Universitat Ramon Llull

TESIS DOCTORAL

Título	**Herramientas de cribado virtual aplicadas a inhibidores de tirosina quinasas. Contribución al desarrollo del programa PRALINS para el diseño de quimiotecas combinatorias.**
Realizada por	Obdulia Rabal Gracia
en el Centro	Escola Tècnica Superior IQS
y en el Departamento	Química Orgánica y Bioquímica
Dirigida por	Dr. Jordi Teixidó i Closa

C. Claravall, 1-3
08022 Barcelona
Tel. 936 022 200
Fax 936 022 249
E-mail: urlsc@sec.url.es
www.url.es

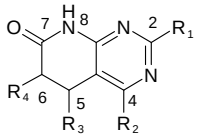
Capítulo 3.

Diseño de una quimioteca de análogos de pirido[2,3-*d*]pirimidinas

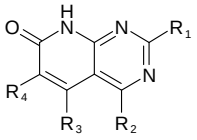
El mencionado proyecto "Diseño y síntesis combinatoria de inhibidores potenciales de tirosina quinasas" (BQU20003-07852), en el que se enmarca la presente tesis, se centra inicialmente en el desarrollo de sistemas pirido[2,3-*d*]pirimidínicos **50** (Figura 3.1). La motivación inicial de la elección de este *scaffold* surge de la confluencia de dos factores:

i) la similitud estructural de estos sistemas **50** con los inhibidores conocidos de estructura pirido[2,3-*d*]pirimidina, **51**, presentados en el capítulo 2, que únicamente difieren en un doble enlace.

ii) la amplia experiencia del Laboratorio de Síntesis del Grupo de Ingeniería Molecular del Instituto Químico de Sarriá en la obtención de estructuras **50**.



50



51

Figura 3.1. Estructuras pirido[2,3-*d*]pirimidínicas.

En este capítulo se aborda uno de los objetivos iniciales del presente trabajo: el diseño de una quimioteca combinatoria de pirido[2,3-*d*]pirimidinas **50,** a partir de la cual se selecciona una subquimioteca combinatoria (*full array*) de compuestos según criterios de diversidad. Dicha subquimioteca, se encuentra sintetizada[361] y su potencial actividad inhibitoria está siendo actualmente testada frente a diversas RTKs.

3.1. Estrategia sintética para la obtención de pirido[2,3-*d*]pirimidinas

Se contempla en una primera aproximación inicial la derivatización de las posiciones C-2, C-4 y C-6 del *scaffold* **50**:

- La posición C-2 se derivatiza con un grupo amino primario sustituido, siguiendo las conclusiones del modelo farmacofórico de interacción propuesto[346], según el cual dicho grupo amino primario participa como dador de puente de hidrógeno en una interacción esencial para la actividad biológica de estos compuestos.

- La posición C-6, implicada en interacciones principalmente hidrofóbicas que contribuyen a la especificidad por un RTK, se derivatiza considerando un amplio espectro de diversidad estructural.

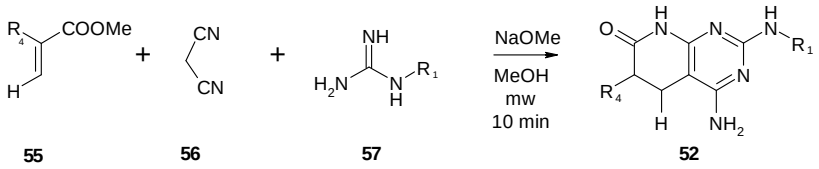
- La posición C-4, para la que no se han concluido implicaciones SAR, se derivatiza según tres posibles sustituciones: 4-amino (**52)**, 4-oxo (**53**) y 4-hidro (**54**); dado que desde un punto de vista sintético resultan inicialmente fácilmente accesibles.

Por otra parte, la posición C-7 ha sido derivatizada en estudios SAR mediante un grupo ceto[342,347] o bien mediante grupos N'-alquilurea[344-345], que si bien no contribuyen a la selectividad por una quinasa, incrementan la potencia de estos inhibidores. En el diseño de la quimioteca, únicamente se considera la derivatización 7-oxo, ya que inicialmente es más accesible según la metodología sintética desarrollada (véase abajo).

En un estudio SAR[347], en el que se ha analizado el efecto de la sustitución del hidrógeno de la posición N-8 por diversos grupos alquilo, se ha encontrado que la sustitución del H por etilo únicamente influye en el incremento de la selectividad frente a PDGFR. Así, inicialmente se construye la quimioteca considerando únicamente estructuras con N-8 sustituidas por un hidrógeno, ya que son las fácilmente asequibles por dicha metodología sintética.

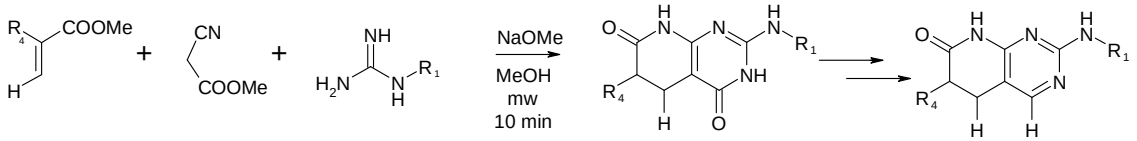
Finalmente, para la posición C-5 tampoco se contempla en esta primera quimioteca su derivatización, manteniéndose R^3=H, análogamente a las series de inhibidores tipo **51** descritas[341-345, 347].

En el laboratorio de síntesis se ha desarrollado una nueva metodología sintética en fase líquida fácilmente automatizable[362-363], basada en una reacción multicomponente asistida por microondas, para la síntesis de pirido[2,3-*d*]pirimidinas 4-amino (**52**) (Figura 3.2) y 4-oxo sustituidas (**53**) (Figura 3.3).



55 **56** **57** **52**

Figura 3.2. Ruta de obtención de sistemas 4-aminopirido[2,3-*d*]pirimidínicos **52**.



55 **58** **57** **53** **54**

Figura 3.3. Ruta de obtención de sistemas 4-oxopirido[2,3-*d*]pirimidínicos **53** y 4-hidropirido[2,3-*d*]pirimidínicos **54**.

En ambas figuras se observa que el grupo R^4 proviene de un éster α,β-insaturado **55** y el sustituyente R^1 es incorporado por una guanidina sustituida **57**.

3.2. Búsqueda de reactivos comerciales

Se realiza una búsqueda estructural con el programa SciFinder®Scholar 2003[364] para seleccionar aquellos monómeros de síntesis o *building blocks* correspondiente a ésteres α,β-insaturados (**55**) y guanidinas (**57**). Se utiliza el programa *ViewerPro*[365] para la visualización de moléculas, conectable con el programa SciFinder®Scholar 2003.

En todos los casos, de las posibilidades de búsqueda de subestructura que ofrece SciFinder® Scholar, listadas en la Tabla 3.1, se escogen los cuatro tipos de sustitución (cursiva) que engloban a los restantes.

Tabla 3.1. Posibilidades de búsqueda de subestructuras disponibles en SciFinder®Scholar 2003.

Tipo de sustituyente	Abreviatura
Cualquier halógeno	X
Cualquier metal	M
Cualquier átomo excepto H	A
Cualquier átomo excepto C o H	*Q*
Cualquier cadena alquílica	*Ak*
Cualquier ciclo	Cy
Cualquier carbociclo	*Cb*
Cualquier heterociclo	*Hy*

En determinados casos, dada la enorme cantidad de posibles candidatos identificados por esta búsqueda, se aplican criterios restrictivos como son:

- La exclusión de todos aquellos compuestos que son isótopos, contienen metales, son mezclas o bien son polímeros.

- La aplicación de un criterio de corte por peso molecular (MW). Dado que se trata de un muestreo previo, el valor del límite de peso molecular se fija en un valor relativamente elevado respecto al valor estimado que aportará cada *building block* de forma que cumpla la regla de Lipinski relativa al peso molecular. Así, en principio se adopta un límite de 500 g/mol, disminuyéndose hasta 350 g/mol en algún caso particular (indicado donde corresponda), debido al elevadísimo número de compuestos identificados.

Posteriormente, se filtran todos los compuestos en función de su disponibilidad comercial. Aunque este paso es automático, se eliminan individualmente todas aquellas sustancias que, pese a ser comerciales, lo son en cantidades muy reducidas (del orden de miligramos) o bien pertenecen a empresas no accesibles (sin representación en Europa, con catálogos comerciales no accesibles…).

Se destaca el hecho de que a veces, la búsqueda por subestructura genera erróneamente resultados que no se ajustan exactamente a la estructura propuesta inicialmente (incluso utilizando la opción *lock out further substitution at position*) y que no son compatibles con la estrategia sintética propuesta, por lo que se descartan.

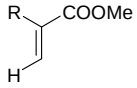
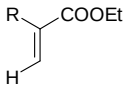
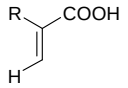
3.2.1. Selección y filtrado de ésteres α,β-insaturados

3.2.1.1. Búsqueda de ésteres α,β-insaturados directamente comerciales

En la Tabla 3.2 se detallan las tres estructuras generales contempladas en la búsqueda junto con el número de *building blocks* inicialmente identificados, su posterior filtraje y el número final seleccionado. R corresponde a cada una de las cuatro posibles sustituciones especificadas en la Tabla 3.1.

En total, el número de compuestos comerciales que rinden el *building block* éster α,β-insaturado **55** es de 42 (15+11+16).

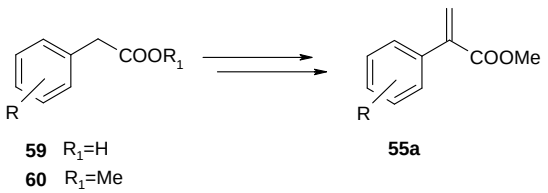
Tabla 3.2. Proceso de selección primaria de monómeros sintéticos directamente comerciales para el *building block* éster α,β-insaturado **55**.

Ésteres α,β-insaturados metílicos	
	R = Q (sin límite de MW Inicial) 1552 total iniciales→42 comerciales→7 cumplen estructura→**5** asequibles
	R = Ak (obliga a MW<500) 7191 total iniciales→23 comerciales→12 cumplen estructura→**8** asequibles
5 + 8 + 1 = 14 Se añade el caso R=H, alcanzándose un total de **15**	R = Cb (obliga a MW<500) 991 total iniciales→3 comerciales→0 cumplen estructura→**0** asequibles
	R = Hy (obliga a MW<500) 715 total iniciales→7 comerciales→1 cumplen estructura→**1** asequible
Ésteres α,β-insaturados etílicos	
	R = Q (sin límite de MW Inicial) 1175 total iniciales→82 comerciales→3 cumplen estructura→**2** asequibles
	R = Ak (obliga a MW<500) 3110 total iniciales→16 comerciales→13 cumplen estructura→**8** asequibles
	R = Cb (sin límite de MW Inicial) 457 total iniciales→1 comercial→1 cumplen estructura→**1** asequible
2 + 8 + 1 = **11**	R = Hy (sin límite de MW Inicial) 467 total iniciales→9 comerciales→0 cumplen estructura→**0** asequibles
Ácidos α,β-insaturados	
	R = Q (obliga a MW<500) 1192 total iniciales→87 comerciales→6 cumplen estructura→**6** asequibles
	R = Ak (obliga a MW<500) 2486 total iniciales→93 comerciales→9 cumplen estructura→**9** asequibles
6 + 9 +1 = **16**	R = Cy (obliga a MW<500) 2164 total iniciales→52 comerciales→1 cumple estructura→**1** asequible

"cumplen estructura" se refiere a si la estructura propuesta se ajusta a la requerida

3.2.1.2. Búsqueda de ésteres α,β-insaturados sintetizables

Puesto que el número de estructuras asequibles es reducido, se decide ampliar la búsqueda con aquellas estructuras que podrían ser precursores sintéticos de ésteres α,β-insaturados. Se plantea el esquema sintético mostrado en la Figura 3.4, a partir de precursores ácidos fenilacéticos **59** o los fenilacetatos **60**.



59 R_1=H
60 R_1=Me

55a

Figura 3.4. Vía de obtención de ésteres α,β-insaturados.

En esta primera quimioteca inicial del proyecto, se consideran únicamente aquellos que aportan un resto R^4 aromático (Figura 3.1), ya que como se ha comentado, los estudios SAR realizados sobre la derivatización de la posición C-6 de **51** han mostrado que esta posición se encuentra implicada principalmente en interacciones de tipo hidrofóbico. Por ello, las búsquedas por subestructura se realizan para los casos de sustituyentes R= Cb ó R= Hy (Tabla 3.1). Además, éstos se restringen según:

- Cb: los anillos carbocíclicos seleccionados para ser incluidos se muestran en la Figura 3.5, ya que de entrada se descartan estructuras con un elevado número de anillos o aquellas no aromáticas.

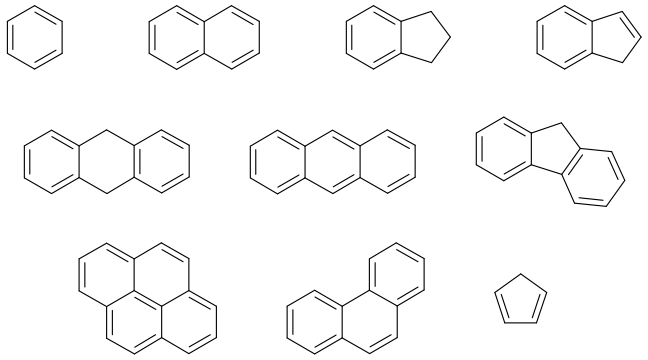
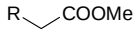
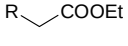


Figura 3.5. Anillos considerados para la sustitución Cb.

- Hy: se focaliza hacia los anillos tipo piridina, pirrol, tiofeno, furano, por comparación con los compuestos descritos en la bibliografía que presentan actividad inhibidora de tirosina quinasa.

Con ello, en la Tabla 3.3 se detallan las tres estructuras generales contempladas en la búsqueda junto con el número de *building blocks* inicialmente identificados, su posterior filtraje y el número final seleccionado.

Tabla 3.3. Proceso de selección primaria de monómeros sintéticos precursores de ésteres α,β-insaturados **55**.

Ésteres metílicos	
	R = Cb (obliga a MW<350) 8225 total iniciales→165 comerciales→descartando aquellos ciclos no aromáticos→**46** asequibles
46 + 12 = **58**	R = Hy (obliga a MW<350) 12761 total iniciales→921comerciales→24 pertenecen a los cuatro ciclos descritos→**12** asequibles
Ésteres etílicos	
	R = Cb (obliga a MW<350) 4670 total iniciales→69 comerciales→descartando aquellos ciclos no aromáticos→**30** asequibles
30 + 13 = **43**	R = Hy (obliga a MW<500) 41 pertenecen a los cuatro ciclos descritos→**13** asequibles

Ácidos	
	R = Cb (obliga a MW<500) 18965 total iniciales→841 comerciales→ aquellos cuyos anillos aparecen en la Figura 3.5, 626→ **220** asequibles
220 + 15 = **235**	R = Hy (obliga a MW<450) 32056total iniciales→2852 comerciales→130 pertenecen a los cuatro ciclos descritos→**15** asequibles

"cumplen estructura" se refiere a si la estructura propuesta se ajusta a la requerida.

Con lo que en total, el número de precursores sintéticos que rinden el *building block* éster α,β-insaturado **55** es de 336 (58+43+235).

3.2.1.3. Filtrado por eliminación de fragmentos repetidos

Los reactivos iniciales identificados en los dos apartados anteriores: 42 (ésteres α,β-insaturados) + 336 (precursores de ésteres α,β-insaturados), se filtran para eliminar aquellos compuestos que aportan el mismo fragmento R^4. Esto puede ser debido a:

- Las distintas formas metílica, etílica o ácida presentan el mismo sustituyente.

- La propia organización de la base de datos. Un mismo sustituyente puede pertenecer a diferentes registros en función del número de acceso (*Registry Number*), ya que el mismo producto se diferencia en distintos catálogos en función de su pureza, formato de venta (cantidades, si se encuentra hidratado, en función del contraión que lo neutraliza si se presenta en forma de sal…).

Para esta identificación se utiliza tanto una herramienta de SciFinder®Scholar 2003 que permite ordenar los compuestos por similitud, como el programa ChemFinder 6.0 para el que se construye un pequeño programa en Visual Basic. Este programa aprovecha la funcionalidad de la búsqueda de una estructura exacta (*kCFFullStructure*), recorriendo una base de datos importada como *sd-file*.

En la Tabla 3.4 se detalla el proceso de filtrado y el número de reactivos que restan en la lista de reactivos, rindiendo un total de 282 compuestos.

Tabla 3.4. Estadísticas después del filtrado por repetición de fragmentos de reactivos α,β-insaturados **55**.

Reactivo	Búsqueda Estructural	Iniciales	Eliminados	Restantes
Ésteres α,β-insaturados comerciales	Ésteres α,β-insaturados metílicos	15	-	15
	Ésteres α,β-insaturados etílicos	11	3	8
	Ácidos α,β-insaturados	16	9	7
Precursores de los ésteres α,β-insaturados	Ésteres metílicos	58	-	58
	Ésteres etílicos	43	23	20
	Ácidos	235	61	174

3.2.1.4. Filtrado por viabilidad sintética, toxicidad y estabilidad

Finalmente, se filtran los 282 compuestos restantes según diversos criterios:

- **Viabilidad sintética.** La reacción multicomponente (Figuras 3.2 y 3.3) se realiza con metóxido sódico, por lo que no es conveniente que existan en R^4 halógenos susceptibles de sufrir sustitución nucleófila. Por lo tanto, se descartan 20 estructuras que contienen bromo y 10 estructuras que contienen yodo. En principio, las 4-aminopirido[2,3-*d*]pirimidinas (**52**) y las 4-oxoaminopirido[2,3-*d*]pirimidinas (**53**) no presentarían problemas por la existencia de cloros, aunque tal vez sí para las 4-hidroaminopirido[2,3-*d*]pirimidinas (**54**). Dado que inicialmente se plantea la síntesis de las estructuras **52** y **53**, se opta por mantener los sustituyentes clorados. También se mantienen los reactivos que contienen flúor. En la Figura 3.6 se muestra la distribución de compuestos halogenados que finalmente se incluyen en la lista de reactivos.

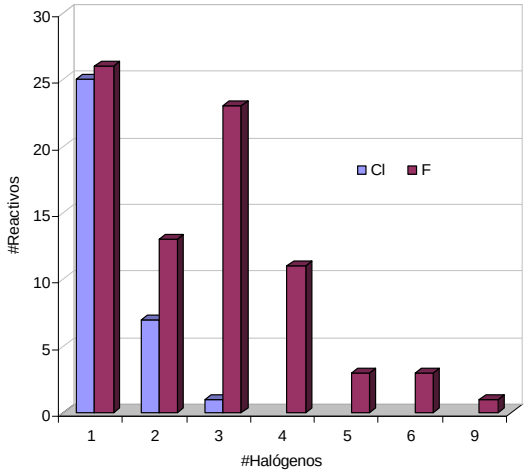
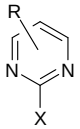
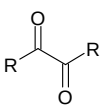
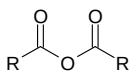
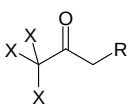
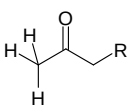
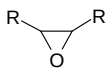
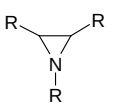
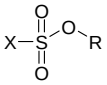
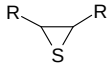
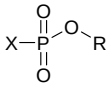
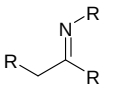
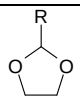
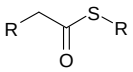
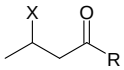
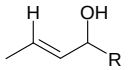
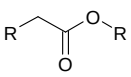
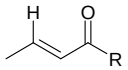
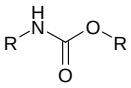


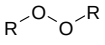
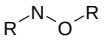
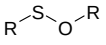
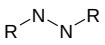
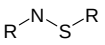
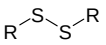
Figura 3.6. Distribución de reactivos halogenados incluidos en la selección final de reactivos α,β-insaturados **55**.

- **Filtrado según criterios de toxicidad, reactividad y estabilidad.** Se eliminan reactivos que generen productos con poca probabilidad de constituir buenos *leads* por presentar subestructuras o grupos funcionales muy reactivos, inestables o tóxicos. En la Tabla 3.5 se muestra un compendio de éstos.[17]

Tabla 3.5. Grupos funcionales no deseables por toxicidad, alta reactividad o inestabilidad [17, 367].

Grupo Funcional	Grupo Funcional	Grupo Funcional
Haluro de Sulfonilo 	Haluro de Acilo 	Halopirimidina 
1,2-dicarbonilo 	Anhídrido 	Perhalo-cetona 
Cetona alifática 	Epóxido 	Aziridina 
Éster sulfónico 	Tioepóxido 	Éster fosfónico 

Imina		Dioxolano		Tioéster alifático	
Carbonilo β-sustituído		Aldehído		Enol y Enolato inestables	
Éster alifático		Cianohidrina		Aceptor de Michael	
Carbamato		Haluro de alquilo		Haluro de fosforilo	$POCl_3$
Isocianato	R-N=C=O	Azida	R-N=N+=N-	Peróxido	R-OOH
Silicato	SiO_4^{4-}	Isotiocianato	R-N=C=S	Trifenilfosfina	Ph_3P
Perácido	R-COOOH	Compuesto diazónico	R-N=N-R		

Enlaces simples heteroátomo-heteroátomo		
		
		

Siguiendo estos criterios, se eliminan un total de 21 estructuras. En particular, para los precursores de ésteres no se consideran aquellos que presentan dos puntos de reactividad, incluso aún siendo simétricos, por posibles problemas en su conversión a ésteres α,β-insaturados.

- **Filtrado por peso molecular.** Partiendo de la referencia de peso molecular total inferior a 500 g/mol establecida por Lipinski y considerando la contribución del *core* más favorable (4-hidroaminopirido[2,3-*d*]pirimidinas, **54**), de 146 g/mol, la contribución "ideal" de cada resto (R^4 y R^1) es aproximadamente de 170 g/mol (2×MW=500-146; MW ~170 g/mol).

  Para cada uno de los reactivos, se establece la aportación del peso molecular del sustituyente R^4 a la molécula final. Sin embargo, a fin de que el filtrado no sea demasiado severo y dado que posteriores revisiones de las reglas de Lipinski han concluido que se puede elevar el criterio de corte de peso molecular, el límite real impuesto de corte de peso molecular aportado se considera de 235 g/mol. Además, se debe tener en cuenta que fijándolo a 170, existirían muchas posibles combinaciones de fragmentos (R^4×R^1) eliminados sin que su peso molecular total superase los 500 g/mol. Con ello, se eliminan un total de 11 productos. En la Figura 3.7 se refleja la distribución de pesos moleculares aportados por los reactivos finalmente seleccionados, donde 17 compuestos superan los 200 g/mol.

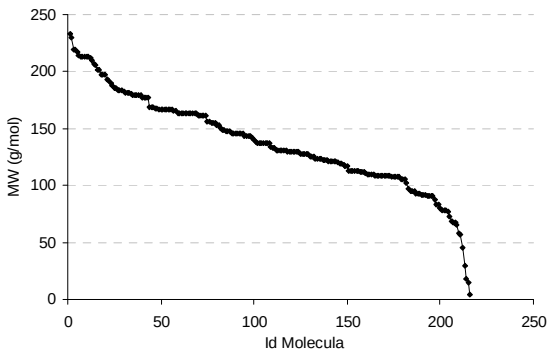


Figura 3.7. Distribución de pesos moleculares aportados por los restos R^4.

- **Filtrado por número de anillos.** Se considera que para que un compuesto posea propiedades *drug-like* éste debe presentar un número de anillos comprendido entre 1 y 4[368]. Dado que el *core* aporta dos anillos, se eliminan todas aquellas estructuras que aporten tres o más anillos, resultando en un total de 4 estructuras descartadas.

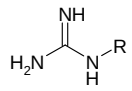
En resumen, de los 282 compuestos previos al filtrado según estos cuatro criterios, se eliminan un total de 66 [30 (viabilidad sintética) + 11 (Peso molecular) + 4 (anillos) +21 (toxicidad, reactividad)]. Así, el número de restos R^4 considerados es de 216, aportados por distintos ésteres α-β insaturados o precursores de los mismos.

3.2.2. Selección y filtrado de guanidinas

En la Tabla 3.6 se detalla la estructura general contemplada en la búsqueda junto con el número de *building blocks* inicialmente identificados, su posterior filtraje y el número final seleccionado. R corresponde a cada una de las cuatro posibles sustituciones especificadas en la Tabla 3.1.

En este caso, el filtrado por peso molecular se realiza directamente desde el programa SciFinder®Scholar 2003. Según lo expuesto en el apartado anterior la aportación de cada resto R^1 debería ser próxima a 170 g/mol. Se considera un límite de 300 g/mol, que al restarle la parte de la guanidina ya considerada en el *core* (43 g/mol) supone una aportación máxima de R^1 próxima a 250 g/mol. Este valor, altamente permisible, se considera adecuado dado que también se incluyen sustituyentes R^4 con aportaciones al peso molecular muy bajas (Figura 3.7).

Tabla 3.6. Proceso de selección primaria de monómeros sintéticos directamente comerciales para el *building block* guanidina sustituida **57**.

Guanidinas sustituidas	
	R = Q (sin límite de MW Inicial) 10115 total iniciales→524 comerciales→233 (MW<300)→**66** asequibles
	R = Ak (obliga a MW<300) 13845 total iniciales→505 comerciales→505 (MW<300)→**89** asequibles
	R = Cb (sin límite de MW Inicial) 7721 total iniciales→68 comerciales→62 (MW<300)→**18** asequibles
66 + 89 + 18 + 21= **194**	R = Hy (sin límite de MW Inicial) 6250 total iniciales→96 comerciales→82 (MW<300)→**21** asequibles

La eliminación de fragmentos repetidos se realiza también directamente desde SciFinder®Scholar 2003.

Por otra parte, de las 194 guanidinas resultantes, se eliminan otras 9 por ser formas enantioméricas de otras incluidas en la búsqueda. El motivo es que los programas usados para el tratamiento de la quimioteca virtual posterior (MOE y Cerius2) ignoran la información estereoquímica, ya sea durante la obtención del modelo tridimensional (Cerius2) como en el mismo momento de la importación de un fichero *sd-file* (MOE).

3.2.2.1. Filtrado por viabilidad sintética, toxicidad y estabilidad

Análogamente al caso anterior, se filtran los 185 compuestos restantes según diversos criterios:

- **Viabilidad sintética.** Para no interferir con una posible sustitución nucleófila, se descartan cuatro compuestos bromados y uno yodado. En la Figura 3.8 se muestra la distribución de compuestos halogenados que finalmente se incluyen en la lista de reactivos.

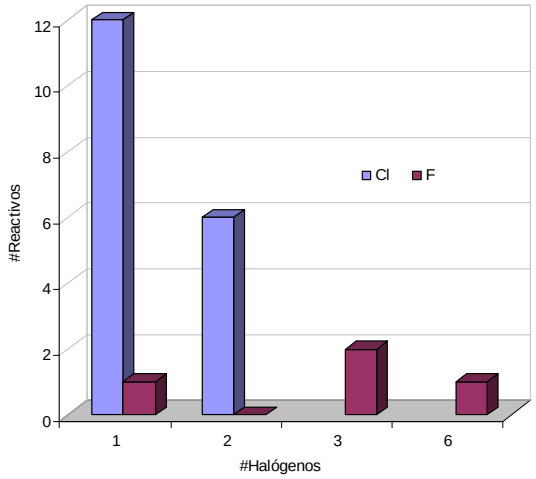


Figura 3.8. Distribución de reactivos halogenados incluidos en la selección final de guanidinas **57**.

- **Filtrado según criterios de toxicidad, reactividad y estabilidad.** No se detectan compuestos que satisfagan los criterios de la Tabla 3.5. Sin embargo, se excluyen 12 estructuras que presentan dos grupos funcionales guanidino.

- **Filtrado por peso molecular.** Pese a que en teoría, el filtrado de SciFinder®Scholar 2003 por peso molecular debería eliminar todas aquellas estructuras con un peso molecular superior a 300 g/mol, se detecta que existen cuatro estructuras que superan levemente este límite. En cualquier caso, el peso real aportado por el sustituyente R^1 es inferior a 300 g/mol, como puede observarse en la Figura 3.9, en la que se detalla la distribución de pesos moleculares aportados por este fragmento R^1. Si se aplicara estrictamente el límite de 170 g/mol, se descartarían 73 estructuras del total de 165 guanidinas finales seleccionadas (44%).

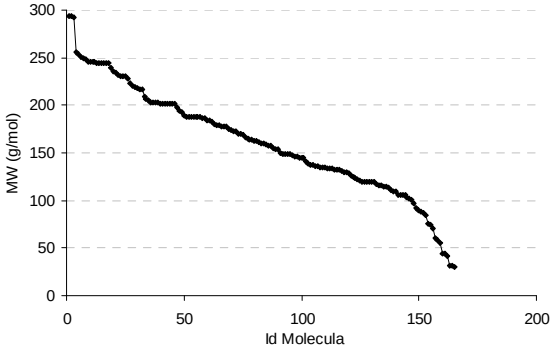


Figura 3.9. Distribución de pesos moleculares aportados por los restos R^1.

- **Filtrado por número de anillos.** Se eliminan dos estructuras que presentan tres anillos y una estructura macrocíclica con un anillo de 12 miembros.

En resumen, de los 185 compuestos previos al filtrado según estos cuatro criterios, se excluyen un total de 20 [5 (viabilidad sintética) + 0 (Peso molecular) + 3 (anillos) +12 (toxicidad, reactividad)]. Así, el número de restos R^1 considerados es de 165, aportados por guanidinas sustituidas comerciales.

3.2.3. Comparación de los restos R^1 y R^4 seleccionados con los restos presentes en inhibidores de tirosina quinasas descritos en la bibliografía

A continuación se presentan aquellos restos descritos por Klutchko[342] y Hamby[344] en las 7-oxopirido[2,3-*d*]pirimidinas descritas por ellos con actividad inhibidora de tirosina quinasa y que se encuentran recogidos en las selecciones de reactivos de la quimioteca virtual.

En la Tabla 3.7 se recogen los correspondientes restos R^4 aportados por los ésteres α,β-insaturados y en la Tabla 3.8, los restos R^1 aportados por las guanidinas sustituidas.

Tabla 3.7. 21 Restos R^4 descritos en 7-oxopirido[2,3-*d*]pirimidinas con actividad inhibidora tirosina quinasa[342, 344].

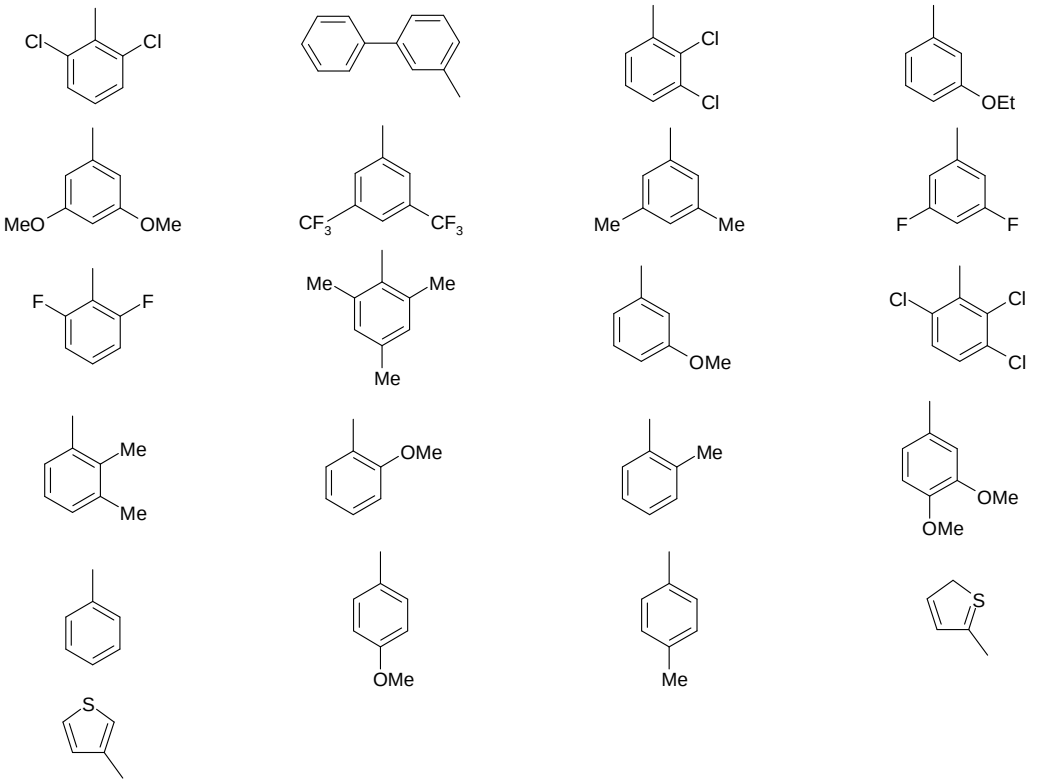
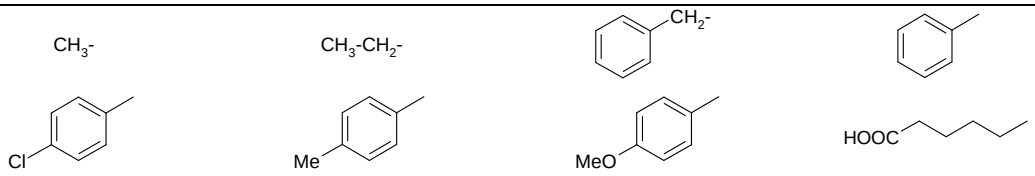


Tabla 3.8. 8 Restos R^1 descritos en 7-oxopirido[2,3-*d*]pirimidinas con actividad inhibidora tirosina quinasa[342, 344] e incluidos en la selección de guanidinas sustituidas.



De un total de 31 restos R^4 recopilados de las referencias [342] y [344], 21 restos se encuentran presentes en esta quimioteca. Por otra parte, la representación de restos R^1 "activos" es mucho menor ya que únicamente se incluyen 8 de los 55 recopilados debido a que los 47 restantes no son comerciales.

3.3. Enumeración de la quimioteca

La selección de reactivos del apartado anterior rinde 165 restos R^1 y 216 restos R^4, que combinados con los 3 posibles restos R^2 (amino, oxo e hidro) generan una quimioteca virtual de 106920 compuestos.

Sin embargo, en la planificación sintética de la misma, se decide inicialmente comenzar a desarrollar los derivados 4-amino, 4-oxo y posteriormente los 4-hidro. Por ello, la construcción de las quimiotecas se mantiene independiente, es decir, se construyen tres quimiotecas combinatorias independientes correspondientes a cada una de las posibles sustituciones en C-4, cada una de las cuales contiene 35640 compuestos. En adelante, se denomina a la quimioteca 4-oxo: BIB_Oxo, a la 4-amino: BIB_Amino y a la 4-hidro: BIB_Hidro.

La enumeración de una quimioteca corresponde a la construcción automatizada de la tabla de conectividades. Esta construcción puede realizarse según dos esquemas:

- **Por fragmentos.** Se identifica el *core* o *scaffold* común de los productos y sus puntos de sustitución, constituyendo el patrón de numeración o *template*. Los reactivos deben modificarse suprimiendo la funcionalidad que quede englobada en el *template* e indicando el átomo a enlazar en los puntos de sustitución de éste.

- **Por reactivos.** La quimioteca se genera a partir de los reactivos inalterados y una descripción de la(s) reaccion(es) químicas implicadas, emulando la síntesis convencional del laboratorio.

