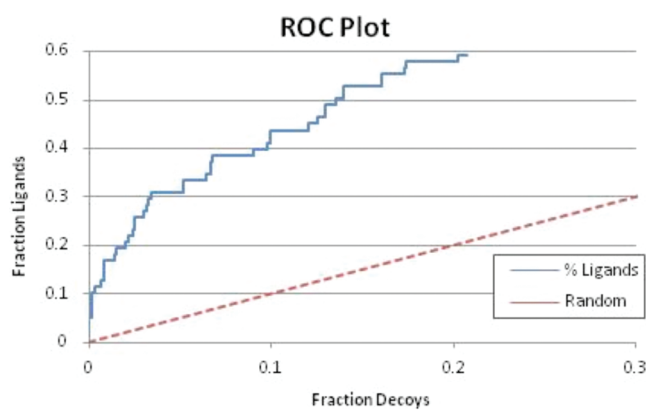


Figure 8. Enrichment plot obtained when the pharmacophore postfilter is applied to the docked poses ranked by Alpha HB scoring function. The curve only shows ligands and decoys whose docked poses match pharmacophoric requirements (22% of the total database). No value is returned for docked poses that do not match the pharmacophore constrains; therefore, they cannot be plotted. On the left, detailed enrichment plot for the first 30% of screened database. In the table, the plots show the EF at the first percentages of database screened.



% True Pos at 1% of False Pos	% True Pos at 5% of False Pos	% True Pos at 10% of False Pos.
16.67%	30.77%	41.03%

Figure 10. ROC plot obtained when postfiltered docked poses are ranked by MACCS similarity to the cognate ligand (Tanimoto index). The curve only shows ligands and decoys whose docked poses match pharmacophoric requirements. In the table, the ROC plot shows early recognition values.

molecules from inactives based on protein–ligand binding interactions. In many cases, docking techniques can predict correct ligand binding modes, but screening between high activity binders and decoys using scoring functions do not always provide good results. In those cases, it is highly convenient to use pharmacophoric filters to set the key interactions needed for a high activity binder and, hence, to discriminate whether a molecule is an active or a decoy, improving in this way ligands from decoy selection. Thus, docked poses well fitted in the binding site which also satisfy the pharmacophoric features ought to be actives. The most difficult aspect is to build a high selective pharmacophore model for filtering the database and discriminate all decoys from the actives. Moreover, if ligand information is added to docking results, such as a similarity search step, then VS is expected to perform better. In this work we combined structure-based and ligand-based information through three steps for VEGFR-2 retrospective VS. It has been shown that when docked poses were ranked by their Alpha HB binding energy values (best MOE scoring function found for this kind of protein–ligand interactions), selection between ligands and decoys was worse than random (AUC = 0.33), and early performance was also very low (true positives at 1% of false positives: 0.0%, see Figure 2). However when poses were reranked by MACCS molecular similarity (Tanimoto index) instead of binding energy, ligand selection improves significantly (AUC = 0.70) and early performance also did (% true positives at 1% of false positives: 14.10%, see Figures 2 and 3). When the pharmacophore postfilter was applied to the docked poses, 59.0% of the ligands and 20.8% of decoys were selected. In this case, the early performance becomes more effective (% true positives at 1% of false positives: 0.0% using only Alpha HB scoring function, and % true positives at 1% of false positives: 10.26% with pharmacophore postfilter, see Figures 2 and 9). ROC curves for pharmacophore postfiltered poses improved selection when the similarity search was also applied as a third step because similar structures are assumed to have similar biological activity. However, the highest improvement was

observed in the early performance values also applying the three steps (% true positives at 1% of false positives: 0.0% using only Alpha HB scoring function versus % true positives at 1% of false positives: 16.67% with pharmacophore postfilter and similarity postprocess, see Figures 2 and 10).

Previous VS approaches applied to the VEGFR-2 DUD test set show that it is difficult to achieve high performance when working with this target.^{17,18,48} Poor overall AUC has been retrieved in many cases: Kirchmair et al. have obtained AUC values from 0.43 to 0.48, using different scoring functions. Tawa et al. achieved AUC values from 0.599 to 0.614 using queries computed by ROCS. Enyedi et al.,³² working with pdb 1YWN and testing an in-house database with different scoring functions, achieved an AUC of 0.78 but using an IC₅₀ value cutoff. Moreover, there are not much early performance data published, and in all these cases very low values were achieved (EF values from 0 to 1.4 at 1%, EF values from 0 to 0.5 at 5%, and EF values from 0.4 to 1.1 at 10% in Kirchmair et al.). Here we show that by applying a three-step VS protocol, i.e., the conventional docking step, a pharmacophore postfilter step, and a similarity search postprocess, we get better early performance than that retrieved in previous work using only the most common scoring functions.

Furthermore, the VS protocol applied is able to better distinguish between correct and incorrect binding modes rather than using only a scoring function, which is logical given that we are using protein–ligand information in the second step. First-ranking binding modes retrieved by the protocol used here keep the main expected interactions for active ligands. Hence, this procedure ensures that selected ligands will have the same binding mode (key interactions) and presumably the same biological response than the query used. It is increasingly the case that pharmaceutical companies have multiple ligands for a given target and these bind in different binding modes to the same pocket. Using this procedure, we ensure that ligands selected bind in the same mode as the query used to select them.

CONCLUSION

Successful docking-based VS approaches must find good binding modes for active ligands and be able to discriminate actives from decoys using a scoring function. Previous VS approaches applied to VEGFR-2 inhibitors show that it is difficult to achieve high performance when working with the VEGFR-2 target. We carried out retrospective VS on the VEGFR-2 DUD data set with the aim to improve ligands from decoy selection. We first validated pose prediction in the VEGFR-2 target using the 1YWN complex (1.71 Å resolution) and the most accurate scoring function that matched the cognate ligand binding mode obtained from docking with the cocrystallized pose. This evaluation was performed by measuring the best rmsd value and best rmsd mean value for the top 10 rmsd poses using several MOE scoring and placement functions. Discrimination of ligands from decoys was then evaluated first through docking into the VEGFR-2 binding pocket, using the selected Alpha HB and Alpha Triangle scoring and placement functions. Docking results were postfiltered according to defined pharmacophore protein–ligand interactions, and the selected poses were again ranked according to their molecular similarity (MACCS fingerprint) to the 1YWN cognate ligand. Results show that by applying this three-step protocol, the overall performance (both AUC and EF) improves, mainly increasing the early performance. Moreover, performance metrics usually retrieve worse

values when an energy scoring function is used for ranking compounds rather than when a similarity search index is used. On the other hand, this VS protocol is better suited for finding correct ligand binding modes than using only a scoring function. First-ranking binding modes retrieved keep the main key protein–ligand interactions needed for active ligands. Hence, this procedure can be useful for prospective VS of novel VEGFR-2 inhibitors, mainly 1YWN VEGFR-2 cognate ligand analogues, with both chemical structure and binding mode similar to that of the 1YWN cognate ligand, and is an ongoing study.

■ ASSOCIATED CONTENT

S **Supporting Information.** Performance of cross-docking validation using all different MOE scoring and placement functions. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +34-93-267.20.00. Fax: +34-93-205.62.66. E-mail: violeta.perez@iqs.url.edu.

■ ACKNOWLEDGMENT

We are grateful to Chemical Computing Group, Inc., for the support provided. V.I.P.-N. thanks the Generalitat de Catalunya – DURSI for a grant within the Formació de Personal Investigador (2008FI) program.

■ REFERENCES

- (1) Pradeep, C. R.; Sunila, E. S.; Kuttan, G. Expression of vascular endothelial growth factor (VEGF) and VEGF receptors in tumor angiogenesis and malignancies. *Integr. Cancer Ther.* **2005**, *4*, 315–321.
- (2) Cee, V. J.; Albrecht, B. K.; Geuns-Meyer, S.; Hughes, P.; Bellon, S.; Bready, J.; Caenepeel, S.; Chaffee, S. C.; Coxon, A.; Emery, M. Alkynylpyrimidine amide derivatives as potent, selective, and orally active inhibitors of Tie-2 kinase. *J. Med. Chem.* **2007**, *4*, 627–640.
- (3) Hasegawa, M.; Nishigaki, N.; Washio, Y.; Kano, K.; Harris, P. A.; Sato, H.; Mori, L.; West, R. L.; Shibahara, M.; Toyoda, H. Discovery of novel benzimidazoles as potent inhibitors of TIE-2 and VEGFR-2 tyrosine kinase receptors. *J. Med. Chem.* **2007**, *18*, 4453–4470.
- (4) Pytel, D.; Sliwinski, T.; Poplawski, T.; Ferriola, D.; Majsterek, I. Tyrosine kinase blockers: new hope for successful cancer therapy. *Anti-Cancer Agent Me.* **2009**, *1*, 66–76.
- (5) Kiselyov, A. S.; Semenova, M.; Semenov, V. V. 1, 2, 3-Triazol-4-yl) benzenamines: Synthesis and activity against VEGF receptors 1 and 2. *Bioorg. Med. Chem. Lett.* **2009**, *5*, 1344–1348.
- (6) Peifer, C.; Selig, R.; Kinkel, K.; Ott, D.; Totzke, F.; Schachtele, C.; Heidenreich, R.; Rocken, M.; Schollmeyer, D.; Laufer, S. Design, synthesis, and biological evaluation of novel 3-aryl-4-(1H-indole-3-yl)-1,5-dihydro-2H-pyrrol-2-ones as vascular endothelial growth factor receptor (VEGF-R) inhibitors. *J. Med. Chem.* **2008**, *13*, 3814–3824.
- (7) Mannion, M.; Raeppl, S.; Claridge, S.; Zhou, N.; Saavedra, O.; Isakovic, L.; Zhan, L.; Gaudette, F.; Raeppl, F.; Déziel, R. N-(4-(6,7-Disubstituted-quinolin-4-yloxy)-3-fluorophenyl)-2-oxo-3-phenylimidazolidine-1-carboxamides: A novel series of dual c-Met/VEGFR2 receptor tyrosine kinase inhibitors. *Bioorg. Med. Chem. Lett.* **2009**, *23*, 6552–6556.
- (8) Ruel, R.; Thibeault, C.; L'Heureux, A.; Martel, A.; Cai, Z. W.; Wei, D.; Qian, L.; Barrish, J. C.; Mathur, A.; D'Arienzo, C. Discovery and preclinical studies of 5-isopropyl-6-(5-methyl-1, 3, 4-oxadiazol-2-yl)-N-(2-methyl-1H-pyrrolo [2, 3-b] pyridin-5-yl) pyrrolo [2, 1-f][1, 2, 4] triazin-4-amine (BMS-645737), an in vivo active potent VEGFR-2 inhibitor. *Bioorg. Med. Chem. Lett.* **2008**, *9*, 2985–2989.
- (9) Polverino, A.; Coxon, A.; Starnes, C.; Diaz, Z.; DeMelfi, T.; Wang, L.; Bready, J.; Estrada, J.; Cattley, R.; Kaufman, S. AMG 706, an oral, multikinase inhibitor that selectively targets vascular endothelial growth factor, platelet-derived growth factor, and kit receptors, potently inhibits angiogenesis and induces regression in tumor xenografts. *Cancer Res.* **2006**, *17*, 8715–8721.
- (10) Pujadas, G.; Vaque, M.; Ardevol, A.; Blade, C.; Salvado, M.; Blay, M.; Fernandez-Larrea, J.; Arola, L. Protein-ligand Docking: A Review of Recent Advances and Future Perspectives. *Curr. Pharm. Anal.* **2008**, *1*, 1–19.
- (11) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2009**, *6*, 1455–1474.
- (12) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.* **2009**, *4*, 1079–1093.
- (13) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *23*, 6789–6801.
- (14) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, *2*, 169–184.
- (15) Graves, A. P.; Brenk, R.; Shoichet, B. K. Decoys for docking. *J. Med. Chem.* **2005**, *11*, 3714–3728.
- (16) MOE (Molecular Operating Environment), 2008.10 release; Chemical Computing Group, Inc.: Montreal, Canada, 2004.
- (17) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How To Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information. *J. Chem. Inf. Model.* **2009**, *3*, 678–692.
- (18) Tawa, G. J.; Baber, J. C.; Humblet, C. Computation of 3D queries for ROCS based virtual screens. *J. Comput.-Aided Mol. Des.* **2009**, *12*, 853–868.
- (19) Cowan-Jacob, S. W. Structural biology of protein tyrosine kinases. *Cell. Mol. Life Sci.* **2006**, *22*, 2608–2625.
- (20) Berman, H.; Battistuz, T.; Bhat, T.; Bluhm, W.; Bourne, P.; Burkhardt, K.; Feng, Z.; Gilliland, G.; Iype, L.; Jain, S. The protein data bank. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *6*, 899–907.
- (21) Harris, P.; Cheung, M.; Hunter, R., 3rd; Brown, M.; Veal, J.; Nolte, R.; Wang, L.; Liu, W.; Crosby, R.; Johnson, J. Discovery and evaluation of 2-anilino-5-aryloxazoles as a novel class of VEGFR2 kinase inhibitors. *J. Med. Chem.* **2005**, *5*, 1610–1619.
- (22) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. 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.* **2008**, *3*, 213–228.
- (23) Tuccinardi, T.; Botta, M.; Giordano, A.; Martinelli, A. Protein Kinases: Docking and Homology Modeling Reliability. *J. Chem. Inf. Model.* **2010**, *8*, 1432–1441.
- (24) Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. Optimal assignment methods for ligand-based virtual screening. *J. Cheminf.* **2009**, *1*, 14–37.
- (25) Peach, M. L.; Nicklaus, M. C. Combining docking with pharmacophore filtering for improved virtual screening. *J. Cheminf.* **2009**, *1*, 6–21.
- (26) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *3*, 793–806.
- (27) Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection?. *J. Comput.-Aided Mol. Des.* **2008**, *3*, 169–178.
- (28) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *3*, 133–139.

- (29) Mackey, M. D.; Melville, J. L. Better than Random? The Chemotype Enrichment Problem. *J. Chem. Inf. Model.* **2009**, *5*, 1154–1162.
- (30) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual screening workflow development guided by the “receiver operating characteristic” curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *7*, 2534–2547.
- (31) Liebeschuetz, J. W. Evaluating docking programs: keeping the playing field level. *J. Comput.-Aided Mol. Des.* **2008**, *3*, 229–238.
- (32) Enyedy, I. J.; Egan, W. J. Can we use docking and scoring for hit-to-lead optimization?. *J. Comput.-Aided Mol. Des.* **2008**, *3*, 161–168.
- (33) Miyazaki, Y.; Matsunaga, S.; Tang, J.; Maeda, Y.; Nakano, M.; Philippe, R. J.; Shibahara, M.; Liu, W.; Sato, H.; Wang, L. Novel 4-amino-furo [2, 3-d] pyrimidines as Tie-2 and VEGFR2 dual inhibitors. *Bioorg. Med. Chem. Lett.* **2005**, *9*, 2203–2207.
- (34) Bostrom, J.; Hogner, A.; Schmitt, S. Do structurally similar ligands bind in a similar fashion?. *J. Med. Chem.* **2006**, *23*, 6716–6725.
- (35) Wolber, G.; Seidel, T.; Bendix, F.; Langer, T. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discovery Today* **2008**, *13*, 23–29.
- (36) Johnson, M. A.; Maggiora, G. M. *Concepts and applications of molecular similarity*; Wiley: New York: 1990; pp 1–13.
- (37) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today* **2007**, *5–6*, 225–233.
- (38) Kortagere, S.; Krasowski, M. D.; Ekins, S. The importance of discerning shape in molecular pharmacology. *Trends Pharmacol. Sci.* **2009**, *3*, 138–147.
- (39) Ghose, A. K.; Herbertz, T.; Salvino, J. M.; Mallamo, J. P. Knowledge-based chemoinformatic approaches to drug discovery. *Drug Discovery Today* **2006**, *23–24*, 1107–1114.
- (40) Neaz, M. M.; Pasha, F. A.; Muddassar, M.; Lee, S. H.; Sim, T.; Hah, J. M.; Cho, S. J. Pharmacophore based 3D-QSAR study of VEGFR-2 inhibitors. *Med. Chem. Res.* **2009**, *2*, 127–142.
- (41) Yu, H.; Wang, Z.; Zhang, L.; Zhang, J.; Huang, Q. Pharmacophore modeling and in silico screening for new KDR kinase inhibitors. *Bioorg. Med. Chem. Lett.* **2007**, *8*, 2126–2133.
- (42) Sammond, D. M.; Nailor, K. E.; Veal, J. M.; Nolte, R. T.; Wang, L.; Knick, V. B.; Rudolph, S. K.; Truesdale, A. T.; Nartey, E. N.; Stafford, J. A. Discovery of a novel and potent series of dianilinopyrimidineurea and urea isostere inhibitors of VEGFR2 tyrosine kinase. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3519–3523.
- (43) Gracias, V.; Ji, Z.; Akritopoulou-Zanze, I.; Abad-Zapatero, C.; Huth, J. R.; Song, D.; Hajduk, P. J.; Johnson, E. F.; Glaser, K. B.; Marcotte, P. A. Scaffold oriented synthesis. Part 2: Design, synthesis and biological evaluation of pyrimido-diazepines as receptor tyrosine kinase inhibitors. *Bioorg. Med. Chem. Lett.* **2008**, *8*, 2691–2695.
- (44) Sridhar, J.; Akula, N.; Sivanesan, D.; Narasimhan, M.; Rathinavelu, A.; Pattabiraman, N. Identification of novel angiogenesis inhibitors. *Bioorg. Med. Chem. Lett.* **2005**, *18*, 4125–4129.
- (45) Harris, P. A.; Boloor, A.; Cheung, M.; Kumar, R.; Crosby, R. M.; Davis-Ward, R. G.; Epperly, A. H.; Hinkle, K. W.; Hunter, R. N., III; Johnson, J. H. Discovery of 5-[[4-[(2,3-Dimethyl-2H-indazol-6-yl)methylamino]-2-pyrimidinyl]amino]-2-methylbenzenesulfonamide (Pazopanib), a Novel and Potent Vascular Endothelial Growth Factor Receptor Inhibitor. *J. Med. Chem.* **2008**, *15*, 4632–4640.
- (46) Potashman, M. H.; Bready, J.; Coxon, A.; DeMelfi, T. M., Jr.; DiPietro, L.; Doerr, N.; Elbaum, D.; Estrada, J.; Gallant, P.; Germain, J. Design, synthesis, and evaluation of orally active benzimidazoles and benzoxazoles as vascular endothelial growth factor-2 receptor tyrosine kinase inhibitors. *J. Med. Chem.* **2007**, *18*, 4351–4373.
- (47) Dakshanamurthy, S.; Kim, M.; Brown, M. L.; Byers, S. W. In-silico fragment-based identification of novel angiogenesis inhibitors. *Bioorg. Med. Chem. Lett.* **2007**, *16*, 4551–4556.
- (48) Zhou, Z.; Felts, A. K.; Friesner, R. A.; Levy, R. M. Comparative performance of several flexible docking programs and scoring functions: enrichment studies for a diverse set of pharmaceutically relevant targets. *J. Chem. Inf. Model.* **2007**, *4*, 1599–1608.

ASE-Alpha PMI

Ligands											
PDB	1Y6A	1YWN	2OH4	2P2H	2QU5	2RL5	3BE2	3C7Q	3CJF	3EWH	3CJG
1Y6A		5.4	6.2	5.3	7.0	5.2	6.2	7.8	4.2	5.0	8.8
1YWN	5.6		4.9	5.1	6.9	1.0	5.6	8.5	4.6	2.1	4.8
2OH4	5.5	6.5		4.8	2.2	2.2	1.0	7.3	4.8	5.2	3.0
2P2H	4.8	6.5	5.6		4.0	5.1	6.3	8.4	6.1	7.8	4.9
2QU5	4.8	1.8	2.8	3.7		1.8	1.9	6.5	4.7	5.8	5.5
2RL5	4.1	4.8	4.9	4.3	10.4		1.3	6.8	5.5	2.0	7.6
3BE2	6.4	5.2	3.4	3.6	6.0	2.9		4.4	2.3	1.9	2.9
3C7Q	4.5	7.1	2.3	4.3	6.7	2.9	4.8		3.9	3.2	4.0
3CJF	3.5	3.0	3.1	3.8	10.2	3.2	7.3	7.9		6.0	3.2
3EWH	3.6	4.7	1.2	4.7	1.6	2.2	5.4	6.9	6.1		3.6
3CJG	6.2	5.9	5.3	5.0	7.1	3.0	6.5	7.7	4.4	8.7	

aRMSD: 4.9

Std Dev: 2.0

num. poses $\leq 2\text{\AA}$: 10

ASE-Alpha Triangle

Ligands											
PDB	1Y6A	1YWN	2OH4	2P2H	2QU5	2RL5	3BE2	3C7Q	3CJF	3EWH	3CJG
1Y6A		5.7	5.8	1.1	7.0	5.7	8.2	5.1	4.3	6.7	3.1
1YWN	3.2		5.9	2.4	7.9	1.9	1.6	6.5	4.7	1.2	3.8
2OH4	3.4	5.5		1.8	2.2	0.8	2.1	6.6	4.4	6.8	1.2
2P2H	1.1	5.4	6.4		4.7	5.9	6.6	1.7	5.2	9.0	2.9
2QU5	4.4	5.2	1.9	1.7		1.1	2.2	2.9	4.7	6.6	2.1
2RL5	3.1	0.9	2.3	4.4	1.8		2.2	2.6	4.1	1.5	1.7
3BE2	1.8	1.6	1.3	1.8	2.8	1.4		2.7	2.3	2.3	3.1
3C7Q	4.9	6.7	2.6	2.6	6.4	1.8	2.8		1.9	2.9	2.5
3CJF	3.7	3.6	2.4	3.2	5.5	2.3	5.1	4.1		4.6	4.0
3EWH	3.1	2.0	1.0	1.8	1.5	1.9	1.4	6.6	4.8		1.3
3CJG	4.6	7.3	5.9	3.9	4.7	3.7	6.3	3.9	4.2	6.6	

aRMSD: 3.7

Std Dev: 2.0

num. poses $\leq 2\text{\AA}$: 29

ASE-Proxy Triangle

Ligands											
PDB	1Y6A	1YWN	2OH4	2P2H	2QU5	2RL5	3BE2	3C7Q	3CJF	3EWH	3CJG
1Y6A		6.1	6.1	2.1	7.2	6.3	5.6	2.1	3.9	5.6	2.0
1YWN	3.3		3.0	3.8	2.4	1.1	1.5	2.4	4.6	2.2	2.3
2OH4	4.2	4.7		1.3	2.1	0.8	1.3	5.8	4.0	1.1	1.7
2P2H	3.8	4.8	5.3		4.8	4.8	6.5	1.9	2.2	5.8	1.9
2QU5	4.5	5.3	2.3	2.2		1.2	2.1	5.3	4.4	2.9	2.2
2RL5	4.1	1.2	1.5	1.5	1.6		1.0	2.4	4.4	1.4	0.8
3BE2	4.5	1.2	1.7	1.7	1.6	1.6		2.6	2.1	1.9	3.1
3C7Q	5.4	7.0	2.8	1.8	2.0	1.8	2.4		3.2	2.3	2.7
3CJF	4.5	3.0	3.7	4.3	5.0	2.4	5.6	5.2		3.1	1.8
3EWH	4.1	1.0	1.8	1.9	2.1	1.2	2.7	6.6	3.6		1.6
3CJG	3.9	5.9	5.1	4.0	5.6	3.3	5.9	3.4	3.6	5.9	

aRMSD: 3.3

Std Dev: 1.7

num. poses $\leq 2\text{\AA}$: 33

ASE-Triangle Matcher

Ligands											
PDB	1Y6A	1YWN	2OH4	2P2H	2QU5	2RL5	3BE2	3C7Q	3CJF	3EWH	3CJG
1Y6A		6.1	6.1	2.1	7.2	4.6	5.6	2.4	3.9	5.6	0.7
1YWN	3.3		3.0	3.8	2.4	1.1	1.5	1.8	5.1	2.1	2.3
2OH4	4.2	4.4		1.6	2.1	0.9	1.3	8.4	4.0	1.1	1.2
2P2H	3.8	4.8	6.0		4.6	4.8	6.2	2.3	4.8	5.8	1.2
2QU5	4.0	5.1	1.9	1.9		1.2	2.1	5.3	4.4	3.3	1.9
2RL5	4.0	1.2	1.4	1.9	2.5		1.6	2.4	4.4	1.4	1.6
3BE2	4.5	1.2	1.0	3.1	1.9	1.6		2.9	1.4	1.9	1.7
3C7Q	5.4	7.2	2.0	1.4	2.0	1.4	1.9		3.2	3.3	3.7
3CJF	3.8	3.0	3.7	1.7	5.2	2.3	5.6	5.2		3.3	1.8
3EWH	4.1	1.0	1.4	1.6	2.1	1.2	1.7	6.6	3.6		1.9
3CJG	3.9	6.6	5.1	4.0	5.6	4.7	5.9	3.4	3.6	5.9	

aRMSD: 3.3

Std Dev: 1.8

num. poses $\leq 2\text{\AA}$: 40

Affinity dG-Alpha PMI

Ligands											
PDB	1Y6A	1YWN	2OH4	2P2H	2QU5	2RL5	3BE2	3C7Q	3CJF	3EWH	3CJG
1Y6A		5.4	6.2	5.3	6.7	5.1	6.2	7.4	5.9	4.7	8.6
1YWN	5.5		1.3	5.4	6.9	1.7	5.2	8.5	5.6	2.1	5.2
2OH4	3.9	6.5		4.8	1.9	2.2	1.0	7.3	6.1	5.2	3.0
2P2H	4.8	6.5	5.6		4.0	5.1	6.3	8.4	6.1	5.7	4.9
2QU5	4.7	1.8	2.8	3.7		1.8	3.6	6.9	4.6	5.8	5.5
2RL5	5.3	5.1	5.1	4.3	10.4		1.3	4.4	5.5	2.0	7.6
3BE2	6.4	5.2	3.4	3.6	6.0	2.9		4.2	2.4	1.9	5.1
3C7Q	4.5	7.9	4.3	4.3	6.4	1.8	4.6		2.1	3.2	4.0
3CJF	3.5	3.0	3.7	3.8	4.5	3.6	7.3	7.8		6.0	5.1
3EWH	3.6	4.7	2.8	4.9	1.6	2.2	5.4	6.9	6.5		3.6
3CJG	5.4	5.9	5.9	5.1	7.1	3.3	6.5	7.7	4.3	8.7	

aRMSD: 4.9

Std Dev: 1.9

num. poses $\leq 2\text{\AA}$: 11**Affinity dG-Alpha Triangle**

Ligands											
PDB	1Y6A	1YWN	2OH4	2P2H	2QU5	2RL5	3BE2	3C7Q	3CJF	3EWH	3CJG
1Y6A		6.6	6.1	2.4	6.7	5.0	6.8	3.5	2.0	6.7	2.5
1YWN	3.2		1.3	2.1	7.9	3.1	9.4	2.6	1.5	2.9	2.5
2OH4	3.4	7.8		1.6	1.9	0.9	3.1	3.4	4.4	7.4	0.9
2P2H	2.5	7.5	6.9		5.2	7.3	8.2	3.1	1.1	6.9	1.5
2QU5	4.5	5.2	1.9	1.7		1.1	2.4	2.9	4.7	8.7	2.1
2RL5	4.0	5.1	3.3	1.4	1.9		2.1	2.6	4.1	3.0	0.7
3BE2	2.8	6.4	2.9	2.8	7.9	3.6		2.8	2.4	8.1	2.5
3C7Q	4.3	7.2	3.9	4.8	7.7	2.1	2.0		1.6	5.3	2.4
3CJF	3.8	6.0	5.3	2.2	6.0	3.6	7.3	3.1		6.1	3.0
3EWH	4.0	2.4	1.0	1.8	1.9	0.9	1.6	3.2	3.7		1.9
3CJG	3.7	6.8	6.5	2.3	6.4	5.8	8.7	2.7	4.8	6.8	

aRMSD: 4.0

Std Dev: 2.2

num. poses $\leq 2\text{\AA}$: 22

Affinity dG-Proxy Triangle

Ligands											
PDB	1Y6A	1YWN	2OH4	2P2H	2QU5	2RL5	3BE2	3C7Q	3CJF	3EWH	3CJG
1Y6A		6.1	6.2	3.6	7.1	6.2	6.8	2.5	1.7	6.0	3.2
1YWN	4.8		2.6	3.1	2.3	1.1	3.5	2.4	3.4	3.7	2.5
2OH4	4.1	6.5		1.2	2.0	1.0	2.9	4.2	3.7	2.5	1.2
2P2H	3.8	5.9	6.3		5.1	4.8	7.5	2.4	2.2	5.7	2.4
2QU5	4.8	8.1	1.4	7.6		1.9	3.1	7.3	4.9	6.2	2.1
2RL5	4.2	6.1	1.1	3.7	1.7		1.9	1.4	4.2	1.6	1.9
3BE2	3.3	1.2	1.5	1.3	1.9	1.5		2.4	3.0	2.2	1.2
3C7Q	5.5	8.5	1.2	4.7	1.3	1.6	3.8		3.2	3.4	1.7
3CJF	3.8	5.7	4.9	1.6	6.6	2.5	6.9	3.1		4.9	2.3
3EWH	3.6	2.4	1.7	2.3	2.2	1.7	2.3	3.4	1.8		1.7
3CJG	4.1	7.1	5.6	3.2	5.7	5.5	7.1	1.2	3.8	6.5	

aRMSD: 3.6

Std Dev: 2.0

num. poses $\leq 2\text{\AA}$: 29**Affinity dG-Triangle Matcher**

Ligands											
PDB	1Y6A	1YWN	2OH4	2P2H	2QU5	2RL5	3BE2	3C7Q	3CJF	3EWH	3CJG
1Y6A		6.3	6.2	3.6	8.0	5.9	6.8	2.5	1.7	6.0	3.2
1YWN	4.8		2.1	3.1	2.3	1.1	3.5	2.4	3.4	3.8	2.5
2OH4	4.1	6.5		1.1	2.0	1.0	2.9	4.2	4.0	0.9	1.1
2P2H	3.8	5.9	6.3		4.8	4.8	7.4	2.4	2.3	5.2	1.9
2QU5	4.8	7.6	1.4	6.6		2.4	2.8	6.7	4.5	5.1	2.1
2RL5	4.2	5.4	1.1	3.6	1.9		2.8	1.9	4.1	3.0	1.7
3BE2	1.7	3.6	1.9	2.5	2.1	0.6		2.4	3.0	2.8	1.9
3C7Q	5.5	8.5	1.2	1.5	1.3	1.6	2.7		3.2	3.3	2.4
3CJF	4.5	5.7	5.4	1.6	6.9	3.9	5.0	3.1		4.1	2.8
3EWH	3.6	2.4	1.7	2.3	2.1	1.5	2.3	3.4	1.8		1.7
3CJG	3.8	6.4	5.6	3.2	5.7	5.5	7.1	1.2	3.7	6.5	

aRMSD: 3.6

Std Dev: 1.9

num. poses $\leq 2\text{\AA}$: 26

Alpha HB-Alpha PMI

Ligands											
PDB	1Y6A	1YWN	2OH4	2P2H	2QU5	2RL5	3BE2	3C7Q	3CJF	3EWH	3CJG
1Y6A		5.4	6.2	5.3	6.7	5.1	6.2	7.4	5.9	4.7	8.6
1YWN	5.5		4.9	5.4	6.9	1.7	5.2	8.5	5.6	2.1	5.2
2OH4	3.9	6.5		4.8	2.2	2.2	1.0	7.3	4.8	5.0	3.0
2P2H	4.8	6.5	5.6		4.0	5.1	6.3	8.4	6.1	5.7	4.9
2QU5	4.7	1.8	2.8	3.7		6.2	3.6	6.9	4.6	5.8	5.5
2RL5	5.3	5.1	5.1	4.3	10.4		2.6	4.4	5.5	2.0	7.6
3BE2	6.4	5.2	3.4	3.6	6.0	2.9		4.4	2.4	1.9	5.0
3C7Q	4.5	7.9	4.3	4.3	6.4	2.9	4.6		2.1	3.2	4.0
3CJF	6.5	9.2	3.6	3.8	4.5	3.6	7.3	7.8		6.0	5.1
3EWH	3.6	4.7	2.8	4.9	10.7	2.2	5.4	6.8	4.1		3.6
3CJG	5.4	5.9	5.9	5.0	7.1	3.0	6.5	5.5	4.3	8.7	

aRMSD: 5.1

Std Dev: 1.9

num. poses $\leq 2\text{\AA}$: 5**Alpha HB-Alpha Triangle**

Ligands											
PDB	1Y6A	1YWN	2OH4	2P2H	2QU5	2RL5	3BE2	3C7Q	3CJF	3EWH	3CJG
1Y6A		5.3	6.4	1.2	7.0	5.5	5.6	3.0	4.9	5.8	1.1
1YWN	3.2		1.2	1.6	2.0	1.4	1.4	1.9	2.9	1.3	1.5
2OH4	3.3	4.9		1.6	1.9	0.9	1.6	2.2	4.3	2.8	1.1
2P2H	4.7	5.1	4.4		4.8	5.5	6.5	2.8	1.7	5.9	1.7
2QU5	4.4	5.8	1.7	3.3		1.7	3.0	3.4	4.6	3.0	1.9
2RL5	4.7	0.8	0.9	1.4	1.8		1.7	2.5	4.5	1.2	0.9
3BE2	1.8	1.8	2.4	1.0	1.3	1.3		4.0	3.1	2.7	1.3
3C7Q	4.4	2.1	2.9	4.8	3.0	1.6	1.7		1.9	2.7	4.5
3CJF	3.1	5.8	4.2	2.5	4.2	4.5	3.6	2.3		3.5	1.6
3EWH	3.2	1.2	1.0	1.3	2.2	1.4	1.3	3.4	4.1		1.6
3CJG	4.2	5.5	1.6	1.5	5.6	3.5	5.4	3.2	1.6	5.1	

aRMSD: 3.0

Std Dev: 1.6

num. poses $\leq 2\text{\AA}$: 45

Alpha HB-Proxy Triangle

Ligands											
PDB	1Y6A	1YWN	2OH4	2P2H	2QU5	2RL5	3BE2	3C7Q	3CJF	3EWH	3CJG
1Y6A		6.3	5.9	5.3	7.0	4.9	5.9	2.1	1.7	5.0	0.8
1YWN	4.3		1.7	2.9	2.5	1.1	2.7	1.4	4.0	2.1	2.1
2OH4	4.1	4.8		2.2	2.1	0.9	1.4	3.6	4.0	1.1	0.9
2P2H	3.8	4.9	5.4		4.6	4.8	6.7	2.1	2.5	5.2	1.6
2QU5	4.8	8.8	1.4	6.1		1.3	2.5	6.4	4.4	4.0	1.9
2RL5	4.1	1.3	1.1	1.4	1.8		2.2	1.7	4.3	2.2	1.5
3BE2	3.8	2.8	1.0	2.1	2.1	1.5		2.4	1.8	2.9	1.9
3C7Q	5.5	3.8	1.4	1.6	1.2	1.7	1.4		2.1	2.5	1.4
3CJF	3.8	3.1	4.2	1.6	5.3	2.5	4.1	3.6		3.8	1.8
3EWH	3.7	1.0	2.6	1.7	2.1	0.8	1.8	1.9	1.6		1.7
3CJG	4.3	6.8	5.1	1.5	5.7	4.2	5.5	1.9	2.3	6.0	

aRMSD: 3.1

Std Dev: 1.7

num. poses $\leq 2\text{\AA}$: 40**Alpha HB-Triangle Matcher**

Ligands											
PDB	1Y6A	1YWN	2OH4	2P2H	2QU5	2RL5	3BE2	3C7Q	3CJF	3EWH	3CJG
1Y6A		6.3	6.0	5.3	7.0	4.9	6.2	2.1	1.7	5.0	1.9
1YWN	3.0		1.4	2.9	2.2	1.1	2.7	1.4	4.0	2.1	2.1
2OH4	4.1	4.8		2.2	2.1	0.9	1.4	3.6	4.0	2.0	0.7
2P2H	3.8	4.9	5.5		4.6	4.8	6.7	1.7	1.6	5.3	1.6
2QU5	4.6	7.6	1.5	6.4		1.3	2.1	6.4	4.6	6.2	2.0
2RL5	3.8	1.1	1.1	1.8	1.5		1.6	2.1	4.6	1.4	1.7
3BE2	1.7	2.1	1.0	1.7	2.0	1.7		2.4	2.8	3.2	1.4
3C7Q	5.5	3.8	1.9	1.3	1.2	1.7	1.4		2.1	3.6	1.4
3CJF	2.1	3.1	4.0	1.6	5.3	3.9	4.1	3.6		4.7	1.8
3EWH	3.7	1.0	2.1	1.7	2.1	1.1	1.8	2.3	1.6		1.7
3CJG	4.3	7.3	5.1	1.5	5.7	4.2	5.5	1.9	1.8	6.0	

aRMSD: 3.1

Std Dev: 1.8

num. poses $\leq 2\text{\AA}$: 44

London dG-Alpha PMI

Ligands											
PDB	1Y6A	1YWN	2OH4	2P2H	2QU5	2RL5	3BE2	3C7Q	3CJF	3EWH	3CJG
1Y6A		5.5	6.2	5.3	6.7	5.1	6.2	7.4	5.9	5.8	8.6
1YWN	5.5		4.9	5.4	1.6	1.7	5.2	8.6	5.5	2.3	5.2
2OH4	3.9	6.5		4.8	2.2	2.2	1.0	7.2	4.8	5.0	3.0
2P2H	4.8	6.5	5.6		4.0	5.1	5.8	7.0	4.1	5.7	4.9
2QU5	4.7	1.8	2.8	3.7		6.2	3.6	6.9	4.6	5.8	5.5
2RL5	5.3	5.1	5.1	4.3	10.4		2.6	4.4	5.5	2.0	7.6
3BE2	6.4	5.2	3.4	3.2	6.0	2.9		4.2	2.4	1.9	5.0
3C7Q	4.5	7.9	4.2	4.2	6.4	2.9	4.6		2.2	3.2	4.0
3CJF	3.5	9.2	3.6	3.8	4.5	3.6	7.3	7.8		6.0	3.1
3EWH	3.6	4.7	2.8	4.9	10.7	2.2	5.4	6.8	4.1		3.0
3CJG	5.4	5.9	5.9	5.0	4.5	3.3	6.5	5.5	4.4	8.7	

aRMSD: 4.9

Std Dev: 1.9

num. poses $\leq 2\text{\AA}$: 6

London dG-Alpha Triangle

Ligands											
PDB	1Y6A	1YWN	2OH4	2P2H	2QU5	2RL5	3BE2	3C7Q	3CJF	3EWH	3CJG
1Y6A		6.6	5.5	1.1	6.3	5.0	5.7	3.2	3.7	5.0	0.9
1YWN	3.2		1.5	2.6	2.1	1.8	2.1	2.4	3.6	1.1	1.3
2OH4	3.4	6.9		1.6	1.9	0.8	3.1	3.1	4.4	2.9	1.7
2P2H	3.8	5.6	4.9		4.9	4.6	7.5	2.1	1.3	6.2	1.1
2QU5	4.2	5.2	1.8	1.7		1.2	2.4	2.9	4.7	8.6	2.1
2RL5	3.1	0.8	1.7	1.4	1.9		1.6	1.9	4.1	3.0	1.6
3BE2	3.0	2.0	1.3	3.6	2.8	1.4		1.5	4.0	3.0	0.8
3C7Q	4.3	5.2	2.3	1.6	3.4	1.9	2.0		2.3	1.6	1.9
3CJF	3.3	3.9	3.8	1.2	6.0	2.5	7.3	2.4		3.4	1.1
3EWH	4.0	2.1	1.0	1.9	1.7	0.9	1.2	4.0	4.3		1.7
3CJG	3.7	6.3	4.4	2.8	5.5	3.6	6.2	2.5	2.0	6.9	

aRMSD: 3.1

Std Dev: 1.8

num. poses $\leq 2\text{\AA}$: 41

London dG-Proxy Triangle

Ligands											
PDB	1Y6A	1YWN	2OH4	2P2H	2QU5	2RL5	3BE2	3C7Q	3CJF	3EWH	3CJG
1Y6A		6.0	5.8	3.6	7.4	5.5	5.6	2.5	1.7	5.0	0.7
1YWN	4.0		1.5	3.7	1.8	1.1	2.5	1.8	3.4	2.3	2.6
2OH4	4.1	4.8		1.0	2.1	0.6	1.3	3.1	4.0	2.0	1.2
2P2H	5.1	3.5	4.9		4.6	4.8	7.0	1.7	2.7	5.2	2.2
2QU5	4.6	6.2	1.4	3.7		1.3	3.4	6.5	4.4	3.2	2.0
2RL5	3.7	1.7	1.1	2.1	1.7		3.0	3.0	4.1	1.4	1.4
3BE2	1.7	1.1	1.0	1.9	2.0	1.8		1.6	2.6	2.0	1.7
3C7Q	5.7	4.3	1.4	1.4	1.3	1.6	3.1		2.1	2.1	1.4
3CJF	3.8	3.1	2.9	1.6	5.3	2.7	4.6	3.6		4.1	1.5
3EWH	1.9	2.9	1.2	1.1	2.0	0.9	1.4	2.3	3.3		1.2
3CJG	4.3	5.7	5.2	4.4	5.7	4.2	5.5	1.9	2.3	5.4	

aRMSD: 3.0

Std Dev: 1.6

num. poses $\leq 2\text{\AA}$: 43

London dG-Triangle Matcher

Ligands											
PDB	1Y6A	1YWN	2OH4	2P2H	2QU5	2RL5	3BE2	3C7Q	3CJF	3EWH	3CJG
1Y6A		6.0	6.1	3.6	7.4	5.0	5.5	2.5	1.7	5.0	0.7
1YWN	4.0		1.5	3.7	1.8	1.1	2.5	1.6	3.4	2.3	2.6
2OH4	4.1	4.8		1.0	2.1	1.1	1.3	3.1	4.0	2.0	0.8
2P2H	5.1	3.5	4.7		4.7	4.8	7.0	1.7	3.1	5.2	1.6
2QU5	4.6	6.2	1.5	6.6		1.3	2.1	6.5	4.3	5.1	2.3
2RL5	3.7	1.5	1.1	3.6	1.5		2.3	3.1	2.9	1.4	1.2
3BE2	4.4	2.8	1.0	1.6	1.7	1.5		2.4	2.6	2.0	1.4
3C7Q	5.7	4.3	1.7	1.6	1.3	1.6	2.5		2.1	3.6	1.2
3CJF	3.8	3.0	4.0	1.6	5.3	2.6	3.9	1.5		3.3	1.5
3EWH	4.1	2.9	1.7	1.1	2.1	0.9	1.7	2.3	3.3		1.2
3CJG	4.3	5.7	5.2	4.4	5.7	4.5	5.5	1.9	2.3	5.4	

aRMSD: 3.1

Std Dev: 1.7

num. poses $\leq 2\text{\AA}$: 40

Table S-1. Cross-docking validation. Best RMSD values obtained when cognate ligands (columns) were docked in the different PDBs protein structures (rows) using the different scoring functions (bold) and placement functions. Protein-ligand complexes were aligned prior to perform cross-dockings and the resulting docked poses were compared with their respective cognate pose.

