

Artículo II

Clustering and Classifying Diverse HIV Entry Inhibitors Using a Novel Consensus Shape-Based Virtual Screening Approach: Further Evidence for Multiple Binding Sites within the CCR5 Extracellular Pocket

Violeta I. Pérez-Nuño,[†] David W. Ritchie,^{*,‡} Jose I. Borrell,[†] and Jordi Teixidó[†]

Grup d'Enginyeria Molecular, Institut Químic de Sarrià (IQS), Universitat Ramon Llull, Barcelona, Spain,
Department of Computing Science, King's College, University of Aberdeen, Aberdeen, U.K.

Received July 28, 2008

HIV entry inhibitors have emerged as a new generation of antiretroviral drugs that block viral fusion with the CXCR4 and CCR5 membrane coreceptors. Several small molecule antagonists for these coreceptors have been developed, some of which are currently in clinical trials. However, because no crystal structures for the coreceptor proteins are available, the binding modes of the known inhibitors within the coreceptor extracellular pockets need to be analyzed by means of site-directed mutagenesis and computational experiments. Previous studies have indicated that there is more than one binding site within the CCR5 extracellular pocket. This article investigates and develops this hypothesis using a novel spherical harmonic-based consensus shape clustering approach. The consensus shape approach is evaluated using retrospective virtual screening of CXCR4 and CCR5 inhibitors. Multiple combinations of CCR5 ligands in multiple trial superpositions are constructed to find consensus queries that give high virtual screening enrichments. Receiver–operator–characteristic performance analyses for both CXCR4 and CCR5 inhibitors show that the new consensus shape matching approach gives better virtual screening enrichments than existing shape matching and docking virtual screening techniques. The results obtained also provide strong evidence to support the notion that there are three main binding sites within the CCR5 extracellular cavity.

INTRODUCTION

Human immunodeficiency virus (HIV) entry inhibitors have emerged as a new generation of antiretroviral drugs which work by blocking interactions between the viral surface gp120 protein and the CXCR4 and CCR5 plasmatic membrane coreceptors of the host cell.^{1–4} A considerable number of small molecule antagonists for CXCR4 and CCR5 have been found to be effective for preventing viral entry, and some of them have been evaluated in clinical trials.^{5–9} However, no crystal structures of these coreceptors or their ligand-bound complexes are available. Consequently, several site-directed mutagenesis (SDM) and computational experiments have been carried out to identify the binding modes of the existing inhibitors. Analysis of the key CXCR4 SDM residues points to a well-defined localized binding cavity,¹⁰ but the CCR5 SDM residues are found to be spatially well-distributed around the pocket within the extracellular loops.¹¹ Moreover, the small-molecule inhibitors for CXCR4 are generally quite similar to each other, whereas CCR5 has many different inhibitors which derive from several diverse scaffold families. Several earlier computational binding experiments have indicated that different CCR5 ligands bind in fundamentally different ways within the CCR5 extracellular pocket.^{12–16} Furthermore, considering that (a) it is very difficult to superpose all the different families of CCR5 active compounds, (b) the results of retrospective virtual screening

(VS) enrichment studies are strongly dependent on the conformation of the query molecule, (c) SDM results suggest a large binding pocket within the extracellular loop region of the CCR5 structure, and (d) not all SDM mutations affect the binding of all ligands, there is good evidence to support a hypothesis that the known binders belong to two or more groups and that the members of each group bind to the same general region of the extracellular pocket. However, it is not clear a priori which actives might belong to which group. For example, different computational binding mode studies of ligands such as Aplaviroc,^{13,15} AD101,^{11,17} SCH-C,^{11,13–15,17} TAK-779,^{11,13,14,17–21} TAK-720,¹⁹ 2-aryl-4-(piperidin-1-yl)-butanamines,^{14,16} and 1,3,4-trisubstituted pyrrolidines piperidines,^{14,16} predict that they each bind in different ways within the CCR5 pocket. Hence it is difficult to obtain a clear picture of how these diverse ligands function.

Here we investigate the multisite binding hypothesis using a new consensus shape matching technique based on spherical harmonic (SH) representations of surface shapes²² to perform rapid and exhaustive comparison and clustering of multiple combinations of ligands in multiple trial superpositions. This novel SH-based shape-matching approach uses one or more “pseudomolecules”, obtained from the consensus shapes of the most active molecules, as VS queries against a database of known actives and decoys. The algorithm has been implemented in the ParaFit module of the ParaSurf suite of programs.²³ The new consensus shape-matching approach has been developed specifically to help analyze targets with

* To whom correspondence should be addressed. Tel: +44 1224 272282. Fax: +44 1224 273422. E-mail: d.w.ritchie@abdn.ac.uk.

[†] Universitat Ramon Llull.

[‡] University of Aberdeen.

large ligand-binding pockets, although it may also be used in conventional VS studies where there are multiple known actives.

This article presents our new consensus shape matching VS approach and applies it to a large database of CXCR4 and CCR5 active inhibitors and comparable inactive decoys which was compiled previously.¹² Several trial CCR5 and CXCR4 consensus shape queries are constructed by clustering and superposing selected known actives, and the utility of each query is assessed using receiver-operator-characteristic (ROC)²⁴ plots. The area under the curve (AUC) of each ROC plot is used to provide an objective measure of the ability of each consensus shape to recognize known actives with similar shapes. To find the best clusters of binders with which to characterize the different groups of CCR5 antagonists, we conduct systematic experiments in which AUCs are compared for clusters formed in different ways. The best consensus queries thus found are then further analyzed in the context of the receptor pocket using rigid-body soft docking²⁵ onto the homology modeled CCR5 receptor.¹² Our virtual screening results show that the CCR5 inhibitors may be clustered into four superconsensus (SC) families, and our docking results show that these may be docked to three overlapping regions within the CCR5 extracellular pocket. These results provide strong evidence to support the notion that there are three main binding sites within the CCR5 extracellular pocket.

METHODS

Shape Representations. We use the ParaSurf and ParaFit modules of CEPOS InSilico Ltd.²³ to calculate and superpose molecular surfaces. ParaSurf calculates molecular shape and electronic properties from semiempirical quantum mechanics theory, and encodes these properties as SH expansions.²⁶ Surface shapes are represented as radial distance expansions of the molecular surface, $r(\theta, \varphi)$, with respect to a selected harmonic coordinate origin (CoH), which is normally set equal to the molecular center of gravity (CoG).²⁷ For example, the radial surface shape of molecule A is represented as

$$r_A(\theta, \varphi) = \sum_{l=0}^L \sum_{m=-l}^l a_{lm} y_{lm}(\theta, \varphi) \quad (1)$$

where θ and φ are the spherical coordinates, $y_{lm}(\theta, \varphi)$ are real spherical harmonics, a_{lm} are the expansion coefficients, and L is the order or highest polynomial power of the expansion. Here, $L = 6$ is used in all calculations. ParaFit calculates superpositions between pairs of molecules by exploiting the special rotational properties of the SH functions.²² For example, rotated SH expansion coefficients for molecule B may be calculated as

$$b'_{lm} = \sum_{m'=-l}^l R_{mm'}^{(l)}(\alpha, \beta, \gamma) b_{lm'} \quad (2)$$

where (α, β, γ) are *zyz* Euler rotation angles and $R_{mm'}^{(l)}(\alpha, \beta, \gamma)$ are real Wigner rotation matrix elements.²² To calculate a superposition between a pair of molecules, the CoH of molecule B is translated to that of the fixed reference molecule A, and a rotational search is then performed to find the rotation which minimizes the distance, D_{AB} , between the corresponding pairs of SH expansions

$$D_{AB} = \int (r_A(\theta, \varphi) - r'_B(\theta, \varphi))^2 d\Omega \quad (3)$$

Thanks to the orthogonality of the basis functions, this expression reduces to

$$D_{AB} = \sum_{l=0}^L \sum_{m=-l}^l a_{lm}^2 + b_{lm}^2 - 2a_{lm}b'_{lm} = |a|^2 + |b|^2 - 2ab' \quad (4)$$

Hence the distance function for any orientation may be calculated very rapidly from the original expansion coefficients. In VS, it is convenient to rearrange and normalize the basic distance expression (eq 4) to give a similarity score. Here, we use the Tanimoto score, S_{AB} , calculated as

$$S_{AB} = \frac{ab'}{(|a|^2 + |b|^2 - ab')} \quad (5)$$

ParaSurf can calculate all necessary SH molecular properties in a matter of a few minutes. Once the surface shapes have been calculated, the ParaFit program can perform on the order of hundreds of molecular comparisons per second. Hence the overall approach is well-suited to tasks that require the calculation of multiple molecular comparisons such as high throughput VS.²⁷

SH Consensus Shape Matching. Using the SH representation, a “consensus shape”, $\bar{r}(\theta, \varphi)$, may be constructed as the average of N individual molecular shape expansion coefficient vectors, a_{lm}^k , for $k = 1, \dots, N$ as

$$\bar{r}(\theta, \varphi) = \frac{1}{N} \sum_{k=1}^N \sum_{l=0}^L \sum_{m=-l}^l a_{lm}^k y_{lm}(\theta, \varphi) \quad (6)$$

However, before computing the average, each molecule in the consensus must first be rotated to minimize the distance between it and the remaining $N - 1$ molecules. In practice, because these rotations are not known a priori, the consensus shape is constructed iteratively as follows. First, all-against-all rotational pairwise superpositions are calculated to find the two most similar surface shapes. Then, the average of these two shapes is taken as the initial seed shape for the consensus, and the remaining $N - 2$ SH shapes are rotated into superposition with the seed shape. The overall average of all SH coefficients is then computed to give the first estimate of the consensus shape. The consensus average is then refined by superposing the member molecular shapes back onto the average and by recalculating a new average shape. This procedure is repeated until convergence to optimal overlap is reached between each molecule and the consensus shape. This protocol is similar to techniques used for refining electron microscopy density images of similar molecules observed in different orientations.²⁸ In the present case, convergence is typically achieved in just three or four cycles. Hence calculating a consensus shape is a quick process. Figure 1 illustrates the overall procedure schematically.

ROC Plot Analyses. Here, all VS results are presented as ROC plots of true positive rate versus false positive rate or equivalently Sensitivity versus $1 - \text{Specificity}$. These quantities are calculated as

$$\text{True Positive Rate} = \text{TP}/(\text{TP} + \text{FN}) = \text{Sensitivity} \quad (7)$$

$$\text{False Positive Rate} = \text{FP}/(\text{TN} + \text{FP}) = 1 - \text{TN}/(\text{TN} + \text{FP}) = 1 - \text{Specificity} \quad (8)$$

where TP represents the number of correctly identified actives (true positives), FP represents the number of inactives

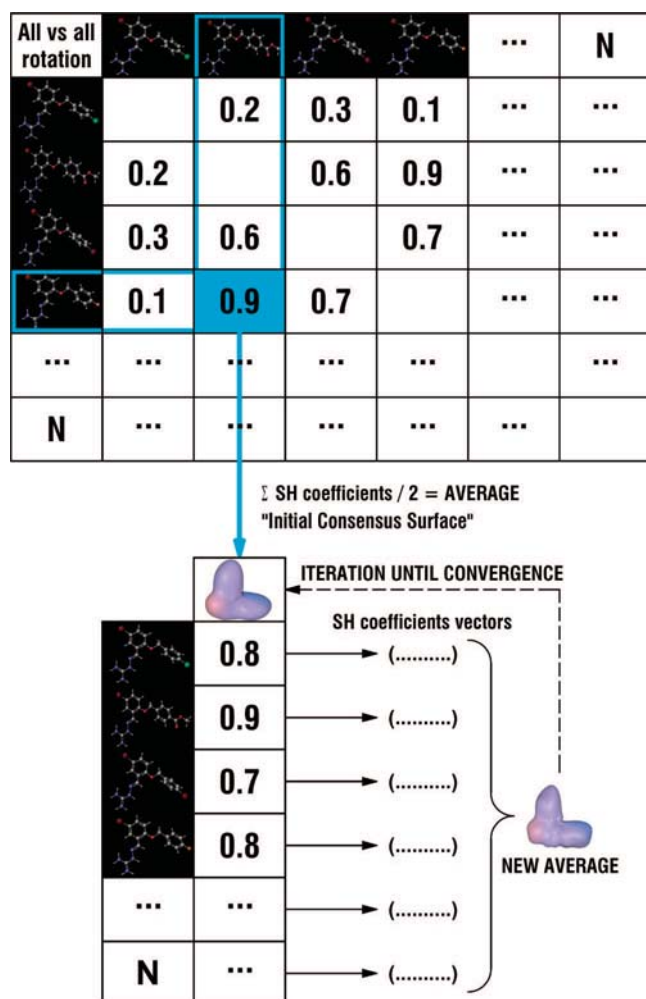


Figure 1. Flow diagram of the consensus shape calculation. First, ParaFit all-against-all rotational superpositions are calculated for the group of *N* molecules that will form the consensus. The two most similar SH shapes are selected and superposed to form a seed consensus shape. Then, all molecules are rotationally superposed onto the seed consensus. A new consensus shape is computed from the average SH coefficients of the superposed shapes. The consensus members are then superposed again onto the consensus average, and the process is iterated until convergence.

incorrectly predicted as active (false positives), TN represents the number of correctly identified inactives (true negatives), and FN represents the number of actives incorrectly predicted to be inactive (false negatives). Each ROC plot is calculated by ranking the database molecules by similarity with the query (or by docking energy with the protein target), and by summing the number of TPs, FPs, TNs, and FNs on either side of each rank position. ROC plots are particularly useful when comparing different VS queries with different numbers of actives and inactives because the AUC of a ROC plot gives an objective measure of query performance which is essentially independent of the actual number of positive and negative instances (i.e., ROC curves do not suffer from “class skew”).²⁹

Consensus Shape-Based Virtual Screening. Obviously, using SH surfaces to compute average shapes will result in some smoothing and loss of detail compared to the original individual molecular shapes. However, this can be considered a desirable property because it provides an unbiased way to combine the most significant features of a related group of molecules. Nonetheless, it is important to select the member shapes carefully to achieve a good balance between capturing

Table 1. Families of CXCR4 and CCR5 Antagonists Used in the Current Study

family	number of compounds	ref
CXCR4 inhibitors		
tetrahydroquinolinamines	123	7, 30–34
KRH derivatives (Kureha Chemical Industries)	23	7, 35–38
macrocycles	4	39
AMD derivatives (AnorMED)	94	7, 39–44
cyclic peptides	2	45
other	2	46
total	248	
CCR5 inhibitors		
SCH derivatives (Schering-Plough)	120	17, 47, 48
diketopiperazines	9	49–53
anilide piperidine <i>N</i> -oxides	22	54
AMD derivatives (AnorMED)	3	44
4-piperidines	10	55, 56
4-aminopiperidine or tropanes	26	55, 57, 58
1,3,4-trisubstituted pyrrolidinepiperidines	9	59
phenylcyclohexilamines	9	60–65
TAK derivatives (Takeda)	66	66, 67
1-phenyl-1,3-propanodiamines	57	68–70
1,3,5-trisubstituted pentacyclics	9	71
<i>N,N'</i> -diphenylureas	4	72
5-oxopyrrolidine-3-carboxamides	5	73
guanyldiazide derivatives	33	74
4-hydroxypiperidine derivatives	36	75
other	6	76
total	424	

the most significant features for binding and smoothing away too much detail. In this work, this balance is achieved, and the overall approach is validated by monitoring the utility of various consensus shapes as VS queries against our database of known CCR5 and CXCR4 binders and decoys.¹² Since this database was first described, some newly published CCR5 inhibitors have been added to the set of actives. Table 1 lists the representative families of CXCR4 and CCR5 inhibitors in the updated database, and Figure 2 shows some representative members of each family. Consensus shape-based VS was applied to these families using query structures constructed from: (a) the consensus shape of the three most active compounds of different scaffolds families in the databases (i.e., an AMD derivative, a macrocycle derivative, and a KRH derivative for CXCR4, and a piperidine derivative, a SCH derivative, and a 1,3,4-trisubstituted pyrrolidine-piperidine derivative for CCR5), and (b) the consensus shape of *all* CXCR4 or CCR5 active inhibitors in the database.

The consensus shape-based approach was also used to investigate the CCR5 multiple binding site hypothesis. First, the CCR5 inhibitors were clustered using Ward's hierarchical clustering method,⁷⁷ as implemented in the JKlustor module of JChem,⁷⁸ using both chemical (topological) fingerprints and two-dimensional pharmacophore fingerprints. The optimal number of clusters to be selected was calculated using Kelley's method,⁷⁹ also implemented in JKlustor, and the consensus shapes of these fingerprint-defined clusters were calculated using ParaFit, as described above. Then, ParaFit was used again to compute all-against-all rotational superpositions of the fingerprint-defined consensus shapes. This produced a shape-based similarity matrix, which was reclustered to identify clusters of similar consensus shapes,⁸⁰ which were again superposed and averaged to compute a small

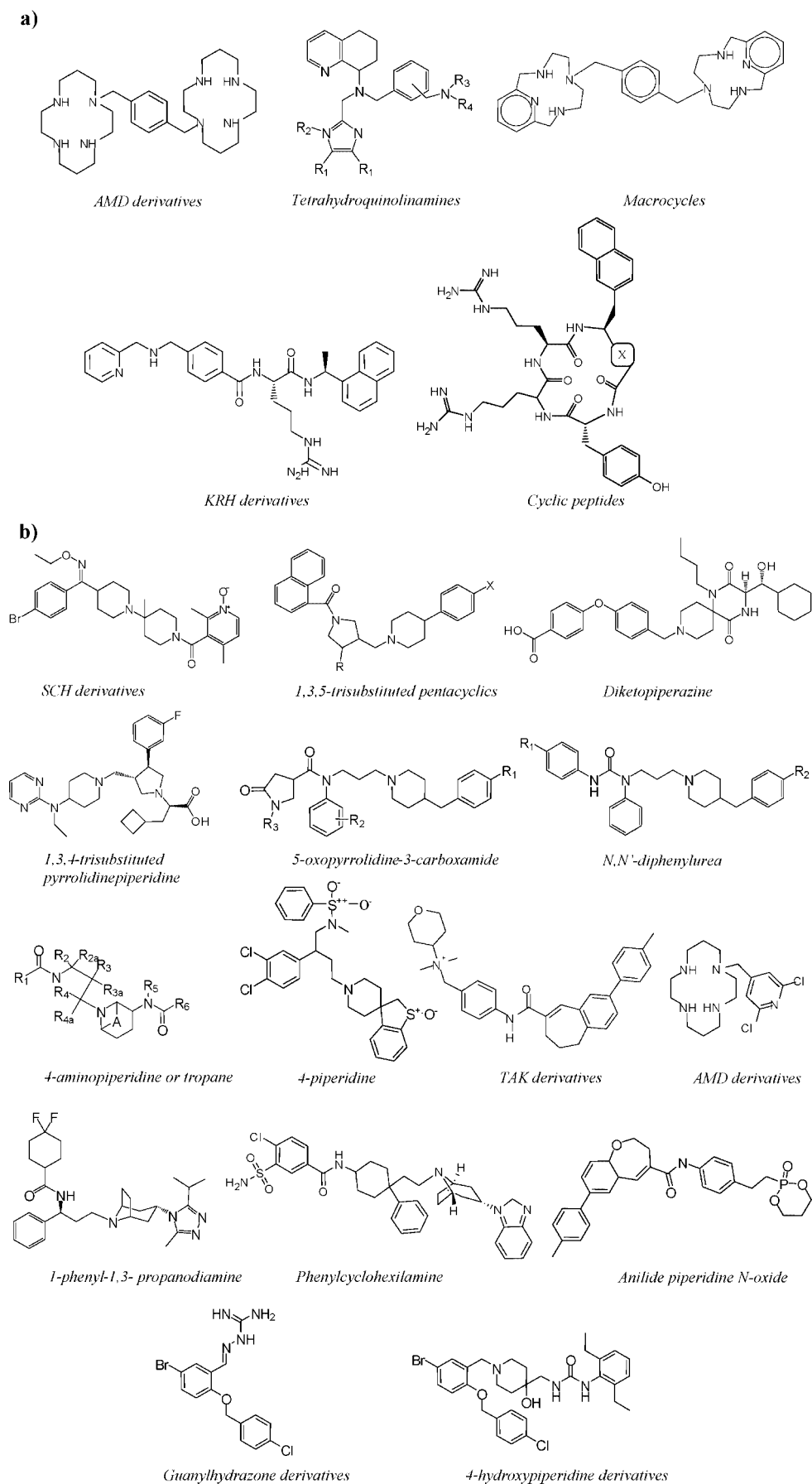


Figure 2. Representative structures of (a) five families of CXCR4 inhibitors and (b) fifteen families of CCR5 inhibitors.

number of SC shapes. In other words, the resulting SC SH shapes correspond to the shapes of pseudomolecules con-

structed from volumetric unions of fingerprint-based and shape-based subclusters of known actives.

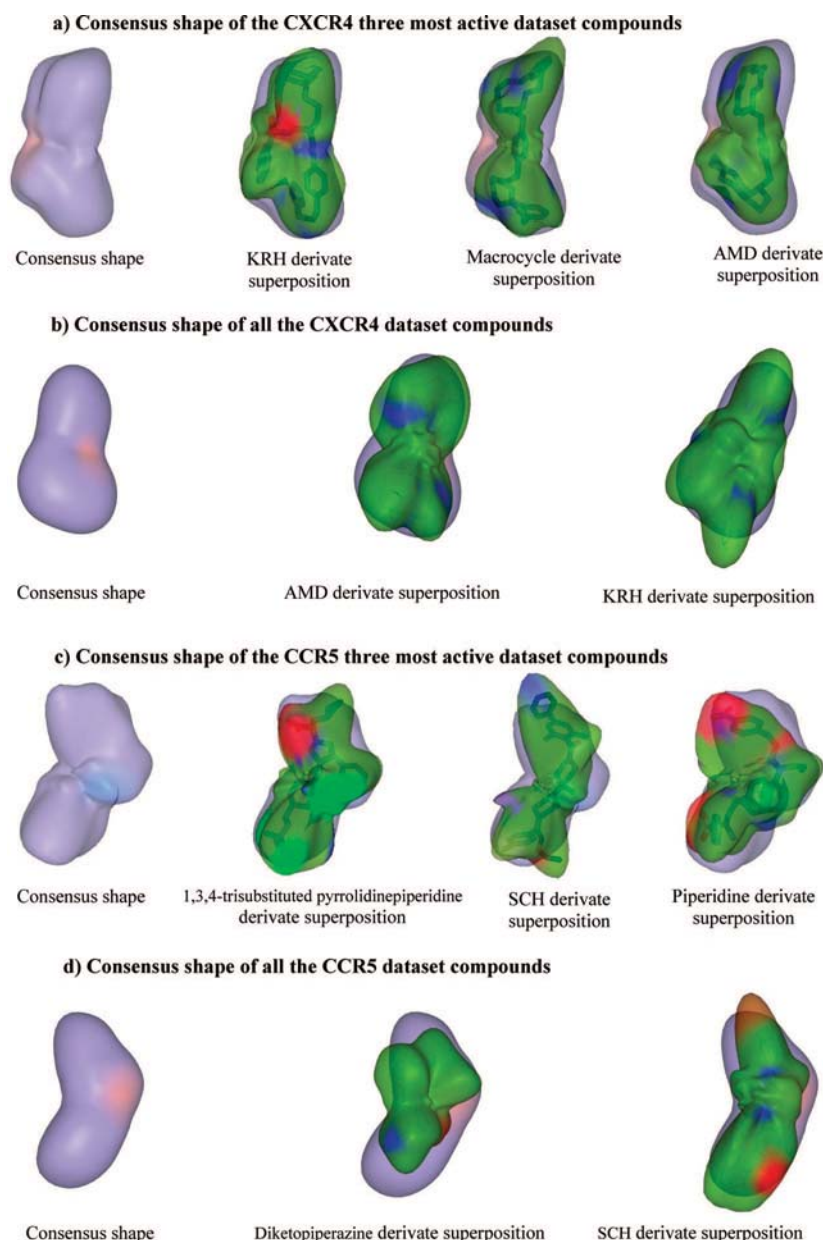


Figure 3. CXCR4 and CCR5 antagonist consensus shapes. (a) The image on the left shows the consensus shape calculated from the three most active compounds of different scaffold families in the CXCR4 inhibitor database: an AMD derivative, a macrocycle derivative, and a KRH derivative. The following three images show the superpositions of these compounds onto the consensus. (b) The consensus shape calculated from all CXCR4 database actives, and example superpositions onto the consensus of two randomly selected compounds (an AMD derivative and a KRH derivative). (c) On the left, the consensus shape calculated from the three most active compounds of different CCR5 database scaffold families: a 1,3,4-trisubstituted pyrrolidinepiperidine derivative, a SCH derivative, and a piperidine derivative. On the right, the superpositions of these compounds onto the consensus. (d) Consensus shape calculated from all CCR5 actives, along with example superpositions onto the consensus of two randomly selected actives (a diketopiperazine derivative and a SCH derivative).

To explore whether the computed SC pseudomolecules are sterically feasible in the context of the CCR5 extracellular pocket, each pseudomolecule was rigidly docked into our model-built CCR5 structure using blind Hex docking with default search parameters.²⁵ This structure was built by homology using bovine rhodopsin as template (PDB code 1HZX: 20% sequence identity and 35% similarity with respect to CCR5), as described previously.¹² To compare quantitatively the ability of the consensus shapes to identify known binders, VS was performed using our ligand database and the screening utility of each query shape was analyzed objectively using ROC analyses. Finally, VS results for CXCR4 and CCR5 consensus shape queries were compared to conventional ROCS 2.2,⁸¹ Hex 4.8,²⁵ and ParaFit 08²²

shape-matching VS, and to rigid-docking-based VS using Hex 4.8, AutoDock 3.0,⁸² GOLD 3.01,⁸³ and FRED 2.2.⁸⁴

RESULTS

CXCR4 and CCR5 Inhibitor Consensus Shapes. Figure 3a shows the consensus shape calculated from the three most active compounds of different scaffold families in the CXCR4 inhibitor database (an AMD derivative, a macrocycle derivative, and a KRH derivative). Figure 3b shows the consensus shape computed from all the CXCR4 inhibitors in our database. Visual inspection of these figures shows that the first consensus shape captures rather well the overall shape of the three selected inhibitors, whereas the all-

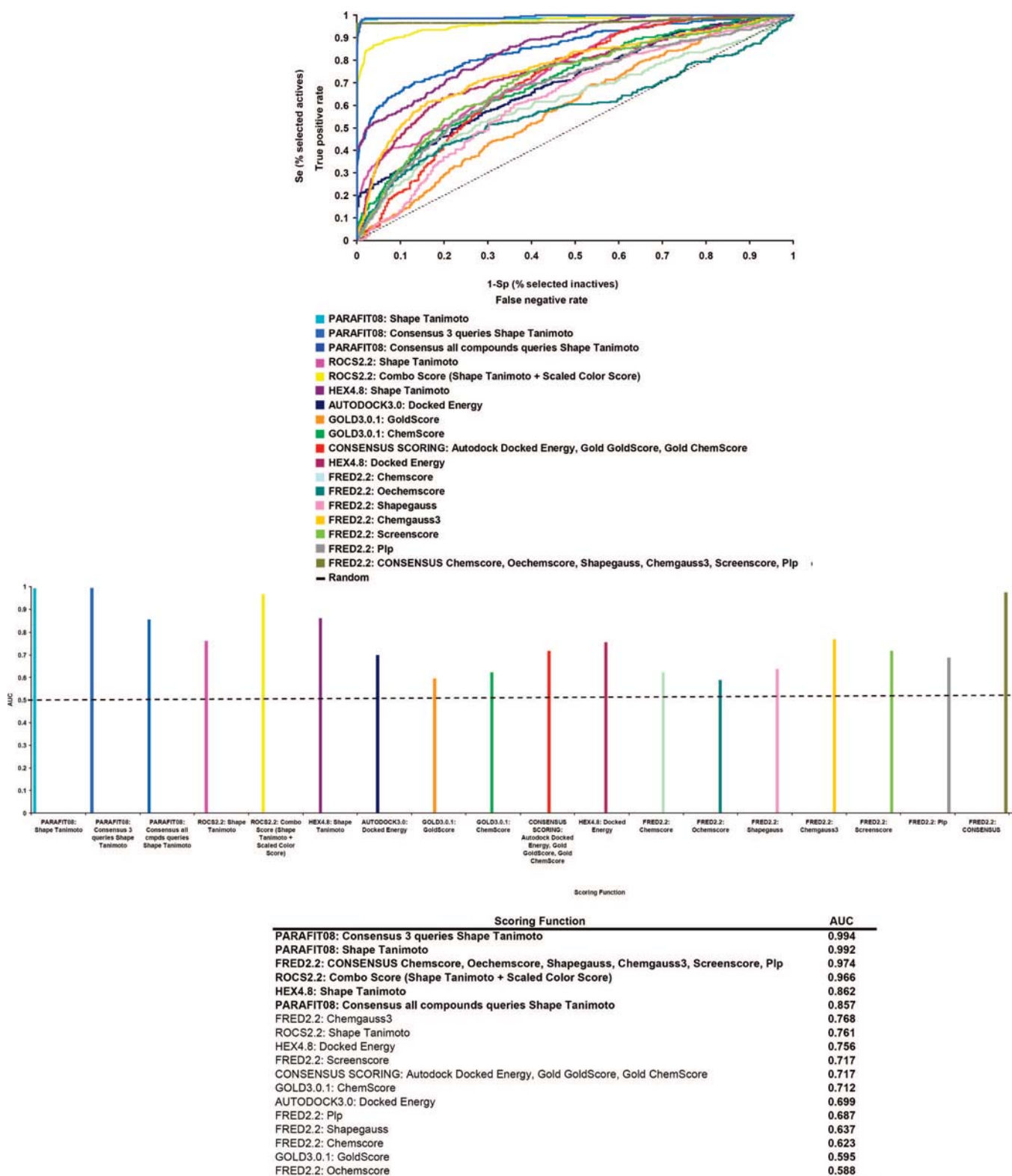





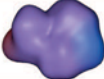






Figure 4. ROC plot validation of various shape-matching and docking VS methods compared to the consensus shape approach applied to CXCR4 antagonists. The dotted black line represents the expected enrichment if actives were selected at random. The lower bar chart and table report the AUC values obtained from the corresponding ROC curves. The scoring functions which give the best VS performance are shown in bold.

molecule consensus has much less local surface detail, yet still broadly retains the gross features of the member shapes. Figure 3c shows the consensus shape calculated for the three most active compounds of different scaffolds families in the CCR5 inhibitor database (a piperidine derivative, a SCH derivative, and a 1,3,4-trisubstituted pyrrolidonepiperidine derivative). Figure 3d shows the consensus shape of *all* the

CCR5 active inhibitors. In this case, it can be seen that using all database compounds to construct the consensus query causes a more spherical average shape than the CXCR4 inhibitors, because of the greater number and diversity of compounds in the CCR5 database.

CXCR4 Virtual Screening. Figure 4 shows the performance of the CXCR4 consensus shaped-based VS queries

Table 2. CCR5 Antagonist Clustering Results Using Ward's Clustering of Chemical Fingerprint Descriptors^a

CLUSTER	Compounds Found	Number of compounds	Consensus Shape
1	(8) <i>1,3,4-trisubstituted pyrrolidinepiperidines</i> (3) <i>1,3,5-trisubstituted pentacyclics</i> (5) <i>5-oxopyrrolidine-3-carboxamides</i> (4) <i>N,N'-diphenylureas</i> (2) TAK derivatives (1) 4-piperidines (1) others (MRK-1 CMPD 167)	24	
2	(1) <i>1,3,4-trisubstituted pyrrolidinepiperidines</i> (6) <i>1,3,5-trisubstituted pentacyclics</i> (13) 1-phenyl-1,3- propanodiamines (3) 4-piperidines (3) <i>AMD derivatives</i> (9) <i>Diketopiperazines</i> (1) SCH derivatives (2) Phenylcyclohexilamines (3) others (GSK, Merck2, Merck3)	41	
3	(22) <i>Anilide piperidine N-oxides</i> (1) TAK derivatives (1) others (1-benzazepine)	24	
4	(21) 1-phenyl-1,3- propanodiamines (5) Phenylcyclohexilamines	26	
5	(11) 1-phenyl-1,3- propanodiamines	11	
6	(12) 1-phenyl-1,3- propanodiamines	12	
7	(26) <i>4-aminopiperidine or tropanes</i> (6) 4-piperidines (2) Phenylcyclohexilamines (1) others (Merck1)	35	
8	(23) SCH derivatives	23	
9	(20) SCH derivatives	20	
10	(37) SCH derivatives	37	
11	(22) SCH derivatives	22	
12	(17) SCH derivatives	17	
13	(19) TAK derivatives	19	
14	(44) TAK derivatives	44	
15	(33) <i>Guanythydrazone derivatives</i>	33	
16	(36) <i>4-hydroxypiperidine derivatives</i>	36	

^a Kelley's method predicts 16 clusters as the optimal number. The number of compounds found for each family in each cluster is specified in parenthesis. The families marked in bold italics comprise the entire family in a unique cluster. The families marked in italics comprise the entire family between two clusters. The ten initial consensus shapes obtained after the grouping of clusters are also shown.

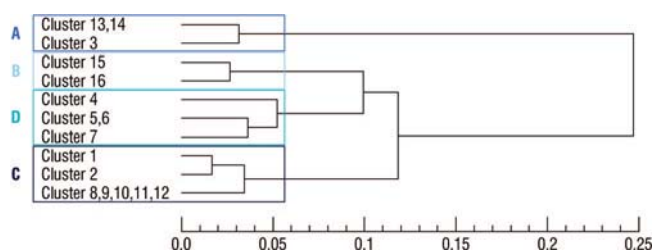
compared to docking-based and shape-based screening using a single high affinity ligand (AMD3100). This figure shows that the consensus shape queries give higher AUCs than the other approaches, although the single-ligand ParaFit query also performs well. As might be expected from consideration of Figure 1, the three-ligand consensus performs considerably better than the all-ligand consensus, due to the high degree of smoothing and loss of surface detail in the all-ligand shape. On the other hand, considering the very good performance of the single high affinity ligand and the marginally superior performance of the three-ligand query suggests that all three ligands share highly similar shapes (as confirmed by Figure 3a), which probably all bind in similar way within the CXCR4 pocket.

Regarding the shape matching approaches, ParaFit Shape Tanimoto, ROCS Combo Score, and Hex Shape Tanimoto all give comparable AUCs to the ParaFit consensus shape query. Of the docking tools, FRED Consensus gives the best AUC, followed by FRED Chemgauss3, Hex Docked Energy, and rank-by-rank docking Consensus Scoring, which gives a better enrichment than the individual Autodock Docked Energy, Gold GoldScore, and Gold ChemScore scoring functions. The ROCS Combo Score and FRED Chemgauss3 scoring functions both include descriptions of shape and molecular chemical properties. If the protein structures contain errors, as is likely with model-built structures, those docking functions that include terms that favor chemical complementarity might be expected to be more resilient to

Table 3. CCR5 Antagonist Clustering Results Using Ward's Clustering of 2D Pharmacophore Fingerprint Descriptors^a

cluster	compounds found	number of compounds
1	(3) <i>1,3,4-trisubstituted pyrrolidinedi-piperidines</i> (7) <i>1,3,5-trisubstituted pentacyclics</i> (38) 1-phenyl-1,3- propanodiamines	48
2	(6) <i>1,3,4-trisubstituted pyrrolidinedi-piperidines</i> (2) <i>1,3,5-trisubstituted pentacyclics</i> (14) 1-phenyl-1,3- propanodiamines (1) 4-aminopiperidine or tropanes (1) others (1-benzazepine)	24
3	(4) 1-phenyl-1,3-propanodiamines (8) 4-piperidines (5) 5-oxopyrrolidine-3-carboxamides (3) <i>Diketopiperazines</i> (5) anilide piperidine <i>N</i> -oxides (3) phenylcyclohexilamines (2) <i>N,N'</i> -diphenylureas (2) SCH derivatives (3) others (MRK-1 CMPD 167, Merck1, Merck2)	35
4	(17) 4-aminopiperidine or tropanes (1) phenylcyclohexilamines (11) SCH derivatives (1) others (Merck3)	30
5	(8) 4-aminopiperidine or tropanes (6) <i>Diketopiperazines</i> (2) anilide piperidine <i>N</i> -oxides	16
6	(15) anilide piperidine <i>N</i> -oxides (5) phenylcyclohexilamines	20
7	(18) SCH derivatives (2) <i>N,N'</i> -diphenylureas (1) 4-piperidines (1) others (GSK 108)	22
8	(59) SCH derivatives (2) TAK derivatives	61
9	(28) SCH derivatives	28
10	(2) SCH derivatives (22) TAK derivatives	24
11	(42) TAK derivatives	42
12	(1) <i>AMD derivatives</i> (1) 1-phenyl-1,3- propanodiamines (1) 4-piperidines (4) <i>guanyldihydrazone derivatives</i> (36) 4-hydroxypiperidine derivatives	43
13	(2) <i>AMD derivatives</i> (29) <i>guanyldihydrazone derivatives</i>	31

^a Kelley's method predicts 13 clusters as the optimal number. The number of compounds found for each family in each cluster is specified in parenthesis. The families marked in bold italics comprise the entire family in a unique cluster. The families marked in italics comprise the entire family between two clusters.

**Figure 5.** Dendrogram of the ten initial CCR5 antagonist groups clustered using Ward's clustering of spherical harmonic distances between the consensus surface shapes of each group. Four main SC groups, labeled A, B, C, and D, are recognized.

structural errors in the receptor. Therefore, it is perhaps not surprising that the FRED Chemgauss3 gives an AUC which is closer to that of the ligand-based scoring functions and

much better than the other docking scoring functions such as FRED Shapegauss, Chemscore, Oechemscore, Shapegauss, Chemgauss3, Screenscore, and Plp.

Overall, the enrichment results for CXCR4 show that ligand-based shape-matching approaches provide better VS performance than structure-based docking tools, except for FRED Consensus Scoring, which gives considerably better enrichments than the other FRED scoring functions. This indicates that inaccuracies probably exist in the homology-built structure of the receptor and that, consequently, the use of ligand-based techniques should provide a more reliable way to identify new inhibitors for this target. It is also worth mentioning that in all docking studies the protein structure was assumed to be rigid, which in reality is not true. The quality of any docking calculation will intrinsically depend on the conformation of the receptor and especially the conformations of the side chains lining the binding region. Hence keeping the protein rigid is potentially a large source of uncertainty that can further influence the performance of docking-based VS calculations.

Clustering Known CCR5 Inhibitor Families. Table 2 shows the result of clustering the CCR5 inhibitors using Ward's clustering of chemical (topological) fingerprints. In this case, Kelley's method gives the optimal number of clusters as 16. Table 3 gives the corresponding results for clustering using 2D pharmacophoric fingerprints, which gives just 13 more tightly grouped clusters according to Kelley's method. Inspection of these clusters shows that the pharmacophoric fingerprint clustering tends to distribute compounds from different chemical families into more different clusters, whereas clustering on chemical fingerprints tends to group the compounds more closely according to the known inhibitor families. For example, all members of two entire families are assigned to a single cluster in two cases (i.e., the 5-oxopyrrolidine-3-carboxamides and *N,N'*-diphenylureas are entirely assigned to cluster 1, and the AMD derivatives and diketopiperazines are entirely assigned to cluster 2), and the members of several other families are entirely assigned to separate clusters (i.e., the members of the anilide piperidine *N*-oxides, 4-aminopiperidine or tropanes, guanyldihydrazone derivatives and 4-hydroxypiperidine derivatives are all assigned to separate clusters). Furthermore, chemical fingerprint clustering nicely separates the 1-phenyl-1,3-propanodiamines and the SCH and TAK families into different clusters depending on their different R-groups. Hence chemical fingerprint-based clustering was selected as the most appropriate point from which to proceed. Further inspection of these clusters shows that by grouping clusters 5 and 6 and similarly grouping clusters 8, 9, 10, 11, and 12, and also clusters 13 and 14, a total of just ten clusters are obtained which correctly groups together all the compounds belonging to a given scaffold family. Hence a total of ten CCR5 inhibitor clusters were selected for further analysis using the consensus shape-based approach.