En este caso la enumeración se realiza con el módulo *Analog Builder* de Cerius2 que sigue el esquema por fragmentos. Para ello, los reactivos deben modificarse a fragmentos, indicando el punto de unión con el *core* mediante un *dummy atom* tipo X. Este módulo no dispone de una función que realice el proceso de manera automática, por lo que en principio debe realizarse manualmente. Para reducir la laboriosidad, se preparan *scripts* en lenguaje TCL (*Tool Command Language*) que manipulan los *sd-files* de los reactivos convenientemente. Finalmente, la enumeración se realiza siguiendo el esquema de la Figura 3.10.

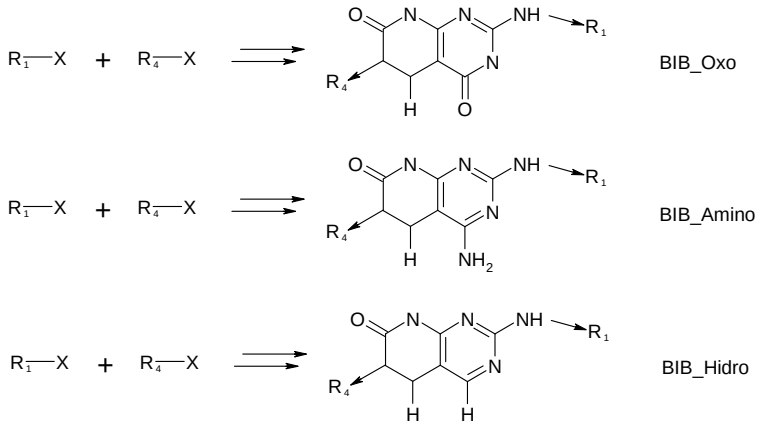


Figura 3.10. Esquema seguido en la enumeración por fragmentos de las tres quimiotecas combinatorias.

La construcción se realiza directamente en un *sd-file* en formato 2D. Alternativamente, Cerius2 dispone de un fichero en formato binario (*bdf*) que permite la manipulación más eficiente de las quimiotecas combinatorias. Sin embargo, no se trata de un formato estándar a todos los programas de diseño molecular.

3.4. Optimización y descripción de las quimiotecas

Cada uno de los ficheros *sd-file* de las quimiotecas se importa al programa MOE, donde se realiza la optimización (estructuras 3D) con el *force field* MMFF94 fijando como condiciones de terminación un gradiente inferior a 0.01.

En el mismo programa se calculan los siguientes descriptores (apartado 1.6):

- **Índices topoestructurales**: índices de Wiener y Zagreb, número de polaridad de Wiener, radio topológico, diámetro topológico, índices de forma de Kier ($^1\kappa$, $^2\kappa$ y $^3\kappa$), índice de Balaban e índice de forma de Petitjean.

- **Índices topoquímicos**: índices de conectividad de orden 0 y 1 ($^0\chi$,$^1\chi$), índices de conectividad de valencia de orden 0 y 1 ($^0\chi^v$,$^1\chi^v$), índices de forma de Kier modificados ($^1\kappa_\alpha$, $^2\kappa_\alpha$ y $^3\kappa_\alpha$) y el índice de flexibilidad de Kier ϕ.

- **Índices topológicos basados en la teoría de la información**: contenido de información de un sistema con *n* elementos (*a_IC*), índice de contenido medio de información de magnitud de adyacencia, índice de contenido medio de información de igualdad de adyacencia, índice de contenido medio de información de igualdad de distancia e índice de contenido medio de información de magnitud de distancia.

- **Descriptores de forma:** globuralidad, momento de inercia principal, radio de giro y volumen molecular de van der Waals calculado mediante una aproximación en mallas.

- **Descriptores fisicoquímicos:** peso molecular, momento dipolar, suma de polarizabilidades atómicas, densidad (*dens*, obtenida a partir de volumen molecular calculado por mallas), logaritmo del coeficiente de partición (SlogP), refractividad molecular (SMR), descriptores derivados de la energía potencial del *force field* (potencial total, componente electrostática de la energía potencial y energía de solvatación), carga formal de la molécula, superficie molecular accesible, área de la superficie polar y descriptores de carga parcial (Q_PC+,Q_PC-,Q_RPC+, Q_RPC-).

- **Descriptores de Carga Parcial y Área de superficie:** DASA, FASA+, FASA-, FASA_H, FASA_P.

La dimensionalidad de cada conjunto se reduce mediante análisis de componentes principales (PCA) mediante el programa MOE, manteniendo todos aquellos PCs que resumen un 90% de la varianza de los datos: 8 PCs para BIB_Oxo, 7 PCs para BIB_Amino y 8 PCs para BIB_Hidro.

En la Figura 3.11 se representan las tres quimiotecas en función de sus tres primeras componentes principales.

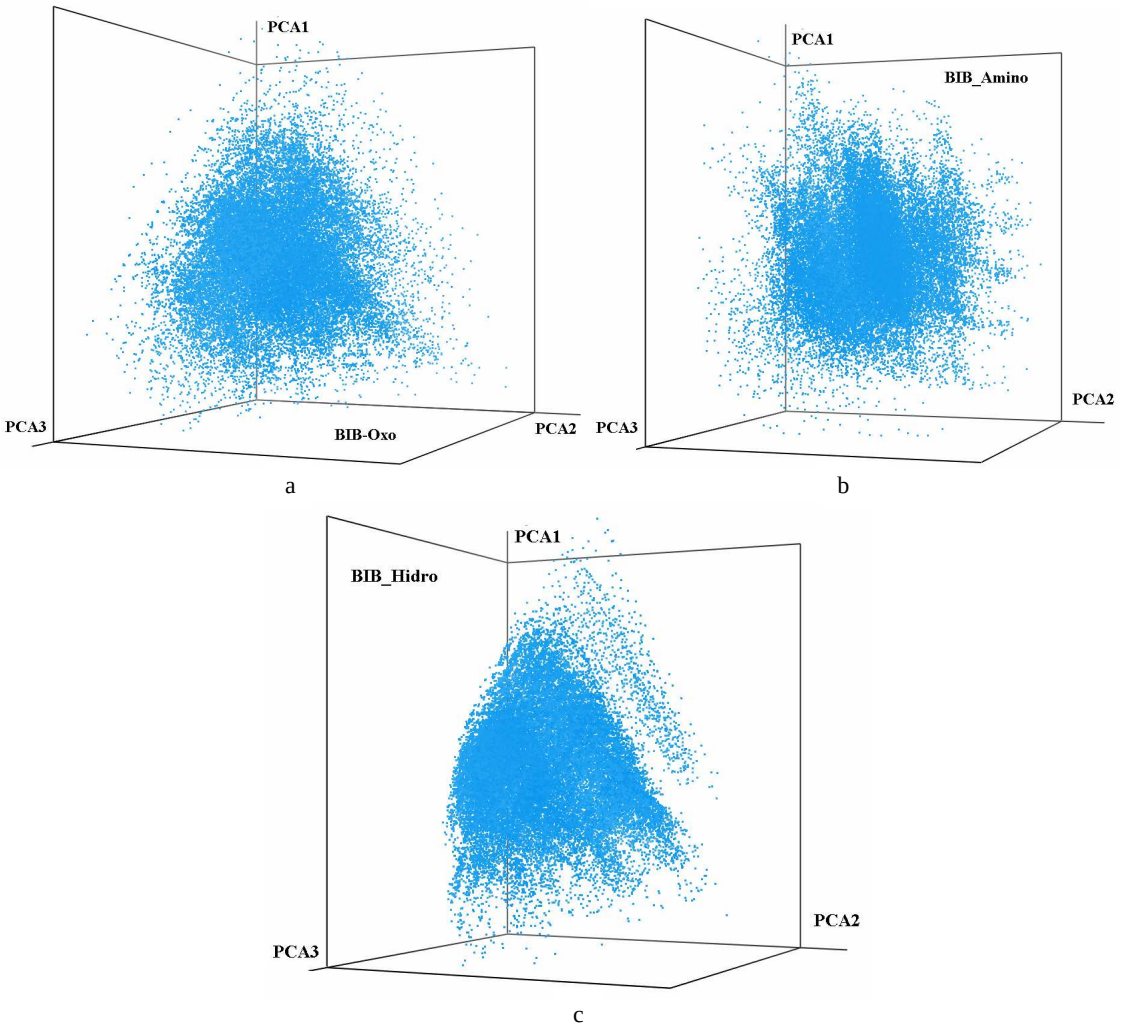


Figura 3.11. Representación de las quimiotecas (a) BIB_Oxo, (b) BIB_Amino y (c) BIB_Hidro en función de sus tres primeras componentes principales, que explican el 75.6 %, 76.5% y 75.7% de la varianza respectivamente.

3.5. Selección de compuestos y análisis de resultados

Para realizar la selección de compuestos se aplican los métodos de partición (*cell-based*) (apartado 1.10.2.3) en formato *full array*, implementados en el programa Cerius2. La partición del espacio se realiza mediante el algoritmo de *Optimum Binning* y como métricas de diversidad se utilizan las cuatro funciones (*cell-based Fraction*, *cell-based Chi2*, *cell-based Entropy* y *cell-based Density*) correspondientes a las ecuaciones [1.103], [1.105]-[1.107] del apartado 1.10.1. Como algoritmo de optimización se usa Monte Carlo, con una temperatura de 300 y un número de ciclos suficientemente alto (500.000).

Se escoge el método *cell-based* como método de selección frente a métodos basados en distancias porque estos últimos tienden a identificar compuestos situados en los extremos de la quimioteca (*edge design*). Por otra parte, el uso de métodos de *clustering*, teóricamente superiores en cuanto a la homogeneidad de la selección, se descarta ya que, por una parte no es posible realizar selecciones *full array* en el programa Cerius2 y, por otra, el tamaño de estas quimiotecas se supera ligeramente el límite de memoria aceptado por los métodos de *clustering* implementados en PRALINS (~30.000 compuestos).

El recubrimiento de las quimiotecas combinatorias obtenidas por cada una de las cuatro funciones objetivo citadas se evalúa también por métodos *cell-based* atendiendo a dos criterios (apartado 1.10.4):

- **Recubrimiento de espacio** (SPC) Relación entre el número de celdas ocupadas por el subconjunto y el número de celdas totales resultantes de una partición del espacio. Este criterio coincide con el criterio a optimizar por la función *cell-based Fraction.*

- **Recubrimiento de población** (POP). Relación entre el número de moléculas presentes en las celdas ocupadas por el subconjunto y el número total de moléculas contenidas en la quimioteca.

Para realizar esta evaluación, se determina primeramente un marco de referencia en el que interesa que la distribución de compuestos sea lo más homogénea posible y en el que se minimice el número de *singletons* presentes (*bins* que contienen un único compuesto).

3.5.1. Elección de un marco de referencia

Para cada una de las quimiotecas se realizan diferentes niveles de partición según el algoritmo de *Optimum Binning*. La distribución de compuestos en las celdas para cada una de las particiones y las tres quimiotecas se muestra en la Figura 3.12.

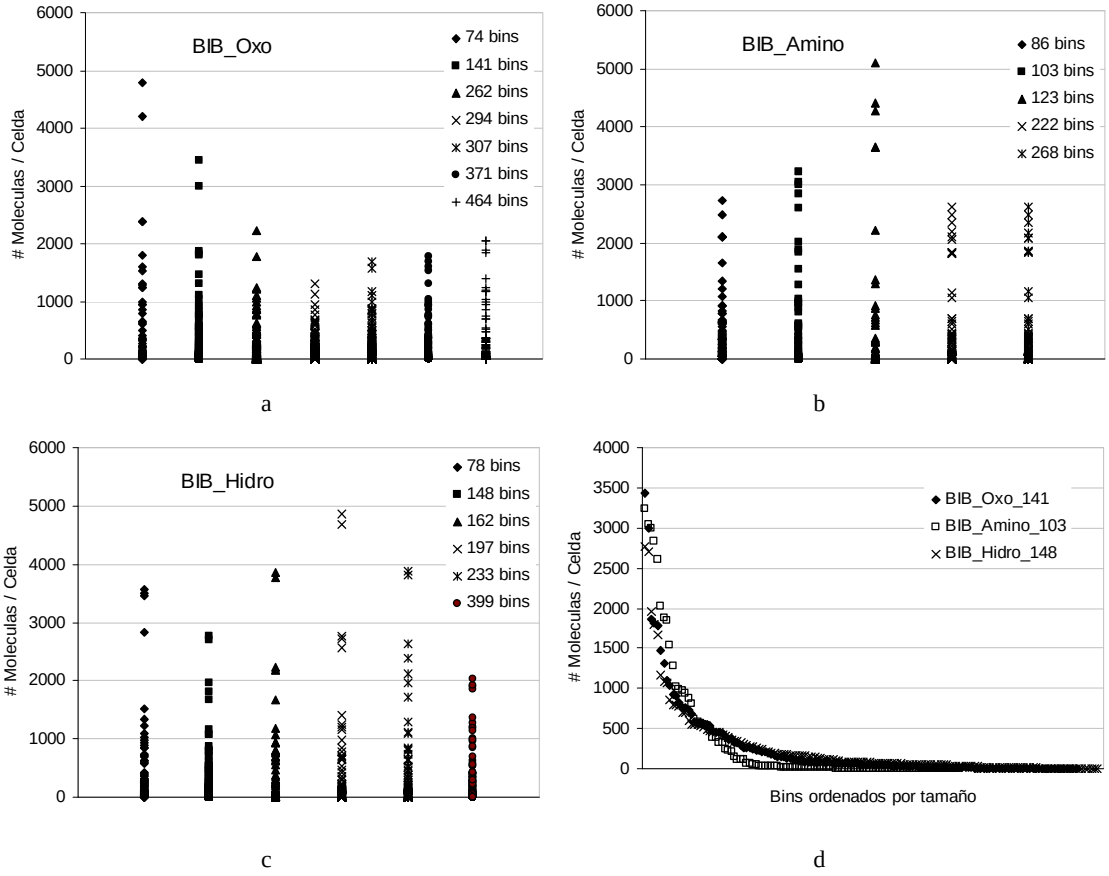


Figura 3.12. Análisis de la uniformidad en la distribución de compuestos de distintas particiones según el algoritmo *Optimum Binning* para cada una de las tres quimiotecas: (a) BIB_Oxo, (b) BIB_Amino y (c) BIB_Hidro. En la leyenda se indica el número total de celdas ocupadas en cada partición. (d) Distribución de compuestos en la partición de referencia seleccionada para cada quimioteca.

Alternativamente, en la Tabla 3.9 se muestra el número de *bins* totales de cada una de las particiones, el número de *bins* ocupados, el número de *singletons* y el número de *bins* con una cantidad de compuestos inferior a diez.

Tabla 3.9. Caracterización de cada una de las particiones para cada quimioteca en función del número total de *singletons* y número de *bins* ocupados por menos de diez compuestos. En negrita aparecen las particiones seleccionadas para cada una de las quimiotecas.

		P1	**P2**	P3	P4	P5	P6	P7
BIB_Oxo	# Bins Totales	96	**192**	384	576	768	1152	1728
	# Bins Ocupados	74	**141**	262	294	307	371	464
	# *Singletons*	2	**9**	20	20	56	75	113
	# Bins <10 compuestos	14	**36**	84	65	149	209	281
BIB_Amino	# Bins Totales	96	**128**	192	384	480		
	# Bins Ocupados	86	**103**	123	222	268		
	# *Singletons*	5	**11**	16	35	49		
	# Bins <10 compuestos	7	**44**	48	109	153		
BIB_Hidro	# Bins Totales	96	**192**	256	384	576	1152	
	# Bins Ocupados	78	**148**	162	197	233	399	
	# *Singletons*	5	**7**	27	38	42	75	
	# Bins <10 compuestos	10	**23**	82	101	115	235	

Considerando la uniformidad de la distribución y priorizando un bajo número de *singletons*, se toma como partición referencia aquella que divide el espacio químico de la quimioteca BIB_Oxo en 141 bins, BIB_Amino en 103 *bins* y BIB_Hidro en 148 *bins*. En la Figura 3.12.d se muestra la distribución de compuestos en las celdas o *bins* para cada una de ellas.

3.5.2. Evaluación de las selecciones según las cuatro funciones objetivo

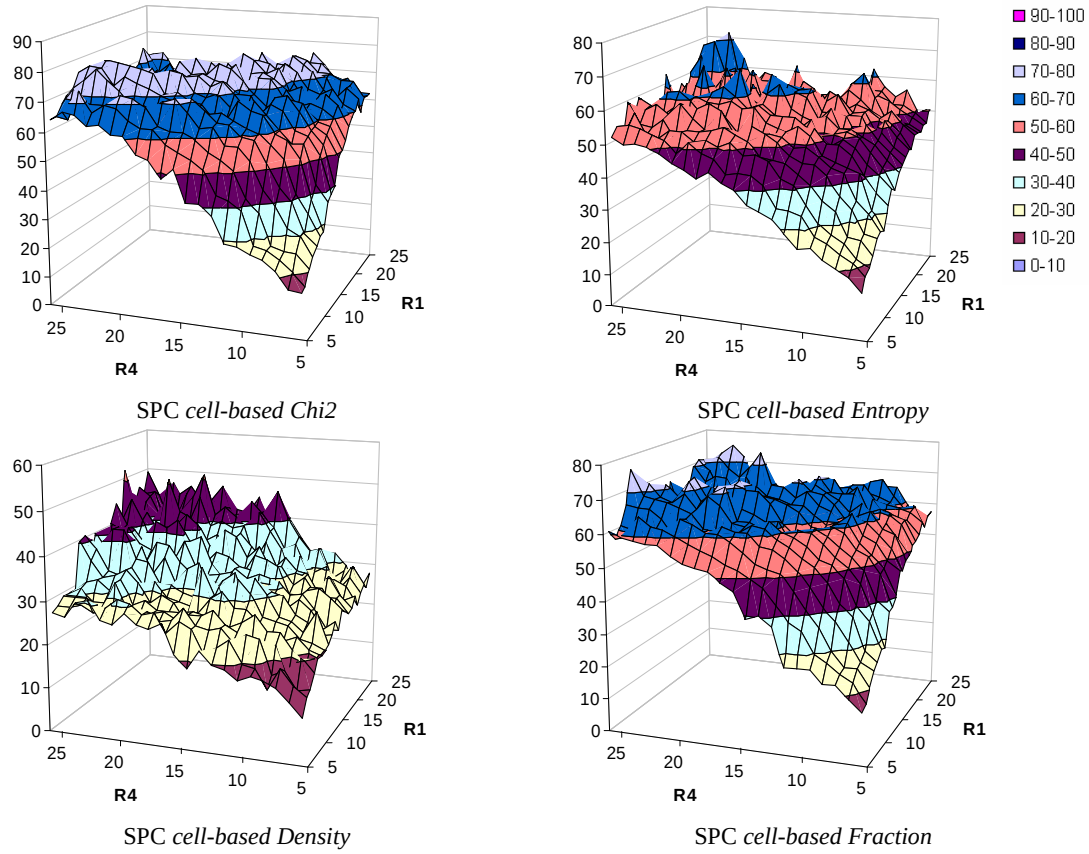
En principio, el tamaño de selección de la quimioteca a sintetizar se plantea de unos 100 compuestos en formato combinatorio. A falta de otros datos, lo más adecuado es que el peso asignado a cada uno de los reactivos que componen el *full array* final sea equilibrado, es decir, realizar una selección 10×10. Para una quimioteca de tamaño *N*, el tamaño de selección usualmente considerado óptimo corresponde a $\sqrt{N}$ para las selecciones *cherry picking*. Para las quimiotecas de 35640 compuestos, este valor se sitúa en torno a 166 compuestos. Aplicando el mismo criterio para cada uno de los puntos de diversidad de la quimioteca, resulta en R^1~13 y R^4~15. Estos valores son ligeramente superiores al formato 10×10, por lo que se estudia el recubrimiento de diversas combinaciones de restos: comprendiendo el rango de 5-25 para cada uno de los restos R^1 y R^4.

En la Tabla 3.10 se muestran los tamaños de selección para cada una de las combinaciones de restos R^1 y R^4 para los que se realizan selecciones *full array*.

Tabla 3.10. Tamaños de selección resultantes de cada una de las configuraciones $R^1 \times R^4$.

	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
5	25																				
6	30	36																			
7	35	42	49																		
8	40	48	56	64																	
9	45	54	63	72	81																
10	50	60	70	80	90	100															
11	55	66	77	88	99	110	121														
12	60	72	84	96	108	120	132	144													
13	65	78	91	104	117	130	143	156	169												
14	70	84	98	112	126	140	154	168	182	196											
15	75	90	105	120	135	150	165	180	195	210	225										
16	80	96	112	128	144	160	176	192	208	224	240	256									
17	85	102	119	136	153	170	187	204	221	238	255	272	289								
18	90	108	126	144	162	180	198	216	234	252	270	288	306	324							
19	95	114	133	152	171	190	209	228	247	266	285	304	323	342	361						
20	100	120	140	160	180	200	220	240	260	280	300	320	340	360	380	400					
21	105	126	147	168	189	210	231	252	273	294	315	336	357	378	399	420	441				
22	110	132	154	176	198	220	242	264	286	308	330	352	374	396	418	440	462	484			
23	115	138	161	184	207	230	253	276	299	322	345	368	391	414	437	460	483	506	529		
24	120	144	168	192	216	240	264	288	312	336	360	384	408	432	456	480	504	528	552	576	
25	125	150	175	200	225	250	275	300	325	350	375	400	425	450	475	500	525	550	575	600	625

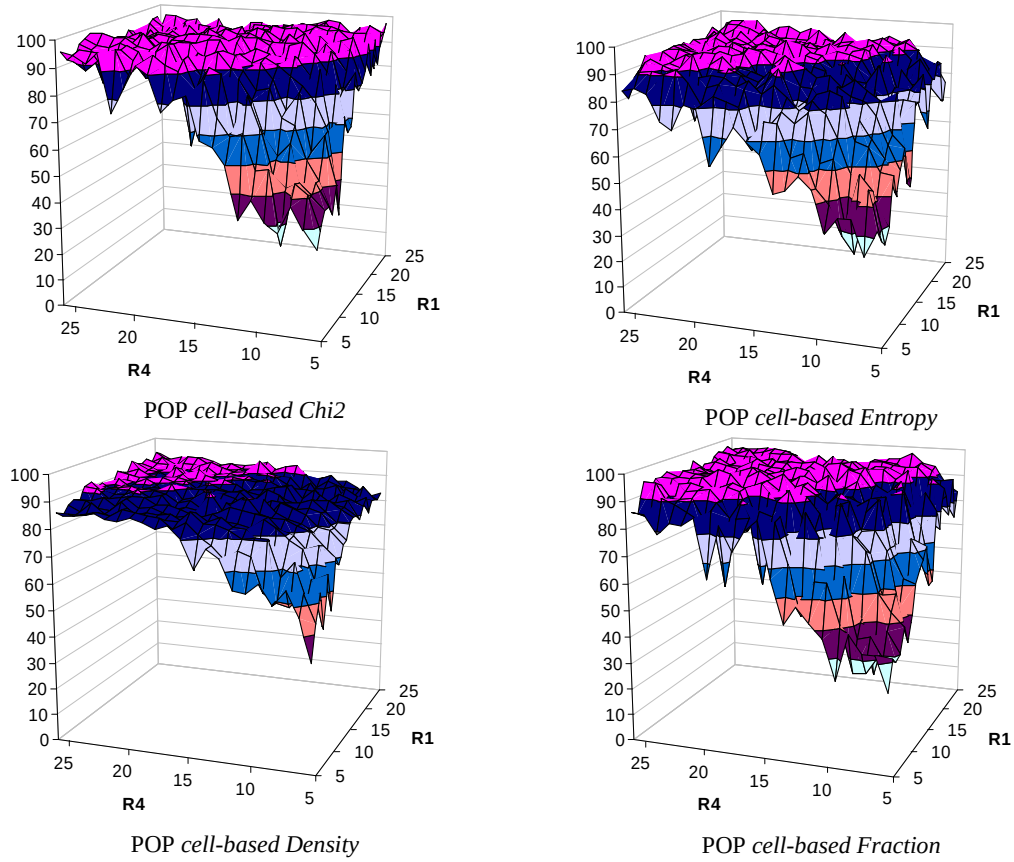
Tanto las selecciones como su evaluación se calculan mediante *scripts* programadas en TCL. Los resultados se muestran en las gráficas tridimensionales de las Figuras 3.13-3.18. Cada conjunto de cuatro gráficas corresponde a la evaluación de las selecciones optimizadas según las cuatro funciones objetivo citadas. Para cada quimioteca, se generan dos cuartetos correspondientes a la evaluación del recubrimiento en términos de espacio y población, respectivamente. Los dos ejes horizontales corresponden al número de reactivos seleccionados para cada punto de diversidad (R^1 y R^4). El eje vertical representa el recubrimiento. La gradación en colores indica el rango de valores en el que se sitúa el recubrimiento de cada punto, según se especifica en la leyenda.



SPC *cell-based Chi2* SPC *cell-based Entropy*

SPC *cell-based Density* SPC *cell-based Fraction*

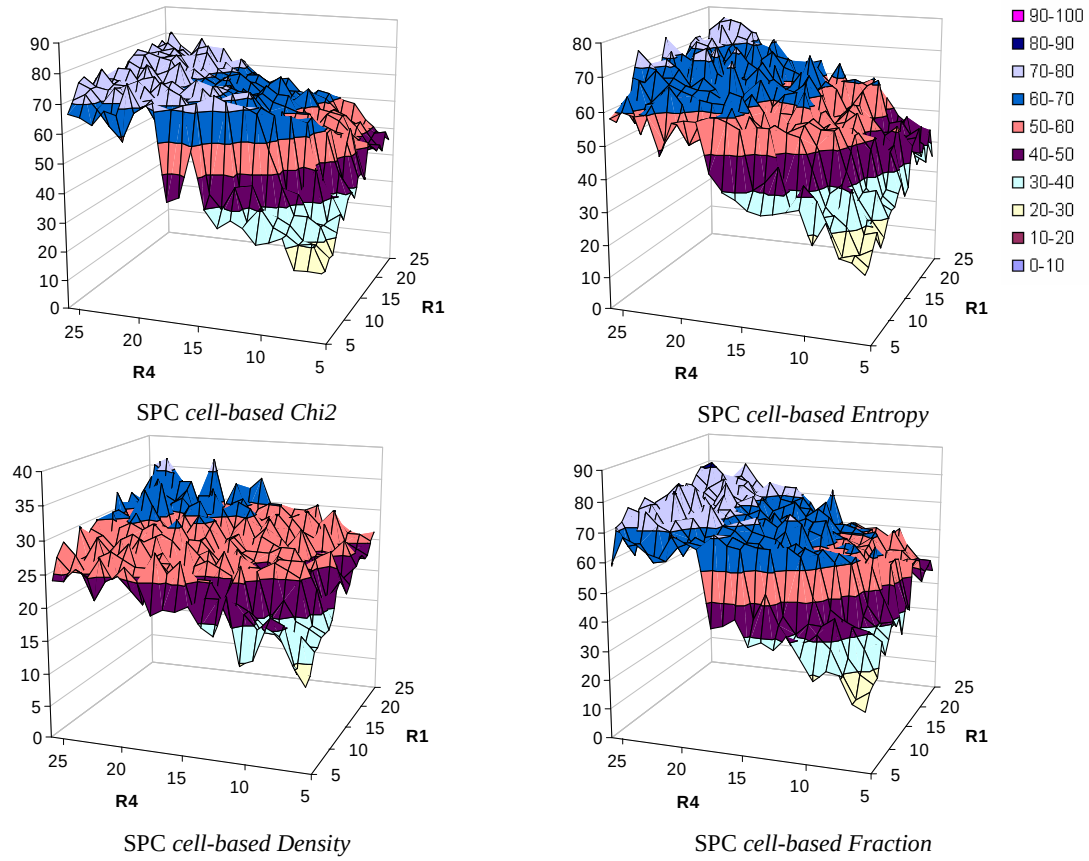
Figura 3.13. Evaluación del recubrimiento en espacio para las cuatro funciones en **BIB_Oxo.**



POP *cell-based Chi2* POP *cell-based Entropy*

POP *cell-based Density* POP *cell-based Fraction*

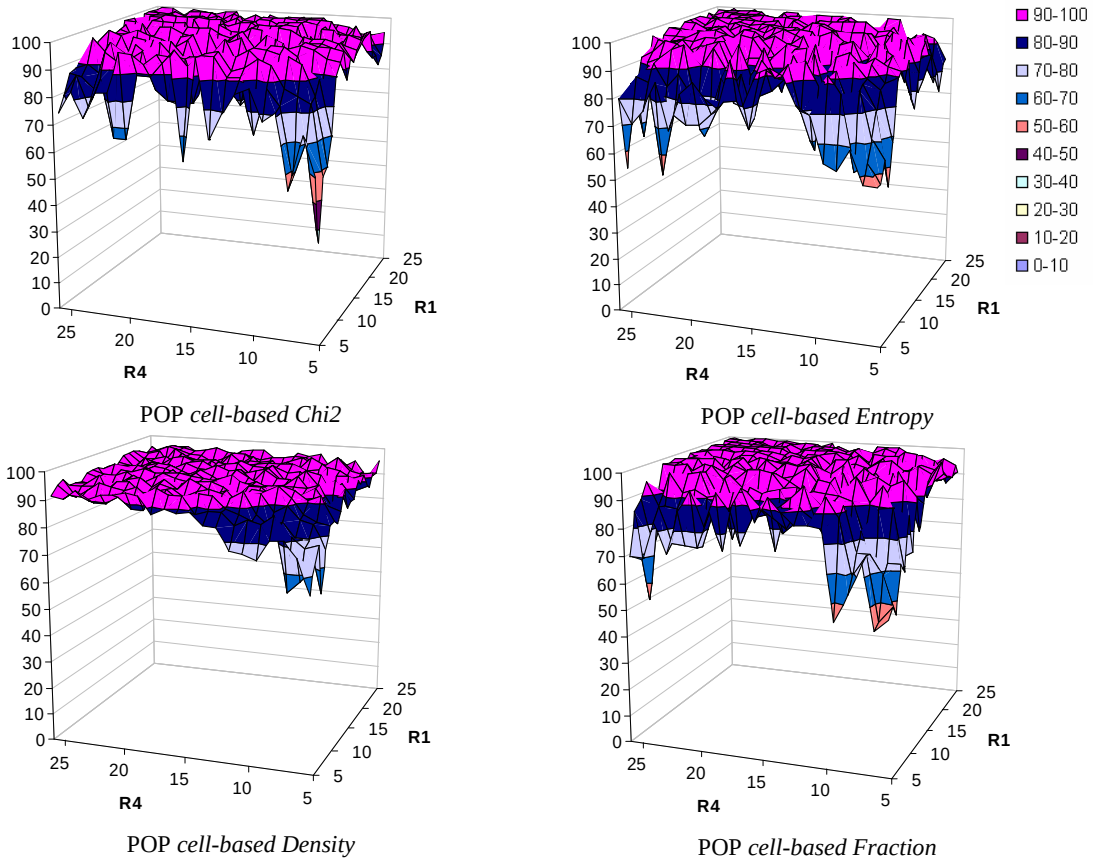
Figura 3.14. Evaluación del recubrimiento en población para las cuatro funciones en **BIB_Oxo**.



SPC *cell-based Chi2* SPC *cell-based Entropy*

SPC *cell-based Density* SPC *cell-based Fraction*

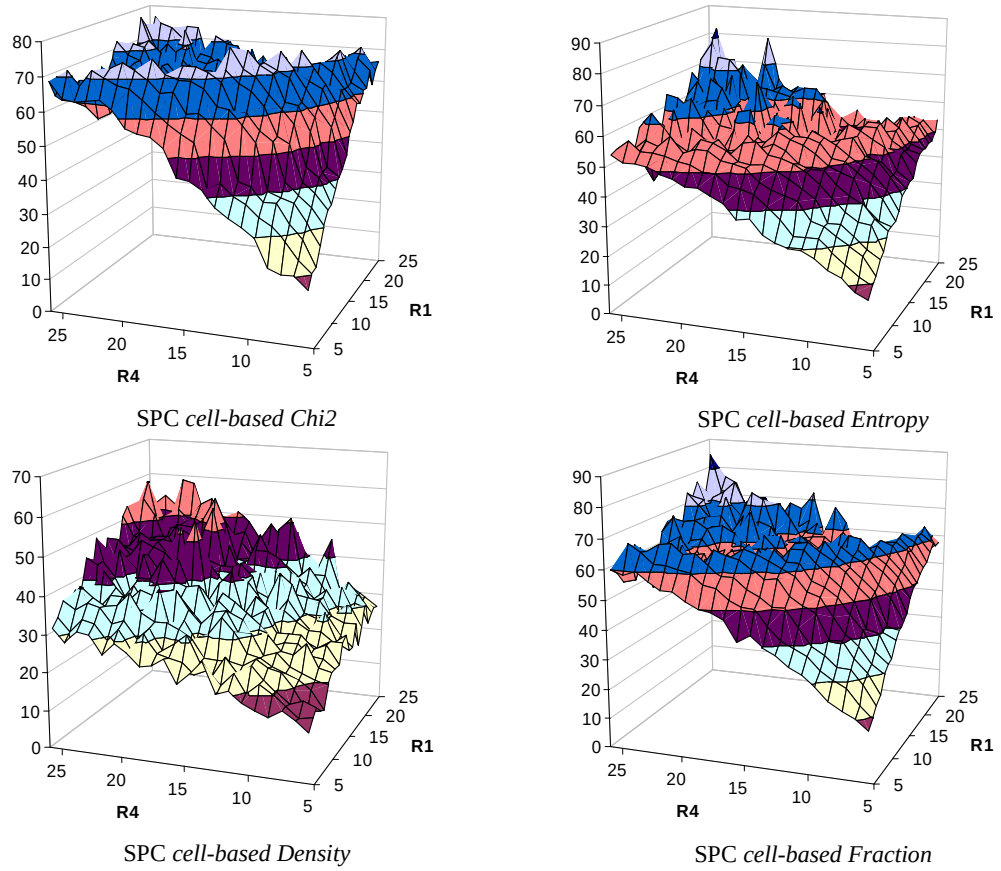
Figura 3.15. Evaluación del recubrimiento en espacio para las cuatro funciones en **BIB_Amino**.



POP *cell-based Chi2* POP *cell-based Entropy*

POP *cell-based Density* POP *cell-based Fraction*

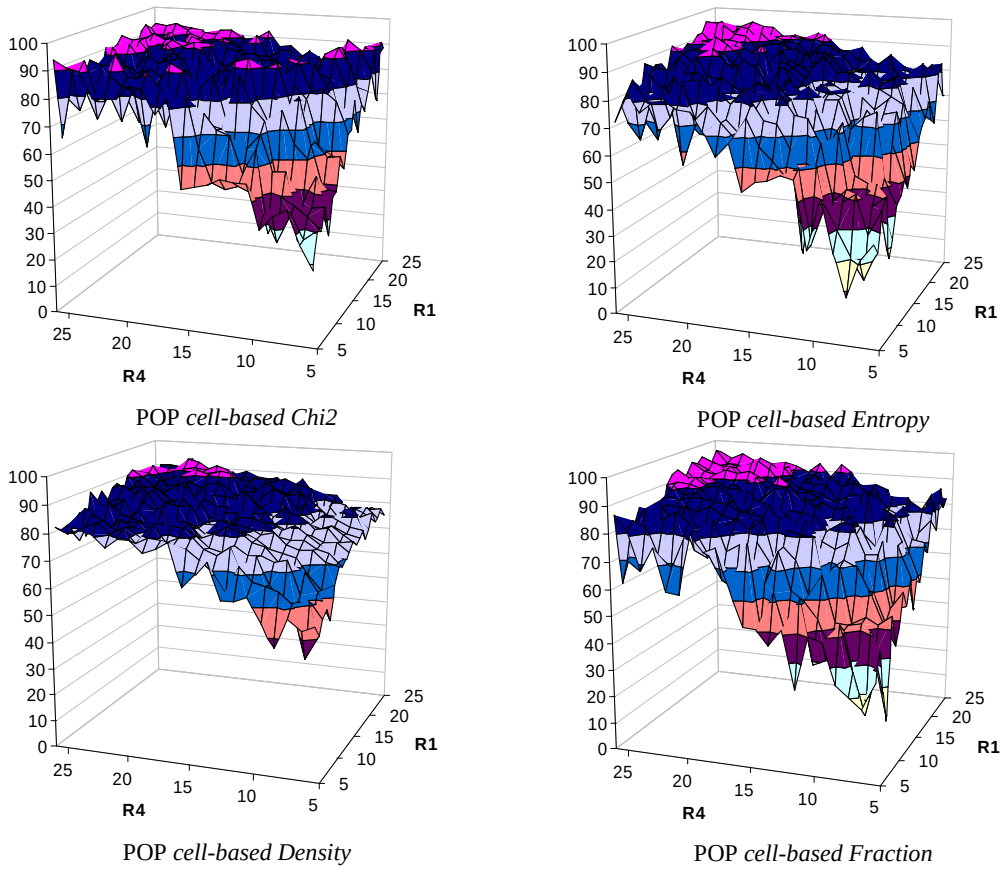
Figura 3.16. Evaluación del recubrimiento en población para las cuatro funciones en **BIB_Amino**.