^a PDB used do not contain allosteric ligands.

^b All values are expressed in \AA .

Artículo 2



Contents lists available at SciVerse ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Impact of the CXCR4 structure on docking-based virtual screening of HIV entry inhibitors

Jesús M. Planesas^{a,1}, Violeta I. Pérez-Nueno^{a,b,*}, José I. Borrell^{a,1}, Jordi Teixidó^{a,1}^a Grup d'Enginyeria Molecular, Institut Químic de Sarrià (IQS), Universitat Ramon Llull, Barcelona, Spain^b INRIA Nancy, LORIA, 615 rue du Jardin Botanique, 54600 Villers-lès-Nancy, France

ARTICLE INFO

Article history:

Accepted 25 June 2012

Available online 3 July 2012

Keywords:

CXCR4

Retrospective docking-based screening

Early recovery

Scaffold retrieval

ABSTRACT

Herein we analyze in depth the receptor-based virtual screening (VS) performance of the five recent crystallized CXCR4 structures along with a CXCR4 rhodopsin-based homology model. All CXCR4 Protein Data Bank (PDB) structures are co-crystallized with a small organic antagonist except structure 3OE0, which is co-crystallized with a cyclic peptide analog. Evaluation of the CXCR4 models was done by retrospective docking-based VS using a test set of 248 known CXCR4 inhibitors from 4 different chemotype families and 4696 different presumed inactives. The performance of the docking protocol using the five different protein structures was assessed in terms of pose prediction and hits detection using 12 different docking scoring functions and a scoring function with rescoring. Results show that 3OE6 structure achieves the highest docking-based performance with an average area under the curve (aAUC) of 0.84 and an average enrichment factor (aEF) of 11.7 at 1% of decoys screened. CXCR4 rhodopsin-like homology model performs comparable to the crystallized structures in the 1% of database screened. Moreover, a detailed analysis of the retrospective docking results using the CXCR4 homology model in Discovery Studio allows us to hypothesize the existence of multiple binding sub-sites in CXCR4 binding pocket.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

HIV cell infection is a multi-stage and complex process. It starts with the virus entry [1] into the host membrane. The first stage of the entry process is the virus attachment to the cell membrane through receptor CD4. Secondly, viral proteins *gp120* [2] and *gp41* bind to the co-receptors CCR5 and CXCR4 [3] to facilitate virus anchoring to the host cellular membrane. Finally, the viral envelope undergoes conformational changes that allow the insertion of the *gp41* fusion peptide into the host membrane.

One of the strategies to block the HIV entry process is inhibiting the binding of the viral protein *gp120* to the co-receptors CXCR4 or CCR5. The first interaction between *gp120* and these co-receptors is through their extracellular loops. CXCR4 and CCR5 are both chemokine receptors which belong to the GPCR superfamily [4] and are implicated in a wide range of human diseases [5]. This sub-family of GPCR proteins is responsible for transmitting signals from

the extracellular environment as well as the recognition of extracellular ligands. All GPCRs have a common characteristic structure relatively hydrophobic with 7 transmembrane domains. The N-terminus is located in the extracellular side and the transmembrane helices are connected by three extracellular loops whereas the C-terminus is located in the intracellular side facing the cytoplasm and three intracellular loops connect the interior domain of the transmembrane helices (Fig. 1). The transmembrane helices are expanded in the plasmatic membrane as anticlockwise bundle of alpha-helices. Therefore extracellular loops fulfill important roles in receptor activation and consequently in ligand binding [6].

An important achievement was accomplished at the end of 2010 when Wu et al. [7] crystallized five independent CXCR4 structures with a resolution of 2.5–3.2 Å (Table 1). These crystal structures contain a small organic molecule derived from isothiourea [8] (IT1t) in Protein Data Bank structures (PDBs) 3ODU, 3OE6, 3OE8 and 3OE9, and a 16-residue cyclic peptide [9] (CVX15) in PDB 3OE0, as antagonists bound to the CXCR4 structure.

In all the PDBs we can observe that the ligand binding site is located in the second extracellular loop, ECL2, close to the N-terminus (Fig. 2). Also, superposition of ligand binding pockets shows similar overlapping volume. The residues that induce common interactions in both ligands are Tyr116, Asp187 and Glu288 (Fig. 3). Those residues are common in CXCR4 antagonist binding as it has been widely discussed in previous literature

* Corresponding author at: INRIA Nancy, LORIA, 615 rue du Jardin Botanique, 54600 Villers-lès-Nancy, France. Tel.: +33 3 83593045/+34 93 267 20 00; fax: +33 3 83413079/+34 93 205 62 66.

E-mail addresses: violeta.pereznueno@inria.fr, violeta.perez@iqs.url.edu (V.I. Pérez-Nueno), jordi.teixido@iqs.edu (J. Teixidó).

¹ Tel.: +34 93 267 20 00; fax: +34 93 205 62 66.

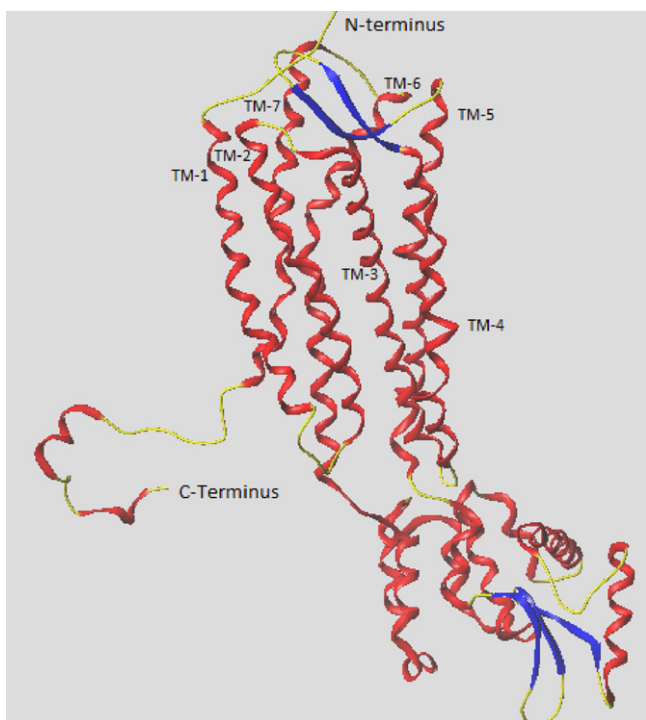


Fig. 1. Backbone scheme of the co-receptor CXCR4 composed by the seven typical GPCR transmembrane domains, the intracellular C-terminus and the extracellular N-terminus with their corresponding triad of extracellular and intracellular loops.

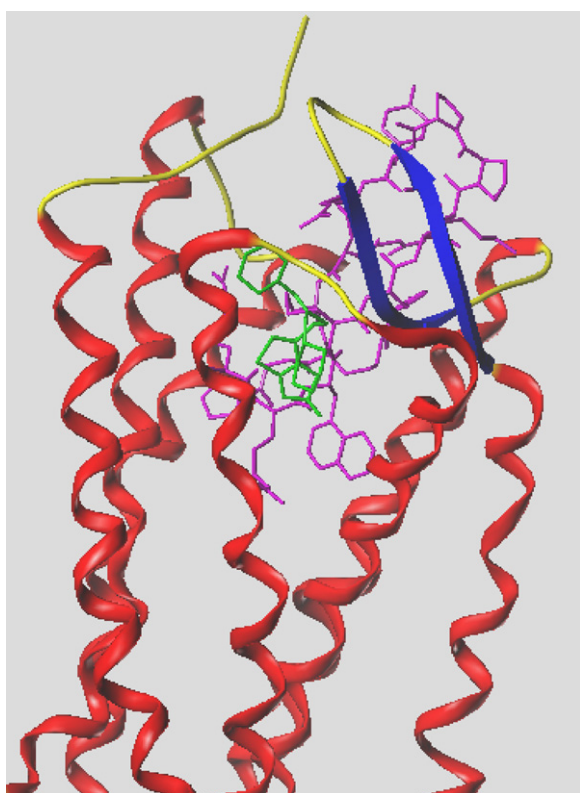


Fig. 2. CXCR4 ligand binding pocket complexed with the small organic antagonist IT1t, colored in green, superimposed with the structure of the peptide ligand CVX15, colored in purple. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Table 1
Resolution and characteristics of CXCR4 structures published by Wu et al.

PDB	Resolution (Å)	Ligand	Receptor units
3ODU	2.5	IT1t	Dimeric CXCR4
3OE0	2.9	CVX15	Monomeric CXCR4
3OE6	3.2	IT1t	Monomeric CXCR4
3OE8	3.1	IT1t	Trimeric CXCR4
3OE9	3.1	IT1t	Dimeric CXCR4

[10,11]. Although the overall orientation differs between ligands, hydrophobic interactions highly contribute in the binding affinity of the majority of CXCR4 antagonists. The residues Ala95, Leu91, Val112, His113, Tyr116, Phe292, Tyr255 and Ile259 play important roles in hydrophobic binding. Moreover, Tyr45 on TM1, Trp94 and Asp97 on TM2, Tyr116 on TM3, Asp171 on TM4, Asp262 on TM6 and His281 and Glu288 on TM7, are critical residues involved in stabilizing interactions with known CXCR4 antagonists [12–14]

Nevertheless, in the case of the peptide ligand it is observed a ligand-induced conformational change in the extracellular side of the helix 5 [7]. Hence, the docking-based virtual screening (VS) of small peptides might be more dependent on the CXCR4 structure used for the screening.

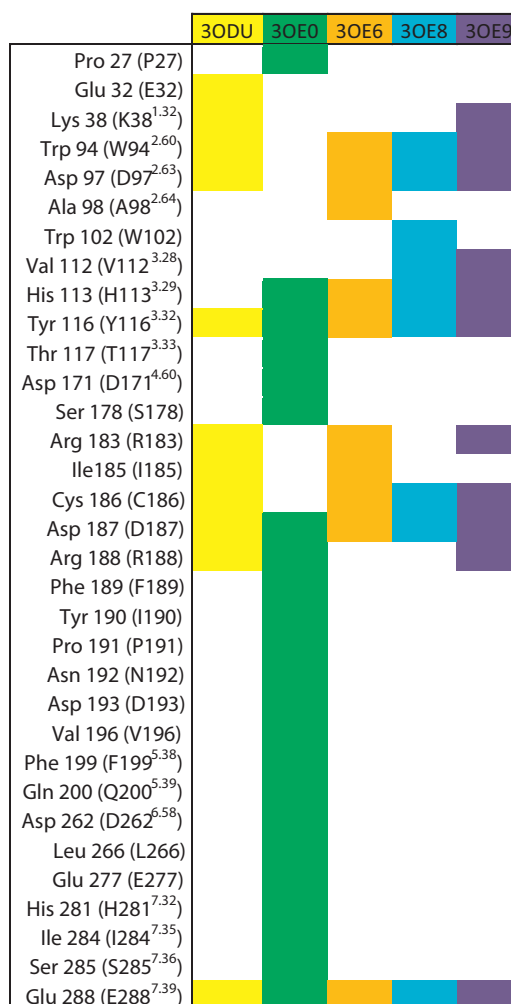


Fig. 3. Protein–ligand interaction map for each complex structure. Structure 3OE0, whose co-crystallized ligand CVX15 has a large volume presents, many protein–ligand interactions in comparison with the rest of complexes whose common co-crystallized ligand IT1t is smaller. Some different interactions can be appreciated between PDBs 3ODU, 3OE6, 3OE8, 3OE9. However, superposition of crystallized IT1t binding modes is very similar in all cases. In brackets Ballesteros–Weinstein residues notation [65].

2. Methods

2.1. Virtual screening data preparation

In the last years GPCR and also chemokine receptors have begun to be determined successfully by protein crystallography. In the case of the chemokine co-receptor CXCR4, Wu et al. obtained five CXCR4 structures at the end of 2010.

We analyzed the five CXCR4 structures published in the Protein Data Bank by retrospective docking-based (VS) using different representative force-field based, empirical, and knowledge-based scoring functions [15]: AutoDock Vina [16] from AutoDock, GoldScore [17], ChemScore [17] and GoldScore rescored with ChemScore from Gold 4.1, Glide 5.5 [18], TotalScore provided in Surflex [19] from Sybyl X 1.1 [20], Ligandfit scoring functions (LigScore1; LigScore2 [21]; PLP1; PLP2 [22,23]; Jain [23], PMF and Dock.Score) implemented in Discovery Studio 2.5.

Moreover, we compared these results with the ones retrieved by our previously published CXCR4 rhodopsin-based homology model built when there were not available CXCR4 crystal structures [24]. This homology model agrees with the folding structure of the crystal PDBs now available. Fig. 4 shows the superposition of our homology model and the crystal CXCR4 structure 3ODU. A detailed structural comparison between homology models and CXCR4 crystal structures was done by Kufareva et al. [25]. Herein, we compare our homology model against the crystal structures in terms of retrospective docking performance, especially the early recovery.

To perform the retrospective docking-based screening we used the test set of Pérez-Nueno et al. [24]. This test set contains 248 inhibitors representative of 4 CXCR4 antagonist scaffold families [8,26] (Fig. 5) and 4696 different presumed inactive extracted [24] from the Maybridge Screening Collection whose physicochemical properties (molecular weight, number of rotatable single bonds,

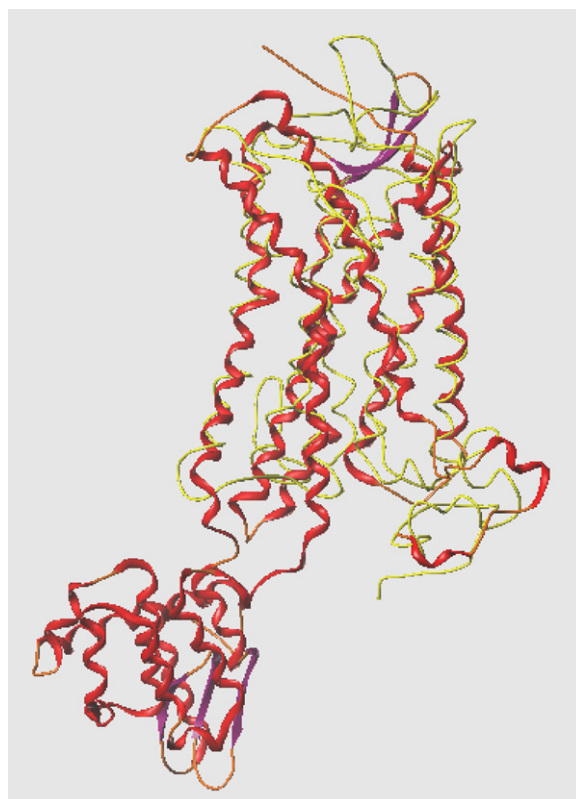


Fig. 4. Superposition of CXCR4 structures PDB 3ODU (red) and the homology model (yellow). Residues Asp171, Asp262 and Glu288 have been used to define superposition. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

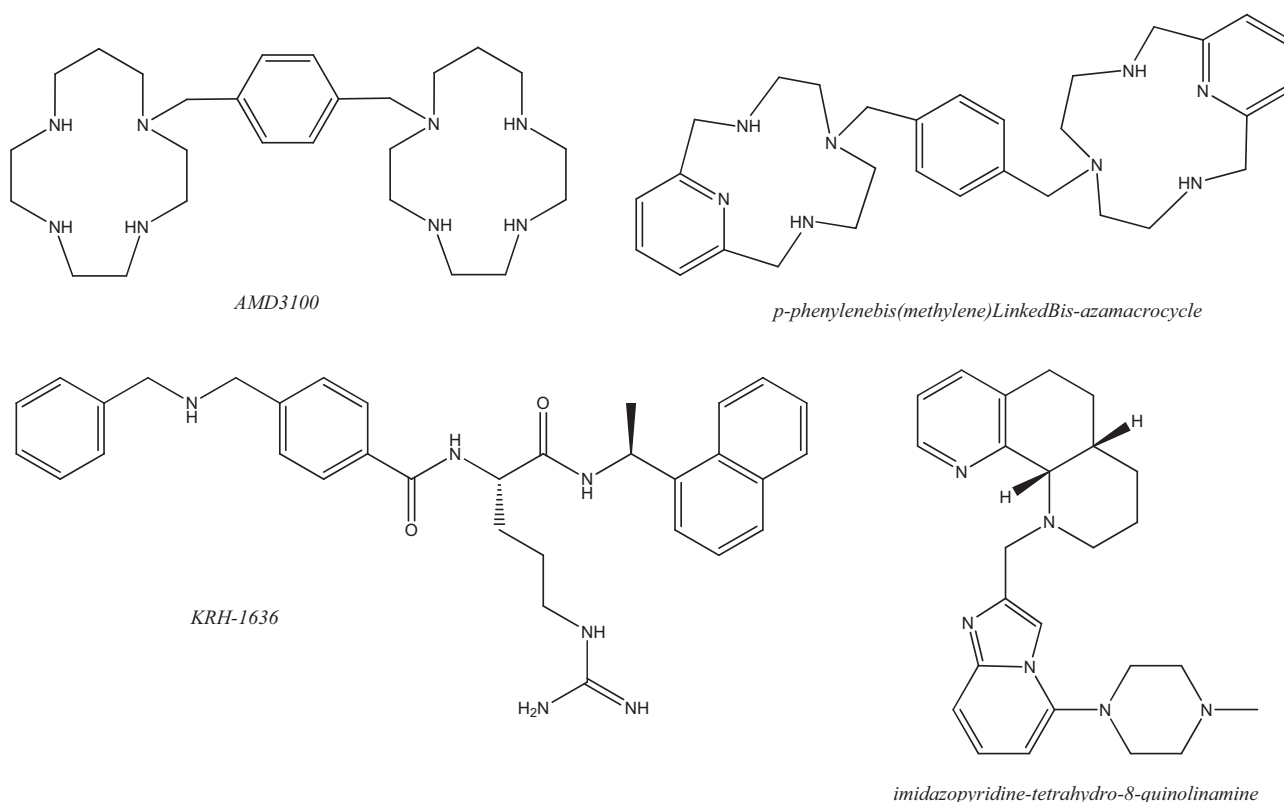


Fig. 5. Representative compounds of the chemotype families used as actives in the test set 24 to guarantee diversity of true binders.

Table 2
Ligands and decoys molecular properties: summary of the 1D physicochemical properties of active and decoy molecules. Standard deviation values in parentheses.

	Ligands: 248	Decoys: 4549
Molecular weight	727.1/306.5 (74.3)	613.9/452.2 (34.9)
Rotatable single bonds	18/0 (4.0)	24/0 (2.2)
Number of hydrogen-bond acceptor atoms	6/1 (1.2)	13/0 (1.6)
Number of hydrogen-bond donor atoms	5/0 (1.3)	9/0 (1.0)
Number of hydrophobic atoms	46/15 (4.6)	43/9 (3.7)
Octanol–water partition coefficient	7.6/–2.8 (2.3)	13.9/–2.8 (1.8)
Polar surface area	485.5/25.8 (60.0)	561.5/0.0 (91.7)
Total positive partial charges	18.58/5.99 (1.24)	12.14/0.69 (2.34)
Total negative partial charges	–4.33/–13.8 (1.16)	–0.69/–10.79 (1.35)

number of hydrogen-bond acceptor atoms, number of hydrogen-bond donor atoms, octanol–water partition coefficient, number of hydrophobic atoms, PSA, total positive partial charges and total negative partial charges) match similar property values than the selected actives [27]. Table 2 shows that the average and standard deviations of these properties calculated by MOE 2009 [28] are quite similar for the active and inactive pools. Supplementary Fig. 1 shows the ROC curves obtained when ligands and decoys were classified by their total positive partial charge (PEOE PC+), total negative partial charge (PEOE PC–), total polar surface area (TPSA) and log octanol/water partition coefficient (log *P*) in order to evaluate the relevance of the decoy set. It can be seen that TPSA, log *P* and PEOE PC– discriminate randomly actives from decoys, as expected in a proper selection of a decoy set. However, the PEOE PC+ ROC curve favors actives retrieval given that the total positive partial charge distribution for the decoy set is a bit lower than the active ligand set.

Both ligands and decoys were protonated at physiological pH, Gasteiger partial charges were assigned and geometric structure was minimized using MMFF94 forcefield implemented in MOE.

Protein structures were prepared prior to their docking, using the own protein preparation module implemented in each docking tool: for AutoDock Vina we used *ADT program* to prepare the corresponding PDBQT–protein file, *Protein Setup* tool was used for Gold, *Protein Preparation Wizard* for Glide, *Prepare Protein Structure* for Sybyl and *Prepare Protein* in the case of Discovery Studio. The protein structure was protonated and minimized using forcefields: OPLS2005, TRIPOS and CHARMM for Glide, Sybyl X and Discovery Studio protocols respectively. His281 and His113, allocated in the near boundaries outside the binding site definition, were protonated equally in all preparation programs. Redundant homodimeric and homotrimeric chains were suppressed. Only the larger chains with less gaps in the binding pocket region were kept in order to avoid gaps inserted between residues known to make important interactions with the bound ligand (for example, in 3ODU structure, the second shortest protein chain lacks the Pro27, which presents a direct interaction with the peptidic ligand in structure 3OE0 (Fig. 3)); a second example is the central monomeric unit in structure 3OE8 which presents gaps corresponding to residues Leu267, Glu268, Ile269, Ile270 and Lys271. These residues do not interact with the small ligand IT1t, but in the case of ligand CVX15 they are allocated much closer to the peptidic ligand and taking into account that CXCR4 has a deep binding pocket it seems convenient not to underestimate. A third example is one of the chains in structure 3OE9 which starts from the residue 35 and presents gaps corresponding to residues Cys28, Phe29, Arg30, Asn31, and gaps as well corresponding to residues Gln66, Lys67, Lys68 and Leu69. Again in this case, residues Cys28, Phe29, Arg30 and Asn31 are placed very close to the ligand CVX15 and in a deep and big binding site these residues could interact with other CXCR4 ligands. Additionally, water molecules present only in PDBs 3ODU and 3OE0, as well as chemical molecules used as crystallization aiders, were eliminated. For structure 3ODU, binding site water molecules were both kept and eliminated for comparison purposes.

CXCR4 binding pocket was defined setting the volume, where the ligands could interact with the protein, delimited by residues Asp171, Asp262 and Glu288 (Fig. 6). Several mutagenic studies confirm interactions between Asp171, Asp262 and Glu288 residues with different ligands such as bicyclam AMD3100 [29–31], monocyclams AMD3465 and AMD3529 [32] or non cyclams like AMD070 [10] in different ways. This triad of acidic residues also interacts with peptidic ligands like T-140 a 14-residue polypeptide [33,34] or cyclopentapeptides antagonists like FC131 and their derivatives [35].

2.2. Retrospective docking-based virtual screening

We docked Pérez-Nueno et al. test set of actives and decoys onto each CXCR4 PDB and the CXCR4 homology model using the aforementioned 12 scoring functions and the scoring function with rescoring. All the screened compounds were ranked by their retrieved scores and the final ranking lists were represented in a ROC plot [36]. The area under the ROC plot (AUC) [36,37] and the normalized partial area under the curve, or partial area index, were quantified to evaluate the VS performance. As high AUC values do not guarantee that the top ranked poses will correspond to active compounds [38], we calculated the partial area under the curve [39,40] (pAUC) at 1%, 5% and 10% of false positives. The corresponding values were normalized to partial area indexes [41].

$$pAUC_{x\%} = \frac{1}{2} \left[1 + \frac{AUC_{x\%} - AUC_{\text{random at } x\%}}{AUC_{\text{perfect at } x\%} - AUC_{\text{random at } x\%}} \right]$$

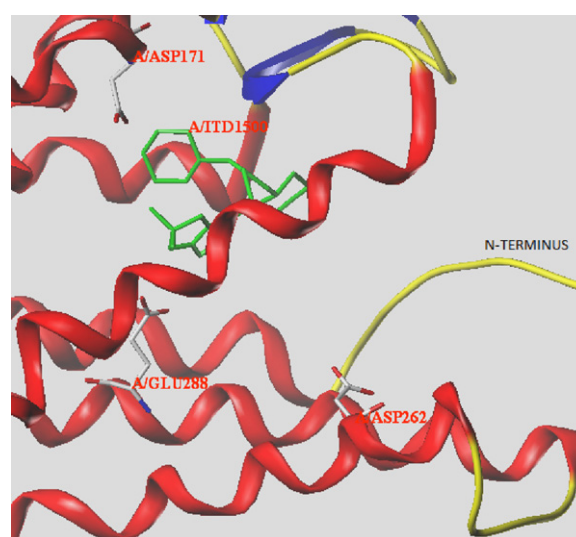


Fig. 6. Definition of the CXCR4 binding site by residues Asp171, Asp262 and Glu288 in the N-terminus region. The small co-crystallized ligand IT1t is shown in green color. It is allocated on one side of the pocket making interactions with Asp171 residue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

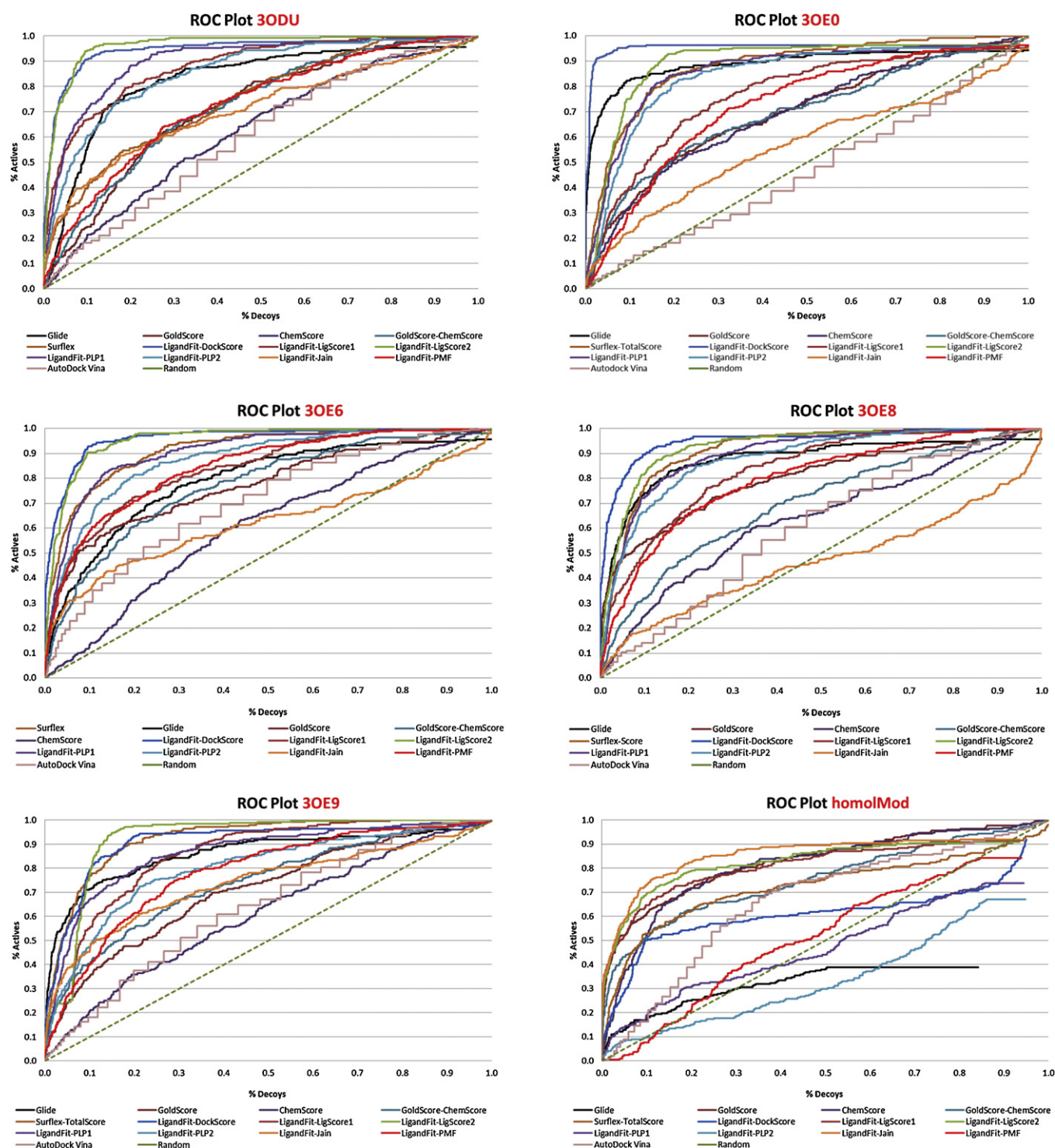


Fig. 7. ROC plots of the retrospective docking-based VS of Pérez-Nueno et al. CXCR4 test set for each of the five experimental CXCR4 structures and the CXCR4 homology model.

In order to compare the overall AUC and the pAUC values obtained for each structure we calculate the average area under the curve (aAUC) as well as the average pAUC of all scoring functions for each PDB.

Furthermore, evaluation [42] of early recovery of actives was assessed by the enrichment factor (EF) [43] at 1%, 5% and 10% of screened database and, additionally, by the percentage of actives retrieved at 1%, 5% and 10% of decoys. Moreover, the average enrichment factor (aEF) at 1%, 5% and 10% of database screened, and the

average of the percentage of true positives retrieved at 1%, 5% and 10% of decoys were also calculated based in all 13 scoring functions used in the analysis.

2.3. Accuracy in pose prediction and docking reliability

The accuracy in pose prediction was evaluated by selfdocking the CXCR4 PDBs containing the same cognate ligand using the aforementioned scoring functions under the same docking

conditions [44]. This allows the analysis of possible geometric and shape differences between CXCR4 structures.

The docking reliability was evaluated for each CXCR4 complex by calculating the root-mean square-deviation (RMSD) between the position of the co-crystallized ligand, taken as the reference, and the binding mode predicted by the docking function, also by the best docked pose (taken as the one with the lowest RMSD with respect to the crystal structure), the average RMSD (aRMSD) [45] of the 10 best poses returned for each docking function and the number of best docked poses with a RMSD value equal or below to 2 Å were calculated.

3. Results

3.1. Pose prediction

The two different CXCR4 co-crystallized ligands, IT1t and CVX15, were selfdocked into their respective protein structures using GoldScore, ChemScore, GoldScore rescored with ChemScore, Glide, Surflex and Ligandfit docking functions. Table 3 shows the best RMSD poses returned for each docking function and each PDB, the calculated average RMSD (aRMSD) of the 10 best poses, and the number of docked poses with a RMSD value equal or below to 2 Å.

The main weakness of using RMSD to evaluate the pose prediction is its dependency to the molecular size [46]. This is well noted for the cyclic peptide ligand, CVX15, in PDB 3OE0. The RMSD values obtained for 3OE0 are clearly higher than the results obtained for the small organic ligand IT1t co-crystallized in the rest of PDBs.

Comparing the results for PDBs whose cognate ligand is IT1t, it can be observed that PDB 3ODU returns the lowest aRMSD value (4.0 Å), as well as the highest number of best docked poses with a RMSD equal or below to 2 Å. In addition, PDB 3ODU is the receptor structure with the highest crystallographic resolution (2.5 Å).

3.2. Docking-based virtual screening performance

Pérez-Nueno et al. test set of actives and decoys was docked against all five CXCR4 PDBs and the CXCR4 homology model using the aforementioned docking functions. Docked poses were ranked according to their docking score. Fig. 7 shows the ROC plots for each PDB and each scoring function. AUC values are shown in Table 5a. Tables 4a and 4b show the early recovery performance in terms of EF at 1%, 5% and 10% of the screened database and the percentage of true positives selected at 1%, 5% and 10% of false positives, respectively.

It can be observed that the ROC curves for the CXCR4 homology model obtained using Ligandfit scoring functions do not arrive until 100% of screened database. Discovery Studio is very sensitive to the size of the ligand, the geometry of the binding pocket, and the definition of the binding pocket sphere to retrieve accurate protein–ligand interaction energies. Hence, if good interactions cannot be found with a given definition of the binding pocket for a determined ligand, the software will not return a docking score for that ligand. When defining the CXCR4 binding pocket as for the rest of the docking software (see Section 2), Discovery Studio did not retrieve results for 200 out of the 248 actives. Hence, in order to improve hits retrieval, the binding site of the CXCR4 homology model was defined first into two and after into three partition levels using Discovery Studio (Fig. 8). Discovery Studio partition levels divide randomly one binding site into a specified number of sub-sites. In this manner, the docking engine Ligandfit explores each sub-site and all the different combinations of them. Consequently when n sub-sites are defined, the test set is docked n factorial times along the binding site. Docking-based VS was carried out exploring in a single run every sub-site, and also the combinations of

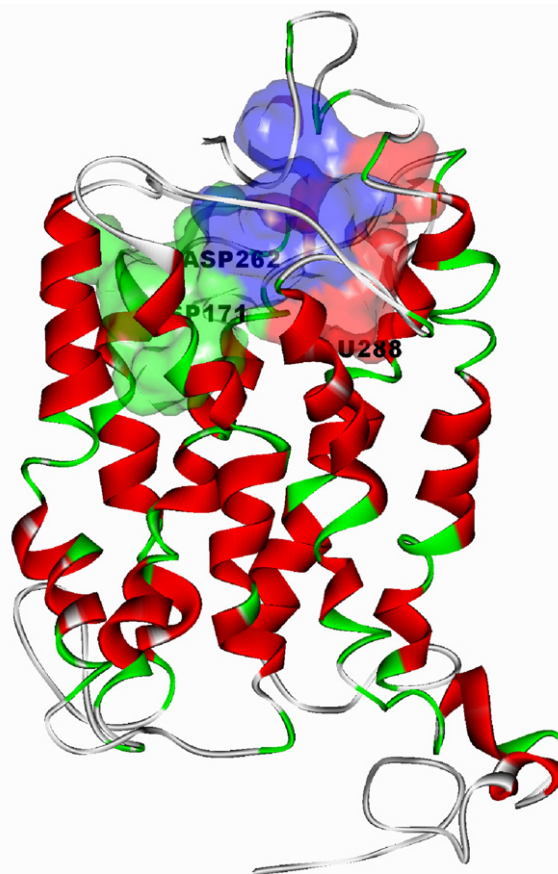


Fig. 8. Definition of three sub-sites (green, blue, and red) in the CXCR4 homology model using Discovery Studio. LigandFit assays to dock every pose into the three sub-sites (green, blue, and red) and also into the combination of 2 sub-sites (green–blue, green–red and blue–red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

two different sub-sites when three sub-sites were defined. Hits retrieval and VS performance improved significantly when defining three sub-sites with respect to defining only one or two sub-sites. The Ligandfit ROC curves shown in Fig. 7 for the CXCR4 homology model represent the results when dividing the pocket in three sub-sites. The fact that dividing the CXCR4 pocket in sub-sites increases the VS performance agrees with recent published literature which support the fact that at least two binding sub-sites exist [12–14] in the main CXCR4 binding pocket located between ECL2 and the N-terminus, and also with the fact that antagonists like the bicyclam AMD3100, the monocyclam AMD3465, and the non-cyclam AMD11070 bind to CXCR4 in three different binding modes [47]. The distribution of actives, when those were docked into the three sub-sites, was mainly between the pockets defined by sub-sites 1–2 and 2–3. 100 actives were docked in sub-site 1–2 and 116 actives were docked in sub-site 1–3. In pockets defined by sub-sites 1–3, 1–1, and 2–2, only 9, 2, and 1 actives were docked, respectively. In addition, if we classify the compounds retrieved in each pocket sub-site by chemotype families, we retrieve 52% of AMD derivatives, 37% of tetrahydro-quinolinamides, 21% of macrocycles and 35% of KRH derivatives in the pocket defined by sub-sites 1–2, and 32% of AMD derivatives, 58% of tetrahydro-quinolinamides, 58% of macrocycles and 35% of KRH derivatives in the pocket defined by sub-sites 2–3.

In general, ROC plot analysis shows that the docking-based VS is selective in choosing ligands from decoys for the retrospective test set. Concretely, comparing early recovery values obtained for all PDBs, it can be observed that PDB 3OE6 obtains the highest top

Table 3

Pose prediction analysis when cognate ligands were selfdocked into their respective protein structures. Data in bold represents the best values obtained for each docking function when comparing all five structures. PDB 3OE0 shows the highest RMSD values due to its higher molecular weight.

Pose prediction	PDB	3ODU	3OE0	3OE6	3OE8	3OE9
GoldScore	Best RMSD pose	1.0	9.2	5.4	6.5	5.7
	aRMSD	3.9	14.0	6.5	7.9	7.3
	No. poses $\leq 2 \text{ \AA}$	3	0	0	0	0
ChemScore	Best RMSD pose	2.2	14.7	4.1	4.6	5.1
	aRMSD	2.7	15.3	6.1	5.2	6.1
	No. poses $\leq 2 \text{ \AA}$	0	0	0	0	0
GoldScore and ChemScore	Best RMSD pose	1.0	9.8	4.2	3.6	6.2
	aRMSD	3.7	13.9	5.4	7.9	7.6
	No. poses $\leq 2 \text{ \AA}$	4	0	0	0	0
Glide	Best RMSD pose	4.3	– ^a	1.5	1.5	3.1
	aRMSD	4.8	–	5.3	3.8	5.1
	No. poses $\leq 2 \text{ \AA}$	0	–	1	2	0
Surflex TotalScore	Best RMSD pose	3.2	13.7	3.8	4.0	2.4
	aRMSD	4.8	22.9	5.2	5.4	4.3
	No. poses $\leq 2 \text{ \AA}$	0	0	0	0	0
LigandFit	Best RMSD pose	2.4	5.3	2.6	3.0	5.3
	aRMSD	4.5	6.0	4.7	5.2	6.0
	No. poses $\leq 2 \text{ \AA}$	0	0	0	0	0
AutoDock Vina	Best RMSD pose	2.3	– ^a	3.2	3.5	5.0
	aRMSD	5.1	–	7.0	5.6	6.6
	No. poses $\leq 2 \text{ \AA}$	0	–	0	0	0
	Mean aRMSD	4.5	14.2	5.7	6.0	6.2

^a Ligand CVX15, co-crystallized in PDB 3OE0, with more than 56 rotatable bonds exceeds Glide and AutoDock Vina docking requirements.

EF values (values shown in bold in Table 4a) and also the highest aEF (11.2 at 1% of screened database, 7.5 at 5% and 5.4 at 10% of database screened) taking into account all scoring functions used, even scoring functions with poorest results. Equally, the average percentage of true positives (Table 4b) returned at 1%, 5% and 10% of false positives is favorable for PDB 3OE6 (19.3% of true positives at 1% of false positives, 44.1% at 5% and 57.2% at 10% of false positives) among all scoring functions.

Comparing AUC and pAUC data values at 1%, 5% and 10% of screened decoys for all PDBs, it can be observed that PDB 3OE6 obtains also the highest top AUC and pAUC values (values shown in bold in Tables 5a and 5b), the highest aAUC value (0.82) and the highest average pAUC values (0.56 at 1%, 0.64 at 5% and 0.69 at 10% of decoys screened, respectively) among all scoring functions.

Enrichment factors, early recovery, AUC and pAUC results for the homology model (HomolMod) are also shown in Tables 4a, 4b, 5a and 5b, respectively.

3.3. Scaffold retrieval analysis

Pérez-Nueno et al. test set contains 248 actives from four chemotype families: AMD derivatives [29,48], tetrahydroquinolinamines [49,50], macrocycles [51,52] and KRH derivatives [53,54]. To evaluate the ability of the docking protocol to retrieve these diverse scaffold families, we analyzed the percentage of actives returned by the scoring functions attending to its intrinsic chemotype classification at 1%, 5%, and 10% of screened database. Moreover, to assess how selective is the scaffold retrieval we chose PDB 3OE6 as CXCR4 working structure, being the most powerful in terms of aAUC, pAUC, and early recovery performance. Fig. 9 shows the scaffold retrieval rates. It can be observed that at 1% of screened database all four CXCR4 scaffolds are found. Only for tetrahydroquinolinamines and for KRH derivatives, one and four scoring functions respectively, did not retrieve any compound at 1% of screened database and at 5% of screened database only one scoring function failed to recover any KRH derivatives. It is also well noted that at 5% of screened database the scoring function

LigScore2 recovered more than 40% of all chemotype CXCR4 families and other scoring functions like Surflex-TotalScore, PLP1, and PMF recovered more than 30% of all CXCR4 scaffolds. Analysis at 10% of screened database revealed that on average all scaffolds were retrieved in more than 50% of their family compounds unless for tetrahydroquinolinamines and KRH derivatives whose average retrieval rate was 47.4% and 46.2%, respectively.

4. Discussion

Docking evaluation of five CXCR4 crystal structures and a CXCR4 homology model was carried out in terms of pose prediction and retrospective docking-based VS. We used representative force-field based scoring functions like GoldScore and DockScore, empirical scoring functions like ChemScore, Surflex-TotalScore, Glide, LigScore1, LigScore2, PLP1, PLP2 and Jain, and also knowledge-based scoring functions such as PMF or the hybrid knowledge-based and empirical scoring function Autodock Vina.