Calculating Consensus and Super-Consensus CCR5 Inhibitor Clusters. SH consensus surface shapes were calculated for each of the ten selected clusters, as described in Methods (eq 6). An all-against-all SH comparison of each consensus surface was calculated using ParaFit, and the resulting pairwise Tanimoto similarity coefficients were used to calculate consensus superclusters using a further round of Ward's hierarchical clustering. Figure 5 shows a dendro-

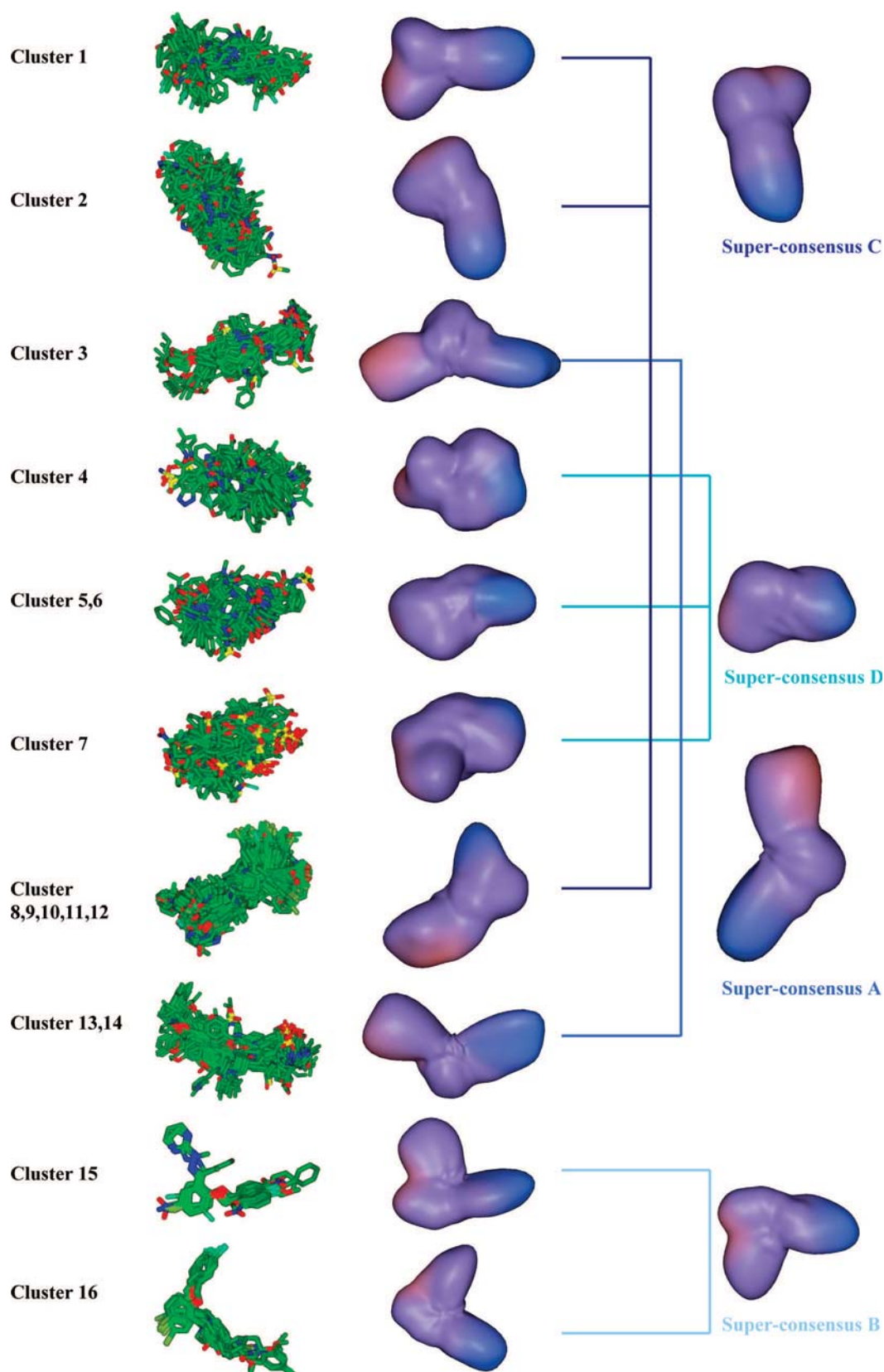
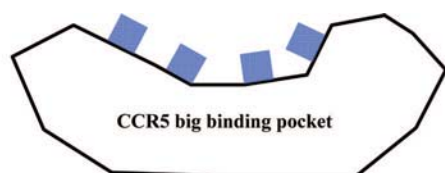


Figure 6. Molecular superpositions and consensus shapes of the ten Ward's clusters used to calculate the final SH SC shapes.

gram of the resulting SCs in which the initial ten consensus shape surfaces are clustered to give four main representative SC groups, A, B, C, and D. Figure 6 shows the 3D molecular overlays, the SH shapes of the 10 fingerprint clusters, and the SC shapes calculated from the clustered consensus surfaces. All molecular orientations shown in this figure were

derived directly from the SH consensus surface shape superposition calculations. If it is supposed that the calculated superclusters correspond to four fundamental families of inhibitors, it might further be hypothesized that these fundamental families bind within different regions of the CCR5 pocket. Figure 7 shows a schematic illustration of this



- SC_A (87 compounds):** TAK derivatives
Anilide piperidine *N*-oxides
- SC_B (69 compounds):** Guanylhydrazone derivatives
4-hydroxypiperidine derivatives
- SC_C (184 compounds):** SCH derivatives
1,3,4-trisubstituted pyrrolidinedipiperidines
1,3,5-trisubstituted pentacyclics
5-oxopyrrolidine-3-carboxamides
N,N'-diphenylureas
Diketopiperazines
AMD derivatives
1-phenyl-1,3-propanodiamines
4-piperidines
- SC_D (84 compounds):** 1-phenyl-1,3-propanodiamines
Phenylcyclohexilamines
4-aminopiperidine or tropanes
4-piperidines

Figure 7. Schematic illustration of the hypothesized binding regions suggested by SC clustering. For each SC, the number of compounds used to construct the consensus and the family to which they belong are given.

hypothesis, along with the calculated scaffold family membership of each fundamental supercluster.

CCR5 Scaffold Family Virtual Screening. SH consensus surface shapes were also calculated for each of the 15 CCR5 scaffold families in the database. Results for the comparison of the VS performance of the SC scaffold queries are shown in Figure 8. It can be seen that the consensus query for each family performs quite well individually, except those for 4-hydroxypiperidine and guanylhydrazone derivatives. These two SC B family queries give poor AUCs because they have somewhat different molecular and consensus shapes than those of the other inhibitor families (see Figure 6).

As can be seen from the lower table in Figure 8, the scaffold family consensus AUCs correspond very well to the proposed SC clusters. More specifically, it can be seen that ordering the individual family consensus groups by AUC is almost sufficient to map each family member directly to its proposed SC. For example, the large group of high AUC scaffolds maps to SC C (namely, the SCH derivatives, 5-oxopyrrolidine-3-carboxamides, diketopiperazines, 1,3,4-trisubstituted pyrrolidinedipiperidines, *N,N'*-diphenylureas, 1,3,5 trisubstituted pentacyclics, 4-piperidines, and AMD derivatives). The next group of good AUC scaffolds maps to SC A (i.e., the anilide piperidine *N*-oxides and the TAK derivatives). Similarly, the final low AUC scaffold families map to SC groups D (the 4-aminopiperidine and 1-phenyl-1,3-propanodiamines derivatives) and B (the 4-hydroxypiperidine and guanylhydrazone derivatives). The only exception to this AUC-based mapping is the phenylcyclohexilamine family which has a very high AUC, but the clustering of which places it in SC group D.

CCR5 Inhibitor Consensus Shape Virtual Screening. Figure 9 shows the VS results obtained using the four SC shapes as queries. It can be seen that SC C gives the best overall VS performance with an AUC of 0.91. It is perhaps not surprising that this SC query performs very well because it includes the three most active compounds in the database and also a large number of other actives (i.e., 184/424) with

similar shapes to the 4-piperidine derivatives, SCH derivatives, and 1,3,4-trisubstituted pyrrolidinedipiperidine derivatives. The SC A query (87/424 actives) also performs rather well with an AUC of 0.79, and the SC D query (84/424 actives) performs reasonably well (AUC = 0.63). However, the ROC plot for SC B shows that this query exhibits good sensitivity and selectivity in the first percentages of the database screened, but the overall AUC is low (0.41) because the database contains relatively few members of the two SC B families (i.e., a total of only 69/424). However, if the members of clusters B and D are grouped together to form a single SC, as might be suggested by the dendrogram in Figure 5, the screening performance becomes essentially random (AUC = 0.51). Thus, despite the small populations of these two groups, their members have significantly different overall shapes, and they should be classified as two distinct structural groups for VS purposes. Performing a similar exercise with other combinations of SC clusters shows similar but less dramatic reductions in AUCs compared to the AUCs of the unmerged clusters. For example, merging A and B gives AUC = 0.65, merging A and D gives AUC = 0.65, and merging A and C gives AUC = 0.87 (compared to the original AUCs of A = 0.79, B = 0.41, C = 0.91, and D = 0.63). This behavior supports the chemical fingerprint clustering results which suggest that the CCR5 inhibitor families may be clustered into no fewer than four main groups. A similar SC cluster analysis was performed for the pharmacophore fingerprint clusters (details not shown), and this also indicated no less than four main SC families with AUCs of 0.79, 0.43, 0.94, and 0.74 for SC clusters A, B, C, and D (the cluster members of B, C, and D differed slightly from the chemical fingerprint analysis). Hence both clustering approaches ultimately indicate that the CCR5 antagonists may be grouped into four main SC families.

Although it is impractical to generate and test large numbers of different potential SC clusters, we wished to ensure that the VS performance of the selected four clusters was not being dominated by a small number of high affinity actives. Hence the AUCs for SC clusters A, B, C, and D were recalculated with the three most active CCR5 inhibitors removed from each cluster. For SC A, the recalculated AUC was unchanged (0.79). For SC C, the AUC was reduced only marginally from 0.91 to 0.90. For SCs B and D, the AUC was reduced from 0.41 to 0.38 and from 0.63 to 0.61, respectively. The different behavior of the A and C clusters compared to the B and D clusters seems to be because the remaining compounds in the B and D clusters have lower activities than those that were removed, whereas the reduced A and C clusters still contain several other molecules with high activities. Nonetheless, because the observed reduction in AUC is either small or modest in all cases, it may be concluded that the original four SC clusters seem to capture very well the general features of many high affinity binders.

Figure 10 shows the ROC plot analysis of the consensus shape-matching VS approach applied to the CCR5 inhibitor database. The consensus shape constructed from the three most active compounds gives the best VS performance (AUC = 0.99), followed by SC C, which comprises the families containing the greatest number of active compounds. As with the CXCR4 inhibitors, the consensus query constructed with the three most active compounds achieves higher perfor-

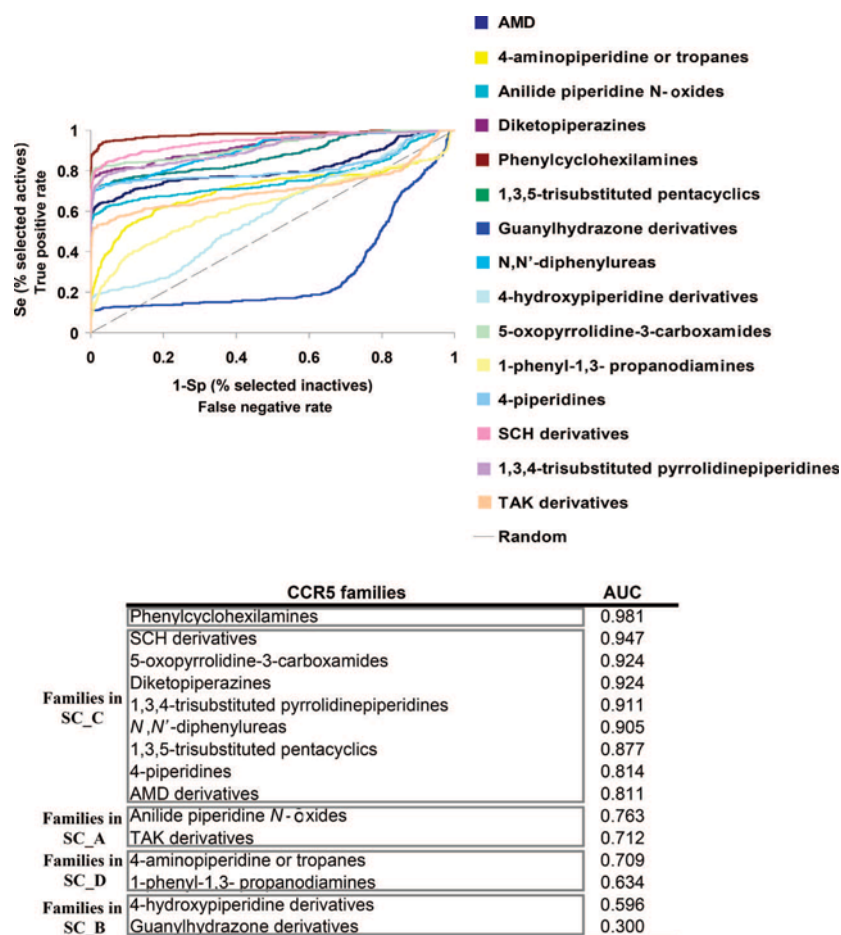


Figure 8. ROC plot evaluation of CCR5 scaffold family consensus shape-matching VS. The dotted black line represents the expected enrichment if actives were selected at random. The lower table reports the AUC values obtained for the consensus ROC curves for the different scaffold families, grouped according to their assigned SC clusters. This shows that the consensus families that give the highest AUC values belong to SC C, followed by those belonging to SC A, D, and B.

mance than the query constructed using all CCR5 inhibitors (AUC=0.87). This supports the notion that using too many molecules to make a consensus shape causes an undesirable degree of surface shape smoothing and the loss of important surface details. Figure 11 shows the ROC plots for all shape-matching VS approaches for the CCR5 inhibitor database. It can be seen that the consensus shape queries generally give larger AUCs than ROCS, Hex, and the single-query ParaFit queries.

Comparing Shape Based and Docking Based CCR5 Virtual Screening. Figure 12 shows the VS performance of selected CCR5 consensus shape queries and several conventional ligand based shape matching and receptor based docking approaches. It can be seen that the three-ligand consensus and SC C queries give the best overall screening performance, along with the FRED consensus scoring and Autodock docking methods, both of which also perform very well even though they do not take into account protein flexibility. As observed previously, the SC A query also performs well, but the SC B and D queries and combinations thereof give rather poor VS results. However, it is interesting to note that the Gaussian-based docking scoring functions in FRED generally give better VS performance than the Gaussian-based superposition scoring functions of ROCS. Comparing Figures 12 and 4 shows that the docking and single-query shape-matching results are generally better for CXCR4 than for CCR5. However, the AUCs for the CCR5

consensus-based query results are almost at the same high level as for CXCR4. Hence, despite CCR5 having a larger binding pocket and a much more diverse set of inhibitors than CXCR4, the use of consensus-based queries can be seen to find many CCR5 inhibitors remarkably well.

Blind Docking Super-Consensus Pseudomolecules. Figure 13 shows the results obtained for blind docking the SC pseudomolecules into the CCR5 extracellular pocket. It can be seen that the SC A pseudomolecule is docked onto one side of the CCR5 binding pocket (Site 1) near residues Ala29, Arg31, Leu33, Tyr37, Thr82, Trp86, Tyr108, and Glu283, delimited by transmembrane (TM) loops 1, 2, 3, and 7, whereas the SC C pseudomolecule is docked onto the opposite side of the pocket (Site 2) near residues Tyr108, Phe113, Ile198, Ile200, Asn252, Glu283, and Glu286, delimited by TM loops 3, 5, and 6. The SC B and SC D pseudomolecules are docked onto the central region of the binding pocket (Site 3) near residues Tyr108, and Glu283, delimited by TMs 3, 6, and 7, thus overlapping the predicted SC A and C binding sites. Figure 14 shows more detailed views of the calculated docking modes. Docking the four SCs derived from pharmacophore fingerprint clustering also gives similar binding modes (details not shown). Thus our docking calculations consistently suggest the existence of two or three main binding sites within the CCR5 pocket. These consensus-based docking predictions are consistent with experimental data.^{11,13,14,16}

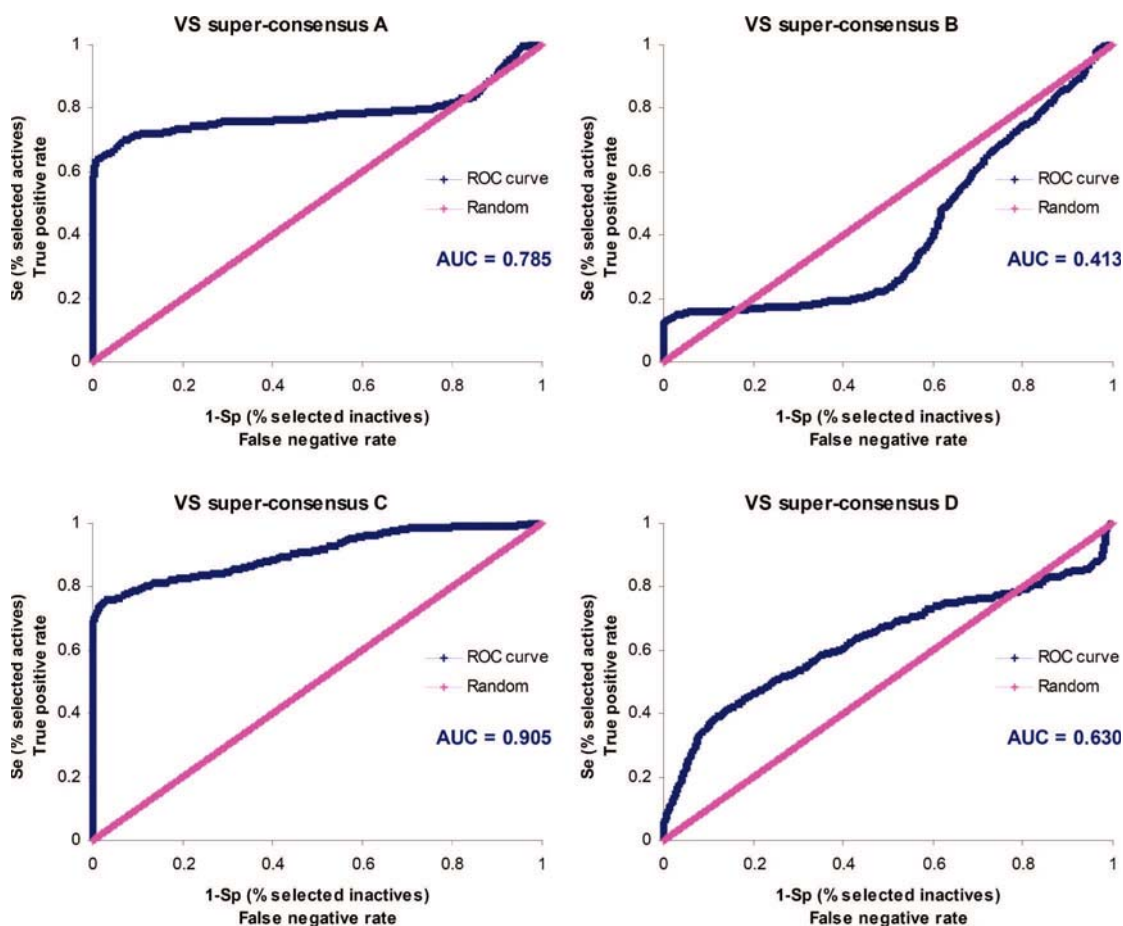


Figure 9. ROC plot validation of the CCR5 inhibitor SC pseudomolecules. The AUC is given for each SC, A, B, C, and D. The diagonal pink line represents the expected enrichment if actives were selected at random.

To confirm that the SC queries are properly matched with their predicted target sites, the three proposed binding sites were each treated as if they were separate targets for docking-based VS using rigid body docking of the corresponding SC pseudomolecules. In other words, when docking to Site 1, compounds belonging to SC A were treated as actives, and compounds belonging to SC B, C, and D were treated as inactives. In a similar manner, when docking to Site 2, compounds belonging to SC C were treated as actives, and compounds belonging to SC A, B, and D were treated as inactives. Similarly for Site 3, compounds belonging to SC B and D were treated as actives, and compounds belonging to SC A and C were treated as inactives. Figure 15 shows the docking VS performance for each of the three proposed CCR5 binding regions. Comparing Figures 15 and 9, it can be observed that docking VS onto Sites 1, 2, and 3 (AUC = 0.83, 0.96, and 0.85, respectively) improves the SC A, C and B/D shape matching AUCs (AUC = 0.79, 0.91, and 0.41/0.63, respectively). Given that SCs A and C already give good shape matching enrichments, reassigning the B and D members as inactives only marginally improves the corresponding AUCs. However, treating the large set of C and A members as inactives for Site 3 gives much higher AUCs for the SC B and D queries, which clearly supports the notion that the CCR5 antagonists bind to at least three main sites within the extracellular pocket.

DISCUSSION

The results of this study show that spherical harmonic consensus shapes can provide effective 3D query structures

for shape-based VS. For the CXCR4 and CCR5 ligands studied here, our results show that well-chosen consensus shape queries can give better (CXCR4) or significantly better (CCR5) virtual screening enrichments than conventional single-molecule VS queries. The CXCR4 results show that consensus shape based queries give higher AUCs (i.e., better enrichments) than conventional ligand-based and rigid-body receptor-based screening approaches. However, for CXCR4, these results are nonetheless broadly similar to the basic ParaFit one-molecule shape-matching approach because the inhibitors for this target share rather similar molecular shapes which individually match quite well the selected query shape. For CCR5, which has a much larger and more diverse set of inhibitor families, the SC family C and the SC all-family queries both give very good overall VS performance. However, this seems to be at least partly because a high proportion of all scaffold families cluster into the family C superconsensus grouping. Hence, by construction, the spherical harmonic consensus shape derived from these family members provides a single representative pseudomolecular shape which recognizes well many of the individual member structures.

Regarding the more challenging problem of understanding how so many diverse inhibitor families might bind within the CCR5 pocket, our consensus shape-based approach provides a straightforward way to identify clusters of inhibitor families from a large set of known actives which is broadly consistent with current experimental SDM data.^{11,13,14,16} More specifically, our clustering results indi-

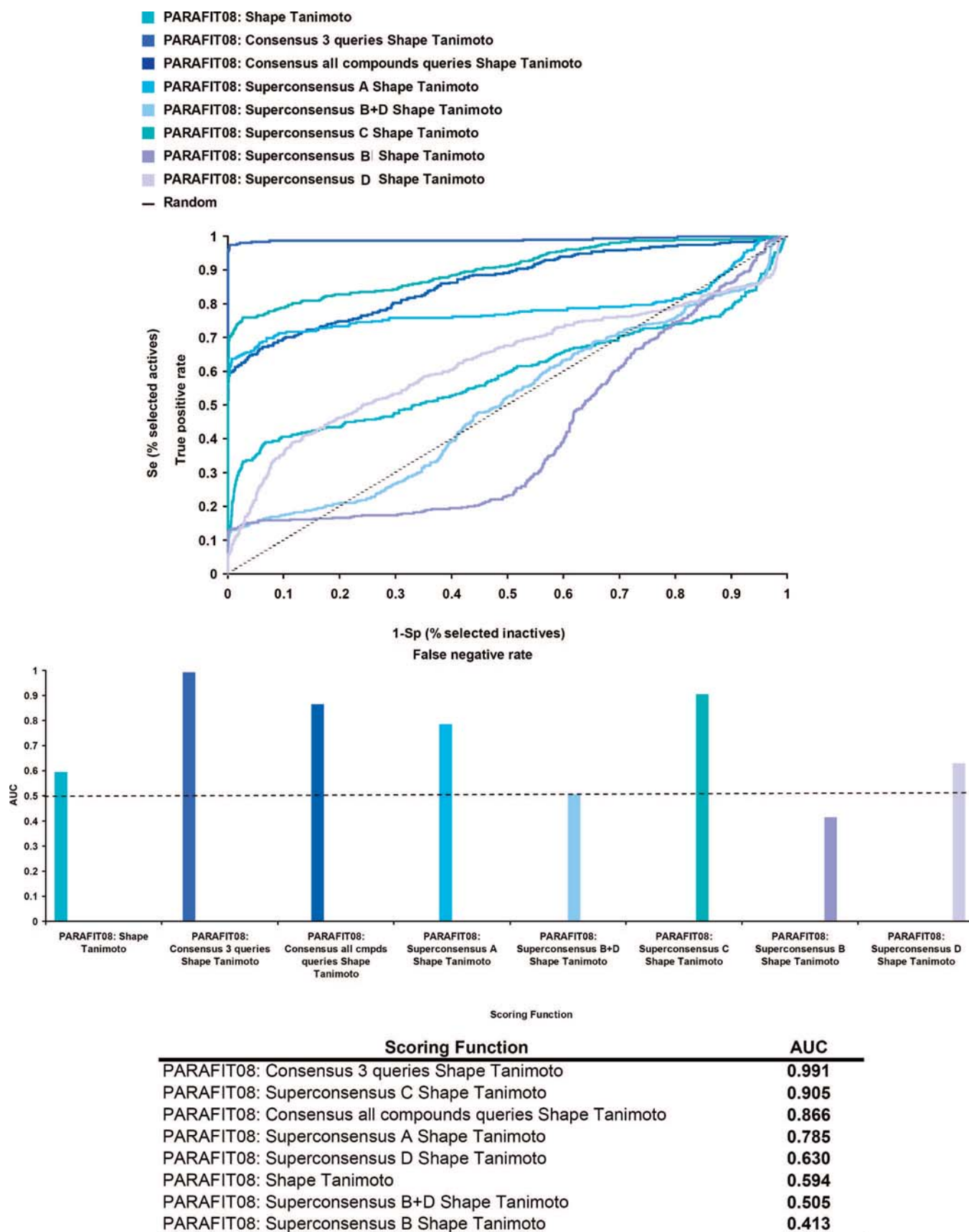


Figure 10. ROC plot evaluation of consensus shape-matching VS for the CCR5 antagonists. The dotted black line represents the expected enrichment if actives were selected at random. The lower bar chart and table report the AUC values of the corresponding VS ROC curves.

cate that the CCR5 inhibitors are in fact described very well by four main consensus families, of which SC family C is the most highly populated. Our docking results suggest that

the families of compounds belonging to SC A bind within Site 1, and this is consistent with SDM-based experimental results for TAK derivative binding.^{11,13,18–21,85} Furthermore,

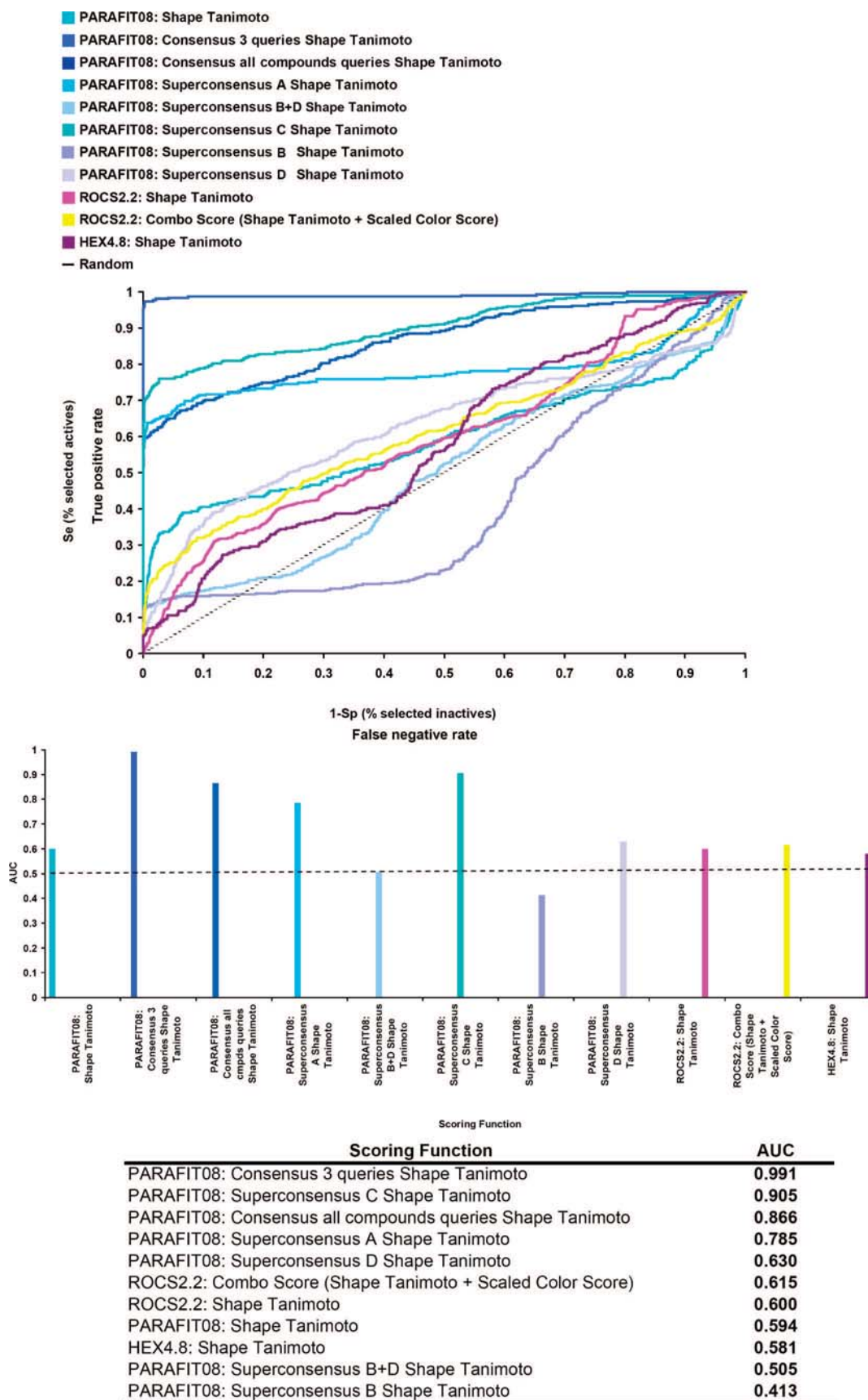


Figure 11. ROC plot comparison of various shape-matching VS methods for the CCR5 antagonists. The dotted black line represents the expected enrichment if actives were selected at random. The lower bar chart and table report the AUC values of the corresponding VS ROC curves.

our docking results suggest that SC C ligands bind within Site 2, and this is consistent with published experimental

results for CCR5 binding of certain SCH derivatives, 1,3,4-trisubstituted pyrrolidinepiperidine derivatives, and diketopi-

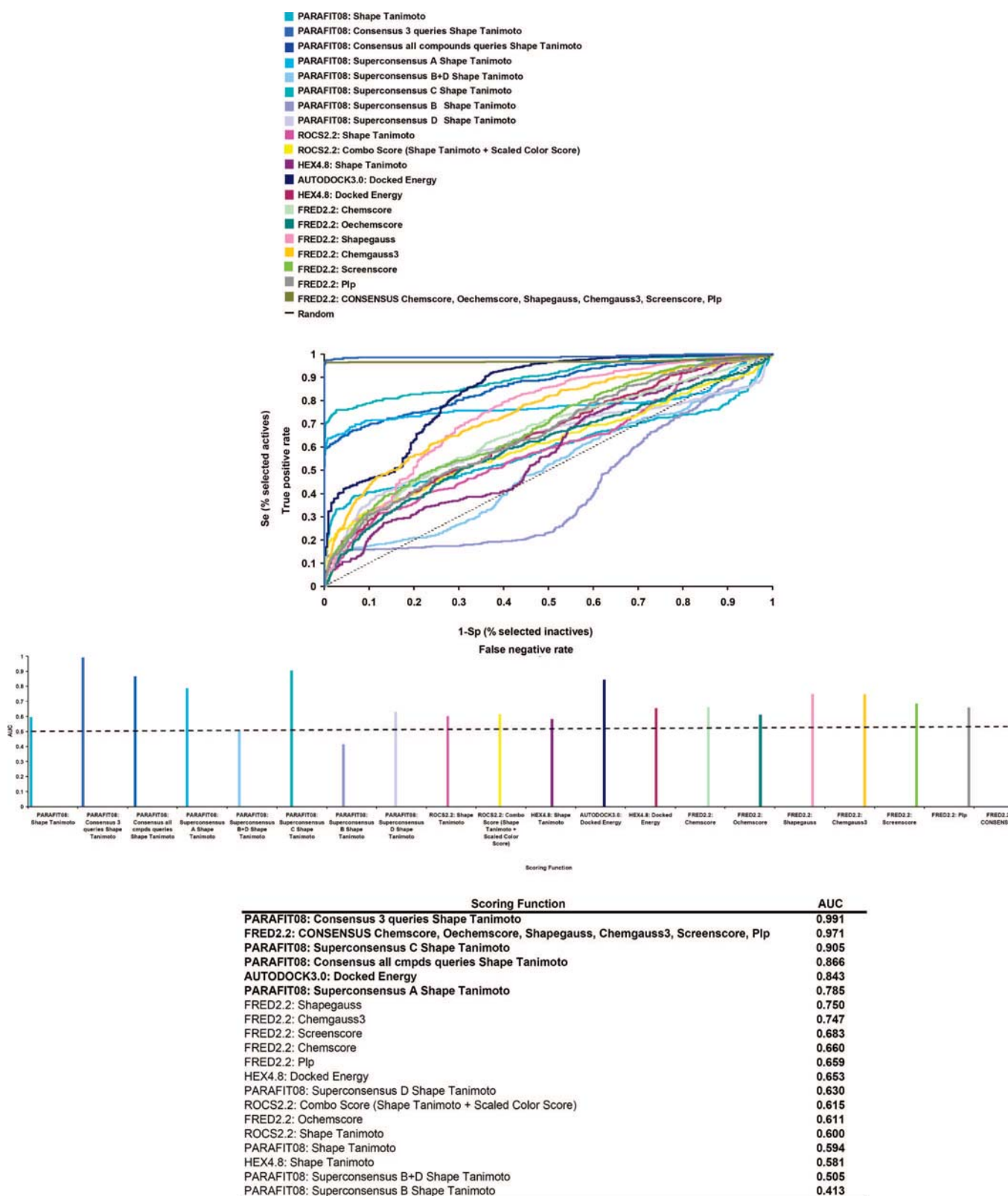


Figure 12. ROC plot validation of various shape-matching and docking VS methods compared to the consensus shape approaches for the CCR5 antagonists. The dotted black line represents the expected enrichment if actives were selected at random. The lower bar chart and table report the AUC values of the corresponding VS ROC curves. The scoring functions which give the best VS performance are shown in bold.

perazines.^{11,13,15–17,85} To our knowledge, there is not yet any experimental SDM evidence to relate compounds belonging to SC B and D to any specific binding site. However, previous docking predictions by Kellenberger et al. suggest a binding mode which includes both Site 1 and

Site 2,¹⁴ and this would be consistent with our prediction of Site 3, which is spatially located between Sites 1 and 2. Overall, the clusters and SC clusters found using our SH-based approach, and the direct correspondence of these with the spatial locations of the three binding sites predicted by

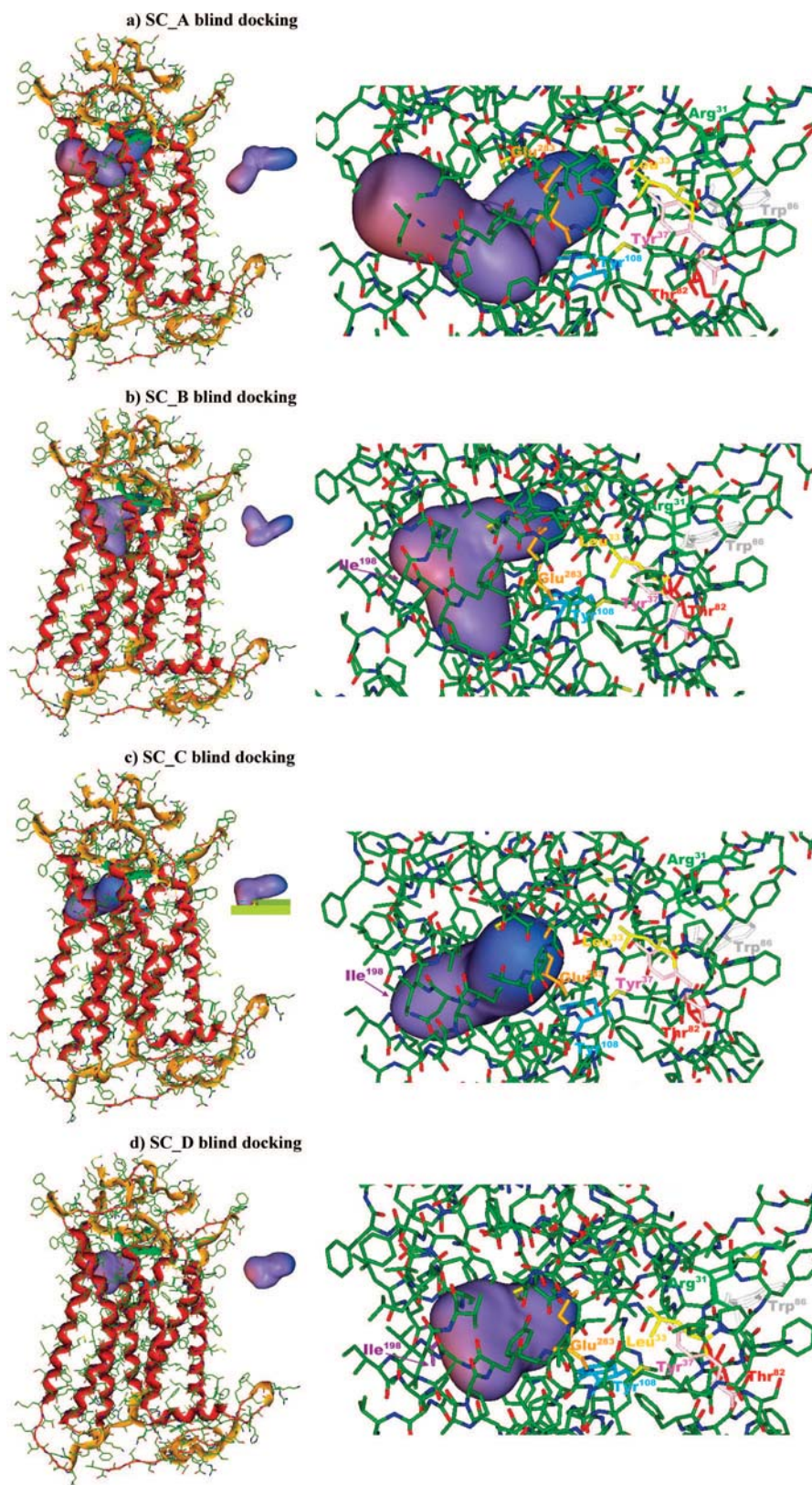


Figure 13. Hex blind docking results for the SC pseudomolecules. The images on the left show the final docked position of the SC pseudomolecules. The images on the right show close-up views of the docked conformations, annotated with the locations of known SDM binding site residues. In each case, the pseudomolecule was initially placed outside the CCR5 receptor pocket, as shown. (a) SC A blind docked onto one side of the CCR5 pocket. (b) SC B blind docked in the middle region of the pocket. (c) SC C blind docked onto the opposite side of the binding pocket. (d) SC D blind docked in the middle region of the pocket.