SPC *cell-based Chi2* SPC *cell-based Entropy*

SPC *cell-based Density* SPC *cell-based Fraction*

Figura 3.17. Evaluación del recubrimiento en espacio para las cuatro funciones en **BIB_Hidro**.



POP *cell-based Chi2* POP *cell-based Entropy*

POP *cell-based Density* POP *cell-based Fraction*

Figura 3.18. Evaluación del recubrimiento en población para las cuatro funciones en **BIB_Hidro**.

En primer lugar, se observa cómo el recubrimiento en términos de población es muy elevado, con una gran parte de las selecciones superando el 90% de recubrimiento. Esto es debido a que la distribución en celdas de los compuestos no es totalmente homogénea, por lo que muestreando un número adecuado de celdas se consigue un buen recubrimiento en población. Esto se refleja, por ejemplo, en los resultados obtenidos para BIB_Amino, en los que al estar la partición realizada en un menor número de celdas, se obtienen en promedio recubrimientos mayores.

Comparando las diferentes funciones, se observa cómo la función *cell-based Density*, que optimiza principalmente la población es precisamente la que muestra menores superficies de bajo recubrimiento. Sin embargo, significativamente para el caso de la quimioteca BIB_Oxo (Figura 3.14), esta función es la que contiene un área más reducida de selecciones con recubrimientos comprendidos entre el 90-100%, comparado con las otras tres funciones restantes. Esto puede deberse a que esta función muestrea una menor región de la quimioteca en términos de espacio, por lo que se le "escapan" celdas que contienen compuestos que pudieran elevar estos resultados a los obtenidos para las otras tres funciones (regiones en rosa en la Figura 3.14).

Precisamente, evaluando el recubrimiento en términos de espacio (Figuras 3.13, 3.15 y 3.17) se observa cómo *cell-based Density* es la función menos efectiva, especialmente para BIB_Oxo y BIB_Hidro, partidas en un mayor número de celdas ocupadas. En general, los recubrimientos en términos de espacio son inferiores (entorno a un máximo del 80%) a los valores obtenidos en términos de población.

Significativamente, destaca el comportamiento ligeramente superior de la función *cell-based Chi2* frente al *cell-based Fraction*. Dado que la función *cell-based Fraction* optimiza precisamente el criterio de evaluación en espacio (ecuación [1.103]) estos resultados parecen contradictorios. Sin embargo, debe tenerse en cuenta el hecho de que la optimización/selección se realiza en una partición cuyo número de celdas "se ajusta" al tamaño de selección requerido, y que por lo tanto difiere de la partición utilizada en la evaluación de recubrimiento (141, 103 y 128 *bins*). Por ejemplo, para seleccionar un *full array* 5×5 el algoritmo *Optimum Binning* divide el espacio químico en un número de celdas ocupadas inferior o igual al número de compuestos a seleccionar. En este caso, correspondería a un máximo de 25 celdas ocupadas, mientras que el recubrimiento se evalúa en particiones de 141, 103 y 128 *bins* para cada quimioteca. Así este criterio es altamente dependiente de la partición realizada. Por otra parte, el criterio *cell-based Chi2* parece mostrarse más robusto frente a la partición, lo que explica que en las tres quimiotecas se obtengan mejores resultados. Así, el ranking relativo de los cuatro criterios en términos de espacio corresponde a: *cell-based Chi2* > *cell-based Fraction* > *cell-based Entropy* > *cell-based Density*. Por ello, se propone la función *cell-based Chi2* como la adecuada para realizar la selección del *full array* a sintetizar. Según este criterio, se alcanzan perfiles de recubrimiento en espacio similares para las tres quimiotecas.

De hecho, la región combinatoria que interesa desde el punto de vista de síntesis es la próxima a los 100 compuestos, por lo que se analizan en detalle las 21 configuraciones combinatorias en esta región, mostradas en la Tabla 3.11.

Tabla 3.11. 21 configuraciones con tamaño de selección próximo a 100, consideradas como posibles elecciones a ser sintetizadas.

10×10 (100)	13×8 (104)	16×6 (96)	19×5 (95)	8×12 (96)	7×15 (105)	5×18 (90)
11×9 (99)	14×7 (98)	17×6 (102)	20×5 (100)	8×13 (104)	6×16 (96)	5×19 (95)
12×8 (96)	15×7 (105)	18×5 (90)	9×11 (99)	7×14 (98)	6×17 (102)	5×20 (100)

En la Tabla 3.12 se muestran el valor promedio y la desviación estándar (en paréntesis) del porcentaje de recubrimiento en población y espacio para las 21 configuraciones mostradas en la Tabla 3.11.

Tabla 3.12. Promedio y desviación estándar (en paréntesis) de los valores de recubrimiento.

	BIB_Oxo		**BIB_Amino**		**BIB_Hidro**	
	%SPC	%POP	%SPC	%POP	%SPC	%POP
Cell-based Chi2	59(2)	87(5)	64(9)	96(5)	61(2)	76(6)
Cell-based Density	26(2)	84(2)	25(2)	93(2)	26(2)	77(2)
Cell-based Entropy	45(2)	76(7)	45(2)	89(6)	48(2)	73(6)
Cell-based Fraction	53(2)	79(9)	53(2)	88(5)	54(2)	71(7)

En términos de población, el recubrimiento promedio de las 21 selecciones es muy similar entre los cuatro métodos, solapándose las desviaciones estándar. En términos de espacio, el criterio *cell-based Chi2* es el que presenta un recubrimiento máximo en todos los casos. Así, cuando la partición del espacio no es excesivamente homogénea, son preferibles las funciones que optimizan el espacio, ya que el recubrimiento en población también se satisface implícitamente.

Por otra parte, en las Figuras 3.13-3.18 se observa cómo no existe una tendencia característica de aumento/disminución del recubrimiento al "desequilibrar" alguno de los puntos de diversidad, R^1 ó R^4. Es decir, no se observan cambios significativos en el caso de tratarse de una selección 20×5 o una 5×20, sino únicamente la tendencia a incrementarse el recubrimiento por efecto de aumentar el tamaño de selección.

Finalmente, se estudia la frecuencia de inclusión en la selección de alguno de los fragmentos "activos" descritos en el apartado 3.2.3, presentes en las quimiotecas BIB_Oxo, BIB_Amino y BIB_Hidro. En estos *full arrays* no se fuerza su selección, pero aún y todo, se observa cómo sí se incluyen algunos de ellos, aunque sea por una cuestión probabilística, ya que la identidad del fragmento “activo” encontrado difiere a lo largo de las selecciones. Para las 21 combinaciones anteriores de la Tabla 3.11 y cada una de las quimiotecas, se muestra en la Tabla 3.13 el promedio de fragmentos "activos".

Tabla 3.13. Promedio del número de fragmentos "activos" incluidos en las 21 selecciones combinatorias.

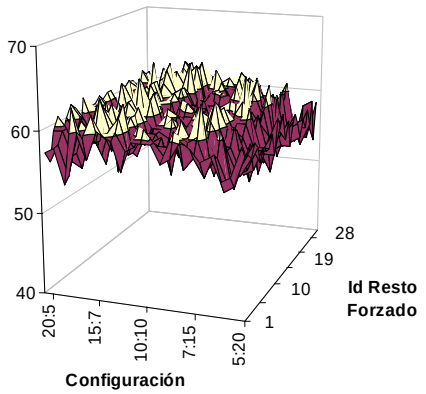
	BIB_Oxo		**BIB_Amino**		**BIB_Hidro**	
	R^1	R^4	R^1	R^4	R^1	R^4
Cell-based Chi2	0.67	0.48	0.14	0.24	0.62	0.43
Cell-based Density	0.57	0.57	0.33	1.14	0.43	1.52
Cell-based Entropy	1	0.52	0.19	0.57	0.28	0.52
Cell-based Fraction	0.86	0.81	0.14	0.43	0.24	0.71

3.5.3. Selecciones con las cuatro funciones objetivo forzando la inclusión de un fragmento "activo"

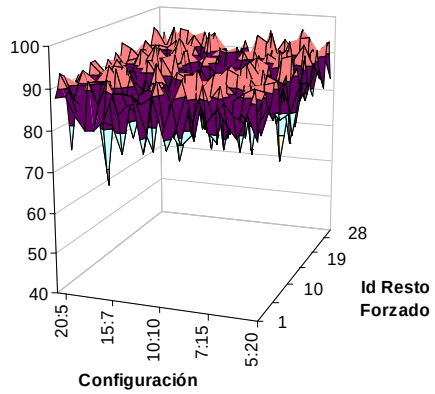
Para las 21 configuraciones *full array* de la Tabla 3.11 se repiten las selecciones forzando la inclusión de cada uno de los posibles fragmentos "activos" mostrados en las Tablas 3.7 y 3.8. Esto es, un total de 29 selecciones posibles para cada configuración (8 restos R^1 y 21 restos R^4).

En la Figura 3.19 se muestran los resultados de recubrimiento en términos de espacio y población para las selecciones optimizadas con el criterio *cell-based Chi2* (únicamente se incluyen las gráficas de este criterio a modo de ejemplo visual y por brevedad). En estas gráficas, los ejes del plano horizontal corresponden a: la configuración en cuestión (de las 21 expuestas en la Tabla 3.11) y la identidad del resto forzado en cada caso (del 1 al 21 para los

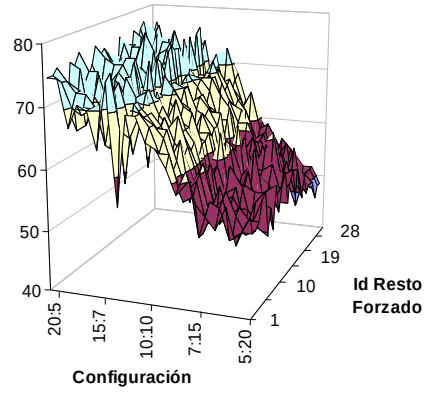
restos R^4 y del 22 al 29 para los restos R^1). El eje vertical representa el recubrimiento. La gradación en colores indica el rango de valores en el que se sitúa el recubrimiento de cada punto, según se especifica en la leyenda.



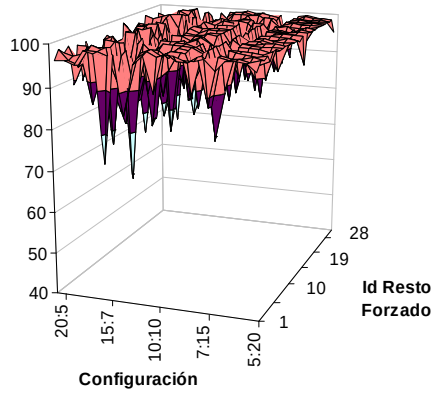
BIB_Oxo. SPC *cell-based Chi2*



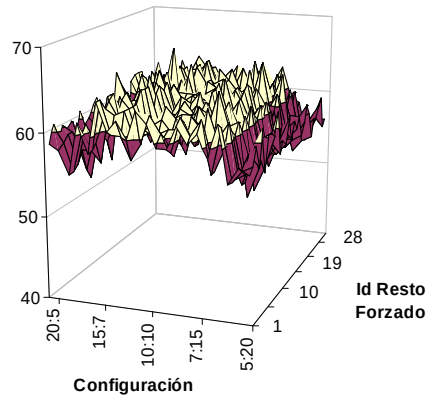
BIB_Oxo. POP *cell-based Chi2*



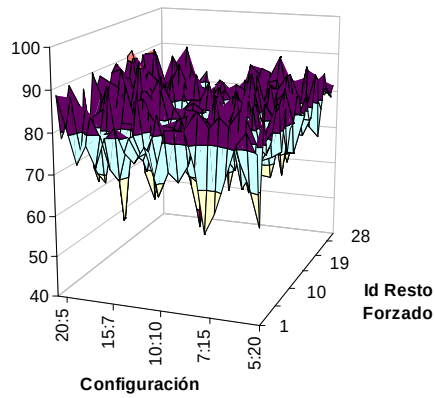
BIB_Amino. SPC *cell-based Chi2*



BIB_Amino. POP *cell-based Chi2*



BIB_Hidro. SPC *cell-based Chi2*



BIB_Hidro. POP *cell-based Chi2*

Figura 3.19. Evaluación del recubrimiento en espacio (SPC) y población (POP) para las 21 selecciones *full array* en las que se ha forzado la inclusión de 1 fragmento "activo" (R^1 ó R^4). *Cell-based Chi2* es el criterio a optimizar en todos los casos.

Se observa cómo no existe una tendencia significativa en el incremento/reducción del recubrimiento por la presencia particular de ninguno de los fragmentos reactivos. Por otra parte, se destaca también la baja influencia en el recubrimiento causada por el desequilibrado de alguno de los puntos de diversidad en las configuraciones para las quimiotecas BIB_Oxo y BIB_Hidro en esta área de selección. Sin embargo, en el caso particular de BIB_Amino, la configuración desequilibrada de R^1 frente a R^4 (20×5 frente a 5×20) incrementa el recubrimiento en términos de espacio.

De hecho, realizando un *zoom* de la gráfica de la Figura 3.15 correspondiente a la evaluación en espacio del criterio *cell-based Chi2*, para esta región de selección, se observa el mismo efecto de aumento en el recubrimiento de espacio al desequilibrar la proporción a favor de un mayor número de guanidinas.

En la Tabla 3.14 se muestran los valores de recubrimiento promedio en espacio y población y sus desviaciones estándar (en paréntesis) para las 21 configuraciones × 29 fragmentos fijados optimizadas por los cuatro criterios y para cada una de las tres quimiotecas.

Tabla 3.14. Promedio y desviación estándar (en paréntesis) de los valores de recubrimiento para las 21 configuraciones × 29 fragmentos de cada una de las quimiotecas.

	BIB_Oxo		**BIB_Amino**		**BIB_Hidro**	
	%SPC	%POP	%SPC	%POP	%SPC	%POP
Cell-based Chi2	59(2)	87(6)	63(8)	95(5)	60(2)	78(7)
Cell-based Density	26(2)	83(2)	25(2)	92(3)	27(2)	77(2)
Cell-based Entropy	45(2)	78(7)	53(6)	88(7)	47(2)	71(7)
Cell-based Fraction	53(2)	80(7)	62(6)	91(7)	54(2)	73(7)

Comparativamente frente a la Tabla 3.12, se observa cómo no existen apenas diferencias en los valores de recubrimiento obtenidos. Sin embargo, de cara a comparar la actividad de la serie de compuestos **50** con la serie **51**, resultaría interesante disponer de productos que presenten sustituyentes idénticos a los encontrados en las moléculas activas de la serie **51**, constituyendo éstos una posible referencia.

Es por ello que se continúa analizando la influencia en la diversidad de la quimioteca al forzar la inclusión de un número mayor de fragmentos "activos" en las selecciones *full array*. Así, basándose en las conclusiones de los estudios SAR expuestas anteriormente, se fuerza la inclusión de los dos restos R^4 procedentes de ésteres α,β-insaturados mostrados en la Figura 3.20. Por otra parte, se fijan también los dos restos R^1 de la Figura 3.20, ya que son las dos derivatizaciones con una mayor actividad en la serie **51** de las ocho guanidinas incluidas en las quimiotecas iniciales.

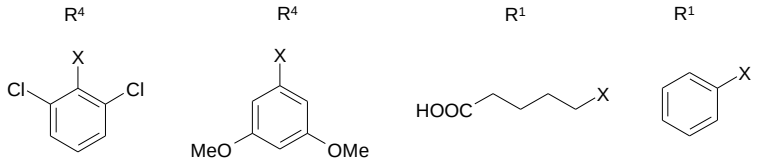


Figura 3.20. Cuatro fragmentos "activos" forzados incluidos en las selecciones.

En la Tabla 3.15 se muestran los valores promedio de recubrimiento de espacio y población junto con sus desviaciones estándar (en paréntesis) para las 21 selecciones *full array* en las que se han fijado los 2×2 fragmentos "activos". Se muestra también el promedio de fragmentos "activos" R^1 y R^4 presentes en estas 21 configuraciones, evidentemente, el valor mínimo es de 2.

Tabla 3.15. Promedio y desviación estándar (en paréntesis) de los valores de recubrimiento para las 21 configuraciones en las que se han fijado los 2×2 fragmentos "activos" de la Figura 3.20. Promedio de fragmentos "activos" R^1 y R^4 presentes en estas 21 configuraciones.

	BIB_Oxo				**BIB_Amino**				**BIB_Hidro**			
	%SPC	%POP	R^1	R^4	%SPC	%POP	R^1	R^4	%SPC	%POP	R^1	R^4
Cell-based Chi2	57(3)	86(5)	2.38	2.28	59(9)	93(9)	2.00	2.29	57(2)	82(6)	2.38	2.28
Cell-based Density	26(2)	82(2)	2.29	3.05	24(2)	92(2)	2.29	2.62	26(2)	77(2)	2.29	3.04
Cell-based Entropy	46(2)	79(6)	2.33	2.56	51(6)	90(6)	2.19	2.43	45(2)	74(4)	2.33	2.52
Cell-based Fraction	53(2)	81(5)	2.19	2.24	59(6)	91(8)	2.05	2.29	52(2)	78(6)	2.19	2.24

De nuevo, no se observa una pérdida de recubrimiento comparado con las selecciones "libres" (Tabla 3.12) o con las selecciones en las que se ha forzado la inclusión de como mínimo un fragmento "activo" (Tabla 3.14).

3.5.4. Selección Final de quimiotecas candidatas a sintetizarse

A partir de las conclusiones de los apartados anteriores, la elección de las quimiotecas combinatorias candidatas a ser sintetizadas para cada quimioteca por separado se basa en:

- Selecciones optimizadas bajo el criterio *cell-based Chi2*, ya que es el que presenta mejores recubrimientos en términos de espacio y población.

- Configuraciones equilibradas. Por un lado, no se ha encontrado una tendencia significativa en el incremento de la diversidad al favorecer la presencia de un determinado número de restos R^1 frente a R^4 (o viceversa), exceptuando el caso de BIB_Amino para la región de interés. Además, desde un punto de vista sintético (accesibilidad y compra de reactivos) resulta más atractivo. Así, se opta por las configuraciones *full array* equilibradas con un tamaño total próximo a 100: 10×10, 11×9 y 9×11.

- De los tres esquemas analizados: i) sin forzar ningún fragmento (libres), ii) forzando un fragmento de un punto de diversidad y iii) forzando 2×2 fragmentos; se consideran los esquemas i) y iii). El segundo se descarta, ya que los recubrimientos obtenidos son muy similares en todos los casos. El primero se mantiene como referencia de diversidad al extrapolar la selección de una quimioteca en otra (véase abajo) y el tercero se mantiene por el interés de disponer de productos "referencia" de actividad comparativa frente a la serie de compuestos **51**.

Con ello, se comparan para cada quimioteca (BIB_Oxo, BIB_Amino y BIB_Hidro) seis selecciones: 10×10 libre, 11×9 libre, 9×11 libre, 10×10 con 2×2 fijados, 11×9 con 2×2 fijados y la 9×11 con 2×2 fijados.

Cada una de las selecciones se ha optimizado de manera independiente en una quimioteca concreta de las tres posibles (BIB_Oxo, BIB_Amino y BIB_Hidro), por lo que los reactivos presentes en la selección óptima sobre una de ellas no tienen porqué coincidir con aquellos contenidos en la selección óptima de otra quimioteca. Con el objetivo de que la lista de reactivos sea común a las tres quimiotecas, en una primera aproximación (véase a continuación), se opta por evaluar el recubrimiento de cada uno de los *full arrays* optimizados para una quimioteca en las otras dos quimiotecas. Estos resultados se muestran en la Tablas 3.16-3.18, en las que también se especifica el número total de fragmentos "activos" que presentan.

Tabla 3.16. Recubrimiento en espacio y población para las 6 selecciones *full array* optimizadas sobre **BIB_Oxo**. Se muestra también el número de fragmentos “activos” presentes en cada una de ellas para R^1 y R^4.

	Selección realizada en el espacio de BIB_Oxo							
	BIB_Oxo		BIB_Amino		BIB_Hidro		R^1	R^4
	%SPC	%POP	%SPC	%POP	%SPC	%POP		
Chi2-10×10	60	87	33	90	34	77	2	-
Chi2-11×9	60	83	17	63	20	58	-	-
Chi2-9×11	60	81	24	65	24	57	-	-
Chi2-10×10 (2×2 fijados)	*58*	*88*	*35*	*94*	*26*	*69*	*4*	*2*
Chi2-11×9 (2×2 fijados)	57	88	34	93	28	72	3	3
Chi2-9×11 (2×2 fijados)	57	86	37	93	32	70	2	2

Tabla 3.17. Recubrimiento en espacio y población para las 6 selecciones *full array* optimizadas sobre **BIB_Amino**. Se muestra también el número de fragmentos "activos" presentes en cada una de ellas para R^1 y R^4.

	Selección realizada en el espacio de BIB_Amino							
	BIB_Oxo		**BIB_Amino**		BIB_Hidro		R^1	R^4
	%SPC	%POP	%SPC	%POP	%SPC	%POP		
Chi2-10×10	43	76	65	99	41	70	-	-
Chi2-11×9	33	64	70	99	43	73	-	-
Chi2-9×11	40	71	65	98	40	61	-	-
Chi2-10×10 (2×2 fijados)	*38*	*77*	*61*	*98*	*49*	*80*	*2*	*2*
Chi2-11×9 (2×2 fijados)	37	73	64	98	42	67	2	3
Chi2-9×11 (2×2 fijados)	37	65	58	99	41	65	2	2

Tabla 3.18. Recubrimiento en espacio y población para las 6 selecciones *full array* optimizadas sobre **BIB_Hidro**. Se muestra también el número de fragmentos "activos" presentes en cada una de ellas para R^1 y R^4.

	Selección realizada en el espacio de BIB_Hidro							
	BIB_Oxo		BIB_Amino		**BIB_Hidro**		R^1	R^4
	%SPC	%POP	%SPC	%POP	%SPC	%POP		
Chi2-10×10	32	70	43	90	62	72	-	1
Chi2-11×9	24	69	34	89	61	79	1	1
Chi2-9×11	26	68	45	90	61	68	1	1
Chi2-10×10 (2×2 fijados)	*29*	*74*	*36*	*97*	*59*	*83*	*2*	*2*
Chi2-11×9 (2×2 fijados)	36	75	53	95	59	74	2	2
Chi2-9×11 (2×2 fijados)	27	72	36	87	58	74	2	2

En primer lugar, se observa cómo la selección *full array* optimizada en una de las tres quimiotecas pierde un recubrimiento significativo en términos de espacio al extrapolarse a las otras dos quimiotecas. En términos de población se encuentran las tres situaciones: i) apenas no varían (por ejemplo, para el caso de las selecciones de BIB_Hidro en BIB_Oxo, Tabla 3.18), ii) se reduce el recubrimiento significativamente (por ejemplo, para el caso de las selecciones de BIB_Amino en las otras dos) ó iii) se incrementa (por ejemplo, para las selecciones de BIB_Hidro en BIB_Oxo o de BIB_Oxo en BIB_Hidro). Dado que la partición en celdas no es especialmente homogénea en ninguno de los casos, la pérdida de recubrimiento en espacio es relativamente tolerable. Además, debe considerarse que el recubrimiento máximo teórico (correspondiente a una selección *cherry picking* de 100 compuestos) en las particiones de evaluación consideradas para cada una de las quimiotecas de 100 compuestos es 71% (100/141) para BIB_Oxo, 97% (100/103) para BIB_Amino y 67% (100/148) para BIB_Hidro.

En cualquier caso, para cada una de las seis selecciones mostradas para cada quimioteca, se opta por las tres en formato 10×10 con los 2×2 restos fijados (mostradas en cursiva en cada una de las tablas). El motivo es que no representan una pérdida de recubrimiento significativo frente a las selecciones en las que no se ha forzado ningún reactivo (libres), priorizándose la inclusión de fragmentos "activos" y el formato más equilibrado del número de reactivos (10×10). Finalmente, entre estas tres candidatas, se escoge la selección realizada sobre BIB_Oxo ya que contiene cuatro fragmentos "activos". Los compuestos que los componen se muestran en el Anexo 1. Para más información del proceso de síntesis, referirse a la quimioteca [361].

El diseño realizado, en el que se optimizan conjuntos de manera independiente para cada quimioteca, presenta dos desventajas: i) los sustituyentes propuestos para cada una de ellas pueden diferir y ii) las diferencias de diversidad observadas en las Tablas 3.16-3.18 al extrapolar una quimioteca a otra.

Una alternativa para superar estas dos desventajas es la optimización simultánea de las tres quimiotecas, cada una en su espacio químico, forzando a que presenten reactivos comunes que maximicen de manera independiente cada recubrimiento. En esta optimización multiobjetivo, la función a optimizar corresponde a la suma ponderada de las contribuciones de cada una de las quimiotecas. En el momento de realizar las selecciones aquí presentadas no se disponía de esta funcionalidad, pero debido a su utilidad, se decide proceder a su implementación en el programa PRALINS y comparar ambas situaciones (no confundir con el criterio de Pareto presentado en el apartado 8.4).

En la Tabla 3.19 se muestran los recubrimientos en términos de espacio para cada quimioteca obtenidos al realizar i) tres selecciones 10×10 optimizándose el criterio *cell-based Fraction* de manera independiente para cada quimioteca (como en el caso anterior) y ii) una selección multiobjetivo 10×10 en las tres quimiotecas bajo el criterio *cell-based Fraction*.

Para facilitar el proceso y dado que únicamente se pretende mostrar los efectos de la optimización multiobjetivo, se realiza el cálculo con los componentes principales calculados sobre el total de la quimioteca (106920 compuestos), por lo que las particiones de referencia obtenidas para cada quimioteca difieren de las anteriormente mostradas y de ahí el hecho que el recubrimiento sea superior al mostrado anteriormente. Así, en este caso el número de celdas ocupadas para cada quimioteca es de 95 (BIB_Oxo), 91 (BIB_Amino) y 93 (BIB_Hidro). Además, el criterio de partición óptima que establecen PRALINS y Cerius2 para realizar la selección difiere. En cualquier caso, se observa cómo la optimización multiobjetivo, si bien pierde algo de recubrimiento respecto a cada una de las quimiotecas individuales (76 *vs* 80, 81 *vs* 83 y 79 *vs* 83), impide la pérdida de recubrimiento al extrapolar las selecciones realizadas sobre una quimioteca concreta a las dos restantes.

Tabla 3.19. Recubrimiento en espacio para tres selecciones independientes realizadas sobre cada una de las quimiotecas y para una selección multiobjetivo que maximiza el recubrimiento en las tres quimiotecas a la vez, escogiendo reactivos comunes.

	SPC (BIB_Oxo)	**SPC (BIB_Amino)**	**SPC (BIB_Hidro)**
Selección Multiobjetivo	76	81	79
Selección óptima en BIB_Oxo	**80**	69	67
Selección óptima en BIB_Amino	66	**83**	67
Selección óptima en BIB_Hidro	65	67	**83**

Capítulo 4.

Cribado por métodos basados en ligandos

Una vez establecida una metodología de selección de compuestos basada en diversidad, se adopta la estrategia *ligand-based* para la identificación de potenciales inhibidores de tirosina quinasas ATP-competitivos. Particularmente, se estudian y comparan tres métodos de búsqueda farmacofórica:

i) Búsquedas de similitud basadas en las representaciones en vectores de correlación de puntos potenciales farmacofóricos, CATS3D (apartado 1.6.5)

ii) Modelo farmacofórico manual generado con el programa MOE (apartado 1.7.1)

iii) Modelo farmacofórico *fuzzy* según la metodología SQUID (apartado 1.7.2)

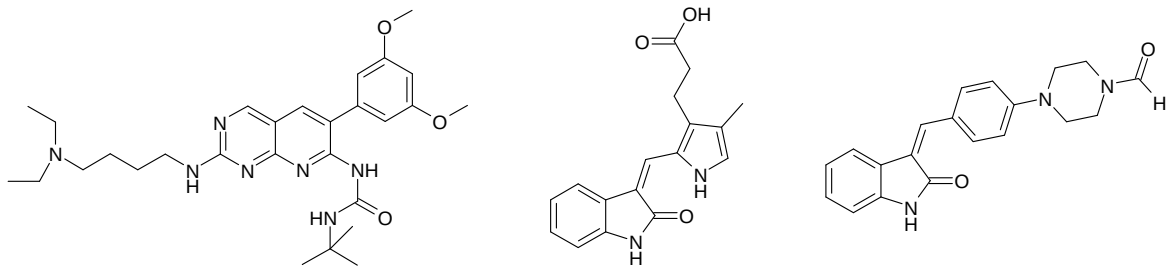
Tras la validación retrospectiva de las tres aproximaciones, éstas se han aplicado en la caracterización de los 106920 compuestos de las tres quimiotecas BIB_Oxo, BIB_Amino y BIB_Hidro descritas en el capítulo anterior.

Actualmente, existen varias estructuras co-cristalizadas de tirosina quinasas con diversos inhibidores en los que éstos se unen a la zona de unión de la adenina del ATP mediante un motivo compuesto de uno, dos o tres puentes de hidrógeno. Este modelo de interacción se ha utilizado en la generación de modelos farmacofóricos sencillos, como el publicado por Lyne *et al*[369] para la identificación de inhibidores de la CheckPoint quinasa 1 (chk 1). Estos autores aplican un filtro basado en la presencia de un dador y un aceptor de puente de hidrógeno comprendidos entre sí en un rango de distancias de 1.35-2.40 Å. Aunque probablemente muy aplicados en VS, no existen muchos modelos farmacofóricos publicados, a excepción de uno presentado en el año 2004 por Aronov *et al*[370]. Este modelo, definido por cinco puntos farmacofóricos identifica inhibidores "promiscuos", aquellos con actividad en el rango nano- y micromolar frente a varias quinasas con diversidad funcional. Además, el modelo distingue también entre estos inhibidores "promiscuos" y otros inhibidores con mayor selectividad frente a quinasas. Sin embargo, los datos disponibles publicados no permiten reproducir este modelo con fines comparativos.

4.1. Bases de Datos utilizadas en la validación retrospectiva

Como se ha expuesto en el capítulo 3, el interés del proyecto se focaliza inicialmente en los sistemas pirido[2,3-*d*]pirimidínicos descritos en el apartado 2.5, y que inhiben principalmente RTKs como FGFR, PDGFR, EGFR y NRTKs como c-Src.

Precisamente, se ha publicado la estructura del FGFR1 co-cristalizado con un derivado pirido[2,3-*d*]pirimidínico, concretamente el PD173074 (**17,** entrada PDB 2fgi) (Figura 4.1)[371]. Por ello, la compilación de un *pool* de activos para validar retrospectivamente los tres métodos citados se focaliza inicialmente en los inhibidores ATP-competitivos de FGFR.



PD 173074 **17** | SU5402 **29** | SU4984 **61**

Figura 4.1. Estructuras de inhibidores co-cristalizados con FGFR.

Se han seleccionado compuestos con actividades en el orden nano y micromolar (IC_{50}<50μM), evitándose los compuestos agregantes descritos en la referencia [360]. En total, la base de datos contiene un total de 288 estructuras, representativas de seis *scaffolds* distintos: pirido[2,3-*d*]pirimidinas[342,347], 7-alquilurea pirido[2,3-*d*]pirimidinas[343-345], naftiridin-2(1*H*)-onas[359], indolin-2-onas[354-355], 1-fenilbenzimidazoles[357] y 4-anilinoquinazolinas[337,372]. Las estructuras de estos *scaffolds*, así como el número total de compuestos de cada uno de ellos (en paréntesis) se muestran en la Figura 4.2. En la referencia [373] puede encontrarse una revisión de inhibidores de FGFR.

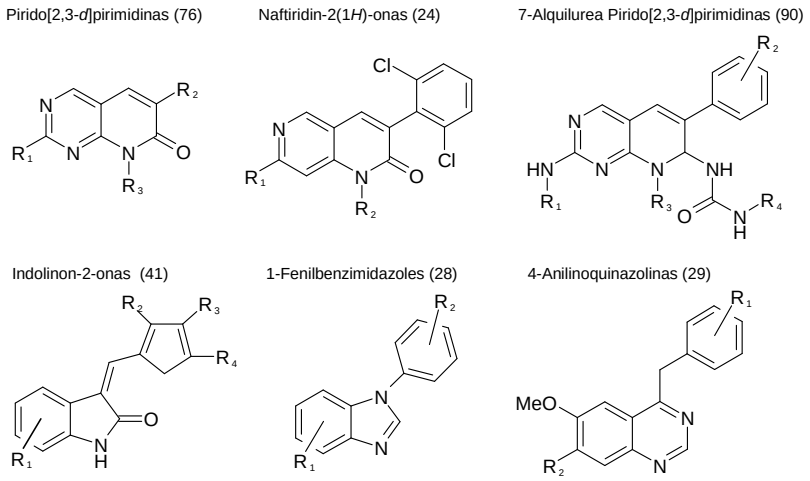


Figura 4.2. Seis *scaffolds* incluidos en la base de datos de inhibidores de FGFR ATP-competitivos.

Estos compuestos no muestran un perfil de selectividad perfecto, por lo que también son inhibidores, con una IC_{50} variable dentro del rango descrito, de PDGFR, EGFR, c-Src y, en algunos casos, de VEGFR. En el Anexo 2 se encuentran tabuladas las 288 estructuras y sus valores de IC_{50} correspondientes. En adelante, este *pool* de activos se denomina ACTIV_1.

Además, se han filtrado 452 estructuras indexadas como inhibidores de tirosina quinasas de la base de datos COBRA, versión 4-3[374]. De ellas, se mantienen únicamente aquellas estructuras que son inhibidores ATP-competitivos de los RTKs y de los NRTKs de la familia Src y que no se encuentran presentes en el *pool* ACTIV_1. Así, se obtiene un segundo *pool* de activos, que se denominará COBRA, formado por de 104 compuestos.

El principal motivo de dividir el set de inhibidores en dos *pools* diferentes reside en que, de este modo, se dispone de un conjunto (ACTIV_1) convenientemente construido para realizar un análisis de diversidad de los compuestos identificados por los tres modelos (véase abajo).