With respect to pose prediction, RMSD results for the top 10 binding poses retrieved for all PDBs using all aforementioned docking functions are not totally satisfactory. Average RMSD values from Table 3 are all higher than 2 Å. Only four aRMSD results reached values between 2 and 4 Å, i.e. with the PDB 3ODU for the scoring functions GoldScore (3.9 Å), ChemScore (2.7 Å) and GoldScore rescored with ChemScore (3.7 Å) and with the PDB 3OE8 using Glide (3.8 Å). Moreover, Table 3 also shows only four binding modes with RMSD values below 2 Å, i.e. for PDB 3ODU the best RMSD pose was 1.0 Å using GoldScore and GoldScore rescored with ChemScore and for PDB 3OE6 and 3OE8 we obtained in both cases 1.5 Å using Glide.

Comparing pose prediction between the PDBs it can be observed that PDB 3ODU obtains on average the most accurate results (mean aRMSD of 4.5 Å) using as scoring functions GoldScore, ChemScore and GoldScore rescored with ChemScore. As we stated before, it is necessary to take into account that RMSD metric depends on the molecular size of the ligand to dock, and therefore results obtained from docking the peptidic ligand CVX15 (molecular weight of

Table 4a
Early recovery performance expressed by the enrichment factor at 1%, 5%, and 10% of the screened database. The best enrichment factor values obtained for each scoring function are shown in bold. PDB 3OE6 retrieves the best enrichment factors across all docking approaches. Moreover, the highest enrichment factor value at 1% of the database screened was also obtained for structure 3OE6 (19.7).

Enrichment factors	% screened database	3ODU	3OE0	3OE6	3OE8	3OE9	HomolMod
GoldScore	1	1.2	4.5	3.3	2.9	1.6	15.3
	5	1.8	3.3	2.5	2.6	1.5	9.0
	10	1.9	3.0	1.9	2.3	1.6	5.7
ChemScore	1	3.6	3.2	10.5	15.3	6.9	7.0
	5	2.6	4.1	7.0	8.4	4.1	5.7
	10	2.3	3.1	5.1	5.2	3.4	4.7
GoldScore and ChemScore	1	3.2	2.8	5.6	6.0	9.3	14.5
	5	3.0	4.3	4.8	4.5	5.3	7.8
	10	2.7	3.6	3.8	3.2	4.0	4.7
Glide	1	2.8	18.9	8.5	8.5	15.7	4.0
	5	4.3	12.4	5.4	5.4	10.3	2.4
	10	4.6	7.6	4.2	4.2	6.5	1.7
Surflex TotalScore	1	9.7	9.2	15.1	8.4	10.1	9.6
	5	5.4	7.2	9.2	7.9	8.7	6.3
	10	3.5	5.9	6.8	6.5	6.6	4.9
LigandFit DockScore	1	18.7	19.3	19.7	19.2	15.8	9.5
	5	11.7	15.5	12.4	12.5	8.3	4.5
	10	8.3	9.4	8.3	8.2	6.2	3.6
LigandFit LigScore1	1	14.7	2.0	19.7	9.4	12.9	18.1
	5	8.6	4.9	12.4	6.6	4.7	9.5
	10	6.1	4.0	8.3	4.5	4.9	6.0
LigandFit LigScore2	1	15.5	6.4	14.8	11.4	12.9	18.5
	5	11.5	7.0	11.3	8.7	4.7	9.3
	10	8.0	6.6	8.0	6.9	5.5	6.2
LigandFit PLP1	1	9.8	8.1	10.3	9.0	8.3	4.1
	5	8.2	6.7	8.5	7.9	6.8	2.8
	10	6.4	5.9	6.7	6.2	5.9	1.7
LigandFit PLP2	1	9.8	5.6	9.4	7.3	7.9	2.9
	5	6.7	5.7	7.4	7.6	5.5	1.6
	10	5.4	5.4	5.8	5.8	4.4	1.0
LigandFit Jain	1	10.2	6.4	12.3	3.7	12.9	19.3
	5	5.7	3.0	5.4	2.6	6.4	9.7
	10	4.0	2.3	3.4	1.9	4.4	6.6
LigandFit PMF	1	5.7	3.6	10.7	7.3	6.2	0.4
	5	3.9	3.2	7.3	5.0	4.3	0.5
	10	3.0	3.0	5.4	4.2	3.6	0.8
AutoDock Vina	1	2.4	0.8	5.1	2.0	2.8	0.4
	5	1.9	1.2	3.6	1.8	2.1	1.6
	10	1.8	1.1	3.0	1.4	1.6	1.6
Average EF for each PDB	1	8.3	6.5	11.2	8.5	9.5	9.5
	5	5.8	5.6	7.5	6.3	5.6	5.4
	10	4.5	4.3	5.4	4.7	4.5	3.8

2138.6 g/mol), co-crystallized in PDB 3OE0, cannot be directly compared with the results obtained for the small ligand IT1t (molecular weight of 406.7 g/mol), co-crystallized in the rest of PDBs.

With regard to the retrospective docking-based VS, we used Pérez-Nueno et al. test set of 248 different actives from four different CXCR4 antagonist families to guarantee chemotype diversity, and 4696 decoys from the Maybridge Screening Collection with similar properties to the actives. It is worth noting the relevance in choosing actives from different chemotype families in order to evaluate well the potency of the scoring functions according to the diversity of ligands [55].

As aforementioned, only PDBs 3ODU and 3OE0 contain cocrystallized water molecules. The convenience of maintaining water molecules into the protein structure was analyzed (see Supplementary Fig. 2 and Tables 1, 2. PDB 3ODU with and without water molecules was compared with PDB 3OE6 (the most selective VS model). Supplementary Fig. 2 shows that ROC plots for LigandFit scoring functions do retrieve neither the 100% of all actives nor

100% of decoys for PDB 3ODU with water molecules. This observation agreed with the VS results obtained for PDB 3OE0 containing water molecules, where no molecule was retrieved for LigandFit Scoring functions.

Comparison of the early recovery data (Supplementary Tables 1 and 2) obtained for structures with and without water molecules shows that GoldScore, GoldScore rescored with ChemScore, Glide, Surflex and Autodock Vina retrieve slightly higher early recovery values for 3ODU with water molecules ($3ODU_{water}$) than when water molecules were removed (3ODU). However, for ChemScore and the seven LigandFit scoring functions, early recovery values for 3ODU outperform $3ODU_{water}$. Hence, removing water molecules from CXCR4 structures 3ODU and 3OE0 does not seem a rough approximation given that the VS early recovery values obtained when comparing $3ODU_{water}$ and 3ODU are nearly equal, or in some cases better, for 3ODU without water molecules.

Analyzing the data obtained from the retrospective docking-based screenings (Fig. 7 and Tables 4a, 4b, 5a and 5b), we can

Table 4b

Early recovery performance expressed as percentage of actives recovered when 1%, 5%, and 10% of decoys were respectively selected. Values in bold represent the best percentages obtained when comparing all structures for each scoring function.

% true positives	% false positive screened	3ODU	3OE0	3OE6	3OE8	3OE9	HomolMod
GoldScore	1	1.6	6.0	3.2	3.2	1.6	34.7
	5	9.3	18.1	12.5	13.3	7.7	52.0
	10	21.4	32.7	19.0	25.0	16.1	61.3
ChemScore	1	4.0	3.2	15.3	30.6	8.9	10.5
	5	14.5	23.8	40.7	48.0	23.8	35.9
	10	24.6	31.9	52.8	54.4	35.5	52.0
GoldScore and ChemScore	1	4.4	2.8	8.1	6.9	12.1	24.2
	5	16.9	24.6	27.0	24.2	29.4	43.1
	10	28.6	39.1	43.1	33.5	40.3	51.2
Glide	1	3.2	54.4	12.1	12.1	34.7	6.0
	5	29.8	75.0	33.9	33.9	60.1	12.9
	10	56.0	83.1	46.4	46.4	71.4	17.3
Surflex TotalScore	1	13.6	13.6	32.5	11.5	18.1	13.2
	5	29.2	49.8	60.9	48.6	56.0	39.1
	10	39.5	65.8	73.7	77.0	77.0	53.9
LigandFit DockScore	1	44.1	77.4	49.6	49.0	26.4	12.1
	5	78.5	93.1	77.4	78.1	54.9	24.6
	10	90.7	96.0	93.5	89.5	79.3	41.9
LigandFit LigScore1	1	26.3	2.0	20.2	17.0	19.1	38.7
	5	55.1	30.6	45.2	36.0	24.4	51.6
	10	66.8	45.6	56.0	50.6	55.7	64.5
LigandFit LigScore2	1	42.1	6.9	34.3	24.7	19.1	35.5
	5	77.3	52.4	76.6	61.5	24.4	57.3
	10	94.3	80.2	90.3	81.8	78.9	69.4
LigandFit PLP1	1	18.6	12.1	18.1	10.9	12.6	4.4
	5	56.7	48.8	57.3	49.4	44.7	13.7
	10	71.3	70.2	75.8	72.5	67.1	19.0
LigandFit PLP2	1	14.2	6.5	17.7	13.8	10.6	3.6
	5	41.7	39.1	47.6	47.8	32.1	8.1
	10	60.7	64.5	65.3	66.4	47.2	9.7
LigandFit Jain	1	14.6	7.3	17.3	4.0	20.3	37.1
	5	30.4	15.7	30.6	13.4	38.2	58.1
	10	41.7	24.6	36.7	19.0	46.7	73.0
LigandFit PMF	1	6.9	6.9	17.7	10.5	8.1	0.4
	5	21.5	21.5	44.8	28.3	26.8	2.4
	10	32.4	32.4	60.5	47.0	39.8	7.7
AutoDock Vina	1	2.4	0.81	5.2	2.0	2.8	0.4
	5	9.7	6.05	19.4	10.5	10.5	8.9
	10	18.2	11.29	30.7	14.1	18.2	16.1
Average % true positives for each PDB	1	15.1	15.4	19.3	15.1	15.0	17.0
	5	36.2	38.4	44.1	37.9	33.3	31.4
	10	49.7	52.1	57.2	52.1	51.8	41.3

Table 5a

AUC values calculated from the ROC plot for all the CXCR4 PDBs and the CXCR4 homology model. Values in bold represent the best AUC values obtained when comparing all structures for each scoring function.

AUC	3ODU	3OE0	3OE6	3OE8	3OE9	HomolMod
GoldScore	0.62	0.70	0.77	0.80	0.60	0.83
ChemScore	0.72	0.70	0.61	0.63	0.71	0.81
GoldScore and ChemScore	0.72	0.69	0.79	0.71	0.74	0.75
Glide	0.80	0.95	0.82	0.91	0.89	0.26
Surflex TotalScore	0.75	0.88	0.91	0.91	0.92	0.74
LigandFit DockScore	0.95	0.99	0.96	0.95	0.91	0.52
LigandFit LigScore1	0.88	0.80	0.84	0.83	0.85	0.75
LigandFit LigScore2	0.96	0.93	0.96	0.93	0.92	0.77
LigandFit PLP1	0.90	0.89	0.91	0.90	0.86	0.43
LigandFit PLP2	0.85	0.88	0.88	0.88	0.92	0.31
LigandFit Jain	0.71	0.59	0.62	0.47	0.74	0.79
LigandFit PMF	0.73	0.76	0.85	0.79	0.78	0.46
AutoDock Vina	0.60	0.47	0.71	0.60	0.64	0.66
aAUC for each PDB	0.78	0.79	0.82	0.79	0.81	0.62

Table 5b

Partial area indexes calculated from the pAUC values at 1%, 5% and 10% of decoys retrieved for all the CXCR4 PDBs and the CXCR4 homology model. Values in bold represent the best partial area indexes obtained when comparing all structures for every scoring function.

pAUC	% false positive Screened	3ODU	3OE0	3OE6	3OE8	3OE9	Homolog. Model
GoldScore	1	0.51	0.51	0.55	0.61	0.52	0.61
	5	0.53	0.55	0.63	0.68	0.56	0.70
	10	0.55	0.58	0.67	0.70	0.59	0.73
ChemScore	1	0.50	0.51	0.50	0.51	0.51	0.53
	5	0.51	0.54	0.52	0.52	0.51	0.59
	10	0.53	0.57	0.53	0.54	0.53	0.64
GoldScore and ChemScore	1	0.51	0.51	0.52	0.52	0.54	0.59
	5	0.54	0.55	0.57	0.56	0.59	0.66
	10	0.56	0.59	0.61	0.59	0.62	0.68
Glide	1	0.50	0.71	0.54	0.59	0.62	0.51
	5	0.55	0.80	0.60	0.70	0.72	0.53
	10	0.62	0.85	0.63	0.76	0.77	0.54
Surflex TotalScore	1	0.54	0.54	0.60	0.53	0.55	0.54
	5	0.59	0.63	0.71	0.64	0.67	0.61
	10	0.62	0.70	0.77	0.72	0.75	0.66
LigandFit DockScore	1	0.66	0.77	0.70	0.70	0.60	0.54
	5	0.79	0.91	0.80	0.81	0.68	0.58
	10	0.86	0.94	0.86	0.86	0.74	0.61
LigandFit LigScore1	1	0.59	0.50	0.56	0.54	0.56	0.65
	5	0.68	0.56	0.64	0.61	0.60	0.71
	10	0.74	0.61	0.69	0.65	0.64	0.74
LigandFit LigScore2	1	0.62	0.52	0.60	0.56	0.57	0.64
	5	0.78	0.61	0.75	0.68	0.60	0.71
	10	0.85	0.72	0.83	0.77	0.67	0.76
LigandFit PLP1	1	0.55	0.53	0.55	0.54	0.54	0.51
	5	0.66	0.61	0.66	0.64	0.62	0.53
	10	0.73	0.70	0.74	0.72	0.70	0.53
LigandFit PLP2	1	0.54	0.52	0.55	0.53	0.53	0.51
	5	0.62	0.58	0.64	0.64	0.59	0.51
	10	0.68	0.66	0.70	0.71	0.63	0.51
LigandFit Jain	1	0.55	0.52	0.56	0.51	0.57	0.67
	5	0.60	0.54	0.60	0.53	0.63	0.73
	10	0.63	0.56	0.62	0.54	0.66	0.77
LigandFit PMF	1	0.52	0.51	0.55	0.53	0.52	0.50
	5	0.55	0.53	0.63	0.58	0.56	0.49
	10	0.58	0.56	0.69	0.62	0.60	0.49
AutoDock Vina	1	0.50	0.50	0.51	0.50	0.51	0.50
	5	0.49	0.50	0.54	0.49	0.52	0.50
	10	0.52	0.50	0.57	0.52	0.53	0.51
Average pAUC for each PDB	1	0.55	0.55	0.56	0.55	0.55	0.56
	5	0.61	0.61	0.64	0.62	0.60	0.60
	10	0.65	0.66	0.69	0.67	0.65	0.63

affirm that retrospective docking results obtained for all CXCR4 PDBs are completely satisfactory in terms of early recovery. All PDBs obtained several EF values equal or higher than 10 at 1% of screened database (Table 4a). Principally, for PDB 3OE6, eight scoring functions reached EF values up to 10 and the average EF of all scoring functions used to dock this PDB was 11.2. Discarding the three scoring functions which retrieve the poorest scores (GoldScore, GoldScore rescored with ChemScore and AutoDock Vina) the aEF rises to 13.1.

Regarding the scoring functions that retrieve the best early recovery performance, scoring function DockScore from LigandFit performed the best, showing EF values close to 20 for all PDBs at 1% of screened database (Table 4a), unless for PDB 3OE9 (15.8). The DockScore scoring function is a force field based approximation that estimates the energy of interaction by summing the internal energy of the ligand and the ligand–receptor interaction energy. In addition, the ligand–receptor interaction energy is taken as the sum of the van der Waals energy and electrostatic energy.

Moreover, the majority of scoring functions from LigandFit, as well as Surflex and Glide retrieved also very good early recovery results. Hence, Ligandfit from Discovery Studio, Surflex from Sybyl and Glide from Schroedinger docking packages are well suited to deal with CXCR4 docking-based VS.

Evaluation of the ROC plots (Fig. 7), and the corresponding AUC data (Table 5a) and pAUC (Table 5b) also show good performance of the different scoring functions for the PDBs complexed with both the peptide antagonist and the IT1t small ligand. LigandFit functions, Glide and Surflex scoring functions returned the most selective ROC plots and the highest AUC values in all retrospective dockings. It should be noted that peptidic ligands often generate induced fit conformational changes into the protein structure, leading, consequently, to structural changes into the binding site. However, docking the test set against PDB 3OE0 and using the scoring function LigandFit, we obtained the best AUC result (0.99) and the best pAUC values at 1% (0.77), at 5% (0.91), and at 10% (0.94) of decoys retrieved. Consistently to these pAUC values, the early

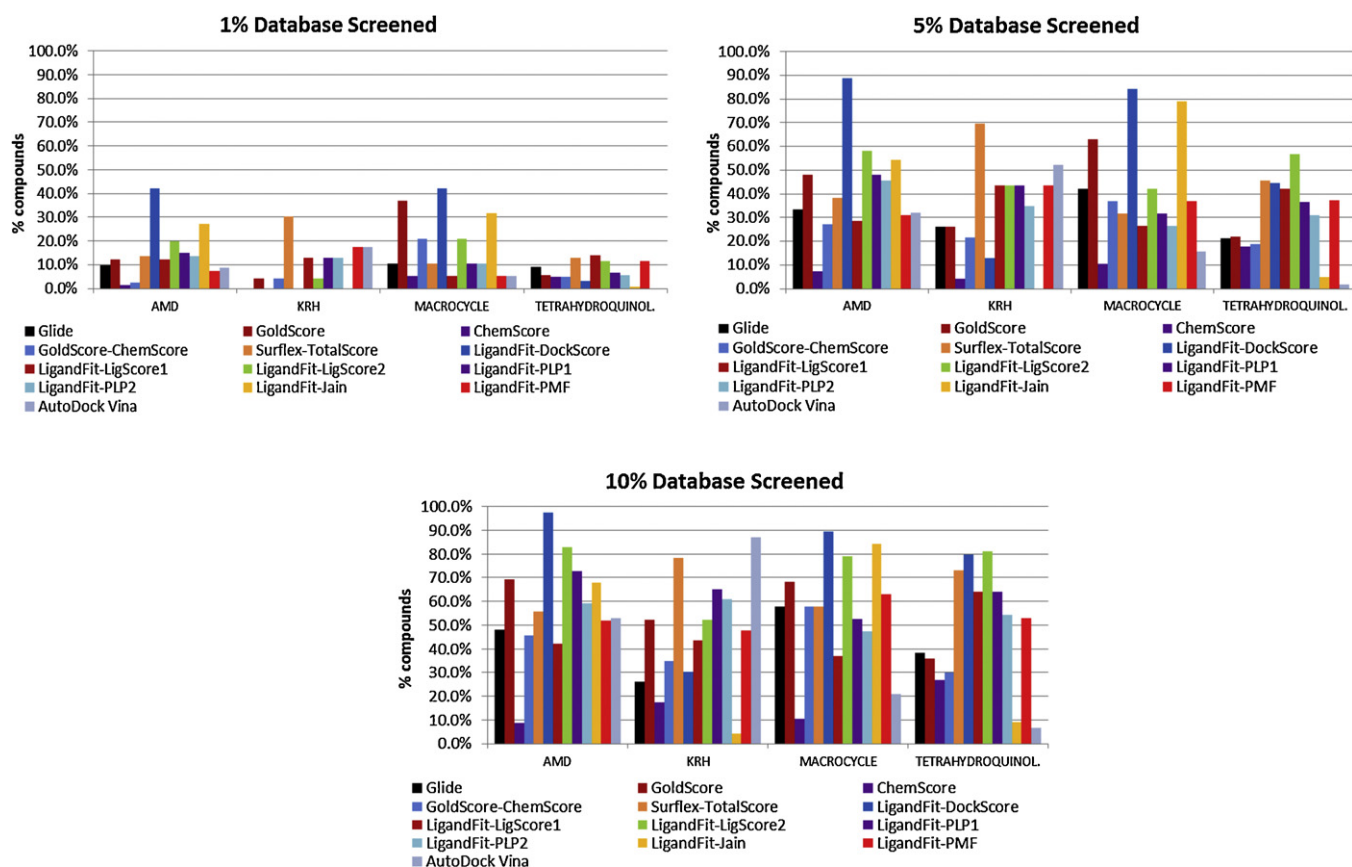


Fig. 9. Scaffold diversity retrieval plots for PDB 30E6. These graphics show the percentage of actives by chemotype family retrieved using all different scoring functions (bars in colors) at 1%, 5% and 10% of the screened database, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

recovery was also the highest (77.4% of actives at 1% of recovered decoys) along all retrospective dockings.

If we compare the ROC plots and their corresponding AUC values by PDB (Table 5a), all PDBs obtained average AUC values (aAUC) around 0.8, corresponding the highest aAUC value to PDB 30E6 (0.82). Also, the PDB 30E6 obtained the best AUC values (Table 5a in bold) when compared with the AUC values achieved for the rest of PDBs. Comparison of the early recovery and AUC values obtained for each scoring function and all the structures shows that PDB 30E6 again returns the most selective results for GoldScore rescored with ChemScore, LigandFit LigScore2, LigandFit PLP1, LigandFit PMF and AutoDock Vina scoring functions. Attending to these results, we chose PDB 30E6 for a scaffold retrieval analysis using the scoring functions selected in this work. This analysis shows that all four chemotype families were retrieved for all scoring functions at 10% of the screened database and only one scoring function did not retrieve KRH compounds at 5% of screened database. Therefore, scaffold retrieval results obtained for PDB 30E6 are also completely satisfactory.

Comparing the docking-based VS results obtained for the CXCR4 homology model with the ones obtained for all the CXCR4 PDBs, it can be observed that the homology model captures as well as the PDB complexes the early recovery performance. Good enrichment factors at 1% of the screened database and good percentage of true positives at 1% of decoys were retrieved for the scoring functions: GoldScore (15.3 and 34.7%) and GoldScore rescored with ChemScore (14.5 and 24.2%), LigandFit LigScore1 (18.1 and 38.7%), LigandFit LigScore2 (18.5 and 35.5%) and LigandFit Jain (19.3 and 37.1%). On average, early recovery results were similar and in some cases even better than those obtained for the crystallized

structures. Nevertheless, generally poor early recovery results were achieved using the scoring functions: Glide, LigandFit PLP1, LigandFit PLP2, LigandFit PMF and AutoDock Vina. Additionally, for the homology model, it was observed that early recovery values became less selective when increasing the percentage of screened database.

In the same way when comparing the ROC plots for the homology model and the crystal structures it can be observed that the highest average pAUC value at 1% of retrieved decoys (Table 5b) was achieved by the homology model as well as by the 30E6 PDB (average pAUC of 0.56). Moreover, average pAUC values at 5% and 10% of retrieved decoys for the homology model were very similar to the values obtained for the crystal structures.

Hence, the homology model performed as well as the crystal structures in the first percentages of database screened, which suggests that using an homology model created from a template crystallized with a representative ligand may be a decent approach to separate members of a particular inhibitor class from inactives in virtual screening experiments. This observation is consistent with several papers that support the use of homology models for GPCR [56,57] predictions and protein kinases VS [58]. Vilar et al. support the notion that homology models often provide a reasonable framework for discussing experimental data. They state that crystal structures are only static representations of proteins and neglecting protein flexibility by using only one rigid representation of the target has been shown to adversely affect virtual screening performance. They claim that receptor ensemble docking, a method where docking is performed on several distinct structures and the results are subsequently combined, has been successfully used to account for protein flexibility in virtual screening campaigns [56].

Hence, combining docking results obtained using different CXCR4 crystal structures or even different CXCR4 homology models could help to account for protein flexibility and to improve VS results. Furthermore, it is also worth mentioning that if a particular protein family has a greater degree of induced fit upon ligand binding, then homology models created from a template with an appropriate ligand bound will produce better docking results than a crystal structure with no ligand or a dissimilar ligand bound [56].

In summary, although the homology model performs as well as the crystal structures in the first percentages of database screened the total VS performance shows a limited predictive power. CXCR4 is generally a target difficult to dock. Retrospective docking results and binding mode prediction are not completely satisfactory. Binding mode analyses of the docked poses onto the CXCR4 homology model allowed us to hypothesize multiple binding sub-sites in CXCR4 binding pocket. This hypothesis could explain the difficulty in retrieving good receptor-based VS results. Recent studies show that there are several notable differences in the binding mode of the monocyclam (such as AMD3465), bicyclam (such as AMD3100) and non-cyclam (such as MX122) antagonists [10,12–14]. Moreover, recent literature suggests that the CXCR4 conformation stabilized *i.e.* by AMD3100 differs slightly from the one stabilized by AMD3465, mainly in the orientation of TM7. This may explain the improved efficacy of AMD3465 over AMD3100 in inhibiting the action of CXCL12 [13].

Furthermore, CXCR4 flexibility makes also difficult to get good docking results. For example, by including receptor backbone flexibility, it is possible to consolidate all the mutation data pertaining to the binding of the cyclam compounds [14]. Techniques that deal with “difficult targets”, that is to say targets with multiple binding sites or multiple binding modes, have recently appeared in order to improve the screening performance, such as consensus shape screening [59–61].

As expected, according to the experiment carried out (selfdocking of the two different CXCR4 co-crystallized ligands, IT1t and CVX15, into their respective protein structures using different scoring functions), the highest resolution structures performed the best in the compound pose prediction analysis. However, regarding the docking-based VS performance, it turned out that the lowest resolution structure, 3OE6, provided the best early enrichment of the actives in the screening dataset. This result leads to the debate of prioritizing conformational selection or induced fit concepts. The majority of active ligands compiled from the literature are just known to bind CXCR4, without an exact idea of their binding mode. Taking this into account, together with the fact that CXCR4 protein is known to be an allosteric transmembrane protein [62,63], leads to the hypothesis that the majority of active ligands in our compiled database bind in a more stable mode within the lowest resolution conformation in the light of our early recovery docking enrichment results. CXCR4 structures can be crystallized in slightly different conformations, regardless of the crystal resolution, which captures in some way the induced fit binding capacity of the protein. We can have two CXCR4 crystal complexes at the same resolution with different interactions in the binding pocket [64] (*i.e.* 3OE8 and 3OE9 structures, see Fig. 3) or two different resolution CXCR4 dimers with different interactions (*i.e.* 3ODU and 3OE9, see Fig. 3) which can give us an idea of the allosteric interactions. Therefore, given the light of the results in this study, we hypothesize that the lowest resolution CXCR4 crystal structure describes better the CXCR4 induced fit binding effect, allowing the wide range of different active ligand scaffold families to be better accommodated in different pocket subsites in 3OE6. Our homology model also shares this behavior of 3OE6, which is able to accommodate the wide range of actives in different pocket subsites retrieving good early recovery values. Hence, it is necessary to take a compromise when performing docking based virtual screening on allosteric proteins between choosing

a higher resolution crystal structure or a lower resolution crystal structure but with a conformation describing better the induced fit mechanism of the protein. Moreover, if the induced fit effect is significantly important, it might be even better to use a good model (or good homology model if crystal structures are not available) built to account for protein flexibility, *i.e.* by consensus shape screening and receptor ensemble docking, using templates with an appropriate ligand bound. Supplementary Fig. 3 shows a comparison between the binding pockets of the highest (3ODU) and lowest (3OE6) resolution CXCR4 crystal structures used in this study in terms of H bonding, lipophilic potential, and electrostatic potential surface maps. Moreover, a detailed binding pocket interaction map showing the different interactions listed in Fig. 3 is shown for both 3ODU and 3OE6 structures. The H bonding surface map is similar for both structures, however the lowest resolution structure seems to have a bigger and more defined hydrophobic binding pocket central region surrounded by high electronegativity areas (see hydrogen bonding surface map on the left), which corresponds to the brown high lipophilic region in the lipophilic potential surface map, and a bigger and central blue region in the electrostatic potential map. The different binding pocket interactions between both structures are mainly between the residues Glu32, Lys38, Ala98, His113 and Arg188.

5. Conclusions

In the present study we have compared and evaluated for the first time all CXCR4 structures published by Wu et al. together with our previously published CXCR4 rhodopsin-based homology model under the point of view of their receptor based virtual screening performance.

With the aim of doing an evaluation that guarantees diversity during the global process we used Pérez-Nueno et al. test set of 248 known actives comprising four different chemotypes [43] of inhibitors. Additionally, we have used different kinds of scoring functions, attending to the nature of the computation algorithm, to dock our previously reported CXCR4 homology model and different CXCR4 structures co-crystallized either with a small organic molecule or a peptide antagonist.

With regard to pose prediction, results are not completely satisfactory. Binding modes obtained from the different dockings seem not to be accurate enough.

Nevertheless, retrospective docking analysis shows a good selection of known CXCR4 binders from decoys when defining a binding pocket delimited by residues Asp171, Asp262, and Glu288. In a similar manner, the homology model shows high selectivity in the classification of actives against decoys mainly in the first percentages of database screened. A detailed analysis of the retrospective docking results using the CXCR4 homology model in Discovery Studio allows us to hypothesize the existence of at least two binding sub-sites in CXCR4 binding pocket.

Summarizing the docking-based VS performance, we found that the crystal structure (3OE6) retrieved the most accurate results, especially regarding the early recovery (aEF of 11.2 at 1% of decoys screened) and the ROC plot analysis (aAUC of 0.82), despite being the lowest resolution structure (3.2 Å). Moreover, scaffold retrieval analysis of PDB 3OE6 confirmed that all chemotype families used in this study were early recovered for all scoring functions. It is also worth noting that enrichment factors and AUC data obtained for PDB 3OE0 (co-crystallized with the peptide antagonist CVX15) are very similar to the rest of PDBs (co-crystallized with the small ligand IT1t). Hence, the fact that peptidic ligands often generate induced fit conformational changes into the protein structure seems not to affect in this case the majority of VS results.

Overall, herein we analyze in depth the receptor-based VS performance of the five recent crystallized CXCR4 structures, along with our previously published CXCR4 rhodopsin-based homology model. We propose PDB 3OE6 as the most suitable crystal structure to carry on new prospective docking analyses with the maximum guarantees. We also show that our previous CXCR4 rhodopsin-based homology model discriminates actives from decoys at the first percentage of database screened as well as the crystal structures do. We hypothesize that these both structures perform well due to their capacity to capture in some way a more general induced fit conformation to which a wide range of ligands can bind due to the known CXCR4 allosteric behavior. These results support the notion that if no information is available, as it was the case of CXCR4 target years ago, homology models can be useful. Moreover, a detailed analysis of the retrospective docking results using the CXCR4 homology model in Discovery Studio allows us to hypothesize the existence of multiple binding sub-sites in CXCR4 binding pocket.

Acknowledgements

The authors thank the Servei de Disseny de Farmacs (SDF) in the Centre de Supercomputació de Catalunya (CESCA) for providing access to the computational software used through. This work is supported by the Programa Nacional de Biomedicina (Ministerio de Educación y Ciencia, SAF2010-21617-CO2-02). V.I.P.N. is grateful for a Marie Curie IEF Fellowship grant reference DOVSA 254128.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jmgm.2012.06.010>.

References

- [1] J.C. Tilton, R.W. Doms, Entry inhibitors in the treatment of HIV-1 infection, *Antiviral Research* 85 (2010) 91–100.
- [2] S.A. Trushin, G.D. Bren, A.D. Badley, CXCR4 tropic HIV-1 gp120 inhibition of SDF-1 α -induced chemotaxis requires Lck and is associated with Cofilin phosphorylation, *Open Virology Journal* 4 (2010) 157.
- [3] P.R. Harrigan, L. Swenson, R.A. McGovern, G.J. Pollock, The determinants and consequences of HIV coreceptor switching, *HIV Medical Update* 4 (2009).
- [4] T. Haga, S. Takeda, G Protein-coupled Receptors: Structure, Function, and Ligand Screening, CRC Press, Boca Raton, 2005.
- [5] W.T. Choi, J. An, Biology and clinical relevance of chemokines and chemokine receptors CXCR4 and CCR5 in human diseases, *Experimental Biology and Medicine* 236 (2011) 637–647.
- [6] M. Peeters, G. van Westen, Q. Li, A. IJzerman, Importance of the extracellular loops in G protein-coupled receptors for ligand recognition and receptor activation, *Trends in Pharmacological Sciences* 32 (2011) 35–42.
- [7] B. Wu, E.Y.T. Chien, C.D. Mol, G. Fenalti, W. Liu, V. Katritch, R. Abagyan, A. Brooun, P. Wells, F.C. Bi, Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists, *Science* 330 (2010) 1066–1071.
- [8] C.A. Mosley, L.J. Wilson, J.M. Wiseman, J.W. Skudlarek, D.C. Liotta, Recent patents regarding the discovery of small molecule CXCR4 antagonists, *Expert Opinion on Therapeutic Patents* 19 (2009) 23–38.
- [9] G. Moncunill, M. Armand-Ugon, I. Clotet-Codina, E. Pauls, E. Ballana, A. Llano, B. Romagnoli, J.W. Vrijbloed, F.O. Gombert, B. Clotet, Anti-HIV activity and resistance profile of the CXCR4 chemokine receptor 4 antagonist POL3026, *Molecular Pharmacology* 73 (2008) 1264.
- [10] S. Pettersson, V.I. Pérez Nuño, M.P. Mena, B. Clotet, J.A. Esté, J.I. Borrell, J. Teixidó, Novel monocyclam derivatives as HIV entry inhibitors: design, synthesis anti HIV evaluation, and their interaction with the CXCR4 co receptor, *ChemMedChem* 5 (2010) 1272–1281.
- [11] V.I. Pérez-Nuño, D.W. Ritchie, J.I. Borrell, J. Teixido, ChemInform abstract: clustering and classifying diverse HIV entry inhibitors using a novel consensus shape-based virtual screening approach: further evidence for multiple binding sites within the CCR5 extracellular pocket, *Journal of Chemical Information and Modeling* 48 (2008) 2146–2165.
- [12] S.P. Kawatkar, M. Yan, H. Gevariya, M.Y. Lim, S. Eisold, X. Zhu, Z. Huang, J. An, Computational analysis of the structural mechanism of inhibition of chemokine receptor CXCR4 by small molecule antagonists, *Experimental Biology and Medicine* 236 (2011) 844–850.
- [13] M.A.C. Neves, S. Simões, M.L. Sá e Melo, Ligand-guided optimization of CXCR4 homology models for virtual screening using a multiple chemotype approach, *Journal of Computer-Aided Molecular Design* 24 (2010) 1023–1033.
- [14] A.R. Lam, S. Bhattacharya, K. Patel, S.E. Hall, A. Mao, N. Vaidehi, Importance of receptor flexibility in binding of cyclam compounds to the chemokine receptor CXCR4, *Journal of Chemical Information and Modeling* 51 (2010) 137–147.
- [15] R. Wang, Y. Lu, X. Fang, S. Wang, An extensive test of 14 scoring functions using the PDB bind refined set of 800 protein–ligand complexes, *Journal of Chemical Information and Computer Science* 44 (2004) 2114–2125.
- [16] O. Trott, A.J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, *Journal of Computational Chemistry* 31 (2010) 455–461.
- [17] M.L. Verdonk, J.C. Cole, M.J. Hartshorn, C.W. Murray, R.D. Taylor, Improved protein–ligand docking using GOLD, *Proteins: Structure, Function, and Bioinformatics* 52 (2003) 609–623.
- [18] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, T. Daniel, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, Glide: a new approach for rapid, accurate docking and scoring 1. Method and assessment of docking accuracy, *Journal of Medicinal Chemistry* 47 (2004) 1739–1749.
- [19] T.A. Pham, A.N. Jain, Customizing scoring functions for docking, *Journal of Computer-Aided Molecular Design* 22 (2008) 269–286.
- [20] SYBYL-X 1.1, Tripos International, 1699 South Hanley Rd., St. Louis, MO 63144, USA.
- [21] A. Krammer, P.D. Kirchhoff, X. Jiang, C.M. Venkatachalam, M. Waldman, LigScore: a novel scoring function for predicting binding affinities, *Journal of Molecular Graphics and Modelling* 23 (2005) 395–407.
- [22] D.K. Gehlhaar, G.M. Verkhivker, P.A. Rejto, C.J. Sherman, D.R. Fogel, L.J. Fogel, S.T. Freer, Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming, *Chemistry and Biology* 2 (1995) 317–324.
- [23] Y. Bustanji, M.O. Taha, A.M. Yousef, A.G. Al-Bakri, Berberine potently inhibits protein tyrosine phosphatase 1B: investigation by docking simulation and experimental validation, *Journal of Enzyme Inhibition and Medicinal Chemistry* 21 (2006) 163–171.
- [24] V.I. Pérez-Nuño, D.W. Ritchie, O. Rabal, R. Pascual, J.I. Borrell, J. Teixido, Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape matching and ligand–receptor docking, *Journal of Chemical Information and Modeling* 48 (2008) 509–533.
- [25] I. Kufareva, M. Rueda, V. Katritch, G. Dock, R.C. Stevens, R. Abagyan, Status of GPCR modeling and docking as reflected by community-wide GPCR Dock 2010 assessment, *Structure* 19 (2011) 1108–1126.
- [26] X. Liang, CXCR4, inhibitors and mechanisms of action, *Chemical Biology and Drug Design* 72 (2008) 97–110.
- [27] A.P. Graves, R. Brenk, B.K. Shoichet, Decoys for docking, *Journal of Medicinal Chemistry* 48 (2005) 3714–3728.
- [28] MOE (Molecular Operating Environment), Chemical Computing Group Inc., Montreal, Canada, 2009.
- [29] S. Hatse, K. Princen, K. Vermeire, L.O. Gerlach, M.M. Rosenkilde, T.W. Schwartz, G. Bridger, E. De Clercq, D. Schols, Mutations at the CXCR4 interaction sites for AMD3100 influence anti-CXCR4 antibody binding and HIV-1 entry, *FEBS Letters* 546 (2003) 300–306.
- [30] L.O. Gerlach, R.T. Skerlj, G.J. Bridger, T.W. Schwartz, Molecular interactions of cyclam and bicyclam non-peptide antagonists with the CXCR4 chemokine receptor, *Journal of Biological Chemistry* 276 (2001) 14153–14160.
- [31] M.M. Rosenkilde, L.O. Gerlach, J.S. Jakobsen, R.T. Skerlj, G.J. Bridger, T.W. Schwartz, Molecular mechanism of AMD3100 antagonism in the CXCR4 receptor, *Journal of Biological Chemistry* 279 (2004) 3033.
- [32] M.M. Rosenkilde, L.O. Gerlach, S. Hatse, R.T. Skerlj, D. Schols, G.J. Bridger, T.W. Schwartz, Molecular mechanism of action of monocyclam versus bicyclam non-peptide antagonists in the CXCR4 chemokine receptor, *Journal of Biological Chemistry* 282 (2007) 27354–27365.
- [33] H. Tamamura, Y. Xu, T. Hattori, X. Zhang, R. Arakaki, K. Kanbara, A. Omagari, A. Otake, T. Ibuka, N. Yamamoto, A low-molecular-weight inhibitor against the chemokine receptor CXCR4: a strong anti-HIV peptide T140, *Biochemical and Biophysical Research Communications* 253 (1998) 877–882.
- [34] J.O. Trent, Z. Wang, J.L. Murray, W. Shao, H. Tamamura, N. Fujii, S.C. Peiper, Lipid bilayer simulations of CXCR4 with inverse agonists and weak partial agonists, *Journal of Biological Chemistry* 278 (2003) 47136.
- [35] J. Våbenø, G.V. Nikiforovich, G.R. Marshall, Insight into the binding mode for cyclopentapeptide antagonists of the CXCR4 receptor, *Chemical Biology and Drug Design* 67 (2006) 346–354.
- [36] T. Fawcett, An introduction to ROC analysis, *Pattern Recognition Letters* 27 (2006) 861–874.
- [37] Y. Jiang, C.E. Metz, R.M. Nishikawa, A receiver operating characteristic partial area index for highly sensitive diagnostic tests, *Radiology* 201 (1996) 745.
- [38] P. Sonogo, A. Kocsor, S. Pongor, ROC analysis: applications to the classification of biological sequences and 3D structures, *Briefings in Bioinformatics* 9 (2008) 198.
- [39] N.A. Obuchowski, ROC analysis, *American Journal of Roentgenology* 184 (2005) 364.
- [40] T.A. Lasko, J.G. Bhagwat, K.H. Zou, L. Ohno-Machado, The use of receiver operating characteristic curves in biomedical informatics, *Journal of Biomedical Informatics* 38 (2005) 404–415.
- [41] D.K. McClish, Analyzing a portion of the ROC curve, *Medical Decision Making* 9 (1989) 190.