rigid-body pseudomolecule docking, are clearly consistent with and add weight to the previous computational predictions of Kellenberger et al. and also recently by Kondru et al.⁸⁵ In these earlier studies, different clinical drug candidates

were used to establish the nature of the binding pocket in CCR5. Although all CCR5 antagonists were predicted to bind to the same main hydrophobic pocket, in agreement with our results, both previous studies indicate that ligands may

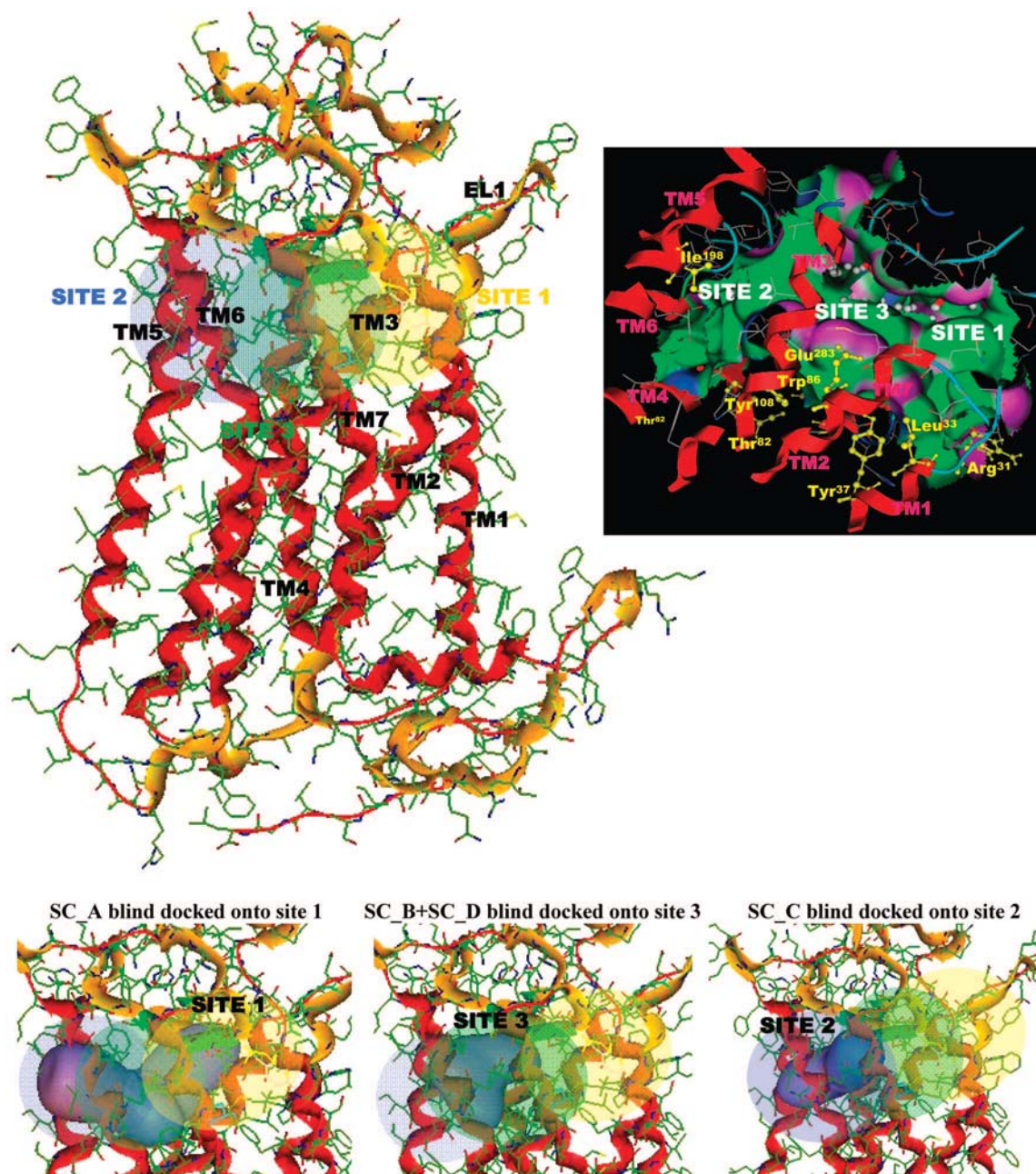


Figure 14. CCR5 binding pocket subsites proposed by the consensus VS and docking results. Here, the SC A pseudomolecule is docked onto the first subsite (Site 1), delimited by TMs 1, 2, 3, and 7. The SC C pseudomolecule is docked onto a second subsite (Site 2), delimited by TMs 3, 5, and 6. The SC B and SC D pseudomolecules are docked onto a third subsite (Site 3), delimited by TMs 3, 6, and 7, and which overlaps the SC A and SC C binding subsites. The top right image shows the CCR5 model with the proposed binding regions specified. On the top left, the van der Waals interaction surface of the CCR5 receptor cavity colored by H-Bonding (purple), hydrophobicity (green), and mild polar (blue) regions. TMs and important binding residues delimiting the three binding regions are shown in red and yellow, respectively. The bottom row of images show close-up views of the SC pseudomolecules in the three proposed subsites.

occupy different subcavities. This is clearly demonstrated by the different CCR5 mutant binding profiles obtained by Kondru et al., which is consistent with the significantly different electrostatic shapes and polarities of the CCR5 antagonists analyzed. Their docking predictions, which are based on SDM and CCR5 homology modeling data, suggest that the CCR5 receptor can accommodate structurally and electrostatically diverse antagonists by utilizing a unique set of interactions for every ligand, which is also consistent with our clustering results.

Nonetheless, the only completely reliable way to verify the validity of docking-based predictions is through comparison with a known crystallographic structure. Clearly, such

a gold standard reference is not available for CCR5. Hence any comparison with previous docking studies can, at best, serve only to add further support to the original prediction. On the other hand, for practical purposes, an unbiased and objective way to validate a structural prediction even when no crystal structure is available is to test its utility in the context of VS. The fact that the VS results obtained here using consensus and four SC shape-based similarity and docking queries give significantly enhanced screening enrichments compared to single-molecule based queries lends very strong support both to the initial validity of the notion of SC structures, and to the hypothesis that the members of

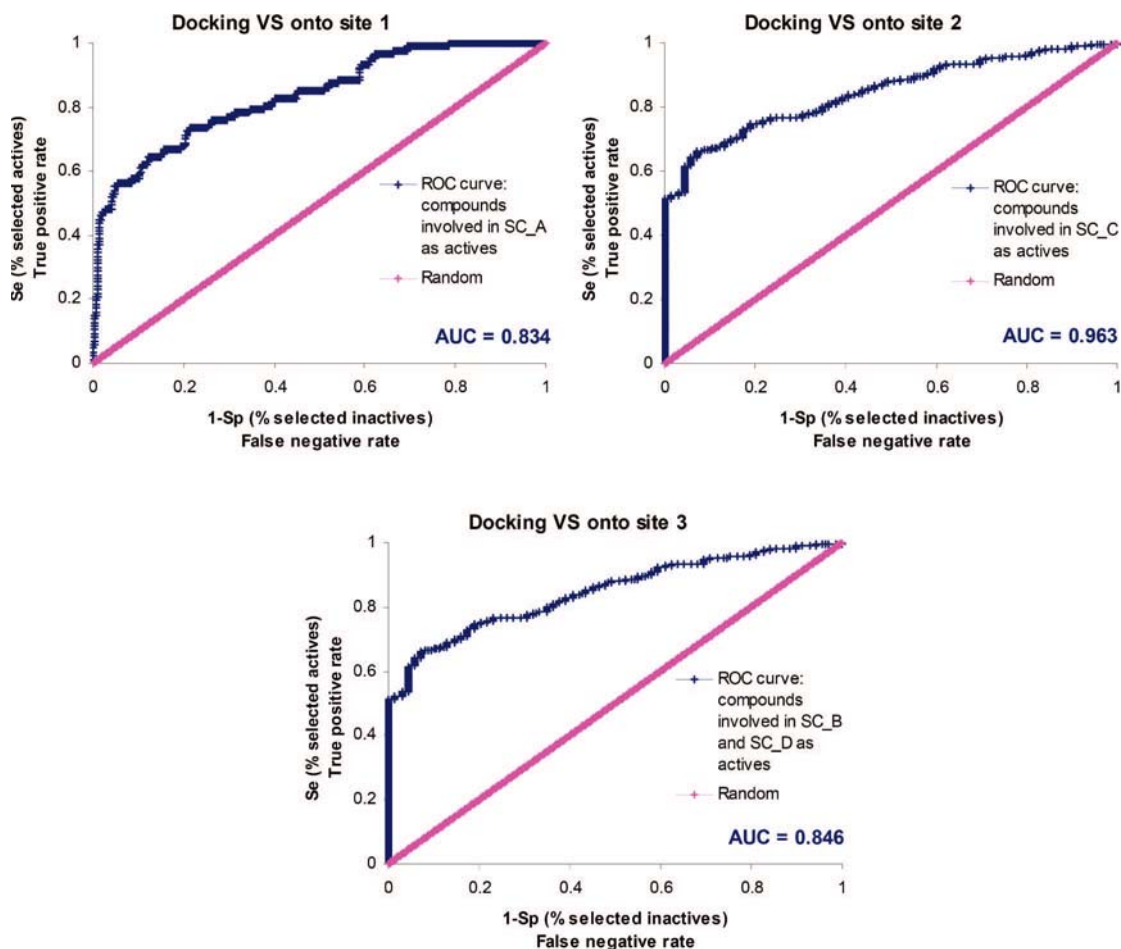


Figure 15. ROC plot validation of docking VS onto the three identified CCR5 subsites for the CCR5 antagonists. The AUC is given for each docking VS (onto Site 1, 2, and 3). The diagonal pink line represents the expected enrichment if actives were selected at random. For docking VS onto Site 1, compounds belonging to SC A are treated as actives, and compounds belonging to SC B, C, and D are treated as inactives. For docking VS onto Site 2, compounds belonging to SC C are treated as actives, and compounds belonging to SC A, B, and D are treated as inactives. For docking VS onto Site 3, compounds belonging to SC B and D are treated as actives, and compounds belonging to SC A and C are treated as inactives.

these SC clusters bind within at least three main sites in the CCR5 extracellular pocket.

CONCLUSION

This study has shown that using spherical harmonic consensus shapes as queries can be a useful strategy to improve hit enrichments in shape-based VS. We have developed a straightforward and fast method to construct consensus molecular shapes from SH surface envelopes. This consensus shape approach has been applied and validated by VS using a database of CXCR4 and CCR5 antagonists. For both receptor targets, ROC plot analyses show an improvement of VS results using the new approach. Moreover, the CCR5 multiple-binding-region hypothesis has been quantitatively explored by constructing different trial SH consensus query structures and by measuring their VS utility against our CCR5 inhibitor database. This study found four main SC clusters whose members are predicted to bind to three different but somewhat overlapping sites within the CCR5 pocket. The good VS results obtained with these virtual structures suggest they may profitably be used to search for novel inhibitors in prospective VS campaigns against other databases. Pseudomolecules corresponding to these SC clusters were docked into the CCR5 pocket, and

the locations of these positions were related to the locations predicted by previous docking studies. Several compounds within each consensus group have experimentally supported or computationally predicted binding modes which are consistent with the locations of the SC clusters docked here. Therefore, the SC structures calculated here provide strong supporting evidence for the CCR5 multiple-ligand-binding-site hypothesis, and help to give a better picture of how the CCR5 antagonists are probably distributed in the CCR5 receptor pocket.

ACKNOWLEDGMENT

The authors are grateful to OpenEye Scientific Software Inc., ChemAxon, and Cepos Insilico Ltd. for providing Academic Licences for ROCS, JChem, and ParaSurf, respectively. V.I.P.N. thanks the Generalitat de Catalunya—DURSI for a grant within the Formació de Personal Investigador (2008FI) program. This work was supported by The TV3 Marathon Foundation (AIDS-2001) promoted by the Catalan Radio and Television Corporation (Corporació Catalana de Ràdio i Televisió, CCRTV) and the Programa Nacional de Biomedicina (Ministerio de Educación y Ciencia, SAF2007-63622-C02-01).

REFERENCES AND NOTES

- (1) Jiang, S.; Lin, K.; Strick, N.; Neurath, A. R. Inhibition of HIV-1 infection by a fusion domain binding peptide from the HIV-1 envelope glycoprotein GP41. *Biochem. Biophys. Res. Commun.* **1993**, *195*, 533–538.
- (2) Kadow, J.; Wang, H. G.; Lin, P. F. Small-molecule HIV-1 gp120 inhibitors to prevent HIV-1 entry: An emerging opportunity for drug development. *Curr. Opin. Invest. Drugs* **2006**, *7*, 721–726.
- (3) Berger, E. A.; Murphy, P. M.; Farber, J. M. Chemokine receptors as HIV-1 coreceptors: Roles in viral entry, tropism, and disease. *Annu. Rev. Immunol.* **1999**, *17*, 657–700.
- (4) Markovic, I.; Clouse, K. A. Recent advances in understanding the molecular mechanisms of HIV-1 entry and fusion: revisiting current targets and considering new options for therapeutic intervention. *Curr. HIV Res.* **2004**, *2*, 223–234.
- (5) De Clercq, E. New antiviral agents in preclinical or clinical development. *Adv. Antiviral Drug Des.* **2004**, *4*, 1–62.
- (6) De Clercq, E. New Anti-HIV Agents and Targets. *Med. Res. Rev.* **2002**, *22*, 531–565.
- (7) Kazmierski, W. M.; Peckman, J. P.; Duan, M.; Kenakin, T. P.; Jenkinson, S.; Gudmundsson, K. S.; Piscitelli, S. C.; Feldman, P. L. Recent progress in the discovery of new CCR5 and CXCR4 chemokine receptor antagonists as inhibitors of HIV-1 entry. Part 2. *Curr. Med. Chem.* **2005**, *4*, 133–152.
- (8) Maeda, K.; Nakata, H.; Ogata, H.; Koh, Y.; Miyakawa, T.; Mitsuya, H. The current status of, and challenges in, the development of CCR5 inhibitors as therapeutics for HIV-1 infection. *Curr. Opin. Pharmacol.* **2004**, *4*, 447–452.
- (9) Palani, A.; Tagat, J. R. Discovery and development of small-molecule chemokine coreceptor CCR5 antagonists. *J. Med. Chem.* **2006**, *49*, 2851–2857.
- (10) Gerlach, L. O.; Skerlj, R. T.; Bridger, G. J.; Schwartz, T. W. Molecular Interaction of Cyclam and Bicyclam Non-peptide Antagonists with the CXCR4 Chemokine Receptor. *J. Biol. Chem.* **2001**, *276*, 14154–14160.
- (11) Seibert, C.; Ying, W.; Gavrilo, S.; Tsamis, F.; Kuhmann, S. E.; Palani, A.; Tagat, J. R.; Clader, J. W.; McCombie, S. W.; Baroudy, B. M.; Smith, S. O.; Dragic, T.; Moore, J. P.; Sakmar, T. P. Interaction of small molecule inhibitors of HIV-1 entry with CCR5. *Virology* **2006**, *349*, 41–54.
- (12) Pérez-Nuño, V. I.; Ritchie, D. W.; Rabal, O.; Pascual, R.; Borrell, J. I.; Teixidó, J. Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape-matching and ligand-receptor docking. *J. Chem. Inf. Model.* **2008**, *48*, 509–533.
- (13) Maeda, K.; Das, D.; Ogata-Aoki, H.; Nakata, H.; Miyakawa, T.; Tojo, Y. Structural and molecular interactions of CCR5 inhibitors with CCR5. *J. Biol. Chem.* **2006**, *281*, 12688–12698.
- (14) Kellenberger, E.; Springael, J.-Y.; Parmentier, M.; Hachet-Haas, M.; Galzi, J.-L.; Rognan, D. Identification of nonpeptide CCR5 receptor agonists by structure-based virtual screening. *J. Med. Chem.* **2007**, *50*, 1294–1303.
- (15) Wang, T.; Duan, Y. Binding modes of CCR5-targeting HIV entry inhibitors: Partial and full antagonists. *J. Mol. Graphics Modell.* **2008**, *26*, 1287–1295.
- (16) Castonguay, L. A.; Weng, Y.; Adolfsen, W.; Di Salvo, J.; Kilburn, R.; Caldwell, C. G.; Daugherty, B. L.; Finke, P. E.; Hale, J. J.; Lynch, C. L.; Mills, S. G.; MacCoss, M.; Springer, M. S.; DeMartino, J. A. Binding of 2-aryl-4-(piperidin-1-yl)butanamines and 1,3,4-trisubstituted pyrrolidines to human CCR5: A molecular modeling guided mutagenesis study of the binding pocket. *Biochemistry* **2003**, *42*, 1544–1550.
- (17) Tsamis, F.; Gavrilo, S.; Kajumo, F.; Seibert, C.; Kuhmann, S.; Ketas, T.; Trkola, A.; Palani, A.; Clader, J. W.; Tagat, J. R.; McCombie, S.; Baroudy, B.; Moore, J. P.; Sakmar, T. P.; Dragic, T. Analysis of the mechanism by which the small-molecule CCR5 antagonists SCH-351125 and SCH-350581 inhibit human immunodeficiency virus type 1 entry. *J. Virol.* **2003**, *77*, 5201–5208.
- (18) Dragic, T.; Trkola, A.; Thompson, D. A.; Cormier, E. G.; Kajumo, F. A.; Maxwell, E.; Lin, S. W.; Ying, W.; Smith, S. O.; Sakmar, T. P.; Moore, J. P. A binding pocket for a small molecule inhibitor of HIV-1 entry within the transmembrane helices of CCR5. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5639–5644.
- (19) Nishikawa, M.; Takashima, K.; Nishi, T.; Furuta, R. A.; Kanzaki, N.; Yamamoto, Y.; Fujisawa, J.-I. Analysis of binding sites for the new small-molecule CCR5 antagonist TAK-220 on human CCR5. *Antimicrob. Agents Chemother.* **2005**, *49*, 4708–4715.
- (20) Paterlini, M. G. Structure modeling of the chemokine receptor CCR5: Implications for ligand binding and selectivity. *Biophys. J.* **2002**, *83*, 3012–3031.
- (21) Fano, A.; Ritchie, D. W.; Carrieri, A. Modelling the structural basis of human CCR5 chemokine receptor function: from homology model-building and molecular dynamics validation to agonist and antagonist docking. *J. Chem. Inf. Model.* **2006**, *46*, 1223–1235.
- (22) Ritchie, D. W.; Kemp, G. J. L. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *J. Comput. Chem.* **1999**, *20*, 383–395.
- (23) *ParaSurf*, version08; CEPOS InSilico Ltd.: Erlangen, Germany, 2008; <http://www.ceposinsilico.de/Pages/Products.html> (accessed May 26, 2008).
- (24) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O. Virtual screening workflow development guided by the “receiver operating characteristic” curve approach. application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534–2547.
- (25) Ritchie, D. W.; Kemp, G. J. L. Protein docking using spherical polar Fourier correlations. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 178–194.
- (26) Lin, J.; Clark, T. An analytical, variable resolution, complete description of static molecules and their intermolecular binding properties. *J. Chem. Inf. Model.* **2005**, *45*, 1010–1016.
- (27) Mavridis, L.; Hudson, B. D.; Ritchie, D. W. Toward high throughput 3D virtual screening using spherical harmonic surface representations. *J. Chem. Inf. Model.* **2007**, *47*, 1787–1796.
- (28) Frank, J. Three-dimensional electron microscopy of macromolecular assemblies. Oxford University Press: Oxford, U.K., 2006.
- (29) Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874.
- (30) Bridger, G.; Skerlj, R.; Kaller, A.; Harwing, C.; Bogucki, D.; Wilson, T. R.; Crawford, J.; McEachern, E. J.; Atsma, B.; Nan, S.; Zhou, Y. World Patent WO 0022600, 2002.
- (31) Bridger, G.; Skerlj, R.; Kaller, A.; Harwing, C.; Bogucki, D.; Wilson, T. R.; Crawford, J.; McEachern, E. J.; Atsma, B.; Nan, S.; Zhou, Y. World Patent WO 0022599, 2002.
- (32) Bridger, G.; Skerlj, R.; Kaller, A.; Harwing, C.; Bogucki, D.; Wilson, T. R.; Crawford, J.; McEachern, E. J.; Atsma, B.; Nan, S.; Zhou, Y. World Patent WO 00234745, 2002.
- (33) Bridger, G.; Skerlj, R.; Kaller, A.; Harwing, C.; Bogucki, D.; Wilson, T. R.; Crawford, J.; McEachern, E. J.; Atsma, B.; Nan, S.; Zhou, Y. World Patent WO 055876, 2003.
- (34) Bridger, G.; Skerlj, R.; Kaller, A.; Harwing, C.; Bogucki, D.; Wilson, T. R.; Crawford, J.; McEachern, E. J.; Atsma, B.; Nan, S.; Zhou, Y.; Smith, C. D.; Di Fluir, R. M. U.S. Patent 0019058, 2004.
- (35) Ichiyama, K.; Yokohama-Kumakura, S.; Tanaka, Y.; Tanaka, R.; Hirose, K.; Bannai, K.; Edamatsu, T.; Yanaka, M.; Niitani, Y.; Miyako-Kurosaki, N.; Takaku, H.; Koyanagi, Y.; Yamamoto, N. A duodenally absorbable CXC chemokine receptor 4 antagonist, KRH-1636, exhibits a potent and selective anti-HIV-1 activity. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 4185–4190.
- (36) Murakami, T.; Yoshida, A.; Tanaka, R.; Mitsushashi, S.; Hirose, K.; Yanaka, M.; Yamamoto, N.; Tanaka, Y. KRH-2731: An Orally Bioavailable CXCR4 Antagonist Is a Potent Inhibitor of HIV-1 Infection. In *2004 Antivirals Pipeline Report*; Proceedings of the 11th Conference on Retroviruses and Opportunistic Infection, San Francisco, CA, Feb 8–11, 2004; Camp, R., Ed.; Treatment Action Group: San Francisco CA, 2004; Abstract 541.
- (37) Yamazaki, T.; Saitou, A.; Ono, M.; Yokohama, S.; Bannai, K.; Hirose, K.; Yanaka, M. World Patent WO 029218, 2003.
- (38) Yamazaki, T.; Kikumoto, S.; Ono, M.; Saitou, A.; Takahashi, H.; Kumakura, S.; Hirose, K. World Patent WO 024697, 2004.
- (39) Bridger, G. J.; Skerlj, R. T.; Padmanabhan, S.; Martellucci, S. A.; Henson, G. W.; Struyf, S.; Witvrouw, M.; Schols, D.; De Clercq, E. Synthesis and structure–activity relationships of phenylenebis(methylene)-linked bis-azamacrocycles that inhibit HIV-1 and HIV-2 replication by antagonism of the chemokine receptor CXCR4. *J. Med. Chem.* **1999**, *42*, 3971–3981.
- (40) De Clercq, E. Inhibition of HIV infection by bicyclams, highly potent and specific CXCR4 antagonists. *Mol. Pharmacol.* **2000**, *57*, 833–839.
- (41) Esté, J. A.; Cabrera, C.; De Clercq, E.; Struyf, S.; Van Damme, J.; Bridger, G.; Skerlj, R. T.; Abrams, M. J.; Henson, G.; Gutierrez, A.; Clotet, B.; Schols, D. Activity of different bicyclam derivatives against human immunodeficiency virus depends on their interaction with the CXCR4 chemokine receptor. *Mol. Pharmacol.* **1999**, *55*, 67–73.
- (42) Egberink, H. F.; De Clercq, E.; Van Vliet, A. L. W.; Balzarini, J.; Bridger, G. J.; Henson, G.; Horzinek, M. C.; Schols, D. Bicyclams, selective antagonists of the human chemokine receptor CXCR4, potentially inhibit feline immunodeficiency virus replication. *J. Virol.* **1999**, *73*, 6346–6352.
- (43) Hatse, S.; Princen, K.; De Clercq, E.; Rosenkilde, M. M.; Schwartz, T. W.; Hernandez-Abad, P. E.; Skerlj, R. T.; Bridger, G. J.; Schols, D. AMD3465, a monomacrocyclic CXCR4 antagonist and potent HIV entry inhibitor. *Biochem. Pharmacol.* **2005**, *70*, 752–761.

- (44) Princen, K.; Hatse, S.; Vermeire, K.; Aquaro, S.; De Clercq, E.; Gerlach, L.-O.; Rosenkilde, M.; Schwartz, T. W.; Skerlj, R.; Bridger, G.; Schols, D. Inhibition of human immunodeficiency virus replication by a dual CCR5/CXCR4 antagonist. *J. Virol.* **2004**, *78*, 12996–13006.
- (45) Tamamura, H.; Araki, T.; Ueda, S.; Wang, Z.; Oishi, S.; Esaka, A.; Trent, J. O.; Nakashima, H.; Yamamoto, N.; Peiper, S. C.; Otaka, A.; Fujii, N. Identification of novel low molecular weight CXCR4 antagonists by structural tuning of cyclic tetrapeptide scaffolds. *J. Med. Chem.* **2005**, *48*, 3280–3289.
- (46) Rosenkilde, M. M.; Gerlach, L. O.; Hatse, S.; Skerlj, R. L.; Schols, D.; Bridger, G.; Schwartz, T. W. Molecular mechanism of action of monociclam versus biciclam non-peptide antagonist in the CXCR4 chemokine receptor. *J. Biol. Chem.* **2007**, *282*, 27354–27365.
- (47) Palani, A.; Shapiro, S.; Clades, J. W.; Greenlee, W. J.; Blythin, D.; Cox, K.; Wagner, N. E.; Strizki, J.; Baroudy, B. M.; Dan, N. Biological evaluation and interconversion studies of rotamers of SCH 351125, an orally bioavailable CCR5 antagonist. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 705–708.
- (48) Billick, E.; Seibert, C.; Pugach, P.; Ketas, T.; Trkola, A.; Endres, M. J.; Murgolo, N. J.; Coates, E.; Reyes, G. R.; Baroudy, B. M.; Sakmar, T. P.; Moore, J. P.; Kuhmann, S. E. The differential sensitivity of human and rhesus macaque CCR5 to small-molecule inhibitors of human immunodeficiency virus type 1 entry is explained by a single amino acid difference and suggests a mechanism of action for these inhibitors. *J. Virol.* **2004**, *78*, 4134–4144.
- (49) Maeda, K.; Yoshimura, K.; Shibayama, S.; Habashita, H.; Tada, H.; Sagawa, K.; Mikayawa, T.; Auki, M.; Fukushima, D.; Mitsuya, H. Novel low molecular weight spirodiketopiperazine derivatives potently inhibit R5 HIV-1 infection through their antagonistic effects on CCR5. *J. Biol. Chem.* **2001**, *276*, 35194–35200.
- (50) Shibayama, S.; Sagawa, K.; Watanabe, N.; Takeda, K.; Tada, H.; Fukushima, D. World Patent WO 2004054616, 2004.
- (51) Takaoka, Y.; Okamoto, M.; Genba, Y. World Patent WO 2004026874, 2004.
- (52) Takaoka, Y.; Nishizawa, R.; Shibayama, S.; Sagawa, K.; Matsuo, M. Y. World Patent WO 2002074770, 2002.
- (53) Imawaka, H.; Shibayama, S.; Takaoka, Y. World Patent WO 2003035074, 2003.
- (54) Cumming, J.; Tucker, H. World Patent WO 2003042177, 2003.
- (55) Cumming, J. World Patent WO 2003042178, 2003.
- (56) Cumming, J. World Patent WO 2003080574, 2003.
- (57) Cumming, J.; Winter, J. World Patent WO 2004018425, 2004.
- (58) Burrows, J.; Cumming, J. World Patent WO 2002076, 2002.
- (59) Willoughby, C. W.; Rosauer, K. G.; Hale, J. J.; Budhu, R. J.; Mills, S. G.; Chapman, K. T.; MacCoss, M.; Malkowitz, L.; Springer, M. S.; Gould, S. L.; DeMartino, J. A.; Siciliano, S. J.; Cascieri, M. A.; Carella, A.; Catver, G.; Colmes, K.; Schlieff, W. A.; Danzeisen, R.; Hazuda, D.; Kessler, J.; Lineberger, J.; Miller, M.; Emini, E. A. 1,3,4 Trisubstituted pyrrolidine CCR5 receptor antagonists bearing 4-aminoheterocycle substituted piperidine side chains. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 427–431.
- (60) Kazmierski, W. M.; Aquino, C. J.; Bifulco, N.; Boros, E. E.; Chauder, B. A.; Chong, P. Y.; Duan, M.; Deanada, F. Jr.; Koble, C. S.; Malean, E. W.; Peckham, J. P.; Perkins, A. C.; Thompson, J. B.; Vanderwall, D. World Patent WO 2004054974, 2004.
- (61) Duan, M.; Kazmierski, W. M.; Aquino, C. J. World Patent WO 200405481, 2004.
- (62) Peckham, J. P.; Aquino, C. J.; Kazmierski, W. M. World Patent WO2004055010, 2004.
- (63) Aquino, C. J.; Chong, P. Y.; Duan, M.; Kazmierski, W. M. World Patent WO 2004055011, 2004.
- (64) Youngman, M.; Kazmierski, W. M.; Yang, H.; Aquino, C. J. World Patent WO 2004055012, 2004.
- (65) Yang, H.; Kazmierski, W. M.; Aquino, C. J. World Patent WO 2004055016, 2004.
- (66) Aramaki, Y.; Seto, M.; Okawa, T.; Oda, T.; Kanzaki, N.; Shiraishi, M. Synthesis of 1-benzothiepine and 1-benzazepine derivatives as orally active CCR5 antagonists. *Chem. Pharm. Bull.* **2004**, *52*, 254–258.
- (67) Seto, M.; Aramaki, Y.; Okawa, T.; Miyamoto, N.; Aikawa, K.; Kanzaki, N.; Shiraishi, M. Orally active antagonists as anti-HIV-1 agents: Synthesis and biological activity of 1-benzothiepine 1,1-dioxide and 1-benzazepine derivatives containing a tertiary amine moiety. *Chem. Pharm. Bull.* **2004**, *52*, 577–590.
- (68) Perros, M.; Price, D. A.; Stammen, B. L. C.; Wood, A. World Patent WO 2003084954, 2003.
- (69) Basford, P. A.; Stephenson, P. T.; Taylor, S. C. J.; Wood, A. World Patent WO 2003084954, 2003.
- (70) Armour, D. R.; Price, D. A.; Stammen, B. L. C.; Wood, A.; Perros, M.; Edwards, M. P. World Patent WO 2000038680, 2000.
- (71) Rusconi, S.; Scozzafava, A.; Mastrolorenzo, A.; Supuran, T. C. New advances in HIV entry inhibitors development. *Curr. Drug Targets Infect. Disord.* **2004**, *4*, 339–355.
- (72) Imamura, S.; Kurasawa, O.; Nara, Y.; Ichikawa, T.; Nishikawa, Y.; Iida, T.; Hashiguchi, S.; Kanzaki, N.; Lizawa, Y.; Baba, M.; Sugihara, Y. CCR5 antagonists as anti-HIV-1 agents. Part 2: Synthesis and biological evaluation of *N*-[3-(4-benzylpiperidin-1-yl)propyl]-*N,N'*-diphenylureas. *Bioorg. Med. Chem.* **2004**, *12*, 2295–2306.
- (73) Imamura, S.; Ishihara, Y.; Hattori, T.; Kurasawa, O.; Matsushita, Y.; Sugihara, Y.; Kanzaki, N.; Lizawa, Y.; Baba, M.; Hashiguchi, S. CCR5 antagonists as anti-HIV-1 agents. 1. Synthesis and biological evaluation of 5-oxopyrrolidine-3-carboxamide derivatives. *Chem. Pharm. Bull.* **2004**, *52*, 63–73.
- (74) Wei, R. G.; Arnaiz, D. O.; Chou, Y.-L.; Davey, D.; Dunning, L.; Lee, W.; Lu, S.-F.; Onuffer, J.; Ye, B.; Phillips, G. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 231–234.
- (75) Lu, S.-F.; Chen, B.; Davey, D.; Dunning, L.; Jaroch, S.; May, K.; Onuffer, J.; Phillips, G.; Subramanyam, B.; Tseng, J.-L.; Wei, R. G.; Wei, M.; Ye, B. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 1883–1887.
- (76) Debnath, A. K. Generation of predictive pharmacophore models for CCR5 antagonists: Study with piperidine- and piperazine-based compounds as a new class of HIV-1 entry inhibitors. *J. Med. Chem.* **2003**, *46*, 4501–4515.
- (77) Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Statist. Assoc.* **1963**, *58*, 236–244.
- (78) *JKlustor*, version 5.0.4; ChemAxon Ltd.: Budapest, Hungary, 2008; <http://www.chemaxon.com/jchem/doc/user/JKlustor.html> (accessed May 26, 2008).
- (79) Kelley, L. A.; Gardner, S. P.; Sutcliffe, M. J. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally-related subfamilies. *Protein Eng.* **1996**, *9*, 1063–1065.
- (80) Kleiweg, P. Data Clustering Software. <http://www.let.rug.nl/~kleiweg/index.html> (accessed May 26, 2008).
- (81) Grant, A. J.; Pickup, B. T. A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653–1659.
- (82) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Hart, W.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.
- (83) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein–ligand docking using GOLD. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 609–623.
- (84) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68*, 76–90.
- (85) Kondru, R.; Zhang, J.; Ji, C.; Mirzadegan, T.; Rotstein, D.; Sankuratri, S.; Dioszegi, M. Molecular interactions of CCR5 with major classes of small-molecule anti-HIV CCR5 antagonists. *Mol. Pharmacol.* **2008**, *73*, 789–800.

CI800257X

Artículo III

Discovery of Novel HIV Entry Inhibitors for the CXCR4 Receptor by Prospective Virtual Screening

*Violeta I. Pérez-Nueno¹, Sofia Pettersson¹, David W. Ritchie²,
José I. Borrell¹ and Jordi Teixidó^{*,1}*

1 Grup d'Enginyeria Molecular, Institut Químic de Sarrià (IQS), Universitat Ramon Llull, Barcelona, Spain
Tel: +34-93-267.20.00. Fax: +34-93-205.62.66.
E-mail: j.teixido@iqs.url.es

2 INRIA Nancy Grant Est, Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA), UMR 7503, BP 239, 54506 Vandoeuvre-les-Nancy, France
Tel: +33 3 83 59 30 00. Fax: +33 3 83 27 83 19.

ABSTRACT

The process of HIV entry begins with the binding of the viral envelope glycoprotein gp120 to both the CD4 receptor and one of CXCR4 or CCR5 chemokine coreceptors. There is currently considerable interest in developing novel ligands which can attach to these coreceptors and hence block virus-cell fusion. This article compares the application of structure-based (docking) and ligand-based (QSAR analyses, pharmacophore modelling, and shape matching) virtual screening tools to find new potential HIV entry inhibitors for the CXCR4 receptor. The comparison is based on retrospective virtual screening of a library containing different known CXCR4 inhibitors from the literature, a smaller set of active CXCR4 inhibitors selected from a large combinatorial virtual library and synthesized by us, and some drug-like presumed inactive molecules as the reference set. The enrichment factors and diversity of the retrieved molecular scaffolds in the virtual hit lists was determined. Once the different virtual screening approaches had been validated and the best parameters had been selected, prospective virtual screening of our virtual library was applied to identify new anti-HIV compounds using the same protocol as in the retrospective virtual screening analysis. The compounds selected using these computational tools were subsequently synthesized and assayed, and showed activity values ranging from 4 to 0.022 µg/ml.

INTRODUCTION

According to the World Health Organization, about 33 million people live with Acquired Immune Deficiency Syndrome (AIDS) ¹. The entry of human immunodeficiency virus (HIV) into the host cell begins with binding of the viral envelope glycoprotein gp120 to both the CD4 cell surface receptor and one of CXCR4 or CCR5 chemokine coreceptors, and leads to fusion of the viral capsid with the cell membrane. Current antiretroviral therapies (ARTs) against AIDS are generally based on reverse transcriptase inhibitors and protease inhibitors. Despite advances in the development of these potent agents which block HIV transcription and assembly, there remain problems regarding drug resistance, latent viral reservoirs, and drug induced toxic effects, which can all compromise effective control of the virus. Hence there is a need to develop new classes of anti-HIV drugs with different modes of action. Several researchers have recognized that knowledge of the mechanism of viral entry into the host cell provides further therapeutic targets against HIV infection ^{2, 3}. To date, at least three subclasses of HIV viral entry/fusion inhibitors have emerged, namely: CD4 binding or attachment inhibitors, which target

initial recognition and binding of the viral glycoprotein gp120 to the cell-surface CD4 antigen; ⁴ chemokine coreceptor binding inhibitors, which target binding of virus to the CCR5 or CXCR4 coreceptor; ⁵ and cell fusion inhibitors, which target the gp41 viral glycoprotein ⁶. Therefore, there is considerable interest in developing novel ligands which can modulate these receptors and block virus-cell fusion ^{7, 8, 9, 10, 11}.

To make progress towards this goal, we compiled a dataset of CXCR4 antagonists from the literature comprising several AMD3100 derivatives, macrocycles, KRH1636 derivatives, dipicolil amine zinc(II) complexes, cyclic peptides, and tetrahydro-quinolinamine derivatives. Several of the AMD3100 derivatives are novel, and have been synthesized in our group ¹². To this set was added some 4700 presumed inactive drug-like compounds from Maybridge Screening Collection ¹³ which have several 1D properties similar to those of the actives. The active molecules synthesized by us belong to a diverse but restricted set of compounds, selected using our PRALINS ¹⁴ program (Program for Rational Analysis of Libraries in Silico) from a large virtual combinatorial library. This library was designed to preserve the main features of AMD3100, i.e. polynitrogenated systems separated by a *p*-phenylene moiety, which is treated as an ideal reference CXCR4 antagonist. The compounds selected by PRALINS showed activities ranging from 20 to 0.008 µg/ml, and experimental binding assays confirmed that their mode of action was indeed to block the CXCR4 receptor ¹².

In order to find other active compounds without having to synthesize the whole of the combinatorial virtual library, ligand-based and structure-based virtual screening tools were used. For ligand-based virtual screening, QSAR analysis was performed with MOE ¹⁵, and a good quantitative structure-activity relationship function was obtained. 3D pharmacophore modelling using MOE and Discovery Studio ¹⁶ was applied in order to study the characteristic features of the actives necessary for interaction with the coreceptor. Shape matching using the PARAFIT ¹⁷, ROCS ¹⁸, and HEX ¹⁹ programs was also carried out to select molecules from the library with similar shapes to known actives. Because the 3D structure of CXCR4 has not yet been solved, a homology model of the protein built previously ²⁰ using bovine rhodopsin ²¹ as the template was used for receptor-based analyses using AUTODOCK ²², GOLD ²³, FRED, ²⁴ and HEX ²⁵.