En ambos casos, el *pool* de compuestos inactivos usado en el ensayo retrospectivo está compuesto por los restantes 4931 compuestos de COBRA no anotados como inhibidores de tirosina quinasas. Como en la mayor parte de validaciones retrospectivas, la inactividad de estos compuestos frente a TKs se presupone. Así, los dos *pools* de activos, ACTIV_1 y COBRA, se diluyen independientemente en este *pool* de inactivos para generar las bases de datos Base_ACTIV_1 y Base_COBRA.

Tal y como se ha comentado en el apartado 1.7.2, el modelo SQUID requiere la optimización del radio de cluster (*cluster radius*) y de una serie de pesos adicionales (*feature-type weights*) para cada caso de estudio. Para agilizar los cálculos, así como para obtener una significatividad estadística, se generan cuatro sets aleatorios a partir de ACTIV_1 de manera que contengan 20 compuestos representativos de cada *scaffold*. Éstos, se denominan Base_ACTIV_1_ale1 … Base_ACTIV_1_ale4. Cada uno de ellos, se diluye en un total de 2000 compuestos escogidos también aleatoriamente del *pool* de 4931 inactivos. Esta proporción se ha escogido de manera que se mantenga aproximadamente la relación inicial en ACTIV_1 (120/2120≈288/5219). Para realizar esta selección aleatoria se prepara un pequeño programa en C.

En la Tabla 4.1 se muestran las relaciones entre el número de activos, inactivos y el número total en cada base de datos.

Tabla 4.1. Resumen de la composición de cada una de las bases de datos.

Base de Datos	#Activos	#Inactivos	Total	Activos/Total
Base_ACTIV_1	288	4931	5219	0.055
Base_COBRA	104	4931	5035	0.021
Cada set aleatorio: Base_ACTIV_1_ale1…Base_ACTIV_1_ale4	120	2000	2120	0.057

En cada uno de los casos, se genera una base de datos multiconformacional con el programa MOE, calculándose hasta un máximo de 50 conformaciones por molécula, utilizando los parámetros del programa por defecto. Cada una de las conformaciones se minimiza con el *force field* MMFF94 y parámetros por defecto. El número de confórmeros totales es: 8958 (*pool* de activos de ACTIV_1), 9109 (*pool* de activos COBRA) y 205529 (*pool* de inactivos). Excepto en los casos indicados, se trabaja siempre con las bases de datos multiconformacionales.

4.2. Plantillas utilizadas en la generación de los modelos farmacofóricos

Además de PD173074, actualmente (mayo 2006) se encuentran depositadas en el *Protein Data Bank*[60] otras dos estructuras tipo indolin-2-ona co-cristalizadas con el dominio TK del FGFR1: SU5402 (entrada PDB 1fgi) y SU4984 (entrada PDB 1agw)[375] (Figura 4.1). Ambos compuestos inhiben la autofosforilación de FGFR inducida por FGF con IC_{50} de 10-20 μM (SU5402) y 20-40 μM (SU4984).

Para analizar la influencia del alineamiento inicial y las plantillas en la diversidad de los *scaffolds* identificados, se planea construir alineamientos farmacofóricos "incrementados": incluyendo consecutivamente una molécula representativa de cada una de las seis clases contenidas en ACTIV_1 (Figura 4.2).

Evidentemente, para la clase 7-alquilurea pirido[2,3-*d*]pirimidina, se escoge la estructura co-cristalizada de PD173074. Dado que SU5402 (**29**, Figura 4.1) y SU4948 (**61**) pertenecen a la misma familia estructural, se selecciona SU5402 debido a su mayor potencia frente a FGFR.

Para las otras cuatro clases restantes sin estructura cristalizada en FGFR, la selección de una plantilla se basa en la actividad de los compuestos y en la presencia de sustituyentes similares a los presentes en las estructuras de los compuestos co-cristalizados. Estas cuatro plantillas se muestran en la Figura 4.3.

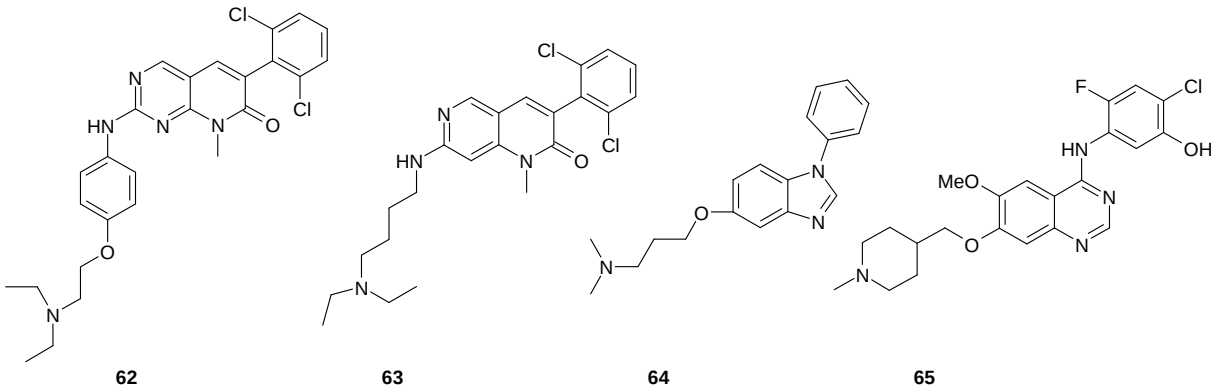


Figura 4.3. Cuatro compuestos incluidos como plantillas en el alineamiento inicial.

4.3. Alineamientos farmacofóricos iniciales

Como se ha mencionado en los apartados 1.7.1 y 1.7.2, el alineamiento inicial de los compuestos permanece "fijo" en la generación de modelos farmacofóricos con MOE y SQUID. Inicialmente, se alinean las estructuras co-cristalizadas de PD173074 y SU5402 por superposición rígida, mediante la herramienta de *homology alignment* de MOE. Para el resto de *scaffolds*, se emplea el módulo de alineamiento flexible de MOE (MOE-FlexAlign[287]), manteniendo PD173074 y SU5402 "fijados", respectivamente. Para ello, se escogen como términos de similitud las opciones *H-bond donor*, H-*bond acceptor*, *Aromaticity*, *Acid/Base*, *Hydrophobe*, *Partial Charge* y *Volume*. Se establece un máximo de 500 iteraciones, manteniéndose el resto de parámetros por defecto.

En la Figura 4.4.a se muestra la superposición de las estructuras cristalinas de PD173074 y SU5402. Para cada molécula, se indican las funcionalidades responsables del modo de unión similar al del ATP. SU5402[375] establece un puente de hidrógeno entre el N-1 de la indolinona y el oxígeno del carbonilo del Glu[562] y entre el O-2 y el nitrógeno amídico de Ala[564]. En el caso de PD173074[371], el N-3 actúa como aceptor del puente de hidrógeno correspondiente con el nitrógeno amídico de Ala[564], mientras que el nitrógeno del grupo butilamino actúa como dador con el oxígeno del carbonilo de Ala[564]. Este alineamiento se denomina en adelante MODEL2PDB. Curiosamente, el alineamiento flexible de SU5402 frente a PD173074 (fijado) resulta en una orientación totalmente opuesta (Figura 4.4.b), debido al peso de las interacciones aromáticas.

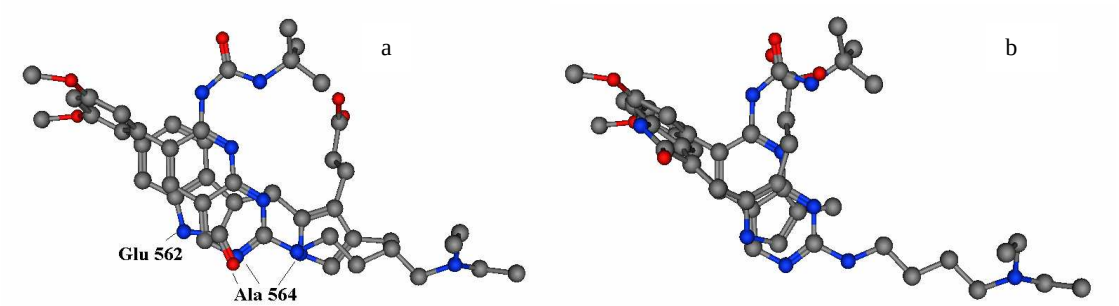


Figura 4.4. (a) Superposición rígida de las estructuras co-cristalizadas de PD173074 y SU5402. Alineamiento inicial MODEL2PDB. (b) Resultado del alineamiento flexible de SU5402 frente a PD173074.

Respecto al alineamiento flexible de los cuatro compuestos de la Figura 4.3 frente a PD173074, únicamente se obtiene un resultado fiable para el caso de los compuestos **62** y **63** (Figura 4.5.a). Por otra parte, el compuesto **64** es el único que se alinea de manera satisfactoria frente a SU5402 (Figura 4.5.b). En el caso del compuesto **65**, el alineamiento es deficiente frente a ambos compuestos, PD173074 y SU5402, respectivamente. El criterio de bondad de alineamiento se basa en la comparación de la superposición de aquellos grupos funcionales que puedan interactuar de manera análoga a como lo hacen las estructuras de referencia. Además, estos resultados se han constatado paralelamente con los obtenidos por *docking*, resultados que se discuten posteriormente en los apartados 5.1.2 y 5.1.3.

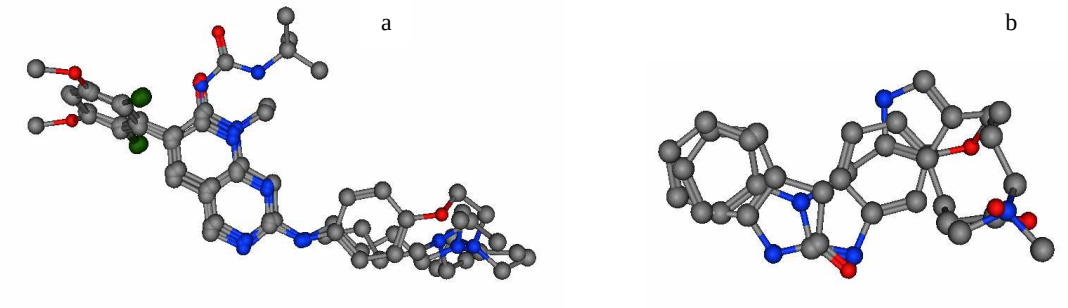


Figura 4.5. (a) Superposición flexible de **62** y **63** en PD173074. Alineamiento inicial MODEL3ALIGNED. (b) Superposición flexible de **64** sobre SU5402.

A la vista de estos resultados, se adopta también como alineamiento inicial el mostrado en la Figura 4.5.a en el que como *scaffolds* participan una pirido[2,3-*d*]pirimidina (**62**), una 7-alquilurea pirido[2,3-*d*]pirimidina (PD173074) y una naftiridina (**63**). En adelante, se refiere a dicho alineamiento como MODEL3ALIGNED.

Finalmente, se propone un tercer alineamiento, MODEL4ALIGNED, resultante de fusionar MODEL2PDB y MODEL3ALIGNED. Este alineamiento, mostrado en la Figura 4.6 se compone de cuatro de los seis *cores* presentes en el *pool* ACTIV_1. Por lo tanto, únicamente se mantienen dos clases (1-fenilbenzimidazoles y 4-anilinoquinazolinas) fuera del alineamiento.

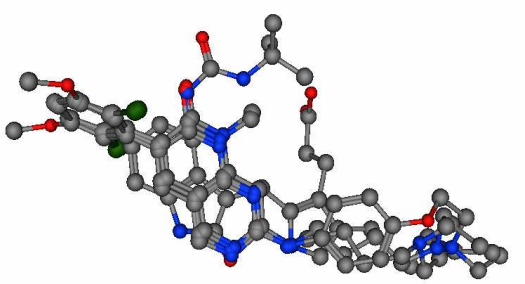


Figura 4.6. Alineamiento inicial MODEL4ALIGNED.

4.4. Métricas utilizadas para evaluar los *hits*

Para evaluar las listas de *hits* y la bondad de los métodos aplicados en cribado virtual se utilizan las métricas[376] mostradas en las ecuaciones [4.1]-[4.4], típicamente aplicadas en búsquedas con modelos farmacofóricos. En todos los casos, *N* corresponde al número de compuestos en la base de datos, *A* es el número de activos en la base de datos, H_t es el número de compuestos en la lista de *hits* (verdaderos positivos + falsos positivos) y H_a es el número de activos en la lista de *hits*. Estos valores pueden calcularse sobre el total de la base de datos, aunque normalmente se calculan para ciertos porcentajes de base de datos escaneada (típicamente para el 5-10% de la base de datos).

- Porcentaje de activos en la lista de *hits* (%Y)

$$\%\mathrm{Y} = \frac{H_a}{H_t} \times 100 \qquad [4.1]$$

A mayor valor, más selectivo es el modelo o hipótesis farmacofórica.

- Porcentaje de activos identificados (%A)

$$\%\mathrm{A} = \frac{H_a}{A} \times 100 \qquad [4.2]$$

A mayor valor, más recubrimiento. Normalmente, no es posible capturar todos los activos, por lo que los valores elevados de recubrimiento se suelen alcanzar a expensas de comprometer la selectividad (%Y).

- Enriquecimiento (*ef*)

$$ef = \left(\frac{H_a / H_t}{A / N} \right) \qquad [4.3]$$

Indica cuántas veces la lista de *hits* está más enriquecida en activos respecto a la base de datos inicial. Por ejemplo, un valor de dos representa que es dos veces más probable que se seleccione aleatoriamente un compuesto de la lista de *hits* que de la base de datos.

Normalmente, este es el parámetro utilizado en los ensayos de cribado virtual. Alternativamente, se suele acompañar de una curva de enriquecimiento correspondiente, en la que se representa el porcentaje de activos identificados en función del porcentaje de base de datos escaneada, ordenada según el valor de la función objetivo en cada caso.

En estas curvas, una línea trazada sobre la diagonal representa un ranking de los compuestos aleatorio.

- Bondad de la lista de *hits* (*Goodness of Hit list*, GH)

$$GH = \frac{H_a(3A + H_t)}{4H_t A} \times \left(1 - \frac{H_t - H_a}{N - A}\right) \qquad [4.4]$$

Este parámetro, introducido por Güner *et al*[376] para la comparación de modelos farmacofóricos, permite evaluar la bondad de la búsqueda farmacofórica estableciendo un compromiso entre las métricas %Y y %A, que normalmente compiten entre sí.

Aparte de evaluar los métodos en términos de enriquecimiento, tasa a la que se identifican los activos, es interesante establecer una medida de la diversidad o la habilidad para identificar *leads* diversos. Para ello, típicamente se determina el porcentaje de base de datos escaneada necesaria para identificar al menos un compuesto representativo de todos y cada uno de los quimiotipos presentes.

4.5. Modelos farmacofóricos del MOE

Para todos los alineamientos estudiados con MOE, se han aplicado dos esquemas farmacofóricos: PCH y PPCH_ALL (apartado 1.7.1). La herramienta *Pharmacophore Consensus* se ha utilizado para agrupar todos los puntos de anotación de los ligandos en *features*, considerando únicamente aquellas totalmente conservadas dentro de una región determinada por un radio de tolerancia. Se han explorado distintos valores de tolerancia.

Inicialmente, en la obtención y refinamiento manual del modelo farmacofórico con MOE se ha trabajado exclusivamente con un único subconjunto aleatorio de Base_ACTIV_1, Base_ACTIV_1_ale1, y con el *pool* de activos COBRA, formado por 104 compuestos.

4.5.1. MODEL2PDB y esquema PCH

En la Figura 4.7 se muestran los modelos obtenidos a partir del alineamiento MODEL2PDB y el esquema PCH: a) con una tolerancia de 1 (MODEL2PDB_tol_1) y b) con tolerancia de 1.2 (MODEL2PDB_tol_1.2). En la Tabla 4.2 se describen sus respectivas *features*.

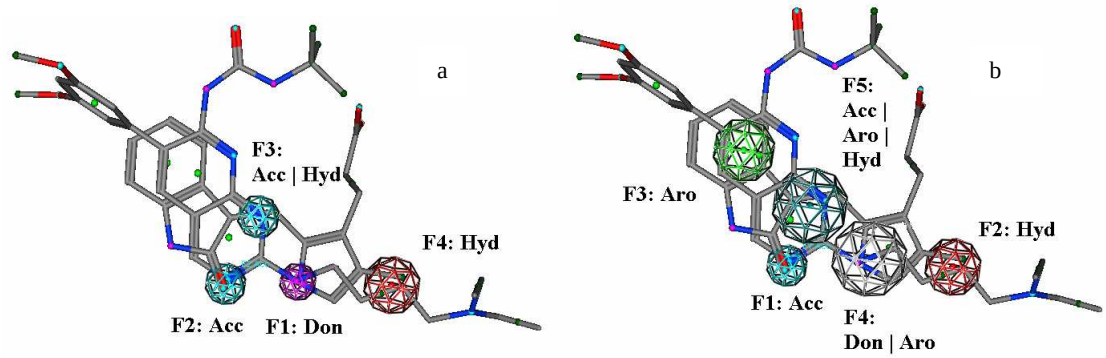


Figura 4.7. Modelos farmacofóricos obtenidos con el alineamiento MODEL2PDB y el esquema PCH con (a) tolerancia de 1 y (b) tolerancia de 1.2.

Tabla 4.2. Listado de *features* que describen los modelos farmacofóricos obtenidos a partir de MODEL2PDB con esquema PCH y con tolerancia de 1 y 1.2.

MODEL2PDB_tol_1			MODEL2PDB_tol_1.2		
ID_Feature	Radio	Expresión	ID_Feature	Radio	Expresión
F1	0.63	Dador	F1	0.70	Aceptor
F2	0.70	Aceptor	F2	0.99	Hidrofóbico
F3	0.72	Aceptor o Hidrofóbico	F3	1.07	Aromático
F4	0.99	Hidrofóbico	F4	1.31	Dador o Aromático
			F5	1.31	Aceptor o Aromático o Hidrofóbico

En la Tabla 4.3 se muestran los valores correspondientes de las métricas descritas en el apartado 4.4 para el cribado de Base_ACTIV_1_ale1. Los resultados se muestran en función del total de moléculas (*hits, H_t*) que satisfacen el modelo correspondiente, sin ordenarse en función de la RMSD (apartado 1.7.1). Se indica también el número de familias de *scaffolds* con al menos un representante incluido en la lista de *hits* (columna "#Sca"), siendo seis el máximo posible. En la columna "COBRA" se muestra el número de compuestos activos identificados del total de 104 inhibidores ATP-competitivos del *pool* COBRA.

Tabla 4.3. Enriquecimiento para los modelos derivados a partir del alineamiento MODEL2PDB y el esquema farmacofórico PCH. *#Sca*: número de familias de *scaffolds* con al menos un representante incluido en la lista de *hits*. *COBRA*: número de compuestos activos identificados del total de 104 inhibidores del *pool* COBRA. En cursiva se muestran las características de la base de datos inicial.

Modelo farmacofórico	Ha	Ht	%Y	%A	*ef*	GH	#Sca	COBRA
Base_ACTIV_1_ale1	*120*	*2120*	*5.66*	*100.00*	*1.00*	*0.00*	*6*	-
MODEL2PDB_tol_1	16	436	3.67	13.33	0.65	0.05	2	15
MODEL2PDB_tol_1.2	25	365	6.85	20.83	1.21	0.09	3	6
MODEL2PDB_tol_1.2_F3Aro|Hyd	45	694	6.48	37.50	1.15	0.10	5	39
MODEL2PDB_tol_1.2_F3Aro|Hyd_F4Donor	27	445	6.07	22.50	1.07	0.08	5	12
MODEL2PDB_tol_1.2_F3Aro|Hyd_F4Donor_NoF2Hyd	79	1247	6.34	65.83	1.12	0.09	6	68

A partir del mejor modelo, MODEL2PDB_tol_1.2, se optimizan diversas *features*:

- MODEL2PDB_tol_1.2_F3Aro|Hyd. Se fuerza a que el carácter de F3 sea Aromático ó Hidrofóbico.
- MODEL2PDB_tol_1.2_F3Aro|Hyd_F4Donor. Como el anterior, pero F4 es exclusivamente dador.
- MODEL2PDB_tol_1.2_F3Aro|Hyd_F4Donor_NoF2Hyd. Como el anterior, pero se ignora F2 (Hidrofóbico).

Los valores de las métricas obtenidas por estos modelos se muestran asimismo en la Tabla 4.3. En todos los casos, el enriquecimiento obtenido es pobre, apenas supera un enriquecimiento aleatorio (*ef* >1). Incrementando la tolerancia a 1.4, los radios de las *features* se amplían considerablemente, por lo que el modelo pierde una gran selectividad, reconociéndose como *hits* un total de 1823 de las 2120 moléculas presentes.

4.5.2. MODEL3ALIGNED y esquema PCH

En la Figura 4.8 se muestran los modelos obtenidos a partir del alineamiento MODEL3ALIGNED y el esquema PCH: a) con una tolerancia de 1 (MODEL3ALIGNED_tol_1) y b) con tolerancia de 1.2 (MODEL3ALIGNED_tol_1.2). En la Tabla 4.4 se describen sus respectivas *features*.

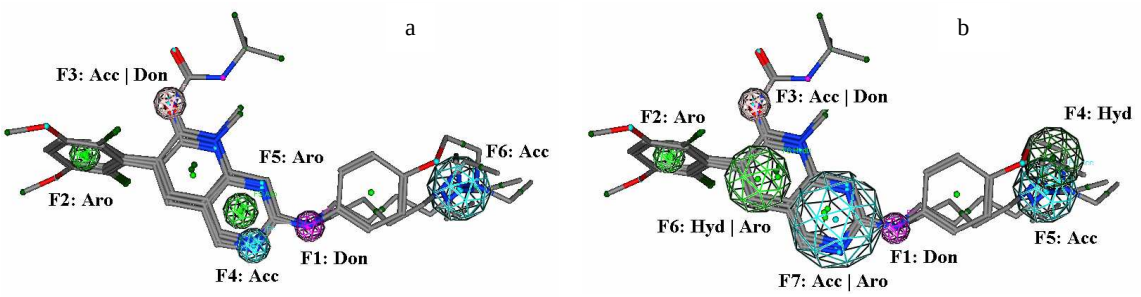


Figura 4.8. Modelos farmacofóricos obtenidos con el alineamiento MODEL3ALIGNED y el esquema PCH con (a) tolerancia de 1 y (b) tolerancia de 1.2.

Tabla 4.4. Listado de *features* que describen los modelos farmacofóricos obtenidos a partir de MODEL3ALIGNED con esquema PCH y con tolerancia de 1 y 1.2.

MODEL3ALIGNED_tol_1			**MODEL3ALIGNED_tol_1.2**		
ID_Feature	Radio	Expresión	ID_Feature	Radio	Expresión
F1	0.59	Dador	F1	0.59	Dador
F2	0.62	Aromático	F2	0.62	Aromático
F3	0.63	Aceptor o Dador	F3	0.63	Aceptor o Dador
F4	0.66	Aceptor	F4	1.19	Hidrofóbico
F5	0.68	Aromático	F5	1.26	Aceptor
F6	1.26	Aceptor	F6	1.28	Hidrofóbico o Aromático
			F7	1.87	Aceptor o Aromático

A partir del mejor modelo, MODEL3ALIGNED_tol_1, se optimizan diversas *features*:

- MODEL3ALIGNED_tol_1_NoF6. Se ignora F6 (aceptor).
- MODEL3ALIGNED_tol_1_NoF6_At_Least4. Como el anterior, pero requiriendo que como mucho se satisfagan cuatro de las cinco *features* presentes.

A partir del mejor modelo, MODEL3ALIGNED_tol_1.2, se optimizan diversas *features*:

- MODEL3ALIGNED_tol_1.2_At_Least_4. Se requiere que como mucho se satisfagan cuatro de las siete *features* presentes.
- MODEL3ALIGNED_tol_1.2_At_Least_5. Se requiere que como mucho se satisfagan cinco de las siete *features* presentes.
- MODEL3ALIGNED_tol_1.2_NoF4_NoF5. Se ignoran las *features* F4 y F5
- MODEL3ALIGNED_tol_1.2_NoF4_NoF5_No_F6. Se ignoran las *features* F4, F5 y F6.

Análogamente a la Tabla 4.3, se muestran en la Tabla 4.5 los resultados de la búsqueda farmacofórica sobre Base_ACTIV_1_ale1 con esta serie de modelos obtenidos a partir de MODEL3ALIGNED y con esquema farmacofórico PCH.

Tabla 4.5. Enriquecimiento para los modelos derivados a partir del alineamiento MODEL3ALIGNED y el esquema farmacofórico PCH. *#Sca*: número de familias de *scaffolds* con al menos un representante incluido en la lista de *hits*. *COBRA*: número de compuestos activos identificados del total de 104 inhibidores del *pool* COBRA. En cursiva se muestran las características de la base de datos inicial. En negrita aparece el modelo seleccionado en principio como óptimo de la serie.

Modelo farmacofórico	Ha	Ht	%Y	%A	*ef*	GH	#Sca	COBRA
Base_ACTIV_1_ale1	*120*	*2120*	*5.66*	*100.00*	*1.00*	*0.00*	*6*	-
MODEL3ALIGNED_tol_1.2	18	18	100.00	15.00	17.67	0.79	3	1
MODEL3ALIGNED_tol_1.2_At_Least_4	104	1774	5.86	86.67	1.04	0.04	6	99
MODEL3ALIGNED_tol_1.2_At_Least_5	**82**	**652**	**12.58**	**68.33**	**2.22**	**0.19**	**6**	**60**
MODEL3ALIGNED_tol_1.2_NoF4_NoF5	59	173	34.10	49.17	6.03	0.36	3	22
MODEL3ALIGNED_tol_1.2_NoF4_NoF5_No_F6	65	413	15.74	54.17	2.78	0.21	4	31
MODEL3ALIGNED_tol_1	20	20	100.00	16.67	17.67	0.79	3	1
MODEL3ALIGNED_tol_1_NoF6	59	62	95.16	49.17	16.81	0.84	3	7
MODEL3ALIGNED_tol_1_NoF6_At_Least4	60	252	23.81	50.00	4.21	0.27	3	28

Los modelos MODEL3ALIGNED_tol_1.2 y MODEL3ALIGNED_tol_1, pese a su óptima selectividad, son excesivamente restrictivos, "capturándose" un mínimo número de compuestos activos: 18 y 20, respectivamente, del total de 120. La eliminación de las *features* referidas a propiedades de las cadenas laterales, incrementa el número de activos identificados (H_a) y con ello el recubrimiento (%A). En este sentido, se considera como mejor modelo de esta serie MODEL3ALIGNED_tol_1.2_At_Least_5, ya que si bien no identifica tantos activos como MODEL3ALIGNED_tol_1.2_At_Least_4, su selectividad es mayor.

4.5.3. MODEL4ALIGNED y esquema PCH

En la Figura 4.9 se muestran los modelos obtenidos con el alineamiento MODEL4ALIGNED y el esquema PCH: a) con una tolerancia de 1 (MODEL4ALIGNED_tol_1) y b) con tolerancia de 1.2 (MODEL4ALIGNED_tol_1.2). En la Tabla 4.6 se describen sus respectivas *features*.

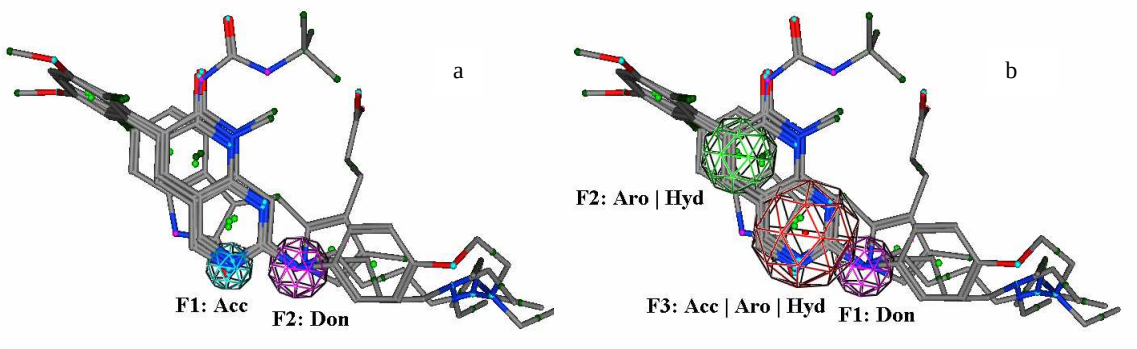


Figura 4.9. Modelos farmacofóricos obtenidos con el alineamiento MODEL4ALIGNED y el esquema PCH con (a) tolerancia de 1 y (b) tolerancia de 1.2.

Tabla 4.6. Listado de *features* que describen los modelos farmacofóricos obtenidos a partir de MODEL4ALIGNED con esquema PCH y con tolerancia de 1 y 1.2.

MODEL4ALIGNED_tol_1			MODEL4ALIGNED_tol_1.2		
ID_Feature	Radio	Expresión	ID_Feature	Radio	Expresión
F1	0.82	Aceptor	F1	1.06	Dador
F2	1.06	Dador	F2	1.34	Aromático o Hidrofóbico
			F3	1.90	Aceptor o Aromático o Hidrofóbico

En la Tabla 4.7 se muestran los resultados de la búsqueda farmacofórica sobre Base_ACTIV_1_ale1. Se observa que ambos modelos son demasiado poco selectivos, ya que contienen pocas *features* y por lo tanto, existe un elevado número de moléculas identificadas como *hits*.

Tabla 4.7. Enriquecimiento para los modelos derivados a partir del alineamiento MODEL4ALIGNED y el esquema farmacofórico PCH. *#Sca*: número de familias de *scaffolds* con al menos un representante incluido en la lista de *hits*. *COBRA*: número de compuestos activos identificados del total de 104 inhibidores del *pool* COBRA. En cursiva se muestran las características de la base de datos inicial. En negrita aparecen los dos modelos seleccionados como óptimos de la serie.

Modelo farmacofórico	Ha	Ht	%Y	%A	*ef*	GH	#Sca	COBRA
Base_ACTIV_1_ale1	*120*	*2120*	*5.66*	*100.00*	*1.00*	*0.00*	*6*	-
MODEL4ALIGNED_tol_1	102	1598	6.38	85.00	1.13	0.07	6	99
MODEL4ALIGNED_tol_1.2	106	1760	6.02	88.33	1.06	0.05	6	99
MODEL4ALIGNED_tol_1_F3Aro|Hyd	101	1541	6.55	84.17	1.16	0.07	6	76
MODEL4ALIGNED_tol_1_F3Aro|Hyd_F4Aro	101	1261	8.01	84.17	1.42	0.11	6	73
MODEL4ALIGNED_tol_1_F3Aro|Hyd_F4Aro_F6Aro|Hyd	66	465	14.19	55.00	2.51	0.20	4	38
MODEL4ALIGNED_tol_1_F3Aro|Hyd_F4Aro_F6Hyd|F7Aro	**65**	**360**	**18.06**	**54.17**	**3.19**	**0.23**	**4**	**31**
MODEL4ALIGNED_tol_1_F3Aro_F4Aro|Hyd_F5Hyd	**81**	**522**	**15.52**	**67.50**	**2.74**	**0.22**	**5**	**30**

Al contrario que en los otros dos alineamientos, en este caso la optimización del modelo se centra en añadir una serie de *features* que permitan aumentar la selectividad. Para ello, se parte de MODEL4ALIGNED_tol_1:

- MODEL4ALIGNED_tol_1_F3Aro|Hyd. Se añade F3 correspondiente a puntos potenciales farmacofóricos (PPPs) aromáticos o hidrofóbicos y con un radio de 1.7.

- MODEL4ALIGNED_tol_1_F3Aro|Hyd_F4Aro. Como el caso anterior, pero se incrementa con dos *features*: F4 (aromático o aceptor, radio 0.68) y F5 (hidrofóbico, radio 0.98). Además, se fija una *constraint* de manera que se satisfaga como mínimo una de estas dos (F4 o F5). (Figura 4.10.a).

- MODEL4ALIGNED_tol_1_F3Aro|Hyd_F4Aro_F6Aro|Hyd. Como el caso anterior, se añade F6 (aromático ó hidrofóbico, con radio 1.8). Además, se establece que como mínimo se satisfagan cinco *features*, siendo F1, F2 (con radio disminuido de 1.06 a 0.59) y F3 (con radio disminuido de 1.7 a 1.28) esenciales. (Figura 4.10.b).

- MODEL4ALIGNED_tol_1_F3Aro|Hyd_F4Aro_F6Hyd|F7Aro. La *feature* F6 añadida en el caso anterior se desglosa en dos términos: F6 (hidrofóbico, radio 1.2) y F7 (aromático, radio 1.2) y se impone la *constraint* de que se satisfaga al menos una de los dos. (Figura 4.10.c).

- MODEL4ALIGNED_tol_1_F3Aro_F4Aro|Hyd_F5Hyd. Presenta las *features* incluidas en la Tabla 4.8. (Figura 4.10.d).

Tabla 4.8. Listado de *features* del modelo MODEL4ALIGNED_tol_1_F3Aro_F4Aro|Hyd_F5Hyd

MODEL4ALIGNED_tol_1_F3Aro_F4Aro|Hyd_F5Hyd			
ID_Feature	Radio	Expresión	%Conservación
F1	0.82	Aceptor	100%
F2	0.59	Dador	100%
F3	0.68	Aromático	75%
F4	1.28	Hidrofóbico o Aromático	50%
F5	0.76	Hidrofóbico	75%

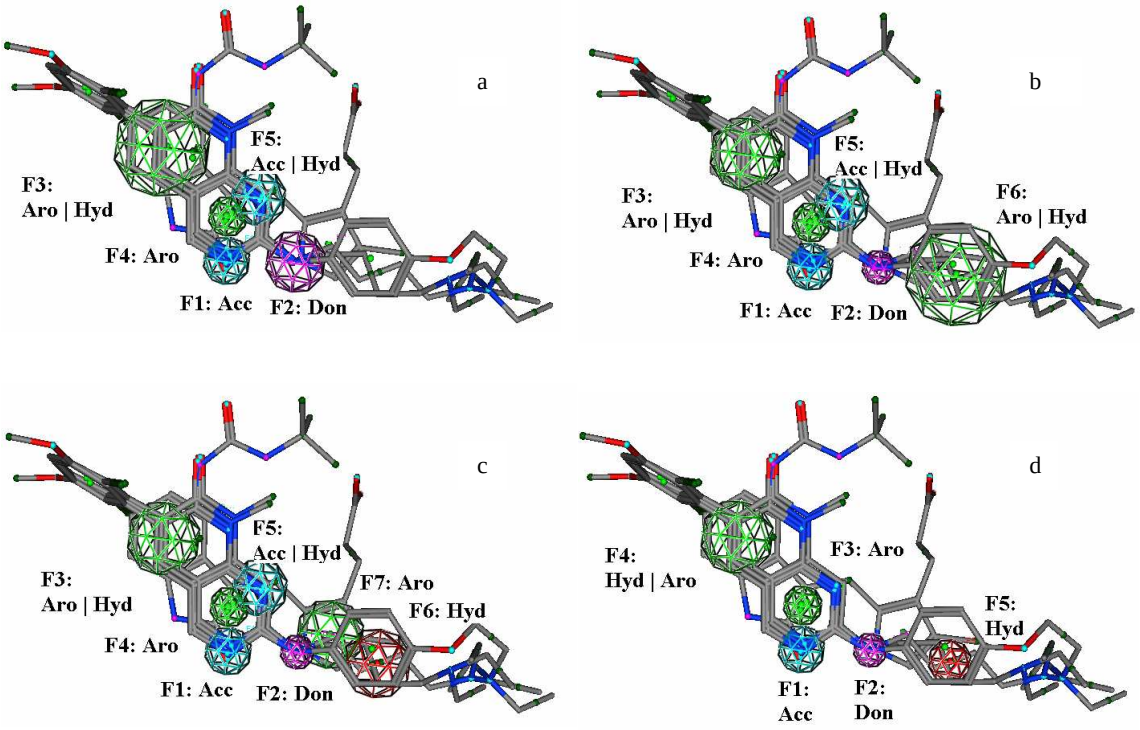


Figura 4.10. Modelos farmacofóricos (a) MODEL4ALIGNED_tol_1_F3Aro|Hyd_F4aro. (b) MODEL4ALIGNED_tol_1_F3Aro|Hyd_F4aro_F6Aro|Hyd. (c) MODEL4ALIGNED_tol_1_F3Aro|Hyd_F4Aro_F6Hyd|F7Aro. (d) MODEL4ALIGNED_tol_1_F3Aro_F4Aro|Hyd_F5Hyd.