- [42] A.N. Jain, Bias, reporting, and sharing: computational evaluations of docking methods, *Journal of Computer-Aided Molecular Design* 22 (2008) 201–212.
- [43] M.D. Mackey, J.L. Melville, Better than random? The chemotype enrichment problem, *Journal of Chemical Information and Modeling* 49 (2009) 1154–1162.
- [44] A.N. Jain, A. Nicholls, Recommendations for evaluation of computational methods, *Journal of Computer-Aided Molecular Design* 22 (2008) 133–139.
- [45] J.B. Cross, D.C. Thompson, B.K. Rai, J.C. Baber, K.Y. Fan, Y. Hu, C. Humblet, Comparison of several molecular docking programs: pose prediction and virtual screening accuracy, *Journal of Chemical Information and Modeling* 49 (2009) 1455–1474.
- [46] J. Kirchmair, P. Markt, S. Distinto, G. Wolber, 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? *Journal of Computer-Aided Molecular Design* 22 (2008) 213–228.
- [47] R.S.Y. Wong, V. Bodart, M. Metz, J. Labrecque, G. Bridger, S.P. Fricker, Comparison of the potential multiple binding modes of bicyclam, monocyclam, and noncyclam small-molecule CXC chemokine receptor 4 inhibitors, *Molecular Pharmacology* 74 (2008) 1485–1495.
- [48] E. De Clercq, Recent advances on the use of the CXCR4 antagonist plerixafor (AMD3100 MozobilTM) and potential of other CXCR4 antagonists as stem cell mobilizers, *Pharmacology and Therapeutics* 128 (2010) 509–518.
- [49] J.G. Catalano, K.S. Gudmundsson, A. Svolto, S.D. Boggs, J.F. Miller, A. Spaltenstein, M. Thomson, P. Wheelan, D.J. Minick, D.P. Phelps, Synthesis of a novel tricyclic 1, 2, 3, 4, 4a, 5, 6, 10b-octahydro-1, 10-phenanthroline ring system and CXCR4 antagonists with potent activity against HIV-1, *Bioorganic and Medicinal Chemistry Letters* 20 (2010) 2186–2190.
- [50] G. Bridger, R. Skerlj, A. Kaller, C. Harwig, D. Bogucki, T.R. Wilson, J. Crawford, E.J. Mceachern, B. Atsma, S. Nan, Chemokine Receptor Binding Heterocyclic Compounds, 2004.
- [51] G.J. Bridger, R.T. Skerlj, S. Padmanabhan, S.A. Martellucci, G.W. Henson, S. Struyf, M. Witvrouw, D. Schols, E. De Clercq, Synthesis and structure–activity relationships of phenylenebis (methylene)-linked bis-azamacrocycles that inhibit HIV-1 and HIV-2 replication by antagonism of the chemokine receptor CXCR4, *Journal of Medicinal Chemistry* 42 (1999) 3971–3981.
- [52] G.J. Bridger, R.T. Skerlj, D. Thornton, S. Padmanabhan, S.A. Martellucci, G.W. Henson, M.J. Abrams, N. Yamamoto, K.D. Vreese, Synthesis and structure–activity relationships of phenylenebis (methylene)-linked bis-tetraazamacrocycles that inhibit HIV replication Effects of macrocyclic ring size and substituents on the aromatic linker, *Journal of Medicinal Chemistry* 38 (1995) 366–378.
- [53] K. Ichiyama, S. Yokoyama-Kumakura, Y. Tanaka, R. Tanaka, K. Hirose, K. Bannai, T. Edamatsu, M. Yanaka, Y. Niitani, N. Miyano-Kurosaki, A duodenally absorbable CXC chemokine receptor 4 antagonist KRH-1636, exhibits a potent and selective anti-HIV-1 activity, *PNAS* 100 (2003) 4185–4190.
- [54] T. Yamazaki, A. Saitou, M. Ono, S. Yokoyama, K. Bannai, K. Hirose, M. Yanaka, Novel Nitrogenous Compound and Use Thereof, 2003.
- [55] G.L. Warren, C.W. Andrews, A.M. Capelli, B. Clarke, J. LaLonde, M.H. Lambert, M. Lindvall, N. Nevins, S.F. Semus, S. Senger, A critical assessment of docking programs and scoring functions, *Journal of Medicinal Chemistry* 49 (2006) 5912–5931.
- [56] S. Vilar, G. Ferino, S.S. Phatak, B. Berk, C.N. Cavasotto, S. Costanzi, Docking-based virtual screening for ligands of G protein-coupled receptors: not only crystal structures but also in silico models, *Journal of Molecular Graphics and Modelling* 29 (2011) 614–623.
- [57] L. Roumen, M. Sanders, B. Vroliing, I.J.P. de Esch, J. de Vlieg, R. Leurs, J.P.G. Klomp, S.B. Nabuurs, C. de Graaf, In silico veritas: the pitfalls and challenges of predicting GPCR–ligand interactions, *Pharmaceuticals* 4 (2011) 1196–1215.
- [58] W.M. Rockey, A.H. Elcock, Structure selection for protein kinase docking and virtual screening: homology models or crystal structures? *Current Protein and Peptide Science* 7 (2006) 437–457.
- [59] V.I. Perez-Nueno, D.W. Ritchie, Using consensus-shape clustering to identify promiscuous ligands and protein targets and to choose the right query for shape-based virtual screening, *Journal of Chemical Information and Modeling* 51 (2011) 1233–1248.
- [60] V.I. Pérez-Nueno, D.W. Ritchie, Applying in silico tools to the discovery of novel CXCR4 inhibitors, *Drug Development Research* 72 (2011) 95–111.
- [61] V.I. Pérez-Nueno, D.W. Ritchie, Identifying and characterizing promiscuous targets: implications for virtual screening, *Expert Opinion on Drug Discovery* 7 (2012) 1–17.
- [62] L.T. May, K. Leach, P.M. Sexton, A. Christopoulos, Allosteric modulation of G protein-coupled receptors, *Annual Review of Pharmacology and Toxicology* 47 (2007) 1–51.
- [63] J.M. Janz, Y. Ren, R.J. Looby, M.A. Kazmi, P. Sachdev, A. Grunbeck, L. Haggis, D. Chinnapan, A.Y. Lin, C. Seibert, Direct interaction between an allosteric agonist pepducin and the chemokine receptor CXCR4, *Journal of the American Chemical Society* 40 (2011) 15878–15881.
- [64] P. Maurice, M. Kamal, R. Jockers, Asymmetry of GPCR oligomers supports their functional relevance, *Trends in Pharmacological Sciences* 32 (2011) 514–520.
- [65] J.A. Ballesteros, H. Weinstein, Integrated methods for the construction of three-dimensional models and computational probing of structure–function relations in G protein-coupled receptors, *Methods in Neurosciences* 25 (1995) 366–428.

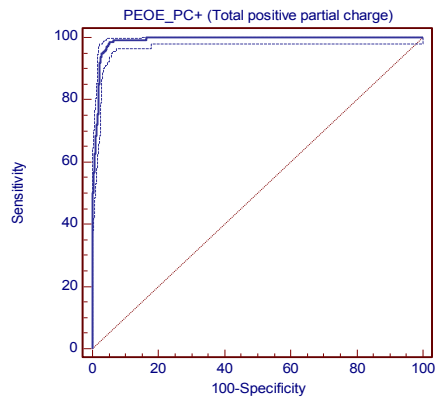
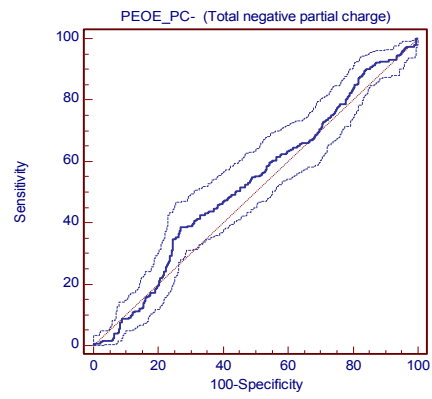
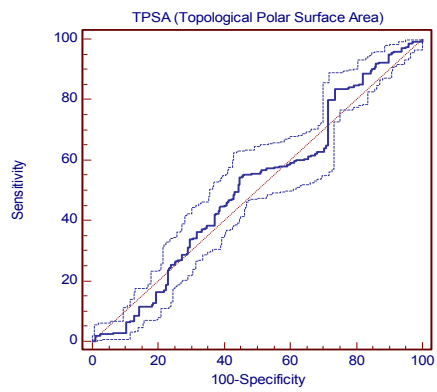
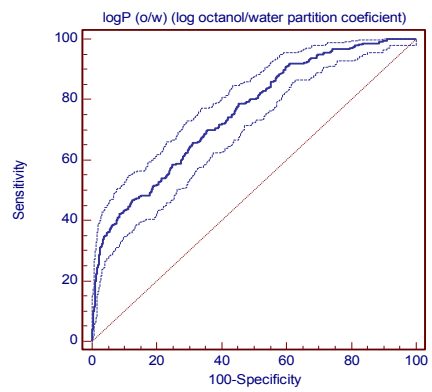
A**B****C****D**

Figure 1 ROC plots obtained when ligands and decoys were classified by their total positive partial charge (PEOE PC+; graphic A), total negative partial charge (PEOE PC-; graphic B), total polar surface area (TPSA; graphic C) and log octanol/water partition coefficient (logP o/w; graphic D).

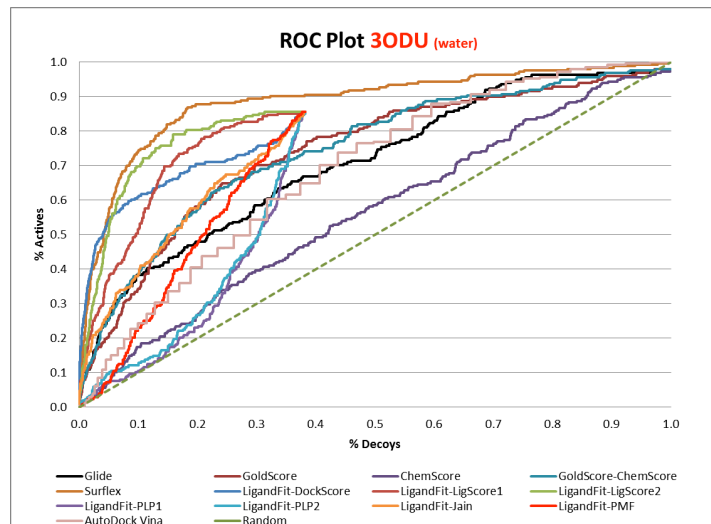


Figure 2 Roc plot of the retrospective docking-based VS of PDB 3ODU without removing cocrystallized water molecules. The graphic shows how the VS does not recover completely the test set (only 1951 molecules recovered) for Ligandfit scoring functions.

Enrichment Factors	% Screened Database	3OE6	3ODU	3ODU _{water}
GoldScore	1%	3.3	1.2	6.8
	5%	2.5	1.8	4.7
	10%	1.9	1.9	3.5
ChemScore	1%	10.5	3.6	2.0
	5%	7.0	2.6	1.5
	10%	5.1	2.3	1.7
GoldScore & ChemScore	1%	5.6	3.2	7.7
	5%	4.8	3.0	3.8
	10%	3.8	2.7	3.3
Glide	1%	8.5	2.8	8.1
	5%	5.4	4.3	4.7
	10%	4.2	4.6	3.5
Surflex TotalScore	1%	15.1	9.7	11.3
	5%	9.2	5.4	8.5
	10%	6.8	3.5	6.7
LigandFit DockScore	1%	19.7	18.7	15.7
	5%	12.4	11.7	9.3
	10%	8.3	8.3	5.8
LigandFit LigScore1	1%	19.7	14.7	9.3
	5%	12.4	8.6	5.8
	10%	8.3	6.1	4.5
LigandFit LigScore2	1%	14.8	15.5	10.5
	5%	11.3	11.5	7.4
	10%	8.0	8.0	6.1
LigandFit PLP1	1%	10.3	9.8	1.6
	5%	8.5	8.2	1.4
	10%	6.7	6.4	1.0
LigandFit PLP2	1%	9.4	9.8	2.0
	5%	7.4	6.7	1.9
	10%	5.8	5.4	1.2
LigandFit Jain	1%	12.3	10.2	8.5
	5%	5.4	5.7	4.9
	10%	3.4	4.0	3.5
LigandFit PMF	1%	10.7	5.7	1.2
	5%	7.3	3.9	1.5
	10%	5.4	3.0	2.1
AutoDock Vina	1%	5.1	2.4	1.2
	5%	3.6	1.9	2.7
	10%	3.0	1.8	2.3

Table 1 Comparison of EF obtained for all scoring functions and PDBs: 3OE6 (the best VS performance model), 3ODU when water molecules were removed from the structure, and PDB 3ODU when all water molecules remained in the protein structure (3ODU_{water}).

pAUC	% False positive Screened	3OE6	3ODU	3ODU_{water}
GoldScore	1%	0.55	0.51	0.52
	5%	0.63	0.53	0.56
	10%	0.67	0.55	0.58
ChemScore	1%	0.50	0.50	0.50
	5%	0.52	0.51	0.51
	10%	0.53	0.53	0.52
GoldScore & ChemScore	1%	0.52	0.51	0.53
	5%	0.57	0.54	0.57
	10%	0.61	0.56	0.60
Glide	1%	0.54	0.50	0.53
	5%	0.60	0.55	0.57
	10%	0.63	0.62	0.60
Surflex TotalScore	1%	0.60	0.54	0.56
	5%	0.71	0.59	0.67
	10%	0.77	0.62	0.74
LigandFit DockScore	1%	0.70	0.66	0.60
	5%	0.80	0.79	0.69
	10%	0.86	0.86	0.73
LigandFit LigScore1	1%	0.56	0.59	0.54
	5%	0.64	0.68	0.60
	10%	0.69	0.74	0.64
LigandFit LigScore2	1%	0.60	0.62	0.54
	5%	0.75	0.78	0.64
	10%	0.83	0.85	0.71
LigandFit PLP1	1%	0.55	0.55	0.50
	5%	0.66	0.66	0.51
	10%	0.74	0.73	0.51
LigandFit PLP2	1%	0.55	0.54	0.51
	5%	0.64	0.62	0.51
	10%	0.70	0.68	0.52
LigandFit Jain	1%	0.56	0.55	0.54
	5%	0.60	0.60	0.58
	10%	0.62	0.63	0.61
LigandFit PMF	1%	0.55	0.52	0.50
	5%	0.63	0.55	0.49
	10%	0.69	0.58	0.52
AutoDock Vina	1%	0.51	0.50	0.50
	5%	0.54	0.49	0.52
	10%	0.57	0.52	0.54

Table 2 Comparison of scoring functions pAUC values obtained for: PDB 3OE6 (the best VS performance model), PDB 3ODU when water molecules were removed from the protein structure, and PDB 3ODU with water molecules (3ODU_{water}) were kept for VS.

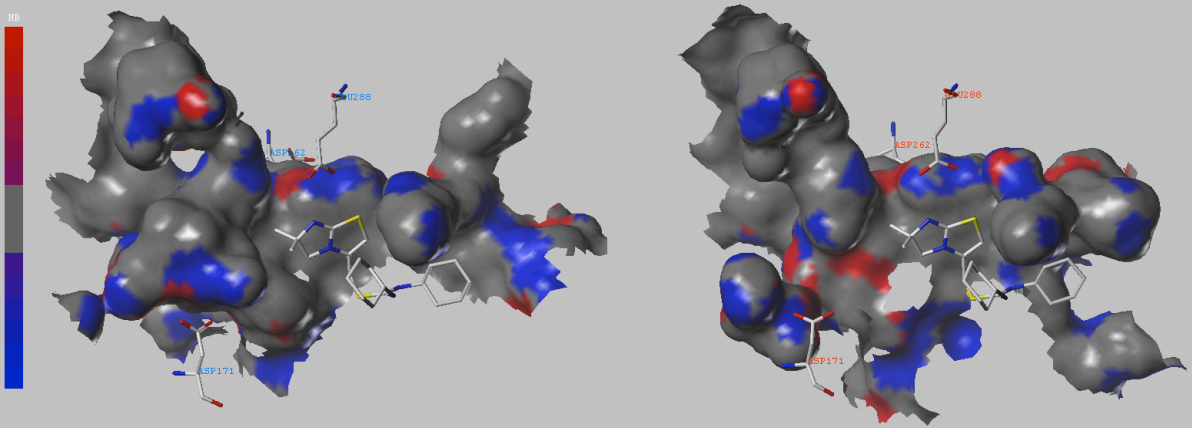
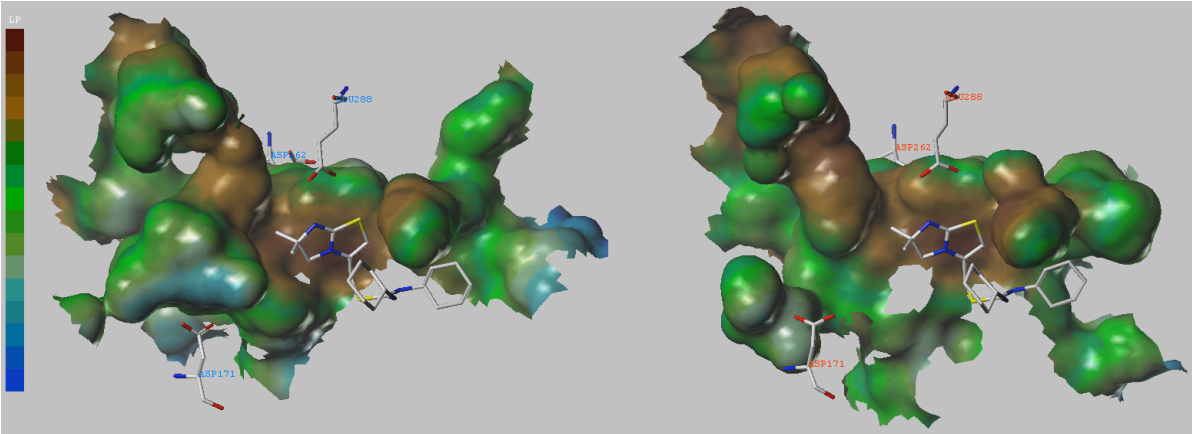
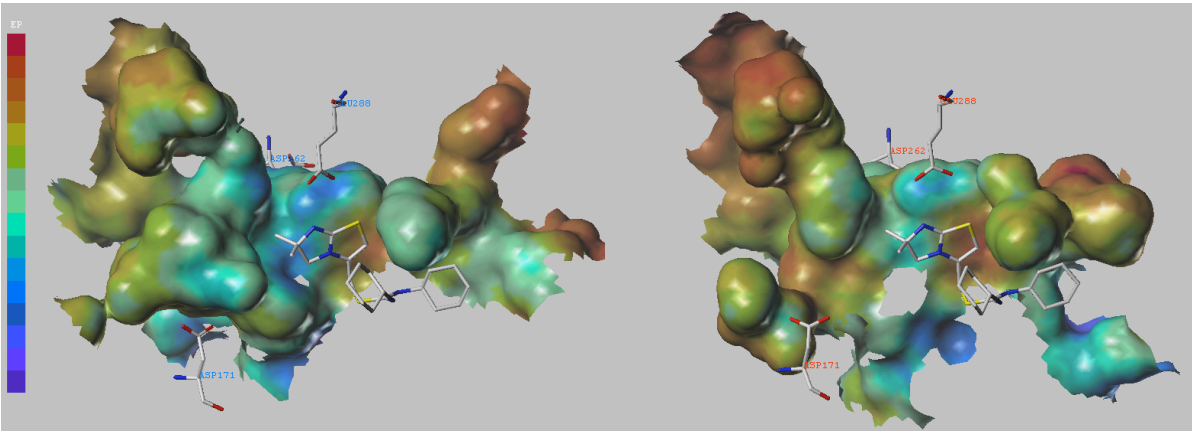
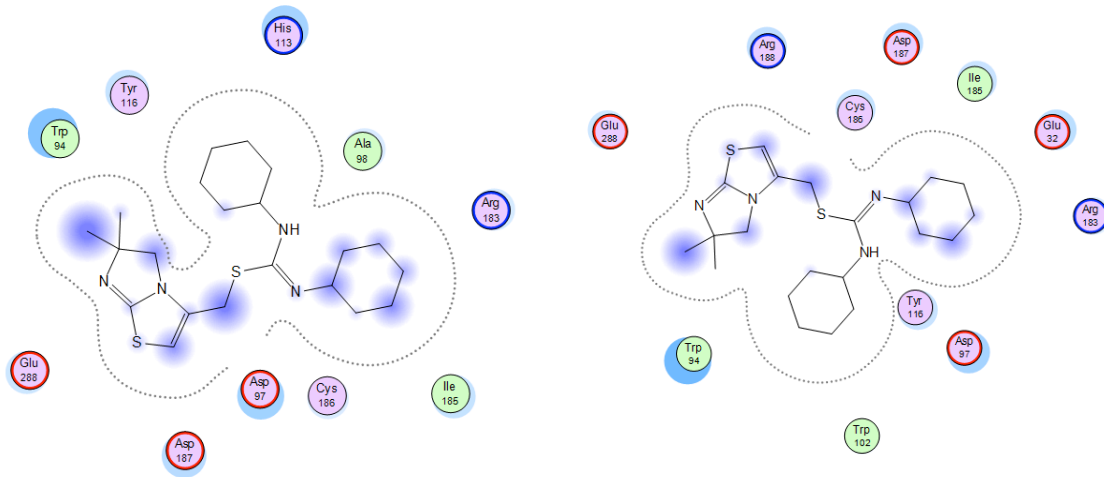
A**B****C****D**

Figure 3 Comparison of surface and interaction maps for PDBs 3OE6 (left) at 3.2 Å crystallographic resolution and 3ODU (right) at 2.5 Å crystallographic resolution. Both PDBs contain the same IT1t ligand.

(A) Hydrogen bonding surfaces map. In red color is shown hydrogen bond donor and low electronegativity, in grey no hydrogen bond donor or acceptor and in blue, hydrogen bond acceptor and high electronegativity areas.

(B) Lipophilic potential surface map. The lipophilic potential is represented in a color scale ranging from brown (the highest lipophilic area of the molecule) to blue (the highest hydrophilic surface area).

(C) Electrostatic potential surface map. The electrostatic potential is represented in a color scale ranging from red (the most positive electrostatic potential) to purple (the most negative electrostatic potential).

(D) Protein-ligand interaction maps for IT1t in their corresponding binding pockets for PDB 3OE6 (left) and 3ODU (right).

Artículo 3

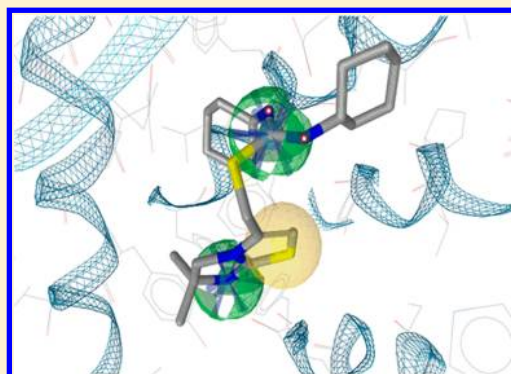
Highly Specific and Sensitive Pharmacophore Model for Identifying CXCR4 Antagonists. Comparison with Docking and Shape-Matching Virtual Screening Performance

Arnaud S. Karaboga,^{†,§} Jesús M. Planesas,^{‡,§} Florent Petronin,[†] Jordi Teixidó,[‡] Michel Souchet,^{*,†} and Violeta I. Pérez-Nueno^{*,†,‡}

[†]Harmonic Pharma, Espace Transfert, 615 rue du Jardin Botanique, 54600 Villers lès Nancy, France

[‡]Grup d'Enginyeria Molecular, Institut Químic de Sarrià (IQS), Universitat Ramon Llull, Barcelona, Spain

ABSTRACT: HIV infection is initiated by fusion of the virus with the target cell through binding of the viral gp120 protein with the CD4 cell surface receptor protein and the CXCR4 or CCR5 coreceptors. There is currently considerable interest in developing novel ligands that can modulate the conformations of these coreceptors and, hence, ultimately block virus–cell fusion. Herein, we present a highly specific and sensitive pharmacophore model for identifying CXCR4 antagonists that could potentially serve as HIV entry inhibitors. Its performance was compared with docking and shape-matching virtual screening approaches using 3OE6 CXCR4 crystal structure and high-affinity ligands as query molecules, respectively. The performance of these methods was compared by virtually screening a library assembled by us, consisting of 228 high affinity known CXCR4 inhibitors from 20 different chemotype families and 4696 similar presumed inactive molecules. The area under the ROC plot (AUC), enrichment factors, and diversity of the resulting virtual hit lists was analyzed. Results show that our pharmacophore model achieves the highest VS performance among all the docking and shape-based scoring functions used. Its high selectivity and sensitivity makes our pharmacophore a very good filter for identifying CXCR4 antagonists.



■ INTRODUCTION

HIV cell infection is a multistage and complex process. It starts with the virus entry¹ into the host membrane. The first stage of the entry process is the virus attachment to the cell membrane through receptor CD4. Second, viral proteins *gp120*² and *gp41* bind to the coreceptors CCR5 and CXCR4³ to facilitate virus anchoring to the host cellular membrane. Finally, the viral envelope undergoes conformational changes that allows the insertion of the *gp41* fusion peptide into the host membrane.

One of the strategies to block the HIV entry process is inhibiting the binding of the viral protein *gp120* to the coreceptors CXCR4 or CCR5. The first interaction between *gp120* and these coreceptors is through their extracellular loops. CXCR4 and CCR5 are both chemokine receptors which belong to the GPCR superfamily⁴ and are implicated in a wide range of human diseases.⁵ This subfamily of GPCR proteins is responsible for transmitting signals from the extracellular environment as well as the recognition of extracellular ligands. All GPCRs have a common characteristic structure relatively hydrophobic with seven transmembrane domains. The N-terminus is located in the extracellular side and the transmembrane helices are connected by three extracellular loops whereas the C-terminus is located in the intracellular side facing the cytoplasm and three intracellular loops connect the interior domain of the transmembrane helices. The transmembrane

helices are expanded in the plasmatic membrane as an anticlockwise bundle of alpha-helices. Therefore extracellular loops fulfill important roles in receptor activation and consequently in ligand binding.⁶

An important achievement was accomplished at the end of 2010 when Wu et al.⁷ crystallized five independent CXCR4 structures with resolution of 2.5–3.2 Å. These crystal structures contain a small organic molecule derived from isothioureia⁸ (IT1t) in Protein Data Bank structures (PDBs) 3ODU, 3OE6, 3OE8, and 3OE9 and a 16-residue cyclic peptide⁹ (CVX15) in PDB 3OE0, as antagonists bound to the CXCR4 structure. As many other chemokine receptors, these CXCR4 structures show different monomer association.¹⁰ PDBs 3OE0 and 3OE6 show monomeric structure. PDBs 3ODU and 3OE9 have been crystallized as homodimers, and PDB 3E80 is a homotrimeric protein. These processes of homo and mainly oligodimerization seem to play an important role in the allostereism process.¹¹

In all the PDBs, we can observe that the ligand binding site is located in the second extracellular loop, ECL2, close to the N-terminus. Also, superposition of ligand binding pockets shows similar overlapping volume. The residues that induce common interactions in both ligands are Tyr 116, Asp 187, and Glu 288.

Received: January 17, 2013

Those residues are common in CXCR4 antagonist binding as it has been widely discussed in previous literature.^{12,13} Although the overall orientation differs between ligands, hydrophobic interactions highly contribute in the binding affinity of the majority of CXCR4 antagonists. The residues Ala95, Leu91, Val112, His113, Tyr116, Phe292, Tyr255, and Ile259 play important roles in hydrophobic binding. Moreover, Tyr45 on TM1, Trp94 and Asp97 on TM2, Tyr116 on TM3, Asp171 on TM4, Asp262 on TM6, and His281 and Glu288 on TM7 are critical residues involved in stabilizing interactions with known CXCR4 antagonists.^{14–16}

Nevertheless, in the case of the peptide ligand a ligand-induced conformational change in the extracellular side of the helix 5⁷ is observed. Hence, the structure-based virtual screening (VS) of small peptides might be more dependent on the CXCR4 structure used for the screening.

Planesas et al.¹⁷ recently compared and evaluated for the first time all CXCR4 structures published by Wu et al. together with their previously published CXCR4 rhodopsin-based homology model from the point of view of their receptor based virtual screening performance. In their work, they showed that the crystal structure 3OE6 retrieved the most accurate results in their docking-based VS study, especially regarding the early recovery (average enrichment factor (EF) of 11.7 at 1% of decoys screened) and the area under the curve (AUC) of the ROC plot (average AUC of 0.84), despite being the lowest resolution structure (3.2 Å). Moreover, scaffold retrieval analysis of PDB 3OE6 confirmed that all chemotype families used in their study were early recovered for all scoring functions. Hence, in the present work, we use PDB 3OE6 as the most suitable crystal structure to carry on the docking benchmarking of our CXCR4 database.

Hence, we used the PDB 3OE6 crystal structure to build a CXCR4 pharmacophore model by incorporating protein–ligand interaction information to define the pharmacophoric features. So far, this is the first CXCR4 pharmacophore published in this way. Here we present the convenience of using this pharmacophore model for identifying new CXCR4 antagonists given its high sensitivity and selectivity, and we compare its performance with classical structure-based and ligand-based methods.

METHODS

CXCR4 Database and Decoy Preparation. We compiled a set of 228 CXCR4 antagonist inhibitors from the literature, which mainly consists of 20 representative families of CXCR4 inhibitors, as listed in Table 1. In the compiled data set, all molecules have activity values of <59 μM against CXCR4. Figure 1 shows some representative members of each family.

To avoid potential bias of the virtual screening results due to large differences in basic properties such as molecular weight, number of rotatable single bonds, number of hydrogen-bond acceptor atoms, number of hydrogen-bond donor atoms, total polar surface area (TPSA), and number of hydrophobic atoms, it is important to work with focused screening libraries.¹⁸ Hence, a set of 4696 presumed inactive compounds was assembled from the Maybridge Screening Collection database¹⁹ in such a way that several physicochemical properties were similar to those of the active compounds. Table 2 shows that the average and standard deviations of these properties calculated by MOE 2009 are quite similar for the active and inactive pools. This decoy compilation follows the observations made by Krüger and Evers²⁰ who propose the use of a

Table 1. Families of CXCR4 Inhibitors Compiled in the Database

family	number of compounds	ref
AMD/bicyclam derivatives	55	69–72, 44, 73
amine–amine/amine–aldehyde derivatives	41	74
AMD070 derivatives	29	75, 76
KRH derivatives/ peptidomimetic Kureha	18	77, 78, 79
monocyclam derivatives	17	44, 80–82, 43, 42, 12, 74
phenantroline derivatives	14	83
bis-azamacrocycles	13	84, 69
(DpaZn) dipicolilamine zinc derivatives	10	85
tetrahydroquinoline imidazopyridine derivatives	8	86, 87
macrocycles	4	84, 69, 88
isothioureia derivatives	3	7, 8, 79, 15
pyrimidine derivatives	3	89
chalcone derivatives	3	90
tetrahydroquinoline benzimidazole derivatives	2	91–93
bis-imidazole derivatives	2	94, 79
ethanediamine derivatives	2	79
chemocentrix	1	95, 15
Garland Marshall diazepine	1	79
spiro-containing compounds	1	96, 79
MSX-112	1	97

commercial database as an effective approach to represent the chemical space of a decoy set.

The 3D structures of all ligands were built using MOE. The structures were protonated at physiological pH (i.e., pH 7), Gasteiger partial charges were assigned, and the geometry was optimized using the MMFF94 force field.

Anti-HIV Data Set Preparation for Analyzing the Sensitivity and Specificity of the Pharmacophore Model. We assembled four data sets of anti-HIV active ligands comprising 928 integrase inhibitors, 2374 protease inhibitors, 2484 reverse transcriptase inhibitors, and 354 CCR5 antagonists, respectively. Ligand structures for the integrase, protease, and reverse transcriptase inhibitor data sets were extracted from Binding DB database²¹ by searching for the “target Biomolecule” of interest and allowing Binding DB to retrieve also results linked to ChEMBL and PubChem databases. The CCR5 antagonists’ data set was assembled from the work of Pérez-Nueno et al.²² The 3D structures of all ligands were protonated at physiological pH (i.e., pH 7), Gasteiger partial charges were assigned, and the geometry was optimized using the MMFF94 force field in MOE.

Table 2 shows the average and standard deviations for several 1D physicochemical properties calculated by MOE 2009 for the integrase, protease, reverse transcriptase, and CCR5 databases, compared to the properties of the CXCR4 antagonists and CXCR4 compiled decoys.

These anti-HIV data sets were used to analyze both the sensitivity and specificity of the pharmacophore model. Specifically, sensitivity was measured by the ability of the model to identify the CXCR4 antagonists in the first ranking positions, while specificity was measured by the ability of the model to identify only the CXCR4 antagonists and nothing else. Hence, we performed two experiments; in the first one, we analyze the sensitivity of the pharmacophore model by

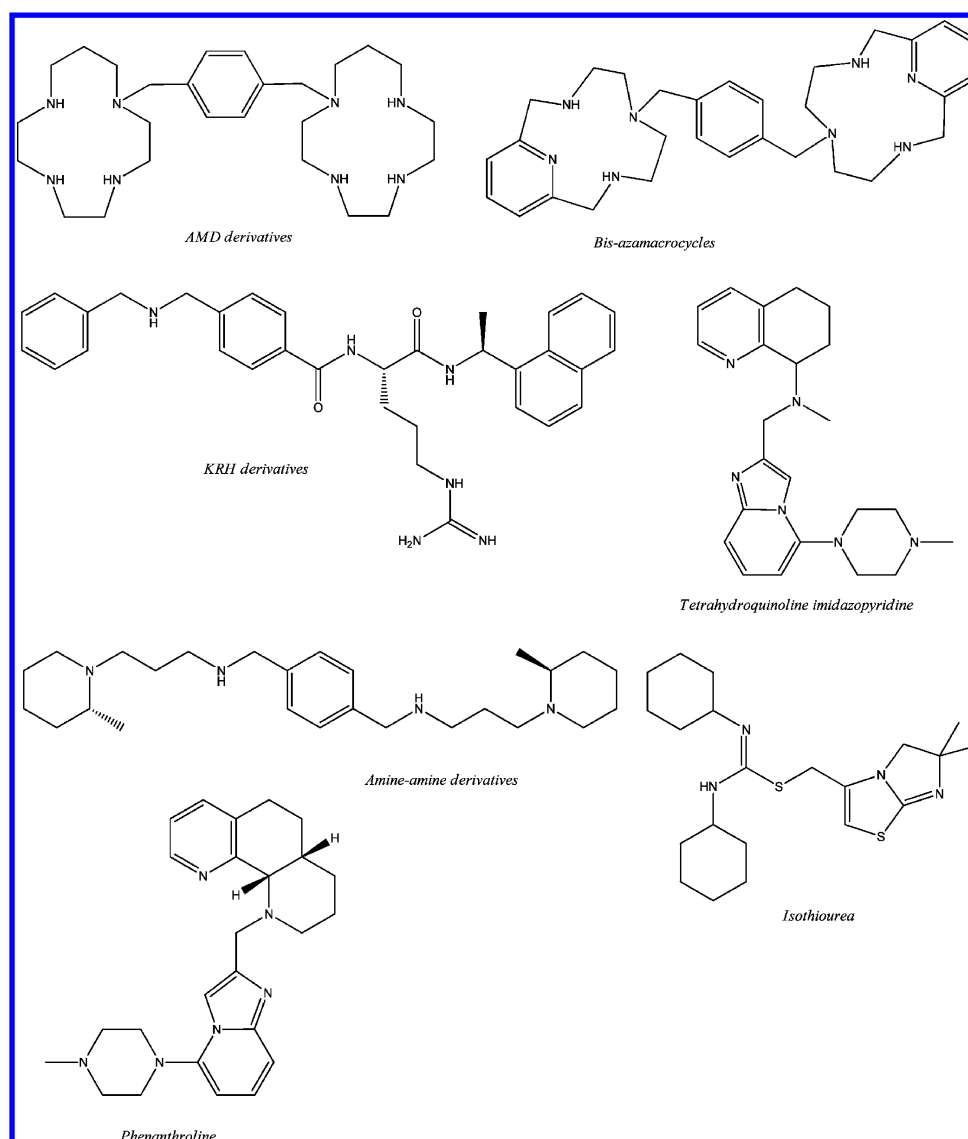


Figure 1. Representative structures of chemotype families in the CXCR4 database.

Table 2. Summary of the 1D Physicochemical Properties of Active and Inactive Molecules in the Screening Database^a

molecule	molecular weight	number of rotatable single bonds b_1rotN	number of hydrogen-bond acceptor atoms, a_acc	number of hydrogen-bond donor atoms, a_don	number of hydrophobic atoms, ahyd	total polar surface area, TPSA
228 CXCR4 antagonists	455.2 (115.5)	7.0 (3.8)	2.4 (1.4)	0.9 (1.1)	24.9 (6.2)	73.5 (32.2)
4696 decoys	497.4 (45.5)	5.8 (2.2)	3.2 (1.6)	0.9 (0.9)	23.7 (4.1)	83.5 (33.4)
354 CCR5 antagonists	560.9 (86.6)	9.3 (3.6)	3.3 (1.6)	0.8 (0.8)	30.3 (4.7)	74.0 (31.5)
928 HIV integrase inh.	465.7 (299.0)	8.1 (10.8)	4.7 (3.4)	2.5 (3.9)	20.4 (3.9)	138.3 (131.1)
2374 HIV protease inh.	597.3 (115.1)	11.69 (4.1)	4.8 (2.1)	3.0 (1.7)	31.0 (5.9)	123.6 (39.8)
2484 HIV reverse transcriptase inh.	388.1 (217.0)	5.9 (6.9)	3.5 (2.2)	1.3 (1.8)	17.7 (10.2)	81.7 (65.0)

^aEach column displays respective average number with standard deviation in parentheses.

considering the different anti-HIV compound databases as actives, and the same data set assembled from Maybridge as decoys; in the second one we analyze the specificity of the pharmacophore model by considering the CXCR4 antagonists as actives and all the other anti-HIV compound databases as decoys.

Sensitivity is generally favored over specificity to prevent the loss of active candidates. Conversely, if the pharmacophore is the ultimate decision manner in a VS cascade, a more specific selection threshold may be set. In any case, it is always needed to make a compromise between sensitivity and specificity of a

model. Improvement of the specificity of a model is always at the cost of losing sensitivity, and vice versa.

Pharmacophore Modeling. A highly specific and sensitive pharmacophore model for identifying CXCR4 antagonists was built using protein–ligand interaction information. We used the automatic pharmacophore generation protocol in LigandScout 3.1 from Inte:Ligand²³ to generate the pharmacophore model directly from the receptor–ligand interactions as revealed in the 3OE6 CXCR4–isothiourea complex. In this first step, seven pharmacophore features of the ligand were identified: (i) two positively charged features located on the two nitrogen atoms of the thiourea group; (ii) two hydrogen-bond donor features respectively centered on the nitrogen of the thiourea and directed toward the backbone carbonyl of Cys286 and on the nitrogen of the aromatic ring to interact with the carboxyl moiety of residue Glu288; (iii) three hydrophobic features positioned on the two gemdimethyl carbons and on the aromatic ring near the thiomethyl linker. In a second step, we manually curate the previous model by keeping and adjusting the features that we considered relevant and mandatory for the protein–ligand interaction, and we removed those considered as unspecific. Eventually, we generated a five points pharmacophore as displayed in Figure 2.

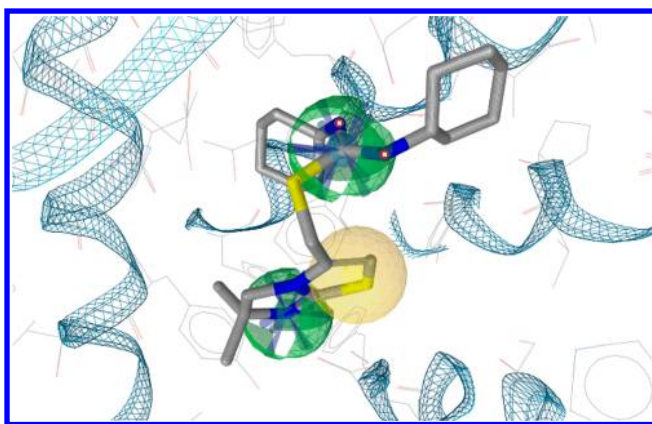


Figure 2. CXCR4 pharmacophore model with a high activity CXCR4 antagonist aligned. Five-featured manually refined final pharmacophore model. The pharmacophore hydrophobic features are shown in yellow. Positively charged features are shown in blue, and hydrogen bond donor features are shown in green.

Our compiled database of actives and decoys was docked onto the CXCR4 3OE6 PDB (see the Docking-Based Virtual Screening section). Given that DockScore scoring function performed the best, the DockScore pose obtained for every active and decoy was used as the starting conformation. These conformations were afterwards converted into LigandScout format using the Idbgen routine implemented in LigandScout 3.1, and 25 conformations per ligand were calculated using the FAST method of OMEGA.²⁴ The routine successfully converted 211 out of 228 actives and 4695 out of 4696 decoys. The remaining 17 CXCR4 inhibitors until the 228 total actives in the database were not considered in the analysis given that LigandScout default force-filed parameters failed to type their atoms (zinc-containing compounds).

Docking-Based Virtual Screening. In the last years GPCR and also chemokine receptors have begun to be determined successfully by protein crystallography. In the case of the chemokine coreceptor CXCR4, Wu et al. obtained five

CXCR4 structures at the end of 2010. According to the work of Planesas et al.,¹⁷ we performed the docking-based virtual screening using 3OE6 CXCR4 crystal structure obtained from the Protein Data Bank (PDB).²⁵ We used different representative force-filed based, empirical, and knowledge-based scoring functions:²⁶ GoldScore,²⁷ ChemScore,²⁷ and GoldScore rescored with ChemScore from Gold 4.1, Glide,²⁸ TotalScore provided in Surflex²⁹ from Sybyl X,³⁰ AutoDock Vina,³¹ and Ligandfit scoring functions (LigScore1-Dreiding; LigScore2-Dreiding;³² PLP1; PLP2;^{33,34} Jain,³⁴ PMF, and Dock_Score³⁵) implemented in Discovery Studio 2.5.

Protein structure 3OE6 is a monomeric protein chain with a crystal resolution of 3.2 Å. It does not contain any crystallized water molecules, and it has a large binding pocket in accordance with the observation that many CXCR4 inhibitors have peptidic or peptidomimetic character. The 3OE6 structure was protonated, histidines were kept in the same protonation state, and chemical molecules used as crystallization aiders were suppressed. This preparation procedure was carried out prior to docking, using the own protein preparation module implemented in each docking tool: the ADT program was used to prepare the corresponding PDBQT-protein file for AutoDock Vina, Protein Preparation Wizard was used for Glide, Protein Setup tool for Gold, Prepare Protein Structure for Sybyl, and Prepare Protein in the case of Discovery Studio. The force fields used to protonate and minimize the protein chain were the following: OPLS-2005,³⁶ TRIPOS,³⁷ and CHARMM³⁸ for Glide, Sybyl X, and Discovery Studio, respectively.