In order to validate the different virtual screening approaches and to set the best parameters for each one, a retrospective virtual screening analysis was performed on the compiled active and inactive datasets. Once the best approaches were selected, prospective analysis of the as yet unsynthesized compounds in

our combinatorial virtual library was applied to establish a ranked list of new candidate CXCR4 inhibitors. A final virtual hit list was obtained from a consensus ranking of the different virtual screening approaches. Five molecules in the resulting hit list were synthesized and tested, and were found to have activity values ranging from 4 to 0.022 $\mu\text{g/ml}$. The most active of these are monocyclams, as might be expected of AMD3100 derivatives^{26, 27, 28}, and these coincided with the compounds in the first ranking positions of our hit list.

METHODS

Library Design

In this study, AMD3100, one of the earliest and still one of the most potent CXCR4 antagonists to be developed, was used as a reference ligand from which a combinatorial library was derived^{12,29}. The compounds in this library were designed in such a way as to retain the main physico-chemical features of this ligand, i.e. a central *p*-phenylene moiety with at least two nitrogen-containing substituents, one in the benzylic position and the other(s) in a heterocyclic system, and with similar distances between such nitrogens as those observed in cyclam. These considerations led us to design target compounds such as the diamines, **1**, as shown in Figure 1. A retrosynthetic analysis of those cases in which $R_1 = R_2$ and the number, *n*, of methyl linkers led to the selection of symmetrical diimines **2** as precursors, which can be extended with further methyls to give terephthalaldehyde **3** and two equivalents of the corresponding amine **4** where $n \geq 1$ (see Figure 2). When $R_1 = R_2$ and $n = 0$, compounds **2** are in fact symmetrical hydrazones which can be obtained by condensation of terephthalaldehyde and the corresponding hydrazine **4** ($n = 0$). These dihydrazones were also included in our library. In order to obtain non-symmetric ($R_1 \neq R_2$) diamines **1** ($n \geq 1$) and dihydrazones **2** ($n = 0$), it was necessary to modify slightly our synthetic approach by using 4-(diethoxymethyl)-benzaldehyde (**5**) as the core precursor. Thus, the intermediate hydrazono and aminobenzaldehydes **6** and **7** allowed such non-symmetric compounds and other non-symmetric aminohydrazones **8** to be included as further compounds in the combinatorial library (see Scheme 1). Overall, the virtual library consists of 66 amino/hydrazono-amine/hydrazone compounds (**1**, **2** and **8**), 11 amino/hydrazono-aldehyde compounds (**6** and **7**), and 11 cyclam-amine/hydrazone compounds (**9** and **10**). Some representative examples of these structures are shown in Figure 3.

Figure 1

Figure 2

Scheme 1

Figure 3

Virtual Screening Datasets

For the retrospective virtual screening analysis, a dataset of 248 CXCR4 antagonists with activity values lower than 0.1 μM against CXCR4 was assembled from the literature. This set was used for receptor-based docking and ligand-based screening analyses. A subset of the 103 most active compounds plus 48 compounds representative of other scaffold classes was then used for pharmacophore modelling. As summarised in Table 1, these compounds mainly belong to seven representative families, i.e., AMD3100 derivatives, macrocycles, KRH1636 derivatives, dipicolil amine zinc(II) complexes, tetrahydroquinolinamine derivatives, cyclic peptides, and also the most active CXCR4 inhibitors from our combinatorial virtual library which had been synthesized by us¹². Figure 4 shows some representative members of each family. These datasets were augmented with two further sets of drug-like presumed inactive compounds from the Maybridge Screening Collection (1462 for pharmacophore modelling and 4696 for the docking and shape matching approaches), selected in such a way that several of their 1D properties were similar to those of the actives (i.e. molecular weight, number of rotatable single bonds, numbers of hydrogen-bond donor and acceptor atoms, number of hydrophobic atoms, and octanol-water partition coefficient), as shown in Table 2.

Table 1

Figure 4

Table 2

For the prospective virtual screening analysis, the same presumed inactive compounds as in the retrospective analysis were used, and a subset of 34 hitherto unsynthesized compounds from the amino/hydrazono-amine/hydrazone (compounds **1**, **2** and **8**), hydrazono/amino-aldehyde (compounds **6** and **7**), and cyclam-hydrazono/amine (**9** and **10**) families were selected from the virtual library for synthesis and testing. The 3D structures of all compounds were protonated at physiological pH, assigned Gasteiger partial charges, and geometry-optimized using the MMFF94 force field. All molecules were aligned with the MOE FlexAlign module⁴⁶ using as superposition template the AMD3100 conformation obtained previously from a CXCR4 docking study²⁰ (see Figure 5).

Figure 5

QSAR Analysis

QSAR analysis applies statistical methods to describe quantitative relationships between chemical structures and biological activities of a series of analogues. The process can be divided into three general steps: 1) dataset selection, 2) data analysis, and 3) model validation. In the present QSAR study, a dataset of 39 compounds with known EC₅₀ activity values consisting of AMD3100 plus 38 further compounds synthesized by us (structures **1**, **2**, **6**, **7** and **8**) was used. This dataset was divided into a training subset of 30 compounds, and an external test set of 9 compounds, as described in Tables 3 and 4. A total of 194 descriptors were calculated with MOE, including 2D and 3D descriptors. These descriptors were then pruned using correlation analysis and forward-selection and backward-elimination methods.

Partial Least Squares (PLS) regression was used to build the QSAR models using the above descriptors as independent variables and using the biological activities as the dependent variables. Model outliers were detected using the Grubbs test, as implemented in MOE, by quantifying how far away the experimental biological activities are from the model by calculating the *Z-SCORE* ratio, defined as the difference between the experimental and model pEC₅₀ values divided by the RMSE (root mean squared error) of the whole dataset. Molecules with *Z-SCOREs* of 2.5 or higher were considered to be possible outliers. The model was then validated using leave-one-out (LOO) cross-validation and validation with an external test set (9 compounds). Several statistical parameters were used to evaluate the performance of the model:

- Correlation coefficient R^2 , cross-validated R^2 and test set validation R^2 against an external dataset, where x is the experimental pEC₅₀ and y is the model value:

$$R = \frac{\sum(x - \bar{x}) \cdot (y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \cdot \sum(y - \bar{y})^2}} \quad \text{Equation 1}$$

- Root mean squared error, *RMSE*, for the model, the cross-validation, and the external test set validation, where *PRESS* is the prediction error sum of squares and n the number of compounds:

$$RMSE = \sqrt{\frac{PRESS}{n}} \quad \text{Equation 2}$$

PRESS is an important cross-validation parameter to measure the accuracy of a model. When *PRESS* is less than *SSY* (sum of the squares deviations for the experimental values from their mean), it indicates that the model is significant and predicts better than chance. Furthermore, a *PRESS/SSY* ratio of less than 0.4, indicates that the model is a reasonable QSAR model⁴⁷.

- Cross-validated R^2 has widely been used as criterion of model robustness and predictive ability, with a threshold of 0.5 (0.6 for model R^2)⁴⁸. Nevertheless, a high cross-validated R^2 is considered a necessary condition for a model to have a high predictive power, but it is not a sufficient condition. Therefore, models are often evaluated with external test sets to estimate their true predictive power. For example, Tropsha *et al.* consider a QSAR model to be predictive if the following conditions are satisfied⁴⁹:

$$\frac{R^2 - R_0^2}{R^2} < 0.1 \quad \text{Equation 3}$$

$$0.85 \leq k \leq 1.15 \quad \text{Equation 4}$$

where R_0 is the correlation coefficient, and k is the value of the slope for the regression line through the origin (i.e. with the intercept set to 0).

- The Fisher test, or F -test, reflects the ratio of the variance explained by the model and the variance due to the error in the model. High values of the F -test indicate the reliability of the QSAR equation.

Table 3

Table 4

Ligand-Based Pharmacophore Modelling

Pharmacophore modelling studies were performed using the MOE and Discovery Studio software suites with four families of known actives from the above virtual screening dataset, namely: AMD3100 derivatives, KRH1636 derivatives, dipicolil amine zinc(II) complexes, and the most active CXCR4 inhibitors from the combinatorial virtual library. 50 conformations and a maximum of 255 conformations of each compound were calculated in MOE (MMFF94 forcefield) and Discovery Studio (Catalyst Confirm algorithm), respectively. The training set consisted of the most active compound from each family of CXCR4 inhibitors. The pharmacophore queries were built on the alignment of these four structures with the FlexAlign module in MOE and using the Common Feature Pharmacophore Generation protocol in Discovery Studio. The pharmacophore scheme of PCH (polarity-charge-hydrophobicity) was applied throughout the MOE study. Chemical features and their tolerance radii were selected between those suggested by MOE to achieve better balance between sensitivity and specificity. Also, in Discovery

Studio, hydrogen bond acceptor, hydrogen bond donor, hydrophobic, ionizable positive, and charged positive pharmacophore features were used. The maximum number of omitted features was set to one.

Ligand-Based Shape Matching Virtual Screening

Shape based virtual screening was performed using PARAFIT 08 Shape Tanimoto, ROCS 2.2 Combo Score and Shape Tanimoto, and HEX 4.8 Shape Tanimoto scores by superposing each database compound onto the docked AMD3100 query conformation. The PARAFIT and HEX superpositions were calculated using the conformation of each database compound that was calculated by MOE FlexAlign. However, as described previously²⁰, the ROCS superpositions used ten further conformations of each molecule calculated by OMEGA⁵⁰. Spherical harmonic consensus shape matching⁵¹ was also performed using PARAFIT 08 by superposing each database compound onto a consensus shape query molecule calculated from three known CXCR4 actives from different scaffold families (an AMD derivative, a macrocycle derivative, and a KRH derivative). Database molecules were ranked according to their shape Tanimoto scores with respect to the query shape. The ROCS calculations also used the “color optimization” mode to maximize both the shape and chemical property overlays (e.g. proton donor/acceptor, cationic/anionic, and hydrophobicity/aromaticity).

Receptor-Based Virtual Screening

Receptor-based screening against CXCR4 was performed using AUTODOCK 3.0, GOLD 3.0.1, FRED 2.2.1, and HEX 4.8. In AUTODOCK and GOLD, ten independent LGA and GA runs were carried out, respectively, using the same protocol as described²⁰. In GOLD, the ligands were constrained to form a hydrogen bond with a carbonyl oxygen of either Glu288, Asp171, or Asp262 which had been identified previously as key binding residues by site-directed mutagenesis (SDM)^{44, 52, 53, 54}. The ligand databases were ranked by AUTODOCK Docked Energy, Gold GoldScore and ChemScore, and a consensus score “Rank-by-Rank”⁵⁵ of these three scoring functions. In FRED, exhaustive rigid body optimization was carried out starting from the ligand conformations aligned to the docked AMD3100 conformation. PLP, Chemgauss3, Shapegauss, OEChemScore, ScreenScore, ChemScore scoring functions, and a consensus combination of these scores were used to rank the ligand databases. In HEX, docking and ranking was performed using a six-dimensional shape-only superposition correlation search with a translational distance range of 10 Å from the SDM-defined active site centre, and Hex Docked energy, respectively.

Analyzing Virtual Screening Hit Lists and Pharmacophores

Before virtual screening protocols and pharmacophoric models may be used prospectively, it is first necessary to validate them by measuring their ability to retrieve actives from a database of compounds with known biological activities. Several formulae have been proposed to score quantitatively the quality of hit lists achieved in this way⁵⁶. For example, for a database of D compounds containing A actives, and where H_t is the number of compounds in a hit list, and H_a is the number of actives in that list, the following terms may be defined⁵⁷:

Percent yield of actives:

$$Y(\%) = \frac{H_a}{H_t} \times 100 \quad \text{Equation 5}$$

Percent ratio of the actives in the hit list:

$$A(\%) = \frac{H_a}{A} \times 100 \quad \text{Equation 6}$$

Enrichment (enhancement):

$$EF = \frac{H_a/H_t}{A/D} = \frac{H_a \times D}{H_t \times A} \quad \text{Equation 7}$$

Goodness of Hit list:

$$GH = \left(\frac{H_a(3A + H_t)}{4H_t A} \right) \times \left(1 - \frac{H_t \times H_a}{D - A} \right) \quad \text{Equation 8}$$

False Negatives:

$$A - H_a \quad \text{Equation 9}$$

False Positives:

$$H_t - H_a \quad \text{Equation 10}$$

For each scoring method, the resulting hit lists were analyzed using the above terms. Following the pharmacophore modelling, shape matching and docking calculations, all compounds were sorted into ranked lists based upon their RMSD, shape matching scores, and docking energies, respectively. These lists were then used to plot the percentage of known actives found *versus* the percentage of the ranked

database screened and to calculate enrichment factors (EFs) at 1%, 5%, and 10% of the screened database.

RESULTS

PLS Analysis and Validation of QSAR Models

After descriptor pruning had been applied, five descriptors were selected to build the QSAR models, namely: VAdjEq, Q_VSA_HYD, dipoleY, SlogP_VSA8 and FASA+. Table 5 shows the correlation analysis for these descriptors. Three QSAR models were calculated as follows:

Model 1 (Figure 6):

$$pEC_{50} = 2.52586 + 0.00940 \cdot (Q_VSA_HYD) + 0.00507 \cdot (SlogP_VSA8) + 0.10611 \cdot (dipoleY)$$

$$N=29, R^2=0.81, RMSE=0.42, F=36.45, R^2_{LOO}=0.75, RMSE_{LOO}=0.49, R^2_{test}=0.69, RMSE_{test}=0.57, n=9$$

$$R_o^2=0.77, (R^2 - R_o^2)/R^2=0.049, k=0.99, PRESS=5.20, SSY=27.93, PRESS/SSY=0.19$$

Model 2:

$$pEC_{50} = 2.52568 + 0.00940 \cdot (Q_VSA_HYD) + 0.10611 \cdot (dipoleY) + 0.00507 \cdot (SlogP_VSA8)$$

$$+ 0.00130 \cdot (FASA+)$$

$$N=29, R^2=0.81, RMSE=0.42, F=26.24, R^2_{LOO}=0.75, RMSE_{LOO}=0.49, R^2_{test}=0.69, RMSE_{test}=0.57, n=9$$

$$R_o^2=0.77, (R^2 - R_o^2)/R^2=0.049, k=0.99, PRESS=5.20, SSY=27.93, PRESS/SSY=0.19$$

Model 3:

$$pEC_{50} = 2.52606 + 0.00940 \cdot (Q_VSA_HYD) + 0.00507 \cdot (SlogP_VSA8) + 0.10611 \cdot (dipoleY) -$$

$$0.00040 \cdot (VAdjEq)$$

$$N=29, R^2=0.81, RMSE=0.42, F=26.24, R^2_{LOO}=0.75, RMSE_{LOO}=0.49, R^2_{test}=0.69, RMSE_{test}=0.57, n=9$$

$$R_o^2=0.77, (R^2 - R_o^2)/R^2=0.049, k=0.99, PRESS=5.20, SSY=27.93, PRESS/SSY=0.19$$

One compound **1**{6,8} was deleted from the training set because it gave a Z-SCORE > 2.5, which indicated it is an outlier. This was confirmed by recalculating the models without it to obtain better overall statistics. All three resulting models showed R^2 values above 0.6 and R^2 for the cross-validation

and external test set validation above 0.5. In all cases, the *PRESS/SSY* ratio was below 0.4, $(R^2 - R_o^2)/R^2$ was less than 0.1, and *k* was between the above thresholds. Because the statistical results were broadly similar for all models, model 1 was selected as the most parsimonious because it used only three descriptors, whereas models 2 and 3 required four descriptors. The use of *dipoleY*, an external 3D descriptor, as independent variable in the three models enhanced the importance of a correct alignment of the molecules in order to obtain a reliable predicted activity value. Prediction of activity values for the training set and the external test set using model 1 are shown in Table 3 and Table 4. Predictions were made for compounds in the virtual library that had not yet been synthesized (**1**, **2** and **8**) and for monocyclams **9** and **10**. These results are shown in Table 6.

Table 5

Figure 6

Table 6

Pharmacophore Hypothesis Generation and Validation

Pharmacophore models were generated and retrospective analyses were performed to select models which achieved a good balance between sensitivity and specificity. Several models were proposed (Table 7), five using MOE (Models 1 to 5) and four using Discovery Studio (Models 6 to 9). Model 1 was built using the MOE Pharmacophore Elucidate module. Features in models 2 and 3 were selected from the consensus analysis performed with MOE Pharmacophore Query module. Models 4 and 5 were manually designed based on the description of the interactions of AMD3100 and CXCR4^{44, 54, 58, 59, 60, 61, 62}. Finally, models 6 to 9 were built in Discovery Studio using the Hypogen and HipHop algorithms to generate hypotheses and to select the best common pharmacophore features produced. The retrospective analysis of the models showed that pharmacophore model 1 (Figure 7) was highly selective with our dataset, giving no false positives and only nine false negatives. This model accurately classified and ranked all the known actives in the dataset, except for the KRH1636 analogues which were positioned at the end of the hit list. Visual inspection of the hit lists in the retrospective analysis showed that the ranking of each compound depended on the model and type of compound. More reliable results were obtained using a consensus of the five MOE models in Table 7.

Figure 7

Table 7

A prospective analysis using the consensus pharmacophore model was then applied to select new compounds for synthesis and testing. These molecules included hitherto unsynthesised compounds from the virtual combinatorial library (i.e. amino/hydrazono-amine/hydrazone compounds **1**, **2**, and **8**, amino/hydrazono-aldehyde compounds **6** and **7** and cyclam-amine/hydrazone compounds **9** and **10**). All of these compounds can be seen to match the pharmacophore model equally well. The screened compounds selected by the consensus of pharmacophore models and their score values are shown in Table 8.

Table 8

Docking Enrichments

In order to analyse the ability of the receptor model structure to discriminate active compounds from decoys, retrospective analysis of docking enrichment curves was performed as described previously.²⁰ Next, enrichment curves for the virtual combinatorial library compounds were calculated using the same protocol. Figure 8 shows the enrichment curves obtained. Inspection of these results shows that the enrichments obtained with the FRED consensus, Consensus scoring (AUTODOCK Docked Energy, GOLD GoldScore and ChemScore), and ChemScore scoring functions are the best, as was observed in the retrospective analysis. Looking at the first percentages of the ranked hit lists, the compounds selected by these three scoring functions can be seen to belong to **9**, **10**, **1**, **2**, and **8**. The compounds found at the top 10% of the ranked hit list using these three scoring functions, as well as AUTODOCK Docked Energy, HEX Docked Energy, and FRED Chemgauss3 are nearly the same. The screened compounds selected by these scoring functions and their score value are shown in Table 9.

Figure 8

Table 9

Shape Matching Enrichments

Because no crystallographic ligand conformation is available for the current system, the SDM-compatible conformation of AMD3100 found previously from computational docking was used as the database query. In order to study the performance of this query structure and the parameters used in the screening protocol, a retrospective analysis of shape matching enrichment curves was first performed²⁰. Next, enrichment curves for the combinatorial virtual library compounds were calculated using the same

protocol. Moreover, a consensus query was built from three different scaffold CXCR4 known actives (an AMD derivative, a macrocycle derivative, and a KRH derivative) and a retrospective analysis was performed. Enrichment curves for the virtual combinatorial library compounds were also calculated showing similar results to the basic PARAFIT AMD3100 query shape Tanimoto score. Figure 9 shows that the ROCS Combo Score and PARAFIT Tanimoto Score and Consensus Shape Tanimoto give the best EFs, as in the retrospective analysis. HEX Shape Tanimoto and ROCS Shape Tanimoto also perform well. Overall, the ligand-based shape matching tools perform better than the docking tools used here. However, looking at the first percentages of the ranked hit lists obtained, the compounds selected by these shape matching methods belong to **9**, **10**, **1**, **2**, and **8**, as found with the docking tools. Molecules found at the top 10% hit ranking list are the same using these different shape matching approaches. The screened compounds selected by these shape-based methods and their score value are shown in Table 10.

Figure 9

Table 10

Hit Selection

A consensus “Rank-by-Vote”⁵⁵ of all the first hit ranking lists compounds found was performed, and five compounds were selected to be synthesized: **1**{7,8}, **8**{2,8}, **8**{1,8}, **10**{11}, and **10**{8}. Both of the cyclam-amine compounds (**10**) were classified in the top of the ranked list in the virtual screenings but we selected the two best ranked and the three best classified amino/hydrazono-amines. Compound **8**{1,8} was toxic at a concentration of 4.1 µg/ml and showed no activity below this concentration (Table11). However, compounds **1**{7,8} and **8**{2,8} showed anti-HIV activity values of 0.6 and 0.4 µg/ml, respectively, and the cyclam-amine compounds **10**{11} and **10**{8} showed the best anti-HIV activities of 0.058 µg/ml and 0.022 µg/ml, respectively.

Table 11

DISCUSSION

A combination of ligand-based and receptor-based screening tools was used to select molecules from the virtual combinatorial library. The different approaches used generally select similar molecules at the first percentages of the ranked hit lists. Compounds selected by the various ligand-based virtual screening tools are practically the same, whereas those selected by the structure-based docking tools also include

some others. All shape-based and pharmacophore ligand-based approaches, and consensus scoring of AUTODOCK and GOLD scoring functions, FRED consensus and Chemgauss3, and the HEX Docked Energy approaches select nearly the same molecules at first percentages of database screened. However, although ligand-based searches give better results than structure-based docking for both retrospective and prospective virtual screening analyses, the pharmacophore models and also AUTODOCK Docked Energy give the best correlation with experimental data. Of the five compounds selected by the Rank-by-Vote consensus, compound **8**{1,8} was toxic below 5 µg/ml, but **1**{7,8} and **8**{2,8}, showed activity values below 1 µg/ml., and the remaining two, **10**{11} and **10**{8}, both of which are monocyclams, showed activity values below 0.06 µg/ml. Our proposed QSAR model agrees well with the experimental results, especially for the non-monocyclam compounds, with predicted activities of 0.66, 1.58, 7.30, and 0.87 µM for **1**{7,8}, **8**{2,8}, **10**{11}, and **10**{8}, which differ by only 0.37, 0.02, 7.17, and 0.82 µM, respectively, from the experimental biological values.

Overall, our screening procedure selects the most active compounds from our combinatorial virtual library (i.e. **1**{8,8} 0.008 µg/ml, **1**{8,9} 0.03 µg/ml, **1**{5,6} 0.2 µg/ml, **1**{8,10} 0.4 µg/ml) in the first ranking positions of the final consensus list. Moreover, the first five unsynthesised compounds which were also predicted to be active were ranked in order of their known activities. Hence our screening procedure can be seen to perform rather well.

CONCLUSION

A database of CXCR4 inhibitors and similar presumed inactive compounds was compiled from the literature in order to perform retrospective virtual screening. This database was used to compare docking-based and ligand-based (i.e. pharmacophore modelling and shape matching) virtual screening approaches. Additionally, a large virtual combinatorial library of candidate CXCR4 antagonists was designed, and the above screening approaches were used to select five compounds for synthesis and testing. The actives identified in this way had activities in the range 20 to 0.008 µg/ml. Experimental binding assays of those compounds confirmed that their mode of action was to block the CXCR4 receptor. Activity values were used for the development of ligand-based QSAR models in order to use them to predict activity of hitherto unsynthesised molecules. Prospective virtual screening, using the same protocol as in retrospective screening analysis, was then used to guide the selection of other molecules from the virtual

combinatorial library. Molecules found at the first positions of the consensus ranked hit list showed activity values in the range from 4 to 0.022 $\mu\text{g/ml}$.

ACKNOWLEDGEMENTS

We are grateful to OpenEye Scientific Software Inc. for providing an Academic Licence for ROCS, and to Cepos Insilico Ltd. for providing PARASURF and PARAFIT. The authors are grateful to José Esté, Imma Clotet-Codina, and Mercedes Armand-Ugón from Laboratori de Reroviologia IrsiCaixa, Hospital Universitari Germans Trias I Pujol, Universitat Autònoma de Barcelona for carrying out the biological activity tests. SP thanks the Institut Químic de Sarrià (IQS) for a predoctoral grant and VIPN thanks the Generalitat de Catalunya – DURSI for a grant within the Formació de Personal Investigador (2008FI) program. This work was supported by The TV3 Marathon Foundation (AIDS-2001) promoted by the Catalan Radio and Television Corporation (Corporació Catalana de Ràdio i Televisió, CCRTV) and the Programa Nacional de Biomedicina (Ministerio de Educación y Ciencia, SAF2007-63622-C02-01).

REFERENCES

1. UNAIDS. AIDS epidemic update: December 2007; <http://www.unaids.org/en/KnowledgeCentre/HIVData/EpiUpdate/EpiUpdArchive/2007/default.asp> (accessed Nov. 11, 2008).
2. De Clercq, E. Emerging anti-HIV drugs. *Expert Opin. Emerg. Drugs* **2005**, *10*, 241-274.
3. De Clercq, E. Anti-HIV chemotherapy: current state of the art. *Med. Chem. Res.* **2004**, *13*, 439-478.
4. Kadow, J.; Wang, H. G.; Lin, P. F. Small-molecule HIV-1 gp120 inhibitors to prevent HIV-1 entry: an emerging opportunity for drug development. *Curr. Opin. Investig. Drugs* **2006**, *7*, 721-726.
5. Berger, E. A.; Murphy, P. M.; Farber, J.M. Chemokine receptors as HIV-1 coreceptors: Roles in viral entry, tropism, and disease. *Ann. Rev. Immunol.* **1999**, *17*, 657-700.
6. Jiang, S.; Lin, K.; Strick, N.; Neurath, A. R. Inhibition of HIV-1 infection by a fusion domain binding peptide from the HIV-1 envelope glycoprotein GP41. *Biochem. Biophys. Res. Commun.* **1993**, *195*, 533-538.
7. De Clercq, E. New antiviral agents in preclinical or clinical development. *Adv. Antiviral Drug Des.* **2004**, *4*, 1-62.
8. De Clercq, E. New Anti-HIV Agents and Targets. *Med. Res. Rev.* **2002**, *22*, 531-565.
9. Bean, P. New Drugs Targets for HIV. *Clin. Infect. Dis.* **2005**, *41*, 96-100.
10. Markovic, I.; Clouse, K. A. Recent advances in understanding the molecular mechanisms of HIV-1 entry and fusion: revisiting current targets and considering new options for therapeutic intervention. *Curr. HIV Res.* **2004**, *2*, 223-34.
11. Kazmierski, W. M.; Peckman, J. P.; Duan, M.; Kenakin, T. P.; Jenkinson, S.; Gudmundsson, K. S.; Piscitelli, S. C.; Feldman, P. L.; Recent Progress in the Discovery of New CCR5 and CXCR4 Chemokine Receptor Antagonists as Inhibitors of HIV-1 Entry. Part 2. *Curr. Med. Chem. – Anti Infect. Agents* **2005**, *4*, 133-152.
12. Pettersson, S.; Pérez-Nuño, V. I.; Ros-Blanco, L.; Puig de la Bellacasa, R.; Rabal, O.; Batllori, X.; Clotet, B.; Clotet-Codina, I.; Armand-Ugón, M.; Esté, J.; Borrell, J. I.; Teixidó, J. Discovery of novel non-cyclam polynitrogenated CXCR4 coreceptor inhibitors. *ChemMedChem* **2008**, *3*, 1549-1557.
13. *Maybride Bringing life to drug discovery*, Maybridge Databases Autumn 2005; Fisher Scientific International: England, 2005.
14. Pascual, R.; Borrell, J. I.; Teixido, J. Analysis of selection methodologies for combinatorial library design. *Mol. Divers.* **2003**, *6*, 121-133.
15. *MOE (Molecular Operating Environment)*, 2006.08 Release; Chemical Computing Group, Inc.: Montreal, Canada, 2004.
16. *Discovery Studio*, version 2.0; Accelrys Software Inc.: San Diego, 2007.
17. Lin, J.; Clark, T. An analytical, variable resolution, complete description of static molecules and their intermolecular binding properties. *J. Chem. Inf. Model.* **2005**, *45*, 1010-1016.
18. Grant, A. J.; Pickup, B. T. A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653-1659.
19. Ritchie, D.W.; Kemp, G. J. L. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *J. Comp. Chem.* **1999**, *20*, 383-395.
20. Pérez-Nuño V. I.; Ritchie, D. W.; Rabal, O.; Pascual, R.; Borrell, J. I.; Teixidó, J. Comparison of Ligand-Based and Receptor-Based Virtual Screening of HIV Entry Inhibitors for the CXCR4 and CCR5 Receptors Using 3D Ligand Shape-matching and Ligand-Receptor Docking. *J. Chem. Inf. Model.* **2008**, *48*, 509-533.
21. Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A.; Le Trong, I.; Teller, D. C.; Okada, T.; Stenkamp, R. E.; Yamamoto, M.; Miyano, M. Crystal structure of rhodopsin: A G-protein-coupled receptor. *Science* **2000**, *289*, 739-745.
22. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Hart, W.; Belew, R.K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639-1662.

23. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. *Proteins: Struct. Funct. Genet.* **2003**, *52*,609-623.
24. McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68*, 76-90.
25. Ritchie, D. W.; Kemp, G. J. L. Protein docking using spherical polar Fourier correlations. *Proteins: Struct. Funct. Genet.* **2000**, *39*, 178-194.
26. Hatse, S.; Princen, K.; De Clercq, E.; Rosenkilde, M. M.; Schwartzb, T. W.; Hernandez-Abad, P. E.; Skerlj, R. T.; Bridger, G. J.; Schols, D. AMD3465, a monomacrocyclic CXCR4 antagonist and potent HIV entry inhibitor. *Biochem. Pharmacol.* **2005**, *70*, 752-761.
27. Princen, K.; Hatse, S.; Vermeire, K.; Aquaro, S.; De Clercq, E.; Gerlach, L.-O.; Rosenkilde, M.; Schwartz, T. W.; Skerlj, R.; Bridger, G.; Schols, D. Inhibition of Human Immunodeficiency Virus Replication by a Dual CCR5/CXCR4 Antagonist. *J. Virol.* **2004**, *78*, 12996-13006.
28. Rosenkilde M. M.; Gerlach L.-O.; Hatse, S.; Skerlj R. L.; Schols D.; Bridger G.; Schwartz T. W. Molecular mechanism of action of monocyclam versus bicyclam non-peptide antagonist in the CXCR4 chemokine receptor. *J. Biol. Chem.* **2007**, *282*, 27354-27365.
29. Teixidó, J.; Borrell, J. I.; Nonell, S.; Pettersson, S.; Ros, L.; Puig de la Bellacasa, R.; Rabal, M. O.; Pérez-Nueno, V. I.; Esté, J.; Clotet-Codina, I.; Armand-Ugón, M.; Nuevos sistemas polinitrogenados como agentes anti-VIH. ES Patent ES200602764, **2006** (filing date: October 26, 2006).
30. Bridger, G.; Skerlj R.; Kaller A.; Harwing C.; Bogucki D.; Wilson T. R.; Crawford J.; McEachern E. J.; Atsma B.; Nan S.; Zhou Y. World Patent WO 0022600, 2002.
31. Bridger, G.; Skerlj R.; Kaller A.; Harwing C.; Bogucki D.; Wilson T. R.; Crawford J.; McEachern E. J.; Atsma B.; Nan S.; Zhou Y. World Patent WO 0022599, 2002.
32. Bridger, G.; Skerlj R.; Kaller A.; Harwing C.; Bogucki D.; Wilson T. R.; Crawford J.; McEachern E. J.; Atsma B.; Nan S.; Zhou Y. World Patent WO 00234745, 2002.
33. Bridger, G.; Skerlj R.; Kaller A.; Harwing C.; Bogucki D.; Wilson T. R.; Crawford J.; McEachern E. J.; Atsma B.; Nan S.; Zhou Y. World Patent WO 055876, 2003.
34. Bridger, G.; Skerlj R.; Kaller A.; Harwing C.; Bogucki D.; Wilson T. R.; Crawford J.; McEachern E. J.; Atsma B.; Nan S.; Zhou Y.; Smith C. D.; Di Fluir R. M. US Patent 0019058, 2004.
35. Ichiyama, K.; Yokohama-Kumakura S.; Tanaka Y.; Tanaka R.; Hirose K.; Bannai K.; Edamatsu T.; Yanaka M.; Niitani Y.; Miyako-Kurosaki N.; Takaku H.; Koyanagi Y.; Yamamoto N.; *Proc. Natl. Acad. Sci.* **2003**, *100*, 4185-4190.
36. Murakami, T.; Yoshida, A.; Tanaka, R.; Mitsuhashi, S.; Hirose, K.; Yanaka, M.; Yamamoto. N.; Tanaka, Y. KRH-2731: An Orally Bioavailable CXCR4 Antagonist Is a Potent Inhibitor of HIV-1 Infection. In *2004 Antivirals Pipeline Report*; Camp, R., Ed.; Proceedings of the 11th Conference on Retroviruses and Opportunistic Infections, San Francisco, CA, Feb. 8-11, 2004; Treatment Action Group: San Francisco, CA, 2004; Abstract No. 541.
37. Yamazaki, T.; Saitou, A.; Ono, M.; Yokohama, S.; Bannai, K.; Hirose, K.; Yanaka, M. World Patent WO 029218, 2003.
38. Yamazaki, T.; Kikumoto, S.; Ono, M.; Saitou, A.; Takahashi, H.; Kumakura, S.; Hirose, K. World Patent WO 024697, 2004.
39. Bridger, G. J.; Skerlj, R. T.; Padmanabhan, S.; Martellucci, S. A.; Henson, G. H.; Struyf, S.; Witvrouw, M.; Schols, D.; De Clercq, E. Synthesis and Structure-Activity Relationships of Phenylenebis(methylene)-Linked Bis-azamacrocycles That Inhibit HIV-1 and HIV-2 Replication by Antagonism of the Chemokine Receptor CXCR4. *J. Med. Chem.* **1999**, *42*, 3971-3981.
40. De Clercq, E. Inhibition of HIV Infection by Bicyclams, Highly Potent and Specific CXCR4 Antagonists. *Mol. Pharmacol.* **2000**, *57*, 833-839.
41. Esté, J. A.; Cabrera, C.; De Clercq, E.; Struyf, S.; Damme, J. V.; Bridger, G.; Skerlj, R. T.; Abrams, M. J.; Henson, G.; Gutierrez, A.; Clotet, B.; Schols, D. Activity of Different Bicyclam Derivatives against Human Immunodeficiency Virus Depends on Their Interaction with the CXCR4 Chemokine Receptor. *Mol. Pharmacol.* **1999**, *55*, 67-73.

42. Egberink, H. F.; De Clercq, E.; Van Vliet, A. L. W.; Balzarini, J.; Bridger, G. J.; Henson, G.; Horzinek, M. C.; Schols, D. Bicyclams, Selective Antagonists of the Human Chemokine Receptor CXCR4, Potently Inhibit Feline Immunodeficiency Virus Replication. *J. Virol.* **1999**, *73*, 6346-6352.
43. Tamamura H.; Araki T.; Ueda S.; Wang Z.; Oishi S.; Esaka A.; Trent J.O.; Nakashima H.; Yamamoto N.; Peiper S.C.; Otaka A.; Fujii N. Identification of novel low molecular weight CXCR4 antagonists by structural tuning of cyclic tetrapeptide scaffolds. *J. Med. Chem.* **2005**, *48*, 3280-3289.
44. Gerlach, L.-O.; Skerlj, R. T.; Bridger, G. J.; Schwartz, T. W. Molecular Interaction of Cyclam and Bicyclam Non-peptide Antagonists with the CXCR4 Chemokine Receptor. *J. Biol. Chem.* **2001**, *276*, 14154-14160.
45. Tamamura, H.; Ojida, A.; Ogawa, T.; Tsutsumi, H.; Masuno, H.; Nakashima, H.; Yamamoto, N.; Hamachi, I.; Fujii, N. Identification of a new class of low molecular weight antagonists against the chemokine receptor CXCR4 having the dipicolylamine-zinc(II) complex structure. *J. Med. Chem.* **2006**, *49*, 3412-3415.
46. Labute, P. Flexible Alignment of Small Molecules. Chemical Computing Group, Inc.: Montreal, Canada, 2004 (Available in Internet at <http://www.chemcomp.com/journal/malign.htm>, accessed June, 2, 2008).
47. Agrawal, V. K.; Singh, J.; Gupta, M.; Jaliwala, Y. A.; Khadikar, P. V.; Supuran, C. T. QSAR studies on benzopyran potassium channel activators. *Eur. J. Med. Chem.* **2006**, *41*, 360-366.
48. Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. -D.; Lee, K. -H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided Mol. Des.* **2003**, *17*, 241-253.
49. Golbraikh, A.; Tropsha, A. Beware of q^2 ! *J. Mol. Graph. Model.* **2002**, *20*, 269-276.
50. OMEGA, version 2.1.0; OpenEye Scientific Software Inc.: Santa Fe, NM., 2006.
51. Pérez-Nueno V. I.; Ritchie, D. W.; Borrell, J. I.; Teixidó, J. Clustering and classifying diverse HIV entry inhibitors using a novel consensus shape based virtual screening approach: Further evidence for multiple binding sites within the CCR5 extracellular pocket. *J. Chem. Inf. Model.* **2008**, *48*, 2146-2165.
52. Hatse, S.; Princes, K.; Vermeire, K.; Gerlach, L.-O.; Rosenkilde, M. M.; Schwartz, T. W.; Bridger, G.; De Clercq, E.; Schols, D. Mutations at the CXCR4 interaction sites for AMD3100 influence anti-CXCR4 antibody binding and HIV-1 entry. *FEBS Letters* **2003**, *546*, 300-306.
53. BreLOT, A.; Heveker, N.; Montes, M.; Alizon, M. Identification of Residues of CXCR4 Critical for Human Immunodeficiency Virus Coreceptor and Chemokine Receptor Activities. *J. Biol. Chem.* **2000**, *275*, 23736-23744.
54. Hatse, S.; Princen, K.; Gerlach, L.-O.; Bridger, G.; Henson, G.; Clercq, E.; Schwartz, T. W.; Schols, D. Mutation of Asp171 and Asp262 of the chemokine receptor CXCR4 impairs its coreceptor function for human immunodeficiency virus-1 entry and abrogates the antagonistic activity of AMD3100. *Mol. Pharmacol.* **2001**, *60*, 164-173.
55. Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422-1426.
56. Güner, O. F.; Henry, D. R. Metric for analyzing hit lists and pharmacophores. Chapter 11. In *Pharmacophore, perception, development and use in drug design*; Edited by Osman F. Güner; International University Line: La Jolla, California, 2000; pp 195-212.
57. Güner, O. F.; Hoffman, R.; Li, H. Techniques and strategies in 3D data mining. In *Report by Wendy A. Warr*; 217th ASC National Meeting and Exposition, Anaheim, California, March 12-25, 1999; Wendy Warr & Associates: London, 1999, 50-53.
58. Hunter, T. M.; McNae, I. W.; Simpson, D. P.; Smith, A. M.; Moggach, S.; White, F.; Walkinshaw, M. D.; Parsons, S.; Sadler, P. J. Configurations of nickel-cyclam antiviral complexes and protein recognition. *Chem. Eur. J.* **2007**, *13*, 40-50.
59. Liang, X.; Parkinson, J. A.; Wishäupl, M.; Gould, R. O.; Paisey, S. J.; Park, H.; Hunter, T. M.; Blindauer, C. A.; Parsons, S.; Sadler, P. J. Structure and dynamics of metallomacrocycles: recognition of zinc xylyxl-bicyclam by an HIV coreceptor. *J. Am. Chem. Soc.* **2002**, *124*, 9105-9112.