Los resultados de la búsqueda en la base de datos Base_ACTIV_1_ale1 se muestran en la Tabla 4.7. Puede observarse cómo el refinamiento manual permite incrementar la selectividad (%Y), aunque a costa de reducirse el recubrimiento de activos. Dentro de esta serie se proponen como posibles modelos MODEL4ALIGNED_tol_1_F3Aro|Hyd_F4Aro_F6Hyd|F7Aro y MODEL4ALIGNED_tol_1_F3Aro_F4Aro|Hyd_F5Hyd, que son los que alcanzan el óptimo de enriquecimiento y GH**.**

4.5.4. MODEL4ALIGNED y esquema PPCH_ALL

El esquema PPCH_ALL genera muchas más *features* farmacofóricas que el esquema anterior, ya que se anotan individualmente los átomos hidrofóbicos. En modelos derivados sobre los alineamientos MODEL2PDB y MODEL3ALIGNED con tolerancias de 1 y 1.2, se observa que este número es tan elevado que incluso supera la capacidad del MOE, no pudiendo realizarse la búsqueda farmacofórica. En ambos casos, los modelos farmacofóricos propuestos son demasiado estrictos y selectivos, por lo que se identifica un reducido número de moléculas (tanto activos como falsos positivos). Por ello, para el esquema PPCH_ALL se decide refinar manualmente únicamente el modelo derivado del alineamiento MODEL4ALIGNED.

En la Figura 4.11 se muestran los modelos obtenidos con el alineamiento MODEL4ALIGNED y el esquema PPCH_ALL: a) con una tolerancia de 1 (MODEL4ALIGNED_PPCH_tol_1) y b) con tolerancia de 1.2 (MODEL4ALIGNED_PPCH_tol_1.2). En la Tabla 4.9 se detallan sus respectivas *features.*

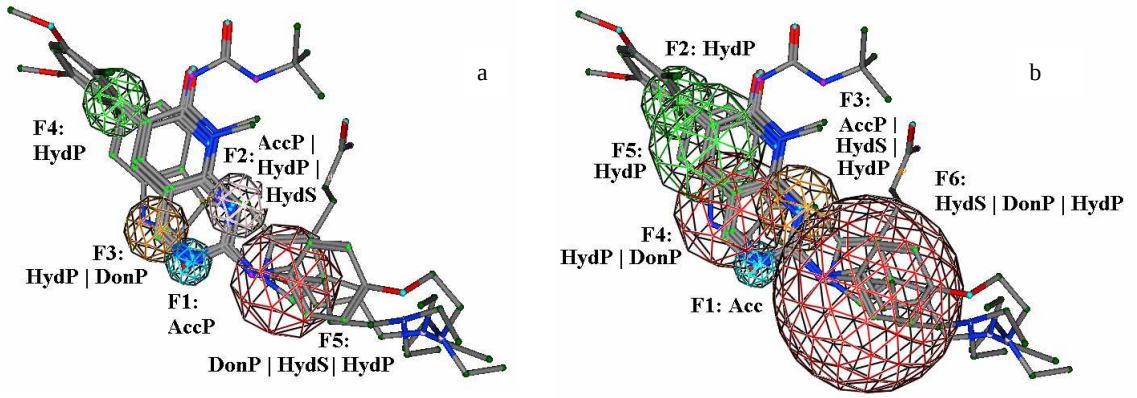


Figura 4.11. Modelos farmacofóricos obtenidos a partir del alineamiento MODEL4ALIGNED y el esquema PCCH_ALL con (a) tolerancia de 1 y (b) tolerancia de 1.2.

Tabla 4.9. Listado de *features* que describen los modelos farmacofóricos obtenidos a partir de MODEL4ALIGNED con esquema PPCH_ALL y con tolerancia de 1 y 1.2.

MODEL4ALIGNED_PPCH_tol_1			**MODEL4ALIGNED_PPCH_tol_1.2**		
ID_Feature	Radio	Expresión	ID_Feature	Radio	Expresión
F1	0.82	Aceptor_planar	F1	0.82	Aceptor planar
F2	1.12	Aceptor planar o Hidrofóbico planar o Hidrofóbico no-planar	F2	1.31	Hidrofóbico planar
F3	1.15	Hidrofóbico planar o Dador planar	F3	1.50	Aceptor planar, Hidrofóbico planar o Hidrofóbico no planar
F4	1.20	Hidrofóbico planar	F4	2.05	Hidrofóbico planar o Dador planar
F5	1.97	Dador planar, Hidrofóbico no planar o Hidrofóbico planar	F5	2.12	Hidrofóbico planar
			F6	3.35	Hidrofóbico no planar, Dador planar ó Hidrofóbico planar

El refinamiento manual de MODEL4ALIGNED_PPCH_tol_1:

- MODEL4ALIGNED_PPCH_tol_1_mod1. Se modifica F5 de manera que únicamente acepte la descripción dador planar, reduciéndose el radio a 1.

- MODEL4ALIGNED_PPCH_tol_1_mod2. A partir del original, se desglosa F5 en dos nuevas *features*: F5 (dador planar, 1) y F6 (Hidrofóbico planar o no planar, 1.2). (Figura 4.12)

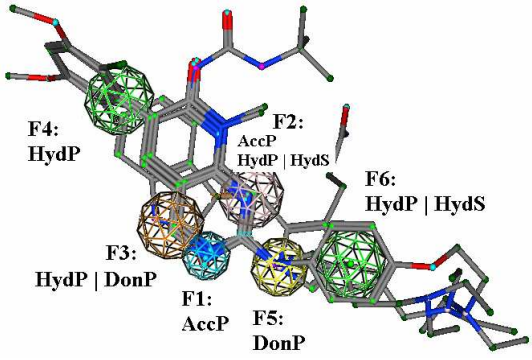


Figura 4.12. Modelo farmacofórico MODEL4ALIGNED_PPCH_tol_1_mod2.

En la Tabla 4.10 se recogen los resultados obtenidos en la búsqueda farmacofórica sobre Base_ACTIV_1_ale1 a partir de los distintos modelos. Se observa cómo inicialmente la selectividad es reducida (bajo %Y) y que el refinamiento cambiando la expresión de F5 permite incrementar la selectividad, aunque a un coste relativo de recubrimiento.

Tabla 4.10. Enriquecimiento para los modelos derivados a partir del alineamiento MODEL4ALIGNED y el esquema farmacofórico PPCH_ALL y tolerancia 1. *#Sca*: número de familias de *scaffolds* con al menos un representante incluido en la lista de *hits*. *COBRA*: número de compuestos activos identificados del total de 104 inhibidores del *pool* COBRA. En cursiva se muestran las características de la base de datos inicial.

Modelo farmacofórico	Ha	Ht	%Y	%A	*ef*	GH	#Sca	COBRA
Base_ACTIV_1_ale1	*120*	*2120*	*5.66*	*100.00*	*1.00*	*0.00*	*6*	-
MODEL4ALIGNED_PPCH_tol_1	115	1226	9.38	95.83	1.66	0.14	6	98
MODEL4ALIGNED_PPCH_tol_1_mod1	72	387	18.60	60.00	3.29	0.24	6	29
MODEL4ALIGNED_PPCH_tol_1_mod2	47	147	31.97	39.17	5.65	0.32	3	29

Por otra parte, MODEL4ALIGNED_PPCH_tol_1.2 se refina manualmente, generándose los siguientes modelos:

- MODEL4ALIGNED_PPCH_tol_1.2_mod1. Se reducen los radios de F3 (de 1.5 a 1), de F5 (de 2.12 a 1) y de F6 (de 3.35 a 2.2).

- MODEL4ALIGNED_PPCH_tol_1.2_mod2. A partir del original, se reduce el radio de F3 (de 1.5 a 1) y se desglosa F6 en dos *features*: F6 (dador planar, 0.59) y F7 (hidrofóbico planar o no-planar, 2). (Figura 4.13.a).

- MODEL4ALIGNED_PPCH_tol_1.2_mod3. A partir del original se reducen los radios de F3 (de 1.5 a 1), de F5 (de 2.12 a 1) y de F6 (de 3.3.35 a 2.2). Además, se reemplaza F4 por la F3 calculada con tolerancia de 1. (Figura 4.13.b).

- MODEL4ALIGNED_PPCH_tol_1.2_mod4. A partir del anterior, se desglosa F6 en dos nuevas *features*: F6 (dador planar, 1.06) y F7 (hidrofóbico planar o no-planar, 1.2). Se impone una *constraint* para que se satisfaga al menos F6 o F7. (Figura 4.13.c).

- MODEL4ALIGNED_PPCH_tol_1.2_mod5. A partir del anterior, pero en lugar de fijar una *constraint*, se etiquetan todas las *features* como esenciales excepto F6 y F7, de donde se debe satisfacer al menos una de ellas.

- MODEL4ALIGNED_PPCH_tol_1.2_mod6. Similar a la anterior, pero el radio de F6 se reduce de 1.06 a 0.59.

- MODEL4ALIGNED_PPCH_tol_1.2_mod7. Similar a la anterior, pero el radio de F2 se reduce a 1.

- MODEL4ALIGNED_PPCH_tol_1.2_mod8. Similar a la anterior, pero se reduce el radio de F4 a 0.9. (Figura 4.13.d).

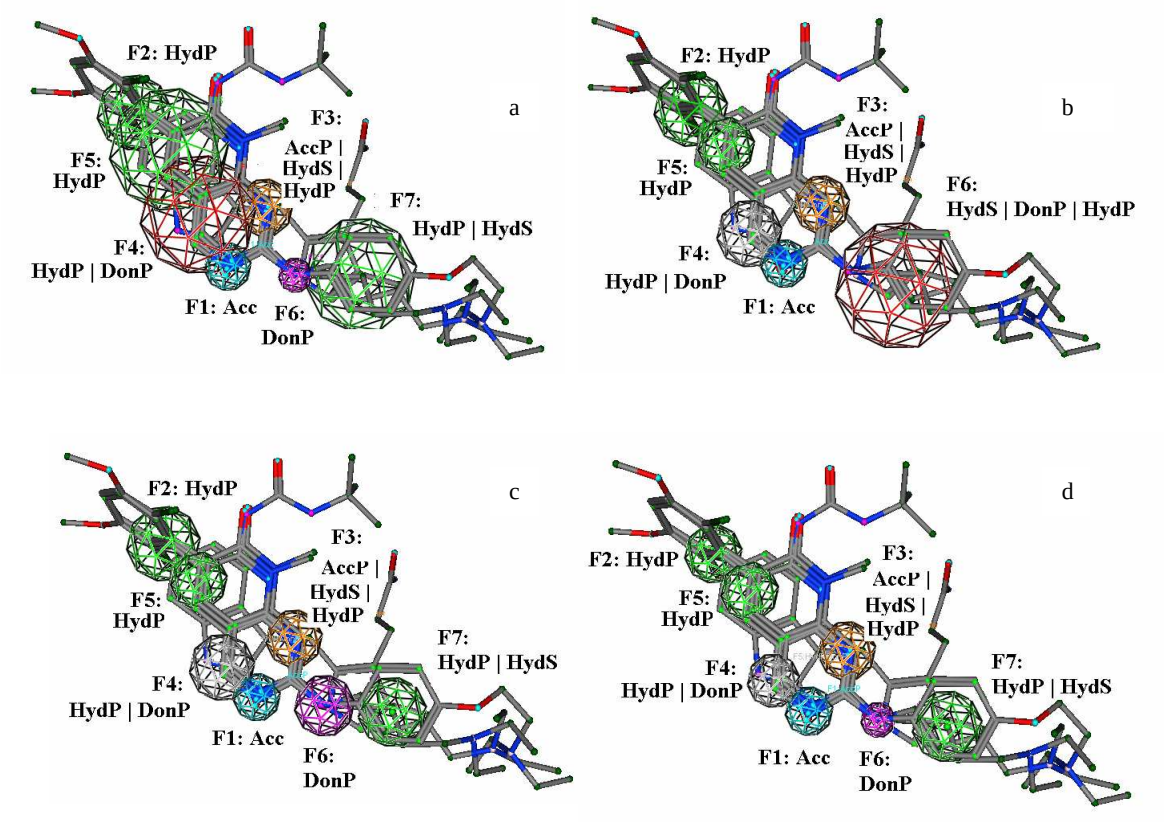


Figura 4.13. Modelos farmacofóricos: a) MODEL4ALIGNED_PPCH_tol_1.2_mod2.
(b) MODEL4ALIGNED_PPCH_tol_1.2_mod3.
(c) MODEL4ALIGNED_PPCH_tol_1.2_mod4.
(d) MODEL4ALIGNED_PPCH_tol_1.2_mod8.

En la Tabla 4.11 se muestran los valores de enriquecimiento obtenidos por estos modelos optimizados a partir de MODEL4ALIGNED_PPCH_tol_1.2. Como en el caso de tolerancia 1, inicialmente el modelo es demasiado poco selectivo y pese a que se capturan los 120 activos de la base de datos, apenas se enriquece. El desglosamiento de F6 en dos nuevas *features*, de manera que se satisfagan opcionalmente una u otra, permite incrementar considerablemente la selectividad sin penalizar significativamente el recubrimiento, a diferencia de lo que ocurre con los modelos derivados con tolerancia de 1.

Tabla 4.11. Enriquecimiento para los modelos derivados a partir del alineamiento MODEL4ALIGNED y el esquema farmacofórico PPCH_ALL y tolerancia 1.2. *#Scaffolds*: número de familias de *scaffolds* con al menos un representante incluido en la lista de *hits*. *COBRA*: número de compuestos activos identificados del total de 104 inhibidores del *pool* COBRA. En cursiva se muestran las características de la base de datos inicial. En negrita aparece el modelo seleccionado en principio como óptimo de la serie.

Modelo farmacofórico	Ha	Ht	%Y	%A	*ef*	GH	*#Scaffolds*	COBRA
Base_ACTIV_1_ale1	*120*	*2120*	*5.66*	*100.00*	*1.00*	*0.00*	*6*	-
MODEL4ALIGNED_PPCH_tol_1.2	120	1598	7.51	100.00	1.33	0.08	6	104
MODEL4ALIGNED_PPCH_tol_1.2_mod1	120	1405	8.54	100.00	1.51	0.11	6	100
MODEL4ALIGNED_PPCH_tol_1.2_mod2	120	1430	8.39	100.00	1.48	0.11	6	101
MODEL4ALIGNED_PPCH_tol_1.2_mod3	116	897	12.93	96.67	2.28	0.21	6	87
MODEL4ALIGNED_PPCH_tol_1.2_mod4	47	114	41.23	39.17	7.28	0.39	3	21
MODEL4ALIGNED_PPCH_tol_1.2_mod5	111	673	16.49	92.50	2.91	0.26	6	70
MODEL4ALIGNED_PPCH_tol_1.2_mod6	115	652	17.64	95.83	3.12	0.27	6	68
MODEL4ALIGNED_PPCH_tol_1.2_mod7	114	553	20.61	95.00	3.64	0.31	6	64
MODEL4ALIGNED_PPCH_tol_1.2_mod8	**101**	**386**	**26.17**	**84.17**	**4.62**	**0.35**	**6**	**63**

En este sentido, se propone el modelo MODEL4ALIGNED_PPCH_tol_1.2_mod8 como el mejor de la serie.

4.5.5. Selección de un modelo final farmacofórico obtenido con MOE

De todos los modelos obtenidos directamente mediante la herramienta *Pharmacophore Consensus*, se aprecian dos tendencias opuestas. Por una parte, se obtienen modelos demasiado restrictivos, como los derivados a partir del alineamiento MODEL3ALIGNED con el esquema PCH, ya que presentan demasiadas *features*. Por otro lado, pueden resultar demasiado generales, como en el caso de los modelos derivados a partir del alineamiento MODEL4ALIGNED y el esquema PCH o PPCH_ALL. Por lo tanto, en todos los casos es necesario un refinamiento manual. En este sentido, se ha encontrado que el esquema PCH es más adecuado para los alineamientos derivados de moléculas muy similares (MODEL3ALIGNED), mientras que el esquema PPCH_ALL, dado el elevado número de puntos potenciales farmacofóricos que detecta, lo es para el caso de plantillas con menor similitud (MODEL4ALIGNED).

De entre los cuatro modelos farmacofóricos seleccionados hasta el momento, recopilados en la Tabla 4.12, se escoge como modelo final MODEL4ALIGNED_PPCH_tol_1.2_mod8, ya que presenta un óptimo de GH, siendo el mejor compromiso en recubrimiento y selectividad. En adelante, se denomina dicho modelo MOE_MODEL. De modo orientativo, se presentan en la Tabla 4.13 valores típicos encontrados para el parámetro GH, extraídos de la referencia [376].

Tabla 4.12. Enriquecimiento para los cuatro modelos farmacofóricos construidos con MOE seleccionados *a priori*. *#Sca*: número de familias de *scaffolds* con al menos un representante incluido en la lista de *hits*. *COBRA*: número de compuestos activos identificados del total de 104 inhibidores del *pool* COBRA.

Modelo farmacofórico	Ha	Ht	%Y	%A	*ef*	GH	#Sca	COBRA
Base_ACTIV_1_ale1	*120*	*2120*	*5.66*	*100.00*	*1.00*	*0.00*	*6*	-
MODEL3ALIGNED_tol_1.2_At_Least_5	82	652	12.58	68.33	2.22	0.19	6	60
MODEL4ALIGNED_tol_1_F3Aro|Hyd_F4Aro_F6Hyd|F7Aro	65	360	18.06	54.17	3.19	0.23	4	31
MODEL4ALIGNED_tol_1_F3Aro_F4Aro|Hyd_F5Hyd	81	522	15.52	67.50	2.74	0.22	5	30
MODEL4ALIGNED_PPCH_tol_1.2_mod8	101	386	26.17	84.17	4.62	0.35	6	63

Tabla 4.13. Ejemplos de valores de GH de diferentes estudios extraídos de [376].

Caso de Estudio	Ha	Ht	%Y	%A	GH
Base de datos 1	244	47926			
Farmacóforo 1	231	4442	5.2	94.7	0.251
Base de datos 2	225	10318			
Farmacóforo 2	174	3772	4.61	77.3	0.147
Base de datos 3	80	10318			
Farmacóforo 3	28	1954	1.43	35.0	0.08

Se destaca el hecho de que este modelo farmacofórico final recoge la ambigüedad de la interacción farmacofórica observada en las estructuras cristalinas de PD173074 y SU5402 (Figura 4.4), en el sentido de que cada una de ellas interacciona del mismo modo pero con distintos residuos de la proteína. Como se ha mencionado, tradicionalmente se requiere la existencia de un dador y un aceptor de puente de hidrógeno en los inhibidores competitivos de ATP. En MOE_MODEL (Figura 4.13.d), el aceptor se recoge en F1, siendo común al O-2 de la indolinona y el N-3 de PD173074. Respecto al dador, existen dos posibilidades no excluyentes: o F6 (como el nitrógeno del grupo butilamino de PD173074) o F4 (como el N-1 de la indolinona SU5402).

Por lo tanto, este modelo es consistente con las interacciones determinadas mediante rayos-X y constituye un ejemplo típico de fármacos que se unen a un receptor común mediante grupos funcionales similares, aunque no precisamente solapados.

Dado que con SQUID no es posible, por el momento, etiquetar un punto potencial farmacofórico (PPP) con múltiples tipos atómicos y, con el objetivo de establecer una comparación con la identificación tradicional de farmacóforos en MOE, se construye un modelo derivado de MOE_MODEL en el que se eliminan las *features* con múltiples tipos atómicos (F3, F4 y F7 en la Figura 4.13.d). En adelante dicho modelo se denomina MOE_MODEL_WM.

Una vez seleccionados los dos modelos obtenidos con MOE (MOE_MODEL y MOE_MODEL_WM), se han aplicado en la identificación retrospectiva de *hits* de: 1) los cuatro subconjuntos aleatorios extraídos de Base_ACTIV_1 y 2) Base_COBRA. Las diferentes relaciones entre el número de *hits* identificados y la base de datos original se muestran en las Tablas 4.14 y 4.15, para cada base de datos, respectivamente. En el caso de los cuatro subconjuntos, dichos resultados corresponden al promedio de todos ellos, mostrándose en paréntesis su desviación estándar. Se observa cómo la eliminación de las *features* con múltiples características reduce notablemente la selectividad del modelo.

Tabla 4.14. Factores de recubrimiento, selectividad, enriquecimiento y GH para el cribado de las cuatro de bases de datos aleatorias extraídas de Base_ACTIV_1, expresados como promedio (desviación estándar en paréntesis). Aparece también el número de *scaffolds* activos identificados (Figura 4.2). Búsquedas realizadas con los modelos farmacofóricos de MOE: MOE_MODEL y MOE_MODEL_WM.

	Ha	**Ht**	**%Y**	**%A**	***ef***	**GH**	**#Scaffolds**
Base_ACTIV_1	*120*	*2120*	*5.67*	*100*	*1.0*	*0*	*6*
MOE MODEL	101 (1.3)	386 (12.1)	25.98 (1.1)	83.96 (1.0)	4.61 (0.2)	0.35 (0.0)	6 (0.0)
MOE_MODEL_WM	120 (0.0)	1674 (17.1)	7.17 (0.1)	100 (0.0)	1.26 (0.01)	0.07 (0.0)	6 (0.0)

Tabla 4.15. Factores de recubrimiento, selectividad, enriquecimiento y GH para el cribado de Base_COBRA. Búsquedas realizadas con los modelos farmacofóricos de MOE: MOE_MODEL y MOE_MODEL_WM.

	Ha	**Ht**	**%Y**	**%A**	***ef***	**GH**
Base_COBRA	*104*	*5035*	*2.07*	*100*	*1.0*	*0*
MOE MODEL	63	775	8.13	60.58	3.94	0.18
MOE_MODEL_WM	104	3933	2.64	100	1.28	0.06

Los resultados de enriquecimiento mostrados hasta el momento corresponden a la clasificación *total* de la base de datos inicial en posibles *hits* o en compuestos inactivos. Sin embargo, como se ha comentado en el apartado 1.7.1, MOE permite también establecer un ranking de los compuestos según su RMSD de superposición entre las *features* presentes en la molécula y aquellas contenidas en la hipótesis. En la Figura 4.14 se muestra la curva de enriquecimiento obtenida sobre las bases de datos Base_ACTIV_1 y Base_COBRA con los dos modelos a distintos porcentajes de base de datos muestreada. En los dos casos se observa cómo el enriquecimiento del modelo generado es muy bueno, con el 80% y 50% de los activos detectados al 10% de base de datos muestreada. Como sería esperable, los rendimientos para Base_ACTIV_1 son mayores que para Base_COBRA, ya que el enriquecimiento inicial es también superior y porque la diversidad de *scaffolds* en esta base de datos está restringida a seis *cores* distintos, por lo que se puede asumir que las moléculas en esta base de datos son más similares entre ellas. En todos los casos, el enriquecimiento es superior a una selección aleatoria (*línea gris*). El modelo MOE_MODEL_WM presenta un comportamiento mucho peor.

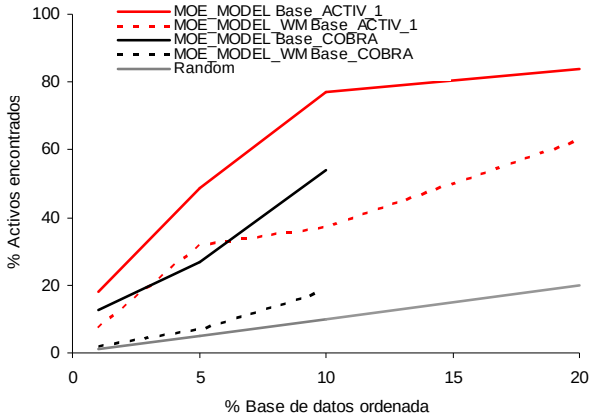


Figura 4.14. Curvas de enriquecimiento obtenidas para las dos bases de datos (Base_ACTIV_1 y Base_COBRA) con los dos modelos farmacofóricos derivados en MOE.

4.6. Búsqueda de Similitud con descriptores CATS3D

Se han realizado búsquedas de similitud utilizando separadamente PD173074 y SU5402 como compuestos *lead* o *focus*. Además, se han analizado dos posibles estados de protonación: compuestos "neutros" y compuestos con carga a pH fisiológico (pH=7), es decir, con las bases protonadas y los ácidos desprotonados. Respecto a las métricas, se han considerado la distancia Manhattan y el coeficiente de Tanimoto (véase Tabla 1.8).

En este caso, la base de datos completa se ordena según el valor de similitud/distancia calculado. Así, el programa no devuelve exclusivamente una lista de *hits* potenciales, como en el caso del farmacóforo derivado con MOE. Por ello, no se han calculado todos los parámetros mostrados en el apartado anterior, en especial GH, sino que se ha restringido el cálculo a los factores de enriquecimiento (ecuación [4.3]) y las curvas de enriquecimiento.

En las Figuras 4.15.a y 4.15.b se muestran las curvas de enriquecimiento obtenidas al focalizar sobre los compuestos PD173074 y SU5402 en estado neutro para las bases de datos Base_ACTIV_1 (promedio de los cuatro subconjuntos aleatorios) y Base_COBRA, respectivamente. El cribado retrospectivo con PD173074 como *focus* supera al correspondiente empleando SU5402 como *query*, que equivale en algún caso a un cribado aleatorio (*línea gris*). Dentro de las métricas, la distancia Manhattan se comporta mejor que el coeficiente de Tanimoto, conclusión también expuesta en un análisis exhaustivo de los descriptores CATS3D.[283]

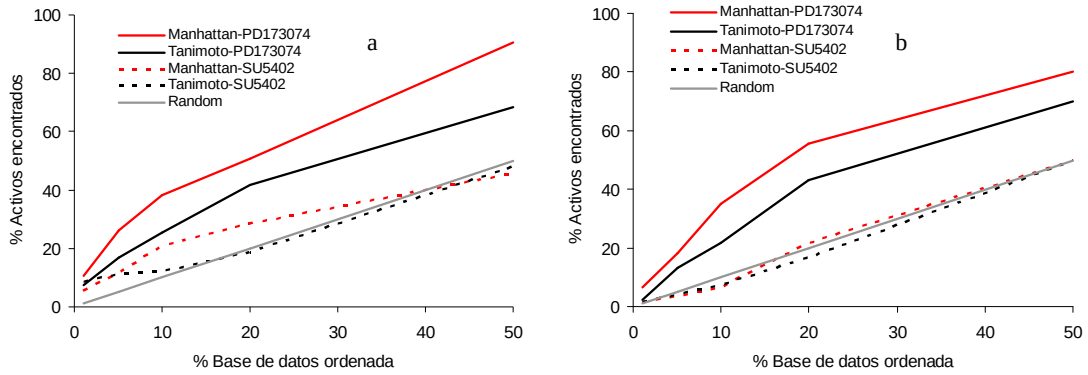


Figura 4.15. Curvas de enriquecimiento obtenidas en las búsquedas de similitud con los descriptores CATS3D para (a) Base_ACTIV_1 (promedio de los cuatro subconjuntos aleatorios) y (b) Base_COBRA.

Sus correspondientes factores de enriquecimiento se tabulan posteriormente, en la comparación de las tres metodologías de búsqueda farmacofórica (véase apartado 4.8.1).

Respecto a la situación de considerar los compuestos a pH fisiológico, no se muestran las curvas de enriquecimiento por una cuestión de espacio y porque, como se verá en el apartado 4.8.1 en el que se tabulan sus correspondientes factores de enriquecimiento, éstos son muy bajos, rindiendo en algunos casos selecciones prácticamente aleatorias.

4.7. Modelos SQUID

En primer lugar, se optimiza el código del programa SQUID, escrito en lenguaje SVL, así como el programa de búsqueda y cálculo de similitud, escrito en C++, incrementándose su eficiencia computacional. Además, se introduce un pequeño entorno gráfico en SVL para el programa SQUID en el entorno de MOE que permite la construcción automática de todos los modelos derivados de un alineamiento para un rango de *cluster radius* definido por el usuario.

Se aplica la metodología SQUID para derivar modelos farmacofóricos para los tres alineamientos descritos: MODEL2PDB (Figura 4.4.a), MODEL3ALIGNED (Figura 4.5.a) y MODEL4ALIGNED (Figura 4.6). Además, se estudia la influencia de las cargas (pH) en los dos primeros alineamientos. Para cada una de las cinco situaciones posibles, se genera una serie de modelos SQUID con valores de *cluster radius* comprendidos entre 0 y 3 Å, en pasos de 0.1 Å (31 modelos).

Una vez creados estos modelos y codificados en el correspondiente vector de correlación, durante la búsqueda de similitud frente a una base de datos caracterizada con descriptores CATS3D, se optimizan los pesos adicionales del modelo o *feature-type weights*. Esta optimización se realiza de manera exhaustiva, variando los valores de los pesos de cada uno de los tipos de interacción generalizada (dador, aceptor…) presentes en el modelo desde 0.1 a 0.5, con pasos de 0.2. Posteriormente, este rango de pesos se amplia hasta 1, sin apreciable variación en los resultados. Así, para cada uno de los alineamientos iniciales se generan aproximadamente unas 3000-7000 combinaciones posibles, dependiendo del número de tipos de interacción generalizada presentes en cada uno de los modelos. De entre todos ellos, se escoge el modelo óptimo según el valor de enriquecimiento *ev* obtenido tras la búsqueda farmacofórica, mostrado en la ecuación [4.5]. Este índice recoge la capacidad de enriquecimiento a lo largo de toda la base de datos, en lugar de fijarse únicamente en el valor del factor de enriquecimiento *ef* (ecuación [4.3]) a un determinado porcentaje de base de datos ordenada.[51]

$$ev = \sum_{i=1}^{100} (101 - i) ef\,(i\%) \qquad [4.5]$$

Este proceso de optimización de los parámetros *cluster radius* y *feature-type weights* se repite para cada subconjunto aleatorio de la Base_ACTIV_1, dado que los parámetros óptimos pueden variar ligeramente entre diferentes bases de datos. Dado el coste computacional que supone, los parámetros óptimos obtenidos a partir del primer conjunto aleatorio (Base_ACTIV_1_ale1), se trasladan a Base_COBRA.

Por una cuestión de espacio, únicamente se muestran los modelos SQUID derivados para cada uno de los cinco alineamientos con el *cluster radius* óptimo obtenido para el subconjunto aleatorio Base_ACTIV_1_ale1 (Figura 4.16). En cualquier caso, las variaciones en el óptimo de *cluster radius* en función del subconjunto considerado de Base_ACTIV_1 son mínimas, observándose tendencias análogas en la variación del enriquecimiento respecto al *cluster radius*.

En el estado neutro, se encuentran tres tipos de interacción generalizada: dadores de puente de hidrógeno (*magenta*), aceptores de puente de hidrógeno (*cian*) y átomos hidrofóbicos (*amarillo*). Además, en los alineamientos MODEL2PDB y MODEL4ALIGNED, la incorporación de SU5402 introduce un tipo de interacción polar (*verde*), debido al grupo carboxilato. Para los dos alineamientos en estado cargado (Figuras 4.16.b y 4.16.c), aparecen otros dos tipos adicionales: uno catiónico (*azul*) correspondiente al grupo amino de la cadena butilamino de PD173074 y otro aniónico (*rojo*) correspondiente al grupo carboxilato de SU5402. En la representación de los modelos SQUID, el radio de las esferas denota la desviación estándar (σ) de la distribución espacial de los átomos de cada punto potencial farmacofórico (PPP). La intensidad del color de cada PPP indica la conservación del mismo entre las distintas moléculas alineadas.

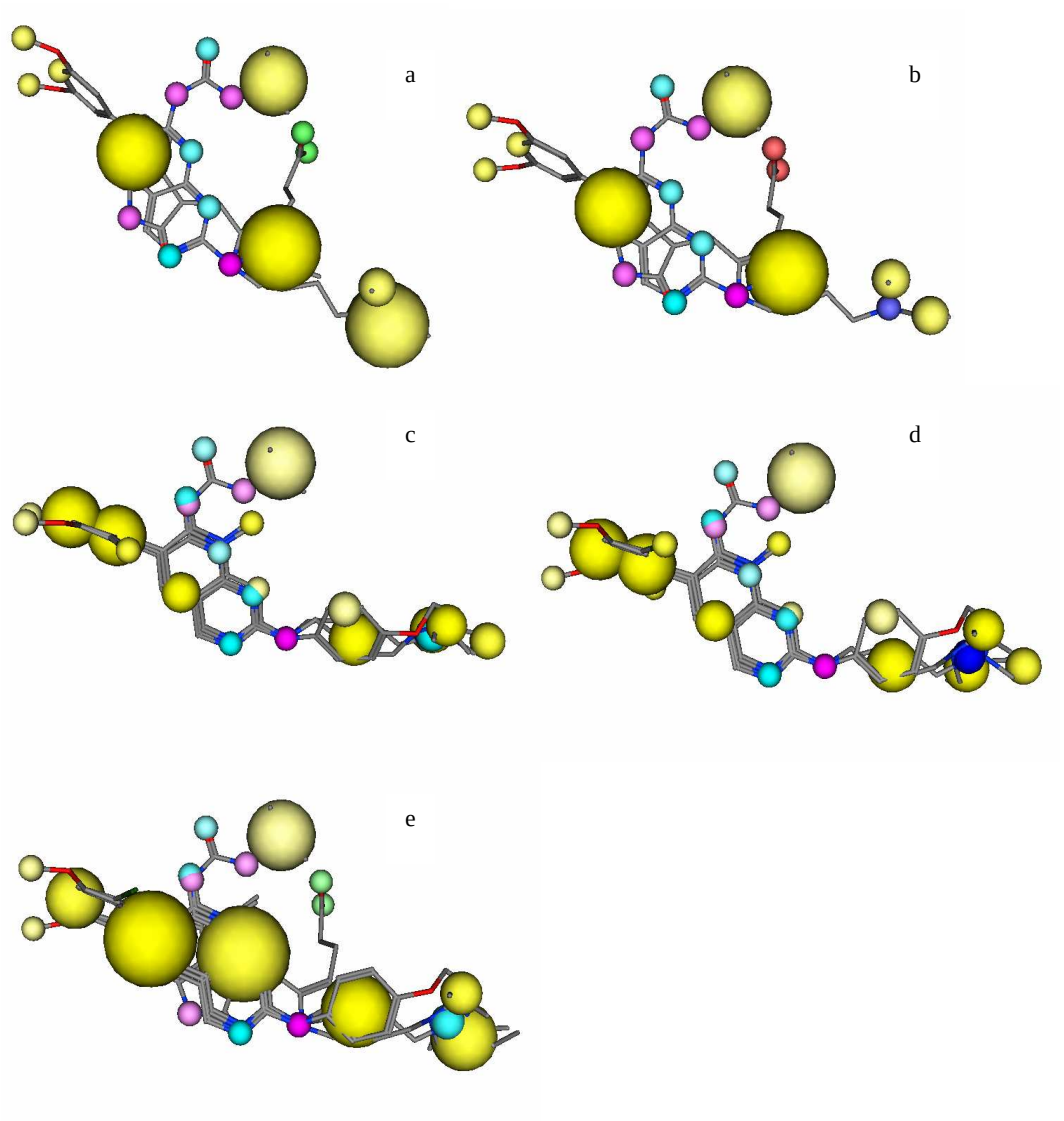


Figura 4.16. Modelos farmacofóricos SQUID seleccionados para cada uno de los cinco alineamientos estudiados. (a) MODEL2PDB neutro con *cluster radius* de 2.2, (b) MODEL2PDB cargado con *cluster radius* de 2.1, (c) MODEL3ALIGNED neutro con *cluster radius* de 1.6, (d) MODEL3ALIGNED cargado con *cluster radius* de 1.6 y (e) MODEL4ALIGNED neutro con *cluster radius* de 1.8. Tipos de interacción generalizada: dadores de puente de hidrógeno (*magenta*), aceptores de puente de hidrógeno (*cian*), átomos hidrofóbicos (*amarillo*), polares (*verde*), cationes (*azul oscuro*) y aniones (*rojo*).