The CXCR4 binding pocket was defined by setting the volume, where the ligands could interact with the protein, delimited by residues Asp171, Asp262, and Glu288, that is in fact the region which comprises the bound ligand IT1t. Several mutagenic studies confirm interactions between Asp171, Asp262, and Glu288 residues with different ligands such as bicyclam AMD3100,^{39–41} monocyclams AMD3465 and AMD3529,⁴² or non cyclams like AMD070^{42–44} in different ways. This triad of acidic residues also interacts with peptidic ligands like T-140, a 14-residue polypeptide^{45,46} or cyclopentapeptides antagonists like FC131 and their derivatives.⁴⁷

Shape-Matching-Based Virtual Screening. Ligand-based virtual screening was performed using PARASURF/PARAFIT,^{48,49} ROCS,⁵⁰ and HPCC.⁵¹ Prior to the screening, OMEGA was used to calculate 25 conformations of each active and decoy according to the same parameters than those used to generate the compound conformations filtered by the pharmacophore model. All shape-based methods were assessed according to their ability to superpose each of the database compounds onto the query isothiourea extracted from the 3OE6 crystal structure. For each compound the conformation showing the best overlay with the query molecule was considered and scored.

PARASURF was used to calculate, from semiempirical quantum mechanics theory, the spherical harmonic (SH) molecular shape of all molecules. PARAFIT was then used to superpose every database compound onto the isothiourea query by exploiting the special rotational properties of the SH expansions. These superpositions used the SH Shape Tanimoto as the objective function, which was also used to rank the ligand database.

Superposition of atom-centered Gaussian functions⁵⁰ was then performed with ROCS to compute shape-based overlays of all conformers of every compound in the database onto the isothiourea query. Database molecules were then ranked

according to their “color optimization” mode (Combo Score) to maximize both the shape and chemical property overlays.

HPCC was used to calculate spherical harmonic polynomial functions derived from atomic coordinates for all conformers of every compound in the database to represent their molecular shape as described in the work of Karaboga et al.⁵² These harmonic based surfaces were discretized as triangle meshes. A chemotype-like property was then assigned to these triangles by using pharmacophoric features (hydrogen-bond donor, hydrogen-bond acceptor, hydrophobic, anion, cation, aromatic, and their combinations), in a way that the similarity calculation between molecules A and B was carried out by comparing the chemotypelike distribution associated to their respective shapes. HPCC was used in “shape plus chemistry” mode (Combo Score) to maximize both the shape and chemical property overlays.

Analyzing Virtual Screening and Pharmacophore Performance. All the screened compounds were ranked by their retrieved scores according to the pharmacophore, shape-matching, and docking approaches (LigandScout Pharmacophore Fit Score, Parafit Shape Tanimoto, ROCS Combo Score and HPCC Combo Score, and the aforementioned docking scoring functions, respectively), and the final ranking lists were represented in a ROC plot.⁵³ The area under the ROC plot (AUC)^{53,54} was quantified to evaluate the VS performance. As high AUC values do not guarantee that the top ranked poses will correspond to active compounds,⁵⁵ we calculated the partial area under the curve^{56,57} (pAUC) at 1%, 2%, and 5% of false positives. The corresponding values were normalized to partial area indexes.⁵⁸

$$\text{pAUC}_{x\%} = \frac{1}{2} \left[1 + \frac{\text{AUC}_{x\%} - \text{AUC}_{\text{random at } x\%}}{\text{AUC}_{\text{perfect at } x\%} - \text{AUC}_{\text{random at } x\%}} \right]$$

Furthermore, evaluation of early recovery of actives⁵⁹ was assessed by the enrichment factor (EF)⁶⁰ at 1%, 2%, and 5% of screened database and, additionally, by the percentage of actives retrieved at 1%, 2%, and 5% of decoys screened.

The comparison of results between the different anti-HIV data sets for analyzing the sensitivity and specificity of the pharmacophore model was performed using the pAUC and also a relative EF, as stated by Venkatraman et al.⁶¹ Because it is not rigorous to compare EF between different data sets, given that EF values are easily influenced by the number of actives in the data set, von Korff et al.⁶² advocate the use of a relative EF (EF_{rel}), where the achieved enrichment for a given cutoff (say $x\%$) is normalized by the maximum possible enrichment in that cutoff.

RESULTS AND DISCUSSION

Pharmacophore Performance. The performance of our five-featured CXCR4 pharmacophore model was validated by evaluating its ability to distinguish between the set of 228 known highly active CXCR4 antagonist inhibitors compiled from the literature and the 4696 compiled Maybridge decoy compounds. In an ideal pharmacophore virtual screening exercise, the ideal pharmacophore model should select all the highly active CXCR4 compounds and none of the decoy compounds. That is to say, it should be 100% sensitive (retrieving all highly active ligands) and 100% specific (retrieving zero false positives). Here we analyze the performance of the proposed pharmacophore model according to these parameters.

Our pharmacophore retrieved 298 hits (Figure 3). From those, 114 out of 4696 were decoys, and 184 out of 211 were

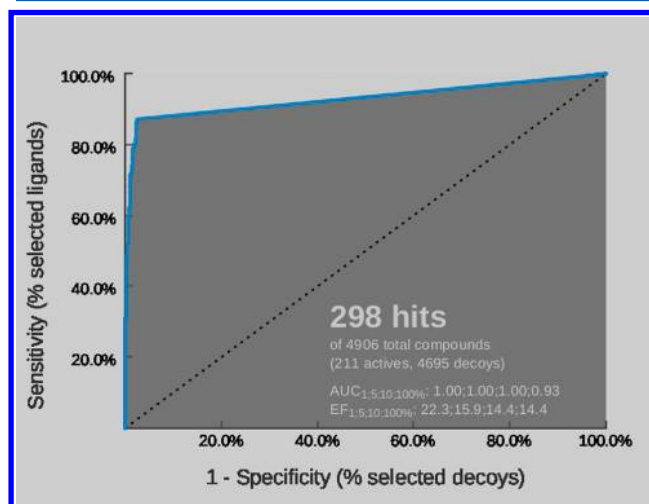


Figure 3. ROC plot validation of the pharmacophore model applied to CXCR4 antagonists. Values of area under the curve (AUC) and enrichment factor (EF) are displayed at 1, 5, 10, and 100% of screened database, respectively. These values highlight the high sensitivity and specificity of the designed pharmacophore model.

CXCR4 active compounds. Hence, the pharmacophore model retrieved the 2.4% of the total decoys and the 83.2% of the total active compounds considered. Therefore, it is very sensitive (retrieves nearly all active ligands) and very specific (retrieves very few false positives), approaching to an ideal pharmacophore. Figure 3 shows the ROC plot validation of the pharmacophore model applied to CXCR4 antagonists, where its high sensitivity and specificity can be appreciated by the early recovery values. At 1%, 2%, and 5% of the screened database, the pharmacophore model retrieves EF values of 22.3, 21.1, and 15.9, respectively, and pAUC values of 0.75, 0.82, and 0.88, respectively. The high specificity is reflected by a total AUC value of 93%. Hence, the designed five-featured pharmacophore model is able to discriminate very well between active and decoys. Nevertheless, it is worth mentioning that the pharmacophore model does not comprise any excluded volume features. Our analyses show that some ligand conformations matching the pharmacophore properly have an internal energy quite high which could lead to putative internal clashes, mainly for the case of big ligands for which a part of the ligand match well the five pharmacophoric features but the rest of the ligand could clash with the protein. Hence, adding excluded volumes to complete the protein–ligand context in the model might be a way to improve further the specificity of the present pharmacophore.

Analyzing the Sensitivity and Specificity of the Pharmacophore Model. The sensitivity and specificity of the pharmacophore model was analyzed by performing two types of experiments. The sensitivity was analyzed by considering the different anti-HIV compound databases as actives, and the same data set assembled from Maybridge as decoys. Results for this analysis are shown in Table 3a. The specificity was analyzed by considering our compiled database of CXCR4 antagonists as actives and all the other anti-HIV compound databases as decoys. Results for this analysis are shown in Table 3b.

Table 3a. Comparison Values for EF_{rel} and pAUC Obtained When the Pharmacophore Is Applied to Different HIV Inhibitors Being Considered As Actives and the Same Data Set Assembled from Maybridge as Decoys

		CXCR4 antagonists + Maybridge decoys	CCR5 antagonists + Maybridge decoys	HIV integrase inh. + Maybridge decoys	HIV protease inh. + Maybridge decoys	HIV reverse transcriptase inh. + Maybridge decoys
EF _{rel}	1%	95.9	60.0	58.9	87.1	35.2
	2%	90.8	56.0	50.0	83.7	30.8
	5%	68.6	- ^a	- ^a	72.2	- ^a
pAUC	1%	0.75	0.55	0.51	0.52	0.50
	2%	0.82	0.58	0.52	0.53	0.50
	5%	0.88	- ^a	- ^a	0.54	- ^a

^aAbsence of hits retrieved by the pharmacophore at the percentage referred. The pharmacophore ROC curve does not arrive until 100% of screened database. LigandScout, as it is the case for all pharmacophore software, does not retrieve values for compounds not matching the pharmacophore over a defined threshold.

Table 3b. Comparison Values for the EF_{rel} and pAUC Obtained When the Pharmacophore Is Applied to Different HIV Inhibitors Being Considered As Decoys and the CXCR4 Antagonists As Actives

		CXCR4 antagonists + CCR5 antagonists	CXCR4 antagonists + HIV integrase inh.	CXCR4 antagonists + HIV protease inh.	CXCR4 antagonists + HIV reverse transcriptase inh.
EF _{rel}	1%	100.0	100.0	96.0	100.0
	2%	100.0	100.0	92.2	100.0
	5%	- ^a	- ^a	68.0	- ^a
pAUC	1%	0.55	0.69	0.66	0.76
	2%	0.58	0.69	0.67	0.83
	5%	- ^a	- ^a	0.70	- ^a

^aAbsence of hits retrieved by the pharmacophore at the percentage referred. The pharmacophore ROC curve does not arrive until 100% of screened database. LigandScout, as it is the case for all pharmacophore software, does not retrieve values for compounds not matching the pharmacophore over a defined threshold.

For the first kind of analysis, we can see in Table 3a that the pharmacophore model is more sensitive to the CXCR4 antagonists. The pharmacophore is able to identify the active CXCR4 ligands with the highest EF_{rel} and pAUC values (EF_{rel} of 95.9, 90.8, and 68.6, and pAUC of 0.75, 0.82, and 0.88 at 1%, 2%, and 5% of database screened, respectively), whereas all the other HIV inhibitors are retrieved with lower early recovery values (EF_{rel} of 60.0, 58.9, 87.1, and 35.2 and pAUC of 0.55, 0.51, 0.52, and 0.50 at 1% of screened database for CCR5 antagonists, HIV integrase inhibitors, HIV protease inhibitors, and HIV reverse transcriptase inhibitors, respectively).

For the second kind of analysis, we can see in Table 3b that the pharmacophore is able to identify only CXCR4 antagonists with very high EF_{rel} and pAUC values and nothing else when using as decoys all the other anti-HIV compound databases. The pharmacophore is able to identify the active CXCR4 ligands in front of the other anti-HIV compounds with very high EF_{rel} and pAUC values (EF_{rel} of 100.0, 100.0, 96.0, and 100.0 at 1% of screened database and EF_{rel} of 100.0, 100.0, 92.2, and 100.0 at 2% of screened database, for CCR5 antagonists, HIV integrase inhibitors, HIV protease inhibitors, and HIV reverse transcriptase inhibitors, respectively).

Overall, in both experiments the highest EF_{rel} and pAUC values are retrieved when screening the CXCR4 antagonists, only slightly approached by the values when screening HIV protease inhibitors. The pharmacophore model is able to discriminate well similar compounds, regarding the average of several 1D physicochemical properties, such as CXCR4 and

CCR5 antagonists (see Table 2). Hence, the superior results for CXCR4 antagonists prove our pharmacophore model to be a filter of great sensitivity and specificity for identifying CXCR4 antagonists.

Comparison with Docking-Based and Shape-Matching Virtual Screening Performance. The CXCR4 compiled data set of actives and decoys was docked against 3OE6 CXCR4 PDB using the aforementioned docking functions. Moreover the isothiurea ligand in 3OE6 crystal structure was used to shape query our database using PARAFIT, ROCS, and HPCC. Figure 4 shows the ROC plot comparing the performance of all of these approaches together with the pharmacophore values obtained. Tables 4a and 4b show the early recovery performance for all approaches in terms of EF at 1%, 2%, and 5% of the screened database, and the percentage of true positives selected at 1%, 2%, and 5% of false positives, respectively. Tables 5a and 5b show the AUC and the pAUC at 1%, 2%, and 5% of decoys retrieved for all the approaches.

As it can be observed in Figure 4, the pharmacophore ROC curve (in magenta) does not arrive until 100% of screened database given that LigandScout, as it is the case for all pharmacophore software, does not retrieve values for compounds not matching the pharmacophore over a defined threshold. Hence, in order to better compare the pharmacophore model with the rest of scoring functions, we deeply analyze early recovery performance until 5% of total database screened.

In general, ROC plot analysis shows that the pharmacophore model obtains the highest AUC, and early recovery performance, expressed as EF, percentage of actives recovered at a percentage of decoys screened, and pAUC (values shown in bold in Tables 5a, 4a, 4b, and 5b, respectively), followed by LigandFit docking scoring functions, the rest of docking scoring functions, and shape-matching scoring functions. It is worth noting that the AUC value for the pharmacophore model is an estimated value (0.93), since its curve does not arrive until 100% of screened database, as mentioned before.

Regarding the EF (Table 4a), it can be observed that apart from the pharmacophore, the LigandFit docking scoring functions DockScore, LigScore2, PLP1, and PLP2 are also selectively choosing ligands from decoys from the compiled data set. At 1% of total database screened, EF values achieved for these LigandFit scoring functions were 22.6, 17.9, 15.5, and 15.1, respectively, and 20.2, 17.2, 14.1, and 13.4, at 2% of screened database, compared with the EF values for the pharmacophore model of 22.3 and 21.1 at 1% and 2% of database screened. Other scoring functions like PMF, LigScore1, Jain, Glide, AutoDock Vina, and Surflex returned lower

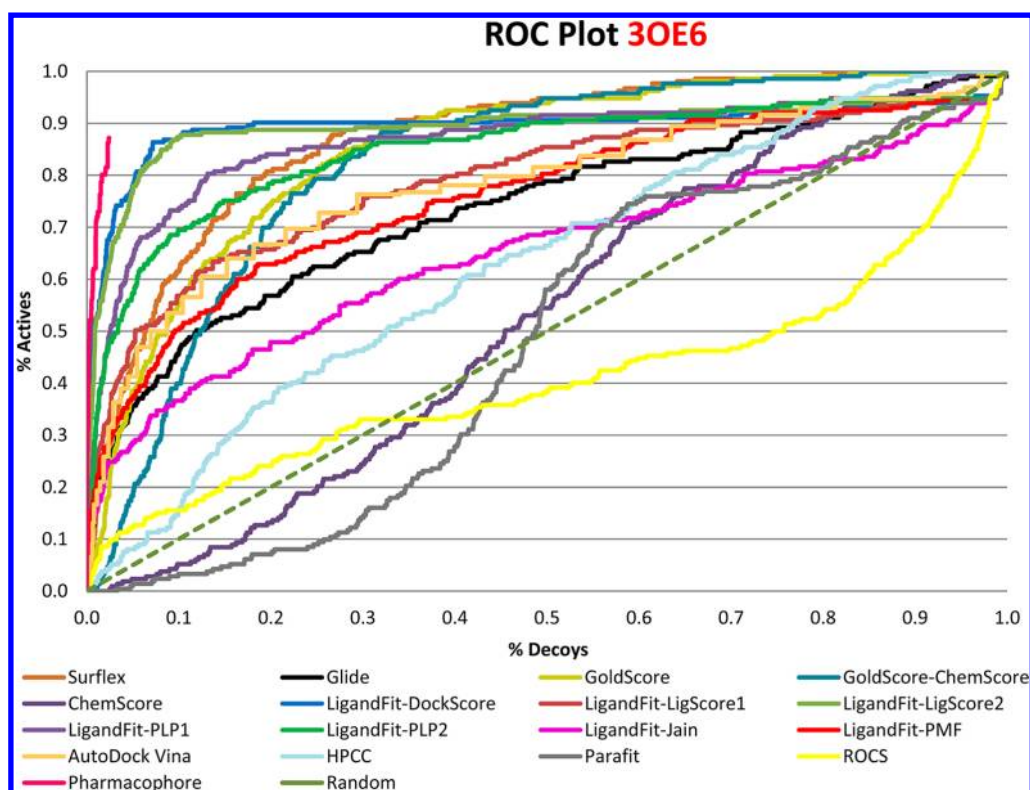


Figure 4. ROC plot validation of the pharmacophore model compared to various docking and shape-matching VS methods applied to CXCR4 antagonists. The dotted green line represents the expected enrichment if actives were selected at random. The pharmacophore ROC plot (curve in magenta) returned only 2.4% of the total decoys set.

but also quite selective EF values at the 1% cutoff (13.6, 13.2, 11.3, 10.4, 9.3, and 8.5, respectively) and also at the 2% cutoff (10.8, 11.8, 8.7, 9.6, 9.7%, and 8.9, respectively). The same trend can be observed for the percentage of actives at a percentage of decoys screened (Table 4b). The pharmacophore model obtains values of 71.1% and 80.6% of actives recovered at 1% and 2% of decoys screened, followed by the LigandFit docking scoring functions DockScore, LigScore2, PLP2, and PLP1, with values of 52.6% and 62.0%, 52.1% and 59.6%, 33.3% and 45.0%, and 31.9% and 45.5% of actives recovered at 1% and 2% of decoys screened, respectively.

Comparing the AUC (Table 5a) and pAUC values at 1%, 2%, and 5% of screened decoys (Table 5b), for all the approaches, it can be observed that the pharmacophore model obtains the highest AUC (0.93) and pAUC values (0.75 at 1%, 0.82 at 2%, and 0.88 at 5% of decoys screened, respectively) with high difference among all scoring functions, only comparable to LigandFit DockScore, with an AUC value of 0.88 and pAUC values of 0.72 at 1%, 0.75 at 2%, and 0.81 at 5% of decoys screened, respectively. LigScore2, PLP1, PLP2, PMF, LigScore1, Jain, Glide, AutoDock Vina, and Surflex show moderately good selectivity (Table 5b). From these scoring functions, LigScore2 achieved the highest pAUC values (0.65 at 1%, 0.71 at 2%, and 0.79 at 5% of screened decoys), followed by PLP1 (values of 0.59%, 0.64%, and 0.74%), PLP2 (values of 0.58%, 0.64%, and 0.70%), PMF (values of 0.57%, 0.59%, and 0.63%), LigScore1 (values of 0.56%, 0.60%, and 0.65%), Jain (values of 0.55%, 0.57%, and 0.60%), Glide (values of 0.54%, 0.57%, and 0.61%), AutoDock Vina (values of 0.54%, 0.57%, and 0.63%), and Surflex TotalScore (values of 0.53%, 0.56%, and 0.62%).

As mentioned above, the ROC plot analysis shows that shape-matching scoring functions perform the poorest in our study. Several papers have dealt with the question if docking performs better or not than shape-matching.^{63,64} A known fact is that the performance of a method is highly dependent on the nature of the target of interest and the data set used to screen, which must be assembled properly as mentioned before. In our case study, several papers have shown that the CXCR4 target has multiple interaction subsites within its pocket,^{65,12,14–17} and this can be easily understood when noticing the wide range of different active scaffold families existing for CXCR4 (Table 1). Hence, it is easy to think that a single shape query would not retrieve good results when performing shape-matching VS on a diverse set of CXCR4 inhibitors, which can bind in different ways (binding subsites) and hence have different active shapes. Therefore, we must take this into account when analyzing our results, which show shape-matching scoring functions performing the poorest (AUC values for shape-matching scoring functions approximately random), since we used the isothiouraea ligand in the 3OE6 structure as a unique representative shape-matching query. Recent studies have shown that using a consensus query, which can average the essential features of several known high-affinity scaffold families to take into account multiple binding subsite shapes, can highly improve shape-matching VS performance.^{13,17,66–68}

Scaffold Retrieval Analysis. The CXCR4 database that we have built contains 228 actives from twenty chemotype families (see Table 1). To evaluate the ability of the different approaches used to retrieve these diverse scaffold families, we analyzed the percentage of actives returned by the scoring functions attending to its intrinsic chemotype classification at 1%, 2%, and 5% of screened database. Figure 5 shows the

Table 4a. Early Recovery Performance Expressed by the Enrichment Factor at 1%, 2%, and 5% of the Screened Database^a

	% screened database	enrichment factors
GoldScore	1%	5.6
	2%	5.2
	5%	6.3
ChemScore	1%	0.0
	2%	0.0
	5%	0.5
GoldScore and ChemScore	1%	0.5
	2%	1.9
	5%	3.3
Glide	1%	10.4
	2%	9.6
	5%	6.3
Surflex	1%	8.5
	2%	8.9
	5%	7.3
TotalScore	1%	22.6
	2%	20.2
	5%	13.4
DockScore	1%	13.2
	2%	11.8
	5%	8.1
LigandFit	1%	17.9
	2%	17.2
	5%	12.6
LigScore1	1%	15.5
	2%	14.1
	5%	10.2
LigandFit PLP1	1%	15.1
	2%	13.4
	5%	9.9
LigandFit PLP2	1%	11.3
	2%	8.7
	5%	5.4
LigandFit Jain	1%	13.6
	2%	10.8
	5%	6.7
LigandFit PMF	1%	9.3
	2%	9.7
	5%	7.3
AutoDock Vina	1%	0.0
	2%	0.0
	5%	0.3
Parafit	1%	4.3
	2%	4.3
	5%	2.4
ROCS	1%	1.4
	2%	1.9
	5%	1.6
HPCC	1%	22.3
	2%	21.1
	5%	15.9

^aValues in bold represent the best enrichment factor values obtained when comparing all scoring functions.

scaffold retrieval rates for the seven most representative chemotype families and the rest of chemotypes are grouped into a generic family named “others”. Therefore, the AMD family includes AMD derivatives, AMD070 derivatives, and

Table 4b. Early Recovery Performance Expressed As Percentage of Actives Recovered When 1%, 2%, and 5% of Decoys Were Respectively Selected^a

	% false positive screened	% true positives
GoldScore	1%	7.0%
	2%	15.0%
	5%	38.5%
ChemScore	1%	0.0%
	2%	0.0%
	5%	2.3%
GoldScore and ChemScore	1%	0.5%
	2%	3.8%
	5%	18.3%
Glide	1%	16.4%
	2%	23.9%
	5%	35.7%
Surflex	1%	14.1%
	2%	23.5%
	5%	43.7%
TotalScore	1%	52.6%
	2%	62.0%
	5%	77.9%
DockScore	1%	23.0%
	2%	32.9%
	5%	45.2%
LigandFit	1%	52.1%
	2%	59.6%
	5%	77.0%
LigScore1	1%	31.9%
	2%	45.5%
	5%	64.3%
LigandFit PLP1	1%	33.3%
	2%	45.0%
	5%	57.3%
LigandFit PLP2	1%	15.5%
	2%	22.1%
	5%	28.6%
LigandFit Jain	1%	19.7%
	2%	27.7%
	5%	37.1%
LigandFit PMF	1%	16.7%
	2%	25.9%
	5%	41.2%
AutoDock Vina	1%	0.0%
	2%	0.0%
	5%	1.3%
Parafit	1%	5.7%
	2%	8.5%
	5%	12.3%
ROCS	1%	2.0%
	2%	3.8%
	5%	8.0%
HPCC	1%	71.1%
	2%	80.6%
	5%	<i>b</i>

^aValues in bold represent the best percentages obtained when comparing all scoring functions. ^bOnly 2.4% of total decoys are retrieved by the pharmacophore.

monocyclam derivatives (see Table 1). Macrocyclic family includes also bis-azamacrocyclics, and tetrahydroquinoline

Table 5a. AUC Values Calculated from the Corresponding ROC Curves for All the Scoring Functions^a

	AUC
GoldScore	0.85
ChemScore	0.53
GoldScore and ChemScore	0.81
Glide	0.74
Surflex TotalScore	0.87
LigandFit DockScore	0.88
LigandFit LigScore1	0.78
LigandFit LigScore2	0.88
LigandFit PLP1	0.86
LigandFit PLP2	0.84
LigandFit Jain	0.64
LigandFit PMF	0.75
AutoDock Vina	0.75
Parafit	0.44
ROCS	0.41
HPCC	0.63
Pharmacophore	0.93^b

^aValues in bold represent the best AUC values obtained when comparing all scoring functions. ^bAUC value is estimated based on a random classification of the actives and decoys not retrieved by the pharmacophore. The pharmacophore ROC curve does not arrive until 100% of screened database given that only values for 2.4% of the decoys set (6.1% of the total database) are retrieved.

family contains tetrahydroquinoline imidazopyridine and tetrahydroquinoline benzimidazole compounds.

Figure 5 shows that at 5% of screened database all the CXCR4 scaffold families were recovered only by the pharmacophore and DockScore function, with an average percentage recovery of 80% and 52%, respectively (average understood as the total percentage of compounds recovered at 5% of screened database divided by 8 scaffold families). At this percentage the pharmacophore retrieves 100% of the macrocycle derivatives and 100% of the tetrahydroquinoline compounds present in the database, while the DockScore scoring function retrieves only 100% of the macrocycle derivatives. Comparing the compounds retrieved for the same scoring function now at 2% of screened database, the unique families with no compounds retrieved were phenanthroline, tetrahydroquinoline, and the generic family “others” for the DockScore function, while the pharmacophore retrieved 30.8% of phenanthroline compounds, 30.0% of tetrahydroquinolines, and 28.6% of “others”. Otherwise the pharmacophore model did not retrieve at 2% any isothiourea compound neither KRH derivatives, but DockScore retrieved 33.3% of the isothiourea chemofamily and 21.1% of the KRH structures. It is worth noting that in Figure 5 the isothiourea family is the shortest chemotype group containing only three structures.

Consequently we can affirm that the pharmacophore model selects actives independently from their chemotype nature in a very effective way, as well as the DockScore function, but also we observe that both scoring functions show indeed, a moderately different and complementary mode of retrieving the CXCR4 chemotype families at early percentages of database screened. This observation evidence the differences in the nature of both methods ranking the same binders.

Table 5b. Partial Area Indexes Calculated from the AUC Values at 1%, 2%, and 5% of Decoys Retrieved for All the Scoring Functions^a

	% false positive screened	pAUC
	1%	0.51
GoldScore	2%	0.53
	5%	0.59
ChemScore	1%	0.50
	2%	0.50
	5%	0.49
	1%	0.50
GoldScore and ChemScore	2%	0.50
	5%	0.52
Glide	1%	0.54
	2%	0.57
	5%	0.61
	1%	0.53
Surflex	2%	0.56
TotalScore	5%	0.62
	1%	0.72
LigandFit	2%	0.75
DockScore	5%	0.81
	1%	0.56
LigandFit	2%	0.60
LigScore1	5%	0.65
	1%	0.65
LigandFit	2%	0.71
LigScore2	5%	0.79
	1%	0.59
LigandFit PLP1	2%	0.64
	5%	0.74
	1%	0.58
LigandFit PLP2	2%	0.64
	5%	0.70
	1%	0.55
LigandFit Jain	2%	0.57
	5%	0.60
	1%	0.57
LigandFit PMF	2%	0.59
	5%	0.63
	1%	0.54
AutoDock Vina	2%	0.57
	5%	0.63
	1%	0.50
Parafit	2%	0.50
	5%	0.49
	1%	0.51
ROCS	2%	0.52
	5%	0.53
	1%	0.50
HPCC	2%	0.51
	5%	0.51
	1%	0.75
Pharmacophore	2%	0.82
	5%	0.88

^aValues in bold represent the best partial area indexes obtained when comparing all the scoring functions.

CONCLUSION

A large database of highly active CXCR4 antagonist inhibitors and similar presumed inactive compounds was compiled from the literature to perform retrospective virtual screening. The set

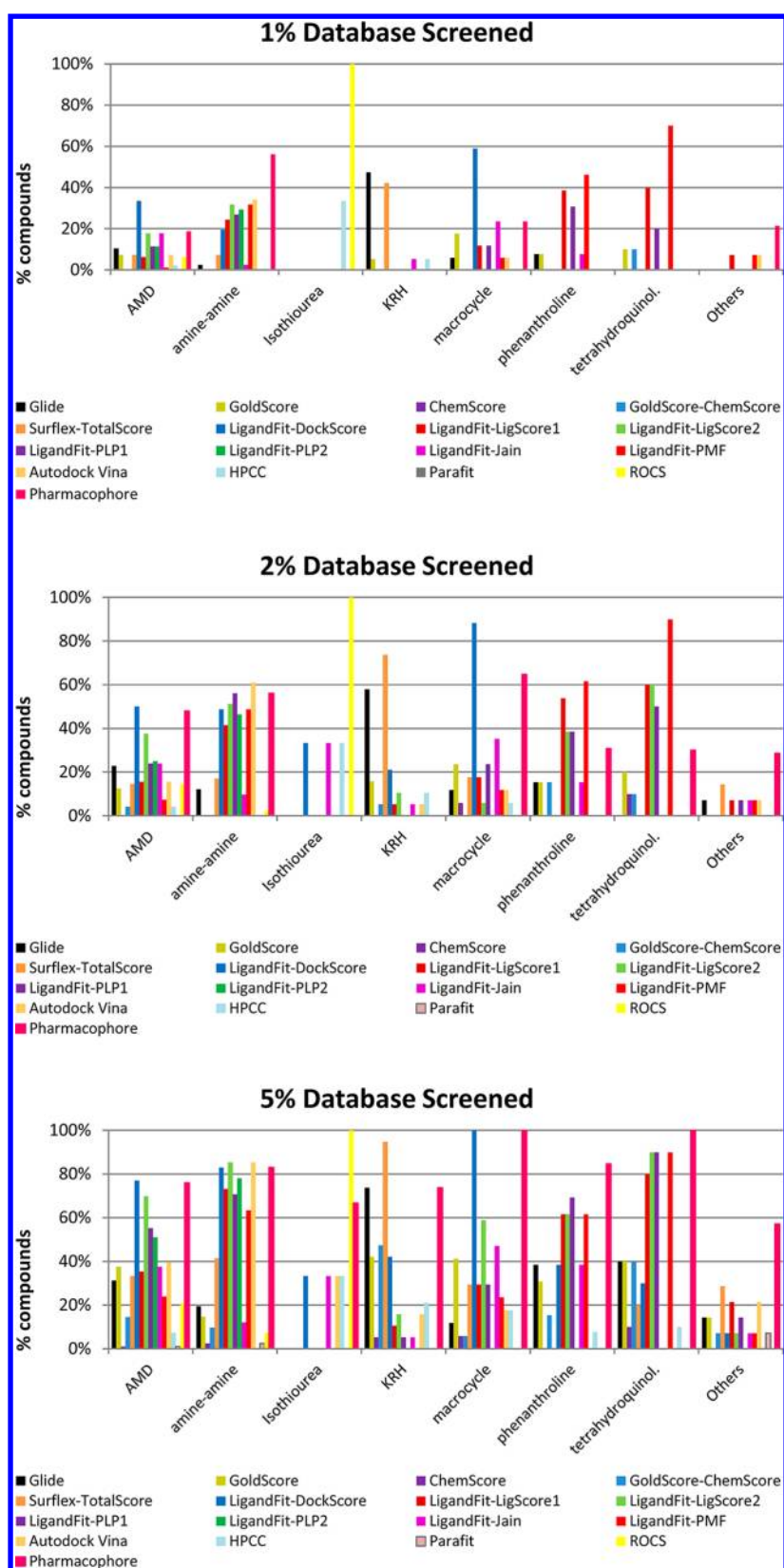


Figure 5. Scaffold diversity retrieval analysis for CXCR4 inhibitors, showing the percentage of actives by chemotype family retrieved using the pharmacophore, docking, and shape-matching scoring functions (colored bars), at 1%, 2%, and 5% of the screened database, respectively.

of active compounds compiled shows a wide scaffold diversity comprising more than 20 different chemotype families. Similarly, the set of decoys selected from the commercial vendor Maybridge possesses the drug-like features required to

be similar to the active set to avoid bias in the screening results. This database was used to validate a new CXCR4 pharmacophore model built by using protein–ligand interaction information. So far, this is the first CXCR4

pharmacophore published in this way. Furthermore, the compiled database was used to perform a benchmark comparison between the pharmacophore model and several docking-based and ligand shape-based virtual screening approaches. The area under de curve of the ROC plot, the early recovery, and the diversity results obtained show that the present pharmacophore model is highly superior to the docking-based and ligand shape-based approaches used in this study, especially the ligand shape-based approaches which were applied using only one shape-query to filter the compiled database. Consequently, the performance of the ligand shape-based approaches appeared to be the poorest. Analysis of these results, and according to other experiments previously carried on CXCR4 by other teams, shows that CXCR4 is a protein target with multiple interaction subsites within the binding pocket. As a matter of fact, the known antagonists distribute over the pocket in at least three different binding modes and that implies that any ligand shape-based screening strategy should be carried out using a multishape query (which takes into account the shape diversity of the several existent CXCR4 scaffold families), in order to be at the performance level of the docking-based approaches. The docking-based approaches and the pharmacophore model already cover in some way the CXCR4 multisite binding information. The docking-based protocols energetically explore the whole pocket region when computing the docking energy of the database compounds. The pharmacophore model takes into account the protein–ligand interaction information to define the essential chemical features needed for the active CXCR4 ligands to bind. These features have wide tolerance radii, therefore the pharmacophore model is also in certain extent representative of the binding of the different scaffold families.

Overall, the total area under de curve of the ROC plot and the early recovery results of the present pharmacophore model show that it is a highly specific and sensitive screening filter, which makes it very appropriate for identifying CXCR4 antagonists. Moreover, the scaffold retrieval analysis shows that the pharmacophore model is able to retrieve a diverse scaffold pool. Hence, using this pharmacophore model to prospectively screen a database of compounds might facilitate the identification of novel lead structures. Therefore, we state that the use of protein–ligand interaction information to define pharmacophoric features and a robust and divers compiled database of very active compounds are essential to built a good-performance model, especially when dealing with targets with multiple interaction sites, as in the case of CXCR4.

AUTHOR INFORMATION

Corresponding Author

*Tel: +33-354 958 604. Fax: +33-383 593 046. E-mail: souchet@harmonicpharma.com (M.S.), pereznueno@harmonicpharma.com (V.I.P.-N.).

Author Contributions

§A.S.K. and J.M.P.: These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank to the Servei de Disseny de Farmacs (SDF) in the Centre de Supercomputació de Catalunya (CESCA) for providing access to some of the computational software used

through. The authors thank OpenEye Scientific Software Inc. for providing a license for ROCS. The authors are also grateful to ChemAxon for the license for the Marvin and Jchem toolkits.

REFERENCES

- (1) Tilton, J. C.; Doms, R. W. Entry inhibitors in the treatment of HIV-1 infection. *Antiviral Res.* **2010**, *1*, 91–100.
- (2) Trushin, S. A.; Bren, G. D.; Badley, A. D. CXCR4 tropic HIV-1 gp120 inhibition of SDF-1 α -induced chemotaxis requires Lck and is associated with cofilin phosphorylation. *Open Virol. J.* **2010**, 157.
- (3) Harrigan, P. R.; Swenson, L.; McGovern, R. A.; Pollock, G. J. The Determinants and Consequences of HIV Coreceptor Switching. *HIV Med. Update* **2009**, *4*, 1–4.
- (4) Haga, T.; Takeda, S. *G protein-coupled receptors: structure, function, and ligand screening*; CRC Press: Boca Raton, 2005; Vol. 6.
- (5) Choi, W. T.; An, J. Biology and clinical relevance of chemokines and chemokine receptors CXCR4 and CCR5 in human diseases. *Exp. Biol. Med.* **2011**, *6*, 637–647.
- (6) Peeters, M.; van Westen, G.; Li, Q.; IJzerman, A. Importance of the extracellular loops in G protein-coupled receptors for ligand recognition and receptor activation. *Trends Pharmacol. Sci.* **2011**, *1*, 35–42.
- (7) Wu, B.; Chien, E. Y. T.; Mol, C. D.; Fenalti, G.; Liu, W.; Katritch, V.; Abagyan, R.; Brooun, A.; Wells, P.; Bi, F. C. Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science* **2010**, 6007, 1066–1071.
- (8) Thoma, G.; Streiff, M. B.; Kovarik, J.; Glickman, F.; Wagner, T.; Beerli, C.; Zerwes, H. G. Orally bioavailable isothioureas block function of the chemokine receptor CXCR4 in vitro and in vivo. *J. Med. Chem.* **2008**, *24*, 7915–7920.
- (9) Moncunill, G.; Armand-Ugon, M.; Clotet-Codina, I.; Pauls, E.; Ballana, E.; Llano, A.; Romagnoli, B.; Vrijbloed, J. W.; Gombert, F. O.; Clotet, B. Anti-HIV activity and resistance profile of the CXCR4 chemokine receptor 4 antagonist POL3026. *Mol. Pharmacol.* **2008**, *4*, 1264.
- (10) Wang, J.; Norcross, M. Dimerization of chemokine receptors in living cells: key to receptor function and novel targets for therapy. *Drug Discov Today* **2008**, 13–14, 625–632.
- (11) Sohy, D.; Parmentier, M.; Springael, J. Y. Allosteric transinhibition by specific antagonists in CCR2/CXCR4 heterodimers. *J. Biol. Chem.* **2007**, *41*, 30062.
- (12) Pettersson, S.; Pérez-Nuño, V. I.; Mena, M. P.; Clotet, B.; Esté, J. A.; Borrell, J. I.; Teixidó, J. Novel monocyclam derivatives as HIV entry inhibitors: design, synthesis, anti-HIV evaluation, and their interaction with the CXCR4 co-receptor. *ChemMedChem* **2010**, *8*, 1272–1281.
- (13) Pérez-Nuño, V. I.; Ritchie, D. W.; Borrell, J. I.; Teixido, J. Clustering and classifying diverse HIV entry inhibitors using a novel consensus shape-based virtual screening approach: further evidence for multiple binding sites within the CCR5 extracellular pocket. *J. Chem. Inf. Model.* **2008**, *11*, 2146–2165.
- (14) Kawatkar, S. P.; Yan, M.; Gevariya, H.; Lim, M. Y.; Eisold, S.; Zhu, X.; Huang, Z.; An, J. Computational analysis of the structural mechanism of inhibition of chemokine receptor CXCR4 by small molecule antagonists. *Exp. Biol. Med.* **2011**, *7*, 844–850.
- (15) Neves, M. A. C.; Simões, S.; Sá e Melo, M. L. Ligand-guided optimization of CXCR4 homology models for virtual screening using a multiple chemotype approach. *J. Comput. Aided Mol. Des.* **2010**, *12*, 1023–1033.
- (16) Lam, A. R.; Bhattacharya, S.; Patel, K.; Hall, S. E.; Mao, A.; Vaidehi, N. Importance of receptor flexibility in binding of cyclam compounds to the chemokine receptor CXCR4. *J. Chem. Inf. Model.* **2010**, *1*, 137–147.
- (17) Planesas, J. M.; Pérez-Nuño, V. I.; Borrell, J. I.; Teixidó, J. Impact of the CXCR4 structure on docking-based virtual screening of HIV entry inhibitors. *J. Mol. Graph. Model.* **2012**, *38*, 123–136.
- (18) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using

protein-ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *3*, 793–806.

(19) Maybridge Bringing life to drug discovery, Maybridge Databases; Fisher Scientific International: England, 2005.

(20) Krüger, D. M.; Evers, A. Comparison of Structure-and Ligand-Based Virtual Screening Protocols Considering Hit List Complementarity and Enrichment Factors. *ChemMedChem* **2009**, *1*, 148–158.

(21) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* **2006**, *suppl 1*, D198.

(22) Pérez-Nueno, V. L.; Ritchie, D. W.; Rabal, O.; Pascual, R.; Borrell, J. I.; Teixido, J. Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape matching and ligand–receptor docking. *J. Chem. Inf. Model* **2008**, *3*, 509–533.

(23) Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *1*, 160–169.

(24) OMEGA, v. 2.1.0.; Open Eye Scientific Software: Santa Fe, NM, 2006.

(25) Berman, H.; Battistuz, T.; Bhat, T.; Bluhm, W.; Bourne, P.; Burkhardt, K.; Feng, Z.; Gilliland, G.; Iype, L.; Jain, S. The protein data bank. *Acta Crystallogr. D Biol. Crystallogr* **2002**, *6*, 899–907.

(26) Wang, R.; Lu, Y.; Fang, X.; Wang, S. An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein–ligand complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *6*, 2114–2125.

(27) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein–ligand docking using GOLD. *Proteins: Structure, Function, Bioinf.* **2003**, *4*, 609–623.

(28) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Daniel, T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *7*, 1739–1749.

(29) Pham, T. A.; Jain, A. N. Customizing scoring functions for docking. *J. Comput. Aided Mol. Des.* **2008**, *5*, 269–286.

(30) SYBYL -X 1.1.; Tripos International, St. Louis, MO, 2010.

(31) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *2*, 455–461.

(32) Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M. LigScore: a novel scoring function for predicting binding affinities. *J. Mol. Graph. Model.* **2005**, *5*, 395–407.

(33) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. R.; Fogel, L. J.; Freer, S. T. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *5*, 317–324.

(34) Bustanji, Y.; Taha, M. O.; Yousef, A. M.; G. Al-Bakri, A. Berberine potently inhibits protein tyrosine phosphatase 1B: Investigation by docking simulation and experimental validation. *J. Enzyme Inhib. Med. Ch.* **2006**, *2*, 163–171.

(35) Feher, M. Consensus scoring for protein–ligand interactions. *Drug Discov. Today* **2006**, *9–10*, 421–428.

(36) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Chem. Inf. Model.* **2001**, *28*, 6474–6487.

(37) Clark, M.; Cramer, R. D., III; Van Opdenbosch, N. Validation of the general purpose Tripos 5.2 force field. *J. Comput. Chem.* **1989**, *8*, 982–1012.

(38) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *2*, 187–217.