60. Hunter, T. M.; McNae, I. W.; Liang, X.; Bella, J.; Parsons, S.; Walkinshaw, M. D.; Sadler, P. J. Protein recognition of macrocycles: Binding of anti-HIV metalocyclams to lysozyme. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 2288-2292.
61. Rosenkilde, M. M.; Gerlach, L.-O.; Jakobsen, J. S.; Skerlj, R. T.; Bridger, G. J.; Schwartz, T. W. Molecular mechanism of AMD3100 antagonism in the CXCR4 receptor. *J. Biol. Chem.* **2004**, *279*, 3033-3041.
62. Valks, G. C.; McRobbie, G.; Lewis, E. A.; Hubin, T. J.; Hunter, T. M.; Sadler, P. J.; Pannecouque, C.; De Clercq, E.; Archibald, S. J. Configurationally restricted bismacrocylic CXCR4 receptor antagonists. *J. Med. Chem.* **2006**, *49*, 6162-6165.

CXCR4 inhibitors for retrospective docking and shape based virtual screening		
<i>Family</i>	<i>Number of compounds</i>	<i>References</i>
Tetrahydroquinolinamine derivatives	123	11, 30, 31, 32, 33, 34
KRH1636 derivatives	23	11, 35, 36, 37, 38
Macrocycles	4	39
AMD3100 derivatives	94	11, 26, 27, 39, 40, 41, 42
Cyclic Peptides	2	43
Other	2	44
Total	248	
CXCR4 inhibitors for retrospective pharmacophore model based virtual screening		
<i>Family</i>	<i>Number of compounds</i>	<i>References</i>
KRH1636 derivatives	13	11, 35, 36, 37, 38
Dipicolil amine zinc(II) complexes	10	45
AMD3100 derivatives and macrocycles	90	11, 26, 27, 39, 40, 41, 42
Active molecules from the combinatorial virtual library (amino-amine, amino-aldehyde)	38	12
Total	151	

Table 1. Summary of the CXCR4 inhibitor families used in the current study.

Comparison of datasets used in pharmacophore modelling	MW	b_1rotN	a_acc	a_don	a_hyd	SlogP
151 CXCR4 actives	485.2 (104.9)	6.9 (3.9)	3.5 (1.5)	1.5 (1.5)	26.3 (5.0)	-0.8 (2.5)
1462 inactives	381.4 (64.9)	5.1 (2.0)	4.0 (1.1)	1.2 (1.1)	16.6 (2.6)	2.6 (0.9)
Comparison of datasets used in docking and shape matching approaches	MW	b_1rotN	a_acc	a_don	a_hyd	SlogP
248 CXCR4 actives	507.3 (74.4)	9.2 (4.9)	4.9 (1.1)	1.7 (1.3)	27.6 (4.2)	4.3 (3.0)
4696 inactives	497.4 (45.6)	6.2 (2.4)	3.6 (1.6)	0.9 (1.0)	21.8 (4.1)	5.5 (1.9)

Table 2. Summary of the 1D physico-chemical properties of active and inactive molecules in the screening databases used in pharmacophore modelling, docking, and shape matching approaches.

This table shows the average and standard deviation (in parenthesis) of the following properties: MW (molecular weight); b_1rotN (number of rotatable single bonds); a_acc (number of hydrogen-bond acceptor atoms); a_don (number of hydrogen-bond donor atoms); a_hyd (number of hydrophobic atoms); S_logP (octanol-water partition coefficient).

	Compound	Name	pEC ₅₀	predicted pEC ₅₀	Residue
1		8 {1,11}	4.512	4.864	-0.352
2		8 {2,4}	4.367	5.303	-0.936
3		8 {2,5}	5.251	5.772	-0.521
4		8 {2,9}	5.005	4.957	0.048
5		8 {3,5}	5.281	4.695	0.586
6		8 {3,6}	4.483	4.578	-0.095
7		8 {3,8}	5.424	5.135	0.289
8		8 {3,9}	4.503	4.238	0.265
9		8 {3,11}	4.647	4.061	0.586
10		1 {4,4}	4.511	4.917	-0.406
11		1 {4,5}	5.299	5.291	0.008
12		1 {4,6}	4.852	4.962	-0.110
13		1 {4,9}	4.659	4.913	-0.254
14		1 {5,5}	5.600	5.548	0.052
15		1 {5,6}	6.250	5.351	0.899
16		1 {5,7}	5.327	5.465	-0.138
17		1 {5,10}	5.177	5.070	0.107
18		1 {6,7}	5.254	5.321	-0.067

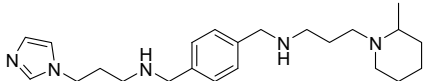
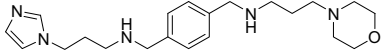
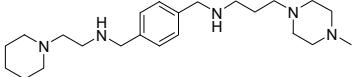
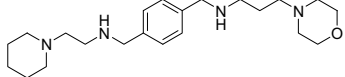
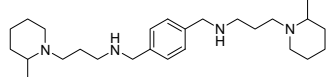
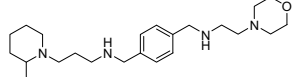
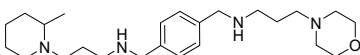
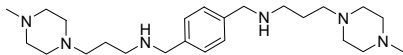
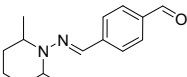
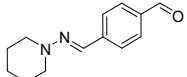
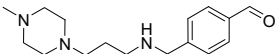
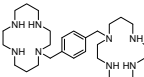
	Compound	Name	pEC ₅₀	predicted pEC ₅₀	Residue
19		1{6,8}	1.106	outlier	outlier
20		1{6,11}	4.305	5.038	-0.733
21		1{7,9}	5.190	5.520	-0.330
22		1{7,11}	5.142	5.105	0.037
23		1{8,8}	7.715	6.843	0.872
24		1{8,10}	5.987	5.599	0.388
25		1{8,11}	5.906	5.890	0.016
26		1{9,9}	4.642	5.047	-0.405
27		6{2}	4.432	4.724	-0.292
28		6{1}	4.235	4.169	0.066
29		7{9}	4.233	3.811	0.422
30		AMD3100	8.688	8.688	0

Table 3. The training set used for the QSAR model building calculations. pEC₅₀ (derived from EC₅₀ in μM) refers to the experimental activity values. The two last columns show the predicted and residual pEC₅₀ values obtained from QSAR model 1. Compound 19 gave a *Z-SCORE* > 2.5, and was therefore considered to be an outlier and was excluded from the training set.

	Compound	Name	pEC ₅₀	predicted pEC ₅₀	Residue
1		8{1,5}	5.079	5.297	-0.218
2		8{2,11}	4.375	5.006	-0.631
3		1{5,8}	6.357	6.045	0.312
4		1{5,9}	5.889	5.336	0.553
5		1{5,11}	5.369	5.239	0.130
6		1{8,9}	7.142	5.783	1.359
7		1{9,11}	4.647	5.101	-0.454
8		7{5}	4.337	4.249	0.088
9		7{8}	5.183	4.962	0.221

Table 4. The external test set used for QSAR model validation. The column headings are described in Table 3.

	pEC ₅₀	VAdjEq	Q_VSA_HYD	dipoleY	SlogP_VSA8	SMR_VSA5	FASA+
pEC ₅₀	100						
VAdjEq	-60	100					
Q_VSA_HYD	84	-68	100				
dipoleY	41	-14	20	100			
SlogP_VSA8	56	-26	41	42	100		
SMR_VSA5	54	-39	37	49	68	100	
FASA+	38	-16	59	-6	-34	-12	100

Table 5. Correlation analysis for the descriptors used in the QSAR models.

Compound	Name	predicted EC ₅₀ (μM)
	1{7,8}	0.66
	10{8}	0.87
	8{2,8}	1.58
	8{2,7}	1.88
	1{4,8}	2.16
	2{1,2}	2.24
	10{7}	2.70
	10{6}	2.87
	1{7,7}	2.93
	8{1,8}	3.70
	9{2}	3.85
	10{9}	4.61
	10{4}	4.75
	10{5}	5.23
	9{1}	5.90
	1{4,7}	5.94
	8{1,7}	5.98
	10{11}	7.30

Compound	Name	predicted EC ₅₀ (μM)
	1{6,6}	7.53
	8{1,4}	9.28
	1{7,10}	9.95
	8{2,10}	10.13
	2{2,3}	10.52
	1{4,10}	17.35
	8{3,7}	18.96
	10{10}	19.70
	8{1,9}	21.50
	9{3}	24.95
	1{10,10}	30.06
	8{1,10}	30.44
	8{3,4}	43.40
	2{1,3}	48.09
	8{3,10}	58.34
	6{3}	> 100

Table 6. Prediction of activity values using QSAR model 1.

Model	H_a	H_t	false +	false -	EF	$Y(\%)$	$A(\%)$	GH
1	142	142	0	9	10.68	100	94	0.99
2	139	140	1	12	10.61	99	92	0.97
3	122	157	35	29	8.30	78	81	0.77
4	133	186	53	18	7.64	72	88	0.73
5	123	168	45	28	7.82	73	81	0.73
6	132	132	0	19	10.68	100	87	0.97
7	96	96	0	55	10.68	100	64	0.91
8	106	107	1	45	10.58	99	70	0.92
9	92	92	0	59	10.68	100	61	0.90

Table 7. Summary of the results obtained for the retrospective screening analysis of the generated pharmacophore models. The quantities H_a , H_t , false +, false -, EF , $Y(\%)$, $A(\%)$, and GH are defined in Equations 5-10. Overall, QSAR model 1 can be seen to give the best statistics.

Compound	Model 1	Model 2	Model 3	Model 4	Model 5
2{1,2}	-	-	-	-	-
2{1,3}	0.9169	0.9086	-	0.9870	-
8{1,4}	0.7352	0.5261	0.5712	0.7153	1.4686
8{1,7}	-	0.5190	0.6307	0.7327	-
8{1,8}	0.6071	0.4723	0.6284	0.9790	-
8{1,9}	0.6714	0.3698	0.5487	0.7998	-
8{1,10}	-	0.5110	0.6307	0.7305	1.2161
2{2,3}	0.7664	0.9129	-	0.8400	-
8{2,7}	-	0.8506	0.6338	0.6678	1.3628
8{2,8}	0.4375	0.4742	0.6199	0.6860	-
8{2,10}	-	0.7040	0.6339	0.6660	0.8185
8{3,4}	0.4417	0.5261	0.5711	0.6690	0.9981
8{3,7}	0.5551	0.4368	0.6306	0.6317	0.8388
8{3,10}	0.5551	0.4289	0.6306	0.6291	0.8327
1{4,7}	0.7565	0.3962	0.5227	0.4716	0.8880
1{4,8}	0.4157	0.3923	0.4166	0.4902	0.8354
1{4,10}	0.7308	0.3284	0.5119	0.5366	0.7563
1{6,6}	0.6027	0.9477	0.4127	-	-
1{7,7}	0.6990	0.5077	0.5190	0.6907	1.0076
1{7,8}	0.4157	0.3625	0.4166	0.5070	0.8043
1{7,10}	0.7308	0.3284	0.5119	0.4728	0.7508
1{10,10}	1.1377	0.4894	0.5079	0.4643	0.7041
6{3}	-	0.9007	-	-	-
9{1}	0.4288	0.4014	0.4803	0.5879	1.1812
9{2}	0.3235	0.3832	0.4742	0.5254	0.8217
9{3}	0.4747	0.4012	0.4805	0.5449	0.5568
10{4}	0.3409	0.4270	0.4640	0.4845	0.6615
10{5}	0.3671	0.3329	0.4250	0.4800	0.7907
10{6}	0.3925	0.4079	0.4250	0.4800	1.1394
10{7}	0.2826	0.3789	0.4512	0.3739	0.5394
10{8}	0.3321	0.4755	0.4248	0.3568	0.5311
10{9}	0.2655	0.3462	0.4248	0.3320	0.5408
10{10}	0.2809	0.3786	0.4512	0.3727	0.5405
10{11}	0.3755	0.2856	0.4248	0.3483	0.5304

Table 8. Pharmacophore-based prospective virtual screening results. This table lists the overall (RMSD) score obtained for each compound using pharmacophore models 1, 2, 3, 4, and 5. Hyphens denote compounds that do not match the pharmacophore model.

Compound	Consensus score	GOLD ChemScore	FRED consensus	AUTODOCK Docked Energy	HEX Docked Energy	FRED Chemgauss3
10{8}	512.70	27.20	446	-16.27	-345.80	-16.97
10{9}	455.30	25.88	408	-17.69	-373.90	-30.96
10{11}	1129.30	8.07	388	-15.76	-396.30	-31.67
10{5}	920.30	15.29	464	-16.59	-343.70	-14.77
9{2}	707	26.42	329	-14.78	-409.50	-29.98
8{2,8}	743.70	35.55	119	-14.59	-374.70	-47.45
1{7,8}	738	31.29	427	-15.14	-407.80	-62.70
8{1,8}	904.30	25.06	212	-14.69	-325.70	-49.30
1{4,10}	922.70	28.29	207	-14.03	-421.60	-46.24
8{2,7}	656	30.80	160	-14.46	-421.20	-57.50
1{4,8}	492.70	33.11	422	-14.92	-443.60	-42.65
1{7,10}	738.30	26.30	341	-14.10	-420.50	-40.59
1{4,7}	683	28.39	185	-15.90	-381.3	-69.14

Table 9. Docking scores of hits from the screened combinatorial library. This table shows compounds found within the top 10% of the ranked hit list using Consensus score (AUTODOCK Docked Energy, GOLD GoldScore and ChemScore), GOLD ChemScore, FRED consensus (PLP, Chemgauss3, Shapegauss, OEChemScore, ScreenScore, ChemScore), AUTODOCK Docked Energy, HEX Docked Energy, and Chemgauss3 scoring functions.

Compound	PARAFIT Shape Tanimoto	PARAFIT Consensus Shape	ROCS Shape Tanimoto	ROCS Combo Score	HEX Shape Tanimoto
10 {11}	0.9725	0.9763	0.5970	0.9010	0.9109
10 {8}	0.9492	0.9506	0.5070	0.6320	0.8479
10 {9}	0.8996	0.9028	0.4820	0.8160	0.8422
10 {5}	0.9193	0.9334	0.5280	0.8720	0.8490
9 {2}	0.9570	0.9507	0.6330	0.6940	0.9145
8 {2,8}	0.9651	0.966	0.5170	0.6840	0.8741
1 {7,8}	0.9597	0.9471	0.5420	0.7320	0.8855
8 {1,8}	0.9393	0.9416	0.5730	0.6070	0.8948
2 {1,2}	0.9230	0.9307	0.5200	0.5580	0.8449
8 {2,7}	0.9101	0.9053	0.4950	0.5480	0.8581
1 {7,10}	0.9498	0.9547	0.4460	0.6240	0.8457
1 {4,7}	0.9015	0.9139	0.4950	0.8100	0.8442

Table 10. Shape matching scores for hits from the screened combinatorial library. This table shows compounds found within the top 10% of the ranked database using PARAFIT Shape Tanimoto and Consensus Shape Tanimoto, ROCS Combo Score and Shape Tanimoto, and HEX Shape Tanimoto scores.

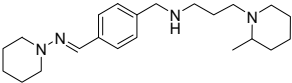
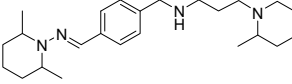
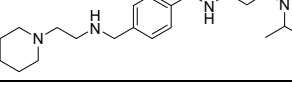
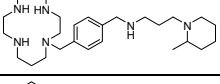
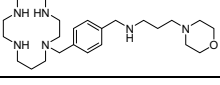
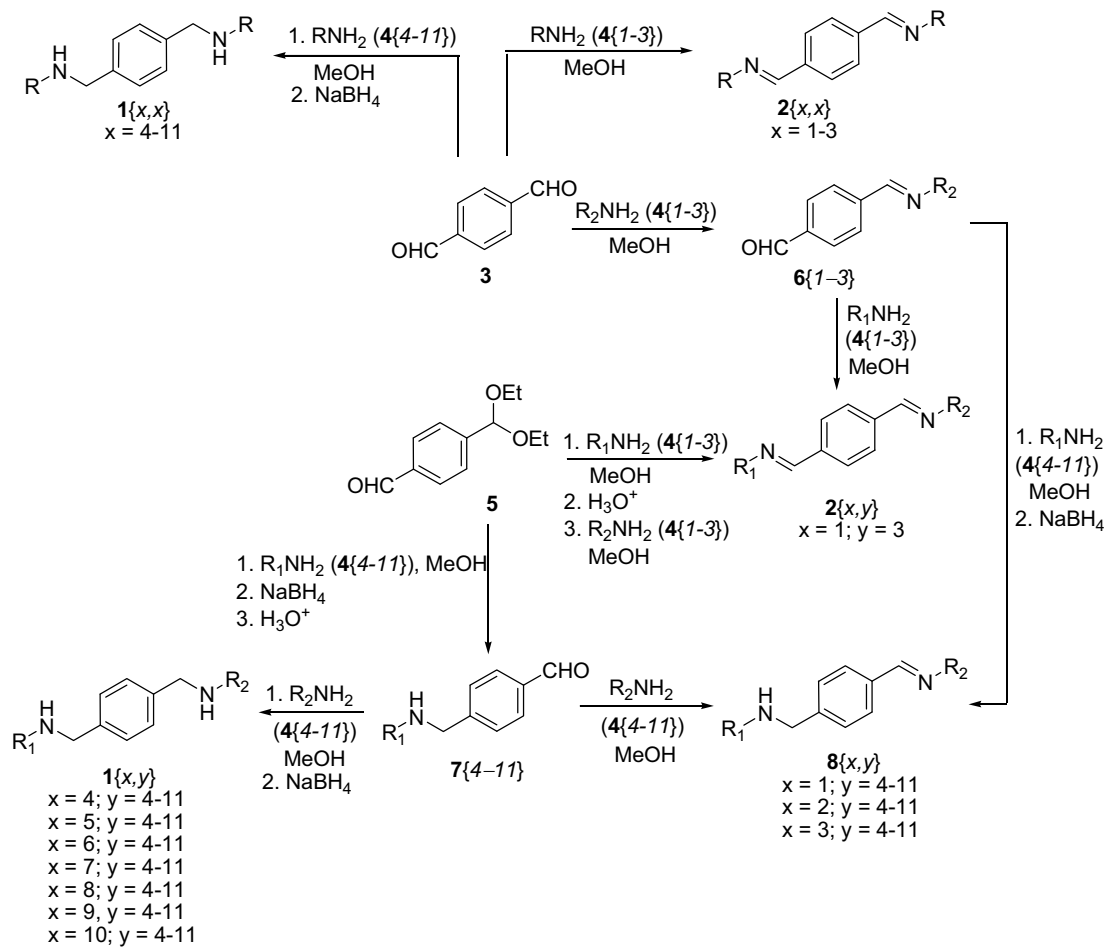
Compound	Name	EC ₅₀ / μg/ml	CC ₅₀ / μg/ml
	8{1,8}	> 4.1	4.1
	8{2,8}	0.6	14.6
	1{7,8}	0.4	> 25
	10{8}	0.022	> 25
	10{11}	0.058	> 25

Table 11. Summary of the five VS-selected hits. EC₅₀ denotes anti-HIV activity, and CC₅₀ is the cytotoxicity value (μM).



Scheme 1. Synthetic scheme for the symmetrical and non-symmetrical diamines **1**, dihydrazones **2**, and aminohydrazones **8**.

The following pages contain NINE Figures for the article.

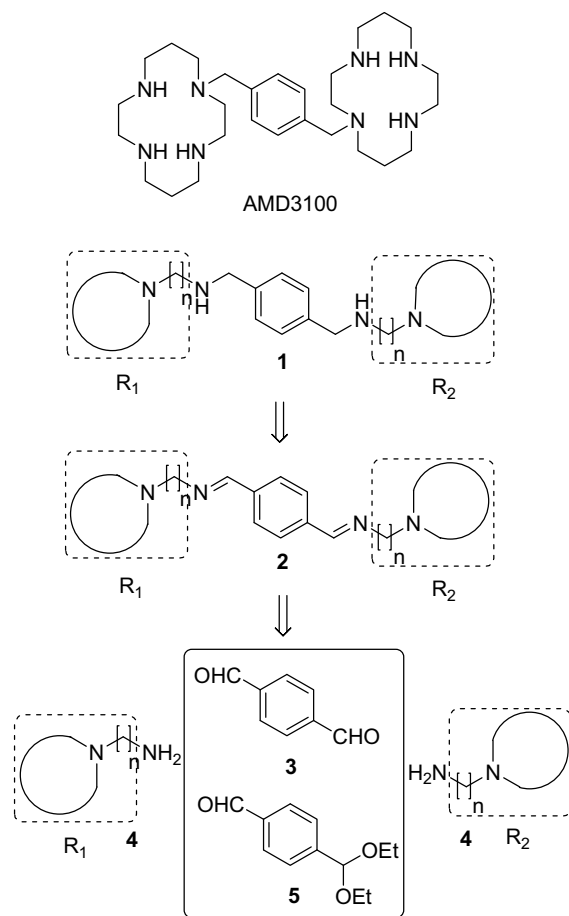


Figure 1. The AMD3100 reference antagonist for CXCR4, and schematic illustration of the target library construction. Top: the AMD3100 reference antagonist for CXCR4, with a *p*-phenyl linker and nitrogen-containing heterocyclic systems on each side of the linker. Bottom: a schematic illustration of the construction of the target library which preserves these features.

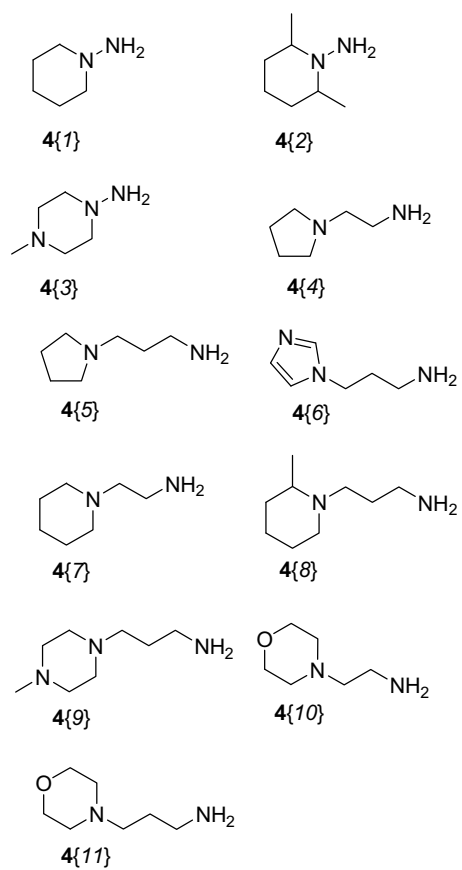


Figure 2. Amine and hydrazone building blocks used for the combinatorial virtual library.

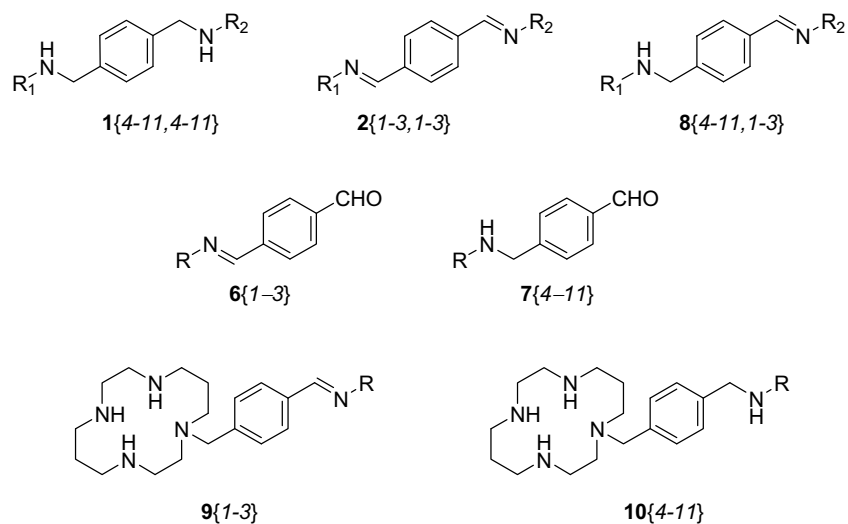


Figure 3. Representative examples of compounds in the combinatorial virtual library. Compounds **1** are symmetrical ($R_1 = R_2$) or non-symmetrical ($R_1 \neq R_2$) diamines, compounds **2** are symmetrical ($R_1 = R_2$) or non-symmetrical ($R_1 \neq R_2$) dihydrazones, compounds **8** are aminohydrazones, compounds **6** and **7** correspond to hydrazonebenzaldehydes and aminobenzaldehydes, respectively, and compounds **9** and **10** are amino or hydrazone substituted monocyclams.

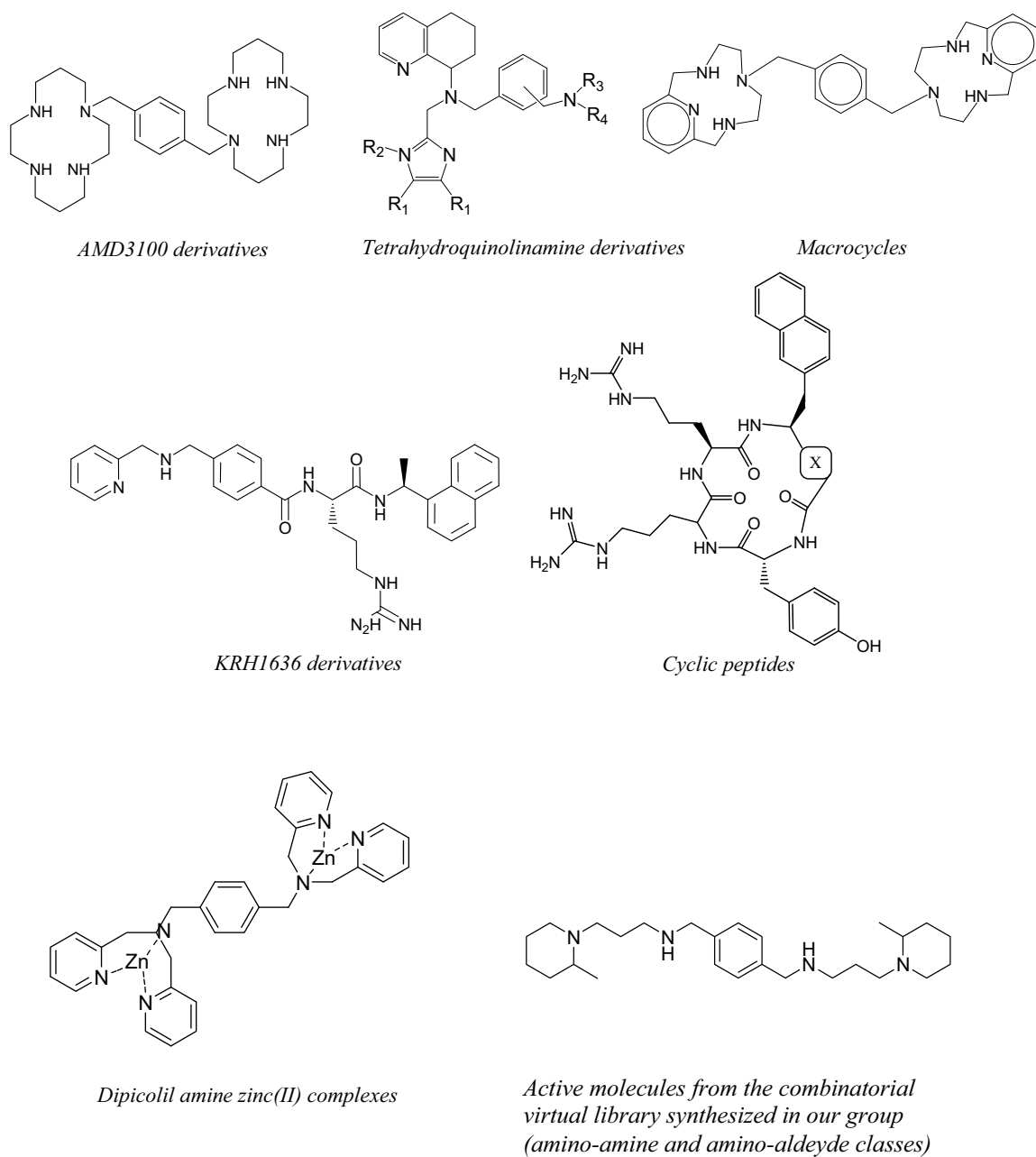


Figure 4. Representative structures of seven families of CXCR4 inhibitor.

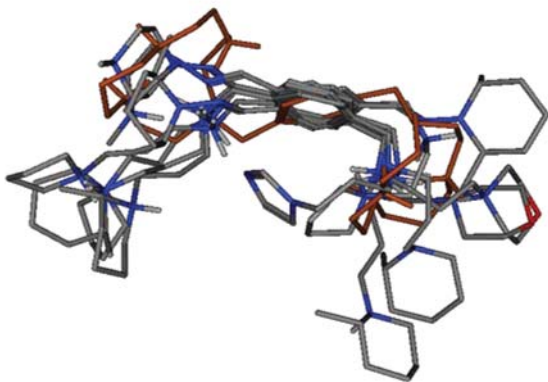


Figure 5. The MOE alignments of active database compounds with AMD3100 (shown in brown).

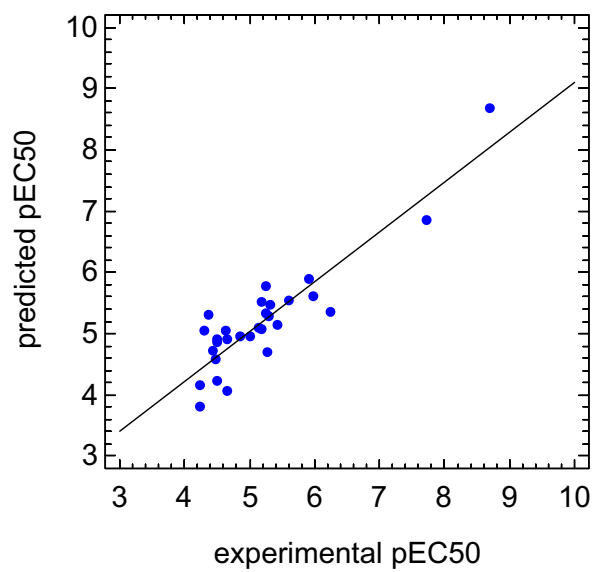


Figure 6. Correlation of experimental *versus* predicted pEC₅₀ for QSAR model 1.

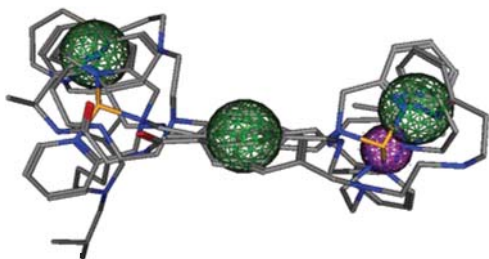


Figure 7. Alignment of the compounds in the training set and the pharmacophore model 1. Hydrophobic and aromatic features (Hyd|Aro) are shown in green. Cationic (Cat) features are shown in purple.

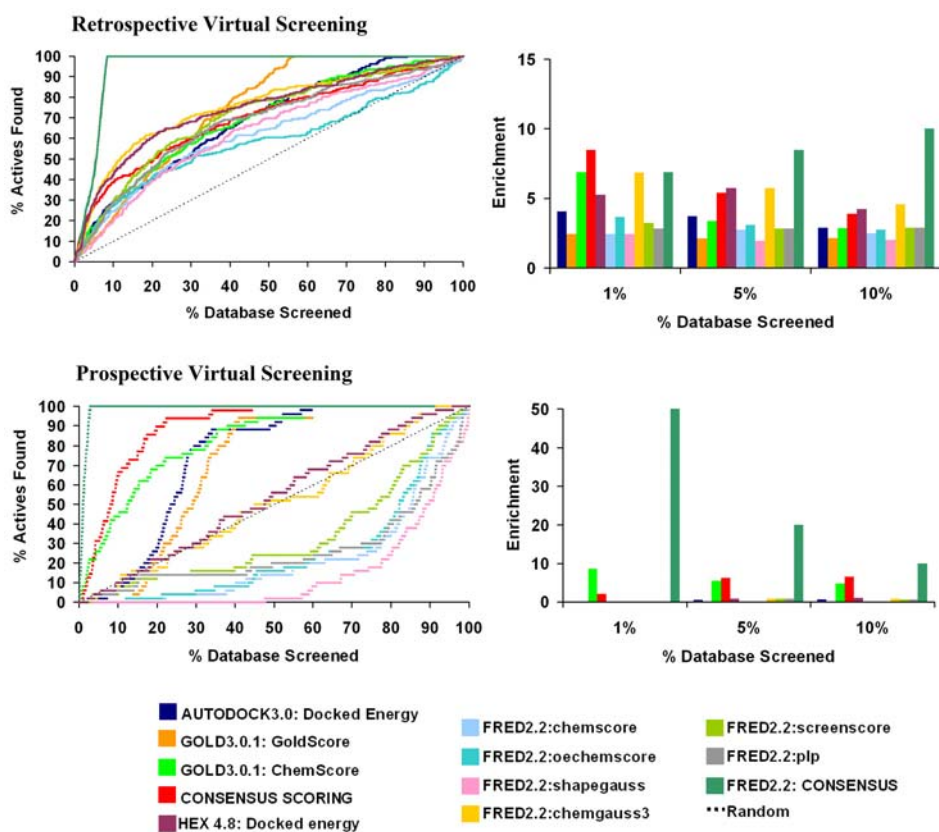


Figure 8. CXC4 docking-based enrichment plots. On the left, enrichment results for several docking protocols for retrospective (top) and prospective (bottom) virtual screening analyses. The dotted black line represents the expected values if actives are selected at random. On the right, enrichment factors for actives found within the top-ranking 1%, 5%, and 10% of the screened inhibitor database (top) and screened virtual combinatorial library (bottom).

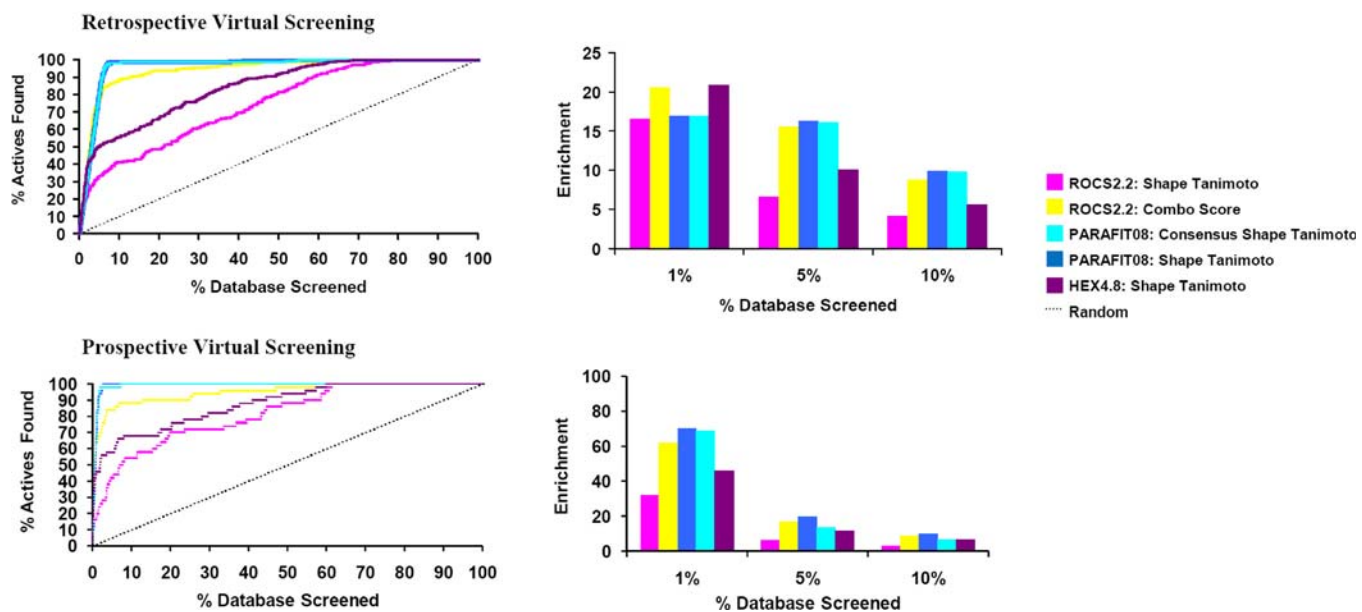
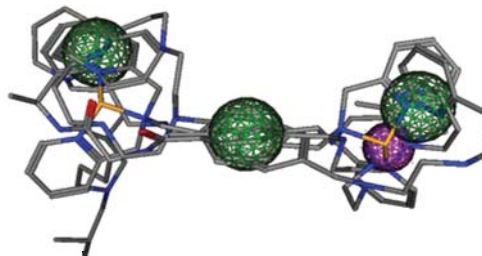


Figure 9. CXCR4 shape matching-based enrichments. On the left, enrichment curves obtained for various shape matching protocols on the known inhibitor database (top) and compounds from the virtual combinatorial library (bottom). The dotted line represents the expected enrichment if actives are selected at random. On the right, enrichment values for actives found within the top-ranking 1%, 5%, and 10% of the screened database (top) and screened virtual combinatorial library (bottom).

For Table of Contents Use Only

Discovery of Novel HIV Entry Inhibitors for the CXCR4 Receptor by Prospective Virtual Screening

Violeta I. Pérez-Nueno, Sofia Pettersson, David W. Ritchie, Jose I. Borrell and Jordi Teixidó*



Caption for Proposed Front Cover Illustration Only (figure provided separately)

This image shows at the top left the proposed MOE pharmacophore model built from the alignment of the four most active structures of four CXCR4 inhibitor families, namely: AMD3100 derivatives, KRH1636 derivatives, dipicolil amine zinc(II) complexes, and the most active CXCR4 inhibitor synthesised from a combinatorial virtual library built by the authors. A polarity-charge-hydrophobicity pharmacophore modelling scheme was used. Chemical features and their tolerance radii were selected between those suggested by MOE to achieve better balance between sensitivity and specificity. Hydrophobic and aromatic features are shown in green. Cationic features are shown in purple. On the bottom left, the MOE FlexAlign alignments of active database compounds found using as superposition template the AMD3100 conformation obtained previously from a CXCR4 docking study (shown in brown). On the right, a depiction of how the selected database compounds fit the calculated pharmacophore model.

Artículo IV

APIF: A New Interaction Fingerprint Based on Atom Pairs and its Application to Virtual Screening

*Violeta I. Pérez-Nueno, Obdulia Rabal, José I. Borrell and Jordi Teixidó**

Grup d'Enginyeria Molecular, Institut Químic de Sarrià (IQS), Universitat Ramon Llull,
Via Augusta 390, 08017 Barcelona, Spain.

Tel: +34-93-267.20.00. Fax: +34-93-205.62.66. E-mail: j.teixido@iqs.url.es

ABSTRACT

A new interaction fingerprint (IF) called APIF (Atom Pairs based Interaction Fingerprint) has been developed for post-processing protein-ligand docking results. Unlike other existing fingerprints which employ absolute locations of individual interactions, APIF considers the relative positions of pairs of interacting atoms. Docking-based virtual screening was performed with GOLD using the crystal structures of trypsin, rhinovirus, HIV protease, carboxypeptidase and estrogen receptor-alpha (ER- α) as targets. A score derived from the similarity of the bit strings for each docking solution to that of a known reference binding mode was obtained. Comparisons between APIF, GoldScore function, and a standard interaction fingerprint (CHIF) scores were performed using enrichment plots. Superior recovery rates were observed in the interaction fingerprints (IFs) score cases. Comparable results were achieved by using either of the two interaction fingerprints, substantially improving GoldScore function enrichment factors. Binding mode analyses were also carried out in order to study the best method for selecting conformations with a binding mode similar to that of the reference crystallized complex. These showed that the first conformations retrieved by interaction fingerprint scores had a more similar binding mode to the reference complex than those retrieved by GoldScore function.