4.8. Comparación del cribado retrospectivo según los tres modelos

4.8.1. Factores de enriquecimiento

En la Tabla 4.16 se listan los factores de enriquecimiento obtenidos para las bases de datos Base_ACTIV_1 (promedio y desviación estándar de los cuatro subconjuntos aleatorios) y Base_COBRA según las tres metodologías (MOE, CATS3D y SQUID) y a diferentes porcentajes de base de datos muestreada.

En primer lugar, se comenta la influencia de considerar las moléculas cargadas a pH fisiológico frente a considerarlas neutras en los resultados de las búsquedas de similitud con CATS3D.

Se observa cómo en general es posible alcanzar un enriquecimiento satisfactorio en activos (*ef*>1) en todos los casos, excepto en la situación cargada frente a Base_COBRA, donde los rendimientos decaen incluso a cero. Así, en general los ligandos cargados se comportan peor que los neutros, excepto en el caso de tomar PD173074 como *query* frente a Base_ACTIV_1, donde el enriquecimiento es similar en los dos estados de protonación. Desde un punto de vista estructural, el grupo butilamino de PD173074 se encuentra cargado en el interior de la proteína FGFR y su función es la de incrementar la solubilidad acuosa.[371,344] Una posible explicación del comportamiento diferencial entre ambas bases de datos es que Base_ACTIV_1 incluye un menor número de *scaffolds* diversos y con moléculas seleccionadas mediante estudios SAR, por lo que se potencia la inclusión de factores como la solubilidad.
En el caso particular de la plantilla SU5402, el único grupo cargado es el carboxietilo unido al C-3' del anillo de pirrol. Estructuralmente, se encuentra unido por puente de hidrógeno a la Asn[568] de FGFR, en un modo similar a como lo hace el anillo de ribosa[375] y se cree que su papel fundamental es el de conferir especificidad.

Tabla 4.16. Factores de enriquecimiento *ef* encontrados para las dos bases de datos Base_ACTIV_1 y Base_COBRA según las tres metodologías (SQUID, CATS3D y los dos modelos farmacofóricos del MOE) a diferentes porcentajes de base de datos muestreada (1%, 5%, 10% y 20%). Los resultados para Base_ACTIV_1 están calculados como el promedio de los cuatro subconjuntos aleatorios, en paréntesis se muestra su desviación estándar. Los modelos SQUID se denominan en función del alineamiento del que son derivados. Las búsquedas CATS3D se distinguen en función del compuesto sobre el que se focaliza la búsqueda y la métrica usada en el cálculo de similitud.

MÉTODO	**MODELO**	***ef* Base_ACTIV_1** % Base de datos muestreada				***ef* Base_COBRA** % Base de datos muestreada			
		1%	5%	10%	20%	1%	5%	10%	20%
SQUID	MODEL2PDB	6.02 (1.91)	3.25 (0.35)	2.08 (0.18)	1.43 (0.09)	1.9	0.96	0.86	1.30
	MODEL3ALIGNED	5.21 (3.11)	2.50 (0.14)	1.98 (0.12)	1.58 (0.12)	0.95	0.96	2.11	2.21
	MODEL4ALIGNED	6.02 (1.04)	3.21 (0.25)	2.06 (0.28)	1.41 (0.09)	0.95	1.15	1.15	1.35
	MODEL2PDB-cargado	4.41 (2.66)	3.54 (0.42)	3.11 (0.38)	2.23 (0.07)	0.00	0.58	0.77	1.20
	MODEL3ALIGNED-cargado	5.42 (2.49)	2.29 (0.58)	1.67 (0.18)	1.70 (0.26)	0.00	0.96	1.73	1.59
CATS3D	Manhattan-PD173074	10.04 (1.39)	5.25 (0.78)	3.84 (0.27)	2.54 (0.27)	7.59	4.23	4.03	3.22
	Tanimoto-PD173074	7.23 (2.18)	3.38 (0.34)	2.52 (0.24)	2.08 (0.13)	2.85	3.07	2.50	2.50
	Manhattan-SU5402	5.22 (0.46)	2.38 (0.21)	2.06 (0.17)	1.42 (0.06)	1.90	0.77	0.77	1.25
	Tanimoto-SU5402	8.43 (0.46)	2.25 (0.09)	1.23 (0.08)	0.94 (0.10)	1.90	0.96	0.86	0.96
	Manhattan-PD173074 cargado	10.84 (1.39)	6.17 (0.49)	4.48 (0.08)	2.88 (0.10)	4.75	5.76	4.13	3.56
	Tanimoto-PD173074 cargado	10.04 (1.04)	3.42 (0.22)	2.57 (0.13)	2.25 (0.03)	0.95	3.37	2.79	2.21
	Manhattan-SU5402 cargado	1.61 (0.66)	1.71 (0.16)	1.21 (0.08)	0.77 (0.07)	0.00	0.38	0.38	0.53
	Tanimoto-SU5402 cargado	1.61 (0.65)	1.69 (0.16)	1.21 (0.08)	0.77 (0.08)	0.00	0.00	0.00	0.00
MOE	MOE_MODEL	17.22 (0.46)	9.73 (0.10)	7.67 (0.08)	4.21 (0.28)	12.34	5.38	5.38	
	MOE_MODEL_WM	7.03 (1.01)	6.38 (0.16)	3.73 (0.04)	3.15 (0.07)	1.90	1.34	1.83	

Respecto a los resultados obtenidos por la metodología SQUID, a primera vista estos parecen más que aceptables para Base_ACTIV_1, ya que son superiores a una selección aleatoria (*ef*>1). Sin embargo, estos resultados no son tan favorables si se comparan con los otros obtenidos mediante el modelo farmacóforico de MOE (MOE_MODEL).

Por ejemplo, para el 1% de Base_ACTIV_1 muestreada, se obtiene un *ef* de 17.22 con MOE_MODEL, mientras que 6.02 es el mejor enriquecimiento obtenido con un modelo SQUID. El enriquecimiento obtenido con el modelo MOE_MODEL decae al eliminar las *features* con múltiple asignación de tipos, alcanzando un valor de 7.03, similar al obtenido con el mejor modelo SQUID.

Una tendencia análoga se obtiene en Base_COBRA. Para el primer 1% de base de datos ordenada, el *ef* de 12.03 obtenido con MOE_MODEl decrece a 1.9 con MOE_MODEL_WM, el mismo alcanzado con el mejor modelo SQUID (a partir de MODEL2PDB). Para esta base de datos, los cinco modelos SQUID se encuentran mayoritariamente por debajo de una selección aleatoria. Este hecho puede deberse en gran parte al hecho de que los pesos *feature-type weights* utilizados para la búsqueda de similitud en esta base de datos no se han optimizado en ella, sino que se han aplicado los mismos que los derivados en la otra base de datos. De hecho, se han realizado pruebas sin aplicar pesos (*feature-type weights*=1) para todos y cada uno de los tipos de interacción generalizada), situándose el enriquecimiento por debajo de los valores mostrados en la Tabla 4.16 (resultados no mostrados).

Respecto a las desviaciones estándar presentadas para Base_ACTIV_1, lógicamente éstas son mayores en el primer 1% de la base de datos muestreada, decayendo conforme se aumenta el conjunto de compuestos seleccionados.

4.8.2. Análisis de Diversidad de *scaffolds* en Base_ACTIV_1

Como se ha comentado, el objetivo de dividir el *pool* de compuestos activos en dos bases de datos es el de permitir un análisis de la diversidad de los *scaffolds* identificados. Dicho análisis se realiza de dos maneras diferentes dependiendo de la base de datos considerada. Para Base_ACTIV_1, contando simplemente el número de moléculas recuperadas correspondientes a cada uno de los seis *cores* presentes en el *pool* de activos (Figura 4.2). Estos resultados se muestran en la Figura 4.17, por brevedad, únicamente para los métodos/modelos más representativos (mejores enriquecimientos) de los mostrados en la Tabla 4.16: MOE_MODEL (*violeta*), búsquedas con CATS3D con PD173074 como *query* y distancia Manhattan (*granate*) y los modelos SQUID construidos a partir de los alineamientos neutros MODEL2PDB (*amarillo*), MODEL3ALIGNED (*verde*) y MODEL4ALIGNED (*naranja*). El eje de abcisas corresponde al *scaffold* y el de ordenadas muestra el promedio de los cuatro subconjuntos aleatorios de Base_ACTIV_1 del número absoluto de compuestos identificados. Por el diseño de dichos subconjuntos, el máximo número de compuestos recuperables para cada *scaffold* es de veinte.

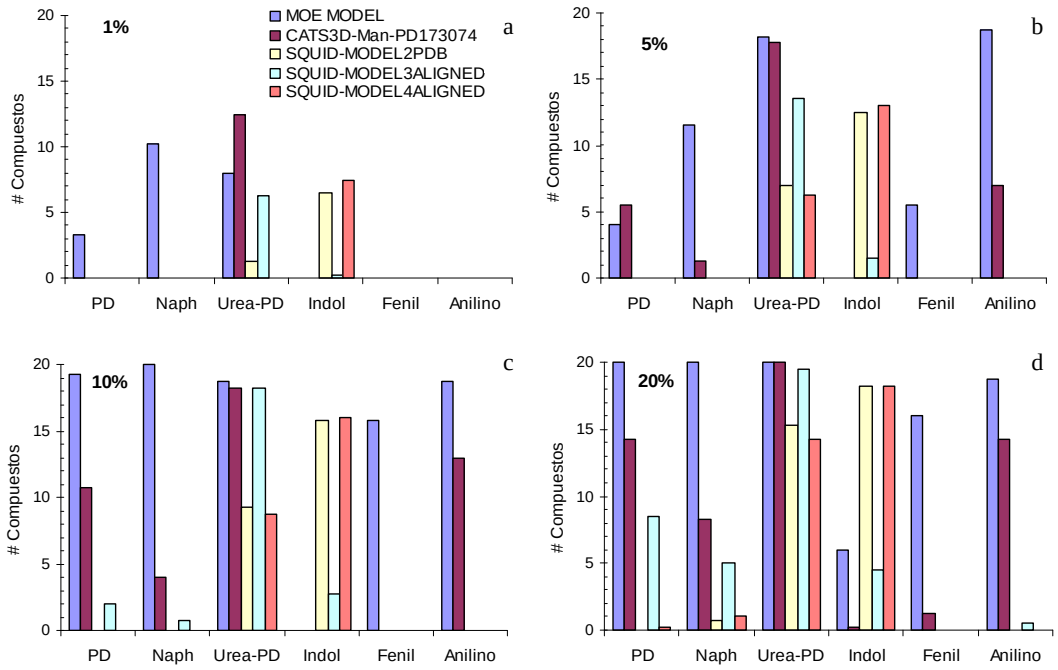


Figura 4.17. Resultados del análisis de diversidad de *scaffolds* para Base_ACTIV_1 a diferentes porcentajes de base de datos muestreada. (a) 1%, (b) 5%, (c) 10% y (d) 20%. En el eje de abcisas se presentan los seis *scaffolds* presentes en la base de datos. *PD*: pirido[2,3-*d*]pirimidinas, *Naph*: naftiridin-2(1*H*)-onas, *Urea-PD*: 7-alquilurea pirido[2,3-*d*]pirimidinas, *Indol*: indolin-2-onas, *Fenil*: 1-fenilbenzimidazoles y *Anilino*: 4-anilinoquinazolinas. En el eje de ordenadas, número absoluto de compuestos recuperados para cada *scaffold*, siendo 20 el máximo posible.

Más que prestar atención a los números absolutos, implícitamente incluidos en los factores de enriquecimiento comentados, es interesante analizar la distribución de los *scaffolds*. Como sería esperable, en el primer 1% de base de datos ordenada, existe una tendencia general a recuperar aquellos *scaffolds* pertenecientes a la misma familia de plantillas incluidas en el alineamiento inicial. Así, MOE_MODEL identifica tres *scaffolds* diferentes, mientras que CATS3D se concentra en las 7-alquilurea pirido[2,3-*d*]pirimidinas, el *core* de PD173074. En este sentido, los modelos SQUID fallan ligeramente ya que, por ejemplo, la búsqueda con el modelo derivado del alineamiento MODEL2PDB (*amarillo*) queda atrapado en las dos familias presentes en el alineamiento prácticamente en todos los casos. La búsqueda farmacofórica con el modelo SQUID derivado del alineamiento MODEL3ALIGNED (*verde*) no detecta las familias pirido[2,3-*d*]pirimidinas y naftiridin-2(1*H*)-onas hasta el 10% de base de datos ordenada, pese a que estas familias se encuentran "representadas" en el alineamiento inicial.

Por otra parte, es destacable cómo MOE_MODEL (*violeta*) no captura la familia indolin-2-onas hasta el 20% de base de datos ordenada, aunque este *core* se encuentra presente en el alineamiento. Para comprobar si este hecho se debe a una desviación hacia los compuestos con mayor tamaño/peso molecular, se ha calculado el peso molecular promedio de Base_ACTIV_1 y de la base de datos de *hits* ordenada por RMSD. La primera tiene un peso molecular promedio de 413.8 ± 138.9 y la segunda de 465.0 ± 142 g/mol. Así, sí que puede existir una cierta desviación hacia las moléculas de mayor tamaño. En este sentido, se encuentra una distribución totalmente opuesta entre MOE_MODEL y el modelo SQUID generado a partir de MODEL4ALIGNED (*naranja*), que captura el *scaffold* indolinona en el primer 1% de base de datos muestreada, pese a que ambos se construyen a partir del mismo alineamiento.

4.8.3. Análisis de Diversidad de *scaffolds* en Base_COBRA

Para esta base de datos se ha determinado el número de *molecular frameworks* diferentes según un programa desarrollado en el grupo del profesor Schneider, implementado en el MOE.[378] Este programa permite obtener el número de *molecular frameworks* de acuerdo a dos criterios: descripción de grafos y descripción atómica, es decir, considerando el número y tipo de átomos que componen el grafo molecular. El *pool* de activos COBRA contiene 43 grafos/57 *frameworks* atómicos diferentes, mientras que la base de datos total, BASE_COBRA, está compuesta de 1851 grafos y 3156 *frameworks* atómicos.

El modelo farmacofórico derivado con MOE, MOE_MODEL, identifica, en el total de *hits*, 28 grafos de compuestos activos (65% del total posibles) y 38 *frameworks* atómicos (67% de los posibles), así como 452 grafos (578 *frameworks* atómicos) correspondientes a los compuestos inactivos (falsos positivos). Por lo tanto, este modelo MOE_MODEL es también satisfactorio desde el punto de vista del enriquecimiento en diversidad. Calculando con la ecuación [4.3] este enriquecimiento, contabilizando *molecular frameworks* en lugar de compuestos, se obtienen unos factores de enriquecimiento de 2.51 ((28/480)/(43/1851)) y 3.42 ((38/616)/(57/3156)) para los grafos y la descripción atómica, respectivamente.

En la Tabla 4.17 se muestra el número de *molecular frameworks* activos (#AM) y totales (#M, verdaderos positivos + falsos positivos) recuperados a diferentes porcentajes de base de datos ordenada según el mejor modelo en términos de enriquecimiento (Tabla 4.16) representativo de cada una de las tres metodologías: modelo MOE_MODEL, búsqueda con CATS3D (PD173074 y distancia Manhattan) y modelo SQUID MODEL2PDB. Para formarse una idea de la diversidad implicada en el número de *frameworks* identificados, se incluye también el número de moléculas activas recuperada (H_a), en paréntesis.

También para esta base de datos se observa como MOE_MODEL rinde una mayor diversidad de *scaffolds* que la búsqueda de similitud con CATS3D, no sólo en los verdaderos positivos encontrados (10 *frameworks* diferentes en 13 compuestos activos comparado frente a 4 *frameworks* de 8 compuestos), sino también en los falsos positivos (41 *molecular frameworks* de 50 moléculas frente a 29 de cada 50). A medida que se incrementa el porcentaje de base de datos muestreada, esta diferencia en el número de *frameworks* identificados desaparece en el *pool* de inactivos (falsos positivos), manteniéndose en el *pool* de activos. El modelo SQUID MODEL2PDB se comporta pobremente, identificando un único *framework* activo entre los 44 *frames* etiquetados como *hits* en el primer 1% de la base de datos muestreada.

Tabla 4.17. Análisis de diversidad de *scaffolds* para Base_COBRA a diferentes porcentajes de base de datos muestreada. *#AM*: número total de *molecular frameworks* presentes en los compuestos activos identificados, H_a: número total de compuestos activos y *#M*: número total de *molecular frameworks* diferentes presentes en los compuestos recuperados (verdaderos positivos + falsos positivos).

	1%				**5%**				**10%**			
	Definición Grafo		Definición Atómica		Definición Grafo		Definición Atómica		Definición Grafo		Definición Atómica	
	#AM (Ha)	#M	#AM (Ha)	#M	#AM (Ha)	#M	#AM (Ha)	#M	#AM (Ha)	#M	#AM (Ha)	#M
MOE_MODEL	10 (13)	41	12	45	16 (28)	170	21	207	24 (56)	325	33 (56)	408
CATS3D-Manhattan-PD173074	4 (8)	29	5	38	13 (22)	169	14	206	20 (42)	318	24 (42)	407
SQUID-MODEL2PDB	1 (2)	44	1	44	4 (5)	163	4	192	6 (9)	286	7 (9)	363

4.9. Modificaciones introducidas en la aplicación del modelo SQUID

Los resultados obtenidos con el modelo SQUID no son satisfactorios si se comparan con los obtenidos en la aplicación original de SQUID[51], ni en términos de enriquecimiento ni en términos de su capacidad para recuperar *scaffolds* diversos. Tampoco lo son si se comparan con los rendimientos obtenidos mediante búsquedas de similitud con los descriptores CATS3D.

En una primera hipótesis, se podría plantear que la falta de implementación de puntos potenciales farmacofóricos (PPPs) con múltiple asignación de tipos de interacción generalizada pueda ser la causa del peor comportamiento del modelo SQUID respecto a los casos de estudio en que ha sido validado. En este sentido, la aplicación de SQUID estaría restringida en casos como el presente, en que el farmacóforo es ambiguo. Sin embargo, éste no puede ser el único motivo, ya que los valores de enriquecimiento se muestran por debajo de los obtenidos con CATS3D, que tampoco utiliza este tipo de etiquetado múltiple.

En este punto, se deciden explorar posibles factores y alternativas que puedan mejorar la bondad del modelo SQUID en esta aplicación.

4.9.1. Cambios en el esquema de *binning*

En la versión original (apartado 1.7.2), la codificación de las distancias entre las PPPs del modelo SQUID se realiza siguiendo un esquema de *binning* o partición de las mismas en 20 rangos equiespaciados de 0 a 20 Å. Así, una determinada distancia añade una cuenta al *bin* en el que se encuentra incluida en el vector de correlación. El mismo esquema se aplica a los descriptores CATS3D frente a los que se compara el vector de correlación de SQUID.

Se modifica dicho esquema implementándose el esquema de partición propuesto por Sheridan[276] (apartado 1.6.5) en el que se incluye la ponderación de la contribución de cada distancia a cada *bin* en función de su cercanía al centro de los *bins* vecinos. El rango de distancias considerado se mantiene de 0 a 20 Å.

En la Tabla 4.18 se muestran los valores de enriquecimiento obtenidos sobre un único subconjunto de Base_ACTIV_1 a partir de los modelos SQUID óptimos derivados de los alineamientos MODEL2PDB, MODEL3ALIGNED y MODEL4ALIGNED en estado neutro. Se puede observar cómo no hay apenas variación en los resultados, por lo que se descarta esta opción, manteniéndose el esquema de *binning* tradicional.

Tabla 4.18. Factores de enriquecimiento para el subconjunto Base_ACTIV_1_ale1 con el modelo SQUID derivado de tres alineamientos distintos utilizando dos esquemas de partición de distancias: el tradicional y el esquema ponderado de Sheridan. En ambos casos, se muestra únicamente el resultado obtenido para los parámetros óptimos (*cluster radius* y *feature-type weights*).

Modelo SQUID	**Esquema Partición Anterior**				**Esquema Sheridan**			
	% base de datos muestreada				% base de datos muestreada			
	1%	5%	10%	20%	1%	5%	10%	20%
MODEL2PDB	8.83	2.83	2.00	1.46	4.82	3.00	2.00	1.50
MODEL3ALIGNED	5.62	2.33	2.08	1.58	5.62	2.50	1.83	1.92
MODEL4ALIGNED	5.62	3.00	2.08	1.42	7.23	2.33	1.83	1.50

4.9.2. Influencia del Escalado de los descriptores CATS3D

En los resultados obtenidos hasta el momento, el descriptor CATS3D de cada una de las moléculas de la base de datos no se encuentra normalizado entre 0 y 1 (dividiendo por el valor máximo encontrado en cada molécula).
Se repite pues el proceso de búsqueda de similitud con los mismos modelos SQUID descritos en el apartado 4.7 y los descriptores CATS3D normalizados. En la Tabla 4.19 se presentan los factores de enriquecimiento obtenidos para el promedio de los cuatro subconjuntos de Base_ACTIV_1 para los modelos SQUID obtenidos a partir de los alineamientos MODEL3ALIGNED y MODEL4ALIGNED. De nuevo, no existe una mejora apreciable.

Tabla 4.19. Factores de enriquecimiento promedio y sus desviaciones estándar (paréntesis) en las búsquedas de similitud de los cuatro subconjuntos de Base_ACTIV_1 obtenidos con el modelo SQUID derivado de los alineamientos MODEL3ALIGNED y MODEL4ALIGNED. Comparación entre descriptores CATS3D normalizados por el valor máximo y sin normalizar para las moléculas de la base de datos.

Modelo SQUID	**CATS3D Sin Normalizar**				**CATS3D Normalizados**			
	% base de datos muestreada				% base de datos muestreada			
	1%	5%	10%	20%	1%	5%	10%	20%
MODEL3ALIGNED	5.21 (3.11)	2.50 (0.14)	1.98 (0.12)	1.58 (0.12)	4.42 (3.38)	2.33 (0.13)	2.04 (0.17)	1.77 (0.29)
MODEL4ALIGNED	6.02 (1.04)	3.21 (0.25)	2.06 (0.28)	1.41 (0.09)	7.02 (0.77)	3.42 (0.52)	2.11 (0.14)	1.39 (0.09)

4.9.3. Introducción de Conservación explícita de *features*

SQUID ha sido diseñado para incluir información "difusa" o *fuzzy* sobre la conservación y tolerancia de las *features* farmacofóricas en el conjunto de moléculas activas sobre las que se deriva el modelo. Esta conservación se calcula según la ecuación [1.86] y queda incluida en la codificación del modelo en un vector de correlación según la ecuación [1.87].

En este punto, se analiza la influencia de considerar la conservación de manera explícita. Es decir, en la derivación del modelo se incluyen únicamente aquellos PPPs con un porcentaje de conservación en las moléculas superior a un valor de *cutoff* determinado por el usuario. Así, se han derivado los siguientes modelos a partir de los alineamientos iniciales, en estado neutro:

- MODEL2PDB: conservación = 1 (presente en las dos moléculas del alineamiento).
- MODEL3ALIGNED: conservación = 0.5 y 1.
- MODEL4ALIGNED: conservación = 0.5, 0.75 y 1.

Los modelos derivados anteriormente en el apartado 4.7 pueden considerarse con conservación nula (todos los PPPs se incluyen). En la derivación de estos modelos con conservación explícita, se repite el proceso de optimización del *cluster radius*, variándolo desde 0.1 a 3 Å con pasos de 0.1 Å y *feature-type weights*, con valores comprendidos desde 0.1 a 0.9 con pasos de 0.2. Únicamente se optimizan los parámetros para la búsqueda en un subconjunto aleatorio de Base_ACTIV_1, Base_ACTIV_1_ale1.

En la Figura 4.18 se muestran los modelos obtenidos para cada uno de los tres alineamientos considerados con sus correspondientes conservaciones y *cluster radius* utilizados para agrupar las *features*.

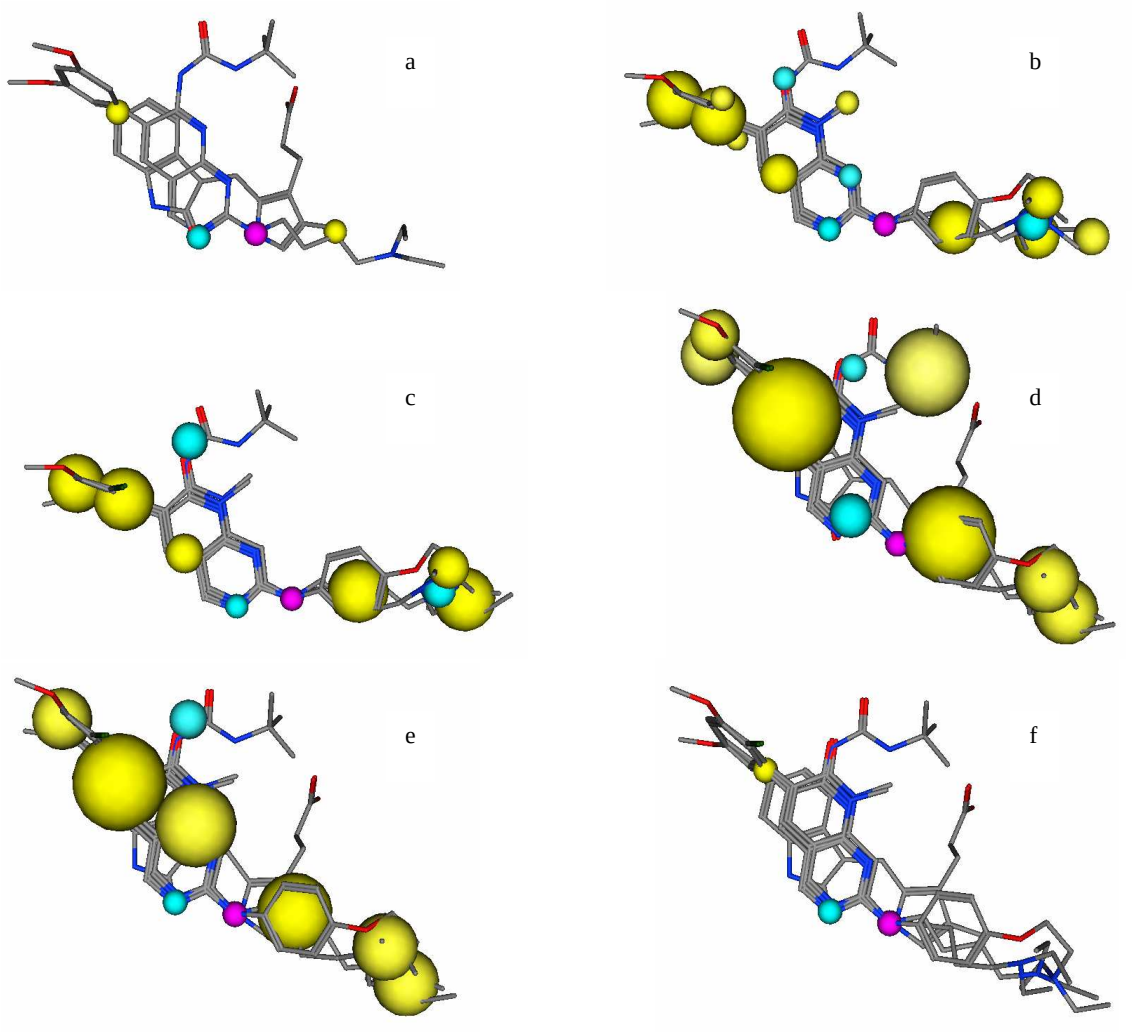


Figura 4.18. Modelos farmacofóricos SQUID seleccionados para cada uno de los tres alineamientos con diferentes grados de conservación. (a) MODEL2PDB con *cluster radius* de 0.6 y conservación 1. (b) MODEL3ALIGNED con *cluster radius* de 1.6 y conservación 0.50. (c) MODEL3ALIGNED con *cluster radius* de 2.0 y conservación 1. (d) MODEL4ALIGNED con *cluster radius* de 2.3 y conservación de 0.50. (e) MODEL4ALIGNED con *cluster radius* de 2.0 y conservación de 0.75. (f) MODEL4ALIGNED con *cluster radius* de 0.9 y conservación de 1. Los resultados de MODEL4ALIGNED con conservación 0.50 y 0.75 parecen contradictorios, ya que se observa una PPP adicional para el caso con conservación 0.50. Sin embargo, hay que considerar la diferencia de *cluster radius* a la que se han calculado uno y otro.

En la Tabla 4.20 se muestran los factores de enriquecimiento obtenidos en el cribado virtual con cada uno de estos modelos introduciendo el concepto de conservación explícitamente. Se muestran también los valores de *cluster radius* y *feature-type weights* utilizados para cada uno de ellos en el cribado retrospectivo del subconjunto Base_ACTIV_1_ale1. Se observa cómo la introducción de esta conservación, que conlleva una elevada reducción del número de PPPs presentes en los modelos SQUID (Figura 4.18), induce una gran pérdida de resolución y selectividad en ellos, alcanzándose factores de enriquecimiento por debajo de una selección aleatoria en la gran parte de los casos. Así, se abandona esta línea.

Tabla 4.20. Efecto de la introducción de conservación explícita en los factores de enriquecimiento obtenidos en el cribado retrospectivo del subconjunto aleatorio ACTIV_1_ale1. Se indican los valores de *cutoff* de conservación, *cluster radius* y *feature-type weights* asignados a cada uno de los tipos de interacción generalizados. +: catiónico, -: aniónico, *P*: polar, *D*: dador de puente de hidrógeno, *A*: aceptor de puente de hidrógeno y *H*: hidrofóbico.

Modelo SQUID	**Conservación**	***Cluster radius***	***Feature-type weights***						**% base de datos muestreada**			
			+	-	P	D	A	H	1%	5%	10%	20%
MODEL2PDB	0	2.2	0	0	0.1	0.5	0.1	0.5	4.82	3.17	2.08	1.46
MODEL2PDB	1	0.6	0	0	0.1	0.1	0.9	0.5	0.00	1.17	1.75	1.42
MODEL3ALIGNED	0	1.6	0	0	0	0.3	0.5	0.3	5.62	2.33	2.08	1.58
MODEL3ALIGNED	0.5	1.6	0	0	0	0.5	0.7	0.3	0.00	2.33	2.33	1.88
MODEL3ALIGNED	1	2	0	0	0	0.5	0.9	0.5	0.80	1.67	2.67	2.00
MODEL4ALIGNED	0	1.8	0	0	0.1	0.5	0.1	0.5	5.62	3.00	2.08	1.42
MODEL4ALIGNED	0.50	2.3	0	0	0.1	0.9	0.7	0.5	0.80	1.83	1.92	2.04
MODEL4ALIGNED	0.75	2	0	0	0.1	0.5	0.5	0.3	0.80	1.83	1.83	2.00
MODEL4ALIGNED	1	0.9	0	0	0.1	0.3	0.1	0.3	0.00	1.17	1.34	1.29

4.9.4. Modificación del Sistema de Asignación de Tipos Atómicos

Como se describe en el apartado 1.7.2, la versión original de SQUID utilizada hasta el momento, caracteriza los tipos atómicos de las moléculas activas del alineamiento según la función *ph4_aType* implementada en MOE.

Se decide analizar también la influencia del sistema de asignación de tipos atómicos, incorporando las funciones de MOE *feature_map_PCH* y *feature_pos*, correspondientes al esquema farmacofórico PCH. En la Figura 4.19 se muestran la asignación de los tipos atómicos de las moléculas PD173074 (*arriba*) y SU5402 (*abajo*) según el esquema *ph4_aType* (*izquierda*) y PCH (*derecha*). La principal diferencia entre ambos esquemas es que el primero utiliza *atom types* polares (*verde*) y asigna individualmente a los átomos hidrófobos (*amarillo*), mientras que PCH no anota la característica polar, pero sí distingue entre átomos hidrófobos (*verde*) y el centro de anillos aromáticos (*amarillo*).

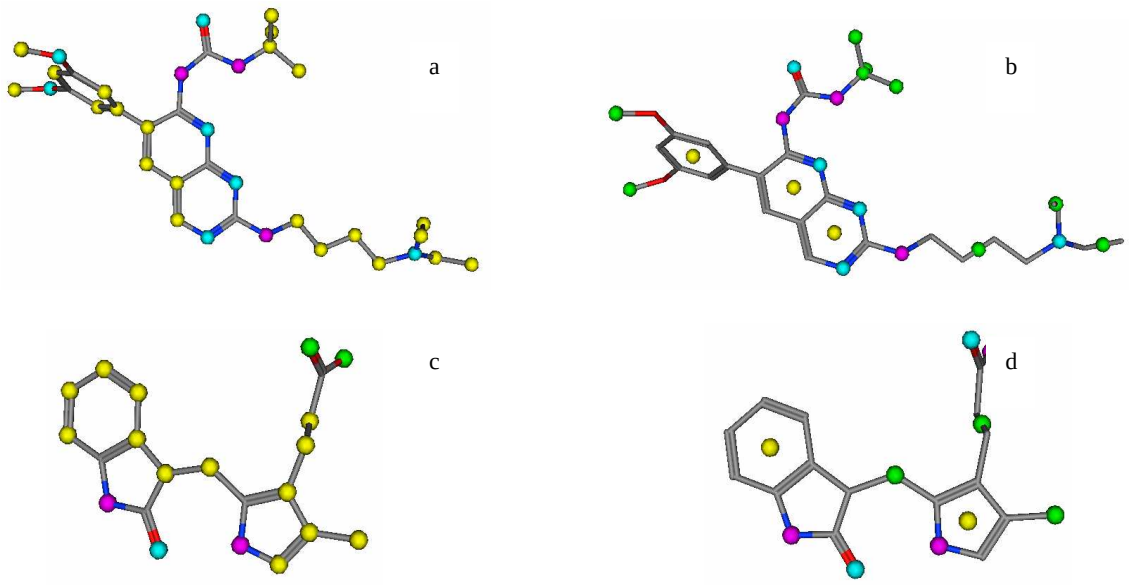


Figura 4.19. Anotación de tipos atómicos para las moléculas PD173074 (*arriba*) y SU5402 (*abajo*) según la función *ph4_aType* (*izquierda*) y el esquema PCH (*derecha*). Tipos de interacción generalizada para *ph4_aType*: aceptor (*cian*), dador (*magenta*), polar (*verde*) e hidrófobo (*amarillo*). Tipos de interacción generalizada para el esquema PCH: aceptor (*cian*), dador (*magenta*), centro de anillo aromático (*amarillo*) y átomo hidrofóbico (*verde*).

El impacto de este cambio se analiza en tres aplicaciones de SQUID diferentes:

- El que nos ocupa, generando los modelos SQUID a partir de los alineamientos iniciales MODEL2PDB, MODEL3ALIGNED y MODEL4ALIGNED y utilizando el subconjunto aleatorio de 2120 moléculas Base_ACTIV_1_ale1.