(39) Hatse, S.; Princen, K.; Vermeire, K.; Gerlach, L. O.; Rosenkilde, M. M.; Schwartz, T. W.; Bridger, G.; De Clercq, E.; Schols, D. Mutations at the CXCR4 interaction sites for AMD3100 influence

anti-CXCR4 antibody binding and HIV-1 entry. *FEBS Lett.* **2003**, *2–3*, 300–306.

(40) Gerlach, L. O.; Skerlj, R. T.; Bridger, G. J.; Schwartz, T. W. Molecular interactions of cyclam and bicyclam non-peptide antagonists with the CXCR4 chemokine receptor. *J. Biol. Chem.* **2001**, *17*, 14153–14160.

(41) Rosenkilde, M. M.; Gerlach, L. O.; Jakobsen, J. S.; Skerlj, R. T.; Bridger, G. J.; Schwartz, T. W. Molecular mechanism of AMD3100 antagonism in the CXCR4 receptor. *J. Biol. Chem.* **2004**, *4*, 3033.

(42) Rosenkilde, M. M.; Gerlach, L. O.; Hatse, S.; Skerlj, R. T.; Schols, D.; Bridger, G. J.; Schwartz, T. W. Molecular mechanism of action of monocyclam versus bicyclam non-peptide antagonists in the CXCR4 chemokine receptor. *J. Biol. Chem.* **2007**, *37*, 27354–27365.

(43) Princen, K.; Hatse, S.; Vermeire, K.; Aquaro, S.; De Clercq, E.; Gerlach, L. O.; Rosenkilde, M.; Schwartz, T. W.; Skerlj, R.; Bridger, G. Inhibition of human immunodeficiency virus replication by a dual CCR5/CXCR4 antagonist. *J. Virol.* **2004**, *23*, 12996–13006.

(44) Hatse, S.; Princen, K.; Clercq, E. D.; Rosenkilde, M. M.; Schwartz, T. W.; Hernandez-Abad, P. E.; Skerlj, R. T.; Bridger, G. J.; Schols, D. AMD3465, a monomacrocyclic CXCR4 antagonist and potent HIV entry inhibitor. *Biochem. Pharmacol.* **2005**, *5*, 752–761.

(45) Tamamura, H.; Xu, Y.; Hattori, T.; Zhang, X.; Arakaki, R.; Kanbara, K.; Omagari, A.; Otaka, A.; Ibuka, T.; Yamamoto, N. A low-molecular-weight inhibitor against the chemokine receptor CXCR4: a strong anti-HIV peptide T140. *Biochem. Biophys. Res. Commun.* **1998**, *3*, 877–882.

(46) Trent, J. O.; Wang, Z.; Murray, J. L.; Shao, W.; Tamamura, H.; Fujii, N.; Peiper, S. C. Lipid bilayer simulations of CXCR4 with inverse agonists and weak partial agonists. *J. Biol. Chem.* **2003**, *47*, 47136–47144.

(47) Våbenø, J.; Nikiforovich, G. V.; Marshall, G. R. Insight into the binding mode for cyclopentapeptide antagonists of the CXCR4 receptor. *Chem. Biol. Drug. Des.* **2006**, *5*, 346–354.

(48) Lin, J. H.; Clark, T. An analytical, variable resolution, complete description of static molecules and their intermolecular binding properties. *J. Chem. Inf. Model.* **2005**, *4*, 1010–1016.

(49) Ritchie, D. W.; Kemp, G. J. L. Protein docking using spherical polar Fourier correlations. *Proteins* **2000**, *2*, 178–194.

(50) Grant, J. A.; Gallardo, M.; Pickup, B. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1998**, *14*, 1653–1666.

(51) Cai, W.; Xu, J.; Shao, X.; Leroux, V.; Beautrait, A.; Maigret, B. SHEF: a vHTS geometrical filter using coefficients of spherical harmonic molecular surfaces. *J. Mol. Model.* **2008**, *5*, 393–401.

(52) Karaboga, A. S.; Petronin, F.; Marchetti, G.; Souchet, M.; Maigret, B. Benchmarking of HPCC: a novel 3D molecular representation combining shape and pharmacophoric descriptors for efficient molecular similarity assessments. *J. Mol. Graph. Model.* **2012**, *41*, 20–30.

(53) Fawcett, T. An introduction to ROC analysis. *Pattern Recog. Lett.* **2006**, *8*, 861–874.

(54) Jiang, Y.; Metz, C. E.; Nishikawa, R. M. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology* **1996**, *3*, 745.

(55) Sonego, P.; Kocsor, A.; Pongor, S. ROC analysis: applications to the classification of biological sequences and 3D structures. *Brief. Bioinform.* **2008**, *3*, 198–209.

(56) Obuchowski, N. A. ROC analysis. *Am. J. Roentgenol.* **2005**, *2*, 364–372.

(57) Lasko, T. A.; Bhagwat, J. G.; Zou, K. H.; Ohno-Machado, L. The use of receiver operating characteristic curves in biomedical informatics. *J. Biomed. Inform.* **2005**, *5*, 404–415.

(58) McClish, D. K. Analyzing a portion of the ROC curve. *Med. Decision Making* **1989**, *3*, 190–195.

(59) Jain, A. N. Bias, reporting, and sharing: computational evaluations of docking methods. *J. Comput. Aided Mol. Des.* **2008**, *3–4*, 201–212.

- (60) Mackey, M. D.; Melville, J. L. Better than Random? The Chemotype Enrichment Problem. *J. Chem. Inf. Model.* **2009**, *5*, 1154–1162.
- (61) Venkatraman, V.; Pérez-Nuño, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods. *J. Chem. Inf. Model.* **2010**, *12*, 2079–2093.
- (62) von Korff, M.; Freyss, J.; Sander, T. Comparison of ligand- and structure-based virtual screening on the DUD data set. *J. Chem. Inf. Model.* **2009**, *2*, 209–231.
- (63) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *4*, 1504–1519.
- (64) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *1*, 74–82.
- (65) Wong, R. S. Y.; Bodart, V.; Metz, M.; Labrecque, J.; Bridger, G.; Fricker, S. P. Comparison of the potential multiple binding modes of bicyclam, monocyclam, and noncyclam small-molecule CXC chemokine receptor 4 inhibitors. *Mol. Pharmacol.* **2008**, *6*, 1485–1495.
- (66) Pérez-Nuño, V. I.; Ritchie, D. W. Using consensus-shape clustering to identify promiscuous ligands and protein targets and to choose the right query for shape-based virtual screening. *J. Chem. Inf. Model.* **2011**, *6*, 1233–1248.
- (67) Pérez-Nuño, V. I.; Ritchie, D. W. Applying in silico tools to the discovery of novel CXCR4 inhibitors. *Drug Dev. Res.* **2011**, *1*, 95–111.
- (68) Pérez-Nuño, V. I.; Ritchie, D. W. Identifying and characterizing promiscuous targets: Implications for virtual screening. *Expert Opin. Drug Discov.* **2012**, *1*, 1–17.
- (69) Bridger, G. J.; Skerlj, R. T.; Padmanabhan, S.; Martellucci, S. A.; Henson, G. W.; Struyf, S.; Witvrouw, M.; Schols, D.; De Clercq, E. Synthesis and structure-activity relationships of phenylenebis (methylene)-linked bis-azamacrocycles that inhibit HIV-1 and HIV-2 replication by antagonism of the chemokine receptor CXCR4. *J. Med. Chem.* **1999**, *19*, 3971–3981.
- (70) De Clercq, E. Inhibition of HIV infection by bicyclams, highly potent and specific CXCR4 antagonists. *Mol. Pharmacol.* **2000**, *5*, 833–839.
- (71) Esté, J. A.; De Vreese, K.; Witvrouw, M.; Schmit, J. C.; Vandamme, A. M.; Anné, J.; Desmyter, J.; Henson, G. W.; Bridger, G.; De Clercq, E. Antiviral activity of the bicyclam derivative JM3100 against drug-resistant strains of human immunodeficiency virus type 1. *Antiviral Res.* **1996**, *2*, 297–307.
- (72) Esté, J. A.; Cabrera, C.; De Clercq, E.; Struyf, S.; Van Damme, J.; Bridger, G.; Skerlj, R. T.; Abrams, M. J.; Henson, G.; Gutierrez, A. Activity of different bicyclam derivatives against human immunodeficiency virus depends on their interaction with the CXCR4 chemokine receptor. *Mol. Pharmacol.* **1999**, *1*, 67–73.
- (73) Egberink, H. F.; De Clercq, E.; Van Vliet, A. L. W.; Balzarini, J.; Bridger, G. J.; Henson, G.; Horzinek, M. C.; Schols, D. Bicyclams, selective antagonists of the human chemokine receptor CXCR4, potentially inhibit feline immunodeficiency virus replication. *J. Virol.* **1999**, *8*, 6346–6352.
- (74) Pérez-Nuño, V. I.; Pettersson, S.; Ritchie, D. W.; Borrell, J. I.; Teixeira, J. Discovery of novel HIV entry inhibitors for the CXCR4 receptor by prospective virtual screening. *J. Chem. Inf. Model.* **2009**, *4*, 810–823.
- (75) Skerlj, R.; Bridger, G.; McEachern, E.; Harwig, C.; Smith, C.; Kaller, A.; Veale, D.; Yee, H.; Skupinska, K.; Wauthy, R. Design of novel CXCR4 antagonists that are potent inhibitors of T-tropic (X4) HIV-1 replication. *Bioorg. Med. Chem. Lett.* **2011**, 1414–1418.
- (76) Bridger, G. J.; Skerlj, R.; Kaller, A.; Harwig, C.; Bogucki, D.; Wilson, T.; Crawford, J.; Mceachern, E.; Atsma, B. *Chemokine Receptor Binding Heterocyclic Compounds with Enhanced Efficacy*. United States Patent WO/2003/055876, 2003.
- (77) Ichiyama, K.; Yokoyama-Kumakura, S.; Tanaka, Y.; Tanaka, R.; Hirose, K.; Bannai, K.; Edamatsu, T.; Yanaka, M.; Niitani, Y.; Miyano-Kurosaki, N. A duodenally absorbable CXC chemokine receptor 4 antagonist, KRH-1636, exhibits a potent and selective anti-HIV-1 activity. *Proc. Natl. Acad. Sci. USA* **2003**, *7*, 4185–4190.
- (78) Yamazaki, T.; Saitou, A.; Ono, M.; Yokoyama, S.; Bannai, K.; Hirose, K.; Yanaka, M. *Novel Nitrogenous Compound and Use Thereof*. World Patent WO/2003/029218, 2003.
- (79) Mosley, C. A.; Wilson, L. J.; Wiseman, J. M.; Skudlarek, J. W.; Liotta, D. C. Recent patents regarding the discovery of small molecule CXCR4 antagonists. *Expert Opin. Ther. Pat.* **2009**, *1*, 23–38.
- (80) Bodart, V.; Anastassov, V.; Darkes, M. C.; Idzan, S. R.; Labrecque, J.; Lau, G.; Mosi, R. M.; Neff, K. S.; Nelson, K. L.; Ruzek, M. C. Pharmacology of AMD3465: a small molecule antagonist of the chemokine receptor CXCR4. *Biochem. Pharmacol.* **2009**, *8*, 993–1000.
- (81) Bridger, G. J.; Boehringer, E. M.; Wang, Z.; Schols, D.; Skerlj, R. T.; Bogucki, D. E. *Methods to treat conditions mediated by chemokine receptors*. United States Patent 7,160,872 B2, 2008.
- (82) Bridger, G. J.; Boehringer, E. M.; Wang, Z.; Schols, D.; Skerlj, R. T.; Bogucki, D. E. *Antiviral compounds*. United States Patent 6,506,770 B1, 2003.
- (83) Catalano, J. G.; Gudmundsson, K. S.; Svolto, A.; Boggs, S. D.; Miller, J. F.; Spaltenstein, A.; Thomson, M.; Wheelan, P.; Minick, D. J.; Phelps, D. P. Synthesis of a novel tricyclic 1, 2, 3, 4, 4a, 5, 6, 10b-octahydro-1, 10-phenanthroline ring system and CXCR4 antagonists with potent activity against HIV-1. *Bioorg. Med. Chem. Lett.* **2010**, *7*, 2186–2190.
- (84) Bridger, G. J.; Skerlj, R. T.; Thornton, D.; Padmanabhan, S.; Martellucci, S. A.; Henson, G. W.; Abrams, M. J.; Yamamoto, N.; Vreese, K. D. Synthesis and structure-activity relationships of phenylenebis (methylene)-linked bis-tetraazamacrocycles that inhibit HIV replication. Effects of macrocyclic ring size and substituents on the aromatic linker. *J. Med. Chem.* **1995**, *2*, 366–378.
- (85) Tamamura, H.; Ojida, A.; Ogawa, T.; Tsutsumi, H.; Masuno, H.; Nakashima, H.; Yamamoto, N.; Hamachi, I.; Fujii, N. Identification of a new class of low molecular weight antagonists against the chemokine receptor CXCR4 having the dipicolylamine-zinc (II) complex structure. *J. Med. Chem.* **2006**, *11*, 3412–3415.
- (86) Jenkinson, S.; Thomson, M.; McCoy, D.; Edelstein, M.; Danehower, S.; Lawrence, W.; Wheelan, P.; Spaltenstein, A.; Gudmundsson, K. Blockade of X4-Tropic HIV-1 cellular entry by GSK812397, a potent noncompetitive cxcr4 receptor antagonist. *Antimicrob. Agents Chemother.* **2010**, *2*, 817.
- (87) Gudmundsson, K. S.; Boggs, S. D.; Catalano, J. G.; Svolto, A.; Spaltenstein, A.; Thomson, M.; Wheelan, P.; Jenkinson, S. Imidazopyridine-5, 6, 7, 8-tetrahydro-8-quinolinamine derivatives with potent activity against HIV-1. *Bioorg. Med. Chem. Lett.* **2009**, *22*, 6399–6403.
- (88) Bridger, G. J.; Skerlj, R. T.; Hernandez-Abad, P. E.; Bogucki, D. E.; Wang, Z.; Zhou, Y.; Nan, S.; Boehringer, E. M.; Wilson, T.; Crawford, J. Synthesis and Structure–Activity Relationships of Azamacrocyclic CXC Chemokine Receptor 4 Antagonists: Analogues Containing a Single Azamacrocyclic Ring are Potent Inhibitors of T-Cell Tropic (X4) HIV-1 Replication. *J. Med. Chem.* **2009**, *3*, 1250–1260.
- (89) Yen, C. F.; Hu, C. K.; Chou, M. C.; Tseng, C. T.; Wu, C. H.; Huang, Y. H.; Chen, S. J.; King, C. H. R. *Pyrimidine compounds*. United States Patent US 7,589,197 B2, 2009.
- (90) Hachet-Haas, M.; Balabanian, K.; Rohmer, F.; Pons, F.; Franchet, C.; Lecat, S.; Chow, K. Y. C.; Dagher, R.; Gizzi, P.; Didier, B. Small neutralizing molecules to inhibit actions of the chemokine CXCL12. *J. Biol. Chem.* **2008**, *34*, 23189–23199.
- (91) Miller, J. F.; Turner, E. M.; Gudmundsson, K. S.; Jenkinson, S.; Spaltenstein, A.; Thomson, M.; Wheelan, P. Novel N-substituted benzimidazole CXCR4 antagonists as potential anti-HIV agents. *Bioorg. Med. Chem. Lett.* **2010**, *7*, 2125–2128.
- (92) Gudmundsson, K. S.; Sebahar, P. R.; Richardson, L. D.; Miller, J. F.; Turner, E. M.; Catalano, J. G.; Spaltenstein, A.; Lawrence, W.; Thomson, M.; Jenkinson, S. Amine substituted N-(1-H-benzimidazol-2-ylmethyl)-5, 6, 7, 8-tetrahydro-8-quinolinamines as CXCR4 antagonists with potent activity against HIV-1. *Bioorg. Med. Chem. Lett.* **2009**, *17*, 5048–5052.

(93) Gudmundsson, K. *Pyranopyridine compounds*. WO/2007/008539, 2007.

(94) Murakami, T.; Kumakura, S.; Yamazaki, T.; Tanaka, R.; Hamatake, M.; Okuma, K.; Huang, W.; Toma, J.; Komano, J.; Yanaka, M. The novel CXCR4 antagonist KRH-3955 is an orally bioavailable and extremely potent inhibitor of human immunodeficiency virus type 1 infection: comparative studies with AMD3100. *Antimicrob. Agents Chemother.* **2009**, *7*, 2940–2948.

(95) Thomas, W. D.; Leleti, M. R.; Pennell, A. M. K. *CXCR4 modulators*. United States Patent 2007/0275965, 2007.

(96) Kokubo, M.; Tanaka, M.; Ochiai, H.; Takaoka, Y.; Shibayama, S. *Compound containing basic group and use thereof*. WO/2007/049771, 2007.

(97) Liang, Z.; Zhan, W.; Zhu, A.; Yoon, Y.; Lin, S.; Sasaki, M.; Klapproth, J. M. A.; Yang, H.; Grossniklaus, H. E.; Xu, J. Development of a unique small molecule modulator of CXCR4. *PLoS One* **2012**, *4*, e34038–e34038.

Artículo 4



Studying the binding interactions of allosteric agonists and antagonists of the CXCR4 receptor



Jesús M. Planesas^{a,1}, Violeta I. Pérez-Nueno^{a,b,*}, José I. Borrell^{a,1}, Jordi Teixidó^{a,1}

^a Grup d'Enginyeria Molecular, Institut Químic de Sarrià (IQS), Universitat Ramon Llull, Barcelona, Spain

^b Harmonic Pharma, Espace Transfert, 615 rue du Jardin Botanique, 54600 Villers lès Nancy, France

ARTICLE INFO

Article history:

Received 22 February 2015

Received in revised form 6 May 2015

Accepted 7 May 2015

Available online 30 May 2015

Keywords:

CXCR4

Allosteric agonists and antagonists

Allosteric binding sites

Docking

ABSTRACT

Several examples of allosteric modulators of GPCRs have been reported recently in the literature, but understanding their molecular mechanism presents a new challenge for medicinal chemistry. For the specific case of the cellular receptor CXCR4, it is known that peptidic fragments of intracellular GPCR loops) such as ATI-2341 modulate CXCR4 activity agonistically via an allosteric mechanism. Moreover, there are also examples of small organic molecules such as AMD11070 and GSK812397 which may also act as allosteric antagonists. However, incomplete knowledge of the ligand-binding sites has hampered a detailed molecular understanding of how these inhibitors work. Here, we attempt to answer this question by analysing the binding interactions between the CXCR4 receptor and the above-mentioned allosteric modulators. We propose two different allosteric binding sites, one located in the intracellular loops 1, 2 and 3 (ICL1, ICL2 and ICL3) which binds the peptidic agonist ATI-2341, and the other at a subsite of the main extracellular orthosteric binding pocket between extracellular loops 1 and 2 and the N-terminus, which binds the antagonists AMD11070 and GSK812397. Allosteric interactions between the CXCR4 and ATI-2341 were predicted by combining different modeling approaches. First, a rotational blind docking search was applied and the best poses were subsequently refined using flexible docking methods and molecular dynamic simulations. For the AMD11070 and GSK812397 antagonists, the entire CXCR4 protein surface was explored by blind docking in order to define the binding region. A second docking analysis by subsites was then performed to refine the allosteric interactions. Finally, we identified the binding residues that appear to be essential for CXCR4 allosteric modulators.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Cellular allosteric regulation is a typical process associated to conformational and functional transitions in protein structures. Allosteric behaviors are found in monomeric proteins, macromolecular complexes, DNA or RNA. Allostery (or allosteric regulation) describes the regulation of the protein function, structure and/or flexibility induced by the binding of a ligand called allosteric modulator or allosteric effector. Allosteric modulators bind to a site on the structure (allosteric site) located in a different topographical region than the catalytic/active binding site (orthosteric site), but

modifying the activity of the catalytic/active binding site [1]. Alteration of the orthosteric ligand affinity is an essential evidence of the presence of allosteric modulators. The result of this alteration is an increment or decrease of the protein–ligand complexation rate [2].

Allosteric regulation is generally due to the non-covalent binding of ions, lipids, proteins, DNA, RNA, or small pharmacological compounds [3,4] to the allosteric sites as activators/agonists, inhibitors/antagonists, allosteric modulators, or other effector types. The receptors when undergo allosteric conformational changes can propagate them to subsequent biochemical processes involving other receptors. Simultaneously, a cascade of allosteric changes generates perturbations or allosteric mutations. These allosteric changes can both cause disease and contribute to development of new therapeutics [5]. Nevertheless, the biocatalytic process of allostery represents an opportunity to regulate abnormal cellular behaviors through the action of ligands playing the role of drug modulators. Chemical allosteric modulators boast several advantages over orthosteric ligands as potential therapeutic agents due to their dormancy in the absence of endogenous–orthosteric activity. Allosteric modulators show greater selectivity as a result

* Corresponding author at: Grup d'Enginyeria Molecular, Institut Químic de Sarrià (IQS), Universitat Ramon Llull, Barcelona, Spain.

Tel.: +34 93 267 20 00/+33 354 958 604; fax: +34 93 205 62 66/+33 383 593 046.

E-mail addresses: violeta.perez@iqs.url.edu, pereznueno@harmonicpharma.com (V.I. Pérez-Nueno), jordi.teixido@iqs.edu (J. Teixidó).

¹ Tel.: +34 93 267 20 00; fax: +34 93 205 62 66.

of higher sequence divergence in the allosteric site. Also, they show limited positive or negative cooperation imposing a ‘ceiling’ on the magnitude of their allosteric effect [6]. In recent years, remarkable progress has been made in the optimization, discovery and clinical development of allosteric drugs in kinases, GPCRs and ion channels by the pharmaceutical industry [6].

Regarding the phenomenon of allostery in GPCR proteins [7], there are more than 100 allosteric modulators reported [6], while currently most of the pharmacological compounds whose target is a GPCR act as orthosteric ligands. In these receptors, the known allosteric sites are structurally and evolutionary less conserved than orthosteric sites. Precisely, Anderson et al. [8] studied the superposition of 60 different crystal structures verifying this high evolutionary conservation in terms of structure and sequence. In general, allosteric sites in GPCRs can be located in the extracellular side of the receptor, in the transmembrane region, as well as in the intracellular side of the protein. It is observed that depending on the GPCR subfamily, allosteric sites are placed in concrete regions [9]. Thus, chemokine receptors usually have two allosteric binding sites, one located in the transmembrane region and the other in the intracellular region inside the cytoplasm [10]. Attending to these observations a wide range of opportunities to modulate proteins like GPCRs opens up in the forthcoming years.

The mechanism of action of allosteric GPCR modulators has been classified by Ballesteros et al. [11] as in-target and off-target, attending to if modulators bind directly to the GPCR. Contrary, if modulators bind to a second protein which binds to the GPCR forming a heterodimer complex.

Focusing on CXCR4 allostery, Sachpatzidis et al. [12] identified two allosteric and agonist peptides (RSVM y ASLW). Both peptides, with 17 residues, do not bind at the same binding site as the small organic antagonist AMD3100 and the antagonist peptide T140 do. Therefore, both peptides exhibit allosteric behavior by binding to CXCR4 in an alternative binding site to the orthosteric site.

Later, Mosi et al. [13] proposed the antagonist AMD11070 (Fig. 1) as potential inhibitor of the interaction CXCR4-SDF-1, the natural ligand of CXCR4 receptor. AMD11070 would act as a selective non-competitive inhibitor of isoforms SDF-1 α and SDF-1 β (or CXCL12a and CXCL12b, respectively) with different modes of action and different IC₅₀ values. Hence, AMD11070 would use a mixed mechanism of inhibition without producing allosteric effects to other chemokine receptors like CXCR3, CCR1, CCR2b, CCR4, CCR5 or CCR7 [14]. These characteristics, such as the selectivity of AMD11070, are typical from allosteric modulators.

Similarly, Boggs et al. [15] synthesized the small organic compound GSK812397 (Fig. 1), a bioavailable and oral non-competitive antagonist of CXCR4, which inhibits HIV entry. This pharmacological compound has a high selectivity and may be able to bind CXCR4 in a distinct and separate site than HIV. Thereby when GSK812397

binds to the receptor may induce conformational changes in CXCR4, inhibiting thermodynamically the entrance of the virus into the cell [16].

Other kind of pharmacological compounds commonly used as allosteric modulators are peptides or their peptidomimetic derivatives. Oishi et al. [17] reported a compilation of different peptides and peptidomimetic ligands of CXCR4. Among them, pepducins [18] are a representative family of intracellular regulators of GPCR signaling [19,20]. Pepducins are peptides with 10–20 residues, whose protein sequences are derived from the intracellular loops of the targeted GPCRs. Additionally, pepducins have in one of their terminal sequences a lipophilic moiety, like palmitate, myristate or lithocholic substituents, linked by peptide bonds [21]. Those substituents provide pepducins the capacity to penetrate through the lipid bilayer of the cytoplasmic membrane and anchor to it to establish interactions with the intracellular face of the target receptor [22]. Pepducins show the characteristic of allosteric modulators of GPCRs [21], they bind to the receptor stabilizing the conformations of GPCRs between active or inactive states, resulting in positive or negative modulation, respectively. Once stabilized, pepducins may block the ability to undergo further conformational changes, inhibiting the receptor to achieve the active state in the case of negative modulators, or contrarily, enhancing the active state for positive modulators. These blocked conformations could disable the process of receptor dimerization or oligomerization [23]. Janz et al. [24] reported three pepducins as allosteric agonists of CXCR4, because they are competitive among themselves but they do not compete with the antagonist peptide T-140 and neither against the endogenous ligand SDF-1. The non-competitive behavior of pepducins with SDF-1 ligand was corroborated using ATI-2766 (Fig. 1), an analogous peptidomimetic derivative of ATI-2341 [25]. ATI-2766 has 15 protein residues, one residue less than ATI-2341, and contains a different lipophilic tail, as well as a modified L-Photo-Leucine residue which exhibits photoaffinity in photo-chemical cross-linking experiments [26,27]. The lipophilic tail of ATI-2341 is a palmitic acid substituent while ATI-2766 has two lipophilic groups, 2-undecyloxyethane and the labeling reagent 5-(6)-carboxy-tetramethyl-rhodamine (TAMRA), which is commonly used in fluorescence assays to monitor biomolecular interactions [28].

In this work, we analyze and propose possible allosteric interactions between the CXCR4 receptor and its allosteric modulators. We deal with two different kinds of allosteric modulators attending to the nature of the compounds identified. On the one hand, pepducin modulators whose allosteric behavior has been experimentally verified [24] and establish peptide–protein interactions with CXCR4. On the other hand, the two small organic molecules, AMD11070 and GSK812397, whose allosteric mechanism is still unclear and requires further experimental investigation [13,16].

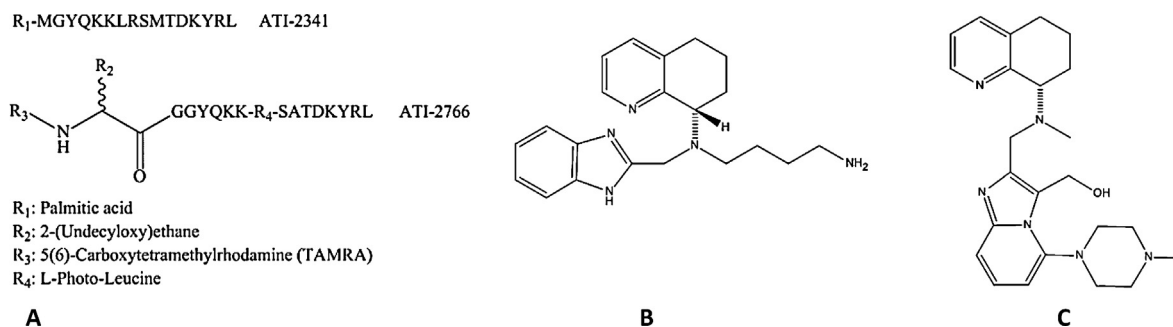


Fig. 1. Structures of CXCR4 allosteric inhibitors. (A) Allosteric peptidomimetic agonists pepducins ATI-2341 and ATI-2766 with their corresponding substituents. (B) Small organic antagonist AMD11070. (C) Small organic antagonist GSK812397.

2. Methods

We first used blind docking in order to establish the region with highest probability of involving allosteric interactions between pepducins and the CXCR4 receptor. We then analyzed pepducins binding modes by flexible protein–protein docking. Since many protein–protein or peptide–protein docking protocols work on the basis of residue–residue interactions, and heteroatoms belonging to non-peptidic substituents (HETATM) [29] are not considered part of the biopolymer chain, and neither their associated interactions, we considered convenient to compare the results obtained by flexible docking with results obtained by molecular dynamic simulations (MD).

The representative pepducin chosen to study CXCR4–pepducin allosteric interactions was ATI-2341, as its lipophilic tail is smaller than the ATI-2766, and the palmitate substituent of ATI-2341 has less HETATMs than the ATI-2766 lipophilic substituents ((2-undecyloxyethane) and TAMRA).

Preparation of pepducin ATI-2341 [24] consisted of protonation, addition of Gasteiger partial charges [30] and minimization of the structure using CHARMM27 [31] force field. Starting from the minimized and energetically stable structure of ATI-2341, a conformational search was applied using the MD protocol *Standard Dynamic Cascade* implemented in Discovery Studio [32]. The conformational search was performed in an implicit solvent model (GBSW) heating from 50K to 300K and using the force field CHARMM22 [33] in a time scale of 1 ns.

Regarding the CXCR4 receptor, we used 3OE6 structure, since in our previous work [34] this structure showed the best virtual screening (VS) performance in comparison with the rest of structures published in the Protein Data Bank [35]. PDB 3OE6 at 3.2 Å resolution presents the intracellular loop ICL3 uncompleted and connected to the protein of fusion T4-lysozyme (T4L). T4L is a protein used to facilitate crystallization of CXCR4 [36] and other GPCRs. The use of the highly crystallizable T4-lysozyme (T4L) is a common strategy in structure determination of protein membranes like GPCRs [37,38]. For example, T4-lysozyme has been used in the structural determination of β_2 adrenergic receptor (β_2 AR) [38,39] and adenosine A_{2A} receptor [40] in addition to CXCR4. This T4-lysozyme protein is fused to the intracellular loop ICL3 of CXCR4 suppressing 33 protein residues from its C-terminal domain [41]. Hence, T4L segments were removed from CXCR4 structure and ICL3 was completed and energetically optimized using Modeller [42].

2.1. Peptidomimetic-protein blind docking using ZDock

The interactions between CXCR4 protein structure and the pepducin ATI-2341 peptidomimetic structure were first approached by blind docking using ZDock [43]. ZDock is a rigid body docking method based on the Fast Fourier Transform [44] that scans the surface of the protein considered as receptor. The protein considered as ligand is moved rotationally and translationally searching for the best tridimensional shape complementarity between receptor and ligand, and the corresponding electrostatic interactions. ZDock is a recommendable protein–protein docking protocol [45–47], mainly when there is no previous information about the residues involved in the interactions between both proteins, neither about the specific region of potential binding sites. In this way, the convenience of using a global systematic search method, like ZDock, in combination with other accurate docking methods has been shown to be a successful strategy to establish protein–protein interactions [48].

The scoring function ZDock Score [45] can be expressed in terms of shape complementarity plus electrostatics and desolvation energy. For the complex CXCR4–ATI-2341, we only consider the shape complementary term, mainly because ZDock was used only as a first approach to search the most favorable binding region.

Detailed energy terms were scored later in the flexible docking and MD steps in order to establish accurate interactions.

ZDock was set with a rotation angle step of 6° and without restrictions for the rotational sampling of the pepducin. For each docking run 2000 clustered poses were returned. Poses were grouped and assigned to the same cluster attending to a RMSD cutoff of 5 Å between poses.

2.2. Peptide–protein flexible dockings

Flexible dockings [49] were performed to refine the pepducin–protein interactions found by blind docking. The best pose clustered in the best binding region returned by the rotational blind docking was used as starting conformation. The use of hybrid analysis of rescoring–refinement has been shown to improve protein–protein docking solutions [50]. For example, Pierce et al. [51] showed the improvement of protein docking solutions when a combination of protocols like ZDock and RosettaDock were used.

As pepducins ATI-2341 and ATI-2766 have 105 and 97 rotatable bonds and protein residues compose their active backbone, the use of protein–protein or peptide–protein docking protocols were considered as the most suitable to search for their key interactions.

Accordingly, flexible docking protocols Rosetta FlexPepDock Web Server [52] and HADDOCK Web Server (High Ambiguity Driven protein–protein DOCKing [53]) were used to study and compare the allosteric interactions retrieved among themselves.

2.2.1. FlexPepDock

Rosetta FlexPepDock is a high resolution protocol that refines protein–peptide complexes through a process which enables full flexibility of the peptide and side-chain flexibility of the protein receptor [54]. FlexPepDock web server is mainly orientated to peptide–protein dockings and classifies the complexes predicted by its Rosetta energy score. The search can be performed in low resolution or in high-resolution phase and, consequently, the energy score calculation varies depending on it. The high resolution phase includes all atoms and is controlled by Van der Waals interaction energy, Gaussian solvation, electrostatic energy and orientation-dependent hydrogen bonding potential [50].

2.2.2. HADDOCK

HADDOCK [55] is a high ambiguity driven protein–protein docking protocol used, essentially, when some information is available regarding the interaction between both proteins [48]. This allows guiding the protein conformational search and, hence, the web server protocol [53] requires for each protein to specify a minimum number of residues involved in these interactions. HADDOCK uses a knowledge-based method for driving the flexible docking which combines conformational selection and induced fit mechanisms [56].

2.2.3. Peptidomimetic-protein MD

MD simulations have been used in multiple cases [57–60] to study protein–ligand interactions as alternative approximations to flexible docking protocols, mainly when reduced sets are tested [61]. In proteins [62], the possibility to design specific force fields to simulate protein motions is also considered as an advantage with respect to flexible docking methods.

Protein–protein flexible docking protocols typically do not incorporate interactions involving organic substituents whose nature is non-peptidic. As pepducins are peptidomimetic compounds whose lipophilic tail is a non-peptide substituent, it was considered relevant to study the CXCR4–ATI-2341 interactions using MD simulations. Comparison with flexible docking results was performed to check if the lipophilic palmitate interacts or

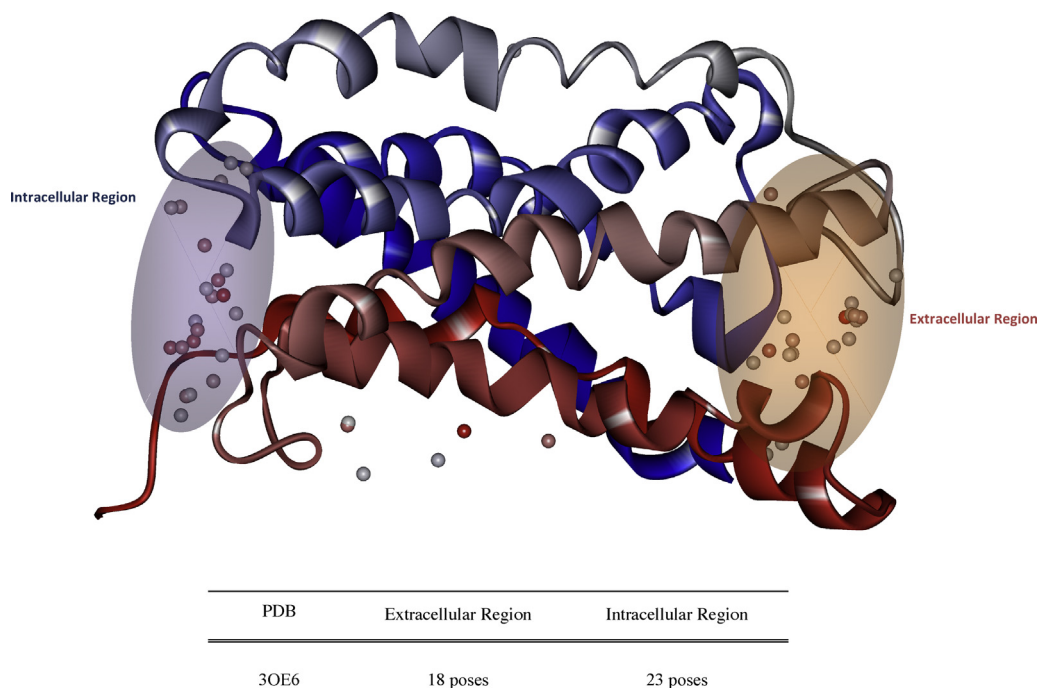


Fig. 2. 3OE6 regions where the best 50 ATI-2341 ZDock poses were distributed. Pepducin poses are represented in balls in colors ranging from red to blue according to ZDock Score (red = best and blue = poor ZDock Scores). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

clashes with CXCR4 or if modifies the interactions found by the flexible docking protocols.

MD simulations were performed using the *Standard Dynamic Cascade* protocol implemented in Discovery Studio [32] in order to calculate motions and trajectories of the pepducin–CXCR4 complex. The *Standard Dynamic Cascade* includes two minimization steps, a heating followed by an equilibration MD step, and finally the production MD step. This standard protocol usually can be used, among others, to test and verify the stability of protein–ligand binding modes obtained by docking [63], to generate stable conformations of peptides [64], or also to predict flexibility in structurally modified enzymes at different temperatures [65].

Regarding CXCR4, as its physiology corresponds to a protein membrane, all MD simulations were performed adding an implicit membrane GBSW [66] and setting an implicit solvent model GBSW [67] to estimate the solvation energy using the force field CHARMM22 [33]. Due to the influence of the membrane thickness in MD results, we also performed simulations at 25.4 Å, 28.5 Å and 30.0 Å in order to find the energetically most stable complex [68].

2.3. Small ligand–protein blind docking

For the small molecules AMD11070 and GSK812397, whose allosteric mechanism is still unclear [13,16], an analogous approach was proposed to study their interactions. The entire CXCR4 protein surface was explored by blind docking in order to define the binding region, and a second docking analysis by subsites was then performed to refine the allosteric interactions.

Initially, both structures were prepared using *LigPrep* [69] from Schrödinger. Force field OPLS.2005 was applied to generate possible ionization states at pH 7 ± 2 with *ionizer*, and also to generate tautomers and stereoisomers. All the resulting structures were then blind docked to 3OE6 CXCR4 crystal structure using AutoDock Vina [70]. In the blind docking the grid box volume (27,000 cubic Angstroms) covered the whole protein structure. The best poses were superposed, analyzed and clustered. Clusters

were re-grouped in regions attending to the region of CXCR4 where they interact using Discovery Studio.

2.3.1. LigandFit docking by subsites

Once established the main region with more docked poses, we analyzed the AMD11070 and GSK812397 binding modes by a subsequent docking analysis by subsites. In a docking by subsites, the binding region to explore is divided sequentially into one, two, three or more subsites depending on the partition level defined, and then each structure is docked into all subsites. Hence, if the partition level is equal to three, every structure is docked 6 times and only the docked pose with the best scoring is retrieved with the associated subsite. This method becomes essential for proteins with big pockets, such as CXCR4.

The use of LigandFit [71] docking functions in our previous work [34] showed very good results when screening a test set of CXCR4 actives and decoys by DockScore scoring function. Moreover, regarding the evaluation of pose prediction, self-docking the cognate ligands showed also LigandFit as one of the best performing functions, principally for structures 3ODU and 3OE6. Hence, LigandFit docking function was used to divide the main binding site applying 3 partition levels, which created 6 binding subsites, according to our previous work [34].

3. Results

3.1. Peptidomimetic-protein blind docking using ZDock

Blind dockings of ATI-2341 to 3OE6 structure generated 2000 poses with their corresponding ZDock Score, grouped in different clusters attending to the RMSD cutoff defined (5 Å). The best 50 docked poses with the highest ZDock values were retained. The clusters associated to these poses were classified attending to the number of docked poses they included and their corresponding ZDock Score. Finally, adjacent clusters sharing the same topographic protein sectors were grouped as regions of high probability of interactions.