INTRODUCTION

Interaction fingerprints (IFs) have been developed to enhance the representation and analysis of three-dimensional protein-ligand interactions^{1,2,3}. In particular, they have proven to be very useful in docking output post-processing as a virtual screening (VS) filter and for binding mode detection⁴. These methods have been developed in order to overcome the known deficiencies in identifying accurately the conformations with closest binding modes to the X-ray structures^{5,6}.

IFs encode the 3D protein-ligand contacts in bit strings of a length derived from the number of residues/atoms in the target protein binding cavity. Typically, each bit denotes either the presence (1) or absence (0) of a particular interaction such as a hydrogen bond, or hydrophobic or van der Waals contact.

The different interaction fingerprint implementations vary depending on the bit string definition and the type of interactions considered. The initial proposal of Deng and co-workers^{1,4} operated at the residue level and considered hydrophobic and hydrogen bond contacts. Following this idea, Kelly and Mancera² transferred the initial concept based on residues to a new one based on atoms for hydrogen bond sites. Moreover, these authors introduced the concept of weighting the importance of the detected interactions.

Recent atom-based IF advances were developed by Mpamhanga and Willet³. These authors encoded hydrogen bonds or/and hydrophobic contacts in a fingerprint of length equal to the number of heavy atoms in the binding site.

In this paper we present a new fingerprint called APIF (Atom Pairs based Interaction Fingerprint) that encodes ligand-protein binding modes in a bit string based on the concept of atom pairs. This approach is widely used in the context of fragment-based similarity searches⁷. APIF encodes ranges of distances between two receptor-ligand interaction points. Each observed distance increases a count in an associated 7-range bin. Depending on the combination of the type of contacts, the corresponding bit is set on.

The three IF approaches previously reported encode ligand-protein interaction information in an absolute manner, i.e., a contact is expected or not at a concrete atom/residue of the protein sequence. On the other hand, APIF considers the relative pairwise position of the interacting atoms rather than their absolute locations. Thus, from our viewpoint, the main novelty of this approach is that our IF encodes the conserved distance between two receptor-ligand interactions rather than requiring a specific atom/residue of the protein.

The performance of this new fingerprint was validated through docking-based VS using both enrichment plots and binding mode analyses. Enrichment results obtained with APIF were compared with those retrieved with the GoldScore function and an in-house implementation of the CHIF fingerprint of Mpamhanga's *et al.* Inspection of the binding modes for the poses selected by these three criteria was carried out in order to analyze their ability to retrieve the closest binding modes to the crystallographic structures within the first top-ranked conformations.

METHODS

Case Studies: Protein and Databases Preparation

To evaluate our approach, we decided to dock several different experimentally determined X-ray co-crystal structures and known inhibitors. Thus, the set of known inhibitors was collected from FlexS-77 dataset⁸ for the trypsin, rhinovirus, HIV protease, and carboxypeptidase targets. The "Bissantz active set"⁹ was used to compile the inhibitors for estrogen receptor-alpha (ER- α) target for comparison purposes with Mpamhanga's work³. The structures of these compounds are shown in Figures 1, 2 and 3. For each target of these sets, the complex with the best crystallographic resolution was selected as the reference for docking. These reference complexes and their corresponding PDB entries are listed on Table 1. For each

target complex, the ligand binding site was defined from the bound ligand using a radius cut-off of 10 Å. Bound waters were removed from the binding sites, and the receptors were protonated at pH 7.

Figure 1

Figure 2

Figure 3

Table 1

In order to perform the virtual screening, the known actives were combined with presumed inactive compounds from the Maybridge Screening Collection database⁹ in such a way that several 1D properties calculated by MOE¹¹ were similar to those of the active compounds (molecular weight, number of rotatable single bonds, number of hydrogen-bond acceptor atoms, number of hydrogen-bond donor atoms, octanol-water partition coefficient and number of hydrophobic atoms)¹². Table 2 shows the average and standard deviations of these properties for the datasets used. It can be seen that they are quite similar for the active and inactive pools.

We would like to remark that for APIF and CHIF (in-house implementation) comparison purposes the ER- α inactive pool of 490 compounds differs from that of Mpamhanga *et al.*³. However, it is worth mentioning that Mpamhanga *et al.* repeated their calculation for three different inactive pools without finding significant differences, so it is not expected that our modified inactive pool will have much influence in reproducing results.

Table 2

Docking Methodology

All the dataset compounds were docked into the aforementioned protein structures using GOLD¹³. In the GOLD runs, the ligand binding site was limited to all protein atoms within 10 Å from the centroid of the binding residues^{14 - 26}. The GOLD cavity-detection algorithm was enabled in order to confine the calculation to concave regions in the vicinity of the binding site. A total of 100 docking runs per experiment (conformations) were performed, with each run consisting of a maximum of 100,000 GA operations. All other GA parameters used default values. Cut-off distances of 2.5 Å for hydrogen-bonds and 4.0 Å for non-bonded contacts were set. In each study all the ligand poses generated were retained for subsequent binding mode analyses. The GoldScore function was used for scoring the docked conformations as the first criterion for VS ranking and for the subsequent respective enrichment plots.

Construction of APIF (Atom Pairs based Interaction Fingerprint)

The algorithm to generate the APIF was implemented in the MOE SVL language ¹¹. First: given a complex, the active site is defined using a radius value (10 Å in the present study). Second: the interactions between the protein and the ligand are detected using the function `pro_Contacts` as implemented in MOE: hydrogen bonds are defined following Stickle446 function and hydrophobic contacts are determined using a cut-off of 4.5 Å. Depending on the type of interaction, both the atoms of the protein (P) and the ligand (L) are labelled as hydrogen bond donor, hydrogen bond acceptor or hydrophobic. This results in six possible types: acceptor-L, acceptor-P, donor-L, donor-P, hydrophobic-L, and hydrophobic-P. Third: all possible pairwise protein-ligand interactions are detected and classified depending on one of the six possible combinations of pairs of interaction contacts. The six possible types are listed in Figure 4a. For each pairwise interaction detected in a complex, the distance between the two receptor atoms (d_1) and the distance between the two ligand atoms (d_2) are measured. Figure 4b shows this process. Each observed distance increases a counter within a *bin* divided into seven ranges, taken from Mason ²⁷, which correspond to distance ranges of (Å): [0-2.5], [2.5-4], [4-6],[6-9], [9-13], [13-18] and [>18]. The two distances taken together define a single bit in a string of 49 bits (enumerated from zero) according to Equation 1.

$$\text{Bit Position} = \text{bin}(d_2) + 7 \cdot \text{bin}(d_1) \quad \text{Equation 1}$$

Fourth: the final fingerprint length corresponds to the number of possible combinations of pairs of interaction contacts (six) and the dimension of the binning partition scheme for the ligand distances (seven) and the receptor distances (seven). In this way, the total fingerprint is composed of $6 \times 7 \times 7 = 294$ bits. Figure 5 illustrates this encoding system. Both raw fingerprints and normalized ones between 0 and 1 were constructed.

Figure 4

Figure 5

In-house CHIF Fingerprint Implementation

Here, the CHIF fingerprint was also implemented in SVL. We followed the CHIF design description from Mpamhanga *et al.*³, although some differences arise in the function used to determine protein contacts. In our case, and as for APIF, the MOE pro_Contacts function was used. This function uses a different hydrogen bond distance threshold and different atom type definitions (donor, acceptor, hydrophobic) from those of Mpamhanga *et al.*³. We also fixed a radius of 10 Å to define the binding site.

Virtual Screening Protocol

After docking the inhibitors against their corresponding target, CHIF and APIF fingerprints were calculated for all the retrieved poses. Similarly, a reference CHIF and APIF fingerprints were directly generated from the crystallographic reference complexes. Then, a similarity search was performed between the fingerprints derived from each docked conformation and the reference fingerprints. Two scoring systems were used to evaluate the conformations for each ligand and to rank the screened list:

- Traditional similarity coefficients²⁸: Euclidean distance, Manhattan distance, Tanimoto coefficient and Simple matching coefficient. These similarity values can be calculated using Equations 2-5 (below), where A and B denote the numbers of bit sets in the two IFs that are being compared and C denotes the number of bits in common. This scoring system will subsequently be called SCORE1, specifying in each case the particular coefficient used (Euclidean, Manhattan, Tanimoto or Simple matching).

$$\text{Euclidean Distance} = \sqrt{A + B - 2C} \quad \text{Equation 2}$$

$$\text{Manhattan Distance} = A + B - 2C \quad \text{Equation 3}$$

$$\text{Tanimoto coefficient} = \frac{C}{A + B - C} \quad \text{Equation 4}$$

$$\text{Simple matching coefficient} = C \quad \text{Equation 5}$$

- Following Mpamhanga's work³, a second kind of score was calculated resulting from the multiplication of the value obtained from the GoldScore function for each solution by the

similarity coefficient (in this case we only considered Tanimoto and Simple matching). This scoring system will subsequently be called SCORE2, specifying in each case the particular coefficient used (Tanimoto or Simple matching).

$$\text{SCORE2} = \text{SCORE1} \times \text{GoldScore function} \quad \text{Equation 6}$$

Finally, the VS was analyzed in terms of enrichment plots using the three criteria: GoldScore function, SCORE1 and SCORE2.

Binding Mode Analyses

IFs provide a good method for analyzing the protein-ligand interactions and optimizing the resulting docking poses. Several studies have been made for analyzing the binding modes obtained from IFs selected conformations^{1, 2, 3, 4, 29, 30}. As many docking validation studies have shown, scoring functions (such as GoldScore) do not always identify within the first ranked conformations the co-crystallized binding modes. This also happens with the poses selected using IF-based similarity scores. However, since IFs take into account experimental data, it is reasonable to suppose that they can select closer poses to the experimental crystallographic complex than using only docking scoring functions. In this work we compared the ability of GoldScore function, APIF-based and CHIF-based similarity fingerprints to retrieve the closest binding modes to the crystallographic structures within the first top-ranked conformations. Binding modes of four out of the five targets used in the enrichment studies (trypsin, rhinovirus, HIV protease and carboxypeptidase) were analyzed. For each target, the RMSD between the crystallographic reference complex and each docked pose was calculated. Results were analyzed using three different plots:

- *Graph 1 (RMSD from crystallographic binding mode vs CHIF-based Tanimoto or APIF-based Tanimoto similarity coefficient)*. Plotting RMS deviations from the X-ray pose versus similarity of IFs expressed by a Tanimoto coefficient, calculated from APIF and CHIF fingerprints, generated by the X-ray and the predicted docking pose.
- *Graph 2 (RMSD from crystallographic binding mode vs GoldScore rank or APIF-based Tanimoto rank or CHIF-based Tanimoto rank)*. Plotting RMS deviations from X-ray pose versus

ranked active ligand docked conformations according to GoldScore function, or APIF-based and CHIF-based similarity Tanimoto scores.

- *Graph 3 (% cases predicted within 2 Å RMSD vs binding mode rank)*. A comparison of the effects of using GoldScore function, and APIF and CHIF IFs to postprocess the docking-generated poses on the likelihood of identifying the crystallographic binding mode within the active docked conformations obtained.

RESULTS AND DISCUSSION

Performance of APIF in Virtual Screening: Database Enrichment

Here we present the VS enrichment plots for the ER- α , trypsin, rhinovirus, HIV protease and carboxypeptidase targets calculated using the GoldScore function, and the CHIF, APIF and normalized APIF based similarity criteria. For the similarity, both the SCORE1 and SCORE2 metrics were used. In order to enhance the first part of the plots, the x axis uses a logarithmic scale for the percent of the database screened plotted against the percent recovery of known active compounds. We also list the enrichment factor values (EFs) for the first 2%, 5% and 10% screened databases.

ER- α recovery plots (Figure 6) show that our in-house CHIF implementation (Figures 6a and 6b) reproduces the results reported for the same case study analyzed in Mpamhanga's work³. Therefore, we have validated our CHIF implementation in spite of the previously mentioned differences determining protein contacts. Regarding the APIF (Figures 6c and 6d) and normalized APIF (Figures 6e and 6f) results, although APIF is able to retrieve compounds over a random selection, the enrichment obtained with the similarity scores is lower than the enrichment achieved using the GoldScore function (Figures 6c and 6e). The combination of the similarity and energetic criteria (SCORE2) achieves higher performance than that the obtained only with the energetic criterion (Figures 6d and 6f). However, even considering SCORE2, APIF does not achieve the high performance achieved by CHIF in the first 1-2% of screened database, although it does at higher percentages. Regarding normalization, no well defined tendency can be found. Whereas for SCORE1 the normalization gives worse results (Figures 6c and 6e), for SCORE2 it has a positive effect (Figures 6d and 6f).

Figure 6

For ER- α , the EFs for the first 2%, 5% and 10% screened database are shown in Table 3. The maximum theoretical value for the EF is 50 (500/10). The maximum value found for each percentage is shown in

bold. It can be observed that CHIF with SCORE2 gives the optimum result. APIF and normalized APIF with SCORE1 are not able to discriminate between active and inactive compounds better than the docking energetic criterion (GoldScore), although its behaviour improves in combination with energetic criterion.

Table 3

Results for trypsin (Figure 7) show that docking does not perform so well in this case. The enrichment obtained with the similarity scores is higher than the enrichment calculated using GoldScore for APIF (Figures 7c, 7d, 7e and 7f), which achieves higher performance than CHIF (Figures 7a and 7b). In both cases, the combination of the similarity and energetic criteria (SCORE2) achieves higher performance than that obtained only with the energetic criterion (Figures 7b, 7d and 7f). APIF normalization does not improve the results in this case (Figures 7e and 7f). Table 4 shows the EFs obtained at the first percentages of database screened. The maximum theoretical value for the EF is 67.7 (474/7). The maximum value found for each percentage is shown in bold.

Figure 7

Table 4

Results for rhinovirus (Figure 8) show that the enrichments given by the APIF-based and CHIF-based similarity scores are higher than those obtained by docking. The CHIF fingerprint achieves higher performance than APIF in all SCORE1 metrics (Figures 8a, 8c and 8e) and simple matching SCORE2 (Figure 8b). APIF achieves higher performance than CHIF in Tanimoto SCORE2 (Figure 8d) and normalized APIF Tanimoto SCORE1 and SCORE2 (Figures 8e and 8f). Table 5 shows in detail some EF values. In this case, the maximum theoretical EF is 63.2 (506/8).

Figure 8

Table 5

Results for HIV protease (Figure 9) show that the enrichments obtained using GoldScore are similar to those from the APIF-based similarity scores (Figures 9c, 9d, 9e and 9f) and higher than those obtained from the CHIF-based similarity scores (Figures 9a and 9b). APIF (Figures 9c and 9d) achieves higher performance than CHIF (Figures 9a and 9b). In both cases Simple matching score gives the best results. In this case, APIF normalization does not improve the APIF results for both SCORE1 and SCORE2 (Figures 9e and 9f). Table 6 shows some EF values. The maximum theoretical value for the EF is 49.9 (499/10).

Figure 9

Table 6

Results for carboxypeptidase (Figure 10) show that the enrichment obtained using GoldScore is similar to the enrichment performed by similarity coefficients. CHIF (Figures 10a and 10b) achieves higher performance than APIF (Figures 10c and 10d). CHIF-based Manhattan and Euclidean SCORE1 and CHIF-based Tanimoto and Simple matching SCORE2 perform better than GoldScore function (Figures 10a and 10b). Regarding the combination of the similarity and energetic criteria (SCORE2), Simple matching gives the best results in all cases. Regarding normalization, the normalized APIF gives similar results for both SCORE1 and SCORE2 than APIF, except for Simple matching score (Figures 10e and 10f), which improves results. Table 7 shows the enrichment values for the first percentages of database screened. The maximum theoretical value for the EF is 55.4 (277/5).

Figure 10

Table 7

Finally, we show the correlation diagram of APIF fingerprint for trypsin, rhinovirus, HIV protease and carboxypeptidase reference complexes (Figure 11). The number of contacts found for each target and the type of protein-ligand interactions are shown: hydrophobic hydrophobic (*HYD HYD*), hydrophobic acceptor (*HYD Acceptor*), hydrophobic donor (*HYD Donor*), donor donor (*Donor Donor*), acceptor acceptor (*Acceptor Acceptor*) or donor acceptor (*Donor Acceptor*).

Figure 11

Summarizing database enrichment analyses, CHIF obtains the best EF values for ER- α , rhinovirus, and carboxypeptidase. For trypsin and HIV protease, APIF achieves the best EFs. Moreover, rhinovirus and carboxypeptidase APIF SCORE2 and normalized APIF SCORE2, respectively, improve CHIF EF results. Furthermore, the APIF Tanimoto and Euclidean similarity scores always return good enrichments even though they do not always achieve the best results. Generally, the combination of the similarity and energetic criteria (SCORE2) achieves higher enrichments than those obtained only with the energetic criterion, except for failed docked conformations or bad scoring function behavior (Tables 3 to 7).

APIF Recognition of The Binding Mode

Here we present the binding mode analyses for trypsin, rhinovirus, HIV protease and carboxypeptidase first-ranked conformations according to the previously described criteria (Figures 12, 13, 14, and 15). For

each target, the docked conformations corresponding to the reference crystallographic ligand are shown in pink colour, whereas those corresponding to the rest of active compounds of the set are shown in blue.

Results for the trypsin target (Figure 12) show a non-well-defined tendency to associate high Tanimoto scores with low RMSD values from the crystal structure (Figure 12a), although this tendency is clearer for the conformations corresponding to the complexed ligand (PDB code: 3PTB). It can be seen that docking performs randomly (broad range of RMSD values), but rather well for some conformations (RMSD < 2 Å). Moreover, IFs capture the basic interactions for the lowest RMSD conformations (Tanimoto score values = 1 for 3PTB conformations). Lower RMSD conformations from the crystallographic binding mode are found in the top CHIF and APIF hitlist ranking positions, especially for CHIF top ranked conformations (Figure 12b). Ligand conformations corresponding to the complexed compound are found in the first ranking positions for CHIF and APIF ranking lists and in the last positions for GoldScore function (Figure 12b). The first 35 CHIF and APIF top-ranked ligand conformations have lower RMSD from the crystallographic binding mode than the first top ranked GoldScore conformations. For the subsequent ranked conformations APIF-based Tanimoto score and GoldScore function give better results (Figure 12c).

Figure 12

Results for the rhinovirus target (Figure 13) show no tendency to associate high Tanimoto scores with low RMSD values from the crystal structure because multiple conformations with the same protein interacting points but different binding mode are found (Figure 14). Two groups of ligand conformations are found, one with low RMSD from crystallographic binding mode (between 0 and 2 Å), and the other with higher RMSD values (between 11 and 14 Å). Moreover, the conformations corresponding to the complexed ligand (PDB code: 2R04) show high Tanimoto score values (Figure 13a), but not exceeding 0.8. These conformations with RMSD < 2 Å do not achieve Tanimoto score values of 1 due to the fact that a hydrophobic interaction present in the crystal reference complex is changed to hydrogen bond contact, and a new hydrophobic interaction is created between the ligand and a neighboring residue to the crystallographic interacting one. Lower RMSD from the crystallographic binding mode ranked conformations alternate with higher RMSD ranked conformations, according to the two binding modes found (Figure 13b). Ligand conformations corresponding to the complexed compound are found in the first ranking positions for CHIF and APIF ranking lists and in the last positions for GoldScore function

(Figure 13b). The first top ranked CHIF and APIF conformations show lower RMSD from the crystallographic binding mode than the first ranked GoldScore function conformations (Figure 13c).

Figure 13

Figure 14

Results for the HIV protease target (Figure 15) show that the docking procedure performs poorly in this case. The RMSD values obtained are generally high, i.e. over 4 Å from the crystallographic binding mode, and the similarity between the docked conformations and the reference complex is always lower than 0.5 (Figure 15a). Both for GoldScore function and for CHIF and APIF IFs, the ligand conformations corresponding to the complexed compound (PDB code: 4PHV) are found randomly along the ranking lists (Figure 15b). Given that the docking procedure cannot find the experimental binding mode, the IFs calculation from the docked poses achieves poor results too (Figure 15c).

Figure 15

Results for the carboxypeptidase target (Figure 16) show that docking performs well (high number of conformations within RMSD < 2 Å). Higher Tanimoto similarity scores correspond to lower RMSD from the crystal structure values (Figure 16a). Moreover, this tendency is emphasized for the ligand conformations corresponding to the complexed compound (PDB code: 2CTC). Ligand conformations with the lowest RMSD from the crystallographic binding mode are found in the first IF ranking positions. However, they are found in the last positions for GoldScore (Figure 16b). The first top-ranking IF ligand conformations show lower RMSD from the crystallographic binding mode than the first GoldScore ranked conformations (Figure 16c).

Figure 16

In order to visualize binding modes, a PCA analysis was performed. Figure 17 shows three-dimensional PCA plots for trypsin, rhinovirus, HIV protease and carboxypeptidase complexes. These plots show that active molecules are located in a different region than the inactive compounds. Corroborating the above-mentioned binding mode results, trypsin and carboxypeptidase seem to best recognize the binding mode closest to the crystallographic structures for the first top-ranked conformations, restricting active molecules to a specific region of space far from inactive compounds in the PCA plot.

Figure 17

Summarizing binding mode analyses, carboxypeptidase and trypsin show the best tendency to associate high Tanimoto scores with low RMSD values from the crystal structure. Rhinovirus and HIV protease do

not follow this tendency because docking is not able to find the correct binding mode conformations. However, all results show that the lowest RMSD conformations are found in the first CHIF and APIF top-ranked positions and in the subsequent ranked GoldScore positions (Figures 12c, 13c, 14c, 15c). Therefore, IFs provide a better method for identifying low RMSD conformations from the crystallographic binding mode than only using a docking scoring function.

CONCLUSION

The analyses in this study indicate that our new interaction fingerprint (APIF) yields satisfactory results, often comparable to our CHIF implementation, and it improves the GoldScore results, inasmuch as our enrichment plots exhibit good recognition of the known actives. Overall, this study shows that APIF has proven to be suitable for ranking and filtering virtual screening docking results. However, the quality of the EFs obtained by APIF scoring strongly depends on docking success. Our results show that if docking is successful, as in the trypsin and carboxypeptidase cases, then APIF scoring retrieves good enrichments, substantially improving the results obtained when using only a docking scoring function. Using APIF is thus a good way to select poses or virtual hits that satisfy a defined ligand-protein interaction reference, which will be useful for receptor-based prospective virtual screening.

ACKNOWLEDGEMENTS

We thank Dave Ritchie for proof-reading the manuscript. VIPN thanks the Generalitat de Catalunya – DURSI for a grant within the Formació de Personal Investigador (2008FI) Program. This work was supported by The TV3 Marathon Foundation (AIDS-2001) promoted by the Catalan Radio and Television Corporation (Corporació Catalana de Ràdio i Televisió, CCRTV) and the Programa Nacional de Biomedicina (Ministerio de Educación y Ciencia, SAF2007-63622-C02-01).

REFERENCES

1. Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337-344.
2. Kelly, M. D.; Mancera, R. L. Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1942-1951.
3. Mpamhanga, C. P.; Chen, B.; McLay, I. M.; Willett, P. Knowledge-based interaction fingerprint scoring: a simple method for improving the effectiveness of fast scoring functions. *J. Chem. Inf. Model.* **2006**, *46*, 686-698.
4. Chuaqui, C.; Deng, Z.; Singh, J. Interaction Profiles of Protein Kinase-Inhibitor Complexes and Their Application to Virtual Screening. *J. Med. Chem.* **2005**, *48*, 121-133.
5. Warren, G. L.; Andrews, C. V.; Capelli, A.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912-5931.
6. Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A Review of Protein-Small Molecule Docking Methods. *J. Comput.-Aided. Mol. Des.* **2002**, *16*, 151-166.
7. Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definitions and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64-73.
8. *FlexS-77 dataset* collected by C. Lemmen / G. Klebe / M. Böhm, first published in: Lemmen, C.; Lengauer, T.; Klebe, G. FlexS: A Method for Fast Flexible Ligand Superposition. *J. Med. Chem.* **1998**, *41*, 4502-4520.
9. Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43*, 4759-4767.
10. *Maybride Bringing life to drug discovery™*, Maybride Databases Autumn 2005; Fisher Scientific International: England, 2005.
11. *MOE (Molecular Operating Environment)*, 2006.08 Release; Chemical Computing Group, Inc.: Montreal, Canada, 2004.
12. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793-806.
13. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 609-623.
14. Marquart, M.; Walter, J.; Deisenhofer, J.; Bode, W.; Huber, R. The Geometry of the Reactive Site and of the Peptide Groups in Trypsin, Trypsinogen and its Complexes with Inhibitors. *Acta Crystallogr., Sect. B* **1983**, *39*, 480-490.
15. Renatus, R.; Bode, W.; Huber, R.; Stürzebecher, J.; Stubbs, M. T. Structural and Functional Analyses of Benzamidine-Based Inhibitors in Complex with Trypsin: Implications for the Inhibition of Factor Xa, tPA, and Urokinase. *J. Med. Chem.* **1998**, *41*, 5445-5456.
16. Böhm, M.; Stürzebecher, J.; Klebe, G. Three-Dimensional Quantitative Structure-Activity Relationship Analyses Using Comparative Molecular Field Analysis and Comparative Molecular Similarity Indices Analysis To Elucidate Selectivity Differences of Inhibitors Binding to Trypsin, Thrombin, and Factor Xa. *J. Med. Chem.* **1999**, *42*, 458-477.
17. Badger, J.; Minor, I.; Oliveira, M.A.; Smith, T.J.; Rossmann, M.G.; Structural analysis of antiviral agents that interact with the capsid of human rhinoviruses. *Proteins* **1989**, *6*, 1-19.
18. Matthews, D. A.; Dragovich, P. S.; Webber, S. E.; Fuhrman, S. A.; Patick, A. K.; Zalman, L. S.; Hendrickson, T. F.; Love, R. A.; Prins, T. J.; Marakovits, J. T.; Zhou, R.; Tikhe, J.; Ford, C. E.; Meador, J. W.; Ferre, R. A.; Brown, E. L.; Binford, S. L.; Brothers, M. A.; DeLisle, D. M.; Worland, S. T. Structure-assisted design of mechanism-based irreversible inhibitors of human rhinovirus 3C protease with potent antiviral activity against multiple rhinovirus serotypes. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 11000-11007.

19. Bone, R.; Vacca, J.P.; Anderson, P.S.; Holloway, M.K. X-Ray Crystal Structure of the HIV Protease Complex with L-700,417, an Inhibitor with Pseudo C2 Symmetry. *J. Am. Chem. Soc.* **1991**, *113*, 9382-9384.
20. Specker, E.; Böttcher, J.; Brass, S.; Heine, A.; Lilie, H.; Schoop, A.; Müller, G.; Griebenow, N.; Klebe, G. Unexpected Novel Binding Mode of Pyrrolidine-Based Aspartyl Protease Inhibitors: Design, Synthesis and Crystal Structure in Complex with HIV Protease. *ChemMedChem* **2006**, *1*, 106 – 117.
21. Specker, E.; Böttcher, J.; Lilie, H.; Heine, A.; Schoop, A.; Müller, G.; Griebenow, N.; Klebe, G. An Old Target Revisited: Two New Privileged Skeletons and an Unexpected Binding Mode For HIV-Protease Inhibitors. *Angew. Chem. Int. Ed.* **2005**, *44*, 3140 –3144.
22. Teplyakov, A.; Wilson, K.S.; Orioli, P.; Mangani, S. High-resolution structure of the complex between carboxypeptidase A and L-phenyl lactate. *Acta Crystallogr. Sect., D* **1993**, *49*, 534-540.
23. Rees, D. C.; Lipscomb, W. N. Binding of ligands to the active site of carboxypeptidase A. *Proc. Natl Acad. Sci. USA* **1981**, *78*, 5455-5459.
24. Kim, H.; Lipscomb, W. N. Crystal Structure of the Complex of Carboxypeptidase A with a Strongly Bound Phosphonate in a New Crystalline Form: Comparison with Structures of Other Complexes. *Biochemistry* **1990**, *29*, 5546-5555.
25. Christianson, D. W.; Lipscomb, W. N. Binding of a possible transition state analogue to the active site of carboxypeptidase A. *Proc. Natl. Acad. Sci. USA* **1985**, *82*, 6840-6844.
26. Shiau, A.K.; Barstad, D.; Loria, P.M.; Cheng, L.; Kushner, P.J.; Agard, D.A.; Greene, G.L. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **1998**, *95*, 927-937.
27. Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251-3264.
28. Willet, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 900-908.
29. Hert, J.; Willet, P.; Wilton, D. J.; Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177-1185.
30. Marcou, G.; Rognan, D.; Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195-207.

Target	Complexed ligand	PDB code	Resolution/Å
Trypsin	Benzamidine inhibitor	3PTB	1.7
Rhinovirus	5-(7-(4-(4,5-dihydro-2-oxazolyl)phenoxy)heptyl)-3-methyl isoxazole	2R04	3
HIV protease	<i>N,N</i> -bis(2-hydroxy-1-indanyl)-2,6-diphenylmethyl- 4-hydroxy-1,7-heptandiamide	4PHV	2.1
Carboxypeptidase	alpha-hydroxy-beta-phenyl-propionic acid	2CTC	1.4
ER- α	4-hydroxy tamoxifene	3ERT	1.9

Table 1. Reference complexes of trypsin, rhinovirus, HIV protease, carboxypeptidase and ER- α targets used in docking and IFs virtual screening.

Trypsin	<i>Weight</i>	<i>b_IrotN</i>	<i>a_acc</i>	<i>a_don</i>
7 Actives	174.8 (113.6)	2.9 (2.5)	0.4 (1.1)	0.1 (0.4)
467 Inactives	238.7 (53.4)	1.9 (1.2)	0.2 (0.7)	0.1 (0.4)

Rhinovirus	<i>Weight</i>	<i>b_IrotN</i>	<i>a_acc</i>	<i>a_don</i>
8 Actives	351.1 (23.8)	9.4 (1.2)	3.0 (0.0)	0.0 (0.0)
498 Inactives	356.2 (20.4)	5.4 (1.0)	3.1 (0.6)	0.2 (0.5)

HIV protease	<i>Weight</i>	<i>b_IrotN</i>	<i>a_acc</i>	<i>a_don</i>
10 Actives	740.2 (88.9)	20.7 (5.8)	7.0 (2.4)	6.5 (2.0)
489 Inactives	609.2 (55.2)	8.4 (3.9)	4.9 (2.3)	1.4 (1.5)

Carboxypeptidase	<i>Weight</i>	<i>b_IrotN</i>	<i>a_acc</i>	<i>a_don</i>
5 Actives	333.3 (183.7)	7.6 (4.7)	1.8 (1.3)	1.4 (0.9)
272 Inactives	281.3 (43.0)	5.5 (0.8)	2.6 (0.5)	1.5 (0.7)

ER-α	<i>Weight</i>	<i>b_IrotN</i>	<i>a_acc</i>	<i>a_don</i>
10 Actives	458.8 (67.1)	11.3 (4.3)	3.6 (0.8)	1.6 (0.7)
490 Inactives	465.1 (16.2)	8.5 (3.1)	4.2 (1.8)	1.0 (0.9)

Table 2. Summary of the 1D physico-chemical properties of active and inactive molecules in the trypsin, rhinovirus, HIV protease, carboxypeptidase and ER- α screening databases. This table shows the average and standard deviation (in parenthesis) of the following properties: *Weight* (molecular weight), *b_IrotN* (number of rotatable single bonds), *a_acc* (number of hydrogen-bond acceptor atoms), *a_don* (number of hydrogen-bond donor atoms).

ESTROGEN RECEPTOR-ALPHA			
	2%	5%	10%
GOLDScore (<i>docking</i>)	15	10	7
CHIF-SCORE1-TANIMOTO	15	10	6
CHIF-SCORE1-SIMPLE_MATCHING	25	12	8
CHIF-SCORE1-EUCLIDEAN	15	10	6
CHIF-SCORE1-MANHATTAN	15	10	6
CHIF-SCORE2-TANIMOTO	25	12	7
CHIF-SCORE2-SIMPLE_MATCHING	30	14	8
APIF-SCORE1-TANIMOTO	10	4	4
APIF-SCORE1-SIMPLE_MATCHING	0	6	4
APIF-SCORE1-EUCLIDEAN	10	4	2
APIF-SCORE1-MANHATTAN	0	4	3
APIF-SCORE2-TANIMOTO	20	10	8
APIF-SCORE2-SIMPLE_MATCHING	20	10	7
NORMALIZED_APIF-SCORE1-TANIMOTO	0	2	2
NORMALIZED_APIF-SCORE1-SIMPLE_MATCHING	10	6	5
NORMALIZED_APIF-SCORE1-EUCLIDEAN	0	4	2
NORMALIZED_APIF-SCORE1-MANHATTAN	0	0	2
NORMALIZED_APIF-SCORE2-TANIMOTO	10	6	5
NORMALIZED_APIF-SCORE2-SIMPLE_MATCHING	20	14	8

Table 3. ER- α enrichment factor values for the first 2%, 5% and 10% screened database.

TRYPSIN			
	2%	5%	10%
GOLDScore (<i>docking</i>)	7	3	1
CHIF-SCORE1-TANIMOTO	0	0	0
CHIF-SCORE1-SIMPLE_MATCHING	0	6	3
CHIF-SCORE1-EUCLIDEAN	0	0	0
CHIF-SCORE1-MANHATTAN	0	0	0
CHIF-SCORE2-TANIMOTO	7	6	4
CHIF-SCORE2-SIMPLE_MATCHING	14	6	7
APIF-SCORE1-TANIMOTO	14	14	7
APIF-SCORE1-SIMPLE_MATCHING	14	9	7
APIF-SCORE1-EUCLIDEAN	14	6	3
APIF-SCORE1-MANHATTAN	14	6	4
APIF-SCORE2-TANIMOTO	29	14	7
APIF-SCORE2-SIMPLE_MATCHING	21	14	7
NORMALIZED_APIF-SCORE1-TANIMOTO	14	11	6
NORMALIZED_APIF-SCORE1-SIMPLE_MATCHING	14	11	6
NORMALIZED_APIF-SCORE1-EUCLIDEAN	14	6	3
NORMALIZED_APIF-SCORE1-MANHATTAN	7	3	1
NORMALIZED_APIF-SCORE2-TANIMOTO	29	14	7
NORMALIZED_APIF-SCORE2-SIMPLE_MATCHING	21	11	7

Table 4. Trypsin enrichment factor values for the first 2%, 5% and 10% screened database.

RHINOVIRUS

	2%	5%	10%
GOLDScore (docking)	6	8	5
CHIF-SCORE1-TANIMOTO	13	5	4
CHIF-SCORE1-SIMPLE_MATCHING	19	8	4
CHIF-SCORE1-EUCLIDEAN	13	5	5
CHIF-SCORE1-MANHATTAN	13	5	4
CHIF-SCORE2-TANIMOTO	13	8	4
CHIF-SCORE2-SIMPLE_MATCHING	13	5	4
APIF-SCORE1-TANIMOTO	6	8	5
APIF-SCORE1-SIMPLE_MATCHING	0	5	3
APIF-SCORE1-EUCLIDEAN	13	8	5
APIF-SCORE1-MANHATTAN	6	5	4
APIF-SCORE2-TANIMOTO	19	10	6
APIF-SCORE2-SIMPLE_MATCHING	6	8	4
NORMALIZED_APIF-SCORE1-TANIMOTO	13	8	6
NORMALIZED_APIF-SCORE1-SIMPLE_MATCHING	6	5	5
NORMALIZED_APIF-SCORE1-EUCLIDEAN	6	10	6
NORMALIZED_APIF-SCORE1-MANHATTAN	0	3	4
NORMALIZED_APIF-SCORE2-TANIMOTO	31	13	8
NORMALIZED_APIF-SCORE2-SIMPLE_MATCHING	13	10	9

Table 5. Rhinovirus enrichment factor values for the first 2%, 5% and 10% screened database.

HIV PROTEASE			
	2%	5%	10%
GOLDScore (<i>docking</i>)	15	8	4
CHIF-SCORE1-TANIMOTO	5	2	1
CHIF-SCORE1-SIMPLE_MATCHING	15	6	4
CHIF-SCORE1-EUCLIDEAN	0	2	1
CHIF-SCORE1-MANHATTAN	0	2	1
CHIF-SCORE2-TANIMOTO	5	2	2
CHIF-SCORE2-SIMPLE_MATCHING	10	4	3
APIF-SCORE1-TANIMOTO	15	10	7
APIF-SCORE1-SIMPLE_MATCHING	15	12	7
APIF-SCORE1-EUCLIDEAN	15	8	4
APIF-SCORE1-MANHATTAN	10	6	4
APIF-SCORE2-TANIMOTO	20	12	8
APIF-SCORE2-SIMPLE_MATCHING	20	10	8
NORMALIZED_APIF-SCORE1-TANIMOTO	10	6	5
NORMALIZED_APIF-SCORE1-SIMPLE_MATCHING	10	4	3
NORMALIZED_APIF-SCORE1-EUCLIDEAN	10	10	6
NORMALIZED_APIF-SCORE1-MANHATTAN	15	8	7
NORMALIZED_APIF-SCORE2-TANIMOTO	10	6	5
NORMALIZED_APIF-SCORE2-SIMPLE_MATCHING	15	6	3

Table 6. HIV protease enrichment factor values for the first 2%, 5% and 10% screened database.

CARBOXYPEPTIDASE			
	2%	5%	10%
GOLDScore (docking)	30	16	10
CHIF-SCORE1-TANIMOTO	10	8	6
CHIF-SCORE1-SIMPLE_MATCHING	20	12	10
CHIF-SCORE1-EUCLIDEAN	30	20	10
CHIF-SCORE1-MANHATTAN	30	20	10
CHIF-SCORE2-TANIMOTO	30	20	10
CHIF-SCORE2-SIMPLE_MATCHING	40	16	10
APIF-SCORE1-TANIMOTO	10	4	6
APIF-SCORE1-SIMPLE_MATCHING	20	12	10
APIF-SCORE1-EUCLIDEAN	10	4	2
APIF-SCORE1-MANHATTAN	10	4	2
APIF-SCORE2-TANIMOTO	30	16	8
APIF-SCORE2-SIMPLE_MATCHING	30	12	10
NORMALIZED_APIF-SCORE1-TANIMOTO	10	4	6
NORMALIZED_APIF-SCORE1-SIMPLE_MATCHING	30	12	10
NORMALIZED_APIF-SCORE1-EUCLIDEAN	10	4	2
NORMALIZED_APIF-SCORE1-MANHATTAN	10	4	2
NORMALIZED_APIF-SCORE2-TANIMOTO	30	12	10
NORMALIZED_APIF-SCORE2-SIMPLE_MATCHING	50	20	10

Table 7. Carboxypeptidase enrichment factor values for the first 2%, 5% and 10% screened database.

The following pages contain SEVENTEEN Figures for the article.