- Búsqueda retrospectiva de inhibidores de la ciclooxigenasa 2 (COX-2).

- Cribado de ligandos de trombina.

Estas dos últimas aplicaciones son las utilizadas en la validación original del modelo SQUID[51], en las que se obtuvieron enriquecimientos notables. Como referencia se han tomado los modelos SQUID descritos en dicha publicación para la notación *ph4_aType*, reconstruyéndolos a partir del mismo alineamiento y condiciones de *cluster radius* y *feature-type weights*. En ambos casos, se trabaja sobre la base de datos COBRA, que contiene 92 moléculas activas y 4611 moléculas inactivas para la ciclooxigenasa 2 (COX-2) y 188 activos y 4517 inactivos para el caso de la trombina. Para más información concerniente a estas dos aplicaciones, consultar la referencia [51].

En todas las aplicaciones, se generan los modelos SQUID con notación PCH a partir del alineamiento inicial explorando los *cluster radius* desde 0 a 3.0 Å con pasos de 0.1. Para cada uno de ellos, se optimizan los *feature-type weights* exhaustivamente con valores de 0.1 a 0.5 y pasos de 0.2. La combinación óptima de *cluster radius* y *feature-type weights* se escoge según el valor máximo de *ev* (ecuación [4.5]). En la Figura 4.20 se muestran los correspondientes modelos SQUID obtenidos con el sistema de anotación propuesto para los cinco casos con sus correspondientes *cluster radius*. En el caso de los inhibidores de COX-2 y trombina, se presentan también los modelos SQUID derivados con la notación *ph4_aType* (referencia [51]).

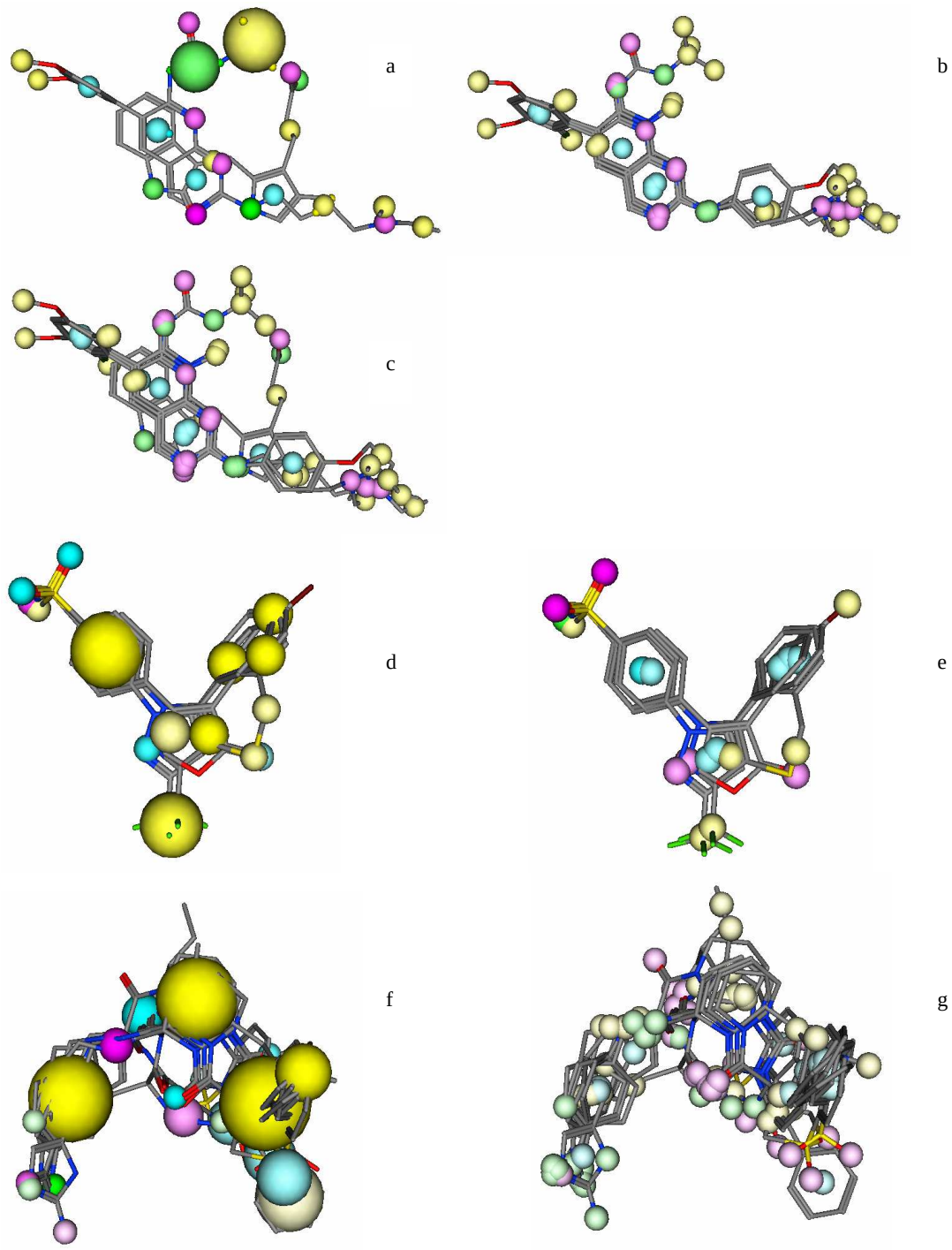


Figura 4.20. Modelos SQUID derivados con el esquema de anotación PCH. (a) MODEL2PDB neutro con *cluster radius* de 2.3. (b) MODEL3ALIGNED neutro con *cluster radius* de 0. (c) MODEL4ALIGNED con *cluster radius* de 0. (d) modelo SQUID derivado para los inhibidores de COX-2 con *cluster radius* de 1.4 y anotación *ph4_aType*, presentado en la referencia [51]. (e) modelo SQUID derivado para los inhibidores de COX-2 con *cluster radius* de 0.3 y anotación PCH, (f) modelo SQUID derivado para los inhibidores de Trombina con *cluster radius* de 2.0 y anotación *ph4_aType*, presentado en la referencia [51]. (g) modelo SQUID derivado para los inhibidores de Trombina con *cluster radius* de 0 y anotación PCH. Leyenda de colores en esquema *ph4_aType*: dadores de puente de hidrógeno (*magenta*), aceptores de puente de hidrógeno (*cian*), átomos hidrofóbicos (*amarillo*), polares (*verde*), cationes (*azul oscuro*) y aniones (*rojo*). La leyenda de colores en esquema PCH: dadores de puente de hidrógeno (*verde*), aceptores de puente de hidrógeno (*cian*), átomos hidrofóbicos (*amarillo*), centro de anillos aromáticos (*cian*), cationes (*azul oscuro*) y aniones (*rojo*).

En la Tabla 4.21 se listan los factores de enriquecimiento obtenidos con cada uno de estos modelos SQUID con anotación PCH, se incluyen también los resultados de los modelos de anotación *ph4_aType*. Se muestran los valores de *cluster radius* y *feature-type weights* utilizados para cada uno de ellos en el cribado retrospectivo del subconjunto Base_ACTIV_1_ale1 y los inhibidores de COX-2 y trombina. En primer lugar, se observa cómo los mejores modelos principalmente se obtienen con *cluster radius* bajos o incluso nulos, ya que la agrupación de átomos hidrófobos de los anillos aromáticos en esta notación se encuentra implícita en la asignación de los centros de dichos anillos.
No existe una tendencia de mejora o empeoramiento claro respecto a la notación anterior. Así, en el caso de los inhibidores de COX-2 se alcanza un ligero incremento de *ef*, pero en los ligandos de trombina y los modelos MODEL3ALIGNED y MODEL4ALIGNED la disminución de *ef* es considerable. Por lo tanto, se decide continuar con el esquema de anotación de la función *ph4_aType*.

Tabla 4.21. Factores de enriquecimiento obtenidos en el cribado retrospectivo del subconjunto aleatorio Base_ACTIV_1_ale1, cribado de inhibidores de COX-2 y ligandos de trombina con dos esquemas de anotación de tipos atómicos: *ph4_aType* y PCH. Se indican los valores de *cluster radius* y *feature-type weights* asignados a cada uno de los tipos de interacción generalizados. +: catiónico, -: aniónico, *P/Ar*: polar (*ph4_aType*) o centro de anillo aromático (PCH), *D*: dador de puente de hidrógeno, *A*: aceptor de puente de hidrógeno y *H*: Hidrofóbico.

Modelo SQUID	**Notación Atom Types**	***Cluster radius***	***feature-type weights***						**% base de datos muestreada**			
			+	-	P/ Ar	D	A	H	1%	5%	10%	20%
MODEL2PDB	*ph4_aType*	2.2	0	0	0.1	0.5	0.1	0.5	4.82	3.17	2.08	1.46
MODEL2PDB	PCH Scheme	2.3	0	0	0.5	0.1	0.1	0.1	4.82	2.33	1.5	1.08
MODEL3ALIGNED	*ph4_aType*	1.6	0	0	0	0.3	0.5	0.3	5.62	2.33	2.08	1.58
MODEL3ALIGNED	PCH Scheme	0	0	0	0.5	0.3	0.5	0.5	0.83	3.17	2.67	1.92
MODEL4ALIGNED	*ph4_aType*	1.8	0	0	0.1	0.5	0.1	0.5	5.62	3.00	2.08	1.42
MODEL4ALIGNED	PCH Scheme	0	0	0	0.5	0.3	0.5	0.5	1.61	2.00	2.00	1.75
COX-2	*ph4_aType*	1.4	0	0	0	0.1	0.4	0.3	40.7	15.5	7.97	4.41
COX-2	PCH Scheme	0.3	0	0	0.3	0.1	0.5	0.3	45.9	15.7	8.39	4.57
Trombina	*ph4_aType*	2.0	0	0	0.4	0.5	0.3	0.5	19.1	9.92	6.66	3.74
Trombina	PCH Scheme	0	0	0	0.4	0.9	0.4	0.4	2.51	4.09	3.38	2.51

4.9.5. Modificación de los descriptores usados en la caracterización de la base de datos: Conexión SQUID-SQUID

Tal y como se ha comentado en el apartado 1.7.2, el vector de correlación en que se encapsula el modelo SQUID se compara en la búsqueda de similitud con los descriptores CATS3D de las moléculas de la base de datos, evitándose así el alineamiento de estas moléculas frente al modelo SQUID. En este punto se decide explorar la posibilidad de describir las moléculas contenidas en la base de datos también de una manera *fuzzy*, es decir, suavizando los diagramas de correlación obtenidos tradicionalmente con CATS3D. Para ello, se describen las moléculas con el vector de correlación calculado aplicando independientemente para cada molécula el modelo SQUID. Es decir, cada compuesto de la base de datos equivale a un alineamiento inicial sobre el que el programa SQUID calcula el correspondiente modelo. En todos los casos, el vector de correlación final obtenido se normaliza entre 0 y 1, dividiendo por el valor máximo.

De este modo, la comparación previa establecida entre gaussianas (modelo farmacofórico SQUID) y picos (descriptores CATS3D para las moléculas de la base de datos), se convierte en una comparación entre gaussianas (modelo farmacofórico SQUID) y gaussianas (descriptores SQUID para las moléculas de la base de datos). En adelante se referirá a este esquema como SQUID-SQUID, en contraposición al anterior SQUID-CATS3D.

Esta comparación aporta otras dos ventajas adicionales:

- Dado que en el esquema previo SQUID-CATS3D no se comparan exactamente los mismos objetos, se utiliza como métrica de similitud el índice de la ecuación [1.88]. El sistema planteado SQUID-SQUID cualifica el uso de las métricas de similitud/distancia comúnmente empleadas, como la distancia Manhattan y la distancia Euclídea (Tabla 1.8).

- Se puede prescindir de utilizar los *feature-type weights* en la comparación SQUID-SQUID, reduciéndose el coste computacional asociado a su optimización.

En principio, se analiza el funcionamiento de este planteamiento SQUID-SQUID en el caso de los farmacóforos de TKs. En apartados posteriores se describe su comportamiento en el caso de las búsquedas retrospectivas de inhibidores de COX-2 y trombina. Las condiciones del análisis son:

- Únicamente se trabaja con los alineamientos iniciales de las moléculas MODEL3ALIGNED y MODEL4ALIGNED. Como *cluster radius* se contemplan las siguientes posibilidades en los dos casos: 0, 1.6, 1.8 y 2.2.

- Se caracterizan los cuatro subconjuntos aleatorios compuestos de 2120 moléculas extraídos de Base_ACTIV_1 y la base de datos Base_COBRA con los vectores de correlación obtenidos a distintos *cluster radius*, los mismos que los utilizados en la derivación del correspondiente modelo farmacofórico SQUID frente al que son comparados.

- Como métricas de similitud se prueban: distancia Manhattan, distancia Euclídea, coeficiente de Tanimoto y el coeficiente del coseno.

En general, en todos los cribados realizados se ha encontrado que estas dos métricas tipo (a-b) se comportan mejor que los coeficientes de asociación (a*b) como Tanimoto y coseno. Por brevedad, se incluyen únicamente los factores de enriquecimiento encontrados en los modelos significativos y concluyentes alcanzados con las métricas Manhattan y Euclídea (Tabla 4.22).

Tabla 4.22. Factores de enriquecimiento, *ef*, encontrados para las dos bases de datos Base_ACTIV_1 y Base_COBRA según la metodología SQUID-SQUID a diferentes porcentajes de base de datos muestreada (1%, 5%, 10% y 20%). Los resultados para Base_ACTIV_1 están calculados como el promedio de los cuatro subconjuntos aleatorios, en paréntesis se muestra su desviación estándar.

MÉTRICA	**MODELO**	***ef* Base_ACTIV_1** % Base de datos muestreada				***ef* Base_COBRA** % Base de datos muestreada			
		1%	5%	10%	20%	1%	5%	10%	20%
Euclídea	MODEL3ALIGNED-0 *cluster radius*	9.24 (1.91)	4.13 (0.70)	3.27 (0.40)	2.22 (0.06)	5.76	4.23	3.17	2.60
	MODEL3ALIGNED-1.6 *cluster radius*	8.23 (2.31)	4.25 (0.44)	3.46 (0.26)	2.32 (0.06)	4.80	3.27	2.69	2.88
	MODEL4ALIGNED-0 *cluster radius*	0.80 (0.00)	2.09 (0.10)	1.28 (0.08)	0.81 (0.14)	0.00	0.19	0.38	0.58
	MODEL4ALIGNED-2.2 *cluster radius*	4.62 (0.40)	1.71 (0.16)	1.75 (0.20)	1.64 (0.10)	0.96	0.96	1.73	1.49
Manhattan	MODEL3ALIGNED-0 *cluster radius*	8.03 (1.74)	4.38 (0.66)	3.38 (0.51)	2.45 (0.05)	5.76	3.85	2.59	2.55
	MODEL3ALIGNED-1.6 *cluster radius*	8.63 (2.65)	4.96 (0.08)	3.52 (0.26)	2.57 (0.07)	1.92	2.31	2.69	2.74
	MODEL4ALIGNED-0 *cluster radius*	0.60 (0.77)	2.04 (0.09)	1.48 (0.08)	1.07 (0.19)	0.00	0.19	1.06	1.30
	MODEL4ALIGNED-2.2 *cluster radius*	4.02 (0.66)	2.79 (0.44)	2.04 (0.22)	1.49 (0.10)	0.00	1.54	1.54	1.63

Atendiendo a MODEL3ALIGNED, se comprueba cómo los factores de enriquecimiento se incrementan con respecto a los obtenidos en la conexión previa SQUID-CATS3D (Tabla 4.16). Por ejemplo, en el primer 1% de base de datos muestreada, el promedio de los cuatro subconjuntos de Base_ACTIV_1 se incrementa desde 5.21 (SQUID-CATS3D) a un ~8-9 en este caso (SQUID-SQUID). El incremento correspondiente en BASE_COBRA es mayor, desde un 0.95 a un 4-5.
Se han incluido los resultados alcanzados con un *cluster radius* nulo y de 1.6 para enfatizar cómo la "difusión" (*fuzziness*) influye sobre la diversidad de *scaffolds* recuperados. En la Figura 4.21 se ilustra la distribución de los seis *scaffolds* en el primer 1% de base de datos Base_ACTIV_1 muestreada con tres modelos derivados del alineamiento MODEL3ALIGNED. Para mostrar el efecto *fuzzy*, se incluye también la distribución obtenida con la búsqueda con descriptores CATS3D focalizando frente a PD173074 (*violeta*), mostrada previamente en la Figura 4.17. Con CATS3D, únicamente se detectan compuestos del mismo *scaffold* que los presentes en PD173074. El modelo SQUID-SQUID con *cluster radius* nulo (*granate*), pese a que no recupera tantos compuestos como CATS3D (*ef* de 8.03 frente a 10.04 en promedio), sí que explora más *scaffolds*. Al incrementar el valor de *cluster radius* a 1.6 (*amarillo*) y con ello el grado de *fuzziness*, aumenta también la diversidad. La distancia Manhattan rinde mejores resultados en términos de diversidad que la distancia Euclídea (*verde*), apreciación también advertida en otros casos no mostrados.

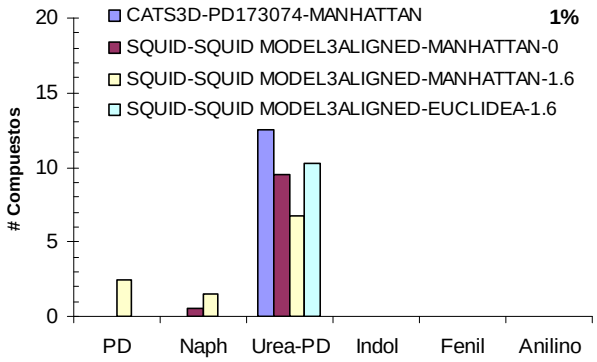


Figura 4.21. Resultados del análisis de diversidad de *scaffolds* para Base_ACTIV_1 en el primer 1% de base de datos muestreada con la conexión SQUID-SQUID. En el eje de abcisas se presentan los seis *scaffolds* presentes en la base de datos. *PD*: pirido[2,3-*d*]pirimidinas, *Naph*: naftiridin-2(1*H*)-onas, *Urea-PD*: 7-alquilurea pirido[2,3-*d*]pirimidinas, *Indol*: indolin-2-onas, *Fenil*: 1-fenilbenzimidazoles y *Anilino*: 4-anilinoquinazolinas. En el eje de ordenadas, número absoluto de compuestos recuperados para cada *scaffold*, siendo 20 el máximo posible.

4.9.6. Modificación de los descriptores usados en la caracterización de la base de datos: SQUID-SQUID *not scaled*

Aunque los resultados obtenidos con la nueva propuesta SQUID-SQUID son mejores que los obtenidos con SQUID-CATS3D para MODEL3ALIGNED, no sucede lo mismo con MODEL4LIGNED (Tabla 4.22). Inspeccionando los vectores de correlación obtenidos para dicho modelo, se encuentra que la codificación del modelo SQUID en el vector de correlación sufre un artefacto generado por el escalado realizado al dividir por el número total de pares posibles de PPPs para dos tipos de interacción generalizados (ecuación [1.87]). MODEL4LIGNED introduce, respecto a MODEL3ALIGNED, la presencia de SU5402, con los dos oxígenos del carboxilato etiquetados como polares (Figura 4.16.e). A bajos *cluster radius*, cada uno de estos átomos compone una PPP y por lo tanto el factor de escalado (#*pairs* TP) de la ecuación [1.87] equivale a dos, maximizándose la influencia de este par de interacción polar-polar frente a otras interacciones encontradas en mayor abundancia que dos (el factor divide).

A medida que se incrementa el *cluster radius*, estos dos átomos se agrupan en una misma PPP, por lo que el factor de escalado pasa a ser uno y no existe tal artefacto. De hecho, los factores de enriquecimiento mejoran. Como puede observarse en la Tabla 4.22, se pasa de un *ef* de 0.8 con *cluster radius* nulo a un *ef* de ~4-5 con un *cluster radius* de 2.2.

Así, se opta por reconstruir los modelos farmacofóricos SQUID y describir las moléculas siguiendo el esquema SQUID, pero sin incluir este factor de escalado (ecuación [4.6]). En adelante, se refiere a esta opción como SQUID-SQUID *not scaled*.

$$\mathrm{CV}_d^{TP} = \sum_{p=1}\sum_{q=1}\frac{1}{2}\delta_{pq}^{TP}\left(\frac{w_p w_q}{\sqrt{2\pi}\left(\sigma_p+\sigma_q\right)}\exp\left(-\frac{1}{2}\frac{\left(D_2(p,q)-centre_d\right)^2\right)}{\left(\sigma_p+\sigma_q\right)^2}\right)\right) \qquad [4.6]$$

Las búsquedas retrospectivas se repiten para los modelos derivados de los alineamientos MODEL3ALIGNED y MODEL4ALIGNED y las cuatro métricas descritas en el apartado anterior. En la Tabla 4.23 se muestran los factores de enriquecimiento obtenidos a diferentes porcentajes de base de datos muestreada para los *cluster radius* óptimos. De nuevo, las métricas Manhattan y Euclídea son las que rinden mayores enriquecimientos. Además, los modelos obtenidos con *cluster radius* comprendido en el márgen de 0 a 1 Å son los que obtienen mayores factores de enriquecimiento, mientras que estos valores decaen a medida que se incrementa el grado de *fuzziness*.

Con esta última aproximación, los factores de enriquecimiento son los mejores obtenidos al aplicar la metodología SQUID de las tres situaciones posibles: SQUID-CATS3D (Tabla 4.16), SQUID-SQUID (Tabla 4.22) y SQUID-SQUID *not scaled* (Tabla 4.23). Así, en el primer 1% de base de datos Base_ACTIV_1 los *ef* promedio aumentan de un 5 a un 8-9 y a un 12-14 con cada una de las tres metodologías a partir del alineamiento MODEL3ALIGNED. Sobre el mismo alineamiento, y con Base_COBRA, *ef* se incrementa desde un 0.95 (SQUID-CATS3D) a un 4-5 (SQUID-SQUID) y finalmente a un 19-20 (SQUID-SQUID *not scaled*).

Respecto a los modelos derivados con MODEL4ALIGNED, en el cribado de Base_COBRA, también se encuentra este gran incremento. Desde lo que era prácticamente una selección aleatoria (SQUID-CATS3D y SQUID-SQUID), la aproximación SQUID-SQUID *not scaled* alcanza *ef* de ~8-11 en el primer 1% de base de datos. Curiosamente, este aumento no es tan notable para el caso de Base_ACTIV_1, al menos para el primer 1% de base de datos muestreada, alcanzándose en los tres casos *ef* de ~5-7.

En este sentido, MODEL3ALIGNED sobrepasa a MODEL4ALIGNED en ambas bases de datos. Con la aproximación SQUID-SQUID *not scaled*, en el primer 1% de base de datos los valores de *ef* decaen de un ~12-14 a un ~5-7 para Base_ACTIV_1 y de un ~19-20 a un ~10-11 para Base_COBRA, al incorporar la molécula SU5402 en el alineamiento. Como se ha mencionado previamente, se sospecha que el motivo de este comportamiento diferencial sea la ambigüedad en la interacción farmacofórica, que SQUID no es capaz de reproducir fielmente dado que carece de múltiple asignación de tipos para un mismo PPP.

Tabla 4.23. Factores de enriquecimiento, *ef*, encontrados para las dos bases de datos Base_ACTIV_1 y Base_COBRA según la metodología SQUID-SQUID *not scaled* a diferentes porcentajes de base de datos muestreada (1%, 5%, 10% y 20%). Los resultados para Base_ACTIV_1 están calculados como el promedio de los cuatro subconjuntos aleatorios, en paréntesis se muestra su desviación estándar.

MÉTRICA	MODELO	*ef* Base_ACTIV_1 % Base de datos muestreada				*ef* Base_COBRA % Base de datos muestreada			
		1%	5%	10%	20%	1%	5%	10%	20%
Euclídea	MODEL3ALIGNED-0 *cluster radius*	12.85 (1.13)	6.50 (0.13)	4.69 (0.22)	2.83 (0.03)	19.20	7.69	4.52	2.88
	MODEL3ALIGNED-0.5 *cluster radius*	12.85 (1.13)	6.42 (0.22)	4.67 (0.16)	2.84 (0.70)	20.17	7.69	4.52	2.88
	MODEL3ALIGNED-1 *cluster radius*	11.04 (1.01)	6.46 (0.34)	4.44 (0.27)	2.82 (0.06)	20.17	7.88	4.52	2.98
	MODEL4ALIGNED-0 *cluster radius*	5.82 (1.20)	4.46 (0.42)	3.77 (0.32)	2.70 (0.07)	11.50	6.15	4.32	2.98
	MODEL4ALIGNED-0.5 *cluster radius*	6.22 (1.01)	4.46 (0.42)	3.77 (0.32)	2.68 (0.07)	10.57	6.15	4.32	2.88
	MODEL4ALIGNED-1 *cluster radius*	4.02 (1.31)	4.13 (0.57)	3.42 (0.24)	2.65 (0.08)	11.50	5.76	4.62	3.22
Manhattan	MODEL3ALIGNED-0 *cluster radius*	14.05 (0.46)	6.34 (0.30)	4.48 (0.17)	2.82 (0.09)	12.49	6.92	5.19	3.22
	MODEL3ALIGNED-0.5 *cluster radius*	14.46 (0.93)	6.17 (0.14)	4.42 (0.15)	2.82 (0.66)	14.41	6.73	4.71	3.17
	MODEL3ALIGNED-1 *cluster radius*	13.05 (0.77)	6.46 (0.44)	4.21 (0.27)	2.82 (0.07)	17.29	6.73	4.52	3.22
	MODEL4ALIGNED-0 *cluster radius*	6.42 (0.66)	4.75 (0.55)	3.63 (0.35)	2.51 (0.07)	8.64	6.73	4.42	2.98
	MODEL4ALIGNED-0.5 *cluster radius*	7.03 (1.01)	4.67 (0.41)	3.61 (0.34)	2.50 (0.10)	8.65	6.53	4.32	2.98
	MODEL4ALIGNED-1 *cluster radius*	5.83 (1.21)	4.59 (0.99)	3.27 (0.37)	2.46 (0.08)	10.57	5.76	4.52	2.79

A partir de este punto se escogen los modelos derivados del alineamiento de MODEL3ALIGNED con *cluster radius* de 0.5 y de MODEL4ALIGNED con *cluster radius* de 0.5. Éstos se muestran en la Figura 4.22.

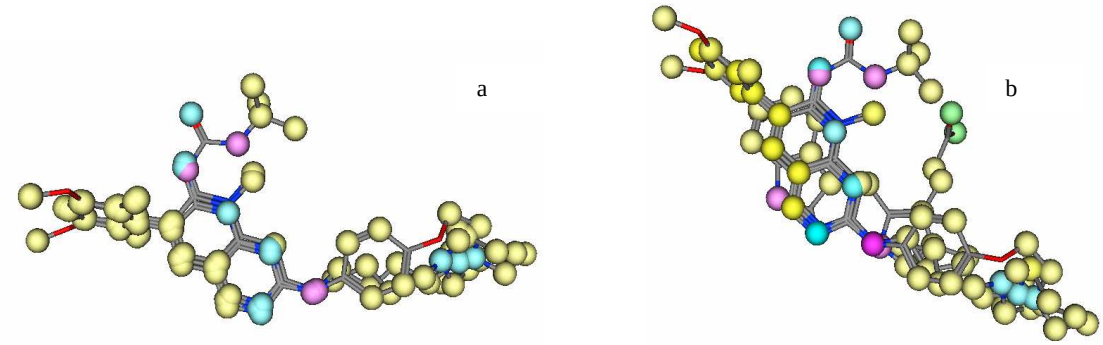


Figura 4.22. Modelos SQUID finales seleccionados aplicados en la aproximación SQUID-SQUID *not scaled*. (a) modelo derivado de MODEL3ALIGNED con *cluster radius* de 0.5 y (b) modelo derivado de MODEL4ALIGNED con *cluster radius* de 0.5.

4.10. Modelos Farmacofóricos finales seleccionados

En total se seleccionan como posibles modelos farmacofóricos o filtros:

- MOE_MODEL obtenido por refinamiento manual en MOE.

- Búsquedas de similitud con descriptores CATS3D utilizando PD173074 como *focus* y con distancia Manhattan.

- Modelo SQUID derivado de MODEL3ALIGNED con *cluster radius* de 0.5. Utilizado en la búsqueda retrospectiva en la conexión SQUID-SQUID *not scaled* con distancia Manhattan. En adelante referido como MODEL3ALIGNED.

- Modelo SQUID derivado de MODEL4ALIGNED con *cluster radius* de 0.5. Utilizado en la búsqueda retrospectiva en la conexión SQUID-SQUID *not scaled* con distancia Manhattan. En adelante referido como MODEL4ALIGNED.

En la Figura 4.23 se muestran las curvas de enriquecimiento obtenidas con cada uno de los modelos para el total de Base_ACTIV_1 (a) y Base_COBRA (b).

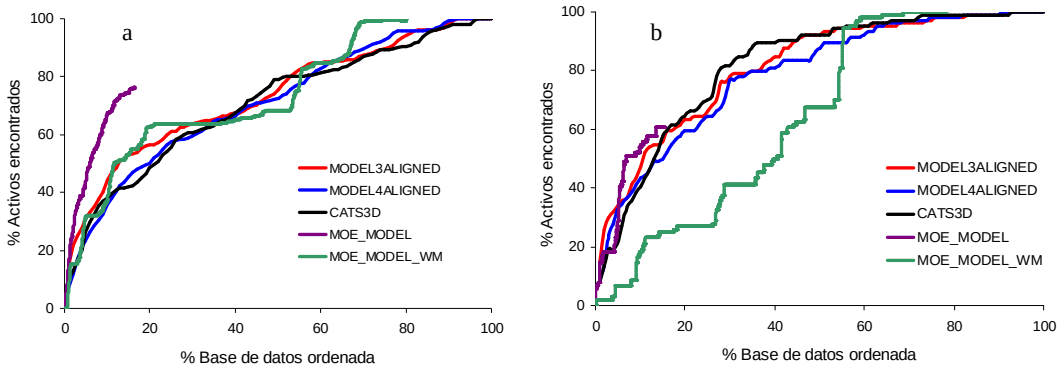


Figura 4.23. Curvas de enriquecimiento obtenidas en las búsquedas de similitud con los modelos finales seleccionados para (a) Base_ACTIV_1 y (b) Base_COBRA.

Para Base_ACTIV_1, MOE_MODEL sobrepasa al resto de procedimientos, principalmente por debajo del primer 10% de base de datos. Sin embargo, este comportamiento empeora al ignorar la asignación múltiple de *features* (MOE_MODEL_WM). MODEL3ALIGNED se comporta ligeramente mejor que CATS3D, mientras que MODEL4ALIGNED es el peor de ellos.

Interesantemente, para Base_COBRA, MODEL3ALIGNED alcanza mejores factores de enriquecimiento que MOE_MODEL hasta el 10% de base de datos muestreada. Así, en el primer 5% de base de datos, MODEL3ALIGNED captura el 34% de los activos mientras que MOE_MODEL el 27%. En el 10% de base de datos, MOE_MODEL sobrepasa a MODEL3ALIGNED (54% *vs* 47%), aunque ambos resultados son altamente favorables.

Todos los modelos se encuentran por encima de una selección aleatoria, incluso para el peor caso, MOE_MODEL_WM, con valores de *ef* entorno al ~1.3-1.9 a lo largo de los diferentes porcentajes de base de datos muestreada.

No debe confundir el hecho de que la curva de enriquecimiento de MOE_MODEL no alcance el 100% de activos, ya que este modelo únicamente devuelve un *score* (RMSD) para aquellas moléculas que clasifica como activas, por lo que un número de activos no se incluyen en esta curva ya que no superan el filtro (falsos negativos).

Análogamente al caso SQUID-CATS3D (apartado 4.8.2), se analiza la diversidad de los *scaffolds* recuperados. Para Base_ACTIV_1, los resultados se muestran en la Figura 4.24. En el primer 1% de base de datos muestreada, MODEL3ALIGNED (*amarillo*) reconoce más *scaffolds* que MOE_MODEL (*violeta*).

Éste recupera representantes de pirido[2,3-*d*]pirimidinas, 7-alquilurea pirido[2,3-*d*]pirimidinas y naftiridin-2(1*H*)-onas, mientras que MODEL3ALIGNED devuelve también 4-anilinoquinazolinas.

En el primer 5% de base de datos muestreada, a pesar del mencionado menor enriquecimiento por parte de MODEL3ALIGNED, tanto MOE_MODEL como MODEL3ALIGNED, se comportan comparablemente en términos de diversidad, recuperándose un representante de los seis *cores*, excepto indolinonas. MODEL4ALIGNED (*verde*) también es inferior a MODEL3ALIGNED en términos de diversidad. Tampoco MODEL4ALIGNED identifica indolinonas hasta el 20% de base de datos muestreada. CATS3D (*granate*) es superado por MODEL3ALIGNED, principalmente en el primer 1% de base de datos muestreada y con respecto a su capacidad par identificar 1-fenilbenzimidazoles.
Así, comparado con la situación anterior SQUID-CATS3D (Figura 4.17), la conexión propuesta SQUID-SQUID constituye una clara mejora en este caso de estudio.

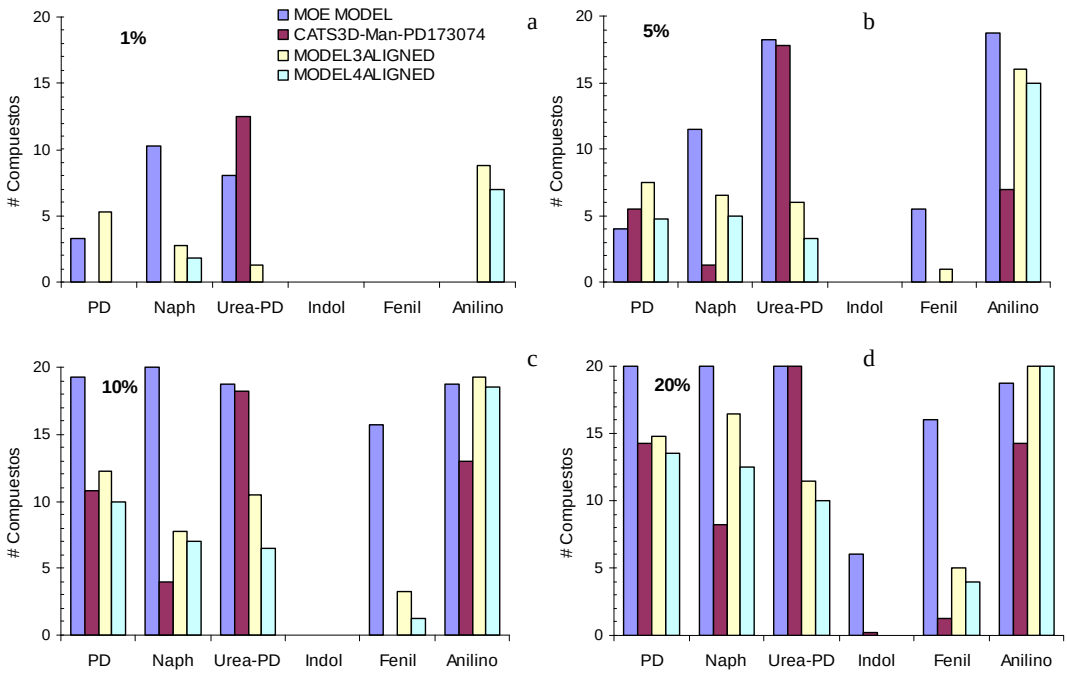


Figura 4.24. Resultados del análisis de diversidad de *scaffolds* para Base_ACTIV_1 a diferentes porcentajes de base de datos muestreada con los modelos farmacóforos finales seleccionados. (a) 1%, (b) 5%, (c) 10% y (d) 20%. En el eje de abcisas se presentan los seis *scaffolds* presentes en la base de datos. *PD*: pirido[2,3-*d*]pirimidinas, *Naph*: naftiridin-2(1*H*)-onas, *Urea-PD*: 7-alquilurea pirido[2,3-*d*]pirimidinas, *Indol*: indolin-2-onas, *Fenil*: 1-fenilbenzimidazoles y *Anilino*: 4-anilinoquinazolinas. En el eje de ordenadas, número absoluto de compuestos recuperados para cada *scaffold*, siendo 20 el máximo posible.