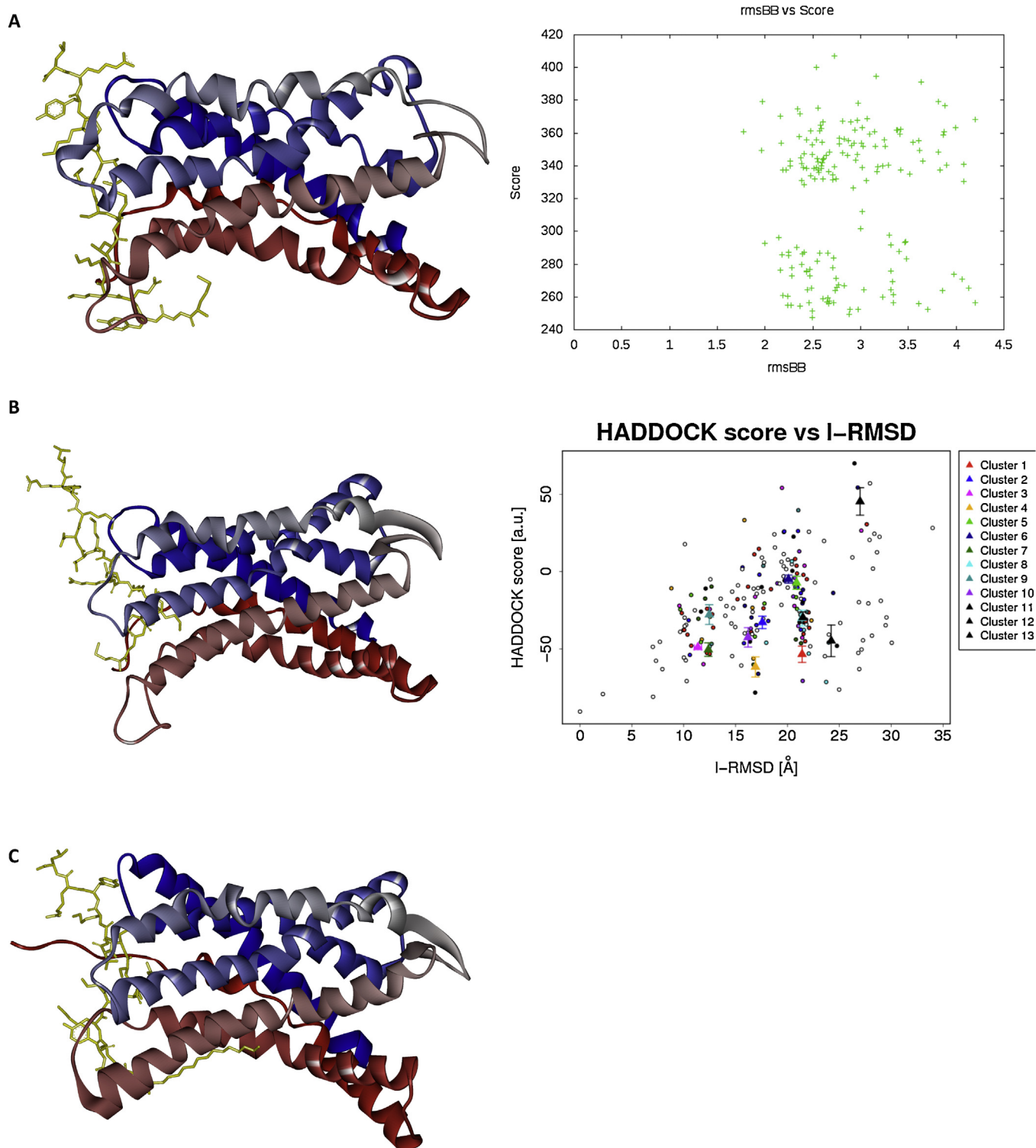
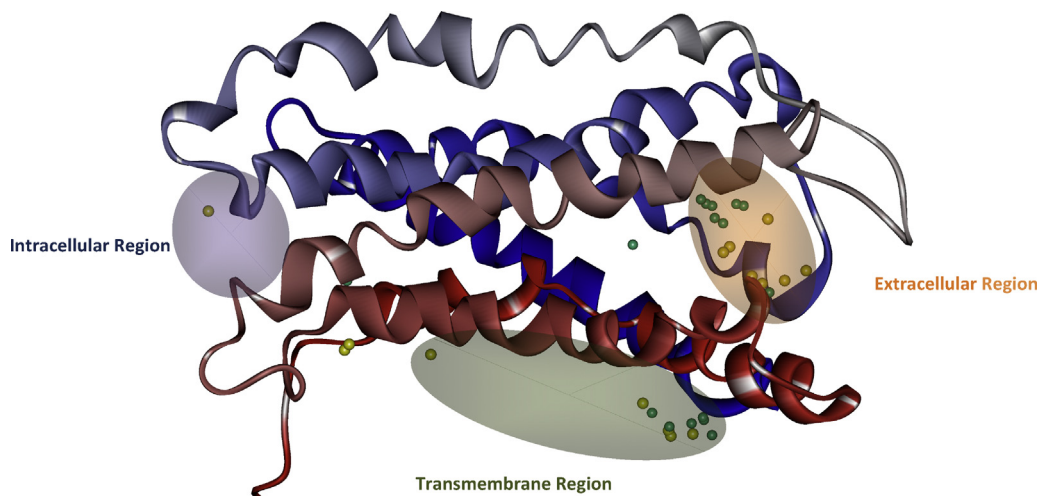


Fig. 3. (A) Best ATI-2341 pose obtained by FlexPepDock, located in ICL1 of PDB 30E6. On the right, the global flexible docking results. Score values are represented versus ligand RMSD over peptide backbone heavy atoms (*rmsBB*) with respect to the initial conformation. (B) HADDOCK best scored pose (left) for 30E6 and the global results grouped in clusters (right). Score values are represented versus the ligand RMSD (I-RMSD) [56]. (C) Final and stabilized conformation of ATI-2341-30E6 obtained by MD simulation.

Results showed two different regions where the best clusters were distributed (see Fig. 2). The first region was located in the N-terminal domain, just in the entrance of the main pocket of CXCR4, and delimited by the three extracellular loops (called *Extracellular Region*). The second region was found in the intracellular domain of

the receptor, which includes the intracellular loops ICL1, ICL2 and ICL3 (called *Intracellular Region*).

As it can be observed in Fig. 2, the majority of the best 50 poses was found in the *Intracellular Region* (23 poses) and secondly in the *Extracellular Region* (18 poses). If we only consider the best



PDB	Extracellular Region	Transmembrane Region	Intracellular Region
	<i>AMD11070/GSK812397</i>	<i>AMD11070/GSK812397</i>	<i>AMD11070/GSK812397</i>
3OE6	8/7 poses	7/4 poses	0/1 poses

Fig. 4. Scheme of the docking regions containing the majority of AMD11070 and GSK812397 blind-docked poses. Transmembrane Region is located on the surface of helix 6 and 7 in the cytoplasmic membrane domain. Intracellular Region is located between ICL1 and ICL3 domains. Extracellular Region covers the main extracellular pocket of CXCR4. The table below shows the number of best poses of AMD11070 and GSK812397 found by region. Values in bold indicate the best energy pose and the region where it was placed. AMD11070 and GSK812397 poses are represented in balls green and yellow, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

10 scored poses, the *Intracellular Region* retains 6 poses and the *Extracellular Region* places only 2 docked poses with the best ZDock score.

3.2. Peptide–protein flexible dockings

According to the blind docking results, we can consider the existence of two different binding sites where pepducin ATI-2341 could bind, one in the extracellular side of the protein and the other in the intracellular domain of CXCR4. Nevertheless, it is relevant to note that all pepducins have lipophilic tails to cross through the plasmatic membrane, hence facilitating their binding in the intracellular face of proteins. Moreover, the conformations retrieved located in ICL1 agree with the fact that pepducins ATI-2341 and ATI-2766 are derived from ICL1 of GPCRs [21]. Attending to these considerations and the fact that the highest number of best ZDock poses was retrieved in the *Intracellular Region*, we explored further the potential binding site located in the intracellular domain of CXCR4. Therefore, the best pose retrieved in the *Intracellular Region* was chosen as starting conformation to further study CXCR4–pepducin interactions.

3.2.1. FlexPepDock

FlexPepDock web server requires one structure containing simultaneously the ligand and the receptor. As input structures, we used the best intracellular ZDock poses of ATI-2341 for 3OE6. The protocol was set to generate 200 structures of high resolution for each FlexPepDock run.

Once completed the simulation, FlexPepDock returned the top 10 models found with their associated scoring values expressed as total Rosetta energy. The energy score obtained was 247.23 and the *rmsBB* was 2.49 (Fig. 3).

Distances between CXCR4 residues and ATI-2341 residues are shown in Supporting Information (Fig. 1) in a phylogram chart.

It is worth noting that the palmitate substituent in the lipophilic tail was absent in the pepducin ATI-2341 conformations retrieved by FlexPepDock.

3.2.2. HADDOCK

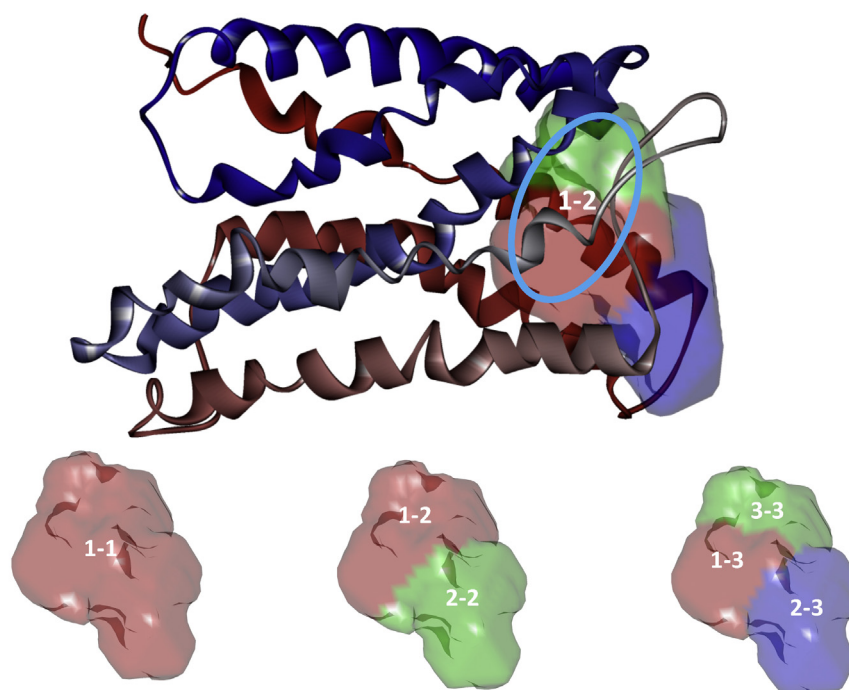
To perform the flexible docking using HADDOCK, the receptor and the ligand must be supplied independently, as well as the starting protein–ligand interaction residues. For the pepducin ATI-2341, all 16 residues were specified as active residues, and for CXCR4 a distance threshold of 5 Å was defined to select the residues that interact with the conformation obtained by ZDock.

The flexible docking performed in 3OE6 returned 10 clusters and 4 best poses in each cluster. The best cluster (cluster 4 in Fig. 3B) had a HADDOCK score of -61.6 ± 13.1 . Distances between these residues and ATI-2341 residues are shown in Supporting Information (Fig. 1) in a phylogram chart.

In the same way than using FlexPepDock, the palmitate substituent in the non-peptidic tail of the pepducin ATI-2341 conformations was not returned using HADDOCK flexible docking.

3.2.3. Peptidomimetic-protein MD

The existence of the lipophilic substituent palmitate in pepducins, which is not taken into consideration by flexible peptide–protein protocols, can introduce reasonable doubts from the standpoint of overall interactions and clashes. The possibility to undergo distinct conformational changes may induce lack of information in the allosteric interaction map. For this reason, MD simulations were performed and compared with flexible



PDB		Subsite					
		1-1	1-2	1-3	2-2	2-3	3-3
3OE6	AMD11070	-	5	11	0	0	0
	GSK812397		4	11	0	1	0

Fig. 5. Representation of the 3 partition levels applied to CXCR4 extracellular pocket in order to dock AMD11070 and GSK812397 by subsites. Subsite 1–1 is generated when only one partition level is applied to the binding site. Subsites 1–2 and 2–2 are generated when two partition levels are applied. Subsites 1–3, 2–3 and 3–3 are created when three partition levels are applied. Applying 3 partition levels create a total of 6 subsites. Hence, an exhaustive site exploration can be performed by docking AMD11070 and GSK812397 into each sub-site. The table below shows the number of best-docked poses located in each subsite for structures AMD11070 and GSK812397. Values in bold indicate the subsite and the structure that returned the best-scored pose.

peptide–protein docking results in terms of interactions and conformational trajectories generated by MD for the whole pepducin structure including organic substituents.

The MD steps (minimization, heating, equilibration, dynamics and production) were set as follows. An initial minimization step was performed using the steepest descent algorithm, defining 1000 minimization cycles and a RMS gradient of 0.1 kcal/mol Å to finish the routine. A second minimization step, was performed defining 2000 minimization cycles and a RMS gradient of 0.0001 kcal/mol Å. The temperature was initially set at 50 K and gradually increased to 300 K during the heating time by increments of 0.001 ps and 2000 iteration steps fixed in a similar way to the equilibrium step at 300 K. The production step was performed at constant temperature of 300 K (canonical ensemble) in time steps of 0.001 ps and 500,000 steps.

Furthermore, MD simulations were performed adding an implicit membrane GBSW to CXCR4 and using the Implicit Solvent Model GBSW. The influence of membrane thickness was tested at 25.4 Å, 28.5 Å and 30.0 Å. Using a thickness of 28.5 Å, the relative position and orientation of the CXCR4 membrane, as well as the results obtained by MD were considered the most suitable.

The MD simulation of the pepducin ZDock poses returned 1000 complex conformations. The energies of the most stable models were equal to 2724.4 kcal/mol at 304.7 K for the complex ATI-2341–3OE6 (Fig. 3). Distances between CXCR4 residues and ATI-2341 residues are shown in Supporting Information (Fig. 1).

It is worth noting that the MD performed kept the complete structure of ATI-2341 including its palmitate substituent.

3.3. Small ligand–protein blind docking

LigPrep was used to prepare AMD11070 and GSK812397 small molecules for blind docking. Each molecule generated globally 16 structures among stereoisomers and tautomeric structures. These resulting structures were blind-docked using AutoDock Vina into the CXCR4 monomeric structure 3EO6. A large search space involving the whole protein was defined to explore completely the CXCR4 structure. The blind docking protocol returned 9 poses for every docked structure, being the best-ranked pose the one with the lowest free energy of binding. All the best-ranked poses were arranged in clusters and clusters were gathered into three main regions (Extracellular Region, Transmembrane Region and Intracellular Region). Fig. 4 shows the docking regions containing the majority of AMD11070 and GSK812397 best blind-docked poses into CXCR4. The majority of poses were distributed in the Extracellular Region and inside the main pocket of CXCR4. Accordingly, this region was selected to analyze and refine possible allosteric interactions between CXCR4 and the small molecules.

3.3.1. LigandFit docking by subsites

The binding site was defined using a total volume of 1520 cubic Angstroms that covered the total extracellular pocket. This volume

ATI-2341	Tyr65	Gln66	Lys67	Lys68	Leu69	Arg70	Ser71	Met72	Thr73	Asp74	Tyr76	Arg77
<i>FlexPepDock (3OE6)</i>				H				H				
<i>Haddock (3OE6)</i>					H				H			
<i>MD (3OE6)</i>									2H			

ATI-2341	Ile130	Asp133	Arg134	Arg148	Lys149
<i>FlexPepDock (3OE6)</i>			H	H	
<i>Haddock (3OE6)</i>	H	H	H		
<i>MD (3OE6)</i>		H	H	2H	

ATI-2341	Gln233	Lys234	Arg235	Lys236	Ala237	Leu238	Lys239	Thr240	Thr241	Val242	Ile243	Leu244	Leu246	Phe249
<i>FlexPepDock (3OE6)</i>				H	H			H						
<i>Haddock (3OE6)</i>					H			H			H			
<i>MD (3OE6)</i>				H	H	H								

ATI-2341	Leu301	Leu302	Ala303	Phe304	Leu305
<i>FlexPepDock (3OE6)</i>					
<i>Haddock (3OE6)</i>					
<i>MD (3OE6)</i>					

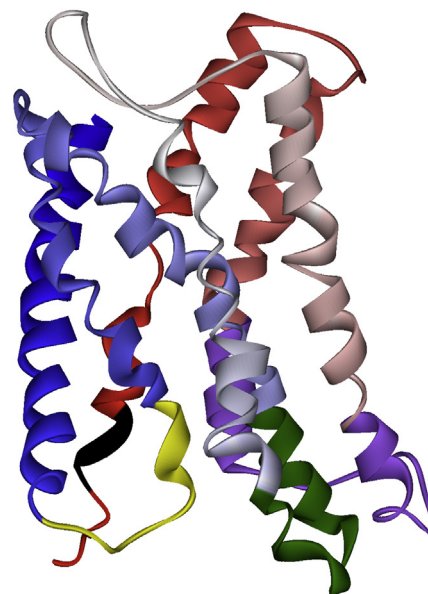


Fig. 6. Illustration of the interactions established between pepducin ATI-2341 and residues of CXCR4 according to each computational method. CXCR4 structure is shown with its 7-helices colored by N-to-C terminal in a continuous gradient from blue at the N-terminus through white to red at the C-terminus. Each table is related to the CXCR4 fragments colored in yellow for ICL1, in dark green for ICL2, in purple for ICL3 and in black for the C-terminal domain. Colored domains correspond to residues of CXCR4 that establish interactions with ATI-2341 according to FlexPepDock, HADDOCK, or MD results. Tables show polar interactions in magenta boxes and hydrophobic interactions in light green boxes. Side chain hydrogen bonds (H) are indicated in red letters and backbone hydrogen bonds in black letters.

was divided into 3 partition levels (1+2+3=6 subsites), where to dock meticulously all AMD11070 and GSK812397 molecules. Fig. 5 shows the defined binding subsites. Partition 1 is the original unpartitioned site. Partition 2 divides the original site into two parts, thereby producing two subsites, and partition 3 divides the

original site into three parts, thereby producing three subsites. Partitioning a site is performed using a cluster analysis algorithm that involves the use of random numbers. Every molecule was docked into each subsite using LigandFit docking function, hence enabling an exhaustive site search. The best binding pose, its DockScore

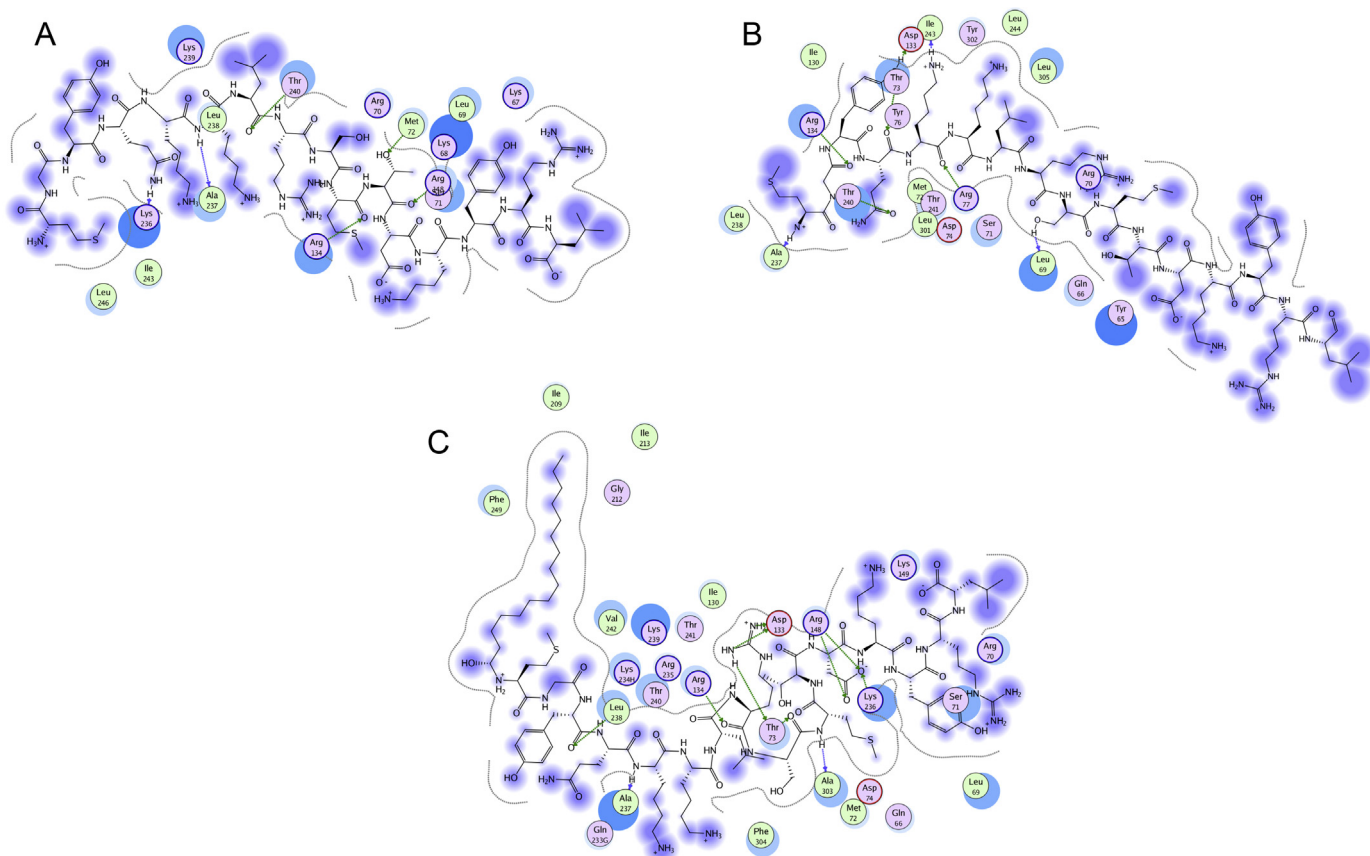


Fig. 7. Protein–ligand interactions established for the complexes CXCR4–ATI-2341 generated by FlexPepDock (A), HADDOCK (B) and MD (C). Circles in magenta represent residues involved in acidic, basic or polar interactions. Circles in light green represent residues involved in non-polar interactions. A blue halo around the residue represents a solvent accessible surface; this halo is proportional to the solvent accessible surface. In blue dashed lines with arrow heads are represented hydrogen bonds with amino acid backbone chains and green dashed lines with arrow heads represent hydrogen bonds with amino acid side chains. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

scoring and the subsite where it was found energetically favorable docked, were kept for each docking. As it can be seen in Fig. 5, docking results obtained for CXCR4 structures showed a distribution of binding poses mainly in binding subsites 1–2 and 1–3. Subsite 1–2 covers the volume defined by subsites 1–3 and 3–3 when the binding partition level passes from 2 to 3 partitions, respectively. The following distribution of docked poses was found: 9 binding poses were located in subsite 1–2 (28.1%), 22 poses in subsite 1–3 (68.8%), and only one pose in subsite 2–3 (3.1%), while subsites 2–2 and 3–3 did not contain any pose. The best-scored pose was found in subsite 1–2.

4. Discussion

Allostery in CXCR4 has been studied using pepducin agonists and small molecules (AMD11070 and GSK812397), which may act as allosteric antagonists. Regarding the pepducins, ATI-2341 was chosen instead of ATI-2766 due to its relative small lipophilic moiety in comparison to ATI-2766.

Blind dockings were performed using ZDock as the first step to approach how the allosteric pepducin ATI-2341 interacts with CXCR4. Rotational and translational blind docking is highly dependent on the ligand conformation that determines clashes and interactions with the protein receptor. For this reason, a starting ATI-2341 conformation for the blind docking was obtained using MD simulation.

Results generated by ZDock showed mainly two different regions where the highest number of the best poses was located.

The extracellular region included poses and clusters that correspond with the orthosteric region of CXCR4 where the endogenous binding site is located, whereas the intracellular region covered intracellular loops of CXCR4, basically the first intracellular loop (ICL1).

Pepducins are protein sequences derived from one of the three intracellular loops [19,21] of different GPCRs, in particular pepducins ATI-2341 and ATI-2766 are protein sequences derived from ICL1. This evidence and the fact that the highest number of best ZDock poses was retrieved in the intracellular domain, makes the *Intracellular Region* the area to explore further pepducin–CXCR4 interactions by flexible protein–protein docking. The role that the lipophilic tail plays in the binding of ATI-2341 to CXCR4 was explored by MD simulations.

Fig. 6 shows the interaction map found using the different computational approaches. It can be seen that the resulting pepducin conformations were moved in all protocols from the initial position in the interface region defined mainly by ICL2 and ICL3, to a displaced position near to ICL1 and establishing interactions with the ICL2. Probably the binding of pepducin modulator to this interface pocket could modify the conformational changes that CXCR4 undergoes when it is activated by the endogenous ligand, altering the corresponding signaling response.

Analyzing the typology of the interactions established between pepducin ATI-2341 and CXCR4 (Figs. 6 and 7) it can be observed that residues Arg70 and Ser71 from ICL1, Arg134 from ICL2, and Thr240 from ICL3 establish polar interactions in all methods used. In addition, residues Leu69 and Met72 from ICL1, and Ala237 and

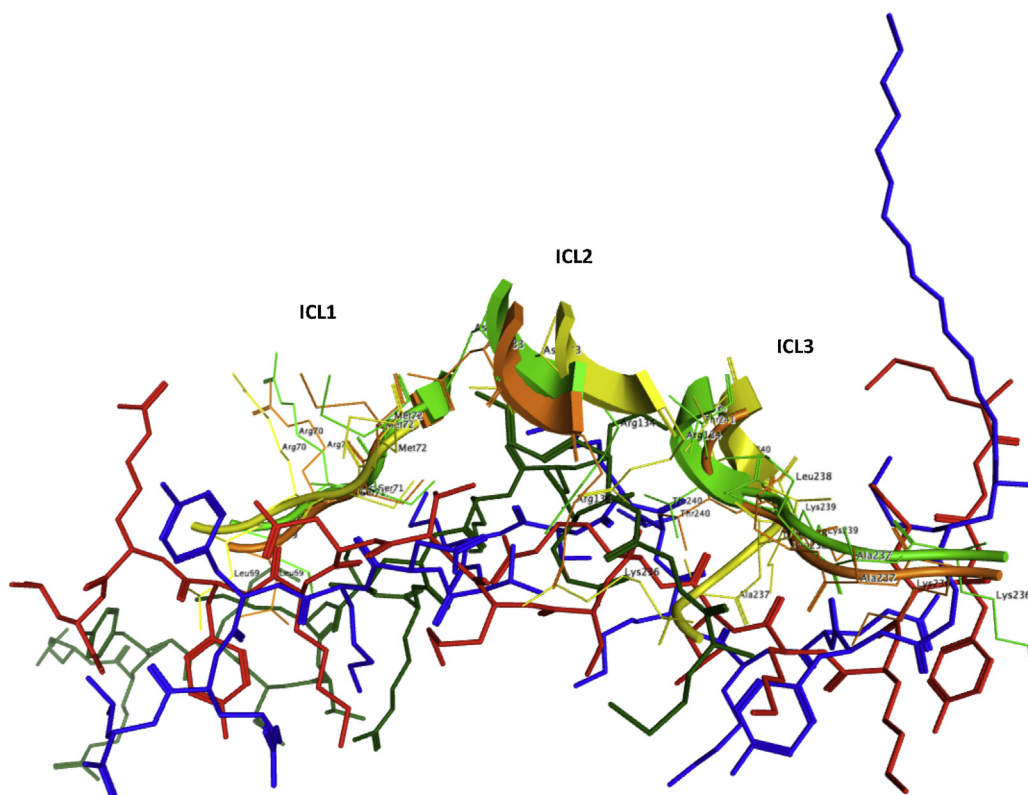


Fig. 8. Superposed pepducin ATI-2341 conformations returned by FlexPepDock in red, by Haddock in green and by MD in blue. Representation generated by protein superposition. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Leu238 from ICL3 establish lipophilic interactions also in all computational protocols. Attending to the hydrogen bond interactions, Arg134 and Ala237 formed hydrogen bonding in all the different methods used, and residues Thr73 from ICL1, Asp133 and Arg148 from ICL2, and Lys236 as well as Thr240 from ICL3 established hydrogen bonding in accordance with two of the three methods used.

Fig. 8 shows the superposition of ATI-2341 conformations obtained by flexible dockings and by MD. It can be observed how the three refining methods present different protein folding for ATI-2341, principally due to the high flexibility of the peptidomimetic compound. Nevertheless, it can be noticed that the three conformers are located in the same CXCR4 binding site and some of the pepducin residues for FlexPepDock and MD show the same interactions with ICL1 and ICL2.

Regarding the small molecules, the blind docking performed by AutoDock Vina on CXCR4, for the overall 32 AMD11070 and GSK812397 isomeric structures, showed mainly 3 different clusters. The main cluster containing the greatest number of docked poses was located in the extracellular side of CXCR4 (*Extracellular Region*). This region corresponds to the main pocket of CXCR4 where the endogenous ligand SDF-1 binds. The other two clusters were placed on the external surface of the transmembrane domain, between helices 6 and 7 (*Transmembrane Region*) similarly to the allosteric binding site recently discovered [72] in purinergic receptor P2Y₁, and the second one, between ICL1 and ICL3, in an equivalent region to the one found for pepducin ZDock dockings (*Intracellular Region*).

According to the distribution of blind docked poses, the wide *Extracellular Region* was considered the region with higher probability to establish interactions between the small ligands AMD11070, GSK812397 and CXCR4 receptor, and therefore the region for docking refinement by subsites.

Dockings performed by subsites located all poses in the homologue subsites 1–2 and 1–3, unless one GSK812397 structure which was found in subsite 2–3 (Fig. 5). These observations show the existence of minimum 2 different binding sites in CXCR4, corresponding approximately to subsites 1–2 and 2–3, which agrees with the experimental observation that co-crystallized ligands IT1t and CVX15 bind to CXCR4 in different domains of the extracellular binding pocket usually known as minor and major binding sites, respectively. In fact, if we superpose the co-crystallized ligand IT1t into our binding subsites, this ligand fills the binding subsite 3–3, suggesting the existence of a third possible binding subsite overlapped [73] to the subsites 1–2 and 1–3 where we found AMD11070 and GSK812397 poses (Supplementary Fig. 2). This hypothesis agrees with the assumptions proposed by Srivastava et al. [74] and Garcia-Perez et al. [75]. Srivastava et al., which reported recently the crystal structure of hGPCR receptor bound to an allosteric ligand. The authors provide insights into the plausible binding of multiple ligands to the receptor hGPR40 and propose three putative binding sites. One of them should be located near the orthosteric site. In addition, Garcia-Perez et al. in their work docked Maraviroc inhibitor into a homology model of CCR5. They obtained multiple poses that roughly delimit three different sites, yet overlapping, in the TM pocket of the protein, in a similar manner that we obtained docked poses by subsites using 3 partitions.

However, it is important to highlight that there were no appreciable differences between the subsite volumes occupied by all the docked poses of AMD11070 and GSK812397 (Fig. 5). The visual inspection of the docked poses returned by each docking was relatively similar, without observing bias or distinct subsite distribution induced by the dissimilar structure of the two chemical candidates. In fact, this observation may be indicative about similar patterns in their interaction with CXCR4 and similarities in their allosteric mechanism.

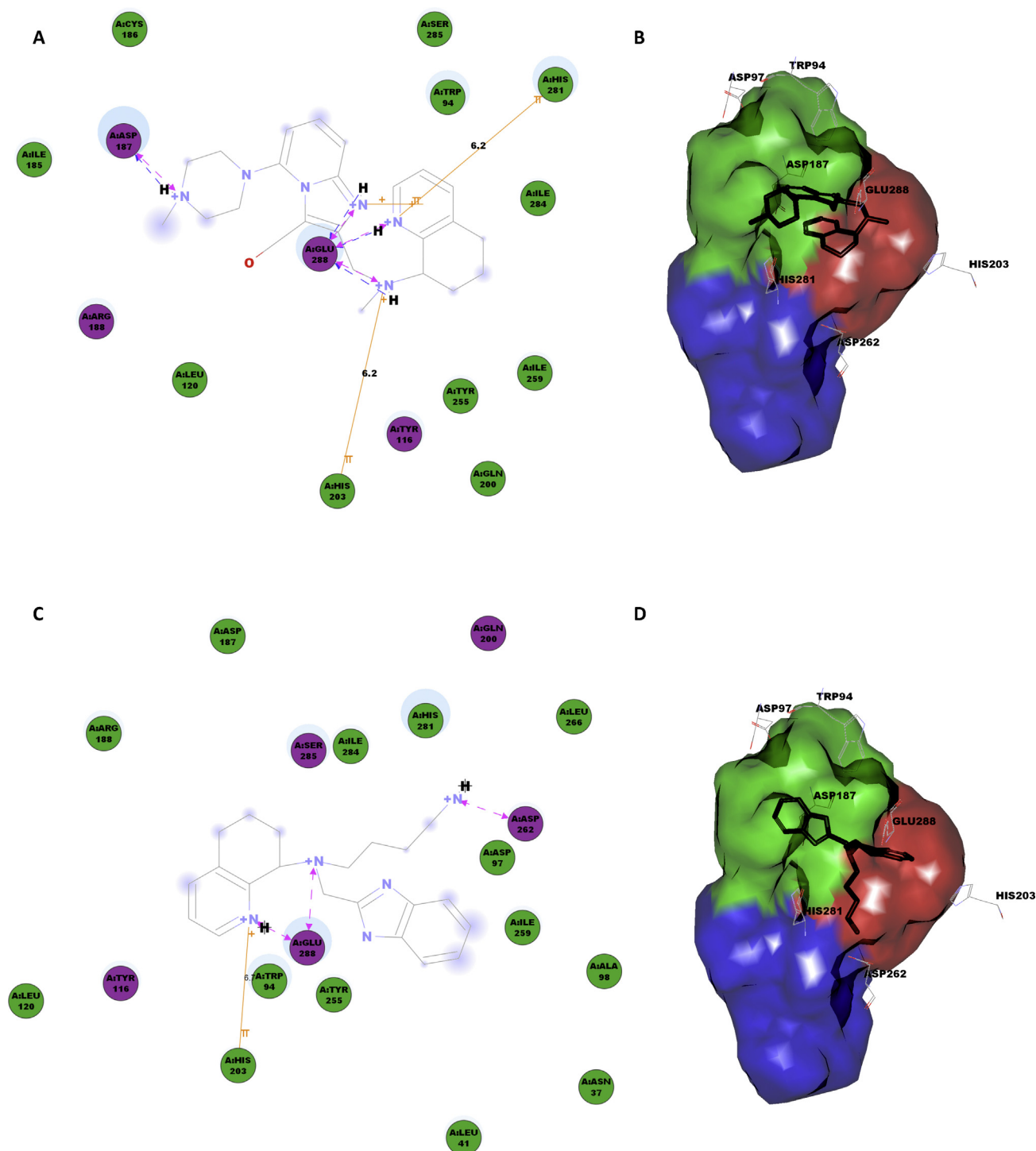


Fig. 9. Protein ligand interaction charts for the best-scored LigandFit-DockScore pose in binding subsite 1–2. The best pose corresponds to structure GSK812397 (A) and AMD11070 (C). The chart shows π interactions in orange lines. Charge interactions are shown in pink dashed lines with arrowheads on both sides. In blue dashed lines with arrow heads toward the electron donor are represented hydrogen bonds. Circles in magenta represent residues involved in hydrogen bond, charge or polar interactions. Circles in green represent residues involved in van der Waals interactions. A blue halo around the residue represents a solvent accessible surface; this halo is proportional to the solvent accessible surface. GSK812397 (B) and AMD11070 (D) best docked pose appeared docked in the same binding subsite 1–2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

We considered that the best AMD11070 and GSK812397 poses found (Fig. 5) should represent better CXCR4–ligand interactions. Indeed, if we analyze these best docked structures we find how they fill the central region of subsite 1–2 (Fig. 9). When protein–ligand interaction diagrams are represented, we observe that residues Asp187, His203, Tyr255, His281 and Glu288 interact directly with GSK812397 and these residues delimit specifically the defined

binding site 1–2 (Supplementary Fig. 2). In this regard, Wong et al. [76] investigated by site-directed mutagenesis key CXCR4 residues that established interactions with AMD11070 and proposed three different binding modes to explain all interactions between AMD11070 and Tyr45, Trp94, Asp97, Asp171, Asp262 and Glu288. Later, Mosi et al. [13] confirmed again those residues as important ones for AMD11070 binding, emphasizing residues

Tyr45, Trp94 and Asp97 among them. All these observations agree with our best model proposed for AMD11070 and GSK812397.

Interestingly, Tian et al. [77] and Zhou et al. [78] concluded in their mutagenesis analyses that several residues of the CXCR4 transmembrane, such as Tyr45, His79, Asp97, Pro163, Trp252, Tyr255, Asp262, Glu288, His294 and Asn298, as well as from the Extracellular loop 2, such as Asp182, Arg183, Tyr184, Asp187, Phe189, Tyr190 and Pro191 interact specifically with HIV-1 without affecting signaling, nor the binding of SDF-1 α . Consequently, it is presupposed their allosteric behavior. The authors noticed specially, how alterations of extracellular loop 2 by mutations had little effect on SDF1- α binding but significantly diminished the activity of HIV-1. Again, this observation agrees with the interactions obtained for GSK812397. Among these listed residues, GSK812397 interacts directly with Asp97, Asp187, Tyr255, Asp262 and Glu288. It is interesting to note how residue Asp262 delimits subsite 1–2 from subsite 2–2 and consequently can establish interactions with ligands that bind in both subsites (Fig. 9). The central position of this residue in the extracellular pocket explains why so many CXCR4 ligands interact with Asp262. However, Asp97 establishes medium-large range interaction with AMD11070. This residue is located at the entrance of subsite 3–3, where cognate ligand IT1t binds, outside from subsite 1–2 (see Fig. 9). Therefore, it is not considered a key residue for the allosteric interactions.

Recently, Thiele et al. [79] studied the binding mode of peptide FC131 with CXCR4 by mutagenesis techniques. The authors confirmed that binding modes of FC131 and CVX15 were similar and located at the same major binding site but FC131 was positioned deeper into the hydrophobic subpocket. The study indicates how mutations in residues Asp187, His203, Tyr255 and His281, among others, did not impair the antagonist potency of FC131. In addition, the author's support that Glu288 affects signaling properties but not the binding of CXCL12, and they propose that residue Glu288 forms indirect interaction with FC131 via water mediated H-bond network. These observations are consistent with the proposal that residues Asp187, His203, Tyr255, His281 and Glu288 may form a different subsite than the major binding subsite, located between the major and minor binding subsites.

Based on these observations and in agreement with the protein–ligand interaction analysis performed, we propose Glu288 as the main residue that establish interactions of allosteric nature with the small CXCR4 antagonists AMD11070 and GSK812397. Also, residue Asp187 and in second term residues His281, Tyr255 and His203, provide short-range hydrophobic interactions and π interactions with both antagonists. This set of residues defines specifically subsites 1–2 and 1–3 (Fig. 9). It is important to note that Asp187 and Glu288, whose interactions are found in both antagonists, also present charge interactions and hydrogen bonding with GSK812397. Charge interactions are considered key and essential in the majority of processes where CXCR4 acts, and therefore residues Asp187 and Glu288 should be considered strategic in CXCR4 allosteric mechanisms.

Overall, we proposed that CXCR4 has minimum two allosteric binding sites in distinct domains attending to the allosteric functionality. Peptiducins bind to an intracellular region while small molecules to an extracellular region. Peptiducins seem to act by blocking the β -hairpin conformational changes in the active state, which would cause α -helices not to be able to make translational movements between them. This seems to be different from what happens with small molecules, which can partially block the active site, allowing the natural CXCL12 ligand to enter but not the HIV. Our binding mode hypothesis for small molecules agrees with Sargand et al. [80] and Scholten et al. [81] works, in which they also propose that, in general, GPCRs such as CXCR4, CXCR2 or CCR4 have two binding pockets in the extracellular side of these receptors. One minor binding pocket formed by residues from helices 1, 2, 3 and 7,

corresponding to our binding subsite 3–3 occupied by IT1t and one major binding pocket formed by residues from helices 3, 4, 5 and 6, corresponding to our binding subsites 1–3 plus 2–3 occupied by peptide CVX15. The overlapping region between subsites 1–3 and 3–3 corresponds to the allosteric subsite occupied by allosteric candidates AMD11070 and GSK812397 (see Supplementary Fig. 2). This agrees with, Scholten et al. work who state that in the case of CXCR4, both pockets are interconnected and ligands bind in both sites, some of them exclusively in the major or in the minor binding pocket, and ligands with partial overlapping binding sites also show non-competitive allosteric behavior.

Hence, the small molecules AMD11070 and GSK812397 could act both as antagonists and allosteric modulators depending on the position occupied when binding the CXCR4 extracellular pocket. The differences in the molecular mechanism of these small molecules could be due to their possible movement from a low affinity site to another interconnected site of higher affinity. For example, between the minor binding pocket (corresponding to our subsite 3–3), which would produce an agonistic effect, and the overlapping region between subsites 1–3 and 3–3, which would produce an allosteric effect. Similarly, Lagane et al. [82] hypothesized for the CCR5 allosteric modulation by Maraviroc (MVC) that “MVC could theoretically pass from the superficial site 1 to the deepest site 3 by a simple translation without necessity of conformational changes for the ligand adaptation to the two sites”. Once the small molecules are bound with high affinity to CXCR4, they may prevent efficient binding of orthosteric ligands to the receptor probably by perturbing the transmembrane–extracellular loops interactions, which is the essential characteristic of allosteric modulation.

5. Conclusion

In the present study, we have analyzed the putative key interactions between chemically different allosteric modulators (peptiducins and small molecules) and the CXCR4 receptor by combining several methods. Peptiducin ATI-2341, a peptidomimetic compound, was selected as the most representative allosteric agonist of CXCR4. AMD11070 and GSK812397, two small organic antagonists of CXCR4 and candidates to be allosteric modulators, were also selected to study their interactions. Given that these are different compounds attending to their chemical nature, structure and mode of action, we proposed different blind docking approximations to locate the main regions where high affinity binding was more favorable. Next, a refining docking process was carried out to establish the putative key allosteric interactions. For peptiducin ATI-2341 two flexible dockings and one MD protocol were performed in order to compare modeling approaches and obtain a consensus result for the key peptiducin–CXCR4 binding interactions. The three methods used confirmed similar interactions between peptiducin ATI-2341 and CXCR4. For the small organic ligands, AMD11070 and GSK812397, docking refinement was performed by subsites. The corresponding protein–ligand interactions were analyzed and compared finding similar interactions and sharing very similar binding positions in the big extracellular pocket of CXCR4.

In conclusion, we propose that CXCR4 has minimum 2 allosteric binding sites in distinct domains attending to the allosteric functionality, such as antagonists to inhibit inflammation and cancer or agonists to enhance angiogenesis in tissue regeneration. One allosteric binding site would be placed in the intracellular region near to ICL1, interacting with ICL2 and in lower mode with ICL3 and the extreme C-terminal. We state that this would be the natural binding site of peptiducin ATI-2341. The allosteric agonist ATI-2341 then would act blocking β -hairpin and impeding the translational movement of the α -helices of CXCR4 in its active state to evolve

to the inactive state. The second allosteric site would be accommodated in the extracellular pocket, in a subsite of the main orthosteric binding pocket, hence being different from the orthosteric minor binding subsite where IT1t binds, but adjacent or overlapped to it. This proposed second binding site would interact with residues that do not affect the binding of the endogenous ligand but in contrast inhibit HIV-CXCR4 binding. Residues Asp187 and Glu 288 that do not affect signaling but inhibit HIV seem to interact in all our proposed models.