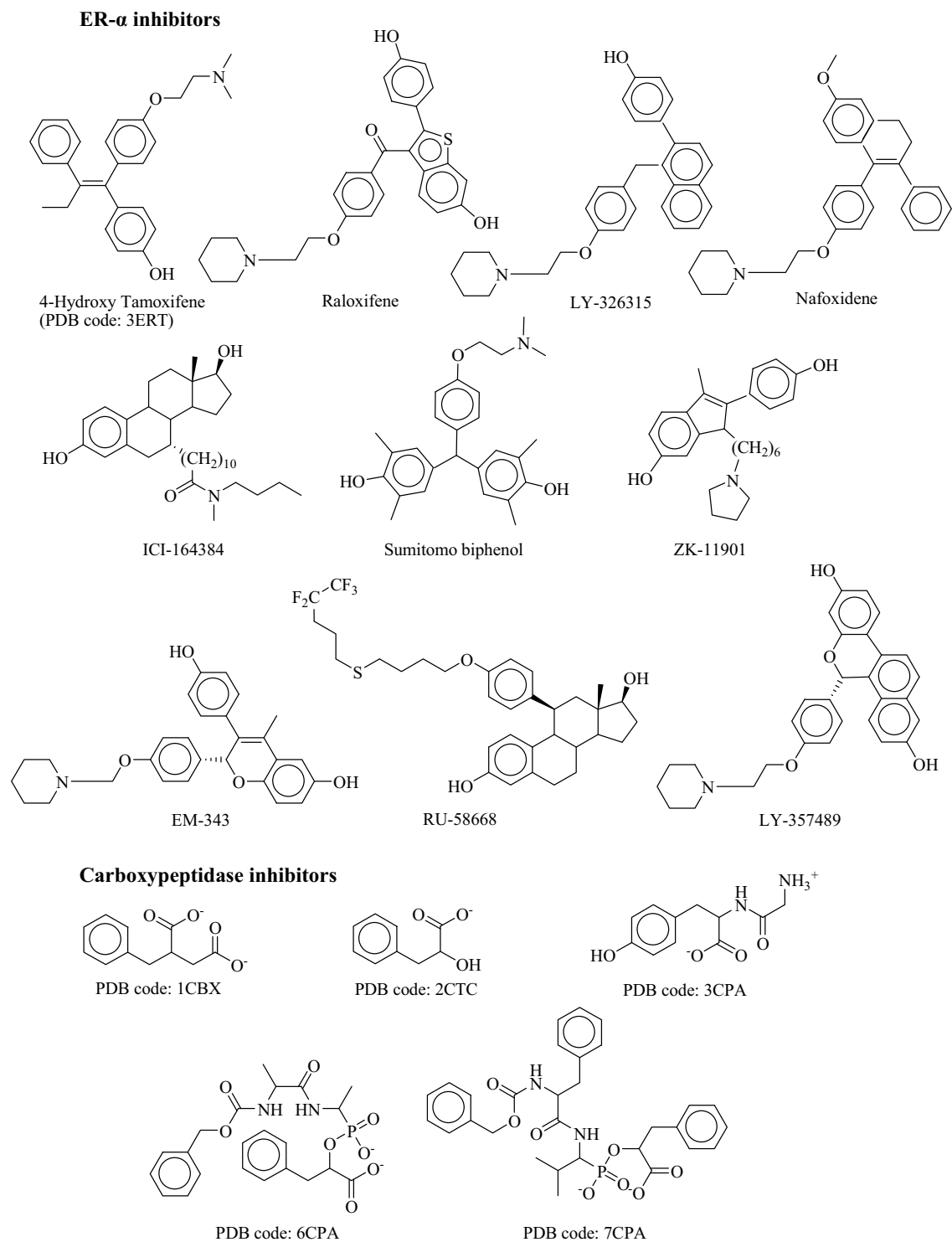
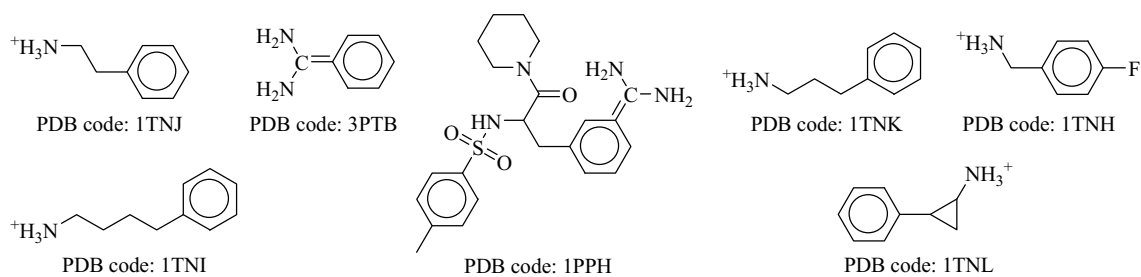


Figure 1. Known ER- α and carboxypeptidase active inhibitors in the virtual screening datasets.

Trypsin inhibitors



Rhinovirus inhibitors

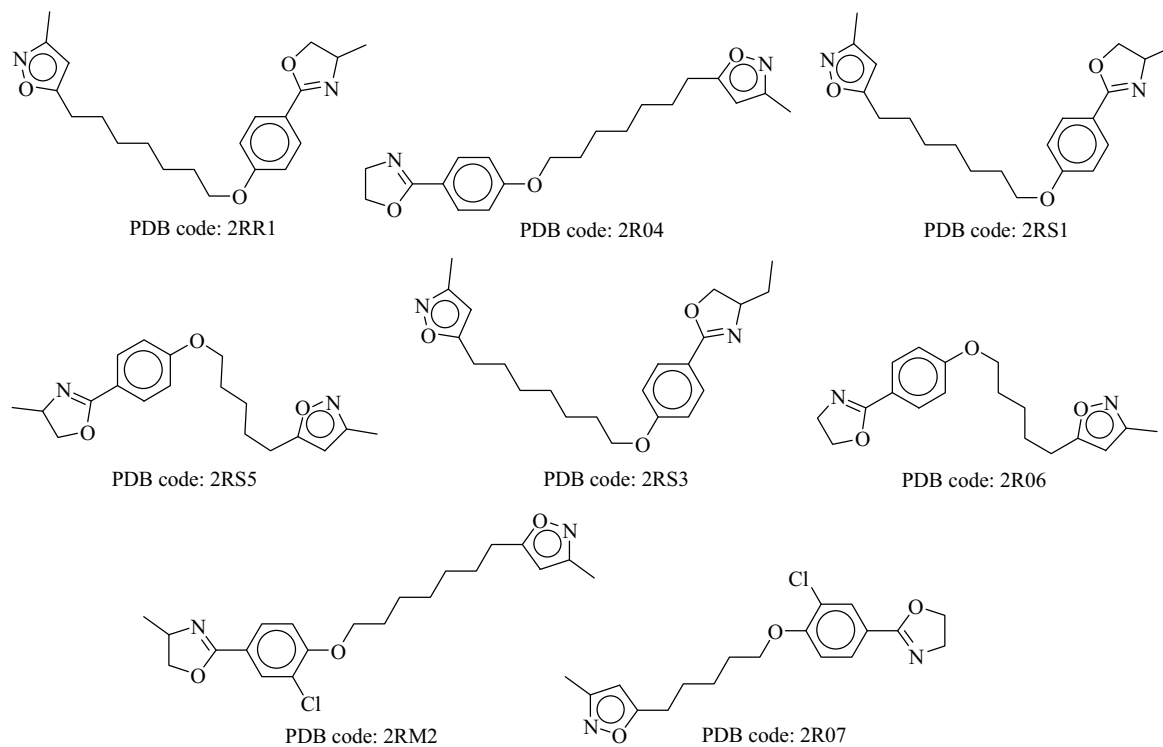


Figure 2. Known trypsin and rhinovirus active inhibitors in the virtual screening datasets.

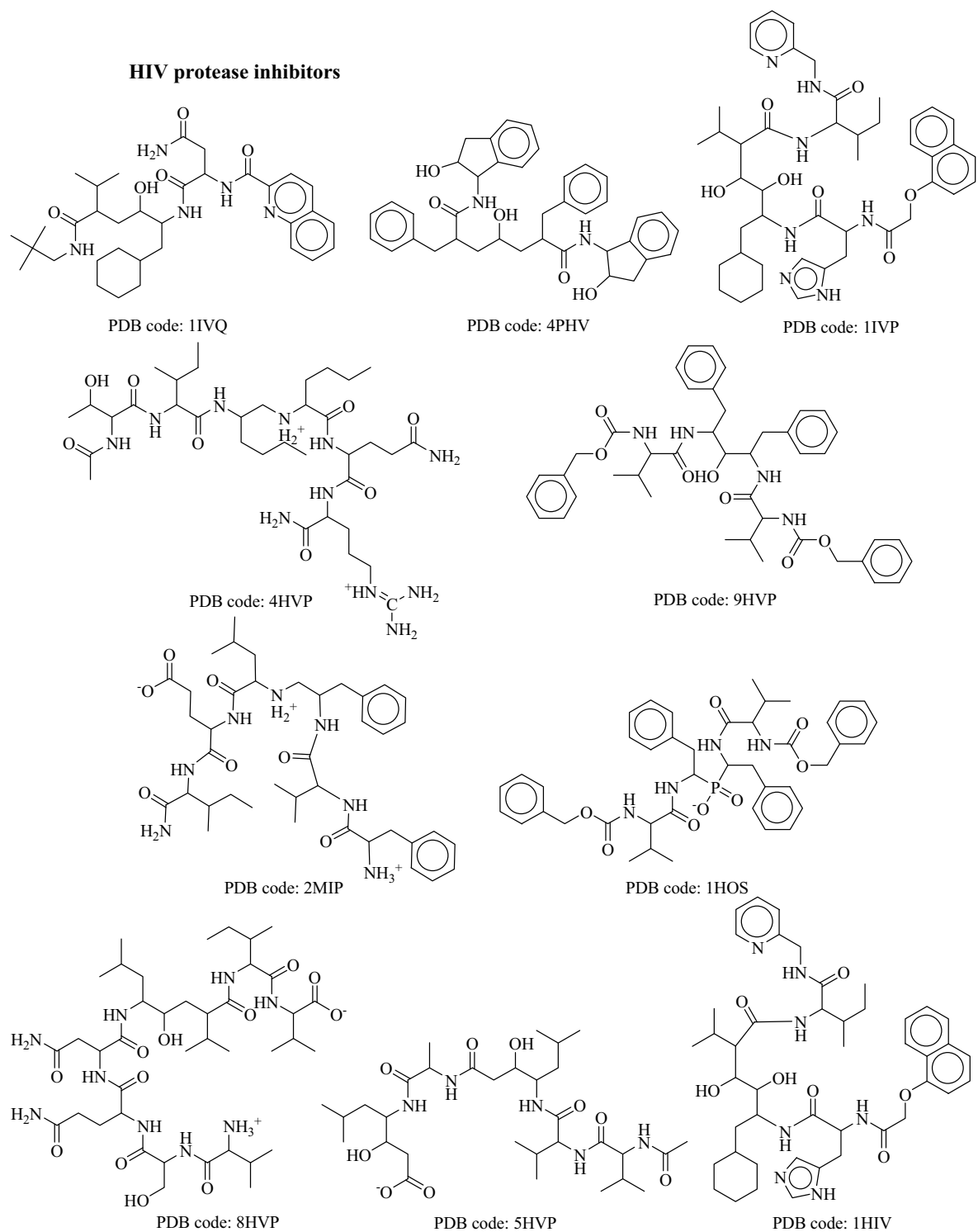


Figure 3. Known HIV protease active inhibitors in the virtual screening dataset.

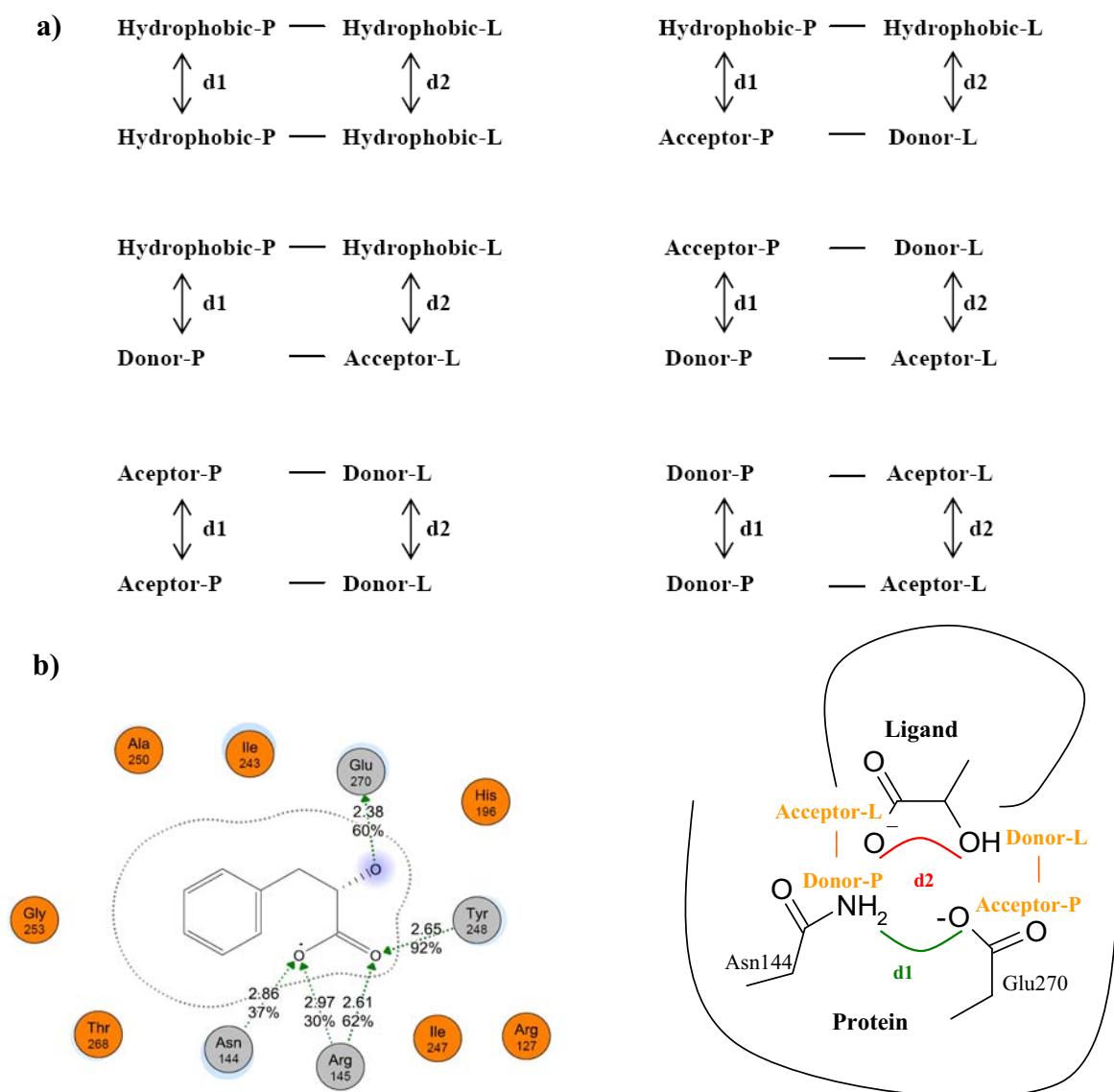


Figure 4. Atom Pairs based Interaction Fingerprint (APIF). (a) Six possible combinations of pairs of interactions defining a set of 49 bits (7 bits for a total of 7 distances). (b) Codification of the pairwise interactions from the distances measured between the two protein (d1) and the two ligand (d2) interacting atoms. This specific case shows the carboxypeptidase in complex with an alpha-hydroxy-beta-phenyl-propionic acid. For a pair of interactions detected, for example a hydrogen bond between ND2 of Asn144 and a negatively charged oxygen of the alpha-hydroxy-beta-phenyl-propionic acid, and a hydrogen bond between OE2 of Glu270 and an oxygen of the alpha-hydroxy-beta-phenyl-propionic acid, the interacting distances between the two protein atoms (d1) and the two ligand atoms (d2) are measured.

Ranks of distances (d1 and d2 in Å):

- 1) [0-2.5]
- 2) [2.5-4]
- 3) [4-6]
- 4) [6-9]
- 5) [9-13]
- 6) [13-18]
- 7) [>18]

Hydrophobic-P ----- Hydrophobic-L Hydrophobic-P ----- Hydrophobic-L (d1_d2) 49 bits	Hydrophobic-P ----- Hydrophobic-L Acceptor-P ----- Donor-L (d1_d2) 49 bits	}																																																																																																		
<table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <tr><td>1 1</td><td>1 2</td><td>1 3</td><td>1 4</td><td>1 5</td><td>1 6</td><td>1 7</td></tr> <tr><td>2 1</td><td>2 2</td><td>2 3</td><td>2 4</td><td>2 5</td><td>2 6</td><td>2 7</td></tr> <tr><td>3 1</td><td>3 2</td><td>3 3</td><td>3 4</td><td>3 5</td><td>3 6</td><td>3 7</td></tr> <tr><td>4 1</td><td>4 2</td><td>4 3</td><td>4 4</td><td>4 5</td><td>4 6</td><td>4 7</td></tr> <tr><td>5 1</td><td>5 2</td><td>5 3</td><td>5 4</td><td>5 5</td><td>5 6</td><td>5 7</td></tr> <tr><td>6 1</td><td>6 2</td><td>6 3</td><td>6 4</td><td>6 5</td><td>6 6</td><td>6 7</td></tr> <tr><td>7 1</td><td>7 2</td><td>7 3</td><td>7 4</td><td>7 5</td><td>7 6</td><td>7 7</td></tr> </table>	1 1		1 2	1 3	1 4	1 5	1 6	1 7	2 1	2 2	2 3	2 4	2 5	2 6	2 7	3 1	3 2	3 3	3 4	3 5	3 6	3 7	4 1	4 2	4 3	4 4	4 5	4 6	4 7	5 1	5 2	5 3	5 4	5 5	5 6	5 7	6 1	6 2	6 3	6 4	6 5	6 6	6 7	7 1	7 2	7 3	7 4	7 5	7 6	7 7	<table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <tr><td>1 1</td><td>1 2</td><td>1 3</td><td>1 4</td><td>1 5</td><td>1 6</td><td>1 7</td></tr> <tr><td>2 1</td><td>2 2</td><td>2 3</td><td>2 4</td><td>2 5</td><td>2 6</td><td>2 7</td></tr> <tr><td>3 1</td><td>3 2</td><td>3 3</td><td>3 4</td><td>3 5</td><td>3 6</td><td>3 7</td></tr> <tr><td>4 1</td><td>4 2</td><td>4 3</td><td>4 4</td><td>4 5</td><td>4 6</td><td>4 7</td></tr> <tr><td>5 1</td><td>5 2</td><td>5 3</td><td>5 4</td><td>5 5</td><td>5 6</td><td>5 7</td></tr> <tr><td>6 1</td><td>6 2</td><td>6 3</td><td>6 4</td><td>6 5</td><td>6 6</td><td>6 7</td></tr> <tr><td>7 1</td><td>7 2</td><td>7 3</td><td>7 4</td><td>7 5</td><td>7 6</td><td>7 7</td></tr> </table>	1 1	1 2	1 3	1 4	1 5	1 6	1 7	2 1	2 2	2 3	2 4	2 5	2 6	2 7	3 1	3 2	3 3	3 4	3 5	3 6	3 7	4 1	4 2	4 3	4 4	4 5	4 6	4 7	5 1	5 2	5 3	5 4	5 5	5 6	5 7	6 1	6 2	6 3	6 4	6 5	6 6	6 7	7 1	7 2	7 3	7 4	7 5	7 6	7 7
1 1	1 2		1 3	1 4	1 5	1 6	1 7																																																																																													
2 1	2 2		2 3	2 4	2 5	2 6	2 7																																																																																													
3 1	3 2		3 3	3 4	3 5	3 6	3 7																																																																																													
4 1	4 2		4 3	4 4	4 5	4 6	4 7																																																																																													
5 1	5 2	5 3	5 4	5 5	5 6	5 7																																																																																														
6 1	6 2	6 3	6 4	6 5	6 6	6 7																																																																																														
7 1	7 2	7 3	7 4	7 5	7 6	7 7																																																																																														
1 1	1 2	1 3	1 4	1 5	1 6	1 7																																																																																														
2 1	2 2	2 3	2 4	2 5	2 6	2 7																																																																																														
3 1	3 2	3 3	3 4	3 5	3 6	3 7																																																																																														
4 1	4 2	4 3	4 4	4 5	4 6	4 7																																																																																														
5 1	5 2	5 3	5 4	5 5	5 6	5 7																																																																																														
6 1	6 2	6 3	6 4	6 5	6 6	6 7																																																																																														
7 1	7 2	7 3	7 4	7 5	7 6	7 7																																																																																														
Hydrophobic-P ----- Hydrophobic-L Donor-P ----- Acceptor-L (d1_d2) 49 bits	Donor-P ----- Acceptor-L Donor-P ----- Acceptor-L (d1_d2) 49 bits	}																																																																																																		
<table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <tr><td>1 1</td><td>1 2</td><td>1 3</td><td>1 4</td><td>1 5</td><td>1 6</td><td>1 7</td></tr> <tr><td>2 1</td><td>2 2</td><td>2 3</td><td>2 4</td><td>2 5</td><td>2 6</td><td>2 7</td></tr> <tr><td>3 1</td><td>3 2</td><td>3 3</td><td>3 4</td><td>3 5</td><td>3 6</td><td>3 7</td></tr> <tr><td>4 1</td><td>4 2</td><td>4 3</td><td>4 4</td><td>4 5</td><td>4 6</td><td>4 7</td></tr> <tr><td>5 1</td><td>5 2</td><td>5 3</td><td>5 4</td><td>5 5</td><td>5 6</td><td>5 7</td></tr> <tr><td>6 1</td><td>6 2</td><td>6 3</td><td>6 4</td><td>6 5</td><td>6 6</td><td>6 7</td></tr> <tr><td>7 1</td><td>7 2</td><td>7 3</td><td>7 4</td><td>7 5</td><td>7 6</td><td>7 7</td></tr> </table>	1 1		1 2	1 3	1 4	1 5	1 6	1 7	2 1	2 2	2 3	2 4	2 5	2 6	2 7	3 1	3 2	3 3	3 4	3 5	3 6	3 7	4 1	4 2	4 3	4 4	4 5	4 6	4 7	5 1	5 2	5 3	5 4	5 5	5 6	5 7	6 1	6 2	6 3	6 4	6 5	6 6	6 7	7 1	7 2	7 3	7 4	7 5	7 6	7 7	<table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <tr><td>1 1</td><td>1 2</td><td>1 3</td><td>1 4</td><td>1 5</td><td>1 6</td><td>1 7</td></tr> <tr><td>2 1</td><td>2 2</td><td>2 3</td><td>2 4</td><td>2 5</td><td>2 6</td><td>2 7</td></tr> <tr><td>3 1</td><td>3 2</td><td>3 3</td><td>3 4</td><td>3 5</td><td>3 6</td><td>3 7</td></tr> <tr><td>4 1</td><td>4 2</td><td>4 3</td><td>4 4</td><td>4 5</td><td>4 6</td><td>4 7</td></tr> <tr><td>5 1</td><td>5 2</td><td>5 3</td><td>5 4</td><td>5 5</td><td>5 6</td><td>5 7</td></tr> <tr><td>6 1</td><td>6 2</td><td>6 3</td><td>6 4</td><td>6 5</td><td>6 6</td><td>6 7</td></tr> <tr><td>7 1</td><td>7 2</td><td>7 3</td><td>7 4</td><td>7 5</td><td>7 6</td><td>7 7</td></tr> </table>	1 1	1 2	1 3	1 4	1 5	1 6	1 7	2 1	2 2	2 3	2 4	2 5	2 6	2 7	3 1	3 2	3 3	3 4	3 5	3 6	3 7	4 1	4 2	4 3	4 4	4 5	4 6	4 7	5 1	5 2	5 3	5 4	5 5	5 6	5 7	6 1	6 2	6 3	6 4	6 5	6 6	6 7	7 1	7 2	7 3	7 4	7 5	7 6	7 7
1 1	1 2		1 3	1 4	1 5	1 6	1 7																																																																																													
2 1	2 2		2 3	2 4	2 5	2 6	2 7																																																																																													
3 1	3 2		3 3	3 4	3 5	3 6	3 7																																																																																													
4 1	4 2		4 3	4 4	4 5	4 6	4 7																																																																																													
5 1	5 2	5 3	5 4	5 5	5 6	5 7																																																																																														
6 1	6 2	6 3	6 4	6 5	6 6	6 7																																																																																														
7 1	7 2	7 3	7 4	7 5	7 6	7 7																																																																																														
1 1	1 2	1 3	1 4	1 5	1 6	1 7																																																																																														
2 1	2 2	2 3	2 4	2 5	2 6	2 7																																																																																														
3 1	3 2	3 3	3 4	3 5	3 6	3 7																																																																																														
4 1	4 2	4 3	4 4	4 5	4 6	4 7																																																																																														
5 1	5 2	5 3	5 4	5 5	5 6	5 7																																																																																														
6 1	6 2	6 3	6 4	6 5	6 6	6 7																																																																																														
7 1	7 2	7 3	7 4	7 5	7 6	7 7																																																																																														
Acceptor-P ----- Donor-L Acceptor-P ----- Donor-L (d1_d2) 49 bits	Acceptor-P ----- Donor-L Donor-P ----- Acceptor-L (d1_d2) 49 bits	}																																																																																																		
<table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <tr><td>1 1</td><td>1 2</td><td>1 3</td><td>1 4</td><td>1 5</td><td>1 6</td><td>1 7</td></tr> <tr><td>2 1</td><td>2 2</td><td>2 3</td><td>2 4</td><td>2 5</td><td>2 6</td><td>2 7</td></tr> <tr><td>3 1</td><td>3 2</td><td>3 3</td><td>3 4</td><td>3 5</td><td>3 6</td><td>3 7</td></tr> <tr><td>4 1</td><td>4 2</td><td>4 3</td><td>4 4</td><td>4 5</td><td>4 6</td><td>4 7</td></tr> <tr><td>5 1</td><td>5 2</td><td>5 3</td><td>5 4</td><td>5 5</td><td>5 6</td><td>5 7</td></tr> <tr><td>6 1</td><td>6 2</td><td>6 3</td><td>6 4</td><td>6 5</td><td>6 6</td><td>6 7</td></tr> <tr><td>7 1</td><td>7 2</td><td>7 3</td><td>7 4</td><td>7 5</td><td>7 6</td><td>7 7</td></tr> </table>	1 1		1 2	1 3	1 4	1 5	1 6	1 7	2 1	2 2	2 3	2 4	2 5	2 6	2 7	3 1	3 2	3 3	3 4	3 5	3 6	3 7	4 1	4 2	4 3	4 4	4 5	4 6	4 7	5 1	5 2	5 3	5 4	5 5	5 6	5 7	6 1	6 2	6 3	6 4	6 5	6 6	6 7	7 1	7 2	7 3	7 4	7 5	7 6	7 7	<table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <tr><td>1 1</td><td>1 2</td><td>1 3</td><td>1 4</td><td>1 5</td><td>1 6</td><td>1 7</td></tr> <tr><td>2 1</td><td>2 2</td><td>2 3</td><td>2 4</td><td>2 5</td><td>2 6</td><td>2 7</td></tr> <tr><td>3 1</td><td>3 2</td><td>3 3</td><td>3 4</td><td>3 5</td><td>3 6</td><td>3 7</td></tr> <tr><td>4 1</td><td>4 2</td><td>4 3</td><td>4 4</td><td>4 5</td><td>4 6</td><td>4 7</td></tr> <tr><td>5 1</td><td>5 2</td><td>5 3</td><td>5 4</td><td>5 5</td><td>5 6</td><td>5 7</td></tr> <tr><td>6 1</td><td>6 2</td><td>6 3</td><td>6 4</td><td>6 5</td><td>6 6</td><td>6 7</td></tr> <tr><td>7 1</td><td>7 2</td><td>7 3</td><td>7 4</td><td>7 5</td><td>7 6</td><td>7 7</td></tr> </table>	1 1	1 2	1 3	1 4	1 5	1 6	1 7	2 1	2 2	2 3	2 4	2 5	2 6	2 7	3 1	3 2	3 3	3 4	3 5	3 6	3 7	4 1	4 2	4 3	4 4	4 5	4 6	4 7	5 1	5 2	5 3	5 4	5 5	5 6	5 7	6 1	6 2	6 3	6 4	6 5	6 6	6 7	7 1	7 2	7 3	7 4	7 5	7 6	7 7
1 1	1 2		1 3	1 4	1 5	1 6	1 7																																																																																													
2 1	2 2		2 3	2 4	2 5	2 6	2 7																																																																																													
3 1	3 2		3 3	3 4	3 5	3 6	3 7																																																																																													
4 1	4 2		4 3	4 4	4 5	4 6	4 7																																																																																													
5 1	5 2	5 3	5 4	5 5	5 6	5 7																																																																																														
6 1	6 2	6 3	6 4	6 5	6 6	6 7																																																																																														
7 1	7 2	7 3	7 4	7 5	7 6	7 7																																																																																														
1 1	1 2	1 3	1 4	1 5	1 6	1 7																																																																																														
2 1	2 2	2 3	2 4	2 5	2 6	2 7																																																																																														
3 1	3 2	3 3	3 4	3 5	3 6	3 7																																																																																														
4 1	4 2	4 3	4 4	4 5	4 6	4 7																																																																																														
5 1	5 2	5 3	5 4	5 5	5 6	5 7																																																																																														
6 1	6 2	6 3	6 4	6 5	6 6	6 7																																																																																														
7 1	7 2	7 3	7 4	7 5	7 6	7 7																																																																																														

Figure 5. Atom Pairs based Interaction Fingerprint encoding system. The six possible combinations of pairs of interaction contacts are each represented with 49 bits. For each pairwise interactions detected in a complex, the distance between the two receptor atoms (d1) and the two ligand atoms (d2) is measured. The results are clustered into seven distance ranges (Å): [0-2.5], [2.5-4], [4-6],[6-9], [9-13], [13-18] and [>18]. Thus, the total fingerprint length is always 6x7x7=294 bits.

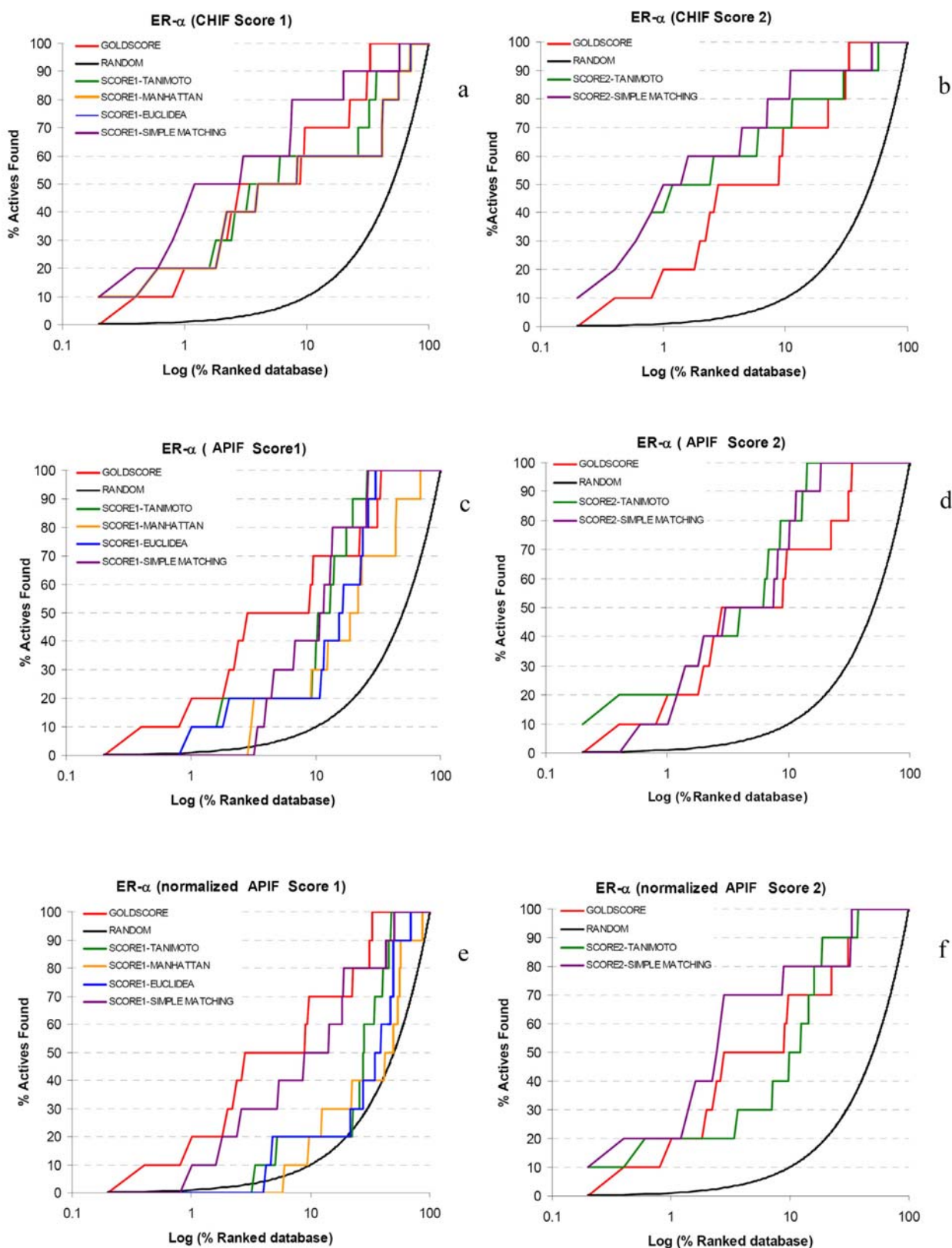


Figure 6. ER- α enrichment plots obtained using (a) CHIF in-house implementation and SCORE1, (b) CHIF in-house implementation and SCORE2, (c) APIF and SCORE1, (d) APIF and SCORE2, (e) Normalized APIF and SCORE1, (f) Normalized APIF and SCORE2. The different similarity scores used correspond to Simple matching (purple), Euclidean distance (blue), Tanimoto coefficient (green) and Manhattan distance (yellow). The enrichment plot obtained using GoldScore is shown in red and a random screening in black. The x axis is the logarithm of the percent of the database screened, plotted against the percent recovery of known active compounds on the y axis.

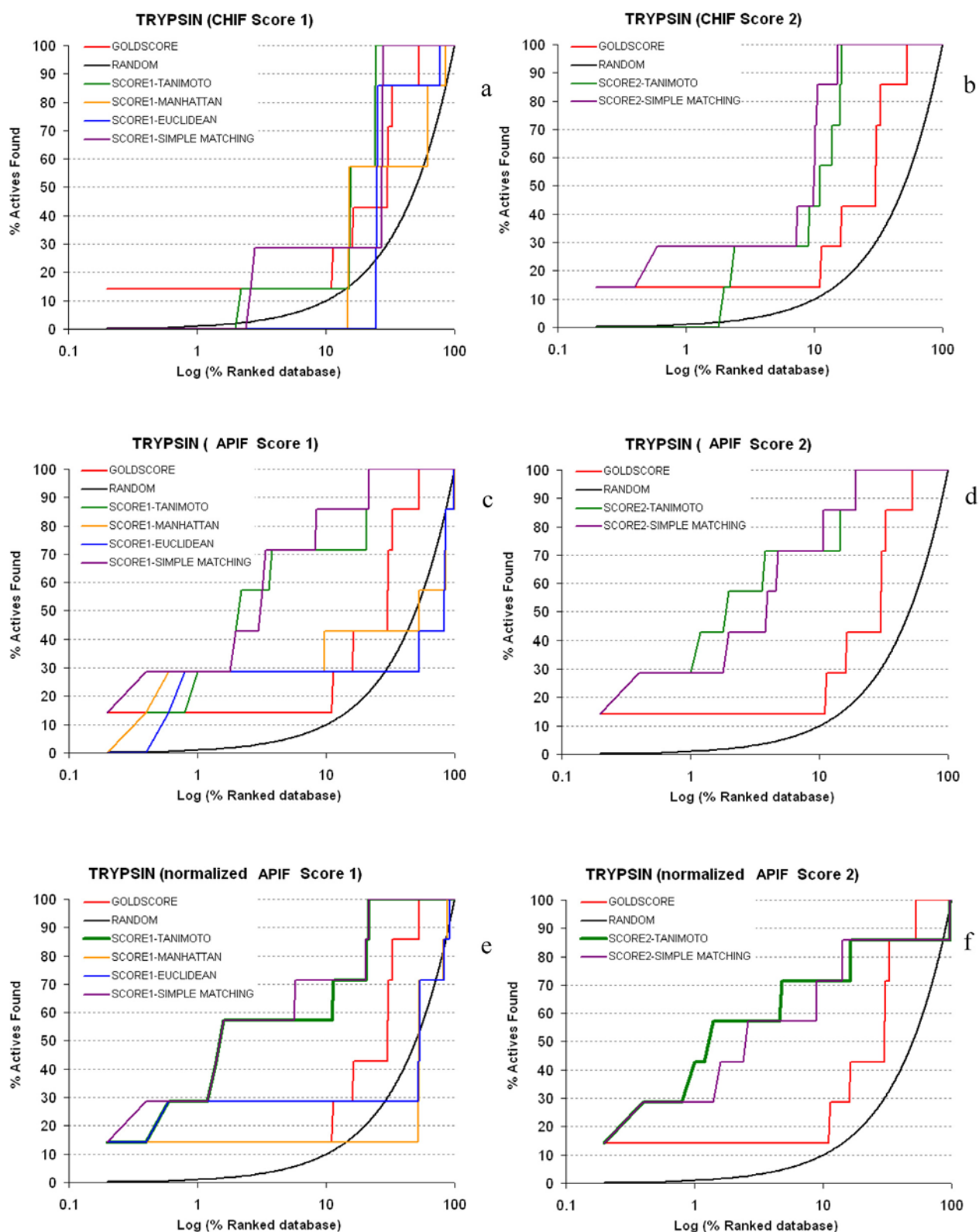


Figure 7. Trypsin enrichment plots obtained using (a) CHIF in-house implementation and SCORE1, (b) CHIF in-house implementation and SCORE2, (c) APIF and SCORE1, (d) APIF and SCORE2, (e) Normalized APIF and SCORE1, (f) Normalized APIF and SCORE2. The different similarity scores used correspond to Simple matching (purple), Euclidean distance (blue), Tanimoto coefficient (green) and Manhattan distance (yellow). The enrichment plot obtained using GoldScore is shown in red and a random screening in black. The x axis is the logarithm of the percent of the database screened, plotted against the percent recovery of known active compounds on the y axis.