Respecto a Base_COBRA, en la Tabla 4.24 se lista el número de *molecular frameworks* activos (#AM) y totales (#M, verdaderos positivos + falsos positivos) recuperados por MODEL3ALIGNED, MODEL4ALIGNED y MOE_MODEL a diferentes porcentajes de base de datos muestreada. En paréntesis se muestra el número de moléculas activas recuperadas (H_a).

Tabla 4.24. Análisis de diversidad de *scaffolds* para Base_COBRA a diferentes porcentajes de base de datos muestreada. *#AM*: número total de *molecular frameworks* presentes en los compuestos activos identificados, H_a: número total de compuestos activos y *#M*: número total de *molecular frameworks* diferentes presentes en los compuestos recuperados (verdaderos positivos + falsos positivos).

	1%				**5%**				**10%**			
	Definición Grafo		Definición Atómica		Definición Grafo		Definición Atómica		Definición Grafo		Definición Atómica	
	#AM (Ha)	#M	#AM (Ha)	#M	#AM (Ha)	#M	#AM (Ha)	#M	#AM (Ha)	#M	#AM (Ha)	#M
MODEL3ALIGNED	8(15)	37	9	41	16(35)	170	20	200	22(49)	335	28	406
MODEL4ALIGNED	6(9)	34	6	40	17(34)	166	21	203	18(45)	323	23	399
MOE_MODEL	10(13)	41	12	45	16(28)	170	21	207	24(56)	325	33	408

En el primer 1% de base de datos muestreada, MOE_MODEL recupera 13 compuestos activos con 10 *frameworks* diferentes (23% del total de *frameworks* activos). Con el mismo número de moléculas ordenadas, MODEL3ALIGNED identifica más compuestos activos, 15, pero el número de *molecular frameworks* es de 8 (19% del total de *frameworks* activos). En este sentido, MODEL3ALIGNED se comporta ligeramente peor que MOE_MODEL. También el *pool* de inactivos identificados es más diverso en *molecular frameworks* para MOE_MODEL que para MODEL3ALIGNED. Una tendencia análoga se encuentra en el primer 5% de base de datos muestreada, donde los dos modelos identifican 16 *molecular frameworks* activos, aunque MOE_MODEL lo hace con 28 moléculas y MODEL3ALIGNED, con 35.

Finalmente, se calcula el solapamiento real entre los activos seleccionados entre estos tres últimos modelos: MODEL3ALIGNED, MODEL4ALIGNED y MOE_MODEL para Base_COBRA (Figura 4.25). En general, el solapamiento entre estos modelos es pequeño, excepto para MODEL3ALIGNED y MODEL4ALIGNED. En el primer 1% de base de datos muestreada, que supone 50 moléculas, hay un único activo compartido entre MODEL3ALIGNED y MOE_MODEL. En el primer 5%, el solapamiento entre estos modelos es de 9 compuestos, mientras que MODEL3ALIGNED contribuye con otros 26 compuestos y MOE_MODEL con 19 activos. Comparando MODEL3ALIGNED y MODEL4ALIGNED, el solapamiento es mucho mayor. Por ejemplo, en el primer 5% de base de datos muestreada, 30 compuestos activos de los 34 identificados por MODEL4ALIGNED son también recuperados por MODEL3ALIGNED.

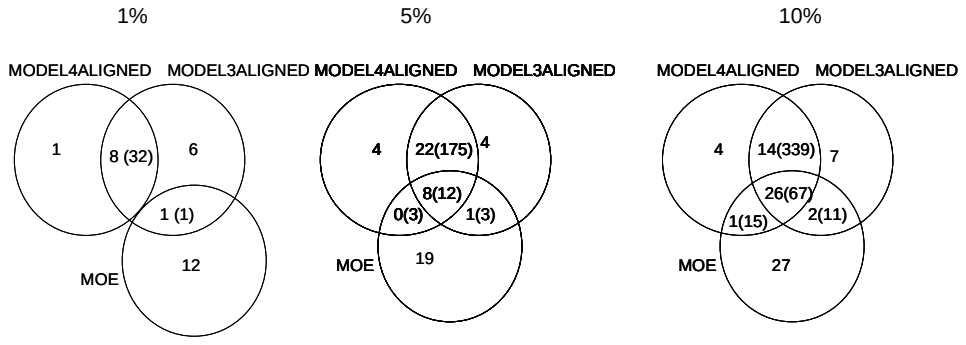


Figura 4.25. Solapamiento entre los subconjuntos seleccionados de Base_COBRA por MOE_MODEL, MODEL3ALIGNED y MODEL4ALIGNED. (a) 1% de base de datos muestreada, (b) 5% de base de datos muestreada y (c) 10% de base de datos. En las áreas de unión, los números en paréntesis corresponden al número total de *hits* compartidos (activos + falsos positivos). El resto corresponde al número de compuestos activos recuperados.

4.11. Modelo SQUID derivado de un único compuesto

Dado que la conexión SQUID-SQUID *not scaled* es exitosa, se ha analizado el comportamiento de aplicar este *fingerprint* farmacofórico *fuzzy* en el caso de focalizar frente a una única molécula diana. Es decir, el modelo SQUID de referencia se construye a partir de una única molécula individual o *focus* y se busca la similitud a éste de todas las moléculas de la base de datos descritas con el *fingerprint* SQUID. Los resultados se comparan frente a los obtenidos con *fingerprints* farmacofóricos tradicionales, como CATS3D, para ver cómo el grado de *fuzziness* influencia el cribado virtual frente a una única molécula diana.

Por cada uno de los métodos (CATS3D y SQUID descriptor), se realizan tres búsquedas de similitud independientes sobre Base_ACTIV_1, en cada una de ellas considerando como *focus* cada uno de los tres compuestos (**62**, **63** y PD173074) presentes en el alineamiento MODEL3ALIGNED (Figura 4.5.a). Como métrica, se escoge la distancia Manhattan. Para la metodología SQUID los mejores *cluster radius* para los compuestos **62** y **63** son nulos, mientras que para PD173074, el *cluster radius* óptimo es de 2.4. Este radio preferido de 2.4 se explica en términos de conservación: permite agrupar el entorno de urea de PD173074, no esencial para la actividad, reduciendo así su contribución al vector de correlación.

En la Figura 4.26 se muestra el análisis de diversidad de *scaffolds* recuperados por cada uno de los métodos en el primer 1% de base de datos muestreada. Comparando los modelos SQUID y CATS3D para cada uno de los compuestos por separado, se observa cómo por una parte se incrementa la diversidad en la identificación de *scaffolds* (por ejemplo, para PD173074 en *negro*) y además se recupera un mayor número de compuestos activos (por ejemplo, para los compuestos **62** y **63** en *azul* y *rojo*, respectivamente). Por motivos de comparación, se incopora también los resultados obtenidos con MODEL3ALIGNED en el apartado anterior (*verde*) para enfatizar cómo la distribución de *features* farmacofóricas individual de cada modelo se agrupa en un modelo *fuzzy* al aplicar la metodología SQUID. Así, MODEL3ALIGNED permite identificar 4 de los 6 *scaffolds*, mientras que con cada uno de los modelos individuales se recuperan unos 2 o 3 *scaffolds*.

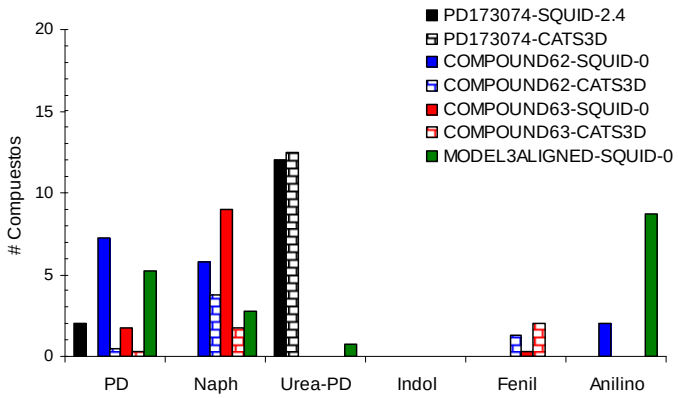


Figura 4.26. Impacto del grado de *fuzziness* en la diversidad de los *scaffolds* recuperados cuando se focaliza sobre un único compuesto activo mediante la metodología SQUID-SQUID *not scaled* y las búsquedas con CATS3D. Los resultados corresponden al primer 1% de Base_ACTIV_1 ordenada (promedio de los cuatro subconjuntos aleatorios). En el eje de abcisas se presentan los seis *scaffolds* presentes en la base de datos. *PD*: pirido[2,3-*d*]pirimidinas, *Naph*: naftiridin-2(1*H*)-onas, *Urea-PD*: 7-alquilurea pirido[2,3-*d*]pirimidinas, *Indol*: indolin-2-onas, *Fenil*: 1-fenilbenzimidazoles y *Anilino*: 4-anilinoquinazolinas. En el eje de ordenadas, número absoluto de compuestos recuperados para cada *scaffold*, siendo 20 el máximo posible.

Aunque no se muestran los resultados obtenidos para Base_COBRA, en el primer 1% de base de datos muestreada (~50 compuestos), MODEL3ALIGNED recupera 20 compuestos activos, de los cuales 11 son comunes a los identificados por los tres modelos SQUID individuales para cada plantilla y 9 son identificados exclusivamente por el modelo resultante de la agrupación de las tres plantillas.

4.12. Influencia de considerar bases de datos uniconformacionales o multiconformacionales

En este punto se analiza la dependencia de la metodología SQUID en el uso de bases de datos uniconformacionales (una única conformación por compuesto) o multioconformacionales. Se analiza sobre Base_COBRA para el conjunto de modelos óptimos encontrados siguiendo las metodologías: CATS3D, SQUID-CATS3D, SQUID-SQUID y SQUID-SQUID *not scaled*. En la Tabla 4.25 se tabula la *diferencia* en los factores de enriquecimiento obtenidos para una base de datos uniconformacional y una base de datos multiconformacional. En general, las diferencias son aceptables, excepto para los modelos SQUID-SQUID *not scaled* utilizando distancia Euclídea, donde claramente los enriquecimientos obtenidos sobre la base de datos uniconformacional son superiores a los obtenidos con una base de datos multiconformacional.

Tabla 4.25. Influencia de considerar bases de datos uniconformacionales o multiconformacionales. El resultado se expresa como la *diferencia* en los factores de enriquecimiento obtenidos para una base de datos uniconformacional y otra multiconformacional.

ALINEAMIENTO	MÉTODO	Base_COBRA			
		% Base de datos muestreada			
		1%	5%	10%	20%
MODEL3ALIGNED	SQUID-CATS3D - 1.6 *radius*	0.00	0.00	-0.38	-0.43
	SQUID-SQUID - 0 *radius* - Euclídea	2.91	0.10	-0.67	-0.25
	SQUID SQUID*not scaled* - 0 *radius* - Euclídea	11.61	0.01	0.20	0.24
	SQUID SQUID*not scaled* - 0.5 *radius* - Euclídea	10.68	0.01	0.20	0.24
	SQUID SQUID*not scaled* - 1 *radius* - Euclídea	6.88	0.39	0.20	0.29
	SQUID SQUID*not scaled* - 0 *radius* - Manhattan	0.15	0.77	0.96	0.10
	SQUID SQUID*not scaled* – 0.5 *radius* - Manhattan	1.12	0.20	0.48	0.09
	SQUID SQUID*not scaled* – 1 *radius* - Manhattan	1.15	0.19	0.20	0.24
MODEL4ALIGNED	SQUID-CATS3D - 1.8 *radius*	1.93	1.35	0.78	0.00
	SQUID-SQUID - 0 *radius* - Euclídea	0.00	0.19	0.77	0.19
	SQUID SQUID*not scaled* - 0 *radius* - Euclídea	2.96	-0.19	0.29	0.10
	SQUID SQUID*not scaled* - 0.5 *radius* - Euclídea	2.03	-0.19	0.29	0.09
	SQUID SQUID*not scaled* - 1 *radius* - Euclídea	2.01	-0.20	0.49	0.24
	SQUID SQUID*not scaled* - 0 *radius* - Manhattan	-0.85	0.77	0.58	-0.10
	SQUID SQUID*not scaled* – 0.5 *radius* - Manhattan	-0.84	0.57	0.48	-0.05
	SQUID SQUID*not scaled* – 1 *radius* - Manhattan	-0.82	0.00	0.10	-0.34
CATS3D	CATS3D PD173074-Manhattan	-1.83	0.96	0.97	0.29

4.13. Aplicación de la conexiones SQUID-SQUID y SQUID-SQUID *not scaled* en otros casos de estudio

Como se ha comentado, las conexiones SQUID-SQUID y SQUID-SQUID *not scaled* se han aplicado también en el cribado retrospectivo de inhibidores de COX-2 y ligandos de trombina, para los que la conexión SQUID-CATS3D se ha validado de manera exitosa.[51]

Los alineamientos iniciales corresponden a los descritos en la referencia [51] y mostrados en las Figuras 4.20.d (COX-2) y 4.20.f (trombina). En la construcción del modelo SQUID (escalado o *not scaled*) se escanean diferentes *cluster radius*, comprendidos entre 0 y 3 Å, con pasos de 0.2 Å. Las bases de datos correspondientes se caracterizan también con la descripción SQUID correspondiente y con los mismos *cluster radius* que los empleados en la generación del modelo SQUID de referencia. Como métricas, se estudian la distancia Manhattan, la distancia Euclídea, el coeficiente de Tanimoto y el del coseno.

De todas las posibles combinaciones de *cluster radius*, se encuentra una preferencia similar a la obtenida para la conexión SQUID-CATS3D: un *cluster radius* de 1.4 Å para COX-2 y un *cluster radius* de 2.4 Å para los ligandos de trombina. De las métricas, la distancia Euclídea es la mejor para COX-2 y la distancia Manhattan para los ligandos de trombina. En la Tabla 4.26 se muestran los factores de enriquecimiento obtenidos en las condiciones óptimas para cada una de las dos familias de compuestos.

Además, como referencia, se realiza una búsqueda de similitud con los descriptores CATS3D individualmente para cada una de las moléculas contenidas en cada uno de los alineamientos: 3 para COX-2 y 7 para los ligandos de trombina. En este caso, la distancia Manhattan es la métrica.

Tabla 4.26. Factores de enriquecimiento obtenidos en las búsquedas retrospectivas para las familias de inhibidores de COX-2 y Trombina con los métodos: SQUID-CATS3D, SQUID-SQUID, SQUID-SQUID *not scaled* y CATS3D.

MÉTODO	**COX-2**				**Trombina**			
	% Base de datos muestreada				% Base de datos muestreada			
	1%	5%	10%	20%	1%	5%	10%	20%
SQUID-CATS3D	40.66	15.48	7.97	4.42	19.10	9.92	6.66	3.74
SQUID-SQUID	17.72	5.51	3.51	3.78	10.05	7.77	5.68	3.41
SQUID-SQUID *not scaled*	25.03	13.79	7.98	4.15	19.1	8.89	5.28	3.33
CATS3D								
Plantilla 1	6.26	7.85	6.70	4.15	4.02	2.15	1.64	1.28
Plantilla 2	32.33	9.76	6.90	4.44	1.51	0.51	0.41	0.36
Plantilla 3	16.68	8.70	6.40	4.31	4.02	3.17	2.72	2.08
Plantilla 4					10.05	2.87	1.95	1.64
Plantilla 5					9.05	2.86	2.00	1.44
Plantilla 6					0.50	0.02	0.31	0.36
Plantilla 7					1.00	0.72	0.51	0.51

En el caso de los inhibidores de COX-2, la conexión original SQUID-CATS3D es la que rinde mayores factores de enriquecimiento. En este caso, la conexión SQUID-SQUID *not scaled* es comparativamente inferior, rindiendo un *ef* de 25 en el primer 1% de base de datos muestreada, frente al 40 alcanzado por la conexión anterior. Incluso, este factor de enriquecimiento se sitúa por debajo del alcanzado con la búsqueda de similitud con CATS3D focalizada sobre uno de los compuestos del alineamiento (*Plantilla 2*, 33), aunque a porcentajes superiores de base de datos muestreada, la conexión SQUID-SQUID *not scaled* supera a CATS3D.

En el caso de los ligandos de trombina, los factores de enriquecimiento obtenidos por las conexiones SQUID-CATS3D y SQUID-SQUID *not scaled* son muy similares. Además, se sitúan claramente por encima de las búsquedas de similitud con los descriptores CATS3D.

Con ello, no es posible concluir totalmente que una de las conexiones, SQUID-CATS3D o SQUID-SQUID *not scaled*, sea mejor que otra. Si bien para TKs, SQUID-SQUID *not scaled* permite alcanzar factores de rendimiento muy buenos, en el caso de los inhibidores de COX-2 la conexión SQUID-CATS3D es superior a la anterior.
Si bien en términos de efectividad no es posible establecer una conclusión contundente, en términos de eficiencia sí que la conexión SQUID-SQUID *not scaled* constituye una mejor alternativa a la conexión SQUID-CATS3D, ya que evita el uso de *feature-typed weights* y su optimización exhaustiva, ahorrando tiempo de cálculo.

4.14. Aplicación del modelo SQUID a un modelo farmacofórico con múltiple asignación de tipos

Finalmente, se ha seleccionado un caso de estudio para testar la hipótesis de la influencia de la presencia de *features* con múltiple asignación de tipos en los modelos farmacofóricos derivados con la metodología SQUID.

Se ha escogido un farmacóforo previamente[378] identificado para los receptores de serotonina 5-HT3 y que se encuentra analizado en detalle en el programa MOE como modelo para la construcción de modelos farmacofóricos con dicho programa. La base de datos se compone de 25 moléculas activas y 250 inactivas seleccionadas aleatoriamente de catálogos comerciales, tal y como proporciona el programa MOE. Como modelo farmacofórico de referencia, se escoge el modelo refinado propuesto en dicho programa, sin incluir *shape constraints* o características direccionales. Dicho modelo se muestra en la Figura 4.27. Está compuesto de cuatro *features*: dador y catiónico (*F1*), aromático (*F2*), aceptor (*F3*) y aromático (*F4*). Por motivos de comparación, se llevan a cabo también búsquedas farmacofóricas con el modelo previo pero sin incluir la *feature* con múltiple asignación de tipos, *F1*.

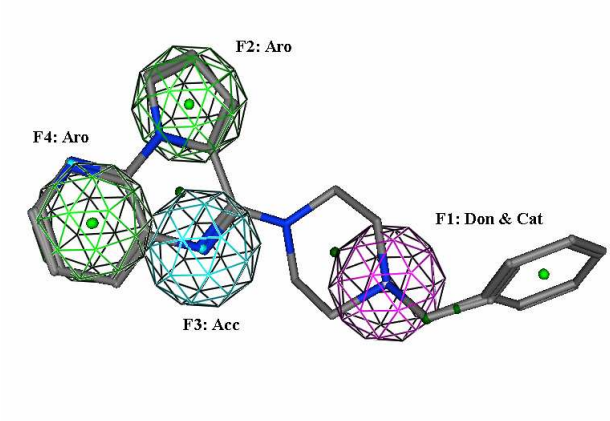


Figura 4.27. Modelo farmacofórico para los receptores de serotonina 5-HT3 identificado con MOE.

Se realizan búsquedas restrospectivas sobre esta misma base de datos con las metodologías SQUID-CATS3D y SQUID-SQUID *not scaled* con los parámetros de *cluster radius* y *feature-typed weight* (SQUID-CATS3D) optimizados. Para SQUID-SQUID *not scaled*, se analizan las cuatro métricas: distancia Manhattan, distancia Euclídea, coeficiente de Tanimoto y coseno. De ellas, la distancia Euclídea es la que rinde mejores factores de enriquecimiento.

En la Tabla 4.27 se muestran los factores de enriquecimiento obtenidos con los cuatro modelos farmacofóricos. La hipótesis farmacofórica formulada con MOE que contiene la múltiple asignación de tipos es la que rinde *ef* superiores por encima del primer 5% de base de datos muestreada.

Se observa cómo la respuesta del modelo al eliminar la *feature F1* es similar a la obtenida con las dos conexiones de SQUID. En este sentido, la *feature F1* determina el límite superior del modelo SQUID. Respecto a las conexiones, SQUID-SQUID *not scaled* y SQUID-CATS3D se comportan similarmente en este caso de estudio.

Tabla 4.27. Factores de enriquecimiento, *ef*, a diferentes porcentajes de base de datos muestreada obtenidos con diferentes metodologías: SQUID-CATS3D, SQUID-SQUID *not scaled*, farmacóforo del MOE y farmacóforo del MOE sin la *feature* con asignación múltiple (F1). *M* y *U* hacen referencia a si la base de datos empleada en la búsqueda retrospectiva es multiconformacional o uniconformacional, respectivamente.

MODELO	**Base**	***ef* SEROTONINA 5-HT3** % Base de datos muestreada			
		1%	5%	10%	20%
SQUID-CATS3D - 2.6 *radius*	M	11.00	10.21	8.25	4.60
	U	11.00	7.86	7.46	4.40
SQUID-SQUID*not scaled* – Euclídea	U(1.4)*	11.00	10.21	8.25	4.60
	M(2.8)	11.00	10.21	7.46	4.60
Farmacóforo del MOE	M	11.00	11.00	9.82	-
Farmacóforo del MOE sin F1	M	11.00	10.21	8.64	-

* valor de *cluster radius* óptimo para esta base de datos.

4.15. Filtrado de las quimiotecas BIB_Oxo, BIB_Amino y BIB_Hidro

Una vez establecidas y validadas las distintas metodologías de filtrado basados en la estructura del ligando, se procede a filtrar las quimiotecas BIB_Oxo, BIB_Amino y BIB_Hidro descritas en el capítulo 3. Para ello se utilizan los cuatro modelos listados en el apartado 4.10 y una búsqueda de similitud mediante descriptores CATS3D focalizando en el compuesto SU5402 con distancia Manhattan.

En la Figura 4.28 se muestra la abundancia de compuestos de cada una de las tres familias, según la sustitución en la posición C-4, que se encuentran en los primeros 100 compuestos ordenados de la base de datos.

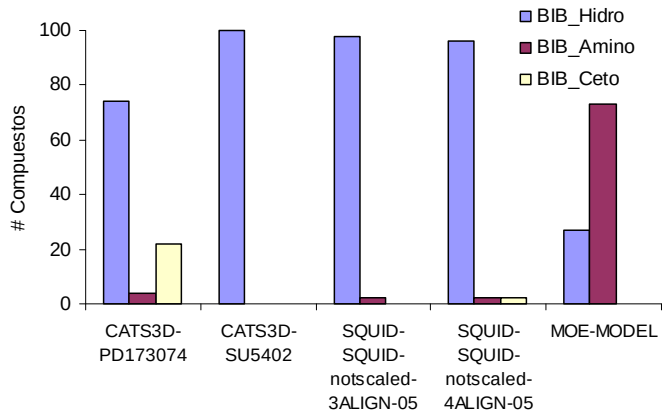


Figura 4.28. Abundancia de compuestos en cada una de las tres quimiotecas para los 100 primeros compuestos ordenados de la quimioteca de 106920 compuestos por cada uno de los cinco modelos de búsqueda farmacofórica.

Se observa cómo los modelos CATS3D y SQUID priorizan los derivados con hidrógeno en posición C-4, mientras que MOE_MODEL detecta principalmente derivados con un grupo amino en dicha posición.

En la Figura 4.29 se muestra el orden de detección preferente de cada una de estas tres familias en función del porcentaje de base de datos analizada.

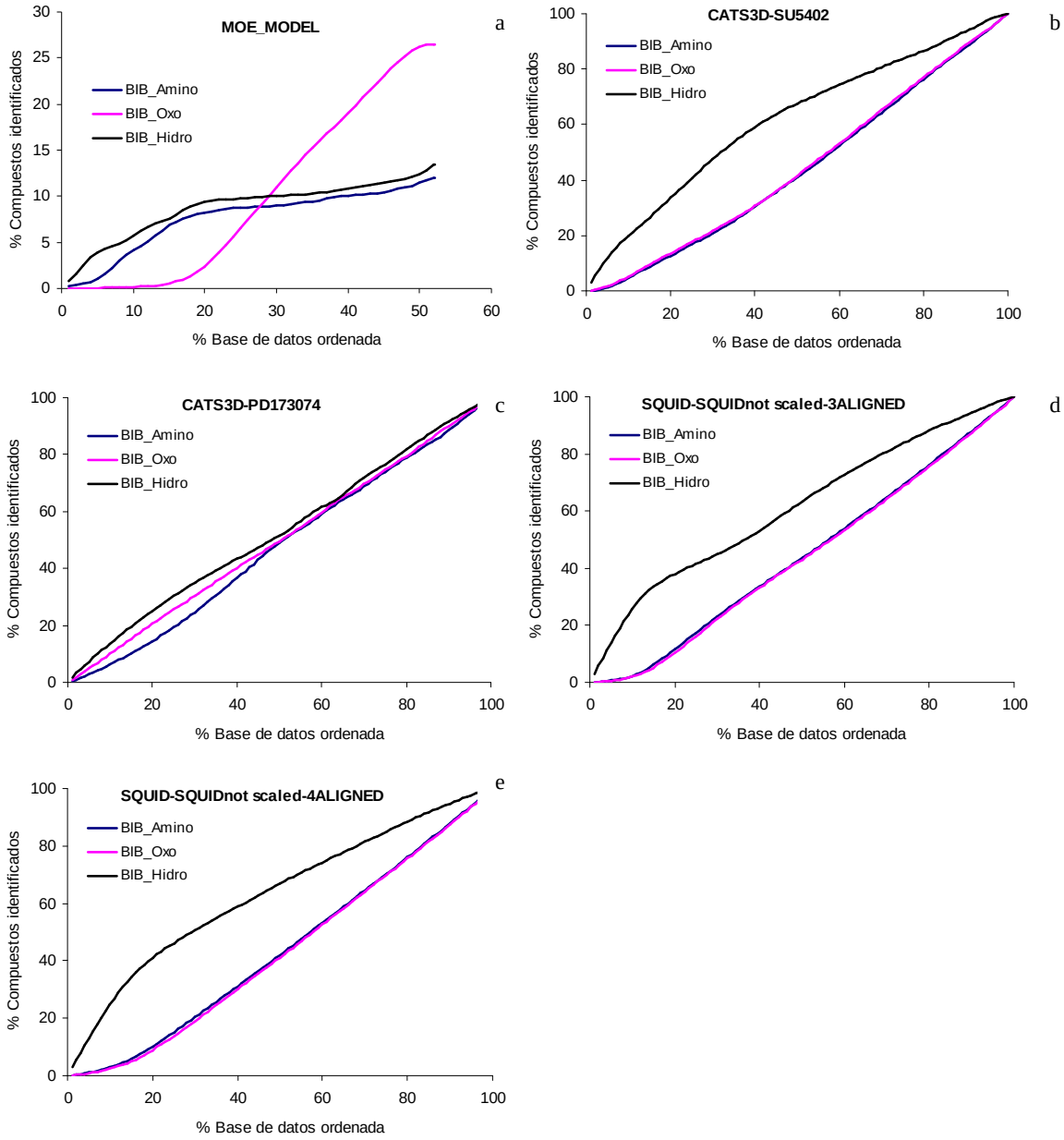


Figura 4.29. Orden de identificación de los compuestos de cada una de las tres familias BIB_Amino, BIB_Oxo y BIB_Hidro en función del porcentaje de base de datos ordenada. (a) Modelo MOE_MODEL. (b) Búsqueda de similitud con CATS3D y SU5402 como *focus* con distancia Manhattan. (c) Búsqueda de similitud con CATS3D y PD173074 como *focus* con distancia Manhattan. (d) Búsqueda de similitud con SQUID-SQUID *not scaled* derivado a partir del alineamiento MODEL3ALIGNED. (e) Búsqueda de similitud con SQUID-SQUID *not scaled* derivado a partir del alineamiento MODEL4ALIGNED.

Se observa cómo MOE_MODEL tiene una preferencia por los compuestos con sustituyentes 4-hidro y 4-amino, apareciendo los compuestos de la quimioteca BIB_Oxo mucho después, aunque están incluidos en mayor número en la lista de *hits*. En la búsqueda de similitud con CATS3D y PD173074 como compuesto *query*, las tres familias se obtienen prácticamente a la par, aunque BIB_Amino tiene una ligera preferencia. En los tres casos restantes, la quimioteca BIB_Hidro es claramente preferente a las otras dos.

Finalmente, se compara el ranking de los 100 primeros candidatos seleccionados de la quimioteca de 106920 compuestos por cada método *versus* el ranking de los compuestos activos, tanto del *pool* ACTIV_1 (288 compuestos, divididos en los seis posibles *scaffolds*) y el *pool* COBRA (104 compuestos) por el mismo método. Los resultados del orden de identificación de todos estos compuestos (288+104+100) en función del porcentaje de base de datos muestreada se ilustran en la Figura 4.30.

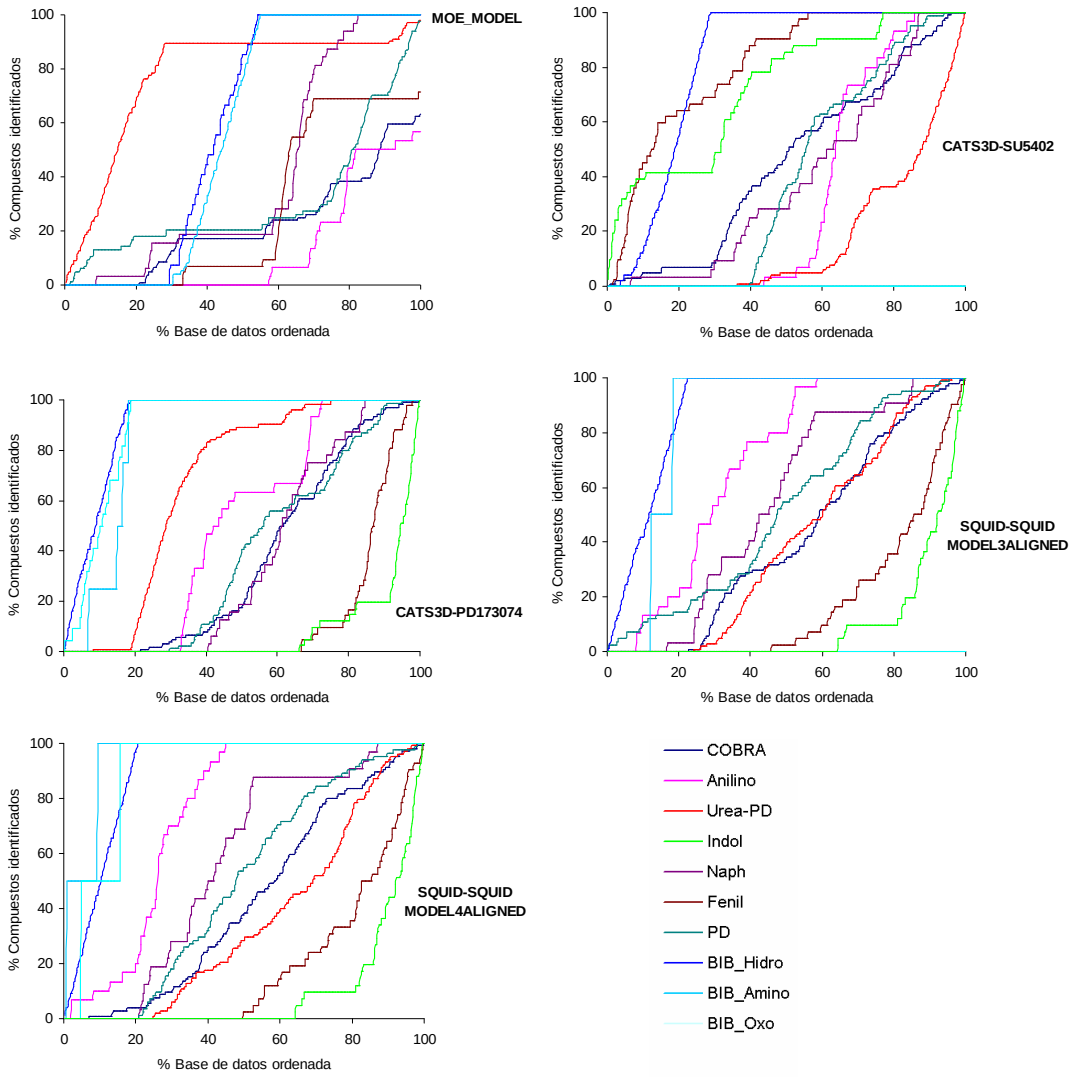


Figura 4.30. Orden de identificación del total de compuestos activos de los dos *pools* de activos considerados (ACTIV_1 y COBRA) más los 100 primeros compuestos seleccionados por cada método de la quimioteca de 106920 compuestos. Los resultados para el *pool* ACTIV_1 y la búsqueda prospectiva se muestran clasificados en función del *scaffold* de la molécula. (a) Modelo MOE_MODEL. (b) Búsqueda de similitud con CATS3D y SU5402 como *focus* con distancia Manhattan. (c) Búsqueda de similitud con CATS3D y PD173074 como *focus* con distancia Manhattan. (d) Búsqueda de similitud con SQUID-SQUID *not scaled* derivado a partir del alineamiento MODEL3ALIGNED. (e) Búsqueda de similitud con SQUID-SQUID *not scaled* derivado a partir del alineamiento MODEL4ALIGNED.

Se observa cómo los modelos SQUID-SQUID *not scaled* y CATS3D ordenan en general a los compuestos de la quimioteca de pirido[2,3-*d*]pirimidinas por delante de otros compuestos con actividad testada, en particular a los BIB_Hidro. Para estos métodos, los 1-fenilbenzimidazoles y las indolin-2-onas adoptan un menor valor de similitud, quedando ordenados en las últimas posiciones. De hecho, únicamente en la búsqueda con CATS3D y SU5402 como *focus* quedan ordenados en las primeras posiciones, ya que, evidentemente el *scaffold* es compartido al tratarse SU5402 de una indolinona. También para este modelo los compuestos de BIB_Hidro quedan ordenados por delante de otros compuestos activos. Finalmente, el modelo MOE_MODEL identifica primeramente compuestos representativos de los *scaffolds* contenidos en el alineamiento del que son derivados (pirido[2,3-*d*]pirimidinas y sus derivados urea y naftiridinas) y posteriormente identifica compuestos de las quimiotecas BIB_Hidro, BIB_Amino y del *pool* de activos de COBRA.

Estos resultados son tal vez demasiado optimistas para la caracterización de la quimioteca de 106920 compuestos. En realidad, no sorprende que los índices de similitud sean tan elevados, ya que el *scaffold* central de la quimioteca de 106920 compuestos únicamente difiere del *scaffold* de pirido[2,3-*d*]pirimidinas incluido en el alineamiento en la existencia o no de un doble enlace (Figura 3.1), no afectando a la asignación de los grupos farmacofóricos esenciales. De hecho, la validación de estos métodos, en especial del modelo farmacofórico del MOE (del que se dispone libremente) podría resultar más útil en aplicaciones futuras de búsqueda de estructuras claramente diferentes del *scaffold* pirido[2,3-*d*]piridimina.