Acknowledgements

The authors thank Accelrys, Chemical Computing Group for providing Academic Licenses for Discovery Studio and MOE, respectively. Glide (Schrodinger) was used through the Servei de Disseny de Farmacs (SDF) in the Centre de Supercomputació de Catalunya (CESCA). This work is supported by the Programa Nacional de Biomedicina (Ministerio de Educación y Ciencia, SAF2010-21617-C02-02).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jmgm.2015.05.004>

References

- [1] K.A. Reynolds, R.N. McLaughlin, R. Ranganathan, Hot spots for allosteric regulation on protein surfaces, *Cell* 147 (2011) 1564–1575.
- [2] L.T. May, K. Leach, P.M. Sexton, A. Christopoulos, Allosteric modulation of G protein-coupled receptors, *Annu. Rev. Pharmacol. Toxicol.* 47 (2007) 1–51.
- [3] Q. Cui, M. Karplus, Allostery and cooperativity revisited, *Protein Sci.* 17 (2008) 1295–1307.
- [4] P. Csermely, R. Palotai, R. Nussinov, Induced fit, conformational selection and independent dynamic segments: an extended view of binding events, *Trends Biochem. Sci.* 35 (2010) 539–546.
- [5] R. Nussinov, C.-J. Tsai, Allostery in disease and in drug discovery, *Cell* 153 (2013) 293–305.
- [6] P. Conn, A. Christopoulos, C. Lindsley, Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders, *Nat. Rev. Drug Discov.* 8 (2009) 41–54.
- [7] M. Spedding, J.-P. Pin, R. Hubbard, J. Kiss, P. Keov, P.M. Sexton, et al., Allosteric modulation of G protein-coupled receptors: a pharmacological perspective, *Neuropharmacology* 60 (2011) 24–35.
- [8] C.-I.A. Wang, R.J. Lewis, Emerging opportunities for allosteric modulation of G-protein coupled receptors, *Biochem. Pharmacol.* 85 (2012) 153–162.
- [9] N. London, B. Raveh, O. Schueler-Furman, Druggable protein–protein interactions: from hot spots to hot segments, *Curr. Opin. Chem. Biol.* 17 (2013) 952–959.
- [10] N. Yanamala, J. Klein-Seetharaman, Allosteric modulation of G protein coupled receptors by cytoplasmic, transmembrane and extracellular ligands, *Pharmaceuticals* 3 (2010) 3325–3342.
- [11] J. Ballesteros, J. Ransom, In-target versus off-target allosteric modulators of GPCRs, *Drug Discov. Today Ther. Strateg.* 3 (2007) 445–450.
- [12] A. Sachpatzidis, B.K. Benton, J.P. Manfredi, H. Wang, A. Hamilton, H.G. Dohlman, et al., Identification of allosteric peptide agonists of CXCR4, *J. Biol. Chem.* 278 (2003) 896–907.
- [13] R.M. Mosi, V. Anastassova, J. Cox, M.C. Darks, S.R. Idzan, J. Labrecque, et al., The molecular pharmacology of AMD11070: an orally bioavailable CXCR4 HIV entry inhibitor, *Biochem. Pharmacol.* 83 (2011) 472–479.
- [14] S. Fricker, R. Mosi, V. Anastassova, J. Labrecque, R. Wong, R. Skerlj, et al., The orally bioavailable allosteric CXCR4 HIV-1 entry inhibitor AMD11070, *Retrovirology* 9 (2012) 7.
- [15] S. Boggs, V.I. Elitzin, K. Gudmundsson, M.T. Martin, M.J. Sharp, Kilogram-scale synthesis of the CXCR4 antagonist GSK812397, *Org. Process Res. Dev.* 13 (2009) 781–785.
- [16] S. Jenkinson, M. Thomson, D. McCoy, M. Edelstein, S. Danehower, W. Lawrence, et al., Blockade of X4-Tropic HIV-1 cellular entry by GSK812397, a potent noncompetitive cxcr4 receptor antagonist, *Antimicrob. Agents Chemother.* 54 (2010) 817–824.
- [17] S. Oishi, N. Fujii, Peptide and peptidomimetic ligands for CXC chemokine receptor 4 (CXCR4), *Org. Biomol. Chem.* 10 (2012) 5720–5731.
- [18] P. Dimond, K. Carlson, M. Bouvier, C. Gerard, L. Xu, L. Covic, et al., G protein-coupled receptor modulation with pepducins: moving closer to the clinic, *Ann. N. Y. Acad. Sci.* 1226 (2011) 34–49.
- [19] J. Lomas-Neira, A. Ayala, Pepducins: an effective means to inhibit GPCR signaling by neutrophils, *Trends Immunol.* 26 (2005) 619–621.
- [20] J. Miller, A. Agarwal, L. Devi, Insider access: pepducin symposium explores a new approach to GPCR modulation, *Ann. N. Y. Acad. Sci.* 1180 (2009) E1–E12.
- [21] K.E. Carlson, T.J. McMurry, S.W. Hunt III, Pepducins: lipopeptide allosteric modulators of GPCR signaling, *Drug Discov. Today Technol.* 9 (2012) e33–e39.
- [22] L. Covic, A.L. Gresser, J. Talavera, S. Swift, A. Kuliopulos, Activation and inhibition of G protein-coupled receptors by cell-penetrating membrane-tethered peptides, *Proc. Natl. Acad. Sci. U. S. A.* 99 (2002) 643–648.
- [23] K. O'Callaghan, A. Kuliopulos, L. Covic, Turning receptors on and off with intracellular pepducins: new insights into G-protein-coupled receptor drug development, *J. Biol. Chem.* 287 (2012) 12787–12796.
- [24] J.M. Janz, Y. Ren, R.J. Looby, M.A. Kazmi, P. Sachdev, A. Grunbeck, et al., Direct interaction between an allosteric agonist pepducin and the chemokine receptor CXCR4, *J. Am. Chem. Soc.* 133 (2011) 15878–15881.
- [25] B. Tchernychev, Y. Ren, P. Sachdev, J.M. Janz, L. Haggis, A. O'Shea, et al., Discovery of a CXCR4 agonist pepducin that mobilizes bone marrow hematopoietic cells, *Proc. Natl. Acad. Sci. U. S. A.* 107 (2010) 22255–22259.
- [26] M. Suchanek, A. Radzikowska, C. Thiele, Photo-leucine and photo-methionine allow identification of protein–protein interactions in living cells, *Nat. Methods* 2 (2005) 261–268.
- [27] M. Trakselis, S. Alley, F. Ishmael, Identification and mapping of protein–protein interactions by a combination of cross-linking, cleavage, and proteomics, *Bioconjug. Chem.* 16 (2005) 741–750.
- [28] P. Kask, K. Palo, Fluorescence-intensity distribution analysis and its application in biomolecular detection technology, *Proc. Natl. Acad. Sci. U. S. A.* 96 (1999) 13756–13761.
- [29] F. Bernstein, T. Koetzle, G. Williams, Protein data bank, atomic coordinate and bibliographic entry format description, *J. Mol. Biol.* 112 (1977) 535–542.
- [30] J. Gasteiger, M. Marsili, Iterative partial equalization of orbital electronegativity – a rapid access to atomic charges, *Tetrahedron* 36 (1980) 3219–3228.
- [31] A.D. MacKerell, M. Feig, C.L. Brooks, Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations, *J. Comput. Chem.* 25 (2004) 1400–1415.
- [32] Discovery Studio, Accelrys, Inc., San Diego, 2009.
- [33] A.D. MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, et al., All-atom empirical potential for molecular modeling and dynamics studies of proteins, *J. Phys. Chem. B* 102 (1998) 3586–3616.
- [34] J.M. Planesas, V.I. Pérez-Nueno, J.I. Borrell, J. Teixidó, Impact of the CXCR4 structure on docking-based virtual screening of HIV entry inhibitors, *J. Mol. Graph. Model.* 38 (2012) 123–136.
- [35] H.M. Berman, T. Battistuz, T.N. Bhat, W.F. Bluhm, P.E. Bourne, K. Burkhardt, et al., The protein data bank, *Acta Crystallogr. D: Biol. Crystallogr.* 58 (2002) 899–907.
- [36] B. Wu, E.Y.T. Chien, C.D. Mol, G. Fenalti, W. Liu, V. Katritch, et al., Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists, *Science* 330 (2010) 1066–1071.
- [37] R.M. Bill, P.J.F. Henderson, S. Iwata, E.R.S. Kunji, H. Michel, R. Neutze, et al., Overcoming barriers to membrane protein structure determination, *Nat. Biotechnol.* 29 (2011) 335–340.
- [38] Y. Zou, W.I. Weis, B.K. Kobilka, N-terminal T4 lysozyme fusion facilitates crystallization of a G protein coupled receptor, *PLoS ONE* 7 (2012) e46039.
- [39] S.G.F. Rasmussen, H.J. Choi, D.M. Rosenbaum, T.S. Kobilka, F.S. Thian, P.C. Edwards, et al., Crystal structure of the human β_2 adrenergic G-protein-coupled receptor, *Nature* 450 (2007) 383–387.
- [40] V.-P. Jaakola, M.T. Griffith, M.A. Hanson, V. Cherezov, E.Y.T. Chien, J.R. Lane, et al., The 2.6 Å crystal structure of a human A2A adenosine receptor bound to an antagonist, *Science* 322 (2008) 1211–1217.
- [41] S.H. Park, B.B. Das, F. Casagrande, Y. Tian, H.J. Nothnagel, M. Chu, et al., Structure of the chemokine receptor CXCR1 in phospholipid bilayers, *Nature* 491 (2012) 779–783.
- [42] A. Šali, T. Blundell, Comparative protein modelling by satisfaction of spatial restraints, *J. Mol. Biol.* 234 (1993) 779–815.
- [43] T. Cheng, X. Li, Y. Li, Z. Liu, R. Wang, Comparative assessment of scoring functions on a diverse test set, *J. Chem. Inf. Model.* 49 (2009) 1079–1093.
- [44] R. Chen, Z. Weng, A novel shape complementarity scoring function for protein–protein docking, *Proteins Struct. Funct. Bioinform.* 51 (2003) 397–408.
- [45] R. Chen, L. Li, Z. Weng, ZDOCK: an initial-stage protein-docking algorithm, *Proteins Struct. Funct. Bioinform.* 52 (2003) 80–87.
- [46] H. Hwang, T. Vreven, B.G. Pierce, J.-H. Hung, Z. Weng, Performance of ZDOCK and ZRANK in CAPRI rounds 13–19, *Proteins Struct. Funct. Bioinform.* 78 (2010) 3104–3110.
- [47] T. Vreven, B.G. Pierce, H. Hwang, Z. Weng, Performance of ZDOCK in CAPRI rounds 20–26, *Proteins* 81 (2013) 2175–2182.
- [48] S. Vajda, D. Kozakov, Convergence and combination of methods in protein–protein docking, *Curr. Opin. Struct. Biol.* 19 (2009) 164–170.
- [49] A.M.J.J. Bonvin, Flexible protein–protein docking, *Curr. Opin. Struct. Biol.* 16 (2006) 194–200.
- [50] S. Lyskov, J.J. Gray, The RosettaDock server for local protein–protein docking, *Nucleic Acids Res.* 36 (2008) W233–W238.
- [51] B. Pierce, Z. Weng, A combination of rescoring and refinement significantly improves protein docking performance, *Proteins Struct. Funct. Bioinform.* 72 (2008) 270–279.
- [52] N. London, B. Raveh, E. Cohen, G. Fathi, O. Schueler-Furman, Rosetta FlexPepDock web server – high resolution modeling of peptide–protein interactions, *Nucleic Acids Res.* 39 (2011) W249–W253.

- [53] S.J. De Vries, M. van Dijk, A.M.J.J. Bonvin, The HADDOCK web server for data-driven biomolecular docking, *Nat. Protoc.* 5 (2010) 883–897.
- [54] B. Raveh, N. London, O. Schueler-Furman, Sub-angstrom modeling of complexes between flexible peptides and globular proteins, *Proteins Struct. Funct. Bioinform.* 78 (2010) 2029–2040.
- [55] C. Dominguez, R. Boelens, A.M.J.J. Bonvin, HADDOCK: a protein–protein docking approach based on biochemical or biophysical information, *J. Am. Chem. Soc.* 125 (2003) 1731–1737.
- [56] M. Trellet, A.S.J. Melquiond, A.M.J.J. Bonvin, A unified conformational selection and induced fit approach to protein–peptide docking, *PLOS ONE* 8 (2013) e58769.
- [57] N. Nakajima, J. Higo, A. Kidera, H. Nakamura, Flexible docking of a ligand peptide to a receptor protein by multicanonical molecular dynamics simulation, *Chem. Phys. Lett.* 278 (1997) 297–301.
- [58] F.L. Gervasio, A. Laio, M. Parrinello, Flexible docking in solution using metadynamics, *J. Am. Chem. Soc.* 127 (2005) 2600–2607.
- [59] M. Mangoni, D. Roccatano, A. Di Nola, Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation, *Proteins Struct. Funct. Bioinform.* 35 (1999) 153–162.
- [60] Z. Huang, C.F. Wong, Docking flexible peptide to flexible protein by molecular dynamics using two implicit-solvent models: an evaluation in protein kinase and phosphatase systems, *J. Phys. Chem. B* 113 (2009) 14343–14354.
- [61] O. Dagliyan, E.A. Proctor, K.M. D’Auria, F. Ding, N.V. Dokholyan, Structural and dynamic determinants of protein–peptide recognition, *Structure* 19 (2011) 1837–1845.
- [62] M. Karplus, J.A. McCammon, Molecular dynamics simulations of biomolecules, *Nat. Struct. Biol.* 9 (2002) 646–652.
- [63] A.H. Mahmoud, M.S.A. Elsayed, M. ElHefnawi, Structure-based predictive model for some benzimidazole inhibitors of hepatitis C virus NS5B polymerase, *Med. Chem. Res.* 22 (2012) 1866–1883.
- [64] P.-C. Shih, M.-S. Yang, S.-C. Lin, Y. Ho, J.-C. Hsiao, D.-R. Wang, et al., A turn-like structure KKPE segment mediates the specific binding of viral protein A27 to heparin and heparan sulfate on cell surfaces, *J. Biol. Chem.* 284 (2009) 36535–36546.
- [65] Q.A.T. Le, J.C. Joo, Y.J. Yoo, Y.H. Kim, Development of thermostable *Candida antarctica* lipase B through novel in silico design of disulfide bridge, *Biotechnol. Bioeng.* 109 (2012) 867–876.
- [66] W. Im, M. Lee, C. Brooks, Generalized born model with a simple smoothing function, *J. Comput. Chem.* 24 (2003) 1691–1702.
- [67] J. Chen, C.L. Brooks III, J. Khandogin, Recent advances in implicit solvent-based methods for biomolecular simulations, *Curr. Opin. Struct. Biol.* 18 (2008) 140–148.
- [68] O. Yuzlenko, T. Lazaridis, Membrane protein native state discrimination by implicit membrane models, *J. Comput. Chem.* 34 (2013) 731–738.
- [69] LigPrep, Version 2.4, Schrödinger, LLC, New York, NY, 2010.
- [70] O. Trott, A.J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, *J. Comput. Chem.* 31 (2010) 455–461.
- [71] C.M. Venkatachalam, X. Jiang, T. Oldfield, M. Waldman, LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites, *J. Mol. Graph. Model.* 21 (2003) 289–307.
- [72] D. Zhang, Z.-G. Gao, K. Zhang, E. Kiselev, S. Crane, J. Wang, et al., Two disparate ligand-binding sites in the human P2Y1 receptor, *Nature* 520 (2015) 317–321.
- [73] B.D. Cox, A.R. Prosser, B.M. Katzman, A.A. Alcaraz, D.C. Liotta, L.J. Wilson, et al., Anti-HIV small-molecule binding in the peptide subpocket of the CXCR4:CVX15 crystal structure, *ChemBioChem* 15 (2014) 1614–1620.
- [74] A. Srivastava, J. Yano, Y. Hirozane, G. Kefala, F. Gruswitz, G. Snell, et al., High-resolution structure of the human GPR40 receptor bound to allosteric agonist TAK-875, *Nature* 513 (2014) 124–127.
- [75] J. Garcia-Perez, P. Rueda, J. Alcami, Allosteric model of maraviroc binding to CC chemokine receptor 5 (CCR5), *J. Biol. Chem.* 286 (2011) 33409–33421.
- [76] R.S.Y. Wong, V. Bodart, M. Metz, J. Labrecque, G. Bridger, S.P. Fricker, Comparison of the potential multiple binding modes of bicyclam, monocylam, and noncylam small-molecule CXC chemokine receptor 4 inhibitors, *Mol. Pharmacol.* 74 (2008) 1485–1495.
- [77] S. Tian, W. Choi, D. Liu, J. Pesavento, Distinct functional sites for human immunodeficiency virus type 1 and stromal cell-derived factor 1 α on CXCR4 transmembrane helical domains, *J. Virol.* 79 (2005) 12667–12673.
- [78] N. Zhou, Z. Luo, J. Luo, D. Liu, J.W. Hall, R.J. Pomerantz, et al., Structural and functional characterization of human CXCR4 as a chemokine receptor and HIV-1 co-receptor by mutagenesis and molecular modeling studies, *J. Biol. Chem.* 276 (2001) 42826–42833.
- [79] S. Thiele, J. Mungalpara, Determination of the binding mode for the cyclopentapeptide CXCR4 antagonist FC131 using a dual approach of ligand modifications and receptor mutagenesis, *Br. J. Pharmacol.* 171 (2014) 5313–5329.
- [80] J.-S. Surgand, J. Rodrigo, E. Kellenberger, D. Rognan, A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors, *Proteins* 62 (2006) 509–538.
- [81] D.J. Scholten, M. Canals, D. Maussang, L. Roumen, M.J. Smit, M. Wijtmans, et al., Pharmacological modulation of chemokine receptor function, *Br. J. Pharmacol.* 165 (2012) 1617–1643.
- [82] B. Lagane, J. Garcia-Perez, E. Kellenberger, Modeling the allosteric modulation of CCR5 function by Maraviroc, *Drug Discov. Today Technol.* 10 (2013) e297–e305.

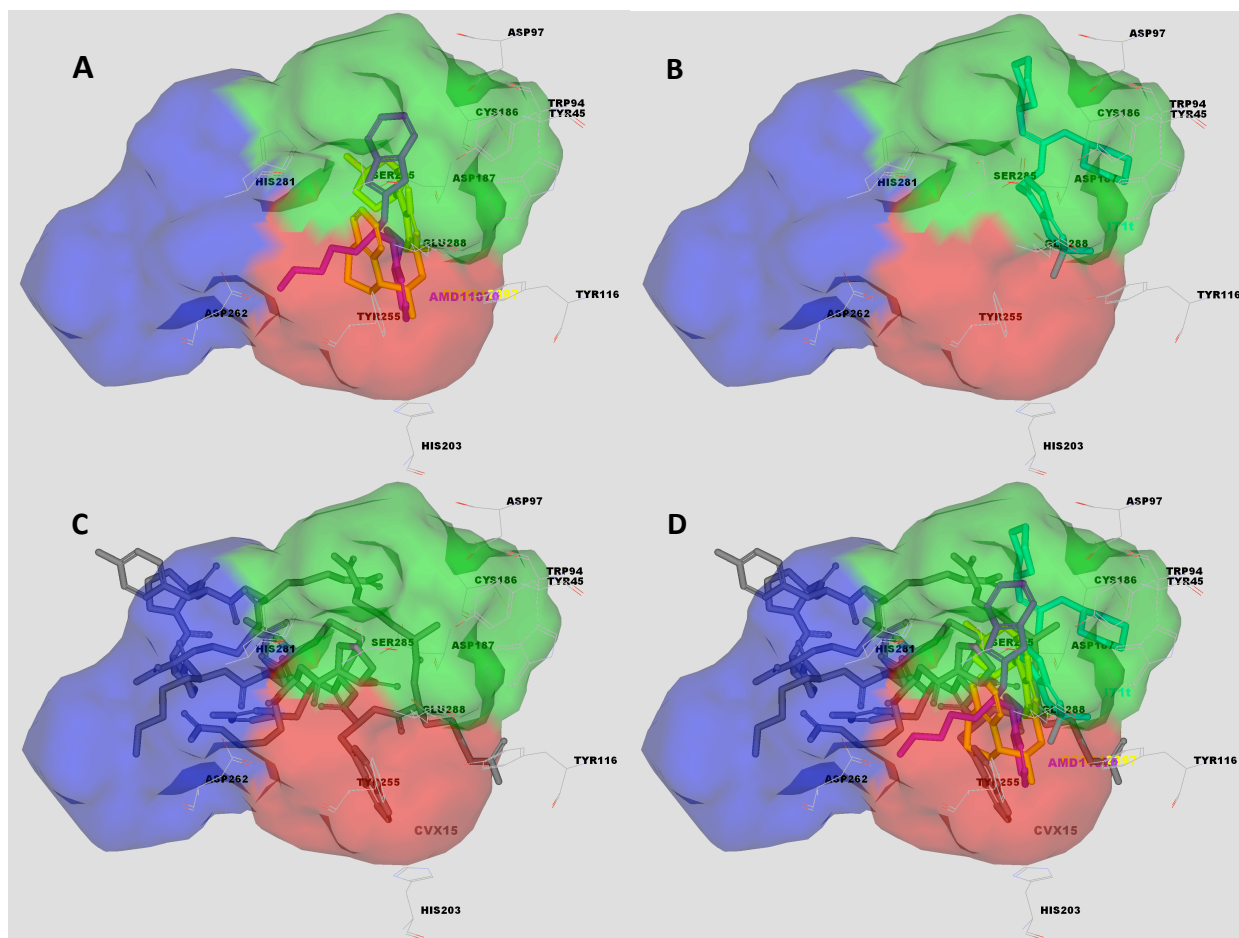


Figure 2 Distribution of the best docked poses of AMD11070 and GSK812397 located in subsite 1-2 (A) in comparison with the co-crystallized poses of antagonist IT1t in the minor subsite 3-3 (B) and with co-crystallized peptide CVX15 in the major binding site (C) and the superposition of all binding poses (D). Peptide CVX15 binds in the central part of the overall binding site, mainly in subsite 2-3 (blue), in part of subsite 1-3 (red) and in the N-terminal domain out of the definition of the binding site. For further details see the video “video_conformations_CXCR4.avi” in Supplementary Information.

APENDICES

- A1.** Póster presentado en **10th International Conference of Computational Methods in Sciences and Engineering-ICCMSE** (04-07 Abril 2014, Atenas, Grecia).
Polypharmacology within CXCR4: Multiple binding sites and allosteric behavior. Planesas, J.M.;Pérez-Nueno, V. I.;Borrell, J. I.; Teixidó, J.(2014) *AIP Conf.Proc.*,**1618**,1036.
- A2.** Abstract comunicación oral presentada en **10th International Conference of Computational Methods in Sciences and Engineering-ICCMSE** (04-07 Abril 2014, Atenas, Grecia.). Artículo: Polypharmacology within CXCR4: Multiple binding sites and allosteric behavior. Planesas, J.M.;Pérez-Nueno, V. I.;Borrell, J. I.; Teixidó, J.(2014) *AIP Conf.Proc.*,**1618**,1036.

Polypharmacology within CXCR4: Multiple Binding Sites and Allosteric Behavior

Jesús M. Planesas^a, Violeta I. Pérez-Nueno^{a,b}, José I. Borrell^a and Jordi Teixidó^a

^aGrup d'Enginyeria Molecular (GEM), IQS, Universitat Ramon Llull, Via Augusta 390, 08017 Barcelona (Spain).

^bHarmonic Pharma, 615 rue du Jardin Botanique, 54600 Villers lès Nancy, France.

E-mail: jesusplanesas@iqs.url.edu, pereznueno@harmonicpharma.com

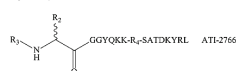
Introduction

CXCR4 is a promiscuous receptor, which binds multiple diverse ligands. As usual in promiscuous proteins, CXCR4 has a large binding site, with multiple subsites, and high flexibility. Hence, it is not surprising that it is involved in the phenomenon of allosteric modulation. However, incomplete knowledge of allosteric ligand-binding sites has hampered an in-depth molecular understanding of how these inhibitors work. For example, it is known that lipidated fragments of intracellular GPCR loops, so called pepducins, such as pepducin ATI-2341, modulate CXCR4 activity using an agonist allosteric mechanism. Nevertheless, there are also examples of small organic molecules, such as AMD11070 and GSK812397, which may act as antagonist allosteric modulators. Here, we give new insights into this issue by proposing the binding interactions between the CXCR4 receptor and the above-mentioned allosteric modulators. We propose that CXCR4 has minimum two topographically different allosteric binding sites. One allosteric site would be in the intracellular loop 1 (ICL1) where pepducin ATI-2341 would bind to CXCR4, and the second one, in the extracellular side of CXCR4 in a subsite into the main orthosteric binding pocket, delimited by extracellular loops n° 1, 2, and the N-terminal end, where antagonists AMD11070 and GSK812397 would bind. Prediction of allosteric interactions between CXCR4 and pepducin ATI-2341 were studied first by rotational blind docking to determine the main binding region and a subsequent refinement of the best pose was performed using flexible docking methods and molecular dynamics. For the antagonists AMD11070 and GSK812397, the entire CXCR4 protein surface was explored by blind docking to define the binding region. A second docking analysis by subsites of the identified binding region was performed to refine the allosteric interactions. Finally, we identified the binding residues that appear to be essential for CXCR4 (agonists and antagonists) allosteric modulators.

Allosteric CXCR4 Modulators

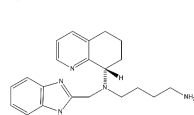
① Pepducins ATI-2341 & ATI2766

R₁-MGVQKRLKSMYDKYRL ATI-2341

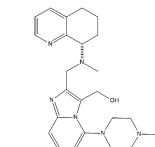


R₂: Palmitic acid
R₁: 2-(Undecyloxy)ethane
R₄: 5(6) Carboxytetramethylrhodamine (TAMRA)
R₄: L-Phots-Leucine

② AMD11070

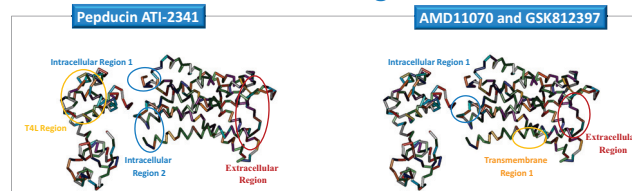


③ GSK812397



Janz *et al.* [1] reported pepducins ATI-2341 and ATI2766 as agonists allosteric modulators due to their non competitive behavior against the endogenous CXCR4 ligand, SDF-1. Small organics antagonists AMD11070 [2] and GSK812397 [3] were considered potential allosteric modulators. AMD11070 is a selective noncompetitive inhibitor of isoforms SDF-1 α and SDF-1 β and GSK812397 may be able to bind CXCR4 in a distinct and separate site than HIV.

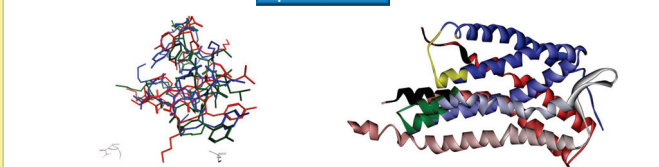
Blind Dockings



- Pepducin ATI-2341: Rotational and translational blind docking performed by Zdock shows 4 main regions with high probability of interactions. Intracellular Region 1 is the smallest region with the greatest number of best docked poses.
- AMD11070 & GSK812397: Blind docking performed by AutoDock Vina for all isomers (protomers, tautomers, etc.) returned 3 main binding regions. The majority of poses were distributed in the Extracellular Region.

Refining Allosteric Interactions

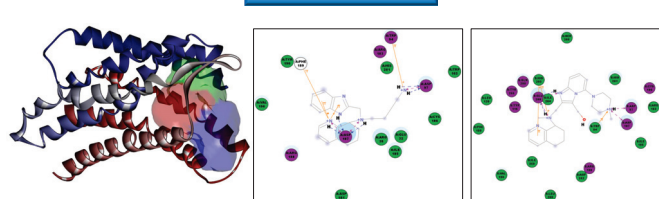
Pepducin ATI-2341



ATI-2341	Met83	Lys87	Lys88	Leu89	Arg10	Ser71	Met72	Thr73	Asp74	Lys75	Arg77
FlexPepDock (DOU)											
FlexPepDock (DOE)											
Haddock (DOU)											
Haddock (DOE)											
MD (DOU)											
MD (DOE)											

- The best ZDock pose of Intracellular Region 1 (yellow) is used for flexible dockings (FlexPepDock and HADDOCK) and a Molecular Dynamic simulation. Similar conformations and interactions are returned by the different methods.

AMD11070 and GSK812397



- A binding site of 1520 Å³ is defined to cover the total extracellular pocket. This binding site is divided in 3 binding subsites to dock all AMD11070 and GSK812397 isomers by subsites using LigandFit docking function[4].
- Subsite 1-2 locates the majority of binding poses and the best poses.
- AMD11070 and GSK812397 protein ligand interactions of the best docked poses of Subsite 1-2 show similar key interactions for residues Trp94, Asp97, Asp187.

Conclusions

- Putative key interactions between allosteric modulators (pepducins and small molecules) and CXCR4 are analyzed combining several methods.
- The **intracellular binding site** proposed for pepducin ATI-2341 shares a common allosteric binding site with CXCR2 and CCR4 [5].
- AMD11070 and GSK812397 are proposed to bind in **extracellular side** of CXCR4 into one of the two interconnected subpockets that form the main pocket.
- CXCR4 is a promiscuous receptor, which binds in different ways to multiple ligands (agonists, antagonists, and allosteric modulators). Different existent allosteric sites might be related to different allosteric mechanisms leading to the same or somewhat different intracellular signal.

References

- [1] J. M. Janz, Y. Ren, R. J. Looby, M. A. Kazmi, P. Sachdev, A. Grunbeck, L. Haggis, D. Chinnapan, A. Y. Lin, C. Seibert, *Journal of the American Chemical Society* 40, 15878–15881 (2011).
- [2] R. M. Mosi, V. Anastassova, J. Cox, M. C. Darks, S. R. Idzan, J. Labrecque, G. Lau, K. L. Nelson, K. Patel, Z. Santucci, *Biochemical pharmacology* 83, 472–479 (2011).
- [3] S. Jenkinson, M. Thomson, D. McCoy, M. Edelstein, S. Danehower, W. Lawrence, P. Wheelan, A. Spaltenstein, K. Gudmundsson, *Antimicrobial Agents and Chemotherapy* 54, 817 (2010).
- [4] J. M. Planesas, V. I. Pérez-Nueno, J. I. Borrell, J. Teixidó, *Journal of Molecular Graphics and Modelling* 38, 123–136 (2012).
- [5] D. J. Scholten, M. Canals, D. Maussang, L. Roumen, M. J. Smit, M. Wijnmans, C. de Graaf, H. F. Vischer, R. Leurs, *British journal of pharmacology* 165, 1617–1643 (2012).

Acknowledgments

This work has been partially supported by the Spanish Ministerio de Educación y Ciencia project SAF2010-21617-C02-02.

Polypharmacology within CXCR4: Multiple Binding Sites and Allosteric Behavior

Jesús M. Planesas^a, Violeta I. Pérez-Nueno^{a,b}, José I. Borrell^a and Jordi Teixidó^a

^a Grup d'Enginyeria Molecular, Institut Químic de Sarrià (IQS), Universitat Ramon Llull, Barcelona, Spain.

^b Harmonic Pharma, 615 rue du Jardin Botanique, 54600 Villers lès Nancy, France.

Abstract. CXCR4 is a promiscuous receptor, which binds multiple diverse ligands. As usual in promiscuous proteins, CXCR4 has a large binding site, with multiple subsites, and high flexibility. Hence, it is not surprising that it is involved in the phenomenon of allosteric modulation. However, incomplete knowledge of allosteric ligand-binding sites has hampered an in-depth molecular understanding of how these inhibitors work. For example, it is known that lipidated fragments of intracellular GPCR loops, so called pepducins, such as pepducin ATI-2341, modulate CXCR4 activity using an agonist allosteric mechanism. Nevertheless, there are also examples of small organic molecules, such as AMD11070 and GSK812397, which may act as antagonist allosteric modulators. Here, we give new insights into this issue by proposing the binding interactions between the CXCR4 receptor and the above-mentioned allosteric modulators. We propose that CXCR4 has minimum two topographically different allosteric binding sites. One allosteric site would be in the intracellular loop 1 (ICL1) where pepducin ATI-2341 would bind to CXCR4, and the second one, in the extracellular side of CXCR4 in a subsite into the main orthosteric binding pocket, delimited by extracellular loops n° 1, 2, and the N-terminal end, where antagonists AMD11070 and GSK812397 would bind. Prediction of allosteric interactions between CXCR4 and pepducin ATI-2341 were studied first by rotational blind docking to determine the main binding region and a subsequent refinement of the best pose was performed using flexible docking methods and molecular dynamics. For the antagonists AMD11070 and GSK812397, the entire CXCR4 protein surface was explored by blind docking to define the binding region. A second docking analysis by subsites of the identified binding region was performed to refine the allosteric interactions. Finally, we identified the binding residues that appear to be essential for CXCR4 (agonists and antagonists) allosteric modulators.

Keywords: Polypharmacology; allostery, CXCR4, binding subsites.

PACS: 87

INTRODUCTION

The study of allostery process without previous knowledge about key interactions becomes an important challenge in drug discovery. For the receptor CXCR4, pepducin ATI-2341 and its photo-derivate ATI-2766 are peptidomimetic derivatives whose modulation as allosteric agonists has been shown[1]. In addition, the small organic antagonists AMD11070 and GSK812397 revealed similar functionalities than allosteric modulators in their pharmacological characterization in CXCR4[2][3].

In a previous work[4] where we virtually screened a set of ligands and decoys, we proposed that small organic molecules such as AMD11070 and GSK812397 bind in the extracellular side of the receptor mainly in two different subsites and probably showing different modes of action.

In the present work we investigated and proposed potential allosteric binding sites by combining different approximations adapted to the chemical typology of the ligands ATI-2341, AMD11070 and GSK812397.

METHODS

The methodology proposed consisted, first to blind dock all ligands all over the protein structure to determine the region with higher allosteric interaction probability. Second different flexible docking methods and molecular dynamics were performed to refine binding poses and interactions. For the blind dockings, all CXCR4 PDBs published were used.

Peptidomimetic-Protein Interactions

A rotational and translational blind docking was performed to determine the CXCR4 region with higher probability to establish interactions with pepducin ATI-2341. ZDock and its ZDock Score function were chosen to dock and rank all different conformations and clusters found.

Two different flexible docking protocols (FlexPepDock and HADDOCK) and a Molecular Dynamics (MD) simulation were performed to refine and compare the allosteric interactions. These flexible docking methods have been successfully tested to dock protein-protein or peptide-protein complexes.

Small ligands-Protein Interactions

AutoDock blind docking function was used to search for the CXCR4 region where antagonists AMD11070 and GSK812397 bind. The region with the greatest number of lowest energy interactions proposed by AutoDock was subsequently divided into 3 different subsites to refine allosteric interactions. LigandFit docking function re-docked all isomers classifying them in subsites and ranked them by DockScore function.

RESULTS

Peptidomimetic-Protein Interactions

Blind dockings performed in all PDBs showed the intracellular region of CXCR4 as the main domain where ATI-2341 established the most important interactions. These interactions were selected as starting contacts to perform FlexPepDock and HADDOCK flexible dockings as well as MD simulation.

Peptide-protein flexible docking and MD results returned similar interactions between ATI-2341 and CXCR4. These interactions were located near the intracellular loop 1 (ICL1), principally within residues Met63 to Leu78. The sequence of residues Met63 to Leu78 matches the ATI-2341 residues sequence.

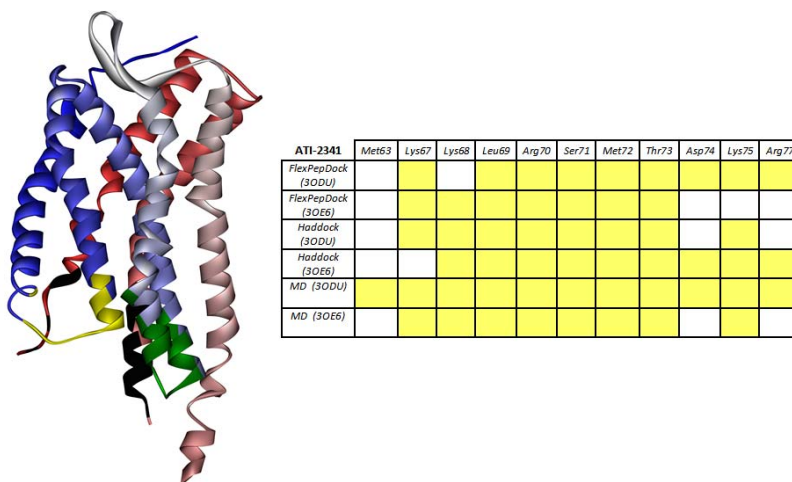


FIGURE 1. Interactions established between pepducin ATI-2341 and CXCR4 classified by the computational methods used and by PDB. Fragments of helices colored in yellow; green and black correspond to residues that interact with ATI-2341 according to FlexPepDock, HADDOCK or MD results. The table shows interactions in ICL1 with residues Met63 to Arg77.

Small ligands-Protein Interactions

Blind dockings performed for all AMD11070 and GSK812397 isomers returned the majority of poses located in the extracellular pocket of CXCR4. This pocket was divided in subsites to redock ligands by LigandFit. The key interactions found were similar between AMD11070 and GSK812397.

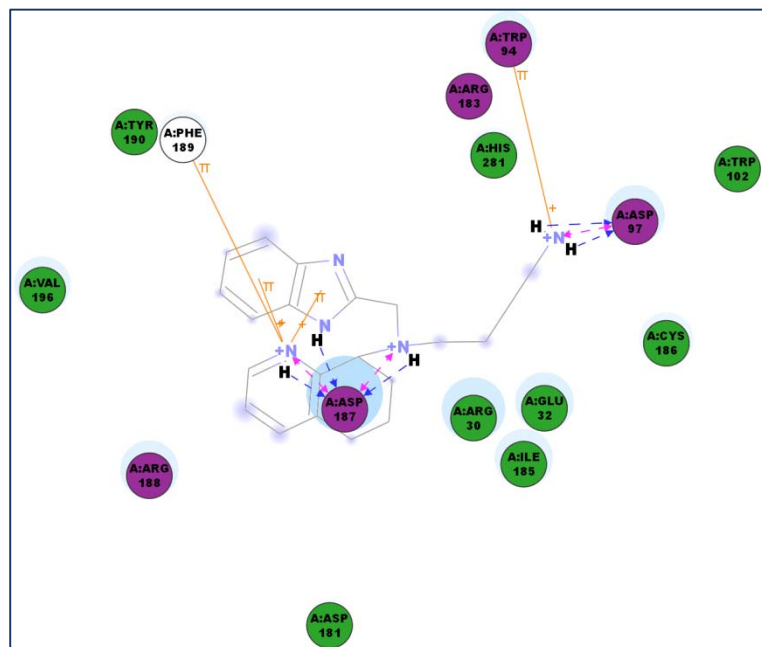


FIGURE 2. Protein-ligand interaction chart for the best scored LigandFit-DockScore pose in binding subsite 1-2, which corresponds to structure AMD11070 in PDB 3ODU. The chart shows π interactions in orange lines with residues Trp94, Phe189, and 2 intermolecular π interactions. Charge interactions are shown in pink dashed lines with arrowheads on both sides. Charge interactions are present in Asp97 and Asp187. In blue dashed lines with arrowheads towards the electron donor are represented hydrogen bonds with amino acid side chains. It shows hydrogen bonds with Asp97 and Asp187. Circles in magenta represent residues involved in hydrogen bond, charge, or polar interactions. Circles in green represent residues involved in van der Waals interactions. A blue halo around the residue represents a solvent accessible surface; this halo is proportional to the solvent accessible surface. Residues Asp97 and Asp187 present solvent accessible surface interaction.

CONCLUSIONS

In the present study, we have analyzed in-depth the putative key interactions between chemically different allosteric modulators (pepducins and small molecules) and the CXCR4 receptor by combining several methods. The allosteric binding sites proposed herein are in accordance with *Scholten et al.*[5] who indicated in a recent review that GPCRs, such as CXCR2 and CCR4, share a common intracellular allosteric binding site as we found for pepducin ATI-2341. Moreover, they also indicated that two interconnected binding subpockets, one major and one minor, can be distinguished in the extracellular side of GPCRs, as well as in CXCR4. This observation agrees with our binding proposal for AMD11070 and GSK812397 and probably their allosteric mode of action.

Overall, we show that CXCR4 is a promiscuous receptor, which binds multiple ligands, such as agonists, antagonists, and allosteric modulators in different ways. The different existent allosteric sites (*i.e.* intracellular region found for pepducins and extracellular region found for small molecules) might be related or not to different allosteric mechanisms that will finally lead to the same or somewhat different intracellular signal.

ACKNOWLEDGMENTS

This work is supported by the Programa Nacional de Biomedicina (Ministerio de Educación y Ciencia, SAF2010-21617-C02-02).

REFERENCES

1. J. M. Janz, Y. Ren, R. J. Looby, M. A. Kazmi, P. Sachdev, A. Grunbeck, L. Haggis, D. Chinnapen, A. Y. Lin, C. Seibert, *Journal of the American Chemical Society* **40**, 15878–15881 (2011).
2. R. M. Mosi, V. Anastassova, J. Cox, M. C. Darkes, S. R. Idzan, J. Labrecque, G. Lau, K. L. Nelson, K. Patel, Z. Santucci, *Biochemical pharmacology* **83**, 472–479 (2011).
3. S. Jenkinson, M. Thomson, D. McCoy, M. Edelstein, S. Danehower, W. Lawrence, P. Wheelan, A. Spaltenstein, K. Gudmundsson, *Antimicrobial Agents and Chemotherapy* **54**, 817 (2010).
4. J. M. Planesas, V. I. Pérez-Nuño, J. I. Borrell, J. Teixidó, *Journal of Molecular Graphics and Modelling* **38**, 123–136 (2012).
5. D. J. Scholten, M. Canals, D. Maussang, L. Roumen, M. J. Smit, M. Wijtmans, C. de Graaf, H. F. Vischer, R. Leurs, *British journal of pharmacology* **165**, 1617–1643 (2012).