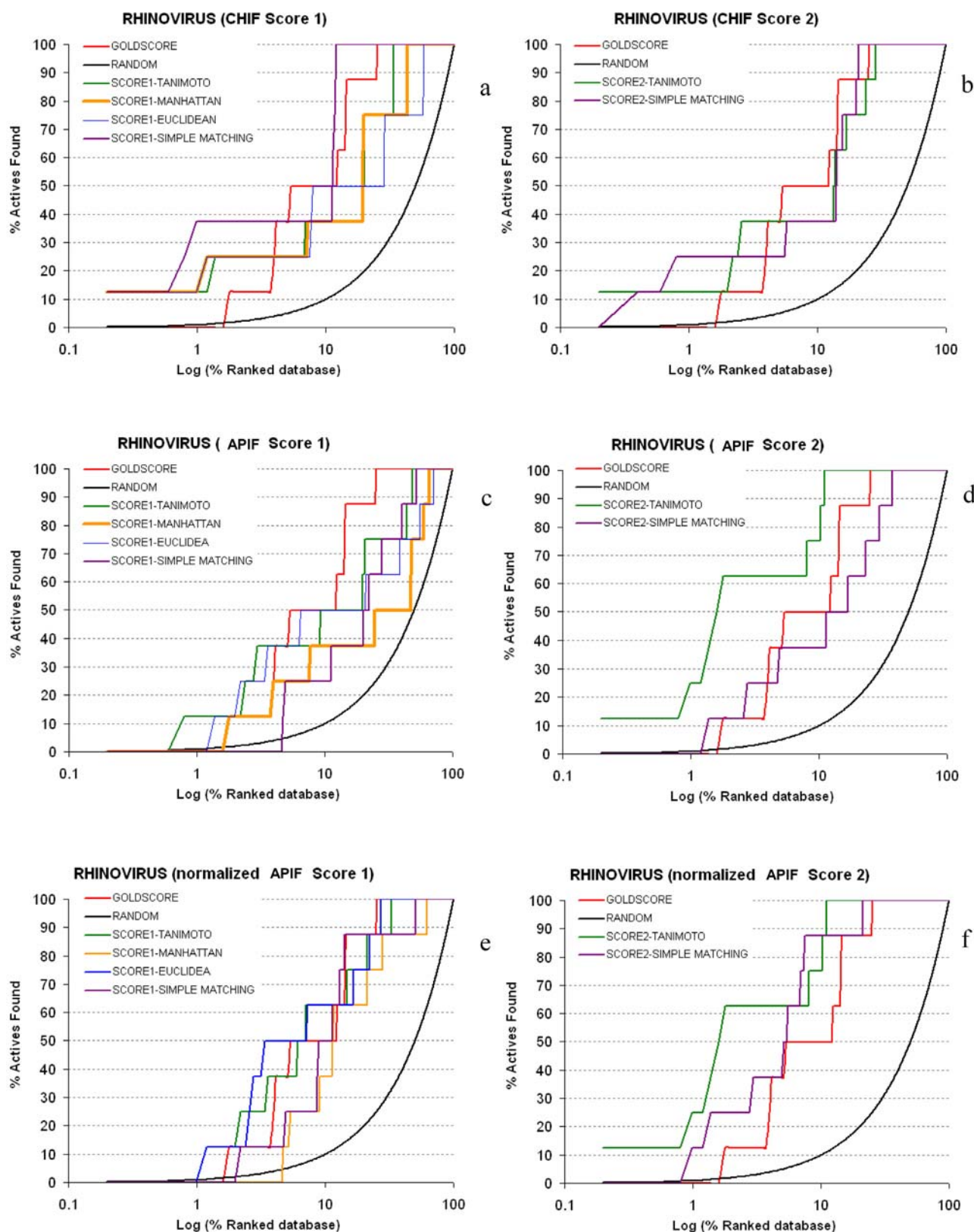


Figure 8. Rhinovirus enrichment plots obtained using (a) CHIF in-house implementation and SCORE1, (b) CHIF in-house implementation and SCORE2, (c) APIF and SCORE1, (d) APIF and SCORE2, (e) Normalized APIF and SCORE1, (f) Normalized APIF and SCORE2. The different similarity scores used correspond to Simple matching (purple), Euclidean distance (blue), Tanimoto coefficient (green) and Manhattan distance (yellow). The enrichment plot obtained using GoldScore is shown in red and a random screening in black. The x axis is the logarithm of the percent of the database screened, plotted against the percent recovery of known active compounds on the y axis.

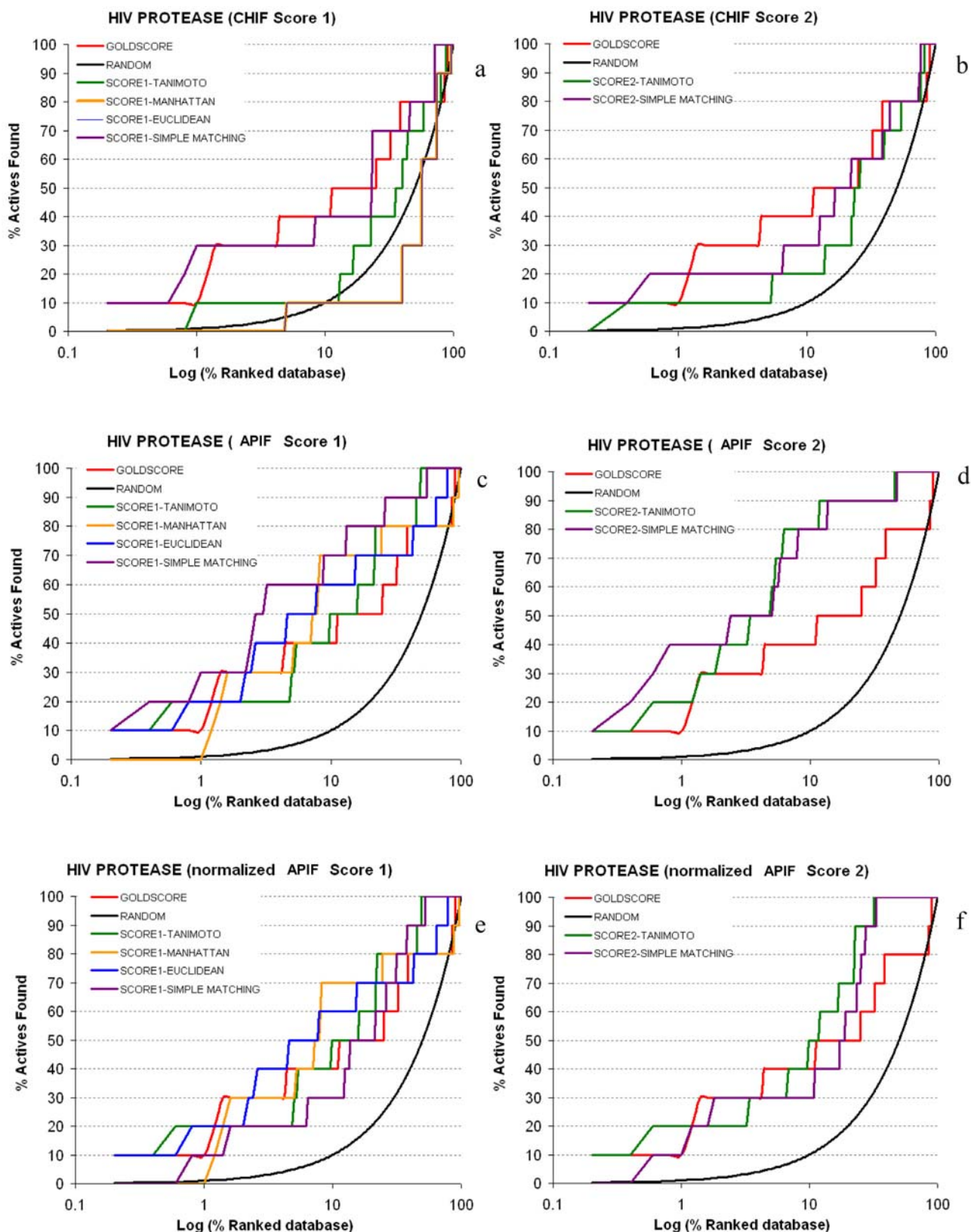


Figure 9. HIV protease plots obtained using (a) CHIF in-house implementation and SCORE1, (b) CHIF in-house implementation and SCORE2, (c) APIF and SCORE1, (d) APIF and SCORE2, (e) Normalized APIF and SCORE1, (f) Normalized APIF and SCORE2. The different similarity scores used correspond to Simple matching (purple), Euclidean distance (blue), Tanimoto coefficient (green) and Manhattan distance (yellow). The enrichment plot obtained using GoldScore is shown in red and a random screening in black. The x axis is the logarithm of the percent of the database screened, plotted against the percent recovery of known active compounds on the y axis.

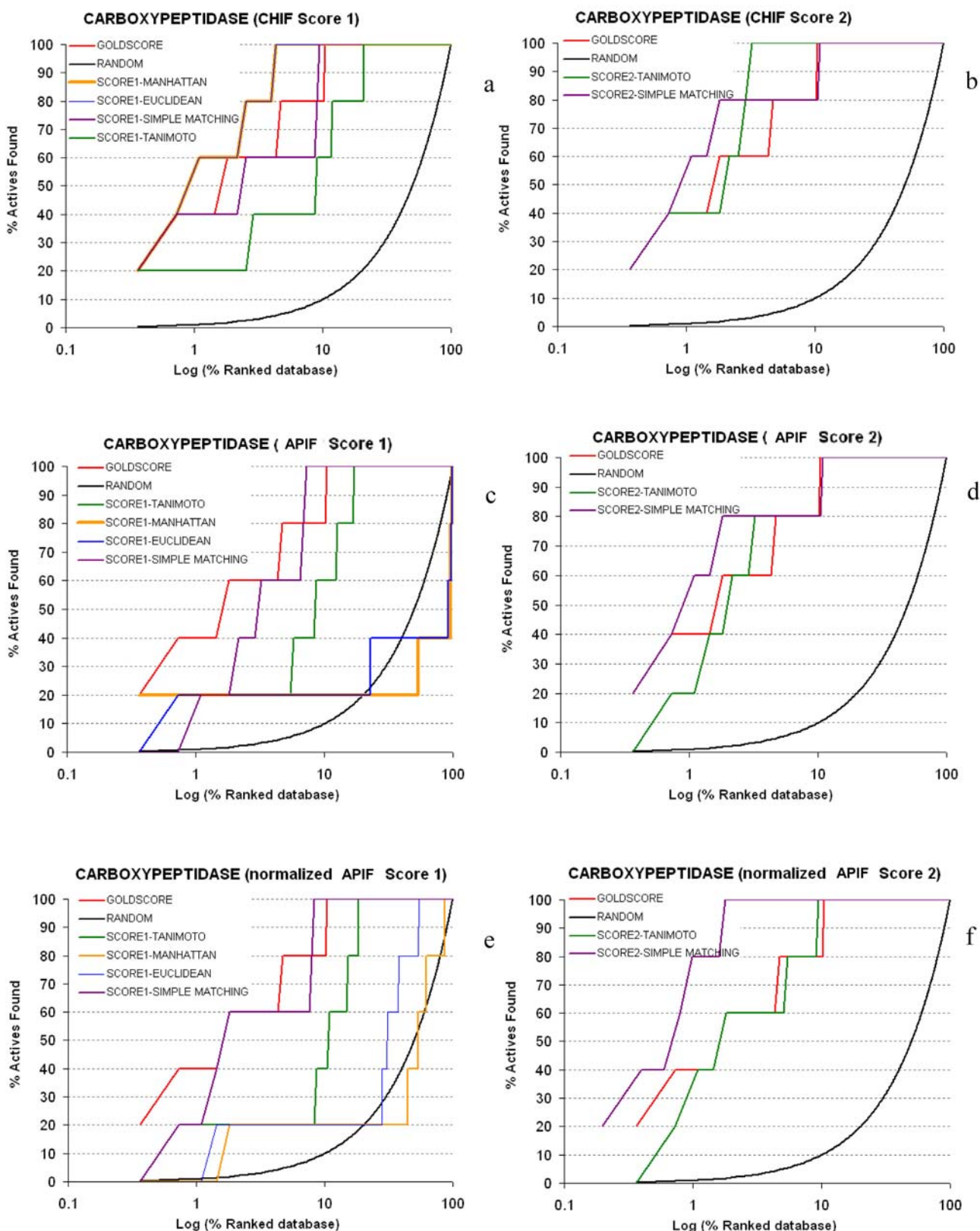


Figure 10. Carboxypeptidase plots obtained using (a) CHIF in-house implementation and SCORE1, (b) CHIF in-house implementation and SCORE2, (c) APIF and SCORE1, (d) APIF and SCORE2, (e) Normalized APIF and SCORE1, (f) Normalized APIF and SCORE2. The different similarity scores used correspond to Simple matching (purple), Euclidean distance (blue), Tanimoto coefficient (green) and Manhattan distance (yellow). The enrichment plot obtained using GoldScore is shown in red and a random screening in black. The x axis is the logarithm of the percent of the database screened, plotted against the percent recovery of known active compounds on the y axis.

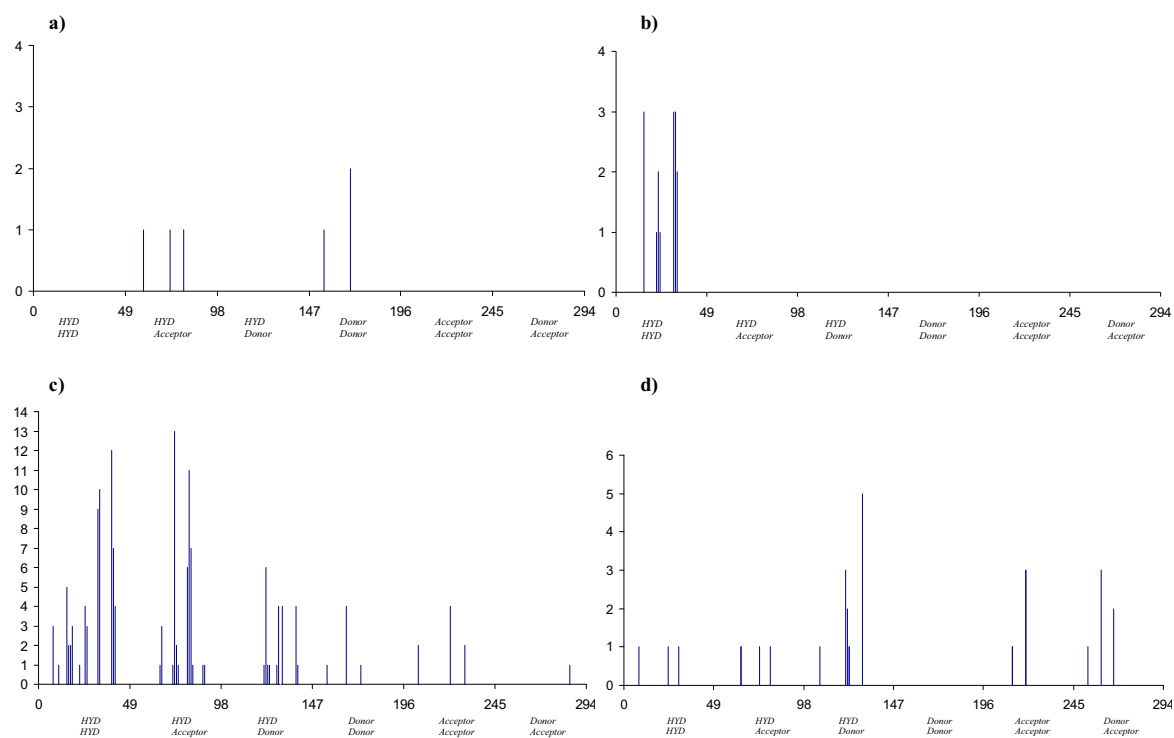


Figure 11. Correlation diagram of the APIF fingerprint for (a) trypsin, (b) rhinovirus, (c) HIV protease and (d) carboxypeptidase complexes. The total fingerprint length is 294 bits, divided in six atom pair contacts: *HYD HYD* referred as hydrophobic hydrophobic protein/ligand interactions, *HYD Acceptor* referred as hydrophobic acceptor protein/ligand interactions, *HYD Donor* referred as hydrophobic donor protein/ligand interactions, *Donor Donor* referred as donor donor protein/ligand interactions, *Acceptor Acceptor* referred as acceptor acceptor protein/ligand interactions, and *Donor Acceptor* referred as donor acceptor protein/ligand interactions.

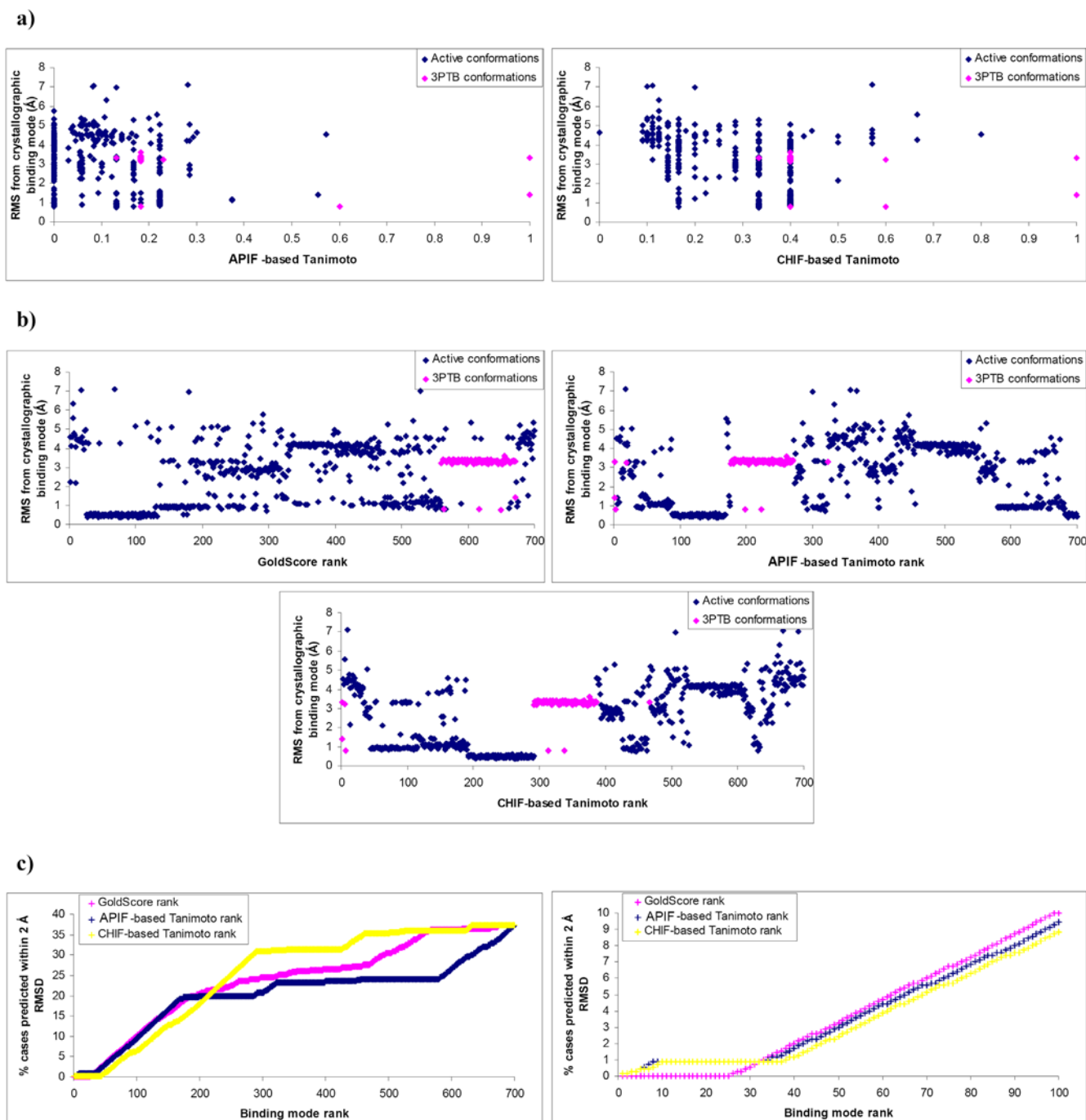
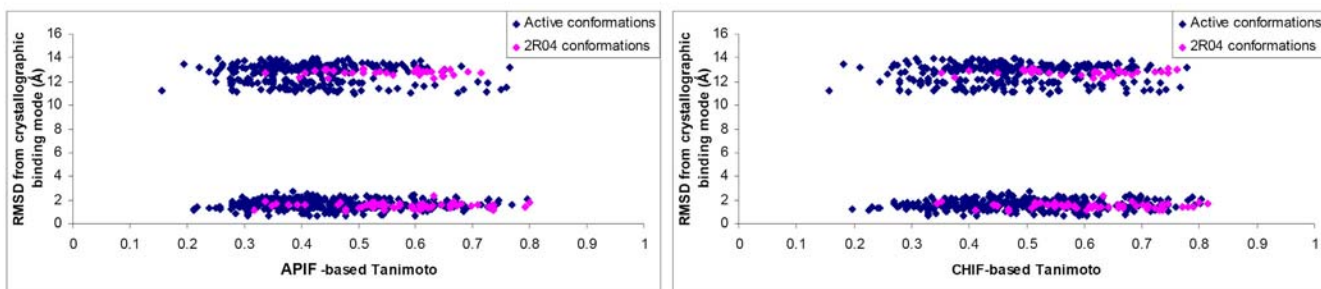
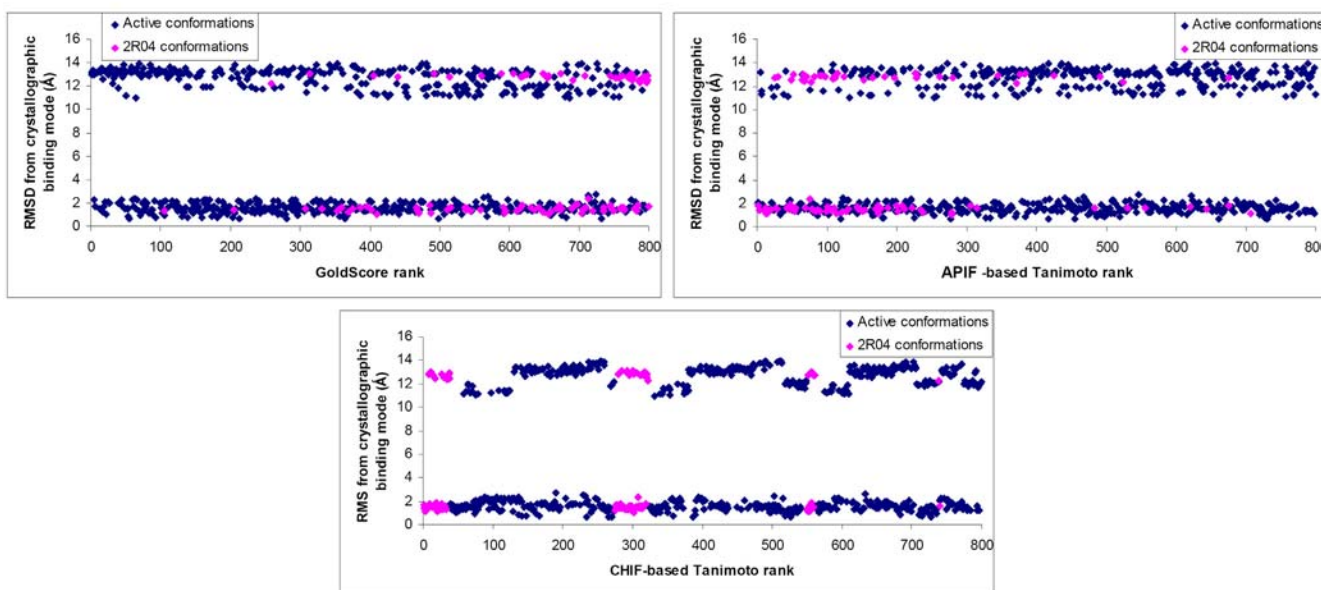


Figure 12. Trypsin binding mode analyses. a) RMSD from crystallographic binding mode (\AA) versus APIF-based Tanimoto (left) and CHIF-based Tanimoto (right) for a set of 7 active ligands. b) RMSD from crystallographic binding mode (\AA) vs GoldScore rank (left), APIF-based Tanimoto rank (right) and CHIF-based Tanimoto rank (centre). c) Fraction of cases (%) predicted within 2 \AA RMSD vs binding mode rank for GoldScore (pink curve), APIF (blue curve), and CHIF (yellow curve) for the 700 active docked conformations obtained (left) and the top 100 ranked solutions (right).

a)



b)



c)

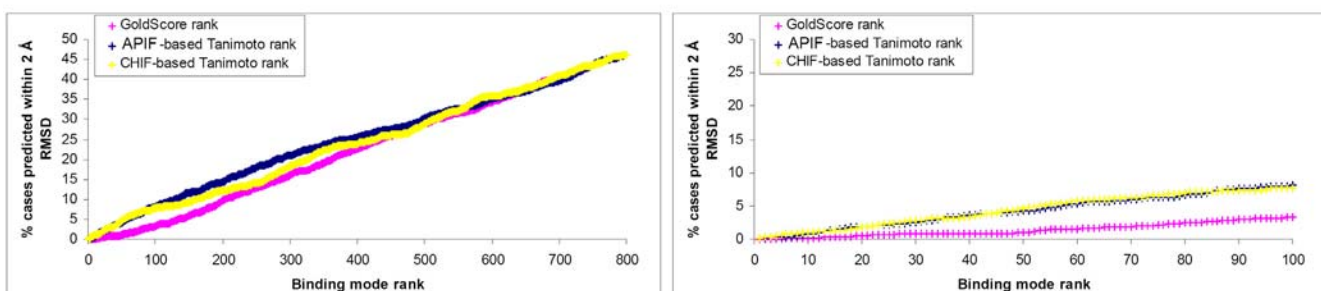


Figure 13. Rhinovirus binding mode analyses. a) RMSD from crystallographic binding mode (\AA) versus APIF-based Tanimoto (left) and CHIF-based Tanimoto (right) for a set of 8 active ligands. b) RMSD from crystallographic binding mode (\AA) vs GoldScore rank (left), APIF-based Tanimoto rank (right) and CHIF-based Tanimoto rank (centre). c) Fraction of cases (%) predicted within 2 \AA RMSD vs binding mode rank for GoldScore (pink curve), APIF (blue curve), and CHIF (yellow curve) for the 800 active docked conformations obtained (left) and the top 100 ranked solutions (right).

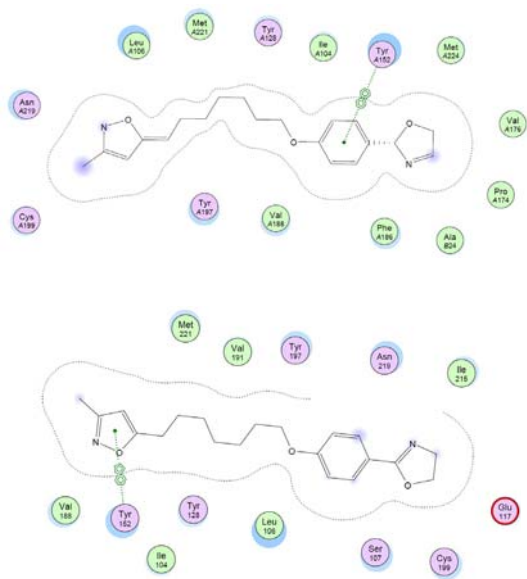


Figure 14. Rhinovirus binding modes. Two binding modes found for rhinovirus conformations interacting with the same protein atoms.

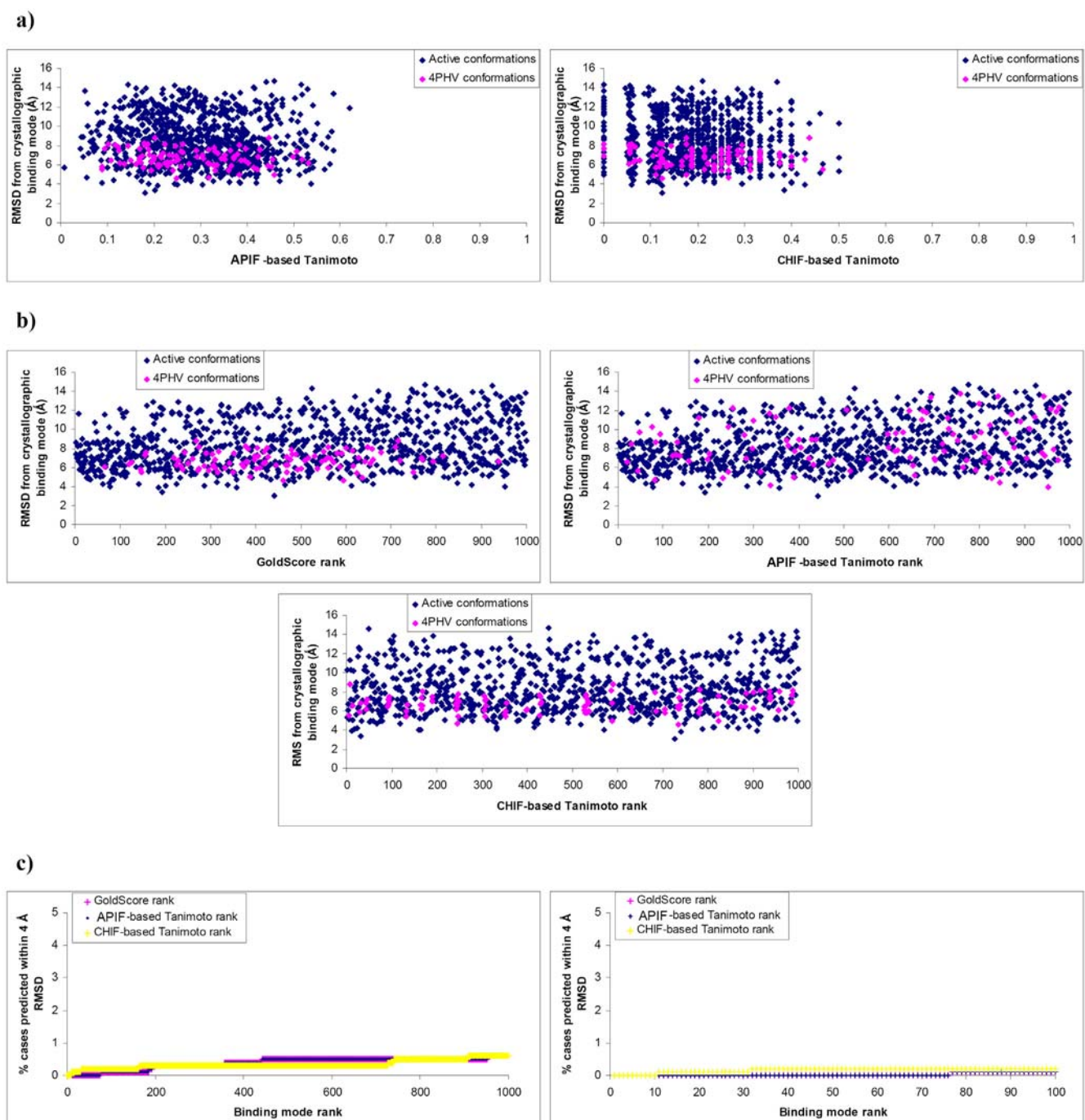


Figure 15. HIV protease binding mode analyses. a) RMSD from crystallographic binding mode (\AA) versus APIF-based Tanimoto (left) and CHIF-based Tanimoto (right) for a set of 10 active ligands. b) RMSD from crystallographic binding mode (\AA) vs GoldScore rank (left), APIF-based Tanimoto rank (right) and CHIF-based Tanimoto rank (centre). c) Fraction of cases (%) predicted within 2 \AA RMSD vs binding mode rank for GoldScore (pink curve), APIF (blue curve), and CHIF (yellow curve) for the 1000 active docked conformations obtained (left) and the top 100 ranked solutions (right).

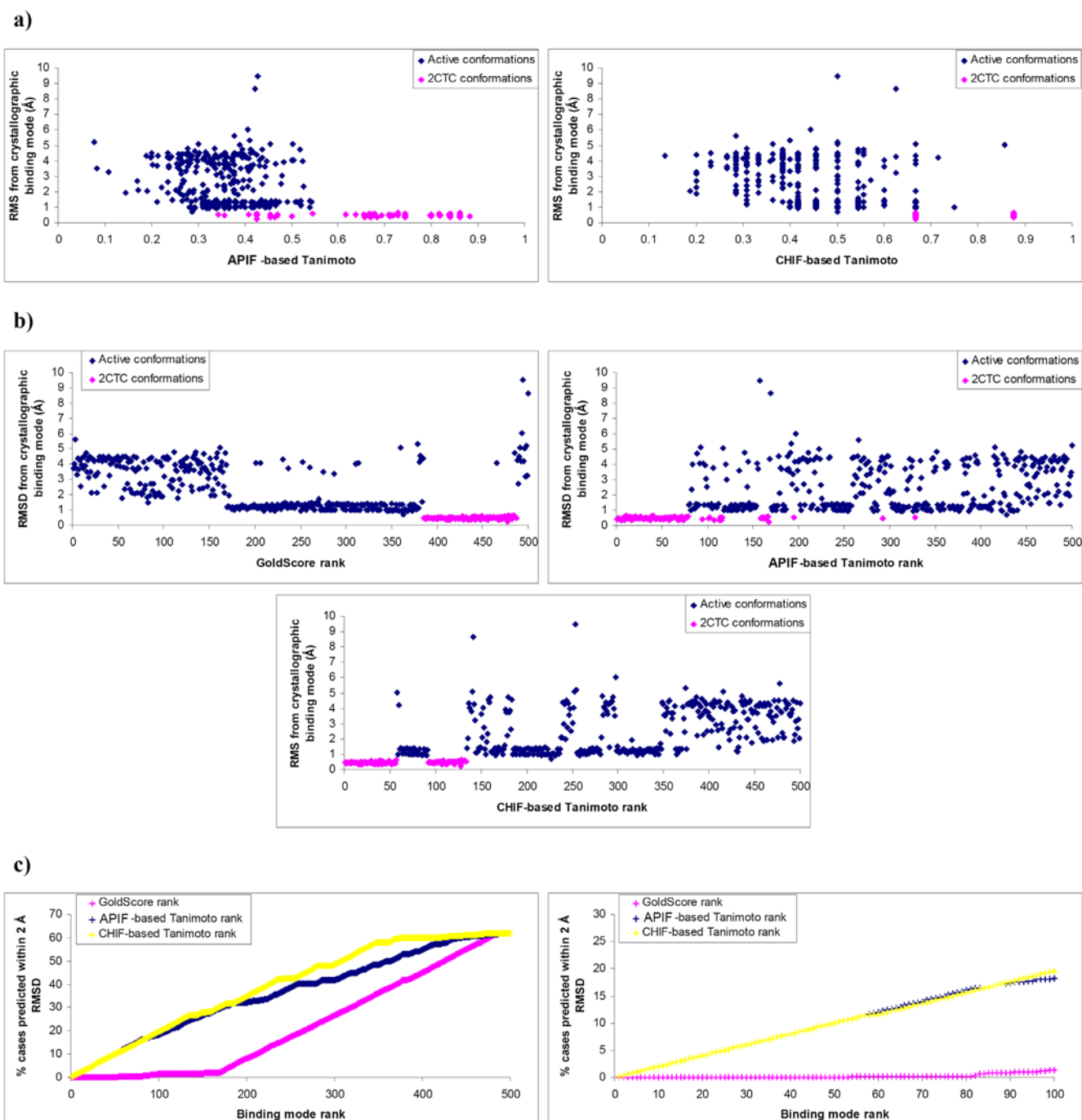


Figure 16. Carboxypeptidase binding mode analyses. a) RMSD from crystallographic binding mode (\AA) versus APIF-based Tanimoto (left) and CHIF-based Tanimoto (right) for a set of 5 active ligands. b) RMSD from crystallographic binding mode (\AA) vs GoldScore rank (left), APIF-based Tanimoto rank (right) and CHIF-based Tanimoto rank (centre). c) Fraction of cases (%) predicted within 2 \AA RMSD vs binding mode rank for GoldScore (pink curve), APIF (blue curve), and CHIF (yellow curve) for the 500 active docked conformations obtained (left) and the top 100 ranked solutions (right).

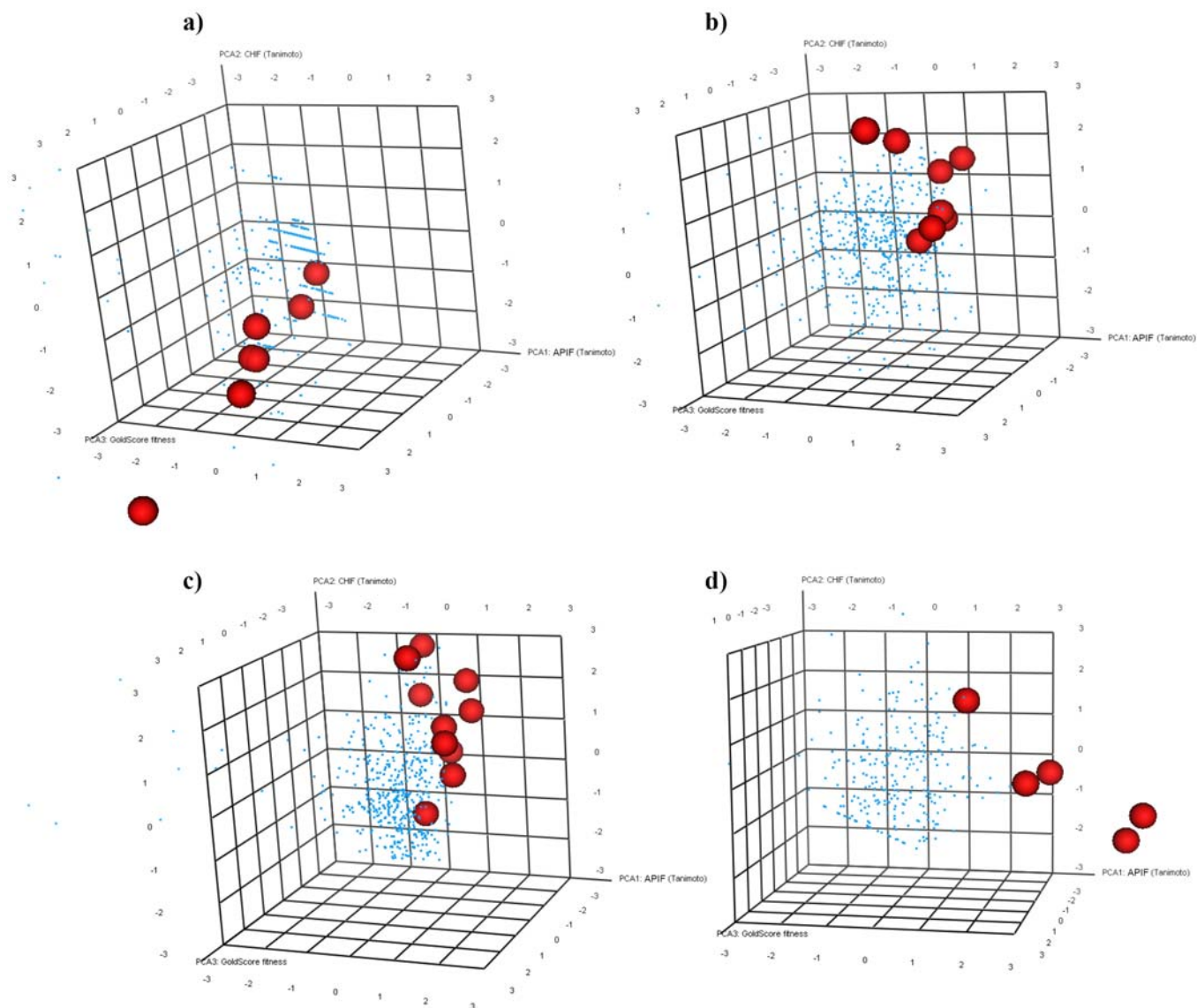


Figure 17. PCA analyses of trypsin, rhinovirus, HIV protease and carboxypeptidase compounds databases, showing the separation of active compounds' IFs into specific eigenvector spaces. a) Trypsin compounds database PCA analysis. b) Rhinovirus compounds database PCA analysis. c) HIV protease compounds database PCA analysis. d) Carboxypeptidase compounds database PCA analysis. In all cases PCA axes correspond to APIF-based Tanimoto score, CHF-based Tanimoto score and GoldScore function. Active compounds are shown in red ball and stick and inactive compounds in blue stick representation.

For Table of Contents Use Only

APIF: A New Interaction Fingerprint Based on Atom Pairs and its Application to Virtual Screening

Violeta I. Pérez-Nueno, Obdulia Rabal, José I. Borrell and Jordi Teixidó*

